ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη Αλγορίθμων Μάθησης για Βελτίωση της Εκπαίδευσης και της Ερμηνείας των Βαθιών Νευρωνικών Δικτύων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Γεωργίου Ιωάννου

Αθήνα, Δεκέμβριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη Αλγορίθμων Μάθησης για Βελτίωση της Εκπαίδευσης και της Ερμηνείας των Βαθιών Νευρωνικών Δικτύων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Γεωργίου Ιωάννου

Συμβουλευτική Επιτροπή:

Ανδρέας - Γεώργιος Σταφυλοπάτης
Διονύσιος - Δημήτριος Κουτσούρης
Γεώργιος Στάμου

Εγκρίθηκε από την επταμελή επιτροπή την 11η Δεκεμβρίου 2023

.....	.....	.....
Ανδρέας - Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π	Διονύσιος - Δημήτριος Κουτσούρης Καθηγητής Ε.Μ.Π	Γεώργιος Στάμου Καθηγητής Ε.Μ.Π
.....	.....	.....
Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π	Κωνσταντίνα Νικήτα Καθηγήτρια Ε.Μ.Π	Αθανάσιος Βουλόδημος Επ. Καθηγητής Ε.Μ.Π

.....
Γεώργιος Αλεξανδρίδης
Επ. Καθηγητής Ε.Κ.Π.Α.

Αθήνα, Δεκέμβριος 2023

.....

Γεώργιος Ιωάννου
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα τελευταία χρόνια το πεδίο της Μηχανικής Μάθησης έχει αναπτυχθεί σε μεγάλο βαθμό. Με την εξέλιξη και την αξιοποίηση σύγχρονων υπολογιστικών συστημάτων και καινούργιων τεχνολογιών ο τομέας της Μηχανικής Μάθησης κατάφερε να παρέχει λύσεις σε προβλήματα διαφόρων επιστημονικών πεδίων, καθώς και να έχει σημαντικό ρόλο στον τομέα της παραγωγής και της εργασίας. Κυριότερα, η ανάπτυξη της Βαθιάς Μηχανικής Μάθησης και των Νευρωνικών Δικτύων ευθύνεται για μέρος αυτής της επιτυχίας. Σε αυτή τη διατριβή ασχοληθήκαμε, κυρίως, με τα Βαθιά Νευρωνικά Δίκτυα και την λειτουργία τους. Αναλύσαμε διάφορους αλγορίθμους μάθησης και εντοπίσαμε προβλήματα που δυσχεραίνουν την καλή επίδοση ενός δικτύου. Μέσω της διατριβής προτείνονται αλγόριθμοι και μέθοδοι μάθησης νευρωνικών δικτύων, οι οποίοι αποσκοπούν στην καλύτερη εκπαίδευση και, κατά συνέπεια, στην βελτίωση των αποδόσεων των Βαθιών Νευρωνικών Δικτύων.

Πιο συγκεκριμένα, στο πρώτο μέρος εξετάσαμε την τεχνική εκπαίδευσης με παρτίδες ενός νευρωνικού δικτύου. Εντρυφήσαμε στο πεδίο της Δυναμικής Επιλογής Παρτίδας και προτείναμε έναν αλγόριθμο που βασίζεται στην Μεροληπτική Δειγματοληψία. Σκοπός του είναι να επιλέγει δείγματα από το σύνολο δεδομένων που εμφανίζουν υψηλές τιμές σφάλματος και να τις εισάγει περισσότερες φορές στην διαδικασία της εκπαίδευσης. Δίνοντας έμφαση στα δύσκολα δείγματα το νευρωνικό δίκτυο καταφέρνει να εκπαιδευτεί γρηγορότερα και να έχει καλύτερες επιδόσεις. Για να αποδειχθεί η χρησιμότητα της μεθόδου, διεξήχθησαν μία σειρά από πειράματα σε διαφορετικά σύνολα δεδομένων. Τα αποτελέσματα δείχνουν ότι ο προτεινόμενος αλγόριθμος βελτιώνει την ταχύτητα σύγκλισης και πολλές φορές την μέγιστη επίδοση του δικτύου. Εκτός αυτού βελτιώνει τον χρόνο εκπαίδευσης και τον αριθμό των υπολογισμών ανά επανάληψη σε σχέση με άλλες τεχνικές της βιβλιογραφίας.

Στο δεύτερο μέρος της διατριβής ασχοληθήκαμε με το πεδίο της Ανισορροπίας δεδομένων. Αυτό το φαινόμενο συναντάται συχνά στα πραγματικά σύνολα δεδομένων και αποτελεί ένα σημαντικό εμπόδιο στην ομαλή εκπαίδευση και γενίκευση των μοντέλων μηχανικής μάθησης. Περιγράψαμε και αναλύσαμε διάφορες μεθόδους και τεχνικές της βιβλιογραφίας πάνω σε αυτό το θέμα. Η μελέτη μας επικεντρώθηκε στις τεχνικές προσαρμογής του αλγορίθμου μάθησης με σκοπό την καταπολέμηση της ανισορροπίας. Προτείναμε την μέθοδο εκπαίδευσης νευρωνικών δικτύων με όνομα Θορυβώδης Επιλογή Παρτίδας με Επανεισαγωγές, η οποία επιλέγει δείγματα από τα δεδομένα με βάση κάποια κριτήρια και προσθέτει κατάλληλο θόρυβο. Με αυτόν τον τρόπο μπορεί το δίκτυο να εκπαιδεύεται εξίσου καλά σε κλάσεις δεδομένων με μι-

κρό αριθμό δειγμάτων επιτυγχάνοντας υψηλότερες επιδόσεις. Μία σειρά από πειράματα σε ανισόρροπα σύνολα δεδομένων έδειξαν την βελτίωση που παρέχει η μέθοδος αυτή σε σχέση με άλλες. Επίσης, δείχνουμε ότι είναι ικανή να λειτουργήσει σε συνδυασμό με άλλες τεχνικές καταπολέμησης ανισορροπίας, όπως τεχνικές μετασχηματισμού δεδομένων.

Μία άλλη θεματική που μελετήθηκε σε αυτή τη διατριβή είναι η ερευνητική περιοχή της βελτιστοποίησης. Στο πλαίσιο της εκπαίδευσης νευρωνικών δικτύων έχουν δημιουργηθεί μία πληθώρα από βελτιστοποιητές, καθένας από τους οποίους έχει τις ιδιαιτερότητές του. Εμβαθύναμε περισσότερο σε προσαρμοστικούς αλγόριθμους και προτείναμε μία μέθοδο βελτιστοποίησης, με όνομα AdaLip, η οποία κατασκευάζει διαφορετικό ρυθμό μάθησης ανά επίπεδο βασισμένη στην σταθερά του Lipschitz. Στοιχεία παρατέθηκαν για την ανάγκη διαφορετικής προσέγγισης των διαφορετικών επιπέδων και υποστηρίχθηκαν πειραματικά. Δοκιμάσαμε την μέθοδο μας σε ένα σύνολο προβλημάτων ταξινόμησης εικόνας και τα αποτελέσματα έδειξαν βελτιώσεις στην ταχύτητα σύγκλισης, στην συνολική επίδοση στο σύνολο εκπαίδευσης αλλά και πιο σταθερή γενίκευση. Η μέθοδος αυτή μπορεί να δουλέψει πάνω από ήδη υπάρχοντες βελτιστοποιητές και να καλυτερέψει τα αποτελέσματά τους. Τέλος, παρατέθηκε θεωρητική απόδειξη σύγκλισης του προτεινόμενου βελτιστοποιητή.

Στο τελευταίο κομμάτι της διατριβής ασχοληθήκαμε με το πεδίο της ερμηνείας των νευρωνικών δικτύων. Η ερμηνευσιμότητα πραγματεύεται με την κατανόηση των νευρωνικών δικτύων και των προβλέψεών τους. Αρχικά, εξερευνούμε διάφορες τεχνικές ερμηνευσιμότητας και συγκρίνουμε τις επιδόσεις τους. Τα πειράματα βασίστηκαν πάνω σε ιατρικές εικόνες για ταξινόμηση του σταδίου της αμφιβληστροειδοπάθειας. Αυτό συνέβαλε στην βαθύτερη κατανόηση της λειτουργίας των μοντέλων αλλά και στην εξήγηση των περιοχών βλάβης των ιατρικών εικόνων. Επίσης, με την χρήση τέτοιων μεθόδων δείξαμε ότι είναι εφικτό να προσεγγιστεί και μία λύση στο πρόβλημα της κατάτμησης εικόνας. Εκτός από αυτό εμβαθύναμε περισσότερο στην λειτουργία των μεθόδων ερμηνευσιμότητας και συγκεκριμένα στις μεθόδους που χρησιμοποιούν σημεία αναφοράς. Δείξαμε ότι η χρήση σημείων αναφοράς εγκυμονεί πολλούς κινδύνους ανακρίβειας των σημασιών των προβλέψεων νευρωνικών δικτύων. Με βάση αυτή την αδυναμία τους προτείναμε ένα νέο επίπεδο που αποσκοπεί στο να βελτιώσει αυτά τα ζητήματα. Το προτεινόμενο Επίπεδο Διανυσματικής Αναπαράστασης με Αντίληψη Βάσης κατασκευάστηκε με σκοπό να ενσωματώνει μέσω της εκπαίδευσης την έννοια της βάσης ή σημείου αναφοράς. Έτσι, οι εκάστοτε αλγόριθμοι ερμηνευσιμότητας που λειτουργούν με σημεία αναφοράς μπορούν να χρησιμοποιούν το παραπάνω επίπεδο στις αρχιτεκτονικές του δικτύου και να δημιουργούν πιο ακριβείς ερμηνείες για τις διάφορες προβλέψεις. Αυτό το δείξαμε πειραματικά πάνω σε 4 σύνολα δεδομένων πινάκων. Τα σύνολα πινάκων επιλέχθηκαν λόγω της μεγάλης ποικιλίας χαρακτηριστικών που διαθέτουν αλλά και επειδή σε αυτά παρατηρείται πιο συχνά το πρόβλημα των σημείων αναφοράς.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Βαθιά Μηχανική Μάθηση, Νευρωνικά Δίκτυα, Συνελικτικά Νευρω-

νικά Δίκτυα, Βελτιστοποίηση, Αλγόριθμοι Μάθησης, Στοχαστικοί Αλγόριθμοι, Δειγματοληψία, Ερμηνευσιμότητα, Επεξηγησιμότητα, Σημεία Αναφοράς, Διανυσματική Αναπαράσταση.

Abstract

In recent years the field of Machine Learning has been developed dramatically. With the progress and use of advanced hardware and computer systems, Machine Learning has given solutions in many scientific problems and is a vital part of some industries. Especially, Deep Learning and Deep Neural Networks are responsible for this great success. In this thesis we got involved, mainly, with Deep Neural Networks and their training process. We analyzed different learning algorithms and pinpointed problems that worsen the performance of neural networks. This dissertation proposes various algorithms and learning methods that intend to improve the training process and the general performance of Deep Neural Networks.

Specifically, the first part revolves around the method of training a network with batches. We focused on the techniques of Online Batch Selection and proposed an algorithm that is based on Biased Sampling. The goal of the algorithm is to select samples with high loss values and add them in the training process more frequently. Emphasizing on the difficult samples the network is trained faster and has a better performance. To prove the usefulness of the proposed method, a series of experiments was inducted on different datasets. The results show that the algorithm improves the convergence speed and the best performance scores of the model. Apart from that, it improves the training time and the number of computations per iteration in comparison to other works in the literature.

In the second part of the dissertation, we delved into the field of Imbalanced Datasets. This phenomenon is encountered often regarding real-world datasets and is a serious obstacle of the training process and the generalization of machine learning models. We described and analyzed various methods and techniques that are popular in the literature. Our work was centered around algorithm-based methods that tackle the problem of imbalance. We proposed a method of training neural networks, called NBSBS-R, that selects samples based on some criteria and adds a proper amount of noise. This way the network can learn the minority class just as well, while achieving better performance. An experimental framework is introduced that uses imbalanced datasets to test the new algorithm. The results showed an improvement in the generalization performance of the networks compared to other methods. Also, the experiments showed that the method is able to work together with other data-transformation techniques in order to build a better

model overall.

Another subject that was studied, was the field of optimization. There is a wide variety of optimizers that can be used to train a neural network, while each one of them has its own intricacies. We dived into adaptive optimizers and proposed an algorithm, called AdaLip, that constructs a learning rate per layer based on the Lipschitz constant. Various reasons were presented to show the need of the different approach of different layers and were supported experimentally. We tested our method on image classification datasets and the results showed imrovements in the convergence speed and the overall training performance. The proposed algorithm can work together with other optimizers and boost their performance scores. Finally, a theoretical proof of convergence of the new optimizer was presented.

In the final part of the thesis, we delved into the field of interpreting neural networks. Interpretability is concerned with understanding neural networks and their predictions. Initially, we explored various interpretability techniques and compared their performances. The experiments were based on medical images for classifying the stages of retinopathy. This contributed to a deeper understanding of the model's functionality in relation to retinal images. We also showed that using interpretability techniques it becomes possible to tackle the problem of image segmantation. Furthermore, we delved deeper into the operation of interpretability methods, specifically those employing reference points. We demonstrated that the use of reference points entails many risks of inaccuracy in interpreting neural network predictions. Based on this limitation, we proposed a new layer aiming to improve these issues. The proposed Baseline-Aware Embedding layer was designed to incorporate the concept of a baseline or reference point through training. Thus, interpretability algorithms that operate with reference points can utilize this layer in network architectures to generate more accurate interpretations for various predictions. We demonstrated this experimentally on four tabular datasets, chosen for their diverse features and the common occurrence of reference point issues.

Keywords

Machine Learning, Deep Learning, Neural Networks, Convolutional Neural Networks, Optimization, Learning Algorithms, Stochastic Methods, Sampling, Interpretability, Explainability, Baselines, Embedding

Ευχαριστίες

Η παρούσα διδακτορική διατριβή εκπονήθηκε στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης της Σχολής των Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου υπό την επίβλεψη του κύριου Ανδρέα-Γεώργιου Σταφυλοπάτη, Καθηγητή του ΕΜΠ. Από τον Νοέμβριο το 2017, ο κύριος Σταφυλοπάτης με καθοδήγησε και βοήθησε στο πορεία των διδακτορικών μου σπουδών μέχρι την ολοκλήρωσή τους. Η υποστήριξή του σε ακαδημαϊκό και ερευνητικό επίπεδο αλλά εξίσου και σε προσωπικό ήταν ένας από τους κύριους παράγοντες που συνέβαλαν στην περάτωση αυτής της διδακτορικής διατριβής και οφείλω να τον ευχαριστήσω για την αδιάκοπη προσπάθεια και όρεξη που κατέβαλε αυτά τα χρόνια.

Οφείλω, επίσης, να ευχαριστήσω τους άλλους δύο καθηγητές της τριμελούς συμβουλευτικής επιτροπής, τον κύριο Δημήτριο-Διονύσιο Κουτσούρη και τον Γεώργιο Στάμου για τις συμβολές και την καθοδήγησή τους αυτά τα χρόνια. Οι προτάσεις και οι παρατηρήσεις τους ήταν πολύτιμες για τα ερευνητικά μου βήματα. Οφείλω επίσης να ευχαριστήσω τα υπόλοιπα μέλη της επταμελούς επιτροπής, τον κύριο Στέφανο Κόλλια, την κυρία Κωνσταντίνα Νικήτα, τον κύριο Αθανάσιο Βουλόδημο και τον κύριο Γεώργιο Αλεξανδρίδη για την παρουσία τους και την συμμετοχή τους στην τελική εξέταση της διατριβής.

Θα ήθελα να ευχαριστήσω όλα τα μέλη του εργαστηρίου για την συνεργασία που είχαμε αυτά τα χρόνια και το ευχάριστο κλίμα που επικρατούσε σε καθημερινή βάση. Πιο συγκεκριμένα, ήθελα να ευχαριστήσω τους Ε.ΔΙ.Π. του εργαστηρίου, κύριο Γεώργιο Σιόλα και κυρία Παρασκευή Τζούβελη, για την αρίστη συνεργασία που είχαμε σε διάφορα ερευνητικά και διδακτικά καθήκοντα. Μία ιδιαίτερη αναφορά αλλά και ευχαριστίες οφείλω να δώσω στους Διδάκτορες και Υποψήφιους Διδάκτορες του εργαστηρίου, που με τον καιρό έγιναν κοντινοί μου φίλοι. Ευχαριστώ τους Θάνο Τάγαρη, Τάσο Παπαγιάννη, Θάνο Τασάκο, Ελένη Βάθη, Άρη Λαναρίδη, Μάρα Σδράκα και Πάνο Κουρή, για την συμπαράσταση και την συνεργασία τους. Τέλος, θα ήθελα να ευχαριστήσω μέσα από τη καρδιά μου τους γονείς μου για την στήριξή τους όλα αυτά τα χρόνια, από την αρχή των προπτυχιακών μου σπουδών μέχρι και το τέλος του διδακτορικού μου. Χωρίς την υποστήριξή τους αυτό το διδακτορικό μου ταξίδι δεν θα ήταν εφικτό.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη είναι ένας σύγχρονος κλάδος της επιστήμης των υπολογιστών, ο οποίος έχει γνωρίσει ραγδαία ανάπτυξη τα τελευταία χρόνια. Ο σκοπός του είναι η δημιουργία μηχανών με ευφυΐα, που μιμείται αυτή των ανθρώπων. Μία ευφυΐα, που αν και κατασκευασμένη, θα μπορεί να διαθέτει κάποια από τα χαρακτηριστικά και τις ικανότητες της ανθρώπινης. Αυτό συνεπάγεται στην κατασκευή ενός συστήματος που να μπορεί να αντιληφθεί το περιβάλλον του και να μπορεί να αποφασίσει τις κινήσεις που θα κάνει, προσπαθώντας να πετύχει τον εκάστοτε στόχο του. Από το 1956 που εδραιώθηκε ο όρος της Τεχνητής Νοημοσύνης (ΤΝ), έχουν υπάρξει πολλές μορφές και προσεγγίσεις της.

Από τότε υπήρξαν προσπάθειες στην προσομοίωση του ανθρωπίνου εγκεφάλου και της ανθρώπινης λογικής. Μετά από πολλά χρόνια έρευνας με επιτυχίες και αποτυχίες έχουμε φτάσει στην σημερινή μορφή της ΤΝ. Πολλές εφαρμογές βασίζονται πάνω σε κάποια έκφανση της τεχνητής νοημοσύνης, όπως η κατανόηση φυσικής γλώσσας, η αναγνώριση εικόνων, αυτόνομη οδήγηση. Σύμφωνα με την ορολογία των [49], μπορούμε να κατατάξουμε αυτές τις εφαρμογές σε τρεις κατηγορίες ΤΝ:

- Περιορισμένη Τεχνητή Νοημοσύνη: Η πρώτη γενιά εφαρμογών τεχνητής νοημοσύνης, οι οποίες εφαρμόζουν ευφυή συστήματα πάνω σε πολύ συγκεκριμένα εργασίες χωρίς κάποιου είδους αυτοματοποίηση.
- Γενική Τεχνητή Νοημοσύνη: Η δεύτερη γενιά τεχνητής νοημοσύνης, η οποία θα εμπεριέχει εφαρμογές που θα είναι ικανές να σχεδιάσουν κάποιο πλάνο ώστε να επιλύουν προβλήματα χωρίς να είναι κατασκευασμένα ειδικά για αυτό το σκοπό. Θα έχουν την δυνατότητα δηλαδή να εφαρμοστούν σε πολλά πεδία και να λειτουργούν με μεγάλη αυτονομία.
- Τεχνητή Ύπερ-Νοημοσύνη: Η τρίτη γενιά τεχνητής νοημοσύνης, η οποία αποτελεί μία εξέλιξη των άλλων δύο και θεωρείται η πραγματική τεχνητή νοημοσύνη. Τέτοιου είδους

ΤΝ θα φτάνει σε στάδια που θα αγγίζει τα όρια της τεχνητής συνείδησης. Συστήματα αυτής της γενιάς θα μπορούν να εφαρμοστούν σε όλους τους κλάδους και να παρουσιάσουν, εκτός από υπολογιστική δύναμη, δημιουργικότητα και εφευρετικότητα.

Οι δύο τελευταίες κατηγορίες είναι ακόμα σε αρκετά πρώιμο στάδιο και θα χρειαστούν χρόνια για την επίτευξή τους αλλά τα τελευταία χρόνια έχουν αποδείξει την δυνατότητα για ραγδαία ανάπτυξη στον τομέα της ΤΝ. Σε αυτή τη διατριβή θα ασχοληθούμε με έναν ιδιαίτερο κλάδο της ΤΝ, ο οποίος κατασκευάζει συστήματα τα οποία εκπαιδεύονται από μόνα τους με την χρήση πραγματικών δεδομένων. Ο κλάδος αυτός της Τεχνητής Νοημοσύνης ονομάζεται Μηχανική Μάθηση και είναι ένας από τους παράγοντες που συνέβαλαν στην ανάπτυξη της ΤΝ τα τελευταία χρόνια.

1.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση ασχολείται με την δημιουργία έξυπνων μηχανών, οι οποίες μαθαίνουν να λύνουν τα εκάστοτε προβλήματα από μόνες τους με τη βοήθεια δεδομένων. Ένας σύγχρονος και πιο αναλυτικός ορισμός [73] της Μηχανικής Μάθησης είναι ο ακόλουθος:

Ορισμός 1.1. *Ένα πρόγραμμα υπολογιστή μαθαίνει από μία εμπειρία E να εκτελεί ένα είδος έργων T με επίδοση P, αν η επίδοση P σε έργα σαν το T βελτιώνεται με την εμπειρία E. Οι τέσσερεις αυτοί όροι μπορούν να συνοψιστούν σε παρακάτω σημεία:*

- *Το πρόγραμμα, που αντιπροσωπεύει τον αλγόριθμο μάθησης.*
- *Την εμπειρία E, η οποία αποτελεί το σύνολο των δεδομένων από τα οποία θα μάθει το πρόγραμμα.*
- *Το έργο T, το οποίο είναι το πρόβλημα προς επίλυση.*
- *Την επίδοση P, με την οποία μετρούμε το πόσο καλά ανταποκρίνεται το πρόγραμμά μας στο έργο T.*

Η Μηχανική Μάθηση χωρίζεται σε τρεις βασικές κατηγορίες ανάλογα με τον τρόπο εκπαίδευσης του μοντέλου και των δυνατοτήτων πυ του δίνονται με βάση τα δεδομένα. Οι κατηγορίες είναι οι εξής:

- Επιβλεπόμενη Μάθηση (Supervised Learning): Το μοντέλο έχει στη διάθεσή του την είσοδο του προβλήματος αλλά και την επιθυμητή έξοδο. Ο σκοπός του είναι να μάθει να αντιστοιχίζει την είσοδο με την αντίστοιχη έξοδο. Είναι από τους πιο πολυχρησιμοποιημένους τρόπους μάθησης και από αυτούς που έχουν παρουσιάσει την μεγαλύτερη επιτυχία. Η επιβλεπόμενη μάθηση περιλαμβάνει αλγόριθμους, όπως ταξινόμηση, παλινδρόμηση, αλλά και εφαρμογές όπως συστήματα συστάσεων.

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Το μοντέλο έχει μόνο στη διάθεσή του την είσοδο του προβλήματος. Δεν χρησιμοποιεί κάποια επιθυμητή έξοδο ή κάποια ετικέτα για το κάθε δείγμα του συνόλου δεδομένων. Σκοπός της είναι η εύρεση της κρυφής δομής των δεδομένων και των σχέσεων των χαρακτηριστικών μεταξύ τους. Χρησιμοποιείται συνήθως σε συσταδοποίηση (χωρισμός των δειγμάτων των δεδομένων σε ομάδες), σε εκμάθηση χαρακτηριστικών ή σε προβλήματα μείωσης της διαστατικότητας.

- Ενισχυτική Μάθηση (Reinforcement Learning): Το μοντέλο αλληλεπιδρά με ένα δυναμικό περιβάλλον και καλείται να επιλέξει σε κάθε κατάσταση του περιβάλλοντος την καλύτερη δυνατή κίνηση για την επίτευξη του στόχου του προβλήματος. Την κίνηση αυτή μαθαίνει να τη βρίσκει με τη βοήθεια ανάδρασης προσπαθώντας να βελτιώσει κάποια συνάρτηση ανταμοιβής. Έχει πολλές εφαρμογές, όπως αυτοοδηγούμενα αυτοκίνητα και πράκτορες διαφόρων παιχνιδιών.

Αυτές οι τρεις κατηγορίες είναι οι κύριες κατηγορίες της Μηχανικής Μάθησης, όμως τα τελευταία χρόνια έχουν αναδειχθεί υποκλάδοι τους που αρχίζουν να έχουν εξίσου υψηλή σημασία. Η ημι-επιβλεπόμενη μάθηση (Semi-Supervised Learning) έχει γνωρίσει σημαντική ανάπτυξη επειδή μπορεί με ένα μικρό σύνολο επισημασμένων δεδομένων μαζί με πολλά δεδομένα χωρίς ετικέτες να καταφέρει να εκπαιδεύσει ανταγωνιστικά μοντέλα. Συνδυάζει χαρακτηριστικά επιβλεπόμενης και μη-επιβλεπόμενης μάθησης, γι΄ αυτό και έχει αποκτήσει αυτό το όνομα. Μία άλλη μέθοδος μάθησης που αναπτύσσεται αρκετά και έχει ελπίδες για μεγαλύτερη βελτίωση είναι η αυτο-επιβλεπόμενη μάθηση (Self-Supervised Learning). Αυτού του είδους η μάθηση βασίζεται στην εκμάθηση ενός ενδιάμεσου έργου, κατασκευασμένο από τεχνητές ετικέτες, με σκοπό την προεκπαίδευση του μοντέλου. Έπειτα, ακολουθεί η εκπαίδευση του κυρίως έργου. Έχει δείξει ότι σε διάφορες εφαρμογές παρουσιάζει εξαίρετες επιδόσεις συγκρινόμενες και με τις τελευταίες τεχνολογίες.

1.3 Βαθιά Μηχανική Μάθηση

Η χρήση και η ανάπτυξη εξελιγμένης τεχνολογίας υλικού (ιδιαίτερα κάρτες γραφικών) οδήγησε στην αποτελεσματική υλοποίηση της Βαθιάς Μηχανικής Μάθησης (Deep Learning) [60]. Συμπεριλαμβάνει αλγορίθμους που εξάγουν χαρακτηριστικά μέσω πολλαπλών επιπέδων μάθησης. Η πληθώρα των επιπέδων δίνει την δυνατότητα στα μοντέλα να εκπαιδεύονται σε πολύπλοκες αναπαραστάσεις δεδομένων και σε αρκετά δύσκολα έργα. Σε συνδυασμό με την διαθεσιμότητα μεγάλου όγκου δεδομένων, η Βαθιά Μηχανική Μάθηση έχει εδραιωθεί τη τελευταία δεκαετία σαν επιλογή σε πολλές εφαρμογές τεχνητής νοημοσύνης, όπως η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας.

Η Βαθιά Μηχανική Μάθηση βασίζεται, κυρίως, στα Τεχνητά Νευρωνικά Δίκτυα. Γι΄ αυτό το λόγο εφαρμόζεται κατά κανόνα σε εφαρμογές επιβλεπόμενης μάθησης. Υπάρχουν, όμως,

εξαιρέσεις με αλγόριθμους βαθιάς μηχανικής μάθησης που εκπαιδεύονται σε δεδομένα με μη-επιβλεπόμενο χαρακτήρα, όπως οι αυτοκωδικοποιητές. Σε αυτή τη διατριβή θα ασχοληθούμε με προβλήματα Βαθιάς Μηχανικής Μάθησης και, συγκεκριμένα με Νευρωνικά Δίκτυα.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό θα αναφέρουμε τις θεωρητικές έννοιες και θα αναλύσουμε το γενικό θεωρητικό υπόβαθρο, το οποίο θα χρειαστεί για το υπόλοιπο την διατριβής.

2.1 Επιβλεπόμενη Μάθηση

Η επιβλεπόμενη μάθηση κατατάσσεται ως η πιο διαδεδομένη μορφή μηχανική μάθησης. Παρακάτω θα ορίσουμε κάποιες έννοιες και θα εισάγουμε συμβολισμούς που θα χρησιμοποιηθούν στην υπόλοιπη διατριβή. Έστω θέτουμε ως $X = \{x_1, ..., x_N\}$ το σύνολο των δεδομένων μας, που αποτελείται από N δείγματα. Το κάθε δείγμα εμπεριέχει M χαρακτηριστικά, που θα τα συμβολίζουμε με $x_i = \{x_i^1, x_i^2, ..., x_i^M\}$. Ο σκοπός της επιβλεπόμενης μάθησης είναι να κατασκευάσει μία αντιστοίχηση μεταξύ της εισόδου και της εξόδου. Η επιθυμητή έξοδος για το i-οστό δείγμα του συνόλου δεδομένων είναι y_i. Επομένως, ορίζουμε ως την αντιστοίχηση (μοντέλο) g που θέλουμε να μάθουμε ώστε να επιλύει το ακόλουθο:

$$g : x_i \rightarrow y_i \tag{2.1}$$

Η παραπάνω σχέση αποτελεί την ιδανική συνάρτηση g, η οποία επιλύει το συγκεκριμένο έργο. Όμως, τέτοια συνάρτηση είναι πολύ δύσκολο να επιτευχθεί σε τέλειο βαθμό, είτε λόγω της αδυναμίας του μοντέλου, είτε λόγω της φύσης των δεδομένων εκπαίδευσης. Για αυτό το λόγο, το ατελές μοντέλο g θα το ονομάζουμε εκτιμητή (εστιματορ), επειδή προσπαθεί να προσεγγίσει αυτή την συνάρτηση. Αντίστοιχα, τις εξόδους του μοντέλου θα τις αποκαλούμε προβλέψεις και θα τις συμβολίζουμε με $\hat{y}_i$. Ως αποτέλεσμα η συνάρτηση του εκτιμητή (μοντέλου) μας θα χαρακτηρίζεται από την ακόλουθη σχέση:

$$\hat{y}_i = g(x_i) \tag{2.2}$$

Υπάρχουν πολλών ειδών μοντέλα επιβλεπόμενης μάθησης, τα οποία βασίζονται στη παραπάνω σχέση. Σε αυτή τη διατριβή θα επικεντρωθούμε στα Νευρωνικά Δίκτυα και θα ανα-

λύσουμε την λειτουργία τους, τους αλγορίθμους εκπαίδευσής τους, καθώς, και τις διάφορες αρχιτεκτονικές που έχουν αναπτυχθεί τα τελευταία χρόνια.

2.2 Νευρωνικά Δίκτυα

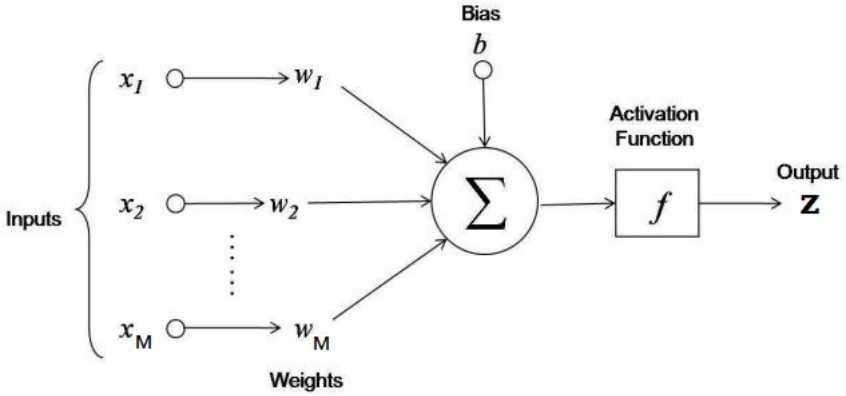
Τα Νευρωνικά Δίκτυα (ή Τεχνητά Νευρωνικά Δίκτυα) είναι ένας υποκλάδος της Μηχανικής Μάθησης και αποτελούν, κατά κύριο λόγο, συστήματα επιβλεπόμενης μάθησης. Έχουν εμπνευστεί από το βιολογικό νευρικό σύστημα του ανθρώπου και τον τρόπο που επικοινωνούν οι βιολογικοί νευρώνες μεταξύ τους. Όμως, για την εκπαίδευσή τους απαιτούνται ένας μεγάλος όγκος δεδομένων. Ευτυχώς, τα τελευταία χρόνια με την χρήση νέων τεχνολογιών είναι εύκολα προσβάσιμος μεγάλος όγκος δεδομένων για πολλά διαφορετικά έργα και εφαρμογές. Αυτό οδήγησε στην ανάπτυξη των Νευρωνικών Δικτύων σε πολλούς τομείς. Πρόσφατα έχουν αναδειχθεί ως η καλύτερη τεχνολογία για συγκεκριμένα έργα, όπως η αναγνώριση εικόνας, η αναγνώριση φωνής, αλλά και συστήματα συστάσεων, όπως ο αλγόριθμος αναζήτησης της Google. Από τι αποτελούνται τα νευρωνικά δίκτυα και πως εκπαιδεύονται μέσω των δεδομένων;

2.2.1 Νευρώνας

Το βασικό δομικό στοιχείο των νευρωνικών δικτύων είναι ο νευρώνας. Κάθε νευρώνας λαμβάνει εισόδους, κάνει υπολογισμούς με αντίστοιχα βάρη και βγάζει μία έξοδο. Ο πιο απλός τύπος νευρώνα (Perceptron) πολλαπλασιάζει ένα διάνυσμα βαρών με την είσοδο και προσθέτει μία μεταβλητή, που ονομάζεται πόλωση. Το αποτέλεσμα, τέλος, το εισάγει σε μία συνάρτηση ενεργοποίησης. Σε μαθηματικό τύπο το παραπάνω μπορεί να εκφραστεί ως:

$$z = f(w \cdot x + b) \tag{2.3}$$

όπου w είναι τα βάρη, x η είσοδος, b η πόλωση (bias). Η f συμβολίζει τη συνάρτηση ενεργοποίησης (activation function), η οποία συνήθως είναι μη-γραμμική. Υπάρχουν διάφορες συναρτήσεις που μπορούν να χρησιμοποιηθούν και η καθεμία είναι ειδική για διαφορετικές λειτουργίες. Σε επόμενο υποκεφάλαιο θα αναφερθούμε πιο αναλυτικά στην επιλογή κατάλληλης συνάρτησης. Σαν z ορίζουμε την έξοδο του νευρώνα. Στο Σχήμα 2.1 μπορούμε να παρατηρήσουμε την λειτουργία του απλού νευρώνα. Το διάνυσμα w έχει όσες διαστάσεις όσες και η είσοδος ($w \in \mathbb{R}^M$), ώστε να είναι εφικτό το εσωτερικό γινόμενο. Η πόλωση είναι ένας πραγματικός αριθμός ($b \in \mathbb{R}$). Το μοντέλο ενός νευρώνα μοιάζει με αυτό της Γραμμικής Παλινδρόμησης.



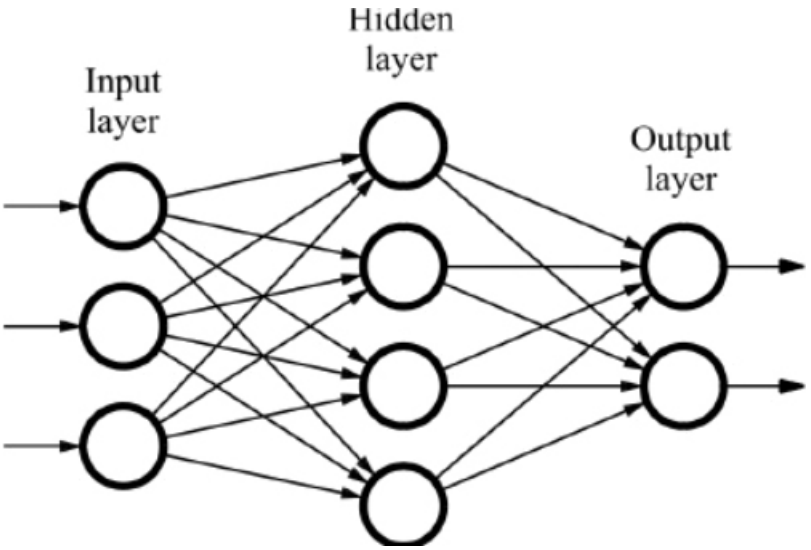
Σχήμα 2.1: Η ανατομία ενός απλού νευρώνα, που αποτελείται από τις εισόδους του, τη πράξη του εσωτερικού γινομένου με τα βάρη, τη πόλωση, τη συνάρτηση ενεργοποίησης και την τελική του έξοδο.

2.2.2 Πολυεπίπεδα Νευρωνικά Δίκτυα

Η δύναμη ενός νευρώνα σαν ένα εργαλείο μάθησης, όμως, είναι περιορισμένη. Για αυτό το λόγο τα Νευρωνικά Δίκτυα ομαδοποιούν τους τεχνητούς νευρώνες σε επίπεδα ή στρώματα (layers). Το επίπεδο αποτελείται από νευρώνες που δέχονται την ίδια είσοδο και, ανεξάρτητα μεταξύ τους, βγάζουν μία έξοδο ο καθένας. Η μαθηματική διατύπωση δεν αλλάζει ριζικά από την εξίσωση 2.3. Στην περίπτωση ενός επιπέδου νευρώνων το w αντικαθίσταται από ένα διάνυσμα βαρών $W \in \mathbb{R}^{M \times K}$, όπου K είναι ο αριθμός των νευρώνων στο επίπεδο. Αντίστοιχα με πριν, η πόλωση θα μετατραπεί σε διάνυσμα K διαστάσεων και θα λαμβάνουμε K εξόδους $(b, z \in \mathbb{R}^K)$. Η πληθώρα νευρώνων σε ένα επίπεδο προσομοιάζει πολλά παράλληλα γραμμικά μοντέλα (π.χ. γραμμική παλινδρόμηση), τα οποία εκπαιδεύονται ανεξάρτητα μεταξύ τους.

Το επόμενο βήμα είναι να ενώσουμε τις εξόδους σε μία συγκεκριμένη αναπαράσταση, όμοια με αυτή που ζητά το κάθε έργο. Για να συνδυάσουμε τις εξόδους, ορίζουμε ένα άλλο επίπεδο νευρώνων, που θα αποκαλείται επίπεδο εξόδου (output layer). Το επίπεδο εξόδου θα δέχεται ως εισόδους τις εξόδους του προηγούμενου επιπέδου, θα κάνει τις απαραίτητες πράξεις (σύμφωνα με την εξίσωση 2.3) και θα βγάζει την τελική πρόβλεψη του δικτύου. Αυτό μπορούμε να το δούμε στο Σχήμα 2.2 που απεικονίζει ένα νευρωνικό δίκτυο με 2 επίπεδα. Το επίπεδο πριν το επίπεδο εξόδου θα το ονομάσουμε κρυφό επίπεδο (hidden layer). Στο σχήμα παρατηρούμε και ένα τρίτο επίπεδο, το επίπεδο της εισόδου (input layer). Αυτό το επίπεδο είναι ένα εικονικό επίπεδο, το οποίο παριστάνει την είσοδο που δέχεται το δίκτυο. Δεν μετράται σαν μέρος του δικτύου, απλά χρησιμοποιείται σχηματικά για αναπαράσταση της εισόδου.

Με βάση αυτή την ιδέα μπορούμε να επεκτείνουμε το παραπάνω δίκτυο σε μία πιο γενικευμένη μορφή με πάνω από ένα κρυφό επίπεδο. Δίκτυα με πολλά κρυφά επίπεδα ονομάζονται Πολυεπίπεδα Νευρωνικά Δίκτυα (Multilayer Perceptron - MLP) και είναι η πιο συνήθης μορ-



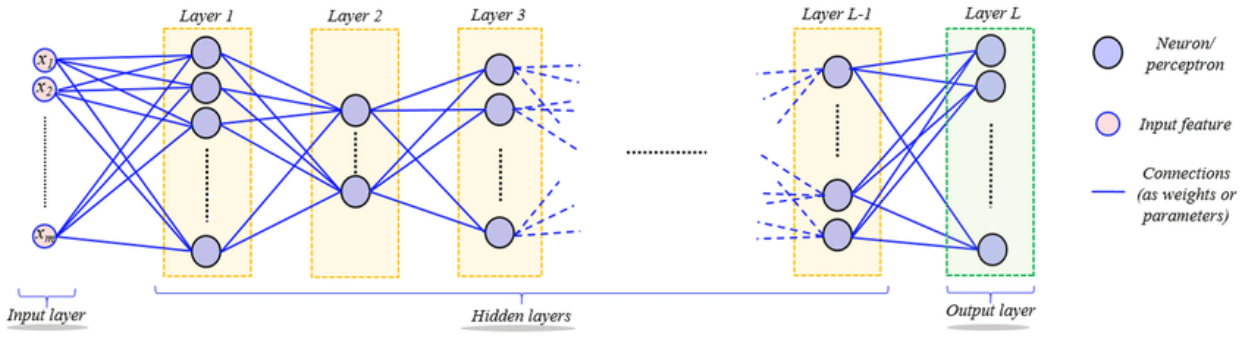
Σχήμα 2.2: Η αναπαράσταση ενός νευρωνικού δικτύου με δύο επίπεδα, το κρυφό επίπεδο και το επίπεδο της εξόδου.

φή νευρωνικού δικτύου στις σύγχρονες εφαρμογές. Στο Σχήμα 2.3 φαίνεται ένα παράδειγμα δικτύου με L επίπεδα. Η έξοδος του $i-1$ επιπέδου δίνεται σαν είσοδο στο i επίπεδο. Η συνολική μαθηματική σχέση περιγράφεται από την παρακάτω σχέση:

$$\hat{y} = z^L(z^{L-1}(...z^1(x))) \tag{2.4}$$

όπου με $\hat{y}$ ορίζουμε την έξοδο όλου του δικτύου, z^i ορίζουμε την έξοδο του i-οστού επιπέδου και x συμβολίζουμε την είσοδο. Η έξοδος z^i περιγράφεται με τον παρακάτω τύπο:

$$z^i = f(W \cdot z^{i-1} + b) \tag{2.5}$$



Σχήμα 2.3: Η γενικευμένη αρχιτεκτονική ενός Πολυεπίπεδου Νευρωνικού Δικτύου με L επίπεδα.

2.2.3 Συναρτήσεις Ενεργοποίησης

Όπως, αναφέρθηκε και προηγουμένως κάθε νευρώνας περνά την έξοδο του από μία συνάρτηση ενεργοποίησης. Ο ρόλος της είναι να μετατρέπει την έξοδο με μη-γραμμικό τρόπο, έχοντας την δυνατότητα να μοντελοποιεί πολύπλοκες αναπαραστάσεις δεδομένων και σχέσεων, τις οποίες κάποια γραμμική συνάρτηση δεν θα ήταν σε θέση να το καταφέρει. Είναι ένα σημαντικό κομμάτι της επιτυχίας των νευρωνικών δικτύων και υπάρχουν διάφορα είδη τέτοιων συναρτήσεων. Παρακάτω θα αναφέρουμε τις πιο δημοφιλείς και την χρησιμότητά τους:

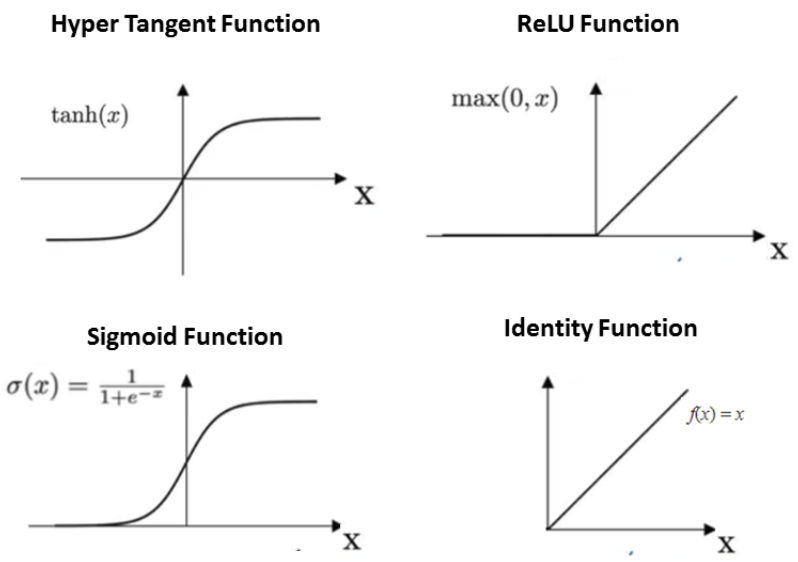
- Γραμμική (Linear, Identity): $f(x) = x$
- Σιγμοειδής (Sigmoid, Logistic): $f(x) = \frac{1}{1+e^{-x}}$
- Υπερβολική Εφαπτομένη (Hyperbolic Tangent): $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
- Διορθωμένη Γραμμική (Rectified Linear Unit or ReLU): $f(x) = max(0, x)$
- Softmax: $f(x_i) = \frac{e_i^x}{\sum_{j=1}^{C} e^{x_j}}$

Ανάλογα την θέση του νευρώνα μέσα στο δίκτυο, αλλάζει και ποια είναι η καταλληλότερη συνάρτηση. Στα ενδιάμεσα επίπεδα (κρυφά) συνήθως χρησιμοποιείται η ReLU συνάρτηση [76] γιατί έχει αποδειχθεί ότι η μη-γραμμικότητα σε συνδυασμό με την απλότητα που προσφέρει ενισχύει αρκετά την δύναμη του δικτύου. Άλλες συναρτήσεις είναι χρήσιμες στο επίπεδο εξόδου ενός δικτύου. Για παράδειγμα, σε ένα σενάριο ταξινόμησης 2 κλάσεων θα βοηθούσε αρκετά η Σιγμοειδής συνάρτηση για να περιορίσει την έξοδο του δικτύου στο εύρος $[0, 1]$. Αυτό είναι βολικό γιατί έτσι αναπαριστούμε την πρόβλεψη, αν ανήκει ένα δείγμα σε μια κλάση, με έξοδο 1, και αντίστοιχα αν δεν ανήκει με 0.

Αν το γενικεύσουμε σε πολλές κλάσεις, για παράδειγμα C κλάσεις, θα κατασκευάσουμε ένα δίκτυο που θα βγάζει C εξόδους στο $[0, 1]$, αναπαριστώντας αντίστοιχα το ανήκει (1) ή δεν ανήκει (0) στην κάθε κλάση-έξοδο. Τέτοιου είδους αναπαράσταση ονομάζεται one-hot encoding και έχει καθιερωθεί ως ο κύριος τρόπος απεικόνισης πολλών κλάσεων. Σε αυτή τη περίπτωση η συνάρτηση Softmax είναι η καταλληλότερη, επειδή μετατρέπει την έξοδο στο εύρος $[0, 1]$ αλλά με την ιδιότητα των πιθανοτήτων (ότι το άθροισμα να είναι 1). Σε περιπτώσεις παλινδρόμησης, αν δεν υπάρχει κάποιος περιορισμός που θέλουμε να επιβάλλουμε στην έξοδο, τότε η γραμμική συνάρτηση χρησιμοποιείται συχνά. Γενικά, αν η έξοδός μας κυμαίνεται γύρω από κάποια όρια, οι συναρτήσεις ενεργοποίησης έχουν την ιδιότητα να περιορίζουν την έξοδο του νευρωνικού, ώστε οι προβλέψεις του δικτύου μας να είναι πάντα έγκυρες. Στο Σχήμα 2.4 φαίνονται οι γραφικές παραστάσεις από κάποιες από τις παραπάνω συναρτήσεις.

2.3 Διαδικασία της Μάθησης

Ένα σημαντικό κομμάτι των Νευρωνικών Δικτύων είναι ο τρόπος εκπαίδευσής τους. Ο σκοπός της διαδικασίας της μάθησης είναι να τροποποιήσει τα βάρη του νευρωνικού (συμπε-



Σχήμα 2.4: Οι απεικονίσεις τεσσάρων συναρτήσεων ενεργοποίησης (Υπερβολική εφαπτομένη, Διορθωμένη Γραμμική, Σιγμοειδής και Γραμμική).

ριλαμβανομένου και των πολώσεων), ώστε να καταφέρει το δίκτυο να βγάζει ως έξοδο την σωστή πρόβλεψη για την εκάστοτε είσοδο. Διαφορετικοί τρόποι μάθησης οδηγούν σε διαφορετικά δίκτυα, τα οποία με τη σειρά τους οδηγούν σε διαφορετικές επιδόσεις. Επομένως, η επιλογή του αλγόριθμου μάθησης είναι εξαιρετικά σημαντική για την επιτυχή εκπαίδευση και την καλή απόδοση του νευρωνικού. Πώς, όμως, γίνεται αυτή η εκπαίδευση·

2.3.1 Συνάρτηση Κόστους

Η μεθοδολογία για να εκπαιδεύσουμε ένα δίκτυο βασίζεται στο να ωθήσουμε την πρόβλεψη του δικτύου όσο πιο κοντά γίνεται στην επιθυμητή έξοδο. Με άλλα λόγια να κατασκευάσουμε ένα νευρωνικό που το $\hat{y}$ θα τείνει στο y. Για να υπολογίσουμε πόσο κοντά βρίσκεται το $\hat{y}$ στην πραγματική τιμή, ορίζουμε μία μετρική που θα καταγράφει το σφάλμα της πρόβλεψής μας. Αυτή τη μετρική θα την ονομάζουμε συνάρτηση κόστους (ή αλλιώς συνάρτηση απώλειας ή Loss function). Επομένως, επικεντρώνουμε την διαδικασία της μάθησης στην ελαχιστοποίηση της συνάρτησης κόστους. Όταν αυτή φτάσει σε πολύ χαμηλά επίπεδα, δηλαδή το δίκτυο κάνει πολύ μικρά σφάλματα, τότε μπορούμε να πούμε πως το μοντέλο μας κατάφερε να εκπαιδευτεί. Παρακάτω, θα παραθέσουμε δύο παραδείγματα συναρτήσεων κόστους στα δύο πιο σημαντικά έργα, ταξινόμηση και παλινδρόμηση.

Για παλινδρόμηση μία αρκετά χρήσιμη συνάρτηση κόστους είναι το μέσο τετραγωνικό σφάλμα ή μέση τετραγωνική απόκλιση (Mean Square Error - MSE). Ο τύπος της είναι ο εξής:

$$J(y, \hat{y}) = MSE(y, \hat{y}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.6}$$

όπου J θα συμβολίζουμε την συνάρτηση κόστους και N θέτουμε ως το πλήθος των δειγμάτων του συνόλου δεδομένων. Μικρό J σημαίνει μικρό σφάλμα μεταξύ y και $\hat{y}$ και, συνεπώς, καλή πρόβλεψη. Αθροίζουμε πάνω σε όλα τα δείγματα των δεδομένων εκπαίδευσης γιατί μας ενδιαφέρει το μοντέλο να τα πηγαίνει καλά σε όλα τα δεδομένα, αλλά και για να έχουμε μία γενική εικόνα της επίδοσης του μοντέλου. Για ταξινόμηση, ένα παράδειγμα συνάρτησης κόστους είναι η διασταυρούμενη εντροπία (cross-entropy). Μπορεί να περιγραφεί από τον παρακάτω τύπο:

$$J(y, \hat{y}) = H(y, \hat{y}) = \sum_{i=1}^{N} \sum_{j=1}^{C} (y_i^j \log \hat{y}_i^j) \tag{2.7}$$

όπου y_i^j συμβολίζει την ετικέτα του δείγματος i ως προς την κλάση j. Αυτή η κωδικοποίηση των ετικετών είναι της μορφής one-hot που προαναφέρθηκε προηγουμένως.

2.3.2 Βελτιστοποίηση

Με βάση την συνάρτηση κόστους, το πρόβλημα της εκπαίδευσης ενός Νευρωνικού Δικτύου μπορεί να μετασχηματιστεί σε ένα πρόβλημα βελτιστοποίησης. Ο στόχος είναι να βρεθούν τα κατάλληλα βάρη, ώστε η συνάρτηση κόστους να ελαχιστοποιηθεί. Αυτό μπορούμε να το γράψουμε με τον ακόλουθο μαθηματικό τύπο:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, J(y, \hat{y}) \tag{2.8}$$

Ο παραπάνω τύπος περιγράφει το πρόβλημα βελτιστοποίησης, το οποίο είναι η εύρεση των παραμέτρων (θ) ώστε να ελαχιστοποιείται το J. Παράμετροι θεωρούνται όλες οι εκπαιδεύσιμες μεταβλητές. Στην δική μας περίπτωση είναι τα βάρη και οι πολώσεις του κάθε νευρώνα. Για την επίλυση τέτοιων ειδών προβλημάτων υπάρχουν διάφορες τεχνικές βελτιστοποίησης. Στα Νευρωνικά Δίκτυα χρησιμοποιούνται μέθοδοι μου βασίζονται στην παράγωγο της συνάρτησης κόστους ως προς τα βάρη. Τέτοιες προσεγγίσεις απαιτούν η συνάρτηση κόστους να είναι παραγωγίσιμη. Ο αλγόριθμος που θεωρείται επικρατέστερος μέχρι στιγμής είναι η κάθοδος κλίσης (gradient descent).

2.3.3 Κάθοδος Κλίσης

Η κάθοδος κλίσης είναι από τους πιο δημοφιλής αλγόριθμους βελτιστοποίησης και ο πιο διαδεδομένος τρόπος εκπαίδευσης των νευρωνικών δικτύων. Τα τελευταία χρόνια, με την

ανάπτυξη των Βαθιών Νευρωνικών Δικτύων, κάθε δίκτυο τελευταίας τεχνολογίας έχει εκπαιδευτεί με αυτόν τον αλγόριθμο ή με κάποια από τις διάφορες παραλλαγές του. Για τις παραλλαγές του αλγορίθμου θα μιλήσουμε σε επόμενες ενότητες. Ο αλγόριθμος της καθόδου κλίσης βασίζεται στην παράγωγο της συνάρτησης κόστους ως προς τις παραμέτρους. Αυτό μπορεί να γραφεί με τη βοήθεια του Ιακωβιανού πίνακα:

$$\nabla_{w_i} J(y, \hat{y}) = \frac{\partial J(y, \hat{y})}{\partial w_i} \tag{2.9}$$

Η κάθοδος κλίσης βασίζεται στην ιδιότητα, ότι όσο προχωρούμε προς την αντίθετη κατεύθυνση της παραγώγου, τόσο θα μειώνεται και η συνάρτηση κόστους. Ο αλγόριθμος αφαιρεί επαναληπτικά από τα βάρη την παράγωγο τους ως προς το κόστος. Με βάση αυτό μπορούμε να κατασκευάσουμε τον κανόνα ως εξής:

$$W_{t+1} = W_t - \alpha * \nabla_{w_t} J(y, \hat{y}) \tag{2.10}$$

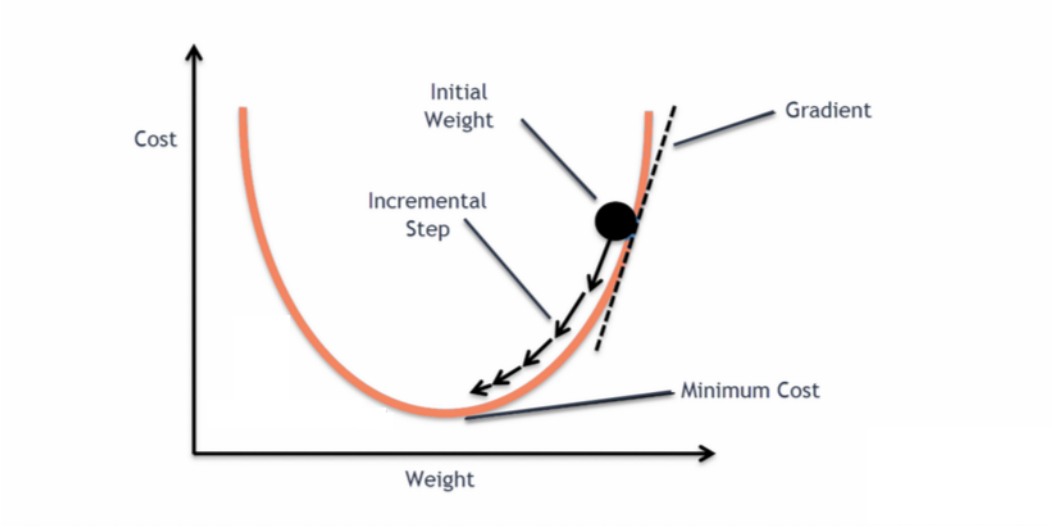
όπου W_t είναι το διάνυσμα των βαρών στην επανάληψη t, W_{t+1} είναι τα καινούργια βάρη και $\alpha \in \mathbb{R}^+$ είναι ο ρυθμός μάθησης (learning rate) ή βήμα μάθησης. Ο ρυθμός μάθησης ελέγχει το μέγεθος του μέτρου της αλλαγής των παραμέτρων σε κάθε επανάληψη. Ο παραπάνω τύπος μπορεί να γραφεί και μεμονωμένα για κάθε βάρος ξεχωριστά, επειδή η αλλαγή ενός βάρους είναι ανεξάρτητη από ένα άλλο. Για παράδειγμα για το w_i έχουμε:

$$w_i^{t+1} = w_i^t - \alpha * \frac{\partial J(y, \hat{y})}{\partial w_i^t} \tag{2.11}$$

όπου w_i^t είναι το i-οστό βάρος κατά την t επανάληψη. Ο σκοπός του αλγορίθμου είναι μετά από αρκετές επαναλήψεις, να φτάσει σε ένα σημείο που η κλίση θα είναι κοντά στο 0 και οι ανανεώσεις των βαρών θα είναι εξαιρετικά μικρές. Επομένως, με το πέρας των επαναλήψεων ο αλγόριθμος ¨καταλήγει' σε ένα σημείο. Το σημείο αυτό ονομάζεται σημείο σύγκλισης και θα λέμε ότι ένα δίκτυο συγκλίνει όταν φτάνει σε ένα σταθερό σημείο που με το χρόνο δεν θα αλλάζει. Στο σημείο που η παράγωγος είναι 0 ονομάζεται τοπικό ελάχιστο και είναι το χαμηλότερο σημείο σε εκείνη την περιοχή. Το χαμηλότερο σημείο ολόκληρης της συνάρτησης κόστους ονομάζεται ολικό ελάχιστο. Στο Σχήμα 2.5 απεικονίζεται η διαδικασία που ακολουθεί ο αλγόριθμος της καθόδου κλίσης για να φτάσει στο ελάχιστο σημείο.

2.3.4 Αλγόριθμος οπισθοδιάδοσης

Ο αλγόριθμος εκπαίδευσης που περιγράψαμε βασίζεται στην χρήση της παραγώγου. Ο υπολογισμός της παραγώγου, όμως, αποτελεί μία χρονοβόρα και δύσκολη δουλειά. Σε δίκτυα που αποτελούνται από πολλούς νευρώνες και πολλά επίπεδα, απαιτείται και ο υπολογισμός ενός μεγάλου αριθμού από παραγώγους, δυσχεραίνοντας την ομαλή και γρήγορη εκπαίδευση. Για αυτόν τον λόγο, είναι απαραίτητος ένας αποδοτικός αλγόριθμος υπολογισμού των παραγώγων των βαρών του δικτύου.



Σχήμα 2.5: Η διαδικασία του αλγορίθμου καθόδου κλίσης σε ένα διάγραμμα της συνάρτησης κόστους ως προς ένα βάρος.

Στα Νευρωνικά Δίκτυα έχει καθιερωθεί για την επίλυση αυτής της διεργασίας ο αλγόριθμος της οπισθοδιάδοσης. Χρησιμοποιεί την ιδιότητα των νευρωνικών δικτύων, ότι τα επίπεδα συνδέονται σειριακά μεταξύ τους και, συνεπώς, οι παράγωγοι θα εξαρτώνται από ένα επίπεδο σε ένα άλλο. Για παράδειγμα, το σφάλμα του προτελευταίο επιπέδου εξαρτάται από το σφάλμα του τελευταίου επιπέδου και αυτό απεικονίζεται και στις παραγώγους. Βασισμένος σε αυτή την ιδιότητα, ο αλγόριθμος οπισθοδιάδοσης χρησιμοποιεί τον κανόνα της αλυσίδας, ώστε να υπολογίσει τις παραγώγους των επιπέδων επαναληπτικά, ξεκινώντας από το τελευταίο επίπεδο και προχωρώντας προς τα πίσω. Με αυτόν τον τρόπο όταν προσπαθεί να υπολογίσει τις παραγώγους τους επιπέδου i, θα είναι ευκολότερο επειδή θα έχει ήδη υπολογισμένες τις παραγώγους του επιπέδου $i+1$. Ο κανόνας της αλυσίδας μπορεί να αναπαρασταθεί με την εξίσωση 2.12, όπου δείχνει την παράγωγο για το βάρος w_i που ανήκει στο επίπεδο i:

$$\frac{\partial J(y,\hat{y})}{\partial w_i} = \frac{\partial J(y,\hat{y})}{\partial z_L} \cdot \frac{\partial z_L}{\partial z_{L-1}} \cdot \ldots \cdot \frac{\partial z_i}{\partial w_i} \tag{2.12}$$

Ο όρος οπισθοδιάδοση προήλθε από την ανάποδη διεργασία που εκτελείται, ξεκινώντας από τις παραγώγους του τελευταίου επιπέδου και συνεχίζοντας προς τα πίσω. Στο παράδειγμα του βάρους w_i παραπάνω, η παράγωγος $\frac{\partial J(y,\hat{y})}{\partial z_L}$ έχει υπολογιστεί από προηγούμενες επαναλήψεις, οπότε εξοικονομούμε υπολογισμούς. Αυτή η μεθοδολογία είναι ένα παράδειγμα από δυναμικό προγραμματισμό.

2.3.5 Εκπαίδευση σε Παρτίδες

Ο αλγόριθμος εκπαίδευσης που αναφέρθηκε μέχρι στιγμής δέχεται όλα τα δεδομένα μαζί (σε μορφή ενός πολυδιάστατου πίνακα). Όμως, σε πραγματικές εφαρμογές Βαθιάς Μηχανικής Μάθησης τα δεδομένα εμπεριέχουν εκατομμύρια δείγματα, δημιουργώντας έναν τεράστιο πίνακα που τα περιέχει όλα μαζί. Εκτός αυτού, κατασκευάζοντας όλο και μεγαλύτερα δίκτυα απαιτείται ένας μεγάλος αριθμός βαρών (της τάξης εκατομμυρίων ή ακόμα και δισεκατομμυρίων). Τα δύο αυτά μαζί μας οδηγούν σε μεγάλες απαιτήσεις σε υπολογιστική μνήμη, η οποία δεν είναι εφικτή. Για αυτό αναπτύχθηκε η μέθοδος εκπαίδευσης σε παρτίδες (batches). Μία παρτίδα (batch) είναι ένα μικρό κομμάτι του συνόλου εκπαίδευσης, το οποίο χρησιμοποιείται για τον υπολογισμό της συνάρτησης κόστους και για την μετέπειτα διαδικασία ανανέωσης των βαρών.

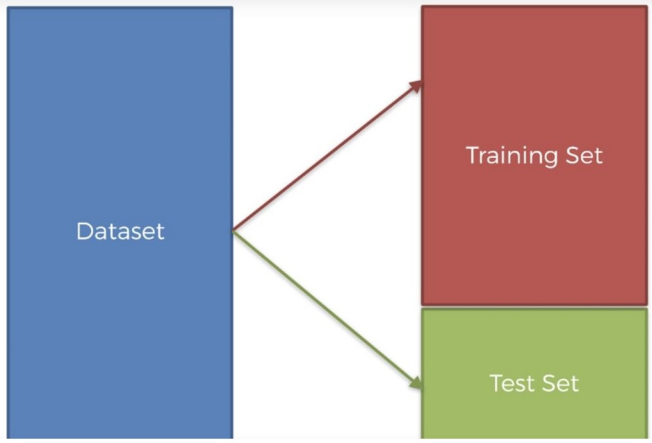
Χρησιμοποιώντας παρτίδες η διαδικασία της εκπαίδευσης αλλάζει λίγο. Μία εποχή τώρα ορίζεται όταν το μοντέλο ΄δει' όλα τα δείγματα του συνόλου δεδομένων μία φορά. Αυτό σημαίνει ότι μία εποχή αποτελείται από τόσες επαναλήψεις όσες και οι παρτίδες που είναι χωρισμένο το σύνολο εκπαίδευσης. Το πλήθος των παρτίδων καθορίζεται από το μέγεθος της κάθε παρτίδας (batch size), το οποίο είναι μία σημαντική υπερπαράμετρος στη διαδικασία της εκπαίδευσης. Για παράδειγμα, έστω θέτουμε μέγεθος παρτίδας σε 128 δείγματα και έχουμε ένα σύνολο δεδομένων με 12800 δείγματα, τότε σε κάθε εποχή χωρίζουμε τα δεδομένα μας σε 100 παρτίδες. Τέλος, η εκπαίδευση σε παρτίδες αλλάζει και τον αλγόριθμο της Καθόδου Κλίσης και μετασχηματίζεται σε αλγόριθμο Στοχαστικής Καθόδου Κλίσης, λόγω της τυχαιότητας που προσδίδει η επιλογή ενός υποσυνόλου των δεδομένων. Ο αλγόριθμος Στοχαστικής Καθόδου Κλίσης, καθώς και οι παραλλαγές του θα αναπτυχθούν στο Κεφάλαιο 5. Εκτός αυτού, η εκπαίδευση σε παρτίδες δίνει πάτημα για ανάπτυξη αλγόριθμων επιλογής της κάθε παρτίδας με συγκεκριμένο σκοπό. Τέτοιοι αλγόριθμοι θα αναφερθούν και θα αναλυθούν στο Κεφάλαιο 3.

2.4 Υπερπροσαρμογή και Υποπροσαρμογή

Στην Μηχανική Μάθηση και, κυρίως, στην επιβλεπόμενη μάθηση είναι πολύ σημαντικό η καλή αξιολόγηση των μοντέλων που εκπαιδεύουμε. Διάφορες μετρικές χρησιμοποιούνται για αυτόν τον σκοπό, κάθε μία από τις οποίες εξειδικεύεται σε κάποιο συγκεκριμένο έργο (π.χ. παλινδρόμηση, ταξινόμηση). Όμως, η αξιολόγηση ενός μοντέλου δεν είναι απλή υπόθεση. Ενώ εκπαιδεύουμε ένα δίκτυο είναι αρκετά εύκολο να μετρήσουμε πόσο καλά τα πάει στα δεδομένα εκπαίδευσης, αλλά αυτό δεν δείχνει την συνολική εικόνα της επίδοσης του μοντέλου μας. Γενικά, σε όλα τα μοντέλα Μηχανικής Μάθησης ο τελικός στόχος τους είναι να μπορούν να επιλύσουν το έργο για οποιοδήποτε σύνολο δεδομένων τους δοθεί. Αυτό είναι λογικό, επειδή χρειαζόμαστε μοντέλα, που να είναι αξιόπιστα και ασφαλή για οποιαδήποτε εφαρμογή εκπαιδευτούν. Συνεπώς, ο στόχος της αξιολόγησης ενός δικτύου μετακινείται στον να έχει καλές επιδόσεις και σε δεδομένα που δεν έχει ξαναδεί. Με άλλα λόγια επιθυμούμε τα δίκτυα μας να έχουν καλές μετρικές σε δεδομένα, στα οποία δεν έχει εκπαιδευτεί το νευρωνικό. Η

ικανότητα ενός νευρωνικού δικτύου να αποδίδει εξίσου καλά σε άγνωστα δεδομένα όσο και στα γνωστά ονομάζεται γενίκευση (generalization).

Για να αξιολογήσουμε καλύτερα ένα μοντέλο υπάρχουν διάφορες μέθοδοι που μπορούν να εφαρμοστούν. Η πιο κοινή είναι ο χωρισμός του συνόλου των δεδομένων μας σε 2 μέρη, στο σύνολο εκπαίδευσης (train set) και στο σύνολο ελέγχου ή δοκιμής (test set). Το σύνολο εκπαίδευσης θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και το σύνολο ελέγχου θα παραμείνει άγνωστο και θα χρησιμοποιηθεί στο τέλος για την αξιολόγηση. Ένα παράδειγμα διαχωρισμού φαίνεται στο Σχήμα 2.6.



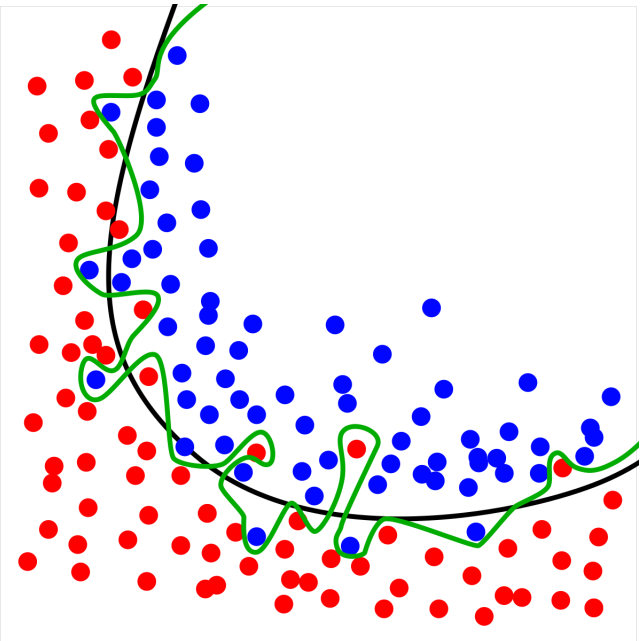
Σχήμα 2.6: Διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου.

Εκπαιδεύοντας πολλά διαφορετικά μοντέλα μπορούμε να λάβουμε πολλές διαφορετικές μετρικές σφάλματος για το σύνολο εκπαίδευσης και το σύνολο ελέγχου αντίστοιχα. Μέσω αυτών των μετρικών μπορούμε να αναγνωρίσουμε δύο είδη σφαλμάτων που μπορεί να υποπέσει ένα μοντέλο μας. Αυτά τα είδη είναι η υπερπροσαρμογή και η υποπροσαρμογή.

2.4.1 Υπερπροσαρμογή

Η υπερπροσαρμογή (overfitting) είναι ένα από τα πιο συνήθη σφάλματα εκπαίδευσης ενός δικτύου. Το πρόβλημα ξεκινά όταν το μοντέλο μαθαίνει πάρα πολύ καλά τα δεδομένα εκπαίδευσης σε βαθμό που κάνει απειροελάχιστα σφάλματα σε αυτά. Αυτή η συμπεριφορά μπορεί να οδηγήσει σε φαινόμενα που ενώ το νευρωνικό παρουσιάζει εξαιρετικές μετρικές στο σύνολο εκπαίδευσης, η απόδοσή του στο σύνολο ελέγχου είναι παραδόξως πολύ χειρότερη. Αυτό συμβαίνει επειδή το μοντέλο μαθαίνοντας ¨απέξω' τα δείγματα εκπαίδευσης, μαθαίνει παράλληλα και τον θόρυβο που εμπεριέχουν. Τέτοιος θόρυβος μπορεί να είναι χαρακτηριστικά των δειγμάτων που φαίνεται να έχουν μεγάλη συσχέτιση με το πρόβλημα ενώ να μην έχουν. Για παράδειγμα, έστω έχουμε ένα σύνολο δεδομένων με εικόνες και προσπαθούμε να αναγνωρίσουμε αν ένα αντικείμενο Α υπάρχει στην εικόνα. Αν όλες οι εικόνες έχουν το ίδιο φόντο (π.χ. ένα μπλε φόντο), τότε το μοντέλο μας κινδυνεύει να συνδυάσει το χρώμα του φόντου με την ύπαρξη του αντικειμένου Α. Αυτό ονομάζεται εκμάθηση του θορύβου και είναι μία αιτία

της υπερπροσαρμογής.



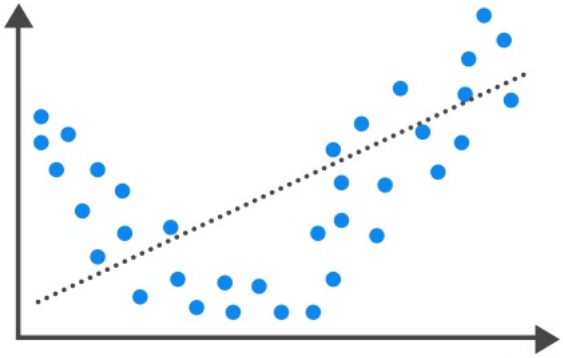
Σχήμα 2.7: Παράδειγμα υπερπροσαρμογής σε ένα πρόβλημα ταξινόμησης 2 κλάσεων. Η μαύρη γραμμή αντιπροσωπεύει ένα ιδανικό μοντέλο και η πράσινη ένα μοντέλο που πάσχει από υπερπροσαρμογή.

Μία άλλη αιτία υπερπροσαρμογής είναι η χωρητικότητα (capacity) του μοντέλου. Η χωρητικότητα εκφράζει την δύναμη του μοντέλου να μάθει περισσότερα δείγματα και περισσότερες και πιο δύσκολες συσχετίσεις. Επομένως, μοντέλα με υψηλή χωρητικότητα έχουν μεγαλύτερη πιθανότητα για υπερπροσαρμογή αφού είναι πιο εύκολο να μάθουν το είδος του θορύβου που δεν είναι επιθυμητός. Η έννοια της υπερπροσαρμογής συνδέεται άμεσα με την έννοια της μεταβλητότητας ενός μοντέλου. Η μεταβλητότητα ενός μοντέλου μετρείται από την διακύμανση των προβλέψεων. Υψηλή διακύμανση και μεταβλητότητα είναι συνώνυμο με υπερπροσαρμογή. Ένα παράδειγμα υπερπροσαρμογής μπορούμε να δούμε στο Σχήμα 2.7. Η πράσινη γραμμή απεικονίζει ένα μοντέλο με υπερπροσαρμογή. Παρατηρούμε ότι αν και πετυχαίνει σωστά όλα τα δεδομένα εκπαίδευσης, το όριο των κλάσεων που δημιουργεί θα παρουσιάσει προβλήματα στην ταξινόμηση άγνωστων δειγμάτων. Η μαύρη γραμμή αντιπροσωπεύει ένα ιδανικό μοντέλο, που παρόλο το ελάχιστα μεγαλύτερο σφάλμα εκπαίδευσης καταφέρνει να διαχωρίσει καλύτερα τις κλάσεις.

2.4.2 Υποπροσαρμογή

Η υποπροσαρμογή είναι κατά μία έννοια το αντίθετο πρόβλημα της υπερπροσαρμογής. Ένα μοντέλο παρουσιάζει υποπροσαρμογή, όταν δεν είναι ικανό να μάθει καλά το έργο για το οποίο εκπαιδεύεται. Συνήθως, το μοντέλο αδυνατεί να κατανοήσει τις σύνθετες σχέσεις των

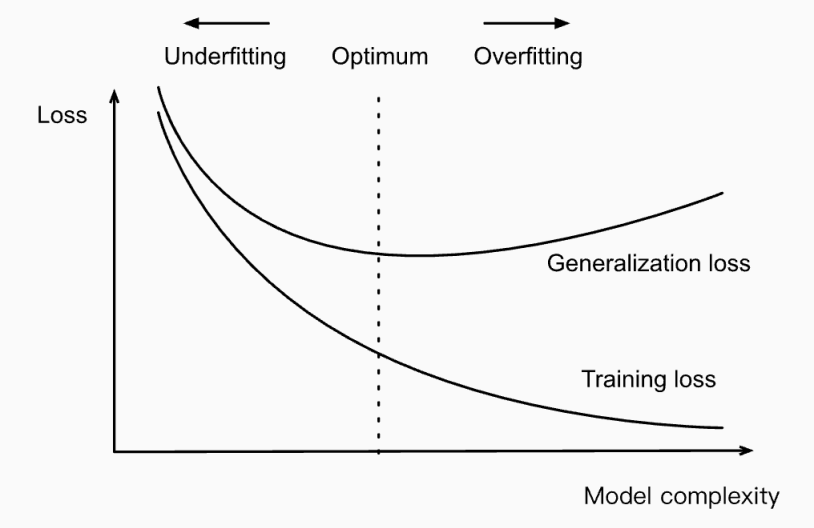
δεδομένων και για αυτό παρουσιάζει μεγάλο σφάλμα εκπαίδευσης. Μοντέλα με χαμηλή χωρητικότητα είναι επιρρεπή σε σφάλμα υποπροσαρμογής. Με την αύξηση της χωρητικότητας, αυξάνουμε την ικανότητα του μοντέλου να μαθαίνει περισσότερο πλήθος δεδομένων αλλά πιο περίπλοκες σχέσεις που τα διέπουν. Στο Σχήμα 2.8 φαίνεται ένα παράδειγμα υποπροσαρμογής σε ένα απλό πρόβλημα παλινδρόμησης. Η διακεκομμένη γραμμή είναι οι προβλέψεις του μοντέλου, οι οποίες υποδηλώνουν ένα απλοϊκό μοντέλο με μεγάλο σφάλμα εκπαίδευσης. Η έννοια της υποπροσαρμογής είναι συνώνυμο της υψηλής πόλωσης των προβλέψεων.



Σχήμα 2.8: Παράδειγμα υποπροσαρμογής σε ένα πρόβλημα παλινδρόμησης. Η πρόβλεψη του μοντέλου υποδεικνύεται με την διακεκομμένη γραμμή έχοντας ως αποτέλεσμα μεγάλο σφάλμα εκπαίδευσης.

2.4.3 Δίλημμα

Ένα από τα προβλήματα που καλούμαστε να λύσουμε είναι να αναγνωρίσουμε αυτά τα δύο λάθη και να τα καταπολεμήσουμε όσο είναι δυνατό. Όμως, υπάρχει μία ιδιαιτερότητα που χαρακτηρίζει τη σχέση μεταξύ αυτών των δύο προβλημάτων. Η σχέση αυτή μπορεί να αναπαρασταθεί καλύτερα από την σχέση μεταξύ των εννοιών πόλωση και μεταβλητότητα, που προαναφέρθηκαν. Η σχέση αυτή μπορεί να περιγραφεί ως ένα δίλημμα. Το δίλημμα πόλωσης-μεταβλητότητας (bias-variance tradeoff) υποδεικνύει ότι ένα μοντέλο προσπαθώντας να βελτιώσει το ένα πρόβλημα από τα δύο, χειροτερεύει την απόδοσή του στο άλλο. Για παράδειγμα, όταν προσπαθούμε να βελτιώσουμε το πρόβλημα της υποπροσαρμογής ή υψηλής πόλωσης, υπάρχει κίνδυνος να εκπαιδεύσουμε το μοντέλο σε βαθμό που να εμφανίζει θέματα υπερπροσαρμογής. Αυτό το παράδειγμα είναι συχνό σε περιπτώσεις που αυξάνουμε την χωρητικότητα ενός μοντέλου με σκοπό να μάθει περισσότερες αναπαραστάσεις δεδομένων αλλά, τελικά, παρατηρούμε χειρότερη απόδοση στο σύνολο ελέγχου. Ωστόσο, αυτό το δίλημμα δεν είναι απόλυτα αυστηρό, δηλαδή μπορεί να υπάρξει μοντέλο με χαμηλή πόλωση και χαμηλή μεταβλητότητα (μικρή υπερπροσαρμογή και υποπροσαρμογή) αλλά είναι αρκετά δύσκολο να επιτευχθεί. Στο Σχήμα 2.9 μπορούμε να δούμε ένα απλό παράδειγμα του διλήμματος. Με την αύξηση της πολυπλοκότητας του μοντέλου μειώνεται το σφάλμα εκπαίδευσης αλλά η γενίκευση του μοντέλου χειροτερεύει.



Σχήμα 2.9: Απεικόνιση της υπερπροσαρμογής και υποπροσαρμογής ενός μοντέλου σε συνάρτηση με την πολυπλοκότητα αυτού.

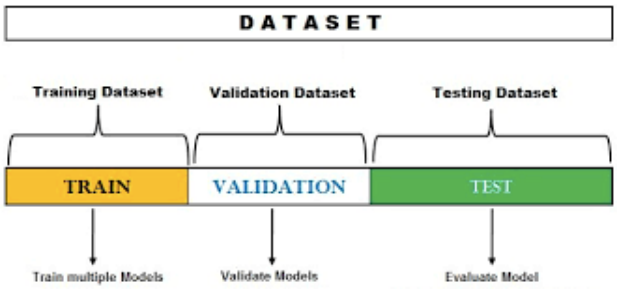
2.5 Αποτελεσματικές μέθοδοι αξιολόγησης

Χωρίζοντας το σύνολο δεδομένων στα δύο, σε σύνολο εκπαίδευσης και ελέγχου, μπορούμε να διακρίνουμε καλύτερα αν το μοντέλο μας κάνει κάποιο από τα παραπάνω σφάλματα. Επομένως, ο στόχος της εκπαίδευσης γίνεται διττός, μειωθούν τα σφάλματα εκπαίδευσης και τα σφάλματα ελέγχου. Όμως, αυτό κρύβει κάποιους κινδύνους. Το σύνολο ελέγχου που διαθέτουμε είναι πεπερασμένο και εμπεριέχει και αυτό διάφορα στοιχεία θορύβου παρόμοια με το σύνολο εκπαίδευσης. Προσπαθώντας να εκπαιδεύσουμε ένα δίκτυο, ώστε να πετυχαίνει την καλύτερη δυνατή απόδοση στο σύνολο ελέγχου, κινδυνεύουμε να υποπέσουμε σε σφάλμα υπερπροσαρμογής στο σύνολο ελέγχου. Για παράδειγμα, αυτό μπορεί να συμβεί όταν ψάχνουμε τις καλύτερες υπερπαραμέτρους και επιλέγουμε αυτές που έχουν καλύτερη απόδοση στο σύνολο ελέγχου. Το πρόβλημα σε αυτό είναι ότι έμμεσα παίρνουμε πληροφορίες από το σύνολο ελέγχου που υποτίθεται είναι άγνωστο. Η αντικειμενικότητα της επίδοσης του μοντέλου μας μειώνεται αρκετά με αυτόν τον τρόπο. Παρακάτω θα αναφέρουμε δύο τεχνικές αξιολόγησης, με τις οποίες καταφέρνουμε αντικειμενικότερη εκτίμηση επίδοσης ενός μοντέλου. Αυτές οι τεχνικές είναι η στρατηγική συγκράτησης και η διασταυρωμένη επικύρωση.

2.5.1 Στρατηγική Συγκράτησης

Η Στρατηγική Συγκράτησης (Hold-out Strategy) βασίζεται στον διαφορετικό διαχωρισμό των δεδομένων. Χωρίζει τα δεδομένα σε τρία μέρη (σε αντίθεση με πριν που ήταν 2), το σύνολο εκπαίδευσης, το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου. Προσθέτει, δηλαδή, ένα ακόμα υποσύνολο, βάσει του οποίου θα επιλέγουμε τις καλύτερες υπερπαραμέτρους. Η διαδικασία που ακολουθείται είναι η εξής: Πρώτα εκπαιδεύεται το μοντέλο στο σύνολο εκ-

παίδευσης, έπειτα βλέπουμε από το σύνολο επικύρωσης μία πρώτη αξιολόγηση σε άγνωστα δεδομένα και κάνουμε τις απαραίτητες αλλαγές στις υπερπαραμέτρους. Όταν βελτιστοποιηθεί η επίδοση στο σύνολο επικύρωσης, τεστάρουμε το μοντέλο στο σύνολο ελέγχου για την τελική αξιολόγηση. Με αυτόν τον τρόπο η τελευταία αξιολόγηση παραμένει τελείως αντικειμενική αφού το σύνολο ελέγχου παρέμεινε άγνωστο. Στο Σχήμα 2.10 παρατηρούμε τον διαχωρισμό του συνόλου δεδομένων στα τρία αυτά μέρη.



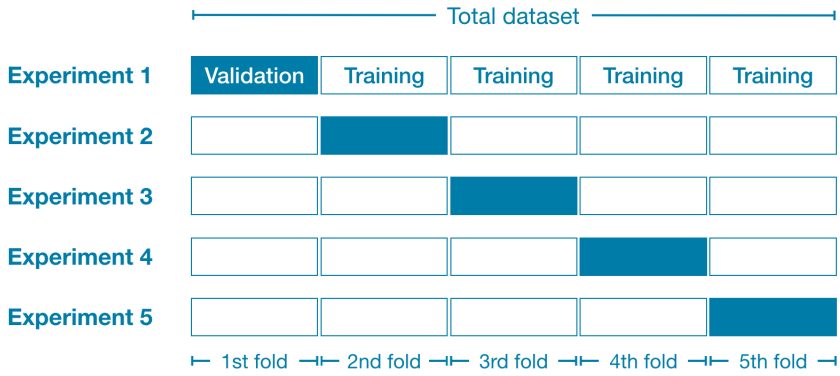
Σχήμα 2.10: Διαχωρισμός των δεδομένων σε τρία μέρη, σύνολο εκπαίδευσης, επικύρωσης και ελέγχου, με σκοπό την καλύτερη εκπαίδευση και αξιολόγηση των μοντέλων.

2.5.2 Διασταυρούμενη Επικύρωση

Μία άλλη χρήσιμη τεχνική για την αποτελεσματικότερη αξιολόγηση μοντέλων είναι η Διασταυρούμενη Επικύρωση. Προσπαθεί να ακολουθήσει μία διαφορετική προσέγγιση και να εξαλείψει τα προβλήματα της Στρατηγικής Συγκράτησης. Θα αναφέρουμε δύο από τα κύρια αρνητικά που επηρεάζουν την Στρατηγική Συγκράτησης. Πρώτο είναι ότι ο διαχωρισμός του συνόλου δεδομένων σε τρία μέρη μειώνει κατά πολύ τα δεδομένα εκπαίδευσης, τα οποία είναι αναγκαία σε μοντέλα σαν τα Νευρωνικά Δίκτυα. Δεύτερο πρόβλημα που δημιουργείται είναι ο παράγοντας της τυχαιότητας. Ο τυχαίος διαχωρισμός των δεδομένων μπορεί να οδηγήσει σε ανεπιθύμητα αποτελέσματα. Για παράδειγμα να χωριστούν τα δεδομένα με τέτοιο τρόπο ώστε η κατανομή του συνόλου εκπαίδευσης ή ελέγχου να μην είναι αντιπροσωπευτική και να χειροτερέψει την αντικειμενικότητα της αξιολόγησης.

Με βάση αυτά τα προβλήματα αναπτύχθηκε η Διασταυρούμενη Επικύρωση (Cross-validation). Η μέθοδος αυτή χωρίζει τα δεδομένα σε k ισομεγέθη μέρη και τρέχει k διαφορετικά πειράματα. Σε κάθε πείραμα επιλέγονται σαν σύνολο εκπαίδευσης $k-1$ από αυτά τα μέρη και ένα σαν σύνολο επικύρωσης, με τέτοιο τρόπο ώστε κάθε υποσύνολο να οριστεί σαν σύνολο επικύρωσης μία φορά. Η διαδικασία γίνεται με κυκλικό τρόπο όπως μπορούμε να δούμε στο Σχήμα 2.11. Με αυτόν το τρόπο, ελέγχουμε το μοντέλο μας πολλές φορές σε διαφορετικά κομμάτια του συνόλου δεδομένων μας, παίρνοντας k διαφορετικές αξιολογήσεις. Ο μέσος όρος τους αποδεικνύεται μία πολύ έμπιστη μετρική για την δύναμη του μοντέλου.

Η τεχνική αυτή ονομάζεται και k-πτυχη Διασταυρούμενη Επικύρωση, ορίζοντας έτσι σε πόσα κομμάτια θα χωριστούν τα δεδομένα. Η μέθοδος αυτή είναι μία από τις κύριες μεθόδους



Σχήμα 2.11: Αναπαράσταση του διαχωρισμού των δεδομένων με βάση την μέθοδο της Διασταυρούμενης Επικύρωσης.

για την ακριβή εύρεση της απόδοσης ενός μοντέλου, αφού τα πολλαπλά πειράματα εξαλείφουν την τυχαιότητα ενός τυχαίου συνόλου ελέγχου. Ένα από τα λίγα προβλήματα που μπορούν να προκύψουν με αυτή τη μέθοδο είναι ο μεγάλος φόρτος εκπαίδευσης που δημιουργείται.

2.6 Συνελικτικά Νευρωνικά Δίκτυα

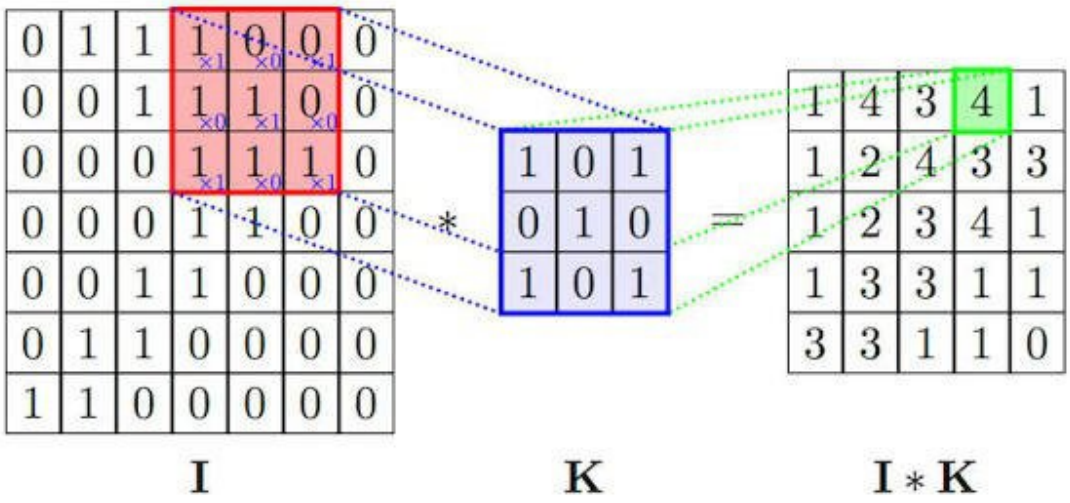
Ο τύπος νευρώνα που έχουμε αναλύσει μέχρι στιγμής είναι ο απλός νευρώνας εμπρόσθιας τροφοδότησης (Perceptron), από το οποίο προκύπτουν τα Πλήρως Συνδεδεμένα Νευρωνικά Δίκτυα (Fully Connected Neural Networks ή FC). Ονομάζονται πλήρως συνδεδεμένα επειδή όλοι οι νευρώνες ενός επιπέδου συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου. Η πλήρης συνδεσιμότητα παρέχει αρκετά θετικά στοιχεία, όπως η υψηλή χωρητικότητα του μοντέλου, αλλά παρόλα αυτά εμφανίζει και κάποια αρνητικά. Ένα από τα αρνητικά είναι η ευκολία στην υπερπροσαρμογή αλλά και ο μεγάλος αριθμός βαρών για εκπαίδευση που περιέχουν. Όσον αφορά τη δεύτερη παρατήρηση, ο αριθμός των παραμέτρων αυξάνεται ακόμα περισσότερο όταν αυξάνεται και η διάσταση της εισόδου. Αυτό το φαινόμενο είναι ιδιαίτερα εμφανές σε δεδομένα εικόνας, όπου ακόμα και μικρής διάστασης εικόνες αποτελούνται από μεγάλο αριθμό χαρακτηριστικών (ο αριθμός χαρακτηριστικών μιας εικόνας είναι ίσος με τον αριθμό των διαφορετικών εικονοστοιχείων (pixels) σε όλα τα κανάλια). Για παράδειγμα, μία εικόνα με διάσταση $224 \times 224 \times 3$ αποτελείται από 150528 χαρακτηριστικά πράγμα που θα δημιουργήσει μία πληθώρα από βάρη σε ένα πλήρως συνδεδεμένο δίκτυο. Πολύ μεγάλα δίκτυα από Πλήρη Συνδεδεμένα Δίκτυα είναι δύσκολο να εκπαιδευτούν σωστά, πέφτοντας συχνά σε σφάλματα υπερπροσαρμογής και παρουσιάζουν προβλήματα υπολογιστικής μνήμης. Εκτός, όμως, αυτού τα δεδομένα εικόνας είναι ένας ιδιαίτερος τύπος δεδομένων. Τα χαρακτηριστικά ενός δείγματος (μιας εικόνας) δεν είναι ανεξάρτητα μεταξύ τους. Υπάρχει μία χωρική εξάρτηση μεταξύ γειτονικών εικονοστοιχείων, κάτι το οποίο τα Πλήρη Συνδεδεμένα Δίκτυα δεν είναι δυνατά να εκμεταλλευτούν. Για αυτό τον λόγο αναπτύχθηκαν τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN) [23], τα οποία παρουσιάζουν μεγάλη βελτίωση σε

έργα που έχουν να κάνουν με δεδομένα εικόνας.

2.6.1 Συνελικτικά Επίπεδα

Τα Συνελικτικά Νευρωνικά Δίκτυα αποτελούνται από τα συνελικτικά επίπεδα. Αυτού του είδους τα επίπεδα χρησιμοποιούνται κυρίως σε εικόνες για δύο λόγους. Ο πρώτος λόγος είναι ότι κάνουν χρήση της τοπικής εξάρτησης των εικόνων και παρουσιάζουν μία τοπική συνδεσιμότητα. Η τοπική συνδεσιμότητα επιτυγχάνεται όταν ένας νευρώνας δέχεται ένα υποσύνολο των εισόδων αντί για όλο το διάνυσμα εισόδου. Το υποσύνολο της εισόδου που έχει μορφή εικόνας είναι μία μικρότερη εικόνα που αποτελείται από γειτονικά εικονοστοιχεία. Έτσι, με αυτόν τον τρόπο ο νευρώνας επεξεργάζεται κάθε φορά ένα μικρό κομμάτι της συνολικής εικόνας και προσπαθεί να εξάγει χρήσιμα χαρακτηριστικά από αυτό. Ο δεύτερος λόγος της δημοφιλούς χρήσης των CNN είναι ο διαμοιρασμός των βαρών. Σε έναν απλό νευρώνα δημιουργείται ένα βάρος για κάθε είσοδο. Με τον διαμοιρασμό των βαρών όλα τα βάρη αντιστοιχούν σε περισσότερες από μία εισόδους, ως αποτέλεσμα να χρειάζονται λιγότερα βάρη ανά νευρώνα.

Ας εξετάσουμε, όμως, πιο αναλυτικά την λειτουργία ενός συνελικτικού επιπέδου και πως εφαρμόζει τις δύο προηγούμενες ιδέες. Ας υποθέσουμε μία τρισδιάστατη είσοδο (υποδηλώνοντας δεδομένα εικόνας) συμβολίζοντας τη ως $x \in \mathbb{R}^{W \times H \times C}$ με W να είναι το πλάτος της εικόνας, H το ύψος και C ο αριθμός των καναλιών. Κάθε νευρώνας ενός συνελικτικού επιπέδου τελεί την πράξη της συνέλιξης μεταξύ της εικόνας της εισόδου και των βαρών (φίλτρα), τα οποία είναι μικρές εικόνες 2 διαστάσεων. Η πράξη της συνέλιξης συντελείται με μετακινούμενα εσωτερικά γινόμενα μεταξύ του φίλτρου και μιας ομοίου διάστασης περιοχής της εικόνας. Αυτό μπορεί να αναπαρασταθεί πιο κατανοητά στο Σχήμα 2.12, όπου I είναι η αρχική εικόνα, K είναι το φίλτρο και $I * K$ είναι παραγόμενος πίνακας της συνέλιξης.



Σχήμα 2.12: Η πράξη της συνέλιξης σε μία εικόνα διάστασης 7×7 με ένα φίλτρο διάστασης 3×3.

Κάθε συνελικτικό επίπεδο διαθέτει έναν αριθμό από διαφορετικά εκπαιδεύσιμα φίλτρα, τα οποία εφαρμόζονται σε όλα τα κανάλια της κάθε εικόνας ή σε όλους τους χάρτες χαρακτηριστικών δημιουργώντας νέες εικόνες (ή χάρτες αντίστοιχα). Η έξοδος του συνελικτικού δικτύου $z \in \mathbb{R}^{W' \times H' \times F}$ έχει ως έξοδο F εικόνες $W' \times H'$ διαστάσεων. Οι διαστάσεις καθορίζονται από το μέγεθος του φίλτρου, από το βήμα μετακίνησης της συνέλιξης και από την χρήση γεμίσματος zero-padding. Στο τέλος εφαρμόζεται μία συνάρτηση ενεργοποίησης, η οποία συνήθως είναι η ReLU.

2.6.2 Επίπεδα Συμψηφισμού

Στα CNN το κύριο επίπεδο που χρησιμοποιείται είναι τα συνελικτικά επίπεδα που προαναφέρθηκαν. Όμως, πολλές φορές χρησιμοποιείται και ένα άλλο είδος επιπέδου που λειτουργεί σε συνδυασμό με τα επίπεδα συνέλιξης, το επίπεδο συμψηφισμού (pooling layer). Το επίπεδο συμψηφισμού πραγματοποιεί μία υποδειγματοληψία στις εικόνες με σκοπό να μικρύνει την διάστασή τους. Αυτό βοηθάει στην μείωση της πολυπλοκότητας και, ως αποτέλεσμα, στην ευκολότερη εκπαίδευση του νευρωνικού. Ο τρόπος με τον οποίο εφαρμόζεται η υποδειγματοληψία είναι παρόμοιος με αυτόν του συνελικτικού επιπέδου. Υπάρχει ένα φίλτρο, το οποίο περνά πάνω από τις εικόνες και εφαρμόζει μία συνάρτηση συμψηφισμού. Συνηθισμένες συναρτήσεις συμψηφισμού είναι η συγκέντρωση μέσου (average pooling) ή η συγκέντρωση μεγίστου (max pooling). Στο Σχήμα 2.13 φαίνονται δύο παραδείγματα για ένα επίπεδο συμψηφισμού μεγίστου και μέσου, όπου σε ένα παράθυρο 2×2 επιλέγεται το εικονοστοιχείο με την μέγιστη ή την μέση τιμή αντίστοιχα.

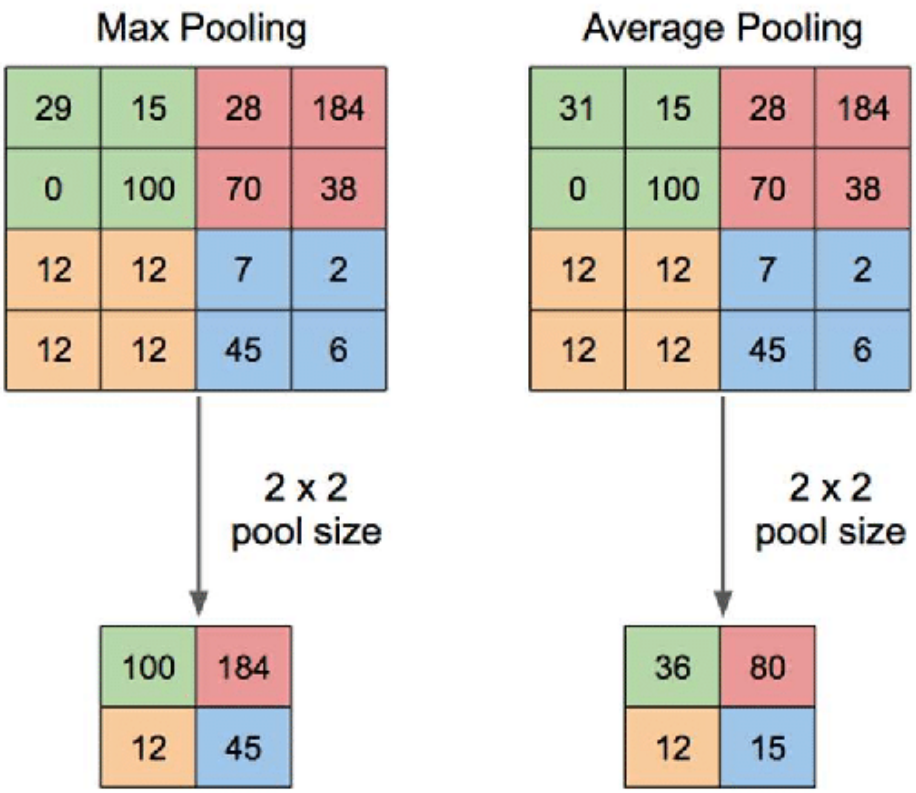
Το επίπεδο συμψηφισμού είναι χρήσιμο και στην καταπολέμηση της υπερπροσαρμογής. Η υποδειγματοληψία βοηθά το νευρωνικό να επηρεάζεται λιγότερο από μικρές αλλαγές στην είσοδο, έχοντας ως αποτέλεσμα το μοντέλο να μάθει να γενικεύει καλύτερα και να μαθαίνει πιο δύσκολα τον ανεπιθύμητο θόρυβο που οφείλεται για την υπερπροσαρμογή.

2.7 Χρήσιμα Βοηθητικά Επίπεδα και Τεχνικές

Τα κύρια δομικά στοιχεία ενός νευρωνικού δικτύου είναι τα επίπεδα των διαφορετικών ειδών νευρώνων καθώς και ο αλγόριθμος μάθησης. Όμως, δεν είναι τα μόνα που συμβάλλουν στην ομαλή εκπαίδευση και καλή επίδοση ενός δικτύου. Σε αυτό το κεφάλαιο θα αναφέρουμε κάποια βοηθητικά επίπεδα και τεχνικές, οι οποίες λειτουργούν ευεργετικά στη διαδικασία της μάθησης του μοντέλου.

2.7.1 Απόσυρση

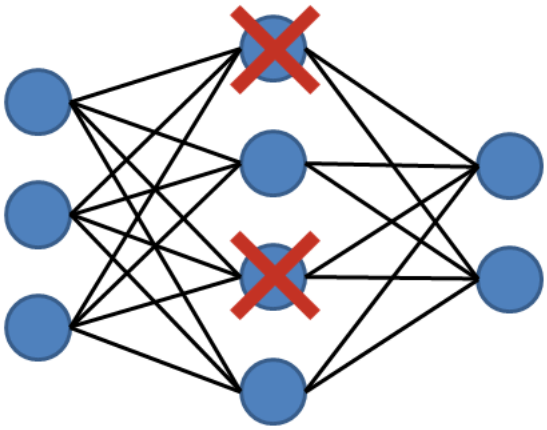
Η Απόσυρση (Dropout) [101] είναι μία τεχνική, η οποία έχει ως σκοπό την ελάττωση της υπερπροσαρμογής του δικτύου στα δεδομένα εκπαίδευσης. Εφαρμόζεται με την μορφή ενός επιπέδου, το οποίο προστίθεται μετά κάποιο πλήρες συνδεδεμένο επίπεδο ή κάποιο συ-



Σχήμα 2.13: Παράδειγμα της λειτουργίας του Επιπέδου Συμψηφισμού Μεγίστου και Μέσου αντίστοιχα.

νελικτικό επίπεδο. Η λειτουργία του βασίζεται στην στοχαστική αφαίρεση νευρώνων με βάση κάποια πιθανότητα. Για παράδειγμα, αν εφαρμόσουμε ένα επίπεδο Απόσυρσης με πιθανότητα απόσυρσης 50% μετά από ένα πλήρες συνδεδεμένο επίπεδο 10 νευρώνων, τότε στην φάση της εκπαίδευσης υπάρχει 50% πιθανότητα να μηδενιστεί η έξοδος καθενός από τους νευρώνες. Η έξοδος των νευρώνων που δεν αποσύρθηκαν ενισχύεται διαιρώντας τη με την πιθανότητα απόσυρσης. Η απόσυρση νευρώνων εκτελείται μόνο στη φάση της εκπαίδευσης, ενώ στη φάση του ελέγχου παραμένει ανενεργό σαν επίπεδο.

Επομένως, αυτό που καταφέρνει η Απόσυρση είναι να αντικαθιστά στην τύχη τις εξόδους των νευρώνων με μηδενικά. Η συμβολή της έχει δύο συνιστώσες. Η πρώτη συμβολή είναι ότι ο συγκεκριμένος νευρώνας δεν θα εκπαιδευτεί σε εκείνη την επανάληψη, το οποίο είναι μία μορφή δυσκολίας που θέτουμε κατά της υπερπροσαρμογής. Η δεύτερη συνιστώσα, και πιο σημαντική, είναι ότι τα επόμενα επίπεδα μαθαίνουν να επιλύουν το πρόβλημα με λιγότερες εισόδους. Αυτό το είδος πρόσθετου θορύβου είναι αυτό που δίνει στο νευρωνικό καλύτερη γενίκευση και απόδοση. Στο Σχήμα 2.14 μπορούμε να δούμε ένα παράδειγμα απόσυρσης.



Σχήμα 2.14: Νευρωνικό Δίκτυο με χρήση της Απόσυρσης στο δεύτερο επίπεδο.

2.7.2 Κανονικοποίηση Παρτίδας

Μία άλλη τεχνική που βοηθά την ομαλή εκπαίδευση των Νευρωνικών Δικτύων είναι η Κανονικοποίηση Παρτίδας (Batch Normalization) [43]. Αυτή η μέθοδος υλοποιείται σε μορφή επιπέδου και έχει ως σκοπό να κανονικοποιεί τις εξόδους του προηγούμενου επιπέδου. Η χρησιμότητα της κανονικοποίησης φαίνεται στο πρόβλημα της εσωτερικής μετατόπισης μεταβλητών (internal covariance shift), το οποίο αποτελεί ένα εμπόδιο στη μάθηση. Η εσωτερική μετατόπιση μεταβλητών συμβαίνει όταν με την πάροδο της εκπαίδευσης οι έξοδοι των επιπέδων παρουσιάζουν μία κατανομή, η οποία αλλάζει και είναι διαφορετική από επίπεδο σε επίπεδο. Αυτή η διαφορά δυσκολεύει τον αλγόριθμο μάθησης να ανανεώσει τα βάρη σε κάποιο καλό σημείο.

Η Κανονικοποίηση Παρτίδας εφαρμόζεται με δύο διαφορετικούς τρόπους ανάλογα αν βρισκόμαστε στη φάση της εκπαίδευσης ή στη φάση του ελέγχου. Στην φάση της εκπαίδευσης το επίπεδο αυτό βρίσκει τον μέσο όρο και την τυπική απόκλιση τις παρτίδας και αφαιρώντας τον και διαιρώντας τη αντίστοιχα μετασχηματίζει την είσοδο σε κανονική κατανομή. Για παράδειγμα ας υποθέσουμε ότι έχουμε μία είσοδο $x \in \mathbb{R}^{B \times F}$, όπου B το μέγεθος της παρτίδας και F η διάσταση των εξόδων του προηγούμενου επιπέδου. Μετά την κανονικοποίηση η έξοδος σταθμίζεται με τους όρους γ και β, οι οποίοι είναι εκπαιδεύσιμοι με τον αλγόριθμο καθόδου κλίσης σαν τα βάρη των νευρώνων. Στις παρακάτω εξισώσεις μπορούμε να δούμε την όλη διαδικασία.

$$\mu_B = \frac{1}{B} \sum_{i=1}^{B} x_i \tag{2.13}$$

$$\sigma_i^2 = \frac{1}{B}\sum_{i=1}^{B}(xi - \mu_B)^2 \tag{2.14}$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{2.15}$$

$$y_i = \gamma\hat{x}_i + \beta \tag{2.16}$$

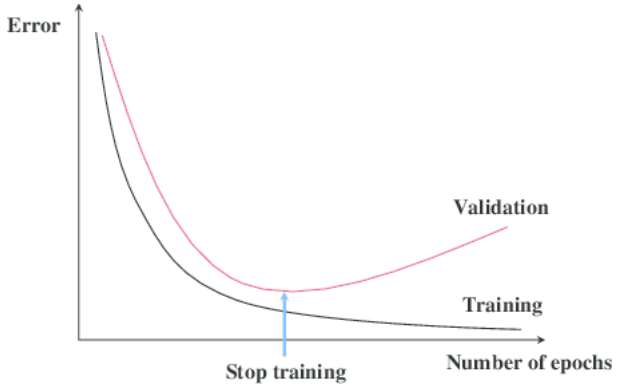
όπου y_i είναι η έξοδος του επιπέδου Κανονικοποίησης Παρτίδας. Αυτό γίνεται κατά την διάρκεια της εκπαίδευσης. Κατά την φάση του ελέγχου, η διαδικασία αλλάζει λίγο επειδή δεν μπορεί να εφαρμοστεί αυτούσια γιατί θα είναι ασταθής. Σε αυτή την περίπτωση υπολογίζονται δύο κινούμενοι μέσοι του μέσου όρου και της διακύμανσης κατά την διάρκεια της εκπαίδευσης και αποθηκεύονται. Όταν έρθει η φάση του ελέγχου τότε χρησιμοποιούνται αυτοί στη θέση των μ και σ των παραπάνω εξισώσεων.

$$E[x] = E_B[\mu_B] \tag{2.17}$$

$$Var[x] = \frac{m}{m-1}E_B[\sigma_B^2] \tag{2.18}$$

2.7.3 Πρόωρη Διακοπή Εκπαίδευσης

Μία πολύ σημαντική τεχνική για την αποτροπή της υπερπροσαρμογής είναι η πρόωρη διακοπή της εκπαίδευσης (early stopping). Η τεχνική αυτή προσπαθεί να σταματήσει την διαδικασία της εκπαίδευσης ενός νευρωνικού δικτύου προτού η γενίκευση του μοντέλου αρχίσει και χειροτερεύει. Όπως αναφέρθηκε και σε προηγούμενο υποκεφάλαιο, η υπερπροσαρμογή οφείλεται στο γεγονός ότι το μοντέλο εκτός από χρήσιμες συσχετίσεις μαθαίνει και τον άχρηστο θόρυβο, που υπάρχει σε όλα τα δεδομένα.



Σχήμα 2.15: Τεχνική της Πρόωρης Διακοπής Εκπαίδευσης με βάση το σύνολο επικύρωσης.

Στο Σχήμα 2.15 μπορούμε να δούμε ένα παράδειγμα εκπαίδευσης, όπου απεικονίζονται οι καμπύλες του σφάλματος εκπαίδευσης και του σφάλματος επικύρωσης. Παρατηρούμε ότι αν το νευρωνικό εκπαιδευτεί παραπάνω εποχές, η επίδοσή του στο σύνολο επικύρωσης αρχίζει και χειροτερεύει. Επομένως, μία στρατηγική πρόωρης διακοπής είναι να σταματήσουμε την εκπαίδευση στην εποχή που η επίδοση στο σύνολο επικύρωσης πάει να χειροτερέψει. Με αυτόν τον τρόπο διασφαλίζουμε ότι το νευρωνικό έχει μάθει τις σημαντικές πληροφορίες των δεδομένων αλλά δεν έχει προλάβει τα αποστηθίσει τον ανεπιθύμητο θόρυβο.

2.7.4 Κυρώσεις Ενημέρωσης

Η κυρώσεις ενημέρωσης (Regularization) είναι τεχνικές που εφαρμόζουν έναν περιορισμό στις ανανεώσεις των παραμέτρων με σκοπό την καταπολέμηση την υπερπροσαρμογής. Υλοποιείται σαν ένας επιπλέον όρος στην συνάρτηση κόστους. Οι πιο συνηθισμένες τεχνικές κυρώσεων είναι βασισμένες σε ποινές πάνω στην νόρμα των βαρών. Οι δύο πιο χρησιμοποιημένες είναι η L1 και L2 νόρμες. Στην παρακάτω συνάρτηση φαίνεται πως προστίθεται η κύρωση ενημέρωσης L2 σε μία συνάρτηση κόστους.

$$J^{reg}(y, \hat{y}) = J(y, \hat{y}) + \lambda \|W_t\|_2 \tag{2.19}$$

Ο όρος λ είναι η υπερπαράμετρος που ελέγχει την ένταση του περιορισμού. Η κύρωση αυτή ουσιαστικά ωθεί το νευρωνικό να βρει τα μικρότερα δυνατά βάρη που επιλύουν το πρόβλημα. Το μοντέλο έτσι οδηγείται σε βάρη κοντά στο 0, το οποίο έχει δείξει ότι είναι ωφέλιμο στο δίκτυο. Κατά τον ίδιο τρόπο και η L1 κύρωση ωθεί το δίκτυο σε μικρότερα βάρη περιορίζοντας την πολυπλοκότητά του.

2.7.5 Ενίσχυση Δεδομένων

Η ενίσχυση δεδομένων είναι μία εξαιρετικά δημοφιλής τεχνική για την καταπολέμηση της υπερπροσαρμογής. Χρησιμοποιεί το σύνολο εκπαίδευσης και δημιουργεί νέα δεδομένα, τα οποία προστίθενται στο σύνολο εκπαίδευσης για να τα μάθει. Αυτά τα νέα δεδομένα είναι παραπλήσια με τα πραγματικά του συνόλου δεδομένων αλλά έχουν κάποιο είδος θορύβου που τα διαφοροποιεί από τα αρχικά. Η ένταση του θορύβου είναι τόσο όσο ώστε να μην παραμορφώνεται σε μεγάλο βαθμό η ουσία του δείγματος παραμένοντας, έτσι, αληθοφανή. Εκπαιδεύοντας το μοντέλο μας με αυτά τα επιπρόσθετα δεδομένα καταφέρνουμε να κάνουμε την διαδικασία της εκπαίδευσης πιο δύσκολη, εμποδίζοντας με αυτόν τον τρόπο την υπερπροσαρμογή. Εκτός αυτού, μαθαίνοντας περισσότερα διαφορετικά δείγματα το δίκτυο είναι ικανό να μάθει και περισσότερες χρήσιμες σχέσεις των χαρακτηριστικών των δειγμάτων, οδηγώντας σε καλύτερη γενίκευση.

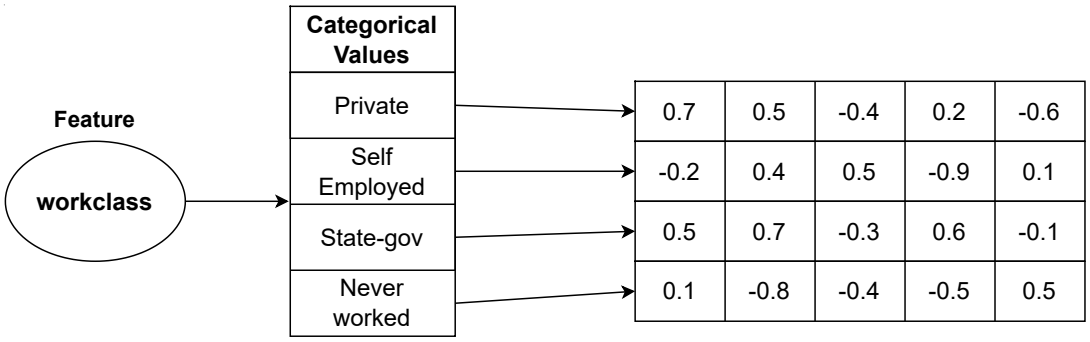
Η ενίσχυση δεδομένων παρατηρείται κυρίως σε δεδομένα εικόνας και αποτελεί μία αναγκαία τεχνική για να επιδιώξει κάποιος να κατασκευάσει ένα μοντέλο τελευταίας τεχνολογίας. Στο

Σχήμα 2.16: Παράδειγμα ενίσχυσης δεδομένων σε μία εικόνα που απεικονίζει έναν σκύλο.

Σχήμα 2.16 παρατηρούμε τις διαφορετικές εικόνες που μπορούν να παραχθούν από ενίσχυση δεδομένων. Στην περίπτωση των εικόνων συχνές τεχνικές για ενίσχυση είναι τυχαίοι μετασχηματισμοί, όπως καθρεπτισμός, προσθήκη θορύβου, αυξομείωση φωτεινότητας ή αντίθεσης. Τέτοιοι μετασχηματισμοί καταφέρνουν να δημιουργούν μία νέα εικόνα που όμως δεν αλλάζει το δείγμα ριζικά. Στο παράδειγμα του σχήματος βλέπουμε ότι υπάρχουν μετασχηματισμοί στην εικόνα αλλά η εικόνα παραμένει σκύλος.

2.7.6 Επίπεδο Διανυσματικής Αναπαράστασης

Μία πολύ σημαντική έννοια στον χώρο της Βαθιάς Μηχανικής Μάθησης είναι η διανυσματική αναπαράσταση (embedding) μεταβλητών. Η διανυσματική αναπαράσταση είναι ένας μετασχηματισμός μιας μεταβλητής ή ενός χαρακτηριστικού σε ένα συνεχές διάνυσμα. Ο σκοπός αυτού του μετασχηματισμού είναι διττός. Πρώτον, δεδομένα που δεν είναι αριθμητικά κάπως πρέπει να διαχειριστούν ώστε να είναι σε θέση να χρησιμοποιηθούν ως αριθμητική είσοδο του νευρωνικού δικτύου. Για παράδειγμα, δεδομένα κειμένου, κατηγορικές μεταβλητές ή χαρακτηριστικά που παριστάνουν αντικείμενα πρέπει με κάποιο τρόπο να μετατραπούν σε αριθμητικά δεδομένα για να είναι εφικτοί οι υπολογισμοί που υφίστανται μέσα στο δίκτυο. Δεύτερον, με την χρήση της διανυσματικής αναπαράστασης γίνεται δυνατή η επιλογή της διάστασης του μετασχηματισμού, το οποίο λύνει διάφορα προβλήματα. Σε εφαρμογές κειμένου, χρησιμοποιείται κατά κόρον ο μετασχηματισμός one-hot δημιουργώντας αραιούς πίνακες μεγάλης διάστασης που χαρακτηρίζονται δύσκολοι στην εκπαίδευση ενός δικτύου. Με την διανυσματική αναπαράσταση μετατρέπουμε τέτοιους πίνακες σε επιθυμητές συμπιεσμένες διαστάσεις κατάλληλες για ευκολότερη εκπαίδευση. Ομοίως, σε περιπτώσεις κατηγορικών μεταβλητών με μεγάλο αριθμό κατηγοριών παρατηρείται το ίδιο πρόβλημα. Επίσης, η μετατροπή αυτή γίνεται με τρόπο



Σχήμα 2.17: Παράδειγμα της διαδικασίας ενός Επιπέδου Διανυσματικής Αναπαράστασης

ώστε ομοιότητες, σχέσεις και πληροφορίες των δεδομένων να διατηρούνται και να αναπαριστώνται με πιο αναγνωρίσιμο τρόπο. Με άλλα λόγια λέξεις παρόμοιας σημασίας ή κατηγορίες παρόμοιας έννοιας θα αναπαριστώνται σε διανύσματα κοντινά μεταξύ τους στο χώρο.

Ως συνέχεια των παραπάνω, στα Νευρωνικά Δίκτυα έχουν αναπτυχθεί τα τελευταία χρόνια τα Επίπεδα Διανυσματική Αναπαράστασης [72]. Αυτά τα επίπεδα χρησιμοποιούν πίνακες αναζήτησης, που αντιστοιχίζουν τα κατηγορικά δεδομένα σε σταθερού μεγέθους εκπαιδεύσιμα διανύσματα. Συνήθως, αυτά τα διανύσματα αρχικοποιούνται με τυχαίες μεταβλητές από διάφορες κατανομές, ομοίως με βάρη άλλων επιπέδων, όπως συνελικτικών ή πρόσθιας τροφοδότησης. Επίσης, εκπαιδεύονται με τον ίδιο τρόπο με άλλα βάρη, συμβάλλοντας στην ελαχιστοποίηση της συνάρτηση κόστους. Στις σύγχρονες εφαρμογές Βαθιάς Μηχανικής Μάθησης τα Επίπεδα Διανυσματικής Αναπαράστασης χρησιμοποιούνται όλο και πιο συχνά και είναι ένα αναπόσπαστο κομμάτι των αρχιτεκτονικών. Εκτός από εφαρμογές κειμένου, τα συγκεκριμένα επίπεδα χρησιμοποιούνται σε εφαρμογές με δεδομένα πινάκων, σε συστήματα συστάσεων, ακόμα και σε δεδομένα εικόνας. Στο Σχήμα 2.17 παρουσιάζεται η διαδικασία στο παράδειγμα του συνόλου δεδομένων Adult [7]. Συγκεκριμένα, βλέπουμε το χαρακτηριστικό 'workclass' με κάποιες από τις δυνατές κατηγορικές τιμές του (π.χ. Private, Self Employed, State-gov, Never worked), καθώς και τα διανύσματα μεγέθους 5 που αντιστοιχούνται στο καθένα.

Κεφάλαιο 3

Δυναμική Επιλογή Παρτίδας

Στο κεφάλαιο αυτό θα ασχοληθούμε με το πεδίο της Δυναμικής Επιλογής Παρτίδας στο πλαίσιο της Βαθιάς Μηχανικής Μάθησης. Αρχικά, θα αναλύσουμε την λειτουργία και την χρησιμότητά της, καθώς και διάφορες τεχνικές και μεθοδολογίες της βιβλιογραφίας. Έπειτα, θα αναπτυχθεί μία καινούργια μεθοδολογία, με όνομα Επιλογή Παρτίδας με Μεροληπτική Δειγματοληψία, αλλά και διάφορες παραλλαγές του βασικού αλγορίθμου.

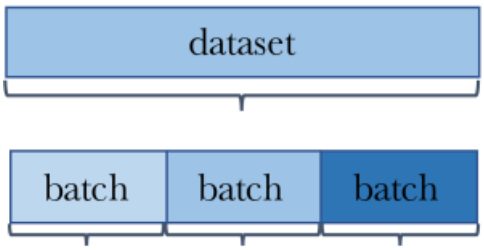
3.1 Εισαγωγή

Σε προηγούμενο κεφάλαιο αναφέρθηκε η έννοια της συνάρτησης κόστους και πως αυτή χρησιμοποιείται για την εκπαίδευση του δικτύου. Οι τύποι 2.6 και 2.7 αφορούσαν την εφαρμογή μίας συνάρτησης κόστους σε όλα τα δεδομένα του συνόλου εκπαίδευσης. Μπορούμε να γράψουμε έναν γενικευμένο τύπο για όλες τις συναρτήσεις κόστους ενός νευρωνικού με τον ακόλουθο τρόπο:

$$\mathbf{J}(X, w) = \frac{1}{N} \sum_{i=1}^{N} J(x_i \; w) \tag{3.1}$$

όπου ορίζει την ολική συνάρτηση κόστους **J** ως ένα άθροισμα των επιμέρους συναρτήσεων κόστους που εφαρμόζονται σε κάθε δείγμα των δεδομένων για συγκεκριμένα βάρη του δικτύου. Ο στόχος της διαδικασίας βελτιστοποίησης είναι να ελαχιστοποιήσει την ολική συνάρτηση κόστους. Αναφέρθηκε, αργότερα, ότι η εκπαίδευση ενός νευρωνικού δικτύου με την χρήση ενός ολόκληρου συνόλου δεδομένων δεν είναι εφικτή σε δεδομένα μεγάλης κλίμακας, αλλά και σε δίκτυα υψηλής χωρητικότητας. Για αυτό εφαρμόζεται η εκπαίδευση σε παρτίδες. Παρτίδα είναι ένα μικρό υποσύνολο των δεδομένων, το οποίο χρησιμοποιείται για μία επανάληψη του αλγορίθμου μάθησης.

Για την επιλογή ενός υποσυνόλου των δεδομένων εφαρμόζεται κάποια μορφή δειγματοληψίας. Λαμβάνοντας υπόψη αυτό, ο σκοπός της βελτιστοποίησης μετασχηματίζεται ώστε



Σχήμα 3.1: Διαχωρισμός του συνόλου εκπαίδευσης σε παρτίδες.

να εφαρμόζεται σε υποσύνολο των δεδομένων. Ο στόχος της βελτιστοποίησης μπορεί να εκφραστεί τώρα με την ακόλουθη σχέση:

$$\min_{w\in\mathbb{R}^n} \mathbb{E}_{x\sim P}\left[\mathbf{J}(X,w)\right] = \min_{w\in\mathbb{R}^n} \int_X \mathbf{J}(X,w)dP(x) \tag{3.2}$$

Επομένως, η διαδικασία της μάθησης μπορεί να χωριστεί σε 2 βήματα:

- Δειγματοληψία μίας παρτίδας μεγέθους $B \subset X$ βασισμένη σε μία κατανομή P.
- Εφαρμόζουμε την ανανέωση των βαρών με τον αντίστοιχο βελτιστοποιητή χρησιμοποιώντας την παράγωγο που υπολογίσαμε από την παρτίδα.

Η στρατηγική της δειγματοληψίας αποτελεί ένα αρκετά σημαντικό βήμα στην εκπαίδευση ενός δικτύου. Επηρεάζει την σύγκλιση και την επίδοση του μοντέλου [66]. Αρκετές προσεγγίσεις έχουν αναπτυχθεί τα τελευταία χρόνια, που υλοποιούν διάφορες τεχνικές δειγματοληψίας με διαφορετικά οφέλη η καθεμία. Στο επόμενο υποκεφάλαιο θα αναφέρουμε μερικές από αυτές τις τεχνικές.

3.2 Βιβλιογραφία

Η πιο γνωστή στρατηγική δειγματοληψίας είναι να επιλέγεται το κάθε δείγμα του συνόλου εκπαίδευσης σύμφωνα με την ομοιόμορφη κατανομή. Αυτό μπορεί να επιτευχθεί με 2 διαφορετικούς τρόπους. Ο πρώτος τρόπος επιλέγει το κάθε δείγμα με την ίδια πιθανότητα ($p_i = \frac{1}{|X|}$), ενώ ο δεύτερος απλά επιλέγει κάθε δείγμα μία φορά μόνο κάθε εποχή. Αυτές τις δύο μεθόδους θα τις αναφέρουμε ως SGD Uniform (SGD-Uni) και SGD-Scan αντίστοιχα, ακολουθώντας την. Η διαφορά τους βασίζεται ότι στην πρώτη περίπτωση το κάθε δείγμα εισέρχεται στην εκπαίδευση κατά μέσο όρο μία φορά, όμως στην πράξη το πόσες φορές μπει κυμαίνεται από 0 μέχρι και πάνω από 2. Η δεύτερη περίπτωση εγγυάται πως όλα τα δείγματα θα επιλεχθούν ακριβώς μία φορά και, έτσι, το δίκτυο θα τα δει σίγουρα στο πέρας μιας εποχής. Συνήθως, παρατηρείται πως η δεύτερη μέθοδος είναι πιο σταθερή και γι΄ αυτό χρησιμοποιείται περισσότερο στην πράξη.

Εμπνευσμένη από την ανθρώπινη συμπεριφορά, η Μάθηση Διδακτικού Προγράμματος (Curriculum Learning - CL) [8] είναι μία μέθοδος, η οποία έχει βελτιώσει την εκπαίδευση των νευρωνικών δικτύων. Η κεντρική ιδέα της βασίζεται στο ότι στις αρχές της εκπαίδευσης επιλέγονται ¨εύκολα' δείγματα και σιγά σιγά εισάγονται όλο και δυσκολότερα. Μιμείται, δηλαδή, τις συνήθειες που θα εμφάνιζε ένας άνθρωπος, ο οποίος μαθαίνει κάτι καινούργιο. Παρά την επιτυχία της μεθόδου, παρατηρήθηκαν κάποια εμπόδια και δυσκολίες στην χρήση της. Πρώτα, δεν είναι γνωστό εξ΄ αρχής ποια είναι τα δύσκολα δείγματα για το δίκτυο και ποια τα εύκολα. Δεύτερο εμπόδιο που εμφανίστηκε αφορά την ταχύτητα της εκπαίδευσης. Δίνοντας έμφαση σε εύκολα δείγματα, συνεπώς και μικρότερου σφάλματος, υπολογίζουμε και μικρότερες παραγώγους. Αυτό οδηγεί σε μικρές ανανεώσεις βαρών και στην πιο αργή εκπαίδευση του δικτύου. Αργότερα, αναπτύχθηκε η Αυτοματοποιημένη Μάθηση Διδακτικού Προγράμματος (Automated Curriculum Learning) [33], η οποία με τη βοήθειας μιας πολιτικής θα προσπαθούσε να μεγιστοποιήσει το ρυθμό μείωσης του συνολικού σφάλματος. Ο ρυθμός αυτός ονομάστηκε ρυθμός μάθησης και οδήγησε σε βελτίωση της αρχικής τεχνικής.

Μία άλλη στρατηγική δειγματοληψίας είναι η Εξόρυξη Δύσκολων Παραδειγμάτων (hard example mining) [97], η οποία βασίζεται στην αντίθετη ιδέα από το CL. Στην ΕΔΠ η δειγματοληψία δίνει έμφαση στα δύσκολα παραδείγματα, ώστε το δίκτυο να εκπαιδευτεί γρηγορότερα, λόγω του μεγέθους των παραγώγων, σε αντίθεση με πριν.

Στο [66] παρουσιάζεται ένας αλγόριθμος, που η πιθανότητα επιλογής ενός δείγματος στην παρτίδα είναι ανάλογη με το σφάλμα που έχουμε στο συγκεκριμένο δείγμα. Με άλλα λόγια, δείγματα που παρουσιάζουν μεγάλο σφάλμα έχουν μεγαλύτερη πιθανότητα να ενταχθούν στην παρτίδα. Για να λειτουργήσει αυτή η μέθοδος, πρέπει να είναι γνωστό το σφάλμα του κάθε δείγματος. Είναι αναγκαίο, λοιπόν, ένα εμπρόσθιο πέρασμα του συνόλου των δεδομένων, ώστε να καταγραφεί το σφάλμα που αντιστοιχεί στο κάθε δείγμα. Αυτό δημιουργεί ένα μεγάλο υπολογιστικό φόρτο στην εκπαίδευση και δεν είναι πρακτικό για εφαρμογές Βαθιάς Μηχανικής Μάθησης. Στο επόμενο μέρος αυτού του κεφαλαίου θα εξετάσουμε πως μπορούμε να το εφαρμόσουμε πιο αποδοτικά χωρίς κάποιο μειονέκτημα στην επίδοση.

Μία άλλη προσέγγιση στην βελτίωση της εκπαίδευσης ενός νευρωνικού είναι η μείωση της διακύμανσης των παραγώγων, που υπολογίζονται από τις παρτίδες των δεδομένων [47, 92, 94, 82]. Αυτό μπορεί να επιτευχθεί από διάφορες τεχνικές, μία από αυτές είναι η Δειγματοληψία Σπουδαιότητας (Importance Sampling). Η Δειγματοληψία Σπουδαιότητας χρησιμοποιεί μία κατανομή Q με πυκνότητα $q = dQ/dP$, η οποία μετασχηματίζει την εξίσωση 3.2 στην νέα μορφή:

$$\min_{w\in\mathbb{R}^n} \int_X \mathbf{J}(X, w) dP(x) = \min_{w\in\mathbb{R}^n} \int_X \frac{\mathbf{J}(X, w)}{q(x)} dQ(x) \tag{3.3}$$

Μία άλλη χρησιμότητα της Δειγματοληψίας Σπουδαιότητας είναι η μείωση του χρόνου εκπαίδευσης ενός νευρωνικού Δικτύου και έχει ερευνηθεί από διάφορες μελέτες [3, 50, 10].

3.3 Δυναμική Επιλογή Παρτίδας με Μεροληπτική Δειγματοληψία

Η μέθοδος που θα προτείνουμε σε αυτό το κεφάλαιο ονομάζεται Επιλογή Παρτίδας με Μεροληπτική Δειγματοληψία. Βασίζεται στην ιδέα ότι τα δυσκολότερα δείγματα για το δίκτυο θα βοηθήσουν περισσότερο την εκπαίδευση. Πιο συγκεκριμένα θα παρατηρήσουμε βελτίωση στην ταχύτητα σύγκλισης και πολλές φορές στην συνολική γενίκευση του μοντέλου. Για ευκολία θα χρησιμοποιηθεί η συντομογραφία ΕΠΜΔ στο υπόλοιπο της διατριβής.

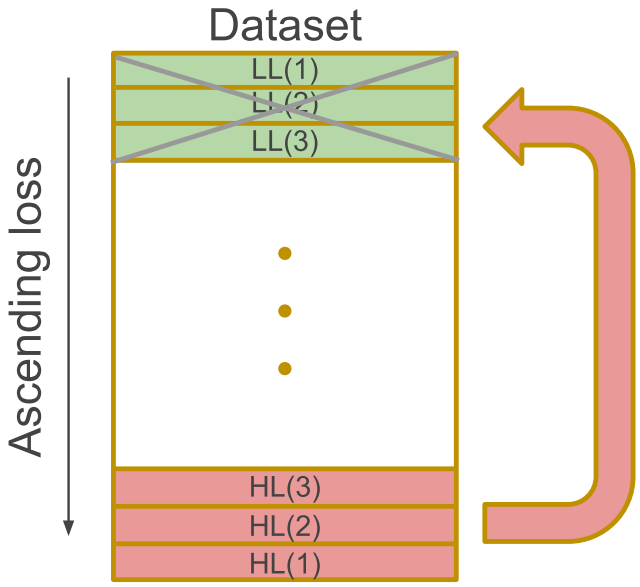
3.3.1 Μεθοδολογία

Έστω ότι έχουμε ένα νευρωνικό δίκτυο, το οποίο επιθυμούμε να εκπαιδεύσουμε. Έστω X είναι το σύνολο δεδομένων που διαθέτουμε για εκπαίδευση. Η διαδικασία της ΕΠΜΔ επιχειρεί την αντικατάσταση των ευκολότερων δειγμάτων με αυτά που το μοντέλο θεωρεί ως πιο δύσκολα. Ευκολότερα δείγματα θεωρούνται αυτά με το μικρότερο σφάλμα, ενώ τα δυσκολότερα αυτά με το μεγαλύτερο. Η προσοχή που θα δώσει το μοντέλο στα δύσκολα δείγματα θα επιταχύνει την σύγκλιση. Στο τέλος κάθε εποχής, η ΕΠΜΔ βρίσκει τα k δείγματα με μικρότερο σφάλμα και τα ανταλλάσσει με τα k δείγματα με το μεγαλύτερο σφάλμα. Πιο συγκεκριμένα, το μοντέλο σε κάθε εποχή θα βλέπει δύο φορές τα δύσκολα δείγματα, ενώ τα εύκολα καμία. Σε αυτό το σημείο θα εισάγουμε κάποιους συμβολισμούς για την καλύτερη κατανόηση. Με $HL^k \subset X$ θα συμβολίζουμε το υποσύνολο του συνόλου των δεδομένων, το οποίο περιέχει τα k δείγματα με το μεγαλύτερο σφάλμα (high loss). Αντίστοιχα $LL \subset X$ θα συμβολίσουμε τα δείγματα με το μικρότερο (low loss). Με αυτόν τον τρόπο σε κάθε εποχή το σύνολο εκπαίδευσης θα αλλάζει. Επομένως, θα υιοθετήσουμε τον συμβολισμό X_t για το σύνολο εκπαίδευσης την εποχή t, ενώ X θα παραμείνει το σύνολο των δεδομένων. Έχοντας αυτούς τους ορισμούς μπορούμε να ορίσουμε το σύνολο εκπαίδευσης ανά εποχή με τον ακόλουθο τύπο:

$$X_t = (X - LL^k_{X_{t-1}}) \cup HL^k_{X_{t-1}} = {LL^{k\prime}}_{X_{t-1}} \cup HL^k_{X_{t-1}} \tag{3.4}$$

όπου ${LL^{k\prime}}_{X_{t-1}}$ είναι τα δείγματα τα οποία δεν ανήκουν στο $LL^k_{X_{t-1}}$. Είναι προτιμότερο να χρησιμοποιούμε το δεύτερο μέρος της σχέσης 3.4, εξαιτίας της ευκολότερης υλοποίησης του. Στο Σχήμα 3.2 παρατηρούμε ένα παράδειγμα της λειτουργίας του αλγορίθμου για $k = 3$ για μία στιγμή κατά τη διάρκεια μίας εκπαίδευσης.

Όμως, για να υπολογίσουμε τα υποσύνολα HL και LL με ακρίβεια πρέπει να κάνουμε ένα εμπρόσθιο πέρασμα όλων των δεδομένων στο τέλος της κάθε εποχής. Αυτό δημιουργεί έναν τεράστιο υπολογιστικό φόρτο στην διαδικασία της εκπαίδευσης και σπαταλά αρκετό χρόνο. Για αυτό το λόγο χρησιμοποιούμε μία προσέγγιση του σφάλματος του συνόλου δεδομένων που θα την αναπτύξουμε παρακάτω.



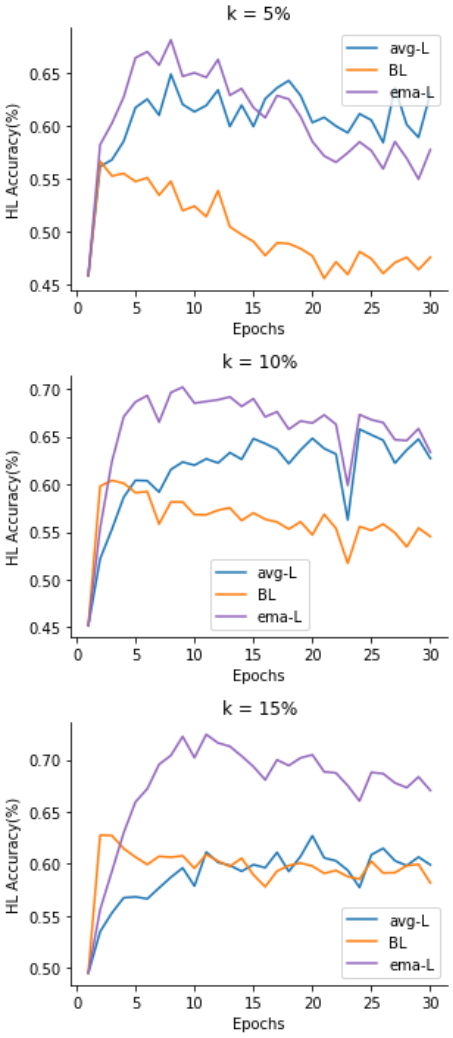
Σχήμα 3.2: Σχηματική αναπαράσταση της διαδικασίας της Επιλογής Παρτίδας με Μεροληπτική Δειγματοληψία για $k = 3$.

3.3.2 Προσέγγιση του σφάλματος

Για να παρακάμψουμε το πρόβλημα που δημιουργείται όταν επιθυμούμε να υπολογίσουμε τα HL και LL, θα προσπαθήσουμε να χρησιμοποιήσουμε τις μετρήσεις σφαλμάτων που είναι διαθέσιμες κατά τη διάρκεια της εκπαίδευσης. Σε κάθε επανάληψη του αλγορίθμου μάθησης επιλέγεται μία παρτίδα δεδομένων, γίνεται ένα εμπρόσθιο πέρασμα στο νευρωνικό για τον υπολογισμό του σφάλματος και, έπειτα, με τον αλγόριθμο της οπισθοδιάδοσης υπολογίζονται οι παράγωγοι και εφαρμόζονται μέσω του βελτιστοποιητή. Τα σφάλματα που παίρνουμε από την κάθε παρτίδα κατά τη διάρκεια μιας εποχής τα ονομάσουμε σφάλματα παρτίδας (Batch Loss) και θα τα συμβολίζουμε ως BL. Τα σφάλματα παρτίδας έχουν το μειονέκτημα ότι είναι παλιές τιμές, οι οποίες με την πάροδο κάποιων ανανεώσεων των βαρών μπορεί να μην ισχύουν. Για παράδειγμα, αν έχουμε ένα σύνολο δεδομένων χωρισμένο σε 100 παρτίδες, τότε σε μία εποχή γίνονται 100 ανανεώσεις βαρών. Τα σφάλματα που θα πάρουμε από τα δείγματα τις πρώτης παρτίδας θα θεωρούνται ξεπερασμένα στο τέλος της εποχής μετά από αυτές τις 100 ανανεώσεις. Από την άλλη, οι παρτίδες που βρίσκονται πιο κοντά στο τέλος της εποχής θα έχουν υπολογίσει πιο ακριβή σφάλματα σε σχέση με τις πραγματικές τιμές σφάλματος. Ας ορίσουμε τις πραγματικές τιμές σφάλματος των δειγμάτων στο τέλος μιας εποχής ως DL.

Σκοπός μας είναι να δείξουμε πόση διαφορά έχουν τα σφάλματα BL από τα πραγματικά DL και αν αυτό μπορεί να βελτιωθεί. Για να μετρήσουμε πόσο κοντά είναι μεταξύ τους οι πίνακες BL και DL υπάρχουν διάφοροι τρόποι. Ο προτεινόμενος τρόπος επιλέχθηκε επειδή ταιριάζει πάνω στην φύση του αλγορίθμου ΕΠΜΔ. Στον ΕΠΜΔ χρειάζονται για την λειτουργία του

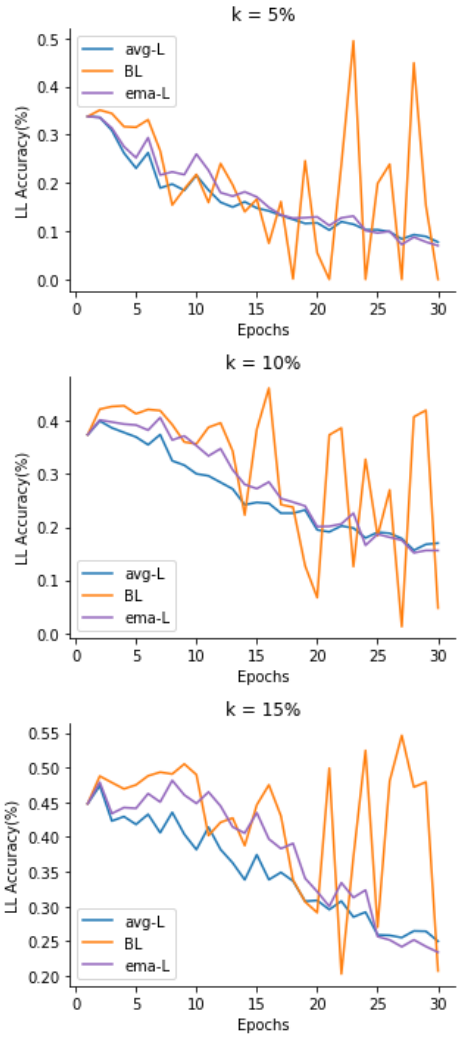
αλγορίθμου μόνο τα k υψηλότερα και χαμηλότερα σε σφάλμα δείγματα. Για αυτό το λόγο δεν είναι απαραίτητο να υπολογίσουμε ακριβώς τις τιμές των σφαλμάτων. Ο υπολογισμός της σειράς των δειγμάτων και η διάταξή τους ανάλογα με το σφάλμα τους αρκεί για την σωστή λειτουργία του αλγορίθμου. Επομένως, το πρόβλημα της ομοιότητας των δύο αυτών σφαλμάτων μπορεί να μετασχηματιστεί στο πόσα δείγματα κατηγοριοποιούνται σωστά στα υποσύνολα HL και LL. Οπότε η μετρική που χρησιμοποιήθηκε είναι η ακρίβεια, δηλαδή πόσα δείγματα ταξινομήθηκαν σωστά στα HL_{BL} και LL_{BL} σύμφωνα με τα HL_{DL} και LL_{DL} αντίστοιχα.



Σχήμα 3.3: Η ακρίβεια του HL συνόλου για διαφορετικές τιμές του k.

Σε αυτό το σημείο εκτελέστηκαν μία σειρά από πειράματα πάνω στο σύνολο δεδομένων του MNIST [61] για 30 εποχές. Για την συνέχεια της διατριβής η υπερπαράμετρος k θα γράφετε σαν ποσοστό του συνολικού αριθμού των δειγμάτων του συνόλου εκπαίδευσης για λόγους ευκολίας. Εξετάστηκαν 3 διαφορετικές τιμές για το k (5%, 10%, 15%). Δύο διαφορετικές προσεγγίσεις του DL εξετάστηκαν στα πειράματα, εκτός από το BL. Η πρώτη είναι ένας

απλός μέσος όρος (avg-L) του σφάλματος κάθε δείγματος για όλες τις εποχές μέχρι εκείνη την στιγμή. Ο δεύτερος τρόπος είναι ένας εκθετικός κινούμενος μέσος (ema-L). Στο Σχήμα 3.3 μπορούμε να δούμε την ακρίβεια των τριών αυτών προσεγγίσεων (BL, avg-L, ema-L) στο σύνολο HL για τις διαφορετικές τιμές του k. Ομοίως, μπορούμε να παρατηρήσουμε το αντίστοιχο Σχήμα 3.4 που αναπαριστά την ακρίβεια του συνόλου LL.



Σχήμα 3.4: Η ακρίβεια του LL συνόλου για διαφορετικές τιμές του k.

Παρατηρώντας τα σχήματα μπορούμε να καταλάβουμε ότι στο HL σύνολο την καλύτερη ακρίβεια την παρουσιάζει ο εκθετικός κινούμενος μέσος και για αυτό τον λόγο θα χρησιμοποιηθεί στον τελικό αλγόριθμο. Η μέγιστη ακρίβεια φτάνει το 70% και βλέπουμε ότι μεγαλώνοντας το k γίνεται και όλο και πιο σταθερή. Από την άλλη, στο Σχήμα 3.4 παρατηρούμε ότι η καλύτερη επιλογή είναι να πάρουμε το σύνολο BL ως έχει χωρίς κάποια αλλαγή. Η ακρίβεια σε αυτό κυμαίνεται γύρω στο 45% με αρκετές διακυμάνσεις. Αυτό μπορεί να αποδοθεί στο ότι στα εύκολα σύνολα δεδομένων υπάρχουν πολλά δείγματα που έχουν σφάλματα κοντά στο 0, έχοντας ως αποτέλεσμα την επιλογή ενός άλλου δείγματος με παραπλήσιο χαμηλό σφάλμα.

Με βάση αυτό μπορούμε να πούμε ότι δεν υπάρχει σοβαρό πρόβλημα στην ταξινόμηση του LL συνόλου. Δηλαδή, αν βγήκε από την εκπαίδευση ένα άλλο εύκολο δείγμα ακόμα και αν δεν ήταν στα k ευκολότερα, δεν αποτυγχάνει η βασική ιδέα του αλγορίθμου. Το πραγματικό πρόβλημα θα παρουσιαζόταν αν ένα δείγμα με χαμηλό σφάλμα επιλεγόταν σαν ένα δύσκολο, ή το αντίθετο. Όμως, το σφάλμα ταξινόμησης ενός δείγματος στο HL ενώ ανήκει στο LL είναι εξαιρετικά μικρό, κάτω από 1%, όπως επίσης, και το ανάποδο. Αυτό σημαίνει ότι υπάρχει πολύ μικρή πιθανότητα να ανταλλαχθεί ένα δύσκολο δείγμα με ένα εύκολο, το οποίο θα εμπόδιζε την εκπαίδευση. Σε σχέση με το [66], η ΕΠΜΔ δεν χρειάζεται επιπλέον εμπρόσθια περάσματα του νευρωνικού αφού χρησιμοποιεί τα σφάλματα τα οποία παίρνει από την κανονική διαδικασία εκπαίδευσης. Στο [66] η ακρίβεια των σφαλμάτων είναι αναγκαία διότι χρησιμοποιούνται για τον υπολογισμό των πιθανοτήτων του κάθε δείγματος. Επειδή, η ΕΠΜΔ βασίζεται στα HL και LL, ο αλγόριθμος απαιτεί λιγότερη ακρίβεια στον υπολογισμό των σφαλμάτων χωρίς να θυσιάζει την ποιότητα της επίδοσης του μοντέλου.

3.3.3 Υλοποίηση του αλγορίθμου

Σε αυτό το υποκεφάλαιο θα ορίσουμε τους τύπους που χρειάζονται για τον αλγόριθμο. Αρχικά, πρέπει να ορίσουμε τον εκθετικό κινούμενο μέσο που προαναφέρθηκε. Ακολουθεί η εξίσωσή του:

$$EMA_{BL_t} = lm * EMA_{BL_{t-1}} + (1 - lm) * BL_t \tag{3.5}$$

όπου t συμβολίζει την εποχή και $lm \in [0, 1]$ είναι η υπερπαράμετρος που ρυθμίζει την ένταση της ορμής του κινούμενου μέσου. Έπειτα, θα πρέπει να ορίσουμε καλύτερα τα υποσύνολα HL και LL. Πρώτα θα εισάγουμε τον ορισμό του υποσυνόλου με τα k μεγαλύτερα (ή μικρότερα) στοιχεία ενός συνόλου.

Ορισμός 3.2. *Έστω S είναι ένα σύνολου με k ή περισσότερα διατεταγμένα στοιχεία. Τότε, το υποσύνολο με τα k υψηλότερα στοιχεία του S θα ορίζεται ως $\hat{H}_k(S)$ και αυτό με τα χαμηλότερα k στοιχεία του S σαν $\hat{L}_k(S)$.*

Παρακάτω, οι εξισώσεις 3.6 και 3.7 δείχνουν την σχέση για την προσέγγιση των LL και HL συνόλων, αντίστοιχα, χρησιμοποιώντας τον Ορισμό 3.2 αλλά και τον εκθετικό μέσο της εξίσωσης 3.5.

$$LL_t^k = \hat{L}_k(BL_t) \tag{3.6}$$

$$HL_t^k = \hat{H}_k(EMA_{BL_t}) \tag{3.7}$$

Algorithm 1: Batch Selection with Biased Sampling

Parameters: dataset **D**, number of epochs **T**, number of datapoints **N**, batch size **B**, loss momentum **lm**, number of datapoints to be swapped **k**, indexes of samples **inds**.

Initializations: **inds**= [0, ... ,N-1], **lm**=0.7

for $t = 0$ *to* T **do**

  $X_t = D[inds]$;

  for $b = 0$ *to* N/B **do**

    $b\_inds \leftarrow SelectBatch(X_t, b)$

    $losses(b\_inds) \leftarrow ForwardPass(X_t[b\_inds])$

    $UpdateWeights(losses)$

    $BL[b\_inds] \leftarrow losses$

    $EMA[b\_inds] = lm * EMA[b\_inds] + (1 - lm) * losses$

  end

  $HL \leftarrow \hat{H}_k(EMA)$

  $LL \leftarrow \hat{L}_k(BL)$

  $inds \leftarrow HL \cup LL'$

end

Για την υλοποίηση του αλγορίθμου δύο πίνακες πρέπει να δημιουργηθούν, ένας για το BL και ένας EMA, όπου θα αποθηκεύουν τις τιμές των σφαλμάτων για το κάθε δείγμα. Η υπερπαράμετρος lm αρχικοποιήθηκε στην τιμή 0.7, ενώ ο πίνακας του εκθετικού κινούμενου μέσου αρχικοποιήθηκε στο 0. Για την εύρεση των k χαμηλότερων και υψηλότερων τιμών ενός συνόλου χρησιμοποιήθηκε ο αλγόριθμος χωρίσματος Introselect [75], ο οποίος χωρίζει ένα σύνολο σε δύο κομμάτια. Το πρώτο κομμάτι εμπεριέχει τα i μεγαλύτερα στοιχεία και το άλλο τα υπόλοιπα. Με αυτόν τον τρόπο μπορούμε να βρούμε το HL χωρίς να ταξινομήσουμε όλα τα στοιχεία του συνόλου. Ως αποτέλεσμα, βελτιώνει αρκετά τον χρόνο υπολογισμού αλλά και την πολυπλοκότητα του συνολικού αλγορίθμου. Ενώ ένας αλγόριθμος ταξινόμησης στοιχείων θα είχε πολυπλοκότητα $O(nlogn)$, τώρα με τον αλγόριθμο Introselect έχει $O(n)$. Αποτελεί σημαντική προσθήκη επειδή δεν αυξάνει την πολυπλοκότητα της συνολικής εκπαίδευσης ενός νευρωνικού. Στον Αλγόριθμο 1 παρατηρείται ο συνολικός αλγόριθμος ΕΠΜΔ (BSBS).

3.3.4 Πειραματική Διαδικασία

Η προτεινόμενη μέθοδος θα εξεταστεί πειραματικά πάνω σε 4 διαφορετικά σύνολα δεδομένων, για ταξινόμηση και παλινδρόμηση. Η σύγκριση θα γίνει ως προς την ταχύτητα σύγκλισης σε σχέση με το SGD-Scan. Τέσσερα διαφορετικά τρεξίματα του ΕΠΜΔ εξετάστηκαν, που διαφοροποιούνται με βάση την υπερπαράμετρο k. Οι τιμές του k που εξετάστηκαν είναι $(5\%, 10\%, 15\%)$ και μία τελευταία ρύθμιση, που το k συνεχώς αυξάνεται σταδιακά κάθε εποχή από το 0% ως το 15%. Η τελευταία επιλογή θα μας βοηθήσει να καταλάβουμε την συμπεριφορά της υπερπαραμέτρου στο πέρας των εποχών. Κάθε μέθοδος θα συγκριθεί στις ίδιες υπερπαραμέτρους και αρχικοποιήσεις βαρών, ώστε να η σύγκριση να παραμείνει όσο πιο δίκαια

γίνεται.

Η αρχιτεκτονική του νευρωνικού δικτύου που επιλέχθηκε για το κάθε σύνολο δεδομένων φαίνεται στον Πίνακα 3.2, ενώ οι υπερπαράμετροι που χρησιμοποιήθηκαν σε κάθε εκπαίδευση παρουσιάζονται στον Πίνακα 3.1. Στις αρχιτεκτονικές το επίπεδο της Απόσυρσης εφαρμόστηκε στα πλήρως συνδεδεμένα επίπεδα μόνο.

Πίνακας 3.1: Υπερπαράμετροι Βελτιστοποίησης.

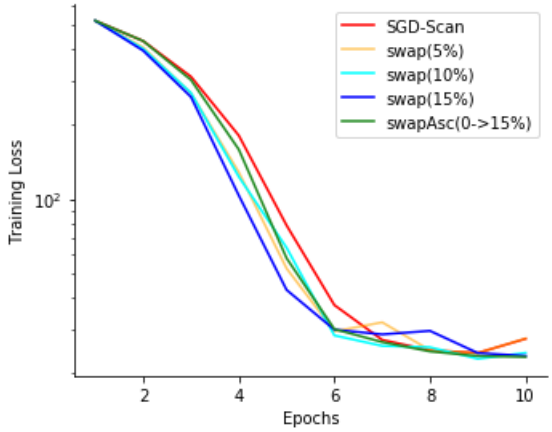
Datasets	Optimizer	Learning rate	Learning rate decay	Batch size
Boston Housing	SGD Momentum	0.001	1	40
MNIST	SGD	0.1	1	128
CIFAR10	Adam[51]	0.001	1	128
CIFAR100	Adam[51]	0.001	0.5 (at 19, 40, 55 epochs)	128

3.3.4.1 Παλινδρόμηση

Το πρώτο πείραμα εφαρμόστηκε πάνω σε πρόβλημα παλινδρόμησης στο σύνολο δεδομένων Boston Housing [38]. Τα δεδομένα αποτελούνται από 404 δείγματα εκπαίδευσης και 102 δείγματα ελέγχου. Κάθε δείγμα αποτελείται από 13 χαρακτηριστικά από σπίτια σε διαφορετικές τοποθεσίες γύρω από τη Βοστόνη. Ο στόχος της πρόβλεψης είναι να βρεθεί ο διάμεσος των τιμών των σπιτιών σε μία περιοχή. Τα χαρακτηριστικά των δεδομένων κανονικοποιήθηκαν (αφαιρέθηκε ο μέσος όρος και διαιρέθηκε η τυπική απόκλιση). Για αυτό το πείραμα, ένα απλό πλήρες συνδεδέμενο δίκτυο με 2 επίπεδα εκπαιδεύτηκε για 10 εποχές. Το μέσο τετραγωνικό σφάλμα (MSE - Mean Square Error) χρησιμοποιήθηκε για την συνάρτηση κόστους. Στο Σχήμα 3.5 βλέπουμε τις καμπύλες εκπαίδευσης σφάλματος, όπου η ΕΠΜΔ επιδεικνύει καλύτερη επίδοση σε σχέση με το SGD-Scan. Το καλύτερο σφάλμα παρατηρήθηκε με την υπερπαράμετρο k να είναι ίση με 10%.

Πίνακας 3.2: Αρχιτεκτονικές των μοντέλων για τα διαφορετικά σύνολα δεδομένων.

Datasets	Conv layers	Filter size	Pooling layers	BN layers	FC layers	Dropout
Boston Housing	-	-	-	-	2	-
MNIST	2	3×3	1	-	2	0.5
CIFAR10	4	3×3	2 (max)	-	2	0.5
CIFAR100	17	(1) 7×7 (16) 3×3	2 (1 max, 1 avg)	yes	1	-



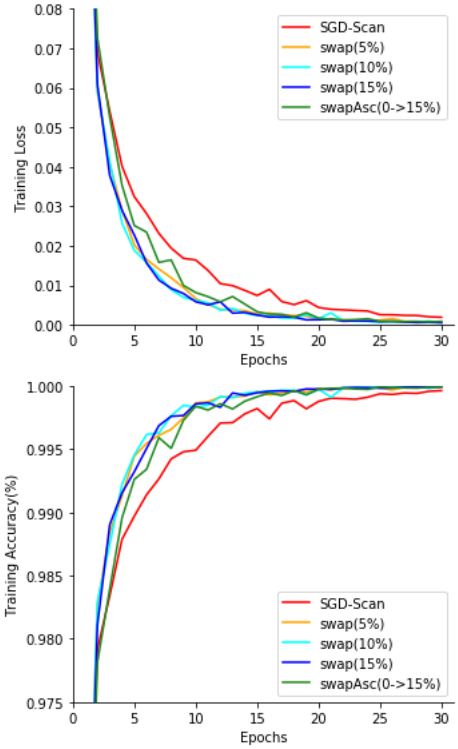
Σχήμα 3.5: Οι καμπύλες σφάλματος της εκπαίδευσης για το σύνολο δεδομένων του Boston Housing.

3.3.4.2 Ταξινόμηση Εικόνας

Όσον αφορά το έργο της ταξινόμησης εικόνας, χρησιμοποιήθηκαν 3 σύνολα δεδομένων με εικόνες για την διεξαγωγή των πειραμάτων. Για όλα τα πειράματα χρησιμοποιήθηκε η κατηγορική σταυροειδής εντροπία ως συνάρτηση κόστους. Οι εικόνες κανονικοποιήθηκαν στο εύρος $[0, 1]$, ενώ καμία άλλη προεπεξεργασία των δεδομένων δεν εφαρμόστηκε. Όλα τα πειράματα αξιολογήθηκαν ως προς την ακρίβεια των προβλέψεων και την ταχύτητα σύγκλισης.

MNIST Ο αλγόριθμος εξετάστηκε στο MNIST [61] σύνολο δεδομένων, το οποίο αποτελείται από εικόνες σε γκρι κλίμακα μεγέθους 28×28 από αριθμούς ζωγραφισμένους με το χέρι. Έχει 60000 εικόνες για εκπαίδευση και 10000 για έλεγχο. Διαθέτει 10 κλάσεις για τα νούμερα $(0-9)$. Ένα απλό Συνελικτικό Δίκτυο επιλέχθηκε για αυτό το πείραμα και εκπαιδεύτηκε για 30 εποχές. Από το Σχήμα 3.6 παρατηρούμε ότι η ΕΠΜΔ οδηγεί σε γρηγορότερη εκπαίδευση σε σχέση με τον SGD-Scan, επειδή φτάνει την καλύτερη ακρίβεια και σφάλμα στις μισές εποχές. Εξαιτίας ότι το MNIST είναι ένα σύνολο δεδομένων που θεωρείται εύκολο στην εκπαίδευση και λόγω της έλλειψης ακραίων δειγμάτων, είναι εμφανές ότι η έμφαση στα δύσκολα δείγματα επιταχύνει την διαδικασία της εκπαίδευσης. Όσον αφορά τις επιδόσεις στα δεδομένα ελέγχου, η ΕΠΜΔ με $k = 15\%$ έδωσε τα καλύτερα αποτελέσματα.

CIFAR10 Το δεύτερο πείραμα εκτελέστηκε πάνω στα δεδομένα του CIFAR10 [55], το οποίο αποτελείται από 60000 έγχρωμες εικόνες (32×32 ανάλυσης). Από αυτές οι 50000 είναι για εκπαίδευση και οι υπόλοιπες για έλεγχο, ενώ είναι χωρισμένες σε 10 κλάσεις. Μία μικρή έκδοση του VGG (που περιγράφεται στον Πίνακα 3.2) εκπαιδεύτηκε για 70 εποχές. Το μεγαλύτερο k που χρησιμοποιήθηκε (15%) απέδωσε καλύτερα στην μετρική της ακρίβειας. Η πόλωση που δημιουργείται στην κάθε παρτίδα φαίνεται να βοηθά στην επιτάχυνση της εκπαίδευσης. Παρόλο που αυτά τα δεδομένα είναι δυσκολότερα από τα προηγούμενα δύο πειράματα,



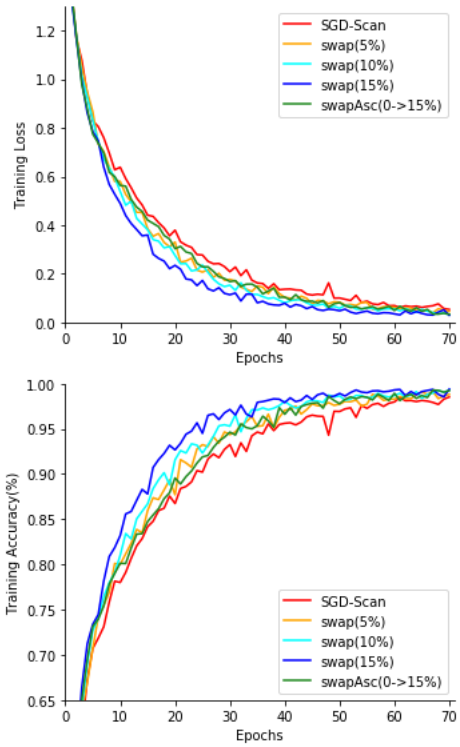
Σχήμα 3.6: Το σφάλμα και η ακρίβεια εκπαίδευσης για το σύνολο δεδομένων MNIST.

παρατηρείται βελτίωση κατά 40% στην ταχύτητα σύγκλισης σε ικανοποιητικό σημείο ακρίβειας.

CIFAR100 Το τελευταίο σύνολο δεδομένων που εξετάστηκε στα πειράματα είναι το CIFAR100 [55]. Όπως και το CIFAR10, αποτελείται από μικρές έγχρωμες εικόνες (32×32) με την ίδια ακριβώς κατανομή σε δείγματα εκπαίδευσης και ελέγχου (50000 και 10000). Η μόνη διαφορά έγκειται στον αριθμό των κλάσεων που τώρα είναι 100. Για αυτά τα δεδομένα χρησιμοποιήθηκε μία αρχιτεκτονική ResNet [39] και εκπαιδεύτηκε για 70 εποχές. Το Σχήμα 3.8 δείχνει ότι στις πρώτες 20 εποχές όλες οι μέθοδοι αποδίδουν παρόμοια. Μετά τις 38 εποχές η ΕΠΜΔ αρχίζει να επιταχύνει την εκπαίδευση και να φτάνει υψηλότερες τιμές ακρίβειας παραμένοντας 5% πάνω από το SGD-Scan σταθερά. Η επίδοση στα δεδομένα ελέγχου δείχνουν ότι το SGD-Scan και η ΕΠΜΔ παρουσιάζουν κοντινές ακρίβειες.

3.3.5 Αποτελέσματα και Παρατηρήσεις

Ο Πίνακας 3.3 συμπεριλαμβάνει όλα τα αποτελέσματα από τις επιδόσεις των μεθόδων στα δεδομένα ελέγχου. Δείχνει την ακρίβεια για τα έργα της ταξινόμησης και το μέσο τετραγωνικό σφάλμα για το πείραμα της παλινδρόμησης. Παρατηρείται ότι η ΕΠΜΔ επιτυγχάνει καλύτερα αποτελέσματα σε αρκετές περιπτώσεις. Είναι σημαντικό να προσθέσουμε ότι οι επιδόσεις γενίκευσης είναι σχετικά χαμηλές σε σχέση με τις βέλτιστες θεωρητικά που θα μπορούσαν να είναι εφικτές. Αυτό εξηγείται στην έλλειψη ενίσχυσης δεδομένων. Όμως, ο σκοπός των

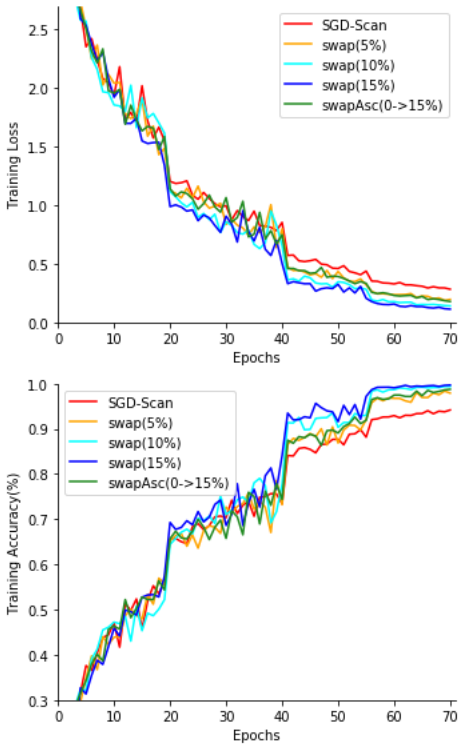


Σχήμα 3.7: Το σφάλμα και η ακρίβεια εκπαίδευσης για το σύνολο δεδομένων CIFAR10.

πειραμάτων είναι η σύγκριση των μεθόδων μεταξύ τους στις ίδιες συνθήκες και όχι η βέλτιστη απόδοση γενικότερα. Με αυτόν τον τρόπο καταφέρνουμε να υπάρχει ένα δίκαιο περιβάλλον σύγκρισης των αλγορίθμων.

Με την χρήση του αλγόριθμου ΕΠΜΔ, εισήχθησαν 2 υπερπαράμετροι, που επηρεάζουν την διαδικασία της μάθησης και την γενικότερη επίδοση του νευρωνικού. Αυτές ήταν ο αριθμός k, που συμβολίζει των αριθμό των δειγμάτων που βγαίνουν εκτός εκπαίδευσης για μία εποχή, και η ορμή lm, που καθορίζει την αλλαγή του εκθετικού μέσου. Η τιμή του k είναι εξαιρετικά σημαντική, επειδή χειρίζεται πόση τεχνητή πόλωση θα προστεθεί στο σύνολο εκπαίδευσης. Όπως, παρατηρήθηκε στα πειράματα αυτή η πόλωση είναι ευεργετική στην διαδικασία της εκπαίδευσης αλλά η υπερβολική χρήση της εγκυμονεί κινδύνους. Είναι σημαντικό το k να μένει κάτω από 30% για την αποφυγή αρνητικών συνεπειών. Για παράδειγμα, με τη χρήση υψηλού k, η γενίκευση του μοντέλου θα χειροτερεύσει, αφού το μοντέλο θα βλέπει ένα μικρό ποσοστό του συνόλου εκπαίδευσης. Εκτός αυτών, κάθε σύνολο δεδομένων έχει την δική του βέλτιστη τιμή για το k, η οποία είναι καλό να εξετάζεται.

Η δεύτερη υπερπαράμετρος είναι η lm, η οποία χρειάζεται λιγότερη λεπτομερή ρύθμιση για την βέλτιστη τιμή. Είναι σημαντική επειδή καθορίζει την τιμή της προσέγγισης του σφάλματος που υπολογίζουμε για το κάθε δείγμα. Επομένως, επηρεάζει ποια δείγματα θα παραμείνουν και ποια όχι στην εκπαίδευση σε κάθε εποχή. Για σύνολα δεδομένων που θεωρούνται εύκολα στην εκμάθηση, το lm θα μπορούσε να πάρει και μικρότερες τιμές από το 0.7. Σε δυσκο-



Σχήμα 3.8: Το σφάλμα και η ακρίβεια εκπαίδευσης για το σύνολο δεδομένων CIFAR100.

λότερα δεδομένα μεγαλύτερες τιμές του lm θα μπορούσαν να βοηθήσουν τη μάθηση. Αυτό συμβαίνει επειδή σε δύσκολα σύνολα δεδομένων τα σφάλματα των δειγμάτων είναι πιο πιθανό να ταλαντεύονται περισσότερο σε κάθε ανανέωση και, με εκθετικό μέσο που ανανεώνεται πιο αργά, θα είναι ευκολότερο να διακρίνουμε τα δυσκολότερα δείγματα.

Πίνακας 3.3: Οι καλύτερες επιδόσεις ακρίβειας (και μέσου τετραγωνικού σφάλματος για Boston Housing) στα σύνολα ελέγχου για κάθε σύνολο δεδομένων των πειραμάτων.

Datasets	Model	SGD_Scan	BSBS(5%)	BSBS(10%)	BSBS(15%)	BSBS(asc)
Boston Housing	FC	21.1266	25.3165	**20.6950**	24.6499	24.5397
MNIST	CNN	0.9919	0.9922	0.9918	**0.9926**	0.9921
CIFAR10	CNN	0.7547	0.7542	0.7520	**0.7573**	0.7542
CIFAR100	ResNet18[39]	0.4838	0.4794	0.4630	0.4706	**0.4849**

Κεφάλαιο 4

Ανισορροπία δεδομένων

Τα τελευταία χρόνια η διαθεσιμότητα δεδομένων έχει ενισχυθεί με την ανάπτυξη της τεχνολογίας. Για την εκπαίδευση Νευρωνικών Δικτύων αυτό είναι αρκετά σημαντικό, αφού όσο περισσότερα δεδομένα διαθέτουμε για εκπαίδευση, τόσο καλύτερες επιδόσεις θα επιτυγχάνει. Όμως, η ποσότητα των δεδομένων δεν παίζει τον μοναδικό καθοριστικό ρόλο για την καλή γενίκευση ενός Νευρωνικού, αλλά και οι ιδιαιτερότητες του κάθε συνόλου δεδομένων. Ένα βασικό χαρακτηριστικό των δεδομένων είναι η υποκείμενη κατανομή των δεδομένων εκπαίδευσης και πόσο διαφορετική είναι από την θεωρητική (τέλεια) κατανομή του έργου που εξετάζουμε. Πολλές φορές η υποκείμενη κατανομή παρουσιάζει κάποια ανισορροπία, η οποία δυσκολεύει την διαδικασία την μάθησης και επηρεάζει αρνητικά την επίδοση γενίκευσης του μοντέλου [45, 35]. Μία κατανομή μπορεί να είναι ισορροπημένη όταν ο αριθμός των δειγμάτων από μοναδικές και διαφορετικές παρατηρήσεις είναι κατανεμημένος ομοιόμορφα. Στην βιβλιογραφία αυτό το πρόβλημα συνήθως αναφέρεται στη μορφή ως ανισορροπία κλάσεων σε προβλήματα ταξινόμησης.

Σε ένα έργο ταξινόμησης υπάρχουν κλάσεις στις οποίες καλούμαστε να κατατάξουμε τα διάφορα αντικείμενα/δείγματα του συνόλου δεδομένων. Στις περισσότερες φορές μπορούμε να ξεχωρίσουμε τις κλάσεις σε δύο κατηγορίες. Η πρώτη κατηγορία ονομάζεται κλάση πλειονότητας και είναι αυτή που παρουσιάζει μία πληθώρα δειγμάτων. Η δεύτερη κατηγορία είναι αυτή που περιέχει πολύ λίγα δείγματα και θα την αποκαλούμε κλάση μειονότητας. Αυτό το φαινόμενο διαχωρισμού κλάσεων συναντάται πολύ συχνά σε πραγματικά σύνολα δεδομένων. Μπορεί να συμβεί για διάφορους λόγους. Ο πιο συχνός λόγος είναι ότι υπάρχουν πολύ λίγες περιπτώσεις εμφάνισης ενός συγκεκριμένου δείγματος σε ένα πραγματικό περιβάλλον στο οποίο θέλουμε να λύσουμε ένα πρόβλημα. Αυτά τα σπάνια δείγματα πρέπει να έχουν ιδιαίτερο ενδιαφέρον στην επίλυση του προβλήματος για να τα ανακηρύξουμε ως κλάση μειονότητας. Για παράδειγμα, στο τομέα του εντοπισμού απάτης σε συναλλαγές μπορούμε να παρατηρήσουμε έναν μεγάλο αριθμό δειγμάτων που είναι κανονικές αλλά μόνο μία μικρή ομάδα από συναλλαγές που να θεωρούνται κακόβουλες. Όμως, αυτές οι περιπτώσεις είναι οι σημαντικές και αφού η αναγνώριση αυτών είναι το κύριο μέλημα αυτού του προβλήματος. Επομένως, βλέπουμε ότι αν και κάποιες κλάσεις μπορεί να παρουσιάζουν χαμηλό αριθμό δειγμάτων, αυτό δεν μεταφράζεται

ως χαμηλή σημασία αλλά το αντίθετο.

Πώς, όμως, ένα μοντέλο καταφέρνει και μαθαίνει εξίσου καλά και τις κλάσεις μειονότητας· Έχουν αναπτυχθεί πολλές τεχνικές και αλγόριθμοι, οι οποίοι επιδιώκουν τον καλύτερο διαχωρισμό των δύο αυτών τύπων κλάσεων και, συνεπώς, στην επίτευξη καλύτερης γενίκευσης των μοντέλων. Στο παρακάτω υποκεφάλαιο θα αναφέρουμε σημαντικές τέτοιες μεθόδους.

4.1 Τεχνικές Καταπολέμησης Ανισορροπίας δεδομένων

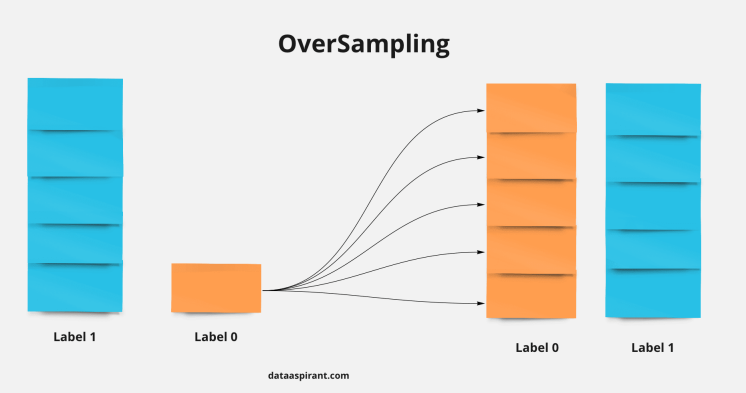
Οι τεχνικές καταπολέμησης της ανισορροπίας των κλάσεων είναι πολλές και καθεμία από αυτές έχει μία διαφορετική προσέγγιση του προβλήματος. Παρόλα αυτά μπορούμε να τις κατηγοριοποιήσουμε σε 2 μεγάλες κατηγορίες. Η πρώτη κατηγορία βασίζεται στον μετασχηματισμό των δεδομένων σε μία πιο ισορροπημένη μορφή. Η δεύτερη κατηγορία βασίζεται στην αλλαγή του αλγορίθμου μάθησης με σκοπό να λύσει την ανισορροπία δεδομένων χωρίς να αλλάξει τα δεδομένων αυτά καθεαυτά. Αυτές οι δύο κατηγορίες μεθόδων είναι εξίσου σημαντικές και έχουν χρησιμοποιηθεί στην βιβλιογραφία πολλές φορές. Κάποιες από τις πιο σημαντικές τεχνικές και μεθόδους θα αναφερθούν παρακάτω.

4.1.1 Μετασχηματισμός των Δεδομένων

Οι πιο συχνές τεχνικές για την καταπολέμηση της ανισορροπίας των κλάσεων είναι οι μέθοδοι μετασχηματισμού των δεδομένων. Ο σκοπός τους είναι μέσω πρόσθεσης, αφαίρεσης ή μετασχηματισμό δειγμάτων από το σύνολο των δεδομένων να βοηθήσει το κάθε μοντέλο να ξεχωρίζει καλύτερα τις κλάσεις μειονότητας. Υπάρχουν τρεις προσεγγίσεις στον μετασχηματισμό δεδομένων: η υπερδειγματοληψία (Oversampling), η υποδειγματοληψία (Undersampling) και ο συνδυασμός των δύο.

Υπερδειγματοληψία

Η υπερδειγματοληψία είναι ο πιο συνήθης τρόπος επίλυσης μη ισορροπημένων δεδομένων. Η Τυχαία Υπερδειγματοληψία (Random Oversampling - ROS) είναι η πιο απλή έκδοσή της. Επιλέγει τα δείγματα της κλάσης μειονότητας και τα αντιγράφει πολλές φορές στο σύνολο δεδομένων μέχρι να φτάσουν σε αριθμό τα δείγματα της κλάσης πλειονότητας. Είναι μία μέθοδος που δεν απαιτεί κάποιο ιδιαίτερο υπολογιστικό κόστος και μπορεί να χρησιμοποιηθεί παράλληλα με βαθιά νευρωνικά δίκτυα αλλά και άλλες τεχνικές μηχανικής μάθησης. Το ROS παρόλο την χρησιμότητά του εμφανίζει κάποια μειονεκτήματα. Ένα από τα κύρια μειονεκτήματα του εμφανίζεται σε περιπτώσεις που η κλάση μειονότητας παρουσιάζει πολύ μικρή ποικιλία. Αυτό συμβαίνει σε περιπτώσεις ακραίας ανισορροπίας, όπου η κλάση μειονότητας έχει πολύ λίγα δείγματα. Η πολλαπλές αντιγραφές των χαμηλής ποικιλίας δειγμάτων δεν προσθέτει επιπλέον γνώση στο σύνολο δεδομένων και μετατρέπει την υποκείμενη κατανομή εξαιρετικά μη ρεαλιστική, οδηγώντας σε υπερπροσαρμογή.



Σχήμα 4.1: Τυχαία Υπερδειγματοληψία για δύο κλάσεις.

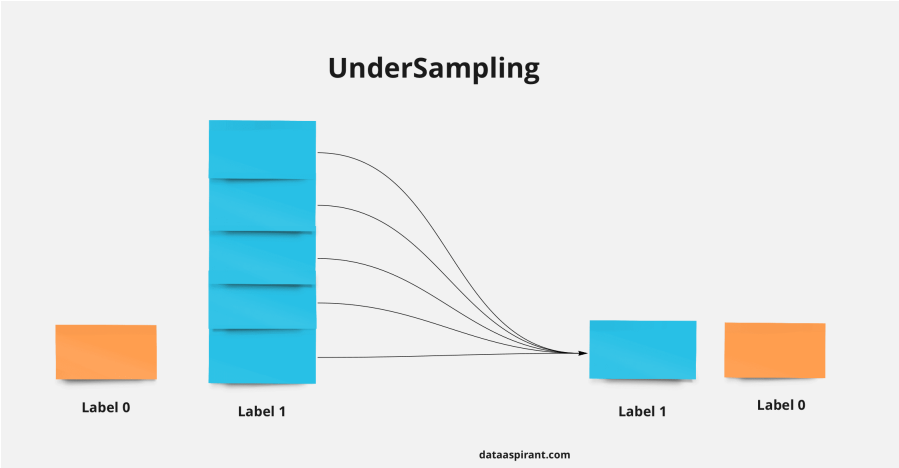
Πολλές τεχνικές έχουν αναπτυχθεί για να επιλύσουν αυτό το πρόβλημα. Μία από τις πιο γνωστές είναι το SMOTE [15]. Αυτή η μέθοδος χρησιμοποιεί τον αλγόριθμο k-Nearest Neighbors - (k-NN) [29] για να επιλέξει παρόμοια δείγματα από την κλάση μειονότητας και να τα συνδυάσει ώστε να δημιουργήσει καινούργια δείγματα. Επομένως, ο SMOTE παράγει καινούργια συνθετικά δεδομένα, τα οποία εμπλουτίζουν το σύνολο εκπαίδευσης. Με αυτόν τον τρόπο η υποκείμενη κατανομή μετατρέπεται σε μία άλλη που αντιπροσωπεύει καλύτερα το πραγματικό έργο.

Από το SMOTE, έχουν εμπνευστεί διάφορες μέθοδοι ανά τα χρόνια. Η τεχνική Borderline-SMOTE [37] είναι μία παραλλαγή, που προσπαθεί να βελτιώσει την υπερδειγματοληψία επιλέγοντας δείγματα που βρίσκονται κοντά στο όριο μεταξύ των κλάσεων. Τα δείγματα που είναι κοντά στο όριο είναι πιο επιρρεπή στο να ταξινομηθούν λάθος από δείγματα που βρίσκονται πιο μακρυά. Επομένως, το BorderlineSMOTE βρίσκει τα δείγματα της κλάσης μειονότητας κοντά στο όριο και τα υπερδειγματοληπτεί. Μία άλλη πιο σύγχρονη μέθοδος βασισμένη στο SMOTE είναι το KMeans-SMOTE [58]. Αυτή η μέθοδος, αρχικά, χρησιμοποιεί τον αλγόριθμο K-Means για να χωρίσει τα δείγματα σε ομάδες. Έπειτα, επιλέγει τις ομάδες που εμφανίζουν τις μεγαλύτερες ανισορροπίες κλάσεων και υπολογίζει βάρη δειγματοληψίας για κάθε ομάδα. Αυτά τα βάρη χρησιμοποιούνται μετά για να υπολογιστούν πόσα καινούργια δείγματα θα προστεθούν στην κλάση μειονότητας για κάθε μία από της ομάδες. Η τελική δειγματοληψία χρησιμοποιεί τον αλγόριθμο SMOTE.

Ένας άλλος αρκετά χρήσιμος αλγόριθμος είναι ο Adasyn [36]. Η κεντρική ιδέα του είναι παραπλήσια με αυτή του SMOTE, επειδή χρησιμοποιεί τον αλγόριθμο k-NN για να βρει k γειτονικά δείγματα σε κάθε δείγμα της κλάσης μειονότητας. Όμως, η διαφορά μεταξύ τους βρίσκεται στον αριθμό των καινούργιων δειγμάτων που δημιουργούνται. Ο Adasyn υπολογίζει τον αριθμό των παραγόμενων δειγμάτων με βάση την κατανομή της γειτονιάς του κάθε δείγματος της μειονότητας. Σε γενικές γραμμές ο Adasyn θεωρείται σαν βελτίωση του αρχικού SMOTE.

Υποδειγματοληψία

Η υποδειγματοληψία είναι η δεύτερη πιο διαδεδομένη τεχνική που μετασχηματίζει τα δεδομένα για να εξαλείψει την ανισορροπία. Σε αντίθεση με την υπερδειγματοληψία, η υποδειγματοληψία αφαιρεί δείγματα από το σύνολο εκπαίδευσης, και ειδικότερα από την κλάση πλειονότητας. Συγκεκριμένα, ο πιο απλός αλγόριθμος υποδειγματοληψίας είναι η Τυχαία Υποδειγματοληψία (Random Undersampling - RUS). Αυτός ο αλγόριθμος επιλέγει τυχαία δείγματα από την κλάση πλειονότητας μέχρι ο αριθμός των δύο ειδών κλάσεων να ισορροπήσει. Στις περισσότερες εφαρμογές, και ειδικότερα σε εφαρμογές βαθιάς μηχανικής μάθησης, θεωρείται πιο ωφέλιμη η υπερδειγματοληψία σε σχέση με την υποδειγματοληψία, επειδή τα περισσότερα δεδομένα βοηθούν αρκετά τέτοιου είδους εφαρμογές. Παρόλα αυτά υπάρχουν περιπτώσεις που η υποδειγματοληψία μπορεί να αποδώσει καλύτερα [24].



Σχήμα 4.2: Τυχαία Υποδειγματοληψία για δύο κλάσεις.

Μία άλλη μέθοδος υποδειγματοληψίας είναι η Edited Nearest Neighbors (ENN) [108], η οποία χρησιμοποιεί σαν βάση του αλγορίθμου τον k-NN. Με αυτόν βρίσκει όλους τους γείτονες των δειγμάτων της κλάσης πλειονότητας. Έπειτα, αφαιρεί τα δείγματα που η πλειοψηφία των γειτόνων τους ανήκουν σε άλλη κλάση. Αφαιρεί, δηλαδή, τα δείγματα της κλάσης πλειονότητας που βρίσκονται κοντά σε δείγματα της κλάσης μειονότητας και, γενικά, κοντά στο όριο μεταξύ των κλάσεων. Αυτό βοηθά την εκπαίδευση επειδή το όριο μεταξύ των κλάσεων γίνεται πιο εύκολο να βρεθεί.

Εμπνευσμένο από τον ENN αναπτύχθηκε ένας άλλος αλγόριθμος υποδειγματοληψίας, με το όνομα Tomek links [1]. Ο σκοπός αυτού του αλγόριθμου είναι να βρει τους συνδέσμους Tomek μέσα στο σύνολο δεδομένων και να τους αφαιρέσει. Ένας σύνδεσμος Tomek είναι ένα ζευγάρι δειγμάτων, που το ένα δείγμα ανήκει στην κλάση μειονότητας και το άλλο στην κλάση πλειονότητας, και τα δύο δείγματα είναι ο κοντινότερος γείτονας του άλλου. Αφαιρώντας τα, τα όρια μεταξύ των κλάσεων γίνονται πιο ευδιάκριτα και ο διαχωρισμός των κλάσεων γίνεται ευκολότερος. Ως αποτέλεσμα βελτιώνεται η επίδοση στην ταξινόμηση και η γενίκευση του

μοντέλου. Μία άλλη μέθοδος υποδειγματοληψίας, η οποία χρησιμοποιεί τεχνικές συσταδοποίησης δεδομένων, είναι η ClustCentroids [113]. Αυτός ο αλγόριθμος επιχειρεί να κρατήσει έναν αριθμό δειγμάτων από την κλάση πλειονότητας με την χρήση του αλγορίθμου K-Means. Ας υποθέσουμε ότι θα κρατήσουμε N δείγματα, τότε ο αλγόριθμος εφαρμόζει τον K-Means χωρίζοντας τα δεδομένα σε N συστάδες (ομάδες). Αυτές οι ομάδες έχουν N κέντρα, τα οποία είναι τα δείγματα, τα οποία κρατάει ο αλγόριθμος ClustCentroids σαν τα καινούργια δείγματα.

Συνδυασμός Υπερδειγματοληψίας και Υποδειγματοληψίας

Μία αρκετά ενδιαφέρουσα ιδέα είναι ο συνδυασμός των δύο τεχνικών που προαναφέρθηκαν, της υπερδειγματοληψίας και της υποδειγματοληψιας. Ο συνδυασμός τους μπορεί να εκμεταλλευτεί τα πλεονεκτήματα και των δύο τεχνικών και να δημιουργήσει ένα σύνολο δεδομένων που να εκπαιδεύει μοντέλα με τέτοιο τρόπο ώστε να μαθαίνουν πιο εύκολα την υποκείμενη κατανομή του προβλήματος. Συγκεκριμένα, ε την χρήση της υπερδειγματοληψίας θα προστεθούν δείγματα της κλάσης μειονότητας για να τα αναγνωρίζει καλύτερα το μοντέλο, ενώ με την χρήση της υποδειγματοληψίας θα αφαιρεθούν δείγματα από την κλάσης της πλειονότητας για να γίνουν πιο έυκολα διαχωρίσιμες οι κλάσεις. Δύο από τις πιο διαδεδομένες τεχνικές, που θα χρησιμοποιηθούν και σε κεφάλαια παρακάτω, είναι οι εξής: SMOTETomek [4] και SMOTEENN [5]. Οι δύο αυτές μέθοδοι χρησιμοποιούν τον αλγόριθμο SMOTE για υπερδειγματοληψία της κλάσης μειονότητας και, έπειτα, χρησιμοποιούν τους αλγόριθμους Tomek links ή ENN, αντίστοιχα, για τον ¨καθαρισμό' των δεδομένων με υποδειγματοληψία. Έχει αποδειχτεί ότι σε διάφορες περιπτώσεις ο συνδυασμός υπερ- και υποδειγματοληψίας αποδίδει καλύτερα σε σχέση με την χρήση μιας από της δύο τεχνικές μεμονωμένα [91].

4.1.2 Προσαρμογή του Αλγορίθμου μάθησης

Η δεύτερη μεγάλη κατηγορία τεχνικών καταπολέμησης της ανισορροπίας δεδομένων είναι μέσω του αλγορίθμου μάθησης. Ο σκοπός αυτού του είδους τεχνικών περιλαμβάνει την κατασκευή ή τροποποίηση αλγορίθμων ώστε να δίνουν περισσότερη σημασία στα δείγματα της κλάσης μειονότητας. Με αυτόν τον τρόπο ισορροπούν την διαφορά που υπάρχει ως προς τον αριθμό δειγμάτων μεταξύ των κλάσεων χωρίς να πειράξουν ή να μετασχηματίσουν τα δεδομένα. Τέτοιοι αλγόριθμοι συνήθως προσεγγίζουν διαφορετικά την διαδικασία της εκπαίδευσης ή τροποποιούν το τελικό μοντέλο στη φάση του ελέγχου με σκοπό να ελαττωθούν οι αρνητικές συνέπειες της ανισορροπίας. Παρακάτω θα δούμε μερικές από τις πιο γνωστές μεθόδους.

Μάθηση Ευαισθησίας Κόστους

Η πλέον πιο διαδεδομένη κατηγορία τέτοιων αλγορίθμων είναι η μάθηση με ευαισθησία στο κόστος [27]. Η κεντρική ιδέα της μάθησης με ευαισθησία στο κόστος είναι η κατασκευή μοντέλων που δίνουν περισσότερη προσοχή στα δείγματα της κλάσης μειονότητας, ώστε να βελτιωθεί η επίδοση του σε αυτά. Αυτό μπορεί να εφαρμοστεί με διάφορους τρόπους ανάλογα

με το είδος και την φύση του κάθε μοντέλου. Όσον αφορά τα νευρωνικά δίκτυα, αυτό συνήθως εφαρμόζεται με τη χρήση βαρών κλάσεων [57]. Κάθε κλάση του προβλήματος συνδέεται μοναδικά με ένα βάρος. Αυτά τα βάρη χρησιμοποιούνται στην συνάρτηση κόστους, ώστε τα δείγματα της κλάσης μειονότητας να επηρεάζουν το συνολικό κόστος περισσότερο. Δηλαδή βάζοντας μεγαλύτερο βάρος στα δείγματα των κλάσεων με λίγες παρατηρήσεις σε σχέση με τα υπόλοιπα, μετασχηματίζουν το συνολικό κόστος σε ένα που δίνει μεγαλύτερη έμφαση σε αυτά παρά τον μικρότερο αριθμό τους. Επιχειρώντας, λοιπόν, μέσω της βελτιστοποίησης να ελαχιστοποιήσουμε αυτή τη συνάρτηση κόστους, ωθούμε το νευρωνικό να μάθει να κάνει λιγότερα λάθη στη κλάση μειονότητας. Εκτός από την συνάρτηση κόστους, η μάθηση με ευαισθησία στο κόστος μπορεί να εφαρμοστεί και σε άλλα κομμάτια των νευρωνικών με παρόμοιο αποτέλεσμα. Για παράδειγμα μπορούν να εφαρμοστούν παρόμοια βάρη κλάσεων στον ρυθμό μάθησης ή στις παραγώγους για να προσαρμοστεί ο βελτιστοποιητής με αυτόν τον τρόπο.

Μέθοδος Μετακίνησης Κατωφλιού

Οι μέθοδοι μετακίνησης κατωφλιού είναι ένα είδος τεχνικής που μεταβάλει το όριο απόφασης ταξινόμησης ενός εκπαιδευμένου ταξινομητή ώστε να βελτιώσει το αποτέλεσμά του. Είναι γνωστές και ως κλιμάκωση μετά την εκπαίδευση. Στην περίπτωση των νευρωνικών δικτύων τέτοιοι αλγόριθμοι αλλάζουν τις πιθανότητες των κλάσεων που βγάζει σαν έξοδο το νευρωνικό βάση ενός κριτηρίου. Αυτό το κριτήριο επιλέγεται ανάλογα με την ανισορροπία που παρουσιάζει το κάθε σύνολο δεδομένων [59]. Σε διάφορες μελέτες [86] εφαρμόζεται υπολογίζοντας τις apriori πιθανότητες και μετασχηματίζοντας την έξοδο του νευρωνικού ανάλογα, υποθέτοντας ότι η έξοδος ακολουθεί Bayesian κατανομή.

Ταξινόμηση μίας κλάσης

Μία άλλη τεχνική της Μηχανικής Μάθησης είναι η Ταξινόμηση μιας κλάσης (One-Class Classification - OCC), η οποία βοηθά στο πρόβλημα της ανισορροπίας των κλάσεων. Η μέθοδος OCC [74, 20], γνωστή και ως ενιαία ταξινόμηση, μαθαίνει να ξεχωρίζει μία συγκεκριμένη κλάση (ή είδος κλάσης) ανάμεσα σε όλα, σε αντίθεση με το να μάθει να ξεχωρίζει όλες τις κλάσεις μεταξύ τους. Είναι εξαιρετικά χρήσιμη τεχνική σε περιπτώσεις ακραίας ανισορροπίας δεδομένων. Από αυτό η Ταξινόμηση μιας κλάσης μπορεί και να κατηγοριοποιηθεί σαν τεχνική Ανίχνευση Ανωμαλίας και κάποιες φορές λέγεται Ανίχνευση Καινοτομίας [44]. Μία απλή εφαρμογή της OCC είναι ο αλγόριθμος Support Vector Data Description (SVDD) [77], ο οποίος προσπαθεί να ταιριάξει τα δεδομένα στην μικρότερη δυνατή υπερσφαίρα, μία ιδέα βασισμένη στις Μηχανές διανυσμάτων υποστήριξης [19].

Δυναμική Δειγματοληψία

Μία ακόμη κατηγορία τεχνικών που βοηθούν στην ανισορροπία των δεδομένων είναι η Δυναμική Δειγματοληψία. Υπάρχουν πολλές προσεγγίσεις και μελέτες, πολλές από τις οπο-

ίες αναφέρθηκαν και αναπτύχθηκαν στο Κεφάλαιο 3. Όμως, εκεί εξετάστηκε ως προς την βελτίωση της διαδικασίας της εκπαίδευσης. Στο επόμενο υποκεφάλαιο θα αναπτύξουμε μία μέθοδο δειγματοληψίας που θα προσπαθήσει να καταπολεμήσει το πρόβλημα της ανισορροπίας.

Υβριδικές Μέθοδοι

Τέλος, είναι σημαντικό να αναφέρουμε ότι είναι δυνατό να συνδυάσουμε διάφορες από τις τεχνικές που περιγράφτηκαν προηγουμένως για να καταφέρουμε ένα καλύτερο αποτέλεσμα. Κάθε μέθοδος προσφέρει τα δικά της πλεονεκτήματα στην διαδικασία της εκπαίδευσης. Ένα παράδειγμα είναι το SMOTEBoost [16], το οποίο είναι ένας συνδυασμός από υπερδειγματοληψία με SMOTE και την τεχνική boosting. Μία άλλη υβριδική μέθοδος είναι το RUSBoost [88], το οποίο με τη σειρά του χρησιμοποιεί τον RUS αλγόριθμο αντί του SMOTE. Εκτός από την τεχνική του boosting, η τεχνική bagging μπορεί να συνδυαστεί με αλγόριθμους κατά της ανισορροπίας δεδομένων [70]. Στην μελέτη [65] παραθέτονται δύο υβριδικοί αλγόριθμοι, ο EasyEnsemble και ο BalanceCascade, που χρησιμοποιούν μία ομάδα από υποδειγματοληπτούμενα σύνολα δεδομένων και εφαρμόζουν έναν αριθμό από ταξινομητές με βάση τις ιδέες του ensembling.

4.2 Θορυβώδης Επιλογή Παρτίδας με Επανεισαγωγές

Στο προηγούμενο κεφάλαιο αναπτύχθηκε η μέθοδος Επιλογή Παρτίδας με Μεροληπτική Δειγματοληψία (Batch Selection with Biased Sampling - BSBS) [42], η οποία χρησιμοποιείται για να βελτιώσει την ταχύτητα σύγκλισης της εκπαίδευσης ενός νευρωνικού. Σε αυτό το κεφάλαιο θα αναπτύξουμε μία τεχνική δειγματοληψίας βασισμένη στο BSBS, που θα είναι εξειδικευμένη σε σύνολα δεδομένων που παρουσιάζουν κάποιου είδους ανισορροπία κλάσεων. Ο αλγόριθμος BSBS παρουσιάζει κάποια προβλήματα σε περιβάλλοντα ανισορροπίας, τα οποία θα εξηγήσουμε παρακάτω. Ο κύριος σκοπός του είναι η επιτάχυνση της εκπαίδευσης που με άλλα λόγια μπορούμε να το περιγράψουμε ως την γρηγορότερη μείωση του σφάλματος εκπαίδευσης σε κάθε εποχή. Αυτή η προσπάθεια, αν και είναι προσοδοφόρα στη σύγκλιση, δεν είναι απαραίτητα ωφέλιμη στην γενίκευση του νευρωνικού.

Κατά την διάρκεια της εκπαίδευσης ο αλγόριθμος βγάζει εκτός παρτίδας τα ¨εύκολα' δείγματα και τα αντικαθιστά με τα ¨δυσκολότερα¨. Όταν ένα δείγμα βγαίνει από την εκπαίδευση, αυτό σημαίνει ότι παρουσίασε χαμηλό σφάλμα. Όταν έχει χαμηλό σφάλμα ένα δείγμα για κάποιες εποχές, τότε ο αλγόριθμος θα το αφαιρέσει συνεχόμενες εποχές από την εκπαίδευση. Αυτό μπορεί να χειροτερέψει την γενίκευση του μοντέλου, αν υπάρχουν κάποια δείγματα που μείνουν εκτός παρτίδας για μεγάλο διάστημα. Η επίδοση της εκπαίδευσης δεν επηρεάζεται αρνητικά από αυτό αλλά σε άγνωστα δεδομένα η οποιαδήποτε αφαίρεση δειγμάτων μπορεί να μειώσει την ποικιλία του συνόλου δεδομένων και, επομένως, να μειώσει την αποδοτικότητά του ως προς την γενίκευση. Εκτός από αυτό, ένα δείγμα μένοντας πολλές συνεχόμενες φορές εκτός εκπαίδευσης έχει ως αποτέλεσμα το αποθηκευμένο σφάλμα του να αρχίζει να μην είναι

έγκυρο. Ο κίνδυνος έγκειται στο ότι με τις πολλές ανανεώσεις του νευρωνικού τα πραγματικά σφάλματα αυτών των δειγμάτων μπορεί να είναι υψηλότερα. Η ερώτηση είναι, όμως, με ποια δείγματα συμβαίνει αυτό το φαινόμενο;

Πολλά δείγματα μπορεί να βγουν από το σύνολο εκπαίδευσης προσωρινά αλλά κάποια στιγμή επιστρέφουν επειδή υπάρχουν άλλα που το νευρωνικό τα μαθαίνει καλύτερα, κάποια που είναι μέσα στην παρτίδα. Οπότε τα περισσότερα δείγματα θα ξεφύγουν από το σύνολο LL. Όμως, υπάρχουν δείγματα σε διάφορα σύνολα δεδομένων τα οποία παρουσιάζουν εξαιρετικά χαμηλές τιμές σφάλματος σε κάποια στιγμή της εκπαίδευσης. Αυτό το φαινόμενο μπορεί να συμβεί συχνότερα σε περιβάλλοντα με ανισορροπία δεδομένων. Για αυτό το λόγο προτείνουμε δύο αλλαγές στον αρχικό αλγόριθμο BSBS, που θα αναλυθούν παρακάτω.

4.2.1 Επιλογή Παρτίδας με Επανεισαγωγές

Βασισμένοι πάνω στο γεγονός ότι κάποια προβλήματα προέρχονται από την απομάκρυνση δειγμάτων για αρκετές εποχές από την εκπαίδευση, θα εισάγουμε την έννοια της επανεισαγωγής. Με άλλα λόγια, δείγματα που κανονικά ο αλγόριθμος θα τα έβγαζε εκτός σε μία συγκεκριμένη εποχή, θα προσπαθήσουμε να τα εντάξουμε πάλι μέσα. Αρχικά, μας ενδιαφέρουν δείγματα, τα οποία έχουν απομακρυνθεί από την εκπαίδευση για πολλές συνεχόμενες εποχές. Αυτά τα δείγματα έχουν μη έγκυρες αποθηκευμένες τιμές για το σφάλμα τους. Για να καταφέρουμε να το ελέγχουμε αυτό την ώρα της εκπαίδευσης, κατασκευάζουμε έναν πίνακα που αποθηκεύει τον αριθμό από συνεχόμενες φορές που ένα δείγμα κατηγοριοποιήθηκε στο σύνολο LL και βγήκε από την εκπαίδευση. Έπειτα, εισάγουμε ένα κατώφλι, το οποίο ελέγχει αν κάποιο δείγματα πρέπει να επανεισαχθεί στις επόμενες παρτίδες.

Έστω, ο πίνακας για τις συνεχόμενες φορές εκτός εκπαίδευσης να συμβολίζεται με ST και αυτό το κατώφλι με E_{sw}. Όταν το ST_i (του $i-$οστού δείγματος) ξεπεράσει το όριο, τότε αλλάζουμε την τιμή του σφάλματος στον BL πίνακα, ώστε να μην καταταχθεί στο LL. Αυτό το σφάλμα θα μπορούσαμε να το θέσουμε σε οποιαδήποτε τιμή μεγαλύτερη από τα k μικρότερα σφάλματα δειγμάτων για να σιγουρευτούμε ότι την επόμενη εποχή θα εισαχθεί στην εκπαίδευση. Σε αυτή την υλοποίηση επιλέξαμε να θέσουμε ως νέα τιμή, την μέση τιμή όλων των σφαλμάτων. Η λογική πίσω από αυτό είναι το νέο σφάλμα να βρίσκεται περίπου στο ενδιάμεσο, ώστε να μην καταταχθεί ούτε στο LL, αλλά ούτε και στο HL. Επομένως, θα εισαχθεί ακριβώς μία φορά στην επόμενη εποχή.

Θα ονομάσουμε αυτόν τον αλγόριθμο Επιλογή Παρτίδας με Επανεισαγωγές (BSBS with re-enters). Ο ολοκληρωμένος αλγόριθμος παρατίθεται στον Αλγόριθμο 2.

4.2.2 Θορυβώδης Επιλογή Παρτίδας

Εκτός από τις επανεισαγωγές, ο αλγόριθμος BSBS μπορεί να βελτιωθεί όσον αφορά την γενίκευση του και από μία διαφορετική σκοπιά. Έχει αναφερθεί σε προηγούμενο υποκεφάλαιο για τις δυσκολίες που παρουσιάζονται σε περιπτώσεις με ακραία ανισορροπία. Σε τέτοιες περι-

Algorithm 2: BSBS with re-enters

Parameters: dataset **D**, number of epochs **T**, number of datapoints **N**, batch size **B**, loss momentum **lm**, number of datapoints to be swapped **k**, indexes of samples **inds**, swapped times **ST**, re-enter threshold $\mathbf{E}_{sw}$.

Initializations: **inds**= [0, ... ,N-1], **lm**=0.7

for $t = 0$ *to* T **do**

$X_t = D[inds]$;

for $b = 0$ *to* N/B **do**

$b\_inds \leftarrow SelectBatch(X_t, b)$

$losses(b\_inds) \leftarrow ForwardPass(X_t[b\_inds])$

$UpdateWeights(losses)$

$BL[b\_inds] = losses$

$EMA[b\_inds] = lm * EMA[b\_inds] + (1 - lm) * losses$

end

$r\_inds \leftarrow select(ST \geq E_{sw})$

$\mu = \frac{\sum_i BL_i}{N}$

$BL[r\_inds] = \mu$

$HL = \hat{H}_k(EMA)$

$LL = \hat{L}_k(BL)$

$inds = HL \cup LL'$

$ST[LL] += 1$

$ST[inds] = 0$

end

πτώσεις υπάρχει πολύ μικρή ποικιλία από δείγματα της κλάσης μειονότητας, το οποίο καθιστά σχεδόν αδύνατο την εκμάθηση μιας κατανομής κοντά στην ιδανική. Εμπνευσμένο από το SMOTE [15] αλλά και τεχνικές ενίσχυσης δεδομένων [95, 56, 21] που χρησιμοποιούνται συχνά σε εφαρμογές Βαθιάς Μηχανικής Μάθησης, προτείνουμε μία παραλλαγή του αλγορίθμου BSBS, που ονομάζουμε Θορυβώδης Επιλογή Παρτίδας (Noisy BSBS).

Η Θορυβώδης Επιλογή Παρτίδας έχει ως στόχο να προσθέσει τεχνητό θόρυβο σε ένα κομμάτι των χαρακτηριστικών των δεδομένων, ώστε να δημιουργηθούν καινούργια δείγματα. Αρχικά, ο αλγόριθμος κατασκευάζει μία μάσκα τυχαίου θορύβου, M_i, για κάθε δείγμα, η οποία αποτελείται από άσσους και μηδενικά ($\{0, 1\}$) με ίση πιθανότητα. Αυτή η μάσκα αντιπροσωπεύει σε ποιο χαρακτηριστικό του καθενός δείγματος θα προστεθεί θόρυβος. Επειδή τα 0 και τα 1 είναι ισοπίθανα, αυτό σημαίνει ότι υπάρχει 50% πιθανότητα να προστεθεί θόρυβος σε κάποιο χαρακτηριστικό ή να μείνει ως έχει. Ο λόγος της προσθήκης του στοχαστικού θορύβου είναι ότι δεν επιθυμούμε να μάθει το μοντέλο έναν σταθερό θόρυβο, που μπορεί να

οδηγήσει σε υπερπροσαρμογή. Ο τύπος θορύβου που επιλέχθηκε είναι Γκαουσιανός θόρυβος (ακολουθώντας την κανονική κατανομή) με μηδενικό μέσο όρο και τυπική απόκλιση ίση με 0.2. Όμως, ο θόρυβος δεν προστίθεται σε όλο το σύνολο εκπαίδευσης. Επιλέγουμε το LL' σύνολο και προσθέτουμε θόρυβο πριν γίνουν οι ανταλλαγές στο τέλος της εποχής. Ως αποτέλεσμα, το τελικό σύνολο εκπαίδευσης για κάθε εποχή αποτελείται από το θορυβώδες LL' σύνολο και το αρχικό HL σύνολο. Αυτό μπορούμε να το αναπαραστήσουμε με την ακόλουθη σχέση:

$$X_t = Noisy(LL') \cup HL \tag{4.1}$$

Algorithm 3: Noisy BSBS

Parameters: dataset **D**, number of epochs **T**, number of datapoints **N**, batch size **B**, loss momentum **lm**, number of datapoints to be swapped **k**, indexes of samples **inds**, standard deviation of noise σ.

Initializations: **inds**= [0, ... ,N-1], **lm**=0.7, σ=0.2

for $t = 0$ *to* T **do**
- $X_t = D[inds]$;
- **if** $t > 0$ **then**
  - $M_t = random_{\{0,1\}}(prob = 0.5)$
  - $noise = sample(\mathcal{N}(0, \sigma^2))$
  - $X_t[LL'] = X_t[LL'] + M_t * noise$
- **end**
- **for** $b = 0$ *to* N/B **do**
  - $b\_inds \leftarrow SelectBatch(X_t, b)$
  - $losses(b\_inds) \leftarrow ForwardPass(X_t[b\_inds])$
  - $UpdateWeights(losses)$
  - $BL[b\_inds] = losses$
  - $EMA[b\_inds] = lm * EMA[b\_inds] + (1 - lm) * losses$
- **end**
- $HL = \hat{H}_k(EMA)$
- $LL = \hat{L}_k(BL)$
- $inds = HL \cup LL'$

end

Η προσθήκη θορύβου στο συγκεκριμένο σύνολο αποσκοπεί στο να μάθει το νευρωνικό και την αρχική μορφή των δύσκολων δειγμάτων του HL αλλά και την θορυβώδη. Έτσι, δημιουργούμε μία τεχνητή ποικιλία στα δύσκολα δείγματα, που αποτρέπει την υπερπροσαρμογή. Ο θόρυβος, όμως, προστίθεται και στα δείγματα μεσαίας δυσκολίας (στα δείγματα που δεν ανήκουν ούτε στο LL, ούτε στο HL). Με αυτόν τον τρόπο αυξάνουμε την πιθανότητα να βρούμε κάποιο άλλο δύσκολο δείγμα ανάμεσα από τα μεσαίας δυσκολίας, το οποίο έτυχε και είχε λίγο καλύτερο σφάλμα. Παρόλα αυτά, η προσθήκη θορύβου στα δεδομένα μπορεί να είναι

επικίνδυνο σε διάφορες περιπτώσεις δεδομένων. Σε διάφορους τύπους δεδομένων ο επιπλέον θόρυβος μπορεί να παραμορφώσει τα δεδομένα με τέτοιο τρόπο που να μην είναι ρεαλιστικά δείγματα. Για αυτόν τον λόγο η ένταση του θορύβου (δηλαδή το πλάτος του) είναι πολλή σημαντική υπερπαράμετρος. Στον Αλγόριθμο 3 παρουσιάζεται ο ολοκληρωμένος αλγόριθμος Θορυβώδης Επιλογή Παρτίδας.

Είναι σημαντικό να επισημάνουμε ότι οι δύο αλγόριθμοι που παρουσιάστηκαν προηγουμένως μπορούν να λειτουργήσουν και μαζί. Ο συνδυασμός τους ονομάζεται Θορυβώδης Επιλογή Παρτίδας με Επανεισαγωγές (Noisy Batch Selection with Biased Sampling with re-enters – NBSBS-R), που εφαρμόζει ανεξάρτητα την μέθοδο των επανεισαγωγών και την μέθοδο του θορύβου. Στο επόμενο υποκεφάλαιο θα αναλυθούν αυτές οι τρεις νέες τεχνικές και θα συγκριθούν μεταξύ τους και σε σχέση με τον αρχικό αλγόριθμο BSBS.

4.2.3 Πειραματική Διαδικασία

Η πειραματική διαδικασία θα επιχειρήσει να συγκρίνει τους αλγόριθμους BSBS αλλά και τις προεκτάσεις του σε περιβάλλοντα ανισορροπίας δεδομένων. Θα χρησιμοποιηθούν 11 τεχνικές μετασχηματισμού δεδομένων, οι οποίες έχουν αναφερθεί σε προηγούμενο υποκεφάλαιο, για να ελεγχθούν οι αλγόριθμοι σε διαφορετικές καταστάσεις. Επίσης, με αυτόν τον τρόπο διασφαλίζεται και η σταθερότητα των μοντέλων και των μεθόδων αντίστοιχα. Οι τεχνικές μετασχηματισμού δεδομένων που επιλέχθηκαν είναι οι εξής: 4 τεχνικές υποδειγματοληψίας (RUS, ENN, ClustCentroids, Tomek links), 5 τεχνικές υπερδειγματοληψίας (ROS, SMOTE, Adasyn, Borderline-SMOTE, KMeans-SMOTE) και δύο συνδυαστικές μέθοδοι (SMOTEENN, SMOTETomek). Ο συνδυασμός των τεχνικών μετασχηματισμού δεδομένων με τους αλγόριθμους Επιλογής Παρτίδας έχει ως σκοπό να δούμε αν η συνεργασία των δύο τύπων μεθόδων οδηγεί σε καλύτερες επιδόσεις γενίκευσης των μοντέλων.

Τα πειράματα έγιναν σε 4 σύνολα δεδομένων, όπου τα τρία από αυτά είναι εξ΄ αρχής ανισόρροπα και το τελευταίο είναι τεχνητά ανισόρροπο. Τα τρία ανισόρροπα σύνολα δεδομένων είναι το Ozone Level Detection Dataset [25], το Adult Dataset [25] και το Default of Creditcard clients [25] σύνολο δεδομένων. Το τεχνητά ανισόρροπο σύνολο δεδομένων, δηλαδή σύνολο δεδομένων που του προκαλέσαμε εμείς ανισορροπία στις κλάσεις του, είναι το MNIST [61]. Κάθε σύνολο δεδομένων έχει διαφορετικό λόγο ανισορροπίας, ώστε να παρατηρήσουμε τις συμπεριφορές των μεθόδων σε διαφορετικές συνθήκες. Ο λόγος ανισορροπίας ορίζεται ως το το κλάσμα του πλήθους των δειγμάτων της κλάσης πλειονότητας ως προς το πλήθος των δειγμάτων της κλάσης μειονότητας. Λεπτομέρειες για το κάθε σύνολο δεδομένων θα παρατεθούν στις παρακάτω παραγράφους. Στο Παράρτημα Α΄ βρίσκονται λεπτομέρειες για τις αρχιτεκτονικές των νευρωνικών δικτύων που χρησιμοποιήθηκαν για τα διαφορετικά σύνολα δεδομένων. Σε κάθε περίπτωση τα βάρη αρχικοποιήθηκαν στις ίδιες τιμές για να είναι δίκαια η σύγκριση των μεθόδων. Τα νευρωνικά εκπαιδεύτηκαν με την χρήση του βελτιστοποιητή Adam [51] με ρυθμό μάθησης ίσο με 0.001. Η συνάρτηση κόστους είναι η κατηγορική σταυροειδής εντροπία. Στα πρώτα τρία σύνολα δεδομένων τα νευρωνικά εκπαιδεύτηκαν για 10 εποχές, ενώ για το

MNIST για 15.

Αξιολόγηση

Για να αξιολογήσουμε καλύτερα και πιο δίκαια τα εκπαιδευμένα μοντέλα χρησιμοποιήθηκαν οι ακόλουθες μετρικές: Ακρίβεια (Accuracy), Ισορροπημένη Ακρίβεια (Balanced Accuracy), F1 macro, ROC AUC macro (Εμβαδόν κάτω από την χαρακτηριστική καμπύλη δέκτη - Receiver Operating Curve - Area Under Curve). Αυτές οι μετρικές μπορούν να περιγράψουν ικανοποιητικά την συνολική επίδοση ενός μοντέλου και πόσο καλά συμπεριφέρεται σε δεδομένα με κάποια ανισορροπία. Για να υπολογίσουμε αυτές τις μετρικές, πρέπει να υπολογίσουμε πρώτα κάποιες βοηθητικές μετρικές:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.4}$$

όπου με TP συμβολίζουμε τα Αληθώς Θετικά (True Positives), FP τα Ψευδώς Θετικά (False Positive), TN τα Αληθώς Αρνητικά (True Negative) και FN είναι τα Ψευδώς Αρνητικά (False Negative). Η ακρίβεια, η ισορροπημένη ακρίβεια και η μετρική F1 μπορούν τώρα να γραφούν με τον παρακάτω τρόπο.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

$$BalancedAccuracy = \frac{1}{2} * (Recall + Specificity) \tag{4.6}$$

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4.7}$$

Για την μετρική ROC AUC πρέπει να υπολογίσουμε το εμβαδόν που περικλείεται κάτω από την καμπύλη του ρυθμού των Αληθώς Θετικών συναρτήσει του ρυθμού των Ψευδώς Θετικών. Χρησιμοποιήσαμε την "macro" έκδοση των μετρικών, γιατί απεικονίζει καλύτερα την επίδοση των μοντέλων σε ανισόρροπα σύνολα δεδομένων. Στα πειράματα που θα ακολουθήσουν η υπερπαράμετρος k θα απεικονίζει το ποσοστό των δειγμάτων που αντικαθιστώνται στο σύνολο εκπαίδευσης. Σε κάθε σύνολο δεδομένων τέσσερεις διαφορετικές τιμές k ελέγχθηκαν $(0, 0.1, 0.15, 0.2)$, με το 0 να σημαίνει ότι δεν χρησιμοποιήθηκε καθόλου ο αλγόριθμος

BSBS. Η πρώτη σύγκριση αφορά την επίδοση του αλγόριθμου BSBS σε σχέση με τις καινούργιες προσθήκες του. Η δεύτερη σύγκριση περιλαμβάνει την συμπεριφορά των καινούργιων αλγορίθμων σε συνδυασμό με τις τεχνικές μετασχηματισμού δεδομένων.

Στους παρακάτω πίνακες θα δείξουμε τα αποτελέσματα για κάθε σύνολο δεδομένων. Για κάθε τεχνική μετασχηματισμού δεδομένων, η πρώτη γραμμή του πίνακα δείχνει ένα τρέξιμο χωρίς BSBS, ενώ η δεύτερες γραμμές δείχνουν το καλύτερο τρέξιμο ανάμεσα σε όλες τις υπερπαραμέτρους και όλες τις παραλλαγές του αλγορίθμου (με επανεισαγωγές ή θορυβώδης).

Ταξινόμηση Στάθμης του Όζοντος

Το πρώτο σύνολο δεδομένων που εξετάστηκε είναι το Ozone Level Dataset [25]. Το σύνολο δεδομένων αποτελείται από 2536 δείγματα με 73 χαρακτηριστικά το καθένα. Υπάρχουν δύο κλάσεις, η ημέρα όζοντος και η κανονική ημέρα, όπου η πρώτη κλάση να είναι η κλάση μειονότητας. Αφού αφαιρέσαμε κάποια δείγματα επειδή είχαν πολλές πεδία που έλειπαν, ο τελικός λόγος ανισορροπίας ήταν γύρω στο 32 : 1. Το σύνολο δεδομένων δεν έχει κάποιο σταθερό διαχωρισμό σε σύνολο εκπαίδευσης και ελέγχου, οπότε χρησιμοποιήσαμε 5-πτυχη Στρωματοποιημένη Διασταυρούμενη Επικύρωση για να την σωστότερη αξιολόγηση. Πριν την εκπαίδευση τα δεδομένα κανονικοποιήθηκαν με μέσο όρο στο 0 και τυπική απόκλιση στο 1. Η αρχιτεκτονική του νευρωνικού φαίνεται στον Πίνακα Α΄.1. Ο Πίνακας 4.1 δείχνει τα αποτελέσματα των πειραμάτων.

Βλέπουμε ότι στις περισσότερες περιπτώσεις ο αλγόριθμος BSBS βελτιώνει τις μετρικές, και κυρίως τις μετρικές F1 και Ισορροπημένη Ακρίβεια (IA), που αφορούν μη ισορροπημένα δεδομένα. Σε 9 από τις 12 περιπτώσεις οι παραλλαγές του BSBS απέδωσαν καλύτερα από τον αρχικό αλγόριθμο. Τα πειράματα με τις μεθόδους ClusterCentroids, Tomek links και Adasyn έδειξαν ότι ο αρχικός αλγόριθμος BSBS είχε λίγο καλύτερες μετρικές. Όσον αφορά τις καλύτερες επιδόσεις των μετρικών, η μέθοδος Tomek links σε συνδυασμό με BSBS(0.15) είχε την υψηλότερη ακρίβεια. Η μέθοδος ROS σε συνδυασμό με την NBSBS-R(0.1) είχε την καλύτερη ισορροπημένη ακρίβεια, ενώ το SMOTE με NBSBS-R(0.2) έφτασε την καλύτερη τιμή της F1 μετρικής.

Σύνολο δεδομένων Adult

Το δεύτερο σύνολο δεδομένων είναι το Adult [25], το οποίο είναι γνωστό και ως το σύνολο δεδομένων εισοδημάτων απογραφής. Ο σκοπός του είναι να προβλέψει αν το ετήσιο εισόδημα ενός ενήλικα είναι πάνω από 50 χιλιάδες ή οχι, χρησιμοποιώντας πληροφορίες από την απογραφή. Αποτελείται από 32561 δείγματα εκπαίδευσης και 16281 δείγματα ελέγχου με 14 χαρακτηριστικά για το κάθε δείγμα. Τα δείγματα που έχουν κενές τιμές σε χαρακτηριστικά τους αφαιρέθηκαν από το σύνολο δεδομένων εξαιτίας του μικρού τους αριθμού. Ένας λογαριθμικός μετασχηματισμός εφαρμόστηκε στα χαρακτηριστικά 'capital-gain' και 'capital-loss' λόγω του μεγάλου εύρους τιμών τους. Έπειτα, τα αριθμητικά χαρακτηριστικά κανονικοποι-

ήθηκαν με την μέθοδο Min-Max, ενώ τα κατηγορικά χαρακτηριστικά μετασχηματίστηκαν με την μέθοδο One-hot. Τα δεδομένα είναι χωρισμένα σε δύο κλάσεις, μία κλάση πλειονότητας και μία μειονότητας. Ο λόγος ανισορροπίας είναι 3 : 1. Η αρχιτεκτονική του δικτύου παρουσιάζεται στον Πίνακα Α΄.1 και είναι η ίδια με αυτή που χρησιμοποιήθηκε στο σύνολο δεδομένων Ozone. Ο Πίνακας 4.2 συνοψίζει τα αποτελέσματα για τα πειράματα αυτών των δεδομένων. Παρατηρούμε πως στις περισσότερες περιπτώσεις οι επεκτάσεις του BSBS επιδεικνύουν υψηλότερη ισορροπημένη ακρίβεια και F1. Η καλύτερη F1 τιμή επιτεύχθηκε από τον NBSBS-R(0.2) χωρίς κάποιο μετασχηματισμό δεδομένων, ενώ όσον αφορά την ισορροπημένη ακρίβεια καλύτερη επίδοση είχε ο NBSBS(0.1) μαζί τον RUS μετασχηματισμό. Οι μετρικές ROC-AUC έχουν αρκετά κοντινές τιμές χωρίς καμία να ξεχωρίζει ιδιαίτερα. Η μόνη περίπτωση που ο αρχικός αλγόριθμος ήταν καλύτερος από τις νέες παραλλαγές του ήταν στην περίπτωση του CluserCentroids μετασχηματισμού. Αυτό αποτελεί και μία ομοιότητα με την αντίστοιχη συμπεριφορά στο σύνολο δεδομένων του Ozone.

Σύνολο δεδομένων για αθέτηση πληρωμής

Ένα άλλο σύνολο δεδομένων που εξετάστηκε είναι το Default of Creditcard clients [25]. Αυτά τα δεδομένα έχουν πληροφορίες από πληρωμές πελατών στην Ταϊβάν και έχει ως σκοπό να προβλέψει την πιθανότητα αθέτησης πληρωμής πιστωτικής κάρτας. Τα δεδομένα αποτελούνται από 30000 δείγματα με 24 χαρακτηριστικά το καθένα. Παρόμοια με το σύνολο δεδομένων Ozone, ούτε αυτό το σύνολο δεδομένων είναι χωρισμένο σε σύνολο εκπαίδευσης και ελέγχου. Για αυτό το λόγο εφαρμόσαμε ξανά 5-πτυχη Στρωματοποιημένη Διασταυρούμενη Επικύρωση. Υπάρχουν πάλι δύο κλάσεις με μεγάλη διαφορά δειγμάτων μεταξύ τους, η οποία διαμορφώνει τον λόγο ανισορροπίας στο 3.5 : 1. Η αρχιτεκτονική του νευρωνικού που χρησιμοποιήθηκε για τα πειράματα παραθέτεται στον Πίνακα Α΄.2. Εφαρμόστηκε στα δεδομένα η κωδικοποίηση One-hot, καθώς και μία κανονικοποίηση. Στον Πίνακα 4.3 μπορούμε να δούμε τα αποτελέσματα από τα πειράματα. Οι παραλλαγές του BSBS φαίνονται να απέδωσαν καλύτερα από τον αρχικό αλγόριθμο, εκτός από την περίπτωση του SMOTE, όπου ο BSBS(0.2) είχε υψηλότερη επίδοση. Σχετικά με τις συνολικές επιδόσεις, ο συνδυασμός RUS με BSBS-R(0.1) έφτασε την υψηλότερη ισορροπημένα ακρίβεια, ενώ ο συνδυασμός ENN με BSBS-R(0.15) είχε το καλύτερο αποτέλεσμα στην F1 μετρική.

MNIST

Το σύνολο δεδομένων, που χρησιμοποιήσαμε για την τεχνητή ανισορροπία, είναι το MNIST [61]. Αποτελείται από 70000 ασπρόμαυρες εικόνες, μεγέθους 28×28, από αριθμούς ζωγραφισμένους με το χέρι. Οι 60000 από αυτές τις εικόνες χρησιμοποιούνται για την εκπαίδευση και οι υπόλοιπες 10000 για τον έλεγχο. Υπάρχουν 10 κλάσεις ($0 - 9$ ψηφία) και όλες είναι σχετικά ισορροπημένες. Για να εισάγουμε ανισορροπία στις κλάσεις ακολουθήσαμε το σχέδιο που παρουσιάστηκε στη μελέτη [11]. Για κάθε μία από τις κλάσεις ένα νέο σύνολο δεδομένων δημιουργήθηκε, όπου από την επιλεγμένη κλάση αφαιρείται το 90% των δειγμάτων. Με άλλα

λόγια, το σύνολο δεδομένων MNIST θα μετασχηματιστεί σε 10 σύνολα, όπου στο i-οστό σύνολο θα παρουσιάζεται μειονότητα στην i-οστή κλάση.

Με την βοήθεια αυτού του πλάνου, τα πειράματα θα αναδείξουν πιο αξιόπιστα αποτελέσματα, αφού θα εξετάσουμε το πρόβλημα της ανισορροπίας από την σκοπιά διαφορετικών κλάσεων. Όσον αφορά την προεπεξεργασία των δεδομένων, οι εικόνες κανονικοποιήθηκαν στον εύρος $[0, 1]$ διαιρώντας τις με 255. Η αρχιτεκτονική του Συνελικτικού Νευρωνικού Δικτύου που χρησιμοποιήθηκε βρίσκεται στον Πίνακα Α΄.3. Στον Πίνακα Β΄.1 φαίνονται τα αποτελέσματα των πειραμάτων. Οι μετρικές που αναγράφονται είναι ο μέσος όρος των τιμών από τα πειράματα που διεξάχθηκαν στα 10 διαφορετικά σύνολα. Παρατηρούμε ότι ο συνδυασμός ROS με τον NBSBS-R(0.1) αποδίδει καλύτερα σε όλες τις μετρικές. Για κάθε μετασχηματισμό που δοκιμάστηκε, οι παραλλαγές του BSBS έδειξαν καλύτερη επίδοση σε σχέση με τον αρχικό. Στις περισσότερες περιπτώσεις ο BSBS-R φαίνεται να είχε υψηλότερες μετρικές σε σχέση με τον NBSBS, εκτός από τις περιπτώσεις των RUS, ClusterCentroids, Tomek links, SMOTETomek.

Πίνακας 4.1: Τα αποτελέσματα των πειραμάτων για το σύνολο δεδομένων Ozone.

Dataset Transformation	BSBS(k)	Re-enters	Noisy	Accuracy	Balanced Accuracy	F1-macro	ROC-AUC
None	-	-	-	96.96	51.79	52.43	88.63
	0.1	✓	✓	97.13	57.26	59.86	89.32
RUS	-	-	-	76.45	78.24	51.87	87.96
	0.1	✓	-	81.97	81.49	81.97	88.06
ENN	-	-	-	96.96	52.45	53.41	88.46
	0.1	✓	-	97.07	62.58	64.67	89.47
ClusterCentroids	-	-	-	64.60	77.24	45.64	86.74
	0.1	-	-	64.71	77.32	45.73	87.14
Tomek Links	-	-	-	96.96	52.42	53.34	88.77
	0.15	-	-	**97.18**	56.73	59.54	89.07
ROS	-	-	-	93.77	83.60	66.79	90.26
	0.1	✓	✓	95.56	**84.21**	67.98	90.30
SMOTE	-	-	-	94.15	82.52	67.38	90.19
	0.2	✓	✓	95.72	83.49	**70.40**	90.31
Adasyn	-	-	-	93.93	83.15	66.79	90.21
	0.2	-	-	94.69	83.79	67.97	90.44
Borderline-SMOTE	-	-	-	94.75	80.12	67.81	**90.61**
	0.2	✓	-	96.10	80.30	70.20	90.40
KMeans-SMOTE	-	-	-	95.13	79.14	67.77	89.60
	0.1	✓	✓	95.94	77.82	69.38	89.66
SMOTEENN	-	-	-	91.34	82.44	64.50	89.86
	0.15	✓	✓	93.50	83.56	67.17	89.92
SMOTETomek	-	-	-	94.21	82.09	66.99	90.29
	0.1	✓	✓	95.18	82.87	68.13	90.32

Πίνακας 4.2: Τα αποτελέσματα των πειραμάτων για το σύνολο δεδομένων Adult.

Dataset Transformation	BSBS(k)	Re-enters	Noisy	Accuracy	Balanced Accuracy	F1-macro	ROC-AUC
None	-	-	-	**84.41**	78.39	78.43	89.82
	0.2	✓	✓	84.01	79.83	**78.52**	89.57
RUS	-	-	-	81.20	81.35	77.21	89.69
	0.1	-	✓	82.35	**81.41**	77.59	89.47
ENN	-	-	-	82.31	81.38	77.83	89.65
	0.1	-	✓	82.51	81.14	77.72	89.57
ClusterCentroids	-	-	-	77.96	79.31	74.08	86.90
	0.15	-	-	78.24	79.22	74.47	86.39
Tomek Links	-	-	-	84.38	78.94	78.39	89.78
	0.15	✓	✓	83.71	80.47	78.10	89.52
ROS	-	-	-	80.74	81.33	76.81	89.84
	0.2	✓	✓	82.01	81.04	77.13	89.63
SMOTE	-	-	-	81.16	81.23	77.14	89.66
	0.2	✓	✓	82.55	81.87	77.26	89.52
Adasyn	-	-	-	79.26	80.96	75.79	89.47
	0.15	✓	✓	81.48	80.92	76.45	89.39
Borderline-SMOTE	-	-	-	78.49	80.91	75.02	89.06
	0.15	✓	-	80.68	81.09	76.08	89.19
KMeans-SMOTE	-	-	-	82.26	80.23	77.42	89.33
	0.15	✓	✓	82.43	80.42	77.58	**89.89**
SMOTEENN	-	-	-	81.36	81.17	76.94	89.53
	0.15	✓	-	80.50	81.28	76.98	89.37
SMOTETomek	-	-	-	80.69	81.33	76.83	89.62
	0.1	✓	-	82.23	81.30	77.34	89.71

Πίνακας 4.3: Τα αποτελέσματα των πειραμάτων για το σύνολο δεδομένων Default of Credit-card clients.

Dataset Transformation	BSBS(k)	Re-enters	Noisy	Accuracy	Balanced Accuracy	F1-macro	ROC-AUC
None	-	-	-	**82.06**	66.00	68.45	77.52
	0.15	-	✓	81.62	67.95	69.56	76.58
RUS	-	-	-	77.66	70.73	69.21	77.37
	0.1	✓	-	77.77	**71.28**	69.02	76.93
ENN	-	-	-	80.36	70.42	70.47	77.50
	0.15	✓	-	80.12	70.61	**70.98**	77.26
ClusterCentroids	-	-	-	47.10	61.06	46.64	73.18
	0.15	-	✓	49.87	61.89	48.94	73.08
Tomek Links	-	-	-	82.00	66.77	69.03	77.51
	0.15	✓	✓	81.74	67.55	69.37	77.15
ROS	-	-	-	77.46	70.67	69.10	**77.54**
	0.1	-	✓	77.45	71.18	69.24	76.74
SMOTE	-	-	-	81.74	67.97	69.72	77.25
	0.2	-	-	81.45	68.30	69.56	76.85
Adasyn	-	-	-	81.75	67.74	69.63	77.03
	0.15	✓	-	81.38	68.48	69.31	76.58
Borderline-SMOTE	-	-	-	81.76	68.05	69.69	77.33
	0.1	✓	-	81.42	68.46	69.71	76.94
KMeans-SMOTE	-	-	-	81.68	67.87	69.53	76.99
	0.1	✓	✓	81.71	68.06	69.69	76.93
SMOTEENN	-	-	-	78.74	70.17	69.32	76.58
	0.2	✓	✓	79.71	70.10	69.81	76.42
SMOTETomek	-	-	-	81.76	67.50	69.37	77.42
	0.2	✓	✓	81.75	67.79	69.59	77.09

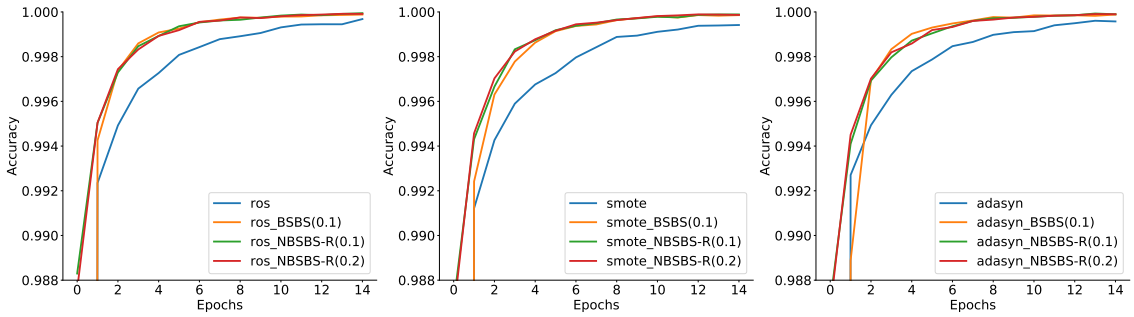
Πίνακας 4.4: Τα αποτελέσματα των πειραμάτων για το σύνολο δεδομένων MNIST.

Dataset Transformation	BSBS(k)	Re-enters	Noisy	Accuracy	Balanced Accuracy	F1-macro	ROC-AUC
None	-	-	-	98.92	98.91	98.91	99.98
	0.1	✓	-	99.09	99.08	99.08	**99.99**
RUS	-	-	-	96.86	96.85	96.85	99.68
	0.15	-	✓	98.08	98.08	98.07	99.97
ENN	-	-	-	89.21	89.24	88.42	95.09
	0.1	✓	-	98.77	98.76	98.76	99.98
ClusterCentroids	-	-	-	96.02	95.99	96.03	99.90
	0.2	-	✓	97.83	97.81	97.83	99.96
Tomek Links	-	-	-	98.90	98.89	98.89	99.98
	0.2	-	✓	99.09	99.08	99.08	**99.99**
ROS	-	-	-	99.00	98.99	98.99	99.98
	0.1	✓	✓	**99.17**	**99.16**	**99.16**	**99.99**
SMOTE	-	-	-	98.97	98.96	98.96	99.98
	0.2	✓	-	99.12	99.11	99.11	**99.99**
Adasyn	-	-	-	98.93	98.92	98.92	99.98
	0.2	✓	✓	99.10	99.09	99.09	**99.99**
Borderline-SMOTE	-	-	-	98.91	98.90	98.90	99.97
	0.2	✓	-	99.07	99.06	99.06	**99.99**
KMeans-SMOTE	-	-	-	98.90	98.89	98.89	99.98
	0.2	✓	-	99.09	99.08	99.08	**99.99**
SMOTEENN	-	-	-	80.36	80.50	78.84	89.84
	0.1	✓	-	98.79	98.78	98.78	99.98
SMOTETomek	-	-	-	98.97	98.96	98.96	99.98
	0.2	-	✓	99.13	99.12	99.12	**99.99**

4.2.4 Επιδράσεις στην σύγκλιση

Ο αλγόριθμος BSBS, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, έχει ως βασικό σκοπό την βελτίωση της ταχύτητας σύγκλισης ενός νευρωνικού δικτύου. Με την προσθήκη των νέων τεχνικών (επανεισαγωγές, θόρυβος) και την δημιουργία νέων αλγορίθμων, είναι αναγκαίο να εξετάσουμε πώς αυτά επηρεάζουν την σύγκλιση του μοντέλου. Για να μετρήσουμε την ταχύτητα σύγκλισης, εξετάζουμε τον αριθμό των εποχών που χρειάζεται το νευρωνικό για να φτάσει μία επιθυμητή τιμή σφάλματος (ή τιμή κάποιας μετρικής, όπως η ακρίβεια). Στην δικιά μας περίπτωση θα εξετάσουμε την σύγκλιση ως προς την ακρίβεια του μοντέλου κατά τη διάρκεια των εποχών.

Για αυτόν τον σκοπό επιλέχθηκαν τρεις από τις τεχνικές μετασχηματισμού δεδομένων (ROS, SMOTE, Adasyn) και χρησιμοποιήθηκαν σε πειράματα ώστε να εξεταστεί η ταχύτητα σύγκλισης. Επιλέχθηκαν αυτές οι τεχνικές επειδή αποτελούν τεχνικές υπερδειγματοληψίας και θα ήταν καταλληλότερες για την καλύτερη σύγκριση λόγω των περισσότερων δειγμάτων. Τέσσερεις διαφορετικές εκδοχές των αλγορίθμων εκτελέστηκαν και είναι οι εξής: BSBS, BSBS(0.1), NBSBS-R(0.1), NBSBS-R(0.2). Στο Σχήμα 4.3 παρατηρούμε τις καμπύλες εκπαίδευσης (ακρίβεια εκπαίδευσης ως προς τις εποχές). Όλες οι εκδοχές του BSBS συγκλίνουν γρηγορότερα σε σχέση με την απλή μέθοδο χωρίς BSBS. Μπορούμε, επίσης, να δούμε ότι στην περίπτωση του SMOTE ότι το NBSBS-R είναι ελάχιστα πιο γρήγορο στις πρώτες εποχές. Αυτό δείχνει ότι οι καινούργιοι αλγόριθμοι είναι ικανοί να βελτιώσουν την ταχύτητα σύγκλισης σε μερικές περιπτώσεις. Συνολικά, είναι εμφανές ότι οι παραλλαγές δεν επηρεάζουν αρνητικά τον αλγόριθμο BSBS όσον αφορά την ταχύτητα σύγκλισης, αλλά, σε αντίθεση, μπορεί να είναι και ευεργετικοί.



Σχήμα 4.3: Οι καμπύλες εκπαίδευσης των νευρωνικών στο MNIST για τρεις διαφορετικές μεθόδους υπερδειγματοληψίας (ROS στα αριστερά, SMOTE στο κέντρο, Adasyn στα δεξιά).

Συζήτηση για τις Υπερπαραμέτρους

Οι υπερπαράμετροι των προτεινούμενων αλγορίθμων είναι οι εξής: το κατώφλι των επανεισαγωγών E_{sw} και η διακύμανση του θορύβου σ. Το E_{sw} αρχικοποιήθηκε στην τιμή 0.2 των συνολικών εποχών σταθερά για όλη τη διάρκεια εκπαίδευσης. Για τα πειράματα που πραγματοποιήθηκαν, αυτές οι τιμές είναι αποτελεσματικές αλλά σε περιπτώσεις με εκπαιδεύσεις

με πολλές εποχές, θα ήταν χρήσιμο να δοκιμαστούν και μικρότερες τιμές της τάξης του 0.1. Όσον αφορά την διακύμανση του θορύβου σ την αρχικοποιήσαμε στην τιμή 0.2. Σε σύνολα δεδομένων που δεν είναι κανονικοποιημένα αυτή η τιμή θα πρέπει να αλλάζει ανάλογα με το μέγεθος του καθενός χαρακτηριστικού. Η τιμή αυτή επιλέχθηκε μετά από δοκιμές.

Κεφάλαιο 5

Προσαρμοστικοί Αλγόριθμοι Βελτιστοποίησης

Στο κεφάλαιο αυτό θα ασχοληθούμε με την διαδικασία της Βελτιστοποίησης των Νευρωνικών Δικτύων. Όπως αναφέρθηκε στο Κεφάλαιο 2 η διαδικασία τη βελτιστοποίησης περιλαμβάνει τους αλγόριθμους με τους οποίους ανανεώνονται τα βάρη ώστε να μειωθεί η συνάρτηση κόστους. Θα αναφερθούμε σε διάφορες τεχνικές και πως αυτές βελτιώνουν την ποιότητα ενός δικτύου.

5.1 Στοχαστική Βελτιστοποίηση

Σε ένα γενικό σενάριο βελτιστοποίησης καλούμαστε να ελαχιστοποιήσουμε (ή να μεγιστοποιήσουμε) μία συνάρτηση. Έστω ότι συμβολίζουμε αυτή τη συνάρτηση ως $f : \Re^d \to \Re$. Σκοπός της βελτιστοποίησης είναι να βρούμε τις κατάλληλες παραμέτρους w της συνάρτησης f δεδομένου κάποιων εισόδων x, ώστε να πετύχουμε την βέλτιστη ελαχιστοποίηση (ή μεγιστοποίηση). Επομένως, ο στόχος αυτός μπορεί να γραφεί ως εξής:

$$w^* = \underset{w}{\operatorname{argmin}} f(w \mid x) \tag{5.1}$$

Το παραπάνω γενικό σχήμα βελτιστοποίησης μπορεί να χρησιμοποιηθεί και για την περίπτωση της βελτιστοποίησης ενός νευρωνικού δικτύου. Τα νευρωνικά δίκτυα χρησιμοποιούν, κατά κύριο λόγο, αλγορίθμους μάθησης, οι οποίοι βασίζονται πάνω στην εύρεση και χρήση της παραγώγου. Ο υπολογισμός της γίνεται ως προς την συνάρτηση κόστους και υπακούει στον τύπο 2.9. Έχοντας την παράγωγο μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο Καθόδου Κλίσης για την ανανέωση των βαρών. Όμως, σε εφαρμογές Βαθιάς Μηχανικής Μάθησης ο αρχικός αλγόριθμος της καθόδου κλίσης δεν είναι εφικτός λόγω της μεγάλης ποσότητας των δεδομένων. Η εκπαίδευση των νευρωνικών σε παρτίδες έχει αλλάξει σε ένα βαθμό την διαδικασία της βελτιστοποίησης αλλά και τους βελτιστοποιητές.

Ο χωρισμός σε παρτίδες μετατρέπει τον στόχο της βελτιστοποίησης σε έναν μερικό στόχο. Με άλλα λόγια η στοχαστική βελτιστοποίηση προσπαθεί να εφαρμόσει τον αλγόριθμο βελτιστοποίησης με δεδομένα ένα μικρό κομμάτι της εισόδου (παρτίδα) με την γνώση ότι κατά μέσο όρο, μετά από πολλές επαναλήψεις, θα συγκλίνει με την ίδια επιτυχία. Αυτό έχει αποδειχθεί ότι ωφελεί την διαδικασία της μάθησης ενός δικτύου. Η Κάθοδος Κλίσης μετασχηματίζεται σε Στοχαστική Κάθοδος Κλίσης (Stochastic Gradient Descent - SGD) και μπορούμε να δούμε τον κανόνα της από κάτω:

$$w_{t+1} = w_t \overbrace{-\alpha_t \underbrace{\nabla f_B(w_t)}_{g_t}}^{u_t} \tag{5.2}$$

όπου α_t είναι ο ρυθμός μάθησης και B είναι η παρτίδα. Για απλότητα θα συμβολίζουμε την παράγωγο στην t επανάληψη του αλγορίθμου ως g_t και την συνολική ανανέωση των βαρών ως u_t.

Τα τελευταία χρόνια ο αλγόριθμος SGD έχει αναπτυχθεί αρκετά και έχουν δημιουργηθεί πολλαπλοί αλγόριθμοι βελτιστοποίησης που βασίζονται πάνω του. Κάθε καινούργια παραλλαγή προσφέρει κάτι νέο στην διαδικασία της εκπαίδευσης. Οι περισσότερες προσεγγίσεις και βελτιώσεις αφορούν μία από τις πιο σημαντικές παραμέτρους της εκπαίδευσης ενός νευρωνικού, τον ρυθμό μάθησης. Ο ρυθμός μάθησης επηρεάζει σε μεγάλο βαθμό την ποιότητα και την επίδοση του μοντέλου. Για παράδειγμα επιλέγοντας έναν υψηλό ρυθμό μάθησης υπάρχει κίνδυνος σε απότομες ταλαντώσεις των καμπυλών μάθησης και, τελικά, σε μη σύγκλιση του μοντέλου. Από την άλλη, πολύ μικρές τιμές ρυθμού μάθησης κάνουν την εκπαίδευση εξαιρετικά αργή. Άλλος κίνδυνος για χαμηλό ρυθμό μάθησης είναι ο εγκλωβισμός σε πιθανά τοπικά ελάχιστα [67, 41].

Μία από τις πιο κοινές πρακτικές επιλογής του ρυθμού μάθησης είναι να αρχικοποιείται σε μία τιμή και να το ελαττώνουμε κατά την διάρκεια της εκπαίδευσης [87, 52]. Όμως, δεν είναι η βέλτιστη τακτική σε κάθε περίπτωση [100]. Μία πληθώρα από μελέτες έχουν δημοσιευθεί για την βέλτιστη τιμή του ρυθμού μάθησης. Κάποιες από τις καλύτερες και πιο χρησιμοποιημένες θα αναφερθούν παρακάτω.

5.2 Προσαρμοστικοί Βελτιστοποιητές

Η πιο γνωστή μέθοδος για την επιλογή του ρυθμού μάθησης είναι με την βοήθεια χρονοδιαγραμμάτων. Ένα χρονοδιάγραμμα είναι μία προκαθορισμένη στρατηγική επιλογής ρυθμού μάθησης για κάθε σημείο της διαδικασίας της εκπαίδευσης. Για κάθε εποχή, δηλαδή, υπάρχει μία καθορισμένη τιμή για τον ρυθμό μάθησης, η οποία ακολουθεί κάποιους προδιαγεγραμμένους κανόνες. Μία από τις πιο παλιές μορφές αυτής της μεθόδου αναφέρεται στο [87]. Επίσης, επισημαίνεται ότι ο ρυθμός μάθησης πρέπει να υπακούει στις δύο παρακάτω εξισώσεις:

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad \text{και} \quad \sum_{t=1}^{\infty} {\alpha_t}^2 < \infty \tag{5.3}$$

Άλλη δημοφιλής στρατηγική χρονοδιαγράμματος είναι να ξεκινά ο ρυθμός μάθησης σε κάποια υψηλή τιμή και να υποδιπλασιάζεται κάθε μερικές επαναλήψεις για να σιγουρέψει ότι θα φτάσει αρκετά κοντά στα βέλτιστα βάρη [9]. Η απλότητα της μεθόδου οδήγησε στο να χρησιμοποιηθεί αρκετά ακόμα και στις πιο πολυχρησιμοποιημένες αρχιτεκτονικές [39]. Μία άλλη αποδοτική στρατηγική είναι να αρχίζει η εκπαίδευση με σταθερό ρυθμό και, όταν το σφάλμα σταματήσει να μειώνεται, να μειώνουμε τον ρυθμό μάθησης κατά έναν σταθερό παράγοντα σε κάθε επανάληψη [17].

Μία υπόθεση που συνήθως παίρνουμε ως δεδομένο σε διάφορους αλγορίθμους βελτιστοποίησης είναι ότι η συνάρτηση βελτιστοποίησης (κόστους) f είναι κυρτή. Σε αυτή την περίπτωση μία καλή επιλογή του ρυθμού μάθησης είναι η ακόλουθη:

$$\alpha_t = \frac{\alpha_0}{\sqrt{t}} \tag{5.4}$$

όπου εγγυάται σύγκλιση της τάξης του $\mathbb{E}[f(w_t) - f^*] \leq \mathcal{O}(\log(t)/\sqrt{t})$ χωρίς κάποια παραδοχή λειότητας, αν η βασική τιμή του ρυθμού α_0 επιλεχθεί καταλλήλως [90]. Σε μία άλλη μελέτη [115] αποδεικνύεται ότι με μία παρόμοια διαδικασία με την εξίσωση 5.4 επιτυγχάνει σύγκλιση της τάξης $\mathcal{O}(1/\sqrt{t})$.

Μία στρατηγική που κερδίζει συνεχώς δημοτικότητα είναι η μέθοδος του χρονοδιαγράμματος με περιοδικό τρόπο. Ο κυκλικός ρυθμός μάθησης (αυξάνεται και ελαττώνεται περιοδικά) έχει αποδειχθεί πειραματικά ότι βελτιώνει την εκπαίδευση σε μεγάλο βαθμό [100, 67, 41]. Αυτό βοηθά το δίκτυο να ξεφύγει από κακά τοπικά ελάχιστα που μπορεί να εγκλωβιστεί.

Ένα από τα αρνητικά που εμφανίζει ο SGD σαν βελτιστοποιητής είναι ότι τελική ανανέωση των βαρών είναι ανάλογη του μεγέθους της παραγώγου. Πολλές φορές αυτό μπορεί να οδηγήσει σε χαμηλή επίδοση του μοντέλου αλλά και αργή σύγκλιση. Ιδανικά, η διαδικασία της βελτιστοποίηση θα έπρεπε να διαλέγει διαφορετικούς ρυθμούς μάθησης για κάθε βάρος ή σετ βαρών. Για παράδειγμα, θα ήταν χρήσιμο να επιλέξουμε υψηλότερους ρυθμούς μάθησης σε περιπτώσεις χαμηλών παραγώγων (ή το αντίθετο), όταν η τοπολογία του σφάλματος το επιτρέπει [109].

Για να ξεπεράσουμε το παραπάνω πρόβλημα διάφορες μέθοδοι έχουν προταθεί ανά τα χρόνια, που προσφέρουν προσαρμοστικό ρυθμό μάθησης. Η πρώτη προσέγγιση εμφανίστηκε με τον αλγόριθμο AdaGrad [26], ο οποίος προσαρμόζει τον ρυθμό μάθησης σύμφωνα με το άθροισμα των τετραγώνων των παραγώγων κατά την διάρκεια της εκπαίδευσης για κάθε παράμετρο ξεχωριστά. Έχει αποδειχθεί ότι βελτιώνει την ταχύτητα σύγκλισης σε μη κυρτά περιβάλλοντα, και κυρίως σε πολύ αραιά δεδομένα, καθώς μειώνει τον ρυθμό μάθησης γρηγορότερα σε παραμέτρους που χρησιμοποιούνται πολύ και πιο αργά σε πιο σπάνιες παραμέτρους. Ένα μεγάλο ζήτημα του AdaGrad είναι ότι στις περισσότερες περιπτώσεις ο ρυθμός μάθη-

σης ελαττώνεται με ραγδαίο ρυθμό, επειδή το άθροισμα των τετραγώνων των παραγώγων ανά επανάληψη μεγαλώνει σε υπερβολικό βαθμό.

Ένας βελτιστοποιητής που προσπάθησε να λύσει τα προβλήματα του AdaGrad είναι ο RMSProp [106]. Αντί να χρησιμοποιεί τα συνολικά αθροίσματα των τετραγώνων των παραγώγων, επιλέγει να υπολογίσει έναν εκθετικό κινούμενο μέσο. Αυτό βοηθά επειδή ο μέσος δεν ξεφεύγει σε πολύ ακραίες τιμές που θα οδηγούσαν σε εξαιρετικά χαμηλούς ρυθμούς μάθησης. Μία άλλη μέθοδος που βελτίωσε δραματικά την επίδοση των νευρωνικών είναι ο Adam [51]. Ο Adam χρησιμοποίησε την ιδέα του RMSProp και εφάρμοσε τον εκθετικό κινούμενο μέσο για να προσαρμόσει κατάλληλα τον ρυθμό μάθησης. Εκτός αυτού, όμως, χρησιμοποίησε και τον μέσο των παραγώγων σαν ένα είδους ορμής, που βοήθησε στην καλύτερη προσαρμογή του ρυθμού μάθησης. Έχει γίνει αρκετά δημοφιλής και τα τελευταία χρόνια επιλέγεται σε πολλές εφαρμογές Βαθιάς Μηχανικής Μάθησης [17, 107] σαν μία τυπική επιλογή ενός βελτιστοποιητή.

Αυτή η οικογένεια από βελτιστοποιητές έχει δεχθεί με την σειρά τους διάφορες κριτικές. Μία από αυτές είναι ότι οδηγούν σε προκαθορισμένες ανανεώσεις βαρών και αντίστοιχες παραγώγους, με αποτέλεσμα να διαφοροποιούν το υποκείμενο πρόβλημα βελτιστοποίησης [109]. Έχει παρατηρηθεί σε διάφορες περιπτώσεις ότι οι λύσεις που καταλήγουν είναι τελείως διαφορετικές από τις λύσεις που θα έφτανε ο SGD [107]. Άλλες τεχνικές έχουν προταθεί τα τελευταία χρόνια, οι οποίες προσπαθούν να επιλύσουν τα παραπάνω προβλήματα. Κάποιες από αυτές είναι οι αλγόριθμοι AMSGrad [83] και AdaBound [69] που παρουσιάζουν πρακτικές βελτιώσεις στην επίδοση των δικτύων αλλά και θεωρητικές αποδείξεις ταχύτερης σύγκλισης. Εκτός αυτών, στην μελέτη [111] προτάθηκε ένας άλλος αλγόριθμος με προσαρμοστικό ρυθμό μάθησης, όπου προσπαθεί να υπολογίσει θεωρητικά την σταθερά Lipschitz για συγκεκριμένα δίκτυα με συγκεκριμένες συναρτήσεις ενεργοποίησης. Μία άλλη προσέγγιση για τον υπολογισμό την σταθεράς Lipschitz επιχειρήθηκε στο [28]. Παρόλα αυτά καμία από τις καινούργιες μεθόδους δεν έχει καταφέρει να ξεπεράσει σε δημοτικότητα τον βελτιστοποιητή Adam.

Μία άλλη κατηγορία μεθόδων βασίζεται στο να εκπαιδεύεται ο ρυθμός μάθησης κατά την διάρκεια της εκπαίδευσης ενός νευρωνικού. Κατά την φάση της οπισθοδιάδοσης είναι δυνατόν να αλλάξουμε τον ρυθμό μάθησης ανάλογα με κάποιο κριτήριο σε κάθε επανάληψη της όλης διαδικασίας. Με αυτόν τον τρόπο μπορούμε να ¨μάθουμε' τον καλύτερο ρυθμό μάθησης βελτιστοποιώντας κάποια συνάρτηση-στόχο. Στην μελέτη [6] αυτή η συνάρτηση είναι η συνάρτηση κόστους και εφαρμόζεται ο αλγόριθμος καθόδου κλίσης στον ρυθμό μάθησης ώστε να την μειώσει. Άλλο παράδειγμα της συνάρτησης-στόχου είναι να μειωθεί μέσω του ρυθμού μάθησης το τετράγωνο της νόρμας των παραγώγων [109]. Η πρώτη τεχνική είναι αρκετά αποτελεσματική αλλά πέφτει και αυτή σε κάποια από τα εμπόδια που έχουμε προαναφέρει.

5.2.1 Θεωρητικό Υπόβαθρο

Προτού εμβαθύνουμε περισσότερο στους διάφορους βελτιστοποιητές, θα εισάγουμε κάποιους συμβολισμούς και ορισμούς για την καλύτερη κατανόηση. Οι ορολογίες και τα σύμβολα που

θα χρησιμοποιηθούν βασίζονται σε συγγράμματα της βιβλιογραφίας, όπως [83, 51]. Έστω $\mathcal{F} \in \Re^d$ θα συμβολίζουμε το σύνολο από τα εφικτά βάρη w_t του νευρωνικού δικτύου. Υποθέτουμε ότι το σύνολο $\mathcal{F}$ έχει φραγμένη διάμετρο $D_\infty \in \Re$, αν $\|x - y\| \leq D_\infty$, για όλα τα $x, y \in \mathcal{F}$. Στο σύνολο $\mathcal{F}$ υποθέτουμε ότι η νόρμα είναι φραγμένη . Μία γενική δομή για έναν προσαρμοστικό αλγόριθμο βελτιστοποίησης παρουσιάζεται στον Αλγόριθμο 4.

Algorithm 4: Generic framework of adaptive optimization methods

Input: $w_t \in \mathcal{F}$, initial step α_0, functions $\{\phi_t, \psi_t\}_{t=1}^T$

for *t=1* ***to*** T **do**

$g_t = \nabla f_t(w_t)$

$m_t = \phi_t(g_1, ..., g_t)$ and $V_t = \psi_t(g_1, ..., g_t)$

$\alpha_t = \alpha_0/\sqrt{t}$

$\hat{w}_{t+1} = w_t - \alpha_t m_t/\sqrt{V_t}$

$w_{t+1} = \Pi_{\mathcal{F}, \sqrt{V_t}}(\hat{w}_{t+1})$

end

Στον Πίνακα 5.1 υπάρχει μία περίληψη από τους πιο γνωστούς αλγόριθμους βελτιστοποίησης που ταιριάζουν στην παραπάνω δομή. Οι κύριες διαφορές επικεντρώνονται γύρω από τις συναρτήσεις ϕ και ψ μέσω των οποίων οι μεταβλητές m_t και V_t ανανεώνονται. Μέσω αυτής της ομαδοποίησης μπορούμε να διακρίνουμε καλύτερα τις διαφορές μεταξύ βελτιστοποιητών.

	SGD	AdaGrad	RMSProp	ADAM
ϕ_t	g_t	g_t	g_t	$(1-\beta_1)\sum_{i=1}^t \beta_1^{t-i} g_i$
ψ_t	$\mathbf{1}$	$\text{diag}(\sum_{i=1}^t g_i^2)$	$(1-\beta_2)\,\text{diag}(\sum_{i=1}^t \beta_2^{t-i} g_i^2))$	$(1-\beta_2)\,\text{diag}(\sum_{i=1}^t \beta_2^{t-i} g_i^2)$

Πίνακας 5.1: Διαφορές μεταξύ δημοφιλών προσαρμοστικών αλγορίθμων βελτιστοποίησης.

Στη συνέχεια θα αναφερθούμε και θα αναλύσουμε την σταθερά του Lipschitz, οπότε σε αυτό το σημείο θα παραθέσουμε κάποιους ορισμούς, οι οποίοι θα χρησιμοποιηθούν παρακάτω.

Ορισμός 5.3. *Μία συνάρτηση $f : \Re^d \to \Re$ είναι L-Lipschitz συνεχής αν για όλα τα $x, y \in \Re^d$, υπάρχει $L > 0$, όπου*[1]

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

και το μικρότερο L, το οποίο ικανοποιεί την παραπάνω εξίσωση, ονομάζεται σταθερά Lipschitz.

Έστω $f : \Re^d \to \Re$ μία συνάρτηση με λεία παράγωγο:

$$\|\nabla f(w_1) - \nabla f(w_2)\| \leq L\|w_1 - w_2\| \quad \forall w_1, w_2 \in \Re^d \tag{5.5}$$

[1] Όλες οι νόρμες που θα αναφερθούν είναι Ευκλείδειες, $\|\cdot\| = \|\cdot\|_2$.

Ας υποθέσουμε ένα πρόβλημα βελτιστοποίησης, που ο στόχος του είναι να βρει ένα σύνολο παραμέτρων w που ελαχιστοποιούν μία συνάρτηση κόστους f μέσω του αλγορίθμου Καθόδου Κλίσης (Gradient Descent):

Λήμμα 5.1. *Δεδομένης μιας κυρτής συνάρτησης κόστους f, με L-Lipschitz συνεχή παράγωγο (Ορ. 5.3) με L να είναι η σταθερά Lipschitz, η βέλτιστη τιμή του ρυθμού μάθησης για τον αλγόριθμο της Καθόδου Κλίσης είναι:*

$$\alpha^* = \frac{1}{L} \tag{5.6}$$

Η απόδειξη του λήμματος μπορεί να βρεθεί στο Παράρτημα Β′. Για να υπολογίσουμε τον βέλτιστο ρυθμό μάθησης, όμως, σε ένα πραγματικό σενάριο χρειάζεται γνώση της συνάρτησης κόστους και της παραγώγου από πριν.

Στην πράξη είναι αρκετά δύσκολο να βρεθεί ένας συνολικά καλός ρυθμός μάθησης για όλα τα σύνολα δεδομένων και όλα τα είδη των δικτύων. Επίσης, χρησιμοποιώντας αλγορίθμους βασισμένους στην εκπαίδευση σε παρτίδες, όπως ο SGD, οι παράγωγοι που υπολογίζονται είναι οι θορυβώδεις εκδοχές τους σε σχέση με τις ολικές παραγώγους. Αυτό σημαίνει ότι η τιμή της σταθεράς L δεν είναι εφικτό να υπολογιστεί με ακρίβεια. Παρόλα αυτά είναι δυνατόν μέσα από την διαδικασία της εκπαίδευσης να ¨μάθουμε' την τιμή της ή να την προσεγγίσουμε αρκετά κοντά στην πραγματική της τιμή. Μέχρι στιγμής αυτό έχει παραμείνει ανοιχτό πρόβλημα έρευνας [109].

5.3 Ο προσαρμοστικός βελτιστοποιητής AdaLip

Σε αυτό το σημείο θα παρουσιάσουμε έναν καινούργιο βελτιστοποιητή, ο οποίος χρησιμοποιώντας μία προσέγγιση της σταθεράς Lipschitz καταφέρνει να κατασκευάσει έναν προσαρμοστικό ρυθμό μάθησης ανά επίπεδο του δικτύου.

5.3.1 Μαθαίνοντας την σταθερά Lipschitz

Έχοντας υπόψη ότι οι ανανεώσεις των βαρών είναι μικρές σε κάθε επανάληψη του αλγορίθμου βελτιστοποίησης, μπορούμε να υπολογίσουμε τη σταθερά L σε μικρό υποχώρο, τον υποχώρο που ο βελτιστοποιητής θα εξερευνήσει. Αυτό σημαίνει ότι οι πληροφορίες που χρειάζονται από τον υποχώρο της συνάρτησης κόστους μπορούν να υπολογιστούν από ένα απλό εμπρόσθιο πέρασμα. Ένα σύστημα από πολλαπλά εμπρόσθια περάσματα θα μπορούσε να χρησιμοποιηθεί υπολογίζοντας τις παραγώγους από πολλές κοντινές κατευθύνσεις, αλλά αυτό απαιτεί μεγάλο υπολογιστικό κόστος. Το κόστος αυτό μεγαλώνει δραματικά με τα βαθιά μοντέλα και τα τεράστια σύνολα δεδομένων καθιστώντας το ανέφικτο. Παρακάτω παρουσιάζουμε μία προσέγγιση του L σε ένα στοχαστικό περιβάλλον. Παίρνοντας την Εξίσωση 5.5, αντικαθιστώντας τα (w_1, w_2) με (w_t, w_{t-1}) έχουμε:

$$\|\nabla f(w_t) - \nabla f(w_{t-1})\| \leq L\|w_t - w_{t-1}\|$$

Από την εξίσωση 5.2:

$$\|\nabla f(w_t) - \nabla f(w_{t-1})\| \leq L\|w_{t-1} - \alpha_t \nabla f(w_{t-1}) - w_{t-1}\|$$

$$\|\nabla f(w_t) - \nabla f(w_{t-1})\| \leq L\alpha_t \|\nabla f(w_{t-1})\|$$

$$L \geq \frac{\|\nabla f(w_t) - \nabla f(w_{t-1})\|}{\alpha_t \|\nabla f(w_{t-1})\|}$$

Μέχρι στιγμής έχει γίνει ανάλυση πάνω στον αλγόριθμο της Καθόδου Κλίσης. Όμως, σε ένα ρεαλιστικό σενάριο η στοχαστική του εκδοχή θα χρησιμοποιείται. Αυτό μεταφράζεται ως την ύπαρξη θορυβωδών παραγώγων, οι οποίες έχουν υπολογιστεί από την χρήση παρτίδων. Αυτές οι παράγωγοι g_t δεν είναι ίσες με την ολική παράγωγο αλλά η αναμενόμενη τιμή τους τείνει σε αυτή. Αυτό μπορεί να εκφραστεί με την παρακάτω σχέση:

$$\mathop{\mathbb{E}}_{x}[g_t] = \mathop{\mathbb{E}}_{x}[\nabla f(w_t|x_t)] = \nabla f(w_t)$$

Αντικαθιστώντας τον συμβολισμό $\nabla f(w_t)$ με g_t και χρησιμοποιώντας την Εξίσωση 5.6, η παραπάνω ανισότητα μετασχηματίζεται στο ακόλουθο:

$$\alpha^* \leq \alpha_t \frac{||g_{t-1}||}{||g_t - g_{t-1}||} \tag{5.7}$$

όπου α^* είναι η βέλτιστη τιμή του ρυθμού μάθησης και α_t είναι ο τωρινός ρυθμός.

Η Εξίσωση 5.7 προσφέρει μία προσέγγιση του βέλτιστου ρυθμού μάθησης α^*. Ιδανικά, θα επιθυμούσαμε να επιλέξουμε σε κάθε στιγμή αυτή την τιμή. Όμως, αυτό είναι προτεινόμενο, καθώς αυτή η εξίσωση παρουσιάζει μεγάλη διακύμανση στον ρυθμό στο πέρασμα των επαναλήψεων. Αυτό θα οδηγούσε σε αρκετά ασταθή εκπαίδευση.

Το βασικό πρόβλημα που δημιουργείται με την παραπάνω σχέση οφείλεται στον παρονομαστή. Ο παρονομαστής αποτελείται από τη νόρμα της διαφοράς των παραγώγων δύο συνεχόμενων επαναλήψεων. Αυτή η νόρμα εκφράζει το μέτρο της αλλαγής της κατεύθυνσης που επιλέγει ο βελτιστοποιητής. Το ζήτημα εμφανίζεται όταν βρισκόμαστε κοντά σε τοπικά ελάχιστα, όπου οι παράγωγοι είναι αρκετά κοντά μεταξύ τους. Σε αυτές τις περιπτώσεις ο παρονομαστής τείνει πολύ κοντά στο 0, εκτινάσσοντας με την σειρά του τον ρυθμό μάθησης. Μία λύση είναι να προσθέσουμε έναν μικρό θετικό όρο c_t στον παρονομαστή, ο οποίος θα αντικρούσει αυτό το φαινόμενο:

$$\alpha^* \leq \alpha_t \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t} \tag{5.8}$$

Αυτή η προσθήκη δημιουργεί ένα κάτω όριο στον παρονομαστή, το οποίο μπορεί να αλλάξει κατά την διάρκεια της εκπαίδευσης. Στη συνέχεια θα δούμε την επίδραση της υπερπαραμέτρου c_t στην διαδικασία της εκπαίδευσης. Παρόλα αυτά η αστάθεια του αλγορίθμου δεν οφείλεται μόνο στο παρονομαστή, αλλά και στην στοχαστική φύση του αλγορίθμου βελτιστοποίησης. Η Εξίσωση 5.8 προσεγγίζει τον βέλτιστο ρυθμό μάθησης του συγκεκριμένο υποχώρου, που δημιουργείται από την κάθε παρτίδα δεδομένων x_t. Ο σκοπός μας είναι να προσεγγίσουμε τον συνολικό βέλτιστο ρυθμό μάθησης της ολικής συνάρτησης κόστους. Για να το πραγματοποιήσουμε, θα εφαρμόσουμε έναν κινούμενο μέσο του α^* για κάθε παρτίδα.

$$S_t = \gamma \cdot S_{t-1} + (1-\gamma) \cdot \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t}$$

με $\gamma \in (0,1)$ να είναι ο συντελεστής του κινούμενου μέσου. Σε κλειστή μορφή η σχέση μπορεί να γραφεί ως:

$$S_t = \gamma^t \cdot S_0 + (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{\|g_{i-1}\|}{\|g_i - g_{i-1}\| + c_i} \tag{5.9}$$

όπου $S_0 = 1$. Επομένως, η προσέγγιση A_t του βέλτιστου ρυθμού μάθησης α^* υπολογίζεται ως ένα γινόμενο του α_t και του παραπάνω εκθετικού κινούμενου μέσου:

$$\begin{aligned} A_t &= \alpha_t \cdot S_t = \\ &= \alpha_t \cdot \left[\gamma^t \cdot S_0 + (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{\|g_{i-1}\|}{\|g_i - g_{i-1}\| + c_i} \right] \end{aligned} \tag{5.10}$$

Λήμμα 5.2. *Έστω f μία συνάρτηση κόστους που ικανοποιεί τον Ορισμό 5.3 με φραγμένες παραγώγους $||\nabla f(w_t)|| \leq G$, $\forall\ w_t \in \Re^d$. Έστω M είναι η ελάχιστη νόρμα των παραγώγων, η οποία είναι διαφορη του μηδενός, τότε ο κινούμενος μέσος S_t της Εξίσωσης 5.9 είναι φραγμένος.*

$$\frac{M(1-\gamma)}{2G + \max_t c_t} \leq S_t \leq 1 + \frac{G}{\min_t c_t}$$

Το παραπάνω Λήμμα αποδεικνύει ότι ο κινούμενος μέσος της Εξίσωσης 5.9 είναι φραγμένος και, συνεπώς, ο ρυθμός μάθησης A_t είναι και αυτός φραγμένος. Αυτό θα είναι χρήσιμο στην απόδειξη του επόμενου θεωρήματος περί της σύγκλισης ενός βελτιστοποιητή SGD με τέτοιου είδους ρυθμό μάθησης. Για την απόδειξη της σύγκλισης θα χρησιμοποιηθεί ο όρος μετάνοιας. Ο όρος μετάνοιας (R) είναι το άθροισμα των προηγούμενων διαφορών μεταξύ της πρόβλεψης $f_t(w_t)$ του δικτύου και της θεωρητικής καλύτερης δυνατής πρόβλεψης $f_t(w^*)$ ενός δικτύου με βέλτιστα βάρη w^*. Η μετάνοια μπορεί να γραφεί με τον ακόλουθο τύπο:

$$R(T) = \sum_{t=1}^{T} [f_t(w_t) - f_t(w^*)] \tag{5.11}$$

Θεώρημα 5.1. *Υποθέτουμε ότι μία συνάρτηση κόστους έχει φραγμένες παραγώγους, $\|g_t\| \leq G$, ελάχιστη μη-μηδενική νόρμα παραγώγων M και τα βάρη βρίσκονται στην σφαίρα $\|w_t\| \leq r$. Έστω $\alpha_t = \alpha_0/\sqrt{t}$ και $\gamma \in (0,1)$, τότε ένας βελτιστοποιητής SGD με την Εξίσωση 5.10 ως τον ρυθμό μάθησής του πετυχαίνει τις ακόλουθες εγγυήσεις για όλα τα $T > 1$ και $c_t \geq \frac{(1-\gamma)\|g_{t-1}\|}{\left(\sqrt{\frac{t}{t-1}}-\gamma\right)S_{t-1}} - \|g_t - g_{t-1}\|$.*

$$\frac{R(T)}{T} \leq \frac{2r^2(2G + c_T)}{\alpha_0 M(1-\gamma)\sqrt{T}} + \frac{G^2\alpha_0}{\sqrt{T}}\left(1 + \frac{G}{c_1}\right)$$

το οποίο οδηγεί στο

$$\frac{R(T)}{T} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \quad \textit{and} \quad \lim_{T\to\infty} \frac{R(T)}{T} = 0$$

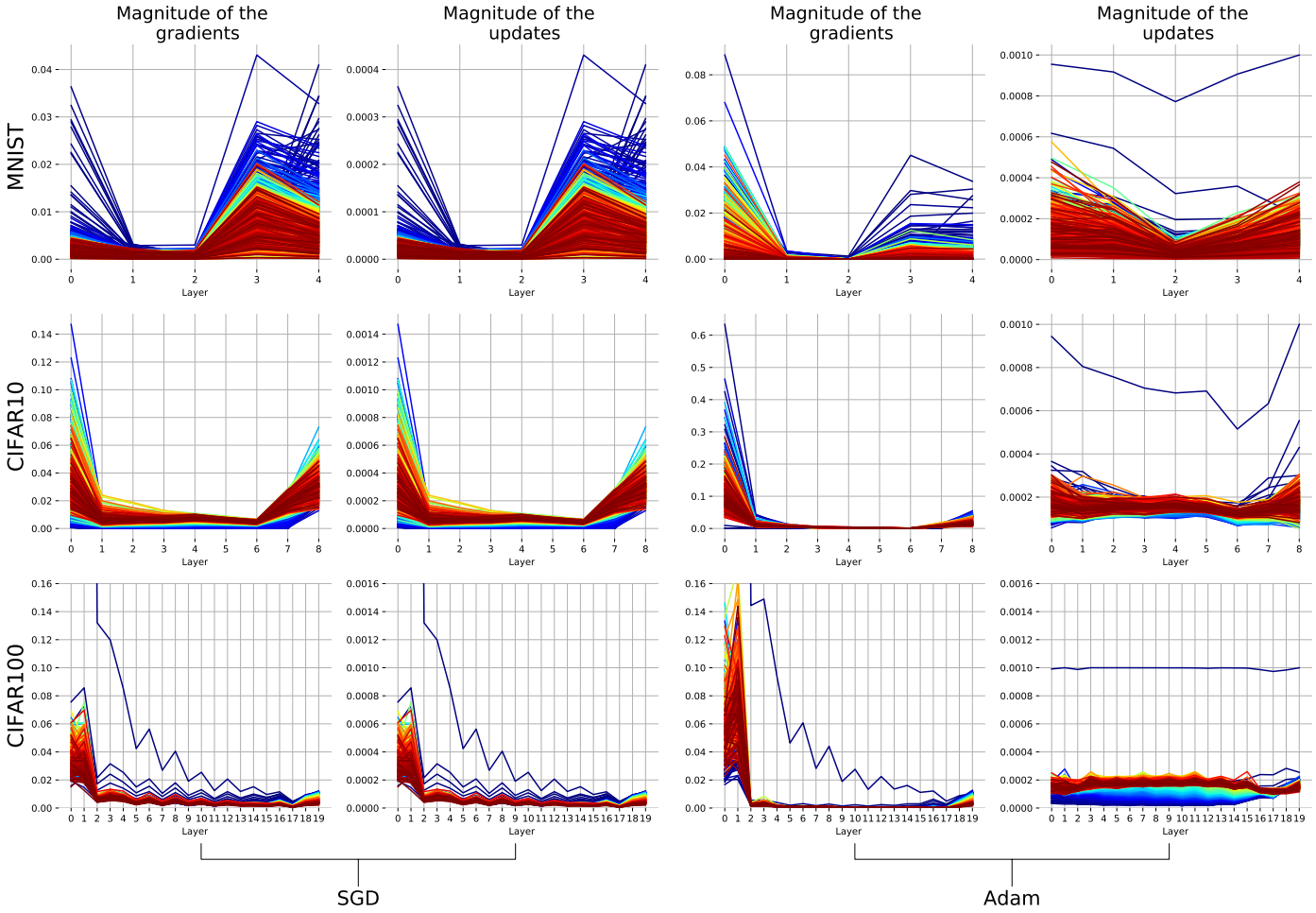
Οι αποδείξεις για το Λήμμα 5.2 και Θεώρημα 5.1 μπορούν να βρεθούν στο Παράρτημα Β′.

5.3.2 Υπολογισμός ανά Επίπεδο

Στα Νευρωνικά Δίκτυα οι παράμετροι του ίδιου επιπέδου έχουν παρόμοιες ιδιότητες. Σε διαφορετικά, όμως, επίπεδα αυτές οι ιδιότητες μπορεί να διαφέρουν. Ένα παράδειγμα του φαινομένου αυτού είναι το ζήτημα των εξαφανιζόμενων κλίσεων [43], όπου τα πρώτα επίπεδα ενός δικτύου δεν εκπαιδεύονται επαρκώς. Αυτό το πρόβλημα έχει αντιμετωπιστεί με διάφορους τρόπους, κάποιοι από αυτούς είναι μη-κορεσμένες συναρτήσεις ενεργοποίησης [76], καλύτερες στρατηγικές αρχικοποίησης βαρών [32] και επίπεδα μετασχηματισμού, όπως η Κανονικοποίηση Παρτίδας [43]. Παρόλα αυτά παρατηρούμε στην πράξη ότι κανένα από αυτές τις μεθόδους δεν λύνουν το πρόβλημα άμεσα.

Για να έχουμε μία καλύτερη εικόνα της διαδικασίας της εκπαίδευσης και της συμπεριφοράς διαφορετικών επιπέδων, θα αναπτύξουμε ένα πειραματικό σχέδιο με το οποίο θα παρατηρήσουμε τα επίπεδα ενός δικτύου κατά τη διάρκεια της εκπαίδευσης. Θα μετρήσουμε το μέτρο των παραγώγων και των ανανεώσεων των βαρών σε ένα δίκτυο για 3 διαφορετικά σύνολα δεδομένων χρησιμοποιώντας 2 βελτιστοποιητές, SGD και Adam, με σταθερό ρυθμό μάθησης. Η μετρική που χρησιμοποιήθηκε είναι ο μέσος όρος των απόλυτων τιμών των παραγώγων και των ανανεώσεων κάθε επιπέδου για κάθε επανάληψη του αλγορίθμου μάθησης. Τα αποτελέσματα φαίνονται στο Σχήμα 5.1. Τα νευρωνικά δίκτυα που χρησιμοποιήθηκαν καταγράφονται στο Παράρτημα Β′. Οι μπλε γραμμές αντιστοιχούν σε αρχικές επαναλήψεις της εκπαίδευσης και καθώς το χρώμα φτάνει το κόκκινο χρώμα πλησιάζουμε τις τελευταίες επαναλήψεις. Τα επίπεδα που παρουσιάζονται είναι Συνελικτικά και Πλήρες Συνδεδεμένα. Οι δύο αριστερές στήλες γραφημάτων δείχνουν τα νευρωνικά που εκπαιδεύτηκαν με SGD. Όπως ήταν αναμενόμενο, το

μέτρο των ανανεώσεων είναι ανάλογο με αυτό των παραγώγων, το οποίο φαίνεται και από το σχήμα τους. Από την άλλη, οι δεξιές στήλες δείχνουν δίκτυα που έχουν εκπαιδευτεί από τον Adam, παρουσιάζοντας καλύτερη ακρίβεια και ταχύτερη σύγκλιση. Είναι εμφανές ότι το μέτρο των ανανεώσεων δεν είναι αναλογικό με αυτό των παραγώγων και δεν μιμείται το σχήμα τους. Αυτό αποτελεί μία εικόνα για το γιατί ο Adam μπορεί να κατασκευάζει ανανεώσεις βαρών κοντινότερες στο βέλτιστο ανεξάρτητα από το μέτρο των παραγώγων. Ο SGD συνήθως πάσχει από αυτό το πρόβλημα, δηλαδή κάποια επίπεδα παρουσιάζουν παραγώγους που δυσκολεύουν την εύρεση της βέλτιστης ανανέωσης βάρους. Αυτό δείχνει την αναγκαιότητα ενός προσαρμοστικού ρυθμού μάθησης ανά επίπεδο, το οποίο θα βοηθήσει διάφορους βελτιστοποιητές να φτάσουν τα βέλτιστα βάρη κλιμακώνοντας αντίστοιχα τις παραγώγους.



Σχήμα 5.1: Το μέτρο των παραγώγων και των ανανεώσεων των βαρών ανά επίπεδο και για κάθε σύνολο δεδομένων (MNIST, CIFAR10, CIFAR100) για δύο διαφορετικούς βελτιστοποιητές (SGD and Adam). Οι μπλε γραμμές δείχνουν τις πρώτες επαναλήψεις. Όσο το χρώμα αλλάζει προς το κόκκινο, το γράφημα δείχνει τις μεταγενέστερες επαναλήψεις.

Όλα τα βάρη και οι παράγωγοι που αναφέρθηκαν στις προηγούμενες εξισώσεις είναι τα ολόκληρα διανύσματα που περιέχουν όλες τις παραμέτρους του δικτύου. Όμως, σε ένα ρεαλιστικό παράδειγμα είναι αρκετά δύσκολο να κατασκευαστεί ένα διάνυσμα τέτοιας διάστασης και

να εκτελέσει πράξεις με αυτό (προσθέσεις, πολλαπλασιασμούς, υπολογισμό νορμών). Αυτός είναι ο δεύτερος λόγος που ενισχύει την ιδέα για ξεχωριστό υπολογισμό ανά επίπεδο των ανανεώσεων των βαρών. Αυτό οδήγησε και στο κίνητρο να κατασκευαστεί ένας βελτιστοποιητής με προσαρμοστικό ρυθμό μάθησης ανά επίπεδο. Στο επόμενο υποκεφάλαιο θα παρουσιαστεί πλήρως ο αλγόριθμος αυτός αλλά και πώς συνδυάζεται με ήδη υπάρχοντες βελτιστοποιητές.

5.3.3 AdaLip

Σε αυτό το σημείο θα παρουσιάσουμε μία τεχνική βελτιστοποίησης, που ονομάζεται AdaLip, η οποία προσαρμόζει τον ρυθμό μάθησης ανά επίπεδο. Αυτό θα δούμε ότι μπορεί να επιλύσει ζητήματα που προκείπτους από υπο-εκπαιδευμένα επίπεδα. Εκτός αυτού ο αλγόριθμος που θα αναπτύξουμε μπορεί να λειτουργήσει μαζί με οποιοδήποτε χρονοδιάγραμμα ρυθμού μάθησης (scheduler), το οποίο θα ελέγχει τον ολικό ρυθμό μάθησης α_t του δικτύου. Επιπλέον, είναι δυνατόν να δουλέψει σε συνδυασμό με ήδη υπάρχοντες βελτιστοποιητές, οι οποίοι χρησιμοποιούν προσαρμοστικό ρυθμό μάθησης ανά παράμετρο, όπως είναι ο Adam. Η μελέτη αυτή επικεντρώνεται γύρω από τον συνδυασμό του AdaLip με τους τρεις πιο γνωστούς βελτιστοποιητές στην Βαθιά Μηχανική Μάθηση: SGD, Adam και RMSProp. Επίσης, μία παραλλαγή του AdaLip θα παρουσιαστεί στο επόμενο υποκεφάλαιο, που περιλαμβάνει κάποια μεταβολή στον κανόνα ανανέωσης των βαρών.

Algorithm 5: AdaLip

Input: $w_t \in \mathcal{F}$, initial step α_0, $\gamma \in (0,1)$, c_t

Initialize: $S_0 = 1$, $g_0 = 0$

for *t=1* ***to*** T **do**

$$g_t = \nabla f_t(w_t)$$
$$\alpha_t = \alpha_0/\sqrt{t}$$
$$S_t = \gamma \cdot S_{t-1} + (1-\gamma)\frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t}$$
$$A_t = \alpha_t \cdot S_t$$
$$\hat{w}_{t+1} = w_t - A_t \cdot g_t$$

end

Στους Αλγόριθμους 5, 6 και 7 παρουσιάζονται οι τρεις νέοι βελτιστοποιητές που δημιουργήθηκαν από αυτή την ένωση. Ο πρώτος είναι η αρχική έκδοση του AdaLip, που ο αλγόριθμος μεταβάλλει τον ρυθμό μάθησης του SGD βελτιστοποιητή. Ο αλγόριθμος AdaLip υπολογίζει με κάποια προσέγγιση την σταθερά Lipschitz για κάθε παρτίδα και, μέσω ενός κινούμενου μέσου, βρίσκει τον βέλτιστο ρυθμό μάθησης A_t (Εξίσωση 5.10). Αυτό πραγματοποιείται για κάθε επίπεδο ξεχωριστά για λόγους που προαναφέρθηκαν. Είναι σημαντικό να αναφερθεί ότι οι νόρμες που εμφανίζονται στους Αλγορίθμους 5, 6 και 7 υπολογίζονται χρησιμοποιώντας τις παραγώγους των βαρών που βρίσκονται στο ίδιο επίπεδο με το βάρος που ανανεώνεται σε εκείνη την στιγμή της επανάληψης.

Η τεχνική AdaLip μπορεί να εφαρμοστεί και σε άλλους βελτιστοποιητές, όπως ο RM-

Algorithm 6: RMSLip

Input: $w_t \in \mathcal{F}$, initial step α_0, $\gamma \in (0,1)$, c_t
Initialize: $S_0 = 1$, $g_0 = 0$, $\nu_0 = 0$
for *t=1* ***to*** T **do**
$g_t = \nabla f_t(w_t)$
$\alpha_t = \alpha_0/\sqrt{t}$
$S_t = \gamma \cdot S_{t-1} + (1-\gamma)\dfrac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t}$
$A_t = \alpha_t \cdot S_t$
$\nu_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) g_t^2$
$\hat{w}_{t+1} = w_t - A_t \cdot g_t/(\sqrt{\nu_t} + \epsilon)$
end

SProp. Ο νέος βελτιστοποιητής που θα παραχθεί θα ονομάζεται RMSLip και περιγράφεται στον Αλγόριθμο 6. Ομοίως, και σε αυτή τη περίπτωση ο ρυθμός μάθησης υπολογίζεται για κάθε επίπεδο ξεχωριστά. Η διαφορά με την απλή περίπτωση του AdaLip είναι ότι κάθε ανανέωση επηρεάζεται με τον κινούμενο μέσο των τετραγώνων των παλαιότερων παραγώγων, όπως είναι και ο αρχικός RMSProp.

Algorithm 7: AdamLip

Input: $x_t \in \mathcal{F}$, initial step α_0, $\gamma \in (0.1)$, c_t
Initialize: $S_0 = 1$, $g_0 = 0$, $m_0 = 0$, $\nu_0 = 0$
for *t=1* ***to*** T **do**
$g_t = \nabla f_t(w_t)$
$\alpha_t = \alpha_0/\sqrt{t}$
$S_t = \gamma \cdot S_{t-1} + (1-\gamma)\dfrac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t}$
$A_t = \alpha_t \cdot S_t$
$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) g_t$
$\nu_t = \beta_2 \cdot \nu_{t-1} + (1-\beta_2) g_t^2$
$\hat{m}_t = m_t/(1-\beta_1^t)$
$\hat{\nu}_t = \nu_t/(1-\beta_2^t)$
$\hat{w}_{t+1} = w_t - A_t \cdot \hat{m}_t/(\sqrt{\hat{\nu}_t} + \epsilon)$
end

Τέλος, θα εξετάσουμε τον συνδυασμό της μεθόδου AdaLip και του βελτιστοποιητή Adam, που θα ονομάζεται ως AdamLip. Ο κανόνας ανανέωσης των βαρών του Adam παραμένει ίδιος, δηλαδή ο κινούμενος μέσος των παραγώγων m_t και των τετραγώνων τους ν_t, αλλά και ο διορθωτής πόλωσης τους). Όμως, στη συγκεκριμένη περίπτωση ο σταθερός ρυθμός μάθησης αντικαθίσταται από το A_t της Εξίσωσης 5.10).

Λεπτομέρειες Υλοποίησης Ο αλγόριθμος AdaLip περιέχει διάφορες υπερπαραμέτρους, οι οποίες θα επεξηγηθούν και θα προταθούν οι αντίστοιχες αρχικοποιήσεις τους. Μία σημαντική υπερπαράμετρος είναι ο όρος c_t. Από το Θεώρημα 5.1 παίρνουμε την βέλτιστη τιμή της, η οποία έχει υπολογιστεί ώστε να εγγυάται την σύγκλιση. Στην πράξη, όμως, τα πειράματα έχουν εκτελεστεί με σταθερή τιμή $c_t = c = 10^{-8}$, η οποία είναι κοντά στην θεωρητική τιμή του c_t. Αυτό θα επεξηγηθεί καλύτερα σε επόμενο υποκεφάλαιο.

Όσον αφορά τον αρχικό ρυθμό μάθησης α_0 επιλέχθηκαν διάφορες αρχικές τιμές για να ελεγχθεί η σταθερότητά του. Η αρχική τιμή του κινούμενου μέσου S_0 ορίστηκε στο 1, επειδή ο ολικός ρυθμός μάθησης θα ξεκινά από τον αρχικό ρυθμό μάθησης α_0.

Μία άλλη σημαντική υπερπαράμετρος είναι ο συντελεστής γ του κινούμενου μέσου. Το εύρος τιμών της κυμαίνεται στα $(0, 1)$ και η προτεινόμενη τιμή της είναι 0.8, η οποίο βρέθηκε εμπειρικά. Χαμηλότερες τιμές επηρεάζουν τον ρυθμό μάθησης κάνοντας τον να αλλάζει δραματικά από επανάληψη σε επανάληψη. Αυτό θα μπορούσε να αποτελεί βοηθητικό παράγοντα σε περιπτώσεις που το δίκτυο εγκλωβίζεται εύκολα σε τοπικά ελάχιστα. Όμως, εγκυμονεί τον κίνδυνο για μία πιο ασταθή διαδικασία εκπαίδευσης. Από την άλλη, υψηλές τιμές θα εξαλείψουν οποιεσδήποτε αλλαγές του ρυθμού μάθησης μετασχηματίζοντάς τον σε σχεδόν σταθερό.

5.3.4 Πειραματική Διαδικασία

Η πειραματική διαδικασία που θα χρησιμοποιηθεί για την αξιολόγηση της καινούργιας τεχνικής βελτιστοποίησης, θα περιλαμβάνει πειράματα σε 3 γνωστά σύνολα δεδομένων, το MNIST [61], το CIFAR10 [54] και το CIFAR100 [54]. Ο αλγόριθμος AdaLip θα συγκριθεί με τους αντίστοιχους βελτιστοποιητές που προσπαθεί να βελτιώσει (δηλαδή ο AdaLip σε σχέση με τον SGD, ο RMSLip με τον RMSProp και ο AdamLip με τον Adam) ως προς την ταχύτητα σύγκλισης και την σταθερότητα τους στην επιλογή της αρχικής τιμή του ρυθμού μάθησης a_0. Ο σκοπός μας είναι να αποφανθούμε αν αυτή η τεχνική βελτιώνει την εκπαίδευση ενός Νευρωνικού Δικτύου.

Λόγω της διαφορετικής δυσκολίας σε κάθε σύνολο δεδομένων, διαφορετικές αρχιτεκτονικές δικτύων επιλέχθηκαν για το καθένα. Οι αρχιτεκτονικές καταγράφονται στους Πίνακες Β′.1, Β′.2 και Β′.3 στο Παράρτημα Β′. Οι πρώτες δύο αρχιτεκτονικές είναι μικρές εκδοχές της αρχιτεκτονικής VGG [99], ενώ η αρχιτεκτονική για το CIFAR100 διαθέτει περισσότερα επίπεδα χρησιμοποιώντας Επίπεδα Κανονικοποίησης [43] καθώς και Απόσυρση [101].

Για την σταθερότητα των αποτελεσμάτων, τα πειράματα του κάθε βελτιστοποιητή εκτελέστηκαν πολλές φορές (25 εκτελέσεις για MNIST και CIFAR10, 10 εκτελέσεις για CIFAR100). Σε κάθε εκτέλεση οι βελτιστοποιητές αρχικοποιήθηκαν με τα ίδια βάρη και υπερπαραμέτρους. Επειδή διαφορετικοί βελτιστοποιητές αποδίδουν καλύτερα με διαφορετικές τιμές για τον αρχικό ρυθμό μάθησης, διάφορες τιμές εξετάστηκαν για κάθε βελτιστοποιητή και για κάθε σύνολο δεδομένων.

Αρχικά, για να αξιολογήσουμε τα αποτελέσματα υπολογίστηκε η υψηλότερη ακρίβεια εκπαίδευσης για κάθε εκτέλεση και, έπειτα, επιλέχθηκε η διάμεσος από όλες τις εκτελέσεις για κάθε αρχικό ρυθμό μάθησης ανεξάρτητα. Αυτό το σύστημα αξιολόγησης επιλέχθηκε επειδή ο καλύτερος βελτιστοποιητής είναι αυτός μου αποδίδει καλύτερα από τους άλλους με συνέπεια ανεξαρτήτων αρχικοποίησης. Το επόμενο βήμα είναι να εξετάσουμε πόσο ευαίσθητος είναι ο κάθε βελτιστοποιητής στην επιλογή του αρχικού ρυθμού μάθησης. Για αυτό το λόγο υπολογίσαμε την μέγιστη (max), την ελάχιστη (min), τη διάμεσο (median) και την τυπική απόκλιση (std) από όλες τις ακρίβειες για όλα τους διαφορετικούς ρυθμούς μάθησης των πειραμάτων. Οι Πίνακες 5.2, 5.3 και 5.4 παρουσιάζουν τα αποτελέσματα για τα 3 σύνολα δεδομένων.

Ένα σημαντικό κομμάτι ενός βελτιστοποιητή είναι η ταχύτητα σύγκλισής του. Αυτό μπορεί να μετρηθεί με διάφορους τρόπους. Σε αυτή τη μελέτη χρησιμοποιήσαμε τον αριθμό των εποχών που χρειάστηκαν για να φτάσει το δίκτυο στο 95% της μέγιστης ακρίβειας εκπαίδευσης. Τα αποτελέσματα υπολογίστηκαν ανάμεσα σε όλες τις εκτελέσεις και όλους τους αρχικούς ρυθμούς μάθησης.

Πίνακας 5.2: Ακρίβεια εκπαίδευσης (πρώτες 4 γραμμές) και Εποχές σύγκλισης (τελευταίες 3 γραμμές) για το σύνολο δεδομένων του MNIST.

	SGD	AdaLip	Adam	AdamLip	RMSProp	RMSLip
max	**1.0**	**1.0**	0.999	**1.0**	**1.0**	**1.0**
median	0.976	**0.996**	**0.999**	**0.999**	**1.0**	**1.0**
mean	0.663	**0.883**	**0.999**	**0.999**	**1.0**	**1.0**
std	0.422	**0.288**	0.001	**0.000**	**0.000**	**0.000**
mean	4.5	**3.1**	1.75	**1.5**	1.5	**1**
median	2.5	**1.5**	2	**1.5**	1.5	**1**
best	2	**1**	**1**	**1**	**1**	**1**

MNIST

Αρχικά, η μέθοδος AdaLip εξετάστηκε στο σύνολο δεδομένων MNIST [61], που, όπως αναφέρθηκε και προηγούμενο κεφάλαιο, αποτελείται από εικόνες χειρόγραφων ψηφίων. Οι εικόνες κανονικοποιήθηκαν στο εύρος $[0, 1]$ και δεν χρησιμοποιήθηκε καμία άλλη προεπεξεργασία. Οι αλγόριθμοι βασισμένοι στον SGD εξετάστηκαν σε 8 διαφορετικούς ρυθμούς μάθησης για αυτό το πρόβλημα. Λεπτομερώς, είναι οι εξής: 0.005, 0.01, 0.05, 0.1, 0.5, 0.7, 1.0, 1.3. Όσον αφορά τους βελτιστοποιητές βασισμένοι στον Adam, εξετάστηκαν σε 4 τιμές: $0.0005, 0.001, 0.005, 0.01$. Τέλος, οι αλγόριθμοι με βάση τον RMSProp εξετάστηκαν σε 6 ρυθμούς: 0.0003, 0.0005, 0.0007, 0.001, 0.005, 0.01.

Ο AdaLip σε συνδυασμό με τους υπόλοιπους βελτιστοποιητές φαίνεται να αποδίδει καλύτερα σε μέγιστη και σε μέση επίδοση. Επίσης, παρουσιάζει την χαμηλότερη τυπική απόκλιση, που σημαίνει ότι τα αποτελέσματά του είναι πιο σταθερά και μεταβάλλονται λιγότερο με τις

αλλαγές του αρχικού ρυθμού μάθησης. Στις εποχές για σύγκλιση παρατηρούμε μία μικρή βελτίωση στο μέσο και στη διάμεσο. Δεδομένου ότι το MNIST είναι ένα αρκετά εύκολο σύνολο δεδομένων και κάθε δίκτυο συγκλίνει αρκετά γρήγορα, είναι εμφανές ότι υπάρχει μικρός χώρος για βελτίωση.

CIFAR10

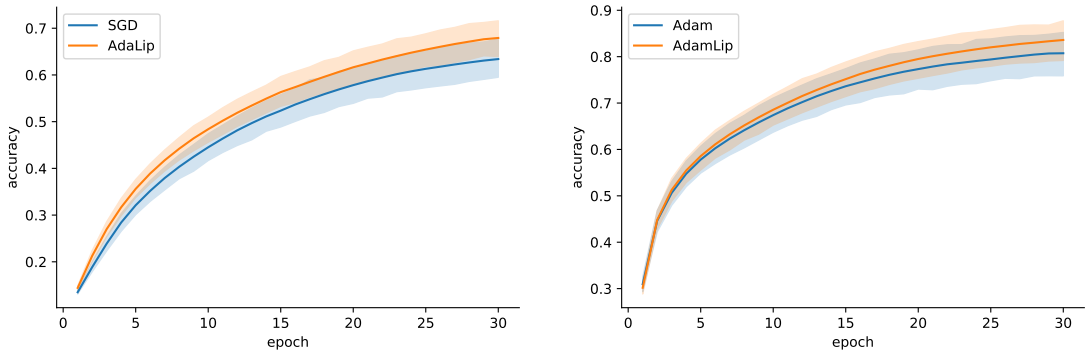
Το δεύτερο σύνολο δεδομένων είναι το CIFAR10 [55], η οποία αποτελείται από χρωματιστές εικόνες χωρισμένες σε 10 κλάσεις. Οι εικόνες κανονικοποιήθηκαν στο εύρος $[0, 1]$ και δεν εφαρμόστηκε άλλη προεπεξεργασία.

Πίνακας 5.3: Ακρίβεια εκπαίδευσης (πρώτες 4 γραμμές) και Εποχές σύγκλισης (τελευταίες 3 γραμμές) για το σύνολο δεδομένων του CIFAR10.

	SGD	AdaLip	Adam	AdamLip	RMSProp	RMSLip
max	**0.945**	0.940	0.978	**0.984**	0.975	**0.987**
median	0.756	**0.838**	0.971	**0.973**	0.968	**0.983**
mean	0.614	**0.675**	0.816	**0.826**	0.910	**0.937**
std	0.346	**0.310**	0.299	**0.298**	0.120	**0.084**
mean	**27.4**	27.6	**22**	22.571	21.2	**20.6**
median	30	**29**	**21**	22	19	**17**
best	**21**	22	**14**	**14**	16	**15**

Σε αυτό το πρόβλημα, οι βελτιστοποιητές βασισμένοι στον SGD εξετάστηκαν σε 10 αρχικούς ρυθμούς μάθησης. Αυτοί είναι οι εξής: 0.005, 0.01, 0.05, 0.07, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0. Για τους βελτιστοποιητές βασισμένους στον Adam εξετάστηκαν 7 τιμές: 0.0002, 0.0003, 0.0005, 0.0007, 0.001, 0.005, 0.01. Για τους βελτιστοποιητές με βάση τον RMSProp εξετάστηκαν 5 ρυθμοί μάθησης: 0.0003, 0.0005, 0.0007, 0.001, 0.005. Η ίδια διαδικασία ακολουθήθηκε και σε αυτό το σημείο, έχοντας παρόμοια αποτελέσματα με αυτά του MNIST.

Ο AdamLip και ο RMSLip απέδωσαν καλύτερα από τους ανταγωνιστές τους στις μέσες και μέγιστες μετρικές. Ο AdaLip είχε 0.5% λιγότερη μέγιστη ακρίβεια, αλλά στις μέσες και διάμεσες ακρίβειες είχε σημαντική βελτίωση από τον SGD. Όσον αφορά τις τυπικές αποκλίσεις, ο RMSLip παρουσιάζει πολύ καλή σταθερότητα, ενώ οι υπόλοιπες τεχνικές έχουν παρόμοια συμπεριφορά. Επίσης, παρατηρούμε ότι ο RMSLip έχει την ταχύτερη σύγκλιση από όλους τους υπόλοιπους. Ο Adam και ο SGD έχουν παρόμοιες ταχύτητες με τις αντίστοιχες AdaLip μεθόδους. Για να αναλύσουμε λίγο καλύτερα την εκπαίδευση, σχεδιάσαμε στο Σχήμα 5.2 τις μέσες καμπύλες εκπαίδευσης από όλους τους ρυθμούς μάθησης για τους βελτιστοποιητές με βάση τον Adam και τον SGD. Η σκίαση δείχνει την διακύμανση σε διαφορετικές επιλογές αρχικού ρυθμού μάθησης. Επιπλέον, φαίνεται στο σχήμα ότι οι καμπύλες του AdaLip συγκλίνουν γρηγορότερα και σε υψηλότερη ακρίβεια,



Σχήμα 5.2: Μέσες καμπύλες μάθησης από όλους τους ρυθμούς μάθησης στο CIFAR10 στους βελτιστοποιητές βασισμένους στον Adam και στον SGD.

CIFAR100

Πίνακας 5.4: Ακρίβεια εκπαίδευσης (πρώτες 4 γραμμές) και Εποχές σύγκλισης (τελευταίες 3 γραμμές) για το σύνολο δεδομένων του CIFAR100.

	SGD	AdaLip	Adam	AdamLip	RMSProp	RMSLip
max	0.965	**0.978**	0.966	**0.967**	0.956	**0.966**
median	0.935	**0.965**	0.956	**0.961**	0.953	**0.962**
mean	0.928	**0.939**	0.952	**0.958**	0.945	**0.957**
std	**0.034**	0.054	0.015	**0.010**	0.016	**0.012**
mean	60	**53.66**	44.75	**42.5**	45.4	**41.6**
median	62.5	**49**	42.5	**40.5**	40	**38**
best	44	**41**	**34**	**34**	37	**34**

Το τελευταίο σύνολο δεδομένων που χρησιμοποιήθηκε για την αξιολόγηση των αλγορίθμων είναι το CIFAR100 [55]. Αποτελείται από τις ίδιες εικόνες το CIFAR10 αλλά είναι χωρισμένο σε 100 κλάσεις, το οποίο το καθιστά αρκετά πιο δύσκολο. Για αυτό το λόγο εφαρμόστηκε ενίσχυση δεδομένων μετά την κανονικοποίηση. Η ενίσχυση δεδομένων έγινε με την χρήση τυχαίων περιστροφών, καθρεπτισμών και μετακινήσεων σε πλάτος/μήκος της εικόνας.

Εξετάστηκαν 6 διαφορετικοί αρχικοί ρυθμοί μάθησης για τους βελτιστοποιητές βασισμένους στον SGD: $0.005, 0.01, 0.05, 0.1, 0.5, 0.7$. Τέσσερις ρυθμούς μάθησης για τους αντίστοιχους του Adam: $0.0005, 0.001, 0.005, 0.01$. Για αυτούς του RMSProp εξετάστηκαν 6, οι εξής: $0.0003, 0.0005, 0.0007, 0.001, 0.005, 0.01$. Παρόμοια με τα προηγούμενα πειράματα, παρατηρούμε ότι οι βελτιστοποιητές με βάση τον AdaLip έχουν βελτιωθεί ως προς τις μέγιστες και μέσες ακρίβειες εκπαίδευσης. Επίσης, επιδεικνύουν χαμηλότερη τυπική απόκλιση στα αποτελέσματά τους. Οι εποχές σύγκλισης μας δείχνουν ότι ο AdaLip αποδίδει καλύτερα από τον SGD, ενώ ο AdamLip συγκλίνει 2 εποχές ταχύτερα κατά μέσο όρο από τον Adam. Από την άλλη, ο RMSLip επιδεικνύει μία γενικότερη βελτίωση και αποτελεί την γρηγορότερη

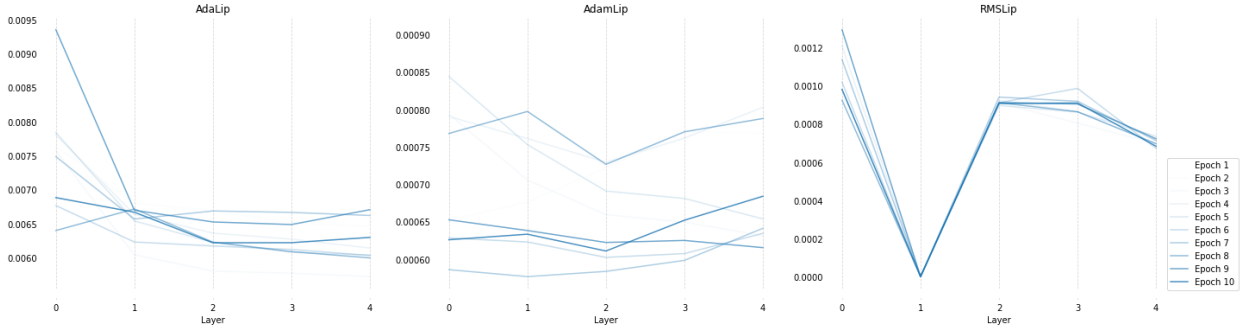
εκπαίδευση από τους υπόλοιπους.

Επιδόσεις Γενίκευσης

Ο σκοπός της μελέτης μέχρι στιγμής έχει γίνει στο πλαίσιο της επίδοσης της εκπαίδευσης, αλλά η γενίκευση ενός νευρωνικού δικτύου είναι εξίσου σημαντική. Για να μετρήσουμε τις επιδόσεις ελέγχου των διαφορετικών βελτιστοποιητών, υπολογίσαμε τον μέσο όρο των μεγίστων ακριβειών κάθε πειράματος. Αυτό σημαίνει ότι για κάθε βελτιστοποιητή και κάθε αρχικό ρυθμό μάθησης συλλέξαμε μία μέση ακρίβεια ελέγχου για τα τρία σύνολα δεδομένων. Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.5. Παρατηρούμε ότι στις περισσότερες περιπτώσεις οι επιδόσεις είναι κοντινές. Πιο συγκεκριμένα, ο AdamLip αποδίδει καλύτερα και στα τρία σύνολα δεδομένων σε σχέση με τον Adam. Ο RMSLip αποδίδει καλύτερα στο MNIST και στο CIFAR100 σε σχέση με τον RMSProp, ενώ βλέπουμε ότι στη περίπτωση του CIFAR10 ο δεύτερος φτάνει σε υψηλότερη ακρίβεια. Τέλος, ο AdaLip φαίνεται να επιδεικνύει καλύτερες ακρίβειες στο CIFAR10 και στο MNIST. Συνολοκά, οι βελτιστοποιητές βασισμένοι στη προσαρμοστική μέθοδο AdaLip δείχνουν μία μικρή βελτίωση στην γενίκευσή τους.

Πίνακας 5.5: Ακρίβειες στα σύνολα ελέγχου για τα τρία σύνολα δεδομένων MNIST, CIFAR10 και CIFAR100.

	MNIST	CIFAR10	CIFAR100
SGD	0.9896	0.6632	0.5706
AdaLip	**0.9909**	**0.6670**	**0.5802**
Adam	0.9910	**0.7545**	0.5937
AdamLip	**0.9911**	**0.7545**	**0.5981**
RMSProp	0.9902	0.7528	0.5868
RMSLip	**0.9909**	**0.7622**	**0.5884**



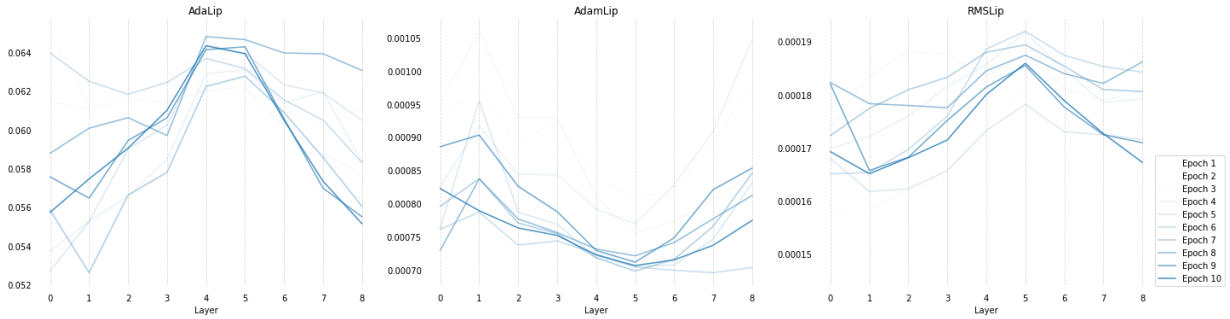
Σχήμα 5.3: Ρυθμοί μάθησης ανά επίπεδο στο σύνολο δεδομένων MNIST.

5.3.5 Σχόλια και Παρατηρήσεις

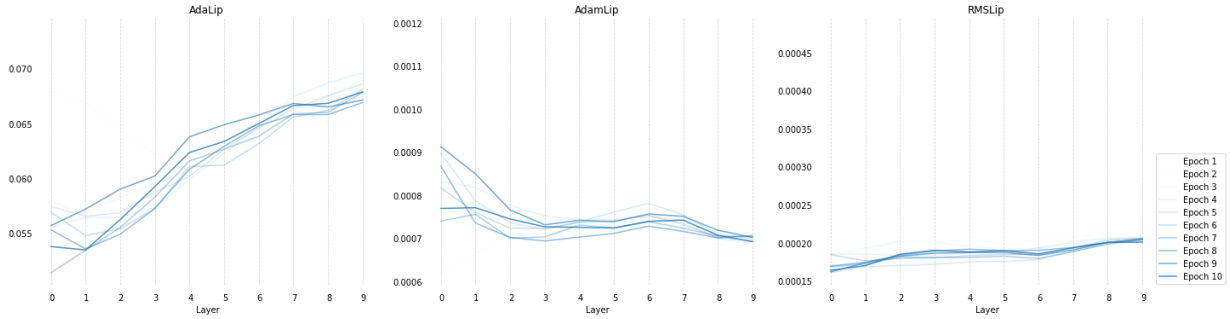
Σε αυτό το σημείο θα παρατεθούν κάποια σχόλια και παρατηρήσεις για τον προτεινόμενο αλγόριθμο αλλά και για την συνολική εκπαίδευση των βαθιών νευρωνικών δικτύων.

Ρυθμοί μάθησης ανά Επίπεδο

Όπως, έχει προαναφερθεί σε προηγούμενο κεφάλαιο, οι νόρμες των βαρών ακολουθούν ένα συγκεκριμένο σχήμα. Συγκεκριμένα, τα πρώτα και τα τελευταία επίπεδα τείνουν να παρουσιάζουν μεγαλύτερες νόρμες. Στα Σχήματα 5.3, 5.4 και 5.5 παρουσιάζονται οι ρυθμοί μάθησης ανά επίπεδο που τελικά εφαρμόζει ο αλγόριθμος. Στις περισσότερες περιπτώσεις, έχει εφαρμοστεί διαφορετικός ρυθμός μάθησης ανά επίπεδο, το οποίο φαίνεται να επικυρώνει την αρχική μας υπόθεση. Είναι σημαντικό να σημειωθεί ότι οι ρυθμοί μάθησης των AdaLip και AdamLip δείχνουν μεγαλύτερη διακύμανση στον χρόνο σε σχέση με τον RMSLip. Μία άλλη παρατήρηση είναι ότι οι ρυθμοί μάθησης δεν έχουν κάποια σταθερή τάση στο μέτρο τους σε αντίθεση με τις νόρμες των βαρών. Αυτό σημαίνει ότι ο ρυθμός μάθησης επιτυχώς προσαρμόζεται στις ανάγκες της κάθε αρχιτεκτονικής και κάθε συνόλου δεδομένων.



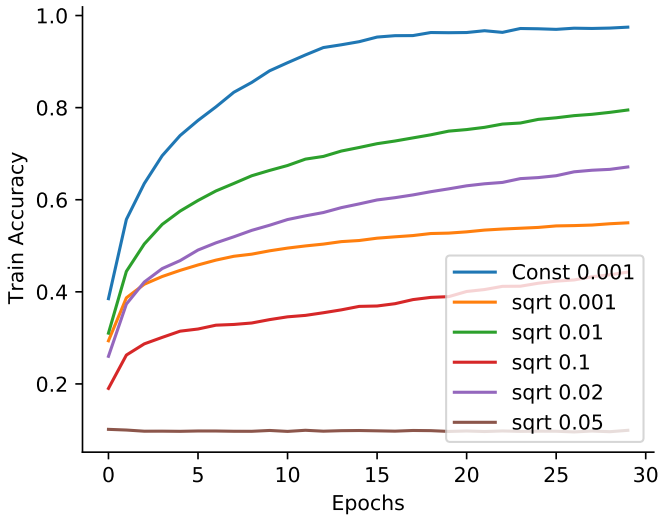
Σχήμα 5.4: Ρυθμοί μάθησης ανά επίπεδο στο σύνολο δεδομένων CIFAR10.



Σχήμα 5.5: Ρυθμοί μάθησης ανά επίπεδο στο σύνολο δεδομένων CIFAR100.

Η Υπερπαράμετρος c_t

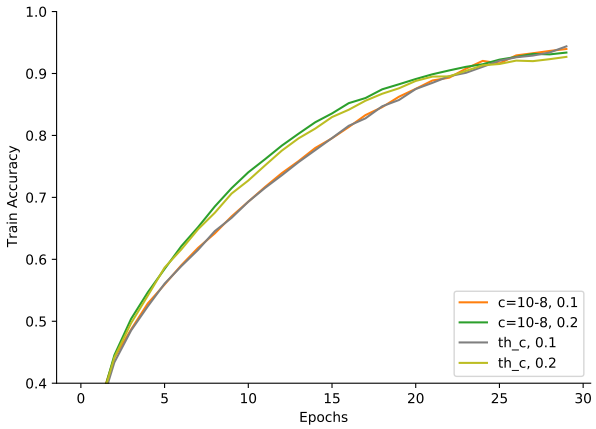
Ένα από τα κύρια σημεία παρατήρησης είναι η επιλογή της υπερπαραμέτρου c_t και πως αυτή συμπεριφέρεται και επηρεάζει τον βελτιστοποιητή αλλά και την εκπαίδευση γενικότερα. Οι δύο συνιστώσες που επηρεάζει πρακτικά είναι η θεωρητική σύγκλιση του αλγορίθμου και η πρακτική επίδοση στα σύνολα δεδομένων. Όπως αναφέρθηκε σε προηγούμενο υποκεφάλαιο, ένα πολυχρησιμοποιημένο χρονοδιάγραμμα ρυθμού μάθησης για ένα νευρωνικό, το οποίο να εγγυάται θεωρητική σύγκλιση, είναι να διαιρούμε τον ρυθμό μάθησης με την ρίζα του αριθμού των επαναλήψεων (βλέπε Εξίσωση ;;). Όμως, στην πράξη αυτή η τεχνική δεν είναι προτεινόμενη. Η πτώση του ρυθμού με την χρήση της ρίζας είναι πολύ μεγάλη για οποιαδήποτε πραγματική εφαρμογή που χρειάζεται πολλές επαναλήψεις για να εκπαιδευτεί. Με τόσες πολλές επαναλήψεις ο ρυθμός μάθησης θα συρρικνωθεί σε τεράστιο βαθμό από τα αρχικά στάδια της εκπαίδευσης, έχοντας ως αποτέλεσμα να εγκλωβιστεί σε κάποιο κακό τοπικό ελάχιστο. Από την άλλη ένας σταθερός ρυθμός μάθησης δείχνει να παράγει καλύτερα αποτελέσματα. Αυτό μπορούμε να το δούμε στο Σχήμα 5.6, που μία εκπαίδευση με σταθερό ρυθμό επιτυγχάνει καλύτερη σύγκλιση από ρυθμό βασισμένο στον θεωρητικό τύπο.



Σχήμα 5.6: Σταθερός ρυθμός μάθησης σε σύγκριση με ρυθμό μάθησης με πτώση τετραγωνικής ρίζας πάνω στον Adam στο CIFAR10.

Τελευταίες μελέτες [100, 67, 41] δείχνουν ότι ένας ρυθμός μάθησης που ταλαντεύεται αποδίδει πολλές φορές καλύτερα από έναν ρυθμό που απλά ελαττώνεται στο πέρας του χρόνου. Η καλύτερη απόδοση τέτοιων τεχνικών βασίζεται στο ότι ο βελτιστοποιητής είναι ικανός με αυτόν τον τρόπο να ξεπεράσει μη-βέλτιστα τοπικά ελάχιστα. Παρομοίως με την περίπτωση του c_t, η θεωρητική τιμή του τιμή έχει παραπλήσια αποτελέσματα με την σταθερή του τιμή, εκτός από κάποια πειράματα που η σταθερή τιμή επιδεικνύει καλύτερη επίδοση. Αυτό μπορούμε να το δούμε στο Σχήμα 5.7, όπου όλα τα πειράματα με σταθερό ρυθμό μάθησης είναι αρκετά κοντά μεταξύ τους, αλλά το σταθερό c_t δείχνει μία βελτίωση. Αυτή η σχέση μεταξύ της θεωρητικής εγγύησης της σύγκλισης και των πειραματικών αποτελεσμάτων είναι πολύ σημαντική και μπορεί

να εκφραστεί και ως η σχέση μεταξύ σταθερότητας και καλύτερης απόδοσης. Ένας από τους λόγους που συμβαίνει αυτό βασίζεται στην φύση του αλγορίθμου SGD. Για να αποδειχθεί θεωρητικά ότι ο όρος της μετάνοιας του SGD (ή κάποιου βελτιστοποιητή βασισμένου σε αυτόν) συγκλίνει στο 0 με το πέρας του χρόνου, διάφορες υποθέσεις γίνονται για την τοπολογία της ολικής συνάρτησης κόστους. Στην πράξη, όμως, η τοπολογία μπορεί να μην είναι ιδανική και η σύγκλιση στο πρώτο τοπικό ελάχιστο να μην είναι ικανοποιητική [18, 62]. Για αυτό τον λόγο μελέτες πάνω στην διαφυγή τοπικών ελαχίστων ή σημείων σέλας έχουν ιδιαίτερη προσοχή στο επιστημονικό πεδίο της βελτιστοποίησης [52, 31, 46].



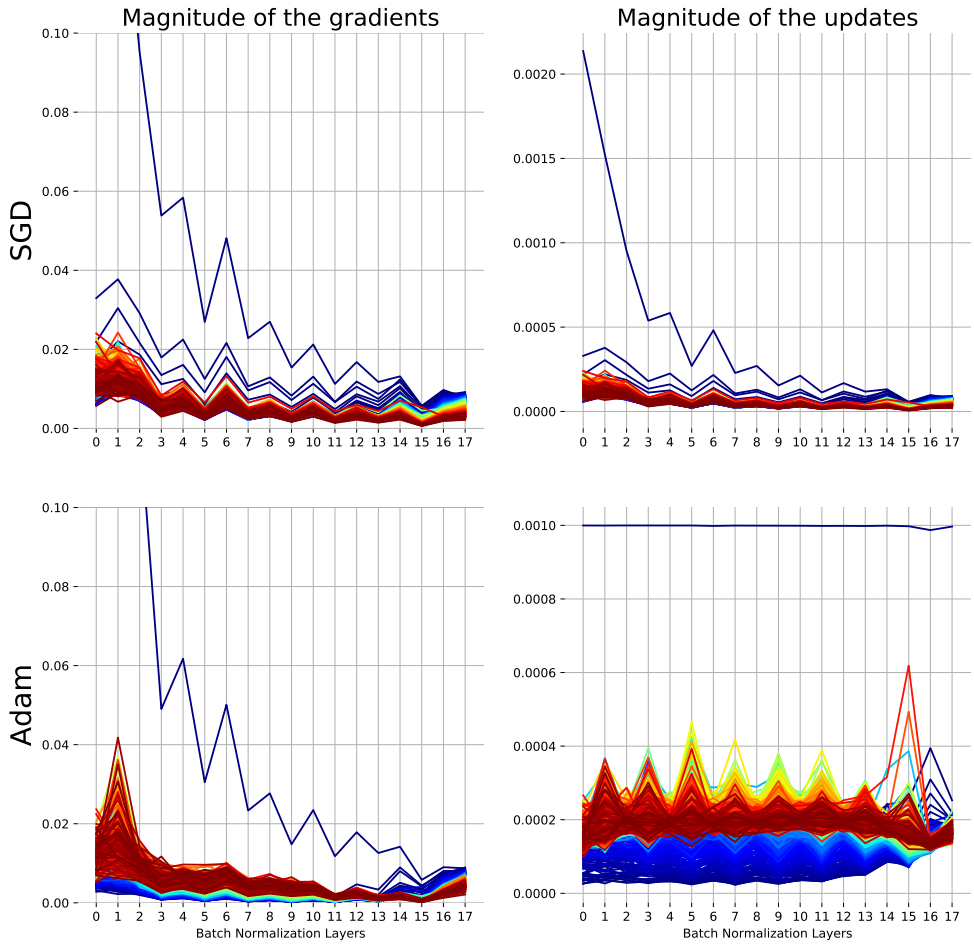
Σχήμα 5.7: Πειράματα με διαφορετικούς ρυθμούς μάθησης $(0.1, 0.2)$ και διαφορετικές τιμές του c_t (θεωρητικό c_t, 10^{-8}) χρησιμοποιώντας τον AdaLip στο CIFAR10.

Βέλτιστος ρυθμός μάθησης με βάση τον SGD

Για να βρούμε τον βέλτιστο ρυθμό μάθησης, χρησιμοποιήθηκε το Λήμμα 5.1, το οποίο είναι βασισμένο στον αλγόριθμο της Καθόδου Κλίσης (Gradient Descent). Θα πρέπει να σημειωθεί, όμως, ότι η εξίσωση βασισμένη στον SGD ελέγχθηκε και δεν χρησιμοποιήθηκε. Διάφορα πειράματα διεξήχθησαν με αυτόν τον τύπο και έδειξαν ότι οδηγεί σε ασταθή διαδικασία μάθησης. Στις περιπτώσεις που το δίκτυο εκπαιδεύτηκε κανονικά, η επιδόσεις του δεν ήταν ανταγωνιστικές με τους άλλους βελτιστοποιητές και η ταχύτητα σύγκλισής του ήταν χαμηλότερη. Στο Παράρτημα Β΄ βρίσκεται το Λήμμα 2.6, που περιγράφει τον τύπο για τον βέλτιστο ρυθμό μάθησης χρησιμοποιώντας την εξίσωση του SGD. Θα είχε αρκετό ενδιαφέρον σαν μελλοντική μελέτη να βρεθεί ένας τρόπος ώστε να εφαρμοστεί αυτός ο τύπος αποδοτικά.

Επίδραση της Κανονικοποίησης Παρτίδας

Μία ενδιαφέρουσα παρατήρηση για τις νόρμες των επιπέδων εμφανίζεται με την χρήση επιπέδων Κανονικοποίησης Παρτίδας. Στο Σχήμα 5.8 μπορούμε να δούμε το μέγεθος των βαρών του επιπέδου αυτού. Είναι εμφανές ότι δεν ακολουθεί το μοτιβό του σχήματος 5.1 που απεικονίζει το μέγεθος των βαρών Συνελικτικών και Πλήρως Συνδεδεμένων Επιπέδων. Αυτό



Σχήμα 5.8: Το μέτρο των παραγώγων και των ανανεώσεων των βαρών των επιπέδων Κανονικοποίησης Παρτίδας (στα γάμμα και βήτα βάρη) στο σύνολο δεδομένων του CIFAR10. Οι μπλε γραμμές δείχνουν τις αρχικές επαναλήψεις, ενώ όσο αλλάζει το χρώμα προς το κόκκινο τις τελευταίες.

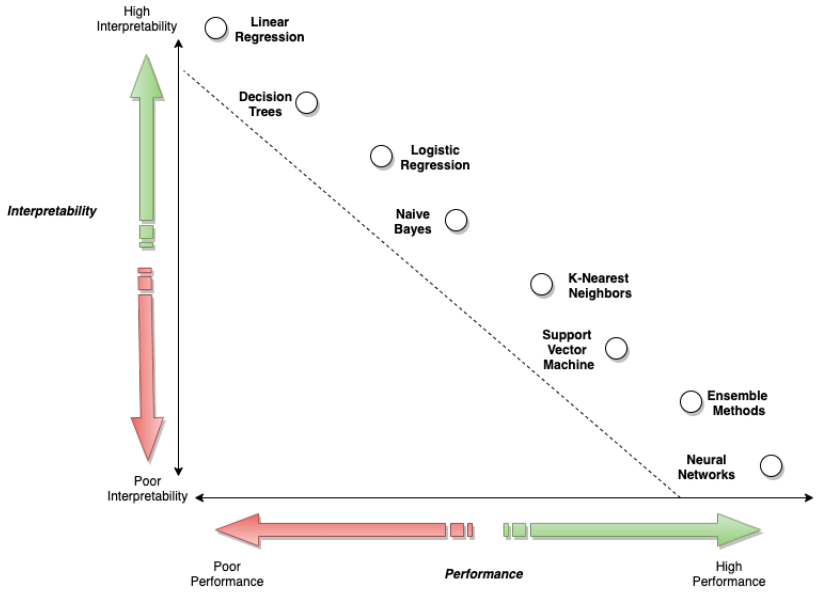
σημαίνει ότι τα βάρη του επιπέδου Κανονικοποίησης Παρτίδας (¨γάμμα' και ¨βήτα¨) συμπεριφέρονται διαφορετικά σε σχέση με τα υπόλοιπα βάρη του δικτύου. Παρόλο που το επίπεδο Κανονικοποίησης Παρτίδας συμβάλει στο να λειαίνει τις παραγώγους άλλων βαρών προς όφελος της εκπαίδευσης [43], παρατηρούμε ότι μπορεί να πάσχει από το ίδιο πρόβλημα στα δικά του βάρη. Αυτός είναι ο λόγος που ο AdaLip μπορεί να κατασκευάσει έναν κατάλληλο ρυθμό μάθησης για όλες τις παραμέτρους διαφορετικών επιπέδων. Αυτό ενισχύεται από το γεγονός ότι στα πειράματα με δίκτυα με Κανονικοποίηση Παρτίδας, όπως στα δεδομένα του CIFAR100, ο βελτιστοποιητής AdaLip απέδωσε αρκετά καλύτερα από άλλους.

Κεφάλαιο 6

Ερμηνευσιμότητα

Τα τελευταία χρόνια, το πεδίο της τεχνητής νοημοσύνης και της μηχανικής μάθησης βίωσε μια εκρηκτική ανάπτυξη. Τα μοντέλα μηχανικής μάθησης έχουν γίνει ένα χρήσιμο εργαλείο σε διάφορες εφαρμογές σε διάφορους επιστημονικούς τομείς. Το μέγεθος και η πολυπλοκότητα τέτοιων μοντέλων είναι οι λόγοι για τους οποίους συχνά θεωρούνται ¨μαύρα κουτιά' [13]. Ένα ¨μαύρο κουτί' είναι ένα μοντέλο του οποίου η διαδικασία λήψης αποφάσεων και ο συλλογισμός του είναι δύσκολα για έναν άνθρωπο να τα κατανοήσει. Ωστόσο, η δυνατότητα ανάλυσης και κατανόησης του τρόπου με τον οποίο ένα μοντέλο κάνει μια πρόβλεψη είναι εξαιρετικά σημαντική. Η εξήγηση των προβλέψεων ενός μοντέλου μπορεί να βοηθήσει στην ενίσχυση της εμπιστοσύνης που έχουν οι άνθρωποι για το μοντέλο, καθώς και στον έλεγχο της αξιοπιστίας του μοντέλου, την ανάλυσης δικαιοσύνης και τη δυνατότητα ενεργοποίησης μιας μορφής εντοπισμού σφαλμάτων του μοντέλου. Έχουν προταθεί διάφορες μέθοδοι [114, 12] για τη βοήθεια των χρηστών στην ερμηνεία των προβλέψεων τέτοιων πολύπλοκων μοντέλων, ειδικά των νευρωνικών δικτύων, τα οποία θα αποτελέσουν το επίκεντρο αυτού του κεφαλαίου.

Πιο συγκεκριμένα, θα εμβαθύνουμε στις έννοιες της ερμηνευσιμότητας και της επεξηγηματικότητας. Η ερμηνευσιμότητα αναφέρεται στη δυνατότητα να κατανοήσουμε την εσωτερική λειτουργία ενός μοντέλου ή αλγορίθμου, επιτρέποντας σε έναν άνθρωπο να κατανοήσει πώς φτάνει σε συγκεκριμένες αποφάσεις ή προβλέψεις. Επικεντρώνεται στη διαφάνεια και την κατανόηση του μοντέλου. Από την άλλη πλευρά, η επεξηγηματικότητα εξετάζει βαθύτερα, αφού όχι μόνο αποκαλύπτει τις εσωτερικές διαδικασίες του μοντέλου αλλά παρέχει επίσης μια αφήγηση ή δικαιολογία για τα αποτελέσματά του. Ουσιαστικά, η επεξηγηματικότητα στοχεύει στο να απαντήσει στο ΄γιατί΄ πίσω από τις αποφάσεις ενός μοντέλου, ενώ η ερμηνευσιμότητα ασχολείται κυρίως με το ΄πώς΄. Και οι δύο ερμηνευσιμότητα και επεξηγηματικότητα είναι ουσιαστικές για την κατασκευή εμπιστοσύνης και τη διασφάλιση της ευθύνης στα συστήματα τεχνητής νοημοσύνης, αλλά αντιπροσωπεύουν διακριτικές πτυχές της διαφάνειας του μοντέλου, αντιμετωπίζοντας τις ερωτήσεις ΄τι΄ και ΄γιατί΄ αντίστοιχα. Αυτές οι έννοιες είναι απαραίτητες όχι μόνο για την κατασκευή εμπιστοσύνης στην τεχνητή νοημοσύνη, αλλά και για την εξασφάλιση ότι τα συστήματα τεχνητής νοημοσύνης μπορούν να αξιοποιηθούν αποτελεσματικά σε ευαίσθητους τομείς όπως η υγεία, οι χρηματοοικονομικές υπηρεσίες ή διάφορα αυτόνομα



Σχήμα 6.1: Δίλημμα Επεξηγησιμότητας - Επίδοσης

συστήματα.

Το Σχήμα 6.1 δείχνει την σχέση που έχουν η επίδοση μοντέλων με την ερμηνευσιμότητά τους. Στον χ-άξονα βλέπουμε τις επιδόσεις των διαφόρων μοντέλων, που διακυμαίνονται από χαμηλή επίδοση σε υψηλή. Αντίστοιχα, στον ψ-άξονα παρατηρούμε την διακύμανση της ερμηνευσιμότητας. Από τα διάφορα μοντέλα και το πως κατανέμονται στο σχήμα βλέπουμε ότι υπάρχει ένα δίλημμα. Μοντέλα που είναι απλά έχουν χαμηλές επιδόσεις αλλά είναι αρκετά ερμηνεύσιμα. Παραδείγματα τέτοιων μοντέλων είναι η γραμμική παλινδρόμηση ή τα δέντρα αποφάσεων. Από την άλλη βλέπουμε μοντέλα που είναι αρκετά πολύπλοκα και καταφέρνουν καλές επιδόσεις, όπως τα νευρωνικά δίκτυα, τα οποία δεν είναι εύκολα ερμηνεύσιμα από τον άνθρωπο. Ο σκοπός της έρευνας στο πεδίο της ερμηνευσιμότητας τα τελευταία χρόνια είναι η δημιουργία τεχνικών που να μπορέσουν να εξηγήσουν τέτοια πολύπλοκα μοντέλα ή να κατασκευάσουν νέα μοντέλα και αρχιτεκτονικές που να είναι ερμηνεύσιμες από μόνες τους χωρίς να θυσιάζουν καθόλου τις επιδόσεις τους. Στα επόμενα υποκεφάλαια θα αναλύσουμε διάφορες δημοφιλείς τεχνικές επεξηγησιμότητας, εφαρμογές σε διαφόρων ειδών δεδομένων, καθώς και θα εμβαθύνουμε περισσότερο στη βελτίωση τέτοιων τεχνικών.

6.1 Τεχνικές Επεξηγηματικότητας

Σε αυτό το υποκεφάλαιο θα περιγράψουμε τις πιο δημοφιλείς τεχνικές επεξηγηματικότητας και θα αναλύσουμε τις λειτουργίες τους καθώς και τα πλεονεκτήματα ή τυχόν μειονεκτήματα που παρουσιάζει η καθεμία. Τα τελευταία χρόνια έχουν αναπτυχθεί μία πληθώρα αλγορίθμων, οι οποίοι βασίζονται σε διαφορετικές ιδέες για το πως μπορεί να μετρηθεί η συνεισφορά ενός χαρακτηριστικού στην πρόβλεψη ενός μοντέλου. Η σημαντικότητα μιας εισόδου έχει χαρακτηριστεί με διάφορες ονομασίες στην βιβλιογραφία του πεδίου της Ερμηνευσιμότητας (π.χ.

importance, attribution, contribution, impact). Σε αυτή τη διατριβή θα αναφερόμαστε σε αυτές τις έννοιες ως Σημασία ενός χαρακτηριστικού και θα την συμβολίζουμε με Φ. Οι αλγόριθμοι αυτοί μπορούν να χωριστούν σε βασικές κατηγορίες ανάλογα με την τεχνική που χρησιμοποιούν. Παρακάτω, αναφέρουμε κάποιες που συναντάμε πολύ συχνά στην βιβλιογραφία.

6.1.1 Τεχνικές με παράγωγο

Αυτές οι μέθοδοι χρησιμοποιούν τις παραγώγους της εξόδου ως προς τις εισόδους για να βρουν την Σημασία των χαρακτηριστικών. Έστω y_i είναι η i-οστή έξοδος ενός δικτύου (συνήθως αντιστοιχίζεται με την i-οστή κλάση σε περίπτωση προβλήματος ταξινόμησης), x είναι μία είσοδος και x_i είναι το i-οστό χαρακτηριστικό αυτής της εισόδου. Τότε συμβολίζουμε την παράγωγο της εξόδου ως προς το i-οστό χαρακτηριστικό ως εξής:

$$S_i = \frac{\partial y_c(x_i)}{\partial x_i} \tag{6.1}$$

Αυτή η τεχνική συναντάται στην βιβλιογραφία σε μία πρώιμη μορφή στους Χάρτες Εστίασης [98], η οποία είναι μία μέθοδος που βρίσκει την Σημασία των εικονοστοιχείων σε δεδομένα εικόνας. Άλλο παράδειγμα τέτοιας μεθόδου είναι ο αλγόριθμος Integrated Gradients - IG (Ολοκληρωμένες Κλίσεις) [103], ο οποίος υπολογίζει το ολοκλήρωμα των παραγώγων ως προς την είσοδο πάνω σε κάποιο μονοπάτι. Βασισμένο στον αλγόριθμο IG έχει αναπτυχθεί η μέθοδος SmoothGrad, η οποία με την προσθήκη θορύβου στα δείγματα καταφέρνει να παράγει Σημασίες που είναι αρκετά πιο ακριβείς. Στο πεδίο της Ερμηνευσιμότητας συγκεκριμένα πάνω σε συνελικτικά επίπεδα (CNN) έχουν αναπτυχθεί 2 αρκετά σημαντικές τεχνικές που χρησιμοποιούν παραγώγους. Αυτές οι μέθοδοι είναι η GradCAM [89] και η επέκτασή της GradCAM++ [14] και προσπαθούν να δώσουν έμφαση στα κομμάτια της εικόνας, στα οποία έδωσε περισσότερη προσοχή το νευρωνικό κατά τη διάρκεια της πρόβλεψης. Συγκεκριμένα, η GradCAM υπολογίζει τις παραγώγους από το τελευταίο συνελικτικό επίπεδο της αρχιτεκτονικής του δικτύου και τις αθροίζει με κατάλληλα βάρη ώστε να φτιάξει έναν θερμικό χάρτη που δείχνει τα σημεία με μεγαλύτερη Σημασία. Από την άλλη ο GradCAM++ επεκτείνει την ιδέα αυτή τροποποιώντας τις παραγώγους με τέτοιο τρόπο ώστε να παράγεται πιο σταθερή και ακριβής απεικόνιση της Σημασίας των εικόνων.

6.1.2 Τεχνικές με διαταραχή εισόδου

Οι τεχνικές με διαταραχή εισόδου βασίζονται στην ιδέα ότι τα χαρακτηριστικά που οφείλονται περισσότερο στην πρόβλεψη του μοντέλου είναι αυτά, τα οποία αλλάζουν την έξοδο του περισσότερο όταν μεταβάλλονται. Για αυτό τέτοιες μέθοδοι αλλάζουν την είσοδο του δικτύου με συγκεκριμένες στρατηγικές για να ανακαλύψουν ποιο χαρακτηριστικό συμβάλλει πιο καθοριστικά στην εκάστοτε πρόβλεψη. Ένα παράδειγμα είναι ο αλγόριθμος SHAP [68], ο οποίος βασίζεται στον υπολογισμό των τιμών Shapley [93, 63]. Περισσότερα για αυτόν τον

αλγόριθμο θα αναφερθούν σε ακόλουθο υποκεφάλαιο. Άλλος πολύ δημοφιλής αλγόριθμος είναι ο LIME [85], που παίρνει δείγματα από το σύνολο δεδομένων, μεταβάλλει την είσοδο (σε μία πιο τοπική περιοχή) και προσπαθεί να εκπαιδεύσει ένα μικρό ερμηνεύσιμο μοντέλο (συνήθως γραμμική παλινδρόμηση) που θα προσομοιάζει το νευρωνικό δίκτυο.

Μία άλλη μέθοδος που βασίζεται στην τροποποίηση της εισόδου είναι ο αλγόριθμος DeepLIFT [96]. Αποτελεί μια μέθοδος ερμηνείας για νευρωνικά δίκτυα, η οποία στοχεύει στο να αναδείξει τη σημαντικότητα κάθε χαρακτηριστικού εισόδου στο σύνολο της πρόβλεψης του μοντέλου. Χρησιμοποιώντας ένα σημείο αναφοράς, συγκρίνει τις εξόδους κάθε νευρώνα του δικτύου και παρατηρεί τυχόν θετικές ή αρνητικές επιδράσεις των χαρακτηριστικών. Είναι σημαντικό να σημειωθεί ότι το DeepLIFT αποτελεί μια διαφορετική προσέγγιση που βασίζεται στις τιμές Shapley, αλλά είναι αρκετά αποτελεσματική και χρησιμοποιείται συχνά για διάφορες εφαρμογές εύρεσης σημασίας χαρακτηριστικών.

6.2 Χρήση Τεχνικών Ερμηνευσιμότητας σε Εφαρμογές Κατάτμησης Εικόνας

Σε αυτό το υποκεφάλαιο θα ασχοληθούμε με την εφαρμογή μεθόδων Ερμηνευσιμότητας και πως αυτές μπορούν να χρησιμοποιηθούν για προβλήματα κατάτμησης εικόνας. Συγκεκριμένα, θα εμβαθύνουμε στο πεδίο της Διαβητικής Αμφιβληστροειδοπάθειας (Diabetic Retinopathy - DR). Η πάθηση αυτή είναι μία ασθένια του οφθαλμού που επηρεάζει τον αμφιβληστροειδή, ο οποίος είναι το επίπεδο του ματιού που μετατρέπει το φως σε ηλεκτρικά σήματα για τον εγκέφαλο. Η πάθηση DR συνήθως προκαλεί θάμπωμα της όρασης, πόνος στο μάτι μέχρι και τύφλωση. Ανάλογα με την ζημιά που έχουν υποστεί τα αγγεία του οφθαλμού, υπάρχουν 4 στάδια της πάθησης: Ήπιο (mild), Μέτριο (moderate), Σοβαρό (severe) και Προχωρημένο (proliferative). Τα πρώτα στάδια είναι αρκετά δύσκολο να διαγνωστούν για αυτό και η διάγνωσή τους έχει ύψιστη σημασία.

Πολλές αρχιτεκτονικές βαθιάς μηχανικής μάθησης έχουν χρησιμοποιηθεί σε εφαρμογές διαβητικής αμφιβληστροειδοπάθειας, οι οποίες συνήθως βασίζονται σε Συνελικτικά επίπεδα [23] με διάφορες παραλλαγές. Σε αυτό το υποκεφάλαιο θα εξερευνήσουμε δημοφιλείς αρχιτεκτονικές, θα τις συγκρίνουμε ως προς τις επιδόσεις τους και θα εφαρμόσουμε πάνω σε αυτές διάφορες τεχνικές Ερμηνευσιμότητας. Σε συνέχεια αυτών θα δείξουμε ότι οι τεχνικές αυτές είναι ικανές με κατάλληλες μετατροπές να βοηθήσουν στην λύση του προβλήματος της κατάτμησης εικόνας. Σε συνδυασμό με τη χρήση επισημειωμένων δεδομένων και την γνώση της πραγματικής κατάτμησης μιας εικόνας μπορούμε να αξιολογήσουμε άμεσα και αντικειμενικά τις επιδόσεις των τεχνικών Ερμηνευσιμότητας.

6.2.1 Βιβλιογραφία

Αρχικά, ας αναφέρουμε κάποιες σημαντικές εργασίες πάνω στη Διαβητική Αμφιβληστοειδοπάθεια. Το 2016 ο Gulshan et al. δημοσιεύσε μία αρχιτεκτονική βασισμένη σε αυτή του Inception-v3 πάνω σε ειδικές εικόνες fundus του αμφιβληστοειδούς [34]. Το μοντέλο αξιολογήθηκε πάνω σε 2 σύνολα δεδομένων με DR εικόνες: το EyePACS [48] και το Messidor [22]. Όσον αφορά τις επιδόσεις το μοντέλο φτάνει 93.4% και 97.5% για τις μετρικές 'Specificity' και 'Sensitivity' αντίστοιχα στο EyePACS σύνολο δεδομένων [48] και 93.9%, 96.1%, αντιστοίχως, στο Messidor [22]. Εκτός αυτού διάφορες μέθοδοι Συστάδων Μοντέλων (Ensembles), που αποτελούνται μοντέλα βαθιάς μάθησης αλλά και κλασσικής μηχανικής μάθησης, έχουν προταθεί σαν βελτιώσεις σε ήδη υπάρχοντα μοντέλα. Στην μελέτη [81], μία ομάδα από μοντέλα βασισμένα σε συνελικτικά δίκτυα (CNN), όπως το ResNet [39], το DenseNet [40], το Inception [104] και το Xception εκπαιδεύτηκαν στο σύνολο δεδομένων του Kaggle και κατάφεραν αποτελεσματικά να ταξινομήσουν εικόνες αμφιβληστροειδούς στα αντίστοιχα στάδια πάθησής τους. Άλλες συστάδες μοντέλων έχουν βασιστεί σε μία μίξη από μοντέλα γραμμικής παλινδρόμησης, μοντέλα (k-NN), δέντρα αποφάσεων και Random Forest στο [84]. Αυτές οι δύο προσεγγίσεις απέδειξαν την χρησιμότητα των συστάδων μοντέλων σε αντίθεση με τα μοντέλα όταν λειτουργούν μόνα τους.

Πρόσφατα, μία αρχιτεκτονική που συνδύασε το ResNet και Random Forest προτάθηκε για το ίδιο πρόβλημα [110]. Σε αυτή τη μέθοδο το επίπεδο συμψηφισμού χρησιμοποιείται για την εξαγωγή χαρακτηριστικών υψηλού επιπέδου. Αυτά τα χαρακτηριστικά μετά δόθηκαν ως είσοδο σε ένα Random Forest για την τελική πρόβλεψη. Αυτή η προσέγγιση είχε καλύτερες επιδόσεις από αρκετούς αλγόριθμους τελευταίας τεχνολογίας επιτυγχάνοντας ακρίβεια 96.0% και 75.09% στο Messidor [22] και στο EyePACS συνόλο δεδομένων [48] αντίστοιχα.

Στάδια της πάθησης	IDRiD		DDR	
	Εκπαίδευση	Έλεγχος	Εκπαίδευση	Έλεγχος
No DR	134	34	3133	1880
Mild	20	5	315	189
Moderate	136	32	2238	1344
Severe	74	19	118	71
Proliferative (PDR)	49	13	456	275
Ungradable	-	-	575	346
Συνολικά	413	103	6835	4105

Πίνακας 6.1: Κατανομή των δειγμάτων στα στάδια της πάθησης για τα 2 σύνολα δεδομένων IDRiD και DDR

Τύποι μασκών	IDRiD	DDR
MA	81	570
HE	80	601
EX	81	486
SE	40	239
OD	81	-
Συνολικά	81	757

Πίνακας 6.2: Κατανομή των διαφορετικών τύπων μασκών κατάτμησης για τα σύνολα δεδομένων IDRiD και DDR

6.2.2 Πειραματική Διαδικασία

Στην πειραματική διαδικασία χρησιμοποιήσαμε 2 από τα πιο γνωστά σύνολα δεδομένων για DR για εκπαίδευση νευρωνικών δικτύων. Στις επόμενες υποενότητες θα χρησιμοποιήσουμε γνωστές αρχιτεκτονικές και θα τις εκπαιδεύσουμε στα δεδομένα. Έπειτα, θα εφαρμόσουμε τεχνικές ερμηνευσιμότητας στα εκπαιδευμένα μοντέλα και θα τις αξιολογήσουμε για την ποιότητα της Σημασίας που δείχνουν ως προς την σωστότερη κατάτμηση εικόνας.

Δεδομένα Διαβητικής Αμφιβληστροειδοπάθειας

Το πρώτο σύνολο δεδομένων που θα χρησιμοποιήσει η μελέτη μας είναι το Indian Diabetic Retinopathy Image Dataset (IDRiD) [80]. Το IDRiD αποτελεί ένα σύνολο fundus εικόνων του αμφιβληστροειδούς της ινδικής πληθυσμιακής ομάδας και περιλαμβάνει τρία κύρια προβλήματα προς λύση: ταξινόμηση της διαβητικής αμφιβληστροειδοπάθειας, κατάτμηση των εικόνων ως προς την πάθηση και εντοπισμό των σημείων που παρατηρείται η πάθηση. Συγκεκριμένα, θα επικεντρωθούμε στα προβλήματα ταξινόμησης της διαβητικής αμφιβληστροειδοπάθειας και κατάτμησης. Το σύνολο δεδομένων ταξινόμησης της διαβητικής αμφιβληστροειδοπάθειας αποτελείται από 516 εικόνες αποθηκευμένες σε μορφή JPEG. Οι εικόνες έχουν ανάλυση 4288×2848 εικονοστοιχείων, που θεωρείται πολύ υψηλής ποιότητας. Τα στάδια διαιρούν το σύνολο σε 5 κλάσεις. Ο βαθμός σοβαρότητας αποτελείται από την κλάση 0 (που υποδηλώνει απουσία εμφανούς διαβητικής αμφιβληστροειδοπάθειας - no DR), την κλάση 1 (ήπια διαβητική αμφιβληστροειδοπάθεια - Mild), κλάση 2 (μέτρια διαβητική αμφιβληστροειδοπάθεια - Moderate), κλάση 3 (σοβαρή διαβητική αμφιβληστροειδοπάθεια - Severe) και, τέλος, κλάση 4 (προχωρημένη διαβητική αμφιβληστροειδοπάθεια - Proliferative). Το σύνολο δεδομένων χωρίζεται σε ένα σύνολο εκπαίδευσης, που αποτελείται από 413 εικόνες, και ένα σύνολο ελέγχου με 103 εικόνες. Ο πίνακας 6.1 δείχνει τη κατανομή των κλάσεων στα δύο αυτά σύνολα. Το πρόβλημα της κατάτμησης αποτελείται από 81 εικόνες, διαιρεμένες σε 54 για εκπαίδευση και 27 για έλεγχο. Υπάρχουν 5 διαφορετικά είδη μασκών κατάτμησης: Microaneurysms (MA), Haemorrhages (HE), Hard Exudates (EX), Soft Exudates (SE) και Optic Disk (OD). Η κατανομή των παραπάνω μπορεί να βρεθεί στον πίνακα 6.2.

Το δεύτερο σύνολο δεδομένων που θα χρησιμοποιηθεί είναι το DDR [64]. Παρόμοια με το IDRiD, υπάρχουν τα ίδια τρία προβλήματα και για αυτό το σύνολο δεδομένων. Το πρόβλημα της ταξινόμησης της διαβητικής αμφιβληστροειδοπάθειας αποτελείται από 13.673 έγχρωμες fundusεικόνες που λήφθηκαν από 147 νοσοκομεία. Οι εικόνες κατηγοριοποιούνται σε 6 κλάσεις, με τις πρώτες 5 να είναι οι ίδιες με το IDRiD και η 6 κλάση να περιλαμβάνει εικόνες χαμηλής ποιότητας που δεν μπορούν να κατηγοριοποιηθούν αλλού. Αυτή η κλάση ονομάζεται ¨Ακατάλληλη για ταξινόμηση - Ungradable' . Η κατανομή των κλάσεων στα σύνολα εκπαίδευσης και ελέγχου φαίνεται στον πίνακα 6.1. Σε αντίθεση με το IDRiD, οι αναλύσεις των εικόνων ποικίλουν. Για παράδειγμα, υπάρχουν εικόνες με αναλύσεις 1137×1470 εικονοστοιχεία, καθώς και εικόνες 3456×5184 εικονοστοιχείων. Από την άλλη, το πρόβλημα κατάτμησης αποτελείται από 757 εικόνες, που παρέχονται με τα ίδια είδη μασκών κατάτμησης με το IDRiD, εκτός από την μάσκα Optic Disk (OD). Η κατανομή των μασκών φαίνεται στον πίνακα 6.2.

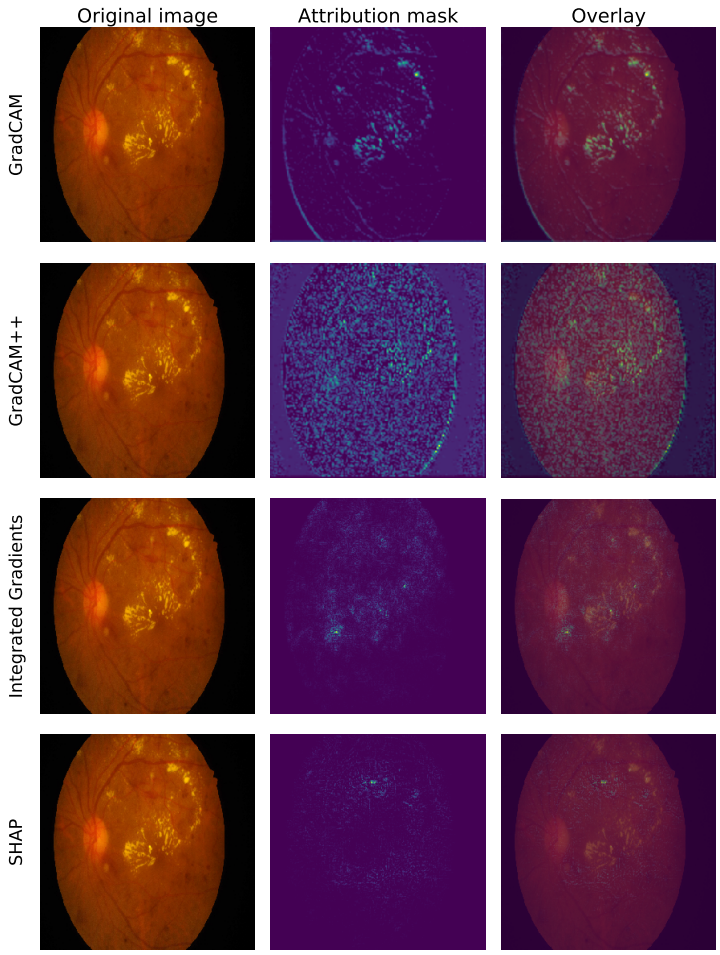
6.2.3 Αρχιτεκτονικές Δικτύων και Ερμηνευσιμότητα

Στην παρούσα εργασία, εξετάσαμε τρία διαφορετικά αρχιτεκτονικά δίκτυα βαθιάς μάθησης CNN προκειμένου να επαληθεύσουμε την αντικειμενικότητα των πειραμάτων μας. Πιο συγκεκριμένα, εκπαιδεύσαμε τα μοντέλα DenseNet [40], InceptionV3 [104] και EfficientNetB0 [105] για ταξινόμηση της διαβητικής αμφιβληστροειδοπάθειας στα σύνολα δεδομένων IDRiD και DDR ξεχωριστά, με αποτέλεσμα την εκπαίδευση 6 μοντέλων. Όλα τα μοντέλα αρχικοποιήθηκαν με προεκπαιδευμένα βάρη από το ImageNet [23]. Η ενίσχυση των εικόνων πραγματοποιήθηκε με τυχαίες μεγεθύνσεις, περιστροφές και αναστροφές. Μετατρέψαμε όλες τις εικόνες από τα δύο σύνολα δεδομένων σε μέγεθος 300×300 εικονοστοιχείων, για να διευκολυνθεί η εκπαίδευση των δικτύων. Τα δίκτυα εκπαιδεύτηκαν για 20 εποχές χρησιμοποιώντας τον βελτιστοποιητή Adam [51].

Τα 6 δίκτυα πέτυχαν την ακρίβεια ελέγχου που φαίνεται στους πίνακες 6.3 και 6.4 για τα δύο σύνολα δεδομένων αντίστοιχα. Είναι αξιοσημείωτο ότι το EfficientNetB0 αποδείχθηκε ως η πιο κατάλληλη αρχιτεκτονική για αυτή το πρόβλημα, με συνολική ακρίβεια $60,19\%$ στο IDRiD και $73,56\%$ στο DDR. Ένα κοινό αδύναμο σημείο των τριών δικτύων φαίνεται να είναι η δυσκολία στον εντοπισμό του πρώτου σταδίου (Ήπιο) της νόσου, όπως συμβαίνει με τις περισσότερες προσεγγίσεις τελευταίας τεχνολογίας, ενώ επιδεικνύουν ικανοποιητική απόδοση

Αρχιτεκτονική	No DR	Mild	Moderate	Severe	PDR	Συνολική Ακρίβεια
DenseNet	0.7647	0.0	0.7812	0.1578	0.1538	0.5436
InceptionV3	0.8235	0.2	0.6250	0.5789	0.0	0.5825
EfficientNetB0	1.0	0.0	0.5937	0.3684	0.1538	0.6019

Πίνακας 6.3: Οι ακρίβειες του συνόλου Ελέγχου για κάθε κλάση και η συνολική ακρίβεια πάνω στις 3 αρχιτεκτονικές που εφαρμόστηκαν στο IDRiD.



Σχήμα 6.2: Μάσκες προσοχής σε εικόνα από το μοντέλο EfficientNet

στα στάδια Μέτρια και Απουσία πάθησης.

Όσον αφορά την ερμηνευσιμότητα, επικεντρωνόμαστε σε τέσσερις τεχνικές για να περιγράψουμε τη λειτουργία των μοντέλων και να εξηγήσουμε τη λογική πίσω από την ταξινόμηση των εικόνων. Συγκεκριμένα, εξετάζονται τέσσερις αλγόριθμοι βασισμένοι σε κλίσεις σε αυτή τη μελέτη, το GradCAM, το GradCAM++, οι Integrated Gradients και το SHAP με τη μέθοδο προσέγγισης αναμενόμενων κλίσεων. Κάθε μέθοδος εφαρμόζεται σε κάθε δίκτυο στην αρχική εικόνα, προκειμένου να παράγει έναν χάρτη προσοχής, υποδεικνύοντας τις περιοχές της εικόνας που συνέβαλαν τα περισσότερο στην πρόβλεψη του δικτύου. Χρησιμοποιήθηκαν επίπεδα διαφορετικού βάθους για την κατασκευή των χαρτών, ανάλογα με την αρχιτεκτονική των δικτύων. Ειδικότερα, τα GradCAM και GradCAM++ εφαρμόστηκαν στο 79ο, 15ο και 15ο στρώμα των DenseNet, Inception και EfficientNet αντίστοιχα. Η μέθοδος των Integrated Gradients είναι ανεξάρτητη από την επιλογή ενός συγκεκριμένου επιπέδου, καθώς οι κλίσεις υπολογίζονται ως προς τις ενδιάμεσες εικόνες, αντίθετα με τα GradCAM και GradCAM++, που χρησιμοποιούν τις εξόδους των επιλεγμένων επιπέδων. Τέλος, ο αλγόριθμος SHAP εφαρμόστηκε στην είσοδο του δικτύου.

Αρχιτεκτονική	No DR	Mild	Moderate	Severe	Proliferative	Ungradable	Συνολική Ακρίβεια
DenseNet	0.8202	0.0	0.6279	0.0	0.2145	0.8063	0.6635
InceptionV3	0.9904	0.0	0.4211	0.1267	0.5636	0.9682	0.7130
EfficientNetB0	0.9574	0.0	0.5610	0.014	0.4727	0.9682	0.7356

Πίνακας 6.4: Ακρίβειες ελέγχου για κάθε κλάση και η συνολική ακρίβεια για τις 3 αρχιτεκτονικές για το σύνολο δεδομένων DDR

Η εικόνα 6.2 δείχνει τις μάσκες προσοχής που παράγονται από τις παραπάνω τεχνικές μετά την εφαρμογή τους στο μοντέλο EfficientNet (εκπαιδευμένο στο IDRiD) για ένα δείγμα από το IDRiD, καθώς και τις επικαλύψεις τους στην αρχική εικόνα. Όσον αφορά τους αλγόριθμους GradCAM και GradCAM++, ο τελευταίος αναδεικνύει μια μεγαλύτερη περιοχή της εικόνας, όπως αναμενόταν, κάτι που δυσκολεύει την εξαγωγή χρήσιμων πληροφοριών. Για τον Integrated Gradients, τείνει να παράγει μικρότερες περιοχές προσοχής, αν και οι κύριες επισημειωμένες περιοχές είναι κοινές με αυτές που υποδεικνύονται από τους προαναφερθέντες αλγορίθμους σε περισσότερες περιπτώσεις. Οι μάσκες SHAP έχουν τιμές χαμηλής τάξης μεγέθους και, συνεπώς, δεν είναι εύκολα συγκρίσιμες με τον υπόλοιπο τρόπο προσέγγισης. Επιπλέον, από την εικόνα 6.2 προκύπτει ότι υπάρχουν ελάχιστες περιοχές προσοχής στις μάσκες SHAP.

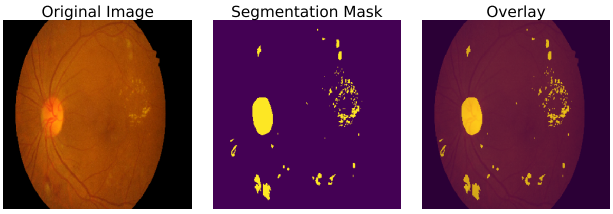
Όλες οι τεχνικές έχουν τα πλεονεκτήματά τους, ωστόσο, από ιατρική άποψη αποτυγχάνουν να εντοπίσουν τις σημαντικές περιοχές του ματιού που μπορεί να είναι χρήσιμες στη συγκεκριμένη διάγνωση της νόσου. Λόγω αυτού του αδύναμου σημείου, θα επικεντρωθούμε στον τρόπο μετασχηματισμού των μασκών της κάθε μεθόδου, προκειμένου να επιτευχθεί μια καλύτερη απόδοση στην ερμηνεία όσον αφορά το πρόβλημα της κατάτμησης των εικόνων.

6.2.4 Προσαρμοστικό Κατώφλι και Κατάτμηση

Όπως παρουσιάστηκε παραπάνω, είναι δύσκολο να αξιολογηθούν οι μάσκες προσοχής στην παρούσα μορφή τους, καθώς δεν είναι συγκρίσιμες μεταξύ τους. Αυτό οφείλεται κυρίως στο γεγονός ότι οι περιοχές που υποδεικνύονται ως κρίσιμες για την ταξινόμηση είναι σημαντικά διαφορετικές σε μέγεθος ανάλογα με τη μέθοδο. Επιπλέον, οι ίδιες τιμές σε διαφορετικές μάσκες προσοχής μπορεί να υποδηλώνουν διαφορετική ένταση Σημασίας. Για παράδειγμα, μια υψηλή τιμή σε μια μάσκα SHAP μπορεί να θεωρηθεί χαμηλή σε σύγκριση με τις τιμές μιας μάσκας GradCAM.

Αν και τα προβλήματα της ταξινόμησης της διαβητικής αμφιβληστροειδοπάθειας και κατάτμησης των βλαβών είναι διαφορετικά, υπάρχει επικάλυψη στην κατηγορία της σοβαρότητας της αμφιβληστροειδοπάθειας με τις βλάβες που βρίσκονται σε περιοχές του ματιού. Κάθε επικάλυψη μεταξύ αυτών των περιοχών των βλαβών και των μασκών προσοχής μπορεί να δείξει με ακρίβεια ποια τεχνική ερμηνευσιμότητας (και, συνακόλουθα, ποιο μοντέλο) μπορεί να ανιχνεύσει τα πιο σημαντικά μέρη του ματιού που μπορούν να κατηγοριοποιήσουν τη βαρύτητα

της αμφιβληστροειδοπάθειας. Προκειμένου να αξιολογηθούν οι μάσκες ανάθεσης, συνδυάσαμε τις μάσκες κατάτμησης για όλες τις πέντε βλάβες σε μια τελική ενοποιημένη μάσκα για κάθε εικόνα στο σύνολο δεδομένων. Αυτό φαίνεται στο Σχήμα 6.3, όπου παρουσιάζεται μια αρχική εικόνα μαζί με τη συνολική μάσκα κατάτμησης (περιλαμβάνοντας όλα τα είδη: MA, HE, EX, SE και OD) και την επικάλυψή τους.



Σχήμα 6.3: Ενοποιημένη κατάτμηση μάσκας εφαρμοσμένη πάνω στην αρχική εικόνα.

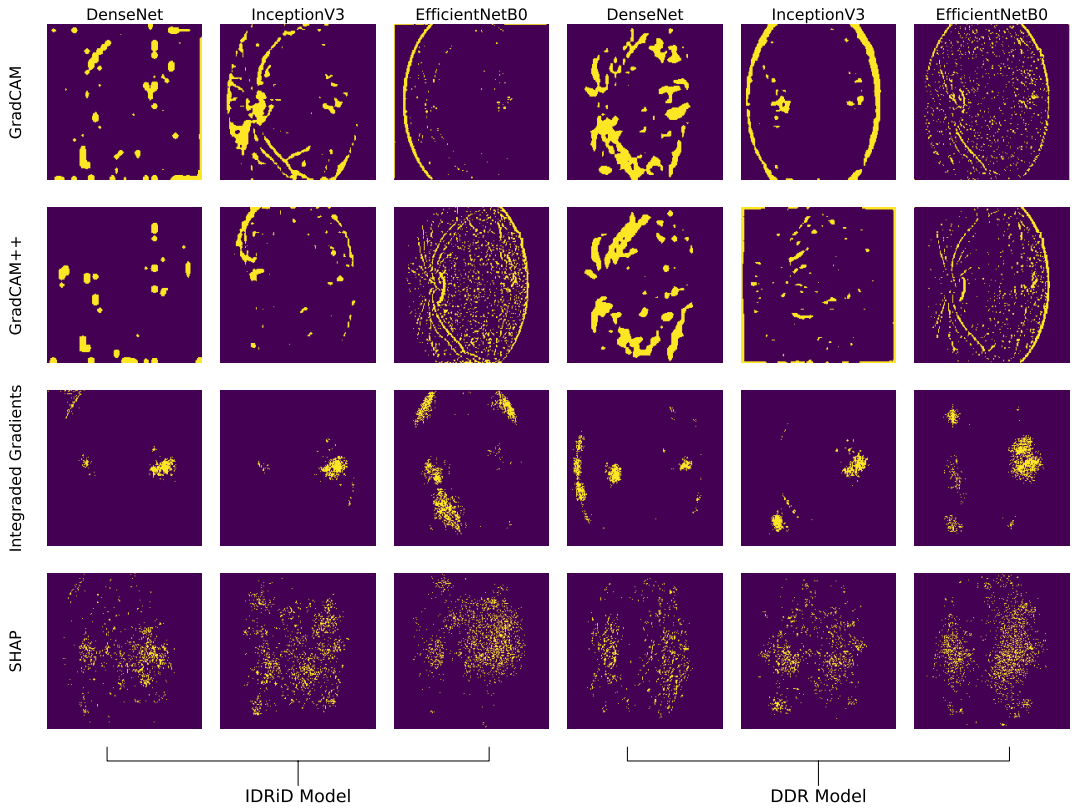
Για να ξεπεράσουμε τις παραπάνω δυσκολίες των διαφορετικών μασκών, προτείνουμε μία μέθοδο κατωφλιού πάνω στις μάσκες προσοχής, η οποία θα μετατρέψει τις τιμές του κάθε εικονοστοιχείου τους μέσα στα όρια $\{0, 1\}$. Αυτό απαιτείται επίσης για τη σύγκριση με τα πραγματικά δεδομένα κατάτμησης, καθώς οι τιμές των μασκών προσοχής είναι συνεχείς στο εύρος $[0, 1]$, ενώ οι μάσκες κατάτμησης αποτελούνται από διακριτές δυαδικές τιμές, όπως φαίνεται στα Σχήματα 6.2 και 6.3. Για αυτό προτείνουμε μια μέθοδο που βασίζεται στις στατιστικές ιδιότητες των εικόνων προκειμένου να παράγει ένα προσαρμοστικό κατώφλι, μεμονωμένα για κάθε μάσκα. Συγκεκριμένα, η τιμή του κατωφλίου καθορίζεται σύμφωνα με την Εξίσωση 6.2:

$$threshold = median + k * std \tag{6.2}$$

όπου η διάμεσος (median) και η τυπική απόκλιση (standard deviation) υπολογίζονται πάνω από τα εικονοστοιχεία της μάσκας. Η παράμετρος k προσαρμόζει το κατώφλι ανάλογα με τη διάμεσο και το ποσοστό των εικονοστοιχείων που αρχικά επισημαίνονται ως σημαντικά για την πρόβλεψη του μοντέλου (δηλαδή μη μηδενικά εικονοστοιχεία), σύμφωνα με τον παρακάτω τύπο:

$$k = [a + \ln(1 + pct)] * (1 - median) \tag{6.3}$$

Στην Εξίσωση 6.3, το pct αντιπροσωπεύει το ποσοστό των μη μηδενικών τιμών σε σχέση με τον συνολικό αριθμό των εικονοστοιχείων, ενώ το a είναι μια σταθερά. Η ιδέα πίσω από το pct είναι ότι η τιμή του κατωφλίου πρέπει να είναι μεγαλύτερη όσο αυξάνεται το ποσοστό των σημαντικών εικονοστοιχείων στη μάσκα, προκειμένου να ισορροπιστεί η ποσότητα των επιλεγμένων εικονοστοιχείων και να καθίστανται συγκρίσιμα μεταξύ τους. Η λογαριθμική συνάρτηση προορίζεται να αποτρέψει το κατώφλι από το να είναι υπερβολικά υψηλό για μεγάλες τιμές ποσοστού, καθώς αυτό θα οδηγούσε σε μάσκες χωρίς πληροφορία που αποτελούνται από πολύ λίγα περιγραφικά σημεία. Όσον αφορά τη συμβολή της διαμέσου, υψηλές τιμές υποδηλώνουν κατανομές με αριστερή στρέβλωση, πράγμα που σημαίνει ότι απαιτείται μικρότερη



Σχήμα 6.4: Οι μάσκες κατάτμησης με το προσαρμοστικό κατώφλι για διαφορετικές τεχνικές ερμηνευσιμότητας και αρχιτεκτονικές.

αύξηση για να αποφευχθεί η αποκοπή ενός πολύ μεγάλου ποσοστού σημείων. Αντίθετα, χαμηλές τιμές της διαμέσου απαιτούν μεγαλύτερη αύξηση ώστε να μην παράγεται μια υπερβολική μάσκα. Τέλος, η σταθερά που καθορίζει τον κάτω όριο της παραμέτρου τέθηκε στο 1.2 μετά από πειραματισμό.

Το Σχήμα 6.4 δείχνει τις μάσκες κατάτμησης που παράγονται από τις μεθόδους ερμηνευσιμότητας στα εκπαιδευμένα μοντέλα μετά την εφαρμογή του προσαρμοστικού κατωφλίου. Η αρχική εικόνα και η μάσκα κατάτμησης της είναι αυτές που απεικονίζονται στο Σχήμα 6.3. Είναι προφανές ότι οι νέες μάσκες είναι σε παρόμοια τάξη μεγέθους και μπορούν να συγκριθούν εύκολα, αντίθετα με τις αρχικές μάσκες, παρόμοιες με αυτές που παρουσιάζονται στο Σχήμα 6.2. Επιπλέον, φαίνεται ότι υπάρχουν πολλές ομοιότητες μεταξύ των μασκών που παράγονται από την ίδια μέθοδο ερμηνευσιμότητας σε διαφορετικές αρχιτεκτονικές. Μεταξύ των μεθόδων, οι GradCAM και GradCAM++ εμφανίζουν οπτικά παρόμοια μοτίβα ως προς τα σημεία με τη μεγαλύτερη επίδραση στην απόφαση των μοντέλων. Αντίθετα, οι μάσκες SHAP φαίνεται να είναι πιο διασκορπισμένες, αλλά φαίνονται αρκετά παρόμοιες με τις μάσκες IG στην πλειοψηφία των εικόνων.

Το γεγονός ότι όλες οι τέσσερις τεχνικές επικεντρώνονται σε αρκετά παρόμοια μέρη των εικόνων και μοιράζονται ορισμένες περιοχές ενδιαφέροντος με τη μάσκα κατάτμησης ενισχύει τις πιθανότητες εντοπισμού των πλέον σημαντικών περιοχών. Στη συνέχεια, υπολογίζουμε

Μέθοδοι Ερμηνευσιμότητας	Δεδομένα Εκπαίδευσης		Συνολικά
	IDRiD	DDR	
GradCAM	0.0563	0.0634	0.060
GradCAM++	0.0547	0.0560	0.055
Integrated Gradients	0.0544	0.0880	0.071
SHAP	0.0763	0.0985	0.087

Πίνακας 6.5: Μετρική Jaccard για τις μάσκες προσοχής

την μετρική Jaccard ή αλλιώς IoU (Intersection over Union) για τις τέσσερις μεθόδους ερμηνευσιμότητας ξεχωριστά, προκειμένου να αξιολογηθεί η απόδοση κάθε τεχνικής μετά την εφαρμογή του προτεινόμενου κατωφλίου. Σε αυτό το σημείο χρησιμοποιήθηκαν 81 εικόνες κατάτμησης από το σύνολο δεδομένων IDRiD. Οι μάσκες κατάτμησης των εικόνων του συνόλου δεδομένων DDR περιείχαν πολύ λίγα και αραιά σημεία και γι΄ αυτό δεν θεωρήθηκαν κατάλληλες για το σκοπό της μελέτης.

Ο Πίνακας 6.5 παρουσιάζει τη μετρική IoU για τις μάσκες που παράγονται από τα μοντέλα εκπαιδευμένα στα σύνολα δεδομένων IDRiD και DDR ξεχωριστά, καθώς και τα συνολικά IoU για κάθε μέθοδο ερμηνευσιμότητας για όλες τις αρχιτεκτονικές. Ανάμεσα στις εξεταζόμενες προσεγγίσεις, το SHAP κατέγραψε ξεκάθαρα υψηλότερο IoU και στις δύο ρυθμίσεις, περίπου 0.087. Όσον αφορά τις άλλες τεχνικές, η απόδοση των Integrated Gradients ήταν κοντά στο SHAP, με συνολικό IoU περίπου 0.071, ενώ οι GradCAM και GradCAM++ έφτασαν το IoU σε 0.06 και 0.055 αντίστοιχα. Φυσικά, οι παραπάνω επιδόσεις επηρεάζονται από τα κατώφλια που εφαρμόστηκαν στις αρχικές μάσκες. Διάφορες μέθοδοι κατωφλίωσης μπορεί να παράγουν τροποποιημένους χάρτες προσοχής, οδηγώντας σε διαφορετικά αποτελέσματα. Παρατηρείται ότι, αν και το σύνολο αξιολόγησης αποτελείται από εικόνες του συνόλου δεδομένων IDRiD, τα μοντέλα που εκπαιδεύτηκαν στο DDR φτάνουν σε υψηλότερα σκορ. Αυτό μπορεί να αποδοθεί στο γεγονός ότι το DDR είναι ένα μεγαλύτερο σύνολο δεδομένων, προσφέροντας πιο ολοκληρωμένα μοντέλα εκπαίδευσης. Διότι τα μοντέλα μας εκπαιδεύτηκαν στο πρόβλημα της ταξινόμησης, οι μετρικές IoU των μεθόδων ερμηνευσιμότητας είναι πολύ χαμηλότερα από οποιαδήποτε προσέγγιση τελευταίας τεχνολογίας στην κατάτμηση εικόνων. Έτσι, τα προηγούμενα αναφερθέντα σκορ δεν είναι συγκρίσιμα με μοντέλα που εκπαιδεύονται απευθείας στην εργασία της κατάτμησης.

6.3 Η Σημασία των Σημείων Αναφοράς

Στις προηγούμενες ενότητες αναφέραμε και περιγράψαμε διάφορες τεχνικές επεξηγησιμότητας καθώς και πως διαφέρουν μεταξύ τους. Οι τεχνικές αυτές βασίζονται σε διάφορες ιδέες, όπως οι διαφορές στις ενεργοποιήσεις, οι διαφορές στις εξόδους ή στις κλίσεις. Ωστόσο, η πλειονότητα αυτών των μεθόδων ερμηνευσιμότητας έχει κάτι κοινό, τη χρήση σημείων αναφοράς (baselines). Αυτά τα σημεία αναφοράς παρέχουν μία βάση από την οποία

μετρώνται οι συνεισφορές των χαρακτηριστικών. Έχοντας μια σταθερή αναφορά, γίνεται δυνατή η σύγκριση της εξόδου του δικτύου, των ενεργοποιήσεων ή των κλίσεων, οι οποίες μπορούν να μεταφραστούν σε σύγκριση της σημασίας των χαρακτηριστικών. Η σημασία των βάσεων γίνεται εμφανής κατά τη διαδικασία επιλογής μιας για ένα συγκεκριμένο πρόβλημα. Η επίδραση των βάσεων στο συνολικό αποτέλεσμα των εξηγήσεων είναι πολύ εμφανής και παίζουν σημαντικό ρόλο στην ακρίβεια, ποιότητα και αξιοπιστία των εξηγήσεων. Σε αυτή την εργασία, παρουσιάζουμε ένα καινοτόμο επίπεδο Διανυσματικής Αναπαράστασης (Embedding), το οποίο σχεδιάστηκε για να μάθει μια αναπαράσταση ενός σημείου αναφοράς μέσω της εκπαίδευσης. Αυτό θα αντιμετωπίσει τα υπάρχοντα προβλήματα με τις βάσεις και θα εξαλείψει την ανάγκη για επιλογή μιας σε δεδομένα πινάκων (tabular data). Συγκεκριμένα:

- Δείχνουμε ότι η ανάγκη για σημεία αναφοράς στην εξαγωγή των σημασιών κρύβει πολλά μειονεκτήματα και υπάρχουν κάποια εγγενή προβλήματα, κυρίως εμφανή σε δεδομένα πινάκων.

- Προτείνουμε ένα καινούργιο Επίπεδο Διανυσματικής Αναπαράστασης με Αντίληψη Βάσης που μπορεί να μάθει να αναπαριστά στον κρυφό χώρο ένα ουδέτερο σημείο που μπορεί να χρησιμοποιηθεί ως βάση χωρίς τα μειονεκτήματα.

- Δείχνουμε τη βελτιωμένη ακρίβεια και σταθερότητα των εξηγήσεων της προτεινόμενης μεθόδου σε 4 διαφορετικά σύνολα δεδομένων πινάκων χρησιμοποιώντας διάφορες τεχνικές ερμηνεύσιμοτητας που απαιτούν τη χρήση βάσεων.

6.3.1 Θεωρητικό Υπόβαθρο

Σε αυτήν την ενότητα παρουσιάζονται ορισμοί και βασικές πληροφορίες, που ακολουθούν τον συμβολισμό της εργασίας [112]. Υποθέτουμε ένα πλαίσιο επιβλεπόμενης μάθησης που αποτελείται από τα εξής: χώρος εισόδου $\mathcal{X} \in \mathbb{R}^d$, χώρος εξόδου $\mathcal{Y} \in \mathbb{R}$ και ένα μοντέλο μηχανικής μάθησης (στην περίπτωσή μας ένα νευρωνικό δίκτυο) $f : \mathbb{R}^d \to \mathbb{R}$. Έστω $x \in \mathcal{X}$ ένα δείγμα από τον χώρο εισόδου, τότε $f(x)$ είναι η πρόβλεψη που πρέπει να εξηγηθεί. Η συνάρτηση που αντιπροσωπεύει τις σημασίες χαρακτηριστικών για ένα συγκεκριμένο μοντέλο και είσοδο θα τη συμβολίζουμε ως $\Phi : f \times \mathbb{R}^d \to \mathbb{R}^d$. Για παράδειγμα, λαμβάνοντας υπόψη ένα μοντέλο f και ένα δείγμα x, τα $\Phi(f, x)$ είναι οι εξηγήσεις χαρακτηριστικών για τα d χαρακτηριστικά του x.

Επιλέξαμε να επικεντρωθούμε στις μεθόδους SHAP και Integrated Gradients στην ανάλυσή μας, επειδή ανήκουν στους πιο αναγνωρίσιμους τύπους μεθόδων επεξηγησιμότητας. Το SHAP είναι μια μέθοδος βασισμένη στην έξοδο, πράγμα που σημαίνει ότι υπολογίζει σημασίες με βάση τις αλλαγές στην έξοδο του μοντέλου. Από την άλλη, το Integrated Gradients ανήκει στην κατηγορία των μεθόδων που βασίζονται στη παράγωγο, καθώς χρησιμοποιεί πληροφορίες γραμμικής κλίσης για την εξήγηση των προβλέψεων του μοντέλου.

6.3.1.1 Τιμές Shapley

Οι τιμές Shapley [93] είναι μία έννοια από τη θεωρία συνεργατικών παιχνιδιών που παρέχει έναν τρόπο για τη δίκαιη κατανομή της συνολικής συμβολής μιας ομάδας παικτών σε ένα συνεργατικό παιχνίδι. Στο πλαίσιο της μηχανικής μάθησης και της επεξηγησιμότητας, οι τιμές Shapley χρησιμοποιούνται για να εξηγήσουν τις προβλέψεις ενός ¨μαύρου κουτιού' μοντέλου αναλύοντας τις συνεισφορές κάθε χαρακτηριστικού στην τελική πρόβλεψη. Παρέχουν έναν ποσοτικό μέτρο της σημασίας ή επιρροής κάθε χαρακτηριστικού στην έξοδο του μοντέλου.

Η ιδέα πίσω από αυτές τις τιμές είναι να υπολογίσουν τη μέση συνεισφορά ενός χαρακτηριστικού σε όλα τα δυνατά υποσύνολα των χαρακτηριστικών. Λαμβάνουν υπόψη την αξία της συμπερίληψης ενός χαρακτηριστικού σε μια πρόβλεψη συγκρίνοντας την πρόβλεψη με και χωρίς αυτό το χαρακτηριστικό, λαμβάνοντας υπόψη όλους τους δυνατούς συνδυασμούς με άλλα χαρακτηριστικά. Η παρακάτω εξίσωση υπολογίζει τις Shapley τιμές του προγνωστικού μοντέλου f με είσοδο το δείγμα x όσον αφορά το χαρακτηριστικό i:

$$\Phi_i(f,x) = \sum_{S \subseteq \mathcal{X} \setminus \{i\}} \frac{|S|!(|\mathcal{X}| - |S| - 1)!}{|\mathcal{X}|!} \left[f(x_{S'}) - f(x_S) \right], \tag{6.4}$$

με το S να είναι ένα υποσύνολο όλων των χαρακτηριστικών $\mathcal{X}$ χωρίς το χαρακτηριστικό i και $S' = S \cup i$. Το άθροισμα επαναλαμβάνεται σε κάθε δυνατό υποσύνολο S που δεν περιλαμβάνει το χαρακτηριστικό i. Το σύμβολο $|\cdot|$ αντιστοιχεί στην πληθικότητα ενός συνόλου, που είναι ισοδύναμο με τον αριθμό των χαρακτηριστικών στο συγκεκριμένο υποσύνολο. Με άλλα λόγια, η εξίσωση 6.4 υπολογίζει ένα σταθμισμένο άθροισμα πάνω από όλες τις δυνατές διαφορές στην έξοδο μεταξύ υποσυνόλων που περιέχουν το συγκεκριμένο χαρακτηριστικό και υποσυνόλων που δεν το περιέχουν.

Ωστόσο, ο υπολογισμός της παραπάνω εξίσωσης είναι συνήθως ακατόρθωτος. Ο αριθμός των επαναλήψεων του αθροίσματος είναι ίσος με τον αριθμό όλων των δυνατών υποσυνόλων $S \subset \mathcal{X}$, που είναι $2^{|\mathcal{X}|}$. Αυτός είναι ο λόγος για τον οποίο χρησιμοποιείται ο αλγόριθμος δειγματοληψίας Shapley values [63] ως ένας αλγόριθμος προσέγγισης, που δειγματοληπτεί ένα διαχειρίσιμο αριθμό υποσυνόλων. Όπως αναφέρεται στο [68], θα αναφερόμαστε στο Shapley values sampling ως SHAP.

Ο υπολογισμός από μόνος του, όμως, δεν είναι το μόνο ζήτημα. Για να αξιολογήσουμε με ακρίβεια την έξοδο του νευρωνικού δικτύου με είσοδο ένα υποσύνολο των $\mathcal{X}$, πρέπει να εκπαιδεύσουμε ένα παρόμοιο μοντέλο $f' : \mathcal{R}^{|S|} \to \mathcal{R}$ για να μάθει τις κρυφές αναπαραστάσεις από την αρχή χωρίς τη γνώση που παρέχουν τα αφαιρούμενα χαρακτηριστικά. Έτσι, χρειάζεται να εκπαιδεύσουμε $2^{|\mathcal{X}|}$ νευρωνικά δίκτυα, πράγμα που το καθιστά αδύνατο σε ένα πραγματικό σενάριο με σύνολα δεδομένων που έχουν εκατοντάδες ή χιλιάδες χαρακτηριστικά. Σε αυτό το σημείο εμφανίζονται οι αναφορές στα σημεία αναφοράς (baselines) [102]. Τα αφαιρούμενα χαρακτηριστικά αντικαθίστανται με συγκεκριμένες αναφορές (βάσεις). Η πιο συνηθισμένη πρακτική για τα baselines του SHAP είναι να θέσουμε τα απουσιάζοντα χαρακτηριστικά στο μηδέν. Άλλη επιλογή είναι να θέσουμε αυτές τις τιμές ίσες με το μέσον ή την διάμεσο τιμή

των χαρακτηριστικών του συνόλου εκπαίδευσης.

6.3.1.2 Integrated Gradients

Ο αλγόριθμος Integrated Gradients [103] είναι μια δημοφιλής τεχνική για την αποτίμηση της επίδρασης των χαρακτηριστικών στις προβλέψεις του μοντέλου. Η μέθοδος IG βασίζεται στις κλίσεις της εξόδου ως προς την είσοδο και πώς αυτές μπορούν να μετρήσουν τη σημασία των χαρακτηριστικών. Συγκεκριμένα, ο IG λαμβάνει υπόψη έναν γραμμικό μονοπάτι από ένα επιλεγμένο σημείο αναφοράς (σημείο στο χώρο) μέχρι την είσοδο μας x. Κατά μήκος αυτού του μονοπατιού, ο Integrated Gradients υπολογίζει τις κλίσεις όλων των σημείων και τις συσσωρεύει. Τυπικά, ο IG ορίζεται ως το ολοκληρούμενο μονοπάτι κατά μήκος της ευθείας γραμμής που αρχίζει από το σημείο αναφοράς x' και φτάνει στην είσοδο x. Αυτό μπορεί να γραφεί ως:

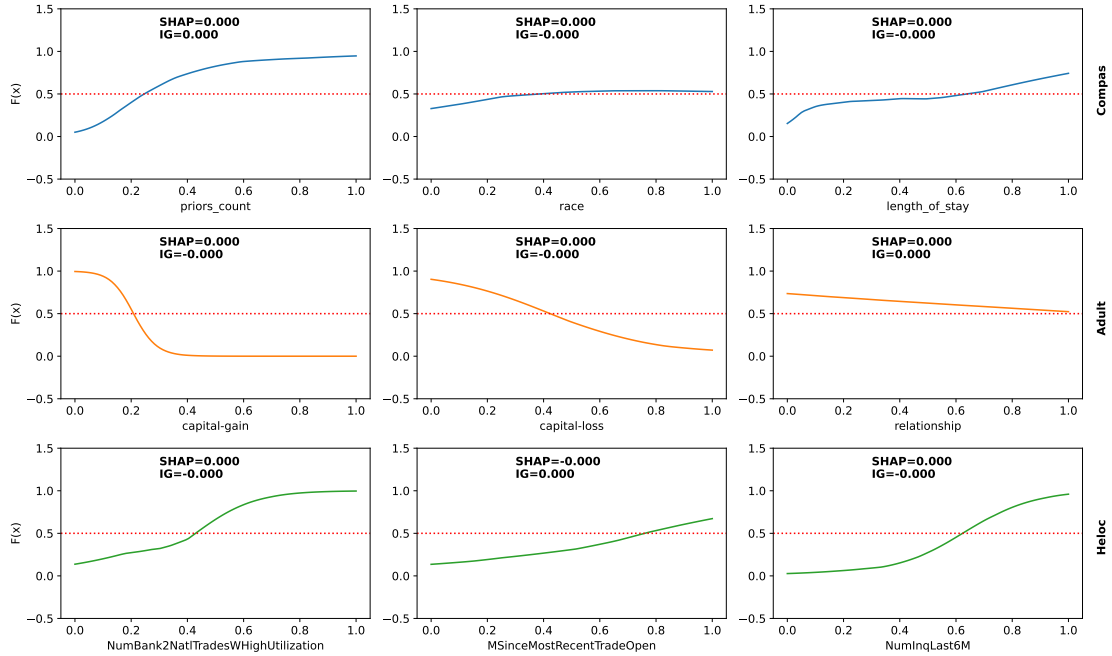
$$IG_i(f,x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{6.5}$$

Στην πράξη, το ολοκλήρωμα υπολογίζεται ως άθροισμα των κλίσεων των επιλεγμένων σημείων που βρίσκονται στην ευθεία γραμμή από το x' στο x. Μπορούμε να δούμε εδώ ξανά την κρίσιμη επίδραση του σημείου αναφοράς και πώς μπορεί να επηρεάσει κάθε δείγμα διαφορετικά, επιδρώντας διαφορετικά στον υπολογισμό των εξηγήσεων των χαρακτηριστικών.

6.3.2 Το πρόβλημα με τα Σημεία Αναφοράς

Η έννοια των σημείων αναφοράς αναφέρεται σε σημεία σύγκρισης που χρησιμοποιούνται για να αξιολογηθεί η σημασία ή η συμβολή των χαρακτηριστικών στην πρόβλεψη ενός μοντέλου. Τα σημεία αναφοράς παρέχουν μια βάση για την κατανόηση της σχετικής επίδρασης διαφόρων χαρακτηριστικών και βοηθούν στην αξιολόγηση της σημασίας των επιπτώσεών τους. Για το υπόλοιπο του κεφαλαίου επικεντρωνόμαστε σε δεδομένα πινάκων.

Η καθορισμός σημείων αναφοράς σε δεδομένα πινάκων εμπεριέχει μοναδικές προκλήσεις σε σύγκριση με άλλους τομείς όπως η όραση υπολογιστών. Σε δεδομένα πινάκων, όπου κάθε δείγμα χαρακτηρίζεται από πολλά χαρακτηριστικά, η επιλογή μιας κατάλληλης βάσης γίνεται μια πολύπλοκη εργασία. Δεν υπάρχει ένας απευθείας αντίστοιχος τρόπος για τον καθορισμό σημείων αναφοράς, όπως στα δεδομένα εικόνας, όπου οι βάσεις μπορούν να δημιουργηθούν χρησιμοποιώντας κενές ή θορυβώδεις εικόνες. Η πολυπλοκότητα προκύπτει από τον ποικίλο χώρο των χαρακτηριστικών και τις περίπλοκες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, καθιστώντας δύσκολο τον καθορισμό ενός ολικού σημείου αναφοράς που να αντικατοπτρίζει την εγγενή διακύμανση κάθε χαρακτηριστικού. Επιπλέον, στην όραση υπολογιστών, οι εικόνες συχνά ακολουθούν ένα σταθερό μορφολογικό πλαίσιο και μπορούν να προσαρμοστούν πιο εύκολα για τη δημιουργία σημείων αναφοράς. Αντίθετα, τα δεδομένα σε πίνακες ποικίλουν σε διαστατικότητα και κλίμακα, απαιτώντας μια πιο λεπτομερής προσέγγιση για την επιλογή της



Σχήμα 6.5: Διάγραμμα μερικής εξάρτησης σε διάφορα χαρακτηριστικά τριών συνόλων δεδομένων (Compas, Adult, Heloc) και οι SHAP/IG σημασίες με μηδενικά σημεία αναφοράς.

βάσης που να είναι προσαρμοσμένη στην εσωτερική κατανομή των δεδομένων.

Το βασικό πρόβλημα εντοπίζεται όταν η σταθερή τιμή που ορίζεται ως βάση είναι μέρος της κατανομής του χαρακτηριστικού που χρησιμοποιείται. Με άλλα λόγια, όταν η βάση είναι μια πραγματική τιμή που ορισμένα από τα δείγματα παρουσιάζουν, οι εξηγήσεις αυτών των χαρακτηριστικών δεν είναι ακριβείς και φαίνεται ότι υποεκτιμούν τη σημασία και τη συνεισφορά των χαρακτηριστικών. Αυτό συμβαίνει συχνά σε δεδομένα πινάκων λόγω της φύσης των κατηγορικών χαρακτηριστικών και της μικρής πληθικότητάς τους. Για παράδειγμα, εάν ένα δείγμα έχει την ίδια τιμή στο κατηγορικό χαρακτηριστικό με το χαρακτηριστικό βάσης, τότε το σκορ σημασίας που υπολογίζεται από το SHAP ή το IG θα είναι αρκετά χαμηλό (ακόμη και μηδέν), παρά τη συνολική συμβολή του χαρακτηριστικού. Εκτός από αυτό, υπάρχει ένα άλλο πρόβλημα που είναι παρόμοιο αλλά αφορά τα αριθμητικά χαρακτηριστικά. Ο υπολογισμός εξηγήσεων χρησιμοποιώντας βάσεις μοιράζεται κάποια χαρακτηριστικά με μετρικές απόστασης, όπου η εγγύτητα μιας βάσης σε ένα δείγμα μπορεί να έχει ως αποτέλεσμα κάποια χαρακτηριστικά να έχουν μειωμένη σημασία, οδηγώντας σε χαμηλότερα σκορ στις εξηγήσεις τους.

Για να δείξουμε αυτό το φαινόμενο, χρησιμοποιήσαμε ένα ευρέως χρησιμοποιούμενο εργαλείο οπτικοποίησης στην ερμηνευσιμότητα, το διάγραμμα μερικής εξάρτησης (Partial Dependence Plot - PDP) [30, 78]. Το PDP παρέχει σημαντικές πληροφορίες σχετικά με τη σχέση μεταξύ ενός συγκεκριμένου χαρακτηριστικού και των προβλέψεων του μοντέλου ενώ κρατούνται σταθερά όλα τα άλλα χαρακτηριστικά. Με την απομόνωση ενός μόνο χαρακτηριστικού, το PDP βοηθά να αποκαλυφθεί πώς οι αλλαγές σε αυτό το χαρακτηριστικό επηρεάζουν την έξοδο του μοντέλου. Για τη δημιουργία ενός PDP, το εξεταζόμενο χαρακτηριστικό μεταβάλ-

λεται σε ένα εύρος τιμών, και οι προβλέψεις του μοντέλου καταγράφονται σε κάθε σημείο. Συνήθως, οι τιμές των χαρακτηριστικών κυμαίνονται μεταξύ του ελάχιστου και του μέγιστου των τιμών που παρατηρούνται στο σύνολο δεδομένων. Το αποτέλεσμα του διαγράμματος απεικονίζει τη μέση επίδραση του χαρακτηριστικού στις προβλέψεις του μοντέλου, αποτυπώνοντας τόσο την κατεύθυνσή του όσο και το μέγεθός του.

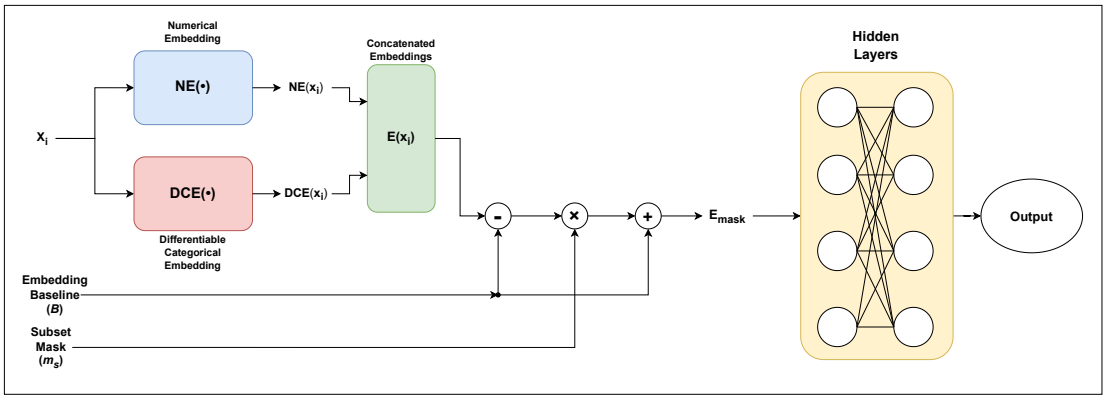
Στο Σχήμα 6.5, παρουσιάζουμε μια συλλογή από PDPs που δείχνουν τη συμπεριφορά διαφόρων χαρακτηριστικών σε 3 διαφορετικά σύνολα δεδομένων: Compas, Adult και Heloc (δείτε την Ενότητα 6.3.4.1 για περισσότερες λεπτομέρειες). Κάθε PDP απεικονίζει το αντίκτυπο ενός συγκεκριμένου χαρακτηριστικού στις προβλέψεις του μοντέλου κρατώντας όλα τα άλλα χαρακτηριστικά σταθερά. Επιπλέον, σε κάθε υπογράφημα εμφανίζονται οι τιμές των σημασιών (σχετικά με SHAP και IG) των επιλεγμένων χαρακτηριστικών και έχουν μηδενική ή σχεδόν μηδενική τιμή. Θα ήταν φυσιολογικό για χαρακτηριστικά με μηδενική απόδοση τα αντίστοιχα PDPs να δείχνουν ελάχιστη αλλαγή στην έξοδο του μοντέλου. Αντιθέτως, στις περισσότερες περιπτώσεις η έξοδος αλλάζει σημαντικά καθώς η τιμή του χαρακτηριστικού κυμαίνεται από το ελάχιστο στο μέγιστό του.

Οι μέθοδοι που περιγράφονται στην Ενότητα 6.3.1 χρησιμοποιούν κάποιου είδους βάσεις (baselines) με διαφορετικούς τρόπους, αλλά ακολουθώντας την ίδια ιδέα. Ο SHAP χρησιμοποιεί τις αναφορές ως έναν τρόπο για να συμπληρώσει την τιμή ενός ¨απουσιάζοντος' χαρακτηριστικού, ενώ το IG χρησιμοποιεί τις αναφορές ως αφετηρία για το μονοπάτι της ολοκλήρωσης.

Είναι προφανές ότι η επιλογή των σημείων αναφοράς έχει σημαντικά μειονεκτήματα που επηρεάζουν την ποιότητα των σημασιών. Αυτή η επιλογή εξαρτάται από το συγκεκριμένο πεδίο προβλήματος και τη μέθοδο ερμηνευσιμότητας που χρησιμοποιείται. Οι βάσεις μπορούν να επηρεάσουν την ερμηνεία και την κατανόηση ενός μοντέλου. Διαφορετικές βάσεις μπορούν να παρέχουν διαφορετικές απόψεις σχετικά με τη σημασία των χαρακτηριστικών. Οι βάσεις μπορεί να ποικίλουν ανάλογα με τον τύπο των χαρακτηριστικών (κατηγορικά ή αριθμητικά), τον χαρακτήρα του προβλήματος (ταξινόμηση ή παλινδρόμηση) ή ακόμη και τον επιθυμητό σημείο αναφοράς (π.χ., τυχαία πρόβλεψη ή μια μέση περίπτωση), κάτι που καθιστά δύσκολο και απαραίτητο να βρεθούν οι καλύτερες για την κάθε περίπτωση. Καθώς τα δεδομένα σε μορφή πίνακα περιέχουν συνήθως εξίσου αριθμητικά όσο και κατηγορικά χαρακτηριστικά, ενώ είναι ανεξάρτητα μεταξύ τους, η εστίαση της παρούσας εργασίας στρέφεται γύρω από αυτού του είδους τα σύνολα δεδομένων.

6.3.3 Επίπεδο Διανυσματικής Αναπαράστασης με Αντίληψη Βάσης

Σε αυτήν την ενότητα παρουσιάζουμε το επίπεδο Διανυσματικής Αναπαράστασης με Αντίληψη Βάσης (BAEM layer), ένα νέο επίπεδο νευρικών δικτύων που σχεδιάστηκε για να αντιμετωπίσει ένα ουσιαστικό ζήτημα που προκύπτει από τη χρήση βάσεων σε μεθόδους ερμηνευσιμότητας για δεδομένα πινάκων. Το επίπεδο BAEM στοχεύει στη βελτίωση της ακρίβειας των ερμηνειών, εκπαιδεύοντας ένα νευρικό δίκτυο έτσι ώστε να μπορεί να μάθει να πραγμα-



Σχήμα 6.6: Σχέδιο του προτεινόμενου επιπέδου διανυσματικής αναπαράστασης και πως μπορεί να χρησιμοποιηθεί σε οποιαδήποτε αρχιτεκτονική.

τοποιεί προβλέψεις σαν να λείπει ένα υποσύνολο των χαρακτηριστικών. Τα δεδομένα σε μορφή πίνακα συνήθως περιλαμβάνουν τόσα κατηγορικά όσο και αριθμητικά χαρακτηριστικά και, ως αποτέλεσμα, οι περισσότερες αρχιτεκτονικές που αντιμετωπίζουν αυτού του είδους τα σύνολα δεδομένων περιλαμβάνουν τόσα κατηγορικές όσο και αριθμητικές διανυσματικές αναπαραστάσεις (embeddings). Το προτεινόμενο επίπεδο χρησιμοποιεί αυτές τις αναπαραστάσεις και κατασκευάζει μια πρόσθετη που αντιπροσωπεύει ένα χαρακτηριστικό που λείπει. Αυτή η αναπαράσταση θα χρησιμοποιηθεί τόσο στο στάδιο της συλλογής των ερμηνειών όσο και στην εκπαίδευση, ώστε το μοντέλο να μπορεί να μάθει ότι αντιστοιχεί σε ένα χαρακτηριστικό που λείπει. Η έννοια των απουσιάζοντων χαρακτηριστικών μπορεί να χρησιμοποιηθεί για να αντιπροσωπεύσει το σημείο αναφοράς που θα χρησιμοποιηθεί από τους αλγόριθμους ερμηνευσιμότητας.

Για να λειτουργήσει κατάλληλα και για τα δύο είδη χαρακτηριστικών, η κατασκευασμένη αναπαράσταση πρέπει να είναι *σταθερή* για όλα τα απουσιάζοντα χαρακτηριστικά, *μοναδική* και *παραγωγίσιμη*. Η πρώτη ιδιότητα εξασφαλίζει ότι η αναπαράσταση παραμένει η ίδια κατά τη διάρκεια της εκπαίδευσης και της συλλογής ερμηνειών, διατηρώντας τη συνοχή στην διανυσματική απεικόνιση. Η δεύτερη εξασφαλίζει ότι καμία άλλη αναπαράσταση χαρακτηριστικού δεν θα είναι πανομοιότυπη με την αναπαράσταση του απουσιάζοντος χαρακτηριστικού. Η τελευταία ιδιότητα μας επιτρέπει να υπολογίσουμε αποτελεσματικά τις εξηγήσεις των μεθόδων που βασίζονται στην κλίση, όπως IG.

Για να επιτευχθεί αυτό, το προτεινόμενο επίπεδο χρησιμοποιεί μία αριθμητική διανυσματική αναπαράσταση με παράμετρο πόλωσης, μία παραγωγίσιμη διανυσματική αναπαράσταση κατηγορικών χαρακτηριστικών και μία επιπρόσθετη είσοδο που καθορίζει ποια χαρακτηριστικά θα χρησιμοποιηθούν σε μια πρόβλεψη. Αυτή η είσοδος ενσωματώνεται στη μορφή μιας '0-1' μάσκας και συμβολίζεται ως $m_S \in \mathbb{R}^d$ (Μάσκα Υποσυνόλου - Subset Mask). Αυτή η μάσκα είναι ένα δυαδικό διάνυσμα με 0 και 1, όπου κάθε στοιχείο αντιστοιχεί σε ένα χαρακτηριστικό στο σύνολο δεδομένων. Η τιμή 1 υποδηλώνει ότι το αντίστοιχο χαρακτηριστικό συμπεριλαμβάνεται στη διαδικασία πρόβλεψης, ενώ η τιμή 0 υποδηλώνει ότι το χαρακτηριστι-

κό αποκλείεται. Χρησιμοποιώντας αυτήν τη μάσκα, το μοντέλο μπορεί να ελέγχει δυναμικά το σύνολο χαρακτηριστικών που χρησιμοποιούνται κατά τη διαδικασία πρόβλεψης. Ωστόσο, η m_S δεν εφαρμόζεται απευθείας στην είσοδο του δικτύου, αλλά στην έξοδο του επιπέδου διανυσματικής αναπαράστασης κατηγορικών και αριθμητικών χαρακτηριστικών.

Πρώτα, θα ασχοληθούμε με τα κατηγορικά επίπεδα. Έστω d ο συνολικός αριθμός χαρακτηριστικών, e η διάσταση της αναπαράστασης και d_n ο αριθμός των αριθμητικών χαρακτηριστικών και d_c ο αριθμός των κατηγορικών χαρακτηριστικών. Έστω $NE : \mathbb{R}^{d_n} \rightarrow \mathbb{R}^{d_n} \times \mathbb{R}^e$ το αριθμητικό επίπεδο αναπαράστασης, που μπορεί να εκφραστεί με την ακόλουθη εξίσωση:

$$NE(x) = x \odot W_{R^{d_n} \times R^{d_e}} + b_e, \tag{6.6}$$

με το $W_{\mathbb{R}^{d_n} \times \mathbb{R}^{d_e}}$ να είναι το εκπαιδεύσιμο βάρος (που πολλαπλασιάζεται κατά στοιχείο [1] με την είσοδο) και b_e είναι το εκπαιδεύσιμο βάρος πόλωσης. Η χρήση της πόλωσης συμβάλλει στην ιδιότητα της μοναδικότητας σε ορισμένες ειδικές περιπτώσεις.

Έστω $DCE : \mathbb{R}^{d_c} \rightarrow \mathbb{R}^{d_c} \times \mathbb{R}^e$ το διαφορήσιμο επίπεδο αναπαράστασης κατηγορικών χαρακτηριστικών. Για να δημιουργήσουμε αυτό το επίπεδο, χρησιμοποιήσαμε τη συνάρτηση ενεργοποίησης LeakyReLU [71], αφού μετατρέψαμε κάθε τιμή κατηγορικού χαρακτηριστικού σε πραγματικό αριθμό από το μηδέν έως το μέγεθος του λεξιλογίου κάθε χαρακτηριστικού $[0, size(voc))$. Εάν το voc_i είναι ο πίνακας λεξιλογίου του i-οστού χαρακτηριστικού, τότε η κατηγορική αναπαράσταση για αυτό το χαρακτηριστικό μπορεί να υπολογιστεί με την ακόλουθη εξίσωση:

$$DCE(x_i) = LeakyReLU\left(1 - (voc_i - x_i)^2\right) \odot W_{R^{d_e}}, \tag{6.7}$$

με $W_{\mathbb{R}^{d_e}}$ να είναι το εκπαιδεύσιμο βάρος διανυσματικής αναπαράστασης. Ο σκοπός αυτού του επιπέδου είναι να υπολογίζει τις κλίσεις όσον αφορά την είσοδο, προκειμένου να χρησιμοποιεί μεθόδους επεξήγησης βασισμένες σε κλίσεις. Τέλος, αναθέτουμε το $\mathcal{B} \in \mathbb{R}^d$ ως τη μοναδική αναπαράσταση σημείου αναφοράς, η οποία είναι μια σταθερή υπερπαράμετρος. Οι παρακάτω εξισώσεις δείχνουν το προτεινόμενο επίπεδο ενσωμάτωσης που λαμβάνει υπόψη τα σημεία αναφοράς (baselines):

$$E(x_i) = concat(DCE(x_i^{cat}), NE(x_i^{num})) \tag{6.8}$$

$$E_{mask}(x_i) = E(x_i) * m_S + (1 - m_S) * \mathcal{B} \tag{6.9}$$

Οι παραπάνω εξισώσεις μπορούν να δείχνονται στο Σχήμα 6.6 σε μορφή διαγράμματος νευρωνικού δικτύου. Όπως φαίνεται στο σχήμα, το BAEMNet (Baseline-Aware Embedding Network) μπορεί να κατασκευαστεί ως ένα νευρωνικό δίκτυο με ένα επίπεδο BAEM

[1] το σύμβολο $\odot$ υποδηλώνει τον κατά στοιχείο πολλαπλασιασμό πινάκων

στην αρχή, ακολουθούμενο από μια συμβατική αρχιτεκτονική δεδομένων πινάκων. Αυτή η αρχιτεκτονική μπορεί να είναι οποιαδήποτε από τις δημοφιλείς και να περιλαμβάνει dense επίπεδα, μετασχηματιστές ή συνδυασμούς, προσαρμοσμένους στις συγκεκριμένες απαιτήσεις του εκάστοτε προβλήματος.

Το BAEMNet λειτουργεί ως μια συνάρτηση f που λαμβάνει δύο εισόδους: την είσοδο δεδομένων x και την είσοδο μάσκας υποσυνόλου m_S. Κατά τη διάρκεια κάθε επανάληψης της εκπαίδευσης, αναθέτουμε μια τιμή στη μάσκα υποσυνόλου από τη κατανομή Bernoulli. Ο στόχος είναι να βελτιστοποιήσουμε τις παραμέτρους του μοντέλου για όσο το δυνατόν περισσότερα διαφορετικά υποσύνολα χαρακτηριστικών. Με αυτόν τον τρόπο, το μοντέλο μαθαίνει να κάνει προβλέψεις με οποιοδήποτε συνδυασμό ¨απουσιάζοντων' χαρακτηριστικών. Η διαδικασία εκπαίδευσης μπορεί να περιληφθεί στον αλγόριθμο 8.

Algorithm 8: Training Algorithm for BAEMNet

Trainset: $D = \{(x_i, y_i)\}_{i=1}^{N}$, **Subset Mask**: m_S, **Probability of feature inclusion**: p, **Number of Epochs**: *Epochs*

Initialize weights θ

for $e = 1$ *to Epochs* **do**

  for $i = 1$ *to* N **do**

    Select Subset: $m_s \sim Bernoulli(p)$

    Forward pass: $y_{\text{pred}} = f(x_i, m_S)$

    Compute loss: $\mathcal{L} = loss(y_{\text{pred}}, y_i)$

    Update parameters: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}$

Υπολογισμός Σημασιών

Για να υπολογίσουμε τις σημασίες με το BAEMNet για το SHAP και το Integrated Gradients, πρέπει να αλλάξουμε λίγο τις αρχικές τους εξισώσεις (Εξίσωση 6.4-6.5) για να ταιριάζουν με το νέο μοντέλο. Για το SHAP, η σημασία για ένα συγκεκριμένο χαρακτηριστικό i υπολογίζεται ως εξής:

$$SHAP(i) = \sum_{S \subseteq \mathcal{X} \setminus \{i\}} \lambda \left[f(x_i, m_{S'}) - f(x_i, m_S) \right], \tag{6.10}$$

με $\lambda = \frac{|S|!(|\mathcal{X}|-|S|-1)!}{|\mathcal{X}|!}$, όπου S είναι ένα υποσύνολο όλων των χαρακτηριστικών $\mathcal{X}$ χωρίς το χαρακτηριστικό i και $S' = S \cup i$. Για το Integrated Gradients (IG), υπολογίζουμε την σημασία για το χαρακτηριστικό i όχι ως ένα ολοκλήρωμα κατά μήκος ενός μονοπατιού από ένα σημείο αναφοράς (baseline) ως την είσοδο, αλλά από μία μηδενική μάσκα m_S σε μια μοναδιαία μάσκα. Η μάσκα $m_S = 0$ συμβολίζει απουσία χαρακτηριστικών και το ολοκληρωμένο μονοπάτι αρχίζει από την πλήρη απουσία χαρακτηριστικών έως την πλήρη συμβολή των χαρακτηριστικών στην πρόβλεψη του συγκεκριμένου δείγματος. Η παρακάτω εξίσωση δείχνει την πρακτική προσέγγιση χρησιμοποιώντας το άθροισμα των κλίσεων ως προς m_S:

$$IG(i) = \frac{1}{K} \sum_{k=0}^{K} \frac{\partial f(x, m_S = k/K)}{\partial m_S}, \tag{6.11}$$

με K να είναι ο αριθμός των βημάτων του υπολογισμού της προσέγγισης Riemman [103].

6.3.4 Πειραματική Διαδικασία

Σε αυτή την ενότητα θα παρουσιαστεί το συνολικό πλαίσιο των πειραμάτων, καθώς και τα αποτελέσματα σχετικά με δημοφιλή σύνολα δεδομένων. Το προτεινόμενο επίπεδο μας εκτιμάται όσον αφορά την ακρίβεια της προσέγγισης των τιμών Shapley σε σύγκριση με τον αρχικό αλγόριθμο SHAP, καθώς και την ποιότητα των εξηγήσεων του IG.

6.3.4.1 Σύνολα Δεδομένων

Τα πειράματα πραγματοποιήθηκαν στα ακόλουθα σύνολα δεδομένων πινάκων:

- **Adult**: Το σύνολο δεδομένων Adult (γνωστό και ως Census Income) χρησιμοποιείται συχνά για προβλήματα δυαδικής ταξινόμησης, και συγκεκριμένα, για τον προσδιορισμό εάν το εισόδημα ενός ατόμου υπερβαίνει τα 50.000 δολάρια το χρόνο. Περιλαμβάνει συνολικά 48.842 παραδείγματα και 14 χαρακτηριστικά. Τα χαρακτηριστικά περιλαμβάνουν δημογραφικές πληροφορίες, όπως ηλικία, επίπεδο εκπαίδευσης, οικογενειακή κατάσταση, επάγγελμα και κέρδη από την πώληση περιουσιακών στοιχείων. Η δυαδική μεταβλητή-έξοδος υποδηλώνει εάν το άτομο έχει εισόδημα μεγαλύτερο από 50.000 δολάρια με '1' ή όχι με '0'.

- **Compas**: Το σύνολο δεδομένων Compas περιλαμβάνει πληροφορίες για άτομα στο σύστημα αστυνομικής δικαιοσύνης και χρησιμοποιείται για την πρόβλεψη της υποτροπής. Περιλαμβάνει 18.876 παραδείγματα με 7 χαρακτηριστικά. Τα χαρακτηριστικά περιλαμβάνουν διάφορες πληροφορίες για τα άτομα, όπως ηλικία, φυλή, φύλο και προηγούμενο ιστορικό ανάκρισης. Η δυαδική ετικέτα υποδηλώνει εάν το άτομο θα διαπράξει νέο έγκλημα, με '1' να υποδηλώνει υποτροπή και '0' την μη.

- **German**: Το German σύνολο δεδομένων ασχολείται με τον έλεγχο πιστοληπτικού κινδύνου και είναι γνωστό και ως (German credit dataset). Περιλαμβάνει 1.000 παραδείγματα με 20 χαρακτηριστικά. Τα χαρακτηριστικά περιλαμβάνουν πληροφορίες για τους υποψήφιους για πίστωση, όπως ηλικία, πιστοληπτική ιστορία, αποταμιεύσεις και επαγγελματική κατάσταση. Η επιθυμητή έξοδος υποδηλώνει τον κίνδυνο πιστοληπτικού, με το '1' να υποδηλώνει καλό πιστωτικό ρίσκο και το '0' κακό πιστωτικό ρίσκο.

- **Heloc**: Το σύνολο δεδομένων Heloc χρησιμοποιείται για την πρόβλεψη της πιθανότητας προβλήματος σε πιστωτικά όρια με βάση τα οικονομικά στοιχεία των δανειοληπτών. Περιλαμβάνει 9.871 παραδείγματα με 23 χαρακτηριστικά. Τα χαρακτηριστικά στο σύνολο

δεδομένων περιλαμβάνουν διάφορα οικονομικά στοιχεία των δανειοληπτών, όπως εισόδημα, αξιοποίηση πίστωσης και ρυθμός καθυστέρησης. Η δυαδική ετικέτα υποδηλώνει εάν ο δανειολήπτης πιθανόν να μην αντεπεξέλθει στην πίστωση με '1' ή '0' για κανένα πρόβλημα πίστωσης.

Για κάθε σύνολο δεδομένων, χρησιμοποιήσαμε την προεπεξεργασία που προτάθηκε στη μελέτη [2]. Αυτή η διαδικασία περιλαμβάνει τη διαχείριση απουσιάζουσων τιμών, τη μετατροπή των κατηγορικών χαρακτηριστικών σε αριθμητικές αναπαραστάσεις, το διαχωρισμό των συνόλων δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, καθώς και την εφαρμογή τεχνικών κανονικοποίησης για τα αριθμητικά χαρακτηριστικά προκειμένου να ετοιμάσουμε τα δεδομένα για εκπαίδευση.

Αυτά τα τέσσερα σύνολα δεδομένων καλύπτουν μια ποικίλη γκάμα εφαρμογών σε δεδομένα πινάκων, επιτρέποντάς μας να αξιολογήσουμε εκτενώς την απόδοση του προτεινόμενου επιπέδου. Αυτό υποστηρίζεται από το γεγονός ότι πρόσφατες εργασίες στον τομέα της επεξηγησιμότητας έχουν χρησιμοποιήσει αυτά τα σύνολα δεδομένων ως αξιολόγηση. Τα πειράματα διεξήχθησαν χρησιμοποιώντας την βιβλιοθήκη PyTorch[2] [79]. Επίσης, χρησιμοποιήθηκε η βιβλιοθήκη Captum [53] για τον υπολογισμό των σημασιών SHAP ή IG.

6.3.4.2 Αποτελέσματα Προσέγγισης των τιμών Shapley

Σε αυτή την υποενότητα, θα παρουσιάσουμε την πειραματική διαδικασία και τα αποτελέσματα της ακρίβειας της προσέγγισης των τιμών Shapley. Τα πειράματα διεξήχθησαν μεταξύ του BAEMNet και μιας παρόμοιας αρχιτεκτονικής με τον ίδιο αριθμό πυκνών επιπέδων, αλλά χωρίς το επίπεδο BAEM. Η αρχιτεκτονική του BAEMNet αποτελείται από τα προτεινόμενα επίπεδα διανυσματικής αναπαράστασης για αριθμητικά και κατηγορικά χαρακτηριστικά, ακολουθούμενα από τρία πλήρως συνδεδεμένα επίπεδα. Το απλό πυκνό δίκτυο αποτελείται από τα ίδια τρία πυκνά επίπεδα. Ο αριθμός των νευρώνων σε κάθε πυκνό επίπεδο και για τα δύο δίκτυα είναι ίσος με 100. Χρησιμοποιήσαμε τον βελτιστοποιητή Adam με ρυθμό μάθησης 0.001 για την εκπαίδευση τόσο του BAEMNet όσο και του απλού δικτύου. Κάθε μοντέλο εκπαιδεύτηκε για 100 εποχές. Όλα τα πειράματα διεξήχθησαν χρησιμοποιώντας τη βιβλιοθήκη βαθιάς μηχανικής μάθησης PyTorch.

Λόγω του ακριβούς υπολογιστικού κόστους των θεωρητικών τιμών Shapley, χρησιμοποιήσαμε ως μετρική προσέγγισης το Μέσο Τετραγωνικό Σφάλμα (MSE) των τυχαία επιλεγμένων διαφορών από τον τύπο του SHAP (Εξίσωση 6.4). Συγκεκριμένα, η θεωρητική τιμή υπολογίζεται εκπαιδεύοντας δύο δίκτυα (ίδια με το απλό πυκνό δίκτυο) χρησιμοποιώντας τα δύο υποσύνολα S και S'. Με αυτόν τον τρόπο, μπορούμε να υπολογίσουμε τη διαφορά $f(x_{S'}) - f(x_S)$ για τη θεωρητική τιμή και να τη συγκρίνουμε με το BAEMNet χρησιμοποιώντας την Εξίσωση 6.10 και για το απλό δίκτυο χρησιμοποιώντας τον αλγόριθμο δειγ-

[2]Ο κώδικας για τα πειράματα μπορεί να βρεθεί στο εναποθετήριο https://github.com/geoioannou/BAEMNet

ματοληψίας SHAP. Το MSE υπολογίζεται έχοντας τις θεωρητικές τιμές ως οι πραγματικές ετικέτες.

Πίνακας 6.6: Μέσο Τετραγωνικό Σφάλμα για τα 4 σύνολα δεδομένων

	Adult	**Compas**	**Heloc**	**German**
Sampling SHAP	0.013938	0.021181	0.011338	0.015214
BAEMNet	**0.008068**	**0.005857**	**0.007851**	**0.001072**

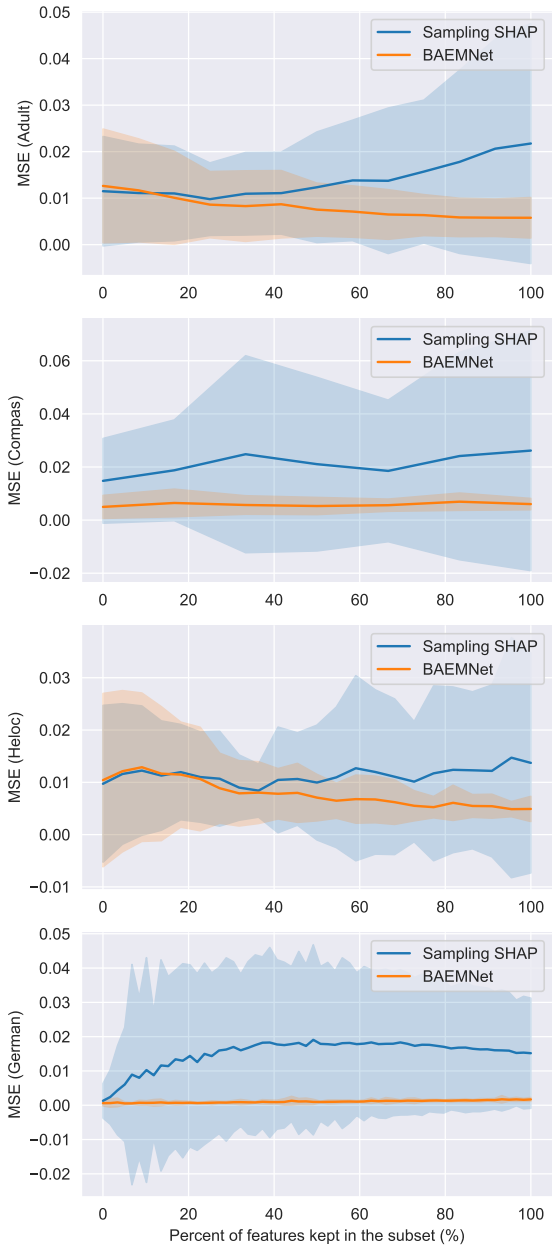
Ο πίνακας 6.6 παρουσιάζει το μέσο MSE σε όλες τις διαφορές που υπολογίστηκαν για όλα τα χαρακτηριστικά όσον αφορά τα τέσσερα σύνολα δεδομένων. Τα πειραματικά αποτελέσματα δείχνουν ότι το BAEMNet συνεχώς ξεπερνά το απλό δίκτυο όσον αφορά το MSE για τις τυχαίες διαφορές του τύπου SHAP. Οι χαμηλότερες τιμές MSE που προκύπτουν για το BAEMNet υποδεικνύουν τη δυνατότητά του να παρέχει πιο ακριβείς εξηγήσεις του αλγορίθμου SHAP. Η Εικόνα 6.7 δείχνει το μέσο σφάλμα προσέγγισης MSE σε όλα τα χαρακτηριστικά για τα τέσσερα σύνολα δεδομένων με βάση το μέγεθος του υποσυνόλου. Ο άξονας x κυμαίνεται από το 0% ως το 100% του αριθμού των χαρακτηριστικών που κρατήθηκαν στο υποσύνολο που χρησιμοποιήθηκε για τον υπολογισμό των διαφορών. Βλέπουμε ότι όταν το υποσύνολο είναι πιο κοντά στο πλήρες (100%), το BAEMNet παρέχει πιο ακριβείς προβλέψεις από τον αλγόριθμο δειγματοληψίας SHAP. Μια εξήγηση είναι ότι όσο μεγαλύτερο το υποσύνολο, τόσο περισσότερα τα πιθανά υποσύνολα με τον ίδιο αριθμό χαρακτηριστικών, όπου οι απλές προσεγγίσεις δεν μπορούν να είναι ακριβείς για το σύνολο αυτών.

6.3.4.3 Αποτελέσματα Σταθερότητας στις εξηγήσεις IG

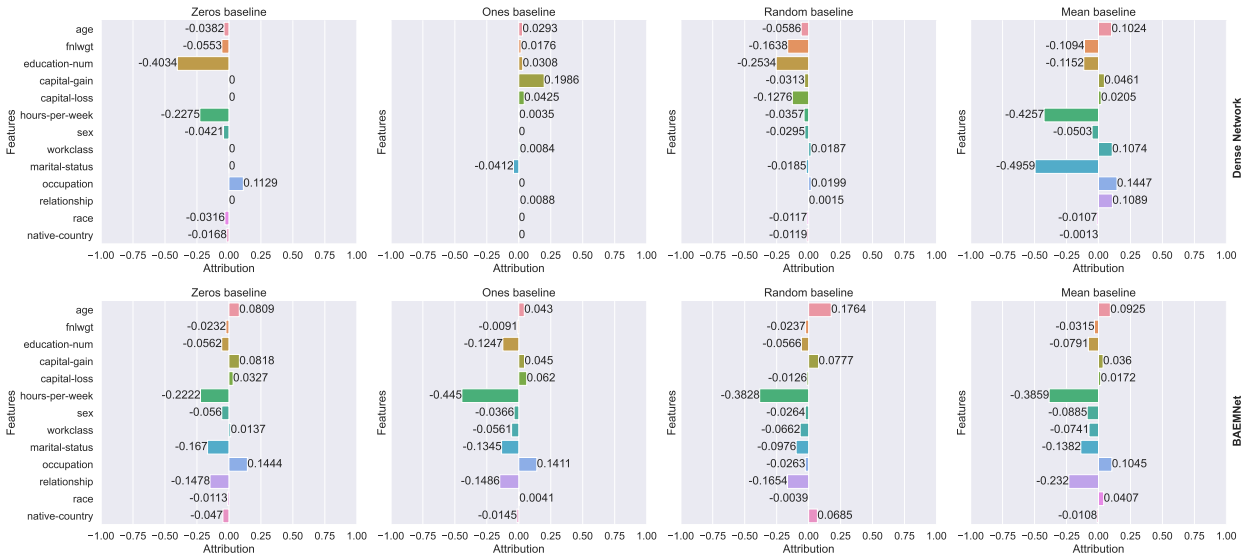
Σε αυτή την υποενότητα εξετάζουμε την σταθερότητα των εξηγήσεων του Integrated Gradients μεταξύ του απλού νευρωνικού δικτύου και του BAEMNet. Η χρήση παραδοσιακών σημείων αναφοράς με τον αλγόριθμο IG μπορεί να οδηγήσει σε σημασίες χαμηλής ποιότητας που μπορεί να αντικρούονται μεταξύ τους. Αυτό καθιστά δύσκολο για διάφορους χρήστες να συγκρίνουν και να αναλύουν τη σημασία των χαρακτηριστικών.

Για να το αποδείξουμε αυτό, εκπαιδεύσαμε ένα πυκνό νευρωνικό δίκτυο (όπως περιγράφεται στην Ενότητα 6.3.4.2) και υπολογίσαμε τις εξηγήσεις του IG χρησιμοποιώντας 4 διαφορετικά σημεία αναφοράς: ένα ΄μηδενικό΄ σημείο αναφοράς (ένα διάνυσμα με μηδενικά για όλα τα χαρακτηριστικά), ένα σημείο αναφοράς που περιέχει μόνο μονάδες, ένα τυχαίο σημείο αναφοράς (που ακολουθεί την κανονική κατανομή $\mathcal{N}(0, 2)$) και ένα μέσο σημείο αναφοράς (ένα διάνυσμα με τη μέση τιμή για κάθε χαρακτηριστικό).

Από την άλλη, εκπαιδεύσαμε το ίδιο δίκτυο BAEMNet με αυτά τα 4 διαφορετικά σημεία αναφοράς ως εκπαιδεύσιμο σημείο αναφοράς $\mathcal{B}$. Η εικόνα 6.8 δείχνει στην επάνω σειρά τις εξηγήσεις των χαρακτηριστικών του συνόλου δεδομένων Adult, ενώ στην κάτω σειρά βρίσκονται αυτές του BAEMNet. Βλέπουμε πολύ μικρές αλλαγές στις εξηγήσεις του BAEMNet μεταξύ διαφορετικών τιμών $\mathcal{B}$ (σημεία αναφοράς του επιπέδου διανυσματικής αναπαράστασης),



Σχήμα 6.7: Μέσο σφάλμα προσέγγισης σε όλα τα χαρακτηριστικά ανά ποσοστό των χαρακτηριστικών που κρατήθηκαν στο επεξεργαζόμενο υποσύνολο. Η σκιασμένη περιοχή αντιστοιχεί στην τυπική απόκλιση των MSE.



Σχήμα 6.8: Οι εξηγήσεις του IG για τα χαρακτηριστικά του συνόλου δεδομένων Adult με διάφορα σημεία αναφοράς. Τα παραδοσιακά σημεία αναφοράς ενός απλού δικτύου εμφανίζονται στην επάνω σειρά, και τα εκπαιδεύσιμα σημεία αναφοράς του BAEMNet βρίσκονται στην κάτω σειρά.

παρουσιάζοντας συνέπεια στις εξηγήσεις. Οι σημασίες του πυκνού δικτύου φαίνεται να διακυμαίνονται σημαντικά όταν αλλάζει το σημείο αναφοράς. Κάποια χαρακτηριστικά δείχνουν μηδενική σημασία με ένα σημείο αναφοράς και μεγάλη σημασία με ένα άλλο. Εάν οι εξηγήσεις εμφανίζουν υπερβολική μεταβλητότητα ή ευαισθησία στην επιλογή του σημείου αναφοράς, αυτό υποδεικνύει ότι η ερμηνεία του μοντέλου θα μπορούσε να αμφισβητηθεί, καθώς μικρές αλλαγές στο σημείο αναφοράς μπορεί να οδηγήσουν σε σημαντικές αλλαγές στην ερμηνεία. Το πλεονέκτημα του BAEMNet είναι εμφανές στην ικανότητά του να παράγει συνεπείς και ποιοτικές εξηγήσεις για τις προβλέψεις του μοντέλου που μπορούν να συγκριθούν και να συζητηθούν από διάφορους χρήστες.

Κεφάλαιο 7

Σύνοψη

Η Μηχανική Μάθηση αποτελεί έναν σύγχρονο τομέα της έρευνας και παρέχει λύσεις και εφαρμογές σε πολλά διαφορετικά επιστημονικά πεδία. Η Μηχανική Μάθηση διαπρέπει σε προβλήματα που βασίζονται σε δεδομένα και οι ντετερμινιστικοί αλγόριθμοι δεν είναι ικανοί να παράγουν σταθερές λύσεις. Έτσι, κατάφερε να εδραιωθεί σε πολλές εφαρμογές ως η καλύτερη προσέγγιση. Τέτοιες εφαρμογές είναι η Όραση Υπολογιστών, η Επεξεργασία και Αναγνώριση Φυσικής Γλώσσας, η πρόβλεψη και κατηγοριοποίηση χρονοσειρών και πολλές άλλες. Σε αυτό βοήθησε και ο υποκλάδος της Βαθιάς Μηχανικής Μάθησης, ο οποίος με την βοήθεια ανεπτυγμένου τεχνολογικού υλικού μπορεί να εκπαιδεύσει μοντέλα μεγάλων διαστάσεων χρησιμοποιώντας τεράστιες ποσότητες δεδομένων. Αυτό έχει ως αποτέλεσμα να βελτιωθούν οι επιδόσεις πολλών προαναφερθέντων εφαρμογών.

Η διατριβή αυτή επικεντρώθηκε στον τομέα της Βαθιάς Μηχανικής Μάθησης και συγκεκριμένα στα Νευρωνικά Δίκτυα. Ασχοληθήκαμε με την ανάπτυξη αλγορίθμων μάθησης και την κατασκευή μεθόδων, που αποσκοπούν στην καλύτερη εκπαίδευση και λειτουργία των νευρωνικών δικτύων. Πιο συγκεκριμένα, στο Κεφάλαιο 2 παρατίθεται ένα απαραίτητο θεωρητικό υπόβαθρο που έχει να κάνει με βασικές έννοιες και αλγορίθμους γύρω από τα νευρωνικά. Πολλές από αυτές τις έννοιες θα χρησιμοποιηθούν και θα αναφερθούν στην υπόλοιπη διατριβή.

Στο πρώτο μέρος της διατριβής ασχοληθήκαμε με μεθόδους που αφορούν την εκπαίδευση των νευρωνικών με παρτίδες. Οι αλγόριθμοι αυτοί καθορίζουν ποια δείγματα και με ποια σειρά θα εισαχθούν στην εκπαίδευση ενός νευρωνικού. Παραθέσαμε διάφορες τεχνικές Επιλογής Παρτίδας και πως αυτές επηρεάζουν την διαδικασία της εκπαίδευσης. Έπειτα, αναπτύξαμε έναν καινούργιο αλγόριθμο Επιλογής Παρτίδας με Μεροληπτική Δειγματοληψία (Batch Selection with Biased Sampling - BSBS). Ο αλγόριθμος επιλέγει ποια δείγματα είναι τα πιο δύσκολα για το δίκτυο και τα εισάγει περισσότερες φορές στην εκπαίδευση. Αυτό γίνεται με βάση την τιμή σφάλματος που έχει επιδείξει το νευρωνικό για το κάθε δείγμα ξεχωριστά. Με αυτόν τον τρόπο ο αλγόριθμος καταφέρνει να επιταχύνει την μάθηση και να αυξήσει την ταχύτητα σύγκλισης. Εκτός αυτού, εφαρμόσαμε μία προσεγγιστική τεχνική υπολογισμού του σφάλματος για να μην έχει επιπλέον υπολογιστικό κόστος η μέθοδός μας. Τα παραπάνω συνοδεύονται από πειράματα

σε 4 σύνολα δεδομένων, όπου η προτεινόμενη μέθοδος βελτιώνει τις επιδόσεις των νευρωνικών δικτύων που χρησιμοποιήθηκαν. Το Κεφάλαιο 3 είναι αφιερωμένο σε αυτή τη θεματική.

Στο Κεφάλαιο 4 ασχοληθήκαμε με τον τομέα της ανισορροπίας δεδομένων. Το φαινόμενο αυτό εμφανίζεται, κυρίως, σε πραγματικά σύνολα δεδομένων που παρουσιάζουν κάποια έλλειψη δειγμάτων ή έλλειψη ποικιλίας αυτών. Αποτελεί ένα σημαντικό ζήτημα για την αποδοτική λειτουργία των νευρωνικών και έχει απασχολήσει την έρευνα για αρκετά χρόνια. Δύο μεγάλες κατηγορίες τεχνικών υπάρχουν για την καταπολέμηση της ανισορροπίας: τεχνικές μετασχηματισμού δεδομένων και τεχνικές προσαρμογής του αλγορίθμου μάθησης. Η διατριβή μας επικεντρώθηκε στην δεύτερη κατηγορία τεχνικών. Αναπτύξαμε μία μέθοδο που χειρίζεται την εκπαίδευση με παρτίδες με τέτοιο τρόπο ώστε το νευρωνικό να δίνει την απαραίτητη προσοχή στα δείγματα που κάνει τα περισσότερα λάθη. Ονομάσαμε τον αλγόριθμο Θορυβώδης Επιλογή Παρτίδας με Επανεισαγωγές, ο οποίος έχει ως σκοπό την βελτίωση της γενίκευσης του μοντέλου στα πλαίσια ανισόρροπων δεδομένων. Παίρνει σαν βάση τον αλγόριθμο BSBS και δειγματοληπτεί τα δεδομένα με βάση την τιμή του σφάλματός τους. Όμως, λαμβάνει υπόψη του και άλλες παραμέτρους, όπως τον αριθμό των φορών που κάποιο δείγμα έχει βγει εκτός εκπαίδευσης. Εκτός αυτού εισάγει επιλεκτικά στοχαστικό θόρυβο στα δεδομένων αυξάνοντας τεχνητά το σύνολο δεδομένων και καθιστώντας την εκπαίδευση πιο σταθερή. Η πειραματική διαδικασία που στηρίζει τα παραπάνω αποτελείται από 4 σύνολα ανισόρροπων δεδομένων, τα οποία εκπαιδεύονται με 11 διαφορετικές τεχνικές μετασχηματισμού δεδομένων. Αυτό επιλέχθηκε για να συγκρίνουμε όλες τις τεχνικές αυτές μεταξύ τους και με την μέθοδό μας, αλλά και να δείξουμε την συμπεριφορά του νευρωνικού στην περίπτωση συνδυασμού τους. Παρατηρήσαμε ότι ο αλγόριθμός μας βελτιώνει τις επιδόσεις σε πολλές από αυτές τις περιπτώσεις, καθιστώντας τον συνδυασμό τους την καλύτερη λύση.

Στο επόμενο κομμάτι της διατριβής εμβαθύναμε στον τομέα της Στοχαστικής Βελτιστοποίησης. Τα νευρωνικά για να ανανεώσουν τα βάρη τους χρησιμοποιούν διάφορους βελτιστοποιητές, οι οποίοι συνήθως είναι παράγωγα του αλγορίθμου Καθόδου Κλίσης. Τα τελευταία χρόνια έχουν καθιερωθεί αλγόριθμοι βελτιστοποίησης με προσαρμοστικό ρυθμό μάθησης. Στην δική μας μελέτη κατασκευάσαμε έναν προσαρμοστικό αλγόριθμο που μεταβάλλει τον ρυθμό μάθησης ανά επίπεδο. Η μέθοδος βασίστηκε πάνω στην σταθερά του Lipschitz και στην θεωρητική βέλτιστη τιμή του ρυθμού μάθησης. Αναπτύξαμε μία προσέγγιση της σταθεράς και κατασκευάσαμε έναν νέο βελτιστοποιητή που χρησιμοποιεί τον παραπάνω ρυθμό μάθησης. Ο νέος αλγόριθμος ονομάζεται AdaLip και ο συνδυασμός του με άλλους βελτιστοποιητές δημιουργεί καινούργιους. Στην παρούσα διατριβή παρουσιάσαμε με δύο ακόμα βελτιστοποιητές, τον AdamLip και τον RMSLip, οι οποίοι αποτελούν τον συνδυασμό του AdaLip με τους Adam και RMSProp, αντίστοιχα. Ο προτεινόμενος αλγόριθμος εξετάστηκε από την πλευρά της ταχύτητας σύγκλισης και την συνολική επίδοση του μοντέλου. Πρώτα αποδείχθηκε θεωρητικά η σύγκλιση του αλγορίθμου. Έπειτα, ακολούθησε η πειραματική διαδικασία σε 3 σύνολα δεδομένων εικόνας. Εξετάστηκαν διάφορες αρχιτεκτονικές, διαφορετικές αρχικοποιήσεις και πολλοί αρχικοί ρυθμοί μάθησης. Σκοπός μας είναι να δείξουμε πειραματικά την καλή επίδοση του βελτιστοποιητή αλλά και την μικρότερη ευαισθησία του στην επιλογή

του αρχικού ρυθμού μάθησης. Τα αποτελέσματα ήταν θετικά και παρατηρήσαμε πολλές βελτιώσεις στην ταχύτητα σύγκλισης αλλά και στη γενίκευση των δικτύων. Αυτή η θεματική βρίσκεται στο Κεφάλαιο 5 της παρούσας διατριβής.

Στο τελευταίο Κεφάλαιο της διατριβής ασχοληθήκαμε με το πεδίο της Ερμηνευσιμότητας των Νευρωνικών Δικτύων, το οποίο πραγματεύεται με την κατανόηση του πώς λειτουργούν τα νευρωνικά δίκτυα και πώς κάνουν τις προβλέψεις τους. Στην αρχή, εξερευνήσαμε διάφορες τεχνικές ερμηνευσιμότητας και συγκρίναμε τις επιδόσεις τους. Τα πειράματα βασίστηκαν σε ιατρικές εικόνες με σκοπό την ταξινόμηση του σταδίου της αμφιβληστροειδοπάθειας. Αυτό μας βοήθησε να κατανοήσουμε βαθύτερα πώς λειτουργούν τα μοντέλα, αλλά και να εξηγήσουμε περιοχές βλάβης σε ιατρικές εικόνες. Επιπλέον, δείξαμε ότι με τη χρήση τέτοιων μεθόδων είναι δυνατό να προσεγγίσουμε μια λύση στο πρόβλημα της κατάτμησης εικόνας. Δείξαμε ότι οι περιοχές προσοχής που βρίσκουμε από τους αλγορίθμους ερμηνευσιμότητας είναι αρκετά κοντινές με τις ετικέτες του προβλήματος κατάτμησης. Κατόπιν, εστιάσαμε περισσότερο στη λειτουργία των μεθόδων ερμηνευσιμότητας που χρησιμοποιούν σημεία αναφοράς. Διαπιστώσαμε ότι η χρήση τέτοιων σημείων μπορεί να οδηγήσει σε ανακρίβειες στις ερμηνείες των προβλέψεων των νευρωνικών δικτύων. Στη συνέχεια, προτείναμε ένα νέο επίπεδο, το Επίπεδο Διανυσματικής Αναπαράστασης με Αντίληψη Βάσης, που αποσκοπεί στη βελτίωση αυτών των θεμάτων. Το προτεινόμενο επίπεδο δημιουργήθηκε με σκοπό να ενσωματώσει την έννοια της βάσης ή του σημείου αναφοράς κατά τη διάρκεια της εκπαίδευσης. Έτσι, οι αλγόριθμοι ερμηνευσιμότητας που λειτουργούν με σημεία αναφοράς μπορούν να χρησιμοποιήσουν αυτό το επίπεδο στην αρχιτεκτονική του δικτύου για πιο ακριβείς ερμηνείες σε διάφορες προβλέψεις. Αυτό επαληθεύτηκε πειραματικά σε 4 διαφορετικά σύνολα δεδομένων πινάκων, τα οποία επιλέχθηκαν λόγω της μεγάλης ποικιλίας των χαρακτηριστικών τους και της συχνής εμφάνισης του προβλήματος των σημείων αναφοράς.

Παράρτημα Α΄

Αρχιτεκτονικές Νευρωνικών Δικτύων

Πίνακας Α΄.1: Αρχιτεκτονική του πρώτου Πλήρως Συνδεδεμένου Νευρωνικού

Layer	Units	Activation
Dense	100	ReLU [76]
Dense	30	ReLU
Dense	2	Softmax

Πίνακας Α΄.2: Αρχιτεκτονική του δεύτερου Πλήρως Συνδεδεμένου Νευρωνικού

Layer	Units	Activation
Dense	100	ReLU
Dense	30	ReLU
Dropout (0.5)	-	-
Dense	2	Softmax

Πίνακας Α΄.3: Αρχιτεκτονική του Συνελικτικού Νευρωνικού Δικτύου

Layer	Units	Activation
Conv2D	32 (3×3)	ReLU
Conv2D	64 (3×3)	ReLU
MaxPooling	-	-
Flatten	-	-
Dense	128	ReLU
Dropout (0.5)	-	-
Dense	10	Softmax

Παράρτημα Β′

Β′.1 Αποδείξεις

Λήμμα 5.1

Απόδειξη. Έστω $w_{t+1} = w_t - \alpha\nabla f(w_t)$ είναι ο κανόνας ανανέωσης. Από την L-Lipschitz συνέχεια έχουμε τα ακόλουθα:

$$\nabla^2 f(w) \preceq LI$$

που σημαίνει ότι οι ιδιοτιμές του Εσσιανού πίνακα είναι φραγμένες στο L. Αυτό μπορεί να γραφεί ως:

$$u^T\nabla^2 f(v)u \leq u^T LIu \tag{Β′.1}$$

Από τις σειρές Taylor της f, συνεπάγεται:

$$\begin{aligned} f(w_{t+1}) = f(w_t) + \nabla f(w_t)^T(w_{t+1} - w_t) + \\ + \frac{1}{2}(w_{t+1} - w_t)^T\nabla^2 f(w_t)(w_{t+1} - w_t) \end{aligned} \tag{Β′.2}$$

Χρησιμοποιώντας την Εξίσωση (Β′.1) και τον κανόνα ανανέωσης στο δεύτερο μέρος της Εξίσωσης (Β′.2):

$$f(w_{t+1}) \leq f(w_t) + \nabla f(w_t)^T(w_{t+1} - w_t) + \frac{1}{2}L\|w_{t+1} - w_t\|^2$$

$$= f(w_t) + \nabla f(w_t)^T(-\alpha\nabla f(w_t)) + \frac{1}{2}L\| - \alpha\nabla f(w_t)\|^2$$

$$= f(w_t) - \alpha\nabla f(w_t)^T\nabla f(w_t) + \frac{1}{2}L\alpha^2\|\nabla f(w_t)\|^2$$

$$= f(w_t) - (\alpha - \frac{\alpha^2 L}{2})\|\nabla f(w_t)\|^2$$

Για να σιγουρέψουμε την μείωση της συνάρτησης f στην διάρκεια των επαναλήψεων, τότε το παρακάτω πρέπει να αληθεύει:

$$(\alpha - \frac{\alpha^2 L}{2}) > 0 => \alpha < \frac{2}{L}$$

Για να μεγιστοποιήσουμε την μείωση της συνάρτησης f ανά επανάληψη, η παράγωγος της f ως προς το α πρέπει να είναι μηδέν:

$$\frac{\partial(\alpha - \frac{\alpha^2 L}{2})}{\partial \alpha} = 0 => \alpha = \frac{1}{L}$$

□

Λήμμα 5.2

Απόδειξη. Από την εξίσωση 5.9, το S_t μπορεί να γραφεί ως:

$$S_t = \gamma^t S_0 + (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{||g_{i-1}||}{||g_i - g_{i-1}|| + c_i}$$

Για το άνω όριο παίρνουμε:

$$S_t \leq S_0 + (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{G}{\min_t(c_t)}$$

$$\leq 1 + \frac{(1-\gamma)G}{\min_t(c_t)} \sum_{i=1}^{t} \gamma^{t-i}$$

Το άθροισμα στα δεξιά είναι το άθροισμα μιας γεωμετρικής σειράς και επειδή $\gamma \neq 1$, τότε μπορεί να μετασχηματιστεί στο:

$$\leq 1 + \frac{(1-\gamma)G}{\min_t(c_t)} \cdot \frac{1-\gamma^t}{1-\gamma}$$

$$\leq 1 + \frac{(1-\gamma^t)G}{\min_t(c_t)} \leq 1 + \frac{G}{\min_t(c_t)}$$

Για κάτω όριο έχουμε:

$$S_t \geq (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{\|g_{i-1}\|}{\|g_i - g_{i-1}\| + c_i}$$

$$\geq (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{M}{2G + \max_t(c_t)}$$

$$\geq (1-\gamma)\frac{M}{2G+\max_t(c_t)}\sum_{i=1}^{t}\gamma^{t-i}$$

Υπολογίζοντας την γεωμετρική πρόοδο παίρνουμε:

$$S_t \geq \frac{M}{2G+\max_t(c_t)}(1-\gamma^t) \geq \frac{M(1-\gamma)}{2G+\max_t(c_t)}$$

το οποίο ολοκληρώνει την απόδειξη για το άνω και κάτω όριο του S_t.

□

Θεώρημα 5.1

Απόδειξη. Από την στοχαστική μέθοδο $w_{t+1} = \Pi_D(w_t - A_t g_t)$, έχουμε:

$$||w_{t+1} - w^*||^2 \leq ||w_t - w^* - A_t g_t||^2$$

$$= ||w_t - w^*||^2 - 2A_t\langle g_t, w_t - w^*\rangle + A_t^2||g_t||^2$$

Από το Λήμμα 2.3 παίρνουμε:

$$\leq ||w_t - w^*||^2 - 2A_t(f_t(w_t) - f_t(w^*)) + A_t^2||g_t||^2$$

Μετασχηματίζοντας το παραπάνω:

$$2A_t(f_t(w_t) - f_t(w^*)) \leq ||w_t - w^*||^2 - ||w_{t+1} - w^*|| + A_t^2||g_t||^2$$

$$f_t(w_t) - f_t(w^*) \leq \frac{1}{2A_t}||w_t - w^*||^2 - \\ - \frac{1}{2A_t}||w_{t+1} - w^*|| + \frac{A_t}{2}||g_t||^2$$

Αθροίζοντας από τ=1,2,..,T ώστε να κατασκευάσουμε τον όρο της μετάνοιας, έχουμε:

$$R(T) = \sum_{t=1}^{T}(f_t(w_t) - f_t(w^*)) \leq$$

$$\leq \sum_{t=1}^{T}\frac{1}{2A_t}||w_t - w^*||^2 - \sum_{t=1}^{T}\frac{1}{2A_t}||w_{t+1} - w^*|| + \\ + \sum_{t=1}^{T}\frac{A_t}{2}||g_t||^2$$

Ξεδιπλώνοντας το άθροισμα:

$$\leq \left(\frac{1}{2A_1}||w_1 - w^*||^2 + \frac{1}{2A_2}||w_2 - w^*||^2 + ... \right.$$
$$\left. + \frac{1}{2A_T}||w_T - w^*||^2\right) -$$

$$-\left(\frac{1}{2A_1}||w_2 - w^*||^2 + \frac{1}{2A_2}||w_3 - w^*||^2 + ... + \frac{1}{2A_T}||w_{T+1} - w^*||^2\right)$$
$$+\sum_{t=1}^{T}\frac{A_t}{2}||g_t||^2$$

$$= \frac{1}{2A_1}||w_1 - w^*||^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2A_{t+1}} - \frac{1}{2A_t}\right)||w_{t+1} - w^*||^2 -$$
$$-\frac{1}{2A_T}||w_{T+1} - w^*||^2 + \frac{1}{2}\sum_{t=1}^{T}A_t||g_t||^2$$

$$\leq \frac{1}{2A_1}||w_1 - w^*||^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2A_{t+1}} - \frac{1}{2A_t}\right)||w_{t+1} - w^*||^2 +$$
$$+\frac{1}{2}\sum_{t=1}^{T}A_t||g_t||^2$$

Επειδή οι παράγωγοι είναι φραγμένες $\|g_t\| \leq G$ και χρησιμοποιώντας το Λήμμα 2.4 παίρνουμε:

$$\leq \frac{4r^2}{2A_1} + \frac{1}{2}\sum_{t=1}^{T}A_tG^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2A_{t+1}} - \frac{1}{2A_t}\right)||w_{t+1} - w^*||^2$$

$$= \frac{2r^2}{A_1} + \frac{G^2}{2}\sum_{t=1}^{T}A_t + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{A_{t+1}} - \frac{1}{A_t}\right)||w_{t+1} - w^*||^2$$

Από το Λήμμα 2.5 αποδεικνύουμε ότι με συγκεκριμένο c_t η διαφορά είναι θετική, οπότε μπορεί να χρησιμοποιηθεί το Λήμμα 2.4 ώστε να πάρουμε:

$$\leq \frac{2r^2}{A_1} + \frac{G^2}{2}\sum_{t=1}^{T}A_t + \frac{4r^2}{2}\sum_{t=1}^{T-1}\left(\frac{1}{A_{t+1}} - \frac{1}{A_t}\right)$$

Το άθροισμα γίνεται τηλεσκοπικό:

$$= \frac{2r^2}{A_1} + \frac{G^2}{2}\sum_{t=1}^{T}A_t + 2r^2\left(\frac{1}{A_T} - \frac{1}{A_1}\right)$$

$$= \frac{2r^2}{A_T} + \frac{G^2}{2}\sum_{t=1}^{T} A_t$$

$$= \frac{2r^2\sqrt{T}}{\alpha_0 S_T} + \frac{G^2\alpha_0}{2}\sum_{t=1}^{T}\frac{S_t}{\sqrt{t}}$$

Η συμπεριφορά του A_t επηρεάζει την σύγκλιση του αλγορίθμου. Ο αριστερός όρος δείχνει ότι αν το A_t μειώνεται πολύ γρήγορα (της τάξης $\mathcal{O}(1/t^2)$) ο αλγόριθμος αποκλίνει. Από την άλλη, αν το A_t είναι σταθερό ή αυξάνεται, τότε ο δεξιός όρος (το άθροισμα) θα αποκίνει. Το Λήμμα 2.5 δείχνει ότι το c_t ελέγχει πόσο δραματικά το A_t μειώνεται. Υπάρχουν πολλές διαφορετικές τιμές c_t που οδηγούν σε σύγκλιση με διαφορετικές ταχύτητες. Εδώ, παρουσιάζουμε μία που ικανοποιεί την παραπάνω εξίσωση:

$$c_t = \max\left\{\frac{(1-\gamma)\|g_{t-1}\|}{\left(\sqrt{\frac{t}{t-1}}-\gamma\right)S_{t-1}} - \|g_t - g_{t-1}\|, \quad c_{t-1}\right\} \tag{Β΄.3}$$

Χρησιμοποιώντας τα όρια S_t από το Λήμμα 5.2 έχουμε:

$$R(T) \leq \frac{2r^2\sqrt{T}}{\alpha_0\frac{M(1-\gamma)}{2G+\max_T(c_T)}} + \frac{G^2\alpha_0}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\left(1+\frac{G}{\min_T(c_T)}\right)$$

Από την Εξίσωση Β΄.3 του c_t είναι εμφανές ότι το c_t είναι μη φθίνον, επομένως, η παραπάνω ανισότητα μετασχηματίζεται σε:

$$R(T) \leq \frac{2r^2(2G+c_T)\sqrt{T}}{\alpha_0 M(1-\gamma)} + \frac{G^2\alpha_0}{2}\left(1+\frac{G}{c_1}\right)\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$$

$$R(T) \leq \frac{2r^2(2G+c_T)\sqrt{T}}{\alpha_0 M(1-\gamma)} + \frac{G^2\alpha_0}{2}\left(1+\frac{G}{c_1}\right)\int_0^T\frac{1}{\sqrt{t}}dt$$

$$R(T) \leq \frac{2r^2(2G+c_T)\sqrt{T}}{\alpha_0 M(1-\gamma)} + G^2\alpha_0\left(1+\frac{G}{c_1}\right)\sqrt{T}$$

Διαιρώντας με T:

$$\frac{R(T)}{T} \leq \frac{2r^2(2G+c_T)}{\alpha_0 M(1-\gamma)\sqrt{T}} + \frac{G^2\alpha_0}{\sqrt{T}}\left(1+\frac{G}{c_1}\right)$$

το οποίο δείχνει ότι:

$$\frac{R(T)}{T} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \quad \text{ανδ} \quad \lim_{T\to\infty}\frac{R(T)}{T} = 0$$

□

Λήμμα 2.3. *Μία συνάρτηση $f : \Re^d \longrightarrow \Re$ είναι κυρτή, τότε για όλα τα $x, y \in \Re^d$,*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Λήμμα 2.4. *Δεδομένου ενός βάρους $w_i \in \Re_d$ και $\|w_i\| \leq r$, τότε $\|w_n - w_m\|^2 \leq 4r^2$, $\forall n, m$.*

Απόδειξη. Από την τριγωνική ανισότητα έχουμε:

$$\|w_n - w_m\|^2 = \|w_n^2 - 2w_n w_m + w_m^2\|$$

$$\leq \|w_n\|^2 + 2\|w_n\|\|w_m\| + \|w_m\|^2$$

$$\leq r^2 + 2rr + r^2 = 4r^2$$

□

Λήμμα 2.5. *Έστω A_t είναι ο ρυθμός μάθησης από την Εξίσωση 5.10, $\alpha_t = \alpha_0/\sqrt{t}$. Τότε $A_t \leq A_{t-1}$ είναι αληθές, αν c_t είναι μία θετική συνάρτηση του t και ικανοποιεί την ακόλουθη ανισότητα:*

$$c_t \geq \frac{(1-\gamma)\|g_{t-1}\|}{\left(\sqrt{\frac{t}{t-1}} - \gamma\right) S_{t-1}} - \|g_t - g_{t-1}\|$$

Απόδειξη.

$$A_t \leq A_{t-1}$$

$$\alpha_t \cdot S_t \leq \alpha_{t-1} \cdot S_{t-1}$$

$$\frac{\alpha_0}{\sqrt{t}} \left(\gamma S_{t-1} + (1-\gamma) \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t} \right) \leq \frac{\alpha_0}{\sqrt{t-1}} S_{t-1}$$

$$\gamma S_{t-1} + (1-\gamma) \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t} \leq \sqrt{\frac{t}{t-1}} S_{t-1}$$

$$(1-\gamma) \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\| + c_t} \leq \left(\sqrt{\frac{t}{t-1}} - \gamma \right) S_{t-1}$$

$$\|g_t - g_{t-1}\| + c_t \geq \frac{(1-\gamma)\|g_{t-1}\|}{\left(\sqrt{\frac{t}{t-1}} - \gamma\right) S_{t-1}}$$

$$c_t \geq \frac{(1-\gamma)\|g_{t-1}\|}{\left(\sqrt{\frac{t}{t-1}}-\gamma\right) S_{t-1}} - \|g_t - g_{t-1}\|$$

□

Λήμμα 2.6. *Δεδομένου μιας κυρτής συνάρτησης κόστους f, με L-Lipschitz συνεχή παράγωγο (Ορ. 5.3), ο βέλτιστος ρυθμός μάθησης για τη Στοχαστική Κάθοδο Κλίσης είναι:*

$$\alpha^* = \frac{\|\nabla f(w_t)\|^2}{L \cdot E\left[\|\nabla f(w_t)\|^2\right]}$$

Απόδειξη. Ο κανόνας ανανέωσης των βαρών του SGD είναι ο ακόλουθος:

$$w_{t+1} = w_t - \alpha \nabla f_t(w_t)$$

όπου f_t είναι η μερική f υπολογισμένη από μία παρτίδα δεδομένων στην επανάληψη t. Ξεκινώντας με την ίδια εξίσωση από το Λήμμα 5.1 έχουμε:

$$f(w_{t+1}) \leq f(w_t) + \nabla f(w_t)^T (w_{t+1} - w_t) + \frac{1}{2} L \|w_{t+1} - w_t\|^2$$

$$f(w_{t+1}) \leq f(w_t) + \nabla f(w_t)^T (-\alpha \nabla f_t(w_t)) + \frac{1}{2} L \| - \alpha \nabla f_t(w_t)\|^2$$

Παίρνοντας την αναμενόμενη τιμή και από τις δύο πλευρές:

$$E[f(w_{t+1})] \leq E\left[f(w_t)\right] - \alpha E\left[\nabla f(w_t)^T \nabla f_t(w_t)\right] + \\ + E\left[\frac{\alpha^2 L}{2} \|\nabla f_t(w_t)\|^2\right]$$

$$E[f(w_{t+1})] \leq f(w_t) - \alpha \nabla f(w_t)^T E\left[\nabla f_t(w_t)\right] + \frac{\alpha^2 L}{2} E\left[\|\nabla f_t(w_t)\|^2\right]$$

Χρησιμοποιώντας το $E[\nabla f_t(w_t)] = \nabla f(w_t)$:

$$E[f(w_{t+1})] \leq f(w_t) - \alpha \|\nabla f(w_t)\|^2 + \frac{\alpha^2 L}{2} E\left[\|\nabla f_t(w_t)\|^2\right]$$

Παίρνοντας την παράγωγο ίση με μηδέν όμοια με το Λήμμα 5.1:

$$\frac{\partial\left(-\alpha \|\nabla f(w_t)\|^2 + \frac{\alpha^2 L}{2} E\left[\|\nabla f_t(w_t)\|^2\right]\right)}{\partial \alpha} = 0$$

$$-\|\nabla f(w_t)\|^2 + \alpha L E\left[\|\nabla f_t(w_t)\|^2\right] = 0$$

$$\alpha = \frac{\|\nabla f(w_t)\|^2}{L \cdot E\left[\|\nabla f_t(w_t)\|^2\right]}$$

□

Β΄.2 Αρχιτεκτονικές Νευρωνικών Δικτύων

Οι λεπτομερείς αρχιτεκτονικές για τα δίκτυα παρουσιάζονται στους παρακάτω πίνακες.

Πίνακας Β΄.1: Αρχιτεκτονική για το μοντέλο του MNIST.

Layer	Units	Padding	Activation
3×3 Conv	32	Valid	Relu
3×3 Conv	64	Valid	Relu
2×2 MaxPool	-	-	-
Flatten	-	-	-
Dense	128	-	Relu
Dense	10	-	Softmax

Πίνακας Β΄.2: Αρχιτεκτονική για το μοντέλο του CIFAR10.

Layer	Units	Padding	Activation
3×3 Conv	32	Valid	Relu
3×3 Conv	32	Valid	Relu
2×2 MaxPool	-	-	-
3×3 Conv	64	Valid	Relu
3×3 Conv	64	Valid	Relu
2×2 MaxPool	-	-	-
3×3 Conv	128	Same	Relu
3×3 Conv	128	Same	Relu
2×2 MaxPool	-	-	-
Flatten	-	-	-
Dense	200	-	Relu
Dense	10	-	Softmax

Πίνακας Β΄.3: Αρχιτεκτονική για το μοντέλο του CIFAR100.

Layer	Units	Padding	Activation
3 × 3 Conv	64	Same	Relu
BatchNorm	-	-	-
2 × 2 MaxPool	-	-	-
3 × 3 Conv	128	Same	Relu
BatchNorm	-	-	-
2 × 2 MaxPool	-	-	-
3 × 3 Conv	256	Same	Relu
BatchNorm	-	-	-
3 × 3 Conv	256	Same	Relu
BatchNorm	-	-	-
2 × 2 MaxPool	-	-	-
3 × 3 Conv	512	Same	Relu
BatchNorm	-	-	-
3 × 3 Conv	512	Same	Relu
BatchNorm	-	-	-
2 × 2 MaxPool	-	-	-
3 × 3 Conv	512	Same	Relu
BatchNorm	-	-	-
3 × 3 Conv	512	Same	Relu
BatchNorm	-	-	-
2 × 2 MaxPool	-	-	-
Flatten	-	-	-
Dense	512	-	Relu
BatchNorm	-	-	-
Dropout(0.5)	-	-	-
Dense	100	-	Softmax

Β΄.3 Επέκταση: AdaLip-U

Εδώ θα παρουσιάσουμε μία παραλλαγή του AdaLip, που ονομάζεται AdaLip-U. Αυτή η παραλλαγή χρησιμοποιεί το γεγονός ότι το μέτρο των παραγώγων των επιπέδων είναι ανάλογο με αυτό των βαρών στην Κάθοδο Κλίσης. Συνεχίζοντας από την Εξίσωση 5.7 και υποθέτοντας ότι $\alpha_{t-1} \approx \alpha_t$, έχουμε:

$$\begin{aligned}\alpha^* &\leq \alpha \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\|} = \alpha \frac{\|g_{t-1}\|}{\|g_t - g_{t-1}\|} \cdot \frac{\alpha_{t-1}}{\alpha_{t-1}} \\ &= \alpha \frac{\|\alpha_{t-1} \cdot g_{t-1}\|}{\|\alpha_{t-1} \cdot g_t - \alpha_{t-1} \cdot g_{t-1}\|} \\ &\approx \frac{\|\alpha_{t-1} \cdot g_{t-1}\|}{\|\alpha_t \cdot g_t - \alpha_{t-1} \cdot g_{t-1}\|}\end{aligned}$$

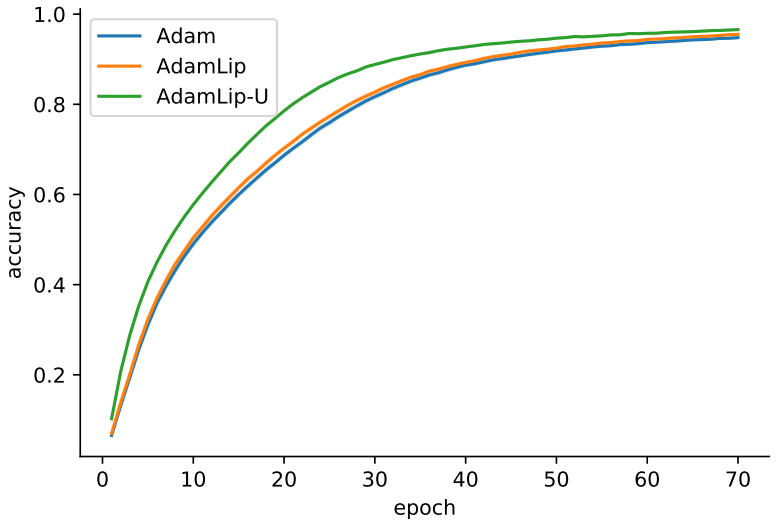
Αυτή η παραλλαγή βασίζεται στην προηγούμενη και στην τωρινή ανανέωση του βελτιστοποιητή, παρά στις παραγώγους. Θέτοντας $u_t = -a_t \cdot g_t$ την ανανέωση του βελτιστοποιητή, έχουμε:

$$\alpha^* \leq \alpha \frac{\|u_{t-1}\|}{\|u_t - u_{t-1}\|} \tag{Β΄.4}$$

Χρησιμοποιώντας αυτής της παρτίδας τον βέλτιστο ρυθμό μάθσης, αντί για την Εξίσωση 5.7, ο ρυθμός μάθησης του AdaLip-U μπορεί να κατασκευαστεί ως:

$$A_t = \alpha_t \cdot S_t = \alpha_t \cdot \left[\gamma^t \cdot S_0 + (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \frac{\|u_{i-1}\|}{\|u_i - u_{i-1}\| + c_i}\right] \tag{Β΄.5}$$

Η προηγούμενη ανανέωση αποθηκεύεται από την προηγούμενη επανάληψη. Η τωρινή ανανέωση είναι η ανανέωση που ο βελτιστοποιητής θα έκανε χωρίς τον AdaLip-U (δηλαδή η ανανέωση τουSGD με τον αρχικό ρυθμό μάθησης). Όπως προηγουμένως με τον AdaLip, αυτή η νέα παραλλαγή γεννά τρεις καινούργιους βελτιστοποιητές βασισμένους στον SGD, RMSProp και Adam, που ονομάζονται AdaLip-U, RMSLip-U και AdamLip-U αντίστοιχα. Στο Σχήμα Β΄.1 βλέπουμε την επίδοση του AdamLip-U και πως σε μερικές περιπτώσεις μπορεί να έχει ταχύτερη σύγκλιση από άλλες μεθόδους. Παρόλα αυτά στις περισσότερες περιπτώσεις η εκπαίδευσή του ήταν πιο ασταθής και αρκετά πιο ευαίσθητη στην επιλογή του αρχικού ρυθμού μάθησης. Στον Αλγόριθμο 9 παρουσιάζεται ο AdaLip-U λεπτομερώς.



Σχήμα Β΄.1: Ο βελτιστοποιητή AdamLip-U σε αντίθεση με τον Adam και τον AdamLip στο σύνολο δεδομένων CIFAR100.

Algorithm 9: AdaLip-U

Input: $w_t \in \mathcal{F}$, initial step α_0, $\gamma \in (0,1)$, c_t

Initialize: $S_0 = 1$, $g_0 = 0$, $u_0 = 0$

for *t=1* ***to*** T **do**

$g_t = \nabla f_t(w_t)$

$\alpha_t = \alpha_0/\sqrt{t}$

$\hat{u}_t = -\alpha_t \cdot g_t$

$S_t = \gamma \cdot S_{t-1} + (1-\gamma)\dfrac{\|u_{t-1}\|}{\|\hat{u}_t - u_{t-1}\| + c_t}$

$A_t = \alpha_t \cdot S_t$

$u_t = -A_t \cdot g_t$

$\hat{w}_{t+1} = w_t + u_t$

end

Βιβλιογραφία

[1] Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, ΣΜ̃-6(11):769–772, 1976.

[2] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik και Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. 2022.

[3] Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron C. Courville και Yoshua Bengio. Variance reduction in SGD by distributed importance sampling. *CoRR*, αβς/1511.06481, 2015.

[4] Gustavo E. A. P. A. Batista, Ana L. C. Bazzan και Maria Carolina Monard. Balancing training data for automated annotation of keywords: a case study. Στο *II Brazilian Workshop on Bioinformatics, December 3-5, 2003, Macaé, RJ, Brazil*Sérgio Lifschitz, Nalvo F. Almeida Jr., Georgios Joannis Pappas Jr. και Ricardo Linden, επιμελητές, σελίδες 10–18, 2003.

[5] Gustavo E. A. P. A. Batista, Ronaldo C. Prati και Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor.*, 6(1):20–29, 2004.

[6] Atilim Gunes Baydin, Robert Cornish, David Martínez-Rubio, Mark Schmidt και Frank Wood. Online learning rate adaptation with hypergradient descent. Στο *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[7] Barry Becker και Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert και Jason Weston. Curriculum learning. Στο *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, σελίδες 41–48, New York, NY, USA, 2009. ACM.

[9] Léon Bottou, Frank E Curtis και Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[10] Guillaume Bouchard, Théo Trouillon, Julien Perez και Adrien Gaidon. Accelerating stochastic gradient descent via online learning to sample. *CoRR*, αβς/1506.09016, 2015.

[11] Mateusz Buda, Atsuto Maki και Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[12] Diogo V. Carvalho, Eduardo M. Pereira και Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019.

[13] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis και Prudhvi Gurram. Interpretability of deep learning models: A survey of results. Στο *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, σελίδες 1–6, 2017.

[14] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader και Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[15] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall και W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002.

[16] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall και Kevin W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. Στο *Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings*Nada Lavrac, Dragan Gamberger, Hendrik Blockeel και Ljupco Todorovski, επιμελητές, τόμος 2838 στο *Lecture Notes in Computer Science*, σελίδες 107–119. Springer, 2003.

[17] Francois Chollet. *Deep Learning with Python*. Manning Publications Co., USA, 1στη έκδοση, 2017.

[18] Anna Choromanska, Yann LeCun και Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. Στο *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*Peter Grünwald, Elad Hazan και Satyen Kale, επιμελητές, τόμος 40 στο *JMLR Workshop and Conference Proceedings*, σελίδες 1756–1760. JMLR.org, 2015.

[19] Corinna Cortes και Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[20] Koby Crammer και Gal Chechik. A needle in a haystack: local one-class optimization. Στο *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*Carla E. Brodley, επιμελητής, τόμος 69 στο *ACM International Conference Proceeding Series*. ACM, 2004.

[21] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan και Quoc V. Le. Autoaugment: Learning augmentation strategies from data. Στο *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, σελίδες 113–123. Computer Vision Foundation / IEEE, 2019.

[22] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton και Jean Claude Klein. Feedback on a publicly distributed database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[23] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li και Li Fei-Fei. Imagenet: A large-scale hierarchical image database. Στο *2009 IEEE conference on computer vision and pattern recognition*, σελίδες 248–255. Ieee, 2009.

[24] Chris Drummond και Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 2003.

[25] Dheeru Dua και Casey Graff. UCI machine learning repository, 2017.

[26] John C. Duchi, Elad Hazan και Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

[27] Charles Elkan. The foundations of cost-sensitive learning. Στο *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*Bernhard Nebel, επιμελητής, σελίδες 973–978. Morgan Kaufmann, 2001.

[28] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari και George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. Στο *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox και Roman Garnett, επιμελητές, σελίδες 11423–11434, 2019.

[29] Evelyn Fix και J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.

[30] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[31] Rong Ge, Furong Huang, Chi Jin και Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. Στο *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*Peter Grünwald, Elad Hazan και Satyen Kale, επιμελητές, τόμος 40 στο *JMLR Workshop and Conference Proceedings*, σελίδες 797–842. JMLR.org, 2015.

[32] Xavier Glorot και Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. Στο *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, σελίδες 249–256, 2010.

[33] Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos και Koray Kavukcuoglu. Automated curriculum learning for neural networks. *CoRR*, αβς/1704.03003, 2017.

[34] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega και Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 2016.

[35] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang και Guangtong Zhou. On the class imbalance problem. Στο *2008 Fourth International Conference on Natural Computation*, τόμος 4, σελίδες 192–201, 2008.

[36] Haibo He, Yang Bai, E. A. Garcia και Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. Στο *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, σελίδες 1322–1328, 2008.

[37] Hui Han, Wenyuan Wang και Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. Στο *Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I*De-Shuang Huang, Xiao-Ping (Steven) Zhang και Guang-Bin Huang, επιμελητές, τόμος 3644 στο *Lecture Notes in Computer Science*, σελίδες 878–887. Springer, 2005.

[38] D. Harrison και D.L. Rubinfeld. Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. Deep residual learning for image recognition. Στο *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, σελίδες 770–778. IEEE Computer Society, 2016.

[40] Gao Huang, Zhuang Liu, Laurensvan der Maaten και Kilian Q. Weinberger. Densely connected convolutional networks. Στο *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, σελίδες 2261–2269. IEEE Computer Society, 2017.

[41] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft και Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[42] George Ioannou, Thanos Tagaris και Andreas Stafylopatis. Improving the convergence speed of deep neural networks with biased sampling. Στο *The 3rd International Conference on Advances in Artificial Intelligence, ICAAI 2019, Istanbul, Turkey, October 26-28, 2019*, σελίδες 35–41. ACM, 2019.

[43] Sergey Ioffe και Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Στο *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*Francis R. Bach και David M. Blei, επιμελητές, τόμος 37 στο *JMLR Workshop and Conference Proceedings*, σελίδες 448–456. JMLR.org, 2015.

[44] Nathalie Japkowicz, Catherine Myers και Mark A. Gluck. A novelty detection approach to classification. Στο *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, σελίδες 518–523. Morgan Kaufmann, 1995.

[45] Nathalie Japkowicz και Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

[46] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade και Michael I. Jordan. How to escape saddle points efficiently. Στο *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*Doina Precup και Yee Whye Teh, επιμελητές, τόμος 70 στο *Proceedings of Machine Learning Research*, σελίδες 1724–1732. PMLR, 2017.

[47] Rie Johnson και Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. Στο *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani και Kilian Q. Weinberger, επιμελητές, σελίδες 315–323, 2013.

[48] Kaggle και EyePacs. Kaggle diabetic retinopathy detection, 2015.

[49] Andreas Kaplan και Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

[50] Angelos Katharopoulos και François Fleuret. Not all samples are created equal: Deep learning with importance sampling. Στο *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*Jennifer G. Dy και Andreas Krause, επιμελητές, τόμος 80 στο *JMLR Workshop and Conference Proceedings*, σελίδες 2530–2539. JMLR.org, 2018.

[51] Diederik P. Kingma και Jimmy Ba. Adam: A method for stochastic optimization. Στο *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*Yoshua Bengio και Yann LeCun, επιμελητές, 2015.

[52] Robert Kleinberg, Yuanzhi Li και Yang Yuan. An alternative view: When does SGD escape local minima? Στο *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*Jennifer G. Dy και Andreas Krause, επιμελητές, τόμος 80 στο *Proceedings of Machine Learning Research*, σελίδες 2703–2712. PMLR, 2018.

[53] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan και Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

[54] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[55] Alex Krizhevsky, Vinod Nair και Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). χ.χ.

[56] Alex Krizhevsky, Ilya Sutskever και Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[57] Matjaz Kukar και Igor Kononenko. Cost-sensitive learning with neural networks. Στο *13th European Conference on Artificial Intelligence, Brighton, UK, August 23-28 1998, Proceedings.*Henri Prade, επιμελητής, σελίδες 445–449. John Wiley and Sons, 1998.

[58] Felix Last, Georgios Douzas και Fernando Bação. Oversampling for imbalanced learning based on k-means and SMOTE. *CoRR*, αβς/1711.00837, 2017.

[59] Steve Lawrence, Ian Burns, Andrew D. Back, Ah Chung Tsoi και C. Lee Giles. Neural network classification and prior class probabilities. Στο *Neural Networks:*

*Tricks of the Trade*Genevieve B. Orr και Klaus-Robert Müller, επιμελητές, τόμος 1524 στο *Lecture Notes in Computer Science*, σελίδες 299–313. Springer, 1996.

[60] Yann LeCun, Yoshua Bengio και Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[61] Yann LeCun και Corinna Cortes. MNIST handwritten digit database. 2010.

[62] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer και Tom Goldstein. Visualizing the loss landscape of neural nets. Στο *Advances in Neural Information Processing Systems*, σελίδες 6389–6399, 2018.

[63] Stan Lipovetsky και Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17:319 – 330, 2001.

[64] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu και Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.*, 501:511–522, 2019.

[65] Xu-Ying Liu, Jianxin Wu και Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B*, 39(2):539–550, 2009.

[66] Ilya Loshchilov και Frank Hutter. Online batch selection for faster training of neural networks. *CoRR*, αβς/1511.06343, 2015.

[67] Ilya Loshchilov και Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[68] Scott M. Lundberg και Su-In Lee. A unified approach to interpreting model predictions. Στο *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*Isabelle Guyon, Ulrikevon Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan και Roman Garnett, επιμελητές, σελίδες 4765–4774, 2017.

[69] Liangchen Luo, Yuanhao Xiong, Yan Liu και Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. Στο *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[70] Yang Lu, Yiu-ming Cheung και Yuan Yan Tang. Hybrid sampling with bagging for class imbalance learning. Στο *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I*James Bailey, Latifur Khan, Takashi Washio, Gillian Dobbie, Joshua Zhexue Huang και Ruili Wang, επιμελητές, τόμος 9651 στο *Lecture Notes in Computer Science*, σελίδες 14–26. Springer, 2016.

[71] Andrew L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[72] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado και Jeffrey Dean. Distributed representations of words and phrases and their compositionality. Στο *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani και Kilian Q. Weinberger, επιμελητές, σελίδες 3111–3119, 2013.

[73] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[74] Mary M. Moya και Don R. Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[75] David R. Musser. Introspective sorting and selection algorithms. *Softw., Pract. Exper.*, 27(8):983–993, 1997.

[76] Vinod Nair και Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. Στο *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*Johannes Fürnkranz και Thorsten Joachims, επιμελητές, σελίδες 807–814. Omnipress, 2010.

[77] Zineb Noumir, Paul Honeine και Cedue Richard. On simple one-class classification methods. Στο *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, σελίδες 2022–2026. IEEE, 2012.

[78] Terence Parr, James D. Wilson και Jeff Hamrick. Nonparametric feature impact and importance. *CoRR*, αβς/2006.04750, 2020.

[79] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai και Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. Στο *Advances in Neural Information Processing Systems 32*, σελίδες 8024–8035. Curran Associates, Inc., 2019.

[80] Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, Tianbo Wu, Jing Xiao, Fengyan Wang, Baocai Yin, Yunzhi Wang, Gopichandh Danala, Linsheng He, Yoon Ho Choi και Fabrice Mériaudeau. Idrid: Diabetic retinopathy - segmentation and grading challenge. *Medical Image Anal.*, 59, 2020.

[81] Sehrish Qummar, Fiaz Gul Khan, Sajid Shah, Ahmad Khan, Shahaboddin Shamshirband, Zia Ur Rehman, Iftikhar Ahmed Khan και Waqas Jadoon. A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, 7:150530–150539, 2019.

[82] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos και Alexander J. Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. Στο *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama και Roman Garnett, επιμελητές, σελίδες 2647–2655, 2015.

[83] Sashank J. Reddi, Satyen Kale και Sanjiv Kumar. On the convergence of adam and beyond. Στο *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[84] G Thippa Reddy, Sweta Bhattacharya, S Siva Ramakrishnan, Chiranji Lal Chowdhary, Saqib Hakak, Rajesh Kaluri και M Praveen Kumar Reddy. An ensemble based machine learning model for diabetic retinopathy classification. Στο *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, σελίδες 1–6, 2020.

[85] Marco Túlio Ribeiro, Sameer Singh και Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, αβς/1602.04938, 2016.

[86] Michael D. Richard και Richard P. Lippmann. Neural network classifiers estimate bayesian *a posteriori* probabilities. *Neural Comput.*, 3(4):461–483, 1991.

[87] Herbert Robbins και Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 1951.

[88] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse και Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part A*, 40(1):185–197, 2010.

[89] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh και Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, αβς/1610.02391, 2016.

[90] Ohad Shamir και Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. Στο *International conference on machine learning*, σελίδες 71–79, 2013.

[91] H. Shamsudin, U. K. Yusof, A. Jayalakshmi και M. N. Akmal Khalid. Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. Στο *2020 IEEE 16th International Conference on Control Automation (ICCA)*, σελίδες 803–808, 2020.

[92] Fanhua Shang, Kaiwen Zhou, James Cheng, Ivor W. Tsang, Lijun Zhang και Dacheng Tao. VR-SGD: A simple stochastic variance reduction method for machine learning. *CoRR*, αβς/1802.09932, 2018.

[93] Lloyd S Shapley. A value for n-person games. Στο *Contributions to the Theory of Games II*Harold W. Kuhn και Albert W. Tucker, επιμελητές, σελίδες 307–317. Princeton University Press, Princeton, 1953.

[94] Zebang Shen, Hui Qian, Tengfei Zhou και Tongzhou Mu. Adaptive variance reducing for stochastic gradient descent. Στο *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*Subbarao Kambhampati, επιμελητής, σελίδες 1990–1996. IJCAI/AAAI Press, 2016.

[95] Connor Shorten και Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

[96] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina και Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, αβς/1605.01713, 2016.

[97] Abhinav Shrivastava, Abhinav Gupta και Ross B. Girshick. Training region-based object detectors with online hard example mining. Στο *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, σελίδες 761–769. IEEE Computer Society, 2016.

[98] Karen Simonyan, Andrea Vedaldi και Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. Στο *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*Yoshua Bengio και Yann LeCun, επιμελητές, 2014.

[99] Karen Simonyan και Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. Στο *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*Yoshua Bengio και Yann LeCun, επιμελητές, 2015.

[100] Leslie N. Smith. Cyclical learning rates for training neural networks. Στο *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, σελίδες 464–472. IEEE Computer Society, 2017.

[101] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever και Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[102] Mukund Sundararajan και Amir Najmi. The many shapley values for model explanation. Στο *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, τόμος 119 στο *Proceedings of Machine Learning Research*, σελίδες 9269–9278. PMLR, 2020.

[103] Mukund Sundararajan, Ankur Taly και Qiqi Yan. Axiomatic attribution for deep networks. Στο *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*Doina Precup και Yee Whye Teh, επιμελητές, τόμος 70 στο *Proceedings of Machine Learning Research*, σελίδες 3319–3328. PMLR, 2017.

[104] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens και Zbigniew Wojna. Rethinking the inception architecture for computer vision. Στο *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, σελίδες 2818–2826. IEEE Computer Society, 2016.

[105] Mingxing Tan και Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. Στο *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 6105–6114. PMLR, 2019.

[106] T. Tieleman και G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[107] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro και Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. Στο *Advances in neural information processing systems*, σελίδες 4148–4158, 2017.

[108] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.*, 2(3):408–421, 1972.

[109] Xiaoxia Wu, Rachel Ward και Léon Bottou. Wngrad: Learn the learning rate in gradient descent. *CoRR*, αβς/1803.02865, 2018.

[110] Muhammad Kashif Yaqoob, Syed Farooq Ali, Muhammad Bilal, Muhammad Shehzad Hanif και Ubaid M Al-Saggaf. Resnet based deep features and random forest classifier for diabetic retinopathy detection. *Sensors*, 21(11):3883, 2021.

[111] Rahul Yedida και Snehanshu Saha. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence, 2019.

[112] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye και Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. Στο *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox και Roman Garnett, επιμελητές, σελίδες 10965–10976, 2019.

[113] Show Jane Yen και Yue Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3, Παρτ 1):5718–5727, 2009.

[114] Yu Zhang, Peter Tiňo, Aleš Leonardis και Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[115] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. Στο *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, σελίδα 928–935. AAAI Press, 2003.

Συντομογραφίες - Αρκτικόλεξα - - Ακρωνύμια

ΒΜΜ	Βαθιά Μηχανική Μάθηση
ΕΜ	Επιβλεπόμενη Μάθηση
ΕΠΜΔ	Επιλογή Παρτίδας με Μεροληπτική Δειγματοληψία
ΜΕΜ	Μη Επιβλεπόμενη Μάθηση
ΜΜ	Μηχανική Μάθηση
ΤΝ	Τεχνητή Νοημοσύνη
ΤΝΔ	Τεχνητά Νευρωνικά Δίκτυα

ANN	Artificial Neural Networks
AUC	Area Under Curve
BL	Batch Loss
BSBS	Batch Selection with Biased Sampling
CL	Curriculum Learning
CNN	Convolutional Neural Network
DL	Dataset Loss
EMA	Exponential Moving Average
ENN	Edited Nearest Neighbors
FC	Fully Connected
GD	Gradient Descent
GPU	Graphics Processing Unit
HL	High Loss
IG	Integrated Gradients
LL	Low Loss
LR	Learning Rate
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error

NBSBS-R	Noisy Batch Selection with Biased Sampling with Re-enters
OCC	One-Class Clasification
ReLu	Rectifies Linear Unit
ROC	Receiver Operating Curve
ROS	Random Oversampling
RUS	Random Undersampling
SGD	Stochastic Gradient Descent
SVDD	Support Vector Data Description

Γλωσσάρι

Activation Function	Συνάρτηση Ενεργοποίησης
Accuracy	Ορθότητα
Adaptive Optimizer	Προσαρμοστικός Βελτιστοποιητής
Artificial Intelligence	Τεχνητή Νοημοσύνη
Artificial Neural Networks	Τεχνητά Νευρωνικά Δίκτυα
Attribution	Συμβολή
Backpropagation	Οπισθοδιάδοση
Baseline	Βάση Αναφοράς
Batch	Παρτίδα
Batch Normalization	Κανονικοποίηση Παρτίδας
Bias	Πόλωση
Cardinality	Πληθικότητα
Classification	Ταξινόμηση
Convolutional Neural Networks	Συνελικτικά Νευρωνικά Δίκτυα
Cross-Entropy	Διασταυρούμενη Εντροπία
Cross-Validation	Διασταυρούμενη Επικύρωση
Deep Learning	Βαθιά Μάθηση
Dropout	Απόσυρση
Embedding	Διανυσματική Αναπαράσταση
Ensemble	Συστάδα Μοντέλων
Explainability	Επεξηγησιμότητα
Feature	Χαρακτηριστικό
Gradient Descent	Κάθοδος Κλίσης
Hidden Layer	Κρυφό Επίπεδο
Hyperbolic Tangent	Υπερβολική Εφαπτομένη
Hyperparameter	Υπερπαράμετρος
Importance	Σημασία
Input Layer	Επίπεδο Εισόδου
Interpretability	Ερμηνευσιμότητα
Layer	Επίπεδο
Learning Rate	Ρυθμός Μάθησης

Loss Function	Συνάρτηση Κόστους
Machine Learning	Μηχανική Μάθηση
Mean Square Error	Μέσο Τετραγωνικό Σφάλμα
Norm	Νόρμα
Optimization	Βελτιστοποίηση
Optimizer	Βελτιστοποιητής
Overfitting	Υπερπροσαρμογή
Partial Dependence Plot	Διάγραμμα Μερικής Εξάρτησης
Performance	Επίδοση
Pixel	Εικονοστοιχείο
Reinforcement Learning	Ενισχυτική Μάθηση
Regression	Παλινδρόμηση
Sample	Δείγμα
Segmentation	Κατάτμηση
Self-Supervised Learning	Αυτο-επιβλεπόμενη Μάθηση
Semi-Supervised Learning	Ημι-επιβλεπόμενη Μάθηση
Sigmoid Function	Σιγμοειδής Συνάρτηση
Supervised Learning	Επιβλεπόμενη Μάθηση
Test Set	Σύνολο Ελέγχου
Threshold	Κατώφλι
Training Set	Σύνολο Εκπαίδευσης
Underfitting	Υποπροσαρμογή
Unsupervised Learning	Μη Επιβλεπόμενη Μάθηση

Βιογραφικό σημείωμα του συγγραφέα

Ο Γεώργιος Ιωάννου έχει λάβει Δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών, με κατεύθυνση την Πληροφορική, από το Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ). Από το 2017 είναι υποψήφιος διδάκτορας στο Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (ΣΗΜΜΥ) του ΕΜΠ. Η Διδακτορική του Διατριβή ολοκληρώθηκε τον Δεκέμβριο του 2023. Στο πλαίσιο των Διδακτορικών του σπουδών έχει δημοσιεύσει μια σειρά από άρθρα σε ερευνητικά περιοδικά και πρακτικά συνεδρίων με κρίση. Έχει πάρει μέρος σε διάφορα ερευνητικά προγράμματα, όπως το «CORTEX – H2020, Ευρωπαϊκό ερευνητικό πρόγραμμα, το οποίο είχε ως σκοπό την ανάπτυξη τεχνικών μηχανικής μάθησης σε συνδυασμό με τεχνικές ανάλυσης σημάτων για τον εντοπισμό ανωμαλιών σε πυρήνες πυρηνικών αντιδραστήρων (2018 – 2021), πρότζεκτ της Ολυμπίας Οδού με σκοπό την ανάπτυξη μοντέλων μηχανικής μάθησης για την πρόβλεψη της κίνησης σε 2 σταθμούς διοδίων (2022), και το ερευνητικό πρόγραμμα με τίτλο «Έξυπνες Συστάσεις Τουριστικών Δράσεων Βασισμένες σε Αποδοτική Εξόρυξη Γνώσης από Ηλεκτρονικές Πλατφόρμες» που έχει ως σκοπό την εξόρυξη πληροφοριών και την εξαγωγή γνώσης από τουριστικά δεδομένα με την χρήση μηχανικής μάθησης (2023).

Έχει συμμετάσχει ως κριτής άρθρων σε ερευνητικά περιοδικά και συνέδρια και έχει παρακολουθήσει πλήθος επιστημονικών συνεδρίων. Έχει προσφέρει επικουρικό έργο στο πλαίσιο του εργαστηριακού μέρους συναφών με τα ερευνητικά του ενδιαφέροντα μαθημάτων του προπτυχιακού και μεταπτυχιακού κύκλου σπουδών του ΕΜΠ και έχει παρακολουθήσει την πορεία διπλωματικών εργασιών που εκπονήθηκαν στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης της ΣΗΜΜΥ του ΕΜΠ. Τα ερευνητικά του ενδιαφέροντα περιλαμβάνουν τις περιοχές της τεχνητής νοημοσύνης, της μηχανικής μάθησης, της ερμηνευσιμότητας και της βελτιστοποίησης.

Σύνδεσμος δημοσιεύσεων:

https://scholar.google.gr/citationsview_op=list_works&hl=el&hl=el&user=4286t10AAAAJ

Κατάλογος δημοσιεύσεων του συγγραφέα

Δημοσιεύσεις σχετικές με τη διατριβή

Περιοδικά με κρίση

- George Ioannou, Thanos Tagaris, Andreas Stafylopatis. AdaLip: An Adaptive Learning Rate Method per Layer for Stochastic Optimization. In Neural Processing Letters, 1-28
- George Ioannou, Georgios Alexandridis, Andreas Stafylopatis. Online Batch Selection for Enhanced Generalization in Imbalanced Datasets. In Algorithms 16 (2), 65, MDPI

Συνέδρια με κρίση

- George Ioannou, Thanos Tagaris, Andreas Stafylopatis. Improving the convergence speed of deep neural networks with biased sampling. In Proceedings of the 3rd International Conference on Advances in Artificial Intelligence, 2019, 35-41
- George Ioannou, Tasos Papagiannis, Thanos Tagaris, Georgios Alexandridis, Andreas Stafylopatis. Visual interpretability analysis of Deep CNNs using an Adaptive Threshold on Diabetic Retinopathy Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 480-486
- George Ioannou, Andreas Stafylopatis. The Issue of Baselines in Explainability Methods. In Proceedings of CXAI 2023 Workshop of 23rd IEEE International Conference on Data Mining (ICDM 2023)

Δημοσιεύσεις εκτός διατριβής

- George Ioannou, Thanos Tagaris, Georgios Alexandridis, Andreas Stafylopatis. Intelligent techniques for anomaly detection in nuclear reactors. In EPJ Web of Conferences 247, EDP Sciences, 2021.
- Thanos Tagaris, George Ioannou, Maria Sdraka, Georgios Alexandridis, Andreas Stafylopatis. Putting together wavelet-based scaleograms and convolutional neural networks for anomaly detection in nuclear reactors. In Proceedings of the 3rd International Conference on Advances in Artificial Intelligence, 237-243, 2019.
- Thanos Tasakos, George Ioannou, Vasudha Verma, Georgios Alexandridis, Abdelhamid Dokhane, Andreas Stafylopatis. Deep learning-based anomaly detection in nuclear reactor cores. In Proceed-

ings of the International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering (M&C 2021), Online, 3-7, 2021.

- Laurent Pantera, Petr Stulík, Antoni Vidal-Ferràndiz, Amanda Carreño, Damián Ginestar, George Ioannou, Thanos Tasakos, Georgios Alexandridis, Andreas Stafylopatis. Localizing perturbations in pressurized water reactors using one-dimensional deep convolutional neural networks. In Sensors, 22, 1, 113, MDPI, 2021
- George Ioannou, Thanos Tasakos, Antonios Mylonakis, Georgios Alexandridis, Christophe Demaziere, Paolo Vinai, Andreas Stafylopatis. Feature extraction and identification techniques for the alignment of perturbation simulations with power plant measurements. In Proceedings of the International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering (M&C 2021), Online, 3-7, 2021.
- Antonios Papaoikonomou, James Wingate, Vasudha Verma, Aiden Durrant, George Ioannou, Tasos Papagiannis, Miao Yu, Georgios Alexandridis, Abdelhamid Dokhane, Georgios Leontidis, Stefanos Kollias, Andreas Stafylopatis. Deep learning techniques for in-core perturbation identification and localization of time-series nuclear plant measurements. In Annals of Nuclear Energy, 178, 109373, 2022.
- Tasos Papagiannis, George Ioannou, Konstantinos Michalakis, Georgios Alexandridis, George Caridakis. Analyzing User Reviews in the Tourism & Cultural Domain – The Case of the City of Athens, Greece. In IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer Nature Switzerland, 284-293, 2023.