Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Κατηγοριοποίηση Ιατρικών Εικόνων με χρήση Υβριδικών CNN-ViT μοντέλων

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ Γ. ΠΑΝΤΕΛΑΙΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής

Αθήνα, Ιούλιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Κατηγοριοποίηση Ιατρικών Εικόνων με χρήση Υβριδικών CNN-ViT μοντέλων

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ Γ. ΠΑΝΤΕΛΑΙΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Ιουλίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας	Γιώργος Στάμου	Αθανάσιος Βουλόδημος
Καθηγητής	Καθηγητής	Επίκουρος Καθηγητής

Αθήνα, Ιούλιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

(Υπογραφή)

.....
Δημήτριος Γ. Παντελαίος

27 Ιουλίου 2023

Περίληψη

Παρακινούμενοι από την επιτυχία των μετασχηματιστών στον τομέα της επεξεργασίας φυσικής γλώσσας, έγιναν προσπάθειες να εφαρμοστούν αντίστοιχα μοντέλα και στον τομέα της όρασης υπολογιστών. Γι' αυτό το λόγο δημιουργήθηκαν οι Vision Transformers, οι οποίοι παρουσιάζουν κορυφαίες επιδόσεις σε τομείς όπως η κατηγοριοποίηση εικόνων. Ωστόσο, οι Vision Transformers συλλαμβάνουν μακρινές καθολικές εξαρτήσεις μέσω των επιπέδων προσοχής, αλλά δεν διαθέτουν επαγωγικές προκαταλήψεις, ώστε να μπορούν να γενικευθούν όταν εκπαιδεύονται σε μικρό σύνολο δεδομένων, με αποτέλεσμα να απαιτούνται μεγαλύτερα σύνολα δεδομένων για την εκπαίδευση τους. Αυτό αποτελεί ένα σημαντικό εμπόδιο στην κατηγοριοποίηση ιατρικών εικόνων, καθώς είναι δύσκολη η εύρεση μεγάλων ιατρικών συνόλων δεδομένων. Η παρούσα μελέτη ασχολείται με την κατηγοριοποίηση ακτινογραφιών θώρακος, που αντιστοιχούν σε διαφορετικές ασθένειες που επηρεάζουν τους πνεύμονες, όπως είναι ο COVID-19. Πιο συγκεκριμένα, COVID-19 είναι μια αρκετά μεταδοτική μολυσματική ασθένεια που προσβάλει το αναπνευστικό σύστημα και οφείλεται στον ιό SARS-CoV-2. Πολλοί ασθενείς που προσβάλλονται από αυτή χρειάζονται άμεση ιατρική βοήθεια και αυτό καθιστά επιτακτική την άμεση ανιχνευση της. Για την επίλυση των παραπάνω προβλημάτων επινοήθηκαν τα υβριδικά μοντέλα, τα οποία προσπαθούν να προσθέσουν κάποια πλεονεκτήματα των συνελικτικών νευρωνικών δικτύων στους Vision Transformer, προκειμένου να γίνει δυνατή η εκπαίδευση των μοντέλων σε μικρότερα σύνολα δεδομένων. Στην μελέτη αυτή επικεντρωνόμαστε στην σύγκριση των υβριδικών μοντέλων προεκπαιδευμένων στο ImageNet-1k με τον παραδοσιακό Vision Transformer προεκπαιδευμένο στο ImageNet-21k, αλλά και στην εκπαίδευση των μοντέλων από την αρχή κάνοντας χρήση τόσο ενός μέρους, όσο και ολόκληρου του διαθέσιμου συνόλου δεδομένων COVID-QU-Ex. Τα αποτελέσματα που προκύπτουν δείχνουν την υπεροχή των υβριδικών μοντέλων τόσο όσον αφορά την ακρίβεια, τον χρόνο εκπαίδευσης, αλλά και τον αριθμό των δεδομένων που απαιτείται για την εκπαίδευση.

Λέξεις Κλειδιά

Vision Transformers, Συνελικτικά Νευρωνικά Δίκτυα, Κατηγοριοποίηση ιατρικών εικόνων, COVID-19, Ιογενής Πνευμονία, Βακτηριακή Πνευμονία, Υβριδικά μοντέλα ViT-CNN, DeiT, CeiT, Compact Transformers, Conformer, LocalViT, Convolutional vision Transformers

Abstract

Inspired by the success of Transformers in the field of Natural Language Processing, efforts have been made to apply similar models in computer vision. This led to the development of Vision Transformers, which have achieved state-of-the-art results in areas such as image classification. However, Vision Transformers capture long-range global dependencies through attention layers but they lack inductive biases, making it challenging for them to generalize when trained on small datasets. As a result, larger datasets are required for their training. This poses a significant obstacle in the classification of medical images since it is difficult to find large medical datasets. This study focuses on the classification of chest X-ray images corresponding to different diseases affecting the lungs, such as COVID-19. Specifically, COVID-19 is a highly contagious and infectious disease that affects the respiratory system and is caused by the SARS-CoV-2 virus. Many patients affected by this disease require immediate medical care, making its timely detection crucial. To address these challenges, hybrid models were devised, aiming to incorporate some advantages of CNNs into Vision Transformers, enabling the training of models on smaller datasets. In this study, we compare the hybrid models pre-trained on ImageNet-1k with the traditional Vision Transformer pre-trained on ImageNet-21k. We also explore training the models from scratch using both a subset and the entire available COVID-QU-Ex dataset. The results obtained demonstrate the superiority of the hybrid models in terms of accuracy, training time and the number of data required for training.

Keywords

Vision Transformers, Convolutional Neural Networks, Medical Image classification, COVID-19, Viral Pneumonia, Bacterial Pneumonia, Hybrid ViT-CNN models, DeiT, CeiT, Compact Transformers, Conformer, LocalViT, Convolutional vision Transformers

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Στέφανο Κόλλια για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης. Επίσης ευχαριστώ ιδιαίτερα την κ.Παρασκευή Τζούβελη και την Παρασκευή-Αντωνία Θεοφίλου για την καθοδήγησή τους και την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου και τα αδέρφια μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Ιούλιος 2023

Δημήτριος Γ. Παντελαίος

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στην Αθήνα, το έτος 2023, στο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης (AILS) που ανήκει στον Τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, με επιβλέποντες τον Καθηγητή κ. Στέφανο Κόλλια, την κ. Παρασκευή Τζούβελη και την Παρασκευή-Αντωνία Θεοφίλου.

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Ο COVID-19 είναι μια πολύ μεταδοτική και μολυσματική ασθένεια που προκαλείται από το ιό Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) και για αυτό το λόγο έχει ανακηρυχθεί πανδημία από τον Παγκόσμιο Οργανισμό Υγείας (WHO). Σε κάποιες περιπτώσεις μπορεί να οδηγήσει σε αναπνευστική ανεπάρκεια και πιθανόν θάνατο, με αποτέλεσμα να καθίσταται αναγκαία η έγκαιρη και αποτελεσματική διάγνωση της νόσου για την επιτυχή αντιμετώπιση της [31]. Μια από τις βασικές μεθόδους διάγνωσης της συγκεκριμένης νόσου είναι το Reverse-Transcription Polymerase Chain Reaction (RT-PCR). Ωστόσο, υπάρχουν και τεχνικές απεικόνισης όπως το Chest Computed Tomography (CT) και οι ακτινογραφίες θώρακα (Chest X-ray ή CXR) που έχουν βρεθεί να έχουν μεγαλύτερη ευαισθησία στον εντοπισμό της νόσου [32]. Πιο συγκεκριμένα, οι ακτινογραφίες θώρακα είναι διαθέσιμες σε χαμηλό κόστος, προκειμένου να συμβάλουν στον εντοπισμό της ασθένειας COVID-19 και μπορούν να παρέχουν πολύτιμες πληροφορίες σχετικά με τις αλλοιώσεις των πνευμόνων που προκαλέι η νόσος.

Η αποδοτικότητα στις εφαρμογές ανάλυσης ιατρικών εικόνων μπορεί να βελτιωθεί αισθητά με την είσοδο της τεχνητής νοημοσύνης (AI) και των τεχνικών βαθιάς μάθησης (DL) στην αντιμετώπιση τους [33]. Τα μοντέλα βαθιάς μάθησης, εκπαιδευμένα σε μεγάλα σύνολα δεδομένων ιατρικών εικόνων, μπορούν να βοηθήσουν στον εντοπισμό μοτίβων και χαρακτηριστικών που μπορεί να μην είναι εύκολα αναγνωρίσιμα από το ανθρώπινο μάτι [34]. Με αυτό τον τρόπο μπορούν να συμβάλουν στην προσπάθεια των ειδικών για πρώιμο εντοπισμό και πρόβλεψη ασθενειών, συμπεριλαμβανομένο και του COVID-19. Γενικά, ο συνδυασμός της ιατρικής επιστήμης και της τεχνολογίας των υπολογιστών μπορούν να βοηθήσουν στην αντιμετώπιση της όλο αυξανόμενης πολυπλοκότητας και των δυσκολιών που εμφανίζονται στον τομέα της υγειονομικής περίθαλψης [35].

Τα τελευταία χρόνια τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) αποτελούσαν ένα από τα πιο συνηθισμένα είδη μοντέλων που χρησιμοποιούταν στις εφαρμογές της όρασης υπολογιστών (computer vision), παρουσιάζοντας σε πολλές περιπτώσεις state-of-the-art αποτελέσματα. Ωστόσο, εμπνευσμένοι από την επιτυχία των μετασχηματιστών (Transformers) στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP) [36], έγιναν προσπάθειες για δημιουργία αντίστοιχων μοντέλων στον τομέα της όρασης υπολογιστών. Οι Vision Transformers (ViTs) [15] είναι ένας τύπος μοντέλων βαθιάς μάθησης που χρησιμοποιούνται όλο και περισσότερο σε

εφαρμογές, όπως κατηγοριοποίηση εικόνων, ανίχνευση αντικειμένων και άλλες περιπτώσεις της όρασης υπολογιστών. Τα ViTs συλλαμβάνουν μακρινές καθολικές εξαρτήσεις, μέσω των self-attention επιπέδων, σε αντίθεση με τα CNNs που συλλαμβάνουν τοπικές πληροφορίες μέσω των τοπικών δεκτικών πεδίων (receptive fields) . Ωστόσο οι Vision Transformers δεν διαθέτουν επαγωγικές προκαταλήψεις (inductive biases) και συνεπώς δεν διαθέτουν την ικανότητα γενίκευσης οταν εκπαιδεύονται σε μικρό σύνολο δεδομένων. Αυτό δημιουργεί την ανάγκη ύπαρξης μεγάλων συνόλων δεδομένων για την εκπαίδευση τους, προκειμένου να πετύχουν state-of-the-art αποτελέσματα στις διάφορες εφαρμογές. Για αυτό το λόγο καθίσταται αναγκαίο να συνδυαστούν τα χαρακτηριστικά των προαναφερθέντων μοντέλων, προκειμένου να αξιοποιηθούν να πλεονεκτήματα και των δύο και να δημιουργηθούν μοντέλα με βελτιωμένα αποτελέσματα που δεν χρειάζονται μεγάλα σύνολα δεδομένων για την εκπαίδευση τους.

Συνεπώς, σκοπός αυτής της εργασίας είναι να μελετήσουμε διάφορα υβριδικά CNN-ViT συστήματα βαθιάς μάθησης και να τα εφαρμόσουμε σε ένα ζωτικής σημασίας πρόβλημα των ημερών μας, τον εντοπισμό του COVID-19 μέσω ακτινογραφιών θώρακα.

Οι κύριες συνεισφορές αυτής της εργασίας είναι:

- Εξετάζουμε μια μεγάλη ποικιλία υβριδικών CNN-ViT μοντέλων για να βελτιώσουμε την απόδοση και τις δυνατότητες της εφαρμογής κατηγοριοποίησης εικόνων.
- Εφαρμόζουμε αυτά τα προσεκτικά επιλεγμένα μοντέλα στο σύνολο δεδομένων COVID-QU-Ex, έναν πολύτιμο δείκτη για την αξιολόγηση της απόδοσης των μοντέλων στην εφαρμογή κατηγοριοποίησης ιατρικών εικόνων και τον εντοπισμό του COVID-19.
- Συγκρίνουμε τις επιδόσεις των CNN-ViT και των απλών ViT όταν γίνονται fine-tuning και όταν εκπαιδεύονται από το μηδέν. Και στις δύο περιπτώσεις, τα υβριδικά μοντέλα επιτυγχάνουν καλύτερα αποτελέσματα όσον αφορά την ακρίβεια, τον χρόνο εκπαίδευσης και το υπολογιστικό κόστος.
- Η μελέτη μας αποδεικνύει την ενισχυμένη ικανότητα εντοπισμού του COVID-19, παρέχοντας αξιόπιστα και ισχυρά (robust) αποτελέσματα.
- Εξετάζουμε την συμπεριφορά των Vision Transformer όταν εκπαιδεύονται με δύο από τις τρεις κλάσεις του συνόλου δεδομένων COVID-QU-Ex, οι οποίες προκύπτουν είτε από συνδυασμό ή διαγραφή κλάσεων. Τα αποτελέσματα σε κάποιες περιπτώσεις ξεπερνούν σε ακρίβεια το 99%.

1.2 Οργάνωση του τόμου

Η εργασία είναι οργανωμένη σε 7 Κεφάλαια, ενώ περιλαμβάνει και 1 Παράρτημα. Το περιεχόμενο των Κεφαλαίων και του Παραρτήματος αναλύεται παρακάτω:

- Το παρών Κεφάλαιο αποτελεί την εισαγωγή της εργασίας.
- Το Κεφάλαιο 2 παρουσιάζει το θεωρητικό υπόβαθρο της μελέτης.

- Το Κεφάλαιο 3 παρέχει εργασίες σχετικές με ιατρικά σύνολα δεδομένων και ιδιαίτερα με την ασθένεια του COVID-19, καθώς και τις ιδιαιτερότητες των συνόλων αυτών.
- Το Κεφάλαιο 4 περιγράφει τα σύνολα δεδομένων που χρησιμοποιήθηκαν.
- Το Κεφάλαιο 5 καθορίζει την αρχιτεκτονική των μοντέλων και την μεθοδολογία που ακολουθήθηκε.
- Το Κεφάλαιο 6 παρουσιάζει τις πειραματικές διαδικασίες και τα αποτελέσματα.
- Το Κεφάλαιο 7 περιγράφει τα συμπεράσματα και τις πιθανές μελλοντικές επεκτάσεις της εργασίας.
- Το Παράρτημα Α΄ αναλύει την δομή του στρώματος προσοχής (attention layer) και παρουσιάζει τις διάφορες παραλλαγές του, ενώ παράλληλα παρουσιάζει τις βασικές συναρτήσεις ενεργοποίησης, με το μεγαλύτερο μέρος εξ αυτών να χρησιμοποιείται στην παρούσα εργασία.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται οι έννοιες της Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και Βαθιάς Μάθησης, κάποιες βασικές εφαρμογές που πραγματοποιούν, καθώς και τα βασικά μοντέλα που χρησιμοποιούνται για την αντιμετώπιση τους.

2.1 Τεχνητή Νοημοσύνη

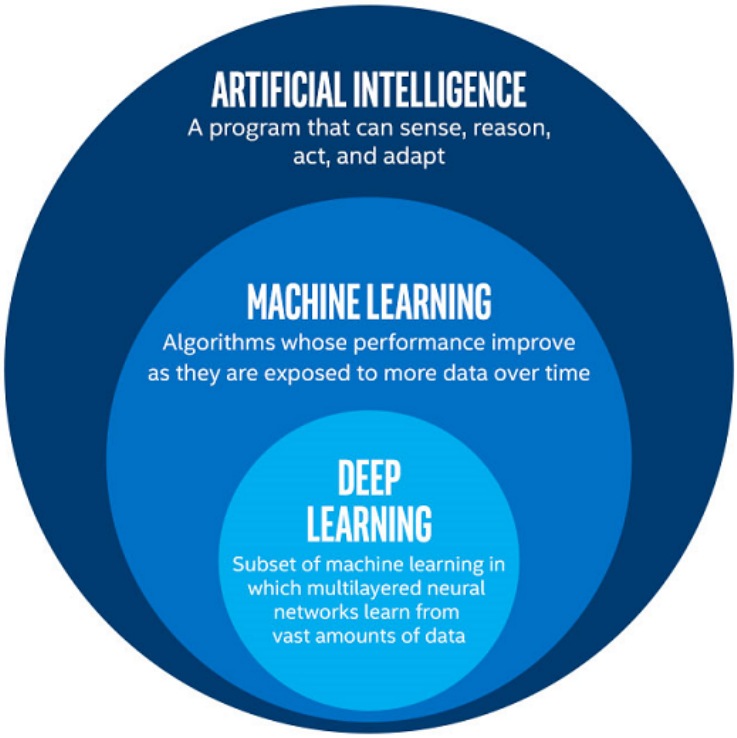
Τεχνητή νοημοσύνη αποτελεί τον τομέα την πληροφορικής που δημιουργεί υπολογιστικά συστήματα, τα οποία προσπαθούν να αναπαραστήσουν την ανθρώπινη νοημοσύνη. Πιο συγκεκριμένα, έχει ως στόχο την αυτοματοποίηση επίλυσης προβλημάτων που ο άνθρωπος επιλύει καλύτερα από τους υπολογιστές. Χαρακτηριστικά παραδείγματα τέτοιων προβλημάτων είναι η μάθηση, η προσαρμοστικότητα, η αναγνώριση προσώπων και η ταξινόμηση σε κλάσεις[37].

Ένα βασικό πλεονέκτημα της τεχνητής νοημοσύνης ειναι η ταχύτητα με την οποία μαθαίνει το εκάστωτε μοντέλο που είναι πολύ μεγαλύτερη από την ταχύτητα εκμάθησης του ανθρώπινου εγκέφαλου με αποτέλεσμα να υπάρχουν πολύ μεγάλες δυνατότητες εξέλιξης. Ακόμη, εξοικονομεί χρόνο στους εργαζομένους εκτελώντας επαναλαμβανόμενες διεργασίες που θα καθιστούσαν λιγότερο ενδιαφέρουσα την δουλεία τους. Τέλος προσφέρει ασφάλεια, καθώς υπάρχει η δυνατότητα κάποιες εργασίες που εμπεριέχουν ρίσκο να επιτελούνται από εκπαιδευμένα ρομπότ[38].

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί υποκατηγορία της τεχνητής νοημοσύνης και αποτελεί έναν από τους πιο γρήγορα αναπτυσσόμενους κλάδους της πληροφορικής. Είναι ένας τομέας που χρησιμοποιεί δεδομένα για την βελτίωση της επίδοσης των υπολογιστών σε συγκεκριμένες εφαρμογές. Πιο συγκεκριμένα, χρησιμοποιεί δεδομένα για την εκπαίδευση των μοντέλων, τα οποια στην συνέχεια μπορούν να χρησιμοποιηθούν για να κάνουν αντίστοιχη πρόβλεψη σε δεδομένα με τα οποία δεν έχουν έρθει σε επαφή. Μερικούς από τους τομείς που χρησιμοποιείται η μηχανική μάθηση είναι η όραση υπολογιστών, η επεξεργασία φυσικής γλώσσας, αναγνώριση φωνής και ανάλυση δεδομένων[39].

Οι βασικοί τύποι μηχανικής μάθησης είναι:



Εικόνα 2.1: *Σχέση τεχνητής νοημοσύνης, μηχανικής μάθησης και βαθιάς μάθησης [22].*

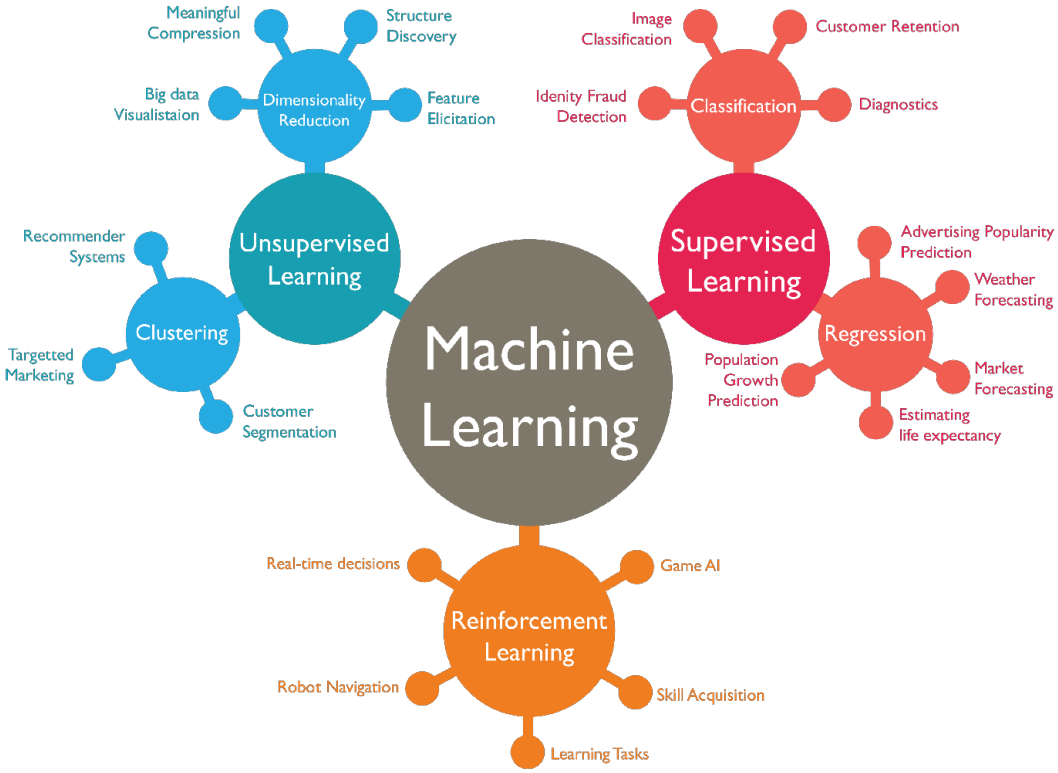
- Επιβλεπόμενη μάθηση (Supervised Learning): Για κάθε δεδομένο εισόδου διαθέτουμε και μία ετικέτα, η οποία δηλώνει την κατηγορία στην οποία ανήκει. Το μοντέλο εκπαιδεύεται με βάση αυτά τα δεδομένα και έχει ως στόχο τον προσδιορισμό των ετικετών των δοκιμαστικών (test) δεδομένων. Μερικές κατηγορίες επιβλεπόμενης μάθησης είναι η κατηγοριοποίση δεδομένων, όπου αποδίδουμε στα δοκιμαστικά δεδομένα κάποια από τις διαθέσιμες ετικέτες και το regression, όπου στόχος είναι η πρόβλεψη της μελλοντικής τιμής μια μεταβλητής εξαρτημένης από άλλες ανεξάρτητες μεταβλητες.

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): χρησιμοποιεί αλγόριθμους για την ανάλυση και ομαδοποίηση δεδομένων με βάση κοινών μοτίβων, καθώς δεν διαθέτουμε ετικετές για αυτά τα δεδομένα. Η μη επιβλεπόμενη μάθηση χρησιμοποιείται κυρίως για τις εξής 3 εφαρμογές[40]:

  1. Clustering, η οποία ομαδοποιεί τα δεδομένα χωρίς ετικέτα με βάση ομοιότητες και διαφορές που παρουσιάζουν μεταξύ τους.
  2. Association, η οποία βασίζεται σε κανόνες για την εύρεση σχέσεων μεταξύ μεταβλητών σε ένα δοσμένο σύνολο δεδομένων. Χρησιμοποιείται συχνά για την ανάλυση της αγοράς, προκειμένου να εντοπισθούν μοτίβα στις αγορές των καταναλωτών και να βρεθούν σχέσεις ανάμεσα σε διαφορετικά προϊόντα.
  3. Dimensionality reduction, μέσω της οποίας μειώνονται οι διαστάσεις των δεδομένων εισόδου, προκειμένου να γίνει ευκολότερη η επεξεργασία τους από τα εκάστωτε μοντέλα. Ωστόσο η μείωση των διαστάσεων θα πρέπει να γίνεται σε

ελεγχόμενο βαθμό, έτσι ώστε να μην επηρεάζεται η αξιοπιστία των δεδομένων. Συνήθως χρησιμοποιείται στο στάδιο προεπεξεργασίας των δεδομένων και η πιο γνωστή τέτοια μέθοδος είναι η Principal Component Analysis (PCA)

- Ενισχυτική μάθηση (Reinforcement Learning): ο πράκτορας ενισχυτικής μάθησης μαθαίνει αλληλεπιδρώντας με το περιβάλλον, παίρνοντας επιβράβευση (reward) για κάθε κίνηση που θεωρείται θετική και τιμωρία (punishment) για κάθε κίνηση που θεωρείται αρνητική.

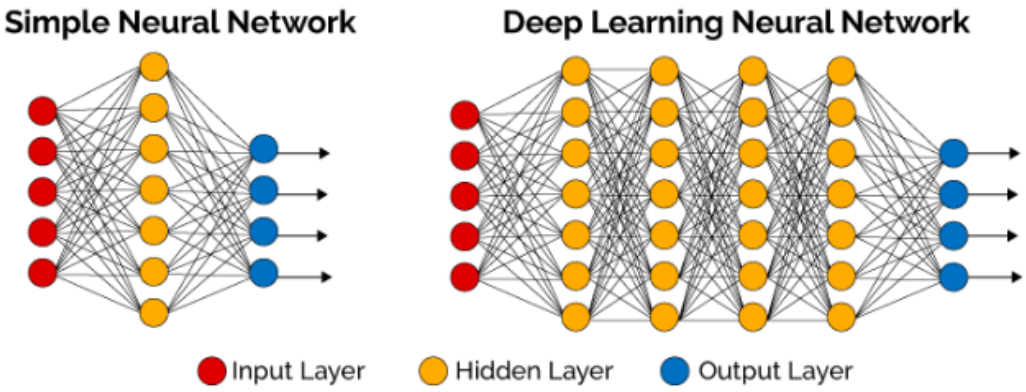


Εικόνα 2.2: *Κλάδοι Μηχανικής Μάθησης [23].*

Μια υποκατηγορία μοντέλων μηχανικής μάθησης αποτελούν τα νευρωνικά δίκτυα 2.3 ή αλλιώς τεχνητά νευρωνικά δίκτυα (artificial neural networks → ANN), τα οποία προσπαθούν να μιμηθούν την λειτουργία του αθρώπινου εγκεφάλου. Τα βασικά επίπεδα των νευρωνικών δικτύων είναι το επίπεδο εισόδου, εξόδου και τα κρυφά επίπεδα (hidden layers). Αυτά τα επίπεδα αποτελούνται απο νευρώνες, βασικά στοιχεία των οποίων είναι οι είσοδοι, τα βάρη, το bias και η έξοδος. Η έξοδος κάθε νευρώνα προκύπτει από τον τύπο $y = \sum_{i=1}^{n} w_i \cdot x_i - bias$, όπου w_i τα βάρη του νευρώνα και x_i οι είσοδοι του νευρώνα. Πιο συγκεκριμένα, ένας νευρώνας ενεργοποιείται όταν το εσωτερικό γινόμενο των βαρών και των εισόδων ξεπεράσει το κατώφλι (threshold ή bias). Η έξοδος αυτών των δικτύων μπορεί να μοντελοποιηθεί, ώστε να αντιστοιχεί σε μια συγκεκριμένη κατηγορία.

2.3 Βαθιά Μάθηση

Τα μοντέλα βαθιάς μάθησης αποτελούν υποσύνολο των νευρωνικών δικτύων και διαθέτουν μεγαλύτερο πλήθος (περισσότερα από 3) ενδιάμεσων επιπέδων (κρυφά επίπεδα) που τα καθιστά πιο ¨βαθιά¨[41]. Η βασική διαφορά των βαθιών νευρωνικών δικτύων με τα μοντέλα μηχανικής μάθησης είναι οτι τα πρώτα μπορούν να επεξεργάζονται δεδομένα στην αρχική τους μορφή, χωρίς περαιτέρω προεπεξεργασία, περιορίζοντας με αυτό τον τρόπο την ανθρώπινη παρέμβαση.



Εικόνα 2.3: *Σύγκριση μοντέλων μηχανικής και βαθιάς μάθησης [24].*

Μερικές κατηγορίες μοντέλων βαθιάς μάθησης είναι τα Συνελικτικά Νευρωνικά Δίκτυα, τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks ή RNN) και οι Μετασχηματιστές (Transformers), οι οποίοι παρουσιάζονται παρακάτω.

2.4 Κατηγοριοποίηση εικόνων (Image Classification)

Κατηγοριοποίηση εικόνων είναι η διαδικασία ανάθεσης ετικετών σε δεδομένα ενός συνόλο με βάση κάποια χαρακτηριστικά που προκύπτουν από αυτά και αποτελεί μια από τις βασικότερες εφαρμογές της τεχνητής νοημοσύνης, της βαθιάς μάθησης και της όρασης υπολογιστών [42]. Κάθε ετικέτα που μπορούμε να αποδόσουμε σε ένα αντικείμενο συγκροτεί μία κλάση. Μερικές βασικές κατηγοριές κατηγοριοποίησης εικόνων είναι η δυαδική ταξινόμηση (binary classification), όταν υπάρχουν μόνο δύο κλάσεις και η πολυταξική ταξινόμηση (multiclass classification), όταν υπάρχουν περισσότερες από δύο κλάσεις. Τα κυριότερα μοντέλα που χρησιμοποιούνται για την συγκεκριμένη εφαρμογή είναι τα Support Vector Machines (SVM), Decision Trees, K Nearest Neighbor, Artificial Neural Networks, Convolutional Neural Networks, καθώς και οι Vision Transformers, οι οποίοι εμφανίστηκαν τα τελευταία χρόνια. Συνήθως, το τελευταίο στάδιο των μοντέλων είναι μία softmax συνάρτηση ενεργοποίησης που λαμβάνει τα δεδομένα που έχουν παραχθεί από το δίκτυο και τα μετατρέπει σε πιθανότητες που αντιστοιχούν στις πιθανές κλάσεις. Πολλές φορές τα δεδομένα εισόδου πρέπει να υποβληθούν σε κάποια προεπεξεργασία, προκειμένου να έρθουν στην κατάλληλη μορφή που απαιτείται για την τροφοδοσία τους στο εκάστωτε μοντέλο. Για παράδειγμα, αν τα δεδομένα εισόδου είναι εικόνες, μπορούν να εφαρμοστούν μέθοδοι, όπως

αλλαγή μεγέθους εικόνας, περιστροφή εικόνας ή και εφαρμογή φίλτρων για την μείωση θορύβου, ενώ αν τα δεδομένα είναι πινακοειδή (tabular) μπορεί να γίνει χειρισμός των τιμών που λείπουν και των κατηγορικών χαρακτηριστικών. Κάποιες βασικές προκλήσεις στον τομέα αυτόν είναι [43]:

- Intra-Class Variation: Μεγάλη διαφοροποίηση μεταξύ αντικειμένων της ίδιας κλάσης.
- Scale Variation: Τα αντικείμενα εμφανίζονται στις εικόνες σε πολύ διαφορετικά μεγέθη.
- View-Point Variation: Εμφάνιση αντικειμένων από διαφορετική οπτική γωνία.
- Occlusion: Κάποια αντικείμενα δεν φαίνονται ολόκληρα, αλλά καλύπτονται από άλλα.
- Illumination: Λόγω διαφοράς φωτισμού το ίδιο αντικείμενο μπορεί να απεικονίζεται με πολύ διαφορετική ένταση μέσω των pixels.
- Background Clutter: Υπάρχουν πολλά διαφορετικά αντικείμενα στην εικόνα, το οποίο καθιστά δύσκολο τον εντοπισμό του αντικειμένου ενδιαφέροντος.

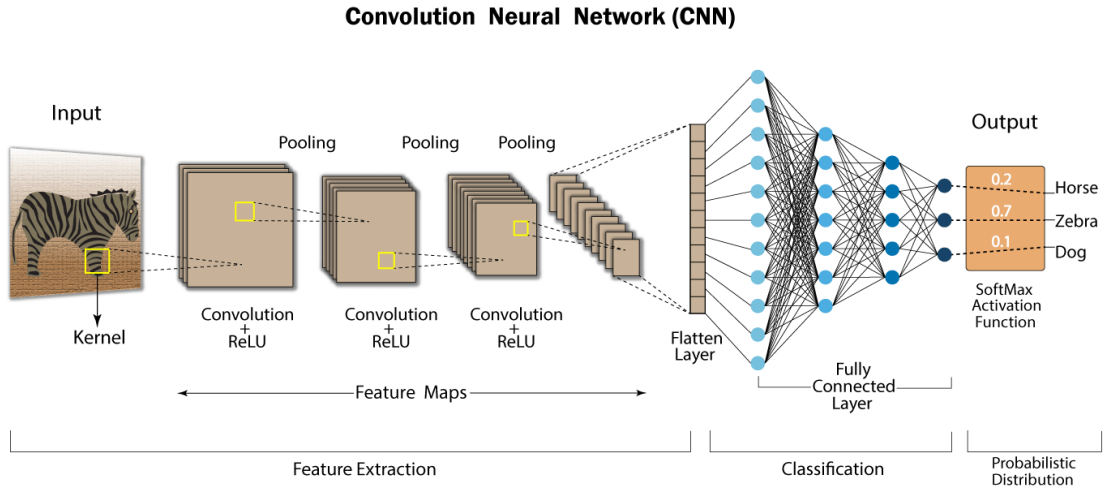
Γίνονται διαρκείς προσπάθειες για βελτίωση όσον αφορά την ακρίβεια, την ευρωστία (robustness) και την δυνατότητα επεξηγησιμότητας, προκειμένου να δημιουργηθούν μοντέλα που αντιμετωπίζουν τα παραπάνω προβλήματα και προοδεύουν τον τομέα της όρασης υπολογιστών.

2.5 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks)

Τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) είναι μία από τις κυρίαρχες αρχιτεκτονικές όσον αφορά τις εφαρμογές που σχετίζονται με την όραση υπολογιστών, όπως κατηγοριοποίηση εικόνων, ανίχνευση αντικειμένων και τμηματοποίηση εικόνων. Αποτελούν ένα αλγόριθμο βαθιάς μάθησης που είναι ειδικά σχεδιασμένος να λαμβάνει ως είσοδο εικονικά δεδομένα, κυρίως εικόνες, και έχει ως στόχο να αναθέσει βάρη και biases στα αντικείμενα στην εικόνα, προκειμένου να μπορέσει να τα διαφοροποιήσει μεταξύ τους [44].

Τα CNNs αποτελούνται κυρίως από 3 επίπεδα, το συνελικτικό επίπεδο (convolutional layer), το επίπεδο ομαδοποίησης (pooling layer) και το πλήρως συνδεδεμένο επίπεδο (fully connected layer ή FC). Τα συνελικτικά επίπεδα εφαρμόζουν φίλτρα στις εικόνες που δέχονται ως είσοδο, επιτρέποντας στο δίκτυο να συλλαμβάνει τοπικά μοτίβα και τοπικές χωρικές σχέσεις. Η έξοδος των επιπέδων αυτών αντιπροσωπεύει συγκεκριμένα χαρακτηριστικά της εικόνας και αποκαλείται χάρτης χαρακτηριστικών (feature map). Όσο το δίκτυο προχωράει τα χαρακτηριστικά αυτά συνδυάζονται για την δημιουργία πολύπλοκων μοτίβων[45]. Στα συνελικτικά επίπεδα χρησιμοποιούνται ακόμη συναρτήσεις ενεργοποίησης με επικρατέστερη την ReLU. Από την άλλη τα επίπεδα ομαδοποίησης χρησιμοποιούνται για την υποδειγματολειψία (downsampling) των χαρτών χαρακτηριστικών, μειώνοντας τις χωρικές διαστάσεις, διατηρώντας παράλληλα τις απαραίτητες πληροφορίες. Με αυτό τον τρόπο μειώνεται το υπολογιστικό κόστος και εξάγονται χαρακτηριστικά που δεν μεταβάλλονται από την θέση που βρίσκονται και από την περιστροφή στην οποία υποβάλλονται. Υπάρχουν 3 είδη επιπέδων

ομαδοποίησης, το max pooling, που επιλέγει την μέγιστη τιμή από το κομμάτι της εικόνας που εφαρμόζεται ο πυρήνας, το average pooling που επιστρέφει τον μέσο όρο των στοιχείων του αντίστοιχου κομματιού και το sum pooling που επιστρέφει το άθροισμα των στοιχείων. Το πλήρως συνδεδεμένο επίπεδο χρησιμοποιείται για την εκμάθηση μη γραμμικών συνδυασμών των χαρακτηριστικών υψηλότερου επιπέδου που προκύπτουν στην έξοδο των συνελικτικών επιπέδων. Στην αρχή των FC επιπέδων η είσοδος γίνεται flatten και μετατρέπεται σε μονοδιάστατο πίνακα. Τέλος για κατηγοριοποίηση της εξόδου το αποτέλσμα των FC επιπέδων περναέι από Softmax ή Sigmoid συνάρτηση ενεργοποίησης. Μια βασική μορφή CNN παρουσιάζεται στο Σχήμα 2.1.



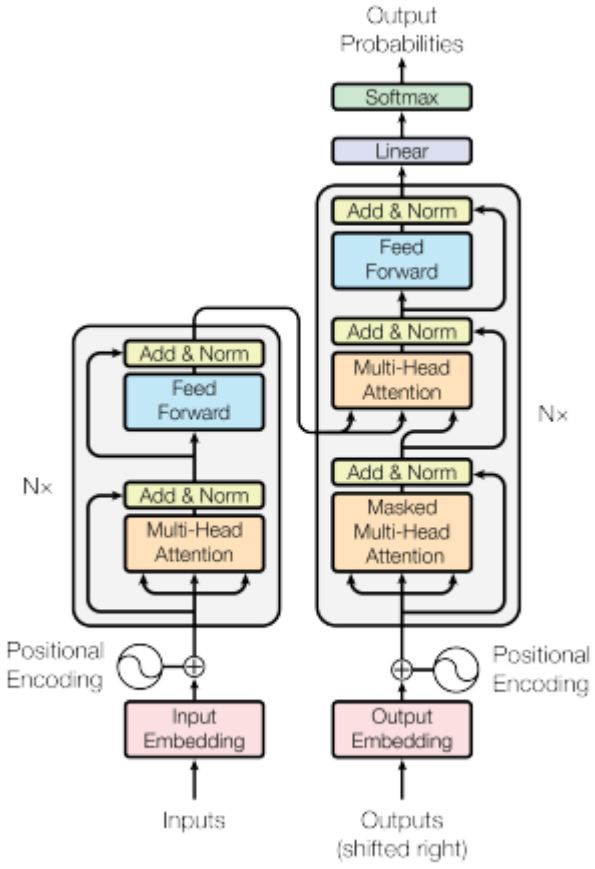
Σχήμα 2.1: *Αρχιτεκτονική απλού CNN [1].*

Μερικά βασικά CNNs είναι το LeNet, το AlexNet, το VGGNet, το ResNet και το GoogleNet ή Inception v1. Το LeNet είναι ένα από τα πρώτα CNNs και αποτελείται από 7 επιπέδα, μεταξύ των οποίων είναι 3 συνελικτικά επίπεδα, 2 average pooling επίπεδα και 2 πλήρως συνδεδεμένα επίπεδα, ενώ ως συνάρτηση ενεργοποίησης χρησιμοποιείται η tanh. Η βασική εφαρμογή του ήταν η απλή αναγνώριση ψηφίων. Το AlexNet χρησιμοποίησε ως συνάρτηση ενεργοποίησης την ReLU, προσθέτοντας μη γραμμικότητα, αυξάνοντας παράλληλα την ταχύτητα του δικτύου και διατηρώντας ίδια ακρίβεια (accuracy). Ακόμη έγινε χρήση dropout αντί για regularization για την αντιμετώπιση της υπερπροσαρμογής (overfitting), ενώ τέλος εφαρμόστηκε επικαλυπτόμενο pooling, μειώνοντας το μέγεθος του δικτύου [46]. Το μοντέλο αυτό νίκησε με μεγάλη διαφορά τον διαγωνισμό ImageNet LSVRC-2012. Το VGGNet με την σειρά του αντικατέστησε τα μεγαλύτερα φίλτρα μεγέθους 5x5 και 11x11 με μικρότερα φίλτρα 3x3, με αποτέλεσμα να μειωθούν οι παράμετροι και να βελτιωθεί ο χρόνος εκπαίδευσης. Παράλληλα σχεδιάστηκαν βαθύτερες εκδόσεις του μοντέλου (VGG-16, VGG-19) για την βελτίωση της απόδοσης. Το ResNet πρόσθεσε υπολειμματικές συνδέσεις (residual connections), επιτρέποντας την χρήση βαθύτερων δικτύων με την συμβολή του στην αντιμετώπιση του προβλήματος των εξαφανιζόμενων παραγώγων (vanishing gradients) στο backpropagation. Τέλος το GoogleNet εισήγαγε παράλληλα φίλτρα διαφορετικών μεγεθών, δημιουργώντας ευρύτερα δίκτυα (wider networks) και μια max pooling διακλάδωση. Ακόμη, έκανε χρήση 1x1 συνελίξεων για την μείωση του βάθους (καναλιών). Τα φίλτρα με

τους μεγάλους πυρήνες συνελάμβαναν καθολικά χαρακτηριστικά, ενώ οι μικροί πυρήνες συγκεκριμένα χαρακτηριστικά περιοχής που κατανέμονταν σε ολόκληρο το πλαίσιο της εικόνας. Σε κάποια αρχικά επίπεδα προστέθηκαν βοηθητικοί ταξινομητές για την επίλυση του προβλήματος των εξαφανιζόμενων παραγώγων.

2.6 Μετασχηματιστές (Transformers)

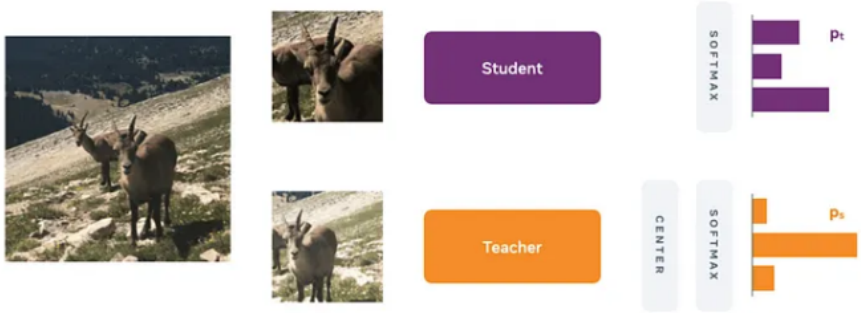
Η αρχιτεκτονική των μοντέλων που βασίζονται στο self-attention και ειδικότερα των Μετασχηματιστών αποτελεί την πιο συνηθισμένη αρχιτεκτονική για εφαρμογές που σχετίζονται με επεξεργασία φυσική γλώσσας και σε πολλές από αυτές αποτελεί το state-of-the-art, επικρατώντας έναντι των αντίστοιχων RNN και LSTM μοντέλων. Ο μηχανισμός self-attention στον οποίο βασίζονται τους επιτρέπει να επικεντρώνονται σε διαφορετικά κομμάτια της εισόδου, συλλαμβάνοντας μακρινές εξαρτήσεις και κατανοώντας τις σχέσεις μεταξύ των λέξεων που οδηγούν σε καλύτερα αποτελέσματα. Τα μοντέλα αυτά βασίζονται στην αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή, όπως φαίνεται στο Σχήμα 2.2 και τα αποδοτικότερα εξ αυτών προκύπτουν μετά από προεκπαίδευση σε μεγάλα σύνολα κειμένων και fine-tuning σε κάποιο σύνολο δεδομένων αρκετά μικρότερης έκτασης από το σύνολο με το οποίο έγινε pre-train. Στον τομέα της φυσικής γλώσσας χρησιμοποιούνται σε εφαρμογές, όπως μετάφραση κειμένου, ανάλυση συναισθημάτων, περίληψη κειμένου, απάντηση ερωτήσεων κ.α.



Σχήμα 2.2: *Αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή [2].*

Ωστόσο, τα τελευταία χρόνια εμφανίζονται όλο και περισσότερες μελέτες που εφαρμόζουν αρχιτεκτονική που βασίζεται στους μετασχηματιστές και στον τομέα της όρασης, δηλαδή σε δεδομένα όπως εικόνες και βίντεο [15]. Η υπολογιστική πολυπλοκότητα και η δυνατότητα επεκτασιμότητας που διαθέτουν οι Transformers επιτρέπουν την εκπαίδευση μοντέλων με πολύ μεγάλο πλήθος παραμέτρων που ξεπερνούν τα 100 δισεκατομμύρια, ενώ παράλληλα παρουσιάζουν βελτιωμένα αποτελέσματα σε σύγκριση με τα CNNs, όταν διαθέτουν τους ίδιους υπολογιστικούς πόρους [15]. Τα μοντέλα αυτά έφεραν την επανάσταση στον τομέα της όρασης υπολογιστών, αντικαθιστώντας τα συνελικτικά επίπεδα με self-attention layers και τα τοπικά receptive πεδία με την δυνατότητα σύλληψης καθολικών χαρακτηριστικών. Αποτελείται από ένα input embedding layer, αρκετους κωδικοποιητές μετασχηματιστή (transformer encoder) και ένα classification head που χρησιμοποιείται για την πρόβλεψη. Τα μοντέλα αυτά χρησιμοποιούνται σε πολλές εφαρμογές της όρασης υπολογιστών συμπεριλαμβανομένης της κατηγοριοποίησης εικόνων, της ανίχνευσης αντικειμένων, της σημασιολογικής τμηματοποίησης και της παραγωγής εικόνας.

Τα Συνελικτικά Νευρωνικά Δίκτυα εμφανίζουν καλύτερες επιδόσεις σε σύγκριση με τους Μετασχηματιστές, όταν εκπαιδεύονται από την αρχή σε κάποιο μικρό σύνολο δεδομένων. Αυτο συμβαίνει, γιατί οι μετασχηματιστές συλλαμβάνουν μακρινές καθολικές εξαρτήσεις, κάτι που οφείλεται στα attention layers και δεν διαθέτουν επαγωγικές προκαταλήψεις και συνεπώς δεν διαθέτουν την ικανότητα γενίκευσης οταν εκπαιδεύονται σε μικρό σύνολο δεδομένων. Όταν όμως τα μοντέλα είναι εκπαιδευμένα σε μεγάλο πλήθος δεδομένων, όπως είναι το ImageNet, τότε οι μετασχηματιστές παρουσιάζουν πολύ μεγάλη βελτίωση και σε κάποιες περιπτώσεις ξεπερνούν τις επιδόσεις των Συνελικτικών Νευρωνικών Δικτύων, που μέχρι πρότινος κυριαρχούσαν στις εφαρμογές της όρασης υπολογιστών. Τέλος μια ακόμη μέθοδος που ευνοεί τις επιδόσεις των μετασχηματιστών είναι η self-supervised προεκπαίδευση μετά την αρχικοποίηση των βαρών που είναι βασισμένη στο σύνολο δεδομένων ImageNet. Πιο συγκεκριμένα, χρησιμοποιεί δεδομένα χωρίς ετικέτα προκειμένου να μάθει αναπαραστάσεις δεδομένων που μπορούν να χρησιμοποιηθούν για ομαδοποίηση (clustering) και συνεπώς στην κατηγοριοποίηση των δεδομένων[3]. Μία από τις πιο χαρακτηριστικές μεθόδους self-supervision είναι η μέθοδος DINO.



Σχήμα 2.3: *Διαδικασία εκπαίδευσης DINO [3].*

Η μέθοδος αυτη, όπως φαίνεται στην Εικόνα 2.3, δίνει στον μαθητή και στον δάσκαλο την

ίδια εικόνα με διαφορετικό μέγεθος επαύξησης και προσπαθεί να τους κάνει να παράγουν παρόμοια αποτελέσματα, προσαρμόζοντας τα βάρη τους ανάλογα με το cross-entropy loss μεταξύ των 2 εξόδων με την μέθοδο backpropagation. Στο τέλος μπορεί να ομαδοποιήσει τα δεδομένα αυτά και να τα κατηγοριοποιήσει [3] [47].

2.7 Υβριδικά μοντέλα ViT - CNN

Προκειμένου να αξιοποιηθούν παράλληλα και οι επαγωγικές προκαταλήψεις των CNNs και οι καθολικές (global) εξαρτήσεις που αιχμαλωτίζουν οι μετασχηματιστές δημιουργήθηκαν κάποια υβριδικά μοντέλα που επιδιώκουν να κάνουν χρήση των καλύτερων χαρακτηριστικών και από τα δύο. Κάποιες τέτοιες περιπτώσεις μοντέλων παρουσιάζονται στις παρακάτω υποενότητες.

2.7.1 Συνελικτικά Νευρωνικά Δίκτυα αντί για patches

Η πιο απλή μορφή υβριδικών μοντέλων που περιλαμβάνουν Vision Transformers και CNNs είναι η χρήση της αρχιτεκτονικής των πρώτων με την διαφορά ότι στην αρχή αντί να χωριστεί η εικόνα σε patches, τροφοδοτείται σε ένα CNN και οι χάρτες χαρακτηριστικών που προκύπτουν τροφοδοτούνται στον μετασχηματιστή με μέγεθος patch, ένα pixel[15].

2.7.2 Προσθήκη συνελικτικών στρωμάτων στην αρχή

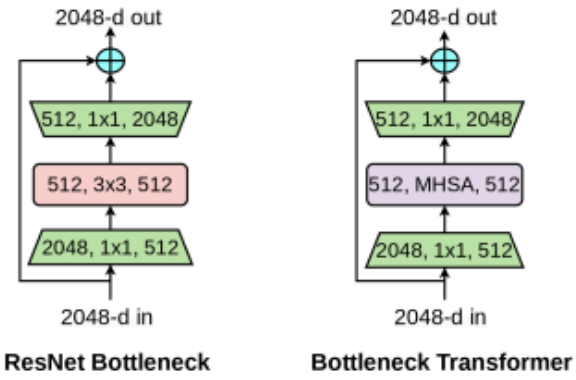
Προσθήκη 5-7 συνελικτικών στρωμάτων στην αρχή και τροφοδότηση της εξόδου που προκύπτει από αυτά στον κωδικοποιήτη του Vision Transformer. Αυτό έχει ως αποτέλεσμα να μπορούν να χρησιμοποιηθούν και ο adam και ο SGD βελτιστοποιητής (optimizer) με την ίδια αποδοτικότητα, κάτι που δεν συνέβαινε προηγουμένως και για αυτό υπήρχε προτίμηση στον adam βελτιστοποιητή. Ακόμη καθίσταται ευκολότερη η βελτιστοποίηση των παραμέτρων του μετασχηματιστή, καθώς μεγαλύτερο πλήθος συνδυασμών παραμέτρων εμφανίζει αποτελέσματα κοντά στα βέλτιστα του κάθε μοντέλου. Τέλος παρατηρείται η γρηγορότερη σύγκλιση του μοντέλου κατά την διάρκεια της εκπαίδευσης [48].

2.7.3 Bottleneck Transformer (BoTNet)

Το μοντέλο BoTNet (Bottleneck Transformers) αποτελείται απο διαδοχικά ResNet blocks, τα 3 τελευταία εξ αυτών περιλαμβάνουν attention. Πιο συγκεκριμένα, ενώ το block αποτελείται από μία 1x1 και μία 3x3 συνέλιξη, στα τελευταία επίπεδα, η 3x3 συνέλιξη αντικαθίσταται από ένα επίπεδο Multi-Head Self Attention[49]. Το μοντέλο χρησιμοποιείται τόσο για κατηγοριοποίηση εικόνων, όσο και για ανίχνευση αντικειμένων (object detection) και για τμηματοποίηση περιπτώσεων (instance segmentation) [4]. Η αρχιτεκτονική του καινούργιου block σε σχέση με το αντίστοιχο ResNet block φαίνεται στο Σχήμα 2.4.

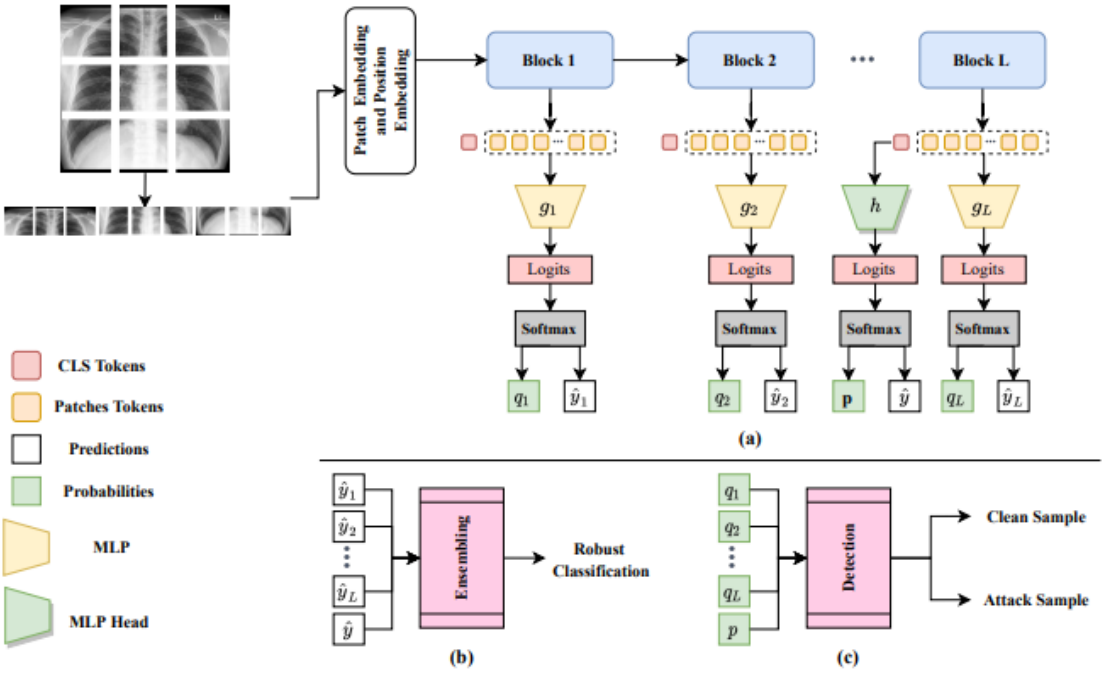
2.7.4 Self-Ensembling Vision Transformer (SeViT)

Ο Self-Ensembling Vision Transformer (SeViT) έχει ως στόχο την αντιμετώπιση των adversarial επιθέσεων, δηλαδή την επιβολή μικρών αλλαγών στις εικόνες με τέτοιο τρόπο,



Σχήμα 2.4: *Bottleneck Transformer block (δεξιά) σε σύγκριση με ResNet Bottleneck block (αριστερά) [4].*

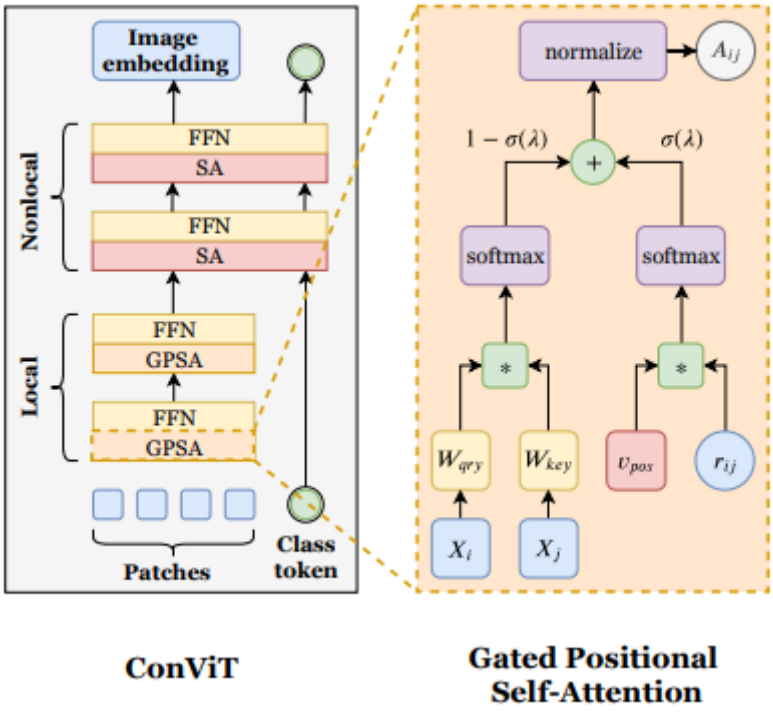
ώστε να αλλάζει η πρόβλεψη του μοντέλου. Για να επιτύχει αυτό τον στόχο εκμεταλλεύεται την ιδιότητα ότι τα αρχικά επίπεδα δεν επηρεάζονται σε μεγάλο βαθμό από αυτές τις διαταραχές. Πιο συγκεκριμένα, προσθέτει Multi Layer Perceptron (MLP) στο τέλος κάθε ενός από τα πρώτα m επίπεδα και συνδυάζει για την τελική πρόβλεψη τις προβλέψεις όλων των MLPs με το MLP που προϋπάρχει στο τέλος του Vision Transformer και στο οποίο τροφοδοτείται το αποτέλεσμα του κωδικοποιητή για το class token. Σε αντίθεση με το τελικό MLP, στα ενδιάμεσα δεν τροφοδοτούμε το class token που παρατηρήθηκε ότι δεν περιέχει αρκετή πληροφορία για να γίνει η διάκριση, αλλά τα patches[5]. Η αρχιτεκτονική του μοντέλου φαίνεται στην Εικόνα 2.5.



Σχήμα 2.5: *SeViT αρχιτεκτονική [5].*

2.7.5 ConViT

Το μοντέλο ConViT χρησιμοποιεί την ίδια αρχιτεκτονική με τον κλασσικό Vision Transformer με την διαφορά ότι στα πρώτα επίπεδα δεν χρησιμοποιεί Self-Attention Layers, αλλά εισάγει ένα καινούργιο είδος επιπέδου, το Gated Positional Self-Attention(GPSA), το οποίο φαίνεται στο Σχήμα 2.6. Το νέο αυτό επίπεδο είναι αρχικοποιημένο με τέτοιο τρόπο, ώστε να προσομοιώνει την τοπικότητα των συνελικτικών επιπέδων, με αποτέλεσμα στην αρχή να υπάρχουν επαγωγικές προκαταλήψεις και να μπορεί το μοντέλο να γενικεύεται ακόμα και όταν υπάρχουν λίγα δεδομένα. Στην συνέχεια δίνεται η δυνατότητα μέσω της παραμέτρου λ να δίνεται προσοχή περισσότερο στο περιεχόμενο και λιγότερο στην θέση, εκμεταλλεύοντας τα πλεονεκτήματα των Vision Transformer[6]. Η αρχιτεκτονική του μοντέλου φαίνεται στην Εικόνα 2.6.

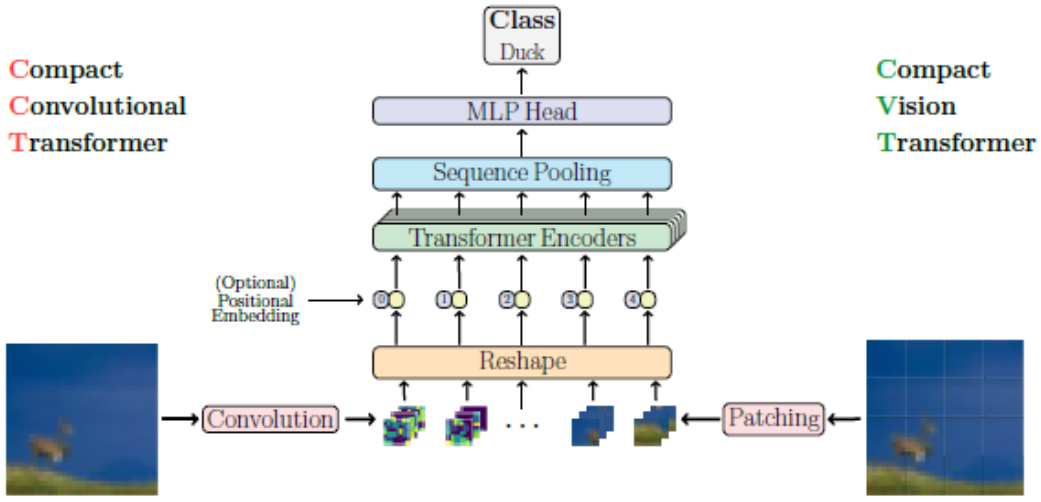


Σχήμα 2.6: *ConViT αρχιτεκτονική [6].*

2.7.6 Compact Convolutional Transformer (CCT)

Για την αντιμετώπιση του προβλήματος ότι για την εκπαίδευση των Vision Transformers χρειάζεται μεγάλο πλήθος δεδομένων, επινοήθηκαν οι Compact Transformers. Αρχικά παρουσιάζεται το μοντέλο ViT Lite, το οποίο χρησιμοποιεί την αρχιτεκτονική του ViT-Base με την διαφορά οτι περιλαμβάνει λιγότερα επίπεδα, λιγότερα heads στο Multi-Head attention επίπεδο, καθώς και μικρότερες διαστάσεις στα κρυφά επίπεδα. Ο Compact Vision Transformer (CVT) είναι παρόμοιος με το μοντέλο ViT-Lite με την αντικατάσταση του class token που χρησιμοποιούταν στην κατηγοριοποίηση από το τελικό MLP με μια δομή Sequence Pooling στο τέλος του Vision encoder και πριν την τροφοδότηση στο τελικό MLP. Αυτή η δομική μονάδα εφαρμόζει pooling στο σύνολο των tokens μετά την επεξεργασία τους από τον κωδι-

κοποιητή του μετασχηματιστή. Η διαφοροποίηση του Compact Convolutional Transformer (CCT) με το προηγούμενο μοντέλο είναι ότι αντί να χωρίζει την εικόνα σε patches, την τροφοδοτεί σε κάποια συνελικτικά επίπεδα προκειμένου να προσθέσει επαγωγικές προκαταλήψεις στο μοντέλο και να αυξήσει την απόδοση του όταν υπάρχει διάθεση λίγο δεδομένων. Όλα τα παραπάνω παρέχουν την δυνατότητα να εκπαιδεύονται τα μοντέλα από την αρχή και να παρουσιάζουν καλά αποτελέσματα και σε σύνολα δεδομένων με μικρό ή μεσαίο μέγεθος, έχοντας παράλληλα μειωμένα υπολογιστικά κόστη και απαιτήσεις μνήμης. Πειράματα εφαρμόζονται στα γνωστά σύνολα δεδομένων CIFAR, MNIST, Flowers και ImageNet-1k χωρίς να έχει προηγηθεί προεκπαίδευση σε μεγαλύτερα σύνολα δεδομένων[7]. Οι αρχιτεκτονικές των CVT και CCT φαίνονται στο παρακάτω Σχήμα 2.7.

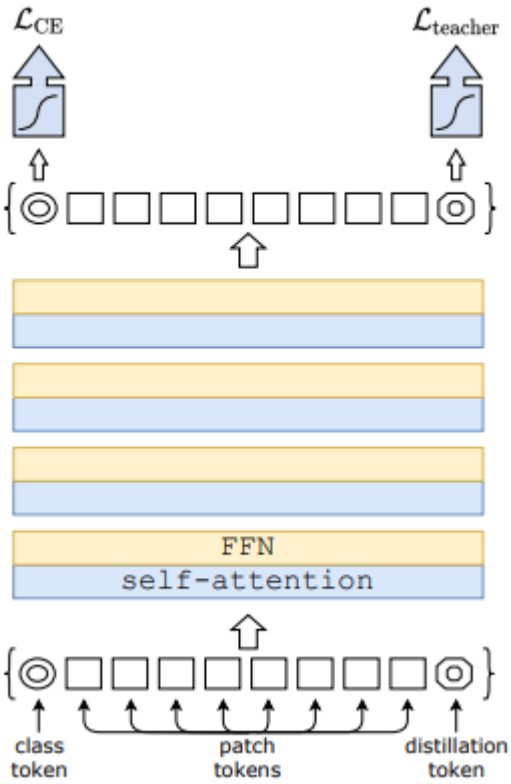


Σχήμα 2.7: *CVT και CCT αρχιτεκτονική [7].*

2.7.7 Data-efficient image Transformer (DeiT)

Το Data-efficient image Transformers (DeiT) είναι ένα μοντέλο που στηρίζεται στην λογική του δάσκαλου-μαθητή. Στις περισσότερες περιπτώσεις ως δάσκαλος χρησιμοποιείται κάποιο CNN και συνήθως το RegNetY-16GF, ενώ ως μαθητής χρησιμοποιείται κάποιος Vision Transformer. Στον μοντέλο του Vision Transformer προστίθεται και ένα distillation token, το οποίο λειτουργεί όπως το class token που υπάρχει ήδη στο μοντέλο με την διαφορά ότι κατά την εκπαίδευση του δεν λαμβάνονται υπόψη οι πραγματικές ετικέτες των εικόνων, αλλά τα αποτελέσματα του αντίστοιχου δασκάλου για αυτές. Η βελτίωση της απόδοσης του μοντέλου με την χρήση CNN ως δάσκαλο πιθανότατα οφείλεται στο γεγονός ότι προσθέτει επαγωγικές προκαταλήψεις μέσω του distillation token. Αν ληφθεί υπόψη μόνο η τελική πρόβλεψη του δασκάλου, η διαδικασία ονομάζεται hard distillation, ενώ αν ληφθεί υπόψη η έξοδος του softmax για όλες τις ετικέτες η διαδικασία ονομάζεται soft distillation. Στο αρχικό μοντέλο δασκάλου-μαθητή μπορεί να λειτουργήσει και το class token ως distillation token. Στο paper [8] γίνονται πειράματα για εκπαίδευση και δοκιμη στο σύνολο δεδομένων ImageNet, fine-tuning, προεκπαδευμένων στο προαναφερθέν σύνολο δεδομένων, μοντέλων

σε μικρότερα σύνολα δεδομένων , καθώς και εκπαίδευση από την αρχή σε μικρότερα σύνολα δεδομένων, χρησιμοποιώντας τεχνικές αύξησης δεδομένων (augmentation) και regularization. Το DeiT παρουσιάζει παρόμοια αποτελέσματα με τα προϋπάρχοντα μοντέλα όσον αφορά το trade-off ακρίβειας και throughput, ενώ παράλληλα χρειάζεται και λιγότερα δεδομένα από του συνηθισμένους Vision Transformers για την εκπαίδευση του. Η αρχιτεκτονική του μοντέλου DeiT με την προσθήκη του distillation token φαίνεται στο Σχήμα 2.8.

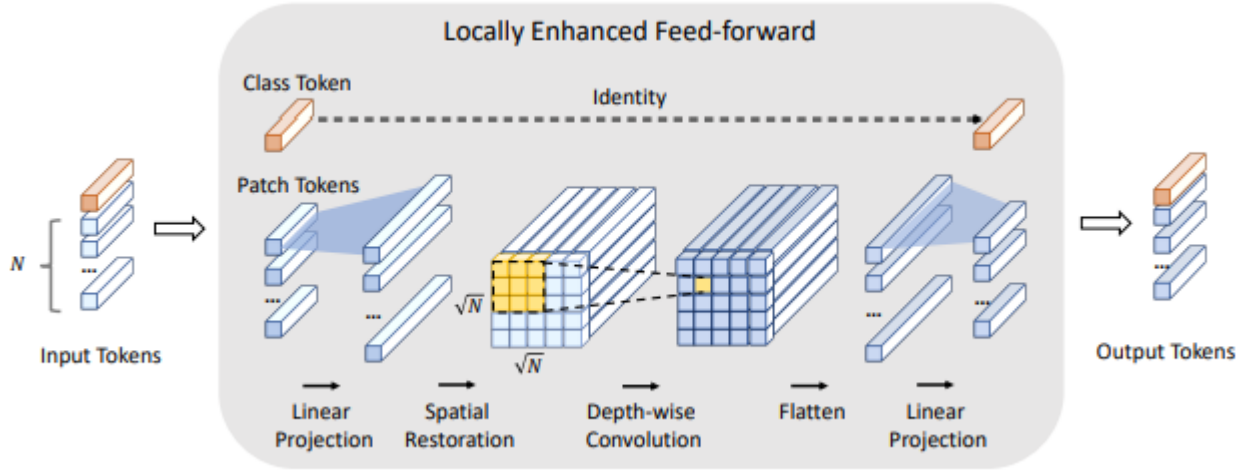


Σχήμα 2.8: *DeiT αρχιτεκτονική μαθητή με την προσθήκη distillation token [8].*

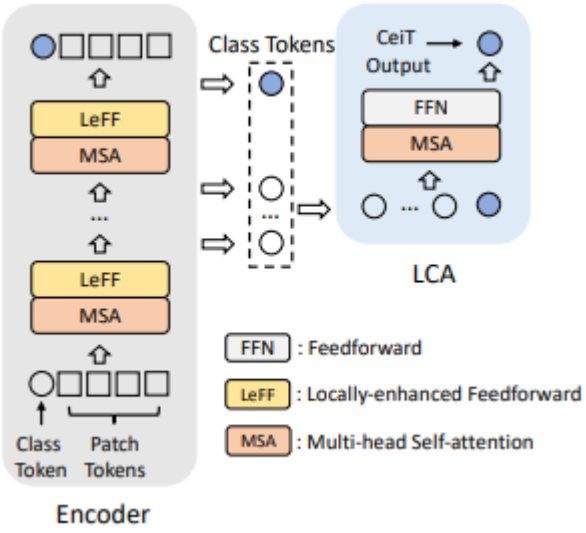
2.7.8 Convolution-Enhanced Image Transformer (CeiT)

Το CeiT (Convolution-Enhanced Image Transformer) έχει 3 βασικές διαφορές με τον απλό Vision Transformer. Αρχικά με μέγεθος patch 16 και 32 που χρησιμοποιείται συνήθως είναι δύσκολο να συλληφθούν χαμηλού επιπέδου χαρακτηριστικά, ενώ παράλληλα τα μοντέλα αυτά διαθέτουν και πολλές παραμέτρους που είναι δύσκολο να βελτιστοποιηθούν. Γι'αυτό για την εξαγωγή των patches χρησιμοποιείται μια μονάδα που λέγεται Image-to-Tokens (I2T), η οποία περιλαμβάνει επίπεδα που είναι βασισμένα σε επίπεδα των CNN, όπως συνελικτικά επίπεδα, max-pooling layer και batch normalization. Ακόμη το MLP σε κάθε επίπεδο του κωδικοποιητή του Vision Transformer αντικαθίσταται από το επίπεδο Locally-enhanced Feed-Forward (LeFF) που κάνει χρήση συνελίξεων κατά βάθος (depth-wise convolutions), ενώ στο τέλος του κωδικοποιητή προστίθεται ένα επίπεδο που ονομάζεται last class token attention (LCA) και συνδυάζει εξόδους από προηγούμενα επίπεδα[49].

Σκοπός του μοντέλου αποτελεί να είναι πιο υπολογιστικά αποδοτικό σε σχέση με τους Vision Transformers και να παρουσιάζει βελτιωμένα αποτελέσματα όταν εκπαιδεύεται σε μικρότερα σύνολα δεδομένων, ενώ παράλληλα επιδιώκει να επιτύχει γρηγορότερα (σε λιγότερες εποχές) σύγκλιση στις βέλτιστες επιδόσεις του. Παρουσιάζονται αποτελέσματα σε σύνολα δεδομένων όπως το ImageNet αλλά και fine-tuning προεκπαιδευμένων στο ImageNet μοντέλων στα Cifar10, Cifar100, Cars, Flowers, Pets και άλλα μικρά σύνολα δεδομένων[9]. Στα παρακάτω Σχήματα 2.9, 2.10 παρουσιάζονται η δομή του επιπέδου LeFF και η δομή του μοντέλου αντίστοιχα.



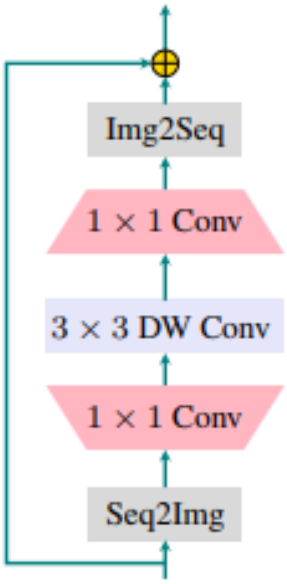
Σχήμα 2.9: *Η αρχιτεκτονική του επιπέδου Locally-enhanced Feed-Forward (LeFF) του μοντέλου CeiT [9].*



Σχήμα 2.10: *Η αρχιτεκτονική του μοντέλου CeiT [9].*

2.7.9 LocalViT (Local Vision Transformer)

Η αρχιτεκτονική του μοντέλου LocalViT (Local Vision Transformer) είναι παρόμοια με την αρχιτεκτονική του κλασσικού Vision Transformer με την διαφορά ότι, επειδή η δομική μονάδα του Multi-Head Attention(MSA) είναι πιο εξειδικευμένη στην εύρεση καθολικών εξαρτήσεων στις εικόνες, το συγκεκριμένο μοντέλο χρησιμοποιεί συνελίξεις κατά βάθος στα Feed-Forward νευρωνικά δίκτυα, προκειμένου να εντάξει τοπικές πληροφορίες των εικόνων στην αρχιτεκτονική του, αλλά και να βελτιώσει την υπολογιστική πολυπλοκότητα και τον αριθμό των παραμέτρων[49]. Αρχικά, η εικόνα αποκτά την δισδιάστατη μορφή της, τροφοδοτείται στο τροποποιημένο Feed-Forward Network (FFN) και στο τέλος γίνεται flatten και δημιουργούνται τα token. Το class token περνάει ξεχωριστά από το FFN και δεν συμμετέχει στον σχηματισμό της δισδιάστατης εικόνας. Στην έρευνα [10] γίνονται πειράματα στο σύνολο δεδομένων ImageNet2012 που δοκιμάζουν την επιρροή του συνελίξεων κατά βάθος στην επίδοση του μοντέλου, την σημασία μη γραμμικών συναρτήσεων ενεργοποίησης μετά από τις συνελίξεις κατά βάθος, καθώς και την αναλογία επέκτασης των διαστάσεων των κρυφών επιπέδων. Η μορφή του Feed-Forward Neural Network είναι παρουσιάζεται στο Σχήμα 2.11.

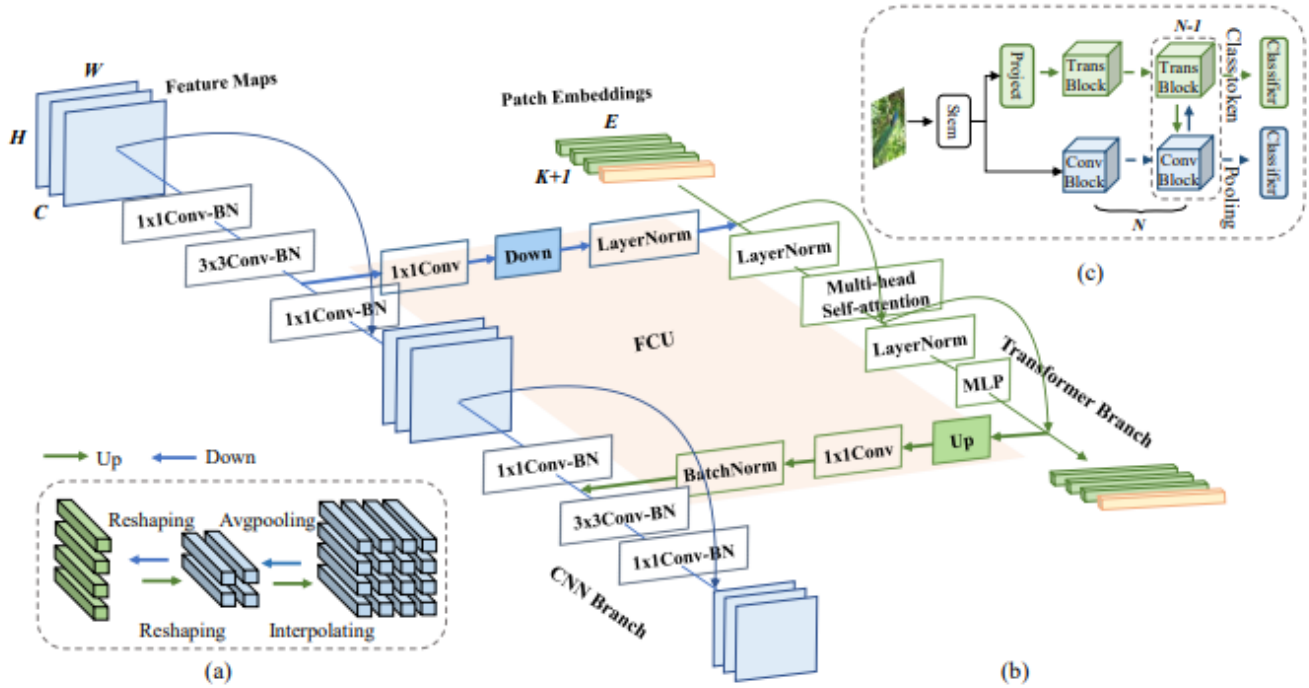


Σχήμα 2.11: *Η αρχιτεκτονική του Feed-Forward Neural Network του μοντέλου LocalViT [10].*

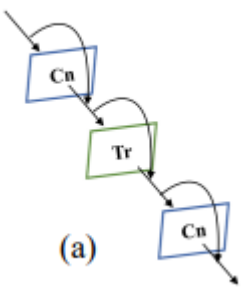
2.7.10 Conformer

Ο Conformer αποτελείται από δύο διακλαδώσεις, την CNN διακλάδωση και την Transformer διακλάδωση. Η πρώτη ακολουθεί την δομή του ResNet, ενώ η δεύτερη την δομή των Vision Transformers. Αρχικά τα δεδομένα περνούν από μία απλή δομή (stem) που εξάγει κάποια τοπικά χαρακτηριστικά και τα τροφοδοτεί στις δύο διακλαδώσεις. Ενδιάμεσα στις δυο διακλαδώσεις (σε όλα τα επίπεδα εκτός από το πρώτο) υπάρχει ένα δομικό στοιχείο που ονομάζεται Feature Coupling Unit(FCU), το οποίο είναι υπεύθυνο για το down-sampling της εξόδου των CNNs κόμβων και το up-sampling της εξόδου των ViT κόμβων, προκειμένου

να μπορούν να χρησιμοποιηθούν τα δεδομένα της μίας διακλάδωσης από την άλλη. Η έξοδος του κάθε επιπέδου ViT χρησιμοποείται από το επόμενο ViT και CNN επίπεδο (μέσω του FCU). Αντίστοιχα η έξοδος του CNN κάθε επιπέδου χρησιμοποείται από το CNN και ViT του επόμενου επιπέδου. Στο τέλος κάθε διακλάδωσης υπάρχει ένας ταξινομητής και η τελική πρόβλεψη προκύπτει από τον συνδυασμό αυτών των δύο[11]. Αφού η CNN διακλάδωση προσθέτει τοπικά χαρακτηριστικά δεν χρειάζονται πλέον positional embeddings. Η αρχιτεκτονική του μοντέλου φαίνεται πιο αναλυτικά στα παρακάτω Σχήματα 2.12, 2.13.



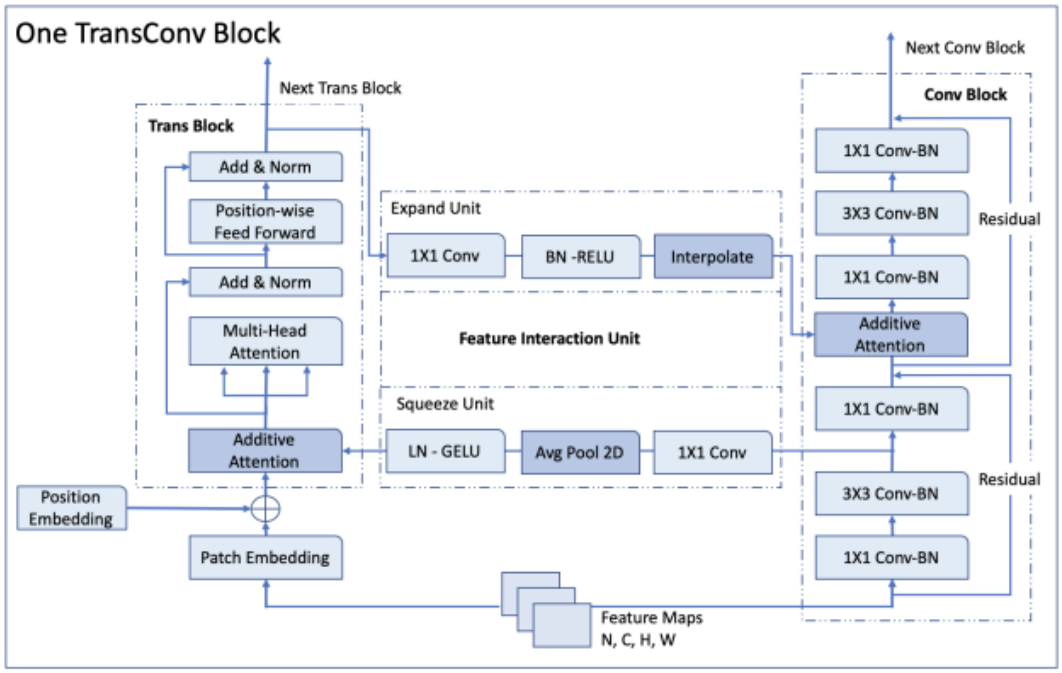
Σχήμα 2.12: *Conformer αρχιτεκτονική (πιο αναλυτικά) [11].*



Σχήμα 2.13: *Conformer αρχιτεκτονική [11].*

2.7.11 RobustVision

Το μοντέλο RobustVision χρησιμοποιεί παρόμοια αρχιτεκτονική με το προαναφερθέν μοντέλο του Conformer, με κάποιες τροποποιήσεις που καθιστούν δυνατή την χρήση προεκπαιδευμένων Vision Transformer και CNN [12]. Αρχικά αφαιρείται το αρχικό στάδιο (stem) που εξάγει κάποια τοπικά χαρακτηριστικά, καθώς οι μετασχηματιστές είναι προεκπαιδευμένοι χρησιμοποιώντας ως είσοδο τα patches των εικόνων και όχι patches από τους χάρτες χαρακτηριστικών που παρέχονται στην περίπτωση του Conformer. Αντίστοιχα σε αυτή την περίπτωση μπορούν να χρησιμοποιηθούν προεκπαιδευμένα μοντέλα CNN. Ακόμη χρησιμοποιούνται L επιπλέον Transformer και M CNN blocks, έτσι ώστε να αποκτήσουν την ίδια δομή με τα αντίστοιχα προεκπαιδευμένα μοντέλα. Ανάμεσα στα πρώτα $K = min(N+L, N+M)$ επίπεδα χρησιμοποιείται η δομική μονάδα Feature Interaction Unit (FIU) για την αλληλεπίδραση μεταξύ των διακλαδώσεων. Στο τέλος υπάρχουν δύο ταξινομητές τα αποτελέσματα των οποίων κατά τη διάρκεια των δοκιμών απλώς προσθέτονται μεταξύ τους. Η αρχιτεκτονική του μοντέλου φαίνεται πιο αναλυτικά στο παρακάτω Σχήμα 2.14.

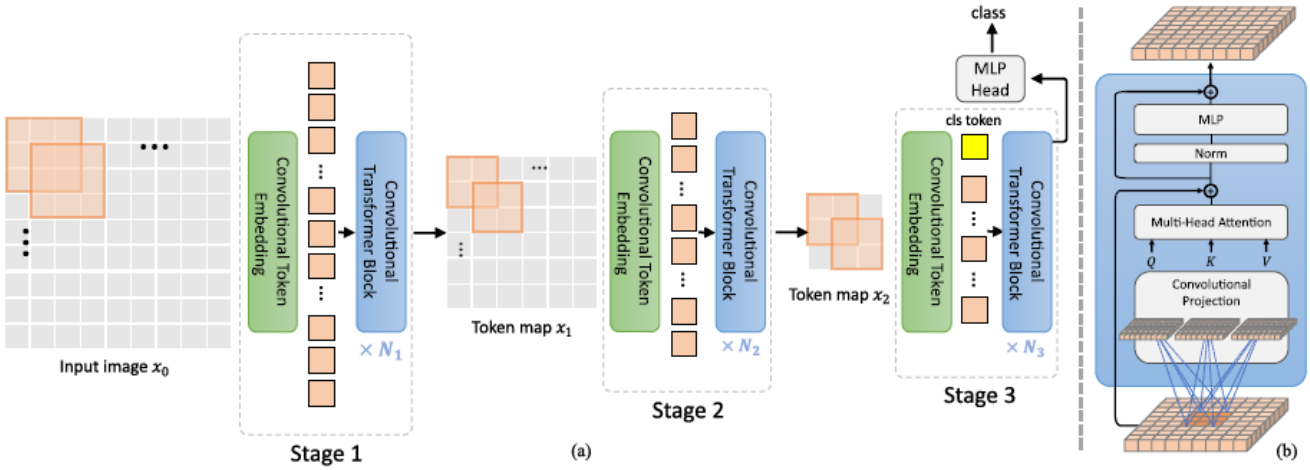


Σχήμα 2.14: *RobustVision αρχιτεκτονική [12].*

2.7.12 Convolutional vision Transformer (CvT)

Το μοντέλο Convolutional vision Transformer (CvT) προσθέτει κάποια συνελικτικά επίπεδα, προκειμένου να βελτιώσει τις επιδόσεις των Vision Transformer σε μικρότερα σύνολα δεδομένων, διατηρώντας παράλληλα τα οφέλη των προεκπαιδευμένων μοντέλων που γίνονται fine-tune σε μικρότερα σύνολα δεδομένων. Οι βασικές προσθήκες αυτού του μοντέλου σε σύγκριση με τον παραδοσιακό Vision Transformer είναι το Convolutional Token και το Co-

nvolutional projection. Το Convolutional Token χρησιμοποιείται για την δημιουργία των patches, πριν από την είσοδο τους στο κάθε Vision Transformer block και αποτελείται από μια δισδιάστατη συνελικτική διαδικασία με μέγεθος πυρήνα s x s και stride s - o. Με αυτόν τον τρόπο μειώνεται το μήκος του συνόλου των token και αυξάνονται οι διαστάσεις χαρακτηριστικών τους, δίνοντας τους την δυνατότητα να αναπαριστούν πιο περίπλοκα πρότυπα πάνω σε μεγαλύτερα χωρικά αποτυπώματα. Το Convolutional projection υπάρχει στην αρχή του κάθε block και έχει ως στόχο την περαιτέρω μοντελοποίηση τοπικά των χωρικών περιεχομένων και να προσφέρει μεγαλύτερη αποδοτικότητα στο μοντέλο υποδειγματολειπτώντας τους K και V πίνακες που χρησιμοποιούνται στο attention layer. Στην αρχή τα token ξαναπαίρνουν την μορφή μιας δισδιάστατης εικόνας, εφαρμόζεται το Convolutional projection και στην συνέχεια η έξοδος γίνεται flattened στην μία διάσταση. Τέλος δεν γίνεται χρήση positional embeddings, καθώς δεν προσφέρουν κάτι παραπάνω στις επιδόσεις του μοντέλου. Στην αντίστοιχη μελέτη [13] παρουσιάζονται αποτελέσματα πειραμάτων για την εκπαίδευση από την αρχή και την δοκιμή στο σύνολο δεδομένων ImageNet, καθώς και εφαρμογή προεκπαιδευμένων μοντέλων σε μικρότερα σύνολα δεδομένων, όπως ImageNet-1k, CIFAR10, CIFAR100, Pets, Flowers102. Η αρχιτεκτονική του μοντέλου CvT φαίνεται στο Σχήμα 2.15.



Σχήμα 2.15: *CvT αρχιτεκτονική. [13]*

Κεφάλαιο 3

Περιγραφή θέματος

Στο Κεφάλαιο αυτό παρουσιάζεται το θέμα της παρούσας εργασίας και πιο συγκεκριμένα η κατηγοριοποίηση ιατρικών εικόνων και ειδικότερα εικόνων που σχετίζονται με την ασθένεια COVID-19, χρησιμοποιώντας μοντέλα που είναι εμπνευσμένα από τα CNNs και τους Vision Transformers, καθώς και η σύγκριση τους με τα μοντέλα αυτά. Ακόμη παρουσιάζονται σχετικές εργασίες που έχουν πραγματοποιηθεί και αφορούν την επιρροή των CNNs και των ViTs στον τομέα της κατηγοριοποίησης ιατρικών εικόνων και συγκεκριμένα με την ανίχνευση της ασθένειας του COVID-19.

3.1 Κατηγοριοποίηση Ιατρικών εικόνων

Η κατηγοριοποίηση ιατρικών εικόνων είναι μια υποκατηγορία της εφαρμογής κατηγοροποίησης εικόνων που ανήκει στον τομέα της όρασης υπολογιστών. Στόχος της παραπάνω εφαρμογής είναι η έγκαιρη και αποτελεσματική αναγνώριση ασθενειών, προκειμένου να καθίσταται ευκολότερη η άμεση αντιμετώπιση τους. Ιδιαίτερα στην εφαρμογή της ανίχνευσης του COVID-19 είναι απαραίτητη η γρήγορη και αποδοτική διάγνωση της ασθένειας, καθώς αποτελεί μία επικίνδυνη νόσο που μεταδίδεται πολύ γρήγορα και μπορεί να βλάψει σε πολύ μεγάλο βαθμό την υγεία των ασθενών.

Υπάρχουν τρεις παράγοντες που καθιστούν την εφαρμογή της κατηγοριοποίησης ιατρικών εικόνων πιο δύσκολη από την κατηγοριοποίηση εικόνων διαφορετικού είδους. Μία από τις ιδιαιτερότητες των ιατρικών εφαρμογών είναι η ανάγκη για άμεση διάγνωση και αντιμετώπιση, όπως αναλύθηκε παραπάνω. Ακόμη, είναι δύσκολη η εύρεση αξιόπιστων και μεγάλων ιατρικών συνόλων δεδομένων. Αυτό συμβαίνει, καθώς λόγω της προστασίας προσωπικών δεδομένων είναι δύσκολη η συλλογή και η χρήση δεδομένων από ασθενείς, αφού πρέπει να εξασφαλιστεί η ιδιωτικότητα του κάθε ανθρώπου. Παράλληλα, είναι απαραίτητη η επιβεβαίωση της αξιοπιστίας των δεδομένων από εξειδικευμένους επαγγελματίες, καθώς η χρήση μη αξιόπιστων και μη έγκυρων δεδομένων για την πρόβλεψη μιας ασθένειας μπορεί να έχει καταστροφικές συνέπειες για την υγεία των ασθενών. Επιπλέον, η συλλογή των δεδομένων γίνεται από διαφορετικά ιδρύματα, τα οποία συλλέγουν τα δεδομένα σε διαφορετικές μορφές και η προσαρμογή τους χρειάζεται χρόνο και εξειδικευμένους ανθρώπους. Επιπροσθέτως, πολλές φορές οι ίδιοι οι οργανισμοί που συλλέγουν τα δεδομένα είναι απρόθυμοι να τα μοιραστούν, είτε για την προστασία των προσωπικών δεδομένων είτε γιατί χρησιμοποιούνται για δική τους έρευνα. Τέλος, μια ακόμη ιδιαιτερότητα των ιατρικών συνόλων δεδομένων

είναι ότι η επεξηγησιμότητα είναι πολύ σημαντική για αυτά, προκειμένου να γίνεται κατανοητή η λογική πίσω από τις προβλέψεις των μοντέλων. Όλα τα παραπάνω δυσχεραίνουν την δημιουργία νέων αξιόπιστων δεδομένων και συνεπώς προκαλούν νέες προκλήσεις στην κατηγοριοποίηση των εικόνων.

Τα τελευταία χρόνια, λόγω της επιτυχίας που είχαν οι Transformers στον τομέα της επεξεργασίας φυσικής γλώσσας [50] σε εφαρμογές όπως μετάφραση ομιλίας [51], παραγωγή φυσική γλώσσας [52], αναγνώριση ομιλίας [53] και αναγνώριση συναισθήματος από ομιλία [54], άρχισαν να δοκιμάζονται τα αντίστοιχα μοντέλα (Vision Transformer) και σε πολλες εφαρμογές στην περιοχή της όρασης υπολογιστών, με την κατηγοριοποίηση εικόνων να μην αποτελεί εξαίρεση. Τα μοντέλα αυτά, μέσω του self-attention, συλλαμβάνουν καθολικές εξαρτήσεις και πολύπλοκα μοτίβα σε αντίθεση με τα προϋπάρχοντα μοντέλα που εξάγουν τοπικά χωρικά μοτίβα και χαρακτηριστικά, με αποτέλεσμα να συμβάλουν σε μεγάλο βαθμό στην ανίχνευση ασθενειών και πιο συγκεκριμένα του COVID-19. Ωστόσο η ανάγκη τους για μεγάλο αριθμό δεδομένων εκπαίδευσης δυσχεραίνεται ακόμη περισσότερο στην κατηγοριοποίηση ιατρικών εικόνων, λόγω της δυσκολίας εύρεσης μεγάλων και αξιόπιστων ιατρικών συνόλων δεδομένων, όπως αναλύθηκε παραπάνω. Για αυτό το λόγο επινοήθηκαν τα υβριδικά CNN-ViT μοντέλα που κάνουν χρήση των πλεονεκτημάτων και των δύο αρχιτεκτονικών που συνδυάζονται και αποτελούν το βασικό αντικείμενο μελέτης της παρούσας εργασίας.

3.2 Σχετικές εργασίες

Στον τομέα της όρασης υπολογιστών και συγκεκριμένα στην εφαρμογή της κατηγοριοποίησης ιατρικών εικόνων υπάρχει πληθώρα ερευνών που ενέπνευσαν και το περιεχόμενο της παρούσας εργασίας. Αρχικά παρουσιάζονται σχετικές εργασίες με αρχιτεκτονικές βαθιών συνελικτικών δικτύων. Στην συνέχεια παραθέτονται έρευνες για COVID-19 σύνολα δεδομένων σε συνδυασμό με CNN και για ιατρικά σύνολα δεδομένων μαζί με ViT. Τέλος γίνεται αναφορά σε COVID-19 σύνολα δεδομένων σε συνδυασμό με ViT μοντέλα.

3.2.1 Βαθιά συνελικτικά δίκτυα

Αρχιτεκτονικές βαθιών νευρωνικών δικτύων έχουν υλοποιηθεί και χρησιμοποιηθεί σε διάφορες εφαρμογές από μέλη του Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης του ΕΜΠ. Ειδικότερα επιβλεπόμενες τεχνικές CNN και CNN-RNN έχουν εφαρμοστεί για κατηγοριοποίηση αντικειμένων, στην ιατρική διάγνωση νευροεκφυλιστικών ασθενειών, όπως της νόσου του Πάρκινσον [55] [56] [57] [58] [59] ή της Covid-19 [60] [61] [62] [63], περιλαμβάνοντας κατάτμηση 2-Δ ή 3-Δ εικόνων. Εμφαση έχει δοθεί στην διαφάνεια και στην προσαρμογή των μοντέλων [64] [65] [66] αλλά και στην ανάπτυξη πλέον σύνθετων αρχιτεκτονικών, μπαϋεσιανών, με κάψουλες και αβεβαιότητα [67] [68] [69] [70]. Βαθιές ημι- και αυτο-επιβλεπόμενες 3-Δ νευρωνικές αρχιτεκτονικές, αλλά και αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή έχουν εφαρμοστεί στην ανίχνευση βλαβών σε πυρηνικούς αντιδραστήρες [71] [72], στην πρόβλεψη της παραγωγής στον αγροτικό τομέα [73] [74] και στην αναγνώριση και σύνθεση συναισθήματος [75] [76] [77], ενώ άλλες εφαρμόζονται σε προβλήματα ανάλυσης εικόνων και αλληλεπίδρασης ανθρώπου-υπολογιστή [78] [79] [80] [81] [82].

3.2.2 COVID-19 σύνολα δεδομένων και CNN

Λόγω της μεγάλης εξάπλωσης του κορονοϊού, πολλές μελέτες έχουν επικεντρωθεί στην αυτόματη ανίχνευση του COVID-19 από ακτινογραφίες θώρακος, προτείνοντας μεθόδους βαθιάς μάθησης με υψηλή απόδοση [83] [84] [85] [86]. Υιοθετούνται ακόμη διαφορετικές τεχνικές μεταφοράς μάθησης χρησιμοποιώντας την extreme έκδοση του μοντέλου Inception (Xception) [87], μοντέλα όπως VGG16, VGG19, ResNet, DenseNet, InceptionV3 [88], CT εικόνες [89] [90], καθώς επίσης γίνεται χρήση των CNNs για αναγνώριση ειδών πνευμονίας συμπεριλαμβανομένης αυτής που προκαλέι ο COVID-19 [91].

Παράλληλα παρουσιάζονται καινοτόμες αρχιτεκτονικές δικτύων, όπως το STM-RENet [92], το COVID-CAPS [93], το GSA-DenseNet121-COVID-19 [94], το VGG16-CNN [95], το Faster R-CNN [96] και άλλα [97] [98]. Τέλος παραθέτονται κάποιες ensemble λύσεις [99], οι οποίες προτάθηκαν για τη βελτίωση των επιδόσεων των δικτύων στην ταξινόμηση του COVID-19, Normal και άλλων πνευμονικών παθήσεων. Η πλειονότητα από αυτές βασίζεται στα CNN λόγω της αποδοτικότητάς τους [100].

3.2.3 Ιατρικά σύνολα δεδομένων και ViT

Παρόλο που τα CNNs θεωρούταν state-of-the-art στον τομέα της όρασης υπολογιστών, οι Vision Transformers [15] γίνονται ολοένα και πιο δημοφιλείς. Όσον αφορά την ανάλυση ιατρικών εικόνων, διάφορες μελέτες έχουν χρησιμοποιήσει τα ViT [101] [102] [103] [47] για ταξινόμηση πολλών κλάσεων και τμηματοποίηση σε εφαρμογές για την ανίχνευση και διάγνωση πνευμονικών, καρδιακών, αμφιβληστροειδών και νευροεκφυλιστικών παθήσεων ή ακόμα και κάποιων μορφών καρκίνου.

Πιο συγκεκριμένα, παρακάτω θα αναλυθούν μοντέλα που χρησιμοποιούνται για την κατηγοριοποίηση εικόνων σε κάποια από τις παραπάνω ιατρικές εφαρμογές. Για την κατηγοριοποίηση όγκων σε καλοήθεις και κακοήθεις γίνεται χρήση του μοντέλου TransMed, το οποίο εκμεταλλέυεται την δομή των ViT [104]. Παρουσιάζει ανταγωνιστικά αποτελέσματα στην ταξινόμηση παρωτιδικού όγκου, αξιοποιώντας διαφορετικές μορφές πληροφορίας (multi-modal), χρησιμοποιώντας μια στρατηγική συγχώνευσης των διαφορετικών πληροφοριών. Αντίστοιχα, για την αναγνώριση ασθενειών του αμφισβληστροειδούς προτάθηκε το MIL-ViT [105], το οποίο είναι προεκπαιδευμένο σε μεγάλα σύνολα δεδομένων και γίνεται fine-tune σε μικρότερα σύνολα σχετικά με τις ασθένειες του αμφισβληστροειδούς. Μερικά ακόμη μοντέλα που χρησιμοποιούνται για την κατηγοριοποίηση ιατρικών εικόνων είναι τα TransMIL [106], Gene Transformer [107] και RadioTransformer [108]. Τέλος, κάποια ενδεικτικά μοντέλα που χρησιμοποιούνται για την τμηματοποίηση ιατρικών εικόνων είναι τα TransUNet [109], TransBTS [110] και Medical Transformer [111].

3.2.4 COVID-19 σύνολα δεδομένων και ViT

Τα τελευταία χρόνια, δημοσιεύονται ολοένα και περισσότερες μελέτες που αφορούν την ανίχνευση της ασθένειας του COVID-19 μέσω ακτινογραφιών θώρακα. Αρχικά, οι μελέτες προσπάθησαν να εξετάσουν τις επιδόσεις των μοντέλων ViT σε σύγκριση με τα CNN, κάνοντας τα fine-tune σε μεγάλα σύνολα δεδομένων και αποδεικνύοντας την ικανότητά τους να

προσφέρουν καλύτερη επεξηγησιμότητα [112] [113] [114]. Άλλες μελέτες πρότειναν τις δικές τους αρχιτεκτονικές βασισμένες στο Vision Transformer, με στόχο να ξεπεράσουν τα παραδοσιακά state-of-the-art ViT [115] [116] [117] [118] [119].

Μερικά μοντέλα που χρησιμοποιούνται για την κατηγοριοποίηση δισδιάστατων εικόνων COVID-19 [103] παρουσιάζονται παρακάτω. Αρχικά προτείνεται το μοντέλο Point-of-Care Transformer (POCFormer) [120], που αξιοποιεί τον Linformer [121], προκειμένου να μειώσει τον χώρο και τον χρόνο του self-attention επιπέδου από τετραγωνικό σε γραμμικό, ενώ αποτελείται από μόλις 2 εκατομύρια παραμέτρους (τις μισές από επίσης μικρά μοντέλα, όπως το MobileNetv2 [122]), καθιστώντας εφικτή την χρήση του ακόμη και σε φορητές συσκευές για την ανίχνευση του COVID-19 σε πραγματικό χρόνο.

Ακόμη παρουσιάζεται ένα καινούργιο μοντέλο, το οποίο προκειμένου να αντιμετωπίσει το γεγονός ότι τα self-attention επίπεδα συλλαμβάνουν καθολικές εξαρτήσεις σε πιο γενικό επίπεδο (coarse level), εντάσσει ένα καινούργιο attention επίπεδο που λέγεται Vision Outlooker (VOLO) [123]. Το επίπεδο αυτό συλλαμβάνει εξαρτήσεις σε πιο λεπτό επίπεδο (finer-level) και τις ενσωματώνει στα token που χρησιμοποιούνται για την αναπαράσταση. Παράλληλα τα μοντέλα Swin Transformer [124] και Transformer-in-Transformer [125] χρησιμοποιούνται για την διάκριση του COVID-19 από ασθενείς με πνευμονία και από υγιείς ασθενείς.

Επειδή οι συνέπειες της ασθένειας μπορεί να βρίσκονται σε διαφορετικά επίπεδα [126], για κάθε ασθενή γίνεται χρήση τρισδιάστατων εικόνων. Προτείνεται ένα υβριδικό μοντέλο που χρησιμοποιεί και δισδιάστατη και τρισδιάστατη πληροφορία [127], αποφασίζοντας για την σημασία της κάθε δισδιάστατης εικόνας με βάση σημαντικά συμπτώματα στις CT εικόνες που προκύπτουν από το Wilcoxon signed-rank test [128] με τον Swin Transformer ως δίκτυο κορμού.

Προκειμένου να αναπτυχθεί ένα πιο αποδοτικό και αξιόπιστο σύστημα για εφαρμογες ανάλυσης ιατρικής εικόνας, έχουν γίνει πολλές προσπάθειες για να συνδυαστούν τα καλύτερα χαρακτηριστικά και από τα δύο, CNNs και ViTs. Συνεπώς, δημιουργούνται υβριδικά μοντέλα CNN-ViT. Είναι, λοιπόν, ενδιαφέρον να δούμε πώς αυτά τα μοντέλα μπορούν να εφαρμοστούν στο πρόβλημα ανίχνευσης του COVID-19.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Δεδομένα

Στο Κεφάλαιο αυτό παρουσιάζονται και αναλύονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν στα πειράματα του Κεφαλαίου 6.

4.1 Σύνολα Δεδομένων

4.1.1 COVID-QU-Ex Dataset

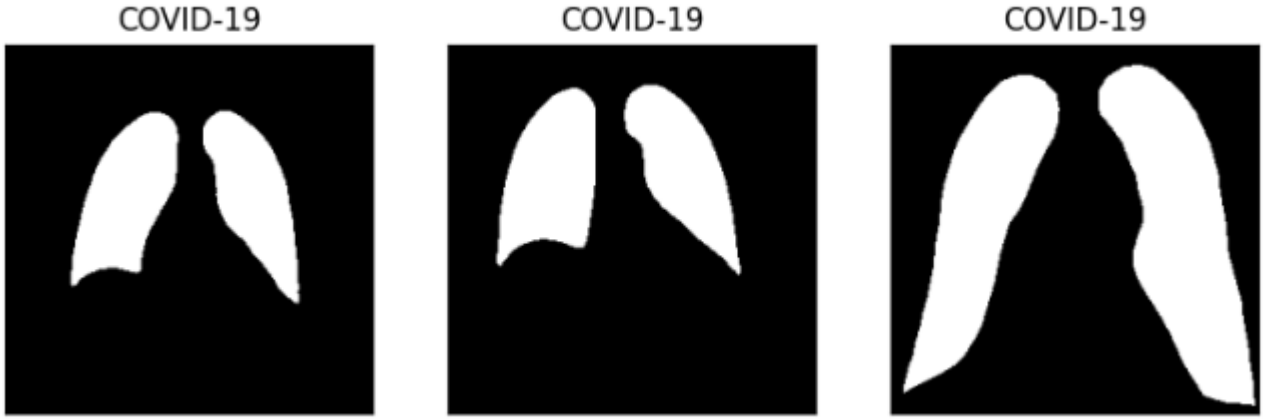
Το Dataset που χρησιμοποιήθηκε στα πειράματα που θα παρουσιαστούν στο Κεφάλαιο 6 είναι το COVID-QU-Ex, το οποίο εισήχθη για πρώτη φορά στην παρακάτω έρευνα [129]. Αποτελείται από 33.920 ακτινογραφίες θώρακος (Chest X-ray), οι οποίες είναι χωρισμένες σε 3 κλάσεις. Πιο συγκεκριμένα, αυτές οι 3 κλάσεις είναι οι COVID-19, Non-COVID που αντιπροσωπεύει ασθενείς που έχουν προσβληθεί από Ιογενή πνευμονία (Viral Pneumonia) ή Βακτηριακή Πνευμονία (Bacterial Pneumonia) και τέλος Normal για τους ανθρώπους που δεν έχουν επηρεαστεί από τις προαναφερθείσες ασθένειες που επηρεάζουν τους πνεύμονες. Ακόμη, για κάθε εικόνα το σύνολο δεδομένων περιλαμβάνει και μια μάσκα, η οποία αντιπροσωπεύει την περιοχή της εικόνας στην οποία βρίσκονται οι πνεύμονες, με αποτέλεσμα να αποτελεί το μεγαλύτερο σύνολο δεδομένων αυτού του είδους. Το σύνολο είναι χωρισμένο σε υποσύνολα Train, Test και Validation τα οποία περιλαμβάνουν 21.715, 6.788 και 5.417 δείγματα αντίστοιχα. Παραδείγματα εικόνων για τις 3 κλάσεις μαζί με τις αντίστοιχες μάσκες για τους πνεύμονες παρουσιάζονται στις Εικόνες 4.1, 4.2, 4.3, 4.4, 4.5 και 4.6.

Ακόμη υπάρχει ένα μικρό υποσύνολο που διαθέτει 1,456 Normal, 1,457 Non-COVID-19 και 2,913 COVID-19 εικόνες μαζί με τις αντίστοιχες μάσκες για τους πνεύμονες, όπως στο αρχικό σύνολο δεδομένων. Η διαφορά, όμως αυτού του υποσυνόλου είναι ότι για τις εικόνες που ανήκουν στην κλάση COVID-19 διαθέτει και μάσκες που αντιπροσωπεύουν την μολυσμένη περιοχή των πνευμόνων. Οι μάσκες που δείχνουν την περιοχή μόλυνσης για τις εικόνες που παρουσιάστηκαν στην Εικόνα 4.1 φαίνονται στην Εικόνα 4.7.

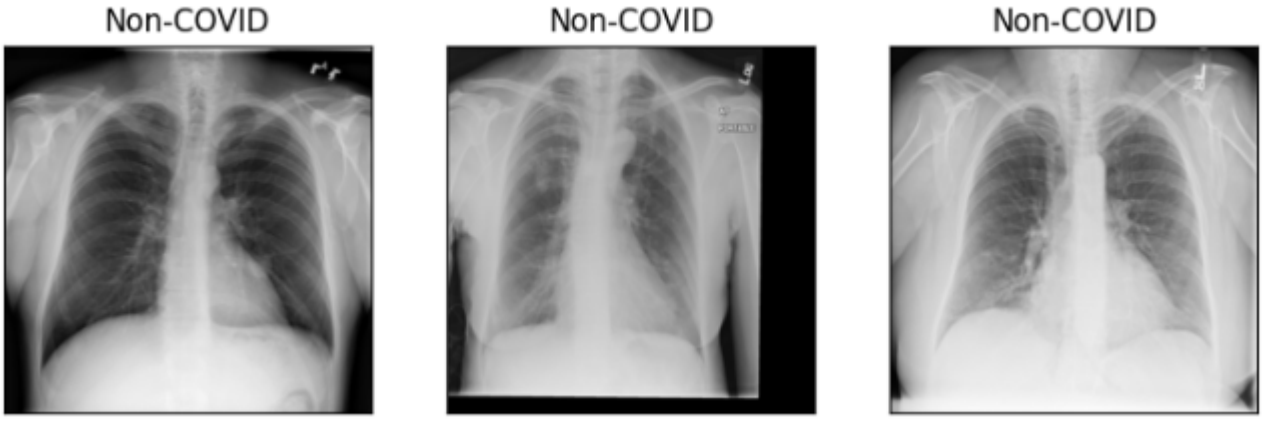
4.1.2 ImageNet

Το ImageNet είναι ένα πολύ μεγάλο σύνολο δεδομένων που αποτελείται από 14,197,122 εικόνες, οι ετικέτες των οποίων έχουν προσθεθεί με το χέρι [130]. Περιλαμβάνει περισσότερες από 21,000 κατηγορίες, για κάθε μία από τις οποίες περιέχει αρκετές εκατοντάδες εικόνες

Εικόνα 4.1: *Εικόνες κλάσης COVID-19.*



Εικόνα 4.2: *Μάσκες πνεύμονα κλάσης COVID-19.*



Εικόνα 4.3: *Εικόνες κλάσης Non-COVID.*

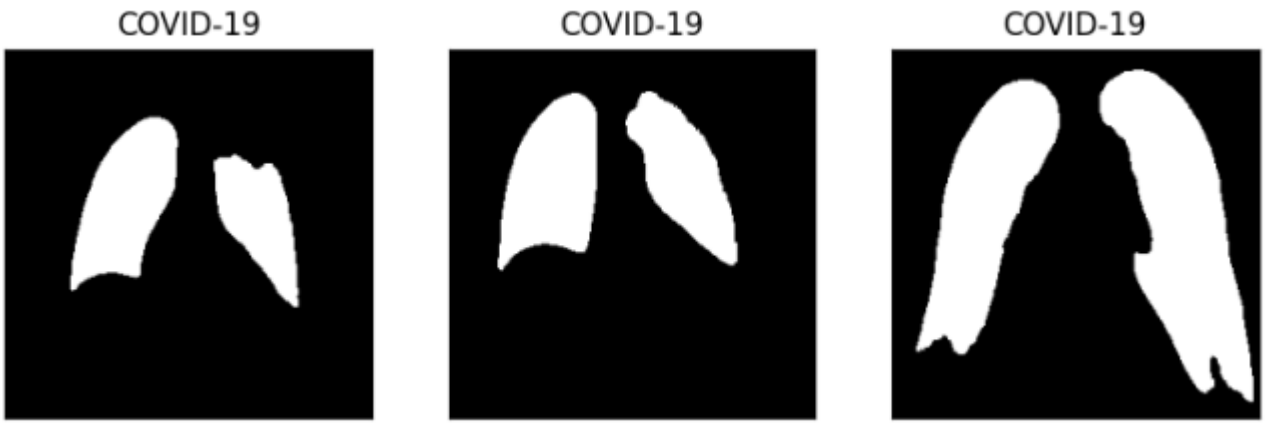
και χρησιμοποιείται σε εφαρμογές όπως η κατηγοριοποίηση εικόνων και ανιχνευση αντικειμένων [131]. Πολλές φορές αναφέρεται ως ImageNet-21k ή ImageNet-22k, λόγω του αριθμού των κλάσεων που περιλαμβάνει. Από το 2010 υποσύνολα του συγκεκριμένου συνόλου δεδομένων χρησιμοποιούνται στην δοκιμάσια ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Ένα υποσύνολο του ονομάζεται ImageNet 2012 ή ImageNet-1k και

Εικόνα 4.4: *Μάσκες πνεύμονα κλάσης Non-COVID.*

Εικόνα 4.5: *Εικόνες κλάσης Normal.*

Εικόνα 4.6: *Μάσκες πνεύμονα κλάσης Normal.*

περιλαμβάνει 1,281,167 εικόνες για την εκπαίδευση, 50,000 εικόνες για το την επικύρωση (validation) και 100,000 εικόνες για τις δοκιμές (testing). Οι εικόνες αυτές είναι χωρισμένες σε 1000 κλάσεις [132].



Εικόνα 4.7: *Μάσκες μόλυνσης κλάσης COVID-19.*

Κεφάλαιο 5

Ανάλυση και σχεδίαση

Στο κεφάλαιο αυτό θα παρουσιαστούν τα μοντέλα που θα χρησιμοποιηθούν στα πειράματα του Κεφαλαίου 6, θα αναλυθεί η αρχιτεκτονική τους, καθώς και η συγκεκριμένη έκδοση που χρησιμοποιήθηκε από το καθένα. Παράλληλα θα αναλυθούν λεπτομέρειες τις υλοποίησης, όπως οι υπερπαράμετροι των μοντέλων, η μετρική αξιολόγησης, το μηχάνημα στο οποίο διεξήχθησαν τα πειράματα, καθώς και η πορεία που ακολουθήθηκε κατά την διάρκεια των πειραμάτων. Τέλος προστίθεται το repository στο github, στο οποίο περιλαμβάνεται ο κώδικας.

5.1 Ανάλυση - Περιγραφή αρχιτεκτονικής - Vision Transformer

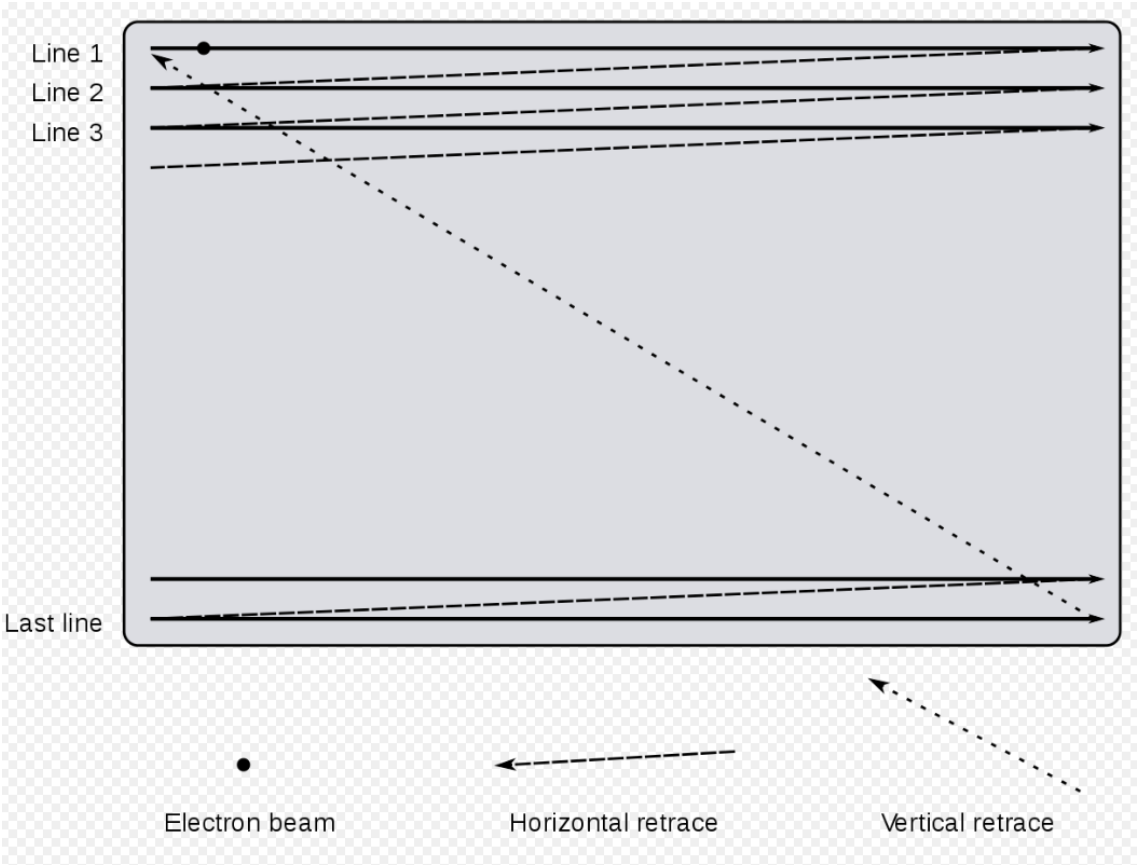
Στην παρακάτω υποενότητα περιγράφεται η δομή των Vision Transformers, καθώς και οι εκδόσεις τους που χρησιμοποιήθηκαν στα πειράματα. Στις περισσότερες περιπτώσεις επιλέχθηκαν μικρότερες εκδόσεις των ViT λόγω περιορισμού υπολογιστικών πορών.

5.1.1 Vision Transformers

Η λογική που ακολουθούν οι Vision Transformers είναι η διαίρεση της εκάστοτε εικόνας σε μη επικαλυπτόμενα κομμάτια σταθερού μεγέθους (patches) και η μετατροπή τους σε μονοδιάστατα embeddings μήκους D (tokens). Στην συνέχεια αυτά τα tokens τροφοδοτούνται στον κωδικοποιητή ενός μετασχηματιστή, ο οποίος τα μεταχειρίζεται όπως τα αντίστοιχα tokens στις εφαρμογές φυσικής γλώσσας. Ακομή προστίθεται ένα class token, για το οποίο η έξοδος του κωδικοποιητή αποτελεί την αναπαράσταση της εικόνας και το οποίο τροφοδοτείται στο MLP για την πραγματοποίηση της πρόβλεψης. Το καθε block του κωδικοποιητή αποτελείται με την σειρά από ένα επίπεδο κανονικοποίησης (Normalization layer) που χρησιμοποιείται για την σταθεροποίηση του δικτύου κάτι το οποίο οδηγεί σε μικρότερους χρόνους εκπαίδευσης και από το Multi-Head Attention Layer, το οποίο είναι υπεύθυνο για την σύλληψη μακρινών καθολικών εξαρτήσεων. Επιπροσθέτως, ακολουθεί πάλι ένα επίπεδο κανονικοποίησης και το Feed-forward network(MLP), που περιλαμβάνει δύο πλήρως συνδεδεμένα επίπεδα, μία μη γραμμική συνάρτηση ενεργοποίησης μεταξύ των επιπέδων (GELU) και χρησιμοποιείται για την εύρεση πιο πολύπλοκων χαρακτηριστικών. Ενδιάμεσα υπάρχουν και 2 υπολειμματικές συνδέσεις που βοηθούν στην βελτίωση της απόδοσης και την μεταφοράς της πληροφορίας.

Όσον αφορά την θέση των patches, η πληροφορία αυτή παρέχεται μέσω των μονοδιάστατων positional embeddings, τα οποία προστίθενται στα embeddings σταθερού μήκους D του κάθε patch και προκύπτουν λαμβάνοντας τα κομμάτια της εικόνας με σειρά raster, όπως φαίνεται στο Σχήμα 5.1. Η αρχιτεκτονική του Vision Transformer φαίνεται αναλυτικά στο Σχήμα 5.2.

Το μοντέλο αυτό αποτελείται από διάφορες εκδόσεις (Base, Large, Huge), οι οποίες εξαρτώνται από διάφορες παραμέτρους που φαίνονται στον Πίνακα 5.1, καθώς και από το μέγεθος των patches (8, 16, 32). Στα πειράματα του Κεφαλαίου 6 χρησιμοποιήθηκε η Base έκδοση των ViT, όπου οι λεπτομέρειες της παρουσιάζονται στον Πίνακα 5.1.

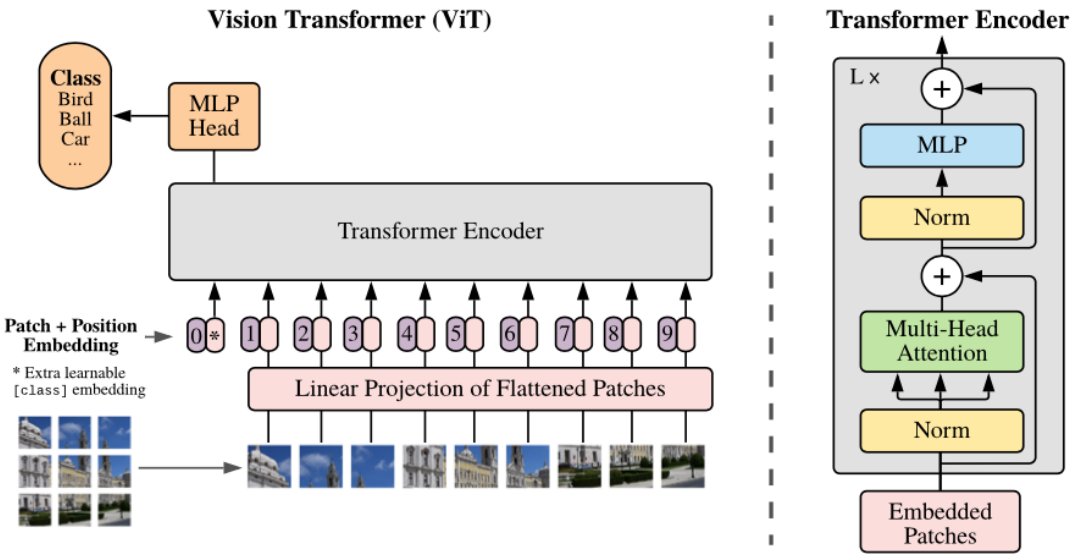


Σχήμα 5.1: *Σειρά Raster [14]*

Μοντέλο	**Αριθμός επιπέδων**	**Hidden size D**	**Μέγεθος MLP**	**Αριθμός Heads**	**Παράμετροι**
ViT-Base	12	768	3072	12	86.000.000
ViT-Large	24	1024	4096	16	307.000.000
ViT-Huge	32	1280	5120	16	632.000.000

Πίνακας 5.1: *Εκδόσεις ViT [15]*

Οι μετασχηματιστές συλλαμβάνουν μακρινές καθολικές εξαρτήσεις, κάτι που οφείλεται στα attention layers και δεν διαθέτουν επαγωγικές προκαταλήψεις και συνεπώς δεν διαθέτουν την ικανότητα γενίκευσης οταν εκπαιδεύονται σε μικρό σύνολο δεδομένων.



Σχήμα 5.2: *Αρχιτεκτονική Vision Transformer [15]*

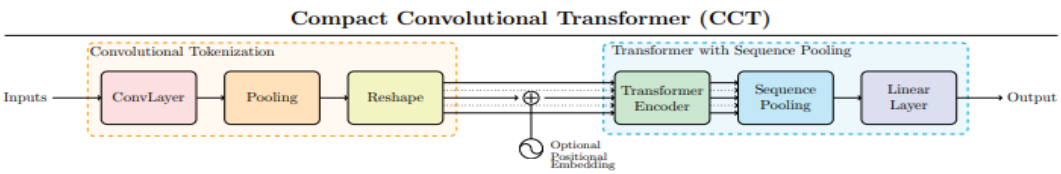
.

5.2 Ανάλυση - Περιγραφή αρχιτεκτονικής - Υβριδικά μοντέλα

Στις παρακάτω υποενότητες παρουσιάζεται η έκδοση κάθε ενός από τα έξι υβριδικά μοντέλα που θα χρησιμοποιηθούν στα πειράματα. Πιο συγκεκριμένα, τα μοντέλα αυτά είναι τα CCT, DeiT, CeiT, LocalViT, Conformer και CvT. Η γενικότερη περιγραφή κάθε ενός από τα έξι μοντέλα παρουσιάζεται στο Κεφάλαιο 2.

5.2.1 cct_14_7x2_224 (CCT)

Όπως φαίνεται και στο Σχήμα 5.3, πριν την είσοδο στον κωδικοποιητή του Transformer, η εικόνα δεν χωρίζεται απευθείας σε patches, αλλά τροφοδοτείται σε συνελικτικά επίπεδα, επίπεδα ομαδοποίησης και μετά αλλάζει το μέγεθος της για να έρθει στην κατάλληλη μορφή. Ο συμβολισμός cct_a/bxc σημαίνει ότι το CCT μοντέλο αποτελείται από a επίπεδα κωδικοποιητή μετασχηματιστή (transformer encoder) και c συνελικτικά επίπεδα με μέγεθος πυρήνα bxb στην αρχική δομή πριν τον κωδικοποιητή. Συνεπώς η έκδοση cct_14_7x2_224 αποτελείται από 14 transformer encoders και η αρχική δομή πριν τον κωδικοποιητή αποτελείται από 2 συνελικτικά επίπεδα με μέγεθος πυρήνα 7x7. Ακόμη περιλαμβάνει 6 heads στο Multi-Head Attention επίπεδο, αναλογία επέκτασης στο MLP ίση με 3 και 384 hidden dimensions D, ενώ παράλληλα δέχεται ως είσοδο εικόνες μεγέθους 224x224.



Σχήμα 5.3: *Δομή CCT μοντέλου [7]*

.

5.2.2 deit_tiny_patch16_224 (DeiT)

Υπάρχουν πολλές διαφορετικές εκδόσεις του μοντέλου που εξαρτώνται από το μέγεθος των embeddings, τον αριθμό των heads και των επιπέδων του κωδικοποιητή μετασχηματιστή, το μέγεθος της εισόδου, αλλά και το μέγεθος των patches. Τρεις βασικές εκδόσεις του μοντέλου DeiT φαίνονται στον Πίνακα 5.2.

Μοντέλο	Αριθμός επιπέδων	Hidden size D	Μέγεθος Εισόδου	Αριθμός Heads	Παράμετροι
DeiT-Ti	12	192	224^2	3	5.000.000
DeiT-S	12	384	224^2	6	22.000.000
DeiT-B	12	768	224^2	12	86.000.000

Πίνακας 5.2: *Εκδόσεις DeiT [8]*

Στα πειράματα του επόμενου Κεφαλαίου γίνεται χρήση του deit_tiny_patch16_224, το οποίο αποτελεί την έκδοση DeiT-Ti του Πίνακα 5.2 με μέγεθος patch ίσο με 16x16 και δάσκαλο το μοντέλο RegNetY-16GF.

5.2.3 ceit_tiny_patch16_224 (CeiT)

Η δομή Image-to-Token αποτελείται από ένα επίπεδο συνέλιξης με μέγεθος πυρήνα 7x7 και βήμα (stride) 2, ένα επίπεδο BatchNorm για σταθερή εκπαίδευση και ένα επίπεδο max-pooling με μέγεθος πυρήνα 3x3 και βήμα 2, δημιουργώντας χάρτες χαρακτηριστικών 4 φορές μικρότερους από την εικόνα εισόδου. Η δομή Locally-enhanced Feed-Forward χρησιμοποιεί αναλογία επέκτασης e ίση με 4 και μέγεθος πυρήνα στις συνελίξεις κατά βάθος ίσο με 3x3. Για το Last Class token Attention κομμάτι, η δομή του ακολουθεί την δομή των επιπέδων του κωδικοποιητή μετασχηματιστή.

Οι διαφορετικές εκδόσεις του CeiT παρουσιάζονται στον Πίνακα 5.3. Το μοντέλο ceit_-tiny_patch16_224 που εμφανίζεται στα πειράματα αποτελεί την έκδοση CeiT-T με μέγεθος patch 16x16 και μέγεθος εισόδου 224x224.

Μοντέλο	Αριθμός επιπέδων	Hidden size D	Μέγεθος Εισόδου	Αριθμός Heads	Παράμετροι
CeiT-T	12	192	224^2	3	6.400.000
CeiT-S	12	384	224^2	6	24.200.000
CeiT-B	12	768	224^2	12	86.600.000

Πίνακας 5.3: *Εκδόσεις CeiT [9]*

5.2.4 Localvit_small_mlp4_act3_r384 (LocalViT)

Οι βασικές εκδόσεις του LocalViT μοντέλου εξαρτώνται από βασικές παραμέτρους, όπως ο αριθμός των επιπέδων του κωδικοποιητή μετασχηματιστή, οι διαστάσεις των embeddings, το μεγεθος των εικόνων εισόδου και ο αριθμό των heads, και φαίνονται στον Πίνακα 5.4.

Μοντέλο	Αριθμός επιπέδων	Hidden size D	Μέγεθος Εισόδου	Αριθμός Heads
LocalViT-Tiny	12	192	224^2	4
LocalViT-Small	12	384	224^2	8

Πίνακας 5.4: *Εκδόσεις LocalViT [25]*

Ακόμη δημιουργούνται μοντέλα για τον πειραματισμό χαρακτηριστικών του καινούργιου Feed-Forward δικτύου, όπως είναι η αναλογία επέκτασης γ (1, 2, 3, 4, 6) και η μη γραμμική συνάρτηση ενεργοποίησης, όπου χρησιμοποιούνται οι ReLU6, h-swish σε συνδυασμό με το squeeze-and-excitation (SE) ή το efficient channel attention (ECA). Παράλληλα η προσθήκη του καινούργιου FFN μπορεί να γίνει σε κάποια από τα επίπεδα transformer encoder, για να προσθέσει τοπικότητα, παράγοντας μοντέλα, όπως το low (προστίθεται στα 4 ή 8 πρώτα επίπεδα), mid (επίπεδα 5-8), high (επίπεδα 9 με 12) και All (όλα τα επίπεδα). Στην δομή squeeze-and-excitation υπάρχει μια παράμετρος που λέγεται reduction και καθορίζει τον ρυθμό μείωσης των διαστάσεων. Οι παραπάνω αλλαγές μπορούν να εφαρμοστούν σε μοντέλα, όπως TNT, PVT, DeiT, T2T, Swin Transformer και να δημιουργήσουν τα αντίστοιχα LocalViT μοντέλα.

Το μοντέλο Localvit_small_mlp4_act3_r384 που χρησιμοποιήθηκε στα πειράματα αποτελεί ένα κλασσικό ViT, με την προσθήκη του αλλαγμένου FFN, με τα αντίστοιχα χαρακτηριστικά του LocalViT-Small στον Πίνακα 5.4. Ακόμει διαθέτει αναλογία επέκτασης γ ίση με 4 και η νέα δομή προστίθεται σε όλα τα επίπεδα. Τέλος το act_3 συμβολίζει ότι ως μη γραμμική συνάρτηση ενεργοποίησης χρησιμοποιείται ο συνδυασμός h-swish και squeeze-and-excitation με reduction διαστάσεις ίσες με 384.

5.2.5 Conformer_small_patch16 (Conformer)

Οι βασικές εκδόσεις του Conformer μοντέλου, όπως παρουσιάζονται στον Πίνακα 5.5, εξαρτώνται από τον αριθμό των επιπέδων του κωδικοποιητή μετασχηματιστή, τις διαστάσεις των embeddings, το μεγεθος των εικόνων εισόδου, τον αριθμό των heads και τον αριθμό των παραμέτρων που περιλαμβάνουν. Ακόμη, διαφορετικά μοντέλα μπορούν να δημιουργηθούν ανάλογα με το μέγεθος των patches (16, 32) στην Transformer διακλάδωση. Παράλληλα δοκιμάζονται διάφορες τιμές για το μέγεθος των embeddings και τον αριθμό των heads στην Transformer διακλάδωση και διάφορες τιμές (C) για τα κανάλια του δεύτερου σταδίου c_2 και τον αριθμό των bottlenecks (n_c) για κάθε επίπεδο στην CNN διακλάδωση.

Μοντέλο	Αριθμός επιπέδων	Hidden size D	Μέγεθος Εισόδου	Αριθμός Heads	Παράμετροι
Conformer-Ti	12	384	224^2	6	23.500.000
Conformer-S	12	384	224^2	6	37.700.000
Conformer-B	12	576	224^2	9	83.300.000

Πίνακας 5.5: *Εκδόσεις Conformer [26]*

Το bottleneck αποτελείται από μία 1x1 down-projection συνέλιξη, μία 3x3 χωρική συνε-

λιξη, μία up-projection συνέλιξη και μία υπολειμματική σύνδεση ανάμεσα στην είσοδο και την έξοδο της δομής. Επιπλέον, πριν την είσοδο των δεδομένων στις δύο διακλαδώσεις υπάρχει μια δομή, που αποτελείται από ένα επίπεδο συνέλιξης με μέγεθος πυρήνα 7x7 και βήμα 2, ένα max-pooling επίπεδο με μέγεθος πυρήνα 3x3 και βήμα 2, η οποία χρησιμοποιείται για εξαγωγή τοπικών χαρακτηριστικών.

Ο Conformer_small_patch16 αποτελεί την Conformer-S έκδοση του Πίνακα 5.5, με μέγεθος των patches 16x16, 2 bottlenecks σε κάθε επίπεδο της CNN διακλάδωσης και αριθμό καναλιών στο δεύτερο στάδιο ίσο με 256. Οι διαστάσεις που προκύπτουν σε κάθε επίπεδο του Conformer-S φαίνονται στον Σχήμα 5.4. Το 56×56,197 σημαίνει ότι ο χάρτης χαρακτηριστικών έχει μέγεθος 56x56 και ο αριθμός των ενσωματωμένων patches είναι 197.

stage	output	CNN Branch	FCU	Transformer Branch	
c1	112×112	7×7, 64, stride 2			
	56×56	3×3 max pooling, stride 2			
c2	56 × 56,197	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	-	4×4, 384, stride 4 $\begin{bmatrix} \text{MHSA-6, 384} \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	×1
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	$[1 \times 1, 384] \longrightarrow$	$\begin{bmatrix} \text{MHSA-6, 384} \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	×3
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	$\longleftarrow [1 \times 1, 64]$		
c3	28 × 28,197	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}$	$[1 \times 1, 384] \longrightarrow$	$\begin{bmatrix} \text{MHSA-6, 384} \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	×4
		$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}$	$\longleftarrow [1 \times 1, 128]$		
c4	14 × 14,197	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$[1 \times 1, 384] \longrightarrow$	$\begin{bmatrix} \text{MHSA-6, 384} \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	×3
		$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$\longleftarrow [1 \times 1, 256]$		
c5	7 × 7,197	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$[1 \times 1, 384] \longrightarrow$	$\begin{bmatrix} \text{MHSA-6, 384} \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	×1
		$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	$\longleftarrow [1 \times 1, 256]$		
classifier	1 × 1, 1	global pooling	-	class token	
		1×1,1000	-	1×1,1000	
Parameters		**37.7 M**			
MACs		**10.6 G**			

Σχήμα 5.4: *Δομή Conformer μοντέλου [11]*

.

5.2.6 cvt-13-224x224 (CvT)

Οι τρεις παραλλαγές του CvT μοντέλου φαίνονται στην Εικόνα 5.1. Conv. Embed. και Conv. Proj. σημαίνουν Convolutional Token Embedding και Convolutional Projection αντίστοιχα, ενώ H_i και D_i εκφράζουν τον αριθμό των heads και τις διαστάσεις των embeddings του i-οστού Multi-Head Self-Attention επιπέδου. Τέλος το R_i εκφράζει την αναλογία επέκτασης των διαστάσεων στο i-οστό MLP επίπεδο.

	Output Size	Layer Name	CvT-13	CvT-21	CvT-W24
Stage1	56×56	Conv. Embed.	7×7, 64, stride 4		7×7, 192, stride 4
	56×56	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 64 \\ H_1 = 1, D_1 = 64 \\ R_1 = 4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 64 \\ H_1 = 1, D_1 = 64 \\ R_1 = 4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 192 \\ H_1 = 3, D_1 = 192 \\ R_1 = 4 \end{bmatrix} \times 2$
Stage2	28×28	Conv. Embed.	3×3, 192, stride 2		3×3, 768, stride 2
	28×28	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 192 \\ H_2 = 3, D_2 = 192 \\ R_2 = 4 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 192 \\ H_2 = 3, D_2 = 192 \\ R_2 = 4 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 768 \\ H_2 = 12, D_2 = 768 \\ R_2 = 4 \end{bmatrix} \times 2$
Stage3	14×14	Conv. Embed.	3×3, 384, stride 2		3×3, 1024, stride 2
	14×14	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 384 \\ H_3 = 6, D_3 = 384 \\ R_3 = 4 \end{bmatrix} \times 10$	$\begin{bmatrix} 3 \times 3, 384 \\ H_3 = 6, D_3 = 384 \\ R_3 = 4 \end{bmatrix} \times 16$	$\begin{bmatrix} 3 \times 3, 1024 \\ H_3 = 16, D_3 = 1024 \\ R_3 = 4 \end{bmatrix} \times 20$
Head	1×1	Linear	1000		
Params			19.98 M	31.54 M	276.7 M
FLOPs			4.53 G	7.13 G	60.86 G

Εικόνα 5.1: *Δομή CvT μοντέλου [13]*

.

Η έκδοση cvt-13-224x224 που χρησιμοποιείται στα πειράματα είναι η CvT-13 έκδοση της Εικόνας 5.1, η οποία διαθέτει 13 επίπεδα κωδικοποιητή μετασχηματιστή και δέχεται ως είσοδο εικόνες μεγέθους 224x224.

5.3 Πληροφορίες υλοποίησης

Στις παρακάτω υποενότητες παρουσιάζονται σημαντικές πληροφορίες για τα πειράματα, όπως είναι η μετρική αξιολόγησης που χρησιμοποιήθηκε, οι υπερπαράμετροι που αξιοποιήθηκαν για την εκπαίδευση των μοντέλων, αλλά και το μηχάνημα στο οποίο εκτελέστηκαν τα πειράματα. Τέλος παρατίθεται σύνδεσμος στο οποίο περιλαμβάνονται τα notebooks και ο κώδικας της παρούσας εργασίας.

5.3.1 Μετρική αξιολόγησης

Η μετρική αξιολόγησης που χρησιμοποιήθηκε σε όλα τα πειράματα είναι η ακρίβεια (accuracy), καθώς ήταν η μετρική που εμφανιζόταν σε όλες τις σχετικές έρευνες των Vision Transformer και των υβριδικών μοντέλων και αποτελούσε το μέτρο σύγκρισης τους. Η ακρίβεια αποτελεί μία από τις βασικότερες μετρικές στον τομέα των βαθιών νευρωνικών δικτύων και γενικότερα της τεχνητής νοημοσύνης και υπολογίζεται από το λόγο του αριθμού των δειγμάτων που ταξινομήθηκαν σωστά προς τον συνολικό αριθμό των δειγμάτων του συνόλου δοκιμής, όπως φαίνεται στην σχέση 5.1.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5.1}$$

Παράλληλα για την σύγκριση μεταξύ των μοντέλων λήφθηκε υπόψη τόσο ο χρόνος που χρειάστηκε για την εκπαίδευση τους όσο και ο αριθμός των παραμέτρων τους. Τέλος, μεγάλη σημασία δόθηκε στην απόδοση των μοντέλων στα μικρότερα σύνολα δεδομένων, καθώς αυτός ηταν ο κύριος στόχος της εργασίας.

5.3.2 Υπερπαράμετροι μοντέλων

Οι βασικοί υπερπαράμετροι των μοντέλων που χρησιμοποιήθηκαν στα πειράματα είναι οι εξής: μέγεθος εισόδου, βελτιστοποιητής (optimizer), αρχικός ρυθμός εκπαίδευσης (LR), τρόπος μεταβολής ρυθμού εκπαίδευσης, αναλογία επέκτασης MLP, αριθμός εποχών προθέρμανσης (χρησιμοποιείται μικρότερο LR για αυτές τις εποχές), αριθμός εποχών χαλάρωσης, weight decay, καθώς και το σύνολο στο οποίο είναι προεκπαιδευμένα (αν είναι) τα μοντέλα. Οι υπερπαράμετροι που χρησιμοποιήθηκαν, τόσο κατά την εκπαίδευση των μοντέλων από την αρχή (from scratch), όσο και κατά το fine-tuning των προεκπαιδευμένων μοντέλων παρουσιάζονται στους Πίνακες 5.6 και 5.7.

Μοντέλο	**Μέγεθος Εισόδου**	**Optimizer**	**Αρχικός LR**	**Μεταβολή LR**	**Αναλογία Επέκτασης στο MLP**
cct_14_7x2_224	224χ224	AdamW	$55 \cdot 10^{-5}$	Συνημιτονοειδής	3
deit_tiny_patch16_224	224χ224	AdamW	$5 \cdot 10^{-4}$	Συνημιτονοειδής	4
ceit_tiny_patch16_224	224χ224	AdamW	$5 \cdot 10^{-4}$	Συνημιτονοειδής	4
Localvit_small_mlp4_act3_r384	224χ224	AdamW	$5 \cdot 10^{-4}$	Συνημιτονοειδής	4
Conformer_small_patch16	224χ224	AdamW	$5 \cdot 10^{-4}$	Συνημιτονοειδής	4
cvt-13-224x224	224χ224	AdamW	$25 \cdot 10^{-5}$	Συνημιτονοειδής	4

Πίνακας 5.6: *Υπερπαράμετροι υβριδικών μοντέλων [27] [28] [29] [25] [26] [30]*

Μοντέλο	**Εποχές προθέρμανσης**	**Εποχές χαλάρωσης**	**Weight decay**	**Προεκπαιδευμένο σε**
cct_14_7x2_224	5	5	$6 \cdot 10^{-2}$	ImageNet-1k
deit_tiny_patch16_224	5	10	0.05	ImageNet-1k
ceit_tiny_patch16_224	5	10	0.05	ImageNet-1k
Localvit_small_mlp4_act3_r384	5	10	0.05	ImageNet-1k
Conformer_small_patch16	5	10	0.05	ImageNet-1k
cvt-13-224x224	5	10	0.1	ImageNet-1k

Πίνακας 5.7: *Υπερπαράμετροι υβριδικών μοντέλων [27] [28] [29] [25] [26] [30]*

5.3.3 Μηχάνημα

Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκε η δωρεάν έκδοση του περιβάλλοντος Google Colab, το οποίο περιέχει ως default CPU την Intel Xeon CPU με δύο vCPUs (virtual CPUs) και 13GB RAM. Όσον αφορά την GPU, περιέχει την Tesla T4 GPU με 16GB VRAM [133].

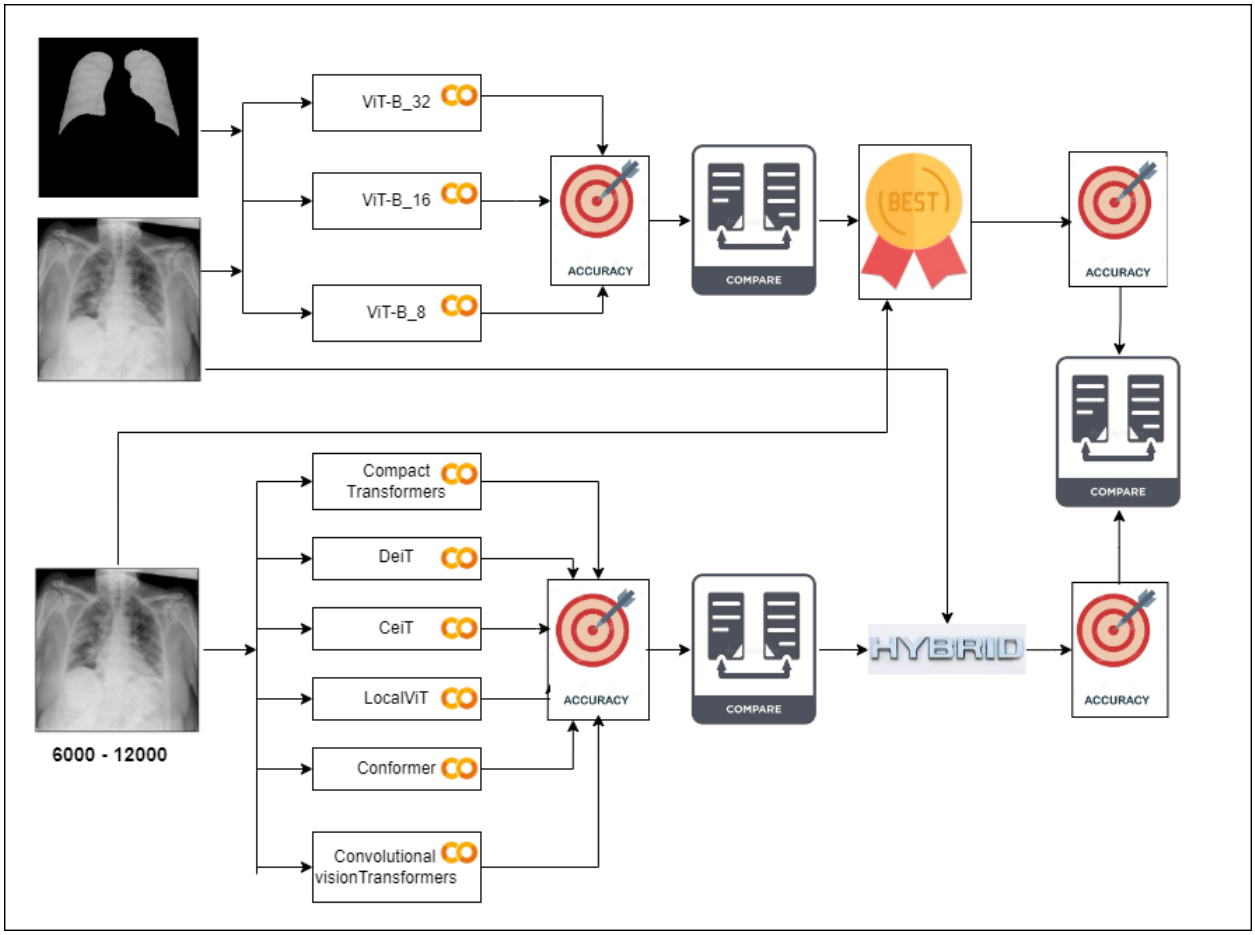
5.3.4 Κώδικας

Για κάθε μοντέλο χρησιμοποιήθηκε για την εκπαίδευση του και την δοκιμή του ένα notebook. Τα notebook αυτά αξιοποιούν τον κώδικα που παρουσιάζεται στις έρευνες που εισήγαγαν τα αντίστοιχα μοντέλα. Οι κώδικες αυτοί είναι τροποποιημένοι σε μικρό βαθμό, ώστε να μπορούν να λαμβάνουν ως είσοδο το σύνολο δεδομένων COVID-QU-Ex. Ακόμη η top-5 ακρίβεια μετατράπηκε σε top-3 ακρίβεια, όπου χρειάστηκε, καθώς το σύνολο δεδομένων διαθέτει μόνο τρεις κλάσεις. Τέλος, στα μοντέλα που το απαιτούσαν, δημιουργήθηκαν τα κατάλληλα configuration αρχεία για το νέο σύνολο δεδομένων, προκειμένου να γίνει ρύθμιση των παραμέτρων εκπαίδευσης και δοκιμής.

Ο κώδικας περιλαμβάνεται στον σύνδεσμο github repository.

5.4 Μεθοδολογία

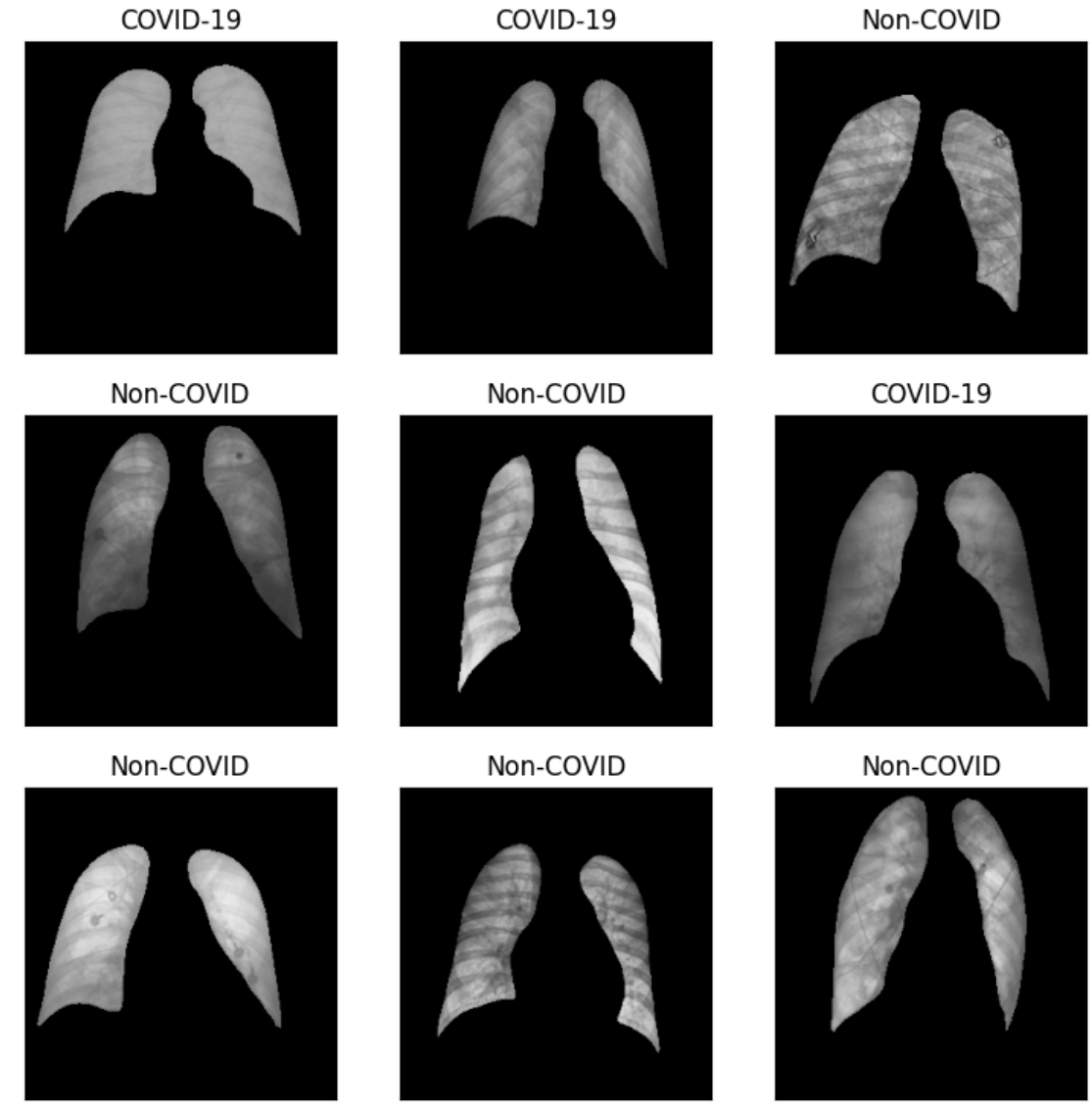
Η ενότητα παρουσιάζει μια σύνοψη της διαδικασίας που ακολουθήθηκε για την διεξαγωγή των πειραμάτων και την σύγκριση μεταξύ των μοντέλων που χρησιμοποιήθηκαν. Στο παρακάτω Σχήμα 5.5 παρουσιάζεται η γενική εικόνα των βημάτων που ακολουθήθηκαν.



Σχήμα 5.5: *Workflow πειραμάτων.*

Αρχικά γίνονται πειράματα πάνω στον παραδοσιακό Vision Transformer, δοκιμάζοντας διαφορετικό μέγεθος για τα patches, διαφορετικά μεγέθη batch αλλά και συνολικό αριθμό

βημάτων που θα γίνουν για το fine-tuning των προεκπαιδευμένων στο ImageNet-21k μοντέλων. Ως είσοδο τα μοντέλα σε κάποια πειράματα λαμβάνουν τις ακτινογραφίες θώρακα όπως παρέχονται στο σύνολο δεδομένων (raw), ενώ σε κάποια άλλα πειράματα λαμβάνουν τις ακτινογραφίες θώρακος, έχοντας εφαρμόσει σε αυτές τις μάσκες που παρέχονται από το σύνολο δεδομένων (masked). Παράδειγμα τέτοιων εικόνων στις οποίες έχει εφαρμοστεί η κατάλληλη μάσκα παρουσιάζονται στην Εικόνα 5.2.



Εικόνα 5.2: *Masked εικόνες.*

Σε κάποια πειράματα, προκειμένου να εισαχθούν επαγωγικές προκαταλήψεις στους Vision Transformers, στην αρχή των μοντέλων αντί να χωρίζεται η εικόνα απευθείας σε patches, γίνεται χρήση είτε μοντέλων ResNet είτε προσθήκη μερικών συνελικτικών επιπέδων.

Στην συνέχεια, αφού όπως αναλύθηκε και στα προηγούμενα κεφάλαια είναι δύσκολο και σπάνιο να βρεθούν μεγάλα ιατρικά σύνολα δεδομένων, γίνονται πειράματα χρησιμοποιώντας ένα μικρό μέρος των συνολικών εικόνων και νέα μοντέλα που αναλύθηκαν στο Κεφάλαιο 2, τα οποία σχεδιάστηκαν για την αντιμετώπιση αυτού ακριβώς του προβλήματος. Δηλα-

δή, προσπαθούν να παρουσιάσουν βελτιωμένη απόδοση για μικρότερα σύνολα δεδομένων διατηρώντας παράλληλα κάποια από τα πλεονεκτήματα των Vision Transformer για τα μεγάλα σύνολα δεδομένων. Τα υβριδικά μοντέλα που χρησιμοποιήθηκαν είναι οι Compact Transformers, DeiT, CeiT, LocalViT, Conformer και CvT. Έγιναν δοκιμές για τα 6 υβριδικά μοντέλα χρησιμοποιώντας αρχικά 6000 εικόνες και 12000 εικόνες, ενώ στο τέλος τα μοντέλα δοκιμάστηκαν σε ολόκληρο το σύνολο δεδομένων.

Έπειτα δοκιμάζεται ο Vision Transformer, που είχε τα καλύτερα αποτελέσματα στα αρχικά πειράματα, για ένα μικρό σύνολο των δεδομένων εκπαίδευσης, προκειμένου να υπάρξει σύγκριση με τα υβριδικά μοντέλα. Έχοντας τα αποτελέσματα από τα πειράματα έγινε σύγκριση ως προς την ακρίβεια και τον χρόνο εκπαίδευσης ανάμεσα στους Vision Transformers και στα υβριδικά μοντέλα, τόσο για μικρό σύνολο του σύνολο δεδομένων, όσο και για ολόκληρο το σύνολο δεδομένων. Στις περιπτώσεις που χρησιμοποιήθηκαν λιγότερες εικόνες από αυτές που διαθέτει το σύνολο δεδομένων, αυτές επιλέχθηκαν ομοιόμορφα από τις 3 κλάσεις.

Επιπροσθέτως, ο Vision Transformer με τα καλύτερα αποτελέσματα και τα υβριδικά μοντέλα που χρησιμοποιήθηκαν παραπάνω εκπαιδεύονται από την αρχή στο σύνολο δεδομένων COVID-QU-Ex για λίγες εποχές, προκειμένου να μελετηθεί η συμπεριφορά των μοντέλων που δεν έχουν προεκπαιδευτεί σε κάποιο σύνολο δεδομένων και η εκπαίδευση τους ξεκινάει από το μηδέν, αλλά και να υπάρχει η δυνατότητα σύγκρισης των μοντέλων μεταξύ τους.

Τέλος γίνονται κάποια δοκιμές για την κατηγοριοποίηση των εικόνων σε 2 κλάσεις από τις 3 που διαθέτει συνολικά το σύνολο δεδομένων. Για να συμβεί αυτό είτε ομαδοποιήθηκαν κάποιες κλάσεις είτε αφαιρέθηκαν κάποιες άλλες. Πειράματα έγιναν μόνο σε εικόνες χωρίς την εφαρμογή της μάσκας, καθώς όπως φάνηκε από τα πειράματα με τις 3 κλάσεις παρουσιάζουν εμφανώς καλύτερα αποτελέσματα από τις εικόνες με μάσκα.

Κεφάλαιο 6

Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα από τα πειράματα που πραγματοποιήθηκαν πάνω στο σύνολο δεδομένων COVID-QU-Ex που παρουσιάστηκε στο Κεφάλαιο 4, αλλά και η σύγκριση αυτών των μοντέλων με βάση την ακρίβεια (χρησιμοποιείται από όλα τα σχετικά μοντέλα στις αντίστοιχες μελέτες) και τον χρόνο εκπαίδευσης. Τα πειράματα πραγματοποιήθηκαν στο περιβάλλον του google colab, κάνοντας χρήση της GPU Tesla T4 με 16GB VRAM.

Τα αποτελέσματα των πειραμάτων με τους Vision Transformer για τις αρχικές εικόνες παρουσιάζονται στον Πίνακα 6.1, ενώ τα αντίστοιχα αποτελέσματα για τις masked εικόνες παρουσιάζονται στον Πίνακα 6.2. Για όλα τα παραπάνω πειράματα έγινε χρήση όλων των εικόνων του συνόλου δεδομένων. Σε κάποιους Vision Transformers, προκειμένου να εισαχθούν επαγωγικές προκαταλήψεις, στην αρχή αντί να χωρίζεται η εικόνα απευθείας σε patches, γίνεται χρήση είτε μοντέλων ResNet είτε προσθήκη μερικών συνελικτικών επιπέδων, όπως φαίνεται και στον Πίνακα 6.1. Η χρησιμοποίηση της Base έκδοσης των Vision Transformer αλλά και batch size ίσο με 32 για τα πειράματα με patch_size=8 ήταν αποτέλεσμα της έλλειψης υπολογιστικών πόρων.

Ωστόσο, από τους Πίνακες 6.1 και 6.2 παρατηρούμε ότι όσο μειώνεται το batch size και αυξάνεται ο αριθμός των steps που χρησιμοποιούνται για το fine-tuning, αυξάνεται παράλληλα και η απόδοση του μοντέλου. Ακόμη, παρατηρούμε ότι όσο μικραίνει το μέγεθος των patches, αυξάνεται η ακρίβεια και ο χρόνος εκπαίδευσης. Αυτό είναι λογικό, καθώς όσο μικρότερο είναι το patch size, τόσο περισσότερα token δημιουργούνται αυξάνοντας την δυνατότητα αναπαράστασης του δικτύου, αλλά και το κόστος επεξεργασίας τους. Επομένως το μοντέλο με την καλύτερη συμπεριφορά ήταν το ViT-B με patch size ίσο με 8. Η αύξηση των steps βοηθούσε σημαντικά την απόδοση μέχρι ένα σημείο κορεσμού περίπου στις 1200 εποχές.

Στην συνέχεια, παρουσιάζονται στον Πίνακα 6.3 τα αποτελέσματα των πειραμάτων που εφαρμόστηκαν στα μοντέλα Compact Transformers, DeiT, CeiT, LocalViT, Conformer και CvT με χρήση 6000 εικόνων για την εκπαίδευσης τους. Στις 6000 εικόνες του συνολου εκπαίδευσης, αλλά και σε κάθε περίπτωση που χρησιμοποιείται υποσύνολο του συνόλου δεδομένων για την εκπαίδευση, περιλαμβάνεται ίδιος αριθμός εικόνων και από τις τρεις κλάσεις. Τα αντίστοιχα αποτελέσματα για 12000 εικόνες παρουσιάζονται στον Πίνακα 6.4. Προκειμένου να έχουμε ολοκληρωμένη εικόνα για την απόδοση των συγκεκριμένων μοντέλων σε

Είσοδος Vision Transformer	Patch size	Batch size	Steps	Accuracy
Patches	32	512	42	0.7997
Patches	32	512	100	0.9097
Patches	32	32	700	0.9428
Patches	32	32	800	0.9435
Patches	32	32	900	0.9467
Patches	32	32	1200	0.9493
Patches	32	16	1600	0.945
Patches	16	256	100	0.9179
Patches	8	32	800	0.9621
Patches	8	32	1200	**0.9658**
7 convolutional layers	32	512	100	0.9023
ResNet-50	16	128	200	0.9338
ResNet-50	16	128	400	0.954
ResNet-50	16	32	700	0.9431

Πίνακας 6.1: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-21k Vision Transformer στις raw εικόνες*

Patch size	Batch size	Steps	Accuracy
32	256	100	0.7946
32	512	100	0.8121
16	256	100	0.8472
8	32	200	0.8548
8	32	600	0.9080
8	32	800	**0.9157**

Πίνακας 6.2: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-21k Vision Transformer στις masked εικόνες*

ολόκληρο το σύνολο δεδομένων και να μπορούμε να συγκρίνουμε με τους παραδοσιακούς Vision Transformer, εκτελούμε τα αντίστοιχα πειράματα σε όλες τις εικόνες του συνόλου εκπαίδευσης. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.5. Για το CCT μοντέλο χρησιμοποιήθηκαν 15000 εικόνες για το fine-tuning λόγω έλλειψης υπολογιστικών πόρων.

Έπειτα δοκιμάστηκε ο Vision Transformer που είχε τα καλύτερα αποτελέσματα για ένα μικρό σύνολο των δεδομένων εκπαίδευσης, προκειμένου να υπάρξει σύγκριση με τα υβριδικά μοντέλα. Τα αποτελέσματα φαίνονται στον Πίνακα 6.6.

Από τα αποτελέσματα που προέκυψαν για τα προεκπαιδευμένα μοντέλα, παρατηρούμε ότι τα υβριδικά μοντέλα διαθέτουν λιγότερες παραμέτρους, προσφέροντας μεγαλύτερη αντίσταση έναντι της υπερπροσαρμογής, ενώ η πλειοψηφία χρειάστηκε λιγότερο χρόνο εκπαίδευσης από τα ViTs. Ακόμη, για 6000 εικόνες εκπαίδευσης, υπάρχουν δύο υβριδικά μοντέλα (CCT και CvT) με μεγαλύτερη ακρίβεια από το ViT-B_8. Τα συγκεκριμένα αυτά μοντέλα συνεχίζουν να έχουν καλύτερες επιδόσεις και όταν χρησιμοποιείται μεγαλύτερο κομμάτι του συνόλου εκπαίδευσης, συμπεριλαμβανομένης και της περίπτωσης που χρησι-

Μοντέλο	Pre-trained	Batch size	Epochs	Παράμετροι	Χρόνος training	Accuracy
cct_14_7x2_224	ImageNet-1k	128	30	21.983.428	**22:05**	**0.948**
deit_tiny_patch16_224	ImageNet-1k	32	30	**5.524.995**	22:48	0.90
ceit_tiny_patch16_224	ImageNet-1k	32	30	6.164.483	26:33	0.893
Localvit_small_mlp4_act3_r384	ImageNet-1k	32	30	22.049.331	36:31	0.89
Conformer_small_patch16	ImageNet-1k	32	30	36.267.654	58:54	0.9177
cvt-13-224x224	ImageNet-1k	32	30	20.000.000	29:32	0.943

Πίνακας 6.3: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-1k Υβριδικών CNN - Vision Transformer μοντέλων σε 6000 εικόνες*

Μοντέλο	Pre-trained	Batch size	Epochs	Παράμετροι	Χρόνος training	Accuracy
cct_14_7x2_224	ImageNet-1k	128	30	21.983.428	**41:10**	0.95
deit_tiny_patch16_224	ImageNet-1k	32	30	**5.524.995**	47:24	0.9342
ceit_tiny_patch16_224	ImageNet-1k	32	30	6.164.483	53:16	0.894
Localvit_small_mlp4_act3_r384	ImageNet-1k	32	30	22.049.331	1:12:59	0.9356
Conformer_small_patch16	ImageNet-1k	32	30	36.267.654	1:43:43	0.9434
cvt-13-224x224	ImageNet-1k	32	30	20.000.000	1:04:28	**0.9605**

Πίνακας 6.4: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-1k Υβριδικών CNN - Vision Transformer μοντέλων σε 12000 εικόνες*

Μοντέλο	Training data	Batch size	Epochs	Παράμετροι	Χρόνος training	Accuracy
cct_14_7x2_224	15000	128	30	21.983.428	**46:30**	0.9663
deit_tiny_patch16_224	21715	32	30	**5.524.995**	1:18:30	0.9492
ceit_tiny_patch16_224	21715	32	30	6.164.483	1:33:57	0.94
Localvit_small_mlp4_act3_r384	21715	32	30	22.049.331	2:02:03	0.9417
Conformer_small_patch16	21715	32	30	36.267.654	2:59:07	0.9487
cvt-13-224x224	21715	32	30	20.000.000	1:36:57	**0.9694**

Πίνακας 6.5: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-1k Υβριδικών CNN - Vision Transformer μοντέλων σε όλες τις εικόνες*

μοποιείται όλο το σύνολο δεδομένων. Η μόνη εξαίρεση αποτελεί το cct_14_7x2_224 για τις 12000 εικόνες, το οποίο παρουσιάζει ελαφρώς μικρότερη ακρίβεια από τον Vision Transformer.

Ύστερα, προσπαθούμε να εκπαιδεύσουμε από την αρχή (from scratch) όλα τα προαναφερθέντα μοντέλα για λίγες εποχές (30) σε ένα μικρό υποσύνολο του συνόλου εκπαίδευσης, ώστε να δούμε την συμπεριφορά τους. Τα αποτελέσματα που προέκυψαν φαίνονται στον Πίνακα 6.7. Και στην εκπαιδευση των μοντέλων από την αρχή, παρόλο που είναι για λίγες

Δεδομένα εκπαίδευσης	Patch size	Batch size	Steps	Παράμετροι	Χρόνος training	Accuracy
6000	8	32	200	86.000.000	**29:33**	0.931
12000	8	32	400	86.000.000	50:47	**0.9536**

Πίνακας 6.6: *Αποτελέσματα προεκπαιδευμένων στο ImageNet-21k Vision Transformer για μικρότερο πλήθος δεδομένων εκπαίδευσης*

Μοντέλο	Batch size	Epochs για Hybrid/ steps για ViT	Παράμετροι	Χρόνος training	Accuracy
ViT-B_8	32	200	86.000.000	26:44	0.6067
cct_14_7x2_224	128	30	21.983.428	**22:02**	**0.821**
deit_tiny_patch16_224	32	30	5.524.995	22:28	0.53
ceit_tiny_patch16_224	32	30	6.164.483	27:23	0.69
Localvit_small_mlp4_ act3_r384	32	30	22.049.331	34:11	0.66
Conformer_small_patch16	32	30	36.267.654	56:11	0.7436
cvt-13-224x224	32	30	20.000.000	26:16	0.6937

Πίνακας 6.7: *Αποτελέσματα όλων των μοντέλων εκπαιδευμένων from scratch σε 6000 εικόνες εκπαίδευσης*

εποχές, παρατηρούμε ότι τα αποτελέσματα των πέντε από τα έξι μοντέλα για τις 6000 εικόνες είναι βελτιωμένα έναντι του ViT, παρουσιάζοντας καλύτερη ακρίβεια. Αυτό μπορεί να σημαίνει είτε ότι μετά από αρκετές εποχές θα συνεχίσουν να παρουσιάζουν καλύτερα αποτελέσματα έιτε ότι συγκλίνουν γρηγορότερα. Επιπλέον, τρία από τα έξι υβριδικά μοντέλα παρουσιάζουν μικρότερους χρόνους εκπαίδευσης από το ViT-B_8.

Τέλος έγιναν κάποια δοκιμές για την κατηγοριοποίηση των εικόνων σε δύο κλάσεις από τις κλάσεις που διαθέτει συνολικά το σύνολο δεδομένων. Τα αποτελέσματα των πειραμάτων παρουσιάζονται στον Πίνακα 6.8. Στις στήλες κλάση 1 και κλάση 2 παρουσιάζονται οι κλάσεις που χρησιμοποιήθηκαν για την ταξινόμηση, οι οποίες προέκυψαν είτε από συγχώνευση δύο κλάσεων είτε από την διαγραφή μίας από τις τρεις κλάσεις του συνόλου δεδομένων.

Τα αποτελέσματα από τα πειράματα με τις δύο κλάσεις δείχνουν ότι οι Vision Transformers μπορούν να επιτύχουν πολύ υψηλές τιμές ακρίβειας. Τα καλύτερα αποτελέσματα παρουσιάζονται όταν γίνεται χρήση των δεδομένων των κλάσεων COVID-19 και Non-COVID παράλληλα με τον συνδυασμό των κλάσεων COVID-19 και Normal. Δηλαδή, υπάρχει η δυνατότητα εντοπισμού ενός ασθενή με COVID-19 από έναν ασθενή με άλλη ασθένεια που επηρεάζει τους πνεύμονες ή στην δεύτερη περίπτωση από ένα υγιή άνθρωπο σε ποσοστά μεγαλύτερα του 99%. Αυτή η τεράστια επιτυχία εντοπισμού είναι καίρια για μία ασθένεια, η οποία όπως έχει αναλυθεί είναι πολυ μολυσματική και επικίνδυνη και σε κάποιες περιπτώσεις πρέπει να εντοπίζεται και να αντιμετωπίζεται γρήγορα και με επιτυχία.

κλάση 1	κλάση 2	Patch size	Batch size	Steps	Accuracy
COVID-19	Normal + Non-COVID	32	512	100	0.9582
COVID-19	Normal + Non-COVID	8	32	800	0.991
COVID-19	Normal	32	512	100	0.9619
COVID-19	Normal	32	32	400	0.9774
COVID-19	Normal	32	32	600	0.976
COVID-19	Normal	8	32	800	**0.992**
COVID-19	Non-COVID	32	512	100	0.9668
COVID-19	Non-COVID	8	32	800	0.9914

Πίνακας 6.8: *Αποτελέσματα fine-tuning προεκπαιδευμένων στο dataset ImageNet-21k Vision Transformer χρησιμοποιώντας 2 κλάσεις*

Μέρος III

Επίλογος

Κεφάλαιο 7

Συμπεράσματα και Μελλοντικές επεκτάσεις

Στο κεφάλαιο αυτό παρουσιάζονται συγκεντρωτικά τα συμπεράσματα που προέκυψαν από τα πειράματα του Κεφαλαίου 6, καθώς και ιδέες για την μελλοντική προέκταση της παρούσας εργασίας.

7.1 Συμπεράσματα

Στο προηγούμενο κεφάλαιο αρχικά παρουσιάστηκαν τα αποτελέσματα που είχαν οι διάφορες εκδόσεις των Vision Transformers όταν γίνονταν fine-tune σε ολόκληρο το σύνολο δεδομένων COVID-QU-Ex Dataset. Χρησιμοποιήθηκαν διάφορες παραλλαγές του παραδοσιακού Vision Transformer με κυριότερες την Base έκδοση του για διαφορετικό μέγεθος patch, αλλά και χρήση παραλλαγών του ResNet πριν την τροφοδοσία των εικόνων στον κωδικοποιητή του μετασχηματιστή. Από τα πρώτα αυτά πειράματα φαίνεται ότι όσο μειώνεται το batch size και αυξάνεται ο αριθμός των steps που χρησιμοποιούνται για το fine-tuning, αυξάνεται και η απόδοση του μοντέλου, ενώ παράλληλα όσο μικραίνει το μέγεθος των patches, αυξάνεται η ακρίβεια και ο χρόνος εκπαίδευσης. Επομένως το μοντέλο με την καλύτερη συμπεριφορά ήταν το ViT-B με patch size ίσο με 8, batch size ίσο με 32 και 1200 βήματα, παρουσιάζοντας ακρίβεια 96.58%. Τα μοντέλα που χρησιμοποίησαν το ResNet-50 πριν την είσοδο στον κωδικοποιητή του Vision Transformer παρουσίαζαν ικανοποιητικές αποδόσεις με μέγιστη ακρίβεια 95.4%, αλλά δεν κατάφεραν να υπερβούν τις επιδόσεις των παραδοσιακών ViT.

Όμως το γεγονός ότι οι Vision Transformers, λόγω του attention layer, συλλαμβάνουν μακρινές καθολικές εξαρτήσεις και δεν διαθέτουν επαγωγικές προκαταλήψεις, προκειμένου να μπορούν να γενικευθούν οταν εκπαιδεύονται σε μικρό σύνολο δεδομένων, σε συνδυασμό με το γεγονός ότι στις περισσότερες ιατρικές εφαρμογές είναι δύσκολη η εύρεση μεγάλων συνόλων δεδομένων μας οδήγησε στην δοκιμή υβριδικών μοντέλων που προσπαθούν να αξιοποιήσουν τα καλύτερα χαρακτηριστικά των CNNs και των Vision Transformers. Για αυτό το λόγο δοκιμάστηκαν συγκεκριμένες εκδόσεις των μοντέλων Compact Transformers, DeiT, CeiT, Conformer, LocalViT και Convolutional vision Transformers τόσο σε υποσύνολο του συνόλου εκπαίδευσης όσο και σε ολόκληρο το σύνολο εκπαίδευσης. Ακόμη ο Vision Transformer με τα καλύτερα αποτελέσματα στα αρχικά πειράματα δοκιμάστηκε και για μικρότερο αριθμό εικόνων εκπαίδευσης, προκειμένου να μπορεί να γίνει η σύγκριση μεταξύ των μοντέλων.

Από τα αποτελέσματα που προέκυψαν παρατηρούμε ότι τα υβριδικά μοντέλα διαθέτουν λιγότερες παραμέτρους και μικρότερο μέγεθος συμβάλοντας στην αντιμετώπιση του προβλήματος της μεγάλης απαίτησης σε μνήμη των ViT. Στα πειράματα με τις 6000 εικόνες εκπαίδευσης υπάρχουν δύο υβριδικά μοντέλα (cct_14_7x2_224 και cvt-13-224x224) με καλύτερη ακρίβεια από το ViT-B_8 (94.8% και 94.3% αντίστοιχα έναντι 93.1%). Στην περίπτωση αυτή τα τέσσερα από τα έξι υβριδικά μοντέλα χρειάστηκαν λιγότερο χρόνο εκπαίδευσης από το ViT. Αντίστοιχα, για 12000 εικόνες το μοντέλο cvt-13-224x224 εξακολουθούσε να παρουσιάζει καλύτερα αποτελέσματα (96.05%), ενώ το cct_14_7x2_224 (95%) είχε μια μικρή πτώση σε σχέση με τα αντίστοιχα αποτελέσματα του ViT (95.36%). Ταυτόχρονα, δύο από τα έξι υβριδικά μοντέλα παρουσίασαν μικρότερο χρόνο εκπαίδευσης. Όσον αφορά ολόκληρο το σύνολο δεδομένων, τα δύο μοντέλα που κυριαρχούσαν στα πειράματα με τις 6000 εικόνες, κυριαρχούν και σε αυτή την περίπτωση με ακρίβεια 96.63% και 96.94% αντίστοιχα έναντι 96.58% του ViT. Η ακρίβεια του δεύτερου μοντέλου για όλο το σύνολο δεδομένων είναι και το καλύτερο που παρουσιάζεται στα πειράματα με τις τρεις κλάσεις. Ωστόσο, και τα υπόλοιπα μοντέλα παρουσιάζουν παραπλήσια αποτελέσματα με το ViT-B_8 και θα μπορούσαν να προτιμηθούν λόγω του πλεονεκτήματος των παραμέτρων, του μεγέθους τους και του μικρότερου χρόνου εκπαίδευσης.

Έπειτα, προσπαθούμε να εκπαιδεύσουμε από την αρχή όλα τα προαναφερθέντα μοντέλα για λίγες εποχές (30) σε ένα μικρό υποσύνολο του συνόλου εκπαίδευσης, ώστε να δούμε την συμπεριφορά τους. Και στην εκπαιδευση των μοντέλων από την αρχή, παρόλο που είναι για λίγες εποχές, παρατηρούμε ότι τα αποτελέσματα για την πλειοψηφία των υβριδικών μοντέλων για 6000 εικόνες είναι βελτιωμένα έναντι του ViT, παρουσιάζοντας καλύτερη ακρίβεια που στην καλύτερη περίπτωση πλησιάζει το 82.1%, σε αντίθεση με το 60.67% του ViT. Όσον αφορά τον χρόνο εκπαίδευσης, τα τρία από τα έξι μοντέλα παρουσιάζουν καλύτερα αποτελέσματα από τον Vision Transformer.

Στο τέλος γίνεται προσπάθεια για την κατηγοριοποίηση των εικόνων σε δύο κλάσεις από τις τρεις που διαθέτει συνολικά το σύνολο δεδομένων. Τα βέλτιστα αποτελέσματα προκύπτουν όταν για την εκπαίδευση και τη δοκιμή χρησιμοποιούνται μόνο εικόνες που ανήκουν στις κλάσεις COVID-19 και Normal ή μόνο εικόνες που ανήκουν στις κλάσεις COVID-19 και Non-COVID, πετυχαίνοντας ακρίβεια ίση με 99.2% και 99.14% αντίστοιχα.

Να σημειωθεί ότι για την προεκπαίδευση όλων των παραπάνω υβριδικών μοντέλων χρησιμοποιήθηκε το σύνολο δεδομένων ImageNet-1k, το οποίο περιέχει μόνο 1000 κλάσεις και είναι υποσύνολο του μεγάλου συνόλου ImageNet-21k, στο οποίο είναι προεκπαιδευμένες όλες οι εκδόσεις του παραδοσιακού Vision Transformer που χρησιμοποιήθηκαν. Το γεγονός ότι τα υβριδικά μοντέλα παρουσίασαν παρόμοια ή καλύτερα αποτελέσματα από τους Vision Transformer, έχοντας προεκπαιδευτεί σε μικρότερα σύνολα δεδομένων δείχνει την συμβολή αυτών των μοντέλων στην αντιμετώπιση ενός βασικού προβλήματος των ViT και πιο συγκεκριμένα την ανάγκη ύπαρξης μεγάλων σύνολων δεδομένων για την εκπαίδευση τους.

7.2 Μελλοντικές επεκτάσεις

Μερικές βασικές μελλοντικές επεκτάσεις θα ήταν η δοκιμή των παραπάνω μοντέλων και σε άλλα ιατρικά σύνολα δεδομένων ή και σε διαφορετικό τύπο εικόνων, προκειμένου να

φανεί αν θα παρουσιάσουν παρόμοια αποτελέσματα. Ακόμη μπορούν να γίνουν δοκιμές σε περισσότερα υβριδικά μοντέλα με διαφορετικούς συνδυασμούς χαρακτηριστικών CNN και Vision Transformer, προκειμένου να επιτευχθεί καλύτερη ακρίβεια, αλλά και να βελτιωθούν άλλες παράμετροι των δικτύων, όπως ο χρόνος εκπαίδευσης, ο αριθμός των παραμέτρων και η ανάγκη για ακόμη μικρότερο σύνολο δεδομένων. Επιπροσθέτως, μέσω της μελέτης που έγινε στα παραπάνω υβριδικά μοντέλα, μπορούν να συνδυαστούν τα βέλτιστα χαρακτηριστικά του κάθε μοντέλου, ώστε να παραχθεί ένα καινούργιο δίκτυο που θα παρουσιάζει ακόμη καλύτερα αποτελέσματα.

Μια ακόμη επιλογή θα ήταν η εφαρμογή των παραπάνω υβριδικών δικτύων ως backbone σε δίκτυα όπως το ViT-NeT [134], τα οποία εκτός από την σωστή κατηγοριοποίηση των εικόνων έχουν ως στόχο και την ικανότητα επεξηγησιμότητας των μοντέλων. Τέλος, αφού οπως αναλύθηκε στο Κεφάλαιο 2, κάποια από τα υβριδικά μοντέλα μπορούν να χρησιμοποιηθούν και για τμηματοποίηση περιπτώσεων, υπάρχει η δυνατότητα να γίνουν τα κατάλληλα πειράματα και να υπάρξει σύγκριση με τα αντίστοιχα state-of-the-art μοντέλα του συγκεκριμένου πεδίου.

Παραρτήματα

Παράρτημα Α΄

Θεωρητικές έννοιες - Αναλυτικότερα

Α΄.1 Attention Mechanism

Οι μηχανισμοί προσοχής (attention mechanisms) στην βαθιά μάθηση βοηθούν τα μοντέλα να επικεντρώνονται σε μέρος της πληροφορίας εισόδου που θεωρείται χρησιμότερη. Με αυτό τον τρόπο επιδιώκουν να επιλύσουν το πρόβλημα του μεγάλου μεγέθους και πολυπλοκότητας των δεδομένων εισόδου που χρησιμοποιούνται [135]. Τα στρώματα προσοχής (attention layers) είναι η βασική δομική μονάδα στην οποία στηρίζονται οι Vision Transformers και γενικότερα οι Transformers. Οι Transformers στον τομέα της επεξεργασίας φυσικής γλώσσας επιδιώκουν τον εντοπισμό μακρυνότερων εξαρτήσεων, αντικαθιστώντας σε μεγάλο βαθμό τα RNNs και LSTMs, τα οποία δεν διαθέτουν την παραπάνω ικανότητα σε ικανοποιητικό. Από την άλλη πλευρά, τα επίπεδα αυτά επιτρέπουν στους Vision Transformers να συλλαμβάνουν μακρινές καθολικές εξαρτήσεις μεταξύ των patches, στα οποία έχει χωριστεί η εικόνα που δίνεται ως είσοδος στο μοντέλο. Διάφορες παραλλαγές στρωμάτων προσοχής είναι οι Generalized Attention, Self-Attention, Additive Attention, Scaled Dot-Product, Multi-Head Attention. Παρακάτω παρουσιάζεται ο τροπος λειτουργίας των παραπάνω δομών, δίνοντας ιδιαίτερη έμφαση στα Scaled Dot-Product Attention και Multi-Head Attention [2].

Α΄.1.1 Generalized Attention

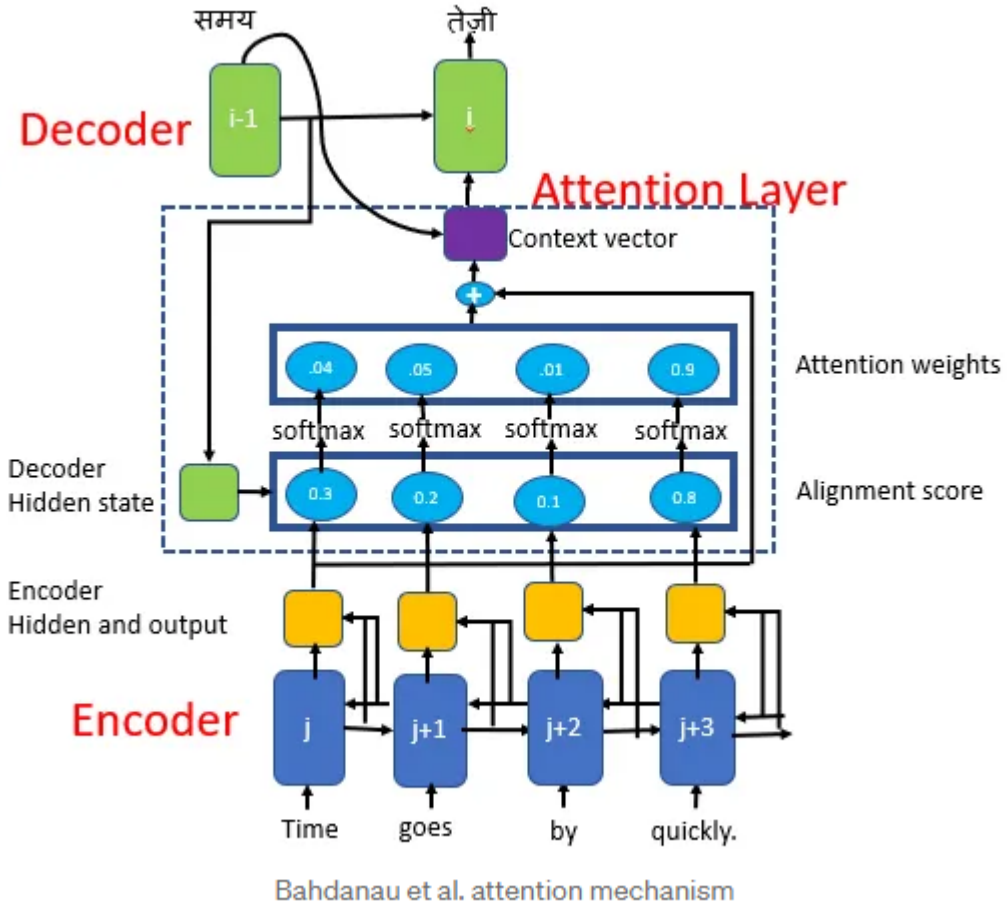
Το είδος προσοχής Generalized Attention λαμβάνει ως είσοδο ακολουθίες λέξεων ή εικόνες και συγκρίνει την ακολουθία εισόδου με την ακολουθία εξόδου. Πιο συγκεκριμένα, με την σύγκριση της εισόδου του κωδικοποιητή με την έξοδου του αποκωδικοποιητή που γίνεται σε κάθε επανάληψη προκύπτουν κάποιες βαθμολογίες που χρησιμοποιούνται από το μοντέλο για να επιλέξει τα κομμάτια της εισόδου που θα δώσει περισσότερη προσοχή.

Α΄.1.2 Self-Attention

Ο μηχανισμός Self-Attention ή Intra-Attention συλλέγει κομμάτια της εισόδου απο διαφορετικές θέσεις και υπολογίζει μια αρχική σύνθεση της ακολουθίας εξόδου, χωρίς να λαμβάνει υπόψη την ακολουθία εξόδου. Χρησιμοποιείται σε μοντέλα όπως το BERT και γενικότερα στα μοντέλα που βασίζονται στους Transformers, παρέχοντας τους την δυνατότητα να συλλαμβάνουν καθολικές εξαρτήσεις μεταξύ εισόδου και εξόδου.

Α΄.1.3 Additive Attention

Το Additive Attention ή Bahdanau Attention χρησιμοποιεί alignment scores που υπολογίζονται σε διαφορετικές θέσεις του δικτύου, προκειμένου να ευθυγραμμίσει τις ακολουθίες εισόδου με τις ακολουθίες εξόδου, βοηθώντας στο να δοθεί προσοχή στις πιο σχετικές πληροφορίες. Η είσοδος συσχετίζεται με τον στόχο ή την ακολουθία εξόδου, αλλά όχι σε ακριβή βαθμό, λαμβάνοντας υπόψη όλες τις κρυφές καταστάσεις του κωδικοποιητή και του αποκωδικοποιητή, προκειμένου να παραχθούν τα διανύσματα συμφραζομένων (context vectors). Το μοντέλο προβλέπει την λεξη στόχο με βάση τα διανύσματα που δημιουργήθηκαν και σχετίζονται με την θέση της πηγής σε συνδυασμό με τις προηγούμενες λέξεις που έχουν παραχθεί. Η αρχιτεκτονική του παραπάνω μηχανισμού φαίνεται στο Σχήμα Α΄.1. Οι σχέσεις από τις οποίες υπολογίζονται τα alignment scores, τα attention weights, αλλά και οι πίνακες με τους οποίου γίνεται η πρόβλεψη παρουσιάζονται στις εξισώσεις Α΄.1, Α΄.2 και Α΄.3 αντίστοιχα. Η πρόβλεψη της λέξης στόχου λαμβάνει υπόψη το δίανυσμα συμφραζομένων, την έξοδο του αποκωδικοποιητή στο προηγούμενο χρονικό βήμα (y_{i-1}), τις προηγούμενες κρυφές καταστάσεις του αποκωδικοποιητή (s_{i-1}) και παρουσιάζεται στην σχέση Α΄.4.



Σχήμα Α΄.1: *Αρχιτεκτονική Bahdanau Attention για μετάφραση προτάσεων [16].*

$$Alignment\_score_{ij} = a(s_{i-1}, h_j) \tag{Α΄.1}$$

$$Attention\_weight_{ij} = \frac{exp(Alignment\_score_{ij})}{\sum_{k=1}^{T_x} exp(Alignment\_score_{ik})} \tag{Α΄.2}$$

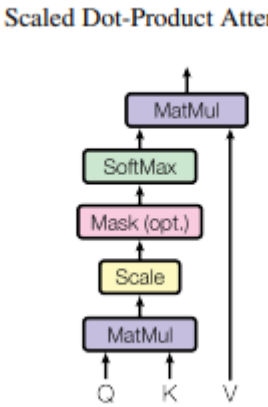
$$Context_i = \sum_{j=1}^{T_x} Attention\_weight_{ij} h_j \tag{Α΄.3}$$

$$s_i = f(s_{i-1}, c_i, y_{i-1}) \tag{Α΄.4}$$

Α΄.1.4 Scaled Dot-Product Attention

Η είσοδος αποτελείται από queries (Q) και keys (K) διαστάσεων d_k και από values (V) διαστάσεων d_u. Το αποτέλεσμα του Scaled Dot-Product Attention προκύπτει πολλαπλασιάζοντας τα βάρη, που δημιουργούνται από το εσωτερικό γινόμενο του κάθε query με όλα τα keys διαιρεμένα με $\sqrt{d_k}$, με τα values. Ο υπολογισμός που αναλύθηκε παραπάνω φαίνεται και στην σχέση Α΄.5. Η αρχιτεκτονική της δομής Scaled Dot-Product Attention φαίνεται στο Σχήμα Α΄.2. Ο υπολογισμός του Dot Product Attention είναι ακριβώς ίδιος με παραπάνω, χωρίς την διαίρεση με τον όρο $\sqrt{d_k}$. Ο μηχανισμός αυτός είναι πολύ πιο γρήγορος και αποτελεσματικός όσον αφορά τον χώρο που καταλαμβάνει από τον μηχανισμο Additive Attention, καθώς μπορεί να χρησιμοποιηθεί βελτιστοποιημένος κώδικας πολλαπλασιασμού πινάκων.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{Α΄.5}$$



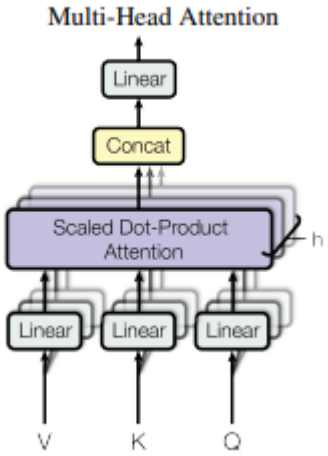
Σχήμα Α΄.2: *Αρχιτεκτονική δομής Scaled Dot-Product Attention [2].*

Α΄.1.5 Multi-Head Attention

Έστω k ο αριθμός των heads. Για κάθε head i διαθέτουμε τους πίνακες W_i^Q, W_i^K και W_i^V, οι οποίοι χρησιμοποιούνται προκειμένου να δημιουργηθούν οι προβολές των πινάκων Q, K, V. Αντί να εφαρμοστεί η συνάρτηση attention στους ήδη υπαρχοντες πίνακες Q, K, V, εφαρμόζεται k φορές στους πίνακες που δημιουργούνται για κάθε head, τροφοδοτώντας τους στην συνάρτηση Scaled Dot-Product Attention, όπως φαίνεται στην σχέση Α΄.6. Οι έξοδοι της συνάρτησης για κάθε head συνενωνονται μεταξύ τους, όπως απεικονίζεται και στην σχέση Α΄.7 και προκύπτει το αποτέλεσμα του Multi-Head Attention. Με αυτό τον τρόπο το μοντέλο επιτρέπει στο κάθε head να επικεντρώνεται σε διαφορετικές θέσεις, δίνει την δυνατότητα να γίνει χρήση πολλών διαφορετικών υποχώρων αναπαράστασης και συνεπώς βελτιώνεται η απόδοση του μοντέλου. Η αρχιτεκτονική της δομής Multi-Head Attention φαίνεται στο Σχήμα Α΄.2.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{Α΄.6}$$

$$MultiHead(Q, K, V) = Concatenate(head_1, ..., head_k)W^o \tag{Α΄.7}$$



Σχήμα Α΄.3: *Αρχιτεκτονική δομής Multi-Head Attention [2].*

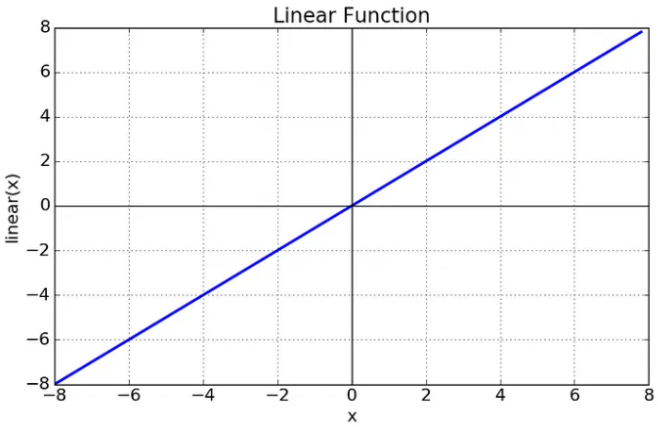
Α΄.1.6 Πλεονεκτήματα Self-Attention

Σε αυτή την υποενότητα θα αναλυθούν τα πλεονεκτήματα του self-attention layer έναντι των recurrent και συνελικτικών στρωμάτων. Τα 3 πλεονεκτήματα είναι η συνολική υπολογιστική πολυπλοκότητα ανά επίπεδο, η υπολογιστική ποσότητα που μπορεί να παραλληλοποιηθεί, η οποία εξαρτάται από τον αριθμό των διαδοχικών ενεργειών που απαιτούνται, καθώς και το μέγεθος των μονοπατιών μεταξύ μακρινών εξαρτήσεων. Τα self-attention layers συνδέουν όλες τις θέσεις με σταθερό αριθμό απαιτούμενων διαδοχικών ενεργειών, ενώ το recurrent layer απαιτεί O(n) αριθμό διαδοχικών ενεργειών. Ακόμη, όσον αφορά την υπολογιστική πολυπλοκότητα, τα επίπεδα self-attention είναι πιο γρήγορα από τα επίπεδα recurrent, όταν το μέγεθος της ακολουθίας n είναι μικρότερο από τις διαστάσεις αναπαράστασης d που απο-

τελεί την πιο συχνή περίπτωση. Όσον αφορά τα συνελικτικά στρώματα με μέγεθος πυρήνα k μικρότερο του n, απαιτούνται O(n/k) συνελικτικά στρώματα στην περίπτωση των συνεχόμενων πυρήνων και O($log_k(n)$) στην περίπτωση των διασταλμένων πυρήνων, μεγαλώνοντας το μέγεθος των μονοπατιών μεταξύ δυο θέσεων. Οι διαχωρίσιμες συνελίξεις μειώνουν την πολυπλοκότητα σε O($k \cdot n \cdot d + n \cdot d^2$), όμως ακόμη και στην περίπτωση του $k = n$ ισοδυναμεί με τον συνδυασμό ενός self-attention επιπέδου με ένα point-wise feed-forward layer. Τέλος, ένα ακόμη πλεονέκτημα του self-attention είναι ότι τα μοντέλα που το χρησιμοποιούν παράγουν πιο επεξηγήσιμα αποτελέσματα [2].

Α΄.2 Γραμμική συνάρτηση ενεργοποίησης

Η γραμμική συνάρτηση ενεργοποίησης (Linear Activation Fuction) ακολουθεί την μορφή που παρουσιάζεται στο Σχήμα Α΄.4. Η παράγωγος της γραμμικής συνάρτησης ενεργοποίησης είναι σταθερή και δεν λαμβάνει υπόψη την τιμή εισόδου της. Αυτό σημαίνει ότι κάθε φορά στην διαδικασία του backpropagation οι παράγωγοι θα ήταν ίδιες, το οποίο θα δημιουργούσε πρόβλημα στην ανανέωση των βαρών και συνεπώς στην εκπαίδευση του μοντέλου [18].



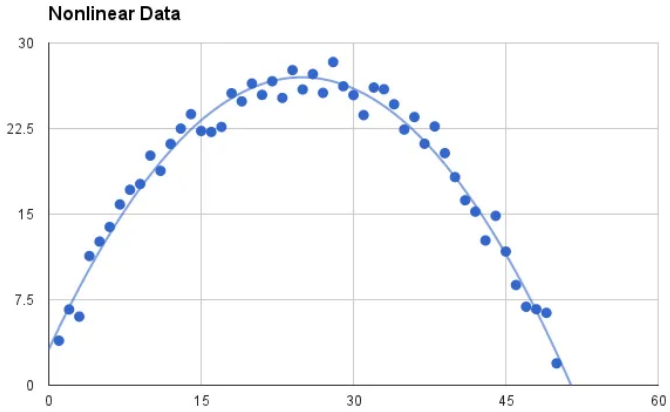
Σχήμα Α΄.4: *Γραμμική συνάρτηση ενεργοποίησης [17].*

Α΄.3 Μη γραμμικές συναρτήσεις ενεργοποίησης

Οι μη γραμμικές συναρτήσεις ενεργοποίησης αποφασίζουν πότε ένας νευρώνας ενεργοποιείται, με σκοπό να προσθέσουν μη γραμμικότητα στο δίκτυο, καθιστώντας εφικτή την εκπαίδευση του και παρέχοντας του την δυνατότητα να μάθει πιο περίπλοκες εργασίες [136]. Στις παρακάτω υποενότητες παρουσιάζονται οι βασικότερες μη γραμμικές συναρτήσεις ενεργοποίησης με τις περισσότερες από αυτές να συναντώνται και στην παρούσα εργασία. Ένα παράδειγμα μη γραμμικής συνάρτησης ενεργοποίησης παρουσιάζεται στο Σχήμα Α΄.5.

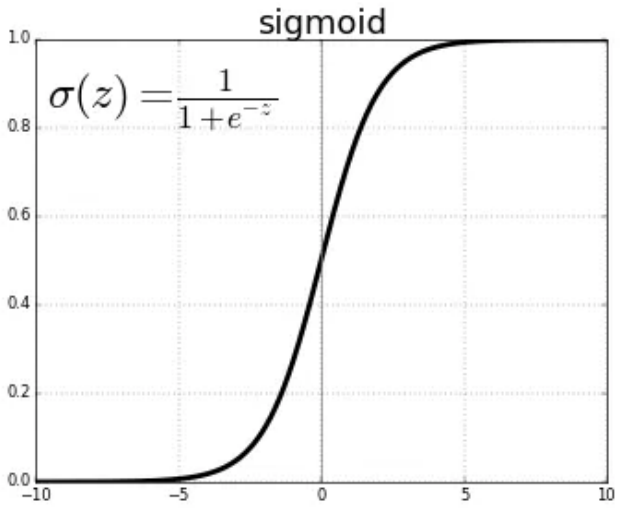
Α΄.3.1 Sigmoid - Softmax συνάρτησεις ενεργοποίησης

Η σιγμοειδής (sigmoid) και η softmax συναρτήσεις ενεργοποίησης σχηματίζουν το γράμμα "S" και η μορφή τους φαίνεται στο Σχήμα Α΄.6. Η διαφορά των δύο συναρτήσεων είναι ότι



Σχήμα Α΄.5: *Παράδειγμα μη γραμμικής συνάρτησης ενεργοποίησης [17].*

η πρώτη δέχεται ως είσοδο μια τιμή και χρησιμοποιείται για την δυαδική ταξινόμηση, ενώ η δεύτερη δέχεται ως είσοδο έναν πίνακα με διαστάση ίση με όσες κλάσεις διαθέτουμε και χρησιμοποιείται για ταξινόμηση περισσότερων των δύο κατηγοριών.

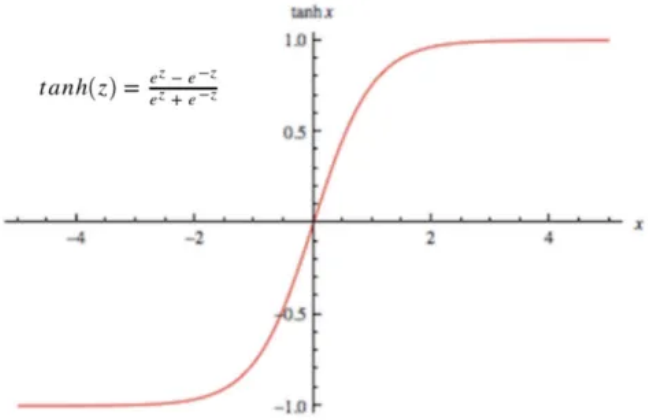


Σχήμα Α΄.6: *Sigmoid - Softmax συναρτήσεις ενεργοποίησης [18].*

Οι παραπάνω συναρτήσεις είναι συνεχώς διαφορίσιμες και χρησιμοποιούνται σε μοντέλα που θέλουν να προβλέψουν πιθανότητες ως έξοδο. Ωστόσο, εμφανίζεται το πρόβλημα της εξαφάνισης των gradients και γενικά αυτές οι συναρτήσεις έχουν αργή σύγκλιση.

Α΄.3.2 Tanh συνάρτηση ενεργοποίησης

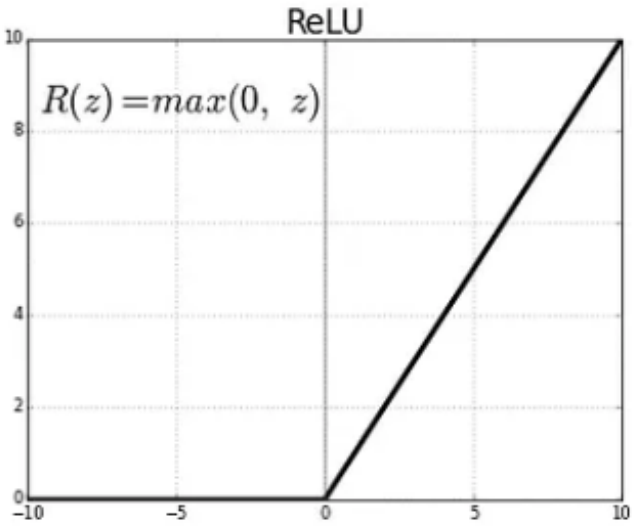
Η μορφή της συνάρτησης ενεργοποίησης tanh φαίνεται στο Σχήμα Α΄.7. Η διαφορά της από την sigmoid είναι ότι το σύνολο τιμών της εκτείνεται από το -1 έως το 1 σε αντίθεση με την δεύτερη που εκτείνεται από το 0 εώς το 1. Η συνάρτηση αυτή λύνει το πρόβλημα ότι όλες οι τιμές είχαν το ίδιο πρόσημο, καθώς οι αρνητικές τιμές αντιστοιχίζονται στα αρνητικά και το μηδέν αντιστοιχίζεται κοντά στο μηδέν. Ωστόσο το πρόβλημα εξαφάνισης των παραγώγων συνεχίζει να υπάρχει, καθώς και οι τιμές των παραγώγων είναι πολυ χαμηλές [18].



Σχήμα Α΄.7: *Tanh συναρτήση ενεργοποίησης [18].*

Α΄.3.3 ReLU (Rectified Linear Unit) συνάρτηση ενεργοποίησης

Η συνάρτηση ενεργοποίησης ReLU έχει την μορφή που φαίνεται στο Σχήμα Α΄.8. Η συνάρτηση αυτή μηδενίζει όλες τις αρνητικές τιμές εισόδου και εφαρμόζει γραμμική συνάρτηση στις θετικές. Είναι μη γραμμική, μπορεί να μεταφέρει τα λάθη προς τα πίσω (backpropagation) και να επιταχύνει σε μεγάλο βαθμό της σύγκλιση του SGD. Ακόμη έχει μικρότερο υπολογιστικό κόστος και δεν ενεργοποιεί όλους τους νευρώνες ταυτόχρονα, κάνοντας το δίκτυο αραιό και αποδοτικό. Ωστόσο, το γεγονός ότι μηδενίζει όλες τις αρνητικές τιμές εισόδου επιβαρύνει την εκπαίδευση του δικτύου, καθώς κάποιες τιμές δεν ενημερώνονται κατά το backpropagation [18].

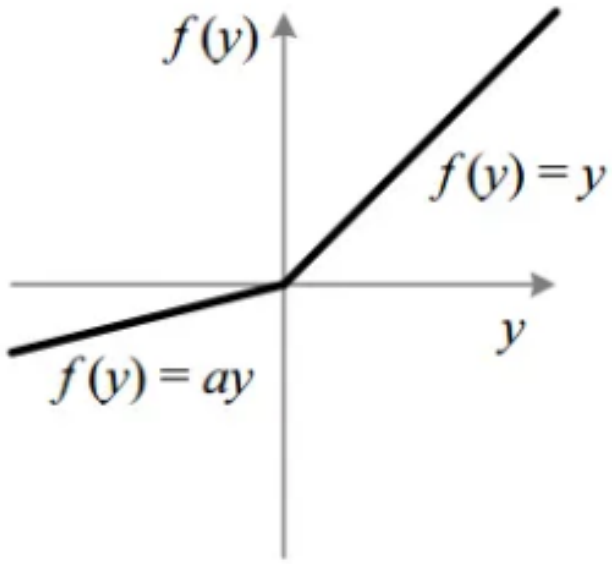


Σχήμα Α΄.8: *ReLU συναρτήση ενεργοποίησης [17].*

Α΄.3.4 Leaky ReLU συνάρτηση ενεργοποίησης

Η μορφή της Leaky ReLU παρουσιάζεται στο Σχήμα Α΄.9. Όπως αναφέρθηκε στην προηγούμενη υποενότητα το πρόβλημα της ReLU είναι ότι μηδενίζει τις αρνητικές τιμές, με α-

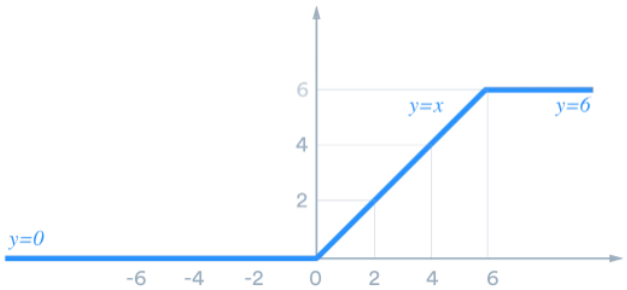
ποτέλεσμα αν η έξοδος ενός νευρώνα είναι συνεχώς αρνητική, η ReLU να δίνει μηδενικό αποτέλεσμα. Με αυτόν τον τρόπο οι παράγωγοι στο backpropagation δεν πηγαίνουν στα προηγούμενα επίπεδα και δεν εκπαιδεύεται το μοντέλο. Για αυτόν τον λόγο η Leaky ReLU δεν μηδενίζει τις αρνητικές τιμές, αλλά τις πολλαπλασιάζει με ένα πολύ μικρό όρο α. Ακόμη η συνάρτηση αυτή κάνει πιο γρήγορη την εκπαίδευση των δικτύων που την χρησιμοποιούν [18].



Σχήμα Α΄.9: *Leaky ReLU συναρτήση ενεργοποίησης [17].*

Α΄.3.5 ReLU6 συνάρτηση ενεργοποίησης

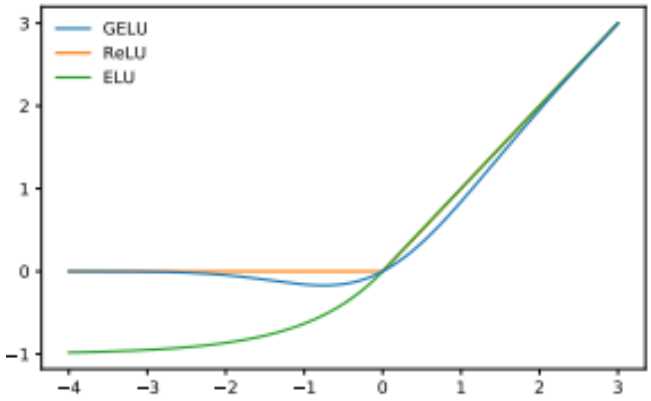
Η μορφή της ReLU6 φαίνεται στο Σχήμα Α΄.10. Η συνάρτηση αυτή μετά την τιμή 6 γίνεται σταθερή και αποκτά άνω όριο, ωθώντας το μοντέλο να μαθαίνει αραιά χαρακτηριστικά απο νωρίς.



Σχήμα Α΄.10: *ReLU-6 συναρτήση ενεργοποίησης [18].*

Α΄.3.6 GeLU (Gaussian-error linear unit) συνάρτηση ενεργοποίησης

Η γραφική παράσταση της συνάρτησης GeLU φαίνεται με μπλε χρώμα στο Σχήμα Α΄.11. Σκοπός της συνάρτησης είναι να συνδυάσει τρία πράγματα. Πρώτος στόχος είναι η απόκτηση της δυνατότητας που έχει η συνάρτηση ενεργοποίησης ReLU να επιτρέπει την γρήγορη και αποδοτική σύγκλιση του δικτύου. Ακόμη ένας στόχος είναι η ενσωμάτωση της ιδιότητας του Dropout να κανονικοποιεί το μοντέλο, μηδενίζοντας τυχαία κάποιους νευρώνες. Τέλος, η GeLU θέλει να αξιοποιήσει το Zoneout που πολλαπλασιάζει στοχαστικά την είσοδο με την τιμή ένα. Ο υπολογισμός της τιμή της GeLU δίνεται από την σχέση Α΄.8 [137].



Σχήμα Α΄.11: *GELU συναρτήση ενεργοποίησης [19].*

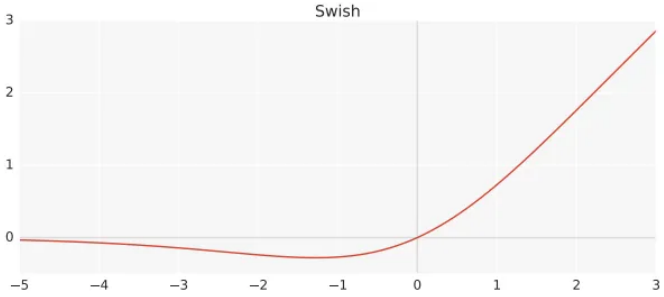
$$GELU(x) = xP(X \leq x) = x\phi(x) \simeq 0.5x(1 + tanh[\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)]) \quad \text{(Α΄.8)}$$

Α΄.3.7 h-swish συνάρτηση ενεργοποίησης

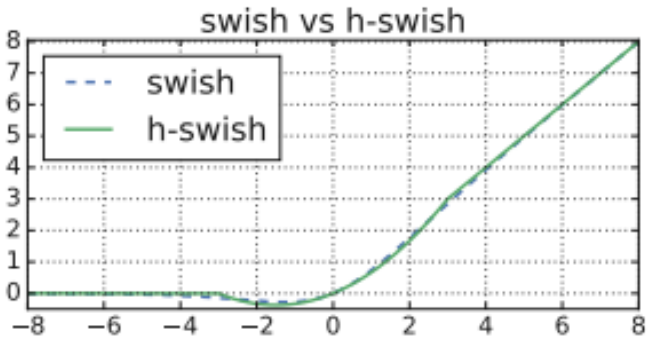
Η μορφή της συνάρτησης ενεργοποίησης swish φαίνεται στο Σχήμα Α΄.12 και ο υπολογισμός των τιμών της δίνεται από τον τύπο Α΄.9. Η συνάρτηση αυτή παρουσιάζει καλύτερα αποτελέσματα σε βαθύτερα δίκτυα από την ReLU, ενώ η απλότητα της καθιστά εύκολη την χρησιμοποίηση της. Είναι οριοθετημένη μόνο από κάτω και δεν είναι μονοτονική, κάτι το οποίο κάνει την διαφορά στην απόδοση της. Τέλος χρησιμοποιεί self-gating, το οποίο απαιτεί βαθμωτή είσοδο, ενώ στην περίπτωση του multi-gating χρειάζονται πολλαπλές είσοδοι με δύο τιμές. Τα παραπάνω είναι εμπνευσμένα από την συνάρτηση sigmoid στα LSTM.

$$Swish(x) = x \cdot sigmoid(x) \quad \text{(Α΄.9)}$$

Στο Σχήμα Α΄.13 φαίνεται η διαφορά στην απεικόνιση της h-swish (hard swish) με την απλή swish, ενώ η σχέση από την οποία υπολογίζεται η πρώτη είναι η Α΄.10. Η h-swish αντικαθιστά την συνάρτηση sigmoid που είναι υπολογιστικά ακριβή με μία γραμμική συνάρτηση που εφαρμόζεται ανά στοιχείο και είναι πιο εύκολο να υπολογιστεί. Συνεπώς, η συνάρτηση h-swish είναι πιο γρήγορη υπολογιστικά και πιο φιλική προς την κβαντοποίηση.



Σχήμα Α΄.12: *swish συναρτήση ενεργοποίησης [20].*



Σχήμα Α΄.13: *Σύγκριση swish με h-swish [21].*

$$H - Swish(x) = x \cdot \frac{ReLU6(x+3)}{6} = x \cdot \frac{min(max(x+3,0),6)}{6} \quad \text{(Α΄.10)}$$

Βιβλιογραφία

[1] *CNN architecture*. https://www.analyticsvidhya.com/blog/2022/03/basics-of-cnn-in-deep-learning/. Ημερομηνία πρόσβασης: 15-07-2023.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser και Illia Polosukhin. *Attention Is All You Need*, 2017.

[3] *dino self-supervised method*. https://towardsdatascience.com/paper-explained-dino-emerging-properties-in-self-supervised-vision-transformers-f9386df266f1. Ημερομηνία πρόσβασης: 24-04-2023.

[4] Aravind Srinivas, Tsung Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel και Ashish Vaswani. *Bottleneck Transformers for Visual Recognition*, 2021.

[5] Faris Almalik, Mohammad Yaqub και Karthik Nandakumar. *Self-Ensembling Vision Transformer (SEViT) for Robust Medical Image Classification*, 2022.

[6] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli και Levent Sagun. *ConViT: improving vision transformers with soft convolutional inductive biases*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114005, 2022.

[7] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li και Humphrey Shi. *Escaping the Big Data Paradigm with Compact Transformers*, 2022.

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles και Hervé Jégou. *Training data-efficient image transformers & distillation through attention*, 2021.

[9] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu και Wei Wu. *Incorporating Convolution Designs into Visual Transformers*, 2021.

[10] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte και Luc Van Gool. *LocalViT: Bringing Locality to Vision Transformers*, 2021.

[11] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao και Qixiang Ye. *Conformer: Local Features Coupling Global Representations for Visual Recognition*, 2021.

[12] Ethan Cheng Haichao Wei και Chunming Peng. *RobustVision: Making CNN and ViT Good Friends with Pre-Trained Vision Model*. http://cs231n.stanford.edu/reports/2022/pdfs/143.pdf. Ημερομηνία πρόσβασης: 23-07-2023.

[13] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan και Lei Zhang. *CvT: Introducing Convolutions to Vision Transformers*, 2021.

[14] *Raster Order*. https://en.wikipedia.org/wiki/Raster_scan. Ημερομηνία πρόσβασης: 09-05-2023.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit και Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021.

[16] *Bahdanau Attention*. https://towardsdatascience.com/sequence-2-sequence-model-with-attention-mechanism-9e9ca2a613a. Ημερομηνία πρόσβασης: 16-07-2023.

[17] *Activation Functions*. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6. Ημερομηνία πρόσβασης: 23-07-2023.

[18] *Activation Functions*. https://prateekvishnu.medium.com/activation-functions-in-neural-networks-bf5c542d5fec. Ημερομηνία πρόσβασης: 23-07-2023.

[19] Dan Hendrycks και Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*, 2023.

[20] *swish Activation Function*. https://medium.com/@neuralnets/swish-activation-function-by-google-53e1ea86f820. Ημερομηνία πρόσβασης: 23-07-2023.

[21] Andrew Howard, Mark Sandler, Grace Chu, Liang Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le και Hartwig Adam. *Searching for MobileNetV3*, 2019.

[22] *Artificial Intelligence - Machine Learning - Deep Learning*. http://danieljhand.com/the-relationship-between-artificial-intelligence-ai-machine-learning-ml-and-deep-learning-dl.html. Ημερομηνία πρόσβασης: 08-05-2023.

[23] *Machine Learning image*. https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications. Ημερομηνία πρόσβασης: 08-05-2023.

[24] *Machine Learning - Deep Learning Comparison*. https://thedatascientist.com/what-deep-learning-is-and-isnt/. Ημερομηνία πρόσβασης: 08-05-2023.

[25] *LocalViT Code*. https://github.com/ofsoundof/LocalViT. Ημερομηνία πρόσβασης: 22-07-2023.

[26] *Conformer Code*. https://github.com/pengzhiliang/Conformer. Ημερομηνία πρόσβασης: 22-07-2023.

[27] *Compact Transformers Code*. https://github.com/SHI-Labs/Compact-Transformers. Ημερομηνία πρόσβασης: 22-07-2023.

[28] *DeiT Code*. https://github.com/facebookresearch/deit. Ημερομηνία πρόσβασης: 22-07-2023.

[29] *CeiT Code*. https://github.com/coeusguo/ceit. Ημερομηνία πρόσβασης: 22-07-2023.

[30] *CvT Code*. https://github.com/leoxiaobin/CvT. Ημερομηνία πρόσβασης: 22-07-2023.

[31] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis και Riaz Agha. *World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)*. *International Journal of Surgery*, 76:71–76, 2020.

[32] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun και Liming Xia. *Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases*. *Radiology*, 296(2):E32–E40, 2020.

[33] Yogesh H Bhosale και K Sridhar Patnaik. *Application of Deep Learning Techniques in Diagnosis of Covid-19 (Coronavirus): A Systematic Review*. *Neural Process Lett*, σελίδες 1–53, 2022.

[34] Paraskevi Antonia Theofilou, Georgios Tsatiris και Stefanos Kollias. *Automatic assessment of Parkinson's patients' dyskinesia using non-invasive machine learning methods*. *2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies (IMET)*, σελίδες 1–4, 2022.

[35] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos D. Kollias και Georgios L. Tagaris. *Deep neural architectures for prediction in healthcare*. *Complex & Intelligent Systems*, 4:119–131, 2018.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser και Illia Polosukhin. *Attention is All you Need*. *Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[37] *Artificial Intelligence*. https://el.wikipedia.org/wiki/%CE%A4%CE%B5%CF%87%CE%BD%CE%B7%CF%84%CE%AE_%CE%BD%CE%BF%CE%B7%CE%BC%CE%BF%CF%83%CF%8D%CE%BD%CE%B7. Ημερομηνία πρόσβασης: 08-05-2023.

[38] *Advantages of Artificial Intelligence*. https://bigblue.academy/gr/ti-einai-i-texniti-noimosuni. Ημερομηνία πρόσβασης: 08-05-2023.

[39] *Machine Learning definition*. https://en.wikipedia.org/wiki/Machine_learning. Ημερομηνία πρόσβασης: 08-05-2023.

[40] *Unsupervised Learning - Main Tasks*. https://www.ibm.com/topics/unsupervised-learning. Ημερομηνία πρόσβασης: 08-05-2023.

[41] *Deep Learning*. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks. Ημερομηνία πρόσβασης: 08-05-2023.

[42] *image classification definition*. https://www.simplilearn.com/tutorials/deep-learning-tutorial/guide-to-building-powerful-keras-image-classification-models. Ημερομηνία πρόσβασης: 24-05-2023.

[43] *Image Classification Challenges*. https://towardsdatascience.com/main-challenges-in-image-classification-ba24dc78b558. Ημερομηνία πρόσβασης: 17-07-2023.

[44] *Convolutional Neural Network*. https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/. Ημερομηνία πρόσβασης: 15-07-2023.

[45] *Convolutional Neural Network_wikipedia*. https://en.wikipedia.org/wiki/Convolutional_neural_network. Ημερομηνία πρόσβασης: 15-07-2023.

[46] *AlexNet*. https://medium.com/@smallfishbigsea/a-walk-through-of-alexnet-6cbd137a5637. Ημερομηνία πρόσβασης: 15-07-2023.

[47] Emerald U. Henry, Onyeka Emebob και Conrad Asotie Omonhinmin. *Vision Transformers in Medical Imaging: A Review*, 2022.

[48] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár και Ross Girshick. *Early Convolutions Help Transformers See Better*, 2021.

[49] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman, Hifsa Asif, Aqsa Asif και Umair Farooq. *A survey of the Vision Transformers and its CNN-Transformer based Variants*, 2023.

[50] Siddique Latif, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat και Junaid Qadir. *Transformers in Speech Processing: A Survey*, 2023.

[51] Zixiu Wu, Ozan Caglayan, Julia Ive, Josiah Wang και Lucia Specia. *Transformer-based Cascaded Multimodal Speech Translation*. *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, 2019. Association for Computational Linguistics.

[52] M. Onat Topal, Anil Bas και Imkevan Heerden. *Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet*, 2021.

[53] Linhao Dong, Shuang Xu και Bo Xu. *Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition*. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 5884–5888, 2018.

[54] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben και Björn W. Schuller. *Dawn of the transformer era in speech emotion recognition: closing the valence gap*, 2022.

[55] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias και Georgios Tagaris. *Deep neural architectures for prediction in healthcare*. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[56] Athanasios Tagaris, Dimitrios Kollias και Andreas Stafylopatis. *Assessment of Parkinson's disease based on deep neural networks*. *International Conference on Engineering Applications of Neural Networks*, σελίδες 391–403. Springer, 2017.

[57] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris και Stefanos Kollias. *Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset*. *International Journal on Artificial Intelligence Tools*, 27(03):1850011, 2018.

[58] Ilianna Kollia, Andreas Georgios Stafylopatis και Stefanos Kollias. *Predicting Parkinson's disease using latent information extracted from deep neural networks*. *2019 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8. IEEE, 2019.

[59] James Wingate, Ilianna Kollia, Luc Bidaut και Stefanos Kollias. *Unified deep learning approach for prediction of Parkinson's disease*. *IET Image Processing*, 14(10):1980–1989, 2020.

[60] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian και Stefanos Kollias. *Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 537–544, 2021.

[61] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *Ai-mia: Covid-19 detection & severity analysis through medical imaging*. *arXiv preprint arXiv:2206.04732*, 2022.

[62] Anastasios Arsenos, Dimitrios Kollias και Stefanos Kollias. *A Large Imaging Database and Novel Deep Neural Architecture for Covid-19 Diagnosis*. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, σελίδες 1–5. IEEE, 2022.

[63] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *AI-MIA: Covid-19 detection and severity analysis through medical imaging*. *Computer Vision-ECCV 2022 Workshops: Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, σελίδες 677–690. Springer, 2023.

[64] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis και Stefanos Kollias. *Adaptation and contextualization of deep neural network models*. *2017 IEEE symposium series on computational intelligence (SSCI)*, σελίδες 1–8. IEEE, 2017.

[65] D Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, I Kollia, L Sukissian, J Wingate και S Kollias. *Deep Transparent Prediction through Latent Representation Analysis*. *arXiv preprint arXiv:2009.07044*, 2020.

[66] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate και S Kollias. *Transparent adaptation in deep medical image diagnosis*. *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, σελίδες 251–267. Springer, 2020.

[67] Fabio De Sousa Ribeiro, Francesco Calivá, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis και Stefanos Kollias. *Deep bayesian self-training*. *Neural Computing and Applications*, 32(9):4275–4291, 2020.

[68] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos Kollias. *Capsule routing via variational bayes*. *Proceedings of the AAAI Conference on Artificial Intelligence*, τόμος 34, σελίδες 3749–3756, 2020.

[69] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos Kollias. *Introducing routing uncertainty in capsule networks*. *Advances in Neural Information Processing Systems*, 33:6490–6502, 2020.

[70] Nikolaos Simou και Stefanos Kollias. *Fire: A fuzzy reasoning engine for impecise knowledge*. Citeseer, 2007.

[71] Francesco Caliva, Fabio Sousa De Ribeiro, Antonios Mylonakis, Christophe Demazi'ere, Paolo Vinai, Georgios Leontidis και Stefanos Kollias. *A deep learning approach to anomaly detection in nuclear reactors*. *2018 International joint conference on neural networks (IJCNN)*, σελίδες 1–8. IEEE, 2018.

[72] Stefanos Kollias, Miao Yu, James Wingate, Aiden Durrant, Georgios Leontidis, Georgios Alexandridis, Andreas Stafylopatis, Antonios Mylonakis, Paolo Vinai και Christophe Demaziere. *Machine learning for analysis of real nuclear plant data in the frequency domain*. *Annals of Nuclear Energy*, 177:109293, 2022.

[73] Bashar Alhnaity, Stefanos Kollias, Georgios Leontidis, Shouyong Jiang, Bert Schamp και Simon Pearson. *An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth*. *Information Sciences*, 560:35–50, 2021.

[74] Bashar Alhnaity, Simon Pearson, Georgios Leontidis και Stefanos Kollias. *Using deep learning to predict plant growth and yield in greenhouse environments*. *International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296*, σελίδες 425–432, 2019.

[75] Andreas Psaroudakis και Dimitrios Kollias. *MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 2367–2375, 2022.

[76] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen και Stefanos Zafeiriou. *Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 5888–5897, 2023.

[77] Dimitrios Kollias και Stefanos Zafeiriou. *Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm*. *2018 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8. IEEE, 2018.

[78] G. Tsechpenakis, K. Rapantzikos, N. Tsapatsoulis και S. Kollias. *A snake model for object tracking in natural sequences*. *Signal processing: image communication*, 19(3):219–238, 2004.

[79] A.D. Doulamis, Y.S. Avrithis, N.D. Doulamis και S.D. Kollias. *Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback*. *Proceedings IEEE International Conference on Multimedia Computing and Systems*, τόμος 2, σελίδες 954–958. IEEE, 1999.

[80] Manolis Wallace, Ilias Maglogiannis, Kostas Karpouzis, George Kormentzas και Stefanos Kollias. *Intelligent one-stop-shop travel recommendations using an adaptive neural network and clustering of history*. *Information Technology & Tourism*, 6(3):181–193, 2003.

[81] Yannis Avrithis, Yiannis Xirouhakis και Stefanos Kollias. *Affine-invariant curve normalization for object shape representation, classification, and retrieval*. *Machine Vision and Applications*, 13:80–94, 2001.

[82] A.N. Delopoulos και S.D. Kollias. *Optimal filter banks for signal reconstruction from noisy subband components*. *IEEE transactions on signal processing*, 44(2):212–224, 1996.

[83] Hanan S Alghamdi, Ghada Amoudi, Salma Elhag, Kawther Saeedi και Jomanah Nasser. *Deep Learning Approaches for Detecting COVID-19 From Chest X-Ray Images: A Survey*. *IEEE Access*, 9:20235–20254, 2021.

[84] Zaid Abdi Alkareem Alyasseri, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A Awadallah, Ammar Kamal Abasi, Sharif Naser Makhadmeh, Osama Ahmad Alomari, Karrar Hameed Abdulkareem, Afzan Adam, Robertas Damasevicius, Mazin Abed Mohammed και Raed Abu Zitar. *Review on COVID-19 diagnosis models based on machine learning and deep learning approaches*. *Expert Syst*, 39(3):ε12759, 2021.

[85] Nandhini Subramanian, Omar Elharrouss, Somaya Al-Maadeed και Muhammed Chowdhury. *A review of deep learning-based detection methods for COVID-19*. *Computers in Biology and Medicine*, 143:105233, 2022.

[86] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui και Adel Mellit. *A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope*. *Sustainability*, 15(7), 2023.

[87] N Narayan Das, N Kumar, M Kaur, V Kumar και D Singh. *Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays*. *Ing Rech Biomed*, 43(2):114–119, 2020.

[88] Kevser Sahinbas και Ferhat Ozgur Catak. *24 - Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images*. *Data Science for COVID-19*Utku Kose, Deepak Gupta, Victor Hugo C. de Albuquerque και Ashish Khanna, επιμελητές, σελίδες 451–466. Academic Press, 2021.

[89] Nirmala Devi Kathamuthu, Shanthi Subramaniam, Quynh Hoang Le, Suresh Muthusamy, Hitesh Panchal, Suma Christal Mary Sundararajan, Ali Jawad Alrubaie και Musaddak Maher Abdul Zahra. *A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications*. *Advances in Engineering Software*, 175:103317, 2023.

[90] Pramit Dutta, Tanny Roy και Nafisa Anjum. *COVID-19 Detection using Transfer Learning with Convolutional Neural Network*. *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, σελίδες 429–432, 2021.

[91] Yusuf Brima, Marcellin Atemkeng, Stive Tankio Djiokap, Jaures Ebiele και Franklin Tchakounté. *Transfer Learning for the Detection and Diagnosis of Types of Pneumonia including Pneumonia Induced by COVID-19 from Chest X-ray Images*. *Diagnostics*, 11(8), 2021.

[92] Saddam Hussain Khan, Anabia Sohail, Asifullah Khan και Yeon Soo Lee. *COVID-19 Detection in Chest X-ray Images Using a New Channel Boosted CNN*. *Diagnostics (Basel)*, 12(2), 2022.

[93] Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N. Plataniotis και Arash Mohammadi. *COVID-CAPS: A Capsule Network-based Framework for Identification of COVID-19 cases from X-ray Images*, 2020.

[94] Dalia Ezzat, Aboul Ella Hassanien και Hassan Aboul Ella. *An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization*. *Applied Soft Computing*, 98:106742, 2021.

[95] Awf A. Ramadhan και Muhammet Baykara. *A Novel Approach to Detect COVID-19: Enhanced Deep Learning Models with Convolutional Neural Networks*. *Applied Sciences*, 12(18), 2022.

[96] Kabid Hassan Shibly, Samrat Kumar Dey, Md Tahzib Ul Islam και Md Mahbubur Rahman. *COVID faster R-CNN: A novel framework to Diagnose Novel Coronavirus Disease (COVID-19) in X-Ray images*. *Informatics in Medicine Unlocked*, 20:100405, 2020.

[97] Asmaa Abbas, Mohammed Abdelsamea και Mohamed Gaber. *Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network*. *Applied Intelligence*, 51:1–11, 2021.

[98] Anubhav Sharma, Karamjeet Singh και Deepika Koundal. *A novel fusion based convolutional neural network approach for classification of COVID-19 from chest X-ray images*. *Biomedical Signal Processing and Control*, 77:103778, 2022.

[99] Daphna Keidar, Daniel Yaron, Elisha Goldstein, Yair Shachar, Ayelet Blass, Leonid Charbinsky, Israel Aharony, Liza Lifshitz, Dimitri Lumelsky, Ziv Neeman, Matti Mizrachi, Majd Hajouj, Nethanel Eizenbach, Eyal Sela, Chedva S Weiss, Philip Levin, Ofer Benjaminov, Gil N Bachar, Shlomit Tamir, Yael Rapson, Dror Suhami, Eli Atar, Amiel A Dror, Naama R Bogot, Ahuva Grubstein, Nogah Shabshin, Yishai M Elyada και Yonina C Eldar. *COVID-19 classification of X-ray images using deep neural networks*. *Eur Radiol*, 31(12):9654–9663, 2021.

[100] Linda Wang, Zhong Qiu Lin και Alexander Wong. *COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images*. *Sci. Rep.*, 10(1):19549, 2020.

[101] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg και Kevin Smith. *Is it Time to Replace CNNs with Transformers for Medical Images?* *CoRR*, αβσ/2108.09038, 2021.

[102] Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali και Muhammad Moazam Fraz. *Vision Transformers in medical computer vision—A contemplative retrospection*. *Engineering Applications of Artificial Intelligence*, 122:106126, 2023.

[103] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan και Huazhu Fu. *Transformers in medical imaging: A survey*. *Medical Image Analysis*, 88:102802, 2023.

[104] Yin Dai και Yifan Gao. *TransMed: Transformers Advance Multi-modal Medical Image Classification*, 2021.

[105] Shuang Yu, Kai Ma, Qi Bi, Cheng Bian, Munan Ning, Nanjun He, Yuexiang Li, Hanruo Liu και Yefeng Zheng. *MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification*. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*Marleende Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng και Caroline Essert, επιμελητές, σελίδες 45–54, Cham, 2021. Springer International Publishing.

[106] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji και Yongbing Zhang. *TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification*, 2021.

[107] Anwar Khan και Boreom Lee. *Gene Transformer: Transformers for the Gene Expression-based Classification of Lung Cancer Subtypes*, 2021.

[108] Moinak Bhattacharya, Shubham Jain και Prateek Prasanna. *RadioTransformer: A Cascaded Global-Focal Transformer for Visual Attention-guided Disease Classification*, 2022.

[109] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille και Yuyin Zhou. *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*, 2021.

[110] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu και Sen Zha. *TransBTS: Multimodal Brain Tumor Segmentation Using Transformer*, 2021.

[111] Eunji Jun, Seungwoo Jeong, Da Woon Heo και Heung Il Suk. *Medical Transformer: Universal Brain Encoder for 3D MRI Analysis*, 2021.

[112] Mohamed Chetoui και Moulay A. Akhloufi. *Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays*. *Journal of Clinical Medicine*, 11(11):3013, 2022.

[113] Koushik Sivarama Krishnan και Karthik Sivarama Krishnan. *Vision Transformer based COVID-19 Detection using Chest X-rays*. *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, σελίδες 644–648, 2021.

[114] Om Uparkar, Jyoti Bharti, R.K. Pateriya, Rajeev Kumar Gupta και Ashutosh Sharma. *Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images*. *Procedia Computer Science*, 218:2338–2349, 2023. Ιντερνατιοναλ ὄνφερενσε ον Μαχηινε Λεαρνινγ ανδ Δατα Ενγινεερινγ.

[115] Luis Balderas, Miguel Lastra, Antonio J. Láinez-Ramos-Bossini και José M. Benítez. *COVID-ViT: COVID-19 Detection Method Based on Vision Transformers*. *Intelligent Systems Design and Applications*Ajith Abraham, Sabri Pllana, Gabriella Casalino, Kun Ma και Anu Bajaj, επιμελητές, σελίδες 81–90, Cham, 2023. Springer Nature Switzerland.

[116] Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla και A. P. Prathosh. *xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography*. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–10, 2022.

[117] Gabriel Iluebe Okolo, Stamos Katsigiannis και Naeem Ramzan. *IEViT: An enhanced vision transformer architecture for chest X-ray image classification*. *Computer Methods and Programs in Biomedicine*, 226:107141, 2022.

[118] Debaditya Shome, T. Kar, Sachi Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang και Abdul Saudagar. *COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare*. *International Journal of Environmental Research and Public Health*, 18(21):11086, 2021.

[119] Hang Yang, Liyang Wang, Yitian Xu και Xuhua Liu. *CovidViT: a novel neural network with self-attention mechanism to detect Covid-19 through X-ray images*. *Int. J. Mach. Learn. Cybern.*, 14(3):973–987, 2023.

[120] Shehan Perera, Srikar Adhikari και Alper Yilmaz. *POCFormer: A Lightweight Transformer Architecture for Detection of COVID-19 Using Point of Care Ultrasound*, 2021.

[121] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang και Hao Ma. *Linformer: Self-Attention with Linear Complexity*, 2020.

[122] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov και Liang Chieh Chen. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, 2019.

[123] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng και Shuicheng Yan. *VOLO: Vision Outlooker for Visual Recognition*, 2021.

[124] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin και Baining Guo. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, 2021.

[125] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu και Yunhe Wang. *Transformer in Transformer*, 2021.

[126] Thomas Kwee και Robert Kwee. *Chest CT in COVID-19: What the Radiologist Needs to Know*. *RadioGraphics*, 40:1848–1865, 2020.

[127] Chih Chung Hsu, Guan Lin Chen και Mei Hsuan Wu. *Visual Transformer with Statistical Test for COVID-19 Classification*, 2021.

[128] R. F. Woolson. *Wilcoxon Signed-Rank Test*, σελίδες 1–3. John Wiley & Sons, Ltd, 2008.

[129] Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed και Tahir Hamid. *COVID-19 infection localization and severity grading from chest X-ray images*. *Computers in Biology and Medicine*, 139:105002, 2021.

[130] *ImageNet*. https://paperswithcode.com/dataset/imagenet. Ημερομηνία πρόσβασης: 14-07-2023.

[131] *ImageNet2*. https://en.wikipedia.org/wiki/ImageNet. Ημερομηνία πρόσβασης: 14-07-2023.

[132] *ImageNet3*. https://huggingface.co/datasets/imagenet-1k. Ημερομηνία πρόσβασης: 14-07-2023.

[133] *Google Colab free edition Specs*. https://saturncloud.io/blog/whats-the-hardware-spec-for-google-colaboratory/. Ημερομηνία πρόσβασης: 22-07-2023.

[134] Sangwon Kim, Jaeyeal Nam και Byoung Chul Ko. *ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder*. *Proceedings of the 39th International Conference on Machine Learning*Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu και Sivan Sabato, επιμελητές, τόμος 162 στο *Proceedings of Machine Learning Research*, σελίδες 11162–11172. PMLR, 2022.

[135] *Attention Mechanism*. https://www.scaler.com/topics/deep-learning/attention-mechanism-deep-learning/. Ημερομηνία πρόσβασης: 16-07-2023.

[136] *Activation Functions' Definition*. https://www.geeksforgeeks.org/activation-functions-neural-networks/. Ημερομηνία πρόσβασης: 23-07-2023.

[137] *GELU Activation Function*. https://medium.com/@shauryagoel/gelu-gaussian-error-linear-unit-4ec59fb2e47c. Ημερομηνία πρόσβασης: 23-07-2023.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

AI	Artificial Intelligence
ANN	Artificial Neural Network
BatchNorm	Batch Normalization
CCT	Compact Convolutional Transformer
CeiT	Convolution-enhanced image Transformer
CVT	Compact Vision Transformer
CvT	Convolutional vision Transformer
DeiT	Data-efficient image Transformer
DL	Deep Learning
FFN	Feed-Forward Network
GELU	Gaussian Error Linear Unit
MHA	Multi-Head Attention
MHSA	Multi-Head Self-Attention
ML	Machine Learning
MLP	MultiLayer Perceptron
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
ViT	Vision Transformer

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
κατώφλι	threshold
συνολο δεδομένων	dataset
εύρωστος	robust
νευρωνικό δίκτυο	neural network
ευρύτερο δίκτυο	wider network
επίπεδο προσοχής	attention layer
συνελικτικό επίπεδο	convolutional layer
χωρικά διαχωρίσιμες συνελίξεις	spatial separable convolutions
κατά βάθος διαχωρίσιμες συνελίξεις	depthwise separable convolutions
συνέλιξη κατά βάθος	depthwise convolution
συνέλιξη κατά σημείο	pointwise convolution
κλειδί	key
τιμή	value
ερώτημα	query
κωδικοποιητής	encoder
αποκωδικοποιητής	decoder
βαθμολογία	score
όραση υπολογιστών	computer vision
βελτιστοποιητής	optimizer
μέγεθος πυρήνας	kernel size
προβολή προς τα κάτω	down-projection
προβολή προς τα πάνω	up-projection
υπολειμματική σύνδεση	residual connection
παράγωγος	gradient
βαθμολογία ευθυγράμμισης	alignment score