Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
τομεας μαθηματικων

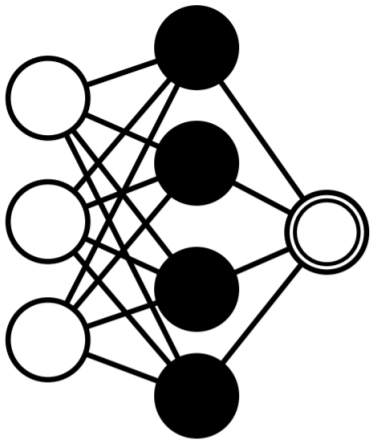
Αντιστάθμιση Ακρίβειας σε Κατανεμημένα Συστήματα Ζεύγους Βαθέων Νευρωνικών Δικτύων

Μελέτη και Ανάπτυξη

Διπλωματικη Εργασια

του

ΤΖΕΒΑΧΙΡΙΔΗ ΑΝΔΡΕΑ



Επιβλέπων: Ιάκωβος Στ. Βενιέρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
τομεας μαθηματικων

Αντιστάθμιση Ακρίβειας σε Κατανεμημένα Συστήματα Ζεύγους Βαθέων Νευρωνικών Δικτύων

Μελέτη και Ανάπτυξη

Διπλωματικη Εργασια

του

ΤΖΕΒΑΧΙΡΙΔΗ ΑΝΔΡΕΑ

Επιβλέπων: Ιάκωβος Στ. Βενιέρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Οκτωβρίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιάκωβος Στ. Βενιέρης	Δήμητρα-Θεοδώρα Κακλαμάνη	Αντώνιος Συμβώνης
Καθηγητής Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
τομεας μαθηματικων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Ανδρέας Τζεβαχιρίδης

4 Οκτωβρίου 2021

Περίληψη

Η σημαντική πρόοδος των τελευταίων ετών στις τεχνολογίες Βαθιάς Μάθησης έχει καταστήσει εφικτή την ανάπτυξη πληθώρας ευφυών κινητών εφαρμογών. Παρόλα αυτά, τα βαθιά νευρωνικά δίκτυα είναι υπολογιστικά απαιτητικά, γεγονός το οποίο, σε συνδυασμό με τους περιορισμένους πόρους των κινητών συσκευών, οδηγεί σε προκλήσεις. Πολλές φορές η τοπική εκτέλεση είναι ανέφικτη ή δεν υπάρχει εγγύηση για την ποιότητα υπηρεσίας. Ένας τρόπος αντιμετώπισης των παραπάνω περιορισμών είναι η κατανεμημένη εκτέλεση.

Ένα κατανεμημένο σύστημα ζεύγους βαθέων νευρωνικών δικτύων αποτελείται από (α) ένα βαθύ νευρωνικό δίκτυο, το οποίο εκτελείται τοπικά, επί της κινητής συσκευής και έχει χαμηλές απαιτήσεις υπολογιστικής ισχύος, μνήμης και ενέργειας, αλλά χαμηλή ακρίβεια, (β) ένα βαθύ νευρωνικό δίκτυο που εκτελείται στο υπολογιστικό νέφος ή στην άκρη του δικτύου με την υποβοήθηση ενός ισχυρού εξυπηρετητή, το οποίο έχει υψηλές απαιτήσεις υπολογιστικής ισχύος, μνήμης και ενέργειας, αλλά υψηλή ακρίβεια και (γ) τον τρόπο με τον οποίο τα δύο δίκτυα επικοινωνούν, καθώς και το είδος των δεδομένων που ανταλλάσσουν. Βασικό χαρακτηριστικό του κατανεμημένου συστήματος είναι η επιλεκτική εκτέλεση είτε στην κινητή συσκευή είτε στον εξυπηρετητή.

Ωστόσο, βασικό περιορισμό των συστημάτων που κάνουν χρήση ενός ζεύγους μοντέλων αποτελεί το γεγονός ότι η ακρίβεια ενός μοντέλου αξιολογείται με βάση τα υπάρχοντα σύνολα δεδομένων. Αυτό έχει ως συνέπεια, στην πράξη, τα μοντέλα να καλούνται να επεξεργαστούν δεδομένα που διαφέρουν δραστικά από τα δεδομένα της εκπαίδευσης, καταλήγοντας σε ανακριβή αποτελέσματα.

Η παρούσα διπλωματική εργασία εστιάζει στο πρόβλημα της ταξινόμησης εικόνας και στόχος της είναι αρχικά η ανάπτυξη αλγορίθμων αξιολόγησης της ακρίβειας του μοντέλου στην κινητή συσκευή και στη συνέχεια η σχεδίαση της κατάλληλης αρχιτεκτονικής του συστήματος ζεύγους νευρωνικών δικτύων, το οποίο μέσω συνέργειας των δύο μοντέλων, θα παρέχει τη δυνατότητα αντιστάθμισης και διόρθωσης των σφαλμάτων του τοπικού μοντέλου σε πραγματικό χρόνο.

Λέξεις Κλειδιά

Βαθιά Μάθηση, Κινητή Συσκευή, Εξυπηρετητής, Κατανεμημένη Εκτέλεση, Αρχιτεκτονική, Κατανεμημένο Σύστημα Ζεύγους Νευρωνικών Δικτύων, Εφαρμογή, Ταξινόμηση Εικόνας, Tensorflow

Abstract

Recent breakthroughs in Deep Learning technologies have enabled numerous intelligent mobile applications. However, deep neural networks are computationally intensive, which in conjunction with the limited resources of mobile devices leads to challenges. In many cases, local inference is unfeasible or quality of service is not guaranteed. One way of handling these limitations is through distributed inference.

A two-DNN distributed system consists of (a) a deep neural network that performs the model inference locally, at the mobile device, having low demand for computational power, memory and energy, but low accuracy, (b) a deep neural network that performs the model inference at the cloud or the edge, supported by a powerful server, which has high demand for computational power, memory and energy, but high accuracy and (c) the ways in which the two networks communicate, as well as the type of data being exchanged. One key characteristic of a two-DNN distributed system is selective inference, either at the mobile device or at the server.

However, an important limitation of two-DNN distributed systems is that model accuracy is calculated with respect to existing datasets. This, in practice, leads to the models' having to process data that differ drastically from training data, resulting in less accurate predictions.

The present diploma thesis is focused on the task of image classification and its objective is initially the development of algorithms for the evaluation of the mobile device's model accuracy and consequently the design of a fitting two-DNN distributed system architecture, which will enable the improvement of the local model's results in real time.

Keywords

Deep Learning, Mobile Device, Server, Distributed Inference, Architecture, two-DNN Distributed System, Application, Image Classification, Tensorflow

Ευχαριστίες

Θα ήθελα καταρχάς να ευχαριστήσω τον καθηγητή κ. Ιάκωβο Στ. Βενιέρη για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα διπλωματική εργασία και να εμβαθύνω πάνω σε ένα θέμα τόσο ενδιαφέρον και επίκαιρο. Επιπλέον, ευχαριστώ την καθηγήτρια κ. Δήμητρα-Θεοδώρα Κακλαμάνη και τον καθηγητή κ. Αντώνιο Συμβώνη για τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαιτέρως τον κ. Ιωάννη Πανόπουλο, υποψήφιο διδάκτορα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ και τον Δρ. Στυλιανό Βενιέρη, ερευνητή στο Κέντρο Τεχνητής Νοημοσύνης της Samsung στο Cambridge. Οι πολύτιμες συμβουλές και η εξαιρετική καθοδήγησή τους συνέβαλαν αδιαμφισβήτητα στην περάτωση της διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους, την οικογένεια και τους στενούς μου ανθρώπους για τη βοήθεια και τη στήριξη που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Οκτώβριος 2021

Ανδρέας Τζεβαχιρίδης

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Ζούμε στην εποχή της άνθισης του τομέα της Τεχνητής Νοημοσύνης (Artificial Intelligence). Η Βαθιά Μάθηση (Deep Learning), μέσω του μεγάλου όγκου των δεδομένων, της εξέλιξης των αλγορίθμων και της υπολογιστικής ισχύος που είναι πλέον διαθέσιμη, έχει σημειώσει σημαντική πρόοδο, κυρίως σε τομείς όπως η Όραση Υπολογιστών (Computer Vision) και η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing). Ως αποτέλεσμα, ένα εύρος από ευφυείς κινητές εφαρμογές έχει εκμεταλλευτεί τα εργαλεία που προσφέρει η σύγχρονη Βαθιά Μάθηση με σκοπό την βελτίωση της ανθρώπινης ζωής. Παραδείγματα τέτοιων εφαρμογών αποτελούν ο εντοπισμός αντικειμένων (object detection), η αναγνώριση προσώπου (facial recognition), η αναγνώριση ομιλίας (speech recognition), η μετάφραση (translation), κ.α.

Ωστόσο, παρόλο που τα βαθιά νευρωνικά δίκτυα είναι φημισμένα για την εξαιρετική απόδοσή τους σε διεργασίες αναγνώρισης, οι δομές αυτές είναι υπολογιστικά απαιτητικές. Το γεγονός αυτό, σε συνδυασμό με το ότι οι κινητές συσκευές διαθέτουν περιορισμένους υπολογιστικούς πόρους, καθιστά την ενσωμάτωση της Βαθιάς Μάθησης στο περιβάλλον των κινητών συσκευών μια διαδικασία συνοδευόμενη από πολλές προκλήσεις. Πέραν της εκπαίδευσης (training), η οποία απαιτεί μνήμη και υπολογιστική ισχύ που συνήθως ξεπερνούν κατά πολύ τους πόρους που διαθέτει μια κινητή συσκευή, η διαδικασία της εκτέλεσης ή αλλιώς συμπερασματολογίας (inference) μπορεί να καταλάβει επίσης σημαντικό ποσοστό των πόρων της συσκευής. Σε πολλές περιπτώσεις η τοπική εκτέλεση ενδέχεται να είναι ανέφικτη, ενώ όταν είναι εφικτή, δεν υπάρχει πάντοτε εγγύηση για την ποιότητα υπηρεσίας.

Μια μέθοδος η οποία φαίνεται να λύνει σε μεγάλο βαθμό αυτά τα ζητήματα είναι η κατανεμημένη εκτέλεση. Τα υβριδικά ή αλλιώς κατανεμημένα συστήματα ζεύγους βαθέων νευρωνικών δικτύων (hybrid/distributed two-DNN systems) παρέχουν τη δυνατότητα της επιλεκτικής εκτέλεσης, είτε μέσω του μοντέλου που βρίσκεται ενσωματωμένο στην κινητή συσκευή, είτε μέσω ενός ισχυρότερου μοντέλου που εδρεύει στο νέφος (cloud) και βασίζεται στη λειτουργία ενός εξυπηρετητή. Αυτή η προσέγγιση οδηγεί τόσο στην απελευθέρωση των περιορισμένων πόρων της συσκευής όταν αυτό είναι απαραίτητο, όσο και στην αξιοποίηση του κατάλληλου μοντέλου ανάλογα με τις ανάγκες του εκάστοτε δείγματος εισόδου.

Λαμβάνοντας το κατανεμημένο σύστημα ζεύγους ως δεδομένο εργαλείο, σημαντικό βήμα είναι να μελετηθούν μέθοδοι με τις οποίες θα γίνεται η κατανομή και η επεξεργασία νέων δειγμάτων εισόδου, έτσι ώστε το σύστημα να πετυχαίνει τη μεγαλύτερη δυνατή ακρίβεια, με την προϋπόθεση ότι δεν γίνεται υπερφόρτωση του εξυπηρετητή.

Οι στόχοι της παρούσας διπλωματικής εργασίας είναι: (α) η ανάπτυξη αλγορίθμων με σκοπό τον εντοπισμό της πραγματικής ακρίβειας του τοπικού μοντέλου, (β) η διερεύνηση μεθόδων με τις οποίες μπορεί να γίνει αποδοτική επιλογή των δειγμάτων που πιθανώς το τοπικό μοντέλο δεν θα ταξινομήσει επιτυχώς και (γ) η διερεύνηση μεθόδων με τις οποίες η ακρίβεια του συστήματος μπορεί να διορθωθεί, μέσω της συνέργειας των δύο μοντέλων.

1.1 Οργάνωση του Τόμου

Η εργασία αυτή είναι οργανωμένη σε οκτώ κεφάλαια: στο κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίζεται η πορεία της εργασίας. Στο κεφάλαιο 3 γίνεται μια παρουσίαση των βασικών γλωσσών προγραμματισμού, βιβλιοθηκών και συνόλων δεδομένων που χρησιμοποιούνται στα πλαίσια της ανάπτυξης. Στο κεφάλαιο 4 γίνεται μια περιγραφή υψηλού επιπέδου των συνιστωσών της ενιαίας εφαρμογής. Έπειτα, στο κεφάλαιο 5 παρουσιάζονται οι μέθοδοι που διευρευνήθηκαν για τους σκοπούς της εργασίας. Στο κεφάλαιο 6 αναλύεται η ανάπτυξη των μοντέλων και των αλγορίθμων που χρησιμοποιεί η εφαρμογή. Στο κεφάλαιο 7 παρουσιάζονται τα αποτελέσματα όλων των μεθόδων που διερευνήθηκαν στο κεφάλαιο 5. Τέλος, στο κεφάλαιο 8 παρατίθενται τα τελικά συμπεράσματα, καθώς και οι μελλοντικές επεκτάσεις της διπλωματικής εργασίας.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο δεύτερο Κεφάλαιο παρουσιάζονται οι θεωρητικές έννοιες τις οποίες οφείλει να έχει κατανοήσει κάποιος πριν συνεχίσει στη μελέτη των επόμενων κεφαλαίων.

2.1 Βαθιά Μάθηση

Το βασικό κίνητρο για την ανάπτυξη της Βαθιάς Μάθησης (Deep Learning) αποτέλεσε η αποτυχία των κοινών αλγορίθμων Μηχανικής Μάθησης (Machine Learning) να λύσουν αποτελεσματικά ένα υποσύνολο των κεντρικών προβλημάτων του κλάδου της Τεχνητής Νοημοσύνης (Artificial Intelligence), όπως η αναγνώριση ομιλίας ή αντικειμένων [1].

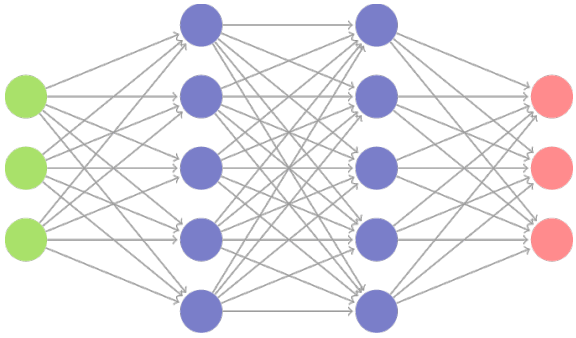
Ένας χρήσιμος ορισμός υπογραμμίζει ότι η Βαθιά Μάθηση εστιάζει σε νευρωνικά δίκτυα με περισσότερα από δύο επίπεδα. Το αρχικό και τελικό επίπεδο ονομάζονται φανερά επίπεδα και αποτελούν το επίπεδο εισόδου και εξόδου του δικτύου αντίστοιχα, ενώ τα ενδιάμεσα επίπεδα συνήθως ονομάζονται κρυφά. Αν και με τον παράπανω ορισμό περιγράφεται καλά η βασική ειδοποιός διαφορά μεταξύ Βαθιάς και Μηχανικής Μάθησης, αυτός δεν μπορεί να θεωρηθεί πλήρης. Πρέπει οπωσδήποτε να σημειωθεί ότι τα νευρωνικά δίκτυα, πριν παράγουν τα θεαματικά αποτελέσματα των τελευταίων ετών, έπρεπε (α) να διαφοροποιηθούν αρχιτεκτονικά από την αρχική μορφή τους και (β) να εκμεταλλευτούν την πολύ περισσότερη υπολογιστική ισχύ που ήταν πλέον διαθέσιμη. Μερικά από τα βασικά στοιχεία της εξέλιξης των νευρωνικών δικτύων είναι: οι περισσότεροι νευρώνες, οι πιο περίπλοκοι τρόποι ένωσης επιπέδων ή νευρώνων, η έκρηξη στην υπολογιστική ισχύ που είναι διαθέσιμη για την εκπαίδευση, η αυτόματη εξαγωγή χαρακτηριστικών και η έκρηξη στον όγκο των διαθέσιμων δεδομένων με ετικέτα [2].

Η σύγχρονη Βαθιά Μάθηση παρέχει πολύ ισχυρά εργαλεία για Επιβλεπόμενη Μάθηση (Supervised Learning). Τα πολύ περισσότερα επίπεδα αλλά και οι πολύ περισσότεροι νευρώνες ανά επίπεδο καθιστούν ένα βαθύ νευρωνικό δίκτυο ικανό να προσεγγίσει αρκετά πολύπλοκες συναρτήσεις. Αυτό σημαίνει ότι με την κατασκευή επαρκώς μεγάλων μοντέλων και την προετοιμασία επαρκώς μεγάλων συνόλων δεδομένων που περιέχουν δείγματα με ετικέτες, οι περισσότερες διεργασίες που αφορούν την απεικόνιση μιας εισόδου σε μια έξοδο μπορούν να γίνουν αποδοτικά μέσω της Βαθιάς Μάθησης [1].

Δύο βασικές αρχιτεκτονικές βαθέων νευρωνικών δικτύων που χρησιμοποιούνται για Επιβλεπόμενη Μάθηση είναι: (α) τα Βαθιά Νευρωνικά Δίκτυα Εμπρόσθιας Τροφοδότησης (Deep Feedforward Neural Networks), στα οποία η πληροφορία ρέει από την είσοδο προς την έξο-

δο με μια μόνο κατεύθυνση και (β) τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks), στα οποία η πληροφορία μπορεί να ρέει αμφίδρομα.

Ο σκοπός ενός Νευρωνικού Δικτύου Εμπρόσθιας Τροφοδότησης είναι η προσέγγιση μιας συνάρτησης f^*. Για παράδειγμα, στην περίπτωση που εξετάζουμε έναν ταξινομητή, η συνάρτηση $y = f^*(x)$ απεικονίζει μια είσοδο x στην κατηγορία y. Ένα Νευρωνικό Δίκτυο Εμπρόσθιας Τροφοδότησης ορίζει μια απεικόνιση $y = f^*(x; \theta)$ και μαθαίνει το σύνολο παραμέτρων θ που καταλήγει στην βέλτιστη προσέγγιση της συνάρτησης [1]. Στην εικόνα 2.1 παρουσιάζεται ένα βαθύ νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης με δυο κρυφά επίπεδα.



Εικόνα 2.1: *Βαθύ νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης*

Όπως ήδη αναφέρθηκε, πέραν της αρχιτεκτονικής αναβάθμισης των Νευρωνικών Δικτύων, σημαντικό ρόλο στην άνθιση της Βαθιάς Μάθησης έπαιξε η έκρηξη σε διαθέσιμη υπολογιστική ισχύ για την εκπαίδευση των μοντέλων. Οι Μονάδες Επεξεργασίας Γραφικών (Graphics Processing Units), οι οποίες έχουν ως βασικό σκοπό την επιτάχυνση της επεξεργασίας γραφικών, μπορούν να επιταχύνουν δραματικά τις υπολογιστικές διεργασίες της Βαθιάς Μάθησης. Αυτό είναι εφικτό λόγω της ιδιότητας των Μονάδων Επεξεργασίας Γραφικών να χρησιμοποιούνται αποδοτικά με παράλληλο τρόπο για υπολογισμούς μεγάλης κλίμακας. Άλλη μια μονάδα υλικού (hardware) που συμβάλλει στην επιτάχυνση των υπολογιστικών διεργασιών της Βαθιάς Μάθησης είναι οι Μονάδες Επεξεργασίας Τανυστών (Tensor Processing Units), οι οποίες κατασκευάστηκαν συγκεκριμένα για τις διεργασίες της Μηχανικής Μάθησης και είναι ενσωματωμένες στο Tensorflow, τη βιβλιοθήκη λογισμικού (framework) ανοιχτού κώδικα της Google.

2.2 Συνελικτικά Νευρωνικά Δίκτυα

Το Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network) είναι μια εξειδικευμένη μορφή Βαθέος Νευρωνικού Δικτύου Εμπρόσθιας Τροφοδότησης που χρησιμοποιείται για την επεξεργασία διανυσμάτων εισόδου με τοπολογία πλέγματος. Παραδείγματα αποτελούν οι χρονοσειρές, οι οποίες μπορούν να θεωρηθούν σαν πλέγμα μιας διάστασης και οι εικόνες, οι οποίες μπορούν να θεωρηθούν σαν πλέγματα δύο ή τριών διαστάσεων. Από το όνομά τους γίνεται αντιληπτό ότι αφορούν μια συγκεκριμένη μαθηματική πράξη, τη συνέλιξη (convolution), η οποία είναι γραμμική [1].

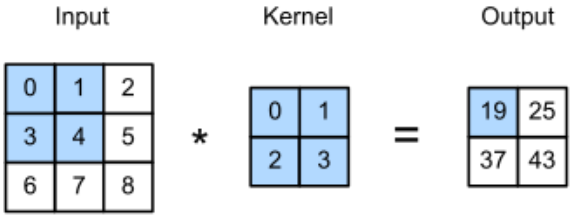
Έστω ότι τα x και w είναι δύο δείγματα διακριτού σήματος. Το x είναι το δείγμα εισόδου, ενώ το w ονομάζεται φίλτρο (filter) ή πυρήνας (kernel). Η συνέλιξη των x και w συμβολίζεται

$x * w$ και στην μονοδιάστατη περίπτωση ορίζεται ως:

$$(x * w)[t] = \sum_{\tau} x[t - \tau]w[\tau]$$

Στη δισδιάστατη περίπτωση, αν και προστίθεται ένας επιπλέον δείκτης, η λογική της πράξης παραμένει η ίδια. Στην εικόνα 2.2 φαίνεται ένα παράδειγμα δισδιάστατης συνέλιξης μιας εικόνας εισόδου 3×3 με ένα φίλτρο 2×2.

$$(x * w)[s, t] = \sum_{\sigma,\tau} x[s - \sigma, t - \tau]w[\sigma, \tau]$$



Εικόνα 2.2: *Δισδιάστατη συνέλιξη*

2.2.1 Επίπεδα

Τα βασικά επίπεδα ενός συνελικτικού νευρωνικού δικτύου χωρίζονται σε τρεις κατηγορίες. Αυτές είναι: το συνελικτικό επίπεδο (convolutional layer), το επίπεδο συγκέντρωσης (pooling layer) και το πλήρως-συνδεδεμένο επίπεδο (fully-connected layer). Στη συνέχεια εξετάζεται ο ρόλος και η λειτουργία του κάθε επιπέδου.

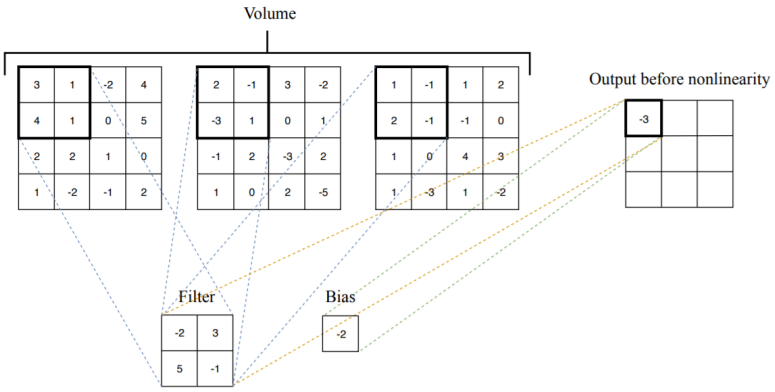
Συνελικτικό Επίπεδο

Οι παράμετροι του συνελικτικού επιπέδου αποτελούνται από ένα σύνολο φίλτρων, τα οποία καθορίζονται από το δίκτυο κατά την διαδικασία της εκπαίδευσης. Καθένα από αυτά τα φίλτρα κινείται κατά μήκος και κατά πλάτος της εικόνας και παράγει έναν δισδιάστατο χάρτη ενεργοποίησης (activation map), ο οποίος περιγράφει το αποτέλεσμα της εφαρμογής του κάθε φίλτρου σε κάθε θέση της εικόνας [3].

Το συνελικτικό επίπεδο καθορίζει την έξοδο των νευρώνων που είναι συνδεδεμένοι σε τοπικές περιοχές της εισόδου μέσω του υπολογισμού του βαθμωτού γινομένου μεταξύ των βαρών και της περιοχής και στη συνέχεια εφαρμόζει μια μη γραμμική συνάρτηση ενεργοποίησης. Στα κρυφά επίπεδα αυτή η συνάρτηση είναι συνήθως η Ανορθωμένη Γραμμική Μονάδα (Rectified Linear Unit).

Σε περίπτωση που ένα επίπεδο l_k έχει n φίλτρα, η έξοδος του l_k αποτελείται από n πίνακες. Συνεπώς, το επίπεδο l_{k+1} δέχεται ως είσοδο μια συλλογή από n πίνακες. Αυτή η συλλογή ονομάζεται όγκος (volume) και η πληθυκότητά της ονομάζεται βάθος (depth). Η συνέλιξη ενός φίλτρου και μιας τοπικής περιοχής ενός όγκου είναι τελικά το άθροισμα των συνελίξεων του φίλτρου και της αντίστοιχης τοπικής περιοχής των πινάκων που αποτελούν τον όγκο. Στην εικόνα 2.3 φαίνεται ένα παράδειγμα συνέλιξης ενός όγκου που αποτελείται

από τρεις πίνακες.



Εικόνα 2.3: *Συνέλιξη όγκου με 3 πίνακες*

Τα συνελικτικά επίπεδα μπορούν να μειώσουν σημαντικά την πολυπλοκότητα του μοντέλου μέσω της βελτιστοποίησης τριών παραμέτρων: του βάθους (depth), του βηματισμού (stride) και του παραγεμίσματος (padding). Αυτές οι παράμετροι ονομάζονται υπερπαράμετροι, καθώς δεν μαθαίνονται από το μοντέλο κατά τη διαδικασία της εκπαίδευσης, αλλά καθορίζονται εκ των προτέρων.

Συνήθως, η βελτιστοποίηση για αυτές τις παραμέτρους γίνεται με τη μέθοδο της διασταυρωμένης επικύρωσης k τμημάτων (k-fold cross-validation), όπου το σύνολο εκπαίδευσης χωρίζεται σε k τμήματα. Τα $k-1$ τμήματα χρησιμοποιούνται για την εκπαίδευση, ενώ το υπολοιπόμενο ένα για την επικύρωση. Αυτή η διαδικασία επαναλαμβάνεται για διαφορετικές τιμές υπερπαραμέτρων, ώστε να βρεθούν οι βέλτιστες τιμές τους, συνήθως μέσω κάποιας αυτοματοποιημένης μεθόδου, όπως είναι η αναζήτηση πλέγματος (grid search).

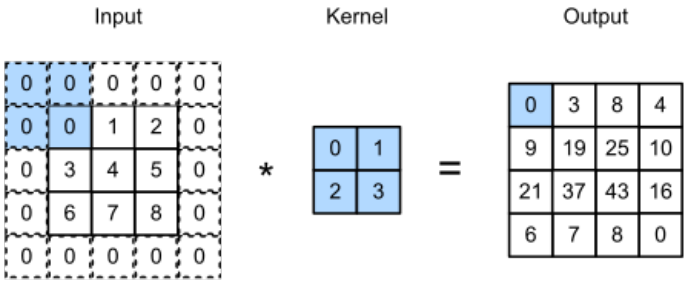
1. Βάθος: ο αριθμός των φίλτρων που θα χρησιμοποιηθούν. Το κάθε φίλτρο εστιάζει σε διαφορετικό χαρακτηριστικό της εισόδου.

2. Βηματισμός: το μέγεθος βήματος του κινούμενου παραθύρου. Η συγκεκριμένη παράμετρος επιτρέπει στο επίπεδο να καθορίσει τις διαστάσεις της εξόδου.

3. Παραγέμισμα: το πλάτος των επιπρόσθετων κελιών τα οποία περιβάλλουν την αρχική εικόνα (ή τον όγκο) πριν γίνει η συνέλιξη με τα φίλτρα. Συνήθως τα επιπρόσθετα κελιά περιέχουν μηδενικά. Αυτή η παράμετρος είναι χρήσιμη γιατί επιτρέπει στο συνελικτικό επίπεδο να ελέγξει τις διαστάσεις του πίνακα εξόδου [4].

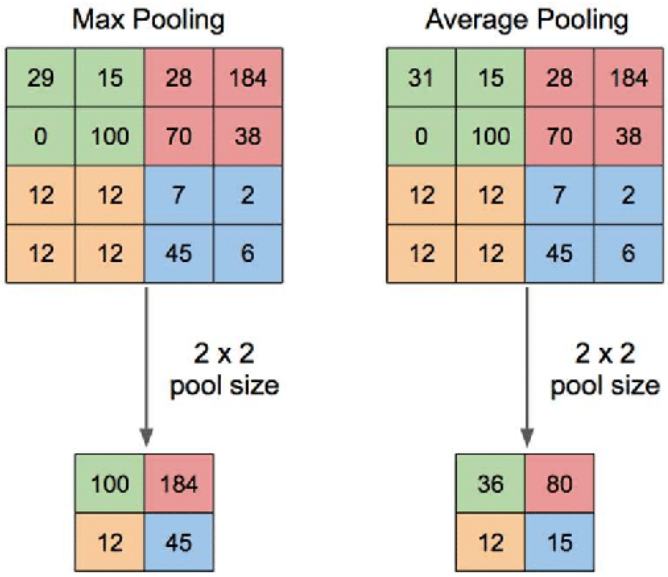
Επίπεδο Συγκέντρωσης

Το επίπεδο συγκέντρωσης έχει παρόμοια λειτουργία με το συνελικτικό επίπεδο, εφόσον και πάλι έχουμε δράση ενός φίλτρου με την προσέγγιση του κινούμενο παραθύρου. Παρόλα αυτά, το φίλτρο σε αυτή την περίπτωση δεν έχει παραμέτρους τις οποίες μαθαίνει το δίκτυο κατά την εκπαίδευση, αλλά προκαθορισμένους τελεστές που εφαρμόζονται κατά το βηματισμό του παραθύρου. Οι συνήθεις τελεστές που χρησιμοποιούνται είναι το μέγιστο (max) και το μέσο (average) [5].



Εικόνα 2.4: *Συνέλιξη με βηματισμό 1 και παραγέμισμα 1*

Λόγω της περιοριστικής φύσης του επιπέδου συγκέντρωσης, υπάρχουν δύο συνήθεις μέθοδοι συγκέντρωσης μεγίστου. Συνήθως ο βηματισμός και το μέγεθος του παραθύρου τίθενται και τα δύο στην τιμή 2, αλλά υπάρχουν και περιπτώσεις που ο βηματισμός είναι 2, ενώ το μέγεθος παραθύρου είναι 3.



Εικόνα 2.5: *Αποτελέσματα συγκέντρωσης με βηματισμό 2*

Ο σκοπός των επιπέδων συγκέντρωσης είναι να μειώνουν σταδιακά τις διαστάσεις του δείγματος, το οποίο οδηγεί σε μείωση των παραμέτρων και συνεπώς μείωση της υπολογιστικής πολυπλοκότητας του μοντέλου.

Επίσης, αξίζει να σημειωθεί ότι τα επίπεδα συγκέντρωσης είναι δυνατόν να εφαρμόζουν γενική συγκέντρωση (general-pooling) με πιο περίπλοκους τελεστές από το μέγιστο και το μέσο. Παράδειγμα αποτελεί η συγκέντρωση με βάση την $L2$-νόρμα (L2-norm pooling), που υπολογίζει την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των τιμών μιας τοπικής περιοχής [4].

Πλήρως-συνδεδεμένο Επίπεδο

Το πλήρως-συνδεδεμένο επίπεδο αποτελείται από νευρώνες που είναι απευθείας συνδεδεμένοι με όλους τους νευρώνες του προηγούμενου, αλλά και του επόμενου επιπέδου. Χρησιμοποιούνται κυρίως μετά από την εφαρμογή αρκετών συνελικτικών επιπέδων και επι-

πέδων συγκέντρωσης και λαμβάνουν το αποτέλεσμα αυτής της επεξεργασίας, έτσι ώστε να το μετασχηματίσουν σε ένα ενιαίο διάνυσμα και στη συνέχεια να εξαγάγουν τις πιθανότητες πρόβλεψης για κάθε κλάση.

2.2.2 Αρχιτεκτονικές

Οι πιο κλασικές αρχιτεκτονικές στον χώρο των συνελικτικών νευρωνικών δικτύων είναι οι εξής:

1. LeNet [6]. Μια από τις πρώτες επιτυχείς αρχιτεκτονικές που σχεδιάστηκε από τον Yann LeCun την δεκαετία του 1990. Τα αντίστοιχα μοντέλα είχαν ως κύριο σκοπό την ανάγνωση ταχυδρομικών κωδίκων ή ψηφίων.

2. AlexNet [7]. Η πρώτη αρχιτεκτονική που έκανε δημοφιλή τα συνελικτικά δίκτυα. Κέρδισε με διαφορά στον διαγωνισμό ILSVRC 2012 του ImageNet και κύριο χαρακτηριστικό της ήταν η ένωση πολλαπλών συνεχόμενων συνελικτικών επιπέδων.

3. GoogLeNet [8]. Κέρδισε στον διαγωνισμό ILSVRC 2014. Η αρχιτεκτονική αυτή βασίστηκε σε ένα Inception Module το οποίο μειώνει καθοριστικά τις παραμέτρους στο δίκτυο. Λόγω της μεγαλης χρησιμότητας των Inception Modules και σε επόμενες αρχιτεκτονικές, πλέον είναι γνωστό ως η πρώτη έκδοση του μοντέλου Inception.

4. VGGNet [9]. Η αρχιτεκτονική αυτή έκανε γνωστό ότι το βάθος ενός νευρωνικού δικτύου παίζει καθοριστικό ρόλο. Περιέχει 16 συνελικτικά και πλήρως-συνδεδεμένα επίπεδα και περιορίζεται σε 3×3 συνελίξεις και 2×2 συγκεντρώσεις.

5. ResNet [10]. Κέρδισε στον διαγωνισμό ILSVRC 2015. Κύριο χαρακτηριστικό της αρχιτεκτονικής είναι η συχνή χρήση επιπέδων κανονικοποίησης και η έλλειψη πλήρως-συνδεδεμένων επιπέδων στο τέλος του δικτύου [3].

Παρόλα αυτά, δυο βασικά ζητήματα στις βαθιές και πλατιές αρχιτεκτονικές είναι το υψηλό υπολογιστικό κόστος και η ανάγκη για μνήμη. Αυτό έχει ως αποτέλεσμα η ενσωμάτωση τέτοιων μοντέλων σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων να παραμένει πρόκληση.

Πιο πρόσφατα έχουν αναπτυχθεί (α) αρχιτεκτονικές ειδικά σχεδιασμένες για πλατφόρμες με περιορισμένους υπολογιστικούς πόρους, όπως η MobileNet, χαρακτηριστικό της οποίας είναι η διαχωρίσιμη κατά βάθος συνέλιξη με σκοπό την κατασκευή νευρωνικών δικτύων με σημαντική μείωση στον αριθμό των υπολογισμών και σημαντικά λιγότερες παραμέτρους, με μικρή μείωση στην ακρίβεια [11], αλλά και (β) αρχιτεκτονικές που έχουν προκύψει μέσω αυτοματοποιημένων μεθοδολογιών Neural Architecture Search (NAS), όπως η NASNet, χαρακτηριστικό της οποίας είναι η δυναμική εκμάθηση της βέλτιστης αρχιτεκτονικής ανάλογα με το σύνολο δεδομένων που χρησιμοποιείται. Επειδή αυτή η μέθοδος είναι κοστοβόρα για μεγάλα σύνολα δεδομένων, όπως το ImageNet, γίνεται πρώτα μια αναζήτηση για το βέλτιστο συνελικτικό επίπεδο στο σύνολο δεδομένων CIFAR-10 και έπειτα εφαρμόζεται στο ImageNet, δημιουργώντας και συνενώνοντας αντίγραφα αυτού του επιπέδου με ξεχωριστές παραμέτρους. Επιπλέον, χρησιμοποιείται μια τεχνική κανονικοποίησης (regularization) με

το όνομα ScheduledDropPath, η οποία βελτιώνει αισθητά τη γενίκευση των μοντέλων NASNet [12].

2.3 Βαθμονόμηση

Τα σύγχρονα βαθιά νευρωνικά δίκτυα έχουν δείξει πολύ σημαντική αύξηση στην ακρίβεια των προβλέψεων τους σε διεργασίες ταξινόμησης. Όμως, εκτός από την ακρίβεια, για εφαρμογές υψηλού ρίσκου, όπως η ιατρική διάγνωση ή τα αυτοοδηγούμενα αυτοκίνητα, είναι πολύ σημαντικό το νευρωνικό δίκτυο να παρέχει μια ένδειξη για την πιθανότητα οι προβλέψεις του να είναι λανθασμένες [13].

Συγκεκριμένα, αυτό επιτυγχάνεται με τη βαθμονόμηση (calibration), μέσω της οποίας το δίκτυο επιστρέφει πλέον μια κατανομή προβλέψεων, δηλαδή ένα διάνυσμα που οι συνιστώσες του δεν περιγράφουν μόνο το επίπεδο της εμπιστοσύνης (confidence) του δικτύου, αλλά πραγματικές πιθανότητες εμφάνισης για την εκάστοτε κλάση.

2.3.1 Ορισμοί

Διαγράμματα Αξιοπιστίας

Τα διαγράμματα αξιοπιστίας (reliability diagrams) αποτελούν μια οπτική αναπαράσταση του επιπέδου της βαθμονόμησης του μοντέλου. Αναπαριστούν την ακρίβεια ανά δείγμα ως συνάρτηση της εμπιστοσύνης. Συγκεκριμένα, τα δείγματα χωρίζονται σε M δοχεία (bins) B_m, όπου περιέχονται τα δείγματα των οποίων η τιμή της εμπιστοσύνης της πρόβλεψής τους είναι στο διάστημα $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$. Η ακρίβεια του δοχείου B_m υπολογίζεται ως:

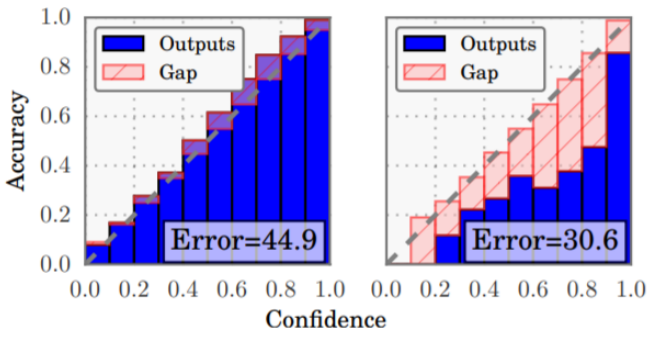
$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

ενώ η εμπιστοσύνη του δοχείου B_m υπολογίζεται ως:

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

όπου $\hat{p}_i$ η τιμή της εμπιστοσύνης για το δείγμα i.

Στην περίπτωση που το μοντέλου είναι απόλυτα βαθμονομημένο, το διάγραμμα αναμένεται να παρουσιάζει την ταυτοτική συνάρτηση (identity function) [13].



Εικόνα 2.6: *Διαγράμματα αξιοπιστίας*

Αναμενόμενο Σφάλμα Βαθμονόμησης

Μια πολύ χρήσιμη βαθμωτή μετρική για το επίπεδο βαθμονόμησης του μοντέλου είναι το Αναμενόμενο Σφάλμα Βαθμονόμησης (Expected Calibration Error - ECE).

Μέσω του διαχωρισμού των δειγμάτων σε M δοχεία ίδιας χωρητικότητας, υπολογίζεται ο σταθμισμένος μέσος όρος (weighted average) της διαφοράς ακρίβειας και εμπιστοσύνης για κάθε δοχείο, όπως φαίνεται στην παρακάτω έκφραση:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

όπου n το πλήθος των δειγμάτων [13].

Μέγιστο Σφάλμα Βαθμονόμησης

Παρόλο που το Αναμενόμενο Σφάλμα Βαθμονόμησης αποτελεί τον βασικό σύμβουλο για τη μέτρηση του επιπέδου βαθμονόμησης ενός μοντέλου, σε εφαρμογές υψηλού ρίσκου, ισχυρή ένδειξη είναι η διαφορά της ακρίβειας και της εμπιστοσύνης στη χειρότερη περίπτωση, ή αλλιώς το Μέγιστο Σφάλμα Βαθμονόμησης (Maximum Calibration Error - MCE) [13].

$$\mathrm{MCE} = \max_{m} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

Αρνητική Λογαριθμική Πιθανοφάνεια

Η Αρνητική Λογαριθμική Πιθανοφάνεια (Negative Log Likelihood - NLL) είναι μια ευρέως διαδεδομένη μετρική της πιθανοτικής ποιότητας ενός μοντέλου, η οποία είναι γνωστή και ως απώλεια διασταυρωμένης εντροπίας (cross-entropy loss) στα πλαίσια της Βαθιάς Μάθησης. Δεδομένων n δειγμάτων και ενός πιθανοτικού μοντέλου $\hat{\pi}(Y|X)$, όπου $\hat{\pi}(y_i|x_i)$ η τιμή εμπιστοσύνης που δίνει το μοντέλο για την πραγματική κλάση του δείγματος x_i, ορίζεται ως:

$$\mathcal{L} = -\sum_{i=1}^{n} \log\left(\hat{\pi}(y_i|x_i)\right)$$

Είναι γνωστό πως η Αρνητική Λογαριθμική Πιθανοφάνεια ελαχιστοποιείται αν και μόνο αν η $\hat{\pi}(Y|X)$ ταυτιστεί με την πραγματική κατανομή $\pi(Y|X)$ [13].

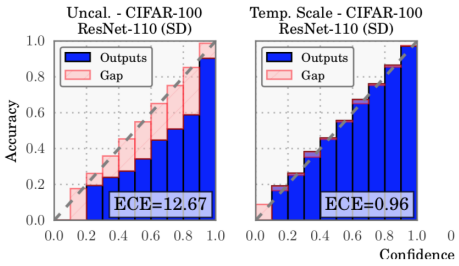
2.3.2 Στάθμιση Θερμοκρασίας

Υπάρχουν αρκετές μέθοδοι που αποσκοπούν στη βαθμονόμηση ενός μοντέλου. Στην παρούσα εργασία, επειδή εστιάζουμε στην περίπτωση της ταξινόμησης πολλαπλών κλάσεων (multiclass classification), θα αναλύσουμε τη Στάθμιση Θερμοκρασίας (Temperature Scaling).

Η συγκεκριμένη μέθοδος εντάσσει μια παράμετρο $T > 0$ (θερμοκρασία) και με δεδομένο το διάνυσμα εξόδου z_i του μοντέλου για την είσοδο x_i πριν την εφαρμογή της συνάρτησης ενεργοποίησης softmax σ, δίνει την βαθμονομημένη προβλέψη ως:

$$\hat{q}_i = \max_{k} \sigma(z_i/T)^{(k)}$$

Η βέλτιστη παράμετρος T υπολογίζεται με βάση την ελαχιστοποίηση της Αρνητικής Λογαριθμικής Πιθανοφάνειας στο σύνολο επικύρωσης (validation set). Επίσης, αξίζει να σημειωθεί ότι η Στάθμιση Θερμοκρασίας δεν επηρεάζει την ακρίβεια του μοντέλου [13]. Στην εικόνα 2.7 παρουσιάζεται το διάγραμμα αξιοπιστίας για το μοντέλο ResNet-110 πριν και μετά από Στάθμιση Θερμοκρασίας.



Εικόνα 2.7: *Στάθμιση θερμοκρασίας για το ResNet-110*

2.4 Κατανεμημένα Ζεύγη Βαθέων Νευρωνικών Δικτύων

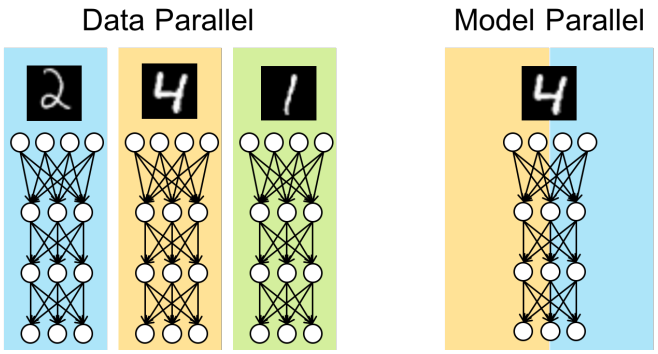
Καθώς ζούμε στην εποχή της πληροφορίας, η ποσότητα των διαθέσιμων δεδομένων μεγαλώνει με ραγδαίους ρυθμούς κάθε μέρα. Πολύ συχνά ένα σύνολο δεδομένων είναι τόσο μεγάλο που καθιστά ασύμφορη την αποθήκευση και την επεξεργασία του από μια μόνο μηχανή. Σε άλλες περιπτώσεις, η πολυπλοκότητα των μοντέλων μπορεί να είναι πολύ αυξημένη, καθιστώντας έτσι ανέφικτη την εκπαίδευση του από μια μόνο μηχανή. Αυτοί ακριβώς οι περιορισμοί δημιούργησαν την ανάγκη για την ανάπτυξη της Κατανεμημένης Μηχανικής Μάθησης (Distributed Machine Learning), στα πλαίσια της οποίας πολλοί εξυπηρετητές επικοινωνούν και ανταλλάσσουν χρήσιμα δεδομένα για την εκπαίδευση.

Ωστόσο, τα τελευταία χρόνια τα κατανεμημένα συστήματα δεν αφορούν πια μόνο ισχυρούς εξυπηρετητές, αλλά έχουν αρχίσει να ενσωματώνουν και τις κινητές συσκευές. Ένα αξιοσημείωτο πλεονέκτημα αυτού είναι η ελάττωση του φορτίου των εξυπηρετητών μέσω της αξιοποίησης των αδρανών πόρων των συσκευών για την εκπλήρωση των απαιτήσεων της εκπαίδευσης. Επιπλέον, εκτός από τη διαδικασία της εκπαίδευσης, τα κατανεμημένα συστήματα μπορούν να χρησιμοποιηθούν και για αποδοτικότερη συμπερασματολογία από την πλευρά των κινητών συσκευών, μέσω της επιλεκτικής εκτέλεσης για την αξιοποίηση των αυξημένων υπολογιστικών πόρων των εξυπηρετητών, όταν αυτό είναι απαραίτητο.

2.4.1 Κατανεμημένα Συστήματα Μηχανικής Μάθησης

Η μεθοδολογία που έρχεται να αντιμετωπίσει το πρόβλημα του αυξημένου όγκου των δεδομένων είναι ο παραλληλισμός σε επίπεδο δεδομένων (data parallelism). Η κύρια ιδέα αυτή της μεθοδολογίας είναι ο διαχωρισμός του ενιαίου συνόλου δεδομένων σε μικρά υποσύνολα και ο διαμοιρασμός αυτών σε ξεχωριστές μηχανές. Κατά αυτόν τον τρόπο, κάθε μηχανή εκπαιδεύει τη δική της εκδοχή του μοντέλου μόνο στα δεδομένα που είναι διαθέσιμα σε αυτή. Όταν οι επι μέρους διαδικασίες εκπαίδευσης ολοκληρωθούν, τα μοντέλα αυτά συνενώνονται σε ένα καθολικό μοντέλο. Αυτή η διαδικασία ονομάζεται συνένωση δεδομένων (data aggregation).

Από την άλλη, η μεθοδολογία που έρχεται να λύσει το πρόβλημα της αυξημένης πολυπλοκότητας των μοντέλων είναι ο παραλληλισμός σε επίπεδο μοντέλου (model parallelism), όπου το νευρωνικό δίκτυο κατανέμεται μεταξύ των διαθέσιμων μηχανών, με την κάθε μηχανή να διατηρεί ένα μέρος του συνολικού δικτύου. Κατά τη διάρκεια της εκπαίδευσης τα δεδομένα πρέπει να περάσουν από όλες τις μηχανές για επεξεργασία. Συνεπώς, είναι πολύ πιθανό η διαδικασία της εκπαίδευσης να πρέπει να γίνει σειριακά, αν η είσοδος κάποιων μηχανών εξαρτάται από την έξοδο κάποιων άλλων [14].



Εικόνα 2.8: *Παραλληλισμός σε επίπεδο δεδομένων και μοντέλου*

2.4.2 Βαθιά Μάθηση στο Νέφος

Η επιτυχία των αλγορίθμων της Βαθιάς Μάθησης κρύβεται πίσω από το γεγονός ότι οι κινητές συσκευές, αλλά και οι συσκευές του διαδικτύου των πραγμάτων (Internet of Things - IoT) συνεχώς παράγουν νέα δεδομένα. Η συνήθης πρακτική στις περισσότερες περιπτώσεις είναι η μετάδοση των παραγόμενων δεδομένων στο νέφος, όπου χρησιμοποιούνται για την εκπαίδευση μοντέλων με τη συμμετοχή ισχυρών εξυπηρετητών. Επίσης, όταν το κεντρικό μοντέλο έχει εκπαιδευτεί, ένας τρόπος να γίνει η συμπερασματολογία είναι η αποστολή του νέου δείγματος στο νέφος και η επιστροφή των αποτελεσμάτων πίσω στην κινητή συσκευή [15].

Παρόλα αυτά, όταν τα προσωπικά δεδομένα των χρηστών μεταδίδονται στο νέφος, η ιδιωτικότητα των χρηστών είναι πολύ πιθανόν να παραβιαστεί μέσω κακόβουλων ενεργειών. Επιπλέον, άλλοι λόγοι που καθιστούν την κεντρική εκπαίδευση και συμπερασματολογία στο νέφος μη συμφέρουσα είναι η καθυστέρηση (latency) της μεταφοράς των δεδομένων, το κόστος της μεταφοράς τους μέσω του διαδικτύου και η απαίτηση για συνεχή σύνδεση στο διαδίκτυο [16].

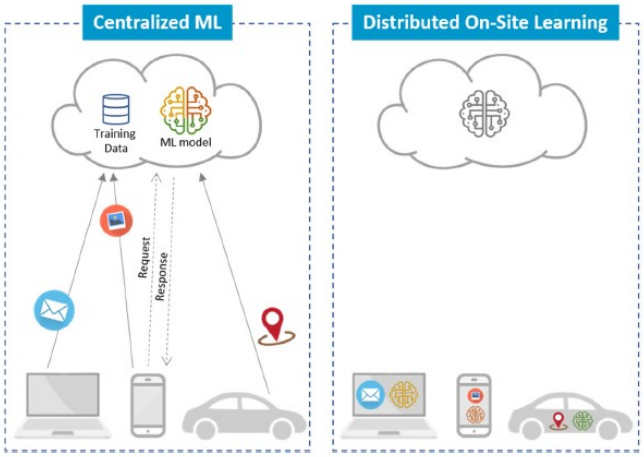
2.4.3 Βαθιά Μάθηση στις Συσκευές

Τα ζητήματα που δημιουργεί η μετάδοση των δεδομένων των χρηστών στο νέφος αποτέλεσαν κίνητρο για να ενσωματωθεί η Βαθιά Μάθηση στις κινητές συσκευές, όπου η εκπαίδευση και οι προβλέψεις βασίζονται στα παραγόμενα δεδομένα της συσκευής.

Σε αντίθεση με την παραδοσιακή πρακτική της μετάδοσης των δεδομένων στο νέφος, οι συσκευές των χρηστών κάνουν αίτηση σε εξυπηρετητή του νέφους για να τους αποσταλεί ένα προεκπαιδευμένο (pre-trained) δίκτυο. Στη συνέχεια, όταν έχουν εγκαταστήσει το μοντέλο,

το εκπαιδεύουν περαιτέρω πάνω σε τοπικά δεδομένα που παράγουν και το χρησιμοποιούν για την εξαγωγή προβλέψεων για νέα δεδομένα. Με αυτόν τον τρόπο τα ζητήματα της ασφάλειας των προσωπικών δεδομένων, της καθυστέρησης και του κόστους επιλύονται, αφού τα δεδομένα δεν αποχωρίζονται την συσκευή και επιπλέον δεν υπάρχει η απαίτηση για συνεχή σύνδεση στο διαδίκτυο [16].

Παρόλα αυτά, εγείρονται άλλα ζητήματα που αφορούν την τοπική εκτέλεση των νευρωνικών δικτύων και αξίζει να σημειωθούν. Αυτά είναι (α) η υψηλή απαίτηση σε χωρητικότητα για την τοπική εγκατάσταση των νευρωνικών δικτύων και (β) η ραγδαία κατανάλωση των ενεργειακών πόρων της συσκευής, λόγω των απαιτήσεων σε επεξεργαστική ισχύ (CPU, GPU) και μνήμη (RAM) [15]. Στην εικόνα 2.9 παρουσιάζεται η διαφορά μεταξύ κεντρικής και κατανεμημένης εκπαίδευσης.



Εικόνα 2.9: *Βαθιά Μάθηση στο νέφος και στις συσκευές*

2.4.4 Ζεύγη Νευρωνικών Δικτύων

Τα Ζεύγη Βαθιών Νευρωνικών Δικτύων (two-DNN Distributed Systems) αποτελούν μια ειδική κατηγορία κατανεμημένων/υβριδικών συστημάτων τα οποία αποτελούνται από:

- Ένα βαθύ νευρωνικό δίκτυο στην κινητή συσκευή, το οποίο εκτελείται τοπικά, χρησιμοποιώντας τους περιορισμένους πόρους της συσκευής.
- Ένα βαθύ νευρωνικό δίκτυο στο υπολογιστικό νέφος ή στην άκρη του δικτύου, το οποίο εκτελείται με την υποβοήθηση ενός ισχυρού εξυπηρετητή.

Χαρακτηριστικό πλεονέκτημα αυτών των συστημάτων είναι η επιλεκτική εκτέλεση, είτε στην κινητή συσκευή, είτε στον εξυπηρετητή, ανάλογα με τις ανάγκες της εκτέλεσης για το εκάστοτε δείγμα εισόδου, αλλά και τις δυναμικές παραμέτρους του συστήματος, όπως είναι η χρήση μνήμης και η μπαταρία.

Κεφάλαιο 3

Τεχνολογίες και Εργαλεία

Στο κεφάλαιο αυτό γίνεται μια παρουσίαση και σύντομη ανάλυση των βασικών τεχνολογιών που χρησιμοποιήθηκαν για τους στόχους της εργασίας.

3.1 Python

Η Python είναι μια διερμηνευόμενη (interpreted), γενικού σκοπού (general purpose) και υψηλού επιπέδου (high-level) γλώσσα προγραμματισμού. Περιέχει αρκετά αποδοτικές δομές δεδομένων υψηλού επιπέδου και δίνει την δυνατότητα ανάπτυξης αντικειμενοστραφούς κώδικα. Επιπλέον, περιέχει ένα εύρος βιβλιοθηκών, οι οποίες επεκτείνονται με νέα κομμάτια γραμμένα σε C ή C++. Η Python είναι λογισμικό ανοιχτού κώδικα και συντηρείται από τον οργανισμό Python Software Foundation (PSF) [17].

Δημιουργήθηκε από τον ολλανδό Guido van Rossum στα τέλη της δεκαετίας του 1980 στο Εθνικό Ινστιτούτο Έρευνας για Μαθηματικά και Επιστήμη Υπολογιστών (Centrum voor Wiskunde en Informatica - CWI) και από τότε έχει γίνει ιδιαίτερα δημοφιλής μεταξύ των προγραμματιστών, λόγω του καθαρού και ευθέος συντακτικού της [18]. Η γλώσσα θεωρείται διάδοχος της ABC και ξεκίνησε ως γλώσσα σεναρίων (scripting language) για το λειτουργικό σύστημα Amoeba.

Με βάση τον ίδιο τον Guido von Rossum, η Python σχεδιάστηκε με τέτοιο τρόπο έτσι ώστε να είναι εύκολα επεκτάσιμη, δίνοντας στους προγραμματιστές τη δυνατότητα να προσθέτουν στοιχεία ανάλογα με τις ανάγκες τους.

Τέλος, η Python εξελίσσεται με γρήγορους ρυθμούς, δημοσιεύοντας νέες εκδόσεις μέσω του εργαλείου Python Enhancement Proposals (PEPs), το οποίο είναι μια συλλογή από τυποποιημένα κείμενα που έχουν ως σκοπό να παρουσιάσουν προτάσεις για νέα χαρακτηριστικά της γλώσσας [17].

3.2 PyCharm

Το PyCharm αποτελεί ένα Ολοκληρωμένο Περιβάλλον Ανάπτυξης (Integrated Development Environment - IDE), το οποίο χρησιμοποιείται για την ανάπτυξη εφαρμογών λογισμικού, κυρίως σε γλώσσα Python και αναπτύχθηκε από την Τσεχική εταιρεία JetBrains. Μεταξύ των εργαλείων που παρέχει βρίσκονται τα εξής: ανάλυση κώδικα (code analysis),

ενσωματωμένο εργαλείο ελέγχου μονάδων (unit tester), ενσωμάτωση με συστήματα ελέγχου εκδόσεων (Version Control Systems - VCS), γραφικό εργαλείο εντοπισμού σφαλμάτων (graphical debugger), αλλά και εργαλεία για ανάπτυξη λογισμικού σε περιβάλλον Επιστήμης Δεδομένων (Data Science) με την κατανομή Anaconda.

3.3 Colaboratory

Το Colaboratory ή Colab είναι προϊόν της Google Research. Επιτρέπει την συγγραφή και την εκτέλεση κώδικα Python μέσω του περιηγητή ιστού (web browser) και θεωρείται κατάλληλο για διεργασίες Μηχανικής Μάθησης, ανάλυσης δεδομένων και εκπαίδευσης. Ειδικότερα, το Colaboratory είναι μια υπηρεσία που βασίζεται στο Jupyter η οποία δεν απαιτεί εγκατάσταση και διαθέτει ελεύθερη πρόσβαση σε υπολογιστικούς πόρους, όπως GPUs και TPUs [19].

3.4 TensorFlow

Το TensorFlow είναι μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα για Μηχανική Μάθηση. Οι λειτουργίες που παρέχει είναι πολλές, αλλά η βασική του συνεισφορά έγκειται στην παροχή ενός άρτια δομημένου πλαισίου για την εκπαίδευση και την εκτέλεση βαθιών νευρωνικών δικτύων.

Το TensorFlow υπήρξε διάδοχος του DistBelief, ενός συστήματος για Μηχανική Μάθηση, βασισμένο σε βαθιά νευρωνικά δίκτυα που αναπτύχθηκε από την Google Brain το 2011. Μετά την επιτυχία του DistBelief, η Google συνέταξε μια ομάδα με κύριο στόχο την αναβάθμιση του DistBelief σε μια πιο γρήγορη και ισχυρή βιβλιοθήκη, η οποία πήρε το όνομα TensorFlow [20].

3.5 Keras

Το Keras είναι μια βιβλιοθήκη ανοιχτού κώδικα που χρησιμοποιείται σαν διεπαφή (interface) για τη βιβλιοθήκη TensorFlow με σκοπό την ανάπτυξη λογισμικού με χρήση βαθιών νευρωνικών δικτύων. Παρέχει μια επεκτάσιμη και φιλική στο χρήστη διεπαφή, μέσω της οποίας γίνεται εφικτή η γρήγορη κατασκευή και αλληλεπίδραση με βαθιά νευρωνικά δίκτυα. Το Keras ήταν αποτέλεσμα του ερευνητικού προγράμματος ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System) με κύριο δημιουργό του τον François Chollet [21][22].

Το Keras διαθέτει ένα εύρος υλοποιημένων χρήσιμων εργαλείων για Βαθιά Μάθηση, όπως: επίπεδα, συναρτήσεις ενεργοποίησης, αλγορίθμους βελτιστοποίησης και εργαλεία για την προεπεξεργασία των δεδομένων εικόνας ή κειμένου.

3.6 ImageNet

Το ImageNet είναι ένα σύνολο δεδομένων, ή αλλιώς μια βάση δεδομένων που περιέχει εικόνες ταξινομημένες με οδηγό την ιεραρχία WordNet για τα ουσιαστικά, με κάθε κλάση να

περιέχει από εκατοντάδες μέχρι χιλιάδες δείγματα. Το σύνολο αυτό έπαιξε πολύ σημαντικό ρόλο στην ανάπτυξη της Βαθιάς Μάθησης και της Όρασης Υπολογιστών (Computer Vision), ενώ διατίθεται δωρεάν για ερευνητικούς σκοπούς [23].

Κάθε έννοια με νόημα, πιθανώς περιγραφόμενη από παραπάνω από μια λέξεις, αποτελεί ενα σύνολο συνωνύμων (synonyn set - synset) στο WordNet. Υπάρχουν περισσότερα από 100.000 σύνολα συνωνύμων, εκ των οποίων τουλάχιστον τα 80.000 είναι ουσιαστικά.

Χαρακτηριστικό του ImageNet είναι ότι διαθέτει κατά μέσο όρο 1000 δείγματα ανά σύνολο συνωνύμων, με εικόνες ελεγχόμενου περιεχομένου και ετικέτες που έχουν οριστεί από ανθρώπους. Στην εικόνα 3.1 παρουσιάζεται ένα αντιπροσωπευτικό υποσύνολο δειγμάτων του συνόλου δεδομένων.

Εικόνα 3.1: *Παραδείγματα δειγμάτων του ImageNet*

Κεφάλαιο 4

Συνιστώσες Εφαρμογής

Στο κεφάλαιο αυτό γίνεται μια περιγραφή υψηλού επιπέδου των συνιστωσών της εφαρμογής. Ο χώρος (space) της εφαρμογής αποτελείται από ένα βαθύ νευρωνικό δίκτυο σε μια κινητή συσκευή και ένα βαθύ νευρωνικό δίκτυο στο νέφος, σχηματίζοντας ένα κατανεμημένο σύστημα ζεύγους.

4.1 Εντοπισμός Ακρίβειας

Γίνεται εύκολα αντιληπτό ότι η μετρική που εκφράζει την ακρίβεια πρώτης πρόβλεψης (top-1 accuracy) δεν δίνει μια πλήρη εικόνα της ακρίβειας ενός νευρωνικού δικτύου. Αυτό είναι γεγονός, διότι η ακρίβεια πρώτης πρόβλεψης προκύπτει μόνο από το όρισμα που μεγιστοποιεί την εμπιστοσύνη του μοντέλου και δεν λαμβάνει καθόλου υπόψη την κατανομή της εξόδου.

Ο Εντοπισμός Ακρίβειας (Accuracy Detection) αποτελεί το πρώτο μέρος της ενιαίας εφαρμογής και έχει ως κύριο στόχο να αξιολογήσει την πραγματική ακρίβεια του μοντέλου με τη χρήση ποικίλων μετρικών. Οι δύο βασικές προσεγγίσεις είναι η αλγοριθμική (algorithmic approach) και η συστημική (system approach).

4.1.1 Αλγοριθμική Προσέγγιση

Κατά την αλγοριθμική προσέγγιση, η προσοχή μας εστιάζεται στην μορφή της κατανομής της εξόδου του δικτύου της κινητής συσκευής και χρησιμοποιούνται μετρικές οι οποίες υποδεικνύουν το πόσο «μυτερή» (spiky) ή διάχυτη (diffuse) είναι, έτσι ώστε να εξάγουμε συμπεράσματα για το πόσο σίγουρο είναι το μοντέλο. Αξίζει να σημειωθεί σε αυτό το σημείο ότι βασικό ρόλο σε αυτή την προσέγγιση παίζει η βαθμονόμηση του μοντέλου, καθώς για να γίνει η θεώρηση ότι η κατανομή του διανύσματος εξόδου είναι καλός σύμβουλος της ακρίβειας του μοντέλου, πρέπει οι συνιστώσες του να είναι πραγματικές πιθανότητες και όχι απλά επίπεδα εμπιστοσύνης.

Μετρικές

Οι κύριες μετρικές που θα χρησιμοποιηθούν στην αλγοριθμική προσέγγιση είναι οι εξής:

1. Ακρίβεια πρώτης πρόβλεψης (Top-1 Accuracy):

$$\text{top-1 acc.} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i = y_i)$$

όπου n το πλήθος των δειγμάτων, $\hat{y}_i$ η πρόβλεψη του μοντέλου και y_i η πραγματική κλάση του δείγματος.

2. Ακρίβεια πρώτων 5 προβλέψεων (Top-5 Accuracy):

$$\text{top-5 acc.} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{Y}_i}(y_i)$$

όπου $\hat{Y}_i$ το σύνολο των πρώτων 5 προβλέψεων του μοντέλου.

3. Εντροπία (Entropy):

$$\text{H} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}\cdot\log(p_{ij})$$

όπου k το πλήθος των κλάσεων και p_{ij} η εμπιστοσύνη που δίνει το μοντέλο στην κλάση j για το i-οστό δείγμα.

4. Διαφορά πρώτης και δεύτερης Εμπιστοσύνης (Best-versus-Second Best - BvSB):

$$\text{BvSB} = \frac{1}{n}\sum_{i=1}^{n}\left(p_i^{(1)} - p_i^{(2)}\right)$$

όπου $p_i^{(1)}$ και $p_i^{(2)}$ η πρώτη και δεύτερη εμπιστοσύνη του μοντέλου για το i-οστό δείγμα αντίστοιχα.

5. Αναμενόμενη Μέγιστη Εμπιστοσύνη (Expected Maximum Confidence):

$$\text{Max conf.} = \frac{1}{n}\sum_{i=1}^{n} p_i^{(1)}$$

6. Αναμενόμενη Ακρίβεια (Expected Accuracy):

$$\text{exp. accuracy} = \frac{1}{n}\sum_{i=1}^{n} q_i$$

όπου q_i η εμπιστοσύνη του μοντέλου για την πραγματική κλάση του i-οστού δείγματος.

7. Δείκτης Gini (Gini Index):

$$\text{G} = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}^2$$

Ο δείκτης Gini, όπως ακριβώς και η Εντροπία, παρέχει μια πληροφορία για τη μορφή της κατανομής εξόδου. Όσο πιο «μυτερή» (spiky) είναι η κατανομή, τόσο πιο κοντά στο μηδέν είναι ο δείκτης Gini. Αντίθετα, όταν η κατανομή είναι διάχυτη και τείνει να γίνει ομοιόμορφη, ο δείκτης Gini τείνει στη μονάδα.

8. Διασταυρωμένη Εντροπία (Cross-entropy):

$$\text{H}_c = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} l_{ij} \cdot \log(p_{ij})$$

όπου το l_{ij} ισούται με μονάδα μόνο αν η πραγματική κλάση του i-οστού δείγματος είναι η j, αλλιώς ισούται με μηδέν.

9. Ζυγισμένη Ακρίβεια (Balanced Accuracy):

$$\text{Balanced acc.} = \frac{1}{k}\sum_{j=1}^{k} a_j$$

όπου a_j η ακρίβεια πρώτης πρόβλεψης περιορισμένη στα δείγματα της κλάσης j.

4.1.2 Συστημική Προσέγγιση

Κατά τη συστημική προσέγγιση, γίνεται η θεώρηση ότι το μοντέλο που βρίσκεται στο νέφος αποτελεί σημείο αναφοράς (reference point), μιας και η ακρίβεια του θα είναι μεγαλύτερη από αυτή του δικτύου της συσκευής. Συνεπώς, χρησιμοποιούνται μετρικές που υποδεικνύουν πόσο κοντά είναι η έξοδος του δικτύου της κινητή συσκευής σε αυτή του δικτύου του νέφους.

Μετρικές

1. Διασταυρωμένη Εντροπία (Cross-entropy):

$$\text{H}_c = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij} \cdot \log(p_{ij})$$

όπου r_{ij}, p_{ij} τα επίπεδα εμπιστοσύνης στην κλάση j για το i-οστό δείγμα του μοντέλου του νέφους και του μοντέλου της κινητής συσκευής αντίστοιχα.

2. Διαφορά πρώτης εμπιστοσύνης (Confidence Difference):

$$\text{Confidence diff.} = \frac{1}{n}\sum_{i=1}^{n}\left|p_i^{(1)} - p_i^{(2)}\right|$$

όπου $p_i^{(1)}$, $p_i^{(2)}$ τα επίπεδα εμπιστοσύνης στην πραγματική κλάση για το i-οστό δείγμα του μοντέλου του νέφους και του μοντέλου της κινητής συσκευής αντίστοιχα.

4.2 Διαδικασία Απόφασης

Ένα ενδιάμεσο στάδιο της εφαρμογής, πριν προχωρήσει στη διόρθωση της ακρίβειας είναι η Διαδικασία Απόφασης (Decision Process). Κατά την διαδικασία αυτή, το σύστημα από μέρους της κινητής συσκευής θα πρέπει με κάποιο τρόπο να αποφασίζει για ποια από τα δείγματα εισόδου που λαμβάνει είναι πιθανό να δώσει λανθασμένη πρόβλεψη, έτσι ώστε να προχωρά σε διόρθωση ακρίβειας.

Ο πιο απλός τρόπος να ληφθεί αυτή η απόφαση είναι μέσω της κατανομής της εξόδου (επίπεδο softmax) του νευρωνικού δικτύου της κινητής συσκευής. Σε περίπτωση που η κατανομή αυτή δεν είναι αρκετά «μυτερή» (spiky), δηλαδή δεν υπάρχει μεγάλη διαφορά μεταξύ της πρώτης εμπιστοσύνης και των υπολοίπων, αυτό αποτελεί ένδειξη ότι η πρόβλεψη του μοντέλου θα είναι λανθασμένη με σημαντική πιθανότητα.

Ένας ακόμη τρόπος που εξετάστηκε για την λήψη της συγκεκριμένης απόφασης είναι η εξέταση των χαρακτηριστικών (features) που μπορούν να εξαχθούν από μια εικόνα μέσω Βαθιάς Μάθησης.

4.3 Αντιστάθμιση Ακρίβειας

Εφόσον αποφασιστεί από την προηγούμενη διαδικασία ποιες εικόνες χρειάζονται διόρθωση ακρίβειας, έρχεται το επόμενο στάδιο, το οποίο ονομάζεται Αντιστάθμιση Ακρίβειας (Accuracy Refinement).

Ο πιο απλός και προφανής τρόπος να διορθωθεί η ακρίβεια είναι η αποστολή των «προβληματικών» δειγμάτων εισόδου στο νέφος, έτσι ώστε η πρόβλεψη να γίνει από το ισχυρότερο μοντέλο. Άλλη μια μέθοδος που διερευνήθηκε είναι η διόρθωση της ακρίβειας μέσω της εξαγωγής της κλάσης υψηλού επιπέδου (high-level) για την εκάστοτε εικόνα.

Κεφάλαιο 5

Διερεύνηση

Στο κεφάλαιο αυτό γίνεται μια περιγραφή των προσεγγίσεων που διερευνήθηκαν προκειμένου να καταλήξουμε στη σχεδίαση του βέλτιστου αλγορίθμου για την αντιστάθμιση της ακρίβειας του συστήματος.

5.1 Μέθοδοι Απόφασης

Στα πλαίσια της παρούσας εργασίας, διερευνήθηκαν πολλές προσεγγίσεις με σκοπό την εύρεση της βέλτιστης μεθόδου ή συνδυασμού μεθόδων. Ως βέλτιστη περιγράφεται η μέθοδος κατά την οποία η ακρίβεια που πετυχαίνει το σύστημα είναι η μεγαλύτερη δυνατή, χωρίς να έχουμε υπερφόρτωση του εξυπηρετητή στο νέφος.

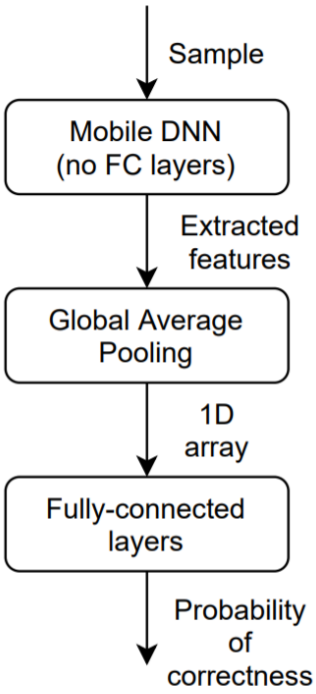
5.1.1 Εξαγωγή Χαρακτηριστικών

Μια από τις μεθόδους που διερευνήθηκε χρησιμοποιεί τα χαρακτηριστικά που μπορούν να εξαχθούν από το δείγμα εισόδου ώστε να παρθεί η απόφαση σχετικά με το αν η εικόνα εισόδου είναι «προβληματική» ή όχι.

Συγκεκριμένα, με τη χρήση ενός βαθέος συνελικτικού νευρωνικού δικτύου αρχιτεκτονικής MobileNet το οποίο είναι εκπαιδευμένο εκ των προτέρων στο σύνολο δεδομένων ImageNet, εφαρμόζουμε αυτόματη εξαγωγή χαρακτηριστικών (feature extraction) στο σύνολο επικύρωσης του ImageNet και έπειτα αυτά τα χαρακτηριστικά τροφοδοντούται σαν είσοδος για την εκπαίδευση ενός δυαδικού ταξινομητή (binary classifier), ο οποίος θα προβλέπει αν ένα δείγμα εισόδου είναι πιθανό να ταξινομηθεί λανθασμένα από το μοντέλο της κινητής συσκευής.

Εκτός από τα χαρακτηριστικά εισόδου, για την εκπαίδευση του δυαδικού ταξινομητή απαιτούνται επίσης ετικέτες για τις πραγματικές κλάσεις των δειγμάτων, οπότε το σύνολο επικύρωσης του ImageNet τροφοδοτείται επίσης στο μοντέλο της κινητής συσκευής. Όταν είναι πλέον γνωστό ποια από τα δείγματα ταξινομήθηκαν ορθώς και ποια όχι, αντιστοιχίζεται σε αυτά η αντίστοιχη ετικέτα: 0 αν το δείγμα ταξινομήθηκε ορθώς και 1 διαφορετικά.

Αφότου εκπαιδευτεί ο δυαδικός ταξινομητής, το ζεύγος feature extractor - binary classifier μπορεί να χρησιμοποιηθεί για να προβλέψει αν το μοντέλο της κινητής συσκευής θα ταξινομήσει σωστά ένα δοθέν δείγμα εισόδου. Στην εικόνα 5.1 παρουσιάζεται το διάγραμμα του ταξινομητή.



Εικόνα 5.1: *Διάγραμμα του ταξινομητή βασισμένου στα χαρακτηριστικά*

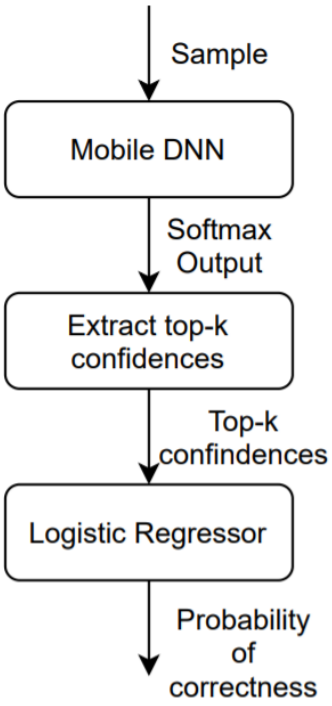
5.1.2 Πρώτα k Επίπεδα Εμπιστοσύνης

Μια άλλη μέθοδος που εξετάστηκε βασίζεται στις πρώτες k καλύτερες τιμές εμπιστοσύνης που προκύπτουν από την έξοδο (επίπεδο softmax) του μοντέλου κινητής συσκευής, με την k να είναι ρυθμιζόμενη (tunable) παράμετρος.

Ειδικότερα, οι πρώτες k καλύτερες τιμές εμπιστοσύνης τροφοδοντούται σαν είσοδος για την εκπαίδευση ενός δυαδικού ταξινομητή (binary classifier), ο οποίος θα προβλέπει αν ένα δείγμα εισόδου είναι πιθανό να ταξινομηθεί λανθασμένα από το μοντέλο της κινητής συσκευής.

Εκτός από τις πρώτες k καλύτερες τιμές εμπιστοσύνης, για την εκπαίδευση του δυαδικού ταξινομητή απαιτούνται επίσης ετικέτες για τις πραγματικές κλάσεις των δειγμάτων, οπότε το σύνολο επικύρωσης του ImageNet τροφοδοτείται επίσης στο μοντέλο της κινητής συσκευής. Όταν είναι πλέον γνωστό ποια από τα δείγματα ταξινομήθηκαν ορθώς και ποια όχι, αντιστοιχίζεται σε αυτά η αντίστοιχη ετικέτα: 0 αν το δείγμα ταξινομήθηκε ορθώς και 1 διαφορετικά.

Αφότου εκπαιδευτεί ο δυαδικός ταξινομητής, μπορεί να χρησιμοποιηθεί για να προβλέψει αν το μοντέλο της κινητής συσκευής θα ταξινομήσει σωστά ένα δοθέν δείγμα εισόδου. Στην εικόνα 5.2 παρουσιάζεται το διάγραμμα του ταξινομητή.



Εικόνα 5.2: *Διάγραμμα του ταξινομητή βασισμένου στις πρώτες k τιμές εμπιστοσύνης*

5.1.3 Μετρικές

Μια πιο άμεση και γρήγορη μέθοδος είναι η χρήση των μετρικών της αλγοριθμικής συνιστώσας του Εντοπισμού Αρκίβειας και η λήψη της απόφασης με βάση κάποια κατώφλια (thresholds).

Οι τιμές των κατωφλιών ρυθμίζονται στις βέλτιστες τιμές τους μέσω του αλγορίθμου Constrained Optimization BY Linear Approximation (COBYLA), μεγιστοποιώντας την ακρίβεια που πετυχαίνει το σύστημα με την αποστολή περιορισμένου πλήθους δειγμάτων στον εξυπηρετητή. Ο συγκεκριμένος αλγόριθμος λειτουργεί επαναληπτικά προσεγγίζοντας το πραγματικό πρόβλημα βελτιστοποίησης μέσω προβλημάτων γραμμικού προγραμματισμού (Linear Programming) και επιλέχθηκε διότι η παράγωγος της συνάρτησης που σκοπεύουμε να μεγιστοποιήσουμε δεν είναι γνωστή.

5.2 Μέθοδοι Αντιστάθμισης

Σε αυτό το σημείο περιγράφονται μέθοδοι που διερευνήθηκαν και έχουν ως στόχο την αντιστάθμιση της ακρίβειας, αφότου ένα δείγμα έχει σημανθεί ως «προβληματικό» μέσω κάποιας από τις μεθόδους απόφασης.

5.2.1 Αποστολή στο Νέφος

Η πιο απλή προσέγγιση είναι τα δείγματα που σημαίνονται ως «προβληματικά» να αποστέλλονται απευθείας στο νέφος. Κατά αυτόν τον τρόπο, η συμπερασματολογία (inference)

για τις συγκεκριμένες εικόνες γίνεται από το ισχυρό μοντέλο στον εξυπηρετητή και τα αποτελέσματα αυτής επιστρέφονται στην κινητή συσκευή.

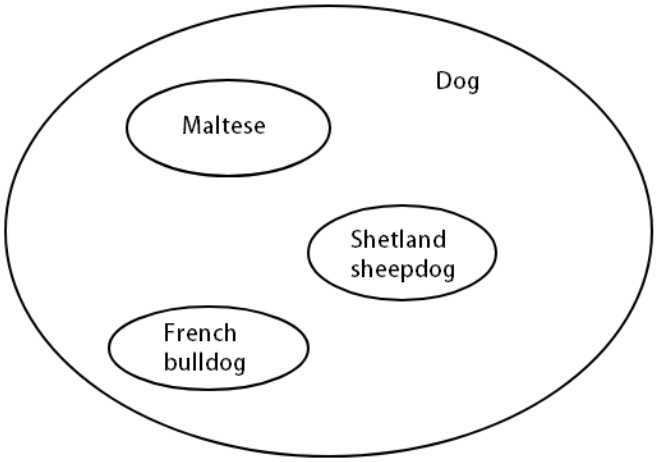
Βέβαια, όπως έχει ήδη αναφερθεί, αυτή η πρακτική, όχι μόνο θέτει σε κίνδυνο την ιδιωτικότητα των δεδομένων του χρήστη, αλλά έχει και άλλα αρνητικά σημεία, όπως η ενδεχόμενη καθυστέρηση, το κόστος της μεταφοράς των δεδομένων και η απαίτηση για συνεχή σύνδεση στο διαδίκτυο.

5.2.2 Εξαγωγή Κλάσης Υψηλού Επιπέδου

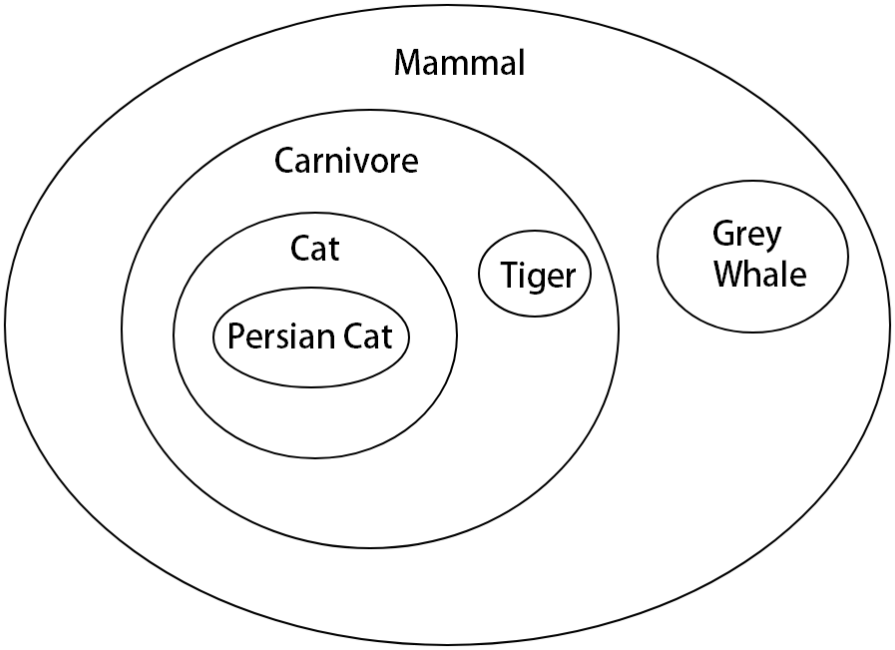
Η μέθοδος της εξαγωγής κλάσης υψηλού επιπέδου (high-level class extraction) αποτελεί μια μέθοδο διόρθωσης της ακρίβειας χωρίς την συμμετοχή του ισχυρού μοντέλου στο νέφος.

Πολλές φορές τα προβλήματα της επικοινωνίας της κινητής συσκευής με το νέφος μπορεί να είναι έντονα. Συγκεκριμένα, η μεγάλη καθυστέρηση, η μη σύνδεση στο διαδίκτυο ή οι πολλές ήδη αιτήσεις στον εξυπηρετητή ενδέχεται να μην τον καθιστούν άμεσα διαθέσιμο για την συμπερασματολογία του δείγματος.

Συνεπώς, με μια κατηγοριοποίηση των κλάσεων χαμηλού επιπέδου σε κλάσεις υψηλού επιπέδου, δίνουμε την δυνατότητα στο τοπικό μοντέλο της κινητής συσκευής να διορθώνει την ακρίβεια του, μέσω της συμπερασματολογίας για την κλάση υψηλού επιπέδου της εικόνας. Δηλαδή, αντί για την λανθασμένη ειδική κλάση, το σύστημα δίνει σαν έξοδο την αντίστοιχη γενική κλάση που, αν και πιο γενική, είναι σωστή. Στις εικόνες 5.3, 5.4 παρουσιάζονται τμήματα της ιεραρχίας του ImageNet.



Εικόνα 5.3: *Παράδειγμα τμήματος της ιεραρχίας του ImageNet*



Εικόνα 5.4: *Παράδειγμα τμήματος της ιεραρχίας του ImageNet*

Κεφάλαιο 6

Ανάπτυξη

Σε αυτό το κεφάλαιο περιγράφεται η ανάπτυξη του συστήματος που βασίζεται στη μελέτη που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Επιπλέον, στο παράρτημα Α΄ παρατίθεται ο πηγαίος κώδικας του συστήματος για περαιτέρω εμβάθυνση.

6.1 Μοντέλα Ζεύγους

Σε αυτό το σημείο γίνεται μια ανασκόπηση των μοντέλων Μηχανικής Μάθησης που χρησιμοποιήθηκαν για την ανάπτυξη του συστήματος.

Για την προσομοίωση του μοντέλου της κινητής συσκευής επιλέχθηκε η αρχιτεκτονική MobileNet. Η συγκεκριμένη αρχιτεκτονική επιλέχθηκε λόγω του μικρού μεγέθους της και της σχετικά χαμηλής ακρίβειας που πετυχαίνει, χαρακτηριστικά τα οποία τονίζουν καλύτερα το πρόβλημα που η εργασία καλείται να λύσει.

Για την προσομοίωση του μοντέλου του νέφους επιλέχθηκε η αρχιτεκτονική NASNet-Large. Η συγκεκριμένη αρχιτεκτονική επιλέχθηκε λόγω του μεγάλου μεγέθους της και της σχετικά υψηλής ακρίβειας που πετυχαίνει, γεγονός που τονίζει καλύτερα τη διαφορά των δύο μοντέλων στην ποιότητα των προβλέψεών τους.

Στον πίνακα 6.1 παρουσιάζονται τα βασικά χαρακτηριστικά των μοντέλων MobileNet και NASNetLarge όπως διατίθενται στην βιβλιοθήκη Keras Applications.

Μοντέλο	Top-1	Top-5	Παράμετροι	FLOPs	Αποτύπωμα Μνήμης	Μέγεθος Εισόδου
MobileNet	0.704	0.895	4 M	569 M	16 MB	224x224
NASNetLarge	0.825	0.960	83 M	24882 M	343 MB	331x331

Πίνακας 6.1: *Χαρακτηριστικά των μοντέλων MobileNet και NASNetLarge*

6.2 Μοντέλα Απόφασης

6.2.1 Ταξινομητής βασισμένος στα χαρακτηριστικά

Το μοντέλο απόφασης το οποίο βασίζεται στη μέθοδο εξαγωγής χαρακτηριστικών απαρτίζεται, όπως έχει ήδη αναφερθεί, από δύο τμήματα:

1. Εξαγωγέας χαρακτηριστικών (feature extractor): Αποτελείται από ένα υποσύνολο της αρχιτεκτονικής του μοντέλου της κινητής συσκευής, δηλαδή της αρχιτεκτονικής Μο-

bileNet. Συγκεκριμένα, περιέχει τα επίπεδα από το επίπεδο εισόδου, έως και το τελευταίο επίπεδο καθολικής μέσης συγκέντρωσης (global average pooling) του MobileNet, το οποίο επιστρέφει ένα μονοδιάστατο διάνυσμα χαρακτηριστικών 1024 στοιχείων.

Το γεγονός ότι χρησιμοποιείται η ίδια αρχιτεκτονική με το μοντέλο της κινητής συσκευής έχει το πλεονέκτημα ότι, αν τελικά δε χρειαστεί να αποσταλεί το δείγμα εισόδου στο νέφος και άρα η συμπερασματολογία εκτελεστεί από το μοντέλο της κινητής συσκευής, τότε αυτό δεν χρειάζεται να γίνει από την αρχή, αλλά από το σημείο που έχει σταματήσει ο εξαγωγέας χαρακτηριστικών.

2. Δυαδικός ταξινομητής (binary classifier): Αποτελείται από πλήρως-συνδεδεμένα επίπεδα και ένα επίπεδο εξόδου με σιγμοειδή (sigmoid) συνάρτηση ενεργοποίησης. Βασική λειτουργία του είναι να λαμβάνει τα χαρακτηριστικά από τον εξαγωγέα και να τα απεικονίζει σε έναν αριθμό από το 0 εώς το 1, ο οποίος αναπαριστά την πιθανότητα σφάλματος του τοπικού μοντέλου.

6.2.2 Ταξινομητής βασισμένος στις τιμές εμπιστοσύνης

Το μοντέλο απόφασης το οποίο βασίζεται στις πρώτες 20 τιμές των επιπέδων εμπιστοσύνης που επιστρέφει το τοπικό μοντέλο, είναι ένα μοντέλο λογιστικής παλινδρόμησης. Η επιλογή του 20 ως τιμή της υπερπαραμέτρου k έγινε μετά από διαπίστωση ότι η ακρίβεια του τελικού μοντέλου μεγιστοποιούνταν για τιμές κοντά στο 20.

Βασική λειτουργία του είναι να λαμβάνει τις πρώτες 20 τιμές εμπιστοσύνης του τοπικού μοντέλου και να τις απεικονίζει σε έναν αριθμό από το 0 εώς το 1, ο οποίος αναπαριστά την πιθανότητα σφάλματος του τοπικού μοντέλου.

6.3 Μονάδες

Σε αυτό το σημείο παρουσιάζονται όλες οι μονάδες (modules) κώδικα της εφαρμογής και αναλύεται η λειτουργία τους. Στις εικόνες 6.1 και 6.2 παρουσιάζονται τα διαγράμματα που περιγράφουν την λειτουργία των μονάδων του συστήματος.

6.3.1 Metrics

Στη μονάδα Metrics γίνεται η ανάπτυξη όλων των μετρικών που κάνει χρήση το σύστημα. Ορίζονται οι μετρικές τόσο για την αλγοριθμική προσέγγιση και συστημική προσέγγιση του Εντοπισμού Ακρίβειας, όσο και για τη βαθμονόμηση του μοντέλου.

6.3.2 Temperature Scaling

Στην κλάση TemperatureScaling αναπτύσσεται η μέθοδος της Στάθμισης Θερμοκρασίας για τη βαθμονόμηση του μοντέλου, η οποία αναλύθηκε στην υποενότητα 2.3.2. Οι κύριες μέθοδοι που την απαρτίζουν είναι οι: __init__, fit, predict και evaluate.

Η μέθοδος __init__ λειτουργεί ως κατασκευαστής (constructor) της κλάσης και αρχικοποιεί την θερμοκρασία ως μονάδα, ενώ παράλληλα ορίζει το πλήθος των επαναλήψεων μέχρι τη σύγκλιση στην βέλτιστη θερμοκρασία και την αριθμητική μέθοδο που θα χρησιμοποιηθεί.

Στη συνέχεια, η μέθοδος fit κάνει χρήση της βιβλιοθήκης βελτιστοποίησης scipy.optimize της Python με σκοπό την εύρεση της θερμοκρασίας που ελαχιστοποιεί την αρνητική λογαριθμική πιθανοφάνεια. Πιο συγκεκριμένα, η αριθμητική μέθοδος που εφαρμόζεται είναι ο αλγόριθμος Broyden–Fletcher–Goldfarb–Shanno (BFGS) με 50 επαναλήψεις.

Κατόπιν, η μέθοδος predict λαμβάνει την βέλτιστη τιμή της θερμοκρασίας, όπως αυτή καθορίστηκε από την fit και υπολογίζει τη βαθμονομημένη έξοδο του μοντέλου.

Τέλος, η μέθοδος evaluate λαμβάνει ως ορίσματα τις βαθμονομημένες προβλέψεις ενός μοντέλου μαζί με τις ετικέτες για ένα σύνολο αξιολόγησης και επιστρέφει ένα διάνυσμα με συνιστώσες τις μετρικές: ECE, MCE και NLL, οι οποίες παρέχουν μια αρκετά πλήρη εικόνα της βαθμονόμησης του μοντέλου.

6.3.3 Accuracy Detection

Η μονάδα Accuracy Detection αποτελείται από δύο συναρτήσεις οι οποίες επιστρέφουν πλαίσια δεδομένων (Dataframes) που περιγράφουν την ακρίβεια του μοντέλου μέσω των μετρικών που έχουν ήδη αναπτυχθεί.

Συγκεκριμένα, η μέθοδος evaluate_accuracy_algo δέχεται ως ορίσματα τις προβλέψεις ενός μοντέλου μαζί με τις ετικέτες για ένα σύνολο αξιολόγησης και επιστρέφει τις μετρικές: top-1 accuracy, top-5 accuracy, entropy, BvSB, maximum confidence, expected accuracy, Gini index, balanced accuracy, cross-entropy.

Από την άλλη, η μέθοδος evaluate_accuracy_sys δέχεται ως ορίσματα τις προβλέψεις και των δύο μοντέλων μαζί με τις ετικέτες για ένα σύνολο αξιολόγησης και επιστρέφει τις μετρικές: cross-entropy, confidence difference.

6.3.4 Threshold Tuning

Η μονάδα Threshold Tuning έχει ως βασικό στόχο τη βελτιστοποίηση της διαδικασίας απόφασης για την αποστολή δειγμάτων εικόνας στο μοντέλο του νέφους.

Ειδικότερα, λαμβάνει υπόψη τις μετρικές BvSB και Max Confidence και υπολογίζει το βέλτιστο κατώφλι (threshold) για την κάθε μια, έτσι ώστε αν οι τιμές τους τα υπερβούν, το δείγμα να αποστέλλεται στον εξυπηρετητή του νέφους. Αξίζει να σημειωθεί ότι, εφόσον η αποστολή και η επεξεργασία στο νέφος είναι κοστοβόρα, η βελτιστοποίηση αυτή έρχεται μαζί με ένα άνω φράγμα στο πλήθος δειγμάτων που μπορούν να αποσταλούν στο νέφος. Η αναπάρασταση του προβλήματος με τη μορφή μαθηματικών εξισώσεων είναι:

$$\max_{\vec{\partial}} a(\vec{\partial}) \quad \text{με τον περιορισμό} \quad n \leq p \cdot N$$

όπου $\vec{\partial}$ το διάνυσμα των μετρικών, $a(\cdot)$ η ακρίβεια του ζεύγους, n ο αριθμός των δειγμάτων που αποστέλλονται στο νέφος, p το άνω φράγμα του ποσοστού των δειγμάτων που μπορούν να αποσταλούν στο νέφος και N ο συνολικός αριθμός των δειγμάτων.

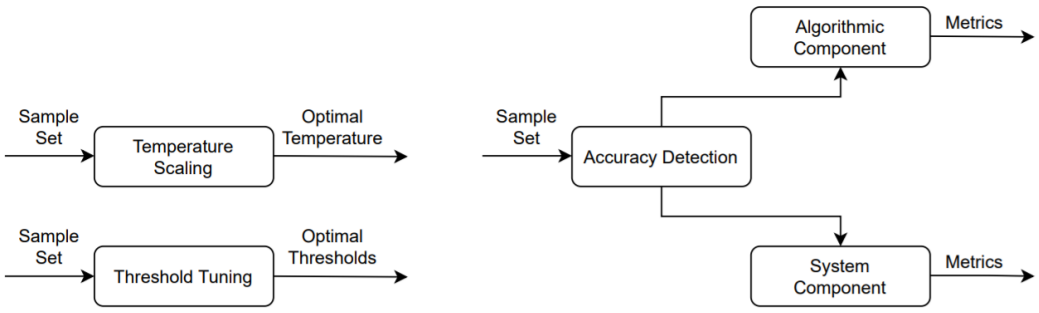
Η μέθοδος optimize_thresholds που εκτελεί τη βελτιστοποίηση κάνει χρήση του αλγόριθμου Constrained Optimization BY Linear Approximation (COBYLA) από τη βιβλιοθήκη scipy.optimize.

6.3.5 Local Accuracy Refinement

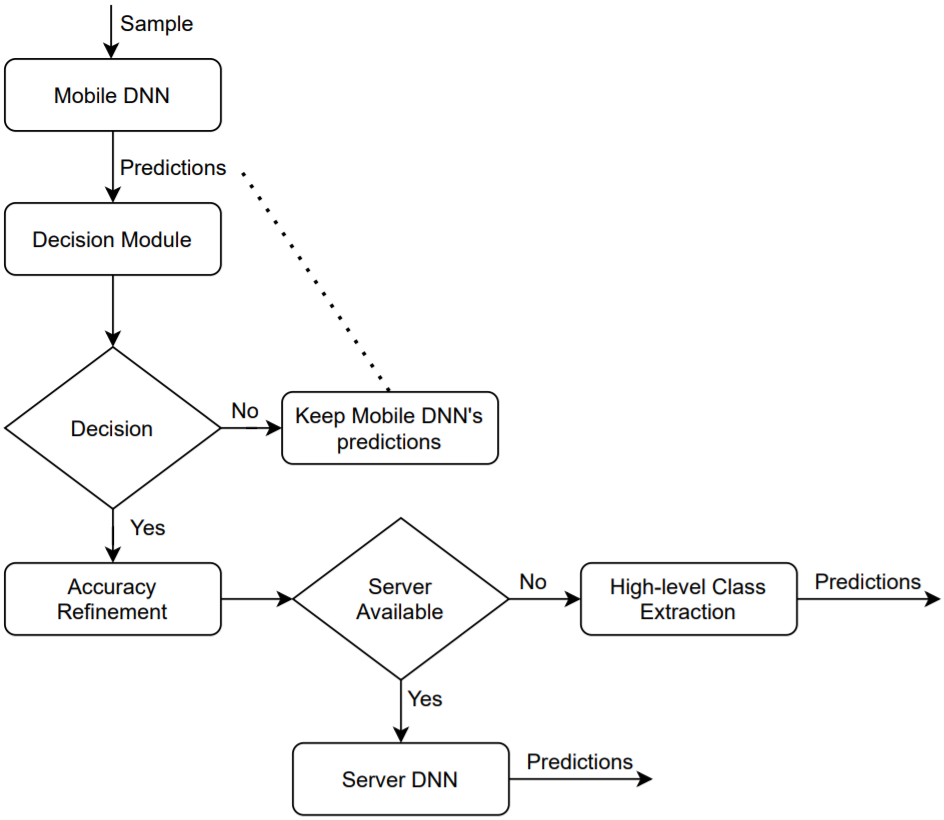
Η μονάδα Local Accuracy Refinement αποτελείται από δύο τμήματα.

Το πρώτο αποσκοπεί σε μια οργάνωση του συνόλου δεδομένων ImageNet σε κλάσεις υψηλού επιπέδου. Κάθε μια από τις χίλιες διαφορετικές κλάσεις του συνόλου δεδομένων απεικονίζεται μέσω μιας δομής λεξικού (dictionary) στην αντίστοιχη κλάση υψηλού επιπέδου. Για την επιλογή των κλάσεων υψηλού επιπέδου, οι κλάσεις που αντιπροσωπεύουν έμβιους οργανισμούς ομαδοποιήθηκαν με βάση κάποια ανώτερη βαθμίδα της συστηματικής ταξινόμησης των έμβιων όντων, ενώ οι κλάσεις που αντιπροσωπεύουν άβια αντικείμενα ομαδοποιήθηκαν με βάση τη λειτουργία τους. Οι κλάσεις υψηλού επιπέδου που τελικά επιλέχθηκαν για το ImageNet είναι οι εξής: {fish, bird, reptile, insect, mammal, marine life, primate, carnivore, dog, cat, instrument, structure, furniture, clothing, shop, kitchen equipment, technology, vehicle, nature}.

Στο δεύτερο, η συνάρτηση predict_superclass λαμβάνει τις πρώτες 3 κλάσεις με τη μεγαλύτερη εμπιστοσύνη με βάση τις προβλέψεις του μοντέλου της κινητής συσκευής και επιστρέφει την κλάση υψηλού επιπέδου στην οποία ανήκουν οι περισσότερες κλάσεις χαμηλού επιπέδου.



Εικόνα 6.1: *Διάγραμμα του συστήματος εκτός σύνδεσης (offline)*



Εικόνα 6.2: *Διάγραμμα του συστήματος στο χρόνο εκτέλεσης*

Κεφάλαιο 7

Αξιολόγηση και Αποτελέσματα

Στο σημείο αυτό θα γίνει μια ανασκόπηση των στόχων της παρούσας διπλωματικής εργασίας. Αρχικά, το πρώτο βήμα ήταν να επιλεχθεί η κατάλληλη αρχιτεκτονική για το ζεύγος μοντέλων, έτσι ώστε να γίνονται αισθητές οι μεταξύ τους διαφορές, τόσο στις απαιτήσεις τους σε υπολογιστικούς πόρους, όσο και στην ακρίβεια των προβλέψεων τους. Ακολούθως, λαμβάνοντας το κατάλληλα σχεδιασμένο ζεύγος ως εργαλείο, το επόμενο βήμα ήταν η διερεύνηση συγκεκριμένων ερευνητικών ερωτημάτων σχετικά με την αποδοτική επεξεργασία και συμπερασματολογία νέων δειγμάτων εισόδου, όπως αυτά τέθηκαν στο πρώτο κεφάλαιο της εργασίας. Το παρόν κεφάλαιο έχει ως στόχο να αξιολογήσει τόσο το σύστημα, όσο και τις επιμέρους μονάδες του και να εξαγάγει συγκεκριμένα ποσοτικά συμπεράσματα.

7.1 Πειραματική Διαδικασία

7.1.1 Κριτήρια Αξιολόγησης

Στο σημείο αυτό θα αναλύσουμε τα κριτήρια με βάση τα οποία θα αξιολογηθεί (α) το σύστημα και (β) οι επιμέρους μονάδες του.

Αρχικά, για την Στάθμιση Θερμοκρασίας (Temperature Scaling) το βασικό κριτήριο θα είναι το πόσο καλά βαθμονομημένο είναι το μοντέλο μετά την εφαρμογή της και αυτό θα κριθεί από τις τιμές των μετρικών: ECE, MCE και NLL.

Έπειτα, όσον αφορά τον Εντοπισμό Ακρίβειας (Accuracy Detection), κύριο κριτήριο για την αξιολόγησή του θα αποτελέσει το αν οι μετρικές που επιστρέφει μπορούν να μας δώσουν χρήσιμες πληροφορίες σχετικά με την πραγματική ακρίβεια του μοντέλου της κινητής συσκευής και κατά πόσο αυτή διαφέρει από την αναμενόμενη ακρίβεια του, που περιγράφεται καλά από την ακρίβεια πρώτης πρόβλεψης (top-1 accuracy).

Στη συνέχεια, κύριο κριτήριο για την αξιολόγηση της Ρύθμισης Κατωφλιών (Threshold Tuning) θα αποτελέσει το αν τα υπολογιζόμενα κατώφλια μπορούν να συμβάλουν στη μεγιστοποίηση της ακρίβειας του ζεύγους τηρώντας τους περιορισμούς που ορίστηκαν στο κεφάλαιο 6.

Όσον αφορά τους δυαδικούς ταξινομητές, αρχικά θα αξιολογηθεί η απόδοση τους σε μη γνωστά δεδομένα με μετρικές όπως η ακρίβεια στο σύνολο επικύρωσης (validation accuracy) και η απώλεια στο σύνολο επικύρωσης (validation loss). Στη συνέχεια, εφόσον διαπιστωθεί ότι μπορούν να γενικεύσουν επαρκώς, θα μελετηθεί η απόδοσή τους συγκριτικά με άλλες

μεθόδους.

Επίσης, για την αξιολόγηση της Τοπικής Διόρθωσης Ακρίβειας (Local Accuracy Refinement) σημαντικό κριτήριο θα αποτελέσει το αν η συμπερασματολογία σχετικά με την κλάση υψηλού επιπέδου των δειγμάτων μπορεί να έχει ανάλογα αποτελέσματα με άλλες μεθόδους στη συνολική ακρίβεια που πετυχαίνει.

Τέλος, για την αξιολόγηση του συνολικού συστήματος στο χρόνο εκτέλεσης κύριο κριτήριο θα αποτελέσει η αλγοριθμική συνιστώσα του Εντοπισμού Ακρίβειας και η σύγκριση των τιμών των μετρικών με αυτές των μεμονωμένων μοντέλων, ώστε να έχουμε μια πλήρη εικόνα του πόσο κοντά μπορεί να φτάσει η πραγματική ακρίβεια του ζεύγους σε αυτή του μοντέλου του νέφους εφαρμόζοντας την διόρθωση ακρίβειας σε ελεγχόμενο πλήθος δειγμάτων.

7.1.2 Πειραματικό Περιβάλλον

Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν κατά βάση το σύνολο επικύρωσης ILSVRC2012 [24]. Συγκεκριμένα, για την εκπλήρωση των αναγκών όλων των μονάδων του συστήματος, το σύνολο αυτό χωρίστηκε σε δύο τμήματα. Το πρώτο τμήμα, που αντιστοιχεί στο 40% του συνόλου χρησιμοποιήθηκε για την αξιόλογηση του Εντοπισμού Ακρίβειας και του συνολικού συστήματος, ενώ το δευτέρο τμήμα, που αντιστοιχεί στο 60% του συνόλου χρησιμοποιήθηκε για την εκπαίδευση των δυαδικών ταξινομητών.

Τέλος, όσον αφορά το υλικό που χρησιμοποιήθηκε, οι μετρήσεις έγιναν σε μηχανή με επεξεργαστή Intel Core i7 CPU και 8,00 GB RAM, ενώ η εκπαίδευση των ταξινομητών έλαβε χώρα στο περιβάλλον του Google Colaboratory με την υποστήριξη της NVIDIA Tesla K80 GPU.

7.2 Εντοπισμός Ακρίβειας

Για ένα σύνολο 20,000 δειγμάτων του ImageNet, υποσύνολο του συνόλου επικύρωσης ILSVRC2012, οι παρακάτω πίνακες παρουσιάζουν τα αποτελέσματα του εντοπισμού της ακρίβειας του συστήματος:

Top-1	Top-5	Exp. accuracy	Entropy	BvSB	Gini	B. accuracy	C. Entropy
0.6970	0.8770	0.5840	1.3000	0.5853	0.4129	0.7011	1.2829

Πίνακας 7.1: *MobileNet: Αλγοριθμικές Μετρικές*

C. Entropy	Confidence Diff.
2.6440	0.2179

Πίνακας 7.2: *MobileNet: Συστημικές Μετρικές αναφορικά με το NASNetLarge*

Temp. Scaling	ECE	MCE	NLL
N	0.0621	0.1606	1.2928
Y	0.0443	0.1429	1.2829

Πίνακας 7.3: *MobileNet: Μετρικές πριν και μετά από Στάθμιση Θερμοκρασίας*

Top-1	Top-5	Exp. accuracy	Entropy	BvSB	Gini	B. accuracy	C. Entropy
0.8040	0.9620	0.7435	0.8413	0.8061	0.2206	0.8052	0.8444

Πίνακας 7.4: *NASNetLarge: Αλγοριθμικές Μετρικές*

Αρχικά, όσον αφορά την Στάθμιση Θερμοκρασίας τα αποτελέσματα φαίνονται στον πίνακα 7.3. Με βάση την τιμή 0.0621 της μετρικής ECE πριν την εφαρμογή της Στάθμισης Θερμοκρασίας φαίνεται ότι η σταθμισμένη μέση διαφορά μεταξύ της ακρίβειας και της τιμής εμπιστοσύνης είναι πολύ μικρή και συνεπώς το MobileNet είναι επαρκώς βαθμονομημένο εκ των προτέρων. Το ίδιο συμπέρασμα μπορεί να εξαχθεί και από την τιμή 0.1606 της μετρικής MCE, η οποία δείχνει πως η διαφορά της ακρίβειας και της τιμής εμπιστοσύνης στη χειρότερη περίπτωση είναι σχετικά μικρή. Μετά την εφαρμογή της Στάθμισης Θερμοκρασίας παρατηρείται μείωση από 10% εώς 30% στις τιμές και των ECE και MCE και μείωση 1% στην τιμή της μετρικής NLL. Παρόλα αυτά, δεδομένου του ότι το συγκεκριμένο μοντέλο είναι ήδη επαρκώς βαθμονομημένο για το συγκεκριμένο σύνολο δεδομένων, η εφαρμογή της Στάθμισης Θερμοκρασίας δεν είχε ιδιαίτερο νόημα.

Έπειτα, αναφορικά με την αλγοριθμική συνιστώσα του Εντοπισμού Ακρίβειας για το μοντέλο της κινητής συσκευής τα αποτελέσματα φαίνονται στον πίνακα 7.1. Οι τιμές 0.6970 και 0.8770 των μετρικών top-1 accuracy και top-5 accuracy αντίστοιχα είναι κοντά στις αναμενόμενες τιμές για το σύνολο επικύρωσης ILSVRC2012 για το MobileNet. Η τιμή 0.5840 της μετρικής expected accuracy μας δείχνει πως πολλές φορές η τιμή της εμπιστοσύνης που επιστρέφει το μοντέλο για την πραγματική κλάση δεν είναι αρκετά υψηλή. Η τιμή 0.7011 της μετρικής balanced accuracy είναι πολύ κοντά στην τιμή της μετρικής top-1 accuracy γεγονός το οποίο δείχνει πως οι προβλέψεις του μοντέλου είναι το ίδιο καλές για την πλειονότητα των κλάσεων. Τέλος, η τιμή 0.5853 της μετρικής BvSB καταδεικνύει πως για αρκετά δείγματα το μοντέλο είναι μεταξύ δύο κλάσεων στις οποίες αποδίδει τιμές εμπιστοσύνης με μικρή διαφορά.

Όσον αφορά την αλγοριθμική συνιστώσα του Εντοπισμού Ακρίβειας για το μοντέλο του νέφους τα αποτελέσματα φαίνονται στον πίνακα 7.4. Και πάλι οι τιμές 0.8040 και 0.9620 των μετρικών top-1 accuracy και top-5 accuracy αντίστοιχα είναι κοντά στις αναμενόμενες τιμές για το σύνολο επικύρωσης ILSVRC2012 για το NASNetLarge. Η τιμή 0.7435 της μετρικής expected accuracy μας δείχνει πως η τιμή της εμπιστοσύνης που επιστρέφει το μοντέλο για την πραγματική κλάση είναι συνήθως αρκετά υψηλή. Τέλος, η τιμή 0.8061 της μετρικής BvSB δείχνει πως κατά βάση το μοντέλο επιστρέφει αρκετά μεγαλύτερη τιμή εμπιστοσύνης για την πρώτη σε σχέση με τη δεύτερη προβλεπόμενη κλάση.

Επίσης, εστιάζοντας στις τιμές των μετρικών entropy και Gini index και το πώς διαφέρουν για τα δύο μοντέλα, βλέπουμε ότι η entropy μειώνεται κατά 35%, ενώ η Gini index κατά 46% όταν η συμπερασματολογία γίνεται από το μοντέλο του νέφους. Η μελέτη των παραπάνω

μετρικών είναι ιδιαίτερα σημαντική καθώς οι τιμές τους μας παρέχουν μια αρκετά πλήρη εικόνα της μορφής της κατανομής εξόδου. Η μείωση και των δύο αυτών μετρικών όταν η συμπερασματολογία των δειγμάτων γίνεται από το μοντέλο του νέφους καταδεικνύει πως η κατανομή εξόδου του είναι αρκετά πιο «μυτερή» (spiky) από την αντίστοιχη κατανομή του μοντέλου της κινητής συσκευής. Επιπλέον, η μείωση της cross-entropy σχεδόν κατά 0.44 είναι μια ένδειξη πως η κατανομή εξόδου του μοντέλου του νέφους είναι συνήθως αρκετά πιο κοντά στην πραγματική κατανομή του δείγματος.

Τέλος, τα αποτελέσματα για την συστημική προσέγγιση του Εντοπισμού Ακρίβειας φαίνονται στον πίνακα 7.2. Η τιμή 2.6440 της μετρικής cross-entropy δείχνει πως η κατανομή εξόδου του μοντέλου της κινητής συσκευής διαφέρει αρκετά από αυτή του μοντέλου του νέφους. Ακόμα, η τιμή 0.2179 της μετρικής confidence difference καταδεικνύει πως κατά μέσο όρο η πρώτη εμπιστοσύνη του μοντέλου της κινητής συσκευής δεν είναι πολύ κοντά σε αυτή του μοντέλου του νέφους.

7.3 Εκπαίδευση Μοντέλων Απόφασης

7.3.1 Ταξινομητής βασισμένος στα χαρακτηριστικά

Ο δυαδικός ταξινομητής που βασίζεται στα χαρακτηριστικά που μπορούν να εξαχθούν από ένα δείγμα εισόδου δεν έδειξε ιδιαίτερα υποσχόμενα αποτέλεσματα.

Κατά την εκπαίδευση του πάνω σε ένα σύνολο 30,000 δειγμάτων του ImageNet, υποσύνολο του συνόλου επικύρωσης ILSVRC2012, παρατηρήθηκε το φαινόμενο της υπερπροσαρμογής (overfitting), με την ακρίβεια εκπαίδευσης (training accuracy) να βελτιώνεται συνεχώς, ενώ η ακρίβεια επικύρωσης (validation accuracy) παρέμενε κοντά στην τιμή 0.55 μετά από αρκετές εποχές (epochs). Αυτό το αποτέλεσμα δεν είναι καλό, εφόσον στην περίπτωση της δυαδικής ταξινόμησης το ποσοστό ακρίβειας 50% είναι ισοδύναμο με το οι προβλέψεις να γίνονται τυχαία. Στις εικόνες 7.1 και 7.2 παρουσιάζονται τα διαγράμματα ακρίβειας και απώλειας του ταξινομητή συναρτήσει των εποχών της εκπαίδευσης.

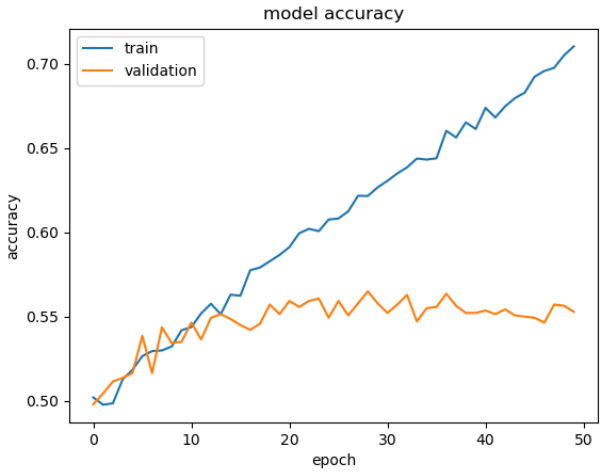
Όπως φαίνεται στις καμπύλες της εικόνας 7.1, η ακρίβεια επικύρωσης ξεκινά από την τιμή 0.5 και αυξάνεται κατά τις πρώτες 20 εποχές, μέχρι να αποκτήσει την τιμή 0.55. Από την εποχή 20 και μετά, η ακρίβεια επικύρωσης φαίνεται να συγκλίνει στην τιμή 0.55, ενώ η ακρίβεια εκπαίδευσης αυξάνεται συνεχώς με κάθε εποχή, αποκτώντας την τιμή 0.72 με το πέρας των 50 εποχών.

Επίσης, όπως φαίνεται στις καμπύλες της εικόνας 7.2, η απώλεια επικύρωσης ξεκινά από την τιμή 0.69 και μειώνεται κατά τις πρώτες 20 εποχές. Από την εποχή 20 και μετά, η απώλεια επικύρωσης φαίνεται να αυξάνεται, ενώ η απώλεια εκπαίδευσης μειώνεται συνεχώς με κάθε εποχή.

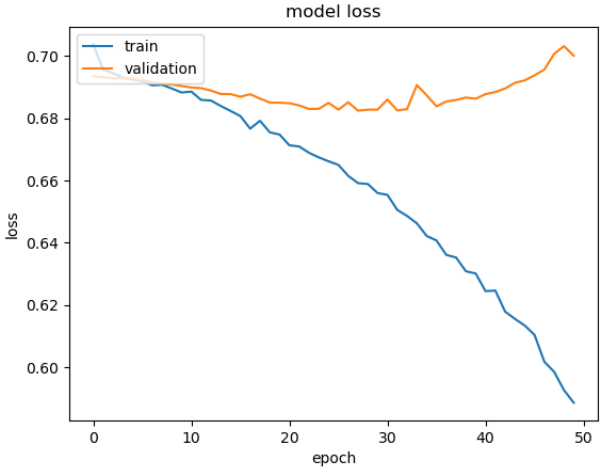
7.3.2 Ταξινομητής βασισμένος στις τιμές εμπιστοσύνης

Ο δυαδικός ταξινομητής που βασίζεται στα πρώτα 20 επίπεδα εμπιστοσύνης του τοπικού μοντέλου έδειξε ικανοποιητικά αποτελέσματα κατά την εκπαίδευση.

Η ακρίβεια επικύρωσης του ταξινομητή μετά από 1000 επαναλήψεις σταθεροποιήθηκε στην τιμή 0.78 με απώλεια 0.45. Η τιμή 0.78 της ακρίβειας αξιόλογησης δείχνει ότι ο ταξι-



Εικόνα 7.1: *Ακρίβεια του ταξινομητή χαρακτηριστικών*



Εικόνα 7.2: *Απώλεια του ταξινομητή χαρακτηριστικών*

νομητής μπορεί να ξεχωρίσει αποδοτικά τα δείγματα τα οποία χρήζουν διόρθωση ακρίβειας. Επίσης, η τιμή 0.45 της απώλειας επικύρωσης δείχνει πως ο ταξινομητής επιστρέφει επαρκώς υψηλή τιμή εμπιστοσύνης για την πραγματική κλάση των δειγμάτων.

7.4 Μετρικές ή Ταξινομητής;

Ο δυαδικός ταξινομητής που βασίζεται στα επίπεδα εμπιστοσύνης του τοπικού μοντέλου δείχνει να μπορεί να ξεχωρίσει αποδοτικά τα «προβληματικά» από τα μη «προβληματικά» δείγματα εισόδου. Το γεγονός αυτό μας οδηγεί στο συμπέρασμα ότι θα είχε ερευνητικό ενδιαφέρον η σύγκριση του ταξινομητή και των μετρικών για την ανάδειξη της μεθόδου ή του συνδυασμού μεθόδων που μεγιστοποιεί την ακρίβεια του συστήματος.

Στους παρακάτω πίνακες φαίνεται αναλυτικά η ακρίβεια που πετυχαίνει το ζεύγος σε κάθε περίπτωση.

Στον πίνακα 7.5 φαίνονται τα αποτελέσματα της κατανεμημένης εκτέλεσης για 1.000

Ποσοστό δειγμάτων στον εξυπηρετητή	BvSB	Decision Model	Entropy	Max Confidence	BvSB + Max Conf.	BvSB + Dec. Model
12%	0.736	0.727	0.728	0.728	0.738	0.736
25%	0.767	0.760	0.762	0.758	0.763	0.761

Πίνακας 7.5: *Αποτελέσματα κατανεμημένης εκτέλεσης 1,000 δειγμάτων*

Ποσοστό δειγμάτων στον εξυπηρετητή	BvSB	Decision Model	Entropy	Max Confidence	BvSB + Max Conf.	BvSB + Dec. Model
12%	0.730	0.7327	0.7326	0.7327	0.7316	0.7315
25%	0.7729	0.7738	0.7722	0.7746	0.7747	0.7747

Πίνακας 7.6: *Αποτελέσματα κατανεμημένης εκτέλεσης 20,000 δειγμάτων*

δείγματα. Στην περίπτωση που έχει τεθεί ο περιορισμός να μην αποστέλλονται πάνω από το 12% των δειγμάτων στο νέφος, φαίνεται ότι την υψηλότερη ακρίβεια πετυχαίνουν οι μετρικές BvSB και max confidence συνδυαστικά, με τιμή 0.738. Ο συνδυασμός BvSB και δυαδικού ταξινομητή (decision model), αλλά και η μετρική BvSB μεμονωμένα έχουν παρόμοιο αποτέλεσμα, ενώ οι μετρικές entropy, max confidence και ο ταξινομητής μεμονωμένα πετυχαίνουν σχεδόν κατά 0.01 χαμηλότερη ακρίβεια. Για την περίπτωση που έχει τεθεί ο περιορισμός να μην αποστέλλονται πάνω από το 25% των δειγμάτων στο νέφος, την υψηλότερη ακρίβεια πετυχαίνει η μετρική BvSB, ενώ οι υπόλοιπες πετυχαίνουν κατά μέσο όρο 0.01 χαμηλότερη ακρίβεια.

Στον πίνακα 7.6 φαίνονται τα αποτελέσματα της κατανεμημένης εκτέλεσης για 20.000 δείγματα. Στην περίπτωση που έχει τεθεί ο περιορισμός να μην αποστέλλονται πάνω από το 12% των δειγμάτων στο νέφος, φαίνεται ότι την υψηλότερη ακρίβεια πετυχαίνουν ο δυαδικός ταξινομητής και η μετρική max confidence μεμονωμένα, με τιμή 0.7327. Ακολουθούν οι συνδυαστικές μέθοδοι και οι μετρικές BvSB και entropy πετυχαίνοντας ελάχιστα χαμηλότερη ακρίβεια. Για την περίπτωση που έχει τεθεί ο περιορισμός να μην αποστέλλονται πάνω από το 25% των δειγμάτων στο νέφος, την υψηλότερη ακρίβεια πετυχαίνουν οι συνδυαστικές μέθοδοι με τιμή 0.7747. Ακολουθεί η μετρική max confidence με τιμή 0.7746, ενώ οι υπόλοιπες πετυχαίνουν σχετικά χαμηλότερη ακρίβεια.

Μέσω των παραπάνω αποτελεσμάτων, παρατηρούμε ότι οι μέθοδοι απόφασης που ξεχωρίζουν είναι οι μετρικές BvSB, max confidence, αλλά και ο δυαδικός ταξινομητής που βασίζεται στις πρώτες τιμές εμπιστοσύνης του τοπικού μοντέλου. Αυτό φαίνεται κυρίως από την πρώτη γραμμή του πίνακα 7.6, όπου ο δυαδικός ταξινομητής και η μετρική max confidence πετυχαίνουν τιμή ακρίβειας 0.7327, αλλά και από τη δεύτερη γραμμή του ίδιου πίνακα, όπου οι συνδυαστικές μέθοδοι με τη μετρική BvSB πετυχαίνουν τιμή ακρίβειας 0.7747.

7.5 Αντιστάθμιση

Για την αποστολή δειγμάτων στο νέφος, επιλέγεται ως μέθοδος απόφασης ο υπολογισμός των μετρικών BvSB και Max Confidence με τα κατώφλια να έχουν καθοριστεί ως 0.165 και 0.54 αντίστοιχα, τιμές οι οποίες υπολογίστηκαν έτσι ώστε το ποσοστό δειγμάτων που

επεξεργάζεται ο εξυπηρετητής να μην υπερβαίνει το 25%. Στον πίνακα 7.7 παρουσιάζεται η αλγοριθμική αξιολόγηση της ακρίβειας του ζεύγους για κατανεμημένη εκτέλεση 20,000 δειγμάτων.

Μοντέλο	Top-1	Top-5	Exp. accuracy	Entropy	BvSB	Gini	B. accuracy	C. Entropy
MobileNet	0.6970	0.8770	0.5840	1.3000	0.5853	0.4129	0.7011	1.2829
Κατ. Ζεύγος	0.7747	0.9334	0.7047	0.8598	0.7714	0.2429	0.7755	1.0338
NASNetLarge	0.8040	0.9620	0.7435	0.8413	0.8061	0.2206	0.8052	0.8444

Πίνακας 7.7: *Σύγκριση κατανεμημένου ζεύγους και μεμονωμένων μοντέλων.*

Από τον πίνακα 7.7 φαίνεται ότι οι τιμές 0.7747, 0.9334 των μετρικών top-1 accuracy και top-5 accuracy για το ζεύγος είναι αρκετά κοντά στις τιμές 0.8040, 0.9620 των αντίστοιχων μετρικών για το μοντέλο του νέφους. Η τιμή 0.7047 της μετρικής expected accuracy είναι κατά 0.12 μεγαλύτερη από την αντίστοιχη 0.5840 για το μοντέλο της κινητής συσκευής, ενώ μόλις κατά 0.04 χαμηλότερη από την αντίστοιχη 0.7435 για το μοντέλο του νέφους. Η τιμές 0.8598 και 0.2429 των μετρικών entropy και Gini index παρουσιάζουν 34% και 41% μείωση αντίστοιχα όταν η συμπερασματολογία δεν γίνεται από την κινητή συσκευή, αλλά από το ζεύγος. Η τιμή 0.7755 της μετρικής balanced accuracy δείχνει να συμβαδίζει με την μετρική top-1 accuracy για το ζεύγος, το οποίο είναι αναμενόμενο, ενώ η τιμή 1.0338 της μετρικής cross-entropy παρουσιάζει 19% μείωση όταν η συμπερασματολογία γίνεται από το ζεύγος.

Επιπλέον, στην περίπτωση που ο εξυπηρετητής δεν είναι διαθέσιμος, η διόρθωση της ακρίβειας μέσω εξαγωγής της κλάσης υψηλού επιπέδου παρουσιάζει εξίσου καλά αποτελέσματα. Με την εξαγωγή της κλάσης υψηλού επιπέδου να εφαρμόζεται μόνο στο 25% των δειγμάτων, με βάση τα κατώφλια των μετρικών, η ακρίβεια πρώτης πρόβλεψης του συστήματος αποκτά την τιμή 0.7744. Αυτή της η επιτυχία βασίζεται κυρίως στον μεγάλο βαθμό ομοιότητας των ειδικών κλάσεων που ανήκουν σε μια γενική κλάση. Το μοντέλο παρόλο που ενδέχεται να επιστρέφει τη λανθασμένη ειδική κλάση, η πλειονότητα των πρώτων του προβλέψεων είναι πολύ πιθανό να είναι υποκλάσεις της σωστής γενικής κλάσης.

Κεφάλαιο 8

Επίλογος

Στο παρόν κεφάλαιο γίνεται μια ανασκόπηση των μεθόδων που προτάθηκαν και δοκιμάστηκαν στα προηγούμενα κεφάλαια με σκοπό την ποιοτική αξιολόγηση των αποτελεσμάτων της κάθε μεθόδου και την πρόταση νέων προσεγγίσεων που αξίζει να μελετηθούν στο μέλλον.

8.1 Συμπεράσματα

Αρχικά, οι τιμές των μετρικών ECE, MCE και NLL όπως παρουσιάζονται στην πρώτη γραμμή του πίνακα 7.3, δείχνουν πως το MobileNet είναι επαρκώς βαθμονομημένο για το σύνολο δεδομένων ImageNet και συνεπώς η εφαρμογή της Στάθμισης Θερμοκρασίας δεν είναι απαραίτητη. Επιπλέον, όπως φαίνεται στην ενότητα 7.2, οι μετρικές Εντοπισμού Ακρίβειας παρέχουν μια καλή και λεπτομερή αξιολόγηση της πραγματικής ακρίβειας του τοπικού μοντέλου. Εκτός από την ακρίβεια πρώτης πρόβλεψης, δίνεται μια αρκετά πλήρης εικόνα της κατανομής εξόδου του μοντέλου και κατ' επέκταση της αξιοπιστίας των προβλέψεών του.

Στη συνέχεια, όσον αφορά τον δυαδικό ταξινομητή που βασίζεται στα χαρακτηριστικά της εικόνας μπορούμε να πούμε ότι η ανάπτυξή του ήταν μη επιτυχής. Η αδυναμία ποιοτικής εκπαίδευσης και η υπερπροσαρμογή δείχνουν ότι τα χαρακτηριστικά που μπορούν να εξαχθούν από μια εικόνα δεν επηρεάζουν απαραίτητα την επιτυχία της ταξινόμησης της εικόνας από το τοπικό μοντέλο. Πιο σημαντικοί παράγοντες που πιθανώς επηρεάζουν άμεσα την επιτυχία της διαδικασίας και αξίζει να διερευνηθούν περαιτέρω είναι (α) ο βαθμός ομοιότητας των κλάσεων χαμηλού επιπέδου που είναι υποσύνολα μιας κλάσης υψηλού επιπέδου (π.χ. δύο εμφανισιακά όμοιες ράτσες σκύλου) και (β) οι επικαλυπτόμενες κλάσεις, δηλαδή η ύπαρξη δύο ή περισσοτέρων κλάσεων που η τομή τους δεν είναι το κενό σύνολο.

Από την άλλη, η ανάπτυξη του δυαδικού ταξινομητή που βασίζεται στα πρώτα επίπεδα εμπιστοσύνης του τοπικού μοντέλου ήταν σε μεγάλο βαθμό επιτυχής. Η ακρίβεια του ήταν αρκετά υψηλή ώστε να μπορεί να διαχωρίζει αποδοτικά τα προβληματικά από τα μη προβληματικά δείγματα εισόδου. Παρόλα αυτά, εφόσον τα αποτελέσματα της ενότητας 7.4 δείχνουν ότι η χρήση των μετρικών BvSB και Max Confidence συνδυαστικά μπορεί να πετύχει την ίδια ακρίβεια, συμπεραίνουμε ότι η εκπαίδευση και η συμπερασματολογία ενός τέτοιου μοντέλου θα ήταν καλύτερο να αποφευχθούν.

Ακολούθως, όταν πλέον η μέθοδος για την απόφαση της κατανομής των δειγμάτων μεταξύ

των δύο μοντέλων έχει καθοριστεί να είναι ο συνδυασμός των μετρικών που αναφέρθηκαν παραπάνω, η διόρθωση της ακρίβειας μέσω της αποστολής των ανεπεξέργαστων δειγμάτων στον εξυπηρετητή καταφέρνει να αυξήσει την ακρίβεια πρώτης πρόβλεψης σχεδόν κατά 10%, αποστέλλοντας μόλις το 25% των δειγμάτων στο νέφος, όπως φαίνεται στην ενότητα 7.5.

Τέλος, ακόμα και στην περίπτωση που ο εξυπηρετητής δεν είναι διαθέσιμος να συμμετάσχει στην συμπερασματολογία, η εξαγωγή της κλάσης υψηλού επιπέδου ως τρόπος διόρθωσης της ακρίβειας από τη μεριά του τοπικού μοντέλου είναι εξίσου επιτυχής, γεγονός το οποίο μας επιστρέπει να την εμπιστευτούμε έως ότου η επικοινωνία με τον εξυπηρετητή αποκατασταθεί. Η τιμή 0.7744 της ακρίβειας δείχνει πως προσωρινά το τοπικό μοντέλο μπορεί να είναι το ίδιο ακριβές με το ζεύγος, με μια απώλεια που εκδηλώνεται στον βαθμό ειδικότητας των προβλέψεών του.

8.2 Μελλοντικές Επεκτάσεις

Το σύστημα που αναπτύχθηκε στα πλαίσια της διπλωματικής εργασίας θα μπορούσε να βελτιωθεί και να επεκταθεί περαιτέρω, τουλάχιστον ως προς επτά κατευθύνσεις. Συγκεκριμένα, αναφέρονται τα ακόλουθα:

- Δείγματα εκτός κατανομής: Μια ενδιαφέρουσα προσέγγιση του προβλήματος της παρούσας διπλωματικής εργασίας θα ήταν η προσομοίωση της πτώσης της ακρίβειας όταν το μοντέλο της κινητής συσκευής δεν έχει εκπαιδευτεί να αναγνωρίζει όλες τις πιθανές κλάσεις. Με αυτόν τον τρόπο, όταν καλείται να εξετάσει πραγματικά δεδομένα, τα οποία πιθανότατα δεν περιλαμβάνονται στην κατανομή εκπαίδευσης, η ταξινόμηση θα είναι ανεπιτυχής και η πραγματική ακρίβεια του θα διαφέρει αρκετά από την αναμενόμενη.

- Πίνακας σύγχυσης: Μια ενδεχομένως χρήσιμη μέθοδος για την διαδικασία απόφασης είναι η μελέτη του πίνακα σύγχυσης του τοπικού μοντέλου και η επιλογή των νέων δειγμάτων που χρήζουν διόρθωση ακρίβειας με βάση την προβλεπόμενη κλάση τους και την ακρίβεια του τοπικού μοντέλου συγκεκριμένα για αυτή την κλάση, όπως υπολογίζεται από τον πίνακα.

- Διεύρυνση συνόλου εκπαίδευσης: Όσον αφορά την εκπαίδευση του ταξινομητή που βασίζεται στην εξαγωγή χαρακτηριστικών, ενδιαφέρουσα επέκταση θα ήταν η προσπάθεια να εκπαιδευτεί πάνω σε ένα πολύ μεγαλύτερο σύνολο εκπαίδευσης, όπως το σύνολο εκπαίδευσης του ILSVRC2012 [24]. Με αυτόν τον τρόπο, υπάρχει πιθανότητα να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής, επιτρέποντας στον ταξινομητή να γενικεύσει επιτυχώς.

- Εκτέλεση με υβριδική αρχιτεκτονική: Ένας επιπλέον τρόπος να γίνει η διόρθωση της ακρίβειας και αξίζει να διερευνηθεί είναι η ταυτόχρονη μερική εκτέλεση και από τα δύο μοντέλα. Για να πετύχει αυτή η προσέγγιση απαιτείται ο σχεδιασμός μιας υβριδικής αρχιτεκτονικής, τέτοιας ώστε να επιτρέπει στο τοπικό μοντέλο να αποστέλλει ενδιάμεσους χάρτες χαρακτηριστικών (feature maps) στον εξυπηρετητή, ώστε το ισχυρό μοντέλο να ολοκληρώνει τη διαδικασία της συμπερασματολογίας. Βέβαια, για να

επιτύχει αυτή η ιδέα, βασική προϋπόθεση είναι η ύπαρξη επιπέδων προσαρμογής (adaption layers), εκπαιδευμένων κατάλληλα έτσι ώστε να μετασχηματίζουν την έξοδο ενός ενδιάμεσου επιπέδου του τοπικού μοντέλου σε είσοδο ενός ενδιάμεσου επιπέδου του μοντέλου του νέφους. Επιπλέον, αυτή η προσέγγιση λύνει το ζήτημα της ιδιωτικότητας των δεδομένων, αφού δεν αποστέλλονται στο νέφος τα ακατέργαστα δεδομένα του χρήστη.

- Προσθήκη επιπλέον περιορισμών: Μια επιπρόσθετη μελλοντική επέκταση αποτελεί η προσθήκη νέων μεταβλητών-περιορισμών στο σύστημα για τη μοντελοποίηση της επικοινωνίας της συσκευής με τον εξυπηρετητή. Τέτοιες παράμετροι μπορεί να είναι το δίκτυο, η καθυστέρηση και το φορτίο του εξυπηρετητή.

- Πραγματική Ανάπτυξη: Πέρα από τη μοντελοποίηση του συστήματος, ενδιαφέρουσα προσέγγιση θα ήταν η ανάπτυξη του σε πραγματικές κινητές συσκευές και εξυπηρετητές βασιζόμενη σε πραγματικά σενάρια εφαρμογών. Με αυτόν τον τρόπο, μπορεί να μελετηθεί το αν και κατά πόσο η αντιστάθμιση της ακρίβειας που προσφέρει το σύστημα ανταποκρίνεται στις ανάγκες του χρήστη, αλλά και να εντοπιστούν προβλήματα που ενδέχεται να προκύψουν μέσω της συμπερασματολογίας των νέων δειγμάτων μηελεγχόμενης ποιότητας που παράγονται από τις κινητές συσκευές.

- Παραλλαγές του χώρου της εφαρμογής: Μια τελευταία και εξίσου ενδιαφέρουσα προσέγγιση θα ήταν η ανάπτυξη του συστήματος με τη χρήση διαφορετικών μοντέλων για το ζεύγος συνοδευόμενη με τη σύγκριση των αποτελεσμάτων τους, αλλά και η ανάπτυξη του συστήματος για διαφορετικές διεργασίες Μηχανικής Μάθησης.

Παραρτήματα

Παράρτημα Α΄

Πηγαίος Κώδικας

Α'.1: *Μέθοδοι υπολογισμού του ECE*

```
def weighted_binwise_confindence_accuracy_diff(preds, confs, low,
high):
    num_of_samples = len(confs)
    indexes = []

    for i in range(num_of_samples):
        if low <= confs[i] <= high:
            indexes.append(i)

    acc_sum, conf_sum = 0, 0

    for i in range(len(indexes)):
        acc_sum += preds[indexes[i]]
        conf_sum += confs[indexes[i]]

    acc = acc_sum/len(indexes) if len(indexes) != 0 else 0
    conf = conf_sum/len(indexes) if len(indexes) != 0 else 0

    return len(indexes) * abs(acc - conf)

def ece(preds, confs, num_of_bins = 10):
    step = 1/num_of_bins
    num_of_samples = len(preds)
    ece_sum = 0

    for i in range(num_of_bins):
        low = i*step
        high = (i+1)*step

        weighted_binwise_acc_conf_diff =
        weighted_binwise_confindence_accuracy_diff(preds, confs,
        low, high)
        ece_sum += weighted_binwise_acc_conf_diff

    return ece_sum/num_of_samples
```

Α΄.2: *Μέθοδοι fit και predict*

```
def fit(self, logits, true):

    opt = minimize(self._loss_fun, x0=1,
    args=(logits, true), options={'maxiter': self.maxiter},
    method=self.solver)
    self.temp = opt.x[0]

    return opt

def predict(self, logits, temp=None):

    if not temp:
        return softmax(logits / self.temp)
    else:
        return softmax(logits / temp)
```

Α΄.3: *Μέθοδος αλγοριθμικής αξιολόγησης ακρίβειας*

```
def evaluate_accuracy_algo(model_name, predictions, labels,
verbose=False):

    df = pd.DataFrame(columns=["Top1", "Top5", "Entropy", "BVSB",
    "Max Conf.", "Exp.accuracy", "Gini", "B.accuracy",
    "C.Entropy"])

    top1_acc = top1_accuracy(predictions, labels)
    top5_acc = topk_accuracy(predictions, labels, 5)
    exp_acc = expected_accuracy(predictions, labels)
    exp_entr = expected_entropy(predictions)
    bvsb_val = expected_bvsb(predictions)
    max_conf = expected_max_confidence(predictions)
    gini = expected_gini_index(predictions)
    cross_entr = expected_cross_entropy(predictions, labels)
    b_acc = balanced_accuracy(predictions, labels)

    df.loc[0] = [top1_acc, top5_acc, exp_entr,
    bvsb_val, max_conf, exp_acc, gini, b_acc, cross_entr]

    if verbose:
        print('{}: '.format(model_name))
        pd.set_option('display.max_columns', None)
        pd.set_option('display.expand_frame_repr', False)
        print(df)

    return df
```

Α'.4: *Μέθοδος βελτιστοποίησης κατωφλιών*

```
def optimize_thresholds(max_samples_to_server, verbose=False):

    def objective(thres, predictions_server, predictions_mobile,
    labels):
        error = objective_calc(thres, predictions_server,
        predictions_mobile, labels)[0]

        return error

    def constraint(thres, predictions_server, predictions_mobile,
    labels):
        count_server = objective_calc(thres,
        predictions_server, predictions_mobile, labels)[1]

        return max_samples_to_server-count_server

    con = {'type': 'ineq', 'fun': constraint, 'args':
    (predictions_server, predictions_mobile, labels)}

    sol = minimize(objective, x0=[0.1, 0.3],
    bounds=((0.0, 1.0), (0.0, 1.0)),
                   args=(predictions_server, predictions_mobile,
                   labels),
                   method='COBYLA', options={'rhobeg': 0.04},
                   constraints=[con])

    if verbose:
        print(sol)

    return sol
```

Α΄.5: *Μέθοδος πρόβλεψης κλάσεων υψηλού επιπέδου*

```
def predict_superclass(pred_distribution):
    top_args = pred_distribution.argsort()[-1:][::-1]
    super_dict = {}

    for class_index in top_args:
        superclass = super_map[class_index]
        if superclass in super_dict:
            super_dict[superclass] += 1
        else:
            super_dict[superclass] = 1

    top_super = {k: v for k, v in sorted(super_dict.items(),
    key=lambda item: item[1], reverse=True)}

    pred_superclass = list(top_super.keys())[0]

    return pred_superclass
```

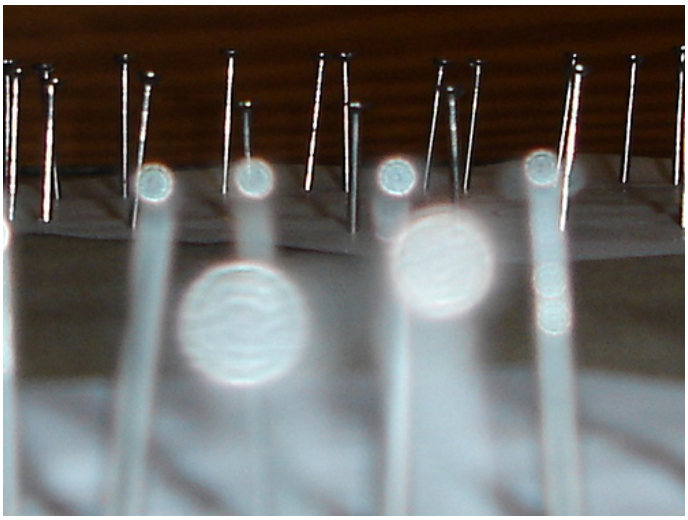
Α΄.6: *Feature-based Classifier*

```
from keras.applications.mobilenet import MobileNet
from keras.models import Sequential
from keras.layers import GlobalAveragePooling2D, Dense

basemodel = MobileNet(input_shape=(224, 224, 3), include_top=False,
weights='imagenet')


for layer in basemodel.layers:
  layer.trainable = False

model = Sequential(
    [
        basemodel,
        GlobalAveragePooling2D(),
        Dense(128),
        Dense(1, activation="sigmoid"),
    ]
)

model.compile(loss="binary_crossentropy", optimizer="adam",
metrics=["acc"])
```

Παράρτημα **Β΄**

Παραδείγματα

Β΄.1 Αποστολή στο Νέφος



Εικόνα Β΄.1: *Δείγμα εισόδου: Καρφί*

Μετρικές: {BVSB: 0.1121, Max Confidence: 0.2666} ⟶ Διόρθωση ακρίβειας

Πρόβλεψη τοπικού μοντέλου: Στοίβα πιάτων ✗

Πρόβλεψη μοντέλου του νέφους: Καρφί ✓

Εικόνα Β΄.2: *Δείγμα εισόδου: Βλαττοειδές*

Μετρικές: {BVSB: 0.1476, Max Confidence: 0.3383} ⟶ Διόρθωση ακρίβειας

Πρόβλεψη τοπικού μοντέλου: Ποντικοπαγίδα ✗

Πρόβλεψη μοντέλου του νέφους: Βλαττοειδές ✓

Εικόνα Β΄.3: *Δείγμα εισόδου: Αγριόχοιρος*

Μετρικές: {BVSB: 0.0170, Max Confidence: 0.2181} ⟶ Διόρθωση ακρίβειας

Πρόβλεψη τοπικού μοντέλου: Έρντεϊλ Τεριέ ✗

Πρόβλεψη μοντέλου του νέφους: Αγριόχοιρος ✓

Β΄.2 Εξαγωγή Κλάσης Υψηλού Επιπέδου

Εικόνα Β΄.4: *Δείγμα εισόδου: Γερμανικός ποιμενικός*

Μετρικές: {BVSB: 0.0080, Max Confidence: 0.3756} ⟶ Διόρθωση ακρίβειας

Πρόβλεψη τοπικού μοντέλου: Αυστραλιανός Κέλπι ✗

Πρόβλεψη υψηλού επιπέδου: Σκύλος ✓

Εικόνα Β΄.5: *Δείγμα εισόδου: Γάτα Τάμπι*

Μετρικές: {BVSB: 0.0828, Max Confidence: 0.4523} ⟶ Διόρθωση ακρίβειας

Πρόβλεψη τοπικού μοντέλου: Γάτα Βεγγάλης ✗

Πρόβλεψη υψηλού επιπέδου: Γάτα ✓

Βιβλιογραφία

[1] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[2] Josh Patterson και Adam Gibson. *Deep Learning: A Practitioner's Approach*. O'Reilly, 2017.

[3] *CS231: Convolutional Neural Networks for Visual Recongition*. https://cs231n.github.io/. Ημερομηνία πρόσβασης: 26-08-2021.

[4] Keiron O'Shea και Ryan Nash. *An Introduction to Convolutional Neural Networks*. *CoRR*, abs/1511.08458, 2015.

[5] Andriy Burkov. *The Hundred-Page Machine Learning Book*. 2019.

[6] Yann Lecun, Leon Bottou, Y. Bengio και Patrick Haffner. *Gradient-Based Learning Applied to Document Recognition*. *Proceedings of the IEEE*, 86:2278 – 2324, 1998.

[7] Alex Krizhevsky, Ilya Sutskever και Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. *Advances in Neural Information Processing Systems 25*, σελίδες 1097–1105. Curran Associates, Inc., 2012.

[8] Christian Szegedy, *et al.* *Going deeper with convolutions*. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, σελίδες 1–9. IEEE Computer Society, 2015.

[9] Karen Simonyan και Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep Residual Learning for Image Recognition*. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, σελίδες 770–778. IEEE Computer Society, 2016.

[11] Andrew G. Howard, *et al.* *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. *CoRR*, abs/1704.04861, 2017.

[12] Barret Zoph, Vijay Vasudevan, Jonathon Shlens και Quoc V. Le. *Learning Transferable Architectures for Scalable Image Recognition*. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22,*

2018, σελίδες 8697–8710. Computer Vision Foundation / IEEE Computer Society, 2018.

[13] Chuan Guo, Geoff Pleiss, Yu Sun και Kilian Q. Weinberger. *On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, τόμος 70 στο *Proceedings of Machine Learning Research*, σελίδες 1321–1330. PMLR, 2017.

[14] Renjie Gu, Shuo Yang και Fan Wu. *Distributed Machine Learning on Mobile Devices: A Survey*. *CoRR*, abs/1909.08329, 2019.

[15] Ji Wang, Bokai Cao, Philip S. Yu, Lichao Sun, Weidong Bao και Xiaomin Zhu. *Deep Learning towards Mobile Applications*. *38th IEEE International Conference on Distributed Computing Systems, ICDCS 2018, Vienna, Austria, July 2-6, 2018*, σελίδες 1385–1393. IEEE Computer Society, 2018.

[16] Sawsan Abdulrahman, *et al.* *A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond*. *IEEE Internet of Things Journal*, PP, 2020.

[17] Νικόλαος Α. Αγγελιδάκης. *Εισαγωγή στον Προγραμματισμό με Python*. Ηράκλειο, 1η έκδοση, 2015.

[18] *The Making of Python, A Conversation with Guido van Rossum, Part I*. https://www.artima.com/articles/the-making-of-python. Ημερομηνία πρόσβασης: 28-08-2021.

[19] *Colaboratory - Frequently Asked Questions*. https://research.google.com/colaboratory/faq.html. Ημερομηνία πρόσβασης: 29-08-2021.

[20] Martin Abadi, *et al.* *TensorFlow: A system for large-scale machine learning*. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, σελίδες 265–283, 2016.

[21] *Keras Documentation*. https://keras.io. Ημερομηνία πρόσβασης: 29-08-2021.

[22] François Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, σελίδες 1800–1807. IEEE Computer Society, 2017.

[23] *ImageNet*. https://www.image-net.org/. Ημερομηνία πρόσβασης: 29-08-2021.

[24] *ILSVRC2012*. https://dbcollection.readthedocs.io/en/latest/datasets/imagenet.html. Ημερομηνία πρόσβασης: 10-09-2021.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

κ.	κύριος, κυρία
κ.α.	και άλλα
π.χ.	παραδείγματος χάρη
BVSB	Best Versus Second Best
CIFAR	Canadian Institute For Advanced Research
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CWI	Centrum voor Wiskunde en Informatica
DNN	Deep Neural Network
ECE	Expected Calibration Error
GPU	Graphics Processing Unit
IDE	Integrated Development Environment
ILSVRC	Imagenet Large Scale Visual Recognition Challenge
MCE	Maximum Calibration Error
N	no
NAS	Neural Architecture Search
NLL	Negative Log Likelihood
PEP	Python Enhancement Proposal
PSF	Python Software Foundation
RAM	Random-access Memory
TPU	Tensor Processing Unit
VCS	Version Control System
Y	yes

Απόδοση Ξενόγλωσσων Όρων

Απόδοση	Ξενόγλωσσος όρος
ακρίβεια	accuracy
Αναδρομικό Νευρωνικό Δίκτυο	Recurrent Neural Network
ανεπεξέργαστος	raw
Ανορθωμένη Γραμμική Μονάδα	Rectified Linear Unit
αντικειμενοστρεφής	object oriented
αξιοπιστία	reliability
Αρνητική Λογαριθμική Πιθανοφάνεια	Negative Log Likelihood
αρχιτεκτονική	architecture
Βαθιά Μάθηση	Deep Learning
βαθμονόμηση	calibration
βάθος	depth
βάρος	weight
βελτιστοποίηση	optimization
βηματισμός	stride
βιβλιοθήκη	library
γενίκευση	generalization
γλώσσα σεναρίων	scripting language
δείγμα	sample
Δίκτυο των Πραγμάτων	Internet of Things
εκτός κατανομής	out of distribution
εμπιστοσύνη	confidence
εξαγωγή χαρακτηριστικών	feature extraction
εξυπηρετητής	server
επανάληψη	iteration
Επεξεργασία Φυσικής Γλώσσας	Natural Language Processing
Επιβλεπόμενη Μάθηση	Supervised Learning
επίπεδο προσαρμογής	adaptation layer
εποχή	epoch
εργαλείο	tool
ετικέτα	label
καθυστέρηση	latency
κανονικοποίηση	regularization
κατανεμημένο σύστημα	distributed system
κατώφλι	threshold

λεξικό	dictionary
λογιστική παλινδρόμηση	logistic regression
μέγεθος	size
μετρική	metric
Μηχανική Μάθηση	Machine Learning
μονάδα	module
Μονάδα Επεξεργασίας Γραφικών	Graphics Processing Unit
Μονάδα Επεξεργασίας Τανυστών	Tensor Processing Unit
μοντέλο	model
νευρώνας	neuron
όγκος	volume
Όραση Υπολογιστών	Computer Vision
παραγέμισμα	padding
παραλληλισμός δεδομένων	data parallelism
παραλληλισμός μοντέλου	model parallelism
πίνακας σύγχυσης	confusion matrix
πλαίσιο δεδομένων	dataframe
μέγεθος	size
πλήρως συνδεδεμένο	fully connected
πρόβλεψη	prediction
προεκπαιδευμένος	pretrained
προεπεξεργασία	preprocessing
πυρήνας	kernel
σιγμοειδής	sigmoid
Στάθμιση Θερμοκρασίας	Temperature Scaling
συγκέντρωση	pooling
συμπερασματολογία	inference
συνάρτηση ενεργοποίησης	activation function
Συνελικτικά Νευρωνικά Δίκτυα	Convolutional Neural Networks
συνέλιξη	convolution
σύνολο επικύρωσης	validation set
σύνολο δεδομένων	dataset
σύνολο εκπαίδευσης	training set
ταξινόμηση	classification
ταξινομητής	classifier
Τεχνητή Νοημοσύνη	Artificial Intelligence
υβριδικός	hybrid
υλικό	hardware
υπερπροσαρμογή	overfitting
φίλτρο	filter
χάρτης ενεργοποίησης	activation map
χρονοσειρά	time series