Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
Τομεας Μαθηματικων

Κατανεμημένο Σύστημα Συμπερασματολογίας

με Ευέλικτες Διασυνδέσεις Ζευγών Νευρωνικών Δικτύων

Διπλωματικη Εργασια

ΑΙΚΑΤΕΡΙΝΗ Κ. ΑΘΑΝΑΣΙΑ

Επιβλέπων: Ιάκωβος Βενιέρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Κατανεμημένο Σύστημα Συμπερασματολογίας

με Ευέλικτες Διασυνδέσεις Ζευγών Νευρωνικών Δικτύων

Διπλωματικη Εργασια

ΑΙΚΑΤΕΡΙΝΗ Κ. ΑΘΑΝΑΣΙΑ

Επιβλέπων: Ιάκωβος Βενιέρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23η Φεβρουαρίου 2023.

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

Αθήνα, Φεβρουάριος 2023


Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

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 Διπλωματικής 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 υπόλογη έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Διπλωματική μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Διπλωματική 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)

.....
Αικατερίνη Αθανασιά

23 Φεβρουαρίου 2023

Περίληψη

Τα τελευταία χρόνια, η χρήση εφαρμογών Τεχνητής Νοημοσύνης έχει αυξηθεί κατακόρυφα, τόσο λόγω της προόδου που έχει σημειωθεί στη Βαθιά Μάθηση, όσο και στην τεχνολογική ανάπτυξη στον τομέα του Κινητού Υπολογισμού και τη διάδοση του Διαδικτύου των Πραγμάτων (IoT). Η βελτίωση της ακρίβειας συμπερασματολογίας οφείλεται στην αύξηση της πολυπλοκότητας των μοντέλων, που συνεπάγεται σημαντική αύξηση στον χρόνο και στους απαιτούμενους πόρους για την εξαγωγή προβλέψεων.

Οι κινητές συσκευές όμως, λόγω των περιορισμένων πόρων τους, δεν επωφελούνται στο μέγιστο από τα αποτελέσματα αυτά. Ο κλάδος της Κατανεμημένης Μηχανικής Μάθησης επιδιώκει να δώσει λύση στο πρόβλημα αυτό, μεταξύ άλλων, με τη χρήση Ζευγών Νευρωνικών Δικτύων, αποτελούμενων από δύο μοντέλα Βαθιάς Μάθησης με διαφορετικά χαρακτηριστικά. Το πρώτο μοντέλο είναι ελαφρύ και με μειωμένη συγκριτικά ακρίβεια, ώστε να μπορεί να εκτελεστεί τοπικά στη συσκευή. Το δεύτερο μοντέλο προσφέρει υψηλή ακρίβεια, η οποία όμως συνοδεύεται από την ανάγκη για πρόσβαση σε αυξημένους υπολογιστικούς πόρους, γι' αυτό και το μοντέλο εκτελείται σε ισχυρό εξυπηρετητή ο οποίος βρίσκεται στα άκρα του δικτύου ή στο νέφος. Μέσω αυτού του Κατανεμημένου Συστήματος Συμπερασματολογίας, επιτρέπεται η παραμετρική βελτιστοποίηση του χρόνου και της ακρίβειας συμπερασματολογίας με συμβιβασμούς στις απαιτήσεις του χρήστη. Ο βέλτιστος συμβιβασμός επιτυγχάνεται με τη δυναμική προσαρμογή του συστήματος τόσο σε αυτές τις απαιτήσεις όσο και στις αλλαγές στο περιβάλλον του προβλήματος, όπως το πλήθος και το είδος των συσκευών.

Στόχος, λοιπόν, της παρούσας διπλωματικής εργασίας, είναι η σχεδίαση και ανάπτυξη ενός ολοκληρωμένου, ευέλικτου, σε σχέση με την στρατηγική εκτέλεσης και του τρόπου διασύνδεσης των μοντέλων, Κατανεμημένου Συστήματος Συμπερασματολογίας, με σκοπό την αποδοτική, ως προς τις ανάγκες του χρήστη και την κατάσταση του συστήματος, εκτέλεση εφαρμογών Βαθιάς Μάθησης σε κινητές συσκευές.

Εκ του αποτελέσματος, παρατηρήθηκε πως πράγματι ένα σύστημα με τις παραπάνω ιδιότητες μπορεί από τη μία να αυξήσει την ακρίβεια σε σχέση με το αν υπήρχε μόνο το μοντέλο της συσκευής, και από την άλλη να επιταχύνει την συμπερασματολογία σε σχέση με την αποστολή όλων των εικόνων σε εξυπηρετητή.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Βαθιά Μάθηση, Κατανεμημένο Σύστημα Συμπερασματολογίας, Ζεύγος Νευρωνικών Δικτύων, Ταξινόμηση Εικόνας, Edge Intelligence

Abstract

In recent years, the use of AI applications has increased dramatically, mainly driven by advances in Deep Learning, technological developments in Mobile Computing and the proliferation of the Internet of Things (IoT). The improvement in inference accuracy is attributed to the increase in model complexity, which in turn results in significant growth in time and resources required to make predictions.

Mobile devices, however, given their limited resources, are not taking full advantage of these achievements. The field of Distributed Machine Learning seeks to provide a solution to this problem, by using, among others, Neural Network Pairs, consisting of two Deep Learning models with different characteristics. The first model is lightweight and has comparatively reduced accuracy so that it can be executed locally on the device, while the second model offers high accuracy, but requires increased computational resources, hence it is executed on a powerful server located at the network edges or in the cloud. Through this Distributed Inference System, parametric optimization of inference time and accuracy is enabled trade-offs to user demands. The optimal trade-off is achieved by dynamically adapting the system to both these demands while considering changes in the problem environment, such as the number and type of devices in the system.

The aim of this thesis is to design and develop an integrated and flexible Distributed Inference System, with respect to its execution strategy and the way the models are interconnected, for the efficient, in terms of user demands and system state, execution of Deep Learning applications on mobile devices.

Indeed, it was observed that a system with the above properties can, on the one hand, increase the accuracy compared to only executing inference on the model on the device, and on the other hand, it speeds up the inference process in relation to sending all the images to the server for inference.

Keywords

Machine Learning, Deep Learning, Distributed Inference System, Neural Network Pair, Image Classification, Edge Intelligence

στους ανθρώπους μου

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ιάκωβο Βενιέρη για την επίβλεψη αυτής της διπλωματικής εργασίας και την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Ευφυών Επικοινωνιών και Δικτύων Ευρείας Ζώνης. Ακόμη, θα ήθελα να ευχαριστήσω τους καθηγητές κ. Δήμητρα-Θεοδώρα Κακλαμάνη και κ. Αντώνιο Συμβώνη για τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή.

Επιπλέον, ευχαριστώ θερμά για τη συνεργασία μας τον Δρ. Στυλιανό Βενιέρη, ερευνητή στο Κέντρο Τεχνητής Νοημοσύνης της Samsung στο Cambridge. Η πολύτιμη γνώση και καθοδήγηση που μου προσέφερε κατά τη διάρκεια εκπόνησης της εργασίας αυτής, καθιστούν τη συμβολή του καθοριστική για την επίτευξη του αποτελέσματος. Θα ήθελα επίσης να ευχαριστήσω ιδιαίτερα τον κ. Σωκράτη Νικολαΐδη και τον κ. Ιωάννη Πανόπουλο, υποψήφιους διδάκτορες ΣΗΜΜΥ ΕΜΠ, για την καθοδήγηση, την υπομονή και την υποστήριξη τους.

Τέλος, καθώς με την ολοκλήρωση αυτής της διπλωματικής εργασίας κλείνει ο κύκλος των προπτυχιακών σπουδών μου, θα ήθελα να ευχαριστήσω όσους ήταν δίπλα μου τα χρόνια αυτά. Ευχαριστώ λοιπόν πολύ τους ανθρώπους μου, την οικογένειά μου, για την ανευ όρων στήριξή τους, την καθοδήγηση και την ηθική συμπαράσταση, και τους φίλους μου, που έκαναν πάντα τις δύσκολες μέρες πιο φωτεινές, για την συνεχή υποστήριξή τους και τις αναμνήσεις που μοιραζόμαστε.

Αθήνα, Φεβρουάριος 2023

Αικατερίνη Αθανασιά

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Με την ανάπτυξη του Κινητού Υπολογισμού και την σημαντική αύξηση του όγκου των δεδομένων που παράγονται στα 'Άκρα' του Δικτύου, δημιουργήθηκε η ανάγκη για τη μεταφορά των υπολογισμών της Τεχνητής Νοημοσύνης σε αυτά. Ωστόσο, οι κινητές συσκευές αδυνατούν να εκμεταλλευτούν πλήρως τα αποτελέσματα που έχει επιφέρει η ραγδαία εξέλιξη στον τομέα της Τεχνητής Νοημοσύνης, λόγω της περιορισμένης διαθεσιμότητας πόρων σε συνδυασμό με την αύξηση στις υπολογιστικές απαιτήσεις για την εκτέλεση Βαθιών Νευρωνικών Δικτύων.

Έχει όμως παρατηρηθεί πως διαφορετικές είσοδοι απαιτούν διαφορετικό ποσό υπολογισμού για να επιτευχθεί μια πρόβλεψη στην οποία να έχει εμπιστοσύνη το εκάστοτε μοντέλο Βαθιάς Μάθησης. Με βάση την παρατήρηση αυτή, με σκοπό την μέγιστη εκμετάλλευση των πλεονεκτημάτων της εκτέλεσης Νευρωνικών Δικτύων κοντά στην πηγή των δεδομένων και της υψηλής υπολογιστικής ισχύος των εξυπηρετητών, διερευνάται η χρήση ενός Κατανεμημένου Συστήματος Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων με σκοπό τη βελτίωση στην ταχύτητα και την ακρίβεια συμπερασματολογίας δειγμάτων εικόνων αποθηκευμένων στις κινητές συσκευές. Τα μικρά μοντέλα, που συνήθως χρησιμοποιούνται από τις κινητές συσκευές, δυσκολεύονται να χειριστούν τις ακραίες περιπτώσεις δειγμάτων, αλλά μπορούν να εκτελέσουν αποτελεσματικά την συμπερασματολογία για την πλειονότητα των άλλων εισόδων. Το προτεινόμενο σύστημα αποτελείται από μια συνεργατική αρχιτεκτονική μεταξύ των άκρων του δικτύου και των κινητών συσκευών, που απαρτίζεται από διαφορετικά μοντέλα Βαθιών Νευρωνικών Δικτύων αποθηκευμένα σε διαφορετικές κινητές συσκευές και έναν εξυπηρετητή, καθένα από τα οποία εκτελείται σε ξεχωριστό υπολογιστικό κόμβο στο κατανεμημένο αυτό περιβάλλον. Για τις 'εύκολες' εισόδους, δηλαδή γι' αυτές στις οποίες το μοντέλο της εκάστοτε συσκευής δείχνει υψηλή εμπιστοσύνη στην συμπερασματολογία τους, δεχόμαστε την έξοδο της συσκευής, ενώ οι 'δύσκολες' προωθούνται στον εξυπηρετητή για εκ νέου εκτέλεση της συμπερασματολογίας στο πιο ισχυρό μοντέλο.

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η σχεδίαση και ανάπτυξη ενός ολοκληρωμένου Κατανεμημένου Συστήματος Συμπερασματολογίας σαν αυτό που περιγράφηκε, το οποίο θα παρέχει ευελιξία ως προς τη διασύνδεση των δύο μοντέλων και τη στρατηγική εκτέλεσης τους, έτσι ώστε να επιτυγχάνεται αποδοτική, σε σχέση με τις ανάγκες του χρήστη και

την κατάσταση του συστήματος, εκτέλεση εφαρμογών Βαθιάς Μάθησης σε κινητές συσκευές. Βασικός στόχος της σχεδίασης του συστήματος είναι να είναι επαρκώς παραμετροποιήσιμο, ώστε να είναι ευέλικτο ως προς το πλήθος και την ετερογένεια των συσκευών στο σύστημα και εύκολα επεκτάσιμο ώστε να μπορεί να λάβει υπόψιν επιπλέον περιορισμούς σε επίπεδο εξυπηρέτησης των χρηστών.

1.2 Δομή κι Οργάνωση Τόμου

Η διπλωματική εργασία είναι οργανωμένη σε 6 κεφάλαια: στο Δεύτερο Κεφάλαιο παρατίθεται το θεωρητικό υπόβραθρο πάνω στο οποίο βασίστηκε η παρούσα εργασία, στο Τρίτο Κεφάλαιο παρουσιάζεται η γλώσσα προγραμματισμού στην οποία γράφτηκε ο απαιτούμενος κώδικας, οι διάφορες τεχνολογίες και το σύνολο δεδομένων που χρησιμοποιήθηκαν στα πλαίσια της ανάπτυξης του συστήματος της εργασίας. Στο Τέταρτο Κεφάλαιο παρουσιάζεται η μοντελοποίηση του συστήματος που σχεδιάστηκε, με την περιγραφή της αρχιτεκτονικής και του τρόπου λειτουργίας του. Στο Πέμπτο Κεφάλαιο παρουσιάζεται ο τρόπος που έγινε η προσομοίωση του συστήματος με σκοπό την αξιολόγησή του. Τέλος, το Έκτο Κεφάλαιο αποτελείται από τα τελικά συμπεράσματα και μελλοντικές επεκτάσεις που θα μπορούσαν να γίνουν πάνω στο σύστημα.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται τα απαραίτητα θεωρητικά θεμέλια για τη μελέτη του θέματος της παρούσας διπλωματικής εργασίας.

2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση αποτελεί ένα σημαντικό υποπεδίο της Τεχνητής Νοημοσύνης και ως όρος επινοήθηκε το 1959 από τον Άρθουρ Σάμιουελ, υπάλληλο της IBM και πρωτοπόρο στον τομέα των ηλεκτρονικών παιχνιδιών και της Τεχνητής Νοημοσύνης [11]. Ως Μηχανική Μάθηση ορίζουμε το σύνολο των υπολογιστικών μεθόδων που χρησιμοποιούν προϋπάρχουσα γνώση, σε μορφή ψηφιακών πειραματικών δεδομένων, με σκοπό να κάνουν προβλέψεις, να βελτιώσουν την απόδοση προγραμμάτων ή να εξάγουν αποφάσεις υπό συνθήκες αβεβαιότητας [12][13].

Το πλεονέκτημα ενός αποτελεσματικού αλγορίθμου Μηχανικής Μάθησης, δηλαδή ενός αλγορίθμου ικανού να μαθαίνει από δεδομένα, έναντι των κλασσικών προσεγγίσεων επίλυσης προβλημάτων, που κάνουν χρήση ειδικά κατασκευασμένων προγραμμάτων των οποίων η συμπεριφορά καθορίζεται ρητά και στατικά από ευρετικές μεθόδους που έχουν παραχθεί χειροκίνητα, είναι σαφές. Αντί για την επίπονη προσέγγιση της δημιουργίας ενός ξεχωριστού, προσαρμοσμένου προγράμματος για την επίλυση κάθε μεμονωμένου προβλήματος σε έναν τομέα, ο ίδιος αλγόριθμος Μηχανικής Μάθησης χρειάζεται απλώς να μάθει, μέσω μιας διαδικασίας που ονομάζεται *εκπαίδευση*, να χειρίζεται κάθε νέο πρόβλημα. Συνεπώς, απαιτείται λιγότερος χρόνος και κόπος από την πλευρά του προγραμματιστή, αφού πλέον δεν απαιτείται ρητός προγραμματισμός για τη διαχείριση κάθε περίπτωσης ξεχωριστά, αλλά αρκεί ένας αλγόριθμος που να γενικεύει με βάση τα δεδομένα [2].

2.1.1 Είδη Μάθησης

Υπάρχουν τρία βασικά είδη μάθησης σε σχέση με την εκπαίδευση αλγορίθμων Μηχανικής Μάθησης: η Επιβλεπόμενη, η Μη Επιβλεπόμενη και η Ενισχυτική Μάθηση. Τα είδη αυτά διαφοροποιούνται ως προς τους τύπους των δεδομένων που χρησιμοποιούνται για την εκπαίδευση, τη σειρά και τον τρόπο που λαμβάνονται τα δεδομένα καθώς και κατά βάση το είδος ανατροφοδότησης που καθορίζει τον τρόπο μάθησης [12][3].

- *Επιβλεπόμενη Μάθηση*

  Σε αυτό το είδος μάθησης, το σύνολο (δεδομένων) εκπαίδευσης αποτελείται από ζεύγη εισόδου-εξόδου και ο αλγόριθμος 'μαθαίνει' μια συνάρτηση που αντιστοιχίζει την είσοδο στην εκάστοτε έξοδο. Η εν λόγω είσοδος αφορά ουσιαστικά παρατηρήσεις, που αποκαλούνται *χαρακτηριστικά*, για παράδειγμα μετρήσεις θερμοκρασίας, πίεσης, υγρασίας κ.τ.λ. Η επιθυμητή έξοδος από την άλλη, γνωστή και ως *ετικέτα* ή *στόχος*, παρέχεται από έναν εξωτερικό εικονικό παρατηρητή-παντογνώστη, γνωστό ως *επόπτη*, που δηλώνει πως οι συγκεκριμένες μετρήσεις αντιστοιχούν σε μια συγκεκριμένη ετικέτα, πως δηλαδή, σε συνέχεια του προηγούμενου παραδείγματος, οι μετρήσεις αυτές προμηνύουν ότι θα βρέξει. Δύο είναι οι πιο γνωστές κατηγορίες επιλύσιμων προβλημάτων με επιβλεπόμενη μάθηση: η ταξινόμηση και η παλινδρόμηση. Στην ταξινόμηση, οι ετικέτες ανήκουν σε ένα αυστηρά καθορισμένο σύνολο διακριτών τιμών, τις κλάσεις. Παράδειγμα προβλήματος ταξινόμησης είναι η εύρεση του τύπου του αντικειμένου που απεικονίζεται σε μια εικόνα. Στην παλινδρόμηση, οι ετικέτες παίρνουν συνεχείς, πραγματικές τιμές. Τέτοιο πρόβλημα αποτελεί για παράδειγμα η πρόβλεψη της ημερήσιας θερμοκρασίας.

- *Μη Επιβλεπόμενη Μάθηση*

  Όπως δηλώνει και το όνομά της, στη Μη Επιβλεπόμενη Μάθηση, το σύνολο εκπαίδευσης αποτελείται μόνο από δεδομένα αντίστοιχα με τα δεδομένα εισόδου που περιγράφηκαν στην Επιβλεπόμενη Μάθηση, χωρίς δηλαδή τις ετικέτες που θα παρείχε ο 'επόπτης'. Ο αλγόριθμος μάθησης συνεπώς, μαθαίνει να αναγνωρίζει μοτίβα και συσχετίσεις στα δεδομένα. Για παράδειγμα, μαθαίνει σταδιακά την έννοια της βροχερής ημέρας βασιζόμενος σε μοτίβα στα δεδομένα, χωρίς να έχει δοθεί από κάποιον *επόπτη* η ετικέτα που αντιστοιχεί στη βροχή. Η πιο χαρακτηριστική μέθοδος που εμπίπτει σε αυτή την κατηγορία είναι η ομαδοποίηση (ή συσταδοποίηση), όπου τα δεδομένα εισόδου χωρίζονται σε ομάδες με βάση τις ομοιότητές τους. Στη μελέτη των κοινωνικών δικτύων, για παράδειγμα, η ομαδοποίηση μπορεί να χρησιμοποιηθεί για την αναγνώριση κοινοτήτων μέσα σε μεγάλες ομάδες ανθρώπων.

- *Ενισχυτική Μάθηση*

  Η ενισχυτική μάθηση διαφοροποιείται αρκετά από τα προηγούμενα δύο είδη μάθησης καθώς, αντί για ένα σύνολο δεδομένων εκπαίδευσης, η εκμάθηση γίνεται αλληλεπιδρώντας με ένα περιβάλλον, μέσω μιας σειράς ενισχύσεων-ανταμοιβών ή τιμωριών, ανάλογα με το αν η εκτέλεση μιας ενέργειας οδήγησε σε μια επιθυμητή ή μη επιθυμητή κατάσταση αντίστοιχα. Παραδείγματος χάριν, αν κάνουμε μια αντιστοιχία μεταξύ αλγορίθμου και οδηγού ταξί, η έλλειψη φιλοδωρήματος στο τέλος της διαδρομής είναι μια ένδειξη πως ο αλγόριθμος-οδηγός έκανε κάποιο λάθος [3].

2.1.2 Βασικοί Αλγόριθμοι Μηχανικής Μάθησης

Παρακάτω παρουσιάζονται κάποιοι βασικοί αλγόριθμοι Μηχανικής Μάθησης. Οι πρώτοι δύο αποτελούν αλγορίθμους Επιβλεπόμενης Μάθησης ενώ οι επόμενοι δύο είναι αλγόριθμοι Μη Επιβλεπόμενης Μάθησης:

k-Πλησιέστεροι Γείτονες

Οι k-Πλησιέστεροι Γείτονες είναι μια οικογένεια μεθόδων που μπορούν να χρησιμοποιηθούν για ταξινόμηση ή παλινδρόμηση. Ως μη παραμετρικός αλγόριθμος μάθησης, δεν περιορίζεται σε ένα σταθερό αριθμό παραμέτρων. Δεν υπάρχει στάδιο εκπαίδευσης ή διαδικασία μάθησης, αντιθέτως όταν θέλουμε να παράγουμε μια έξοδο, δεδομένης μιας εισόδου ελέγχου, ο αλγόριθμος βρίσκει τους k-πλησιέστερους γείτονες της εισόδου, συνήθως σε σχέση με την ευκλείδια απόστασή τους, στα δεδομένα εκπαίδευσης. Ο αλγόριθμος αυτός μπορεί να επιτύχει υψηλή ακρίβεια δεδομένου ενός μεγάλου συνόλου εκπαίδευσης, ωστόσο αυτό γίνεται με υψηλό υπολογιστικό κόστος και δεν μπορεί να γενικεύσει καλά αν το σύνολο εκπαίδευσης είναι μικρό.

Μηχανές Διανυσμάτων Υποστήριξης

Το μοντέλο αυτό κάνει χρήση μιας γραμμικής συνάρτησης με σκοπό την ταξινόμηση ενός δείγματος όταν υπάρχουν συνολικά δύο κλάσεις. Όμως πολυπλοκότερα προβλήματα, στα οποία υπάρχουν πολυδιάστατα διανύσματα χαρακτηριστικών και πολλαπλές κλάσεις, αντιμετωπίζονται κάνοντας χρήση (μη γραμμικών) συναρτήσεων πυρήνα, η δράση των οποίων είναι ισοδύναμη με τον μετασχηματισμό των δεδομένων και στη συνέχεια την προσαρμογή ενός γραμμικού μοντέλου στον μετασχηματισμένο χώρο, στον οποίο πλέον οι κλάσεις είναι γραμμικά διαχωρίσιμες ανά δύο. Η χρήση των συναρτήσεων πυρήνα παρουσιάζει δύο σημαντικά οφέλη. Πρώτον, είναι δυνατό να εκπαιδευτούν μοντέλα τα οποία αποτελούν μη γραμμική συνάρτηση των εισόδων κάνοντας χρήση τεχνικών κυρτής βελτιστοποίησης που είναι εγγυημένο ότι συγκλίνουν αποτελεσματικά. Δεύτερον, η συνάρτηση πυρήνα επιλύει το πρόβλημα με σημαντικά πιο υπολογιστικά αποδοτικό τρόπο. Ένα σημαντικό μειονέκτημά τους είναι το υψηλό υπολογιστικό κόστος εκπαίδευσης όταν το σύνολο δεδομένων είναι μεγάλο.

Αλγόριθμος k-Μέσων

Ο αλγόριθμος αυτός είναι αλγόριθμος συσταδοποίησης που χωρίζει το σύνολο εκπαίδευσης σε k διαφορετικές συστάδες δειγμάτων που βρίσκονται το ένα κοντά στο άλλο. Λειτουργεί αρχικοποιώντας k κεντροειδή με διαφορετικές τιμές οι οποίες ανανεώνονται μέχρι η μέθοδος να συγκλίνει.

Ανάλυση σε Κύριες Συνιστώσες

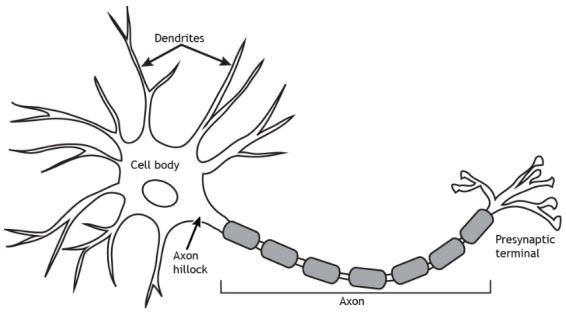
Ο αλγόριθμος Ανάλυσης σε Κύριες Συνιστώσες μαθαίνει έναν ορθογώνιο, γραμμικό μετασχηματισμό της εισόδου σε μια αναπαράσταση μικρότερης διάστασης με στοιχεία που δεν είναι γραμμικώς συσχετισμένα [14].

2.2 Τεχνητά Νευρωνικά Δίκτυα

Στο πλαίσιο της μελέτης της Μηχανικής Μάθησης, υπάρχει το υποπεδίο των Τεχνητών Νευρωνικών Δικτύων, που όπως είναι εμφανές από το όνομά του, είναι εμπνευσμένο από τη

δομή και τον τρόπο που αντιλαμβανόμαστε πως λειτουργεί ο εγκέφαλος. Καθώς ο εγκέφαλος είναι πλέον ο καλύτερος 'υπολογιστής' που γνωρίζουμε για τη μάθηση και την επίλυση προβλημάτων, είναι ένα προφανές σημείο αναζήτησης μιας μεθόδου μάθησης.

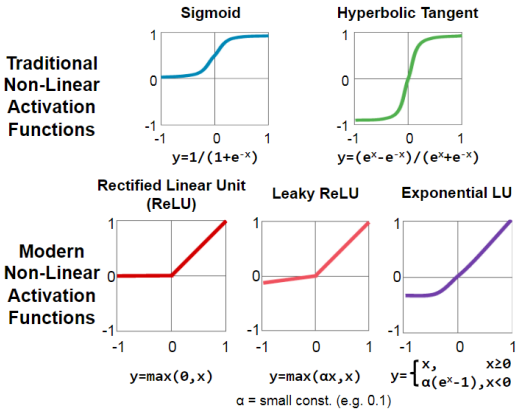
Ο *νευρώνας* θεωρείται το πρωταρχικό υπολογιστικό στοιχείο του εγκεφάλου. Οι ίδιοι οι νευρώνες συνδέονται μεταξύ τους με δομές που ονομάζονται *δενδρίτες*. Από κάθε νευρώνα, εξέρχεται ένας *άξονας* ο οποίος συνδέεται με δενδρίτες γειτονικών νευρώνων. Στον νευρώνα εισέρχονται σήματα μέσω των δενδριτών στα οποία εφαρμόζονται υπολογισμοί και το νέο σήμα εξέρχεται στον άξονα. Αυτά τα σήματα, εισόδου και εξόδου, ονομάζονται *ενεργοποιήσεις*. Μια *σύναψη* είναι ο σύνδεσμος μεταξύ μιας αξονικής διακλάδωσης και ενός δενδρίτη και μπορεί να κλιμακώσει το σήμα που τη διασχύει πολλαπλασιάζοντάς το με ένα παράγοντα, γνωστό ως *βάρος*. Επομένως, διαφορετικά βάρη οδηγούν σε διαφορετική αντίδραση δεδομένης κάποιας εισόδου. Αυτό είναι από τα βασικότερα χαρακτηριστικά της σύναψης καθώς το παραπάνω συνεπάγεται ότι η μάθηση δεν είναι τίποτα άλλο παρά η κατάλληλη προσαρμογή των βαρών αυτών ως απόκριση σε εξωτερικά ερεθίσματα, ενώ η εγκεφαλική δομή μένει απαράλλακτη. Αυτό



Εικόνα 2.1: *Νευρώνας εγκεφάλου [1]*

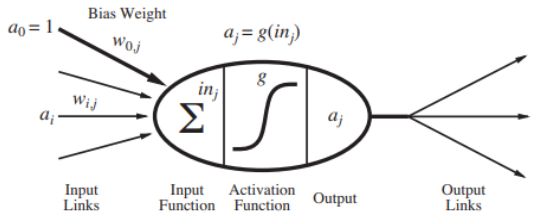
έρχεται σε πλήρη αντιστοιχία με τα Τεχνητά Νευρωνικά Δίκτυα, στα οποία επίσης ο τεχνητός νευρώνας αποτελεί θεμελιώδες στοιχείο τους. Εισέρχονται λοιπόν κάποιες τιμές στον νευρώνα οι οποίες όπως περιγράφηκε και στην παραπάνω παράγραφο, αποτελούν συνήθως τις εξόδους προηγούμενων νευρώνων. Στη συνέχεια, μέσα στον νευρώνα γίνεται ένας υπολογισμός και πιο συγκεκριμένα, υπολογίζεται το σταθμισμένο άθροισμα των εισόδων στο οποίο προστίθεται ένας παράγοντας, γνωστός ως *μεροληψία*. Το ποσό αυτό στη συνέχεια αποτελεί είσοδο σε μια συνάρτηση, τη *συνάρτηση ενεργοποίησης*, ώστε να παραχθεί η έξοδος του νευρώνα. Επιθυμούμε η συνάρτηση ενεργοποίησης να είναι μη γραμμική, καθώς μόνο έτσι επιτυγχάνεται η δημιουργία ενός μη γραμμικού μοντέλου και συνεπώς η επίτευξη μη γραμμικών αντιστοιχίσεων από τις εισόδους στις εξόδους. Αυτό ισχύει καθώς διαφορετικά τα Νευρωνικά Δίκτυα θα ήταν γραμμικός (λόγω του μετασχηματισμού μέσω γραμμικής συνάρτησης ενεργοποίησης) συνδυασμός γραμμικών μοντέλων (αφού το σταθμισμένο άθροισμα εξόδων προηγούμενων νευρώνων είναι επίσης γραμμικός μετασχηματισμός) και επομένως θα μπορούσαν να ανταποκριθούν μόνο σε γραμμικές απεικονίσεις των εισόδων στις αντίστοιχες εξόδους. Οι πιο χαρακτηριστικές μη γραμμικές συναρτήσεις ενεργοποίησης παρουσιάζονται στην Εικόνα 2.2.

Τη διαδικασία αυτή μπορούμε να την αντιμετωπίσουμε διαισθητικά ως εξαγωγή ενός συμπεράσματος με βάση την αξιολόγηση των πληροφοριών που λαμβάνει ο εκάστοτε νευρώνας [2]. Στην Εικόνα 2.3 παρουσιάζεται ένα απλό μαθηματικό μοντέλο ενός νευρώνα στα Τεχνητά



Εικόνα 2.2: *Γνωστές μη γραμμικές συναρτήσεις ενεργοποίησης [2]*

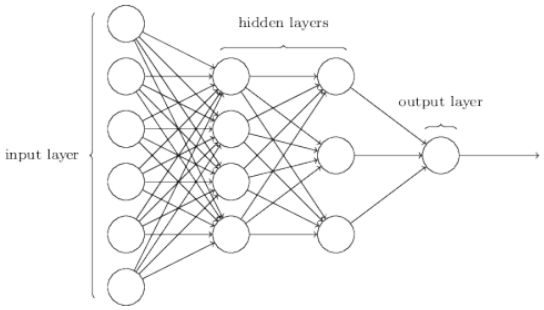
Νευρωνικά Δίκτυα. Η έξοδος είναι η $a_j = g(\sum_{i=0}^{n} w_i a_i)$ όπου w_0 η μεροληψία, g η συνάρτηση ενεργοποίησης, a_i η έξοδος του νευρώνα i και w_i το βάρος της σύναψης που ενώνει τον προηγούμενο νευρώνα i με αυτόν τον νευρώνα j.



Εικόνα 2.3: *Απλό μαθηματικό μοντέλο ενός νευρώνα [3]*

Έχοντας περιγράψει το θεμελιώδες δομικό στοιχείο των Τεχνητών Νευρωνικών Δικτύων, μπορούμε να αναφερθούμε τώρα στον τρόπο με τον οποίο μοντελοποιούνται. Τα Τεχνητά Νευρωνικά Δίκτυα δεν είναι παρά γραφήματα των οποίων οι κόμβοι είναι νευρώνες και οι κατευθυνόμενες ακμές αντιστοιχούν στους δενδρίτες και τους άξονες. Στις αρχιτεκτονικές Τεχνητών Νευρωνικών Δικτύων που αναφέρονται στην παρούσα εργασία, οι νευρώνες είναι οργανωμένοι σε στρώματα, με τον πιο συνηθισμένο τύπο να είναι τα *πλήρως συνδεδεμένα*, όπου οι νευρώνες ενός στρώματος είναι συνδεδεμένοι με όλους τους νευρώνες του προηγούμενου και του επόμενου στρώματος και δεν υπάρχουν συνδέσεις μεταξύ νευρώνων που ανήκουν στο ίδιο στρώμα [15].

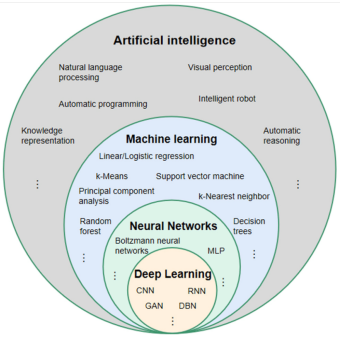
Ανάλογα με την θέση τους στο δίκτυο, τα στρώματα κατατάσσονται σε τρεις κατηγορίες: το πρώτο στρώμα στο οποίο εισάγονται τα δεδομένα, γνωστό ως στρώμα εισόδου, το τελευταίο στρώμα, που ονομάζεται στρώμα εξόδου, από το οποίο εξάγεται η συμπερασματολογία και τα (προαιρετικά) ενδιάμεσα στρώματα, τα κρυφά στρώματα. Η ροή της πληροφορίας στο δίκτυο είναι η εξής: οι νευρώνες στο στρώμα εισόδου λαμβάνουν κάποιες τιμές και τις μεταδίδουν στους νευρώνες του πρώτου κρυφού στρώματος του δικτύου. Τα μετασχηματισμένα σταθμισμένα αθροίσματα από ένα ή περισσότερα κρυφά στρώματα διαδίδονται τελικά στο στρώμα εξόδου, το οποίο παρουσιάζει τις τελικές εξόδους του δικτύου στον χρήστη [2].



Εικόνα 2.4: *Αρχιτεκτονική Νευρωνικού Δικτύου [4]*

2.3 Βαθιά Μάθηση

Η Βαθιά Μάθηση (επίσης γνωστή ως Βαθιά Μηχανική Μάθηση) είναι η μελέτη των Τεχνητών Νευρωνικών Δικτύων, που περιέχουν περισσότερα από ένα κρυφά στρώματα.



Εικόνα 2.5: *Σχέση μεταξύ της Τεχνητής Νοημοσύνης, της Μηχανικής Μάθησης, των Τεχνητών Νευρωνικών Δικτύων και της Βαθιάς Μάθησης [5]*

Ένα από τα σημαντικά πλεονεκτήματα της Βαθιάς Μάθησης είναι η αντικατάσταση της 'χειρονακτικής' εξαγωγής χαρακτηριστικών, που γινόταν δηλαδή με ανθρώπινη παρέμβαση, με αποδοτικούς αλγορίθμους.

Τα Βαθιά Νευρωνικά Δίκτυα:

- Χρησιμοποιούν μια σειρά από πολλά στρώματα μη γραμμικών μονάδων επεξεργασίας, των νευρώνων, για εξαγωγή χαρακτηριστικών και μετασχηματισμούς. Κάθε επόμενο στρώμα χρησιμοποιεί την έξοδο του προηγούμενου ως είσοδο, ενώ οι αλγόριθμοι είναι είτε επιβλεπόμενης είτε μη επιβλεπόμενης μάθησης.
- Βασίζονται στη μάθηση πολλαπλών επιπέδων χαρακτηριστικών ή αναπαραστάσεων των δεδομένων. Υψηλότερου επιπέδου χαρακτηριστικά προκύπτουν από τον μετασχηματισμό χαρακτηριστικών χαμηλότερου επιπέδου σχηματίζοντας μια ιεραρχική αναπαράσταση. Μαθαίνουν συνεπώς πολλά επίπεδα αναπαραστάσης που αντιστοιχούν σε διαφορετικά επίπεδα αφαιρετικότητας και τα επίπεδα αυτά σχηματίζουν μια εννοιολογική ιεραρχία.
- Είναι μέρος ενός ευρύτερου τομέα Μηχανικής Μάθησης, της εκμάθησης αναπαραστάσεων των δεδομένων.

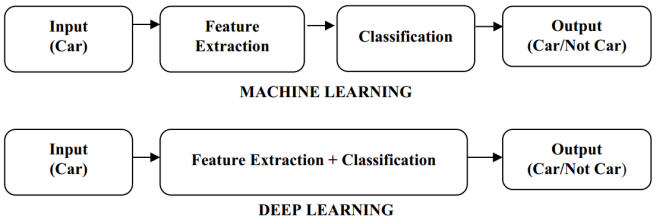
Σε ένα Βαθύ Νευρωνικό Δίκτυο υπάρχουν πολλά στρώματα μεταξύ της εισόδου και της εξόδου, επιτρέποντας στον αλγόριθμο να χρησιμοποιεί πολλαπλά στρώματα επεξεργασίας, αποτελούμενα από πολλαπλούς γραμμικούς και μη γραμμικούς μετασχηματισμούς [16].

Δεδομένου ότι τα Βαθιά Νευρωνικά Δίκτυα είναι μια περίπτωση αλγορίθμου μηχανικής μάθησης, το βασικό μοντέλο δεν αλλάζει καθώς μαθαίνει να εκτελεί τα καθήκοντα που του έχουν ανατεθεί. Στη συγκεκριμένη περίπτωση των Βαθιών Νευρωνικών Δικτύων, η εκμάθηση αυτή περιλαμβάνει τον καθορισμό της τιμής των παραμέτρων στο δίκτυο, και αναφέρεται ως *εκπαίδευση* του δικτύου. Αφού εκπαιδευτεί, ο αλγόριθμος μπορεί να εκτελέσει το έργο του υπολογίζοντας την έξοδο του δικτύου, χρησιμοποιώντας τις παραμέτρους που καθορίστηκαν κατά τη διαδικασία εκπαίδευσης. Η εκτέλεση του μοντέλου με αυτές τις παραμέτρους αναφέρεται ως *συμπερασματολογία* [2].

2.3.1 Διαφορές Βαθιάς Μάθησης - Μηχανικής Μάθησης

Η πολυεπίπεδη αρχιτεκτονική, όπως ήδη αναφέρθηκε, διευκολύνει την αντιστοίχιση της εισόδου σε υψηλότερο επίπεδο αναπαράστασης. Αυτό όμως σημαίνει ότι τα μοντέλα Βαθιάς Μάθησης είναι μεγαλύτερα, με πολλές παραμέτρους και περισσότερα στρώματα σε σχέση με τα ρηχά μοντέλα Μηχανικής Μάθησης. Οι κύριες διαφορές που εντοπίζονται μεταξύ της Βαθιάς και της Μηχανικής Μάθησης λόγω των παραπάνω είναι οι εξής:

- Η σύνθετη και πολύπλοκη φάση της μηχανικής χαρακτηριστικών εξαλείφεται στη Βαθιά Μάθηση σε αντίθεση με τη Μηχανική Μάθηση, αφού στη Βαθιά Μάθηση δημιουργούνται νέα χαρακτηριστικά με τις δικές της διαδικασίες και τεχνικές, ενώ στην περίπτωση της Μηχανικής Μάθησης, τα χαρακτηριστικά αναγνωρίζονται με ακρίβεια από τους χρήστες.
- Η Βαθιά Μάθηση απαιτεί μεγάλο όγκο δεδομένων, ενώ η Μηχανική Μάθηση χρειάζεται μικρότερο όγκο δεδομένων για να λειτουργήσει και να φτάσει σε ένα συμπέρασμα.
- Η Βαθιά Μάθηση απαιτεί υλικό με πολύ υψηλές επιδόσεις. Χρειάζονται υψηλών προδιαγραφών μονάδες γραφικής επεξεργασίας (GPU) οι οποίες είναι ακριβές και είναι επιδέξιες σε επαρκή χρόνο με μεγάλα δεδομένα.
- Η Βαθιά Μάθηση λύνει το πρόβλημα από άκρη σε άκρη, ενώ η Μηχανική Μάθηση το επιλύει αποσυνθέτοντας το σε μικρότερα υποπροβλήματα και στη συνέχεια συνδυάζει τα αποτελέσματα.
- Τα Βαθιά Δίκτυα είναι δίκτυα 'μαύρα κουτιά' και η λειτουργία τους είναι πολύ δύσκολο να κατανοηθεί από τους χρήστες εξαιτίας των υπερπαραμέτρων και του πολύπλοκου σχεδιασμού του δικτύου.
- Η απαίτηση χρόνου για εκπαίδευση είναι πολύ μεγαλύτερη στη Βαθιά Μάθηση από ότι στη Μηχανική Μάθηση.
- Το ποσοστό ακρίβειας που επιτυγχάνεται με τη Βαθιά Μάθηση είναι πολύ ικανοποιητικό σε σύγκριση με τη Μηχανική Μάθηση [6].



Εικόνα 2.6: *Βασική διαφορά μεταξύ Μηχανικής Μάθησης και Βαθιάς Μάθησης [6]*

2.3.2 Συνελικτικά Νευρωνικά Δίκτυα

Στη Μηχανική Μάθηση, ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN) είναι ένας τύπος Τεχνητού Νευρωνικού Δικτύου στο οποίο το μοτίβο συνδεσιμότητας μεταξύ των νευρώνων του είναι εμπνευσμένο από την οργάνωση του οπτικού φλοιού των ζώων. Οι επιμέρους νευρώνες του φλοιού ανταποκρίνονται σε ερεθίσματα σε μια περιορισμένη περιοχή του χώρου, γνωστή ως *δεκτικό πεδίο*. Τα δεκτικά πεδία διαφορετικών νευρώνων επικαλύπτονται εν μέρει, έτσι ώστε να καλύπτουν ολόκληρο το οπτικό πεδίο. Η απόκριση ενός μεμονωμένου νευρώνα σε ερεθίσματα εντός του δεκτικού του πεδίου μπορεί να προσεγγιστεί μαθηματικά με την πράξη της συνέλιξης, από την οποία πήρε και ο συγκεκριμένος τύπος Νευρωνικών Δικτύων το όνομά του [16].

Η πράξη της συνέλιξης μεταξύ δύο συναρτήσεων μιας διάστασης συνεχούς χρόνου, της εισόδου ($Input$) έστω I και του πυρήνα ($Kernel$) ή φίλτρου, έστω K, συμβολίζεται με $I * K$ και ορίζεται ως εξής:

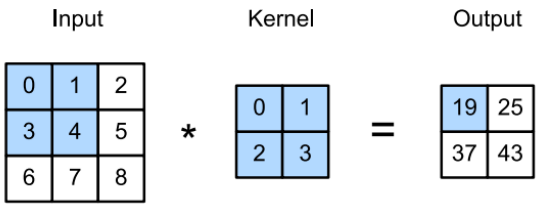
$$(I * K)(t) = \int I(a)K(t-a)\,da$$

Στην πραγματικότητα όμως αν ως είσοδο έχουμε μια δισδιάστατη εικόνα, δεν μπορούμε να μιλήσουμε πλέον για συνεχή αλλά για διακριτά δεδομένα και συνεπώς το ολοκλήρωμα μετατρέπεται σε άπειρο άθροισμα. Ύστερα απο μετασχηματισμούς και εκμεταλλευόμενοι την αντιμεταθετική ιδιότητα της, προκύπτει ο παρακάτω ορισμός της δισδιάστατης συνέλιξης μεταξύ μιας δισδιάστατης εικόνας I κι ενός δισδιάστατου πυρήνα K, που είναι αυτός που χρησιμοποιείται ευρέως στη Μηχανική Μάθηση.

$$(I * K)(t), = \sum_m \sum_n I(i+m, j+n)K(m,n)$$

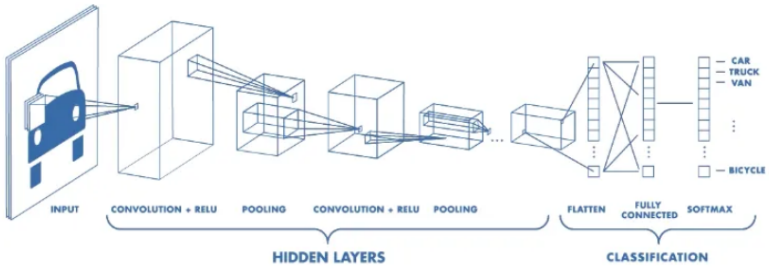
Ο ορισμός αυτός αντιστοιχεί στην διασταυρούμενη συσχέτιση αλλά οι όροι χρησιμοποιούνται ισοδύναμα [14]. Στην Εικόνα 2.7 παρουσιάζεται ένα παράδειγμα δισδιάστατης συνέλιξης. Τα χρωματισμένα τμήματα αντιστοιχούν στο πρώτο στοιχείο εξόδου, καθώς επίσης και στα στοιχεία εισόδου και πυρήνα που χρησιμοποιούνται για τον υπολογισμό της.

Λόγω της αρχιτεκτονικής τους, τα Συνελικτικά Νευρωνικά Δίκτυα προσφέρονται για διάφορες εφαρμογές σε αναγνώριση ομιλίας, αναγνώριση εικόνων και βίντεο και επεξεργασία φυσικής γλώσσας. Αυτά τα οποία χρησιμοποιούνται για επίλυση προβλημάτων ταξινόμησης αποτελούνται συνήθως από τρεις τύπους στρωμάτων, τα *συνελικτικά στρώματα*, τα *στρώματα συγκέντρωσης* και τα *πλήρως συνδεδεμένα στρώματα*. Οι δύο πρώτοι τύποι, εξάγουν χαρακτηριστικά ενώ ο τρίτος αντιστοιχίζει τα εξαγόμενα χαρακτηριστικά στην τελική έξοδο.



Εικόνα 2.7: *Δισδιάστατη συνέλιξη [7]*

Καθώς ένα επίπεδο τροφοδοτεί την έξοδό του στο επόμενο επίπεδο, τα εξαγόμενα χαρακτηριστικά μπορούν να γίνουν ιεραρχικά και προοδευτικά πιο πολύπλοκα. Μια τυπική αρχιτεκτονική αποτελείται από επαναλήψεις μιας στοίβας συνελικτικών στρωμάτων και ενός στρώματος συγκέντρωσης, ακολουθούμενα από ένα ή περισσότερα πλήρως συνδεδεμένα στρώματα [17].



Εικόνα 2.8: *Βασική αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου για ταξινόμηση [8]*

Πιο αναλυτικά για τα στρώματα, το **συνελικτικό στρώμα** είναι το βασικότερο στρώμα στα Συνελικτικά Νευρωνικά Δίκτυα. Ουσιαστικά συνελίσσει τον πίνακα με τα pixel που δημιουργείται για τη δεδομένη εικόνα με ένα φίλτρο που δρα ως ανιχνευτής χαρακτηριστικών ώστε να παράγει έναν χάρτη ενεργοποίησης, γνωστό και ως *χάρτη χαρακτηριστικών* για τη δεδομένη εικόνα [18]. Ο χάρτης ενεργοποίησης είναι ουσιαστικά ένας 'χάρτης'-απεικόνιση των σημαντικών για την πρόβλεψη περιοχών της εισόδου [17]. Το κύριο πλεονέκτημα του είναι ότι αποθηκεύει όλα τα διακριτικά χαρακτηριστικά μιας δεδομένης εικόνας, ενώ ταυτόχρονα μειώνει τον όγκο των δεδομένων που πρέπει να υποβληθούν σε επεξεργασία [18]. Η εφαρμογή πολλαπλών πυρήνων σχηματίζει έναν αριθμό χαρτών ενεργοποίησης των πινάκων εισόδου. Οι διαφορετικοί πυρήνες λοιπόν μπορούν να θεωρηθούν διαφορετικοί εξαγωγείς χαρακτηριστικών [17].

Το **στρώμα συγκέντρωσης** μειώνει τη διάσταση των χαρτών χαρακτηριστικών, ώστε η αποτύπωση του αντικειμένου στον χάρτη χαρακτηριστικών να είναι αναλλοίωτη ως προς μικρές μετατοπίσεις και παραμορφώσεις και μειώνει τον αριθμό των επακόλουθων παραμέτρων προς εκμάθηση [17]. Η συγκέντρωση επιτρέπει στο Συνελικτικό Νευρωνικό Δίκτυο να ενσωματώνει όλες τις διαφορετικές διαστάσεις μιας εικόνας, ώστε να αναγνωρίζει με επιτυχία το συγκεκριμένο αντικείμενο, ακόμη και αν το σχήμα του είναι λοξό ή σε διαφορετική γωνία. Υπάρχουν διάφοροι τύποι συγκέντρωσης, όπως η μέγιστη συγκέντρωση, η μέση συγκέντρωση, η στοχαστική συγκέντρωση, η συγκέντρωση με χωρική πυραμίδα, κ.α. Η πιο δημοφιλής είναι η μέγιστη συγκέντρωση η οποία παίρνει την υψηλότερη τιμή από κάθε υποπίνακα του χάρτη

ενεργοποίησης και σχηματίζει έναν ξεχωριστό πίνακα. Με αυτόν τον τρόπο διασφαλίζεται ότι τα χαρακτηριστικά προς εκμάθηση παραμένουν περιορισμένα σε αριθμό, ενώ παράλληλα διατηρούνται τα βασικά χαρακτηριστικά κάθε εικόνας [18].

Το τελευταίο στρώμα του δικτύου είναι το **πλήρως συνδεδεμένο στρώμα**. Οι χάρτες χαρακτηριστικών εξόδου του τελικού συνελικτικού στρώματος ή στρώματος συγκέντρωσης μετατρέπονται σε μονοδιάστατο πίνακα αριθμών (ή διάνυσμα), και συνδέονται σε ένα ή περισσότερα πλήρως συνδεδεμένα στρώματα στα οποία κάθε είσοδος συνδέεται με κάθε έξοδο μέσω ενός προς εκμάθηση βάρους. Μόλις εξαχθούν τα χαρακτηριστικά από τα συνελικτικά στρώματα και μειωθεί η διάστασή τους από τα στρώματα συγκέντρωσης, αντιστοιχίζονται από ένα υποσύνολο πλήρως συνδεδεμένων στρωμάτων στις τελικές εξόδους του δικτύου, όπως για παράδειγμα στις πιθανότητες για κάθε κλάση σε προβλήματα ταξινόμησης. Το τελικό πλήρως συνδεδεμένο στρώμα έχει τον ίδιο αριθμό κόμβων εξόδου με τον αριθμό των κλάσεων [17].

Καθώς στην εργασία αυτή, το πρόβλημα που θα μας απασχολήσει είναι η Κατηγοριοποίηση Εικόνας, παρακάτω ακολουθούν λίγα λόγια για τις αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων που χρησιμοποιήθηκαν (με χρονολογική σειρά):

- **Inception** (09/2014): Γνωστό και ως GoogLeNet. Βελτίωσε την αξιοποίηση υπολογιστικών πόρων εντός του δικτύου μέσω ενός σχεδιασμού που επιτρέπει την αύξηση του βάθους και του πλάτους του δικτύου, διατηρώντας σταθερό τον υπολογιστικό προϋπολογισμό [19].

- **MobileNet** (04/2017): Αποδοτικό για κινητές και ενσωματωμένες εφαρμογές όρασης. Χρησιμοποιεί διαχωρίσιμες κατά βάθος συνελίξεις για τη δημιουργία ελαφρών βαθιών νευρωνικών δικτύων. Εισάγει υπερ-παραμέτρους που αντισταθμίζουν τον χρόνο απόκρισης με την ακρίβεια [20].

- **NASNet** (07/2017): Χρησιμοποιεί μια μέθοδο αναζήτησης ενισχυτικής μάθησης για τη βελτιστοποίηση της αρχιτεκτονικής. Η αρχιτεκτονική που μαθαίνεται είναι αρκετά ευέλικτη, καθώς μπορεί να κλιμακωθεί όσον αφορά το υπολογιστικό κόστος και τις παραμέτρους για να αντιμετωπίσει εύκολα μια ποικιλία προβλημάτων [21].

- **EfficientNet** (05/2019): Κλιμακώνει ομοιόμορφα όλες τις διαστάσεις, βάθος, πλάτος, ανάλυση, χρησιμοποιώντας έναν απλό αλλά αποτελεσματικό σύνθετο συντελεστή [22].

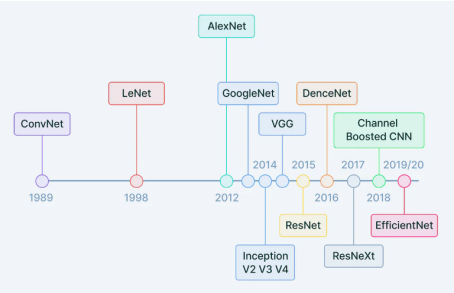
Για ιστορικούς λόγους, άξιες αναφοράς είναι επίσης οι εξής αρχιτεκτονικές:

- **LeNet** (11/1998): Αναπτύχθηκε από τον Yann LeCun το 1998 για να αναγνωρίζει χειρόγραφους αριθμούς [23].

- **AlexNet** (2012): Η πρώτη αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου που νίκησε την πρόκληση εικονικής αναγνώρισης μεγάλης κλίμακας ImageNet [24].

- **VGG** (09/2014): Έδειξε ότι η απόδοση μπορεί να βελτιωθεί σημαντικά με την αύξηση του βάθους [25].

- **ResNet** (12/2015): Ευκολότερο να βελτιστοποιηθούν και μπορούν να αυξήσουν την ακρίβεια μέσω αύξησης του βάθους [26].



Εικόνα 2.9: *Ιστορική εξέλιξη των αρχιτεκτονικών των Συνελικτικών Νευρωνικών Δικτύων [6]*

2.4 Κατανεμημένη Μηχανική Μάθηση

Όπως αναφέρθηκε νωρίτερα, οι αλγόριθμοι Μηχανικής Μάθησης κάνουν χρήση μεγάλου όγκου δεδομένων για την εκπαίδευση ενός μοντέλου με ικανοποιητική ακρίβεια σε προβλέψεις πάνω σε άγνωστα δείγματα. Με την πάροδο του χρόνου το φαινόμενο αυτό εντατικοποιείται περαιτέρω, καθώς ο διαθέσιμος όγκος πληροφορίας συνεχώς αυξάνεται και το μέγεθος των μοντέλων συνεχώς κλιμακώνεται.

Για την αύξηση της ταχύτητας και της αποδοτικότητας της επεξεργασίας των δεδομένων, δεδομένης της έλευσης του Κινητού Υπολογισμού, προτάθηκε η ενσωμάτωση της Τεχνητής Νοημοσύνης στις συσκευές στα 'άκρα' του δικτύου, που πλέον είναι αυτές που δημιουργούν την πλειονότητα των δεδομένων, από τα οποία οι χρήστες των συσκευών επιθυμούν να καταλήξουν σε κάποιο συμπέρασμα ή να ενεργήσουν με βάση αυτά. Στο πλαίσιο αυτό, αναπτύχθηκε η Κατανεμημένη Μηχανική Μάθηση που αφορά τη μεταφορά των υπολογισμών της Τεχνητής Νοημοσύνης έξω από τα κέντρα δεδομένων, στα οποία μέχρι πριν κάποια χρόνια ήταν 'αποκλεισμένη'. Όμως, λόγω της αυξημένης πολυπλοκότητας των μοντέλων Βαθιάς Μάθησης και των περιορισμένων πόρων των κινητών συσκευών, δεν είναι πάντα αποδοτική η εκτέλεση των μοντέλων αυτών στις συσκευές. Από την άλλη, η μεταφορά των υπολογισμών σε κάποιον εξυπηρετητή νέφους συνεπάγεται επιπλέον καθυστέρηση για τη μεταφορά των δεδομένων ενώ επίσης δεν εξασφαλίζεται πλέον η ιδιωτικότητα των δεδομένων του εκάστοτε χρήστη. Συνεπώς, η βέλτιστη λύση φαίνεται να είναι ένας συνδυασμός αυτών.

2.4.1 Κινητός Υπολογισμός και Τεχνητή Νοημοσύνη

Η ανάπτυξη του Κινητού Υπολογισμού οφείλεται στην ραγδαία εξέλιξη του διαδικτύου, των ασύρματων δικτύων και τον υπολογιστικών μηχανών, μεταξύ άλλων.

Ως έννοια, αναφέρεται στο σύνολο των τεχνολογιών, προϊόντων και υπηρεσιών που επιτρέπουν στους τελικούς χρήστες να έχουν πρόσβαση σε υπολογισμούς, πληροφορία και πόρους ενώ βρίσκονται σε κίνηση. Η κινητικότητα συνήθως ορίζεται ως η πρόσβαση εν κινήσει, όπου ο χρήστης δεν περιορίζεται σε μια συγκεκριμένη γεωγραφική θέση. Η κινητή πρόσβαση μπορεί επίσης να αναφέρεται σε πρόσβαση σε σταθερή τοποθεσία μέσω εξοπλισμού που οι χρήστες μπορούν να μετακινούν ανάλογα με τις ανάγκες, ο οποίος όμως είναι σταθερός κατά τη λειτουργία του [27].

Ο Κινητός Υπολογισμός, ως όρος, περιλαμβάνει τους τομείς του Υπολογισμού στο Νέφος, του Υπολογισμού 'Ομίχλης' και του Υπολογισμού στα 'Άκρα' του Δικτύου, όπως αυτοί ο-

ρίζονται παρακάτω.

- **Υπολογισμός στο Νέφος**

  Ο Υπολογισμός στο Νέφος ορίστηκε το 2011 από το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας (NIST) ως ένα μοντέλο που επιτρέπει την καθολική, εύχρηστη, κατά παραγγελία δικτυακή πρόσβαση σε μια κοινή συλλογή υπολογιστικών πόρων (π.χ. δίκτυα, εξυπηρετητές, αποθηκευτικός χώρος, εφαρμογές και υπηρεσίες) που μπορούν να παρέχονται και να αποδεσμεύονται γρήγορα με ελάχιστη προσπάθεια διαχείρισης ή αλληλεπίδραση με τον πάροχο υπηρεσιών. Με βάση αυτό, ο υπολογισμός στο Νέφος αναφέρεται τόσο στις εφαρμογές που παρέχονται ως υπηρεσίες μέσω του Διαδικτύου όσο και στο υλικό και το λογισμικό συστημάτων στα κέντρα δεδομένων που παρέχουν αυτές τις υπηρεσίες.

  Το υλικό και το λογισμικό του κέντρου δεδομένων είναι αυτό που ονομάζεται Υπολογιστικό Νέφος.

- **Υπολογισμός 'Ομίχλης'**

  Για να αντιμετωπιστούν ορισμένοι από τους περιορισμούς του Υπολογιστικού Νέφους, η ερευνητική κοινότητα πρότεινε την έννοια του υπολογισμού 'Ομίχλης' με στόχο να φέρει τα χαρακτηριστικά των υπηρεσιών νέφους πιο κοντά σε αυτό που αναφέρεται ως 'πράγματα', στο Διαδίκτυο των Πραγμάτων, συμπεριλαμβανομένων αισθητήρων, ενσωματωμένων συστημάτων, κινητών τηλεφώνων, αυτοκινήτων, κ.λπ.

  Ο πρώτος επίσημος ορισμός του υπολογισμού 'Ομίχλης' διατυπώθηκε το 2012 από την CISCO ως εξής: 'Η υπολογιστική ομίχλης είναι μια εικονική πλατφόρμα που παρέχει υπηρεσίες υπολογισμού, αποθήκευσης και δικτύωσης μεταξύ των τελικών συσκευών και των παραδοσιακών κέντρων δεδομένων υπολογιστικού νέφους, τα οποία συνήθως, αλλά όχι αποκλειστικά, βρίσκονται στα άκρα του δικτύου'.

- **Υπολογισμός στα 'Άκρα' του Δικτύου**

  Τα 'Άκρα' του Δικτύου ουσιαστικά αποτελούν τόσο οι δρομολογητές, οι κινητοί σταθμοί βάσης που δρομολογούν την κυκλοφορία στο δίκτυο όσο και οι συσκευές που έχουν απευθείας σύνδεση με το διαδίκτυο, όπως τα έξυπνα τηλέφωνα, μια πύλη δικτύου σε ένα έξυπνο σπίτι κ.λπ.

  Στον Υπολογισμό στα 'Άκρα' του Δικτύου τα δεδομένα επεξεργάζονται μακριά από τον συγκεντρωτικό χώρο αποθήκευσης, διατηρώντας πληροφορίες στα τοπικά τμήματα του δικτύου, τις συσκευές στα άκρα του δικτύου και τις πύλες. Η έννοια αυτή τοποθετεί τις εφαρμογές, τα δεδομένα και την επεξεργασία στα λογικά άκρα ενός δικτύου αντί να τις συγκεντρώνει. Η τοποθέτηση των δεδομένων και των εφαρμογών που κάνουν έντονη χρήση δεδομένων στα 'άκρα' μειώνει τον όγκο και την απόσταση που πρέπει να διακινηθούν τα δεδομένα αυτά και συνεπώς επιτρέπει την άμεση απόκριση και παρέχει πρωτοφανή ταχύτητα [28][29].

Ο συνδυασμός της Τεχνητής Νοημοσύνης και του υπολογισμού στα άκρα του δικτύου δημιούργησε ένα νέο ερευνητικό πεδίο το οποίο εκμεταλλεύεται τα δεδομένα που παράγονται

από τις συσκευές στα άκρα με αποτελεσματικό, οικονομικά προσιτό και προσβάσιμο τρόπο ενώ ταυτόχρονα αυξάνεται η ποικιλομορφία σε σχέση με τις εφαρμογές της Τεχνητής Νοημοσύνης [10].

Πιο συγκεκριμένα, η Τεχνητή Νοημοσύνη στα άκρα του δικτύου αναφέρεται στις τεχνολογίες που επιτρέπουν την εκτέλεση υπολογισμών στα άκρα του δικτύου, σε δεδομένα downstream για λογαριασμό υπηρεσιών νέφους και upstream δεδομένα για λογαριασμό υπηρεσιών του Διαδικτύου των Πραγμάτων. Εδώ ορίζουμε ως 'άκρα' όλους τους υπολογιστικούς και δικτυακούς πόρους κατά μήκος της διαδρομής μεταξύ πηγών δεδομένων και των κέντρων δεδομένων νέφους [30].

Ανάλογα με την ποσότητα των δεδομένων προς μεταφόρτωση και το μήκος της διαδρομής που πρέπει να ακολουθήσουν, η Τεχνητή Νοημοσύνη στα άκρα του δικτύου διαχωρίζεται σε έξι επίπεδα, με αυτά να ορίζονται ως εξής:

1. **Συνεργατική συμπερασματολογία ανάμεσα στο νέφος και τα άκρα του δικτύου, εκπαίδευση στο νέφος**

   Εκπαίδευση του μοντέλου Βαθέος Νευρωνικού Δικτύου στο νέφος, αλλά με τη συμπερασματολογία να εκτελείται με εναν συνεργατικό τρόπο ανάμεσα στα άκρα του δικτύου και το νέφος, πιο συγκεκριμένα με μερική μεταφόρτωση των δεδομένων στο νέφος.

2. **Συμπερασματολογία στα άκρα του δικτύου, εκπαίδευση στο νέφος**

   Εκπαίδευση του μοντέλου Βαθέος Νευρωνικού Δικτύου στο νέφος, αλλά με τη συμπερασματολογία να εκτελείται στα άκρα του δικτύου, η οποία μπορεί να πραγματοποιηθεί με την πλήρη ή μερική μεταφόρτωση των δεδομένων στους κόμβους στα άκρα ή σε κοντινές συσκευές.

3. **Συμπερασματολογία στη συσκευή, εκπαίδευση στο νέφος**

   Εκπαίδευση του μοντέλου Βαθέος Νευρωνικού Δικτύου στο νέφος, αλλά εκτέλεση της συμπερασματολογίας αποκλειστικά στη συσκευή, χωρίς δηλαδή μεταφόρτωση δεδομένων.

4. **Συνεργατική συμπερασματολογία και εκπαίδευση στο νέφος και τα άκρα του δικτύου**

   Εκπαίδευση και συμπερασματολογία για το Βαθύ Νευρωνικό Δίκτυο με συνεργατικό τρόπο στα άκρα του δικτύου και το νέφος.

5. **Εξ΄ ολοκλήρου στα άκρα του δικτύου**

   Εκπαίδευση και συμπερασματολογία για το Βαθύ Νευρωνικό Δίκτυο αποκλειστικά στα άκρα του δικτύου.

6. **Εξ΄ ολοκλήρου στη συσκευή**

   Εκπαίδευση και συμπερασματολογία για το Βαθύ Νευρωνικό Δίκτυο αποκλειστικά στη συσκευή.

Όπως παρατηρεί κανείς, όσο ανεβαίνει το επίπεδο, η ποσότητα των δεδομένων προς μεταφόρτωση και το μήκος της διαδρομής που πρέπει να ακολουθήσουν μειώνεται. Συνεπώς, η καθυστέρηση μετάδοσης της μεταφόρτωσης δεδομένων μειώνεται, η ιδιωτικότητα των δεδομένων βελτιώνεται και το κόστος εύρους ζώνης μειώνεται. Ωστόσο, αυτό επιτυγχάνεται με το κόστος της αυξημένης καθυστέρησης υπολογισμών και της κατανάλωσης ενέργειας. Δεδομένων των παραπάνω, δεν υπάρχει απόλυτα βέλτιστο επίπεδο, παρά μόνο βέλτιστο σε σχέση με τις απαιτήσεις που έχει κανείς από μια συγκεκριμένη εφαρμογή [10].

2.4.2 Κατανεμημένη Εκπαίδευση

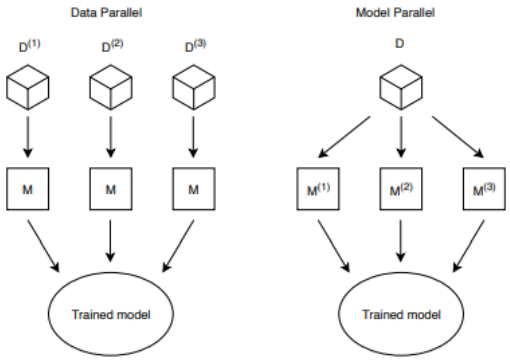
Με την αύξηση του διαθέσιμου όγκου δεδομένων και την βελτίωση της ακρίβειας των μοντέλων, που προκύπτει από τη χρήση πολυπλοκότερων αρχιτεκτονικών, ήταν ανέφικτο για έναν μόνο εξυπηρετητή να επιτύχει σύνθετες εργασίες εκμάθησης με συγκεντρωτικό τρόπο.

Παράλληλη Εκπαίδευση

Ως λύση στο πρόβλημα αυτό αρχικά προτάθηκε η έννοια της Παράλληλης Εκπαίδευσης, σύμφωνα με την οποία πολλαπλοί εξυπηρετητές, συνδεδεμένοι με κοινούς διαύλους δεδομένων ή σε ένα γρήγορο τοπικό δίκτυο, ανταλλάσσουν βασικές πληροφορίες που έχουν αποκτηθεί κατά τη διάρκεια της εκπαίδευσης με σκοπό τη συνεργατική εκπαίδευση ενός μοντέλου [31]. Πιο συγκεκριμένα, αναπτύχθηκαν δύο διαφορετικοί τρόποι κατανομής της εκπαίδευσης σε πολλαπλές μηχανές/εξυπηρετητές, ο *παραλληλισμός σε επίπεδο δεδομένων* και ο *παραλληλισμός σε επίπεδο μοντέλου*. Οι παραπάνω μέθοδοι μπορούν επίσης να εφαρμοστούν ταυτόχρονα.

Στον **παραλληλισμό σε επίπεδο δεδομένων**, τα δεδομένα διαμερίζονται τόσες φορές όσοι είναι οι κόμβοι εργασίας στο σύστημα και όλοι οι *κόμβοι-εργάτες* (μηχανές) εφαρμόζουν στη συνέχεια τον ίδιο αλγόριθμο σε διαφορετικά σύνολα δεδομένων. Το ίδιο μοντέλο είναι διαθέσιμο σε όλους τους κόμβους-εργάτες, το οποίο όμως ο καθένας εκπαιδεύει με διαφορετικά δεδομένα, έτσι ώστε τελικά, ύστερα από συνένωση των επιμέρους μοντέλων, να προκύπτει ένα καθολικό μοντέλο. Η τεχνική μπορεί να χρησιμοποιηθεί με κάθε αλγόριθμο Μηχανικής Μάθησης υπό την προϋπόθεση ότι τα διαφορετικά σύνολα δεδομένων είναι ανεξάρτητα και προέρχονται από την ίδια κατανομή. Συνεπώς, είναι εύκολα αντιληπτό πως αυτή η μέθοδος, λόγω της παράλληλης επεξεργασίας δεδομένων, επιταχύνει την εκπαίδευση και αποτελεί λύση στο πρόβλημα του αυξημένου όγκου δεδομένων.

Ο **παραλληλισμός σε επίπεδο μοντέλου** επιλύει το πρόβλημα της αυξημένης πολυπλοκότητας των μοντέλων ως εξής: ακριβή αντίγραφα ολόκληρων των συνόλων δεδομένων επεξεργάζονται από τους κόμβους-εργάτες που λειτουργούν σε διαφορετικά τμήματα του μοντέλου. Ο παραλληλισμός σε επίπεδο μοντέλου δεν μπορεί να εφαρμοστεί σε κάθε αλγόριθμο Μηχανικής Μάθησης, επειδή οι παράμετροι του μοντέλου δεν μπορούν πάντα να διαχωριστούν. Μια λύση στο πρόβλημα αυτό είναι να εκπαιδευτούν διαφορετικά 'στιγμιότυπα' του ίδιου ή παρόμοιου μοντέλου και να αθροιστούν οι έξοδοι όλων των εκπαιδευμένων μοντέλων. Επιπλέον, δεδομένου ότι συχνά η είσοδος κάποιων κόμβων εξαρτάται από την έξοδο κάποιων άλλων, η εκπαίδευση πρέπει να γίνει σειριακά [9].



Εικόνα 2.10: *Παραλληλισμός στην εκπαίδευση σε επίπεδο δεδομένων και μοντέλου [9]*

Βαθιά Μάθηση στις συσκευές

Με την πάροδο του χρόνου και την ταχεία εξάπλωση και ανάπτυξη των κινητών συσκευών, εμφανίστηκε η ανάγκη της πραγματοποίησης εργασιών Μηχανικής Μάθησης σε κινητές συσκευές και να μην είναι περιορισμένες πλέον στους εξυπηρετητές. Για παράδειγμα, εφαρμογές όπως η αναγνώριση προσώπου και ομιλίας, βασίζονται στη Μηχανική Μάθηση και εκτελούνται συχνά στις κινητές συσκευές. Για την υποστήριξη αυτών των εφαρμογών, αναπτύχθηκε ο τομέας της Βαθιάς Μάθησης στις συσκευές, στον οποίο συνήθως εκπαιδεύεται πρώτα ένα μοντέλο Μηχανικής Μάθησης πλήρους μεγέθους σε εξυπηρετητές χρησιμοποιώντας μεγάλες ποσότητες δεδομένων, και στη συνέχεια προσαρμόζεται και παραδίδεται στις κινητές συσκευές για να γίνουν τοπικά η συμπερασματολογία και οι προβλέψεις [31].

Κατανεμημένη Εκπαίδευση στα 'άκρα' του δικτύου

Τα τελευταία χρόνια, με την περαιτέρω ανάπτυξη του Κινητού Υπολογισμού και της Τεχνητής Νοημοσύνης στα άκρα του δικτύου, όπως περιγράφηκαν παραπάνω, αναπτύχθηκε η Κατανεμημένη Εκπαίδευση Βαθιών Νευρωνικών Δικτύων στα 'άκρα' του δικτύου. Παρακάτω παρουσιάζονται οι σχετικές αρχιτεκτονικές και μετρικές απόδοσης. Σχηματικά οι κατηγορίες αυτές απεικονίζονται στην Εικόνα 2.12.

Οι βασικές αρχιτεκτονικές Κατανεμημένης Εκπαίδευσης Βαθιών Νευρωνικών Δικτύων εμπορούν να χωριστούν σε τρεις τύπους, τη συγκεντρωτική, την αποκεντρωμένη και την υβριδική.

- **Συγκεντρωτική**

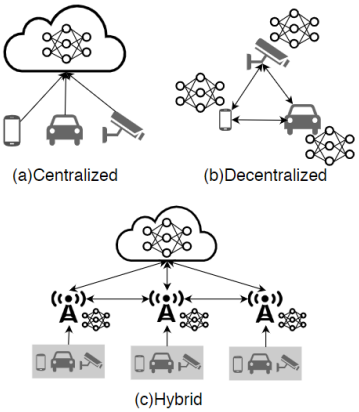
  Το Βαθύ Νευρωνικό Δίκτυο εκπαιδεύεται με δεδομένα που έχουν παραχθεί και συλλεχθεί από κατανεμημένες τελικές συσκευές, όπως είναι οι κινητές συσκευές για παράδειγμα, σε ένα κέντρο δεδομένων στο νέφος. Τα συστήματα που κατατάσσονται σε αυτόν τον τύπο, αντιστοιχούν σε κάποιο από τα επίπεδα ένα έως τρία της Τεχνητής Νοημοσύνης στα άκρα του δικτύου που παρουσιάστηκαν παραπάνω ή στην Τεχνητή Νοημοσύνη στο νέφος.

- **Αποκεντρωμένη**

  Κάθε υπολογιστικός κόμβος εκπαιδεύει τοπικά το δικό του Βαθύ Νευρωνικό Δίκτυο με τα τοπικά διαθέσιμα δεδομένα, διατηρώντας την ιδιωτικότητα τους. Καθένας από αυτούς για να αποκτήσει το ολικό μοντέλο διαμοιράζεται με τους υπόλοιπους κόμβους την τοπική βελτίωση στην εκπαίδευση και τις ενημερώσεις του τοπικού μοντέλου. Δεδομένου ότι όπως είναι αντιληπτό, τα κέντρα δεδομένων στο νέφος δεν είναι απαραίτητα για την εκπαίδευση, ο τύπος αυτός αντιστοιχεί στο επίπεδο πέντε της Τεχνητής Νοημοσύνης στα άκρα του δικτύου.

- **Υβριδική**

  Συνδυάζει τους δύο παραπάνω τύπους, καθώς οι εξυπηρετητές που βρίσκονται στα άκρα του δικτύου μπορούν να εκπαιδεύσουν το Βαθύ Νευρωνικό Δίκτυο είτε με αποκεντρωμένες ενημερώσεις μεταξύ τους είτε με αποκεντρωμένη εκπαίδευση με το κέντρο δεδομένων νέφους. Από τα παραπάνω, συμπεραίνει κανείς πως αντιστοιχεί στα επίπεδα τέσσερα και πέντε της Τεχνητής Νοημοσύνης στα άκρα του δικτύου [10].



Εικόνα 2.11: *Αρχιτεκτονικές Κατανεμημένης Εκπαίδευσης [10]*

Για την αξιολόγηση της εκάστοτε κατανεμημένης μεθόδου εκπαίδευσης, χρησιμοποιούνται οι εξής βασικοί δείκτες επίδοσης: η απόκλιση ανάμεσα στις προβλεπόμενες και τις πραγματικές τιμές του συνόλου δεδομένων εκπαίδευσης, το αν και πόσο γρήγορα μια αποκεντρωμένη μέθοδος συγκλίνει σε σχέση με το αποτέλεσμα της εκπαίδευσης, η ιδιωτικότητα των δεδομένων, το κόστος επικοινωνίας μεταξύ των υπολογιστικών κόμβων, ο χρόνος απόκρισης και η ενεργειακή απόδοση της [10].

Κάποιες χαρακτηριστικές τεχνολογίες για τη βελτιστοποίηση των δεικτών αυτών είναι οι κάτωθι:

- **Συνεργατική Μάθηση**

  Η Συνεργατική Μάθηση είναι ένα είδος Κατανεμημένης Μάθησης που επιτρέπει σε σύνολα εκπαίδευσης και μοντέλα να βρίσκονται σε διαφορετικές, αποκεντρωμένες θέσεις, και η μάθηση μπορεί να συμβεί ανεξάρτητα από το χρόνο και το χώρο. Αυτή η αρχιτεκτονική εκπαίδευσης προτάθηκε για πρώτη φορά από την Google, η οποία επιτρέπει στις κινητές

συσκευές να μαθαίνουν συνεργατικά ένα κοινό μοντέλο με τα τοπικά δεδομένα εκπαίδευσης, αντί να μεταφορτώνουν όλα τα δεδομένα σε ένα κεντρικό εξυπηρετητή νέφους [32].

- **Μεταφορά Γνώσης**

  Στη Μεταφορά Γνώσης, τα χαρακτηριστικά που μαθαίνονται σε προηγούμενα μοντέλα μπορούν να χρησιμοποιηθούν από άλλα μοντέλα, μειώνοντας σημαντικά τον χρόνο εκπαίδευσης. Οι Valery et al. πρότειναν τη μεταφορά χαρακτηριστικών από το εκπαιδευμένο μοντέλο σε τοπικά μοντέλα, τα οποία θα επανεκπαιδευτούν με τα τοπικά δεδομένα εκπαίδευσης [32].

- **Διαχωρισμός Βαθέως Νευρωνικού Δικτύου**

  Ο στόχος του Διαχωρισμού Βαθιών Νευρωνικών Δικτύων είναι η προστασία της ιδιωτικότητας. Συγκεκριμένα, η ιδιωτικότητα του χρήστη προστατεύεται με τη μετάδοση μερικώς επεξεργασμένων δεδομένων αντί για τη μετάδοση των αρχικών ανεπεξέργαστων. Για να καταστεί δυνατή η εκπαίδευση των μοντέλων στα άκρα του δικτύου με διατήρηση της ιδιωτικότητας, ο διαχωρισμός Βαθιών Νευρωνιών Δικτύων πραγματοποιείται μεταξύ των τελικών συσκευών και του εξυπηρετητή στα άκρα. Αυτό βασίζεται στο γεγονός ότι ένα μοντέλο Βαθέως Νευρωνικού Δικτύου μπορεί να χωριστεί εσωτερικά μεταξύ δύο διαδοχικών στρωμάτων με δύο τμήματα εγκατεστημένα σε διαφορετικές τοποθεσίες, χωρίς μείωση της ακρίβειας [10].

2.4.3 Κατανεμημένη Συμπερασματολογία

Στα πλαίσια της Βαθιάς Μάθησης στις συσκευές, καθώς αρχικά οι πόροι τους ήταν αρκετά περιορισμένοι, δεν γινόταν η εκπαίδευση παρά μόνο η συμπερασματολογία στις κινητές συσκευές.

Συμπερασματολογία σε κινητές συσκευές

Τα πλεονεκτήματα της Κατανεμημένης Συμπερασματολογίας στην κινητή συσκευή είναι πολλά, δεδομένου ότι κατά τη διάρκεια της διαδικασίας εξαγωγής συμπερασμάτων, η κινητή συσκευή δεν επικοινωνεί με κάποιον εξυπηρετητή. Πρώτον, εξοικονομείται εύρος ζώνης επικοινωνίας, καθώς όσο περισσότεροι υπολογισμοί γίνονται στην κινητή συσκευή, τόσο λιγότερα δεδομένα χρειάζεται να αποστέλλονται στον εξυπηρετητή. Επιπλέον, μειώνεται ο χρόνος απόκρισης, αφού εάν όλοι οι υπολογισμοί μπορούν να εκτελεστούν τοπικά, δεν θα υπάρχει επιβάρυνση στο χρόνο επικοινωνίας ή οποιεσδήποτε ανησυχίες για την αξιοπιστία του εξυπηρετητή. Τέλος, με αυτόν τον τρόπο τα ευαίσθητα προσωπικά δεδομένα διατηρούνται τοπικά στη συσκευή, βελτιώνοντας έτσι σημαντικά την ιδιωτικότητα των δεδομένων του χρήστη [33].

Όμως, με αυτόν τον τρόπο αυξάνονται πολύ οι απαιτήσεις πόρων στις μονάδες επεξεργασίας, στη μνήμη και στη μπαταρία στην κινητή συσκευή και επομένως η απόδοση εξαρτάται από την ίδια τη συσκευή [10].

Συμπερασματολογία στα 'άκρα' του δικτύου

Η Κατανεμημένη Συμπερασματολογία στις κινητές συσκευές όμως δεν είναι παρά μια υποκατηγορία της Κατανεμημένης Συμπερασματολογίας στα άκρα του δικτύου. Σε αυτή, ένα εκπαιδευμένο μοντέλο χρησιμοποιείται για την συμπερασματολογία ενός αγνώστου δείγματος μέσω της προώθησης του υπολογισμού της εξόδου σε συσκευές και εξυπηρετητές στα άκρα του δικτύου [32].

Εκτός από τις πιο συνηθισμένες αρχιτεκτονικές συμπερασματολογίας στο νέφος και με συνεργασία νέφους και συσκευής, ορίζονται κάποιες ακόμα σημαντικές αρχιτεκτονικές για συμπερασματολογία στα άκρα του δικτύου οι οποίες κατηγοριοποιούνται στις εξής κατηγορίες:

- **Βασισμένες στα άκρα του δικτύου**

  Η κινητή συσκευή λαμβάνει τα δεδομένα εισόδου και τα στέλνει στον εξυπηρετητή άκρου. Στη συνέχεια, η συμπερασματολογία του μοντέλου εκτελείται στον εξυπηρετητή αυτό και τα αποτελέσματα της πρόβλεψης στέλνονται πίσω στη συσκευή.

- **Βασισμένες στη συσκευή**

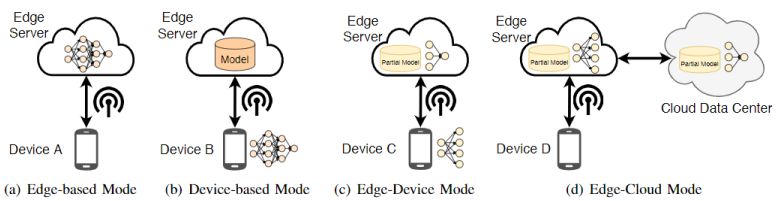
  Η κινητή συσκευή λαμβάνει το μοντέλο από τον εξυπηρετητή στο άκρο του δικτύου και πραγματοποιεί τοπικά την συμπερασματολογία. Κατά την εκτέλεση της συμπερασματολογίας, η συσκευή δεν επικοινωνεί με τον εξυπηρετητή.

- **Βασισμένες στη συνεργασία της συσκευής με τα άκρα του δικτύου**

  Η κινητή συσκευή πρώτα χωρίζει το μοντέλο σε διαφορετικά κομμάτια με βάση κάποια από τα τρέχοντα χαρακτηριστικά του συστήματος, όπως το εύρος ζώνης του δικτύου, οι πόροι της συσκευής και ο φόρτος εργασίας του εξυπηρετητή άκρου. Ύστερα, η συσκευή εκτελεί το μοντέλο μέχρι ένα συγκεκριμένο στρώμα και στέλνει τα ενδιάμεσα δεδομένα στον εξυπηρετητή άκρου. Ο εξυπηρετητής εκτελεί τα υπόλοιπα στρώματα και στέλνει τις προβλέψεις στη συσκευή.

- **Βασισμένες στη συνεργασία του νέφους με τα άκρα του δικτύου**

  Η συσκευή είναι υπεύθυνη για τη συλλογή των δεδομένων εισόδου και το μοντέλο εκτελείται μέσω συνεργασίας του νέφους με τα άκρα του δικτύου [10].



Εικόνα 2.12: *Αρχιτεκτονικές Κατανεμημένης Συμπερασματολογίας [10]*

Για την αξιολόγηση της εκάστοτε κατανεμημένης μεθόδου συμπερασματολογίας, χρησιμοποιούνται οι εξής βασικοί δείκτες επίδοσης: ο χρόνος απόκρισης, η ακρίβεια, η κατανάλωση ενέργειας, η ιδιωτικότητα, η επιβάρυνση λόγω επικοινωνίας και το αποτύπωμα μνήμης [10].

Κάποιες χαρακτηριστικές τεχνολογίες για τη βελτιστοποίηση των δεικτών αυτών είναι οι κάτωθι:

- **Συμπίεση Μοντέλου**

  Για την εφαρμογή των μοντέλων σε συσκευές στα άκρα του δικτύου χωρίς παροχή ρεύματος, έχουν γίνει σημαντικές προσπάθειες για τη συμπίεση των μοντέλων. Η συμπίεση μοντέλων αποσκοπεί να ελαφρύνει το μοντέλο, να βελτιώσει την ενεργειακή απόδοση και να επιταχύνει τη συμπερασματολογία σε συσκευές στα άκρα του δικτύου με περιορισμένους πόρους, χωρίς μείωση της ακρίβειας [32].

  Έχουν προταθεί ποικίλες τεχνικές συμπίεσης Βαθιών Νευρωνικών Δικτύων, όπως το Κλάδεμα Βαρών, η Κβαντοποίηση και ο σχεδιασμός συμπαγούς αρχιτεκτονικής [10].

- **Διαμοιρασμός Μοντέλου**

  Ο Διαμοιρασμός Μοντέλου χρησιμοποιείται κυρίως για τα ζητήματα της καθυστέρησης, της ενέργειας και της ιδιωτικότητας και χωρίζεται σε δύο τύπους, τον διαμοιρασμό ανάμεσα σε εξυπηρετητή και συσκευή και τον διαμοιρασμό μεταξύ συσκευών.

- **Πρόωρη Έξοδος**

  Για την επιτάχυνση της συμπερασματολογίας του μοντέλου, η μέθοδος πρόωρης εξόδου του μοντέλου αξιοποιεί τα δεδομένα εξόδου ενός πρώιμου στρώματος για να λάβει το αποτέλεσμα ταξινόμησης, πράγμα που σημαίνει ότι η διαδικασία συμπερασματολογίας ολοκληρώνεται με τη χρήση μερικού μοντέλου Βαθέως Νευρωνικού Δικτύου. Συνεπώς, ο στόχος της Πρόωρης Εξόδου είναι η μείωση του χρόνου απόκρισης [10].

2.4.4 Κατανεμημένα Συστήματα Συμπερασματολογίας Νευρωνικών Δικτύων

Όπως παρουσιάστηκε στην παραπάνω υποενότητα, ο συνήθης τρόπος Κατανεμημένης Συμπερασματολογίας βασισμένης στη συνεργασία της συσκευής με τα άκρα του δικτύου, είναι ο Διαμοιρασμός Μοντέλου, στον οποίο η συσκευή εκτελεί το μοντέλο μέχρι ένα συγκεκριμένο στρώμα και στέλνει τα ενδιάμεσα δεδομένα στον εξυπηρετητή, στον οποίο εκτελούνται στη συνέχεια τα υπόλοιπα στρώματα του μοντέλου.

Έχει παρατηρηθεί όμως πως, σε γενικές γραμμές, τα μικρά μοντέλα δυσκολεύονται να χειριστούν τις ακραίες περιπτώσεις, με την πλειονότητα των 'εύκολων' δειγμάτων να κατηγοριοποιείται σωστά. Κατά συνέπεια, με βάση την παρατήρηση αυτή, είναι δυνατόν να χρησιμοποιηθεί μια αρχιτεκτονική συνεργασίας κινητής συσκευής και εξυπηρετητή που επεξεργάζεται τις 'εύκολες' εισόδους με ένα σχετικά μικρό Βαθύ Νευρωνικό Δίκτυο στη συσκευή και μεταφέρει τις 'δύσκολες' εισόδους στο ισχυρό και βαθύτερο Νευρωνικό Δίκτυο στον εξυπηρετητή. Μια τέτοια αρχιτεκτονική διευκολύνει την επίτευξη ενός καλού συμβιβασμού ενέργειας-χρόνου-ακρίβειας για την συμπερασματολογία με Βαθιά Νευρωνικά Δίκτυα σε σύγκριση με μονομερείς λύσεις, που κάνουν χρήση μόνο της κινητής συσκευής ή μόνο του εξυπηρετητή [34].

Αυτά τα συστήματα συμπερασματολογίας αποτελούν μια ευρέως μελετημένη σχεδιαστική προσέγγιση στη Μηχανική Μάθηση. Στόχος τους είναι η ελαχιστοποίηση του χρόνου υπολογισμού ανά ταξινόμηση εκμεταλλευόμενα την ιδιότητα ότι διαφορετικές είσοδοι απαιτούν διαφορετικό ποσό υπολογισμού για να επιτευχθεί μια αξιόπιστη πρόβλεψη. Μια τέτοια δομή σχηματίζεται συνήθως ως αρχιτεκτονική πολλαπλών σταδίων, με τη σύνδεση ταξινομητών αυξανόμενης πολυπλοκότητας. Σε κάθε στάδιο, με βάση την εμπιστοσύνη της πρόβλεψής του, το σύστημα μπορεί είτε να αποφασίσει να 'δεχτεί' την συμπερασματολογία του δείγματος εισόδου από τον τρέχοντα ταξινομητή και να τερματίσει την εκτέλεση ή να το περάσει στο επόμενο στάδιο, δηλαδή στον επόμενη ταξινομητή. Ο τύπος αυτός Κατανεμημένου Συστήματος Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων ονομάζεται *cascade* και είναι και αυτός που μελετάται στην παρούσα εργασία [35].

Κεφάλαιο 3

Τεχνολογικά Εργαλεία

Στο κεφάλαιο αυτό παρουσιάζονται τα βασικά εργαλεία και οι τεχνολογίες που χρησιμοποιήθηκαν για την εκπόνηση της παρούσας διπλωματικής εργασίας.

3.1 Python

Η Python αναπτύχθηκε το 1989 από τον Guido van Rossum στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) και πήρε το όνομά της από το "Monty Python's Flying Circus".

Πρόκειται για μια εύκολη στην εκμάθηση, ισχυρή γλώσσα προγραμματισμού. Διαθέτει αποδοτικές δομές δεδομένων υψηλού επιπέδου και μια απλή, αλλά αποτελεσματική, προσέγγιση για τον αντικειμενοστραφή προγραμματισμό, ενώ μεταξύ άλλων υποστηρίζει επίσης διαδικαστικό και συναρτησιακό προγραμματισμό. Το κομψό συντακτικό και το ότι πρόκειται για μια διαδραστική γλώσσα διερμηνέα που ενσωματώνει δυναμική τυποποίηση, δυναμικούς τύπους δεδομένων υψηλού επιπέδου, κλάσεις, δομοστοιχεία κι εξαιρέσεις, την καθιστούν ιδανική γλώσσα για την ταχεία συγγραφή κώδικα κι ανάπτυξη εφαρμογών σε πολλούς τομείς στις περισσότερες πλατφόρμες [36][37].

Σε σύγκριση με άλλες γλώσσες προγραμματισμού, όπως η Java και η C++, η Python είναι ιδιαίτερα χρήσιμη για την εκμάθηση προγραμματισμού υπολογιστών. Μεταξύ των τριών, η C++ εφευρέθηκε νωρίτερα. Οι εντολές της ήταν οι πιο κοντινές στις εντολές υλικού. Ως αποτέλεσμα, τα προγράμματα της C++ εκτελούνται πολύ γρήγορα και μπορούν να βελτιστοποιηθούν με πολλούς τρόπους. Ωστόσο, η εκμάθηση της C++ απαιτεί γνώσεις στο υλικό των υπολογιστών. Η Java εφευρέθηκε δεκαετίες μετά τη C++. Προσφέρει υψηλότερα επίπεδα αφαιρετικότητας σε σύγκριση με τη C++, καθιστώντας ευκολότερο τον προγραμματισμό. Συνεπώς, τα προγράμματα Java εκτελούνται πιο αργά από τη C++. Η Python κάνει ένα βήμα παραπέρα στην αφαιρετικότητα σε σχέση με λεπτομέρειες που αφορούν το υλικό, προσφέροντας ένα εννοιολογικά πολύ απλό σύστημα εκτέλεσης εντολών. Ως αποτέλεσμα, η Python είναι η πιο εύκολη στην εκμάθηση, αλλά τα προγράμματα Python είναι συνήθως τα πιο αργά στην εκτέλεσή τους. Όσον αφορά τις λειτουργικότητες, η Python είναι εξίσου ισχυρή με τη Java και τη C++. Υποστηρίζει όλες τις λειτουργίες που υποστηρίζει το σύνηθες υλικό του υπολογιστή. Επιπλέον, οι βιβλιοθήκες που παρέχονται από την Python και όσους συνεισφέρουν στον ανοιχτό κώδικα της εμπλουτίζουν την ισχύ της, καθιστώντας την την πιο

βολική επιλογή για πολλές εφαρμογές, όπως η επεξεργασία κειμένου. Οι βιβλιοθήκες της, όπως οι numpy και scipy, υλοποιούνται σε C++ σε χαμηλότερο επίπεδο, προσφέροντας πολύ γρήγορους τρόπους εκτέλεσης επιστημονικών υπολογισμών με τη χρήση της Python [38].

3.2 Kaggle

Το Kaggle, θυγατρική της Google LLC, είναι μια διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών στη Μηχανική Μάθηση που προσφέρει τη δυνατότητα σε χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να γράφουν και να εκτελούν κώδικα Python στο νέφος σε ένα περιβάλλον βασισμένο στο Jupyter ή ακόμα και να βρίσκουν προγράμματα άλλων χρηστών με τους οποίους μπορούν επίσης να συνεργαστούν και να συμμετέχουν σε διαγωνισμούς για την επίλυση προκλήσεων της επιστήμης δεδομένων. Επιπλέον, δεν απαιτεί εγκατάσταση, και παρέχει ελεύθερη πρόσβαση σε Μονάδες Γραφικής Επεξεργασίας (GPU) και Μονάδες Επεξεργασίας Τανυστών (TPU) [39].

3.3 TensorFlow

Το TensorFlow είναι μια ολοκληρωμένη πλατφόρμα ανοικτού κώδικα για Μηχανική Μάθηση. Διαθέτει ένα ολοκληρωμένο, ευέλικτο οικοσύστημα εργαλείων, βιβλιοθηκών και πόρων της κοινότητας που επιτρέπει στους ερευνητές να προωθήσουν την τελευταία λέξη της τεχνολογίας στην Μηχανική Μάθηση και στους προγραμματιστές να δημιουργήσουν και να αναπτύξουν εύκολα εφαρμογές που υποστηρίζονται από την Μηχανική Μάθηση. Αναπτύχθηκε αρχικά από ερευνητές και μηχανικούς που εργάζονταν στην ομάδα Google Brain στο πλαίσιο του οργανισμού Machine Intelligence Research της Google για τη διεξαγωγή ερευνών Μηχανικής Μάθησης και Βαθιών Νευρωνικών Δικτύων. Παρέχει σταθερές διεπαφές σε Python και C++ [40]. Τέλος, προσφέρει πρόσβαση σε ένα αποθετήριο εκπαιδευμένων μοντέλων Μηχανικής Μάθησης, το TensorFlow Hub [41].

3.4 Keras

Το Keras είναι μια διεπαφή γραμμένη σε Python για την πλατφόρμα Μηχανικής Μάθησης TensorFlow με σκοπό την ανάπτυξη λογισμικού με χρήση Βαθιών Νευρωνικών Δικτύων και αναπτύχθηκε αρχικά στο πλαίσιο της ερευνητικής προσπάθειας του έργου ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System). Διαθέτει ένα εύρος υλοποιημένων εργαλείων για Βαθιά Μάθηση, όπως στρώματα, αλγορίθμους βελτιστοποίησης, μετρικές, συναρτήσεις απώλειας και εργαλεία για την προεπεξεργασία διαφόρων τύπων συνόλων δεδομένων, ενώ παρέχει πρόσβαση επίσης ακόμα και σε υλοποιημένα μοντέλα Συνελικτικών Νευρωνικών Δικτύων [42].

3.5 ImageNet

Το ImageNet είναι ένα σύνολο δεδομένων εικόνων οργανωμένο σύμφωνα με την ιεραρχία του WordNet. Κάθε έννοια με νόημα στο WordNet, που ενδεχομένως περιγράφεται από

πολλαπλές λέξεις ή φράσεις λέξεων, ονομάζεται 'σύνολο συνωνύμων'. Υπάρχουν περισσότερα από 100.000 σύνολα συνωνύμων στο WordNet με την πλειονότητά τους να είναι ουσιαστικά (80.000+). Στο ImageNet, παρέχονται κατά μέσο όρο 1000 εικόνες για την απεικόνιση κάθε τέτοιου συνόλου. Οι εικόνες κάθε έννοιας περνάνε από ανθρώπινο ποιοτικό έλεγχο και τους προστίθενται ετικέτες. Η ιδέα του ImageNet βασίστηκε σε δύο σημαντικές ανάγκες της έρευνας στην Όραση Υπολογιστών, πρώτον την αυξανόμενη ζήτηση για ένα συγκριτικό σημείο αναφοράς για την κατηγοριοποίηση αντικειμένων υψηλής ποιότητας με σαφώς καθορισμένες μετρικές αξιολόγησης και δεύτερον την ανάγκη για περισσότερα δεδομένα ώστε να καταστεί δυνατή η εφαρμογή πιο γενικευμένων μεθόδων Μηχανικής Μάθησης. Τέλος, να σημειωθεί πως το σύνολο επικύρωσης του ImageNet, που είναι και αυτό που θα χρησιμοποιηθεί, περιέχει συνολικά 50000 εικόνες [43].

Κεφάλαιο 4

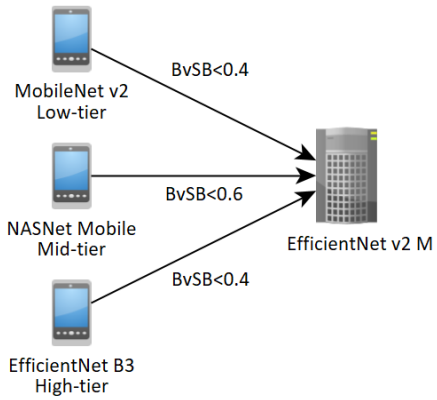
Μοντελοποίηση Συστήματος

Στο κεφάλαιο αυτό παρουσιάζεται η μοντελοποίηση του Κατανεμημένου Συστήματος Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων, τύπου Cascade.

4.1 Περιγραφή Συστήματος

Στην εργασία αυτή με βάση τα κίνητρα που περιγράφηκαν στην Υποενότητα 2.4.4, αναπτύσσεται ένα Κατανεμημένο Σύστημα Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων. Το σύστημα απαρτίζεται από (α) κινητές συσκευές, κάθε μία από τις οποίες μπορεί να έχει διαφορετικά χαρακτηριστικά και διαφορετικό μοντέλο με το οποίο εκτελεί τοπικά την συμπερασματολογία και (β) έναν κεντρικό εξυπηρετητή, στον οποίον εκτελείται εκ νέου συμπερασματολογία για τις εικόνες στις οποίες δεν έχει εξασφαλιστεί το κατώτατο όριο εμπιστοσύνης που επιθυμούμε από το τοπικό μοντέλο.

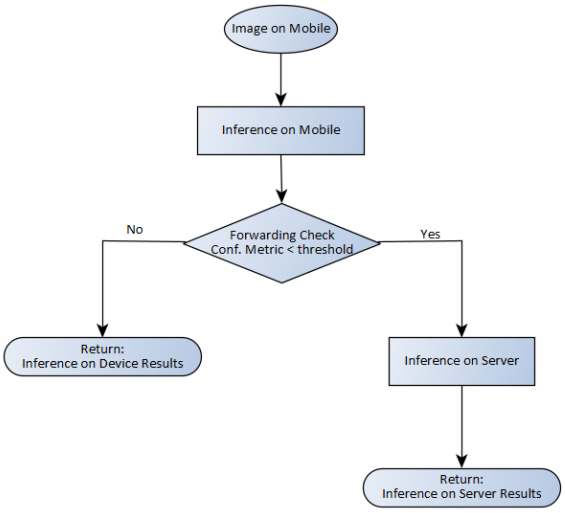
Το σύστημα αυτό είναι παραμετροποιήσιμο σε σχέση με το πλήθος, τον τύπο και το μοντέλο των συσκευών, το μοντέλο του εξυπηρετητή, το μέγεθος δέσμης προς συμπερασματολογία στον εξυπηρετητή και το κατώφλι προώθησης του εκάστοτε δείγματος στον εξυπηρετητή. Μια εικόνα ενός τέτοιου συστήματος φαίνεται στην Εικόνα 4.1 κι ο τρόπος λειτουργίας του στην Εικόνα 4.2. Είσοδος στο σύστημα είναι οι εικόνες προς συμπερασματολογία και έξοδός του ο χρόνος που χρειάστηκε μέχρι να εξέλθει η κάθε εικόνα από το σύστημα και η ακρίβειά του.



Εικόνα 4.1: *Σύστημα τριών συσκευών*

Επιπλέον, προς διευκόλυνση της μοντελοποίησης και μελέτης του συστήματος υποθέτουμε ότι οι εικόνες προς συμπερασματολογία είναι αποθηκευμένες στη συσκευή και συνεπώς δεν υπήρχε αναμονή για τη συλλογή τους.

Η επεξεργασία δειγμάτων εισόδου σε δέσμες, δηλαδή η παράλληλη εκτέλεση συμπερασματολογίας για παραπάνω από ένα δείγματα, αποτελεί μια ευρέως διαδεδομένη τεχνική στις βιβλιοθήκες Μηχανικής Μάθησης για την αύξηση της απόδοσης του συστήματος, καθώς αξιοποιεί καλύτερα τη δυνατότητα παράλληλου υπολογισμού στις Μονάδες Επεξεργασίας [44].



Εικόνα 4.2: *Τρόπος λειτουργίας συστήματος*

Σε σχέση με την έννοια της *εμπιστοσύνης* που αναφέρθηκε νωρίτερα, στα σύγχρονα συστήματα λήψης αποφάσεων, είναι σημαντικό τα μοντέλα ταξινόμησης όχι μόνο να είναι ακριβή, αλλά και να είναι σε θέση να υποδείξουν πότε είναι πιθανό να είναι λανθασμένα, τόσο για λόγους ασφάλειας όσο και για την ερμηνευσιμότητα των αποτελεσμάτων. Παραδείγματος χάριν, σε σχέση με την ασφάλεια, στην αυτοματοποιημένη ιατρική περίθαλψη, σε περίπτωση που το μοντέλο δεν είναι σίγουρο για τη διάγνωση μιας ασθένειας, θα πρέπει να ελεγχθεί από ιατρικό προσωπικό. Όσο για την ερμηνευσιμότητα, δεδομένου ότι οι άνθρωποι έχουν μια φυσική γνωστική αντίληψη για τις πιθανότητες, καλές εκτιμήσεις της εμπιστοσύνης, δηλαδή εκτιμήσεις που να αντιστοιχούν στην πραγματική πιθανότητα μιας κλάσης, συμβάλλουν στο να θεωρείται το μοντέλο αξιόπιστο από τους χρήστες. Εν ολίγοις, ένα μοντέλο πρέπει να είναι σε θέση να παρέχει, εκτός από την ίδια την πρόβλεψη, ένα βαθμονομημένο μέτρο εμπιστοσύνης γι΄ αυτή, δηλαδή θα πρέπει η πιθανότητα που σχετίζεται με την προβλεπόμενη ετικέτα κλάσης να ανταναχλά την πραγματική πιθανότητα ορθότητας της πρόβλεψης. Δυστυχώς όμως, με την αύξηση της πολυπλοκότητας τους, τα σύγχρονα μοντέλα παρόλο που έχουν αυξημένη ακρίβεια, δεν είναι πλέον καλά βαθμονομημένα [45].

Αυτή η παρατήρηση όμως δυσκολεύει την ανάπτυξη του συστήματος καθώς πρέπει να βρεθεί ένας αποτελεσματικός τρόπος ώστε να μπορεί να κρίνει πότε το τοπικό μοντέλο της κάθε συσκευής δεν έχει εμπιστοσύνη στην πρόβλεψή του, δεδομένου ότι για τυχαίες εικόνες των συσκευών δεν μπορούμε να γνωρίζουμε εκ των προτέρων την πραγματική πιθανότητα μιας

κλάσης. Γι΄ αυτό προσπαθούμε να εκτιμήσουμε την εμπιστοσύνη, προσδίδοντας στην έξοδο του μοντέλου ιδιότητες κατανομής πιθανότητας και στη συνέχεια μέσω μετρικών να αξιολογήσουμε την πραγματική ακρίβεια του μοντέλου αλγοριθμικά με βάση την μετασχηματισμένη κατανομή της εξόδου του.

Για τον μετασχηματισμό της εξόδου, προσθέτουμε στο μοντέλο ένα τελευταίο στρώμα Softmax που ουσιαστικά περνάει την προηγούμενη έξοδο του μοντέλου από τη συνάρτηση ενεργοποίησης Softmax. Η τιμή της συνάρτησης Softmax που αντιστοιχεί στην κλάση i για την έξοδο ενός μοντέλου ορίζεται ως εξής:

$$softmax(\mathbf{x})_i = \frac{exp(x_i)}{\sum_j^k exp(x_j)}$$

όπου k το πλήθος των κλάσεων και $\mathbf{x}$ το διάνυσμα εξόδου του μοντέλου. Η συνάρτηση αυτή μετασχηματίζει το διάνυσμα εξόδου ώστε πλέον να αντιστοιχεί στην εμπιστοσύνη του μοντέλου για την κάθε κλάση, έχοντας δηλαδή πλέον ιδιότητες μέτρου πιθανότητας. Πιο συγκεκριμένα, όλα του τα στοιχεία να είναι μη αρνητικά και να αθροίζουν στη μονάδα, όπως ήταν επιθυμητό.

Με σκοπό την επιλογή του κατάλληλου κριτηρίου εμπιστοσύνης με βάση το οποίο οι εικόνες θα προωθούνται στον εξυπηρετητή, έγινε σύγκριση μεταξύ τεσσάρων μετρικών εμπιστοσύνης ως προς την απόδοση του συστήματος, για την αλγοριθμική προσέγγιση της πραγματικής ακρίβειας του μοντέλου. Οι μετρικές που συγκρίθηκαν είναι οι εξής:

1. **Μέγιστη Εμπιστοσύνη**

$$\text{Max conf.} = \mathbf{p}^{(1)}$$

όπου $\mathbf{p} = softmax(\mathbf{x})$ και $\mathbf{p}^{(1)} = max(\mathbf{p})$.

2. **Εντροπία**

$$H = -\sum_{j=1}^{k} \mathbf{p}_j \cdot \log \mathbf{p}_j$$

όπου k το πλήθος των κλάσεων και $\mathbf{p}_j$ η εμπιστοσύνη που δίνει το μοντέλο στην κλάση j. Η εντροπία είναι μέτρο αβεβαιότητας μιας τυχαίας μεταβλητής αφού μεγάλες τιμές εντροπίας υποδεικνύουν μεγαλύτερη αβεβαιότητα στην κατανομή. Συνεπώς, αν μια έξοδος έχει υψηλή εντροπία, ο ταξινομητής δεν είναι βέβαιος για την επιλογή του. Η εντροπία παίρνει τιμές που ανήκουν στο διάστημα $[0, \log k]$, όσο πιο 'μυτερή' είναι η κατανομή τείνει στο 0 ενώ όσο πλησιάζει στην ομοιόμορφη τείνει στο $\log k$, και με βάση αυτό στη συνέχεια την κανονικοποιούμε ώστε να εξετάζουμε όλες τις μετρικές με όρια εμπιστοσύνης που ανήκουν στο $[0, 1]$.

Γενικά, αν το μοντέλο είναι καλά βαθμονομημένο και η εμπιστοσύνη του στις κλάσεις στην έξοδο είναι κοντά στην πραγματική πιθανότητα, επιθυμούμε η κατανομή εξόδου του να είναι 'μυτερή' γιατί και διαισθητικά μπορούμε να καταλάβουμε ότι όσο περισσότερο 'ξεχωρίζει' η μέγιστη εμπιστοσύνη σε σχέση με την εμπιστοσύνη σε άλλες κλάσεις, τόσο πιο βέβαιο είναι το μοντέλο ότι το δείγμα ανήκει σε αυτή έναντι των άλλων.

3. **Διαφορά πρώτης και δεύτερης Εμπιστοσύνης**

$$\text{BvSB} = \mathbf{p}^{(1)} - \mathbf{p}^{(2)}$$

όπου $\mathbf{p}^{(2)}$ η δεύτερη μεγαλύτερη εμπιστοσύνη του μοντέλου. Το μεγάλο πλεονέκτημα αυτής της μετρικής έναντι της εντροπίας είναι ότι δεν επηρεάζεται από τις τιμές εξόδου που δίνει σε μη πιθανές κλάσεις το μοντέλο.

4. **Δείκτης Gini**

$$G = 1 - \sum_{j=1}^{k} \mathbf{p}_j^2$$

Η συγκεκριμένη μετρική, όπως και η εντροπία, υποδεικνύουν πόσο 'μυτερή' ή διάχυτη είναι η κατανομή. Παίρνει τιμές που ανήκουν στο διάστημα $[0, 1]$ και όταν η κατανομή τείνει στην ομοιόμορφη, ο δείκτης Gini τείνει στη μονάδα [46].

4.2 Μέθοδος

Αρχικά εξετάστηκε πειραματικά πώς επηρεάζεται ο χρόνος απόκρισης και ο ρυθμός διέλευσης των εικόνων για τα διάφορα μοντέλα στον εξυπηρετητή για διαφορετικά μεγέθη δέσμων. Στη συνέχεια, μοντελοποιήθηκε ο χρόνος απόκρισης σε μια απλή εκδοχή του συστήματος που περιγράφηκε παραπάνω, για τον υπολογισμό των παραμέτρων της οποίας χρειάστηκε να ελεγχθεί το κριτήριο προώθησης του εκάστοτε δείγματος στον εξυπηρετητή. Τέλος, έγινε μια πειραματική προσομοίωση του συστήματος για διάφορες τιμές των παραμέτρων του.

4.2.1 Προδιαγραφές Υλικού

Όλοι οι πειραματικοί έλεγχοι και δοκιμές έγιναν στο Kaggle (Ενότητα 3.2) κάνοντας χρήση τόσο της διαθέσιμης CPU όσο και της GPU οι τύποι και τα χαρακτηριστικά των οποίων παρουσιάζονται στον Πίνακα 4.1 παρακάτω.

Σε σχέση με τις κινητές συσκευές, χρησιμοποιήθηκαν τρεις κατηγορίες-παραδείγματα συσκευών, υψηλού, μέσου και χαμηλού επιπέδου απόδοσης GPU αντίστοιχα. Η απόδοσή τους φαίνεται στον Πίνακα 4.2.

Πίνακας 4.1: *Μονάδες επεξεργασίας του Kaggle*

Μονάδα Επεξεργασίας	**Μοντέλο**	**Απόδοση FP32** (TFLOPS)
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz	1.12
GPU	NVIDIA Tesla P100-PCIE	9.3

Πίνακας 4.2: *Απόδοση συσκευών ανά τύπο*

Κατηγορία	**Παράδειγμα Μοντέλου**	**Απόδοση** (TFLOPS)
High-Tier	Adreno 660 on Qualcomm (QC) Snapdragon (SDM) 888	1.72
Mid-Tier	Adreno 630 on QC SDM845	0.727
Low-Tier	Adreno 619 on QC SDM480	0.465

4.2.2 Χαρακτηριστικά Μοντέλων

Παρακάτω παρουσιάζονται τα μοντέλα για τα οποία έγιναν μετρήσεις μαζί με τα χαρακτηριστικά τους, σύμφωνα με τις μετρήσεις που παρουσιάζονται στην ιστοσελίδα του Kerasapplications [42]. Πιο συγκεκριμένα, τα MAC (multiply-accumulate operations), δηλαδή οι πράξεις πολλαπλασιασμού και συσσώρευσης, που απαιτούνται για την εξαγωγή συμπερασματολογίας, οι παράμετροι, οι διαστάσεις της εικόνας που δέχεται το εκάστοτε μοντέλο ως είσοδο, η top-1 ακρίβεια τους στο σύνολο επικύρωσης του ImageNet, δηλαδή το ποσοστό των εικόνων στις οποίες η κλάση στην οποία το κάθε μοντέλο είχε μέγιστη εμπιστοσύνη ταυτίζονταν με την πραγματική κλάση στην οποία ανήκε και η top-5 ακρίβεια του, δηλαδή το ποσοστό των εικόνων στις οποίες η πραγματική κλάση ανήκε στις 5 κλάσεις με τη μεγαλύτερη εμπιστοσύνη.

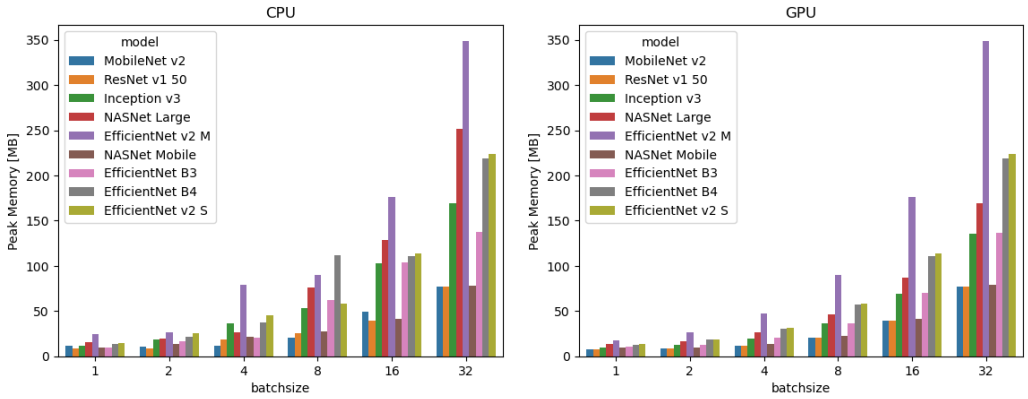
Πίνακας 4.3: *Μοντέλα και τα χαρακτηριστικά τους*

Μοντέλο	**MAC (B)**	**Parameters (M)**	**Input Size**	**Top 1 Acc.**	**Top 5 Acc.**
MobileNet v2	0.3	3.5	224 x 224	71.3	90.1
NASNet Mobile	0.564	5.3	224 x 224	74.4	91.9
EfficientNet B3	1.8	12.3	300 x 300	81.6	95.7
ResNet v1 50	3.8	25.6	224 x 224	74.9	92.1
EfficientNet B4	4.2	19.5	380 x 380	82.9	96.4
Inception v3	5	23.9	299 x 299	77.9	93.7
EfficientNet v2 S	8.4	21.6	384 x 384	83.9	96.7
NASNet Large	23.8	88.9	331 x 331	82.5	96.0
EfficientNet v2 M	24.7	54.4	480 x 480	85.3	97.4

4.3 Χαρακτηρισμός Εξυπηρετητή

Για τον χαρακτηρισμό του εξυπηρετητή έγιναν μετρήσεις για τα διάφορα μοντέλα στο Kaggle με και χωρίς χρήση της διαθέσιμης GPU σε 300 τυχαία παραγμένες εικόνες. Πιο συγκεκριμένα, μετρήθηκε η μέγιστη μνήμη που χρησιμοποιήθηκε, ο μέσος και ο συνολικός χρόνος απόκρισης και ο ρυθμός διέλευσης εικόνων από το σύστημα για μεγέθη δέσμης ίσα με 2,4,8,16 και 32. Για κάθε μοντέλο παρουσιάζεται πρώτα ο χαρακτηρισμός του εξυπηρετητή με χρήση της CPU και ύστερα με χρήση της GPU.

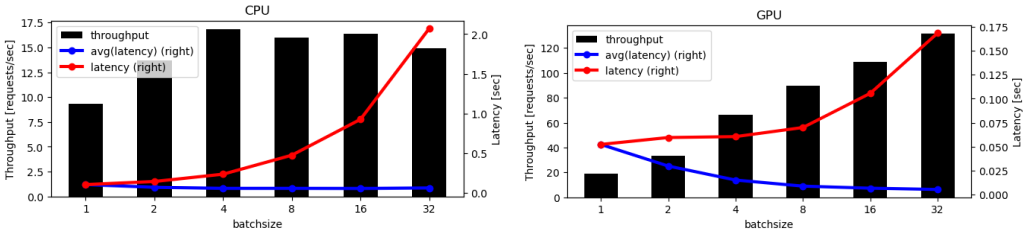
Αρχικά παρουσιάζεται η μέγιστη χρήση μνήμης κατά την εκτέλεση της συμπερασματολογίας για τα διαφορετικά μεγέθη δέσμης για κάθε μοντέλο.



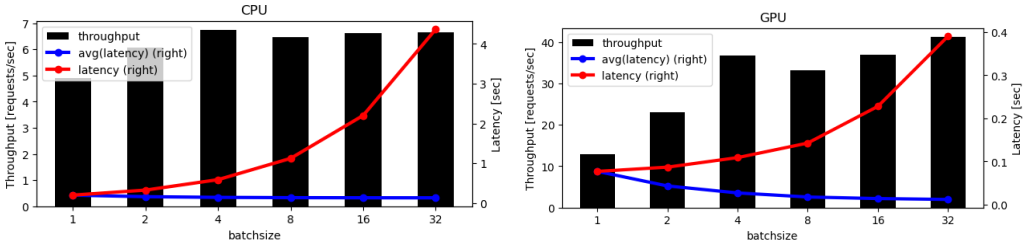
Εικόνα 4.3: *Μέγιστη χρήση μνήμης κατά την εκτέλεση της συμπερασματολογίας*

Παρατηρούμε ότι δεδομένης της παράλληλης επεξεργασίας των εικόνων στον εξυπηρετητή, όσο αυξάνεται το μέγεθος της δέσμης, αυξάνονται και οι απαιτήσεις σε μνήμη, με τις απαιτήσεις να κλιμακώνονονται περισσότερο στα 'βαριά' μοντέλα, όπως το EfficientNet v2 M, από ότι στα πιο 'ελαφριά', όπως το NASNet Mobile. Επιπλέον λόγω της αρχιτεκτονικής τους, οι υπολογισμοί για κάποια μοντέλα φαίνεται να επιδέχονται μεγαλύτερη παραλληλοποίηση με χρήση της GPU.

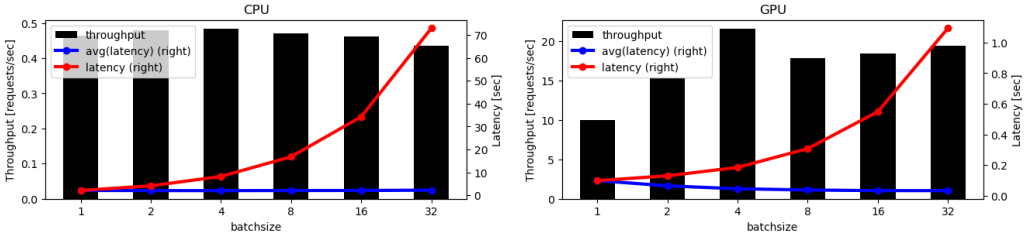
Παρακάτω παρουσιάζεται για κάθε μοντέλο ο ρυθμός διέλευσης των εικόνων (throughput), ο μέσος χρόνος απόκρισης ανά δέσμη (latency) κι ο μέσος χρόνος απόκρισης ανά δέσμη για κάθε δείγμα (avg(latency)) για τα διάφορα μεγέθη δέσμεων. Ο ρυθμός διέλευσης ορίζεται ως το συνολικό πλήθος εικόνων για τις οποίες εκτελέστηκε συμπερασματολογία προς τον συνολικό χρόνο που απαιτήθηκε γι' αυτή. Ως μέσος χρόνος απόκρισης ανά δέσμη ορίζεται ο χρόνος που απαιτείται για τη συμπερασματολογία μίας δέσμης εικόνων και ταυτίζεται με τον χρόνο που απαιτείται για την συμπερασματολογία μίας εικόνας, αφού όλες οι εικόνες εξέρχονται από το μοντέλο ύστερα από τη συμπερασματολογία της δέσμης, ταυτόχρονα. Τέλος, ο μέσος χρόνος απόκρισης ανά δέσμη για κάθε δείγμα είναι ίσος με τον χρόνο απόκρισης της κάθε δέσμης διαιρεμένος με το μέγεθος της δέσμης.



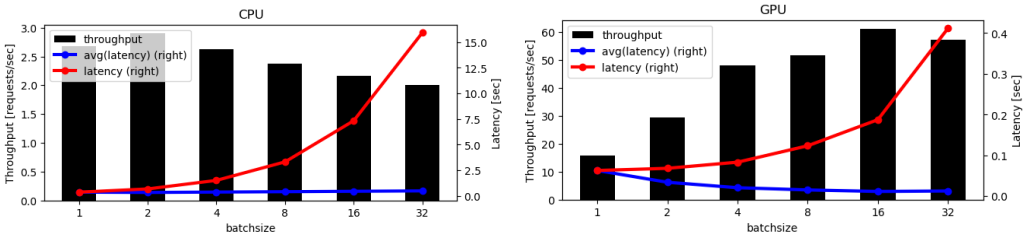
Εικόνα 4.4: *Χαρακτηρισμός εξυπηρετητή για το MobileNet v2*



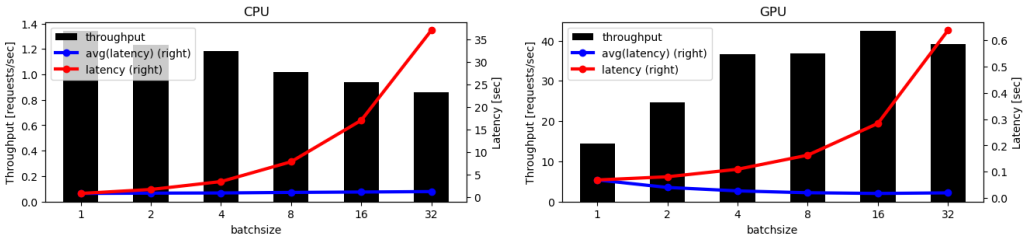
Εικόνα 4.5: *Χαρακτηρισμός εξυπηρετητή για το NASNet Mobile*



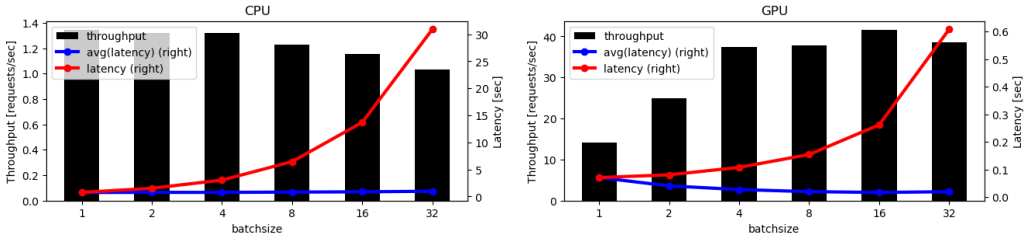
Εικόνα 4.6: *Χαρακτηρισμός εξυπηρετητή για το NASNet Large*



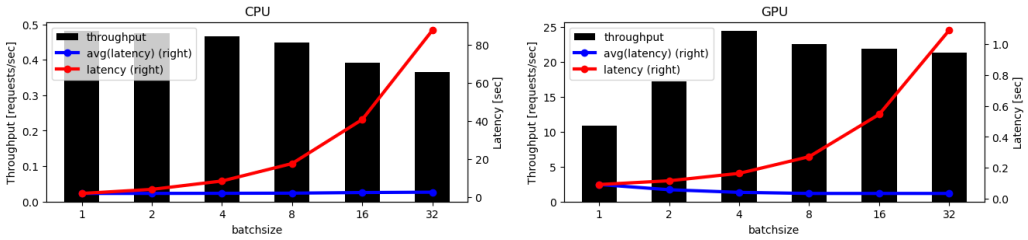
Εικόνα 4.7: *Χαρακτηρισμός εξυπηρετητή για το EfficientNet B3*



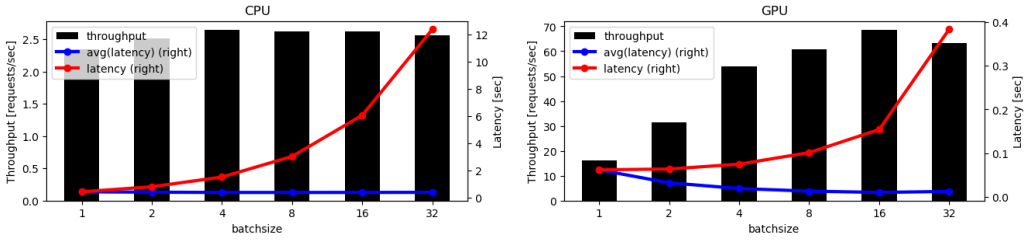
Εικόνα 4.8: *Χαρακτηρισμός εξυπηρετητή για το EfficientNet B4*



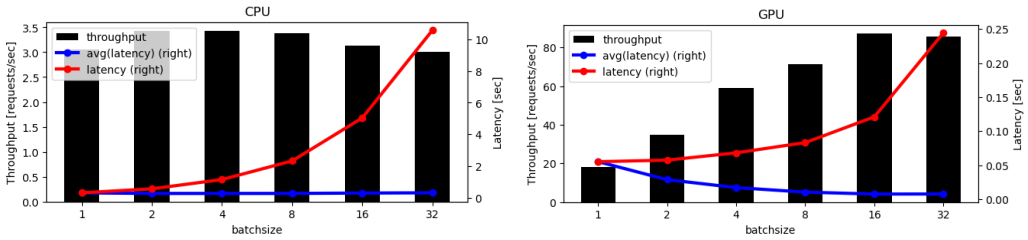
Εικόνα 4.9: *Χαρακτηρισμός εξυπηρετητή για το EfficientNet v2 S*



Εικόνα 4.10: *Χαρακτηρισμός εξυπηρετητή για το EfficientNet v2 M*



Εικόνα 4.11: *Χαρακτηρισμός εξυπηρετητή για το Inception v3*



Εικόνα 4.12: *Χαρακτηρισμός εξυπηρετητή για το ResNet v1 50*

Από τις παραπάνω γραφικές γίνεται εύκολα αντιληπτό πως για κάθε μοντέλο, η χρήση δέσμεων, κυρίως όταν γίνεται χρήση GPU για τη συμπερασματολογία, αυξάνει το ρυθμό διέλευσης αλλά και τον χρόνο απόκρισης. Επομένως, ανάλογα με τις απαιτήσεις που έχουμε από την εφαρμογή και τον ρυθμό άφιξης των εικόνων, και κατ' επέκταση την καθυστέρηση για τη δημιουργία της δέσμης, αλλάζει και το βέλτιστο μέγεθος δέσμης για το εκάστοτε μοντέλο.

4.4 Μοντελοποίηση Χρόνου Απόκρισης

Για τη μοντελοποίηση του χρόνου απόκρισης για τη συμπερασματολογία μιας εικόνας στο σύστημα, αρκεί να 'χωρίσουμε' την πιθανή 'διαδρομή' μιας τυχαίας εικόνας στο σύστημα στα τρία στάδια διέλευσης: την εκτέλεση συμπερασματολογίας στη συσκευή, τη μεταφορά

της εικόνας από την συσκευή στον εξυπηρετητή και την εκτέλεση συμπερασματολογίας στον εξυπηρετητή, αν αυτή προωθηθεί.

Στην μοντελοποίηση που ακολουθεί, έχουμε υποθέσει ότι οι εικόνες προς συμπερασματολογία είναι αποθηκευμένες στη συσκευή και πως η συμπερασματολογία δεν γίνεται σε δέσμες. Στη γενική περίπτωση όμως, θα έπρεπε να συνυπολογιστεί ο χρόνος για τη συλλογή εικόνων και τη δημιουργία των δεσμών.

Με βάση τα παραπάνω, για τον χρόνο συμπερασματολογίας τ_i μιας εικόνας j από τη συσκευή i, ισχύει πως:

$$\tau_{i,j} = t_{i,j} + p_i(t_{trans_{i,j}} + T_{s_j})$$

οπου $t_{i,j}$ ο χρόνος συμπερασματολογίας για την εικόνα j στη συσκευή i, p_i η πιθανότητα να προωθηθεί η εικόνα στον εξυπηρετητή, $t_{trans_{i,j}}$ ο χρόνος μεταφοράς της εικόνας από τη συσκευή στον εξυπηρετητή και T_{s_j} ο συνολικός χρόνος, αναμονής και συμπερασματολογίας, στον εξυπηρετητή.

Για τον χρόνο μεταφοράς t_{trans} της εικόνας j στη συσκευή i ισχύει:

$$t_{trans_{i,j}} = \frac{d_j}{BW_{av_i}} + v_i$$

όπου d_j το μέγεθος του αρχείου της εικόνας σε bits, BW_{av_i} το διαθέσιμο εύρος ζώνης και v_i η καθυστέρηση του διαύλου που χρησιμοποιεί η συσκευή i, που υπολογίζεται ως μέση καθυστέρηση από παρόχους.

Για τον χρόνο στον εξυπηρετητή T_s, γίνεται χρήση της Θεωρίας Ουρών Αναμονής και δεδομένων των υποθέσεων που έχουμε κάνει εως τώρα για το σύστημα και απλοποιώντας το περαιτέρω θεωρώντας τον χρόνο μεταφοράς σταθερό, προκύπτει πως με βάση το σύστημα ουράς αναμονής M/M/1, με ρυθμό αφίξεων $\lambda = \sum_{i=1}^{N} \frac{p_i}{t_i}$ και ρυθμό εξυπηρέτησης $\mu = \frac{1}{T}$ όπου T ο χρόνος συμπερασματολογίας στον εξυπηρετητή και N το πλήθος συσκευών στο σύστημα, ο χρόνος T_s δίνεται από τον τύπο:

$$T_s = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{T} - \sum_{i=1}^{N} \frac{p_i}{t_i}}$$

Ο λόγος που επιλέχθηκε το σύστημα M/M/1 έγκειται στο γεγονός ότι οι ιδιότητες του συστήματος συμπερασματολογίας που περιγράφεται στην παρούσα εργασία ταιριάζουν με τις ιδιότητες του συστήματος M/M/1. Ειδικότερα, το σύστημα M/M/1 είναι ένα σύστημα ουράς αναμονής ενός εξυπηρετητή, με μη ντετερμινιστικούς και συγκεκριμένα εκθετικά κατανεμημένους χρόνους μεταξύ των αφίξεων και εξυπηρέτησης με παραμέτρους λ και μ αντίστοιχα, οπότε έχουν τη σημαντική ιδιότητα απώλειας μνήμης. Σύμφωνα με την ιδιότητα αυτή, οι χρόνοι μεταξύ αφίξεων και οι χρόνοι εξυπηρέτησης είναι ανεξάρτητοι από τους προηγούμενους αντίστοιχους χρόνους. Αποτελούν την απλούστερη Μαρκοβιανή ουρά, δεν υπάρχει κανένα όριο στη χωρητικότητα του συστήματος και οι πελάτες εξυπηρετούνται με τη σειρά που φτάνουν στην ουρά. Σημειώνεται επίσης πως αφού οι χρόνοι μεταξύ των αφίξεων είναι

εκθετικά κατανεμημένοι, οι χρόνοι αφίξεων ακολουθούν την κατανομή Poisson.

Στο σύστημα της εργασίας, δεδομένων των παραδοχών που έχουν γίνει, ο χρόνος μεταξύ δύο διαδοχικών αφίξεων στον εξυπηρετητή και ο χρόνος συμπερασματολογίας-εξυπηρέτησης σε αυτόν δεν είναι ντετερμινιστικοί, αλλά τυχαίες μεταβλητές οι οποίες μάλιστα είναι μεταξύ τους ανεξάρτητες και έχουν την ιδιότητα απώλειας μνήμης. Οι μόνες συνεχείς τυχαίες μεταβλητές με την ιδιότητα της απώλειας μνήμης είναι αυτές που ακολουθούν την εκθετική κατανομή, οπότε η επιλογή φαίνεται να ευσταθεί [47].

Όσον αφορά στον ρυθμό άφιξης στον εξυπηρετητή, με βάση ό,τι έχει περιγραφεί για το σύστημα της παρούσας εργασίας έως τώρα, μπορεί κανείς να αντιληφθεί πως ο ρυθμός άφιξης στον εξυπηρετητή ταυτίζεται με το άθροισμα των ρυθμών εξόδου των εικόνων από τις συσκευές που, με βάση κάποιο κριτήριο προώθησης το οποίο διαμορφώνει την πιθανότητα προώθησης p_i, προωθείται στον εξυπηρετητή για συμπερασματολογία. Αυτό ισχύει καθώς, οι χρόνοι άφιξης των εικόνων στον εξυπηρετητή ισούνται με τον χρόνο εξόδου τους από την εκάστοτε συσκευή, μετατοπισμένους κατά t_{trans} που έχουμε θεωρήσει σταθερά, συνεπώς ο ρυθμός εξόδου των εικόνων προς συμπερασματολογία στον εξυπηρετητή, ταυτίζεται με τον ρυθμό άφιξης τους σε αυτόν. Επιπλέον, για τους λόγους που περιγράφηκαν στην παραπάνω παράγραφο μπορούμε να θεωρήσουμε ότι οι αφίξεις από την κάθε συσκευή στον εξυπηρετητή ακολουθούν διαδικασία Poisson με ρυθμό $\lambda = \frac{1}{t_i} \cdot p_i$, αφού ο χρόνος συμπερασματολογίας στη συσκευή και επομένως ο χρόνος εξόδου από αυτή ισούται με t_i και η πιθανότητα προώθησης από αυτή στον εξυπηρετητή ισούται με p_i. Τέλος, γνωρίζουμε ότι το άθροισμα διαδικασιών Poisson είναι επίσης διαδικασία Poisson με ρυθμό ίσο με το άθροισμα των ρυθμών των επιμέρους διαδικασιών, που μας οδηγεί στο ζητούμενο.

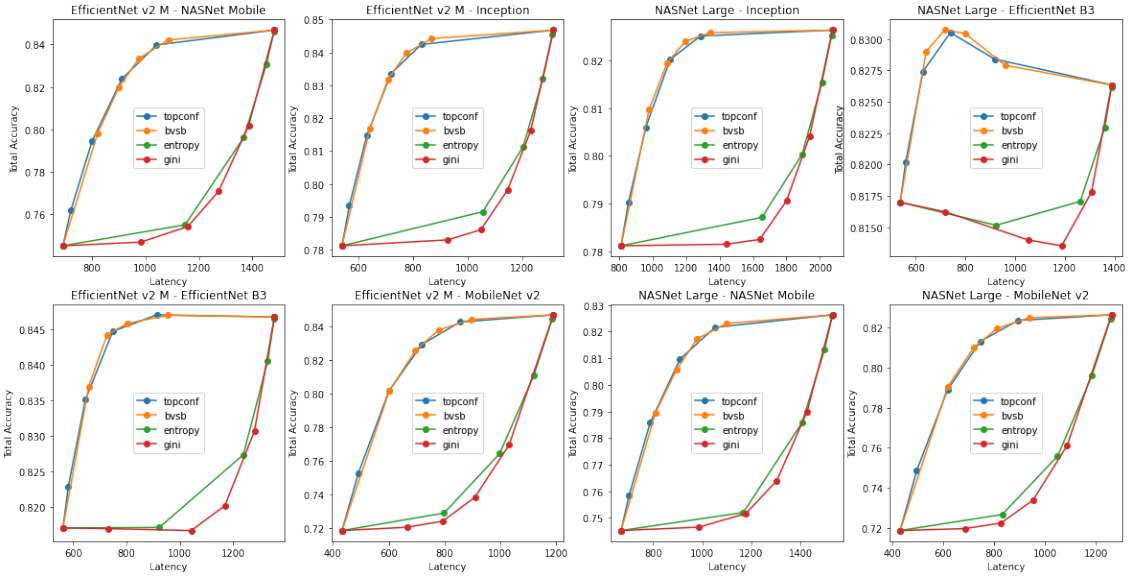
4.5 Κριτήριο Προώθησης

Στη συνέχεια, όπως αναφέρθηκε, έγινε διερεύνηση για το κριτήριο προώθησης της εκάστοτε εικόνας στον εξυπηρετητή, δηλαδή, για την εύρεση της κατάλληλης μετρικής και κατωφλιού αυτής για την εξασφάλιση ικανοποιητικής εμπιστοσύνης της πρόβλεψης των μοντέλων λαμβάνοντας υπόψιν την επιθυμία για συμβιβασμό της ακρίβειας με τον χρόνο. Οι εικόνες με τιμή της μετρικής μικρότερη από το κατώφλι αυτό, θα αποστέλλονται στον εξυπηρετητή.

Η διαδικασία που ακολουθήθηκε ήταν αρχικά να παρθούν 3 υποσύνολα του συνόλου επικύρωσης του ImageNet χωρίς επικάλυψη, το καθένα μεγέθους ίσου με 20% του συνολικού με σκοπό να γίνουν μετρήσεις για την καθυστέρηση και την ακρίβεια της συμπερασματολογίας του συστήματος για 8 συνδυασμούς μοντέλων για τα διαφορετικά κριτήρια προώθησης (μετρικές και κατώφλια πεποίθησης). Ο λόγος που επιλέχθηκαν αυτοί οι 8 συνδυασμοί είναι για να γίνει μια επαρκής αλλά ενδεικτική μελέτη πάνω στα κριτήρια προώθησης. Τα συγκεκριμένα μοντέλα από τον Πίνακα 4.3 επιλέχθηκαν έτσι ώστε να υπάρχει μια κλιμάκωση στο μέγεθος και την ακρίβειά τους. Επίσης, πάρθηκε ένα τέταρτο, μη επικαλυπτόμενο, υποσύνολο, ίσο με το 10% του συνολικού, ώστε να γίνει έλεγχος για το αν οι μετρήσεις που έγιναν ανταποκρίνονται σε όλο το σύνολο δεδομένων, μελετώντας την απόκλιση του μέσου όρου των μετρήσεων στα 3 επιμέρους υποσύνολα από το τέταρτο υποσύνολο ελέγχου. Οι εν λόγω αποκλίσεις για τα διάφορα μοντέλα και μετρικές παρουσιάζονται στο Παράρτημα Α, Εικόνα Α΄.1 και Εικόνα Α΄.2. Τέλος, σημειώνεται πως οι μετρήσεις έγιναν με τα ζεύγη μοντέλων και τις εικόνες να

βρίσκονται όλα στον εξυπηρετητή, με χρήση GPU καθώς ο σκοπός των μετρήσεων αυτών ήταν για την εύρεση του βέλτιστου συμβιβασμού μεταξύ ακρίβειας και χρόνου σε κάθε περίπτωση.

Παρακάτω, λοιπόν, παρουσιάζονται οι γραφικές παραστάσεις της ακρίβειας συναρτήσει του χρόνου απόκρισης και του κατωφλιού αντίστοιχα για οκτώ ζεύγη μοντέλων, κάνοντας χρήση των μετρικών πεποίθησης που παρουσιάστηκαν στην Ενότητα 4.1 για κατώφλια (όρια) που ανήκουν στο σύνολο {0,0.2,0.4,0.6,0.8,1}. Κάθε σημείο των γραφικών παραστάσεων αντιστοιχεί σε ένα από αυτά τα κατώφλια, με τη σειρά που εμφανίζονται. Για κατώφλι ίσο με μηδέν, αφού όπως είδαμε οι μετρικές είναι θετικές και συνεπώς όχι μικρότερες του μηδενός, η συμπερασματολογία εκτελείται εξολοκλήρου στις συσκευές, ενώ από την άλλη αν είναι ίσο με τη μονάδα, επειδή όπως αναφέρθηκε στην παρουσίαση μετρικών στην Ενότητα 4.1 χρησιμοποιούμε την κανονικοποιημένη τους μορφή, και επομένως είναι πάντα μικρότερες της μονάδας, η συμπερασματολογία επαναλαμβάνεται για όλες τις εικόνες στον εξυπηρετητή. Γι΄ αυτό παρατηρούμε πως για κατώφλι ίσο με μηδέν, ο συνολικός χρόνος απόκρισης είναι μικρός όπως και η ακρίβεια, αφού πρόκειται για τα 'ελαφριά' αλλά λιγότερο αποτελεσματικά μοντέλα των συσκευών, ενώ για κατώφλι ίσο με τη μονάδα, έχουμε μεγαλύτερη ακρίβεια αλλά και μεγαλύτερο συνολικό χρόνο απόκρισης.



Εικόνα 4.13: *Απόδοση συστημάτων ζεύγους Νευρωνικών Δικτύων για διαφορετικά όρια στις μετρικές εμπιστοσύνης*

Από τις παραπάνω γραφικές παρατηρούμε πως η καλύτερη μετρική για βελτιστοποίηση του χρόνου απόκρισης και της ακρίβειας είναι η διαφορά πρώτης και δεύτερης εμπιστοσύνης (BvSB).

Όπως ήταν αναμενόμενο, σχεδόν σε όλους τους συνδυασμούς μοντέλων παρατηρείται μεγιστοποίηση της ακρίβειας στον μέγιστο χρόνο απόκρισης, όταν δηλαδή η συμπερασματολογία, όπως ειπώθηκε, επαναλαμβάνεται για όλες τις εικόνες στον εξυπηρετητή. Άξιο αναφοράς είναι όμως το γεγονός ότι στους συνδυασμούς μοντέλων εξυπηρετητή-συσκευής NASNet Large - EfficientNet B3 και EfficientNet v2 M - EfficientNet B3 επιτυγχάνεται μεγιστοποίηση της ακρίβειας σε μικρότερο συνολικό χρόνο απόκρισης σε σχέση με αυτόν αν επαναλαμβανόταν

η συμπερασματολογία για όλες τις εικόνες στον εξυπηρετητή. Συγκεκριμένα, στην πρώτη περίπτωση παρατηρείται μεγιστοποίηση της ακρίβειας για κατώφλι ίσο με 0.4 και ακρίβεια ίση με 83.07% έναντι της 82.63% που επιτυγχάνεται μόνο με τον εξυπηρετητή και στη δεύτερη περίπτωση η ακρίβεια μεγιστοποιείται για κατώφλι ίσο με 0.8 και ακρίβεια ίση με 84.70% έναντι της 84.68% που επιτυγχάνεται μόνο με τον εξυπηρετητή. Αυτό συμβαίνει πιθανότατα γιατί λειτουργούν 'καλά' συνδυαστικά, δηλαδή λόγω αρχιτεκτονικής, οι προβλέψεις για τις οποίες το τοπικό μοντέλο έχει χαμηλή εμπιστοσύνη και κάνει πράγματι λάθος πρόβλεψη, φαίνεται να είναι αυτές που προβλέπει σωστά το μοντέλο του εξυπηρετητή, ενώ δε φαίνεται να συμβαίνει το ίδιο για αυτές στων οποίων την πρόβλεψη το τοπικό μοντέλο έχει υψηλή εμπιστοσύνη.

Σε σχέση με την επιλογή βέλτιστου κατωφλιού, καθώς πρόκειται για πρόβλημα βελτιστοποίησης, σύμφωνα με τη βιβλιογραφία χρησιμοποιήθηκε η εξής μεθοδολογία: Η δυναμική μεταξύ ακρίβειας και χρόνου απόκρισης καθορίζουν πόσο επιπλέον κόστος σε χρόνο επιτρέπουμε να 'πληρώσουμε' για κάθε ποσοστιαία μονάδα (percentage point - p.p.) κέρδους ακρίβειας. Επιθυμούμε δηλαδή να ελαχιστοποιηθεί ο χρόνος απόκρισης με ανοχή στην πτώση ακρίβειας έως και 1 ή 2 ποσοστιαίες μονάδες [48]. Η μόνη εξαίρεση ήταν ο συνδυασμός μοντέλων NASNet Large - EfficientNet B3 στον οποίο παρατηρήθηκε σημαντική αύξηση στην ακρίβεια στο κατώφλι που αναφέρθηκε παραπάνω.

Οι επιλογές για το κατώφλι της μετρικής BvSB για τους 8 συνδυασμούς μοντέλων, για πτώση της ακρίβειας κατά μία και δύο ποσοστιαίες μονάδες, παρουσιάζονται στον Πίνακα 4.4.

Πίνακας 4.4: *Επιλογή Κατωφλιών για τη μετρική BvSB*

Μοντέλο Συσκευής	Μοντέλο Εξυπηρετητή			
	NASNet Large		EfficientNet v2 M	
	p.p. = 1	p.p. = 2	p.p. = 1	p.p. = 2
MobileNet v2	0.6	0.4	0.6	0.6
NASNet Mobile	0.6	0.6	0.8	0.6
EfficientNet B3	0.4	0.4	0.2	0.2
Inception	0.4	0.2	0.6	0.4

4.6 Πειραματική Προσομοίωση του Συστήματος

Για την πειραματική προσομοίωση του συστήματος, επαναλήφθηκε 3 φορές η διαδικασία που ακολουθεί, σε κάθε μία από αυτές ως σύνολο εικόνων προς συμπερασματολογία ήταν ένα από τα υποσύνολα του ImageNet όπως αυτά περιγράφηκαν στην Ενότητα 4.5. Αρχικά, πάρθηκαν μετρήσεις σε σχέση με την ακρίβεια και διάφορες μετρικές της απόδοσης του συστήματος που θα περιγραφούν σε επόμενη ενότητα, που βασίζονται στον χρόνο απόκρισης του, αυξάνοντας σταδιακά το πλήθος και την ετερογένεια των συσκευών του συστήματος, κάθε μία από τις οποίες είχε αποθηκευμένο το ίδιο σύνολο εικόνων για συμπερασματολογία. Οι μετρήσεις δεν έγιναν σε πραγματικές συσκευές, αλλά σε εξυπηρετητή και στη συνέχεια έγινε προσομοίωση του χρόνου συμπερασματολογίας σε αυτές με τρόπο που θα περιγραφεί στην Υποενότητα 4.6.1.

Σε σχέση με την επιλογή των μοντέλων, δοκιμάστηκαν δύο μοντέλα στον εξυπηρετητή, το NASNet Large και το EfficientNet v2 M, ένα κάθε φορά. Τα διαθέσιμα μοντέλα συσκευών είναι τα: MobileNet v2, NASNet Mobile και EfficientNet B3 τα οποία τοποθετήθηκαν σε συσκευές απόδοσης Low-Tier, Mid-Tier και High-Tier αντίστοιχα. Η διανομή μοντέλων ανά τύπο συσκευής έγινε με τρόπο ώστε οι λιγότερο αποδοτικές συσκευές να είναι εξοπλισμένες με πιο ελαφριά μοντέλα, ενώ συσκευές με καλύτερα χαρακτηριστικά που μπορούν να υποστηρίξουν πιο βαριά μοντέλα, να είναι εξοπλισμένα με αυτά.

Αρχικά μετρήθηκε ο χρόνος για τη συμπερασματολογία των εικόνων των 3 υποσυνόλων, όπως επίσης και η τιμή της μετρικής BvSB των προβλέψεων αυτών, για τα 3 μοντέλα συσκευών που αναφέρθηκαν παραπάνω και μετρήθηκε επίσης ο χρόνος για τη συμπερασματολογία των εικόνων για τα 2 μοντέλα εξυπηρετητή. Όλες οι μετρήσεις έγιναν στο Kaggle με χρήση της διαθέσιμης GPU. Σημειώνεται επίσης ότι δεν συνυπολογίστηκε ο χρόνος μεταφοράς.

Από τις μετρήσεις αυτές έχουν αποθηκευτεί τα εξής:

- **acc$_\mathbf{m}$**: λίστα 0, 1 ανάλογα με το αν η αντίστοιχη εικόνα έχει ταξινομηθεί σωστά με βάση το μοντέλο της συσκευής
- **acc$_\mathbf{s}$**: λίστα 0, 1 ανάλογα με το αν η αντίστοιχη εικόνα έχει ταξινομηθεί σωστά με βάση το μοντέλο του εξυπηρετητή
- **time$_\mathbf{m}$**: λίστα με τους χρόνους συμπερασματολογίας για το μοντέλο της συσκευής (δημιουργούνται με βάση την υποενότητα 4.6.1)
- **time$_\mathbf{s}$**: λίστα με τους χρόνους συμπερασματολογίας για το μοντέλο του εξυπηρετητή
- **bvsb**: λίστα με τις τιμές της μετρικής BvSB για τη συμπερασματολογία στη συσκευή

4.6.1 Χρόνος Συμπερασματολογίας Συσκευών

Για τις μετρήσεις, όπως αναφέρθηκε, δεν χρησιμοποιήθηκαν πραγματικές συσκευές αλλά έγινε προσομοίωση του χρόνου συμπερασματολογίας τους κλιμακώνοντας τον χρόνο που χρειάστηκε ο εξυπηρετητής (Kaggle) πολλαπλασιάζοντάς τον με έναν παράγοντα κλιμάκωσης χρόνου (scaling factor):

$$sf = \frac{\text{απόδοση επεξεργαστή εξυπηρετητή}}{\text{απόδοση επεξεργαστή συσκευής}}$$

με βάση την απόδοση της GPU του εξυπηρετητή (Kaggle) και την απόδοση των τριών τύπων GPU των συσκευών (βλ. Πίνακα 4.2).

4.6.2 Ουρά στον Εξυπηρετητή

Για την προσομοίωση της ουράς στον εξυπηρετητή, αρχικά μετασχηματίστηκαν οι χρόνοι συμπερασματολογίας στις εικονικές συσκευές όπως περιγράφηκε στην προηγούμενη υποενότητα. Οι χρόνοι αυτοί είναι αποθηκευμένοι σε λίστες, μία για κάθε μία από τις συσκευές του συστήματος. Οι λίστες αυτές αποτελούν τα στοιχεία του πίνακα χρόνων συμπερασματολογίας

στις συσκευές, $time_m$. Στη συνέχεια, για την προσομοίωση του χρόνου εξόδου τους από την εκάστοτε συσκευή, δημιουργούμε ένα νέο διάνυσμα ανά συσκευή, κάθε στοιχείο του οποίου ισούται με το άθροισμα των στοιχείων που προηγούνταν στο αντίστοιχο διάνυσμα του πίνακα $time_m$. Ο πίνακας αυτός είναι ο $timestamps_m$ και ουσιαστικά αποτελεί τον πίνακα στον οποίο αποθηκεύεται ο 'απόλυτος χρόνος' εξόδου από την κάθε συσκευή, δηλαδή η χρονοσφραγίδα τη στιγμή της εξόδου της κάθε εικόνας, αν θεωρήσουμε ότι ο χρόνος ξεκινάει από όταν ξεκινάει η εκτέλεση της συμπερασματολογίας της πρώτης εικόνας στις συσκευές. Να σημειωθεί επίσης ότι θεωρούμε ότι όλες οι συσκευές ξεκινάνε ταυτόχρονα τη συμπερασματολογία. Για τυχαίο στοιχείο j του $timestamps_m[i]$, της λίστας δηλαδή του πίνακα $timestamps_m$ που αντιστοιχεί στη συσκευή i έχουμε:

$$timestamps_m[i][j] = \sum_{k=1}^{j} time_m[i][k]$$

Μετά, με βάση τις τιμές της μετρικής BvSB που είναι αποθηκευμένες στον πίνακα *bvsb* με όμοιο τρόπο με τους χρόνους στον $time_m$, δημιουργούμε έναν νέο πίνακα, τον $server_{ind}$ ο οποίος περιέχει τους δείκτες θέσης των εικόνων που θα προωθούνταν στον εξυπηρετητή για συμπερασματολογία εκ νέου, ο οποίος επίσης αποτελεί λίστα λιστών της κάθε συσκευής. Χρησιμοποιώντας την $server_{ind}$, δημιουργούμε τη λίστα $time_{ms}$ στην οποία αποθηκεύουμε τις τιμές των στοιχείων του $timestamps_m$ που ανήκουν στην $server_{ind}$. Ύστερα, δημιουργούμε τη λίστα ms_{ind} που περιέχει τους δείκτες της $time_{ms}$ που θα την διατάσσανε, εν ολίγοις αποτελεί την ουρά των εικόνων στον εξυπηρετητή.
Συνοπτικά, έχουν δημιουργηθεί επιπλέον οι εξής λίστες:

- **server$_{\text{ind}}$**: λίστα με τους δείκτες των εικόνων που πρέπει να προωθηθούν στον εξυπηρετητή
- **time$_{\text{ms}}$**: λίστα με τις χρονοσφραγίδες εξόδου από τις συσκευές των εικόνων προς συμπερασματολογία στον εξυπηρετητή
- **ms$_{\text{ind}}$**: λίστα με τους δείκτες των εικόνων με τη σειρά που εισέρχονται στην ουρά του εξυπηρετητή

Διατάσσουμε την $time_{ms}$ με βάση τους δείκτες που είναι αποθηκευμένοι στην ms_{ind}, δηλαδή με αύξουσα σειρά εισόδου στον εξυπηρετητή. Τελικά διατρέχουμε και μετασχηματίζουμε τη διατεταγμένη πλέον $time_{ms}$ με βάση τον εξής αλγόριθμο:

```
timestamp = 0
for img in ms_ind:
    if (time_ms[img]-timestamp)>=0:
        time_ms[img] += time_s[server_ind[img]]
    else:
        time_ms[img] = timestamp + time_s[server_ind[img]]
    timestamp = time_ms[img]
```

Ουσιαστικά δηλαδή, ξεκινώντας από την αρχή του διατεταγμένου $time_{ms}$ ελέγχουμε για κάθε εικόνα αν την στιγμή που φτάνει έχει τελειώσει στην συμπερασματολογία της προηγούμενης. Αν έχει τελειώσει, προσθέτουμε στον χρόνο άφιξης της στον εξυπηρετητή, τον χρόνο που χρειάζεται για συμπερασματολογία σε αυτόν. Αν όχι, ο χρόνος εξόδου της από τον εξυπηρετητή, ισούται με τον χρόνο εξόδου της προηγούμενης από τον εξυπηρετητή αν σε αυτόν προσθέσουμε τον χρόνο συμπερασματολογίας αυτής στον εξυπηρετητή. Η διαδικασία αυτή επαναλαμβάνεται για καθένα από τα 3 υποσύνολα του ImageNet και για τα πλήθη συσκευών που ανήκουν στο σύνολο $\{1, 2, 3, 4, 5, 6, 7, 8, 16, 32, 64, 128\}$

Κεφάλαιο 5

Αξιολόγηση

Στο κεφάλαιο αυτό παρουσιάζονται οι μετρικές και τα αποτελέσματα των μετρήσεων για την αξιολόγηση της απόδοσης του συστήματος όταν το πλήθος και η ετερογένεια των συσκευών του μεταβάλλονται.

5.1 Μετρικές Ελέγχου

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης του συστήματος είναι οι εξής:

- **System Accuracy**: Η ακρίβεια του συστήματος.
- **Normalized Turnaround Time (NTT)**: Ποσοτικοποιεί την καθυστέρηση στη συμπερασματολογία που οφείλεται στην ύπαρξη πολλών συσκευών. Το Normalized Turnaround Time για τη συσκευή i ενός συστήματος ορίζεται ως:

$$NTT_i = \frac{T_i^{MD}}{T_i^{SD}}$$

όπου T_i^{MD} ο χρόνος για την εκτέλεση συμπερασματολογίας όλων των εικόνων της συσκευής i στο σύστημα ως έχει και T_i^{SD} ο χρόνος για την εκτέλεση συμπερασματολογίας όλων των εικόνων της συσκευής i σε ένα σύστημα που αποτελείται μόνο από την συσκευή i και τον εξυπηρετητή. Για τις τιμές της μετρικής αυτής ισχύει πως $NTT \in [1, N]$ όπου N το συνολικό πλήθος συσκευών. Όσο η τιμή της μετρικής αυτής τείνει στη μονάδα, τόσο πιο αποδοτικό είναι το σύστημα, αφού αυτό συνεπάγεται ότι η αύξηση των συσκευών στο σύστημα δεν το καθυστερεί.

- **Average Normalized Turnaround Time (ANTT)**: Το μέσο NTT των συσκευών του συστήματος που ορίζεται ως:

$$ANTT = \frac{1}{N} \sum_{i=1}^{N} NTT_i$$

Όσο μικρότερη είναι η τιμή του, τόσο το καλύτερο, καθώς αυτό συνεπάγεται πως η μέση καθυστέρηση των συσκευών λόγω της παράλληλης αποστολής εικόνων στον εξυπηρετητή είναι επίσης μικρή.

- **Max Normalized Turnaround Time (MNTT)**: Το μέγιστο NTT του συστήματος, ή:

$$MNTT = \max_{i=1}^{N} NTT_i$$

- **SLA Violation Rate**: Το ποσοστό παραβίασης του ορίου του χρόνου απόκρισης που έχει τεθεί για την εξασφάλιση μιας καθορισμένης ποιότητας εξυπηρέτησης του συστήματος. Το αποδεκτό όριο καθυστέρησης στην παρούσα εργασία ισούται με $c \cdot t_i$ όπου c μια σταθερά που παίρνει τις τιμές $\{1, 10, 20, 100\}$ για να ορίσουμε διαφορετικά όρια σε σχέση με την ποιότητα εξυπηρέτησης, και t_i ο χρόνος συμπερασματολογίας της συσκευής για την εικόνα i.

- **System Throughput (STP)**: Ποσοτικοποιεί το πλήθος ολοκληρωμένων συμπερασματολογιών ανά μονάδα χρόνου. Συγκεκριμένα, ποσοτικοποιεί τη συσσωρευμένη πρόοδο συμπερασματολογίας της κάθε συσκευής στο σύστημα. Μαθηματικά ορίζεται ως:

$$STP = \sum_{i=1}^{N} \frac{T_i^{SD}}{T_i^{MD}}$$

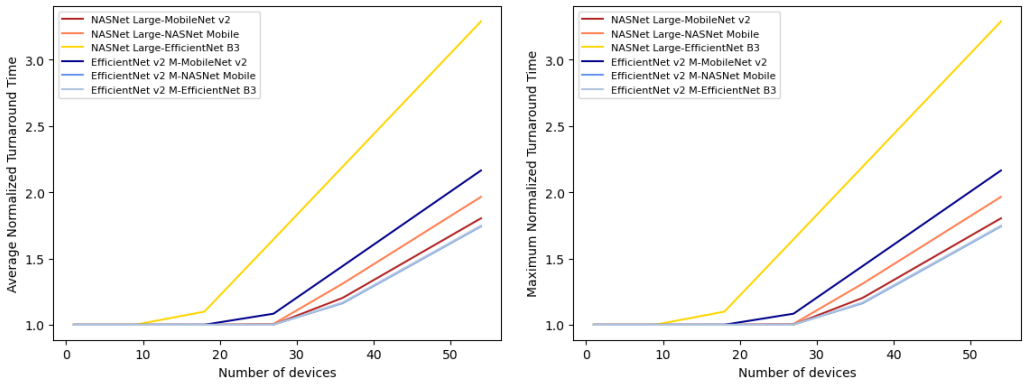
  Η μετρική αυτή παίρνει τιμές που ανήκουν στο σύνολο [1,N] και όσο μεγαλύτερη είναι η τιμή της, τόσο πιο αποδοτικό είναι το σύστημα.

5.2 Απόδοση Συστήματος

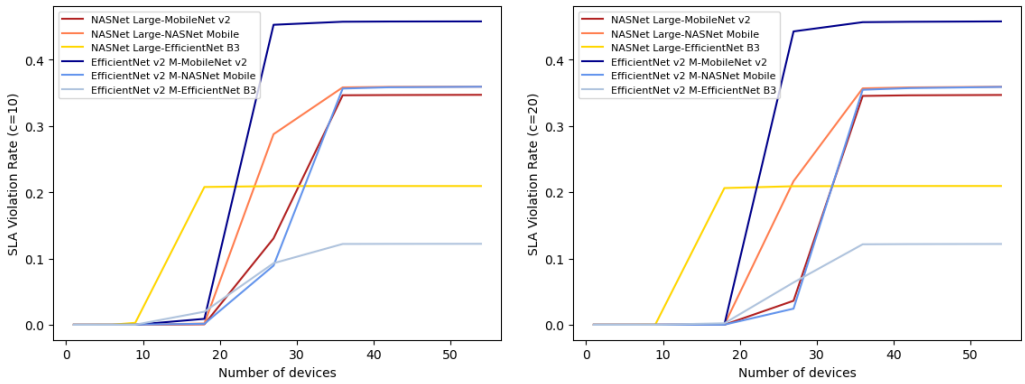
Οι παραπάνω μετρικές χρησιμοποιήθηκαν για τη μελέτη του τρόπου με τον οποίο η αύξηση στην ετερογένεια και το πλήθος των συσκευών επιδρά στην απόδοση του συστήματος. Πάρθηκαν μετρήσεις για πλήθη συσκευών που ανήκουν στο σύνολο {1,2,3,4,6,9,18,27,36,54}, χρησιμοποιώντας ως κριτήριο προώθησης στον εξυπηρετητή την τιμή της μετρικής BvSB της τοπικής συμπερασματολογίας της κάθε εικόνας, με τα κατώφλια που καθορίστηκαν στην Ενότητα 4.5 για ανοχή στην πτώση της ακρίβειας της τάξεως των 2 ποσοστιαίων μονάδων (2 p.p.).

Για τη μελέτη της επίδρασης της κλιμάκωσης στο πλήθος των συσκευών στο σύστημα, όλες οι συσκευές ήταν του ίδιου τύπου, με το ίδιο μοντέλο, βάσει των συνδυασμών που παρουσιάστηκαν στην Ενότητα 4.6.

Όπως παρατηρούμε από την Εικόνα 5.1, οι μετρικές Average Turnaround Time και Maximum Turnaround Time δεν διαφοροποιούνται μεταξύ τους για δεδομένο σύστημα. Επίσης, παρατηρούνται μεγαλύτερες τιμές των μετρικών αυτών στην περίπτωση που το μοντέλο του εξυπηρετητή είναι το NASNet Large, που σημαίνει πως μάλλον με την επιλογή των κατωφλιών που έχει γίνει, στέλνονται αρκετές εικόνες στον εξυπηρετητή και δημιουργείται συμφόρηση, κυρίως στην περίπτωση του συστήματος με μοντέλο στη συσκευή το EfficientNet B3. Επιπλέον, αξίζει να σημειωθεί πως μέχρι τις 9 συσκευές, δεν επηρεάζεται ιδιαίτερα η τιμή των μετρικών και συνεπώς η αύξηση του πλήθους των συσκευών μέχρι το σημείο αυτό δεν επιδρά σημαντικά στην απόδοση του συστήματος.



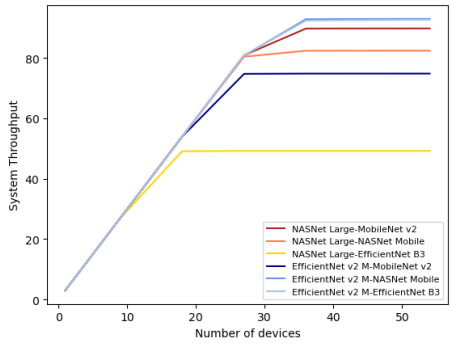
Εικόνα 5.1: *ANTT και MNTT συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*



Εικόνα 5.2: *Service Level Agreement Violation Rates συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*

Από την Εικόνα 5.2 μπορεί κανείς να δει πως και για τις δύο σταθερές ($c = 10$ και $c = 20$), για πλήθος συσκευών μεγαλύτερο από 36, η απόδοση του συστήματος ως προς το SLA Violation Rate μένει σταθερή. Μέχρι αυτή την τιμή όμως, παρόλο που οι τιμές της μετρικής διαφέρουν ελαφρώς για διαφορετική τιμή της σταθεράς, οι σχέσεις τιμών της μεταξύ των διαφορετικών ως προς τα μοντέλα συστημάτων είναι κοινές. Για όλους τους συνδυασμούς μοντέλων εξυπηρετητή-συσκευών, με εξαίρεση τον NASNet Large-EfficientNet B3 παρατηρείται πως τα μοντέλα με χαμηλές τιμές των μετρικών MNTT και ANTT έχουν χαμηλό SLA Violation Rate, όπως άλλωστε ήταν αναμενόμενο με βάση την Εικόνα 5.1, καθώς καθυστέρηση στη συμπερασματολογία ανά δείγμα είναι λογικό να προκαλεί καθυστέρηση συνολικά στο σύστημα. Η εξαίρεση για τον συνδυασμό NASNet Large-EfficientNet B3 που έχει υψηλότερες τιμές των μετρικών MNTT και ANTT σε σχέση με τα υπόλοιπα μοντέλα, αλλά χαμηλότερο SLA Violation Rate. Αυτό θα μπορούσε να δικαιολογείται στην περίπτωση που στέλνονται λιγότερες εικόνες, σε σχέση με τους υπόλοιπους συνδυασμούς μοντέλων, στον εξυπηρετητή, οπότε να μην καθυστερεί η συμπερασματολογία για πολλές εικόνες, αλλά η συμπερασματολογία των εικόνων στον εξυπηρετητή να διαρκεί αρκετά παραπάνω σε σχέση με τον τοπικό χρόνο απόκρισης, και συνεπώς το σύστημα συνολικά να καθυστερεί.

Στην Εικόνα 5.3 παρουσιάζονται οι ρυθμοί διέλευσης των ομοιογενών συστημάτων. Από



Εικόνα 5.3: *System Throughput συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*

την εικόνα αυτή μπορεί κανείς να συμπεράνει πως η αποδοτικότητα του συστήματος αυξάνεται με την αύξηση του πλήθους των συσκευών μέχρι και τις 18 συσκευές για τον συνδυασμό μοντέλων εξυπηρετητή-συσκευών NASNet Large-EfficientNet B3, μέχρι και τις 27 συσκευές για τον συνδυασμό EfficientNet v2 M-MobileNet v2 και μέχρι και τις 36 συσκευές για τους υπόλοιπους. Μετά από αυτό, το σύστημα έρχεται σε κορεσμό και η απόδοση του σταθεροποιείται όπως παρατηρήθηκε και στην Εικόνα 5.2.

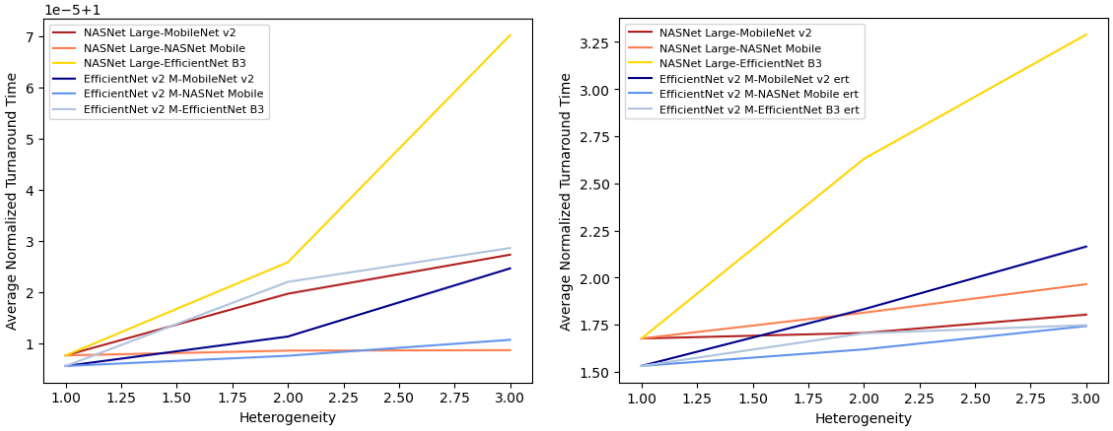
Στην περίπτωση ομοιογένειας, ανεξάρτητα από την αύξηση του πλήθους των συσκευών, η συνολική ακρίβεια του συστήματος παραμένει σταθερή δεδομένου ότι κάθε συσκευή έχει το ίδιο μοντέλο και τις ίδιες εικόνες προς συμπερασματολογία, και στέλνει τις εικόνες για επανάληψη της συμπερασματολογίας τους στον εξυπηρετητή με βάση το ίδιο κριτήριο. Η ακρίβεια αυτή, είναι η ίδια με αυτή που παρουσιάστηκε στην Εικόνα 4.13 για το δεδομένο ποσοστιαίο όριο ανοχής, και παρουσιάζεται για τους διάφορους συνδυασμούς στον Πίνακα 5.1.

Πίνακας 5.1: *Ακρίβεια συνδυασμών μοντέλων εξυπηρετητή-συσκευής για επιλογή κατωφλιών με 2p.p. ανοχή*

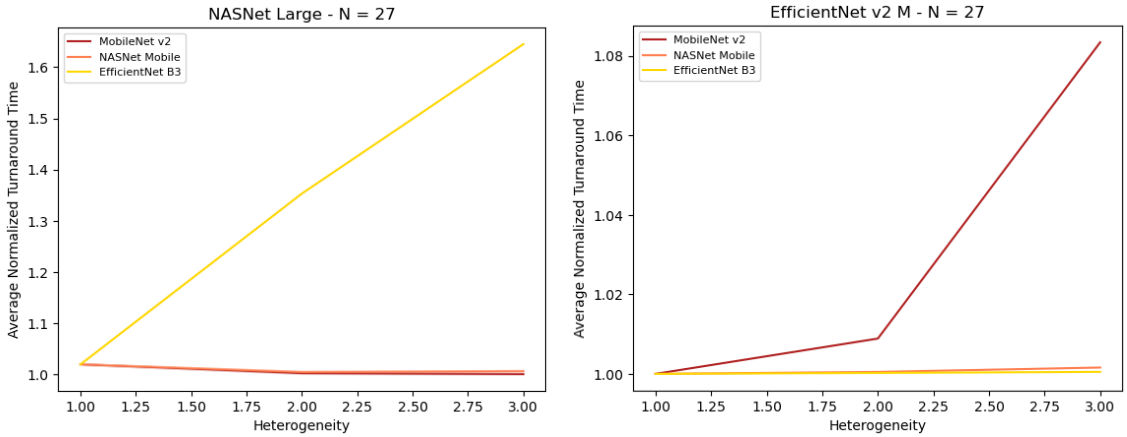
Μοντέλο Εξυπηρετητή	**Μοντέλο Συσκευής**	**Κατώφλι 2 p.p.**	**Ακρίβεια**
NASNet Large	MobileNet v2	0.4	81.00%
NASNet Large	NASNet Mobile	0.6	81.73%
NASNet Large	EfficientNet B3	0.4	83.07%
EfficientNet v2 M	MobileNet v2	0.6	83.79%
EfficientNet v2 M	NASNet Mobile	0.6	83.34%
EfficientNet v2 M	EfficientNet B3	0.2	83.70%

Ακολουθεί η αξιολόγηση της μεταβολής της απόδοσης του συστήματος ως συνάρτηση της ετερογένειάς του. Οι μετρήσεις έγιναν για συστήματα με τα ίδια πλήθη συσκευών όπως και στη μελέτη για την κλιμάκωση του πλήθους. Οι τιμές της ετερογένειας στην Εικόνα 5.4, έχουν την εξής αντιστοιχία: Για κάθε συνδυασμό μοντέλου, που αντιστοιχεί σε μία γραμμή, κι εφόσον έχουν οριστεί 3 τύποι συσκευών, ετερογένεια ίση με τη μονάδα σημαίνει απόλυτη ετερογένεια, δηλαδή στο σύστημα υπάρχουν ίσου πλήθους συσκευές για κάθε μοντέλο. Για παράδειγμα, για πλήθος συσκευών ίσο με 9, υπάρχουν 3 συσκευές με το μοντέλο MobileNet v2, 3 με το μοντέλο NASNet Mobile και 3 με το μοντέλο EfficientNet B3. Για ετερογένεια ίση με 2, τα

2/3 του συνόλου των συσκευών έχουν το μοντέλο που αντιστοιχεί στο μοντέλο συσκευής της γραμμής ενώ το υπόλοιπο 1/3 είναι διαφορετικο μοντέλο. Πιο συγκεκριμένα, για παράδειγμα στη γραμμή που αφορά τον συνδυασμό μοντέλων EfficientNet v2 M-MobileNet v2, για την μέτρηση που αντιστοιχεί σε ετερογένεια ίση με 2, στην περίπτωση 9 συσκευών, οι 6 συσκευές έχουν το μοντέλο MobileNet v2. Η τιμή της μετρικής ANTT σε αυτή την περίπτωση είναι η μέση τιμή της ανάμεσα, (α) στο σύστημα με 6 συσκευές με MobileNet v2 και 3 συσκευές με EfficientNet B3 και (β) στο σύστημα με 6 συσκευές με MobileNet v2 και 3 συσκευές με NASNet Mobile. Τέλος, για ετερογένεια ίση με 3, αντιστοιχεί σε απόλυτη ομοιογένεια, το σύστημα δηλαδή αποτελείται μόνο από συσκευές με το μοντέλο που αντιστοιχεί στη γραμμή.



Εικόνα 5.4: *Επιρροή της μείωσης της ετερογένειας του συστήματος στη μετρική ANTT (Αριστερά: Πλήθος Συσκευών = 9, Δεξιά: Πλήθος Συσκευών = 54)*



Εικόνα 5.5: *Επιρροή της μείωσης της ετερογένειας του συστήματος στη μετρική ANTT για πλήθος συσκευών ίσο με 27*

Βασιζόμενοι στις Εικόνες 5.4 και 5.5, γίνεται αντιληπτό πως όσο πιο ετερογενές είναι το σύστημα, όσο δηλαδή οι συσκευές τείνουν να είναι ίδιου πλήθους ανά είδος, τόσο πιο αποδοτικό είναι. Αυτό οφείλεται στο ότι όταν το σύστημα, με βάση τις υποθέσεις που έχουμε κάνει, είναι πλήρως ομοιογενές, οι συσκευές στέλνουν όλες ταυτόχρονα στο σύστημα επιβαρύνοντάς το επιπλέον, και από μία αρχική τιμή τείνουν στην τιμή της μετρικής ANTT του αντίστοιχου συνδυασμού για το εκάστοτε πλήθος συσκευών.

Κεφάλαιο 6

Επίλογος

Στο κεφάλαιο αυτό παρουσιάζονται συμπεράσματα και βασικές παρατηρήσεις που έγιναν κατά τη διάρκεια της εκπόνησης της διπλωματικής εργασίας, καθώς επίσης και μελλοντικές επεκτάσεις της.

6.1 Συμπεράσματα

Το Κατανεμημένο Σύστημα Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων που αναπτύχθηκε και μοντελοποιήθηκε στο πλαίσιο της παρούσας εργασίας, παρέχει ευελιξία στον τρόπο διασύνδεσης των μοντέλων αλλά και την στρατηγική εκτέλεσής τους μέσω παραμέτρων όπως το μέγεθος δέσμης, το κριτήριο προώθησης στον εξυπηρετητή, οι τύποι και το πλήθος των συσκευών του συστήματος.

Μέσω των μετρήσεων που πάρθηκαν από την προσομοίωση ενός τέτοιου συστήματος, έγινε αντιληπτό ότι η ύπαρξη του δύναται από τη μία να αυξήσει την ακρίβεια σε σχέση με το αν υπήρχε μόνο το μοντέλο της συσκευής, και από την άλλη να επιταχύνει την συμπερασματολογία σε σχέση με την αποστολή των εικόνων σε εξυπηρετητή. Αυτές του οι ιδιότητες σε συνδυασμό με τη δυναμικότητα που του προσφέρει η παραμετροποίηση που επιδέχεται, καθιστούν το εν λόγω σύστημα ιδανικό για εφαρμογές στις οποίες υπάρχει διακύμανση στην κατάσταση του δικτύου και στην κατανομή του πλήθους και των χαρακτηριστικών των συνδεδεμένων συσκευών, ώστε να μπορεί να ανταποκρίνεται στις απαιτήσεις του χρήστη αποτελεσματικά.

Άξιο αναφοράς είναι επίσης ότι για συγκεκριμένους συνδυασμούς μοντέλων εξυπηρετητή-συσκευής, παρατηρείται αύξηση της ακρίβειας ακόμα και σε σχέση με το μοντέλο του εξυπηρετητή. Πιο συγκεκριμένα, για τον συνδυασμο NASNet Large - EfficientNet B3 επιτεύχθηκε αύξηση της ακρίβειας της συμπερασματολογίας σε 83.1% από 81.7% αν εκτελούνταν μόνο στη συσκευή και από 82.6% αν εκτελούνταν μόνο στον εξυπηρετητή σε λιγότερο μάλιστα χρόνο, και για τον συνδυασμό EfficientNet v2 M - EfficientNet B3 σε 84.7% από 81.7% αν εκτελούνταν μόνο στη συσκευή ενώ με συμπερασματολογία αποκλειστικά στο μοντέλο του εξυπηρετητή επιτυγχάνεται περίπου η ίδια ακρίβεια αλλά σε αισθητά μεγαλύτερο χρόνο. Ο τελευταίος συνδυασμός, εκτός από την αξιοσημείωτη ακρίβεια, ξεχωρίζει και για την αποδοτικότητά του σε σχέση με μετρικές που αφορούν την επιβάρυνση και την καθυστέρηση του συστήματος.

6.2 Μελλοντική Εργασία

Σε αυτή τη διπλωματική εργασία τέθηκαν τα θεμέλια για τη δημιουργία ενός ολοκληρωμένου και πλήρως παραμετροποιήσιμου Κατανεμημένου Συστήματος Συμπερασματολογίας Ζεύγους Νευρωνικών Δικτύων. Παρόλα αυτά υπάρχουν κατευθύνσεις που δεν έχουν διερευνηθεί επαρκώς, και άλλες που δεν λήφθηκαν γενικότερα υπόψιν. Αναλυτικότερα:

- Πειραματική προσομοίωση του συστήματος και για την περίπτωση στην οποία οι εικόνες δεν είναι αποθηκευμένες στις συσκευές και επομένως ο ρυθμός άφιξης των εικόνων σε αυτές να μην είναι σταθερός

- Μελέτη της επίδρασης του μεγέθους δέσμης της εισόδου που δέχεται το μοντέλο του εξυπηρετητή

  Γνωρίζουμε ότι, λόγω παραλληλοποίησης, η δημιουργία δεσμών επιταχύνει την εξαγωγή συμπερασματολογίας. Από την άλλη, το σύστημα μπορεί να καθυστερεί λόγω αναμονής μέχρι να σχηματιστεί η δέσμη. Επομένως θα ήταν χρήσιμο να γίνει μια διερεύνηση σε σχέση με την επίδραση της δημιουργίας δεσμών στον χρόνο συμπερασματολογίας και τον βέλτιστο τρόπο με τον οποίο δύναται να γίνει η επιλογή του μεγέθους της δέσμης.

- Μελέτη του πλήθους των εικόνων που στέλνονται στον εξυπηρετητή για διαφορετικά κατώφλια

  Με αυτόν τον τρόπο θα μπορούσε να διερευνηθεί περαιτέρω η συμφόρηση στον εξυπηρετητή και να γίνει καλύτερη, ενδεχομένως, επιλογή κατωφλιών με βάση το πλήθος εικόνων που στέλνονται κατά μέσο όρο ώστε να υπάρχει και μια εποπτία σε σχέση με τη συμφόρηση στον δίαυλο.

- Διερεύνηση της επίδρασης της καθυστέρησης της μεταφοράς των εικόνων από τις συσκευές στον εξυπηρετητή και μοντελοποίησή της

  Στην εργασία αυτή, στα πλαίσια της μοντελοποίησης, ο χρόνος μεταφοράς θεωρήθηκε σταθερός, μια υπόθεση που όμως δεν αντιστοιχεί σε πραγματικές συνθήκες. Επιπλέον, δεν συμπεριλήφθηκε καθόλου στην προσομοίωση του συστήματος.

- Μοντελοποίηση του χρόνου απόκρισης του συστήματος λαμβάνοντας υπόψιν περισσότερες παραμέτρους, ενδεχομένως χωρίς απλοποίηση του συστήματος, με σκοπό την κατασκευή πληρέστερης αντικειμενικής συνάρτησης για το πρόβλημα βελτιστοποίησης που θέτει το σύστημα

- Ανάπτυξη του συστήματος κάνοντας χρήση πραγματικών συσκευών

  Στην εργασία αυτή, όλες οι μετρήσεις έγιναν στον εξυπηρετητή, κάνοντας κλιμάκωση του χρόνου ώστε αυτός να προσομοιάζει τον χρόνο που θα προέκυπτε από πραγματικές συσκευές. Η σχέση όμως ανάμεσα στον χρόνο που χρειάζεται ο εξυπηρετητής και η συσκευή δεν είναι απαραίτητα γραμμική.

- Αναζήτηση μεθόδων για επίτευξη ιδιωτικότητας των δεδομένων των συσκευών του συστήματος

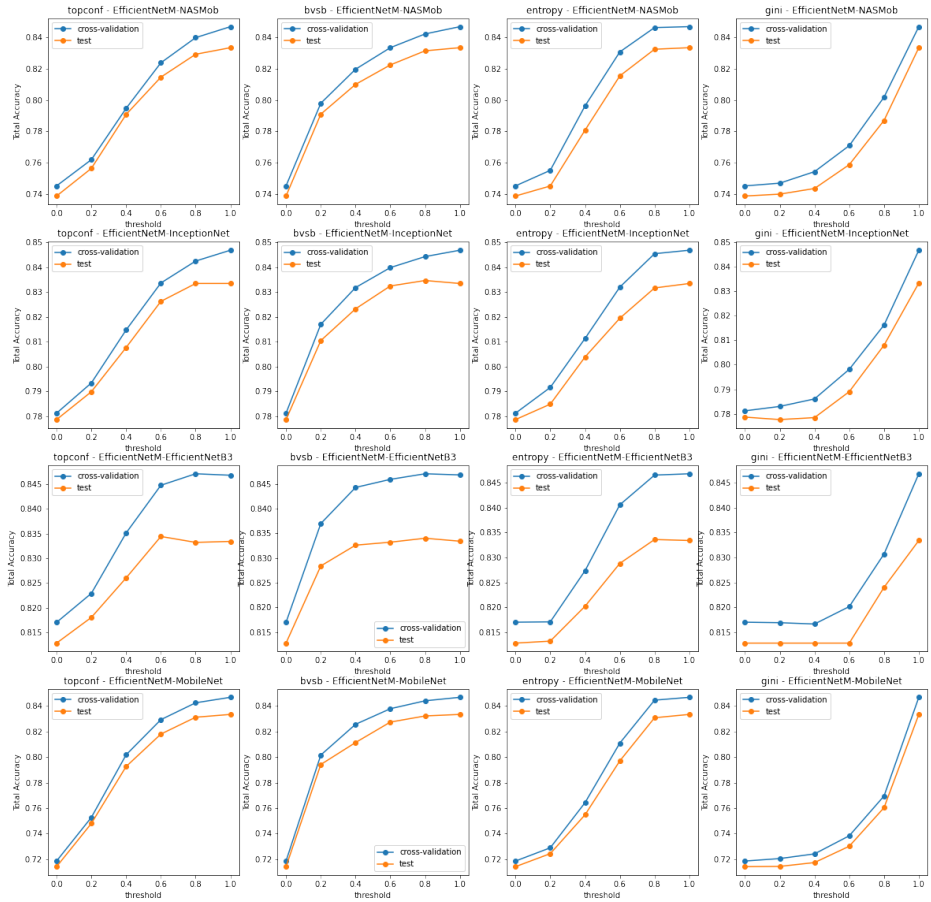
Παραρτήματα

Παράρτημα Α΄

Απόκλιση Μεθόδου cross-validation

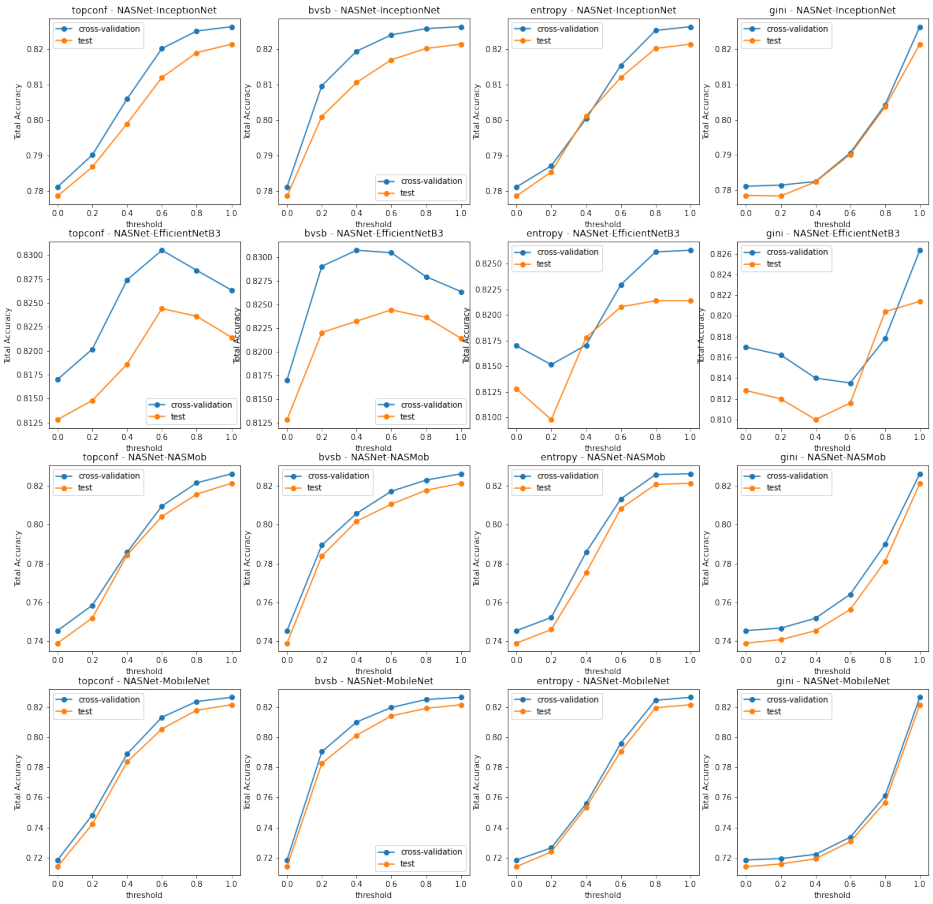
Παρακάτω παρουσιάζονται οι αποκλίσεις του μέσου όρου των μετρήσεων στα 3 υποσύνολα του συνόλου επικύρωσης του ImageNet από τις αντίστοιχες μετρήσεις στο τέταρτο υποσύνολο του ImageNet που χρησιμοποιήθηκε ως σύνολο ελέγχου. Σε όλες τις περιπτώσεις παρατηρούμε ότι η απόκλιση των μετρήσεων που πάρθηκαν με χρήση της μεθόδου δεν απέχουν σημαντικά από τις αντίστοιχες μετρήσεις ενός τυχαίου συνόλου.

Α΄.1 Μοντέλο Εξυπηρετητή EfficientNet v2 M



Εικόνα Α΄.1: *Απόκλιση μεθόδου cross-validation - Εξυπηρετητής EfficientNet v2 M*

Α΄.2 Μοντέλο Εξυπηρετητή NASNet Large

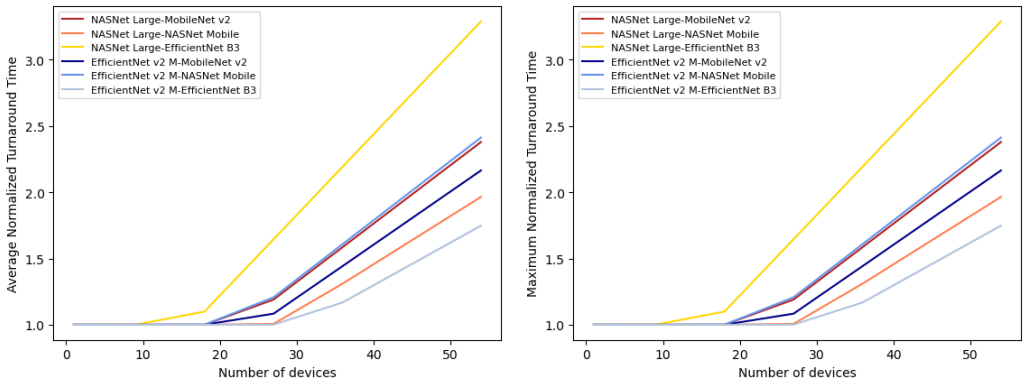


Εικόνα Α΄.2: *Απόκλιση μεθόδου cross-validation - Εξυπηρετητής NASNet Large*

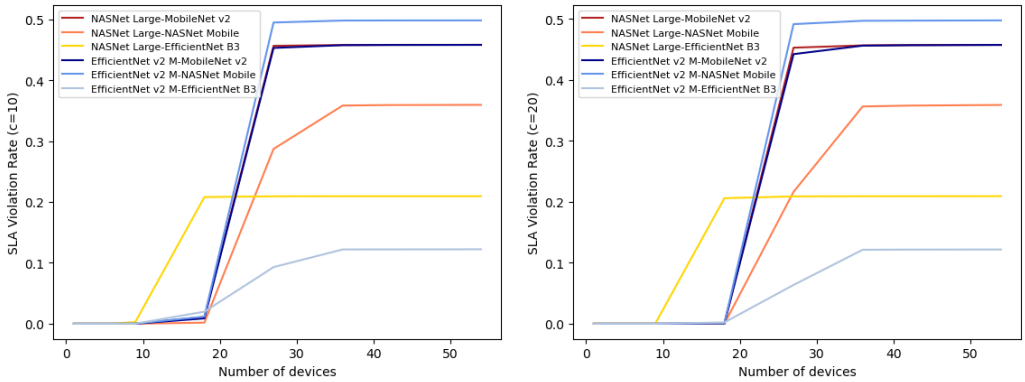
Παράρτημα Β′

Μετρικές Απόδοσης Συστήματος (ανοχή 1 p.p.)

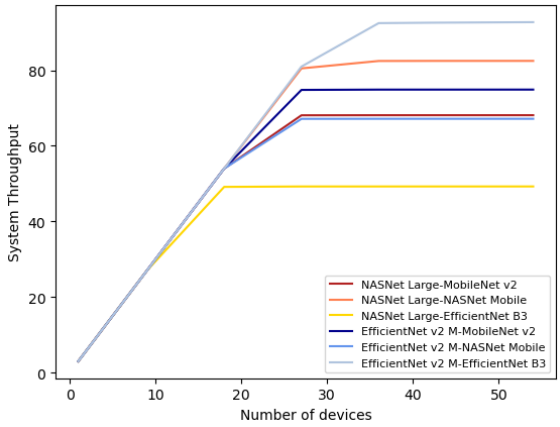
Παρακάτω παρουσιάζονται οι γραφικές παραστάσεις των μετρικών απόδοσης του συστήματος με βάση τις μετρήσεις για πλήθη συσκευών που ανήκουν στο σύνολο {1,2,3,4,6,9,18,27,36,54} χρησιμοποιώντας ως κριτήριο προώθησης στον εξυπηρετητή την τιμή της μετρικής BvSB της τοπικής συμπερασματολογίας της κάθε εικόνας, με κατώφλι που καθορίστηκε στην Ενότητα 4.4 για ανοχή στην πτώση ακρίβειας της τάξεως της 1 ποσοστιαίας μονάδας (1 p.p.).



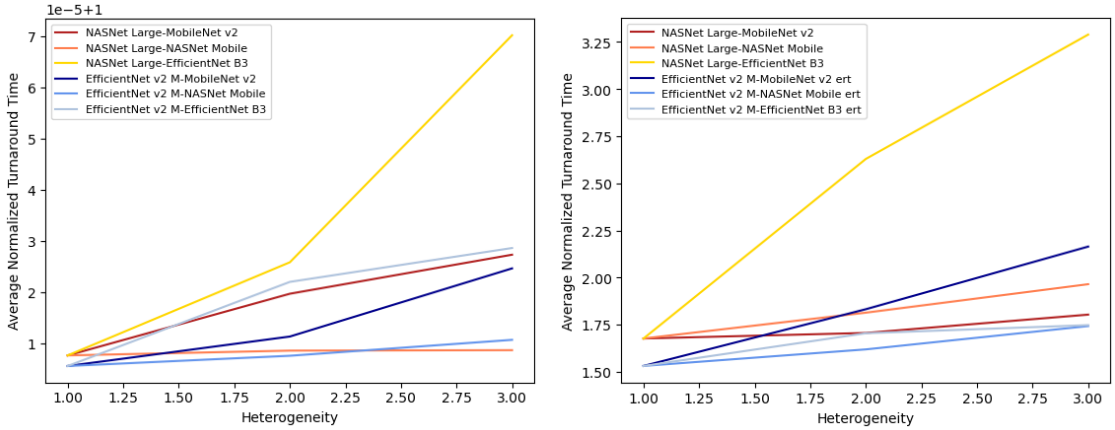
Εικόνα Β′.1: *ANTT και MNTT συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*



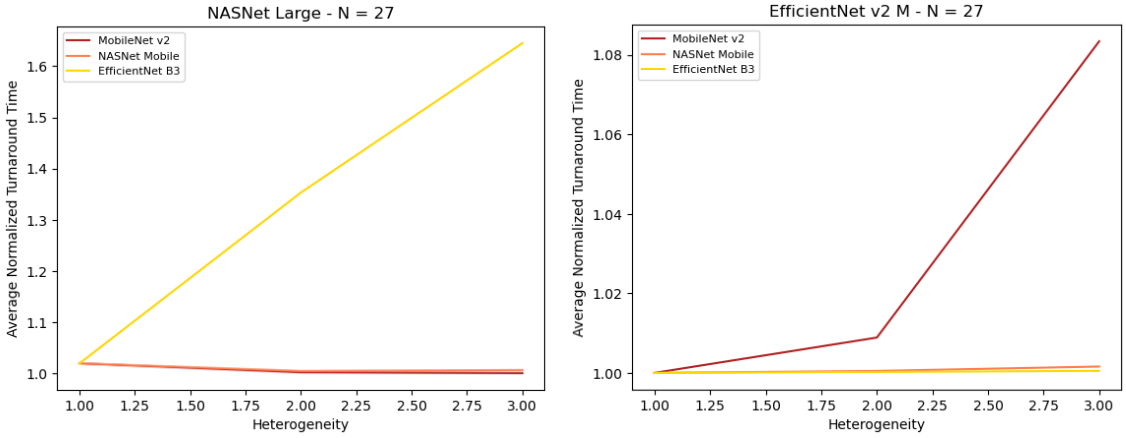
Εικόνα Β′.2: *Service Level Agreement Violation Rates συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*



Εικόνα Β΄.3: *System Throughput συστημάτων αυξανόμενου πλήθους συσκευών - απολύτως ομοιογενών*



Εικόνα Β΄.4: *Επιρροή της μείωσης της ετερογένειας του συστήματος στη μετρική ANTT (Αριστερά: Πλήθος Συσκευών = 9, Δεξιά: Πλήθος Συσκευών = 54)*



Εικόνα Β΄.5: *Επιρροή της μείωσης της ετερογένειας του συστήματος στη μετρική ANTT για πλήθος συσκευών ίσο με 27*

Όπως παρατηρούμε ισχύουν οι ίδιες παρατηρήσεις που έγιναν στην Ενότητα 5.2, με μόνη διαφορά να αποτελεί η σχέση μεταξύ των συνδυασμών μοντέλων εξυπηρετητή-συσκευής.

Παρακάτω παρουσιάζεται η ακρίβεια συνδυασμών μοντέλων εξυπηρετητή-συσκευής για επιλογή κατωφλιών για ανοχή στην πτώση ακρίβειας της τάξεως της 1 ποσοστιαίας μονάδας (1 p.p.)

Πίνακας Β′.1: *Ακρίβεια συνδυασμών μοντέλων εξυπηρετητή-συσκευής για επιλογή κατωφλιών με 1p.p. ανοχή*

Μοντέλο Εξυπηρετητή	**Μοντέλο Συσκευής**	**Κατώφλι 1 p.p.**	**Ακρίβεια**
NASNet Large	MobileNet v2	0.6	81.96%
NASNet Large	NASNet Mobile	0.6	81.73%
NASNet Large	EfficientNet B3	0.4	83.07%
EfficientNet v2 M	MobileNet v2	0.6	83.79%
EfficientNet v2 M	NASNet Mobile	0.8	84.22%
EfficientNet v2 M	EfficientNet B3	0.2	83.70%

Βιβλιογραφία

[1] *Libraries: Michigan State University, Neuroscience.* `openbooks.lib.msu.edu/neuroscience/chapter/the-neuron/`. Ημερομηνία πρόσβασης: 22-1-2023.

[2] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang και Joel S. Emer. *Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proc. IEEE*, 105(12):2295–2329, 2017.

[3] Stuart Russell και Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3η έκδοση, 2010.

[4] M.A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. `neuralnetworksanddeeplearning.com/`.

[5] Aston Zhang, Zachary C. Lipton, Mu Li και Alexander J. Smola. *Dive into Deep Learning. arXiv preprint arXiv:2106.11342*, 2021.

[6] Dargan Shaveta, Kumar Munish, Ayyagari Maruthi Rohit και Kumar Gulshan. *A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. Archives of Computational Methods in Engineering*, 27:1071–1092, 2020.

[7] Aston Zhang, Zachary C. Lipton, Mu Li και Alexander J. Smola. *Dive into Deep Learning. CoRR*, abs/2106.11342, 2021.

[8] Mayank Mishra. *Convolutional Neural Networks, Explained.* `www.towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939`, 2020.

[9] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen και Jan S. Rellermeyer. *A Survey on Distributed Machine Learning. ACM Comput. Surv.*, 53(2):30:1–30:33, 2021.

[10] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo και Junshan Zhang. *Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. Proc. IEEE*, 107(8):1738–1762, 2019.

[11] A. L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development*, 3(3):210–229, 1959.

[12] Mehryar Mohri, Afshin Rostamizadeh και Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2η έκδοση, 2018.

[13] Kevin P. Murphy. *Machine learning : A Probabilistic Perspective*. The MIT Press, Cambridge, Mass. [u.a.], 2012.

[14] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. `www.deeplearningbook.org`.

[15] *Notes for the Stanford CS class "CS231n: Convolutional Neural Networks for Visual*. `cs231n.github.io/`. Ημερομηνία πρόσβασης: 23-1-2023.

[16] Pariwat Ongsulee. *Artificial intelligence, machine learning and deep learning. 2017 15th International Conference on ICT and Knowledge Engineering (ICTKE)*, σελίδες 1–6, 2017.

[17] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do και Kaori Togashi. *Convolutional neural networks: an overview and application in radiology. Insights into Imaging*, 9(4):611–629, 2018.

[18] Arohan Ajit, Koustav Acharya και Abhishek Samanta. *A Review of Convolutional Neural Networks. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, σελίδες 1–5, 2020.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke και Andrew Rabinovich. *Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1–9, 2015.

[20] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto και Hartwig Adam. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR*, abs/1704.04861, 2017.

[21] B. Zoph, V. Vasudevan, J. Shlens και Q. V. Le. *Learning Transferable Architectures for Scalable Image Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 8697–8710, Los Alamitos, CA, USA, 2018. IEEE Computer Society.

[22] Mingxing Tan και Quoc Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning*, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 6105–6114. PMLR, 2019.

[23] Y. Lecun, L. Bottou, Y. Bengio και P. Haffner. *Gradient-based learning applied to document recognition. Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[24] Alex Krizhevsky, Ilya Sutskever και Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25*.

[25] 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.

[26] 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)*, σελίδες 770–778, 2016.

[27] Ben Lutkevich. *TechTarget*. `www.techtarget.com/searchmobilecomputing/definition/nomadic-computing`, 2022. Ημερομηνία πρόσβασης: 20-2-2023.

[28] P. J. Escamilla-Ambrosio, A. Rodríguez-Mota, E. Aguirre-Anaya, R. Acosta-Bermejo και M. Salinas-Rosales. *Distributing Computing in the Internet of Things: Cloud, Fog and Edge Computing Overview*, σελίδες 87–115. Springer International Publishing, Cham, 2018.

[29] Digiteum Team. *Digiteum*. `www.digiteum.com/cloud-fog-edge-computing-iot/#4`, 2022.

[30] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li και Lanyu Xu. *Edge Computing: Vision and Challenges. IEEE Internet of Things Journal*, 3(5):637–646, 2016.

[31] Renjie Gu, Chaoyue Niu, Fan Wu, Guihai Chen, Chun Hu, Chengfei Lyu και Zhihua Wu. *From Server-Based to Client-Based Machine Learning: A Comprehensive Survey. ACM Comput. Surv.*, 54(1), 2021.

[32] Dianlei Xu, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma και Pan Hui. *A Survey on Edge Intelligence. CoRR*, abs/2003.12172, 2020.

[33] Yunbin Deng. *Deep learning on mobile devices: a review. Mobile Multimedia/Image Processing, Security, and Applications 2019*, τόμος 10993, σελίδα 109930A. International Society for Optics and Photonics, SPIE, 2019.

[34] Min Li, Yu Li, Ye Tian, Li Jiang και Qiang Xu. *AppealNet: An Efficient and Highly-Accurate Edge/Cloud Collaborative Architecture for DNN Inference. 58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021*, σελίδες 409–414. IEEE, 2021.

[35] Alexandros Kouris, Stylianos I. Venieris και Christos Savvas Bouganis. *CascadeCNN: Pushing the Performance Limits of Quantisation in Convolutional Neural Networks. 28th International Conference on Field Programmable Logic and Applications (FPL)*, σελίδες 155–1557, 2018.

[36] Guido Van Rossum και Fred L Drake. *An introduction to Python*. Network Theory Ltd. Bristol, 2003.

[37] *General Python FAQ*. `docs.python.org/3/faq/general.html`. Ημερομηνία πρόσβασης: 5-2-2023.

[38] Yue Zhang. *An Introduction to Python and Computer Programming*, σελίδες 1–11. Springer Singapore, Singapore, 2015.

[39] *Kaggle*. `www.kaggle.com/`. Ημερομηνία πρόσβασης: 5-2-2023.

[40] *TensorFlow*. `github.com/tensorflow/tensorflow`. Ημερομηνία πρόσβασης: 5-2-2023.

[41] *TensorFlow Hub*. `www.tensorflow.org/hub`. Ημερομηνία πρόσβασης: 5-2-2023.

[42] *Keras*. `keras.io/`. Ημερομηνία πρόσβασης: 5-2-2023.

[43] *ImageNet*. `www.image-net.org/about.php`. Ημερομηνία πρόσβασης: 5-2-2023.

[44] Y. Choi, Y. Kim και M. Rhu. *Lazy Batching: An SLA-aware Batching System for Cloud Machine Learning Inference*. *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, σελίδες 493–506, Los Alamitos, CA, USA, 2021. IEEE Computer Society.

[45] 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 - Volume 70*, ICML'17, σελίδα 1321–1330. JMLR.org, 2017.

[46] Α. Τζεβαχιρίδης. *Αντιστάθμιση ακρίβειας σε κατανεμημένα συστήματα βαθέων νευρωνικών δικτύων*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2021.

[47] Moshe Zukerman. *Introduction to Queueing Theory and Stochastic Teletraffic Models*, 2013.

[48] Stefanos Laskaridis, Stylianos I. Venieris, Hyeji Kim και Nicholas D. Lane. *HAPI: Hardware-Aware Progressive Inference*. *Proceedings of the 39th International Conference on Computer-Aided Design*, ICCAD '20, New York, NY, USA, 2020. Association for Computing Machinery.

Απόδοση Ξενόγλωσσων Όρων

Απόδοση	Ξενόγλωσσος όρος
Τεχνητή Νοημοσύνη	Artificial Intelligence
Βαθιά Μάθηση	Deep Learning
Κινητός Υπολογισμός	Mobile Computing
κινητή συσκευή	mobile device
Διαδίκτυο των Πραγμάτων	Internet of Things
συμπερασματολογία	inference
Κατανεμημένη Μηχανική Μάθηση	Distributed Machine Learning
ζεύγος Νευρωνικών Δικτύων	Neural Network pair
μοντέλο	model
εξυπηρετητής	server
άκρα του δικτύου	network edge
νέφος	cloud
είσοδος	input
πρόβλεψη	prediction
εμπιστοσύνη	confidence
αρχιτεκτονική	architecture
Βαθιά Νευρωνικά Δίκτυα	Deep Neural Networks
Μηχανική Μάθηση	Machine Learning
εκπαίδευση	training
Επιβλεπόμενη Μάθηση	Supervised Learning
Μη Επιβλεπόμενη Μάθηση	Unsupervised Learning
Ενισχυτική Μάθηση	Reinforcement Learning
σύνολο δεδομένων	dataset
σύνολο εκπαίδευσης	training set
χαρακτηριστικό	feature
ετικέτα	label
επόπτης	supervisor
ταξινόμηση	classification
παλινδρόμηση	regression
κλάση	class
ομαδοποίηση	clustering
κ-Πλησιέστεροι Γείτονες	k-Nearest Neighbours
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines
Αλγόριθμος κ-Μέσων	k-means algorithm

Ανάλυση σε Κύριες Συνιστώσες	Principal Components Analysis
Τεχνητά Νευρωνικά Δίκτυα	Artificial Neural Networks
νευρώνας	neuron
δενδρίτης	dendrite
άξονας	axon
συνάρτηση ενεργοποίησης	activation function
σύναψη	synapse
βάρος	weight
μεροληψία	bias
πλήρως συνδεδεμένο	fully-connected
στρώμα	layer
δεδομένα	data
στρώμα εισόδου	input layer
στρώμα εξόδου	output layer
κρυφό στρώμα	hidden layer
Μονάδα Επεξεργασίας	Processing Unit
παράμετρος	parameter
Συνελικτικά Νευρωνικά Δίκτυα	Convolutional Neural Networks
δεκτικό πεδίο	receptive field
πυρήνας	kernel
φίλτρο	filter
συνέλιξη	convolution
κατηγοριοποίηση εικόνας	image classification
κέντρο δεδομένων	data center
Υπολογισμός στο Νέφος	Cloud Computing
Υπολογισμός 'Ομίχλης'	Fog Computing
Υπολογισμός στα ' Ακρα' του Δικτύου	Edge Computing
Τεχνητή Νοημοσύνη στα άκρα του δικτύου	Edge Intelligence
μεταφόρτωση	offloading
κόμβος	node
έυρος ζώνης	bandwidth
Κατανεμημένη Εκπαίδευση	Distributed Training
παραλληλισμός σε επίπεδο δεδομένων	data parallel
παραλληλισμός σε επίπεδο μοντέλου	model parallel
συγκεντρωτική	centralized
αποκεντρωμένη	decentralized
υβριδική	hybrid
ιδιωτικότητα	privacy
χρόνος απόκρισης	latency
Συνεργατική Μάθηση	Federated Learning
Μεταφορά Γνώσης	Transfer Learning
Διαχωρισμός Βαθέως Νευρωνικού Δικτύου	DNN Splitting
Κατανεμημένη Συμπερασματολογία	Distributed Inference

Συμπίεση Μοντέλου	Model Compression
Κλάδεμα Βαρών	pruning
Κβαντοποίηση	quantization
Διαμοιρασμός Μοντέλου	model partition
Πρόωρη Εξοδος	Early Exit
δέσμη	batch
μέγεθος δέσμης	batchsize
κατώφλι	threshold
εντροπία	entropy
ρυθμός διέλευσης	throughput