ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

mfol.ece.ntua.gr
icbnet.ntua.gr

ΟΚΤΩΒΡΙΟΣ 2023

Για περισσότερες πληροφορίες επικοινωνήστε στα dkaklam@mail.ntua.gr (Καθ. Δ.-Θ. Κακλαμάνη, Θέματα 1-3), venieris@cs.ntua.gr (Καθ. Ι. Στ. Βενιέρης, Θέματα 4-9).

1. Αντιμετώπιση προβλημάτων Δυναμικής Ανάθεσης Ραδιοπόρων σε Δίκτυα Επόμενης Γενιάς (Beyond 5G) με Συνεργατική Μάθηση (Federated Learning - FL) (1 Άτομο)

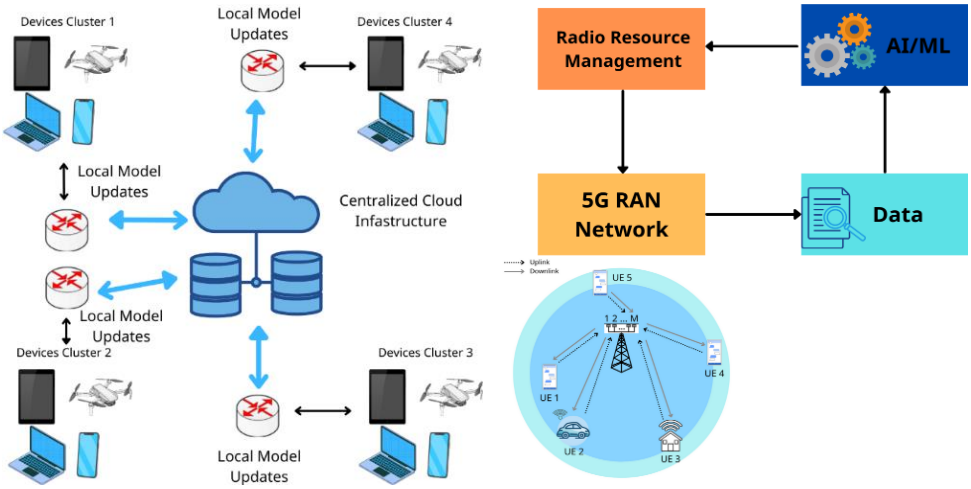
Ο μεγάλος αριθμός χρηστών σε δίκτυα νέας γενιάς (Beyond 5G - B5G) και η ολοένα αυξανόμενη απαίτησή τους για υψηλούς ρυθμούς μετάδοσης και υψηλά επίπεδα ποιότητας υπηρεσίας και εμπειρίας (Quality of Service - QoS και Quality of Experience - QoE) επιβάλλουν την ανάπτυξη προηγμένων μεθόδων πρόσβασης στο φυσικό μέσο και αποδοτικών τεχνικών μετάδοσης δεδομένων. Συγκεκριμένα, η χρήση πολύ μεγάλου πλήθους κεραιών στο σταθμό βάσης (Massive Multiple Input Multiple Output - mMIMO), η προσαρμοστική κωδικοποίηση (Adaptive Modulation Coding - AMC), τα προηγμένα σχήματα πολλαπλής πρόσβασης στο μέσο (όπως η μη-ορθογώνια πολλαπλή πρόσβαση (Non-Orthogonal Multiple Access - NOMA) και οι τοπολογίες χωρίς κυψέλες (cell-free topologies) χρίζουν περαιτέρω αξιοποίησης.

Στο ίδιο πλαίσιο, η χρήση κόμβων αναμετάδοσης (Relay Nodes - RN), αποσκοπεί στην εύκολη επεκτασιμότητα και διεύρυνση του δικτύου, χωρίς την εγκατάσταση επιπλέον σταθμών βάσης (Base Station - BS). Όταν δε η ενεργοποίηση RN συνδυάζεται με τις προαναφερθείσες τεχνολογίες, οδηγεί σε αυξημένη ενεργειακή και φασματική απόδοση, αλλά και στη δημιουργία πλήθους ασυσχέτιστων καναλιών, τα οποία μπορούν να διατεθούν σε διαφορετικές ομάδες χρηστών. Συνεπώς, ο αριθμός τον εξυπηρετούμενων χρηστών και η περιοχή κάλυψης αυξάνονται, διατηρώντας σταθερά επίπεδα QoS. Η Μηχανική Μάθηση (Machine Learning - ML) υπόσχεται ακόμα μεγαλύτερα οφέλη, χάρις στην ικανότητά της να επιλύει πολυπαραμετρικά προβλήματα, με ταυτόχρονη μείωση της υπολογιστικής πολυπλοκότητας. Ωστόσο, οι κλασσικές ML τεχνικές δεν συνεκτιμούν την εισαγωγή μεγάλης υπολογιστικής πολυπλοκότητας λόγω της πολυπαραμετρικής φύσης των προβλημάτων ανάθεσης ραδιοπόρων (Radio Resource Management - RRM) σε B5G συστήματα. Η Συνεργατική Μάθηση (Federated Learning - FL) είναι μια ML τεχνική, η οποία εκτελεί ML εργασίες (εκπαίδευση, εκτέλεση τμημάτων αλγορίθμων) σε πολλαπλές αποκεντρωμένες δικτυακές τοποθεσίες (συσκευές, εξυπηρετητές), σε κάθε μία από τις οποίες διατηρούνται τοπικά δεδομένα (local datasets). Με αυτόν τον τρόπο επιτυγχάνεται ο διαμοιρασμός του υπολογιστικού φόρτου όσον αφορά τα RRM προβλήματα σε B5G δίκτυα.

Στόχος της παρούσης διπλωματικής εργασίας είναι η μελέτη και αξιοποίηση FL αλγορίθμων σε υποδομές B5G για την αποδοτικότερη και δυναμική αντιμετώπιση RRM προβλημάτων (όπως η κατανομή υποφερόντων σε χρήστες, η τοποθέτηση και επιλογή RN, κ.α.). Κύριες μετρικές απόδοσης των παραπάνω αλγορίθμων θα αποτελέσουν η ενεργειακή και φασματική αποδοτικότητα (Energy Efficiency - EE και Spectral Efficiency - SE).

Η εκπόνηση της διπλωματικής εργασίας περιλαμβάνει, αρχικά, μελέτη τόσο του θεωρητικού υπόβαθρου φυσικού επιπέδου των συστημάτων B5G, όσο και των βασικών αρχών και τεχνικών FL. Εν συνεχεία, ο σπουδαστής θα εφαρμόσει μεθόδους και αλγορίθμους FL, παραμετροποιώντας B5G προσομοιωτές επιπέδου ζεύξης, οι οποίοι έχουν ήδη υλοποιηθεί στο εργαστήριο. Η διπλωματική εργασία θα

ολοκληρωθεί με τη συγκριτική αποτίμηση των χρησιμοποιούμενων παραλλαγών των FL μοντέλων, καθώς και με σύγκριση αυτών με ένα σύστημα αναφοράς χωρίς χρήση ML.



Απαραίτητες γνώσεις: *Βασικές γνώσεις κινητών επικοινωνιών, Βασικές γνώσεις Python*
Επιθυμητές γνώσεις: *Αρχές και αρχιτεκτονικές Μηχανικής Μάθησης, MATLAB, Python βιβλιοθήκες (Keras/TensorFlow)*

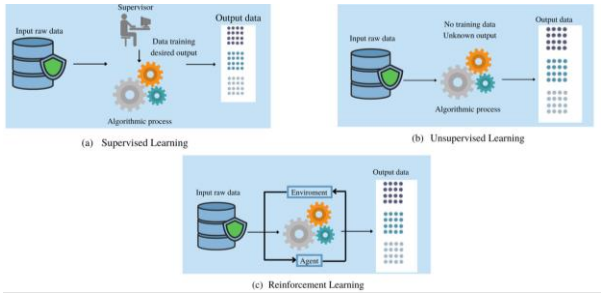
2. *Βελτιστοποίηση τεχνικών Ασφάλειας Φυσικού Στρώματος (Physical Layer Security) με τεχνικές Μηχανικής Μάθησης (Machine Learning - ML) σε Ετερογενή Δίκτυα επόμενης γενιάς (Beyond 5G) (1 Άτομο)*

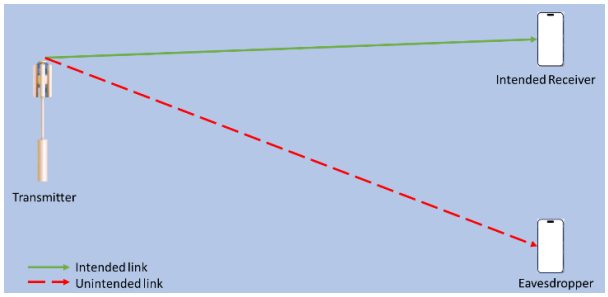
Η γεωμετρική αύξηση της ταυτόχρονης παρουσίας ενεργών χρηστών σε ασύρματα δίκτυα νέας γενιάς (Beyond 5G - B5G), καθώς και οι ολοένα αυξανόμενες απαιτήσεις τους για υψηλούς ρυθμούς μετάδοσης και ελάχιστη καθυστέρηση, καθιστούν αναγκαία την ανάπτυξη προηγμένων μεθόδων πρόσβασης στο φυσικό μέσο και αποδοτικών τεχνικών μετάδοσης δεδομένων για τη διασφάλιση της ποιότητας υπηρεσίας και εμπειρίας (Quality of Service - QoS και Quality of Experience - QoE), αλλά και στην αύξηση της φασματικής απόδοσης των B5G συστημάτων.

Ωστόσο, εξαιτίας της μετάδοσης μέσω της διεπαφής του αέρα, το ασύρματο μέσο μεταφοράς -ειδικά σε περιβάλλονται μαζικής πρόσβασης όπως τα B5G- είναι προσβάσιμο τόσο σε εξουσιοδοτημένους όσο και σε μη-νόμιμους χρήστες. Ως εκ τούτου, το ανοικτό περιβάλλον επικοινωνίας καθιστά τις ασύρματες μεταδόσεις περισσότερο ευάλωτες από τις ενσύρματες σε επιθέσεις ασφαλείας. Δεδομένου ότι οι χρήστες (είτε οικιακοί είτε εταιρικοί είτε κρατικοί) βασίζονται στα δίκτυα ασυρμάτων επικοινωνιών για τη μετάδοση σημαντικών και ιδιωτικών πληροφοριών (όπως πχ. οι τραπεζικές συναλλαγές, οι πληρωμές λογαριασμών, η είσοδος σε διαδικτυακές πλατφόρμες, κ.α.), η ασφάλεια των ασύρματων επικοινωνιών (Physical Layer Security - PLS) είναι ένας τομέας κριτικής σημασίας με πολλές προκλήσεις σε B5G τοπολογίες. Αντικείμενό της είναι η αντιμετώπιση δυσμενών καταστάσεων, οι οποίες σχετίζονται με την απώλεια ακεραιότητας δεδομένων, τις υποκλοπές και παρεμβολές μεταδιδόμενης πληροφορίας, την είσοδο και μετάδοση στο δίκτυο ψευδών ή τροποποιημένων μηνυμάτων πληροφορίας και την κατασπατάληση πόρων.

Η μηχανική μάθηση (Machine Learning - ML) έχει αποδειχθεί μια αποτελεσματική λύση για τη βελτιστοποίηση απόκρισης σε πολυπαραμετρικά προβλήματα, μειώνοντας συγχρόνως σημαντικά την υπολογιστική πολυπλοκότητα. Ωστόσο, στο πεδίο των ασυρμάτων επικοινωνιών με έμφαση στο δίκτυο πρόσβασης, η ύπαρξη πολλαπλών διασυνδεδεμένων συσκευών και η πολυπλοκότητα του καναλιού μετάδοσης δυσχεραίνουν ακόμα περισσότερο το πρόβλημα της διασφάλισης του απορρήτου των επικοινωνιών σε περιβάλλοντα πολλαπλής πρόσβασης. Συνεπώς, η Βαθιά (Deep Learning - DL) προτείνεται ως η αποτελεσματικότερη κατηγορία ML αλγορίθμων για προβλήματα PLS.

Στόχος της παρούσης διπλωματικής εργασίας είναι η μελέτη και αξιοποίηση DL αλγορίθμων σε υποδομές B5G για την ανάπτυξη προηγμένων σχημάτων PLS. Η εκπόνηση της διπλωματικής εργασίας περιλαμβάνει, αρχικά, μελέτη τόσο του θεωρητικού υπόβαθρου του φυσικού επιπέδου των συστημάτων B5G, των αρχών και αλγορίθμων DL όσο και των χρησιμοποιούμενων σχημάτων PLS. Στη συνέχεια, ο σπουδαστής θα εφαρμόσει μεθόδους και αλγορίθμους DL, παραμετροποιώντας B5G προσομοιωτές επιπέδου ζεύξης, οι οποίοι έχουν ήδη υλοποιηθεί στο εργαστήριο και χρησιμοποιώντας τα αντίστοιχα σύνολα δεδομένων (datasets). Η διπλωματική εργασία θα ολοκληρωθεί με τη συγκριτική αποτίμηση των χρησιμοποιούμενων παραλλαγών των DL μοντέλων για PLS. Θα χρησιμοποιηθούν τόσο ML μετρικές (accuracy, RMSE, f1-score, κ.λπ.) όσο και δικτυακές μετρικές ασφάλειας (secrecy throughput, secrecy capacity, Quality of Security (QoSec), κ.λπ.).





Απαραίτητες γνώσεις: *Βασικές γνώσεις κινητών επικοινωνιών, Βασικές γνώσεις Python*
Επιθυμητές γνώσεις: *Αρχές και αρχιτεκτονικές Μηχανικής Μάθησης, MATLAB, Python βιβλιοθήκες (Keras/TensorFlow)*

3. Πλάνο Διαχείρισης Μη-Ορθογώνιων Πόρων σε Ετερογενή Κατανεμημένα massive MIMO Συστήματα (1 Άτομο)

Οι τεχνικές πολλαπλής πρόσβασης επιτρέπουν σε πολλούς τελικούς χρήστες να χρησιμοποιούν τους ίδιους πόρους για τη λήψη πληθώρας υπηρεσιών. Οι προγενέστερες τεχνολογικές γενιές κυψελωτών δικτύων (1G-4G) χαρακτηρίζονται από την ορθογωνιότητα μεταξύ των σημάτων, κατανέμοντας τους διαθέσιμους πόρους (frequency, time, code, space) σε διαφορετικούς τελικούς χρήστες. Η νέα τεχνολογική γενιά των ασύρματων δικτύων 5^{ης} γενιάς (5G) αναμένεται να υποστηρίζει έναν ακόμη μεγαλύτερο αριθμό συνδέσεων διαφορετικών απαιτήσεων (throughput, latency) και γενικά να παρέχει υπηρεσίες σε δίκτυα εκατονταπλάσιας σχεδόν πυκνότητας σε σχέση με την 4G. Για να ικανοποιηθούν αυτού του είδους οι απαιτήσεις, τα 5G κυψελωτά δίκτυα υιοθετούν νέες τεχνολογίες, οι οποίες έχουν αναπτυχθεί την τελευταία δεκαετία. Μεταξύ αυτών συγκαταλέγεται η μη-ορθογώνια πολλαπλή πρόσβαση (Non Orthogonal Multiple Access - NOMA). Η NOMA μπορεί να συνδυαστεί εύκολα και με άλλες υφιστάμενες αλλά και νέες τεχνολογίες, όπως αυτές των πολλαπλών κεραιοστοιχείων μεγάλης κλίμακας (massive MIMO) και των επικοινωνιών χιλιοστομετρικής μετάδοσης (mmWave), με σκόπο την αύξηση της απόδοσης του συστήματος γενικότερα.

Συγκεκριμένα, οι κεραιοστοιχείες μεγάλης κλίμακας οι οποίες αποτελούνται από δεκάδες εκατοντάδες/χιλιάδες κεραιοστοιχεία στο σταθμό βάσης, αυξάνουν το πλήθος των εξυπηρετούμενων χρηστών και εξομαλύνουν τις H/M ομοδιαυλικές παρεμβολές. Από την άλλη, τα ετερογενή δίκτυα (HetNets) ενσωματώνουν στη δομή τους μεγάλης πυκνότητας μικρές κυψέλες, με σκοπό τη δημιουργία κοντινότερων ζεύξεων σταθμού βάσης - χρήστη, καθώς και την αποφόρτιση των μεγαλύτερων κυψελών. Αυτό έχει ως αποτέλεσμα τη μείωση της καταναλισκόμενης ισχύος, την αύξηση της χωρητικότητας και τη βελτίωση της χωρικής επαναχρησιμοποίησης συχνοτήτων.

Στα πλαίσια της διπλωματικής εργασίας, αρχικά θα σχεδιαστεί ένα υβριδικό ετερογενές δίκτυο πολλαπλών κυψελών (macro, pico) στοχαστικής γεωμετρίας. Δεδομένου ότι υπάρχει η δυνατότητα επικάλυψης της υψήλης ισχύος macro κυψέλης με χαμηλής ισχύος pico κυψέλες, (α) οι σταθμοί βάσης

των macro κυψελών θα είναι εξοπλισμένοι με massive MIMO κεραιοστοιχεία, ενώ οι σταθμοί βάσης των pico κυψελών και οι συσκευές των τελικών χρηστών θα είναι εξοπλισμένοι με μία απλή κεραία, (β) η σύνδεση των πολλαπλών χρηστών με τους pico σταθμούς βάσης θα πραγματοποιείται με μετάδοση NOMA, ενώ με τους macro σταθμούς βάσης στο ίδιο resource block (π.χ. time/frequency/code). Επίσης, στις κυψέλες υψηλής ισχύος θα υιοθετηθούν τεχνικές μετάδοσης και προεπεξεργασίας σήματος, ενώ στις κυψέλες χαμηλής ισχύος θα ενσωματωθούν και τεχνικές δίκαιης κατανομής πόρων. Η επίδοση των τεχνικών αυτών θα μελετηθεί πολύπλευρα και θα αξιολογηθεί κατόπιν αμοιβαίας σύγκρισης. Η εργασία θα ολοκληρωθεί με την συνολική αποτίμηση των αποτελεσμάτων προσομοίωσης.

Απαραίτητες γνώσεις: *Βασικές γνώσεις ασύρματων ζεύξεων και διάδοσης, MATLAB*

4. Κατασκευή Μοντέλων για την Πρόβλεψη του Χρόνου Εκτέλεσης Βαθιών Νευρωνικών Δικτύων σε Κινητές Συσκευές (1 Άτομο)

Τα βαθιά νευρωνικά δίκτυα (deep neural networks, DNNs) έχουν γίνει ο ακρογωνιαίος λίθος των σύγχρονων εφαρμογών τεχνητής νοημοσύνης (artificial intelligence, AI), που κυμαίνονται από την όραση υπολογιστών έως την επεξεργασία φυσικής γλώσσας. Η αποτελεσματική ανάπτυξη DNNs για ένα ευρύ φάσμα κινητών συσκευών αποτελεί σοβαρή πρόκληση λόγω της εγγενούς ποικιλομορφίας των χαρακτηριστικών υλικού τους, συμπεριλαμβανομένων των επεξεργαστών, της χωρητικότητας μνήμης και του χώρου αποθήκευσης. Ωστόσο, αυτή η προσπάθεια είναι υψίστης σημασίας, καθώς είναι απαραίτητη για την απελευθέρωση του πλήρους δυναμικού των εφαρμογών τεχνητής νοημοσύνης σε κινητές συσκευές.

Μία από τις βασικές μετρικές απόδοσης είναι ο χρόνος εκτέλεσης (latency), ο οποίος έχει άμεση επίδραση την ποιότητα εμπειρίας (Quality of Experience, QoS) των χρηστών. Στο πλαίσιο πολλών περιστάσεων, απαιτείται η ακριβής εκτίμηση του χρόνου εκτέλεσης ενός DNN σε διάφορους επεξεργαστές κινητών συσκευών και διαμορφώσεις. Αυτό είναι ιδιαίτερα σημαντικό, για παράδειγμα, στην βελτιστοποίηση της επιλογής του κατάλληλου μοντέλου ή σε διαδικασίες αναζήτησης αρχιτεκτονικής μοντέλου που λαμβάνουν υπόψη τα χαρακτηριστικά υλικού (hardware-aware neural architecture search).

Σκοπός της παρούσης διπλωματικής εργασίας είναι η ανάπτυξη μοντέλων που θα προβλέπουν τον χρόνο εκτέλεσης ενός DNN σε διάφορους επεξεργαστές κινητών συσκευών και διαμορφώσεις (π.χ. αριθμός νημάτων CPU, αριθμητική ακρίβεια). Τα εν λόγω μοντέλα μπορεί να ανήκουν σε δύο κατηγορίες: (α) βασισμένα σε κανόνες (rule-based), που χρησιμοποιούν προκαθορισμένες ευριστικές (heuristics) και μαθηματικά μοντέλα, ή (β) βασισμένα στη μάθηση (learning-based), όπως είναι οι αλγόριθμοι μηχανικής μάθησης και τα νευρωνικά δίκτυα.

Η διπλωματική εργασία περιλαμβάνει τα ακόλουθα στάδια: (α) ανάπτυξη mobile εφαρμογής (Android) για τη συλλογή μετρήσεων του χρόνου εκτέλεσης διαφόρων DNNs σε διάφορους επεξεργαστές κινητών συσκευών, (β) προσδιορισμός των χαρακτηριστικών που επηρεάζουν τον χρόνο εκτέλεσης, όπως η πολυπλοκότητα του μοντέλου, οι προδιαγραφές υλικού και τα χαρακτηριστικά εισόδου, (γ) ανάπτυξη των μοντέλων πρόβλεψης, και (δ) χρήση διαφόρων τεχνικών επικύρωσης (π.χ. cross-validation) για την αξιολόγηση των δυνατοτήτων γενίκευσης των μοντέλων και εκτεταμένες δοκιμές με διαφορετικά μοντέλα και διαμορφώσεις υλικού.

Απαραίτητες γνώσεις προγραμματισμού: *Python, Java*
Επιθυμητές γνώσεις: *Android mobile app development, Deep Learning frameworks (TensorFlow, TFLite)*

5. Εκπαίδευση Βαθιών Νευρωνικών Δικτύων σε Κινητές Συσκευές και Αξιολόγηση της Ποιότητας των Δεδομένων (1 Άτομο)

Η ανάπτυξη βαθιών νευρωνικών δικτύων (deep neural networks, DNNs) σε κινητές συσκευές, όπως είναι τα smartphones, IoT συσκευές και ενσωματωμένα συστήματα, έχει συγκεντρώσει αρκετό ενδιαφέρον τα τελευταία χρόνια λόγω της δυνατότητας της για ενεργοποίηση εξατομικευμένων εφαρμογών τεχνητής νοημοσύνης (artificial intelligence, AI) πραγματικού χρόνου. Ωστόσο, η επίτευξη αποτελεσματικής εκπαίδευσης σε μια συσκευή με περιορισμένους πόρους θέτει μοναδικές προκλήσεις.

Η παρούσα διπλωματική εργασία στοχεύει στη διερεύνηση της εκπαίδευσης DNNs σε κινητές συσκευές (on-device training) για ταξινόμηση εικόνων. Παραδοσιακά, το επιλεγμένο μοντέλο εκπαιδεύεται εκτός συσκευής (π.χ. σε έναν ισχυρό διακομιστή) για έναν αριθμό από κατηγορίες (κλάσεις) αντικειμένων και στη συνέχεια τοποθετείται εντός συσκευής, όπου ακολουθεί η ρύθμισή του (fine-tuning) σε έναν μικρό αριθμό επιπρόσθετων κατηγοριών (1-2).

Τα δεδομένα που θα χρησιμοποιηθούν για τη ρύθμιση του μοντέλου θα προέρχονται είτε (α) από το αρχικό σύνολο δεδομένων, είτε (β) από την ίδια την κινητή συσκευή (real-life δεδομένα). Με αυτόν τον τρόπο, γίνεται δυνατή η αξιολόγηση της ποιότητας των δεδομένων ρύθμισης και η διερεύνηση της επίδρασης των διαφορετικών κατανομών των δεδομένων στην απόδοση του μοντέλου.

Τα βήματα της διπλωματικής εργασίας περιλαμβάνουν: (α) ανάπτυξη mobile εφαρμογής (Android) για την εκπαίδευση DNNs σε κινητές συσκευές, (β) αρχική εκπαίδευση του μοντέλου εκτός συσκευής και καταγραφή των απαιτήσεων εκπαίδευσης, (γ) τοποθέτηση του μοντέλου στη συσκευή, ρύθμιση στις νέες κατηγορίες και καταγραφή των απαιτήσεων εκπαίδευσης, (δ) σύγκριση των δύο τρόπων εκπαίδευσης, και (ε) συγκριτική αξιολόγηση της ποιότητας των δύο πηγών δεδομένων ρύθμισης με βάση την τελική ακρίβεια ταξινόμησης του μοντέλου ή άλλων μετρικών ενδιαφέροντος.

Απαραίτητες γνώσεις προγραμματισμού: *Python, Java*
Επιθυμητές γνώσεις: *Android mobile app development, Deep Learning frameworks (TensorFlow, TFLite)*

6. Διασύνδεση Βαθιών Νευρωνικών Δικτύων για Αποτελεσματική Χρήση Πόρων (1 Άτομο)

Η ταχεία ανάπτυξη των βαθιών νευρωνικών δικτύων (deep neural networks, DNNs) έχει οδηγήσει στην ευρεία χρήση τους σε ποικίλες εφαρμογές που καλύπτουν διάφορα υπολογιστικά περιβάλλοντα, από διακομιστές και υπολογιστές, μέχρι κινητές συσκευές και συσκευές IoT. Ωστόσο, η ταυτόχρονη ανάπτυξη πολλαπλών DNNs (multi-DNN), ειδικά σε πλατφόρμες με περιορισμένους πόρους, όπως οι κινητές συσκευές, θέτει σημαντικές προκλήσεις λόγω περιορισμών στην υπολογιστική ισχύ, τη μνήμη, κ.λπ.

Στόχος της παρούσης διπλωματικής εργασίας είναι η διερεύνηση μιας νέα προσέγγισης που ονομάζεται «Διασύνδεση Βαθιών Νευρωνικών Δικτύων» (DNN Linking), όπου τα DNNs που πρέπει να εκτελεστούν αντιμετωπίζονται ως «μαύρα κουτιά» και ένας αριθμός από συνδέσμους (links) -που είναι μικρά νευρωνικά δίκτυα- εκπαιδεύονται ώστε να αντιστοιχίζουν αποτελεσματικά τις εξόδους τους, μειώνοντας αποτελεσματικά τις απαιτήσεις πόρων.

Για παράδειγμα, θεωρείστε μια εφαρμογή όρασης που απαιτεί την ταυτόχρονη εκτέλεση τριών δικτύων για κάθε εικόνα εισόδου, τα οποία εκτελούν τις ακόλουθες διεργασίες: (α) αναγνώριση σκηνής (scene recognition), (β) ανίχνευση αντικειμένων (object recognition), και (γ) κατάτμηση εικόνας (image segmentation). Σε αυτό το σενάριο, ένα μοντέλο διασύνδεσης θα μπορούσε να μάθει την αντιστοίχιση (mapping) μεταξύ της εξόδου του τρίτου μοντέλου (που αποτελείται από θερμικούς χάρτες ανά εικονοστοιχείο) και της εξόδου του πρώτου μοντέλου (που αποτελείται από ένα διάνυσμα πεποιθήσεων που σχετίζεται με ένα σύνολο σκηνών). Σε αυτήν τη διάταξη, το μοντέλο αναγνώρισης σκηνής αντικαθίσταται αποτελεσματικά από το (μικρότερο) μοντέλο σύνδεσης, μειώνοντας αποδοτικά τις υπολογιστικές απαιτήσεις.

Η διπλωματική εργασία περιλαμβάνει τις ακόλουθες φάσεις: (α) καθορισμός σεναρίων πραγματικών multi-DNN εφαρμογών και συλλογή των αντίστοιχων προ-εκπαιδευμένων μοντέλων, (β) κατασκευή,

εκπαίδευση και αξιολόγηση μικρών DNNs που θα λειτουργούν ως μοντέλα διασύνδεσης, και (γ) ανάπτυξη mobile εφαρμογής (Android) για την αξιολόγηση των συνδέσμων σε κινητές συσκευές.

Απαραίτητες γνώσεις προγραμματισμού: *Python, Java*
Επιθυμητές γνώσεις: *Android mobile app development, Deep Learning frameworks (TensorFlow, TFLite)*

7. Αξιολόγηση Αρχιτεκτονικών Transformers σε Κατανεμημένα Συστήματα Βαθιάς Μάθησης για Εφαρμογές Κινητών Συσκευών (1 Άτομο)

Η εξέλιξη των κινητών και IoT συσκευών τα τελευταία χρόνια έχει επιφέρει μια μεταβατική εποχή στον τομέα του υπολογισμού, με επίκεντρο τις παρυφές του δικτύου (edge). Με το μεγαλύτερο μέρος των δεδομένων να παράγεται στις παρυφές, υπάρχει αυξανόμενη ανάγκη για την τοπική επεξεργασία τους. Αυτό έχει προκαλέσει ενδιαφέρον για εκτέλεση συμπερασματολογίας (inference) μοντέλων βαθιάς μάθησης απευθείας στις συσκευές, ελαχιστοποιώντας την ανάγκη μεταφοράς και επεξεργασίας δεδομένων στο νέφος (cloud). Ωστόσο, αυτό το σενάριο απαιτεί την ανάπτυξη μοντέλων ειδικά προσαρμοσμένων στις υπολογιστικά περιορισμένες συσκευές, τα οποία έχουν μειωμένη ακρίβεια.

Για τη διατήρηση της ακρίβειας και την αποφόρτιση του υπολογιστικού φόρτου από τις συσκευές, έχει αναπτυχθεί η ιδέα της κατανεμημένης συμπερασματολογίας με υποβοήθηση από εξυπηρετητή (server) ο οποίος είναι τοποθετημένος στις παρυφές του δικτύου ώστε να βρίσκεται κοντά στις συσκευές. Μια διαδεδομένη κατανεμημένη αρχιτεκτονική είναι αυτή της αλληλουχίας (cascade) μοντέλων. Η αρχιτεκτονική αυτή βασίζεται στην ιδέα ότι το μεγαλύτερο μέρος των δεδομένων είναι απλά και μπορούν να επεξεργαστούν σωστά από ελαφριά μοντέλα με μειωμένο υπολογιστικό κόστος, ενώ τα πιο δύσκολα δείγματα μπορούν να προωθηθούν ώστε να επεξεργαστούν από ένα πιο περίπλοκο μοντέλο με αυξημένες υπολογιστικές ανάγκες, διατηρώντας έτσι υψηλή, state-of-the-art, ακρίβεια.

Κατά τη διάρκεια των τελευταίων ετών, με την κυκλοφορία εφαρμογών όπως το ChatGPT, έχει δημιουργηθεί τεράστιο ενδιαφέρον γύρω από τα μοντέλα αρχιτεκτονικής Transformers, τα οποία είναι και το κύριο συστατικό της επιτυχίας αυτών των εφαρμογών. Η ομαλή ενσωμάτωση τέτοιων μοντέλων σε κινητές και IoT συσκευές είναι καίρια για την ευρεία διάδοση των αυξημένων δυνατότητων τεχνητής νοημοσύνης που προσφέρουν. Καθώς τα μοντέλα αυτά τείνουν να είναι εξαιρετικά πολύπλοκα, δημιουργείται η ανάγκη για κατανεμημένη εκτέλεση.

Στόχοι της παρούσης διπλωματικής εργασίας αποτελούν: (α) αξιολόγηση Transformer μοντέλων ως προς την ακρίβεια και την αξιοποίηση πόρων, καθώς και η σύγκριση τους με καθιερωμένες αρχιτεκτονικές όπως τα CNNs, (β) επιλογή κατάλληλων μοντέλων για την ενσωμάτωση τους σε ένα σχήμα αλληλουχίας, (γ) βελτιστοποίηση του συστήματος ως προς την καθυστέρηση (latency), την ακρίβεια (accuracy), τη χρήση πόρων (resource utilization), κ.α., και (δ) εξερεύνηση της ανθεκτικότητας του κατανεμημένου συστήματος σε ακραίες καταστάσεις.

Απαραίτητες γνώσεις προγραμματισμού: *Python*
Επιθυμητές γνώσεις: *Deep Learning frameworks (TensorFlow, Keras, PyTorch)*

8. Ανάπτυξη Συστήματος Κατανεμημένης Μηχανικής Μάθησης για Εκφόρτωση Δεδομένων στις Παρυφές του Δικτύου (1 Άτομο)

Τα τελευταία χρόνια, η ταχεία ανάπτυξη των κινητών συσκευών σε συνδυασμό με την εξαιρετική επίδοση των βαθιών νευρωνικών δικτύων στην επίλυση πολύπλοκων προβλημάτων (κατηγοριοποίηση εικόνας, εντοπισμός αντικειμένων, αναγνώριση φωνής, μοντελοποίηση κειμένου) έχουν δημιουργήσει την ανάγκη για ευφυείς εφαρμογές κινητών συσκευών (smart mobile apps) που σέβονται την ιδιωτικότητα του χρήστη και παρέχουν την απαιτούμενη ποιότητα υπηρεσίας.

Η εκτέλεση νευρωνικών δικτύων στα πλαίσια τέτοιων εφαρμογών εμπεριέχει δύο βασικές προσεγγίσεις: (α) τοπικά, χρησιμοποιώντας τους περιορισμένους υπολογιστικούς πόρους της κινητής συσκευής του

χρήστη, και (β) στο υπολογιστικό νέφος ή στις παρυφές του δικτύου (edge) με την υποβοήθηση ενός ισχυρού εξυπηρετητή. Αν επιλεγεί η τοπική εκτέλεση, τότε το βασικό μειονέκτημα είναι ότι οι πόροι της κινητής συσκευής μπορεί να μην είναι πάντοτε επαρκείς και επομένως να μην μπορεί να διατηρηθεί η ποιότητα υπηρεσίας. Αντίθετα, με την απομακρυσμένη εκτέλεση, η επιπρόσθετη καθυστέρηση που εισάγεται λόγω της μεταφοράς των δεδομένων μπορεί να είναι απαγορευτική για την εύρυθμη λειτουργία της εφαρμογής. Μια λύση στα παραπάνω ζητήματα είναι η επιλεκτική κατανεμημένη εκτέλεση ανάλογα με τις συνθήκες και τα δυναμικά χαρακτηριστικά τόσο του απομακρυσμένου εξυπηρετητή και της κινητής συσκευής, όσο και της σύνδεσης μεταξύ τους.

Η αρχιτεκτονική της εκφόρτωσης (offloading) κάνει χρήση της τεχνικής της τομής μοντέλων, όπου γίνεται μερική εκτέλεση του μοντέλου στη συσκευή και προώθηση των ενδιάμεσων χαρακτηριστικών στον εξυπηρετητή ώστε να ολοκληρωθεί η επεξεργασία τους. Η τεχνική αυτή παρουσιάζει πολλές δυναμικές προκλήσεις, όπως είναι η επιλογή του βέλτιστου σημείου τομής ανάλογα με τις εκάστοτε συνθήκες και ανάγκες του χρήστη, ή ο βέλτιστος τρόπος συμπίεσης και κωδικοποίησης των ενδιάμεσων χαρακτηριστικών.

Η προτεινόμενη διπλωματική εργασία περιλαμβάνει τις ακόλουθες φάσεις: (α) ανάπτυξη και αξιολόγηση στρατηγικών τομής μοντέλων για την αποδοτική εκτέλεση εφαρμογών βαθιάς μάθησης, (β) εξερεύνηση και αξιολόγηση τεχνικών διατήρησης της ιδιωτικότητας των δεδομένων, (γ) αξιολόγηση της ανθεκτικότητας του συστήματος σε ακραίες καταστάσεις.

Απαραίτητες γνώσεις προγραμματισμού: *Python*
Επιθυμητές γνώσεις: *Deep Learning frameworks (TensorFlow, Keras, PyTorch)*

9. Ενίσχυση της Αποδοτικότητας Βαθιών Νευρωνικών Δικτύων μέσα από Βελτιστοποιημένες Στρατηγικές Πρόωρης Εξόδου (Early Exit) (1 Άτομο)

Η βαθιά μάθηση (deep learning, DL) έχει αναμφισβήτητα μεταμορφώσει το τοπίο της τεχνητής νοημοσύνης, οδηγώντας σε δραματικές προόδους σε ποικίλους τομείς, από την όραση υπολογιστών μέχρι την επεξεργασία φυσικής γλώσσας, αλλά και σε διαφορετικά περιβάλλοντα, που κυμαίνονται από αυτόνομα οχήματα, που απαιτούν αντίληψη σε πραγματικό χρόνο, μέχρι τις διασυνδεδεμένες συσκευές του διαδικτύου των πραγμάτων (Internet of Things, IoT). Παρόλα αυτά, οι εξαιρετικά υψηλές υπολογιστικές απαιτήσεις που συχνά έχουν τα βαθιά νευρωνικά δίκτυα (deep neural networks, DNNs) απαιτούν εκτεταμένη χρήση υπολογιστικών πόρων και οδηγούν σε χρονοβόρα συμπερασματολογία (inference). Κατά συνέπεια, προκύπτει μια ανάγκη ανάπτυξης δυναμικών στρατηγικών για την αποβάρυνση του υπολογιστικού φόρτου που δημιουργούν τα μοντέλα βαθιάς μάθησης, ειδικά σε περιβάλλοντα περιορισμένων πόρων.

Μια οικογένεια στρατηγικών που έχει προταθεί για την αντιμετώπιση των παραπάνω προκλήσεων είναι οι στρατηγικές πρόωρης εξόδου (early exit), που προσφέρουν μια δίοδο για τη βελτίωση της αποτελεσματικότητας, της αποκριτικότητας και της πρακτικότητας των μοντέλων βαθιάς μάθησης. Επιτρέποντας στα μοντέλα να παράγουν προβλέψεις ή να παίρνουν αποφάσεις πριν την ολοκλήρωση των υπολογισμών, οι στρατηγικές πρόωρης εξόδου καθιστούν τα μοντέλα πιο ευέλικτα κατά τη διάρκεια της συμπερασματολογίας. Η ενσωμάτωση τέτοιων στρατηγικών σε εφαρμογές βαθιάς μάθησης όχι μόνο έχει τη δυνατότητα να ενισχύσει την αποτελεσματικότητα, αλλά διευρύνει επίσης τους ορίζοντες των μοντέλων, διευκολύνοντας την ανάπτυξή τους σε σενάρια πραγματικού χρόνου και ενισχύοντας έτσι την ανάπτυξη πιο έξυπνων συστημάτων ταχείας απόκρισης.

Η παρούσα διπλωματική εργασία περιλαμβάνει τα ακόλουθα στάδια: (α) μελέτη και ανάπτυξη στρατηγικών πρόωρης εξόδου για βαθιά νευρωνικά δίκτυα, βάσει παραγόντων όπως η αρχιτεκτονική του δικτύου, το βάθος των επιπέδων και τα χαρακτηριστικά των δεδομένων, (β) εκτενής αξιολόγηση των στρατηγικών για την εκτίμηση του αντίκτυπου που έχουν στην απόδοση, την ακρίβεια και τη χρήση υπολογιστικών πόρων, (γ) βελτιστοποίηση τεχνικών με δυναμικά κατώφλια (thresholds) για την

υιοθέτηση στρατηγικών πρόωρης εξόδου σε πραγματικό χρόνο με βάση τη συμπεριφορά του μοντέλου, την πολυπλοκότητα των δεδομένων και τα επιθυμητά επίπεδα ακρίβειας, και (δ) εξερεύνηση σεναρίων πρακτικών εφαρμογών τα οποία μπορούν να επωφεληθούν από στρατηγικές πρόωρης εξόδου σε τομείς όπως η αναγνώριση εικόνας, η επεξεργασία φυσικής γλώσσας, η αναγνώριση αντικειμένων ή τα συστήματα συστάσεων.

Απαραίτητες γνώσεις προγραμματισμού: *Python*
Επιθυμητές γνώσεις: *Deep Learning frameworks (TensorFlow, Keras, PyTorch)*