Ταξινόμηση χημικών δομών με χρήση βαθιάς μηχανικής μάθησης

Διπλωματική εργασία

ΔΠΜΣ Εφαρμοσμένων Μαθηματικών Επιστημών

Καρατζάς Παντελής

Επιβλέπων καθηγητής

Στεφανέας Πέτρος

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Εφαρμοσμένων Μαθηματικών
Αθήνα
Μάιος 2020

Contents

1 Εισαγωγή

Τα τελευταία χρόνια οι εξελίξεις στον τομέα της μηχανικής μάθησης και ειδικά όσον αφορά τα νευρωνικά δίκτυα είναι τεράστιες. Νέες αρχιτεκτονικές και μεθοδολογίες μας επιτρέπουν να χρησιμοποιούμε δεδομένα όπως ήχος και εικόνες και να εξάγουμε συμπεράσματα από αυτά.

Μέχρι πριν λίγα χρόνια δεν υπήρχε αυτή η δυνατότητα. Προβλήματα υπολογιστικής όρασης αντιμετοπίζονταν κατα βάση με κανόνες και όχι με μεθοδολογίες μηχανικής μάθησης. Με τις εξελίξεις στα νευρωνικά δίκτυα μπορούμε να επιλύσουμε προβλήματα όπως αναγνώριση, κατηγοριοποίηση φωτογραφίας και αναγνώριση αντικειμένου σε αυτές με εκπληκτικά καλά αποτελέσματα, σε κάποιες περιπτώσεις καλύτερα και απο τον ίδιο τον άνθρωπο.

Οι εξελίξεις στις αρχιτεκτονικές των νευρωνικών δικτύων συνεχίζουν να υπάρχουν. Το κομμάτι όμως των εφαρμογών αυτών γνωρίζει απίστευτη άνθιση. Ο συνδιασμός της πρόσβασης σε δεδομένα και υπολογιστική δύναμη έχει βοηθήσει ερευνητές και προγραμματιστές στο να βρίσκουν εφαρμογές νευρωνικών δικτύων με απίστευτους ρυθμούς.Έτσι και εμείς προχωρήσαμε σε κάποιες μοντελοποιήσεις που δεν θα ήταν εφικτό να γίνουν αν δεν υπήρχαν αυτές οι εξελίξεις στον τομέα της μηχανικής μάθησης και της επιστήμης των υπολογιστών.

Εδώ και πολλά χρόνια γίνονται προσπάθειες αποτύπωσης χημικών δομών με διάφορους τρόπους. Μια κοινά αποδεκτή από επιστημονικές κοινότητες είναι η επιλογή των 'Smiles'. Το σύστημα απλοποιημένης μοριακής γραμμικής γραφής (simplified molecular-input line-entry system ή SMILES) είναι μια προδιαγραφή με τη μορφή μιας γραμμικής σημειογραφίας για την περιγραφή της δομής των χημικών ειδών χρησιμοποιώντας σύντομες συμβολοσειρές ASCII. Οι συμβολοσειρές SMILES μπορούν να εισαχθούν από τους περισσότερους μοριακούς επεξεργαστές για μετατροπή πάλι σε δισδιάστατα σχέδια ή τρισδιάστατα πρότυπα των μορίων.

Η δισδιάστατη αποτύπωση αυτών, δηλαδή η περιγραφική φωτογραφία μιας χημικής δομής θα αποτελέσει και το αντικείμενο μελέτης της επικείμενης διπλωματικής. Οι φωτογραφίες των χημικών δομών θα χρησιμοποιηθούν σαν είδοσος σε νευρωνικά δίκτυα που θα προσπαθήσουν να μοντελοποιήσουν την απόκριση αυτών σε ανθρώπινους ιστούς και κύταρα. Σκοπός ήταν το να δούμε το αν η συγκεκριμένη πληροφορία μπορεί να χρησιμοποιηθεί σε μοντελοποιήσεις.

Μετά από μελέτες αποκρίσεις χημικών ή φαρμάκων μετρώνται με αναλυτικούς τρόπους μέσα από κοστοβόρες διαδικασίες. Ποιοτικά μοντέλα πρόβλεψης ή προσομοιώσεις μπορούν να μειώσουν αυτό το κόστος. Την ίδια στιγμή μπορούμε πλέον να μιλάμε σίγουρα για μείωση των πειραμάτων σε ζώα και σύντομα για την εξάλειψη αυτών.

2 Νευρωνικά Δίκτυα

Τα τεχνητά νευρονικά δίκτυα (ANN) παρέχουν μια γενική, πρακτική μέθοδο για την μάθηση μάθηση και την επίλυση προβλημάτων παλινδρόμησης και ταξινόμησης. Αλγόριθμοι εκπαίδευσης όπως το BACK-PROPAGATION επιτρέπουν την δημιουργία νευρωνικών δικτύων με πολλά επίπεδα. Βρίσκοντας την κλίση (gradient) αναπροσαρμόζονται οι παραμέτροι του δικτύου μαθαίνοντας ένα σετ δεδομένων από των ζεύγη εισόδου-εξόδου. Τα νευρωνικά δίκτυα είναι ισχυρά σε προβλήματα γνωστικών λειτουργιών και έχουν εφαρμοστεί με επιτυχία σε προβλήματα όπως η διερμηνεία οπτικών σκηνών, την αναγνώριση ομιλίας και την εκμάθηση στρατηγικών ελέγχου.

2.1 Αλγόριθμοι Σύγκλισης και Εκπαίδευση Νευρωνικών Δικτύων

2.2 Εισαγωγή

Υπάρχουν κατηγορίες προβλημάτων που δεν μπορούν να διατυπωθούν σαν αλγόριθμοι. Τέτοια προβλήματα εξαρτώνται από αρκετά παραμέτρους, οι οποίες συνήθως μεταβάλλονται με τον χρόνο. Χαρακτηριστικό παράδειγμα είναι η πρόβλεψη του καιρού, ένα χαοτικό σύστημα του οποίου η συμπεριφορά εξαρτάται από πολλές δυναμικές παραμέτρους, όπως η πίεση, το υψόμετρο, η τοποθεσία κ.α. Ένας άνθρωπος που έχει ζήσει αρκετά πάνω στην γη ξέρει ότι όταν δει πολλά σύννεφα στον ουρανό τότε κατά πάσα πιθανότητα θα βρέξει (χωρίς απαραίτητα να ξέρει τί προκάλεσε την δημιουργία των συννέφων ή για ποιο λόγο αρχίζει η βροχή) και γι'αυτό προετοιμάζεται κατάλληλα παίρνοντας ομπρέλα ή/και φορώντας περισσότερα ρούχα. Έχει *μάθει* δηλαδή να συσχετίζει τα σύννεφα με την βροχή, μέσα από τις εμπειρίες του, και έτσι μπορεί και προσαρμόζεται στις ενδεχόμενες αλλαγές του καιρού. Ένας υπολογιστής όμως απαιτεί έναν αλγόριθμο για να μπορέσει να κάνει το ίδιο, δηλαδή να συσχετίσει αλλαγές στο περιβάλλον με αλλαγές στον καιρό. Αυτή είναι και η ριζική διαφορά μεταξύ του ανθρώπινου εγκεφάλου και ενός υπολογιστή. Ότι ο ανθρώπινος εγκέφαλος *μαθαίνει* και *προσαρμόζεται*. Οι υπολογιστές μπορούν να εκτελέσουν πολύ σύνθετους υπολογισμούς σε πολύ λίγο χρόνο όμως δεν είναι καθόλου προσαρμοστικοί. Θα ήταν λοιπόν πολύ χρήσιμο αν μπορούσαμε να κάνουμε έναν υπολογιστή να *σκέφτεται* όπως ακριβώς σκέφτεται και ο ανθρώπινος εγκέφαλος. Στην επιθυμία αυτή βασίζεται η επιστήμη των νευρωνικών δικτύων. Θα θέλαμε δηλαδή να δημιουργήσουμε μία ψηφιακή δομή στον υπολογιστή η οποία να επεξεργάζεται δεδομένα με παρόμοιο τρόπο που τα επεξεργάζεται και ο ανθρώπινος εγκέφαλος. Δίνουμε λοιπόν έναν πρωταρχικό ορισμό ενος νευρωνικού δικτύου [4]:

Ορισμός: Ένα (τεχνητό) *νευρωνικό δίκτυο* είναι ένα διασυνδεδεμένο συγκρότημα από απλά στοιχεία επεξεργασίας, τους *κόμβους* ή *νευρώνες*, η λειτουργία των οποίων είναι βασισμένη στους νευρώνες των ανθρώπων. Η ικανότητα επεξεργασίας ενός νευρωνικού δικτύου είναι αποθηκευμένη στους ενδονευρωνικούς συνδέσμους, τα λεγόμενα *βάρη*, οι οποίοι προσδιορίζονται από μία διαδικασία προσαρμογής (ή *μάθησης*) σε ένα σύνολο από μοτίβα εκπαίδευσης.

Θα συνεχίσουμε κάνοντας μία μικρή αναφορά στα βασικά στοιχεία της νευροβιολογίας. Ο ανθρώπινος εγκέφαλος αποτελείται από περίπου 10^{11} (100 δισεκατομμύρια) νευρικά κύτταρα που λένονται *νευρώνες* [4]. Οι νευρώνες επικοινωνούν μεταξύ τους μέσω ηλεκτρικών μηνυμάτων τα οποία αποτελούν εφήμερους παλμούς της ηλεκτρικής τάσης του κυτταρικού τοιχώματος. Ανάμεσα στις ενδονευρωνικές συνδέσεις παρεμβάλλονται ηλεκτροχημικές διασυνδέσεις που λέγονται *συνάψεις*. Βρίσκονται στα "κλαδία" του νευρικού κυττάρου, τους λεγόμενους *δενδρίτες*. Κάθε νευρώνας δέχεται αρκετές χιλιάδες συνδέσεις από άλλους νευρώνες και συνεπώς δέχεται σε κάθε στιγμή μία πληθώρα από εισερχόμενα ηλεκτρικά σήματα, τα οποία εν τέλει φθάνουν στο κυρίως σώμα του κυττάρου. Εκεί, αθροίζονται με κάποιον τρόπο και, χονδρικά, αν το τελικό σήμα είναι μεγαλύτερο από μία συγκεκριμένη τιμή τότε ο νευρώνας θα παράγει έναν ηλεκτρικό παλμό. Ο παλμός αυτός τότε θα διαδοθεί σε άλλους νευρώνες μέσω μίας ινώδους δομής γνωστής ώς *άξονας* ή *νευράξονας*. Η σχηματική αναπαράσταση ενός νευρικού κυττάρου φαίνεται παρακάτω

Το αν ένα νευρικό κύτταρο θα εκπέμψει ένα ηλεκτρικό σήμα εξαρτάται από τα εισερχόμενα ηλεκτρικά σήματα. Κάποια από αυτά παράγουν μία παρεμποδιστική δράση και τείνουν να εμποδίσουν την εκπομπή του σήματος. Κάποια άλλα δρουν διεγερτικά και προάγουν την εκπομπή του ηλεκτρικού σήματος. Τελικά

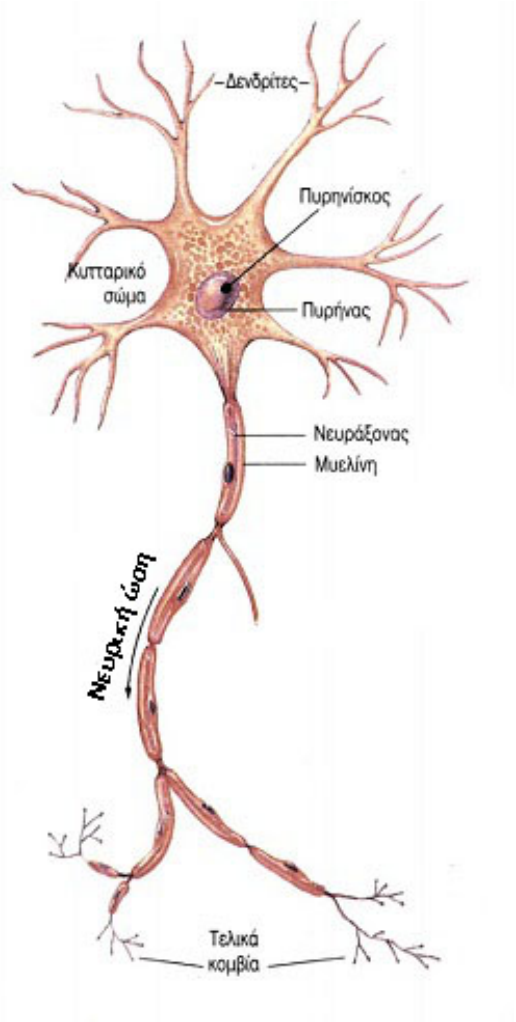


Figure 1: Σχηματική Αναπαράσταση ενός Νευρικού Κυττάρου

η ικανότητα του κάθε νευρικού κυττάρου να επεξεργάζεται δεδομένα βασίζεται στον τύπο (παρεμποδιστικό ή διεγετρικό) και στην δύναμη των συναπτικών συνδέσεων με άλλους νευρώνες.

Αυτού του είδους η αρχιτεκτονική και στυλ επεξαργασίας δεδομένων επιθυμούμε να ενσωματώσουμε στα τεχνητά νευρωνικά δίκτυα. Το τεχνητό ισοδύναμο των βιολογικών νευρώνων είναι οι κόμβοι που εμφανίστηκαν στον αρχικό ορισμό του νευρωνικού δικτύου. Οι συνάψεις μοντελοποιούνται με έναν πραγματικό αριθμό, το *βάρος* (weight), έτσι ώστε κάθε είσοδος να πολλαπλασιάζεται με ένα βάρος πριν εισέλθει στο τεχνητό ισοδύναμο με το κυρίως σώμα του νευρικού κυττάρου. Τα σταθμισμένα σήματα εισόδου αθροίζονται ακρίβως όπως γνωρίζουμε ότι αθροίζονται οι πραγματικοί αριθμοί και στην συνέχεια επεξεργάζονται από ειδικές συναρτήσεις, οι οποίες δίνουν το όριο με βάση το οποίο το κύτταρο αποφασίζει αν θα εκπέμψει το ηλεκτρικό του σήμα ή όχι.

2.3 Η μαθηματική δομή ενός Νευρωνικού Δικτύου

Η μαθηματική μοντελοποίηση ενός νευρωνικού δικτύου γίνεται με ένα *συνεκτικό κατευθυνόμενο γράφημα* (connected oriented graph). Αποτελείται δηλαδή από κόμβους (nodes), οι οποίοι καλούνται *νευρώνες* (neurons) και από *ακμές* (edges), οι οποίες συνδέουν τους νευρώνες μεταξύ τους. Η έννοια του συνεκτικού γραφήματος σημαίνει ότι κάθε νευρώνας θα πρέπει να συνδέεται με τουλάχιστον έναν άλλο νευρώνα μέσω μίας ακμής. Οι νευρώνες δεν τοποθετούνται τυχαία στο γράφημα αλλά έχουν μία συγκεκριμένη *δομή*. Πιο συγκεκριμένα, οι κόμβοι του γραφήματος κατανέμονται στα λεγόμενα *στρώματα* (layers) του δικτύου. Ένα τέτοιο στρώμα αποτελείται από κόμβους οι οποίοι *δεν επικοινωνούν μεταξύ τους* (δηλαδή δεν συνδέονται με ακμές) αλλά δέχονται πληροφορία από προηγούμενα στρώματα. Μπορούμε να θεωρήσουμε ένα στρώμα του νευρωνικού δικτύου με n νευρώνες ώς ένα διάνυσμα του $\mathbb{R}^n$.
Σε ένα νευρωνικό δίκτυο μπορούμε να διακρίνουμε τρία είδη στρωμάτων [6].

Στρώμα Εισόδου (Input Layer) : Το στρώμα αυτό αποτελεί την είσοδο του νευρωνικού δικτύου, με την έννοια ότι στο στρώμα αυτό τοποθετούνται τα δεδομένα προς εκπαίδευση. Τα δεδομένα αυτά, τα οποία μπορεί να προέρχονται από κάποια βάση δεδομένων (φωτογραφίες, κομμάτια ήχου κ.α.) ή ακόμη και να αποτελούν συναρτήσεις, προωθούνται στα επόμενα στρώματα του δικτύου με σκοπό να αρχίσει η εκπαίδευσή τους.

Κρυφά Στρώματα (Hidden Layers) : Τα στρώματα αυτά αποτελούν την καρδιά ενός νευρωνικού δικτύου καθώς σε αυτά γίνεται το μεγαλύτερο μέρος της εκπαίδευσης. Κάθε νευρώνας ενός κρυφού στρώματος συμβολίζει μία *συνάρτηση ενεργοποίησης* (activation function), οι οποίες θα αναλυθούν στην συνέχεια. Κάθε ακμή συμβολίζει το σήμα που μεταδίδεται από έναν νευρώνα στον επόμενο. Ο αριθμός των κρυφών στρωμάτων και ο αριθμός των νευρώνων σε κάθε κρυφό στρώμα αποτελεί θέμα κυρίως εμπειρικό και δεν υπάρχει ακόμη ιδιαίτερα κατατοπιστική μεθοδολογία για τον προσδιορισμό ενός βέλτιστου αριθμού κρυφών στρωμάτων και νευρώνων.

Στρώμα Εξόδου (Output Layer) : Στο στρώμα αυτό εισέρχονται τα εκπαιδευμένα δεδομένα που έχουν εξέλθει από το τελευταίο κρυφό στρώμα. Ο αριθμός των νευρώνων σε αυτό το στρώμα εξαρτάται από το εκάστοτε πρόβλημα.

Εφόσον το νευρωνικό δίκτυο σχηματίσει επιτυχώς το στρώμα εξόδου, σκοπός μας είναι να *συγκρίνουμε* τα δεδομένα του στρώματος εξόδου με τις προβλέψεις μας, δηλαδή με τις επιθυμητές τιμές. Αυτό επιτυγχάνεται με τη βοήθεια μιας ειδικής *συνάρτησης κόστους* (Loss Function). Η συνάρτηση αυτή μας βοηθά να ποσοτικοποιήσουμε το σφάλμα ή την απόκλιση των δεδομένων του στρώματος εξόδου σε σχέση με τις επιθυμητές τιμές. Στόχος μας προφανώς είναι να ελαχιστοποιήσουμε το σφάλμα αυτό και σε αυτήν την ελαχιστοποίηση συνίσταται η *εκπαίδευση* ενός νευρωνικού δικτύου, όπως θα αναλυθεί παρακάτω.

2.4 Εκπαίδευση ενός Νευρωνικού Δικτύου

Με τον όρο *εκπαίδευση* του νευρωνικού δικτύου εννοούμε την *ρύθμιση* ή *προσαρμογή* των ειδικών *παραμέτρων* του δικτύου, που ονομάζονται *βάρη* (weights) και *μεροληψίες* (στο εξής bias). Οι παράμετροι

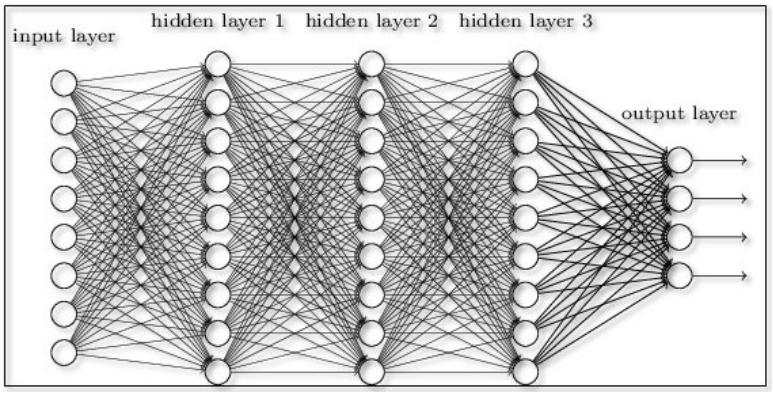


Figure 2: Αναπαράσταση ενός νευρωνικού δικτύου με τρία κρυφά στρώματα

αυτοί αρχικοποιούνται τυχαία και με βάση τον αλγόριθμο που θα περιγραφεί παρακάτω προσπαθούμε να βρούμε τις βέλτιστες δυνατές τιμές αυτών.

Ας υποθέσουμε ότι έχουμε ένα νευρωνικό δίκτυο με N κρυφά στρώματα, δηλαδή συνολικά το δίκτυο έχει $N+2$ στρώματα. Σε κάθε στρώμα αντιστοιχίζονται δύο *πίνακες*, ο πίνακας των βαρών, $\mathbf{W}^l_{jk}$ και ο πίνακας των biases, $\mathbf{b}^l_j$. Με $\mathbf{W}^l_{jk}$ συμβολίζουμε το βάρος για την σύνδεση του k-οστού νευρώνα του (l-1)-οστού στρώματος στον j-οστό νευρώνα του l-οστού στρώματος . Ως σύμβαση το στρώμα εισόδου θα έχει $l=0$, τα κρυφά στρώματα θα έχουν $l=1,\ldots,N$ ενώ το στρώμα εξόδου θα έχει $l=L$. Κάθε στοιχείο του πίνακα των βαρών μεταφράζεται ως η συνεισφορά ενός συγκεκριμένου νευρώνα στο υπολογισμό της εξόδου του επόμενου νευρώνα με τον οποίο συνδέεται. Το μοναδικό στρώμα για το οποίο δεν έχει νόημα να αντιστοιχίσουμε πίνακα βαρών είναι το στρώμα εξόδου αφού η προώθηση τελειώνει εκεί. Με $\mathbf{b}^l_j$ συμβολίζουμε το bias του j-οστού νευρώνα στο l-οστό στρώμα. Η διαστάσεις του πίνακα των βαρών και των biases καθορίζονται από την διάσταση (τον αριθμό των νευρώνων) των στρωμάτων στα οποία αναφέρεται. Ένας πίνακας με biases αντιστοιχίζεται σε όλα τα στρώματα εκτός από το στρώμα εισόδου και έχει διάσταση $1\times n$, όπου n ο αριθμός των νευρώνων του αντίστοιχου στρώματος.

Για παράδειγμα ας υποθέσουμε ότι το στρώμα εισόδου ενός νευρωνικού δικτύου έχει 5 νευρώνες και το αμέσως επόμενο στρώμα, δηλαδή το πρώτο κρυφό στρώμα, έχει 10 νευρώνες (νούμερα πάρα πολύ μικρά για ένα τυπικό νευρωνικό δίκτυο!). Τότε ο πίνακας $\mathbf{W}^1_{jk}$ είναι ο πίνακας των βαρών που δίνει την συνεισφορά των στοιχείων του στρώματος εισόδου στον υπολογισμό της εξόδου των στοιχείων του πρώτου κρυφού στρώματος. Η διάσταση του πίνακα αυτού είναι $5\times 10=50$, περιέχει δηλαδή 50 βάρη. Άρα προς το παρόν έχουμε 50 μεταβλητές. Επίσης για το πρώτο κρυφό στρώμα έχουμε επίσης 10 biases, άρα συνολικά 60 μεταβλητές. Τέλος ας υποθέσουμε ότι το νευρωνικό μας δίκτυο έχει μόνο ένα κρυφό στρώμα και το στρώμα εξόδου έχει 3 νευρώνες. Άρα ο πίνακας $\mathbf{W}^L_{jk}$ θα έχει 30 βάρη και επίσης στο στρώμα εξόδου θα αντιστοιχισθούν 3 βάρη. Συνολικά δηλαδή το νευρωνικό μας δίκτυο αποτελεί μία συνάρτηση $60+30+3=93$ μεταβλητών, οι οποίες όχι μόνο πρέπει να υπολογισθούν αλλά να βρεθεί και η βέλτιστη τιμή αυτών. Όπως είναι προφανές, ακόμα και για ένα πολύ απλό νευρωνικό δίκτυο, η εκπαίδευση αποτελεί μία πολύ περίπλοκη διαδικασία βελτιστοποίησης μιας συνάρτησης πάρα πολλών μεταβλητών.

Περιληπτικά η μεθοδολογία της εκπαίδευσης είναι η εξής: Η πληροφορία των δεδομένων εισόδου *προωθείται* μέσα στο νευρωνικό δίκτυο και επεξεργάζεται μέσω της διαδικασίας *feed forward*. Μόλις η πληροφορία φθάσει στο στρώμα εξόδου συγκρίνεται με τις αντίστοιχες επιθυμητές τιμές (που μπορεί να είναι για παράδειγμα πειραματικά δεδομένα ή δικές μας προβλέψεις). Η σύγκριση αυτή γίνεται με την βοήθεια μιας συνάρτησης κόστους. Πιο συγκεκριμένα λαμβάνεται η *μερική παράγωγος* της συνάρτησης κόστους ως προς όλες τις παραμέτρους του στρώματος εξόδου (βάρη και biases). Ανάλογα με το πρόσημο της κάθε μερικής παραγώγου το νευρωνικό δίκτυο καταλαβαίνει με ποιον τρόπο θα πρέπει να αλλάξει η κάθε παράμετρος *προκειμένου να ελαχιστοποιηθεί η συνάρτηση κόστους* που έχουμε επιβάλλει. Στην συνέχεια η πληροφορία αυτή *διαδίδεται προς τα πίσω* μέσω της διαδικασίας της *Προς-τα-πίσω διάδοσης* (Back-propagation). Η διαδικασία αυτή αποτελεί μία συνεχόμενη εφαρμογή του *κανόνα της αλυσίδας* και

βοηθά το νευρωνικό δίκτυο να καταλάβει με ποιον τρόπο πρέπει να αλλάξουν όλες οι παράμετροι όλων των υπόλοιπων στρωμάτων του δικτύου. Μία επανάληψη του κύκλου που αποτελείται από μία διαδικασία feed-forward και μία διαδικασία back-propagation καλείται *εποχή* (epoche). Το νευρωνικό δίκτυο διανύει όσες εποχές χρειάζεται έτσι ώστε να βρεθεί ένα ελάχιστο (το οποίο τις περισσότερες φορές είναι τοπικό και πάρα πολύ σπάνια ολικό) της συνάρτησεις κόστους.

Συνεχίζουμε παρουσιάζοντας το πρώτο στάδιο της εκπαίδευσης ενός νευρωνικού δικτύου, την προώθηση της πληροφορίας.

2.5 Προώθηση (Feed Forward)

Το πρώτο στάδιο της εκπαίδευσης ενός νευρωνικού δικτύου είναι η εισαγωγή των δεδομένων μας σε αυτό, μέσω του στρώματος εισόδου, και η επεξεργασία τους από τα κρυφά στρώματα. Η διαδικασία είναι η εξής [5]:
Ένα δεδομένο $\mathbf{x}_k$ που εξέρχεται από έναν νευρώνα του στρώματος εισόδου πολλαπλασιάζεται με το στοιχείο $\mathbf{W}^1_{jk}$ και στο αποτέλεσμα προστίθεται το αντίστοιχο βάρος $\mathbf{b}^1_k$. Σχηματίζεται έτσι η ποσότητα:

$$\psi^1_{jk} = \mathbf{W}^1_{jk}\mathbf{x}_k + \mathbf{b}^1_k$$

$\forall k = 1, 2, \ldots, N$ όπου N ο αριθμός των νευρώνων του στρώματος εισόδου. Στην συνέχεια αθροίζουμε πάνω στον δείκτη k και παίρνουμε έτσι ένα *σταθμισμένο άθροισμα* των δεδομένων εισόδου. Το άθροισμα αυτό θα αποτελέσει την είσοδο του j νευρώνα του πρώτου κρυφού στρώματος. Όπως προαναφέρθηκε, κάθε νευρώνας ενός κρυφού στρώματος αντιπροσωπεύει μία συνάρτηση ενεργοποίησης. Συνεπώς η έξοδος του j νευρώνα του πρώτου κρυφού στρώματος θα είναι το κανονικοποιημένο σταθμισμένο άθροισμα:

$$\alpha^j_1 = \sigma\left(\sum_k \psi^1_{jk}\right) = \sigma\left(\sum_k \mathbf{W}^1_{jk}\mathbf{x}_k + \mathbf{b}^1_k\right) = \sigma\left(z^1_j\right) \tag{1}$$

Από το σημείο αυτό και μέχρι το στρώμα εξόδου η διαδικασία συνεχίζεται με παρόμοιο τρόπο. Στο πρώτο κρυφό στρώμα θα αντιστοιχισθούν οι πίνακες βαρών και biases. Για κάθε νευρώνα του πρώτου κρυφού στρώματος θα σχηματισθεί το σταθμισμένο άθροισμα και τελικά το άθροισμα αυτό θα περάσει σε όλους του νευρώνες του δεύτερου κρυφού στρώματος, όπου και θα δράσει η αντίστοιχη συνάρτηση ενεργοποίησης. Η έξοδος του κάθε νευρώνα του τελευταίου κρυφού στρώματος θα πολλαπλασιασθεί με το αντίστοιχο στοιχείο του τελευταίου πίνακα βαρών, θα προστεθεί το τελευταίο διάνυσμα biases και το τελικό αυτό αποτέλεσμα θα αποτελέσει το στρώμα εξόδου. Παίρνουμε έτσι τον αναδρομικό τύπο των τιμών ενεργοποίησης

$$\alpha^l_j = \sigma(z^l_j) = \sigma\left(\sum_k W^l_{jk}\alpha^{l-1}_k + b^l_j\right) \tag{2}$$

Η ποσότητα α^l_j ονομάζεται *τιμή ενεργοποίησης* (activation value). Το παραπάνω άθροισμα είναι πάνω σε όλους τους νευρώνες του $(l-1)$-οστού στρώματος.

Συνεχίζουμε την διερεύνηση της εκπαίδευσης ενός νευρωνικού δικτύου με μία μικρή παράκαμψη. Όπως αναφέρθηκε προηγουμένως τα δεδομένα του δικτύου επεξεργάζονται με την βοήθεια των λεγόμενων συναρτήσεων ενεργοποίησης. Θα εξετάσουμε τις πιθανές υποψήφιες κλάσσεις τέτοιων συναρτήσεων, ποιες είναι κατάλληλες και ποιες όχι.

2.6 Συναρτήσεις Ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης είναι ο τρόπος με τον οποίο η πληροφορία των σημάτων εισόδου περνούν από τον έναν νευρώνα στον επόμενο. Σηματοδοτούν δηλαδή πότε ένας νευρώνας θα "ενεργοποιηθεί" και θα μεταβιβάσει την εκάστοτε πληροφορία. Υπάρχουν πολλές υποψήφιες επιλογές συναρτήσεων ενεργοποίησης και η επιλογή της "καταλληλότερης" συνάρτησης για τον κάθε νευρώνα (η συνηθέστερα για το κάθε κρυφό στρώμα) είναι ως επί το πλείστον εμπειρική και αποτελεί πηγή πολλών ερευνητικών

δυνατοτήτων. Για να καταλάβουμε τί είδους συναρτήσεις είναι κατάλληλες για την μεταφορά πληροφορίας μεταξύ των νευρώνων μπορούμε να ξεκινήσουμε μελετώντας ποιες συναρτήσεις *δεν είναι κατάλληλες* για αυτόν τον σκοπό.

Η μελέτη των συναρτήσεων ενεργοποίησης ξεκίνησε με την εξής απλή σκέψη: Εφόσον ένας νευρώνας είτε θα μεταβιβάσει είτε δεν θα μεταβιβάσει την πληροφορία στον επόμενο νευρώνα, τι πιο φυσικό από το να χρησιμοποιήσουμε την *βηματική συνάρτηση* (step function):

$$\sigma(x) = \begin{cases} 1 & , x > 0 \\ 0 & , x < 0 \end{cases}$$

Παρόλο που η παραπάνω συνάρτηση φαίνεται διαισθητικά σωστή, στην πραγματικότητα εμφανίζει πολύ άσχημα αποτελέσματα στην πράξη της εκπαίδευσης των δικτύων. Αυτό οφείλεται στο γεγονός ότι ο ορισμός της παραπάνω συνάρτησης θα επιτρέψει στον νευρώνα είτε να μεταδώσει πλήρως την πληροφορία του είτε καθόλου, ενεργεί δηλαδή με δυαδικό τρόπο. Στην πράξη παρόλα αυτά μία καλή εκπαίδευση θα πρέπει να περιλαμβάνει και "ποσοστά μετάδοσης" πληροφορίας. Με αυτόν τον τρόπο το δίκτυό μας έχει περισσότερους βαθμούς ελευθερίας για να προσαρμόσει τις προς προσδιορισμό παραμέτρους του και αυτό προσφέρει μεγαλύτερη ευρωστία στην αρχιτεκτονική του.

Συνεπώς θα πρέπει να επιλέξουμε μία συνάρτηση που να δίνει και ενδιάμεσες τιμές ενεργοποίησης. Η πρώτη σκέψη είναι μία γραμμική συνάρτηση:

$$\sigma(x) = \alpha x$$

Μια συνάρτηση τέτοιας μορφής μπορεί να δώσει όντως μια μεγάλη ποικιλία τιμών ενεργοποίησης, οι οποίες τιμές είναι ανάλογες της εκάστοτε εισόδου. Παρόλα αυτά οι γραμμικές συναρτήσεις έχουν δύο σοβαρά μειονεκτήματα. Το πρώτο είναι το γεγονός ότι έχουν *σταθερή βαθμίδα* (*ή κλίση*). Αυτό σημαίνει ότι αν υπάρχει σφάλμα στην πρόβλεψη, δηλαδή η έξοδος του νευρωνικού δικτύου διαφέρει από την επιθυμητή τιμή, τότε η διόρθωση και οι αλλαγές που θα γίνονται με την προς-τα-πίσω μετάδοση θα είναι σταθερές και δεν θα εξαρτώνται από το σφάλμα αυτό. Το δεύτερο μειονέκτημα δεν είναι τόσο προφανές. Αν όλες οι συναρτήσεις ενεργοποίησης επιλεγούν γραμμικές τότε η τελική έξοδος του νευρωνικού δικτύου θα είναι και αυτή μία γραμμική συνάρτηση της εισόδου. Στην ουσία το σύνολο του νευρωνικού δικτύου (που μπορεί να περιέχει μέχρι και δεκάδες κρυφά στρώματα!) θα είναι ισοδύναμο με ένα νευρωνικό δίκτυο με ένα μόνο κρυφό στρώμα, εφόσον η σύνθεση γραμμικών συναρτήσεων είναι μία επίσης γραμμική συνάρτηση. Αυτό στην πράξη σημαίνει χειρότερη εκπαίδευση.

Το επόμενο βήμα δεν είναι ιδιαίτερα προφανές. Χρειαζόμαστε μία συνάρτηση που να έχει τουλάχιστον τις εξής ιδιότητες:

- Μη-γραμμική
- Φραγμένη (προτιμέστερα μεταξύ του 0 και του 1, χωρίς να έχει ιδιαίτερη σημασία)
- Να είναι τουλάχιστον μια φορά συνεχώς παραγωγίσιμη

Μία πολύ καλή υποψήφια συνάρτηση είναι η λεγόμενη *λογιστική συνάρτηση*:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Η παραπάνω συνάρτηση έχει όλες τις παραπάνω επιθυμητές ιδιότητες συν άλλη μία η οποία δεν είναι άμεσα προφανής. Η λογιστική συνάρτηση έχει κλίση που δίνεται από τον τύπο:

$$\nabla\sigma(x) = \frac{e^x}{(e^x + 1)^2}$$

Στο διάστημα [-2,2] η κλίση της λογιστικής συνάρτησης είναι αρκετά μεγάλη, όπως φαίνεται και από την γραφική της παράσταση. Μεγάλη κλίση σημαίνει πως μικρές μεταβολές των μεταβλητών εισόδου (μικρά

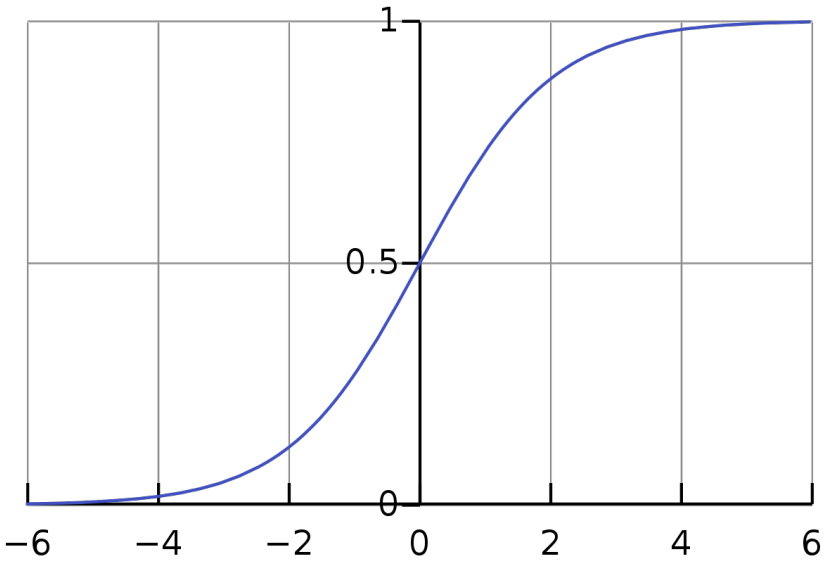


Figure 3: Γραφική παράσταση της λογιστικής συνάρτησης

Δx) οδηγούν σε μεγάλες μεταβολές των μεταβλητών εξόδου. Υπάρχει δηλαδή αρκετά μεγάλη ευαισθησία σε μικρές διαταραχές του σήματος που εισέρχεται στον νευρώνα και αυτό συνεπάγεται πιο αποδοτική εκπαίδευση με την χρήση της προς-τα-πίσω μετάδοσης.

Η λογιστική συνάρτηση είναι όντως μία από τις ευρέως χρησιμοποιούμενες συναρτήσεις ενεργοποίησης στην τεχνολογία των νευρωνικών δικτύων λόγω των προαναφερθέντων ιδιοτήτων της. Στο σημείο αυτό όμως θα πρέπει να αναφερθεί και το μοναδικό μειονέκτημα της εν λόγω συνάρτησης που οφείλεται στην ίδια την φύση της. Παρατηρούμε από την παραπάνω γραφική παράσταση ότι στο σύνολο $[-\infty, -2] \cup [2, +\infty]$ η κλίση της λογιστικής συνάρτησης πρακτικά μηδενίζεται (η συνάρτηση τείνει να γίνει οριζόντια). Παρουσιάζεται δηλαδή το φαινόμενο των *εξαφανισμένων βαθμίδων* (vanishing gradients). Πολύ μικρές βαθμίδες στα διαστήματα αυτά σημαίνει πως μόλις η εκπαίδευση μας οδηγήσει στα "άκρα" της συνάρτησης τότε η εκπαίδευση επιβραδύνεται με πολύ μεγάλο ρυθμό και πρακτικά σταματά.

Το επόμενο βήμα αποτελεί μία βελτίωση της λογιστικής συνάρτησης και ονομάζεται *υπερβολική εφαπτομένη*:

$$\sigma(x) = \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Συγκρίνοντας τις κλίσεις της λογιστικής συνάρτησης και της υπερβολικής εφαπτομένης παρατηρούμε ότι στο κοινό διάστημα [-2,2] η υπερβολική εφαπτομένη έχει μεγαλύτερη κλίση. Κατά τ'άλλα παρουσιάζει τα ίδια μειονεκτήματα και πλεονεκτήματα με την λογιστική συνάρτηση. Η επιλογή ανάμεσα στις δύο αυτές συναρτήσεις εξαρτάται από το πόσο μεγάλες κλίσεις θέλουμε να επιβάλλουμε στο νευρωνικό μας δίκτυο. Αποτελεί προς το παρόν θέμα κυρίως εμπειρικό και εξαρτάται από το εκάστοτε πρόβλημα.

Εναλλακτικές επιλογές συναρτήσεων που παρουσιάζουν παρόμοια χαρακτηριστικά με την λογιστική συνάρτηση και την υπερβολική εφαπτομένη είναι οι εξής συναρτήσεις:

$$f(x) = \arctan x \tag{3}$$

$$f(x) = \frac{x}{1 + |x|} \tag{4}$$

$$f(x) = \frac{x}{\sqrt{1 + \alpha x^2}} \tag{5}$$

Όλες οι παραπάνω μη-γραμμικές συναρτήσεις ανήκουν στην ειδικότερη κατηγορία των *σιγμοειδών* συναρτήσεων (sigmoid functions) και είναι η πιο ευρέως χρησιμοποιούμενη κατηγορία συναρτήσεων ενεργοποίησης στα νευρωνικά δίκτυα λόγω των ιδιοτήτων τους. Τα τελευταία χρόνια όμως έχει τεθεί σε

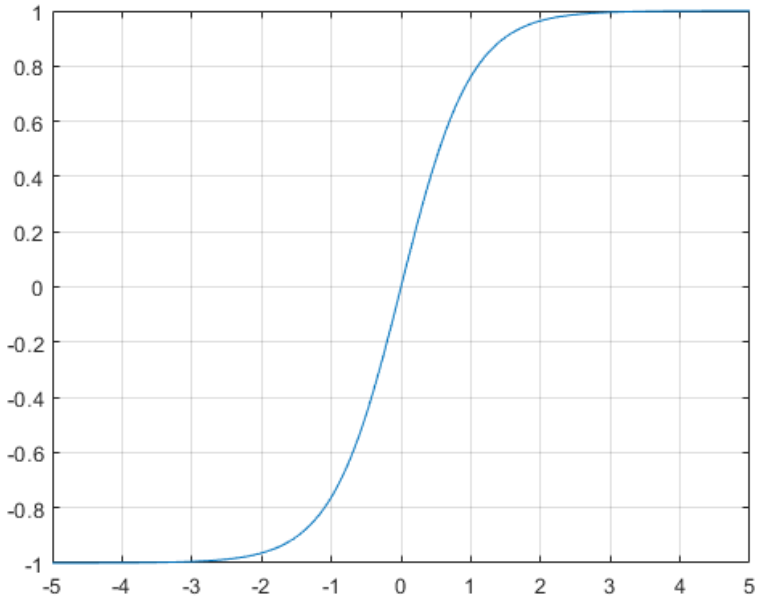


Figure 4: Γραφική παράσταση της υπερβολικής εφαπτομένης

εφαρμογή μια συνάρτηση που δεν είναι σιγμοειδής και ονομάζεται **ReLU** (Rectified Linear Unit). Δίνεται από την σχέση:

$$R(x) = \begin{cases} x & , x > 0 \\ 0 & , x < 0 \end{cases}$$

Η ReLU μπορεί να ορισθεί και ως $\mathrm{R(x)} = \max(0, x)$. Παρά το γεγονός ότι με μία πρώτη ματιά η ReLU φαίνεται γραμμική, στην πραγματικότητα δεν είναι. Μια γραμμική απεικόνιση $T : \mathbb{R} \to \mathbb{R}$ ικανοποιεί τις εξής σχέσεις:

$$T(x + y) = T(x) + T(y)$$
$$T(\lambda x) = \lambda x$$

Είναι προφανές ότι η ReLU δεν ικανοποιεί την πρώτη συνθήκη των γραμμικών συναρτήσεων. Ως παράδειγμα έχουμε:
$R(1) = \max(0, 1) = 1$. Όμως $1 = 2 - 1$ και αν υποθέσουμε ότι η ReLU είναι γραμμική τότε θα ισχύει ότι: $R(2 - 1) = R(2) + R(-1) = 2 = R(1) = 1$ καί έτσι καταλήγουμε σε άτοπο και άρα στο συμπέρασμα ότι η ReLU δεν είναι γραμμική.

Ένα ενδιαφέρον χαρακτηριστικό της ReLU είναι ότι μπορεί να δράσει πολύ καλά ως συναρτησιακός προσεγγιστής, με την έννοια ότι μπορούμε να προσεγγίσουμε όσο καλά θέλουμε μία συνεχή συνάρτηση με την χρήση γραμμικού συνδυασμού συναρτήσεων ReLU. Πιο φορμαλιστικά, για κάθε $\epsilon > 0$ και για κάθε συνεχή συνάρτηση f υπάρχει ένας φυσικός N, πραγματικοί αριθμοί α_i και συναρτήσεις ReLU ορισμένες σε κατάλληλο εκάστοτε διάστημα, ώστε :

$$\left| f - \sum_{i=1}^{N} \alpha_i R_i(x) \right| < \epsilon \tag{6}$$

Στις περισσότερες περιπτώσεις η ReLU χρησιμοποιείται σε συνδυασμό με άλλες συναρτήσεις ενεργοποίησης, συνήθως σιγμοειδείς. Η παρουσία της προσφέρει ένα σημαντικό πλεονέκτημα. Λόγω της μορφής της δεν επιτρέπει σε όλους τους νευρώνες να ενεργοποιηθούν. Αν φανταστούμε ένα νευρωνικό δίκτυο με εκατοντάδες ή και χιλιάδες νευρώνες τότε προφανώς ο υπολογιστικός χρόνος της εκπαίδευσης μπορεί να είναι πολύ μεγάλος για να είναι πρακτικός. Η χρήση της ReLU λοιπόν μας βοηθά να δημιουργήσουμε ένα νευρωνικό δίκτυο πιο αραιό και ανάλαφρο και έτσι να εξοικονομήσουμε πολύτιμο υπολογιστικό χρόνο. Αυτό είναι το μεγάλο πλεονέκτημά της. Το πρόβλημα που δημιουργεί είναι προφανές στο σημείο

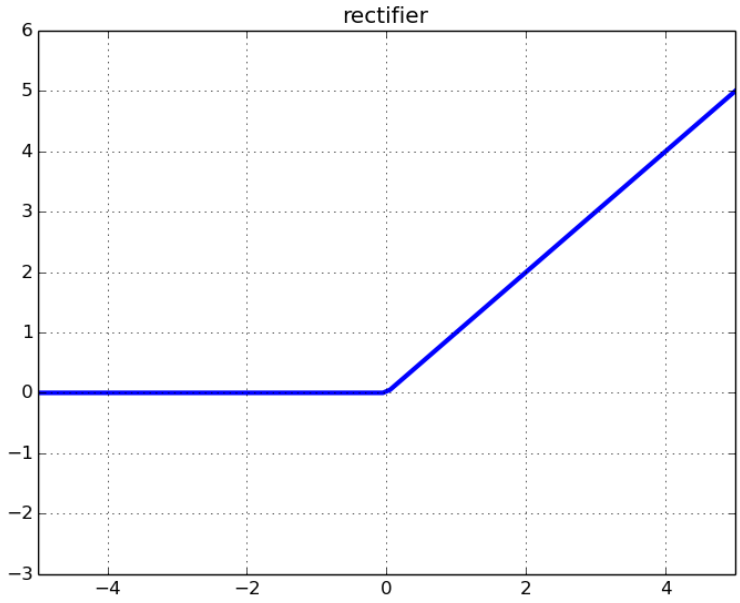


Figure 5: Γραφική παράσταση της ReLU

αυτό και είναι η μηδενική κλίση (δηλαδή τερματισμός της εκπαίδευσης) για αρνητικές τιμές εισόδου. Το πρόβλημα αυτό λύνεται με μικρές παραλλαγές της ReLU που παίρνουν την μορφή:

$$R_\epsilon(x) = \begin{cases} x & , x > 0 \\ \epsilon x & , x < 0 \end{cases}$$

όπου $\epsilon << 1$. Η μικρή παράμετρος ϵ μπαίνει έτσι ώστε να υπάρχει μη-μηδενική κλίση για αρνητικές τιμές εισόδου. Τέτοιες παραλλαγές της ReLU είναι γνωστές ως **Leaky ReLU**.

2.7 Συναρτήσεις Κόστους

Η ανάγκη για την ύπαρξη μίας συνάρτησης κόστους στο νευρωνικό δίκτυο έγκειται στο γεγονός ότι χρειαζόμαστε ένα κριτήριο με βάση το οποίο θα γίνει η εκπαίδευση του δικτύου. Οι συναρτήσεις κόστους μας βοηθούν να συγκρίνουμε τις τιμές του στρώματος εξόδου με τις επιθυμητές τιμές (προβλέψεις ή πειραματικά δεδομένα). Είναι δηλαδή ο τρόπος με τον οποίο βοηθούμε τον υπολογιστή να καταλάβει αν η διαδικασία της προώθησης ήταν επιτυχής. Για λόγους που δεν είναι άμεσα προφανείς και σχετίζονται με την διαδικασία του back-propagation οι συναρτήσεις κόστους πρέπει να πληρούν τις δύο εξής προϋποθέσεις [6]:

- Η συνάρτηση κόστους θα πρέπει να μπορεί να γραφεί ως ένας μέσος όρος $C = \frac{1}{n}\sum_x C_x$ συναρτήσεων κόστους για κάθε μεμονωμένο αντικείμενο εκπαίδευσης, x. Ο λόγος που χρειαζόμαστε αυτήν την προϋπόθεση είναι επειδή στον αλγόριθμο back-propagation απαιτείται ο υπολογισμός μερικών παραγώγων ως προς όλα τα αντικείμενα προς εκπαίδευση.

- Η συνάρτηση κόστους θα πρέπει να μπορεί να γραφτεί συναρτήσει των εξόδων του νευρωνικού δικτύου $C = C(\alpha^l)$

Μερικές από τις ευρέως χρησιμοποιούμενες συναρτήσεις κόστους είναι οι εξής:

Συνάρτηση μέσου τετραγωνικού σφάλματος (Mean Square Error Function):

$$C = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

όπου y_i οι τιμές του στρώματος εξόδου του δικτύου και $\hat{y}_i$ οι προβλέψεις μας με τις οποίες τις συγκρίνουμε. Η συνάρτηση αυτή χρησιμοποιείται ευρέως σε προβλήματα παλινδρόμησης (regression) όμως έχει ένα σοβαρό πρόβλημα. Σε περίπτωση που υπάρχουν δεδομένα (παρατηρούμενες τιμές) που απέχουν πολύ από όλα τα υπόλοιπα, τα λεγόμενα outliers, τότε η εφαρμογή της συνάρτησης τετραγωνικού σφάλματος μπορεί να έχει πολύ μεγάλα σφάλματα στους υπολογισμούς. Για τον σκοπό αυτό επινοήθηκε η συνάρτηση σφάλματος κατά Huber.

Συνάρτηση Huber (Huber Loss):

$$C = \begin{cases} \frac{1}{2}\sum_i (y_i - \hat{y}_i)^2 & , |y_i - \hat{y}_i| \leq \delta \ \ \forall i \\ \delta \sum_i |y_i - \hat{y}_i| - \frac{1}{2}\delta^2 & , \alpha\lambda\lambda\iota\omega\varsigma \end{cases}$$

Η συνάρτηση Huber είναι πολύ λιγότερο ευαίσθητη σε outliers.

Συνάρτηση Cross Entropy:

$$C = -\sum_{i=1}^{n} y_i \log \hat{y}_i$$

Η Cross Entropy αποτελεί μια από τις σημαντικότερες συναρτήσεις για το κλάδο τον Μαθηματικών με όνομα Θεωρία Πληροφορίας.

Όλες οι μορφές των παραπάνω συναρτήσεων αφορούν φυσικά διακριτό χώρο δειγματοληψίας. Αν οι κατανομές μας αποτελούν συνεχείς συναρτήσεις τότε το άθροισμα αντικαθιστάται με ολοκλήρωμα.

2.8 Κανονικοποίηση Κόστους

Η κανονικοποίηση κόστους (ή κανονικοποίηση συνάρτησης κόστους) είναι μία τεχνική που χρησιμοποιούταν αρκετές δεκαετίες πριν την ανάπτυξη της εκπαίδευσης νευρωνικών δικτύων βάθους. Γραμμικά μοντέλα όπως η γραμμική και η λογιστική παρεμβολή δίνουν την ικανότητα εφαρμογής απλών και αποτελεσματικών τεχνικών κανονικοποίησης.

Οι εφαρμογές κανονικοποίησης βασίζονται στον περιορισμό των ικανοτήτων και της χωρητικότητας του μοντέλου (είτε αυτό είναι νευρωνικά δίκτυα ή κάποιου είδους παρεμβολή) προσθέτοντας μια παραμετρικοποιημένη ποινή $\Omega(\boldsymbol{\vartheta})$ στην αντικειμενική συνάρτηση κόστους $\mathcal{J}$. Συμβολίζουμε την κανονικοποιημένη συνάρτηση κόστους ώς

$$\widetilde{\mathcal{J}} = \mathcal{J}(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}) + \alpha\Omega(\boldsymbol{\vartheta}) \tag{7}$$

όπου η πολλαπλασιαστική σταθερά $\alpha \in [0, \infty)$ είναι μία υπερπαράμετρος που σταθμίζει την σχετική συνεισφορά του όρου της ποινής, Ω, στην συνάρτηση κόστους, $\mathcal{J}$. Φυσικά θέτοντας $\alpha = 0$ έχει ως αποτέλεσμα να μην έχουμε καμία κανονικοποίηση. Όσο το α αυξάνεται τόσο αυξάνεται και η συνεισφορά της ποινής στην αντικειμενική συνάρτηση. Όταν ο αλγόριθμος εκπαίδευσής μας προσπαθεί να ελαχιστοποιήσει την αντικειμενική συνάρτηση $\widetilde{\mathcal{J}}$, τότε θα μειώνει και την συνάρτηση $\mathcal{J}$ αλλά και μία συνάρτηση της οποίας το μέτρο θα εξαρτάται καθαρά και μόνο από τις παραμέτρους $\boldsymbol{\vartheta}$. Διαφορετικές επιλογές της συνάρτησεις Ω θα έχουν φυσικά διαφορετική επίδραση στην εκπαίδευση. Στο σημείο αυτό αξίζει να σημειώσουμε ότι στις περισσότερες εφαρμογές των νευρωνικών δικτύων επιβάλλουμε ποινή μόνο στα βάρη του δικτύου και όχι στα biases [7]. Κάθε βάρος προσδιορίζει τον τρόπο που δύο μεταβλητές αλληλεπιδρούν. Η προσαρμογή ενός βάρους λοιπόν απαιτεί την παρατήρηση και των δύο αυτών μεταβλητών σε πολλές περιστάσεις. Εν αντιθέσει, ένα bias ελέγχει μόνο μία μεταβλητή. Αυτό σημαίνει ότι δεν εισάγουμε πολύ διακύμανση αφήνοντας τα βάρη μη-κανονικοποιημένα. Για το υποκεφάλαιο αυτό θα συμβολίζουμε με $\mathbf{w}$ το διάνυσμα των βαρών ενός νευρωνικού δικτύου στα οποία θα επιβάλλεται μία ποινή κανονικοποίησης και με $\boldsymbol{\vartheta}$ θα συμβολίζουμε όλες τις παραμέτρους του δικτύου (όλα τα βάρη και όλα τα

biases). Συνεχίζουμε με τα διάφορα είδη κανονικοποίησης ποινής.

- $\mathcal{L}^2$ - Κανονικοποίηση ή Κανονικοποίηση *Tykhonov*

Πρόκειται δηλαδή για την $\mathcal{L}^2$ νόρμα του διανύσματος των βαρών στα οποία επιβάλλουμε την επιπλέον ποινή. Η ποινή κόστους τύπου $\mathcal{L}^2$ είναι η πιο κοινή στην πράξη. Βοηθά στο να κρατά τα βάρη πιο κοντά στο 0 [8], όπως θα δείξει η μαθηματική ανάλυση που θα ακολουθήσει. Η συνάρτηση κανονικοποίησης Ω δίνεται από την σχέση

Με βάση την παραπάνω ποινή η αντικειμενική συνάρτηση γράφεται ως

$$\widetilde{\mathcal{J}} = \mathcal{J}(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}) + \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w} \tag{8}$$

Η συνάρτηση αυτή έχει βαθμίδα ως προς τα βάρη

$$\nabla_{\mathbf{w}}\widetilde{\mathcal{J}} = \alpha\mathbf{w} + \nabla_{\mathbf{w}}\mathcal{J}(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y})$$

Με βάση αυτήν την βαθμίδα ένα βήμα του αλγορίθμου gradient descent επικαιροποιεί το διάνυσμα των παραμέτρων $\mathbf{w}$ ως εξής

$$\mathbf{w} \leftarrow \mathbf{w} - \tau(\alpha\mathbf{w} + \nabla_{\mathbf{w}}\mathcal{J}(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}))$$

ή, αν αναδιατάξουμε τους όρους

$$\mathbf{w} \leftarrow (1 - \tau\alpha)\mathbf{w} - \tau\nabla_{\mathbf{w}}\mathcal{J}(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}) \tag{9}$$

Όπως φαίνεται από την παραπάνω εξίσωση η εισαγωγή της ποινής $\mathcal{L}^2$ τροποποίησε τον κανόνα επικαιροποίησης. Σε κάθε επανάληψη της διαδικασίας το διάνυσμα των βαρών συρρικνώνεται κατά έναν σταθερό παράγοντα. Στην συνέχεια θα εξετάσουμε την επίδραση της $\mathcal{L}^2$ ποινής κατά την διάρκεια όλης της εκπαίδευσης. Προς απλοποίηση των πράξεων του συλλογισμού που θα ακολουθήσει θα κάνουμε μία τετραγωνική προσέγγιση (στην ουσία ανάπτυγμα Taylor δευτέρου βαθμού) της μη-κανονικοποιημένης αντικειμενικής συνάρτησης $\mathcal{J}$ γύρω από το σημείο που την ελαχιστοποιεί, $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} \mathcal{J}(\mathbf{w})$. Γράφουμε δηλαδή

$$\hat{\mathcal{J}}(\mathbf{w}) = \mathcal{J}(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\intercal\mathbf{H}(\mathbf{w} - \mathbf{w}^*)$$

όπου $\mathbf{H}$ είναι ο Εσσιανός πίνακας της $\mathcal{J}$ ως προς το διάνυσμα $\mathbf{w}$ υπολογισμένος στο σημείο $\mathbf{w}^*$. Προφανώς αν η συνάρτηση κόστους είναι τετραγωνική (που στις περισσότερες εφαρμογές είναι) τότε η προσέγγιση συμπίπτει με την ίδια την συνάρτηση. Το ελάχιστο της συνάρτησης $\hat{\mathcal{J}}(\mathbf{w})$ λαμβάνεται εκεί όπου η βαθμίδα

$$\nabla_{\mathbf{w}}\hat{\mathcal{J}}(\mathbf{w}) = \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$$

είναι ίση με 0. Θα συμβολίσουμε το σημειό στο οποίο ικανοποιείται η παραπάνω συνθήκη με $\hat{\mathbf{w}}$. Για να εξετάσουμε την επίδραση της ποινής $\mathcal{L}^2$ θα πρέπει να προσθέσουμε στην παραπάνω συνθήκη και την βαθμίδα του όρου της ποινής, $\alpha\hat{\mathbf{w}}$. Λαμβάνουμε έτσι

$$\begin{aligned}
\alpha\hat{\mathbf{w}} + \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \\
(\mathbf{H} + \alpha\mathbf{I})\hat{\mathbf{w}} = \mathbf{H}\mathbf{w}^* \\
\hat{\mathbf{w}} = (\mathbf{H} + \alpha\mathbf{I})^{-1}\mathbf{H}\mathbf{w}^*
\end{aligned}$$

Παρατηρούμε ότι καθώς $\alpha \to 0$, $\hat{\mathbf{w}} \to \mathbf{w}^*$. Επειδή ο πίνακας $\mathbf{H}$ είναι πραγματικός και συμμετρικός μπορούμε να τον γράψουμε ως γινόμενο ενός διαγώνιου πίνακα $\boldsymbol{\Lambda}$ και έναν πίνακα $\mathbf{Q}$ που αποτελείται από μία ορθοκανονική βάση ιδιοδιανυσμάτων, ώστε $\mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal$ [7]. Με βάση αυτή την αποσύνθεση λαμβάνουμε

$$\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal + \alpha\mathbf{I})^{-1}\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal\mathbf{w}^* \\
&= (\mathbf{Q}(\boldsymbol{\Lambda} + \alpha\mathbf{I})\mathbf{Q}^\intercal)^{-1}\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal\mathbf{w}^* \\
&= \mathbf{Q}(\boldsymbol{\Lambda} + \alpha\mathbf{I})^{-1}\boldsymbol{\Lambda}\mathbf{Q}^\intercal\mathbf{w}^*
\end{aligned}$$

Παραρηρούμε ότι η επίδραση της ποινής $\mathcal{L}^2$ είναι ακριβώς ότι ανακλιμακώνει (rescales) το διάνυσμα $\mathbf{w}^*$ κατά μήκος των αξόνων που ορίζονται από τα ιδιοδιανύσματα του πίνακα $\mathbf{H}$.

Συνεχίζουμε τώρα με την ποινή τύπου $\mathcal{L}^1$.

- $\mathcal{L}^1 - Κανονικοποιηση$

Στην περίπτωση της κανονικοποίησης τύπου $\mathcal{L}^1$ η συνάρτηση Ω δίνεται από την σχέση

$$\Omega(\boldsymbol{\vartheta}) = \|\mathbf{w}\|_1 = \sum_i |w_i| \tag{10}$$

Πρόκειται δηλαδή για την $\mathcal{L}^1$ νόρμα του διανύσματος των βαρών στα οποία επιβάλλουμε την επιπλέον ποινή. Τώρα η κανονικοποιημένη αντικειμενική συνάρτηση έχει την μορφή

$$\widetilde{\mathcal{J}}(\mathbf{w}; \mathbf{x}, \mathbf{y}) = \alpha\|\mathbf{w}\|_1 + \mathcal{J}(\mathbf{w}; \mathbf{x}, \mathbf{y}) \tag{11}$$

όπου, ομοίως με πριν, η σταθερά αείναι μία υπερπαράμετρος που καθορίζει το μέγεθος της επίδρασης της κανονικοποίησης. Η βαθμίδα της κανονικοποιημένης αντικειμενικής συνάρτησης είναι

$$\nabla_{\mathbf{w}}\widetilde{\mathcal{J}}(\mathbf{w}; \mathbf{x}, \mathbf{y}) = \alpha sign(\mathbf{w}) + \nabla_{\mathbf{w}}\mathcal{J}(\mathbf{w}; \mathbf{x}, \mathbf{y}) \tag{12}$$

Από την παραπάνω βαθμίδα βλέπουμε ότι η επίδραση της $\mathcal{L}^1$ κανονικοποίησης είναι διαφορετική από αυτήν της $\mathcal{L}^2$. Πιο συγκεκριμένα η συνεισφορά της κανονικοποίησης στην βαθμίδα δεν αυξομειώνεται γραμμικά με την κάθε συνιστώσα w_i. Είναι απλώς μία πολλαπλασιαστική σταθερά αμε πρόσημο που εξαρτάται από το $sign(w_i)$, δηλαδή από το πρόσημο της συνιστώσας w_i. Μία συνέπεια αυτής της μορφής βαθμίδας είναι ότι μία τετραγωνική προσέγγιση της αντικειμενικής συνάρτησης δεν θα μπορεί να δώσει απαραίτητα κάποια αναλυτική λύση όπως στην περίπτωση της κανονικοποίησης $\mathcal{L}^2$. Ο κανόνας επικαιροποίησης του διανύσματος των βαρών θα είναι τώρα

$$\mathbf{w} \leftarrow \mathbf{w} + \tau(\alpha sign(\mathbf{w}) + \nabla_{\mathbf{w}}\mathcal{J}(\mathbf{w}; \mathbf{x}, \mathbf{y})) \tag{13}$$

2.9 Κατάβαση Βαθμίδας (Gradient Descent)

Η διαδικασία του Gradient Descent είναι ένας αλγόριθμος με τον οποίο μπορούμε να βρούμε ένα *τοπικό ακρότατο* μιας πολυμεταβλητής συνάρτησης. Η ιδέα πίσω από τον αλγόριθμο είναι απλή. Ξεκινώντας από ένα αυθαίρετο σημείο πάνω στην (υπερ)επιφάνεια που ορίζει η συνάρτηση μας τότε αν θέλουμε να κάνουμε ανάβαση της επιφάνειας, δηλαδή να πάμε προς το πιο κοντινό μέγιστο, τότε θα πρέπει να ακολουθήσουμε την κατεύθυνση που ορίζει η *βαθμίδα* (gradient) της συνάρτησης στο σημείο αυτό. Αυτό είναι και το φυσικό νόημα του gradient. Σε κάθε σημείο της επιφάνειας μας λέει προς τα που πρέπει να κινηθούμε έτσι ώστε να έχουμε την μεγαλύτερη αύξηση *ή κλίση*. Συνεπώς, αν επιθυμούμε να κάνουμε κατάβαση της επιφάνειας τότε θα πρέπει να ακολουθήσουμε πορεία *αντίθετη της βαθμίδας*. Σε ακριβώς αυτό έγκειται ο αλγόριθμος του gradient descent, δηλαδή στην επικαιροποίηση ενός τυχαίου σημείου της επιφάνειας με βήμα ανάλογο της βαθμίδας.

Ας υποθέσουμε λοιπόν ότι μας δίνεται ένα συνεκτικό ανοιχτό χωρίο $D \subset \mathbb{R}^d$ και μία πραγματική συνάρτηση $f \in C^1(D)$, η οποία θα αποκαλείται *αντικειμενική συνάρτηση* (objective function)

$$f : D \to \mathbb{R}$$

την οποία επιθυμούμε να ελαχιστοποιήσουμε. Ψάχνουμε δηλαδή ένα $\mathbf{x}^*$, τέτοιο ώστε:

$$\mathbf{x}^* = \underset{\mathbf{x}\in D}{\operatorname{argmin}}\, f(\mathbf{x})$$

Προφανώς επίσης θα ισχύει

$$\nabla f(\mathbf{x}^*) = 0$$

Ο αλγόριθμος απαιτεί να του ορίσουμε το αρχικό σημείο πάνω στην επιφάνεια από το οποίο θα αρχίσει η κατάβαση (ή η ανάβαση), το *βήμα επικαιροποίησης* t_k και ένα *όριο ανοχής* (tolerance) το οποίο θα καθορίσει πότε θα σταματήσει ο αλγόριθμος. Συνοπτικά ο αλγόριθμος gradient descent φαίνεται στον παρακάτω πίνακα.

Αλγόριθμος Gradient Descent
1: Αρχικοποίηση τυχαίου σημείου $\mathbf{x}^{(0)}$, βήματος t_k και tolerance ϵ
2: **while** $\|\|\nabla f(\mathbf{x}^{(k)})\|\| \geq \epsilon$ **do** :
3: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)})$
4: $k \leftarrow k + 1$
5: **end while**
6: **return** $\mathbf{x}^{(k)}$

Table 1: Αλγόριθμος Gradient Descent

Μερικά σχόλια για τον αλγόριθμο και τις παραμέτρους του. Καταρχάς ο παραπάνω αλγόριθμος αναφέρεται σε προβλήματα βελτιστοποίησης *χωρίς περιορισμούς*, δηλαδή θεωρούμε ότι όλα τα σημεία του χώρου που ορίζονται από την επιφάνεια που ορίζει η αντικειμενική συνάρτηση είναι προσβάσιμα. Ο αλγόριθμος τροποποιείται για προβλήματα που περιέχουν περιορισμούς όμως έχει την ίδια επιτυχία.

Στο τρίτο βήμα του αλγορίθμου παρατηρούμε ότι χρησιμοποιούμε το πρόσημο "-" για να κάνουμε επικαιροποίηση του σημείου της επιφάνειας. Ο λόγος που επιλέγουμε αυτό το πρόσημο είναι ακριβώς επειδή θέλουμε να ακολουθήσουμε μία πορεία που είναι *αντίθετη* της βαθμίδας, με σκοπό να οδηγηθούμε στο ελάχιστο της συνάρτησης. Αν, αντιθέτως, χρησιμοποιούσαμε το πρόσημο "+" τότε θα ακολουθούσαμε την πορεία της ίδιας της βαθμίδας και θα πηγαίναμε προς το μέγιστο της συνάρτησης. Στην περίπτωση αυτή ο αλγόριθμος λέγεται *ανάβαση βαθμίδας* (Gradient Ascent).

Το βήμα επικαιροποίησης t_k μπορεί να είναι είτε ένας σταθερός αριθμός είτε να εξαρτάται από το βήμα στο οποίο βρισκόμαστε. Η δεύτερη περίπτωση είναι πιο αποδοτική αν το βήμα επιλεγεί με έναν σχετικά έξυπνο τρόπο. Το κόστος φυσικά είναι αρκετά μεγαλύτεροι υπολογιστικοί χρόνοι. Η λογική πίσω από την επιλογή του βήματος είναι η σκέψη ότι θέλουμε να ελαχιστοποιεί την αντικειμενική μας συνάρτηση σε μια περιοχή του βήματος στο οποίο πρόκειται να βρεθούμε μετά από το εκάστοτε βήμα. Σε μαθηματικούς όρους, το *βέλτιστο βήμα* t_k^* θα ικανοποιεί [9]:

$$t_k^* = \underset{t \geq 0}{\operatorname{argmin}} f(\mathbf{x}^{(k+1)}) = \underset{t \geq 0}{\operatorname{argmin}} f(\mathbf{x}^{(k)} - t \nabla f(\mathbf{x}^{(k)})$$

Φυσικά είναι εξαιρετικά λίγες οι περιπτώσεις όπου το βήμα αυτό μπορεί να υπολογισθεί αναλυτικά μιας και στις περισσότερες εφαρμογές οι αντικειμενικές συναρτήσεις είναι μέχρι και δεκάδων μεταβλητών.

Το όριο ϵ είναι το κριτήριο τερματισμού του αλγορίθμου και είναι ένας πραγματικός αριθμός πολύ μικρότερος της μονάδας και φυσικά θετικός. Όταν η νόρμα (δηλαδή το μέγεθος) της βαθμίδας, $||\nabla f||$, πέσει κάτω από ϵ μετά την κ-οστή επανάληψη τότε αυτό σημαίνει ότι σε περιοχή του σημείου $\mathbf{x}^{(k)}$ η κλίση της συνάρτηση είναι σχεδόν μηδενική και άρα ως γνωστόν θα είμαστε πολύ κοντά στο ακρότατο. Θεωρητικά μπορούμε να προσεγγίσουμε το ακρότατο αυτό όσο καλά θέλουμε μικραίνοντας όλο και πιο πολύ το ϵ.

Φυσικά, επειδή καταφέραμε να βρούμε έναν τρόπο να προσεγγίσουμε ένα τοπικό ακρότατο μίας συνάρτησης δεν σημαίνει ότι μπορούμε πάντα να εντοπίσουμε το *ολικό ελάχιστο* μιας συνάρτησης. Στην πραγματικότητα, ο προσδιορισμός του ολικού ελαχίστου μια πολυμεταβλητής συνάρτησης είναι ένα υπέρμετρα δύσκολο μαθηματικό πρόβλημα. Για τον λόγο αυτό στις περισσότερες εφαρμογές η εύρεση ενός τοπικού ακροτάτου της αντικειμενικής μας συνάρτησης θεωρείται επιτυχία. Επίσης κρίνεται σκόπιμο να αναφερθεί ότι στην περίπτωση που η αντικειμενική μας συνάρτηση είναι κυρτή (αντ. κοίλη) τότε γνωρίζουμε από τον μαθηματικό λογισμό ότι θα παρουσιάζει μοναδικό ελάχιστο (αντ. μέγιστο). Συνεπώς στις περιπτώσεις

αυτές ο αλγόριθμος gradient descent μας εγγυάται σύγκλιση αφού η βαθμίδα πάντα θα οδηγεί προς το μέγιστο ή το ελάχιστο

2.10 Σύγκλιση του αλγορίθμου Gradient Descent

Ένα εύλογο ερώτημα που προκύπτει είναι κατά πόσο ο αλγόριθμος gradient descent δουλεύει, δηλαδή κατά πόσο συγκλίνει. Η απάντηση δεν είναι απλή και χρειαζόμαστε, μεταξύ άλλων, τον παρακάτω ορισμό:

Ορισμός: Έστω ένα κυρτό σύνολο $D \subset \mathbb{R}^d$, μία κυρτή συνάρτηση $f : D \to \mathbb{R}$. Η f θα λέγεται **ισχυρώς κυρτή** (strongly convex) αν υπάρχει θετικός πραγματικός αριθμός m τέτοιος ώστε για κάθε $x, y \in D$:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m \parallel x - y \parallel^2$$

όπου $\| \cdot \|$ μία οποιαδήποτε νόρμα. Αποδεικνύεται [10] ότι η παραπάνω ανισότητα ισοδυναμεί με την ύπαρξη θετικών πραγματικών αριθμών m και M για τους οποίους ισχύει:

$$mI \leq \mathcal{H}_f(\mathbf{x}) \leq MI$$

όπου $\mathcal{H}_f(\mathbf{x})$ ο εσσιανός πίνακας της f στο $\mathbf{x}$ και I ο μοναδιαίος πίνακας. Σημειώνεται ότι για δύο πίνακες A και B ίδιων διαστάσεων $A \geq B$ σημαίνει ότι ο πίνακας A - B είναι θετικά ημι-ορισμένος. Μπορούμε έυκολα να δείξουμε ότι μία ισχυρώς κυρτή συνάρτηση είναι και αυστηρά κυρτή [10]. Παίρνοντας το ανάπτυγμα Taylor δεύτερης τάξης για την f έχουμε ότι για $x, y \in D$ μπορούμε να βρούμε ένα $z \in [x, y]$ τέτοιο ώστε:

$$f(y) = f(x) + \nabla f(x)^\intercal (y - x) + \frac{1}{2}(y - x)^\intercal \mathcal{H}_f (y - x)$$

και λόγω της παραπάνω διπλής ανισότητα παίρνουμε:

$$f(y) \geq f(x) + \nabla f(x)^\intercal (y - x) + \frac{m}{2} \parallel y - x \parallel^2$$

και επειδή $\frac{m}{2} \parallel y - x \parallel^2 > 0$:

$$f(y) > f(x) + \nabla f(x)^\intercal (y - x)$$

Έχοντας τώρα την ιδιότητα της ισχυρής κυρτότητας παίρνουμε το εξής θεώρημα [9]:

Θεώρημα: Έστω $f : D \to \mathbb{R}$ μία ισχυρώς κυρτή συνάρτηση με παραμέτρους m και M, όπως και στην παραπάνω διπλή ανισότητα, και $\alpha^* = \min_{\mathbf{x} \in D} f(\mathbf{x})$. Τότε για κάθε $\epsilon > 0$ μπορούμε να πετύχουμε $f(\mathbf{x}^{(k^*)}) - \alpha^* \leq \epsilon$ μετά από k^* επαναλήψεις για κάποιο κ^* που ικανοποιεί:

$$k^* \geq \frac{\log\left(\frac{f(\mathbf{x}^{(0)}) - \alpha^*}{\epsilon}\right)}{\log\left(\frac{1}{1 - \frac{m}{M}}\right)}$$

Το παραπάνω θεώρημα μας εξασφαλίζει ότι μετά από έναν αριθμό επαναλήψεων που εξαρτάται από την φύση της ίδιας της συνάρτησης (τις σταθερές m και M) αλλά και από την ακρίβεια που επιθυμούμε μπορούμε να φτάσουμε όσο κοντά θέλουμε στο ακρότατο της f. Προφανώς όταν $\epsilon \to 0$ τότε $k^* \to +\infty$. Επίσης παρατηρούμε ότι καθώς $m \to M$ απαιτούνται όλο και λιγότερες επαναλήψεις για να συγκλίνει ο αλγόριθμος.

Προφανώς ο ρόλος του αλγορίθμου του gradient descent στα πλαίσια των νευρωνικών δικτύων είναι η ελαχιστοποίηση των συναρτήσεων κόστους ως προς τις παραμέτρους του δικτύου, δηλαδή τα βάρη και τα biases. Συνεπώς ένα κρίσιμο βήμα της εκπαίδευσης είναι ο υπολογισμός των βαθμίδων της συνάρτησης κόστους ως προς όλες αυτές τις παραμέτρους ξεχωριστά. Στο σημείο αυτό όμως, έχοντας αρχίσει να εφαρμόζουμε τον αλγόριθμο gradient descent, παρατηρούμε ότι παρουσιάζεται ένα μείζονος σημασίας πρόβλημα και αυτό είναι οι τεράστιοι υπολογιστικοί χρόνοι που απαιτούνται για να τερματίσει ο αλγόριθμος. Όπως

είδαμε και στην ενότητα **4.6** μία τυπική συνάρτηση κόστους έχει την γενική μορφή:

$$\mathbf{J}(\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{\vartheta})$$

όπου $\mathcal{L}$ μία συνάρτηση σφάλματος (π.χ. η διαφορά των τετραγώνων) και $\boldsymbol{\vartheta}$ το διάνυσμα των παραμέτρων ως προς τις οποίες επιθυμούμε να γίνει η βελτιστοποίηση, δηλαδή η εύρεση των σημείων μηδενισμού της βαθμίδας

$$\nabla_{\boldsymbol{\vartheta}}\mathbf{J}(\boldsymbol{\vartheta}) \ = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\vartheta}}\mathcal{L}(\boldsymbol{\vartheta})$$

Ο υπολογιστικός χρόνος (ή υπολογιστικό κόστος) της παραπάνω πράξης είναι $\mathcal{O}(n)$. Συνεπώς όταν ο αριθμός των δεδομένων, n, γίνεται πάρα πολύ μεγάλος και ειδικά αν το πρόβλημά μας αφορά μη-κυρτή βελτιστοποίηση τότε η εφαρμογή του αλγορίθμου gradient descent είναι απολύτως μη-πρακτική. Φαίνεται λοιπόν ότι χρειαζόμαστε μία τροποποίηση προκειμένου να είμαστε σε θέση να λύσουμε αποδοτικά ένα πρόβλημα μηχανικής μάθησης.

2.11 Στοχαστικός Αλγόριθμος Gradient Descent και Mini-Batch Gradient Descent

Ο αλγόριθμος του στοχαστικού gradient descent είναι μία επέκταση του κανονικού αλγορίθμου που περιγράφηκε στην ενότητα **4.8**. Η ανάγκη της ανάπτυξης του εν λόγω αλγορίθμου έγκειται στις δυσκολίες που παρουσιάζονται στην εφαρμογή του απλού gradient descent όταν ο αριθμός των δεδομένων εκπαίδευσης είναι πάρα πολύ μεγάλος. Πιο συγκεκριμένα, όπως είδαμε στην ενότητα **4.6**, οι συναρτήσεις κόστους αποτελούν αθροίσματα (τετραγωνικών ή και άλλου είδους) σφαλμάτων μεταξύ των δεδομένων εξόδου και των προβλέψεών μας. Όταν έχουμε στην διάθεσή μας έναν υπερβολικά μεγάλο αριθμό δεδομένων (της τάξεως των εκατομμυρίων) τότε προφανώς ένα μονάχα βήμα του αλγορίθμου gradient descent απαιτεί πάρα πολύ μεγάλο υπολογιστικό χρόνο.

Η καρδιά του στοχαστικού gradient descent είναι η εφαρμογή του κλασσικού gradient descent με την παραλλαγή ότι σε κάθε επανάληψη θα υπολογίζουμε την βαθμίδα *για ένα μόνο δεδομένο*. Με άλλα λόγια, στην i-οστή επανάληψη του αλγορίθμου, αντί να υπολογίσουμε την βαθμίδα $\sum_{i=1}^{n}\nabla_{\theta}\mathcal{L}_i$ θα πρέπει μόνο να υπολογίσουμε την βαθμίδα $\nabla_{\theta}\mathcal{L}_i$. Με την μέθοδο αυτή ο υπολογιστικός χρόνος μειώνεται σημάντικα αφού σε κάθε επανάληψη θα πρέπει να υπολογίσουμε πολύ πιο απλές βαθμίδες σε σχέση με την κλασσική έκδοση του αλγορίθμου.

Μία μικρή παραλλαγή του στοχαστικού gradient descent είναι ο λεγόμενος αλγόριθμος *mini-batch gradient descent*. Στην περίπτωση αυτή κάνουμε μία τυχαία επιλογή *παρτίδων* (mini-batches) σταθερού μεγέθους $|\mathcal{B}|$ από τα δεδομένα μας και εφαρμόζουμε τον αλγόριθμο gradient descent σε αυτά. Σε κάθε επανάληψη δηλαδή θα πρέπει να υπολογισθεί η βαθμίδα

$$\nabla_{\boldsymbol{\vartheta}}\mathbf{J}(\boldsymbol{\vartheta}) \ = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\nabla_{\boldsymbol{\vartheta}}\mathcal{L}(\boldsymbol{\vartheta})_i$$

Ο υπολογιστικός χρόνος σε αυτήν την περίπτωση είναι $\mathcal{O}(\mathcal{B})$.

2.12 Προς-τα-πίσω Διάδοση (Back Propagation)

Έχοντας τώρα τον αλγόριθμο που θα χρησιμοποιήσουμε (gradient descent) για την ελαχιστοποίηση του κόστους θα πρέπει να σκεφτούμε πώς θα υπολογίσουμε τις μερικές παραγώγους (βαθμίδες) που απαιτεί. Η μεθοδολογία υπολογισμού αυτών των μερικών παραγώγων ονομάζεται *προς-τα-πίσω διάδοση ή back-propagation* και επιτρέπει την ροή πληροφορίας που παίρνουμε από τη συνάρτηση κόστους *προς τα πίσω* κατά μήκος του νευρωνικού δικτύου. Η ροή αυτή μοντελοποιείται με βάση τον γνώστο από τον

Μαθηματικό Λογισμό *κανόνα της αλυσίδας*. Έστω $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, $g : \mathbb{R}^m \to \mathbb{R}^n$ και $f : \mathbb{R}^n \to \mathbb{R}$. Απαιτούμε $g \in C^1(\mathbb{R}^m)$ και $f \in C^1(\mathbb{R})^n$. Αν $\mathbf{y} = g(\mathbf{x})$ και $z = f(\mathbf{y})$ τότε:

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j}\frac{\partial y_j}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j}\mathcal{J}_{ij}$$

όπου $\mathcal{J}_{ij}$ ο *Ιακωβιανός πίνακας* της g. Σε διανυσματική μορφή ο κανόνας της αλυσίδας γράφεται ως

$$\nabla_{\mathbf{x}} z = \left(\frac{\partial \mathbf{y}^\intercal}{\partial \mathbf{x}}\right) \nabla_{\mathbf{y}} z$$

Δηλαδή η μερική παράγωγος ως προς μία μεταβλητή $\mathbf{x}$ μπορεί να υπολογισθεί πολλαπλασιάζοντας τον Ιακωβιανό πίνακα $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ με την βαθμίδα $\nabla_{\mathbf{y}} z$. Ο αλγόριθμος back-propagaton δεν εφαρμόζεται καθαρά και μόνο σε διανύσματα αλλά και γενικά σε τανυστές οποιασδήποτε διάστασης. Πριν συνεχίζουμε με την περιγραφή του αλγορίθμου back-propagation στα πλαίσια των συμβολισμών ενός νευρωνικού δικτύου θα πρέπει να εισάγουμε έναν νέο συμβολισμό, το *γινόμενο Hadamar*.

Ορισμός: Για δύο πίνακες $\mathbf{A}$ και $\mathbf{B}$ ίδιας διάστασης $m \times n$, το *γινόμενο Hadamar* $\mathbf{A} \odot \mathbf{B}$ είναι ένας πίνακας ίδιας διάστασης με στοιχεία που δίνονται από την σχέση

$$(\mathbf{A} \odot \mathbf{B})_{ij} = (\mathbf{A})_{ij}(\mathbf{B})_{ij}$$

Πρόκειται δηλαδή για πολλαπλασιασμό ανά στοιχείο. Για παράδειγμα για δύο 2×2 πίνακες το γίνομενο Hadamar ορίζεται ως

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{bmatrix} \tag{14}$$

Το γινόμενο Hadamar είναι αντιμεταθετικό, προσεταιριστικό και επιμεριστικό με πράξη την πρόσθεση πινάκων.

Ο αλγόριθμος back-propagation μας δείχνει πώς αλλάζοντας τα βάρη και τα biases ενός νευρωνικού δικτύου αλλάζει η συνάρτηση κόστους. Αυτό εν τέλει σημαίνει προφανώς πως θα πρέπει να υπολογίσουμε τις μερικές παραγώγους $\frac{\partial C}{\partial w^l_{jk}}$ και $\frac{\partial C}{\partial b^l_j}$. Για να υπολογίσουμε αυτές τις παραγώγους θα πρέπει να κάνουμε άλλον έναν ενδιάμεσο υπολογισμό. Ο υπολογισμός αυτός περιλαμβάνει το λεγόμενο *σφάλμα* (error), δ^l_j. Η ποσότητα αυτή αποτελεί το σφάλμα στον j-οστό νευρώνα του l-οστού στρώματος. Ο αλγόριθμος back-propagation μας δίνει ακριβώς έναν τρόπο για να υπολογίσουμε το σφάλμα δ^l_j, το οποίο σχετίζεται άμεσα με τις μερικές παραγώγους $\frac{\partial C}{\partial w^l_{jk}}$ και $\frac{\partial C}{\partial b^l_j}$.

Ας υποθέσουμε ότι z^l_j είναι η σταθμησμένη είσοδος (weighted input) του j-οστού νευρώνα του l-οστού στρώματος. Τότε το σφάλμα ορίζεται ως

$$\delta^l_j \equiv \frac{\partial C}{\partial z^l_j}$$

Με βάση τις γνωστές συμβάσεις θα συμβολίσουμε με δ^l το διάνυσμα των σφαλμάτων που σχετίζεται με το στρώμα l. Η εξίσωση που δίνει το σφάλμα για το στρώμα εξόδου είναι η εξής:

$$\delta^L_j = \frac{\partial C}{\partial \alpha^L_j}\sigma'(z^L_j) \tag{15}$$

Ο πρώτος όρος στο δεξί μέλος, $\frac{\partial C}{\partial \alpha^L_j}$, μετράει πόσο γρήγορα αλλάζει η συνάρτηση κόστους ως συνάρτηση της j-οστής ενεργοποίησης εξόδου. Ο δεύτερος όρος , $\sigma'(z^L_j)$ μετράει πόσο γρήγορα η συνάρτηση ενεργοποίησης σαλλάζει στο z^L_j. Η έκφραση αυτή δεν είναι παρά ο κάνονας της αλυσίδας. Πράγματι

$$\delta^L_j = \frac{\partial C}{\partial z^L_j} = \frac{\partial C}{\partial \alpha^L_j}\frac{\partial \alpha^L_j}{\partial z^L_j}$$

Όμως $\alpha_j^L = \sigma(z_j^L)$ και άρα $\frac{\partial \alpha_j^L}{\partial z_j^L} = \sigma'(z_j^L)$. Έτσι παίρνουμε την παραπάνω έκφραση. Η εξίσωση (5.0.2) μπορεί να γραφτεί σε μία πιο συμπυκνωμένη μορφή πινάκων ως εξής

$$\delta^L = \nabla_\alpha C \odot \sigma'(z^L) \tag{16}$$

όπου $\nabla_\alpha C$ είναι το διάνυσμα του οποίου οι συνιστώσες είναι οι μερικές παράγωγοι $\frac{\partial C}{\partial \alpha_j^L}$. Εκφράζει τον ρυθμό μεταβολής της συνάρτησης C ως προς τις τιμές ενεργοποίησης εξόδου. Παρακάτω δίνουμε την σχέση που εκφράζει το σφάλμα δ^l συναρτήσει του σφάλματος του επόμενου στρώματος δ^{l+1}:

$$\delta^l = \left[(w^{l+1})^\intercal \delta^{l+1}\right] \odot \sigma'(z^l) \tag{17}$$

όπου $(w^{l+1})^\intercal$ είναι ο ανάστροφος πίνακας του πίνακα βαρών w^{l+1} του $(l+1)$-οστού στρώματος. Μία ίσως διαισθητική εξήγηση της παραπάνω εξίσωσης είναι η εξής [6]: Ας υποθέσουμε ότι γνωρίζουμε το σφάλμα δ^{l+1}. Τότε πολλαπλασιάζοντας από τα αριστερά με τον πίνακα w^{l+1} μεταφέρουμε το σφάλμα προς τα πίσω κατά μήκος του δικτύου, παίρνοντας έτσι μία εκτίμηση του σφάλματος της εξόδου του l-οστού στρώματος. Έπειτα παίρνουμε το γινόμενο Hadamar $\odot\sigma'(z^l)$. Η πράξη αυτή μεταφέρει το σφάλμα κι άλλο προς τα πίσω δια μέσου της συνάρτησης ενεργοποίησης του l-οστού στρώματος, δίνοντάς μας έτσι το σφάλμα δ^l της σταθμησμένης εισόδου του l-οστού στρώματος.

Με την βοήθεια των εξισώσεων (4.11.2) και (4.11.4) μπορούμε να υπολογίσουμε το σφάλμα δ^l για κάθε στρώμα του δικτύου ξεκινώντας από το στρώμα εξόδου του δικτύου και ταξιδεύοντας προς τα πίσω, προς το στρώμα εισόδου του δικτύου. Εξ'ου και το όνομα *προς-τα-πίσω* διάδοση. Στο σημείο αυτό είμαστε έτοιμοι τώρα να δούμε πως θα υπολογίσουμε τις μερικές παραγώγους-στόχους, δηλαδή τις μερικές παραγώγους ως προς τις παραμέτρους του δικτύου, δηλαδή τα βάρη και τα biases. Ξεκινάμε δίνοντας την εκπληκτικά απλή εξίσωση που δίνει την μερική παράγωγο της συνάρτησης κόστους ως προς τα βάρη του δικτύου

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \tag{18}$$

Η μερική παράγωγος ως προς τα βάρη είναι, προς μεγάλη μας έκπληξη και χαρά το σφάλμα, για το οποίο έχουμε ήδη βρει τρόπο υπολογισμού. Αν δεν υπάρχει σύγχηση ως προς τους δείκτες μπορούμε να γράψουμε την παραπάνω εξίσωση σε μία πιο συμπυκνωμένη μορφή

$$\frac{\partial C}{\partial b} = \delta \tag{19}$$

όπου φυσικά το σφάλμα δ υπολογίζεται στον ίδιο νευρώνα με το bias b. Συνεχίζουμε με την εξίσωση που δίνει την μερική παράγωγο της συνάρτηση κόστους ως προς ένα βάρος του δικτύου

$$\frac{\partial C}{\partial w_{jk}^l} = \alpha_k^{l-1} \delta_j^l \tag{20}$$

Φυσικά γνωρίζοντας ήδη πως να υπολογίσουμε τις ποσότητες α_k^{l-1} και δ_j^l μπορούμε εύκολα να υπολογίσουμε και την μερική παράγωγο $\frac{\partial C}{\partial w_{jk}^l}$. Και πάλι, μπορούμε να γράψουμε την παραπάνω εξίσωση στην πιο συμπυκνωμένη και πιο διαισθητική μορφή

$$\frac{\partial C}{\partial w} = \alpha_{in} \delta_{out} \tag{21}$$

όπου φυσικά η α_{in} είναι η ενεργοποίηση του νευρώνα που έχει σαν είσοδο το βάρος w και δ_{out} είναι το σφάλμα του νευρώνα που έχει σαν έξοδο το βάρος w. Παρατηρούμε ότι όταν η τιμή α_{in} είναι πολύ μικρή, δηλαδή $\alpha_{in} \approx 0$, η μερική παράγωγος $\frac{\partial C}{\partial w}$ θα είναι εξίσου μικρή. Τότε λέμε ότι αυτό το βάρος *μαθαίνει αργά*, που σημαίνει ότι δεν αλλάζει πολύ κατά την διάρκεια του gradient descent. Στον παρακάτω πίνακα παρουσιάζονται συγκεντρωμένες οι εξισώσεις που συνιστούν την διαδικασία του back-propagation [6]:

Οι εξισώσεις του αλγορίθμου Back-Propagation,

$$\delta^L = \nabla_\alpha C \odot \sigma'(z^L) \tag{1}$$

$$\delta^l = \left[(w^{l+1})^\intercal \delta^{l+1}\right] \odot \sigma'(z^l) \tag{2}$$

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j \tag{3}$$

$$\frac{\partial C}{\partial w^l_{jk}} = \alpha^{l-1}_k \delta^l_j \tag{4}$$

Οι εξισώσεις αυτές φυσικά αποδεικνύονται με αυστηρό μαθηματικό τρόπο. Οι αποδείξεις είναι απλές και αποτελούν απλή και προσεχτική εφαρμογή του κανόνα της αλυσίδας. Οι αποδείξεις παρατίθενται στο παράρτημα.

Έχοντας τώρα περιγράψει τόσο την διαδικασία της προώθησης όσο και της προς-τα-πίσω διάδοσης, δηλαδή μία πλήρη εποχή της εκπαίδευσης ενός νευρωνικού δικτύου, μπορούμε εύκολα να την συνοψίσουμε σέ έναν μικρό αλγόριθμο. Στον παρακάτω πίνακα δίνεται συνοπτικά ο συνολικός αλγόριθμος της προώθησης και της προς-τα-πίσω διάδοσης [6].

Αλγόριθμος Προώθησης και Προς-τα-πίσω Διάδοσης
1. **Διάνυσμα Εισόδου x**: Υπολογίζουμε το αντίστοιχο διάνυσμα ενεργοποιήσεων α^1 του στρώματος εισόδου
2. **Προώθηση**: Για κάθε $n = 1, 2, \cdots, L$ υπολογίζουμε τις ποσότητες $z^n = w^n \alpha^{n-1} + b^n$ και $\alpha^n = \sigma(z^n)$
3. **Σφάλμα Εξόδου** δ^L: Υπολογίζουμε το διάνυσμα $\delta^L = \nabla_\alpha C \odot \sigma'(z^L)$
4. **Προς-τα-πίσω διάδοση του σφάλματος**: Για κάθε $l = L-1, L-2, \cdots, 1$ υπολογίζουμε το σφάλμα $\delta^l = \left[(w^{l+1})^\intercal \delta^{l+1}\right] \odot \sigma'(z^l)$
5. **Έξοδος**: Υπολογίζουμε τις μερικές παραγώγους της συνάρτησης κόστους από τις σχέσεις $\frac{\partial C}{\partial b^l_j} = \delta^l_j$ και $\frac{\partial C}{\partial w^l_{jk}} = \alpha^{l-1}_k \delta^l_j$

Table 2: Αλγόριθμος Feed-Forward και Back-Propagation

Θα λήξουμε το κεφάλαιο με ένα πολύ σημαντικό θεώρημα της επιστήμης των νευρωνικών δικτύων που μας δείχνει και την ευρωστία και τις ικανότητες εν γένει της αρχιτεκτονικής των νευρωνικών δικτύων να δρουν ως συναρτησιακοί προσεγγιστές.

2.13 Το Θεώρημα Καθολικής Προσέγγισης (Universal Approximation Theorem)

Ένας τεράστιος αριθμός σύγχρονων τεχνολογικών και επιστημονικών εφαρμογών βασίζονται στην ικανότητα των νευρωνικών δικτύων να συσχετίζουν σχεδόν κάθε είδους δεδομένων εισόδου με μία επιθυμητή έξοδο, δηλαδή στο να βρίσκουν μία συνάρτηση που να τα συσχετίζει. Πιο σωστά, τα νευρωνικά δίκτυα δρούν ως *συναρτησιακοί προσεγγιστές* (function approximators), με την έννοια ότι μπορούν να "πλησιάσουν" μια επιθυμητή συνάρτηση με πολύ μεγάλη ακρίβεια. Πώς όμως μπορεί κανείς να είναι σίγουρος ότι η αρχιτεκτονική των νευρωνικών δικτύων θα είναι όντως ικανή να επιλύσει το εκάστοτε πρόβλημα; Στα πλαίσια των πραγματικών εφαρμογών οι συναρτήσεις που επιθυμούμε να προσεγγίσουμε μπορεί να είναι πολύ περίπλοκες ή ακόμη και να μην έχουν σαφή (explicit) μορφή.

Στο ερώτημα αυτό δίνει μία πολύ ικανοποιητική απάντηση το πανίσχυρο *Θεώρημα Καθολικής Προσέγγισης για Νευρωνικά Δίκτυα* (Universal Approximation Theorem for Neural Networks) [11]. Το θεώρημα αυτό έχει μία πολύ καθορισμένη και αυστηρή Μαθηματική θεμελίωση. Μας λέει πως μπορούμε να προσεγγίσουμε όσο καλά θέλουμε μια συνεχή συνάρτηση ορισμένη σε ένα συμπαγές σύνολο με ένα προωθητικό νευρωνικό δίκτυο που περιέχει *μονάχα ένα κρυφό στρώμα* το οποίο θα αποτελείται από *πεπερασμένους νευρώνες*. Η απόδειξη του θεωρήματος δεν είναι απλή και χρησιμοποιεί πολλά βαριά εργαλεία της συναρτησιακής ανάλυσης και της θεωρίας μέτρου, όπως το θεώρημα Hahn-Banach και το

Θεώρημα Κυρίαρχης Σύγκλισης. Παρόλα αυτά μας εξασφαλίζει την αδιαμφησβήτητη ικανότητα των νευρωνικών δικτύων να μας παρέχουν, με μία σχετικά απλή αρχιτεκτονική, οποιαδήποτε ομοιόμορφα συνεχή συνάρτηση θέλουμε. Πρίν δοθεί η πλήρης Μαθηματική περιγραφή του θεωρήματος θα χρειαστούμε μερικούς ορισμούς.

Ορισμός: Μία συνάρτηση σ: $\mathbb{R} \to \mathbb{R}$ θα λέγεται *σιγμοειδής* (sigmoidal) αν είναι γνήσια μονότονη και ικανοποιεί:

$$\sigma(t) = \begin{cases} \alpha & ,t \to +\infty \\ \beta & ,t \to -\infty \end{cases}$$

για κάποιους πραγματικούς αριθμούς α και β. Δηλαδή μία σιγμοειδής συνάρτηση είναι φραγμένη. Στις περισσότερες εφαρμογές και πολύ συνχά στην βιβλιογραφία οι σιγμοειδείς συναρτήσεις επιλέγονται με τέτοιο τρόπο ώστε να είναι κανονικοποιημένες, δηλαδή α = 1 και β = 0 ή α = 1 και β = -1.

Παραδείγματα σιγμοειδών συναρτήσεων αποτελούν οι εξής συναρτήσεις :

- *Λογιστική Συνάρτηση*, σ(x) = $\frac{1}{1+e^{-x}}$
- *Υπερβολική Εφαπτομένη* σ(x) = $\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
- *Συνάρτηση Σφάλματος* σ(x) = $\frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

Ορισμός:Έστω I_n ο μοναδιαίος n-διάστατος κύβος. Μία συνάρτηση σ θα λέγεται *μεροληπτική* (discriminatory) εάν ισχύει η συναπαγωγή:

$$\int_{I_n} \sigma(\mathbf{w}_i^{\intercal}\mathbf{x} + b_i) d\mu(x) = 0 \Rightarrow \mu(x) = 0 \tag{22}$$

$\forall$ $\mathbf{w}_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, $\mu \in M(I_n)$.

Η συνάρτηση μ(x) αποτελεί ένα κάπως τεχνικό εργαλείο και καλείται *μέτρο*. Είναι μια συνάρτηση που ορίζεται σε ένα σύνολο και μας επιτρέπει να αντιστοιχίσουμε σε κάθε υποσύνολο του συνόλου αυτού έναν θετικό πραγματικό αριθμό, ο οποίος διαισθητικά μπορεί να ερμηνευθεί ώς το *μέγεθος* του υποσυνόλου αυτού. Το σύνολο $M(I_n)$ μπορούμε να το σκεφτούμε ως ένα σύνολο στο οποίο έχει νοήμα να ορίσουμε συναρτήσεις-μέτρα. Είναι το σύνολο των μέτρων ορισμένα στον n-διάστατο μοναδιαίο κύβο. Σε κάθε πιθανό υποσύνολο του κύβου αυτού αντιστοιχούμε ένα μέτρο, το οποίο μας δίνει τον n-διάστατο όγκο του. Ο παραπάνω ορισμός μας λέει πως μία μεροληπτική συνάρτηση μπορεί να δράσει *μη-καταστρεπτικά* σε έναν γραμμικό συνδυασμό πραγματικών αριθμών με εξαίρεση ένα υποσύνολο που έχει μέτρο ίσο με το μηδέν (μ= 0). Διαισθητικά, ένα σύνολο έχει μέτρο ίσο με το μηδέν όταν ο πληθάριθμός του είναι πεπερασμένος. Συνεπώς μια μεροληπτική συνάρτηση επιτρέπει στην πληροφορία να περάσει από τον έναν νευρώνα στον επόμενο χωρίς να χαθεί η πληροφορία της εισόδου.

Έστω A $\subset \mathbb{R}$ ένα συμπαγές σύνολο (κλειστό και φραγμένο). Θα συμβολίζουμε το σύνολο των συνεχών, μεροληπτικών και μη γραμμικών σιγμοειδών συναρτήσεων ορισμένες στο A με παραμέτρους α και β ώς **Sig(A,α,β)**. Επειδή το A είναι συμπαγές και οι συναρτήσεις συνεχείς σημαίνει ότι το Sig(A,α,β) αποτελείται από ομοιόμορφα συνεχείς -και άρα ολοκληρώσιμες- συναρτήσεις.

Παρακάτω δίνουμε έναν μαθηματικό ορισμό ενός νευρωνικού δικτύου με ένα μόνο κρυφό στρώμα:

Ορισμός: Ένα *νευρωνικό δίκτυο* N νευρώνων (ή κόμβων) διατεταγμένων σε ένα μόνο κρυφό στρώμα είναι μία συνάρτηση $\Psi : \mathbb{R}^n \to \mathbb{R}$ που δίνεται από την σχέση:

$$\Psi(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i \sigma(\mathbf{w}_i^{\intercal}\mathbf{x} + b_i) \tag{23}$$

όπου $\mathbf{w}_i, \mathbf{x} \in \mathbb{R}^n, \lambda_i, b_i \in \mathbb{R}$ και σ∈ Sig(I_n,α,β). Το σύνολο των νευρωνικών δικτύων της παραπάνω μορφής που αποτελούνται από ένα κρυφό στρώμα θα συμβολίζεται με $\aleph_1$. Ο όρος που υπάρχει μέσα στο άθροισμα του παραπάνω ορισμού μας θυμίζει φυσικά την *έξοδο ενός νευρώνα*, όπως παρουσιάσθηκε στην ενότητα **4.4**. Η παραπάνω συνάρτηση Ψ λοιπόν δεν είναι παρά ένας γραμμικός συνδυασμός των σημάτων εξόδου όλων των νευρώνων του μοναδικού κρυφού στρώματος του δικτύου. Το άθροισμα αυτό αποτελεί φυσικά την έξοδο του δικτύου.

Θεώρημα (Καθολικής Προσέγγισης για Νευρωνικά Δίκτυα) [11]: Έστω $\sigma : \mathbb{R} \to \mathbb{R}$ με σ∈ Sig(I_n,α,β) και C(I_n) το σύνολο των συνεχών συναρτήσεων στο I_n (ή σε οποιοδήποτε συμπαγές υποσύνολο του $\mathbb{R}^n$. Τότε το σύνολο $\aleph_1$ είναι *πυκνό* στο C(I_n). Δηλαδή για κάθε $\epsilon > 0$ και κάθε συνάρτηση $f \in \mathrm{C}(I_n)$ υπάρχει ένα νευρωνικό δίκτυο της μορφής (4.12.2) τέτοιο ώστε:

$$|\Psi(x) - f(x)| < \epsilon$$

$\forall x \in I_n$.

Από το παραπάνω κεντρικό θεώρημα φαίνεται ότι η αρχιτεκτονική των νευρωνικών δικτύων είναι παραπάνω από ικανή για να δράσει ώς συναρτησιακός προσεγγιστής. Ένα δίκτυο με ένα μόνο κρυφό στρώμα αρκεί για να μας εξασφαλίσει όλες τις συνεχείς και φραγμένες συναρτήσεις, πόσο μάλλον ένα "βαθύτερο" δίκτυο με παραπάνω κρυφά στρώματα.

Στο σημείο αυτό έχουμε ολοκληρώσει το μεγάλο αυτό κεφάλαιο του Deep Learning. Όπως είναι προφανές η εκπαίδευση ενός τεχνητού νευρωνικού δικτύου δεν είναι εύκολη υπόθεση και αποτελεί ένα πρόβλημα μη-κυρτής βελτιστοποίησης που όμως βρίσκει πάρα πολλές εφαρμογές στην σύγχρονη τεχνολογία. Ακολουθεί ένα κεφάλαιο εξίσου σημαντικό, αυτό της *ενισχυτικής μάθησης*. Είναι μία από τις δύο θεωρίες (μαζί με το Deep Learning) που όταν συνδυαστούν μας δίνουν ένα πολύ χρήσιμο εργαλείο της τεχνητής νοημοσύνης, την λεγόμενη *ενισχυτική μάθηση βάθους*.

3 Συνελικτικά Νευρωνικά Δίκτυα και Βαθιά Μηχανική Μάθηση

Τα συνελικτικά νευρωνικά δίκτυα, είναι ένα εξειδικευμένο είδος νευρωνικού δικτύου για την επεξεργασία δεδομένων που έχει μια γνωστή τοπολογία σαν το πλέγμα. Παραδείγματα περιλαμβάνουν δεδομένα χρονοσειράς, τα οποία μπορούν να θεωρηθούν ως ένα μονοδιάστασο πλέγμα λαμβάνοντας δείγματα σε κανονικά χρονικά διαστήματα ή δεδομένα εικόνας, τα οποία μπορούν να θεωρηθούν ως ένα δισδιάστατο πλέγμα pixel, αλλά ακόμα και τρισδιάστατα πλέγματα όπως ακτινογραφίες. Το όνομα “συνελικτικά νευρωνικά δίκτυα” υποδηλώνει ότι το δίκτυο χρησιμοποιεί μια μαθηματική λειτουργία που ονομάζεται συνέλιξη.

Η συνέλιξη είναι ένα εξειδικευμένο είδος γραμμικής λειτουργίας. Τα συνελικτικά δίκτυα είναι απλά νευρωνικά δίκτυα που χρησιμοποιούν σύμπτυξη στη θέση γενικής μήτρας πολλαπλασιασμού σε τουλάχιστον ένα από τα στρώματά τους. Σε αυτό το κεφάλαιο, περιγράφουμε καταρχάς ποια είναι η συνέλιξη. Στη συνέχεια, εξηγούμε το κίνητρο πίσω από τη χρήση της συνέλιξης σε ένα νευρωνικό δίκτυο. Στη συνέχεια, περιγράφουμε τη λειτουργία pooling, την οποία χρησιμοποιούν σχεδόν όλα τα συνελικτικά δίκτυα. Συνήθως, η λειτουργία που χρησιμοποιείται σε ένα συνελικτικό νευρωνικό δίκτυο δεν αντιστοιχεί ακριβώς στον ορισμό της συνέλιξης όπως χρησιμοποιείται σε άλλους τομείς, όπως τη μηχανική ή τα μαθηματικά. Δείχνουμε επίσης πώς μπορεί να εφαρμοσθεί η μετατροπή σε πολλά είδη δεδομένων, με διάφορους αριθμούς διαστάσεων. Τα συνελικτικά δίκτυα αποτελούν παράδειγμα των αρχών νευροεπιστημών που επηρεάζουν τη βαθιά μάθηση. Συζητάμε αυτές τις αρχές νευροεπιστημών και στη συνέχεια ολοκληρώνουμε με σχόλια για το ρόλο που συνέβαλαν τα συνελικτικά δίκτυα στην ιστορία της βαθιάς μάθησης.

3.1 Συνελικτική διαδικασία

Η διακεκριμένη συνέλιξη μεταξύ δύο συναρτήσεων f και g ορίζεται ως

$$(f * g)(x) = \sum_t f(t)g(x + t) \tag{24}$$

Για δισδιάστατα σήματα, όπως εικόνες με τις οποίες και ασχολούμαστε, θεωρούμε τις δισδιάστατες συνελίξεις .

$$(K * I)(i, j) = \sum_{m,n} K(m, n)I(i + n, j + m) \tag{25}$$

Κ είναι ένας πυρήνας συνέλιξης που εφαρμόζεται σε μια εικόνα Ι. Κατά την συνέλιξη σύρεται ένα kernel / φίλτρο στην εικόνα. Σε κάθε θέση, έχουμε την περιστροφή μεταξύ του φίλτρου και του τμήματος της εικόνας που αντιμετωπίζεται. Στη συνέχεια, το φίλτρο μετακινείται με αριθμό s pixel, s καλείται η επικάλυψη / stride. Όταν το βήμα είναι μικρό, παίρνουμε περιττές πληροφορίες. Μερικές φορές, προσθέτουμε και ένα zero padding, το οποίο είναι ένα περιθώριο μεγέθους ρ που περιέχει μηδενικές τιμές γύρω από την εικόνα και το φίλτρο προκειμένου να ελέγξει το μέγεθος της εξόδου. Αν εφαρμοστεί kernels C0 (που ονομάζονται επίσης φίλτρα), κάθε καμπύλη k * k σε μια εικόνα.

Ο όγκος της εξόδου $W_i * H_i * C_i$ (W_iδηλώνει το εύρος, το H_i το ύψος ,και το C_i το μέγεθος των καναλιών, τυπικά $C_i = 3$) , ο όγκος εξόδου είναι $W_0 * H_0 * C_0$, όπου C_0 αντιστοιχεί στον αριθμό των φίλτρων που θεωρούμε και

$$W_0 = \frac{W_i - k + 2p}{s} + 1 \tag{26}$$

$$H_0 = \frac{H_i - k + 2p}{s} + 1 \tag{27}$$

Εάν η εικόνα έχει 3 κανάλια και εάν $K_i(i = 1, ..., C0)$ δηλώνουν 5 * 5 * 3 φίλτρα (όπου 3 αντιστοιχεί στον αριθμό των καναλιών της εικόνας εισόδου), η συνέλιξη με την εικόνα Ι με τον πυρήνα K_i αντιστοιχεί στον τύπο:

$$K_l * I(i, j) = \sum_{c=0}^{2} \sum_{n=0}^{4} \sum_{m=4}^{2} K_l(n, m, c)I(i + n - 2, i + m - 2, c) \tag{28}$$

Για εικόνες με κανάλια C_i, το σχήμα του φίλτρου είναι (k, k, C_i, C_0)όπου C_0 είναι ο αριθμός των καναλιών εξόδου (αριθμός φίλτρων) που θεωρούμε. Ο αριθμός παραμέτρων που σχετίζεται με έναν φίλτρο του σχήματος (k, k, C_i, C_0) είναι $(k * k * C_i + 1) * C_0$. Οι λειτουργίες συνέλιξης συνδυάζονται με τη συνάρτηση ενεργοποίησης φ(γενικά Relu συνάρτηση ενεργοποίησης). Αν θεωρήσουμε φίλτρο Κ με μέγεθος $k * k$, αν το x είναι $k * k$ patch εικόνας, η ενεργοποίηση επιτυγχάνεται μετατοπίζοντας το παράθυρο $k * k$ και υπολογισμό $z(x) = \phi(K * x + b)$ όπου το b είναι πόλωση. Σχ.65 Οι μονάδες ανταποκρίνονται στην ίδια θέση αλλά με διαφορετικά βάθη, κάθε μονάδα χρησιμοποιεί διαφορετικό φίλτρο (kernel) στο ίδιο patch της εικόνας.

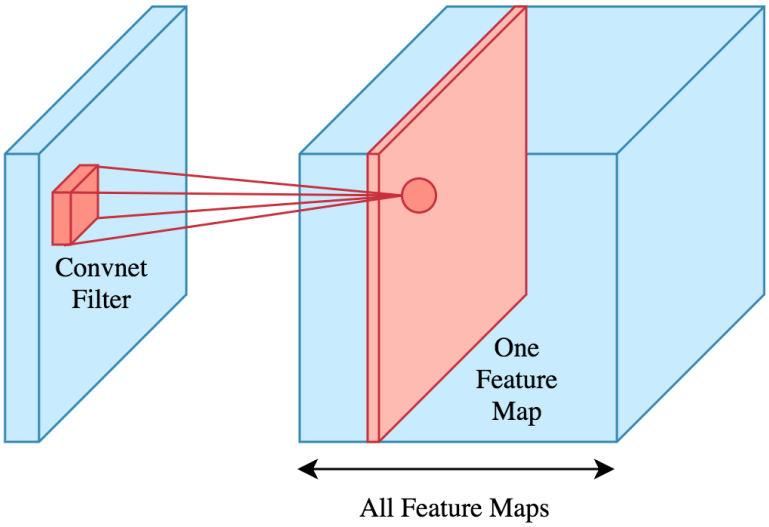


Figure 6: Συνελικτικό επίπεδο

Αυτό είναι στο στρώμα συνέλιξης. Το επίπεδο αυτό του νευρωνικού δικτίου θα "μάθει" τα φίλτρα που είναι πιο χρήσιμα για το έργο που πρέπει να κάνουμε (όπως ταξινόμηση) κατα την διαδικασία τση εκπαίδευσης. Διάφοροι τύποι συνελίξεων μπορούν αν εφαρμοστούν. Η έξοδο της συνέλιξης περνάει από διαδικασία pooling που θα περιγράψουμε παρακάτω και έπειτα συνεχίζει σε όσες συνελίξεις επιθυμούμε.

3.2 Pooling layers

Τα επίπεδα συγκέντρωσης (pooling layers) στα συνελικτικά νευρωνικά δίκτυα, συνοψίζουν τις εξόδους νευρώνων εντός ενός παραθύρου (patch) με μια αντιπροσωπευτική τιμή. Τα γειτονικά παράθυρα δεν επικαλύπτονται. Πρόκειται ουσιαστικά για μια διαδικασία υπο-δειγματοληψίας των δεδομένων. Για καλύτερη κατανόηση της διαδικασίας μπορούμε να φανταστούμε ένα επίπεδο pooling σαν ένα “πλέγμα” pooling νευρώνων τοποθετημένων σε απόσταση pixels, καθένας από τους οποίους συνοψίζει μια περιοχή με κέντρο τον ίδιο τον νευρώνα. Το pooling αποτελεί μια πολύ βασική λειτουργία για κάθε CNN, αφού απλοποιεί πολύ τη διαδικασία λόγω της σημαντικής μείωσης των δεδομένων κι επομένως του αριθμού των απαιτούμενων πράξεων. Αυτό μειώνει κατά πολύ την διαδικασία εκπαίσευσης ενός νευρωνικου δικτύου. Οι επικρατέστερες κατηγορίες του pooling είναι το max, sum και average pooling, ενώ μπορεί τα παράθυρα που χρησιμοποιούνται να επικαλύπτονται ή και όχι ανάλογα με τις ανάγκες του προβλήματος. Max pooling έχουμε όταν τελικά θα περάσει η μέγιστη τιμή του pooling παραθύρου. Sum όταν προσθέτουμε τις τιμές και average όταν δημιουργούμε μέσους όρους. Η διαδικασία του pooling, εκτός από τη μείωση του μεγέθους των δεδομένων, μας δίνει τη δυνατότητα προσθήκης περισσότερης πληροφορίας στην αρχική εικόνα μέσω των αρχικών διαστάσεων ενώ είναι ανεξάρτητο μικρών μετασχηματισμών.

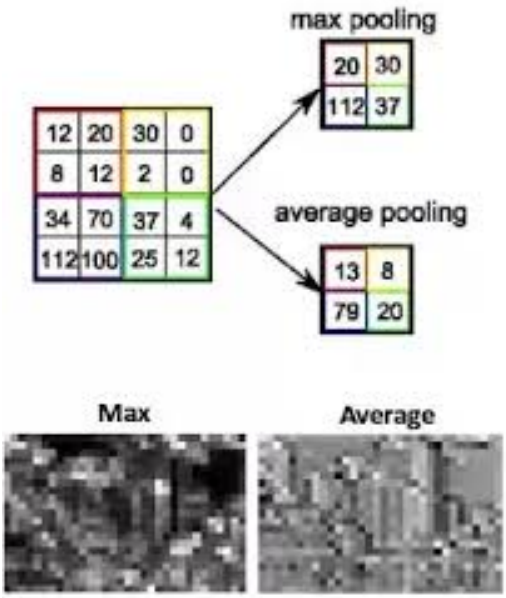


Figure 7: max pooling και sum pooling πάνω σε εικόνα

Όταν έχουμε πλέον λάβει τα χαρακτηριστικά διανύσματα που περιγράψαμε νωρίτερα, αποφασίζουμε για το μέγεθος των παραθύρων, ας πούμε που θα κάνουμε pool τα δεδομένα. Τότε διαιρούμε τα χαρακτηριστικά μας σε περιοχές και παίρνουμε το μέγιστο (max) ή τον μέσο όρο (mean) του παραθύρου, το οποίο αποτελεί το νέο χαρακτηριστικό μας. Αυτές οι pooled περιοχές μπορούν πλέον να χρησιμοποιηθούν για ταξινόμηση. Μετά την διαδικασία του pooling και αφού έχουμε λάβει τα χαρακτηριστικά, συνήθως θέλουμε να τα χρησιμοποιήσουμε για ταξινόμηση. Στη θεωρία, μπορούμε να χρησιμοποιήσουμε όλα τα χαρακτηριστικά σε έναν ταξινομητή όπως ο softmax classifier, ωστόσο αυτό είναι υπολογιστικά δύσκολο.

Ας αναλογιστούμε την περίπτωση εικόνων pixels και 400 χαρακτηριστικά από inputs. Κάθε συνέλιξη φέρει ως αποτέλεσμα μια έξοδο μεγέθους , και εφόσον έχουμε 400 χαρακτηριστικά, προκύπτει ένα διάνυσμα χαρακτηριστικά ανά δείγμα. Η εκμάθηση ενός ταξινομητή με περισσότερο απο 3 εκατομμύρια χαρακτηριστικά, μπορεί να γίνει πολύ επίπονη και υπέρ-εξειδικευμένη (over-fitting). Προκειμένου να το διευθετήσουμε, ας ανακαλέσουμε το γεγονός ότι αποφασίσαμε να λαμβάνουμε convolved εικόνες επειδή έχουν την ιδιότητα της "στατικότητας", δηλαδή ότι τα χαρακτηριστικά που είναι χρήσιμα σε μια περιοχή της εικόνας, είναι χρήσιμα σε κάθε άλλη περιοχή της. Έτσι, προκειμένου να περιγράψουμε μια μεγάλη εικόνα θα λάβουμε ένα μίγμα στατιστικών των χαρακτηριστικών των παραπάνω εικόνων. Για παράδειγμα, θα μπορούσαμε να λάβουμε έναν μέσο όρο ή τη μέγιστη τιμή ενός συγκεκριμένου χαρακτηριστικού για ένα patch της εικόνας. Έτσι, ελαττώνονται σημαντικά οι διαστάσεις των δεδομένων και πετυχαίνουμε μικρότερο over-fitting.

Εάν κάποιος επιλέξει οι περιοχές που γίνεται το pooling να είναι συνεχόμενες στην εικόνα και λαμβάνει χαρακτηριστικά μόνο από τους ίδιους κρυφούς νευρώνες, τότε αυτοί οι pooling νευρώνες θα γίνουν "translation invariant" ή αλλιώς ανεξάρτητα μετατοπίσεων της εικόνας. Αυτό σημαίνει ότι το ίδιο χαρακτηριστικό θα ενεργοποιείται ακόμα και όταν η εικόνα υπόκειται μικρές μετατοπίσεις. Συνήθως γι αυτό το λόγο επιδιώκεται τα χαρακτηριστικά μας να είναι ανεξάρτητα μετατοπίσεων, όπως για παράδειγμα στην ανίχνευση αντικειμένων ή την αναγνώριση ήχου.

3.3 Μέθοδοι αποφυγής overfitting

3.3.1 Dropout

Υπάρχουν διάφορες μέθοδοι για να μειώσουμε τα σφάλματα εκπαίδευσης ενός Νευρωνικού Δικτύου, όπως για παράδειγμα να συγκρίνουμε τις προβλέψεις πολλών και διαφορετικών μοντέλων. Για μεγάλα Νευρωνικά Δίκτυα των οποίων η εκπαίδευση μπορεί να κρατήσει αρκετές ημέρες αυτή η μέθοδος είναι χρονοβόρα και σε ορισμένες περιπτώσεις ίσως και αδύνατη. Προς αντιμετώπιση αυτού του προβλήματος έχει αναπτυχθεί κάποια μέθοδος σύγκρισης μοντέλων αρκετά αποτελεσματική, το υπολογιστικό κόστος της οποίας είναι πολύ χαμηλό. Η μέθοδος αυτή, η οποία ονομάζεται "Dropout" συνίσταται στην την ανάθεση ως "0" της

εξόδου κάθε κρυφού νευρώνα με πιθανότητα 0.5. Οι νευρώνες που συμμετέχουν με αυτό τον τρόπο στο "Dropout" δεν συνεισφέρουν στη διάδοση προς τα εμπρός των σημάτων εκπαίδευσης και δεν συμμετέχουν στην διαδικασία του back-propagation. Έτσι, κάθε φορά που μια είσοδος παρουσιάζεται στο Δίκτυο, εκείνο χρησιμοποιεί διαφορετική αρχιτεκτονική αλλά όλες αυτές οι αρχιτεκτονικές μοιράζονται τα ίδια βάρη. Αυτή η τεχνική μειώνει τις περίπλοκες συν-προσαρμογές των νευρώνων κι έτσι το Δίκτυο αποκτά τη δυνατότητα εκμάθησης ισχυρότερων χαρακτηριστικών. Κατά τη διαδικασία της επαλήθευσης (test) χρησιμοποιούμε όλους τους νευρώνες πολλαπλασιάζοντας όμως τις εξόδους τους με 0.5 προκειμένου να πάρουμε τον γεωμετρικό μέσο των κατανομών πρόβλεψης. Τέλος, αναφέρουμε ότι η χρήση του Dropout σχεδόν διπλασιάζει τον αριθμό των επαναλήψεων που απαιτούνται για σύγκλιση.

3.3.2 Data augmentation

Μια πιό απλή μέθοδος είναι η τεχνητή αύξηση των δεδομένων εκπαίδευσης του νευρωνικού δικτύου. Στην περίπτωση που αυτά είναι φωτογραφίες μπορούμε με απλές τεχνικές να αυξήσουμε τον όγκο αυτόν. Περιστροφή των εικώνων ή πρόσθεση φιλτρων / θορύβου σε αυτές είναι κάποιες από αυτές.

4 Τα δεδομένα

4.1 Smiles

Το SMILES (Simplified Molecular Input Line Entry System) είναι ένα σύστημα χημικής σηματοδότησης σχεδιασμένο για σύγχρονη επεξεργασία χημικών πληροφοριών. Με βάση τις αρχές της θεωρίας των μοριακών γραφημάτων, Το SMILES επιτρέπει αυστηρές προδιαγραφές δομών με τη χρήση μιας πολύ μικρής και φυσικής γραμματικής. Το σύστημα συμβολισμού SMILES είναι επίσης κατάλληλο για επεξεργασία μηχανής υψηλής ταχύτητας. Το αποτέλεσμα η ευκολία χρήσης από τον φαρμακοποιό και η συμβατότητα με το μηχάνημα επιτρέπουν πολλά υψηλής απόδοσης χημικά εφαρμογές ηλεκτρονικών υπολογιστών που θα σχεδιάζονται, συμπεριλαμβανομένης της δημιουργίας μιας μοναδικής συμβολής, σταθερής ταχύτητας ανάκτησης βάσης δεδομένων, ευέλικτη αναζήτηση υποδομής και μοντέλα πρόβλεψης ιδιοτήτων.

4.2 Περιγραφή και κανόνες

4.2.1 Άτομα

Τα άτομα αντιπροσωπεύονται από την συνήθη συντομογραφία των χημικών στοιχείων, σε αγκύλες, όπως το [Au] για το χρυσό. Οι υποστηρίξεις μπορούν να παραλειφθούν στην κοινή περίπτωση ατόμων που: βρίσκονται στο "οργανικό υποσύνολο" των B, C, N, O, P, S, F, CI, Br ή I και δεν έχουν επίσημη κατηγορία, και έχουν τον αριθμό των υδρογόνων που συνδέονται με το μοντέλο σθένους SMILES (τυπικά το κανονικό τους σθένος, αλλά για τα N και P είναι 3 ή 5 και για το S είναι 2, 4 ή 6) και είναι τα φυσιολογικά ισότοπα. Όλα τα άλλα στοιχεία πρέπει να περικλείονται σε παρένθεση και να έχουν σαφώς αναγραφόμενα φορτία και υδρογόνα. Για παράδειγμα, τα SMILES για νερό μπορούν να γραφτούν ως είτε O είτε [OH2]. Το υδρογόνο μπορεί επίσης να γραφτεί ως ξεχωριστό άτομο. το νερό μπορεί επίσης να γραφεί ως [H] O [H].

Όταν χρησιμοποιούνται βραχίονες, προστίθεται το σύμβολο H αν το άτομο σε παρένθεση συνδέεται με ένα ή περισσότερα υδρογόνα, ακολουθούμενο από τον αριθμό των ατόμων υδρογόνου εάν είναι μεγαλύτερο από 1, τότε με το σύμβολο + για θετικό φορτίο ή με - για αρνητικό χρέωση. Για παράδειγμα, [NH4 +] για αμμώνιο (NH + 4). Εάν υπάρχουν περισσότερες από μία χρεώσεις, συνήθως γράφονται ως ψηφία. Ωστόσο, είναι επίσης δυνατό να επαναλάβετε το σύμβολο όσες φορές το ιόν έχει φορτίσεις: μπορεί κανείς να γράψει είτε [Ti + 4] είτε [Ti ++++] για τιτάνιο (IV) Ti4 +. Έτσι, το ανιόν υδροξειδίου (OH-) αντιπροσωπεύεται από [OH-], το κατιόν υδρονίου (H 3O +) είναι [OH3 +] και το κατιόν κοβαλτίου (III) (Co3 +) είναι είτε [Co + 3] είτε [Co +++].

4.2.2 Δεσμοί

Ένας δεσμός αντιπροσωπεύεται χρησιμοποιώντας ένα από τα σύμβολα. - = $: / \ .

Οι δεσμοί μεταξύ αλειφατικών ατόμων θεωρούνται ότι είναι μονές αν δεν ορίζεται διαφορετικά και υπονοούνται από την παρακέντηση στη συμβολοσειρά SMILES. Αν και οι απλοί δεσμοί μπορούν να γραφτούν ως -, αυτό συνήθως παραλείπεται. Για παράδειγμα, τα SMILES για την αιθανόλη μπορούν να γραφτούν ως C-C-O, CC-O ή C-CO, αλλά συνήθως γράφονται CCO. Οι διπλοί, τριπλοί και τετραπλοί δεσμοί αντιπροσωπεύονται από τα σύμβολα =, #, και \ αντίστοιχα, όπως απεικονίζεται από το SMILES O = C = O (διοξείδιο του άνθρακα CO 2), C # N (κυανιούχο υδρογόνο HCN) και [Ga-] $ [As +] (αρσενίδιο του γαλλίου). Ένας επιπλέον τύπος δεσμού είναι ένας "μη δεσμός", που υποδεικνύεται με., Για να υποδείξει ότι δύο μέρη δεν είναι συνδεδεμένα μεταξύ τους. Για παράδειγμα, το υδατικό χλωριούχο νάτριο μπορεί να γραφεί ως [Na +] [Cl-] για να δείξει τη διάσταση. Ένας αρωματικός "ενάμισι" δεσμός μπορεί να υποδεικνύεται με :. Οι απλοί δεσμοί δίπλα σε διπλούς δεσμούς μπορεί να αντιπροσωπεύονται χρησιμοποιώντας / ή \ για να δείξει στερεοχημική διαμόρφωση.

4.2.3 Δαχτύλιοι

Οι δακτυλιοειδείς δομές γράφονται με το σπάσιμο κάθε δακτυλίου σε ένα αυθαίρετο σημείο (αν και κάποιες επιλογές θα οδηγήσουν σε ένα πιο ευανάγνωστο SMILES από άλλους) για να δημιουργήσουν μια ακυκλική δομή και να προσθέσουν αριθμητικές ετικέτες κλεισίματος δακτυλίου για να δείξουν συνδεσιμότητα μεταξύ μη γειτονικών ατόμων. Για παράδειγμα, το κυκλοεξάνιο και το διοξάνιο μπορούν να γραφτούν ως C1CCCCC1 και O1CCOCC1 αντίστοιχα. Για ένα δεύτερο δακτύλιο, η ετικέτα θα είναι 2. Για παράδειγμα, η δεκαλίνη (δεκαϋδροναφθαλίνη) μπορεί να γραφεί ως C1CCCC2C1CCCC2. Το SMILES δεν απαιτεί να χρησιμοποιούνται οι αριθμοί δακτυλίων σε οποιαδήποτε συγκεκριμένη σειρά και επιτρέπει τον αριθμό δακτυλίου μηδέν, αν και σπάνια χρησιμοποιείται. Επίσης, επιτρέπεται η επαναχρησιμοποίηση των αριθμών δακτυλίων μετά το κλείσιμο του πρώτου δακτυλίου, αν και αυτό συνήθως καθιστά τους τύπους πιο δύσκολο να διαβαστούν. Για παράδειγμα, το δικυκλοεξύλιο συνήθως γράφεται ως C1CCCCC1C2CCCCC2, αλλά μπορεί επίσης να γραφεί ως C0CCCCC0C0CCCCC0. Πολλαπλά ψηφία μετά από ένα μόνο άτομο υποδεικνύουν πολλαπλούς δεσμούς κλεισίματος δακτυλίου. Για παράδειγμα, μια εναλλακτική σημείωση SMILES για τη δεκαλίνη είναι C1CCCC2CCCCC12, όπου ο τελικός άνθρακας συμμετέχει και στους δύο δεσμούς κλεισίματος 1 και 2. Εάν απαιτούνται διψήφιοι αριθμοί δακτυλίων, η ετικέτα προηγείται από το έτσι το C 12 είναι ένα μονό δακτυλιοειδές δεσμό του δακτυλίου 12. Το ένα ή και τα δύο ψηφία μπορεί να προηγείται από έναν τύπο δεσμού που υποδεικνύει τον τύπο του δεσμού που κλείνει δακτύλιο. Για παράδειγμα, το κυκλοπροπένιο είναι συνήθως γραμμένο C1 = CC1, αλλά αν ο διπλός δεσμός επιλέγεται ως δεσμός κλεισίματος δακτυλίου, μπορεί να γραφεί ως C = 1CC1, C1CC = 1 ή C = 1CC = 1. (Η πρώτη μορφή προτιμάται.) Το C = 1CC-1 είναι παράνομο, καθώς ορίζει ρητά τους αντικρουόμενους τύπους για τον δεσμό κλεισίματος δακτυλίου. Οι δεσμοί κλεισίματος δακτυλίων δεν μπορούν να χρησιμοποιηθούν για να δηλώσουν πολλαπλούς δεσμούς. Για παράδειγμα, το C1C1 δεν αποτελεί έγκυρη εναλλακτική λύση έναντι του C = C για το αιθυλένιο. Ωστόσο, μπορούν να χρησιμοποιηθούν με μη δεσμούς. Το C1.C2.C12 είναι ένας ιδιότυπος αλλά νόμιμος εναλλακτικός τρόπος για να γράψετε προπάνιο, συνηθέστερα γραμμένο CCC. Η επιλογή ενός σημείου διακοπής δακτυλίου δίπλα σε συσχετισμένες ομάδες μπορεί να οδηγήσει σε μια απλούστερη μορφή SMILES αποφεύγοντας κλάδους. Για παράδειγμα, η κυκλοεξανο-1,2-διόλη είναι απλούστερα γραμμένη ως OC1CCCCC10. η επιλογή μιας διαφορετικής θέσης θραύσης του δακτυλίου δημιουργεί μια διακλαδισμένη δομή που απαιτεί τη συμπλήρωση των παρενθέσεων.

4.2.4 Αρωματικότητα

Οι αρωματικοί δακτύλιοι όπως το βενζόλιο μπορούν να γραφτούν σε μία από τις τρεις μορφές:

Σε μορφή Kekulé με εναλλασσόμενους μονούς και διπλούς δεσμούς, π.χ. C1 = CC = CC = C1, Χρησιμοποιώντας το σύμβολο αρωματικού δεσμού: π.χ. C: 1: C: C: C: C: C1, ή Συνήθως, γράφοντας τα συστατικά B, C, N, O, P και S άτομα σε μικρές περιπτώσεις β, ο, η, ο, ρ και s, αντιστοίχως. Στην τελευταία περίπτωση, δεσμοί μεταξύ δύο αρωματικών ατόμων θεωρούνται (αν δεν φαίνονται ρητά) ότι είναι αρωματικοί δεσμοί. Έτσι, το βενζόλιο, η πυριδίνη και το φουράνιο μπορούν να αναπαρασταθούν αντιστοίχως από τα SMILES c1ccccc1, n1ccccc1 και o1cccc1. Το αρωματικό άζωτο που συνδέεται με το υδρογόνο, όπως βρίσκεται στο πυρρόλιο, πρέπει να αντιπροσωπεύεται ως [nH]. έτσι η ιμιδαζόλη γράφεται στη σημείωση SMILES ως n1c [nH] cc1. Όταν αρωματικά άτομα συνδέονται μεμονωμένα μεταξύ τους, όπως σε διφαινύλιο, ένας απλός δεσμός πρέπει να φαίνεται ρητά: c1ccccc1-c2ccccc2. Αυτή είναι μία από τις

λίγες περιπτώσεις όπου απαιτείται το σύμβολο απλής ομολογίας. (Στην πραγματικότητα, το μεγαλύτερο μέρος του λογισμικού SMILES μπορεί σωστά να συμπεράνει ότι ο δεσμός μεταξύ των δύο δακτυλίων δεν μπορεί να είναι αρωματικός και έτσι θα αποδεχθεί τη μη τυπική μορφή c1ccccc1c2ccccc2.) Οι αλγόριθμοι Daylight και OpenEye για την παραγωγή κανονικών SMILES διαφέρουν ως προς την αρωματοθεραπεία τους.

4.2.5 Διακλάδωση

Οι κλάδοι περιγράφονται με παρενθέσεις, όπως στο CCC (= O) O για το προπιονικό οξύ και στο FC (F) F για το φθοροφόρμιο. Το πρώτο άτομο εντός των παρενθέσεων και το πρώτο άτομο μετά την παρένθετη ομάδα παρέχονται ταυτόχρονα με το ίδιο άτομο σημείου διακλάδωσης. Το σύμβολο δεσμού εμφανίζεται μέσα στις παρενθέσεις. Το CCC = (O) O είναι άκυρο. Υποκατεστημένοι δακτύλιοι μπορούν να γραφτούν με το σημείο διακλάδωσης στον δακτύλιο όπως απεικονίζεται από τα SMILES COc (c1) cccc1C # N (βλέπε απεικόνιση) και COc (cc1) ccc1C # N (βλέπε απεικόνιση) που κωδικοποιούν τα ισομερή 3 και 4-κυανοανισόλης. Γράφοντας SMILES για υποκατεστημένα δαχτυλίδια με αυτόν τον τρόπο μπορούν να γίνουν πιο ανθρώπινα αναγνώσιμα. Τα υποκαταστήματα μπορούν να γραφτούν με οποιαδήποτε σειρά. Για παράδειγμα, το βρωμοχλωροδιφθορομεθάνιο μπορεί να γραφεί ως FC (Br) (Cl) F, BrC (F) (F) Cl, C (F) (Cl) (F) Br ή τα παρόμοια. Γενικά, μια μορφή SMILES είναι ευκολότερη στην ανάγνωση αν ο απλούστερος κλάδος έρχεται πρώτο, με το τελικό, μη αφαιρούμενο τμήμα να είναι το πιο πολύπλοκο. Οι μόνες προειδοποιήσεις σε τέτοιες αναδιατάξεις είναι: Εάν οι αριθμοί κλήσης επαναχρησιμοποιηθούν, αντιστοιχίζονται ανάλογα με τη σειρά εμφάνισής τους στη συμβολοσειρά SMILES. Κάποιες ρυθμίσεις ενδέχεται να απαιτούνται για τη διατήρηση της σωστής αντιστοίχισης. Εάν καθορίζεται στερεοχημεία, πρέπει να γίνουν προσαρμογές. βλ. Στερεοχημεία Η μία μορφή διακλάδωσης που δεν απαιτεί παρενθέσεις είναι δεσμοί κλεισίματος δακτυλίου. Η επιλογή των δεσμών κλεισίματος δακτυλίων κατάλληλα μπορεί να μειώσει τον αριθμό των παρενθέσεων που απαιτούνται. Για παράδειγμα, το τολουόλιο γράφεται κανονικά ως Cc1ccccc1 ή c1ccccc1C, αποφεύγοντας τις παρενθέσεις που απαιτούνται εάν είναι γραμμένες ως c1ccc (C) ccc1 ή c1ccc (ccc1) C.

4.2.6 Στερεοχημεία

Το SMILES επιτρέπει, αλλά δεν απαιτεί, προδιαγραφή των στερεοϊσομερών.

Η διαμόρφωση γύρω από τους διπλούς δεσμούς καθορίζεται χρησιμοποιώντας τους χαρακτήρες / και για να δείξουν κατευθυντικούς απλούς δεσμούς δίπλα σε διπλό δεσμό. Για παράδειγμα, το F / C = C / F (βλέπε απεικόνιση) είναι μία αναπαράσταση του trans-1,2-διφθοροαιθυλενίου, όπου τα άτομα φθορίου βρίσκονται στις απέναντι πλευρές του διπλού δεσμού (όπως φαίνεται στο σχήμα) C = C F (βλέπε απεικόνιση) είναι μία πιθανή αναπαράσταση του cis-1,2-διφθοροαιθυλενίου, όπου τα φθόρια είναι στην ίδια πλευρά του διπλού δεσμού.

Τα σύμβολα κατεύθυνσης πρόσφυσης έρχονται πάντα σε ομάδες τουλάχιστον δύο, εκ των οποίων η πρώτη είναι αυθαίρετη. Δηλαδή, το F C = C F είναι το ίδιο με το F / C = C / F. Όταν υπάρχουν εναλλασσόμενοι μονο-διπλοί δεσμοί, οι ομάδες είναι μεγαλύτερες από δύο, με τα σύμβολα μέσης κατεύθυνσης να γειτονεύουν με δύο διπλούς δεσμούς. Για παράδειγμα, η κοινή μορφή του (2,4) -εξαδιενίου είναι γραμμένη C / C = C / C = C / C.

Το βήτα-καροτένιο, με τα έντεκα διπλά ομόλογα επισημαίνονται. Ως περισσότερο περίπλοκο παράδειγμα, το βήτα-καροτένιο έχει πολύ μακρύ σκελετό εναλλασσόμενων μονών και διπλών δεσμών, οι οποίοι μπορεί να είναι γραμμένοι CC1CCC / C (C) = C1 / C = C / C (C) = C / C = C / C (C) = C / C = C / C = C (C) / C = C / C = C (C)

Η διαμόρφωση σε τετραεδρικό άνθρακα καθορίζεται από @ ή @@. Εξετάστε τα τέσσερα ομόλογα με τη σειρά που εμφανίζονται, από αριστερά προς τα δεξιά, στη φόρμα SMILES. Κοιτώντας προς τον κεντρικό άνθρακα από την οπτική του πρώτου δεσμού, οι άλλες τρεις είναι είτε δεξιόστροφα είτε αριστερόστροφα. Αυτές οι περιπτώσεις υποδεικνύονται με @@ και @, αντίστοιχα (επειδή το ίδιο το σύμβολο @ είναι μια στροφή προς τα αριστερά).

Για παράδειγμα, εξετάστε το αμινοξύ αλανίνη. Μία από τις μορφές της SMILES είναι NC (C) C (= O) O, πληρέστερα γραμμένη ως N [CH] (C) C (= O) O. Η L-αλανίνη, το πιο συνηθισμένο εναντιομερές, γράφεται ως N [C'H] (C) C (= O) O (βλέπε απεικόνιση). Κοιτώντας από τον δεσμό αζώτου-άνθρακα,

οι ομάδες υδρογόνου (H), μεθυλίου (C) και καρβοξυλικού (C (= O) O) εμφανίζονται δεξιόστροφα. Η D-Αλανίνη μπορεί να γραφεί ως N [CH] (C) C (= O) O (βλέπε απεικόνιση).

Ενώ η εντολή είναι ποιες κλάσεις καθορίζονται σε SMILES είναι συνήθως ασήμαντη, σε αυτή την περίπτωση έχει σημασία? η εναλλαγή οποιωνδήποτε δύο ομάδων απαιτεί αντιστροφή του δείκτη chirality. Εάν τα κλαδιά αντιστρέφονται έτσι ώστε η αλανίνη να γράφεται ως NC (C (= O) O) C, τότε η διαμόρφωση αναστρέφεται επίσης. Η L-αλανίνη γράφεται ως N [CH] (C (= O) O) C (βλέπε απεικόνιση). Άλλοι τρόποι συγγραφής αυτού περιλαμβάνουν το C [CH] (N) C (= O) O, OC (= O) [CHH] (N) C και OC (= O) [CH]) N.

Συνήθως, ο πρώτος από τους τέσσερις δεσμούς εμφανίζεται στα αριστερά του ατόμου άνθρακα, αλλά εάν οι SMILES γράφονται αρχίζοντας με τον ασύμμετρο άνθρακα, όπως το C (C) (N) C (= O) O, τότε και οι τέσσερις είναι το δικαίωμα, αλλά το πρώτο που εμφανίζεται (ο δεσμός [CH] σε αυτή την περίπτωση) χρησιμοποιείται ως αναφορά για την τάξη των ακόλουθων τριών: L-αλανίνη μπορεί επίσης να γραφεί [C @ H] (C) = O) O.

Η προδιαγραφή SMILES περιλαμβάνει επεξεργασίες στο σύμβολο @ για να δείξει στερεοχημεία γύρω από πιο περίπλοκα χειρομορφικά κέντρα, όπως τριγωνική διπυραμιδική μοριακή γεωμετρία.

4.2.7 Ισότοπα

Τα ισότοπα καθορίζονται με έναν αριθμό ίσο προς την ακέραια ισοτοπική μάζα που προηγείται του ατομικού συμβόλου. Το βενζόλιο στο οποίο ένα άτομο είναι άνθρακας-14 γράφεται ως [14c] 1ccccc1 και το δευτεριοχλωροφόρμιο είναι [2H] C (Cl) (Cl) Cl.

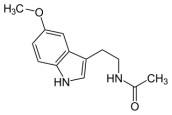
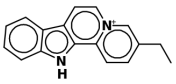
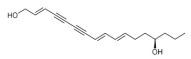
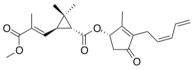
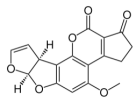
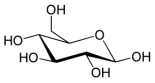
Vanillin		`O=Cc1ccc(O)c(OC)c1` `COCc1cc(C=O)ccc1O`
Melatonin ($C_{13}H_{16}N_2O_2$)		`CC(=O)NCCC1=CNc2c1cc(OC)cc2` `CC(=O)NCCc1c[nH]c2ccc(OC)cc12`
Flavopereirin ($C_{17}H_{15}N_2$)		`CCc(c1)ccc2[n+]1ccc3c2[nH]c4c3cccc4` `CCc1c[n+]2ccc3c4ccccc4[nH]c3c2cc1`
Nicotine ($C_{10}H_{14}N_2$)		`CN1CCC[C@H]1c2cccnc2`
Oenanthotoxin ($C_{17}H_{22}O_2$)		`CCC[C@@H](O)CC\C=C\C=C\C#CC#C\C=C\CO` `CCC[C@@H](O)CC/C=C/C=C/C#CC#C/C=C/CO`
Pyrethrin II ($C_{22}H_{28}O_5$)		`CC1=C(C(=O)C[C@@H]1OC(=O)[C@@H]2[C@H](C2(C)C)/C=C(\C)/C(=O)OC)C/C=C\C=C`
Aflatoxin B_1 ($C_{17}H_{12}O_6$)		`O1C=C[C@H]([C@H]1O2)c3c2cc(OC)c4c3OC(=O)C5=C4CCC(=O)5`
Glucose (glucopyranose) ($C_6H_{12}O_6$)		`OC[C@@H](O1)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)1`

Figure 8: Παραδείγματα Smiles

5 Μοντελοποιήσεις και μεθοδολογίες

5.1 Ποσοτικά μοντέλα σχέσης δομής-δραστηριότητας (QSAR)

Τα ποσοτικά μοντέλα σχέσης δομής-δραστηριότητας (μοντέλα QSAR) είναι μοντέλα παλινδρόμησης ή ταξινόμησης που χρησιμοποιούνται στις χημικές και βιολογικές επιστήμες και στη μηχανική. Όπως και άλλα μοντέλα παλινδρόμησης, τα μοντέλα παλινδρόμησης QSAR συνδέουν ένα σύνολο μεταβλητών «πρόβλεψης» (X) με την ισχύ της μεταβλητής απόκρισης (Υ), ενώ τα μοντέλα ταξινόμησης QSAR συνδέουν τις μεταβλητές πρόβλεψης με μια κατηγορική τιμή της μεταβλητής απόκρισης.

Στο μοντέλο QSAR, οι προγνωστικοί παράγοντες συνίστανται από φυσικοχημικές ιδιότητες ή από θεωρητικούς μοριακούς περιγραφείς χημικών. η μεταβλητή απόκρισης QSAR θα μπορούσε να είναι μια βιολογική δραστηριότητα των χημικών ουσιών. Τα μοντέλα QSAR συνοψίζουν πρώτα μια υποτιθέμενη σχέση μεταξύ των χημικών δομών και της βιολογικής δραστηριότητας σε ένα σύνολο χημικών ουσιών. Δεύτερον, τα μοντέλα QSAR προβλέπουν τις δραστηριότητες των νέων χημικών ουσιών.

Οι σχετικοί όροι περιλαμβάνουν ποσοτικές σχέσεις δομής-ιδιότητας (QSPR) όταν μια χημική ιδιότητα διαμορφώνεται ως μεταβλητή απόκρισης. [3] [4] (QSRRs), ποσοτικές σχέσεις δομής-χρωματογραφίας (QSCRs) και σχέσεις ποσοτικής δομής-τοξικότητας (QSTRs), ποσοτική δομή (QSERs) και σχέσεις ποσοτικής δομής-βιοδιασπασιμότητας (QSBRs).

Ως παράδειγμα, η βιολογική δραστικότητα μπορεί να εκφραστεί ποσοτικά ως η συγκέντρωση μιας ουσίας που απαιτείται για να δώσει μια ορισμένη βιολογική απόκριση. Επιπλέον, όταν οι φυσικοχημικές ιδιότητες ή οι δομές εκφράζονται με αριθμούς, μπορεί να βρεθεί μια μαθηματική σχέση, ή μια ποσοτική σχέση δομής-δραστηριότητας, μεταξύ των δύο. Η μαθηματική έκφραση, εάν επικυρωθεί προσεκτικά [7] [8] [9] μπορεί στη συνέχεια να χρησιμοποιηθεί για την πρόβλεψη της προσομοιωμένης απόκρισης άλλων χημικών δομών.

5.2 Μοριακός περιγραφέας χημικής δομής

Οι μοριακοί περιγραφείς διαδραματίζουν θεμελιώδη ρόλο στη χημεία, στις φαρμακευτικές επιστήμες, στην πολιτική προστασίας του περιβάλλοντος και στις έρευνες υγείας, καθώς και στον ποιοτικό έλεγχο, είναι ο τρόπος με τον οποίο τα μόρια μετατρέπονται σε αριθμούς επιτρέποντας κάποια μαθηματική επεξεργασία

της χημικής ουσίας ως πληροφορίες που περιέχονται στο μόριο. Αυτό ορίστηκε από τον Todeschini και τον Consonni ως εξής:

"Ο μοριακός περιγραφέας είναι το τελικό αποτέλεσμα μιας λογικής και μαθηματικής διαδικασίας που μετατρέπει τις χημικές πληροφορίες που κωδικοποιούνται μέσα σε μια συμβολική αναπαράσταση ενός μορίου σε έναν χρήσιμο αριθμό ή το αποτέλεσμα κάποιου τυποποιημένου πειράματος." [1]

Με αυτόν τον ορισμό, οι μοριακοί περιγραφείς διαιρούνται σε δύο κύριες κατηγορίες: πειραματικές μετρήσεις, όπως log P, μοριακή διαθλαστικότητα, διπολική ροπή, πολικότητα και, γενικά, πρόσθετες φυσικοχημικές ιδιότητες και θεωρητικοί μοριακοί περιγραφείς, οι οποίοι προέρχονται από μια συμβολική αναπαράσταση του μορίου και μπορεί να ταξινομηθεί περαιτέρω σύμφωνα με τους διαφορετικούς τύπους μοριακής αναπαράστασης.

5.3 Ποσοτικές περιγραφές μοριακής δομής

Ο αριθμητικός χαρακτηρισμός της μοριακής δομής είναι ένα πρώτο βήμα σε πολλές υπολογιστικές αναλύσεις δεδομένων χημικής δομής. Αυτές οι αριθμητικές αναπαραστάσεις, που ονομάζονται περιγραφείς, έρχονται σε πολλές μορφές, που κυμαίνονται από απλές μετρήσεις ατόμων και μεταβλητές μοριακού γραφήματος σε κατανομή ιδιοτήτων, όπως φορτίο, σε μια μοριακή επιφάνεια..

Οι υπολογιστικές μέθοδοι διαδραματίζουν σημαντικό ρόλο σε πολλές χημικές επιστήμες που κυμαίνονται από την ανακάλυψη φαρμάκων μέχρι την επιστήμη των υλικών. Υπάρχει μια πληθώρα τεχνικών που διαφέρουν όσον αφορά την υπολογιστική πολυπλοκότητα, τις απαιτήσεις χρόνου και ούτω καθεξής. Ωστόσο, η κοινή απαίτηση που βασίζεται σε όλες αυτές τις μεθόδους είναι μια τυπική περιγραφή μιας μοριακής δομής. Υπάρχουν πολλοί τρόποι να "περιγράψουμε" ένα μόριο. Μια κοινή προσέγγιση είναι να περιγράψουμε τη συνδεσιμότητα, λαμβάνοντας υπόψη τους τύπους ατόμων και δεσμών. Με άλλα λόγια, ρητές αναπαραστάσεις της χημικής δομής, όπως SMILES, αρχεία MDL / Symyx SD και ούτω καθεξής. Παρόλο που αυτές οι περιγραφές είναι ζωτικής σημασίας για τα σύγχρονα συστήματα χημικών πληροφοριών, δεν επιτρέπουν απαραιτήτως την άμεση εφαρμογή υπολογιστικών τεχνικών σε αυτά.

Μέθοδοι που στοχεύουν στην πρόβλεψη χημικών και βιολογικών ιδιοτήτων γενικά απαιτούν μια αριθμητική περιγραφή των χημικών δομών. Τέτοιες αριθμητικές μορφές κυμαίνονται από μια σειρά από 3D συντεταγμένες που συνδυάζονται με τους κατάλληλους τύπους ατόμων, είναι επαρκείς για μεθόδους όπως οι κβαντικές μηχανικές προσεγγίσεις και η πρόσδεση σε πιο αφηρημένες αριθμητικές περιγραφές που προέρχονται από 2D ή 3D αναπαραστάσεις οι οποίες μπορούν να είναι χρήσιμες στις στατιστικές προσεγγίσεις. Τώρα είναι δυνατόν να εκτιμηθούν χιλιάδες αριθμητικοί περιγραφείς της χημικής δομής. Όπως θα συζητηθεί αργότερα, πολλοί από αυτούς τους περιγραφείς σχετίζονται στενά ή καταγράφουν τις ίδιες πληροφορίες, επιτρέποντας σε κάποιον να αντικαταστήσει έναν άλλο. Η επιλογή των σχετικών περιγραφέων είναι ένα πολύ γνωστό πρόβλημα και δεδομένης μιας μεγάλης συλλογής αυτών, οι προσεγγίσεις για τον προσδιορισμό ενός κατάλληλου υποσυνόλου έχουν συζητηθεί εκτενώς στη βιβλιογραφία [**?**, **?**]. Οι μοριακοί περιγραφείς μπορούν να υπολογιστούν για πολλές χημικές οντότητες, όχι μόνο για μικρά οργανικά μόρια.

Εκτός από την ύπαρξη πολλών περιγραφικών δεικτών που ορίζονται στη βιβλιογραφία, υπάρχουν επίσης πολλαπλές υλοποιήσεις ενός περιγραφικού δελτίου. Αυτές οι υλοποιήσεις είναι διαθέσιμες με τη μορφή βιβλιοθηκών (οι οποίες απαιτούν τη σύνταξη ενός προγράμματος για τη χρήση τους) ή πλήρεις εφαρμογές (γραφικό περιβάλλον χρήστη ή γραμμή εντολών). Ως αποτέλεσμα, όχι μόνο πρέπει να επιλέξετε έναν ή περισσότερους περιγραφείς που σχετίζονται με το πρόβλημα, αλλά πρέπει να ανησυχείτε για τον τρόπο με τον οποίο υπολογίζονται και αν ένας υπολογισμός μπορεί να αναπαραχθεί σε διαφορετικές υλοποιήσεις αυτών των περιγραφών. Είναι εύκολο να καταλάβουμε γιατί δύο υλοποιήσεις του ίδιου περιγραφέα μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα. Οι κύριοι λόγοι είναι οι διαφορές στο μοντέλο χημείας του πλαισίου ή του εργαλείου που χρησιμοποιείται για την υλοποίηση του περιγραφικού εγγράφου. Για παράδειγμα, ένας περιγραφέας που υπολογίζει τον αριθμό αρωματικών ατόμων μπορεί να εφαρμοστεί χρησιμοποιώντας δύο ομάδες εργαλείων με διαφορετικά μοντέλα αρωματικότητας και επομένως είναι πιθανό οι τιμές που παράγονται από τις δύο υλοποιήσεις να διαφέρουν. Άλλες πηγές διαφορών περιλαμβάνουν παραμέτρους που μπορεί να εμπλέκονται στον υπολογισμό του περιγραφικού υποδείγματος και στις τιμές δεδομένων αναφοράς (όπως είναι οι ατομικές ακτίνες, οι τιμές ηλεκτροναυτιμότητας) που χρησιμοποιούνται κατά τον υπολογισμό του περιγραφικού. Ενώ οι περισσότερες εφαρμογές θα χρησιμοποιούν τις ίδιες πηγές δεδομένων για τυπικές έννοιες (π.χ. ατομικά βάρη), μικρές διαφορές σε αυτούς τους τύπους δε-

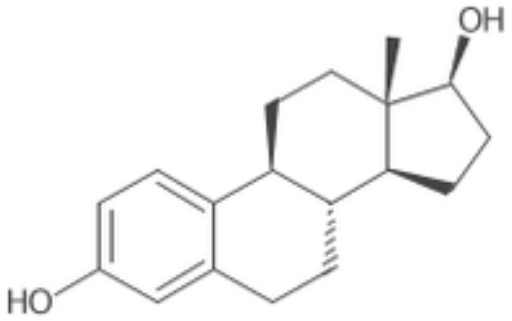


Figure 9: Φωτογραφία "estradiol"

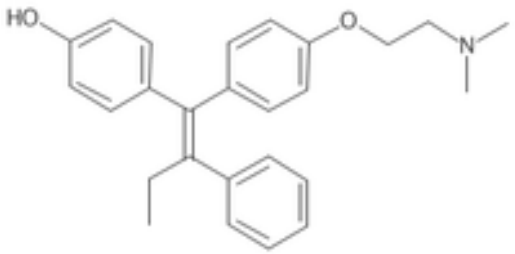


Figure 10: Φωτογραφία "4-hydroxytamoxifen"

δομένων εισόδου μπορούν να οδηγήσουν σε διαφορές στην τελική τιμή περιγραφέων . Ως αποτέλεσμα, στις περισσότερες περιπτώσεις, δύο υλοποιήσεις ενός περιγραφικού στοιχείου δεν δίνουν συνήθως την ίδια ακριβώς τιμή, αν και είναι συνήθως αρκετά παρόμοιες. Η ρητή εξήγηση των διαφορών μπορεί ή όχι να είναι δυνατή (συνήθως είναι πιο δύσκολη στις περιπτώσεις εμπορικών υλοποιήσεων για τις οποίες δεν υπάρχει διαθέσιμος πηγαίος κώδικας).

5.4 Περιγραφή μοριακής δομής με φωτογραφία

Για να προχωρήσουμε στις μοντελοποιήσεις χρησιμοποιήσαμε φωτογραφίες των smiles. Όπως έχει περιγραφεί παραπάμω μια χημική δομή μπορεί να περιγραφεί με μια φωτογραφία. Τα νευρωνικά δίκτυα και ειδικά τα νευρωνικά δίκτυα με συνελικτικές διαδικασίες τα τελευταία χρόνια έχουν αποτελέσει την πιο σύγχρωνη τεχνολογία στην επεξεργασία φωτογραφίας με πάρα πολύ καλά αποτελέσματα. Έχουν καταφέρει να επιλύσουν με τον καλύτερο τρόπο μέχρι στιγμής προβλήματα ταξινόμησης, αναγνώρισης, ανίχνευσης αντικειμένων σε φωτογραφίες. Έτσι οι φωτογραφίες των χημικών δομών θα χρησιμοποιηθούν ώς η είσοδος στα νευρωνικά δίκτυα με τελικό σκοπό την ταξινόμηση αυτών σε κλάσεις τοξικότητας. Παρακάτω υπάρχουν παραδείγματα αυτών των φωτογραφιών και της πληροφορίας που περιέχεται σε αυτές.

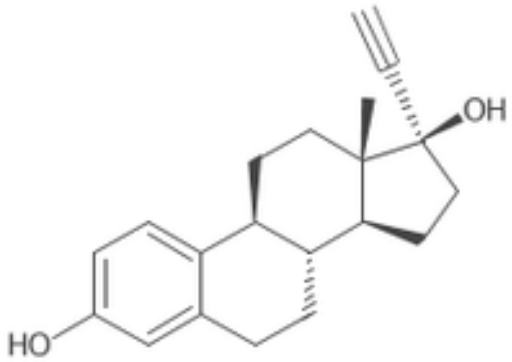


Figure 11: Φωτογραφία "ethinyl estradiol"

6 Τα δεδομένα

6.1 Ενδοκρινικοί Διαταράκτες

Οι ενδοκρινικοί διαταράκτες είναι χημικές ουσίες που μπορούν να επηρεάσουν τα ενδοκρινικά (ή ορμονικά) συστήματα σε ορισμένες δόσεις. Αυτές οι διαταραχές μπορούν να προκαλέσουν καρκινικούς όγκους, γενετικές ανωμαλίες και άλλες αναπτυξιακές διαταραχές. [1] Οποιοδήποτε σύστημα στο σώμα που ελέγχεται από ορμόνες μπορεί να εκτροχιασθεί από διαταραχές ορμονών. Συγκεκριμένα, οι ενδοκρινικοί διαταράκτες μπορεί να σχετίζονται με την ανάπτυξη μαθησιακών δυσκολιών, σοβαρή διαταραχή έλλειψης προσοχής, γνωστικά προβλήματα και προβλήματα ανάπτυξης του εγκεφάλου. παραμορφώσεις του σώματος (συμπεριλαμβανομένων των άκρων) καρκίνο του μαστού, καρκίνο του προστάτη, θυρεοειδή και άλλους καρκίνους. προβλήματα σεξουαλικής ανάπτυξης, όπως γυναικεία αρσενικά ή ανδρικές επιδράσεις στις γυναίκες κ.λπ.

Βρίσκονται σε πολλά οικιακά και βιομηχανικά προϊόντα. Οι ενδοκρινικοί διαταράκτες είναι ουσίες που "παρεμβαίνουν στη σύνθεση, την έκκριση, τη μεταφορά, τη δέσμευση, τη δράση ή την αποβολή φυσικών ορμονών στο σώμα που είναι υπεύθυνες για την ανάπτυξη, τη συμπεριφορά, τη γονιμότητα και τη συντήρηση της ομοιόστασης. Η ποικιλία των όρων που χρησιμοποιούνται για την περιγραφή αυτών των ουσιών αντικατοπτρίζει όχι μόνο ένα εύρος εννοιών αλλά και μια σειρά από υπονοήσεις, με τον ενδοκρινικό διαταράκτη να δίνει έμφαση στις επιβλαβείς επιπτώσεις, ενώ ο ορμονικά ενεργός παράγοντας ή η ξενορμόνη είναι πιο ουδέτεροι.

Δρούν μέσω πολλαπλών μηχανισμών, κυρίως ως συνδέτες σε ορμονικούς υποδοχείς, με αποτέλεσμα να παρεμβαίνουν στη λειτουργία των φυσικών ορμονών του ενδοκρινικού συστήματος: στη σύνθεση, στο μεταβολισμό και στη χημική σηματοδότηση, με καθοριστικές επιδράσεις στο αναπαραγωγικό και νευρικό σύστημα του ανθρώπινου οργανισμού. Ένας έμμεσος μηχανισμός είναι παραδείγματος χάριν η σύνδεση της διοξίνης ως αγωνιστής του υποδοχέα υδρογονάνθρακα αρυλίου, η οποία προκαλεί ενεργοποίηση του παράγοντα πυρηνικής μεταγραφής για την επαγωγή του γονιδίου CYP1A1 σε καρκινικά κύτταρα. Ένας άμεσος τρόπος σύνδεσης, αφορά στην πρόσδεση των PCBs (φθαλικοί εστέρες, οργανοχλωριωμένοι υδρογονάνθρακες). Η έκθεση του ανθρώπου στις χημικές αυτές ουσίες αρχίζει ήδη από την ενδομήτρια, εμβρυϊκή ζωή και είναι καθοριστική καθ' όλα τα αναπτυξιακά στάδια, αλλά και υπεύθυνη για συγγενείς ανωμαλίες και νευροαναπτυξιακές διαταραχές. Οι ενδοκρινικοί διαταράκτες επιφέρουν αλλαγές στο επιγενετικό προφίλ του ανθρώπου στην ενήλικη ζωή του, αλλαγές δηλαδή στη γενετική έκφραση και ρύθμιση στο άτομο, γεγονός που ενδέχεται να επιβαρύνει (μηχανισμοί μεταλλαξιγένεσης) και τις επόμενες γενιές.

Όλες οι χημικές ενώσεις, είτε υπάρχουν στη φύση, είτε είναι συνθετικές, που μπορούν να επηρεάσουν το ορμονικό σύστημα του ανθρώπου ή των ζώων αναφέρονται ως ενδοκρινικοί διαταράκτες (Endocrin Disruptors Compounds, EDCs), όρος που χρησιμοποιείται για πρώτη φορά το 1992 από τους Theo Colborn και Peter Thomas. Στη συνέχεια, ο όρος αποσαφηνίζεται στο Ευρωπαϊκό Workshop (European Commission, 1996) ως "Weybridge" και περιλαμβάνει τις δύο παρακάτω κατηγορίες: Δυνητικός ενδοκρινικός διαταράκτης είναι μία εξωγενής ουσία ή μίγμα ουσιών που ενδέχεται να προκαλέσει ενδοκρινική διαταραχή

σε έναν οργανισμό ή στους απογόνους του ή σε (ύπο)πληθυσμούς. Ενδοκρινικός διαταράκτης είναι μία εξωγενής ουσία ή μίγμα ουσιών που αλλοιώνει τη λειτουργία ή τις λειτουργίες του ενδοκρινικού συστήματος και, ως εκ τούτου, προκαλεί ανεπιθύμητες δράσεις σε έναν οργανισμό ή στους απογόνους του ή σε (ύπο)πληθυσμούς. Τέλος, η Διεύθυνση Περιβαλλοντικής Προστασίας των Η.Π.Α. (U.S. Environmental Protection Agency, EPA, 1997) προτείνει έναν ακόμη πληρέστερο ορισμό του ενδοκρινικού διαταράκτη: ενδοκρινικός διαταράκτης είναι ένας εξωγενής παράγοντας που παρεμβαίνει στη σύνθεση, απέκκριση, μεταφορά, πρόσδεση, κίνηση ή εξάλειψη των φυσικών ορμονών στο σώμα και είναι υπεύθυνος για τη μεταβολή της αναπαραγωγής, της ανάπτυξης ή/και της συμπεριφοράς.

6.2 Binding affinity / Συγγένεια πρόσδεσης

Η συγγένεια πρόσδεσης αναφέρεται στην ισχύ της πρόσδεσης διάφορων, συνθετικών, ακόμη και φαρμακευτικών μορίων με εκτεθειμένα στοιχεία σε ικριώματα ιστών δεδομένης αρχιτεκτονικής. Διαδραματίζει καθοριστικό ρόλο κατά τη φαρμακοκινητική. Πιο συγκεκριμένα, η συγγένεια πρόσδεσης διαφόρων ικριωμάτων ιστών με φαρμακευτικούς παράγοντες, δε θα πρέπει να είναι ούτε ιδιαίτερα χαμηλή, ούτε ιδιαίτερα υψηλή. Αυτό κυρίως γιατί αν η συγγένεια πρόσδεσης του παραπάνω συστήματος είναι ιδιαίτερα χαμηλή, υπάρχει ο κίνδυνος ο φαρμακευτικός παράγοντας, αντί να προσδεθεί να διαχυθεί στους γειτονικούς ιστούς με μη ελεγχόμενο τρόπο, ώστε τελικά να παράξει ανεπιθύμητα, τοξικά αποτελέσματα (Tarun et al., 2012). Εάν πάλι η συγγένεια πρόσδεσης είναι ιδιαίτερα υψηλή, μία ανεπαρκής ποσότητα φαρμάκου ενδέχεται να απελευθερώνεται από το σύστημα ικριώματος-φαρμάκου στο απαιτούμενο χρονικό διάστημα.

Σχεδόν κάθε διαδικασία στη βιολογία μπορεί να αποδοθεί σε αλληλεπίδραση μεταξύ μορίων. Οι επιστήμονες χρησιμοποιούν μετρήσεις (το Kd) για να προσδιορίσουν ή "να ταξινομήσουν δεσμευτικές αντιδράσεις που συχνά μεταφράζονται σε βιολογική λειτουργία ή να αποκαλύψουν τη συνάφεια των στόχων που εξετάζονται. Όσα περισσότερα γνωρίζουμε για αυτές τις αλληλεπιδράσεις, τόσο περισσότερο καταλαβαίνουμε τα βιολογικά συστήματα στα οποία λειτουργούν με το περίπλοκο δίκτυο μοριακών οδών που ελέγχουν διάφορες κυτταρικές διεργασίες. Ο ακριβής χαρακτηρισμός των βιομοριακών αλληλεπιδράσεων σε ένα βιολογικό σύστημα αποτελεί σημαντικό ακρογωνιαίο λίθο στη βασική έρευνα. Στην εφαρμοσμένη επιστήμη, η μέτρηση της δεσμευτικής συγγένειας των αλληλεπιδράσεων είναι απαραίτητη προϋπόθεση για την ανάπτυξη νέων προϊόντων, όπως φάρμακα, ένζυμα ή βιοδείκτες. Η μέτρηση της συγγένειας δέσμευσης έχει πολλές εφαρμογές, συμπεριλαμβανομένης της αναγνώρισης και διαλογή μικρών και / ή μεγάλων μορίων, παρακολούθηση της ρύθμισης των κυτταρικών οδών, δοκιμές σχέσεων δομής-λειτουργίας, και βελτιστοποίηση της ανάπτυξης δοκιμασιών που εξετάζουν το αλληλεπίδραση δύο μορίων.

Την συγκεκριμένη ιδιότητα θα προσπαθήσουμε και να μοντελοποίσουμε. Συγκεκριμένα συγκεντρώθηκαν τα αποτελέσματα απο περιπου 1400 in vivo πειράματα με υπολογισμένη την τιμή log RBA (λογαριθμοποιημένη συγγένια πρόσδεσης).

Το σύνολο των δεδομένων αποτελείται από 1.459 χημικές δομές. Σε αυτές έχουν δημιουργηθεί οι φωτογραφικές αποτυπώσεις αυτών. Για να προχωρήσουμε σε μοντελοποίηση κατηγοριοποιούμε τα δεδομένα σε 3 κλάσεις με βάση την πειραματική τους απόκριση. Συγκεκριμένα η πρώτη κλάση περιέχει τις δομές που έχουν τιμές απόκρισης στο διάστημα [-3.328, -0.26]. Η δεύτερη κλάση έχει τιμές απόκρισης [-0.259, 0.824] και η τρίτη έχει [0.826, 2.857]. Παρακάτω υπάρχουν παραπάνω λεπτομέριες για το πως διαχειρίστηκαν αυτά απο τα μοντέλα.

7 Μεθοδολογία και αρχιτεκτονικές μοντέλων

Παραπάνω έχουν αναλυθεί οι βασικοί αλγόριθμοι εκπαίδευσης και οι βασικές αρχιτεκτονικές νερωνικών δικτύων. Στο κεφάλαιο θα παρουσιαστούν οι αρχιτεκτινικές που χρησιμοποιήθηκαν σε λεπτομέρια καθώς και η διαδικασία εκπαίδευσης.

7.1 Αρχιτεκτονικές δικτύων

Τα τελευταία χρόόνια έχουν υπάρχει πολλές αρχιτεκτονικές δικτύων αναγώρισης φωτογραφίας. Όλες βασίζονται στην πρώτη αρχιτεκτονική LeNet [?] και επόμενες όπως Inception και ResNet προσπαθούν να επεκτείνουν το βάθος τον νευρωνικών δικτύων με περισσότερα στρώμματα. Πολές τέτοιες θα χρησιμοποιηθούν και θα συγκριθούν παρακάτω.

7.2 LeNet / ImageNet

7.2.1 Γιατί είναι σημαντικό

Το LeNet-5 χρησιμοποιήθηκε σε μεγάλη κλίμακα για την αυτόματη ταξινόμηση των χειρόγραφων ψηφίων σε τραπεζικούς ελέγχους στις Ηνωμένες Πολιτείες. Αυτό το δίκτυο είναι ένα συνελικτικό νευρωνικό δίκτυο (CNN). Τα CNN αποτελούν το θεμέλιο του σύγχρονου οράματος υπολογιστών βασισμένου σε βάθος μάθησης. Τα δίκτυα αυτά βασίζονται σε 3 βασικές ιδέες: τοπικά πεδία ευαισθητοποίησης, κοινά βάρη και υποδειγματοληψία χώρου. Τα τοπικά πεδία υποδοχής με κοινό βάρη αποτελούν την ουσία του

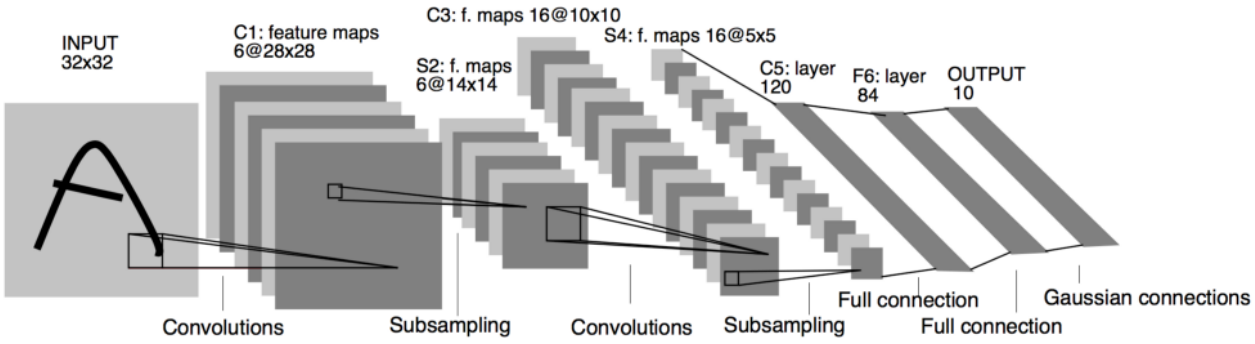


Figure 12: LeNet

συνελικτικού στρώματος και οι περισσότερες αρχιτεκτονικές που περιγράφονται παρακάτω χρησιμοποιούν συνελικτικά στρώματα σε μία ή την άλλη μορφή.

Ένας άλλος λόγος για τον οποίο το LeNet είναι μια σημαντική αρχιτεκτονική είναι ότι πριν από την ανακάλυψή του, η αναγνώριση χαρακτήρων έγινε κυρίως με τη χρήση μηχανικών χαρακτηριστικών με το χέρι, ακολουθούμενη από ένα μοντέλο μηχανικής μάθησης για να μάθουμε να ταξινομούμε χειροκίνητα χαρακτηριστικά. Το LeNet κατέστησε περιττό τα μηχανικά χαρακτηριστικά του μηχανήματος, επειδή το δίκτυο μαθαίνει αυτόματα την καλύτερη εσωτερική αναπαράσταση από τις πρώτες εικόνες.

7.2.2 Σύντομη περιγραφή

Με τα σύγχρονα πρότυπα, το LeNet-5 είναι ένα πολύ απλό δίκτυο. Έχει μόνο 7 στρώματα, μεταξύ των οποίων υπάρχουν 3 στρώματα περιελίξεων (C1, C3 και C5), 2 στρώματα υπο-δειγματοληψίας (pooling) (S2 και S4) και 1 πλήρως συνδεδεμένο στρώμα (F6), τα οποία ακολουθούνται από την έξοδο στρώμα. Τα στρώματα της περιστροφής χρησιμοποιούν 5 έως 5 συνέλιες με το βήμα 1. Τα επίπεδα υπο-δειγματοληψίας είναι 2 έως 2 μέσου όρου συγκεντρώσεων. Οι ενεργοποιήσεις σιγμοειδούς Tanh χρησιμοποιούνται σε όλο το δίκτυο. Υπάρχουν πολλές ενδιαφέρουσες αρχιτεκτονικές επιλογές που έγιναν στο LeNet-5 οι οποίες δεν είναι πολύ συχνές στη σύγχρονη εποχή της βαθιάς μάθησης.

Πρώτον, μεμονωμένοι σπειροειδείς πυρήνες στο στρώμα C3 δεν χρησιμοποιούν όλα τα χαρακτηριστικά που παράγονται από το στρώμα S2, το οποίο είναι πολύ ασυνήθιστο με το σημερινό πρότυπο. Ένας λόγος γι 'αυτό είναι να γίνει το δίκτυο λιγότερο απαιτητικό από υπολογιστική άποψη. Ο άλλος λόγος ήταν να κάνει τους συνελικτικούς πυρήνες να μάθουν διαφορετικά πρότυπα. Αυτό έχει νόημα: εάν οι διαφορετικοί πυρήνες λαμβάνουν διαφορετικές εισόδους, θα μάθουν διαφορετικά πρότυπα.

Δεύτερον, το στρώμα εξόδου χρησιμοποιεί 10 νευρώνες ευκλείδειας ακτινικής βάσης που υπολογίζουν την απόσταση L2 μεταξύ του διανύσματος εισόδου της διάστασης 84 και των χειροκίνητα προκαθορισμένων διανυσμάτων βάρους της ίδιας διάστασης. Ο αριθμός 84 προέρχεται από το γεγονός ότι ουσιαστικά τα βάρη αντιπροσωπεύουν μια δυαδική μάσκα 7 * 12, μία για κάθε ψηφίο. Αυτό αναγκάζει το δίκτυο για να μετασχηματίσει την εικόνα εισόδου σε μια εσωτερική αναπαράσταση που θα κάνει τις εξόδους του στρώματος F6 όσο το δυνατόν πλησιέστερα στα χέρια κωδικοποιημένα βάρη των 10 νευρώνων του στρώματος εξόδου.

Το LeNet-5 ήταν σε θέση να επιτύχει ποσοστό σφάλματος κάτω από το 1% στο σύνολο δεδομένων MNIST, το οποίο ήταν πολύ κοντά στην τεχνολογία της εποχής εκείνης (που παράγεται από ένα ενισχυμένο σύνολο τριών δικτύων LeNet-4)

[?]

7.3 AlexNet / ImageNet

AlexNet AlexNet Αρχιτεκτονική [1] AlexNet Layers Λεπτομέρειες [2] Το AlexNet είναι η πρώτη αρχιτεκτονική νευρωνικών δικτύων μεγάλης κλίμακας που κάνει καλά την ταξινόμηση του ImageNet. Το AlexNet μπήκε στον διαγωνισμό και κατάφερε να ξεπεράσει όλα τα προηγούμενα μη βαθιά μοντέλα που βασίζονται στη μάθηση με σημαντικό περιθώριο.

Η αρχιτεκτονική AlexNet είναι ένα στρώμα μετατροπής ακολουθούμενο από τη συγκέντρωση στρώματος, την ομαλοποίηση, τον κανόνα conv-pool, και στη συνέχεια μερικά ακόμη στρώματα conv, ένα στρώμα pooling και στη συνέχεια αρκετά πλήρως συνδεδεμένα στρώματα μετά. Στην πραγματικότητα μοιάζει πολύ με το δίκτυο LeNet. Υπάρχουν συνολικά μόνο περισσότερα στρώματα. Υπάρχουν πέντε από αυτά τα στρώματα μετατροπής και δύο πλήρως συνδεδεμένα στρώματα πριν το τελικό πλήρως συνδεδεμένο στρώμα πηγαίνει στις κλάσεις εξόδου.

Το AlexNet εκπαιδεύτηκε στο ImageNet, με εισόδους σε εικόνες μεγέθους 227 x 227 x 3. Αν δούμε αυτό το πρώτο στρώμα που είναι ένα στρώμα μετατροπής για το AlexNet, είναι 11 x 11 φίλτρα, 96 από αυτά εφαρμόστηκαν στο βήμα 4. Είχα 55 x 55 x 96 στην παράμετρο εξόδου και 35K σε αυτό το πρώτο στρώμα. Το δεύτερο στρώμα είναι ένα στρώμα συγκέντρωσης και σε αυτή την περίπτωση έχουμε 3 φίλτρα 3 x 3 που εφαρμόζονται στο βήμα 2. Ο όγκος εξόδου του στρώματος συγκέντρωσης είναι 27 x 27 x 96 με και την παράμετρο 0 για μάθηση. Το στρώμα συγκέντρωσης δεν μαθαίνει τίποτα επειδή οι παράμετροι είναι τα βάρη που προσπαθούν να μάθουν. Τα περιστροφικά στρώματα έχουν βάρη που μαθαίνουμε, αλλά η συγκέντρωση όλων που κάνουμε είναι ένας κανόνας, κοιτάζουμε την περιοχή συγκέντρωσης και παίρνουμε το μέγιστο. Έτσι, δεν υπάρχουν παράμετροι που έχουν μάθει.

Υπάρχουν 11 x 11 φίλτρα στην αρχή, στη συνέχεια πέντε με πέντε και μερικά τρία με τρία φίλτρα. Στο τέλος, έχουμε δύο πλήρως συνδεδεμένα στρώματα μεγέθους 4096 και τέλος, το τελευταίο στρώμα, το FC8 πηγαίνει στο softmax, το οποίο πηγαίνει στις 1000 κατηγορίες ImageNet. Αυτή η αρχιτεκτονική είναι η πρώτη χρήση της μη γραμμικότητας του ReLu.

Υπερπαραμετρική: Αυτή η αρχιτεκτονική είναι η πρώτη χρήση της μη γραμμικότητας του ReLU. Το AlexNet χρησιμοποιεί επίσης ένα επίπεδο ομαλοποίησης. Στην αύξηση των δεδομένων, η ALexNet χρησιμοποίησε το flipping, jittering, περικοπή, κανονικοποίηση των χρωμάτων και αυτά τα πράγματα. Άλλες παράμετροι είναι η απόρριψη με 0,5, η SGD Momentum με 0,9, ο ρυθμός αρχικής εκμάθησης 1e-2 και πάλι μειώνεται κατά 10 όταν η ακρίβεια επικύρωσης είναι επίπεδη. Η τακτοποίηση που χρησιμοποιείται σε αυτό το δίκτυο είναι L2 με αποσύνθεση βάρους 5e-4. Εκπαιδεύτηκε σε GPU GTX580 που περιέχει 3GB μνήμης.

Έχει ποσοστό σφάλματος 16,4 στην προβολή Visual Visual Recognition Large Image Challenge (ILSVRC).

Ο AlexNet ήταν ο νικητής της ταξινόμησης του δείκτη αναφοράς ImageNet Large Scale Visual Recognition Challenge (ILSVRC) το έτος αναφοράς το 2012.

VGGNet Αρχιτεκτονική VGG16 [3] Λεπτομέρειες στρώσεων VGG 16 και VGG 19 [2] Το 2014 υπάρχουν δύο αρχιτεκτονικές που ήταν πιο διαφορετικές μεταξύ τους και έκαναν άλλο άλμα στην απόδοση και η κύρια διαφορά με αυτά τα δίκτυα με τα βαθύτερα δίκτυα.

Το VGG 16 είναι αρχιτεκτονική 16 επιπέδων με ένα ζευγάρι στρώματα συνέλιξης, στρώματα συγκεντρώσεων και στο τέλος πλήρως συνδεδεμένο στρώμα. Το δίκτυο VGG είναι η ιδέα των πολύ βαθύτερων δικτύων και με πολύ μικρότερα φίλτρα. Το VGGNet αύξησε τον αριθμό των στρώσεων από οκτώ επίπεδα στο AlexNet. Αυτή τη στιγμή είχε μοντέλα με παραλλαγές 16 έως 19 στρώσεων του VGGNet. Ένα βασικό στοιχείο είναι ότι αυτά τα μοντέλα κράτησαν πολύ μικρά φίλτρα με 3 x 3 conv σε όλη τη διαδρομή, το οποίο είναι βασικά το μικρότερο μέγεθος φίλτρου μετατροπής που κοιτάζει λίγο από τα γειτονικά εικονοστοιχεία. Και κράτησαν αυτή την πολύ απλή δομή των 3 x 3 convts με την περιοδική συγκέντρωση σε όλη τη διαδρομή μέσω του δικτύου.

Η VGG χρησιμοποίησε μικρά φίλτρα λόγω λιγότερων παραμέτρων και συγκέντρωσε περισσότερα από αυτά, αντί να έχει μεγαλύτερα φίλτρα. Το VGG έχει μικρότερα φίλτρα με μεγαλύτερο βάθος αντί για μεγάλα φίλτρα. Έχει καταλήξει να έχει το ίδιο αποτελεσματικό πεδίο δεκτικότητας σαν να έχετε μόνο ένα 7 x 7 στρώματα περιελίξεων.

Το VGGNet έχει στρώματα μετατροπής και ένα στρώμα συγκέντρωσης μερικά ακόμα στρώματα μετατροπής, συγκέντρωση στρώματος, μερικά ακόμη στρώματα μετατροπής και ούτω καθεξής. Η αρχιτεκτονική VGG έχει τον 16 συνολικό αριθμό συνθετικών και πλήρως συνδεδεμένων στρωμάτων. Έχει 16 σε αυτή την περίπτωση για VGG 16, και 19 για VGG 19, είναι απλώς μια πολύ παρόμοια αρχιτεκτονική, αλλά με μερικά ακόμα στρώματα conv σε εκεί.

VGG16 Παράμετροι [2] Έτσι, αυτό είναι πολύ δαπανηρό υπολογισμό με 138M συνολική παράμετρο και κάθε εικόνα έχει μια μνήμη 96MB που είναι τόσο πολύ μεγάλη από μια κανονική εικόνα. Έχει μόλις 7,3 ποσοστό σφάλματος στην πρόκληση ILSVRC.

Το VGGNet ήταν ο επιλαχόντης της ταξινόμησης του πρότυπου οπτικής αναγνώρισης ImageNet μεγάλης κλίμακας (ILSVRC) το έτος αναφοράς το 2014.

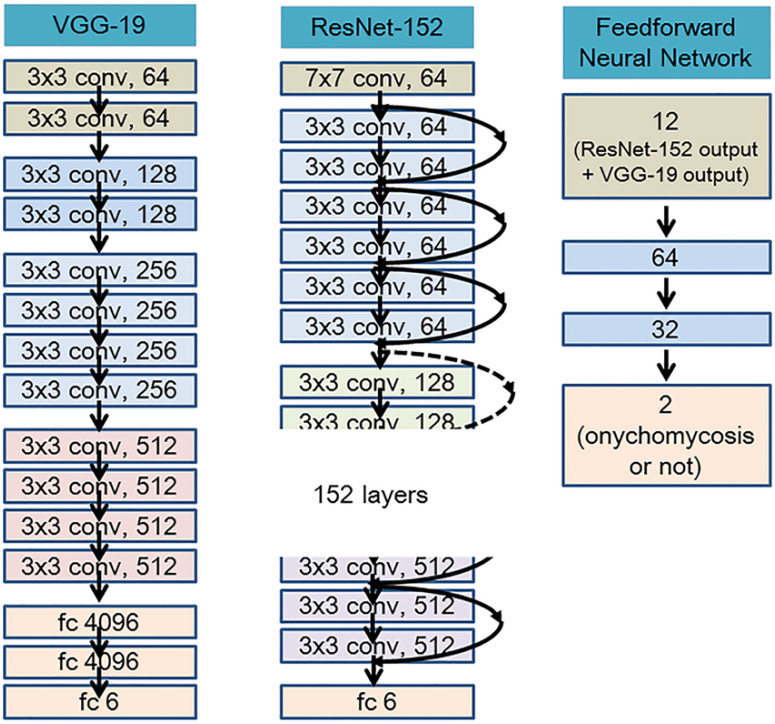


Figure 13: ResNet modules

7.4 ResNet

Αρχιτεκτονική ResNet και λεπτομέρειες στρώματος Το βασικό στοιχείο βάσης του ResNet είναι το μπλοκ υπολοίπου. Καθώς μπαίνουμε βαθύτερα στο δίκτυο με μεγάλο αριθμό στρώσεων, ο υπολογισμός γίνεται πιο περίπλοκος. Αυτά τα στρώματα τοποθετούνται το ένα πάνω στο άλλο και κάθε στρώμα προσπαθεί να μάθει κάποια υποκείμενη χαρτογράφηση της επιθυμητής λειτουργίας και αντί να έχει αυτά τα μπλοκ, προσπαθούμε να τοποθετήσουμε μια υπολειπόμενη χαρτογράφηση.

Με λίγα λόγια, καθώς μπαίνουμε βαθύτερα στο δίκτυο, είναι τόσο δύσκολο να μάθουμε το H (X) καθώς έχουμε μεγάλο αριθμό στρώσεων. Έτσι εδώ χρησιμοποιήσαμε τη σύνδεση παράκαμψης και μάθαμε F (x) την άμεση είσοδο του x ως την τελική έξοδο. Έτσι F (x) ονομάζεται ως υπολειπόμενο.

Στο ResNet, συγκεντρώνει όλα αυτά τα μπλοκ μαζί πολύ βαθιά. Ένα άλλο πράγμα με αυτήν την πολύ βαθιά αρχιτεκτονική είναι ότι επιτρέπει έως και 150 στρώματα βαθιά από αυτό, και τότε αυτό που κάνουμε είναι να στοιβάζουμε όλα αυτά τα στρώματα περιοδικά. Επίσης, διπλασιάζουμε τον αριθμό των φίλτρων και μειώνουμε το διάστημα χρησιμοποιώντας το δεύτερο βήμα. Στο τέλος, μόνο το πλήρως συνδεδεμένο στρώμα 1000 με τις κλάσεις εξόδου.

Υπερπαραμέτρους: Στο ResNet, χρησιμοποιεί Κανονικοποίηση παρτίδας μετά από κάθε στρώμα conv. Χρησιμοποιεί επίσης την αρχικοποίηση Xavier με SGD + Momentum. Ο ρυθμός εκμάθησης είναι 0,1 και διαιρείται με 10, καθώς το σφάλμα επικύρωσης γίνεται σταθερό. Επιπλέον, το μέγεθος παρτίδας είναι 256 και η αποσύνθεση βάρους είναι 1e-5. Το σημαντικό είναι ότι δεν υπάρχει εγκατάλειψη στο ResNet.

Η ResNet εξασφάλισε την 1η θέση στον ανταγωνισμό ILSVRC και COCO 2015 με ποσοστό σφάλματος μόλις 3,6% του ποσοστού σφάλματος. (Καλύτερη από την ανθρώπινη απόδοση !!!)

[22]

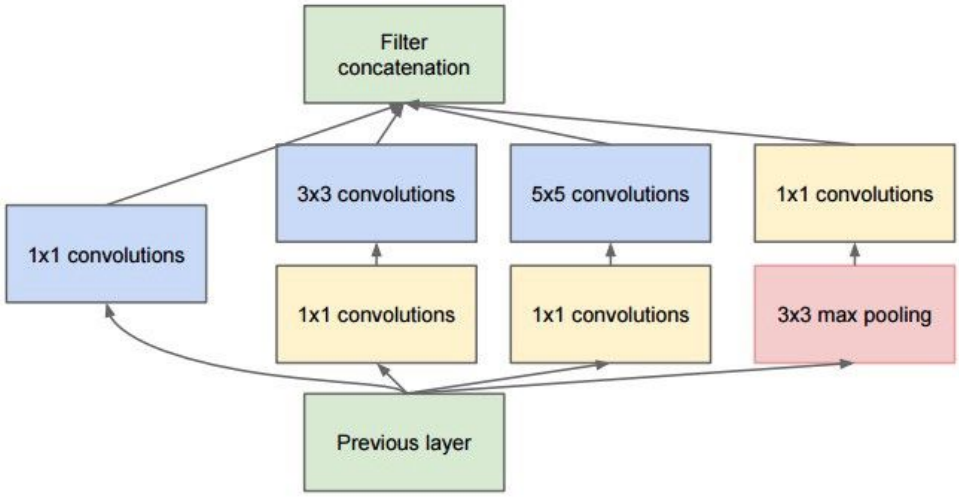


Figure 14: ResNet modules

7.5 Inception net

Η αρχιτεκτονική inception v3 είναι ένα ευρέως χρησιμοποιούμενο μοντέλο αναγνώρισης εικόνων που έχει αποδειχθεί ότι επιτυγχάνει ακρίβεια μεγαλύτερη από 78,1% στο σύνολο δεδομένων ImageNet. Το μοντέλο είναι ο συνδυασμός πολλών ιδεών που αναπτύχθηκαν από πολλούς ερευνητές κατά τη διάρκεια των ετών.

Το ίδιο το μοντέλο αποτελείται από συμμετρικά και ασυμμετρικά δομικά στοιχεία, συμπεριλαμβανομένων συνελισσών, μέσου όρου συγκέντρωσης, μέγιστης συγκέντρωσης, εγκατάλειψης και πλήρως συνδεδεμένων στρωμάτων. Το Batchnorm χρησιμοποιείται εκτενώς σε όλο το μοντέλο και εφαρμόζεται σε εισόδους ενεργοποίησης. Το συνολικό λάθος υπολογίζεται μέσω Softmax.

Έναρξη εργασίας με Factorizing Convolutions. Συγκεντρώσεις συντεταγμένων που χρησιμοποιούνται για τη μείωση του αριθμού των συνδέσεων και των παραμέτρων που θα μάθουν. Αυτό θα αυξήσει την ταχύτητα και θα δώσει καλή απόδοση. [**?**]

7.6 Συναρτήσεις λάθους

Μια βασική συνάρτηση λάθους που χρησιμοποιείται για την εκαπίδευση νευρωνικών δικτύων ειναι η συνάρτηση cross entropy. Στη θεωρία των πληροφοριών, η διασταυρούμενη εντροπία (cross entropy) μεταξύ δύο κατανομών πιθανοτήτων p και q πάνω από το ίδιο υποκείμενο σύνολο συμβάντων μετράει τον μέσο αριθμό των bits που χρειάζονται για την ταυτοποίηση ενός συμβάντος που προέρχεται από το σύνολο εάν ένα σχήμα κωδικοποίησης που χρησιμοποιείται για το σετ είναι βελτιστοποιημένο για μια εκτιμώμενη κατανομή πιθανότητας q, αντί για την πραγματική κατανομή p.

$$H(p, q) = -E_p[\log q], \tag{29}$$

8 Σύνολο Δεδομένων

Το σύνολο των δεδομένων αποτελείται από 1.459 χημικές δομές. Σε αυτές έχουν δημιουργηθεί οι φωτογραφικές αποτυπώσεις αυτών. Για αυτές υπάρχουν οι υπολογισμένες απο πειράματα τιμές για το Relative Binding Affinity σε λογαριθμοποιημένη κλίμακα. Για να προχωρήσουμε σε μοντελοποίηση κατηγοριοποιούμε τα δεδομένα σε 3 κλάσεις με βάση την πειραματική τους απόκριση. Συγκεκριμένα η πρώτη κλάση περιέχει τις δομές που έχουν τιμές απόκρισης στο διάστημα [-3.328, -0.26]. Η δεύτερη κλάση έχει τιμές απόκρισης [-0.259, 0.824] και η τρίτη έχει [0.826, 2.857]. Παρακάτω παρατίθεται ο πίνακας με τις τιμές και τις κλάσεις.

	Κατώτατη τιμή	Ανώτατη τιμή
Κλάση 0	-3.328	-0.26
Κλάση 1	-0.259	0.824
Κλάση 2	0.826	2.857

Οι κλάσεις για να μπορέσουν να χρησιμοποιηθούν στην εκπαίδευση ενός νευρωνικού δικτύου έχουν κωδικοποιηθεί ως One Hot Encoded. Συγκεκριμένα η πρώτη κλάση έχει κωδικοποιηθεί με το διάνυσμα $[1., 0., 0.]$. Αντίστοιχα η δεύτερη και η τρίτη με τα διανύσματα $[0., 1., 0.]$ και $[0., 0., 1.]$ αντίστοιχα. Με αυτό τον τρόπο μπορούμε να υπλογίσουμε την τιμη της συνάρτησης λάθους που έχει περιγραφεί παραπάνω.

Τα δεδομένα για να προχωρήσει η εκπαίδευση συγκεντρώνονται σε τεμάχια φωτογραφιών και κλάσης ανάλογα με τις ανάγκες της εκπαίδευσης. Καθοριστικός ρόλος και το μέγεθος της μνήμης του εξοπλισμού διαθέσιμο στον χρήστη. Στην δικιά μας περίπτωση χρησιμοποίσαμε τυχαία δειγματοληψεία απο το σύνολο των φωτογραφιών σε τεμάχια των 42. Τα τεμάχια "προωθούνται" στο νευρωνικό δίκτυο για να προχωρήσει η διαδικασία εκπαίδευσης.

Επίσης να επισημάνουμε την κανονικοποίηση των φωτογραφιών. Οι τιμές των pixel αυτών μετασχηματίστηκαν με σκοπό να έχουν μέση τιμή 0 και διασπορά 1.

9 Αποτελέσματα και συζήτηση

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της προσέγγισης μας. Προχωρήσαμε με δύο αρχιτεκτονικές νευρωνικών δικτύων. Στην μια περίπτωση χρησιμοποιήθηκαν αρχιτεκτονικές τύπου ImageNet όπως περιγράφηκαν παραπάνω και στην δεύτερη νευρωνικά δίκτυα με Residual blocks.

9.1 Μοντελοποιήσεις με ImageNet

Τα δεδομένα όπως παρουσιάστηκαν αποτελούντε απο εικόνες χημικών μορίων αποτυπωμένα σε δισδιάστατη μορφή. Ανάλογα με το μέγεθος της φωτογραφίας χρησιμοποιήθηκε και αντίστοιχο νευρωνικό δίκτυο. Οι φωτογραφίες σαν είσοδο είχαν διάσταση 128 * 128 ή 200 * 200 και τέλος 256 * 256. Αντίστοιχα διαμορφώθηκε το νευρωνικό δίκτυο. Τρία συνελικτικά επίπεδα ακολοθούσαν την είδοδο του δικτύου. Μετά απο κάθε συνελικτικό επίπεδο ακολουθεί ένα 2 * 2 επίπεδο pooling που δημιουργεί το down sampling της ειδόσου του δικτύου. συνεπώς ανάλογα με την είσοδο έχουμε τελευταίο συνελικτικό επίπεδο με φίλτρα 16 * 16 επί το βάθος των φίλτρων (παραδίγματος χάρην 56) , 25 * 25 επί το βάθος των φίλτων και τέλος στην περίπτωση της 256 * 256 εισόδου 32 * 32 επί το βάθος των φίλτρων. Ακολουθεί η διαδικασία της επιπεδοποίησης και τα δύο τελευταία επίπεδα του δικτύου τα πλήρως συνδεδεμένα επίπεδα που στις μοντελοποίησεις μας αποτελούταν απο 1024 κόμβους. Στην έξοδο του δικτύου έχουμε τις τρείς κλάσεις που θέλουμε να προβλέψουμε. Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε σε όλα τα νευρωνικά δίκτυα είναι η ReLu (rectified linear unit (ReLU).

$$f(x) = max(0, x) \tag{30}$$

Οι συναρτήσεις βελτιστοποίησης ήταν 2. Χρησιμοποίθηκε η κλασική μέθοδος Gradient descent και ο AdamOptimizer. Η τιμή learning rate ήταν σχετικά μεγάλη και σταθερή, συγκεκριμένα 0.3 . Επίσης σε κάθε επίπεδο του δικτύου χρησιμοποίθηκε η μέθοδος dropout για αποφυγή φαινομένων overfitting του μοντέλου.

Κατα την εκπαίδευση παρατηρήθηκε πως ιδιαίτερο ρόλο έπαιξε η τιμή learning rate. Όσο μικρότερη τόσο δυσκολότερη η σύγκλιση του μοντέλου και για να ξεκινήσει η εκπαίδευση και να προχωράει η σύγκλιση θα πρέπει να ξεκινήσει η διαδικασία εκπαίδευσης απο σχετικά μεγάλη τιμή, τουλάχιστον 0.3 . Το μέγεθος της εικόνας εισόδου δεν έπαιξε σημαντικό ρόλο καθότι η ακρίβεια του μοντέλου άλλαξε κατά το πολύ 1 % χωρίς να παρατηρείται κάποιο πατερν.

Παρακάτω εμφανίζονται τα διαγράμματα ακρίβειας της εκπαίδευσης. Στο πρώτο βλέπουμε την ακρίβεια όπως διαμορφώνεται κατα την εκπαίδευση στα δεδομένα εκπαίδευσης. Στα δεδομένα πάνω στα οποία δηλαδή εκπαιδεύουμε τα ίδια τα μοντέλα. Όπως περιμέναμε από κάποιο σημείο και μετά (πιο συγκεκριμένα μετά το 5000 βήμα) η ακρίβεια πλησιάζει το 1.

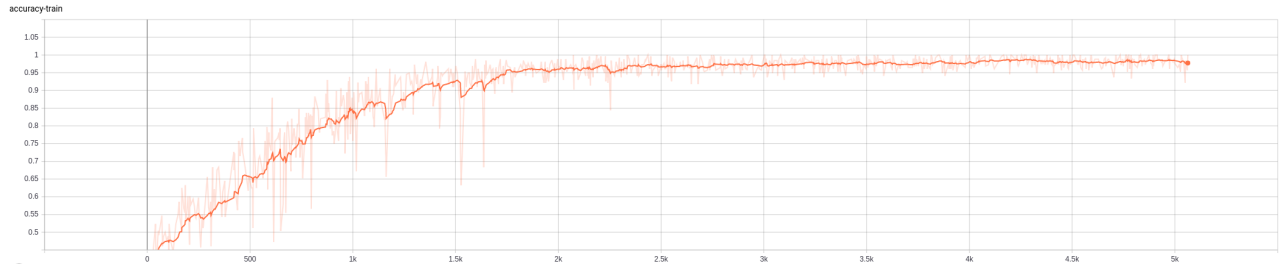


Figure 15: Ακρίβεια μοντέλου στα δεδομένα εκπαίδευσης

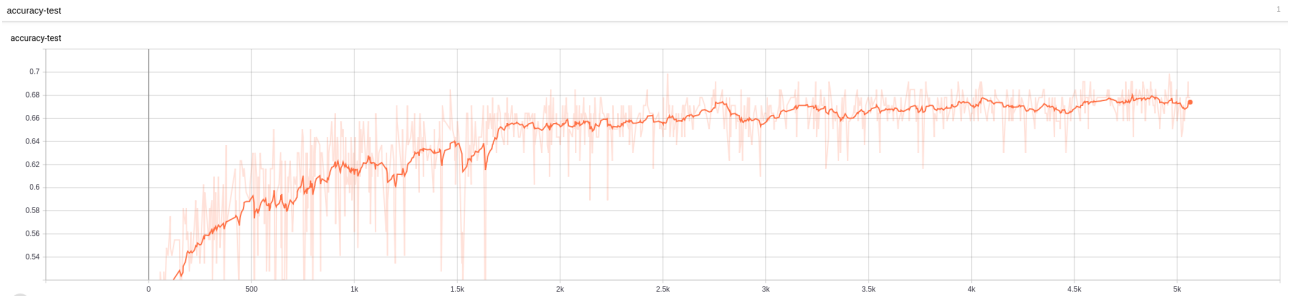


Figure 16: Ακρίβεια μοντέλου στα δεδομένα επαλήθευσης

Στο παραπάνω γράφημα παρουσιάζεται το γράφημα ακρίβειας για ένα μοντέλο στα δεδομένα επαλήθευσης. Στα δεδομένα πάνω στα οποία δεν έχει εκπαιδευτεί δηλαδή το μοντέλο το ίδιο. Σε αυτή την περίπτωση βλέπουμε πως η ακρίβεια φτάνει το εβδομήνα της εκατό . Αυτή είναι και μια απο τις καλύτερες αποδώσεις μοντέλου που είδαμε κατα την εκπαίδευση και ήρθε απο μοντέλα τύπου Image Net.

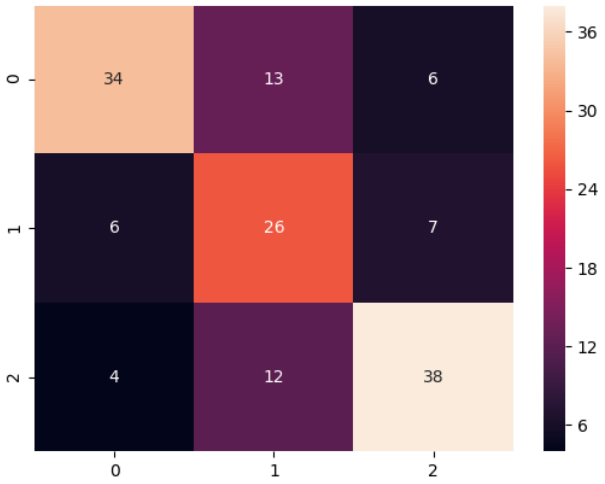


Figure 17: Πίνακας σύγχυσης προβλέψεων

Στον τομέα της μηχανικής μάθησης και συγκεκριμένα το πρόβλημα της στατιστικής ταξινόμησης, ένας πίνακας σύγχυσης, επίσης γνωστός ως πίνακας σφάλματος, [7] είναι μια συγκεκριμένη διάταξη πίνακα που επιτρέπει την οπτικοποίηση της απόδοσης ενός αλγορίθμου, συνήθως μια εποπτευόμενη μάθηση (σε μη επιτηρούμενη μάθηση συνήθως ονομάζεται αντίστοιχος πίνακας). Κάθε σειρά του πίνακα αντιπροσωπεύει τις εμφανίσεις σε μια προβλεπόμενη κλάση, ενώ κάθε στήλη αντιπροσωπεύει τις παρουσίες σε μια πραγματική κλάση (ή το αντίστροφο). [2] Το όνομα προέρχεται από το γεγονός ότι καθιστά εύκολο να διαπιστωθεί εάν το σύστημα προκαλεί σύγχυση σε δύο κατηγορίες (δηλ. Συνήθως εσφαλμένη σήμανση μεταξύ τους.

Παρακάτω βλέπουμε τον πίνακα σύγχυσης (Confussion matrix) των προβλέψεων.

Ο συντελεστής συσχέτισης Matthews (MCC) χρησιμοποιείται στη μηχανική μάθηση ως μέτρο της ποιότητας των δυαδικών ταξινομήσεων (δύο κατηγοριών), που εισήχθη από τον βιοχημικό Brian W. Matthews το 1975. Αν και το MCC είναι ισοδύναμο με τον συντελεστή ph του Karl Pearson, που αναπτύχθηκε δεκαετίες νωρίτερα, ο όρος MCC χρησιμοποιείται ευρέως στον τομέα της βιοπληροφορικής.

Ο συντελεστής λαμβάνει υπόψη αληθινά και ψευδώς θετικά και αρνητικά και θεωρείται γενικά ως ισορροπημένο μέτρο που μπορεί να χρησιμοποιηθεί ακόμη και αν οι τάξεις έχουν πολύ διαφορετικά μεγέθη. Το MCC είναι ουσιαστικά ένας συντελεστής συσχέτισης μεταξύ των παρατηρούμενων και των προβλεπόμενων δυαδικών ταξινομήσεων. Επιστρέφει μια τιμή μεταξύ μείον ένα και ένα. Ένας συντελεστής ένα αντιπροσωπεύει μια τέλεια πρόβλεψη, το μηδέν δεν είναι καλύτερο από την τυχαία πρόβλεψη και το μείον ένα δείχνει συνολική διαφωνία μεταξύ της πρόβλεψης και της παρατήρησης. Για τις προβλεψεις του μοντέλου έχουμε τιμή MCC = 0.51 .

$$MCC = 0.51 \tag{31}$$

9.2 Μοντελοποιήσεις με Residual nets

Μια εναλλακτική αρχιτεκτονική αποτελούν τα νευρωνικά δίκτυα residual nets. Στην δικιά μας περίπτωση δοκιμάσαμε να προχωρήσουμε και με τα συγκεκριμένα καθότι έχουν πάρα πολύ καλά αποτελέσματα σε μια σειρά απο μοντελοποιήσεις. Δοκιμάσαμε διάφορα επίπεδα στο δίκτυο.

Παρακάτω εμφανίζονται τα διαγράμματα ακρίβειας της εκπαίδευσης.

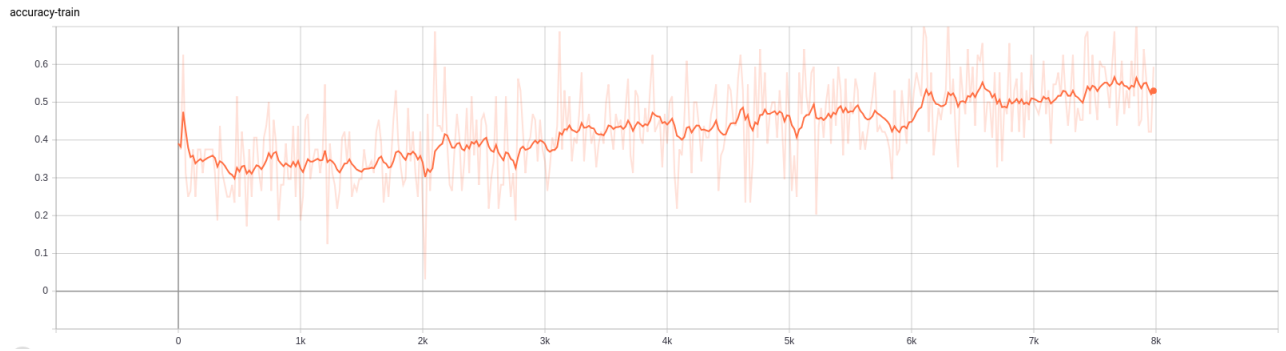


Figure 18: Ακρίβεια μοντέλου στα δεδομένα εκπαίδευσης

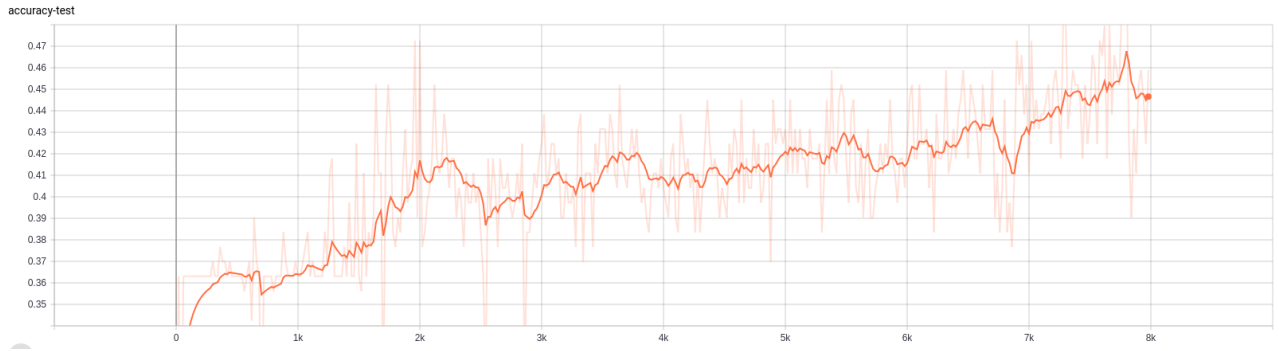


Figure 19: Ακρίβεια μοντέλου στα δεδομένα επαλήθευσης

Στο παρακάτω γράφημα παρουσιάζεται το γράφημα ακρίβειας για ένα μοντέλο στα δεδομένα επαλήθευσης.

Παρατηρούμε από τα διαγράμματα πως η εκπαίδευση δεν έχει τελειώσει για τα συγκεκριμένα μοντέλα αλλά δεδομένο των κατα πολλών περισσότερων παραμέτρων ο χρόνος εκπαίδευσης τείνει να γίνει απαγορευτικός. Συγκεκριμένα με μια κάρτα γραφικών με δυνατότητα μνήμης 8 gb χρειάστηκαν περίπου 6 ώρες για να φτάσουμε στο 8000 βήμα που βλέπουμε παρακάτω. Επίσης η ακρίβεια που διαμορφώνεται δεν φαίνεται να δημιουργεί αισιοδοξία για πολύ καλύτερα αποτελέσματα απο τα Image Net νευρωνικά δίκτυα. Για να επιβεβαιώσουμε τον ισχυρισμό θα χρειαστούμε μεγαλύτερο υπολογιστή και περισσότερους πόρους.

9.3 Περαιτέρω συζήτηση και έρευνα

Παραπάνω δείξαμε πως η δισδιάστατη δομή μιας χημικής ένωσης μπορεί να αποτελέσει σημαντική πληροφορία στις QSAR μοντελοποιήσεις. Επίσης καταφέραμε να βρούμε κάποιες αρχιτεκτονικές δικτύων που μπορούν να χρησιμοποιήσουν την συγκεκριμένη πληροφορία με τον καλύτερο τρόπο.

Σε επόμενα βήματα μπορεί να προχωρήσει μια έρευνα για το ποιές άλλες αποκρίσεις μπορούν να μοντελοποιηθούν με την συγκεκριμένη διαδικασία καθώς και η μεγαλύτερη συλλογή δεδομενων καθότι όσα περισσότερο τόσο το καλύτερο για τις μοντελοποιήσεις με νευρωνικά δίκτυα.

Γνωρίζουμε επίσης πως η είσοδος ενός νευρωνικού δικτύου μπορεί να έχει ότι σχήμα θέλουμε και πέρα απο την δισδιάστατη δομή υπάρχει και η τρισδιάστατη δομή του μορίου που περιέχει πολύ περισσότερη πληροφορία όπως την θέση ενός ατόμου στον χώρο σε σχέση με τα υπόλοιπα άτομα στην δομή. Κατάλληλες κωδικοποιήσεις που θα επιτρέψουν την διαμόρφωση των δεδομένων να χρησιμοποιηθούν απο νευρωνικά δίκτυα πιθανά θα παράγουν και καλύτερα αποτελέσματα. Επίσης η δημιουργία μεγαλύτερων σετ δεδομένων θα δώσουν περισσότερη πληροφορία και καλύτερες προβλέψεις στα μοντέλα.

Παρακάτω φαίνεται ένα κομμάτι μιας τρισδιάστατης δομής μιας χημικής ένωσης.

10.6127	-7.8570	0.5943	O	0	0	0	0	0	0	0	0	0	0	0	0
11.3760	-7.3991	-0.5165	C	0	0	0	0	0	0	0	0	0	0	0	0
6.9800	-5.0975	-1.1041	C	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0									
7	1	1	0	0	0	0									
8	7	2	0	0	0	0									

References

[1] Timothy P. Lillicrap, Jonathan H. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver and Daan Wierstra. *Continuous Control with Deep Reinforcement Learning*, 2016.

[2] *Introduction to Machine Learning: The Wikipedia Guide*, available at http://www.datascienceassn.org/sites/default/files/Introduction%20to%20Machine%20Learning.pdf, pp. 1-4.

[3] Vincent Francois-Lavet, Riashat Islam, Joelle Pineau, Peter Henderson and Marc G. Bellemare. *An Introduction to Deep Reinforcement learning*, Foundations and Trends in Machine Learning Vol. 11, No 3-4, pp. 15-19, 24-25 2018.

[4] Kevin Gurney. *An introduction to neural networks*, UCL Press, pp.12-16, 1997.

[5] David Kriesel. *A brief introduction to neural networks*, available at http://www.dkriesel.com, 2007.

[6] Michael A. Nielsen. *Neural Networks and Deep Learning*, available at http://neuralnetworksanddeeplearning.com/, Determination Press, pp. 42-49 2015.

[7] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*, MIT Press, available at http://www.deeplearningbook.org, pp. 226-231, 2016.

[8] Steven Spielberg Pon Kumar.*Deep Reinforcement Learning Approaches for Process Control*, Master Thesis, pp.14-15, 77, 2017.

[9] Yaron Singer. *AM 221: Advanced Optimization, Lecture 9*, pp. 3-5, 2016.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*, Cambridge University Press, pp. 459-461, 2004.

[11] Balázs Csanád Csáji. *Approximation with Artificial Neural Networks*. MSc Thesis, pp. 11, 2001.

[12] Csaba Szepesvári. *Algorithms for Reinforcement Learning*, Morgan and Claypool, pp. 12-15, 2009.

[13] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction, second edition*, MIT Press, pp. 50-51, 2017.

[14] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariance Shift*, 2015.

[15] Diederik P. Kingma and Jimmy Lei Ba. *Adam: A Method for Stochastic Optimization*, 2015.

[16] Matthews, B. W. *"Comparison of the predicted and observed secondary structure of T4 phage lysozyme". Biochimica et Biophysica Acta (BBA) - Protein Structure. 405 (2): 442–451.*, 1975.

[17] Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. L. *"The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics". Journal of Chemical Information and Computer Sciences. 43 (2): 493–500.*, 2004.

[18] Tom M. Mitchell *Machine Learning*, 1997.

[19] Rajarshi Guha and Egon Willighagen *A Survey of Quantitative Descriptions of Molecular Structure*, 2012.

[20] Kohavi R, John G. *Wrappers for Feature Subset Selection*, 1997.

[21] Christian Szegedy and Wei Liu and Yangqing Jia and Pierre Sermanet and Scott Reed and Dragomir Anguelov and Dumitru Erhan and Vincent Vanhoucke and Andrew Rabinovich *Going Deeper with Convolutions*, 2015.

[22] K. He, X. Zhang, S. Ren and J. Sun *Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2016.

[23] Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, Zoeller RT, Gore AC *"Endocrine-disrupting chemicals: an Endocrine Society scientific statement"*, 2009.