ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

“ Μέθοδοι μηχανικής μάθησης για την υποβοήθηση της ιατρικής διάγνωσης πνευμονικής εμβολής με αξιοποίηση κλινικών και απεικονιστικών δεδομένων ”

ΜΑΡΙΑ ΘΩΜΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Μιχάλης Σαβελώνας

Επίκουρος Καθηγητής

Λαμία, 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

“ Μέθοδοι μηχανικής μάθησης για την υποβοήθηση της ιατρικής διάγνωσης πνευμονικής εμβολής με αξιοποίηση κλινικών και απεικονιστικών δεδομένων ”

ΜΑΡΙΑ ΘΩΜΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Μιχάλης Σαβελώνας

Επίκουρος Καθηγητής

Λαμία, 2024

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 03 / 07 / 2024

Η Δηλούσα

Θωμά Μαρία

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

“ Μέθοδοι μηχανικής μάθησης για την υποβοήθηση της ιατρικής διάγνωσης πνευμονικής εμβολής με αξιοποίηση κλινικών και απεικονιστικών δεδομένων ”

ΜΑΡΙΑ ΘΩΜΑ

Τριμελής Επιτροπή:

Μιχάλης Σαβελώνας, Επίκουρος Καθηγητής (επιβλέπων)

Δημήτρης Ιακωβίδης, Καθηγητής

Παναγιώτης Βαρθολομαίος, Επίκουρος Καθηγητής

Στους φίλους και την οικογένεια μου

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν στην υλοποίηση της πτυχιακής μου εργασίας. Πρώτα και κύρια, τον επιβλέποντα καθηγητή μου κ. Μιχάλη Σαβελώνα, επίκουρο καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, για όλο το χρόνο που διέθεσε, την εμπιστοσύνη και τη στήριξη που κατέδειξε καθ' όλη τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας, ακόμα και όταν υπήρχαν δυσκολίες. Επιπλέον, εξίσου μεγάλη ήταν η συνεισφορά των φίλων μου, που με τον δικό τους τρόπο ο κάθε ένας ξεχωριστά με υποστήριζε και μου έδινε δύναμη να συνεχίσω. Η ενθάρρυνσή τους και η υπομονή τους όλο αυτό το διάστημα, κυρίως όταν μοιραζόμουν μαζί τους αρνητικές σκέψεις και δυσκολίες έπαιξαν καθοριστικό ρόλο στο να αντλήσω την απαραίτητη δύναμη για να συνεχίσω. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, που αν και δεν είχε επακριβή γνώση του αντικειμένου με υποστήριζε ψυχολογικά όλο αυτό το διάστημα και ήταν πάντα δίπλα μου να με ενθαρρύνει μέχρι το τέλος.

Μαρία Θωμά,

Λαμία 2024

Περίληψη

Η Πνευμονική Εμβολή (ΠΕ) είναι μια πάθηση των πνευμόνων που χαρακτηρίζεται από μεγάλη νοσηρότητα και θνησιμότητα, ενώ η μη έγκαιρη διάγνωσή της μπορεί να αποβεί θανατηφόρα. Ωστόσο, συχνά η διάγνωση της ΠΕ δυσχεραίνεται, λόγω της ομοιότητας των συμπτωμάτων που παρουσιάζει με άλλες πνευμονικές παθήσεις, όπως είναι η πνευμονία. Η εξέταση της αξονικής υπολογιστικής πνευμονικής αγγειογραφίας (Computed Tomography Pulmonary Angiogram – CTPA) είναι η κύρια επιλογή για τη διάγνωση της ασθένειας, αφού είναι σύντομη και προσφέρει 3Δ εικόνες υψηλής ανάλυσης. Ωστόσο, η συνολική διαδικασία της διάγνωσης παραμένει χρονοβόρα, γεγονός που επιβαρύνει την υγεία του ασθενή. Επιπλέον, η ορθότητα της διάγνωσης εξαρτάται από την ειδίκευση και την εμπειρία του ιατρού. Αυτό συνεπάγεται την αυξημένη πιθανότητα λάθους, ιδιαίτερα σε δύσκολες περιπτώσεις. Τέλος, η εξουθένωση των ιατρών εξαιτίας του μεγάλου φόρτου εργασίας σε ώρες εφημερίας αυξάνει περαιτέρω την πιθανότητα λάθους.

Η ανάγκη για την αντιμετώπιση των παραπάνω προκλήσεων είναι επιτακτική, λόγω της σοβαρότητας της ασθένειας. Στα πλαίσια αυτά, υπάρχει έντονο ερευνητικό ενδιαφέρον στους τομείς της ανάλυσης βιοϊατρικών εικόνων και της Τεχνητής Νοημοσύνης (ΤΝ) για την ανάπτυξη υπολογιστικών συστημάτων υποβοήθησης διάγνωσης (Computer Aided Detection systems - CAD) με στόχο την ενίσχυση της αντικειμενικότητας, της ακρίβειας, της ευαισθησίας και της ειδικότητας των διαγνώσεων, καθώς και τη μείωση του συνολικού χρόνου της διάγνωσης.

Τα τελευταία χρόνια, διερευνώνται νέες μέθοδοι που βασίζονται στην βαθιά μάθηση για την ανάπτυξη συστημάτων CAD για τη διαχείριση περιστατικών ΠΕ. Οι μέθοδοι που βασίζονται σε αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks - CNN) είναι οι ευρύτερα διαδομένες. Πιο πρόσφατα, διερευνάται η εφαρμογή αρχιτεκτονικών με βάση τον οπτικό μετασχηματιστή (Vision Transformer - ViT). Αυτή είναι η προσέγγιση που υιοθετείται και στην παρούσα εργασία.

Σε πρώτο στάδιο, επιχειρείται η εκπαίδευση του μοντέλου ViT για την πρόβλεψη της ΠΕ σε επίπεδο τομών CTPA. Σε δεύτερο στάδιο, οι αναπαραστάσεις των τομών αξιοποιούνται για την εκπαίδευση ενός μοντέλου που βασίζεται σε δίκτυα μακράς - βραχείας μνήμης (Long - Short Term Memory - LSTM), τα οποία είναι σε θέση να συσχετίσουν τις τομές της εξέτασης για την τελική ταξινόμηση του περιστατικού. Για την αντιμετώπιση της ανισορροπίας του συνόλου των δεδομένων σε επίπεδο τομών και

σε επίπεδο εξετάσεων, δοκιμάστηκε η χρήση διαφορετικών μεθόδων δειγματοληψίας. Η μέθοδος της υπό – δειγματοληψίας αποδίδει εξίσου καλά με τις υπόλοιπες μεθόδους, παρά την απώλεια πληροφορίας ως προς την κλάση πλειονότητας. Επιπλέον, δοκιμάστηκε η εφαρμογή της μεθόδου προσαρμογής χαμηλής τάξης (Low - Rank Adaptation - LoRA), η οποία είναι μια πρόσφατη, αποδοτική μέθοδος fine-tuning μεγάλων προεκπαιδευμένων μοντέλων, όπως αυτό που χρησιμοποιείται στην παρούσα εργασία, με τα αποτελέσματα να είναι ιδιαίτερα ενθαρρυντικά και συγκρίσιμα με αυτά εργασιών στην τεχνολογική αιχμή. Τέλος, στο δεύτερο επίπεδο δοκιμάστηκε η εφαρμογή διαφορετικών μεθόδων για την εκπαίδευση του μοντέλου LSTM με ακολουθίες τομών μεταβλητού μεγέθους, όπως συμβαίνει με τις εξετάσεις CTPA. Οι δοκιμές κατέδειξαν ότι η μέθοδος packed padded sequence για την εκπαίδευση των ακολουθιακών μοντέλων, όπως είναι το LSTM, η οποία δε λαμβάνει υπόψη τις μηδενικές τιμές συμπλήρωσης (padding), οδηγεί σε καλύτερα αποτελέσματα σε σύγκριση με άλλους δύο μεθόδους συμπλήρωσης που χρησιμοποιήθηκαν.

Τα πειραματικά αποτελέσματα καταδεικνύουν τη σημασία της μεθόδου LoRA στην ενίσχυση της ανίχνευσης της ΠΕ. Η μέθοδος LoRA επιτρέπει την εκπαίδευση παραμέτρων των κρυφών επιπέδων της αρχιτεκτονικής του ViT, ενώ η ρύθμιση μόνο του 0.17% των παραμέτρων του ViT είναι επαρκής για την επίτευξη συγκρίσιμων τιμών ακρίβειας με αντίστοιχα μοντέλα που εκπαιδεύτηκαν σε όλο το σύνολο των παραμέτρων. Αυτό είναι ιδιαίτερα σημαντικό σε περιπτώσεις περιορισμένης πρόσβασης σε υπολογιστικούς πόρους. Τέλος, η εκπαίδευση με χρήση υπό-δειγματοληψίας τομών είναι συγκρίσιμη ως προς την ακρίβεια με την εκπαίδευση στο σύνολο των τομών. Αυτό μπορεί να εξηγηθεί από την ύπαρξη υψηλής συσχέτισης μεταξύ των διαδοχικών τομών κάθε εξέτασης, έτσι ώστε λίγες αλλά αντιπροσωπευτικές τομές, αρκούν για την επαρκή εκπαίδευση του μοντέλου.

Λέξεις κλειδιά: Πνευμονική Εμβολή, Αξονική Υπολογιστική Πνευμονική Αγγειογραφία, Βαθιά Μάθηση, Οπτικός Μετασχηματιστής, Δίκτυα Μακράς-Βραχείας Μνήμης, Προσαρμογή Χαμηλής Τάξης

Abstract

Pulmonary Embolism (PE) is a lung condition characterized by high morbidity and mortality, and its untimely diagnosis can be fatal. However, diagnosing PE is often challenging due to the similarity of its symptoms to other pulmonary conditions, such as pneumonia. The Computed Tomography Pulmonary Angiogram (CTPA) is the primary imaging choice for PE diagnosis, since it is efficient and provides high-resolution 3D images. Nonetheless, the overall diagnostic process remains time-consuming, which adversely affects patients' health. Moreover, the diagnostic accuracy depends on the MD's specialization and experience, increasing the likelihood of errors, especially in difficult PE cases. Additionally, MD burnout due to high workloads during on-call hours further increases the probability of errors.

Addressing these challenges is imperative due to the severity of PE. In this context, there is significant research activity in the fields of biomedical image analysis and Artificial Intelligence (AI) for the development of Computer Aided Detection (CAD) systems to enhance the objectivity, accuracy, sensitivity, and specificity of diagnoses, as well as to reduce the overall diagnostic time cost.

In recent years, new methods based on deep learning have been explored for developing CAD systems, in order to manage PE cases. Methods based on Convolutional Neural Networks (CNN) architectures are the most widely used. More recently, the application of Vision Transformer (ViT) architectures has been investigated. This is the approach adopted in this work as well.

In the first stage, the ViT model is trained to predict PE at the level of CTPA slices. In the second stage, the slice representations are used as input to train a model based on Long-Short Term Memory (LSTM) networks, which can correlate subsequent slices, in order to classify each case. To address dataset imbalance at both slice and examination levels, various sampling methods were tested. The under-sampling method is comparable to other methods, despite the loss of information regarding the majority class. Additionally, the application of Low-Rank Adaptation (LoRA), a recent, efficient fine-tuning method for large pre-trained models, as the one used in this work, was tested, yielding particularly encouraging results, which are comparable to state-of-the-art. Finally, in the second stage, different methods were tested for training the LSTM model with variable-length slice sequences, as occurs with CTPA examinations. The tests showed that the packed padded sequence method for training sequential models,

such as the LSTM networks, leads to better results, when compared to other two padding methods.

The experimental results demonstrate the importance of LoRA in enhancing PE recognition. LoRA allows training of the hidden layers of ViT, while adjusting only 0.17% of ViT parameters and leads to sufficient training and comparable accuracy with the one obtained by training on the entire parameter set. This is particularly important is settings with limited computational resources. Finally, training using under-sampling of slices is comparable with respect to accuracy with training using the entire set of slices. This can be explained by the high correlation between successive slices of each examination, where fewer but more representative samples are sufficient for adequate model training.

Keywords: Pulmonary Embolism, Computed Tomography Pulmonary Angiogram, Deep Learning, Vision Transformer, Long-Short Term Memory, Low-Rank Adaptation.

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

1. Εισαγωγή

Οι πνεύμονες συγκαταλέγονται στα βασικότερα ζωτικά όργανα του ανθρώπινου οργανισμού. Είναι υπεύθυνοι για μία από τις κυριότερες και σημαντικότερες λειτουργίες για τον άνθρωπο: την λειτουργία της αναπνοής. Διαμέσου της αναπνοής επιτυγχάνεται η πρόσληψη του οξυγόνου και η αποβολή του διοξειδίου του άνθρακα από τον ανθρώπινο οργανισμό. Η Πνευμονική Εμβολή (ΠΕ) αποτελεί μία πάθηση των πνευμόνων που είναι αρκετά απειλητική για την ζωή του ατόμου. Δημιουργείται όταν ένας θρόμβος αίματος, καταλήξει στα αιμοφόρα αγγεία των πνευμόνων δια μέσου της κυκλοφορίας του αίματος [1]. Ο θρόμβος, ο οποίος συνήθως αποσπάται από κάποιον άλλον θρόμβο που έχει σχηματιστεί στην περιοχή της πυέλου ή των κάτω άκρων, έχει σαν αποτέλεσμα την απόφραξη των αιμοφόρων αγγείων των πνευμόνων που συνεπάγεται τη διαταραχή της απρόσκοπτης λειτουργίας τους. Αναλόγως της σημασίας της αρτηρίας που προσβάλλεται, μπορεί η ΠΕ να αποβεί θανατηφόρα [1]. Η ΠΕ αποτελεί την τρίτη πιο συχνή αιτία καρδιαγγειακού θανάτου παγκοσμίως, μετά το εγκεφαλικό επεισόδιο και το έμφραγμα του μυοκαρδίου [1]. Επιπλέον, αποτελεί νόσο που χαρακτηρίζεται από υψηλή νοσηρότητα και θνησιμότητα, συμβάλλοντας σημαντικά στη παγκόσμια βαρύτητα ασθενειών (Global Burden of Disease - GBD) [2,3,4]. Παραθέτοντας στοιχεία από επιδημιολογικές μελέτες, ο αριθμός των ασθενών που πάσχουν από ΠΕ κυμαίνεται από 39-115 ανά 100.000 ετησίως [5]. Ανάλογη μελέτη που διενεργήθηκε στην Ελλάδα, έδειξε ότι η ετήσια συχνότητα εμφάνισης της ΠΕ παρουσίασε ανοδική τάση που κυμαίνεται από 14 (1999) έως 30 (2012) ανά 100.000 πληθυσμού [6]. Τα στοιχεία που προκύπτουν από κλινικές μελέτες που έχουν διενεργηθεί σε παγκόσμιο επίπεδο, υποδηλώνουν υψηλή συσχέτιση μεταξύ της λοίμωξης από τον ιό SARS-CoV-2 (COVID-19) και τον αυξημένο κίνδυνο πνευμονικής εμβολής (ΠΕ) [12].

Οι παρατηρούμενες ανοδικές τάσεις της ΠΕ μπορούν εν μέρει να ερμηνευθούν από τη συνεχή βελτίωση των διαγνωστικών στρατηγικών που υιοθετούνται για την ανίχνευσή της. Η ευρεία χρήση της αξονικής πνευμονικής αγγειογραφίας (Computed Tomography Pulmonary Angiogram - CTPA), η υιοθέτηση όλο και περισσότερων διαγνωστικών διαδικασιών, η μεγαλύτερη ευαισθητοποίηση των κλινικών ιατρών σχετικά με τη νόσο και το υψηλό ποσοστό τυχαίας διάγνωσης (όταν δηλαδή η αξονική τομογραφία διενεργείται για άλλους σκοπούς π.χ. σταδιοποίηση του καρκίνου) μπορεί να δικαιολογήσει τον εντοπισμό όλο και περισσότερων περιστατικών ΠΕ [6,13].

Η διάγνωση της ΠΕ μπορεί να αποτελέσει πρόκληση καθώς τα συμπτώματά της δεν είναι συγκεκριμένα αλλά ποικίλλουν και εμφανίζουν ομοιότητες με άλλες ασθένειες όπως η πνευμονία [3]. Ασθενείς με ΠΕ μπορεί να εμφανίζουν δύσπνοια, πόνο στο στήθος, προσυγκοπή, συγκοπή ή αιμόπτυση. Τα συμπτώματα εξαρτώνται από το μέγεθος και τον αριθμό των αρτηριακών κλάδων του πνευμονικού αγγειακού δικτύου που αποφράσσονται [7].

Η ασθένεια φέρει σημαντική νοσηρότητα και θνησιμότητα. Αν δεν αντιμετωπιστεί, σχετίζεται με ποσοστό θνησιμότητας 30%. Σε περίπτωση έγκαιρης διάγνωσης και θεραπείας, η θνησιμότητά της μειώνεται σε 8% [14]. Η διενέργεια της εξέτασης CTPA αποτελεί το χρυσό κανόνα για την διάγνωση της νόσου έναντι άλλων απεικονιστικών μεθόδων (π.χ πνευμονική αγγειογραφία), αφού προσφέρει παρόμοια διαγνωστική ακρίβεια, αποτελεί μη επεμβατική μέθοδο εξέτασης και είναι σύντομη προσφέροντας σε μικρό χρονικό διάστημα εικόνες υψηλής ανάλυσης [1,3]. Παρόλα αυτά, η διαδικασία της διάγνωσης είναι χρονοβόρα και η επιτυχία της εξαρτάται από την ειδίκευση του ακτινολόγου. Αυτό οδηγεί στη καθυστέρηση της διάγνωσης και στην αύξηση της πιθανότητας λάθους αντίστοιχα [15]. Μελέτη που περιελάβανε 241 ασθενείς με συμπτώματα ΠΕ, οι οποίοι εισήχθησαν στο τμήμα επειγόντων περιστατικών, έδειξε σημαντική συσχέτιση ανάμεσα στον χρόνο διάγνωσης της ΠΕ και στην πιθανότητα θανάτου εντός 30 ημερών. Διαπιστώθηκε, επίσης, ότι ασθενείς με χρόνο διάγνωσης μεγαλύτερο των 12 ωρών μετά το συμβάν της ΠΕ είχαν υψηλότερο ποσοστό θνησιμότητας εντός 30 ημερών συγκριτικά με ασθενείς που διαγνώστηκαν νωρίτερα. Αυτό το εύρημα υπογραμμίζει τη σημασία της έγκαιρης διάγνωσης της ΠΕ [16]. Η γρήγορη διάγνωση επιτρέπει την έγκαιρη έναρξη της κατάλληλης θεραπείας, η οποία μπορεί να βελτιώσει την πρόγνωση του ασθενούς και να μειώσει τον κίνδυνο θανάτου.

Επιπλέον, ο αριθμός εξετάσεων CTPA που διενεργούνται σε ώρες εφημερίας έχει αυξηθεί κατά 500% την περίοδο 2006-2020, με τις εξετάσεις που αφορούν τη διάγνωση ΠΕ να έχουν αυξηθεί σε ποσοστό 1360%. Όλα τα παραπάνω έχουν ως συνέπεια, ο φόρτος εργασίας συνεχώς να αυξάνεται γεγονός που αποτελεί ακόμη έναν παράγοντα που μπορεί να οδηγήσει σε λανθασμένη διάγνωση [17]. Παράλληλα, μία ακόμη μελέτη που εξέτασε τη συσχέτιση της διάρκειας των εφημεριών με την απόδοση των γνωματεύσεων έδειξε ότι όσο μεγαλύτερες είναι οι βάρδιες των ακτινολόγων, τόσο

αυξάνονται τα ποσοστά σφαλμάτων, με τα περισσότερα να συμβαίνουν προς το τέλος της βάρδιας, συνήθως μετά από 10-12 ώρες εργασίας [18].

Η χρήση Τεχνητής Νοημοσύνης (ΤΝ), μπορεί να συνεισφέρει στην αντιμετώπιση αυτών των προκλήσεων. μέσω της μείωσης του χρόνου αναμονής, της πιθανότητας λανθασμένης διάγνωσης και κατ' επέκταση της αποτελεσματικότερης διαχείρισης των περιστατικών ΠΕ. Πολλές είναι οι μελέτες οι οποίες έχουν αναδείξει την συμβολή των εφαρμογών ΤΝ στη μείωση του χρόνου αναμονής. Ειδικότερα, η εφαρμογή ενός συστήματος διαλογής ΤΝ για την ανίχνευση της ΠΕ σε εξετάσεις CTPA οδήγησε σε σημαντική βελτίωση του χρόνου που μεσολαβούσε από την ολοκλήρωση της εξέτασης έως την ερμηνεία της από τον ακτινολόγο [19]. Επιπλέον, ανέδειξε την αποτελεσματικότητα της χρήσης λογισμικού ΤΝ στον επαναπροσδιορισμό των προτεραιοτήτων στη λίστα εργασιών του ακτινολογικού τμήματος. Η χρήση της ΤΝ μείωσε το χρόνο αναμονής στο ένα τρίτο (15 λεπτά έναντι 44 λεπτών). Αυτό σημαίνει ότι η ΤΝ μπορεί να συμβάλει στην ταχύτερη εξέταση των CTPA σε ασθενείς που παρουσιάζουν μεγαλύτερη πιθανότητα εμφάνισης ΠΕ.

Ήδη έχουν διενεργηθεί ορισμένες μελέτες που προβάλλουν τη σημασία των αλγορίθμων ΤΝ στη μείωση των λανθασμένων διαγνώσεων ΠΕ. Μία από αυτές κατέδειξε ότι ο αλγόριθμος επέδειξε σημαντικά υψηλότερη διαγνωστική ακρίβεια στον εντοπισμό της ΠΕ σε CTPA πιθανών ασθενών, σε σύγκριση με την έκθεση του θεράποντα ακτινολόγου. Επιπλέον, τόνισε ότι ο συνδυασμός της εμπειρίας των ακτινολόγων με την ΤΝ μπορεί να οδηγήσει σε πιο ακριβείς διαγνώσεις στην κλινική πράξη [20].

Συμπερασματικά, η ΤΝ αποτελεί ένα χρήσιμο εργαλείο υποβοήθησης των ακτινολόγων, συμβάλλοντας στη βελτίωση της ακρίβειας και της ευαισθησίας των διαγνώσεών τους, προσδίδοντας τους μεγαλύτερη αυτοπεποίθηση στη διάγνωσή τους [20]. Η χρήση ΤΝ μπορεί να μειώσει τα ποσοστά των λανθασμένων αρνητικών αποτελεσμάτων, τα οποία αντιστοιχούν σε περιπτώσεις που η ΠΕ δεν εντοπίζεται από τον ακτινολόγο [20].

Η παρούσα εργασία, έχει ως στόχο την ανάπτυξη ενός μοντέλου τεχνητής νοημοσύνης, βασιζόμενο σε μοντέλα βαθιάς μάθησης, για την ταξινόμηση εξετάσεων CTPA σε θετικές ή μη, αναφορικά με το αν υπάρχουν ή όχι αντίστοιχα, ενδείξεις ΠΕ στην εξέταση. Στα πλαίσια της εργασίας , μελετήθηκε η χρήση της αρχιτεκτονικής του οπτικού μετασχηματιστική (Vision Transformers – ViT). Η αρχιτεκτονική που

παρουσιάστηκε το 2021 σε μια ερευνητική εργασία με τίτλο «An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale», που δημοσιεύτηκε στο International Conference on Learning Representations (ICLR) 2021, είναι εμπνευσμένη από την επιτυχία των μετασχηματιστών (transformers) στον τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP). Στο τέλος της εργασίας, αξιολογείται κατά πόσο η νέα αυτή αρχιτεκτονική μπορεί να ανταγωνιστεί την καλή απόδοση των κλασσικών αρχιτεκτονικών βαθιάς μάθησης στον τομέα της ταξινόμησης ιατρικών εικόνων και συγκεκριμένα στην εξέταση της ΠΕ.

2. Βιβλιογραφική Ανασκόπηση

Οι δυσκολίες στη διάγνωση της ΠΕ οδηγούν σε θνησιμότητα > 30% [71]. Η έγκαιρη διάγνωση και κατ' επέκταση η έναρξη φαρμακευτικής αγωγής περιορίζει τη θνησιμότητα στο 5% [72]. Η αξονική αγγειογραφία πνευμόνων CTPA είναι η πιο διαδεδομένη απεικονιστική εξέταση για τη διάγνωση της ΠΕ. Ωστόσο, όπως έχει ήδη αναφερθεί η αξιολόγησή της από τους ακτινολόγους μπορεί να είναι χρονοβόρα, γεγονός που επιβαρύνει την υγεία του ασθενή. Σε αυτό το σημείο η ανάπτυξη υπολογιστικών εργαλείων υποβοήθησης της ιατρικής διάγνωσης με αξιοποίηση τεχνολογιών ΤΝμπορεί να μειώσει σημαντικά το χρόνο διάγνωσης της ασθένειας. Για τον λόγο αυτό έχουν μελετηθεί διάφορες μέθοδοι που βασίζονται στη βαθιά μάθηση.

Οι πρώτες εφαρμογές ΤΝ βασίστηκαν σε νευρωνικά δίκτυα που λαμβάνουν σαν είσοδο εικόνες που προκύπτουν με τη μέθοδο του σπινθηρογραφήματος αιμάτωσης-αερισμού πνευμόνων. Η τελευταία αποτελεί απεικονιστική εξέταση για την ανίχνευση πνευμονολογικών παθήσεων, όπως είναι η ΠΕ [54,55,56,57]. Ωστόσο, αυτές οι προσεγγίσεις βασίστηκαν σε χειροκίνητη εξαγωγή χαρακτηριστικών από τις εικόνες, γεγονός που τις καθιστά ιδιαίτερα χρονοβόρες και υπολογιστικά περίπλοκες. Η πρώτη εφαρμογή μοντέλων βαθιάς μάθησης πραγματοποιήθηκε από τους Tajbakhsh *et al.* [58]. Συγκεκριμένα, χρησιμοποίησαν μοντέλο CNN για την ανίχνευση της ΠΕ από εξετάσεις CTPA. Η μέθοδός τους ξεπέρασε τις παραδοσιακές τεχνικές μηχανικής μάθησης, επιτυγχάνοντας την ανίχνευση μεμονωμένων εμβολών με ευαισθησία 0.83. Οι Xing Yang *et al.* [59] προκειμένου να ανιχνεύσουν την ΠΕ σε εξετάσεις CTPA επέλεξαν τη χρήση CNN σε δύο στάδια. Στο πρώτο στάδιο, εφάρμοσαν 3Δ CNN το οποίο ανιχνεύει ένα σύνολο κύβων (cubes) που είναι υποψήφιοι να περιέχουν ΠΕ. Το δεύτερο στάδιο, έχει στόχο την εξάλειψη των ψευδώς θετικών που προέκυψαν στο

πρώτο στάδιο. Για τον σκοπό αυτό, χρησιμοποίησαν το μοντέλο ResNet-18 στα υποψήφια τμήματα που προέκυψαν από το πρώτο στάδιο, επιτυγχάνοντας ευαισθησία 0.75. Ακόμα μία προσέγγιση που βασίστηκε σε 3Δ CNN δίκτυα αποτελεί το μοντέλο PENet, που πρότειναν οι Huang *et al.* [60]. Το μοντέλο αποτελείται από 77 στρώματα και έχει προεκπαιδευτεί στο σύνολο δεδομένων kinetics-600 [73]. Οι Huang et al εκπαίδευσαν περαιτέρω το μοντέλο με δεδομένα από εξετάσεις CTPA με τη μέθοδο fine-tuning. Επιπλέον, οι Rajan *et al.* [61] πρότειναν τη χρήση αρχιτεκτονικής που αποτελείται από δύο τμήματα. Στο πρώτο τμήμα αξιοποιείται το μοντέλο U-Net για την τμηματοποίηση της εικόνας εισόδου σε πιθανά τμήματα που περιέχουν ΠΕ. Στο δεύτερο τμήμα χρησιμοποιείται ένα μοντέλο συνελικτικού LSTM, το οποίο εκπαιδεύεται με την τεχνική Machine Instance Learning (MIL), προκειμένου να ανιχνεύσει την ΠΕ, επιτυγχάνοντας AUC 0.85. Ομοίως, και οι Suman et al [62] χρησιμοποίησαν δύο στάδια εκπαίδευσης προκειμένου να προβλέψουν την ανίχνευση της ΠΕ τόσο σε επίπεδο εικόνας όσο και σε επίπεδο εξέτασης. Χρησιμοποίησαν ένα μοντέλο CNN στο πρώτο στάδιο με σκοπό την εξαγωγή χαρακτηριστικών από τις εικόνες. Στο δεύτερο στάδιο χρησιμοποίησαν την τεχνική MIL.

Οι πιο επιτυχημένες αρχιτεκτονικές του διαγωνισμού Pulmonary Embolism Detection Challenge της Radiological Society of North America (RSNA) [49] υιοθετούν επίσης την προσέγγιση των δύο σταδίων εκπαίδευσης. Η εργασία που κέρδισε την πρώτη θέση χρησιμοποιεί ένα μοντέλο CNN για την πρόβλεψη της ΠΕ σε επίπεδο εικόνας. Στη συνέχεια, τα χαρακτηριστικά που προκύπτουν από το πρώτο στάδιο εισέρχονται σε ένα μοντέλο GRU για την πρόβλεψη της ΠΕ στο επίπεδο την εξέτασης.

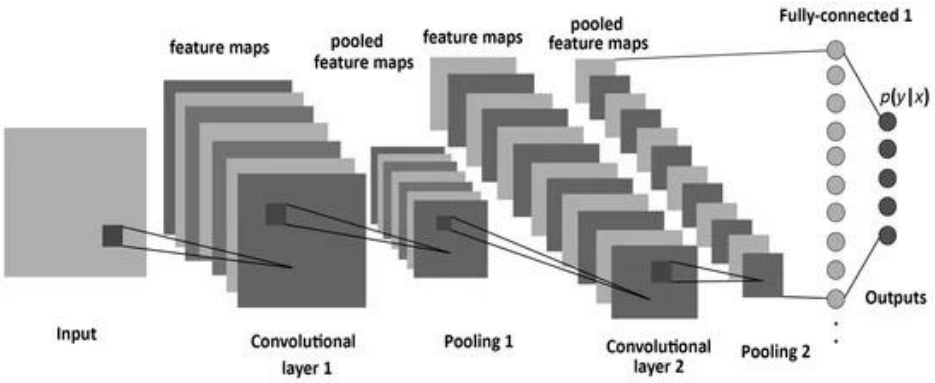
Χρησιμοποιώντας, το σύνολο δεδομένων RSNA Pulmonary Embolism οι Islam et al [63] συγκρίνουν την απόδοση διαφορετικών αρχιτεκτονικών CNN και ViT για την πρόβλεψη της ΠΕ σε επίπεδο εικόνας. Επιπλέον, διερευνούν την επίπτωση της αυτό-εποπτευόμενης μάθησης (self-supervised learning) έναντι της εποπτευόμενης μάθησης (supervised learning). Επιπλέον, εξετάζουν την επιρροή της μεταφοράς μάθησης (transfer learning) έναντι της εκπαίδευσης του μοντέλου από την αρχή (from scratch). Σε επίπεδο εξέτασης προτείνουν την αρχιτεκτονική E-ViT έναντι των παραδοσιακών μοντέλων RNN. Το E-ViT χρησιμοποιεί τις αναπαραστάσεις εικόνων από το πρώτο επίπεδο. Το μοντέλο βασίζεται στην αρχιτεκτονική ViT με τη διαφορά ότι δεν ενσωματώνει στις αναπαραστάσεις την πληροφορία της θέσης. Με τη χρήση

αναπαραστάσεων που προκύπτουν από τα μοντέλα Xception και SeResNext50 το μοντέλο E-ViT πετυχαίνει AUC 0.89. Οι αναπαραστάσεις από μοντέλα που βασίζονται στην αρχιτεκτονική του ViT αυξάνουν την τιμή της AUC σε 0.90στο επίπεδο της διάγνωσης της εξέτασης.

3. Θεωρητικό Υπόβαθρο

3.1 Συνελικτικά Νευρωνικά Δίκτυα

Η εφαρμογή νευρωνικών δικτύων σε δεδομένα εικόνας απαιτεί τον μετασχηματισμό της εικόνας εισόδου σε διάνυσμα. Με αυτό τον τρόπο, όμως, χάνεται η χωρική πληροφορία της εικόνας. Αντίθετα, τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks – CNN) περιλαμβάνουν στρώματα συνελίξεων (convolutional layers), στα οποία εφαρμόζονται φίλτρα σε όλη την εικόνα, λαμβάνοντας υπόψη τη χωρική πληροφορία. Στα πρώτα στρώματα συνέλιξης, εντοπίζονται βασικά μοτίβα της εικόνας, όπως για παράδειγμα οριζόντιες ή κάθετες ακμές. Σε υψηλότερα επίπεδα εντοπίζουν μοτίβα σε υψηλότερο επίπεδο αφαίρεσης, όπως για παράδειγμα αντικείμενα και πρόσωπα, ανάλογα με το πρόβλημα στο οποίο εκπαιδεύονται. Η ικανότητά τους να διαχειρίζονται αποτελεσματικά δεδομένα εικόνων, προσέδωσε στα CNN πρωταγωνιστικό ρόλο σε εφαρμογές οπτικής όρασης (computer vision).

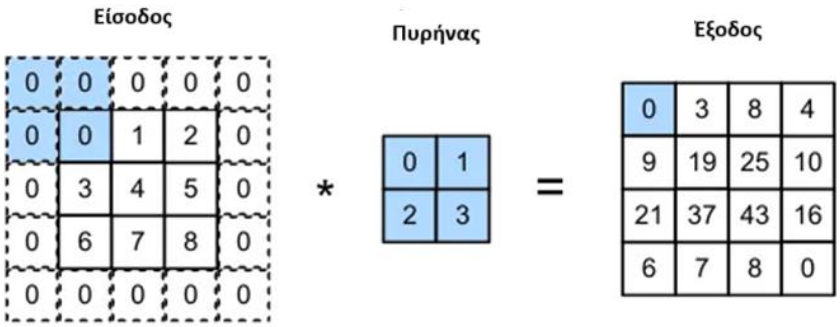


Εικόνα 1 - Δίκτυο CNN (Πηγή https://doi.org/10.3390/e19060242)

3.1.1 Στρώμα Συνέλιξης

Κάθε αρχιτεκτονική δικτύων CNN αποτελείται από στρώματα συνέλιξης και επίπεδα υποδειγματοληψίας (pooling layers). Το επίπεδο συνέλιξης είναι υπεύθυνο για την εξαγωγή χαρακτηριστικών από διάφορες περιοχές της εικόνας. Αποτελείται από ένα σύνολο πυρήνων συνέλιξης (convolutional kernels) ή φίλτρων συνέλιξης

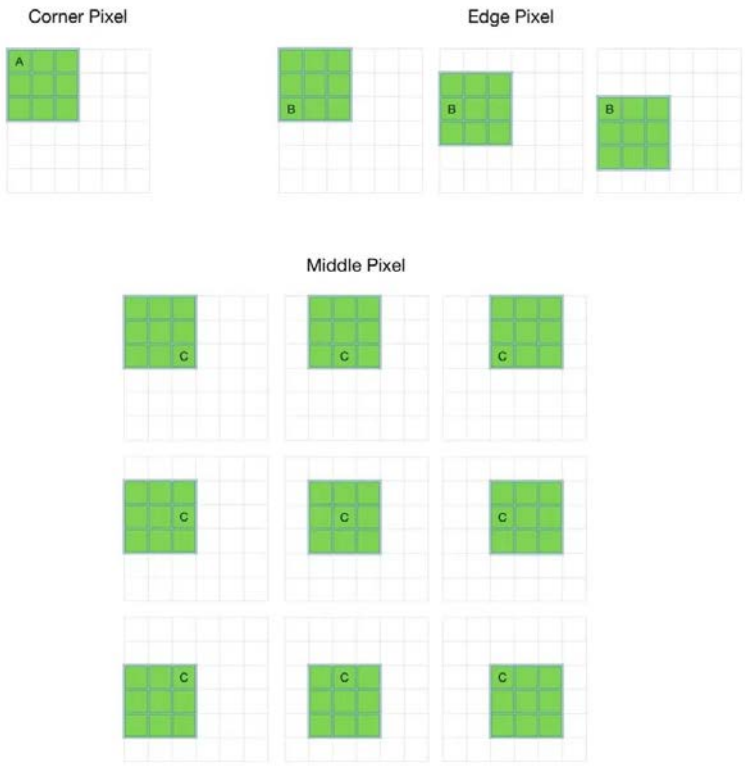
(convolutional filters) που εφαρμόζονται στην εικόνα εισόδου, προκειμένου να προκύψουν οι χάρτες ενεργοποίησης (activation maps) ή χάρτες χαρακτηριστικών (feature maps). Η είσοδος κάθε επιπέδου είναι η έξοδος του προηγούμενου, όπως παρουσιάζεται και στην Εικόνα 1. Ένα φίλτρο το οποίο έχει μέγεθος $F \times F$ και εφαρμόζεται σε είσοδο που περιέχει C κανάλια, αποτελεί μια οντότητα μεγέθους $F \times F \times C$ που εκτελεί την πράξη της συνέλιξης σε μία είσοδο μεγέθους $I \times I \times C$. Το αποτέλεσμα της συνέλιξης έχει μέγεθος $O \times O \times 1$ και όπως αναφέρθηκε ονομάζεται χάρτης χαρακτηριστικών. Το επίπεδο συνέλιξης περιέχει K φίλτρα τα οποία εφαρμόζονται στην είσοδο. Κατά αυτό τον τρόπο προκύπτουν $O \times O \times K$ χάρτες χαρακτηριστικών.



Εικόνα 2- Συνέλιξη με πυρήνα 2x2

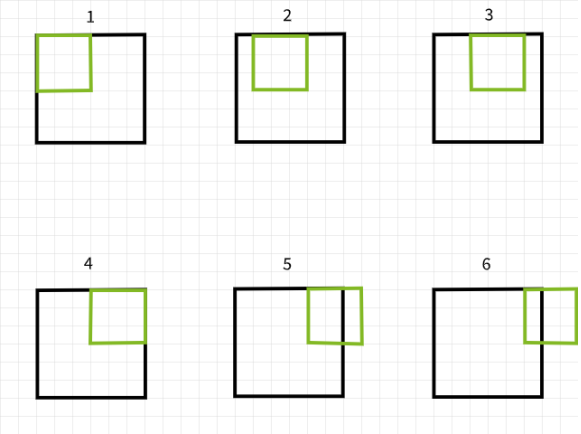
Για παράδειγμα, έστω ότι η είσοδος στο δίκτυο είναι μία ασπρόμαυρη εικόνα, η οποία περιέχει ένα κανάλι. Στο πρώτο επίπεδο συνέλιξης, κάθε πυρήνας ή φίλτρο έχει διάσταση $F \times F \times 1$. Κάθε φίλτρο θα διατρέξει ολόκληρη την εικόνα εκτελώντας σε κάθε βήμα την πράξη της συνέλιξης μεταξύ του φίλτρου και ενός τμήματος της εικόνας. Έστω ότι το παράδειγμα αναφέρεται στην Εικόνα 2. Η εικόνα εισόδου έχει μέγεθος 5×5 και στο πρώτο επίπεδο συνέλιξης εφαρμόζεται ένα φίλτρο 2×2. Σε κάθε βήμα υπολογίζεται το άθροισμα του κατά στοιχείο γινομένου μεταξύ των εικονοστοιχείων του τμήματος της εικόνας με τις αντίστοιχες τιμές του φίλτρου. Το αποτέλεσμα κάθε βήματος αποτελεί και ένα στοιχείο του χάρτη χαρακτηριστικών. Στην Εικόνα 2 , το πρώτο βήμα της συνέλιξης αποτυπώνεται με γαλάζιο χρώμα. Ο χάρτης χαρακτηριστικών, αποτελεί ουσιαστικά το αποτέλεσμα της συνέλιξης του φίλτρου με την εικόνα. Τελικά, για κάθε φίλτρο υπολογίζεται ο αντίστοιχος πίνακας χαρακτηριστικών. Κάθε στοιχείο του συνδέεται με ένα μόνο τμήμα της εικόνας εισόδου και αυτό ονομάζεται δεκτικό πεδίο (receptive field).

Στο παράδειγμα της Εικόνας 2 παρατηρείται η συμπλήρωση μηδενικών γύρω από τα όρια της εικόνας. Αυτή η τεχνική ονομάζεται συμπλήρωμα (padding) και συμβάλει στη μείωση της απώλειας της πληροφορίας που υπάρχει στα όριατης εικόνας εισόδου (αν πρόκειται για το επίπεδο εισόδου) ή των χαρτών χαρακτηριστικών (αν πρόκειται για κρυφό επίπεδο). Στην Εικόνα 3 εμφανίζεται ένα σχετικό παράδειγμα, όπου φαίνεται ότι το εικοστοιχείο *A* που βρίσκεται στη γωνία της εικόνας συμμετέχει σε μία μόνο συνέλιξη με το φίλτρο.



Εικόνα 3 - Πρόβλημα με τη πράξη της συνέλιξης (Πηγή https://www.geeksforgeeks.org/cnn-introduction-to-padding/)

Το εικονοστοιχείο *B*, που βρίσκεται στην άκρη λαμβάνεται υπόψη σε τρεις συνελίξεις, ενώ το εικονοστοιχείο *C*, που βρίσκεται στη μέση της εικόνας, συμμετέχει σε περισσότερες συνελίξεις. Συνεπώς, η πληροφορία από τα όρια της εικόνας φαίνεται ότι δεν διατηρείται κατά την συνέλιξη στο βαθμό που διατηρείται για τα ενδιάμεσα εικονοστοιχεία. Επιπλέον, υπάρχει ακόμη μία περίπτωση απώλειας της πληροφορίας. Κατά τη διαδικασία της συνέλιξης, το φίλτρο "ολισθαίνει" στην εικόνα, εξάγοντας χαρακτηριστικά. Ωστόσο, αν το φίλτρο υπερβεί τα όρια της εικόνας, όπως φαίνεται στην Εικόνα 4, τα εικονοστοιχεία σε αυτό το τμήμα της εικόνας δε θα χρησιμοποιηθούν στην συνέλιξη, οδηγώντας σε μικρότερο πίνακα χαρακτηριστικών.



Εικόνα 4 - Το φίλτρο καταλήγει εκτός των ορίων της εικόνας (Πηγή https://blog.paperspace.com/padding-in-convolutional-neural-networks/)

Η χρήση του padding σε αυτή την περίπτωση μπορεί να εξασφαλίσει ότι το φίλτρο παραμένει εντός των ορίων και όλα τα εικονοστοιχεία λαμβάνουν μέρος στη συνέλιξη. Το padding συνδέεται με μια υπερπαράμετρο (εύρος εικονοστοιχείων), η οποία ορίζεται από τον χρήστη.

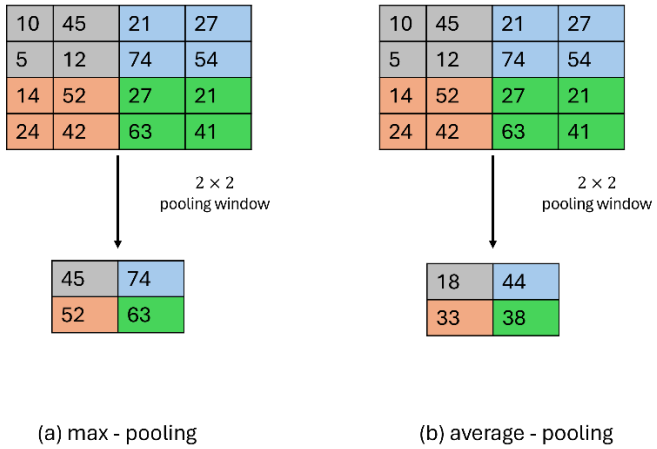
Άλλη μία υπερπαράμετρος που αφορά τη διαδικασία της συνέλιξης και πρέπει να οριστεί από το χρήστη αποτελεί η τιμή του διασκελισμού (stride). Η τιμή αυτή υποδηλώνει τον αριθμό των εικονοστοιχείων που διασχίζει το φίλτρο πριν από κάθε εφαρμογή του. Για παράδειγμα, όταν η τιμή είναι 1, το φίλτρο θα κινείται σε κάθε βήμα κατά ένα εικονοστοιχείο. Η χρήση μεγάλων τιμών stride οδηγεί σε μείωση της διάστασης του χάρτη χαρακτηριστικών. Επιπλέον, όσο μεγαλύτερη τιμή έχει το stride, τόσο λιγότερες είναι οι πράξεις συνέλιξης που θα εκτελέσει το φίλτρο με την εικόνα, γεγονός που μειώνει το υπολογιστικό κόστος. Παρόλα αυτά, το φίλτρο παραβλέπει ορισμένα εικονοστοιχεία με αποτέλεσμα να υπάρξει απώλεια πληροφορίας.

3.1.2 Στρώμα Δειγματοληψίας

Η κύρια λειτουργία του στρώματος δειγματοληψίας είναι η υπό-δειγματοληψία των χαρτών χαρακτηριστικών που προκύπτουν από το επίπεδο συνέλιξης. Συνήθως, μετά από κάθε επίπεδο συνέλιξης εφαρμόζεται ένα επίπεδο δειγματοληψίας. Στο επίπεδο αυτό, κάθε χάρτης χαρακτηριστικών υπό-δειγματοληπτείται με τέτοιο τρόπο ώστε να διατηρεί τη σημαντική πληροφορία των χαρτών χαρακτηριστικών. Κατά τη διαδικασία της δειγματοληψίας, οι χάρτες χαρακτηριστικών διαιρούνται σε τμήματα και εκτελείται σε κάθε ένα από αυτά η συνάρτηση δειγματοληψίας. Μέσω της δειγματοληψίας, παράγεται για κάθε τμήμα μια αντιπροσωπευτική τιμή. Όπως και στο επίπεδο

συνέλιξης έτσι και σε αυτή την περίπτωση, χρησιμοποιείται η τιμή stride και το μέγεθος του φίλτρου. Τα φίλτρα στο επίπεδο αυτό δεν περιέχουν βάρη και ορίζουν το παράθυρο στο οποίο θα επενεργήσει η συνάρτηση δειγματοληψίας. Οι πιο συχνές συναρτήσεις δειγματοληψίας είναι η max pooling και η average pooling.

Η χρήση της συνάρτησης max pooling είναι η συνάρτηση δειγματοληψίας κατά την οποία επιλέγεται το μεγαλύτερο στοιχείο από την περιοχή του χάρτη χαρακτηριστικών που επενεργεί το φίλτρο. Έτσι, η έξοδος μετά το επίπεδο max-pooling θα είναι ένας χάρτης χαρακτηριστικών που περιέχει τα πιο σημαντικά χαρακτηριστικά του προηγούμενου χάρτη χαρακτηριστικών. Η εφαρμογή average pooling λαμβάνει το μέσο όρο των στοιχείων του τμήματος του χάρτη χαρακτηριστικώνπου επενεργεί το φίλτρο. Η εφαρμογή των δύο αυτών συναρτήσεων φαίνονται στην Εικόνα 5.



Εικόνα 5 - (a) Αποτέλεσμα Max-pooling, (b) Αποτέλεσμα Average-pooling

Τα επίπεδα δειγματοληψίας συγκεντρώνουν την πιο σημαντική πληροφορία, μειώνοντας σημαντικά τον αριθμό των παραμέτρων στο δίκτυο και συμβάλουν με αυτό τον τρόπο στην αποφυγή της υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης.

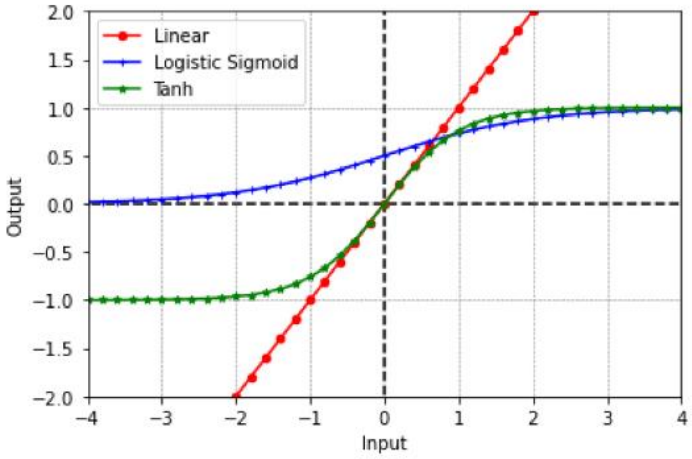
3.1.3 Συναρτήσεις Ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης (activation functions) αποτελούν σημαντικό στοιχείο στα δίκτυα CNN καθώς εισάγουν τη μη γραμμικότητα στο μοντέλο. Η μη γραμμικότητα είναι απαραίτητη στα δίκτυα βαθιάς μάθησης, καθώς οι εφαρμογές του πραγματικού κόσμου χαρακτηρίζονται από μεγάλη πολυπλοκότητα και ένα δίκτυο που χρησιμοποιεί μόνο γραμμικές σχέσεις δεν είναι ικανό να προσαρμόζεται σε αυτές της εφαρμογές. Η σιγμοειδής συνάρτηση (sigmoid function), η υπερβολική συνάρτηση

εφαπτομένης (tanh), η συνάρτηση SoftMax , συνάρτηση ReLU και η συνάρτηση Leaky ReLU είναι οι πιο διαδεδομένες συναρτήσεις ενεργοποίησης.

Σιγμοειδής Συνάρτηση

Η σιγμοειδής συνάρτηση, όπως φαίνεται και στην Εικόνα 6 (μπλε χρώμα) , λαμβάνει τιμές μεταξύ 0 και 1, καθιστώντας την ιδανική για δυαδικά προβλήματα ταξινόμησης. Παρόλα αυτά , μπορεί να οδηγήσει στο πρόβλημα της εξαφάνισης κλίσης (vanishing gradient) που συνεπάγεται τη μη ανανέωση των παραμέτρων κατά τη διάρκεια της εκπαίδευσης του μοντέλου, αποτρέποντάς το από το να εκπαιδευτεί. Επιπλέον, η συνάρτηση δεν είναι συμμετρική ως προς το 0 με αποτέλεσμα να οδηγεί σε ασταθή και δύσκολη εκπαίδευση [30].



Εικόνα 6- Συναρτήσεις ενεργοποίησης [30]

Συνάρτηση Υπερβολικής Εφαπτομένης

Παρόμοια με τη σιγμοειδή συνάρτηση, η συνάρτηση υπερβολικής εφαπτομένης, η οποία απεικονίζεται στην Εικόνα 6 (πράσινο χρώμα), λαμβάνει τιμές μεταξύ -1 και 1, παρόλα αυτά είναι συμμετρική ως προς το 0. Το πρόβλημα της εξαφάνισης της κλίσης και της υπολογιστικής πολυπλοκότητας εμφανίζεται και στη συνάρτηση tanh [30].

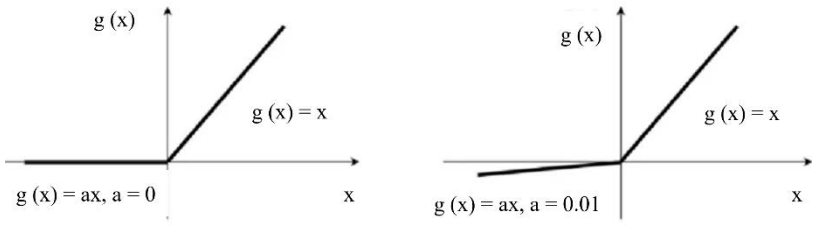
Συνάρτηση SoftMax

Η συνάρτηση SoftMax χρησιμοποιείται σε προβλήματα όπου απαιτείται ταξινόμηση σε πολλές διαφορετικές κλάσεις (multi-class classification) [30].

Εφαρμόζεται στο διάνυσμα που προκύπτει από το επίπεδο εξόδου και μετατρέπει τις τιμές σε πιθανότητες που έχουν άθροισμα τη μονάδα. Σε ένα πρόβλημα ταξινόμησης, δηλαδή, οι τιμές που προκύπτουν μετά την εφαρμογή της συνάρτησης περιγράφουν την πιθανότητα πρόβλεψης για κάθε κατηγορία. Η συνάρτηση SoftMax αποτελεί επέκταση της σιγμοειδούς συνάρτησης. Όταν ο αριθμός των κατηγοριών είναι δύο, τότε η συνάρτηση SoftMax ταυτίζεται με τη σιγμοειδή συνάρτηση.

Ανορθωμένη Γραμμική Μονάδα

Η συνάρτηση ανορθωμένης γραμμικής μονάδας (Rectified Linear Unit – ReLU) αντιμετωπίζει τα προβλήματα της σιγμοειδούς συνάρτησης και της υπερβολικής εφαπτομένης. Όταν η είσοδος είναι θετική τότε η παράγωγος ισούται με 1 και αυτό βελτιώνει το πρόβλημα της εξαφάνισης κλίσης, με αποτέλεσμα η εκπαίδευση να συγκλίνει πιο γρήγορα. Επιπλέον, καθώς περιλαμβάνει μόνο γραμμικές πράξεις είναι υπολογιστικά πιο αποδοτική από τις άλλες δύο συναρτήσεις. Ωστόσο, όταν η είσοδος είναι αρνητική τότε η παράγωγος είναι 0 δυσχεραίνοντας τη διαδικασία της εκπαίδευσης. Το φαινόμενο αυτό ονομάζεται “Dying ReLU” [67]. Μία εναλλακτική είναι η χρήση της συνάρτησης LeakyReLU [67]. Η γραφική παράσταση της συνάρτησης ενεργοποίησης ReLU φαίνεται Εικόνα 7 (αριστερά).



Εικόνα 7- (Αριστερά) Συνάρτηση ReLU , (Δεξιά) Συνάρτηση LeakyReLU

Συνάρτηση LeakyReLU

Η συνάρτηση LeakyReLU προσπαθεί να αντιμετωπίσει το πρόβλημα που δημιουργεί η ReLU με την προσθήκη μικρής θετικής κλίσης στις αρνητικές τιμές εισόδου, όπως φαίνεται και από τη γραφική της παράσταση που αποτυπώνεται στην Εικόνα 7 (δεξιά). Με αυτό τον τρόπο δίνει μία τιμή στην παράγωγο και επιτρέπει την οπισθοδιάδοση (backpropagation) ακόμα και για αρνητικές τιμές εισόδου. Παρόλα

αυτά δεν έχει αποδειχθεί πλήρως ότι η συνάρτηση LeakyReLU είναι πάντα αποδοτικότερη από τη συνάρτηση ReLU.

3.1.4 Πλήρες Συνδεδεμένο Στρώμα

Το πλήρες συνδεδεμένο στρώμα (fully connected layer) τοποθετείται στο τέλος του δικτύου και είναι υπεύθυνο για τη διαδικασία της ταξινόμησης. Σε ένα πλήρες συνδεδεμένο στρώμα, κάθε νευρώνας συνδέεται με όλους τους νευρώνες του προηγούμενου στρώματος. Η έξοδος από το τελευταίο pooling layer, είναι χάρτες χαρακτηριστικών που αποτυπώνουν τα υψηλά χαρακτηριστικά που έχουν προκύψει από την εικόνα εισόδου. Οι τιμές των χαρτών χαρακτηριστικών «απλώνονται» (flatten) σε ένα διάνυσμα το οποίο στη συνέχεια τροφοδοτείται στο πλήρες συνδεδεμένο στρώμα. Το στρώμα αυτό «ενώνει» την έξοδο από το τελευταίο pooling layer με το στρώμα εξόδου. Το πλήθος νευρώνων του στρώματος εξόδου ισούται με τον αριθμό των κατηγοριών, αν πρόκειται για πρόβλημα ταξινόμησης και παράγει την έξοδο του δικτύου. Συνήθως, η έξοδος αυτή ακολουθείται από μία συνάρτηση ενεργοποίησης προκειμένου να παραχθούν πιθανότητες. Για παράδειγμα, στο πρόβλημα ταξινόμησης σε πολλές κατηγορίες επιλέγεται η χρήση της συνάρτησης SoftMax.

3.1.5 Στρώμα Απόρριψης

Η τεχνική του στρώματος απόρριψης (dropout) χρησιμοποιείται για την αντιμετώπιση της υπερπροσαρμογής (overfitting) στα δεδομένα εκπαίδευσης [21]. Στις περισσότερες περιπτώσεις εφαρμόζεται στο πλήρες συνδεδεμένο στρώμα ενισχύοντας την απόδοση των δικτύων [21,22,23]. Κατά τη διαδικασία της εκπαίδευσης, επιλέγεται, με τυχαίο τρόπο, η απενεργοποίηση ενός ποσοστού νευρώνων, με μία προκαθορισμένη πιθανότητα. Η πιθανότητα ορίζεται από το χρήστη και συνεπώς αποτελεί υπερπαράμετρο του δικτύου. Η απενεργοποίηση των νευρώνων επιτυγχάνεται με το μηδενισμό των εξόδων τους. Αυτό σημαίνει ότι οι νευρώνες που απενεργοποιήθηκαν δεν θα ληφθούν υπόψη στα επόμενα επίπεδα. Σε κάθε πρόσθιο πέρασμα (forward pass) επιλέγεται διαφορετικός αριθμός νευρώνων που θα απορριφθεί. Έτσι, οι νευρώνες δεν βασίζονται πάντα σε συγκεκριμένες συνδέσεις για την παραγωγή της εξόδου, μαθαίνοντας με αυτό τον τρόπο πιο γενικές αναπαραστάσεις.

3.1.6 Κανονικοποίηση Ομάδας

Είναι ήδη γνωστό ότι η κανονικοποίηση των δεδομένων εισόδου συμβάλει στην ταχύτερη και αποδοτικότερη εκπαίδευση των μοντέλων νευρωνικών δικτύων. Για αυτό το λόγο αποτελεί απαραίτητο βήμα της προεπεξεργασίας των δεδομένων πριν την τροφοδότησή ενός νευρωνικού δικτύου. Η κανονικοποίηση των δεδομένων εισόδου εξαλείφει τις διαφορές της κλίμακας μεταξύ διαφορετικών χαρακτηριστικών, βελτιώνοντας έτσι τη διαδικασία μάθησης [74]. Η κατανομή της εισόδου σε κάθε επίπεδο αλλάζει επειδή οι παράμετροι στο προηγούμενο επίπεδο μεταβάλλονται. Το φαινόμενο αυτό ονομάζεται εσωτερική μεταβλητή μετατόπιση (internal covariate shift) και καθυστερεί σημαντικά την διαδικασία μάθησης καθώς τα επίπεδα, κάθε φορά, πρέπει να αναπροσαρμόζονται στην καινούρια κατανομή. Απαιτεί επίσης τη χρήση χαμηλού ρυθμού μάθησης (learning rate) και πιο προσεκτική αρχικοποίηση παραμέτρων [74]. Ο περιορισμός του φαινομένου της εσωτερικής μεταβλητής μετατόπισης θα επιταχύνει τη διαδικασία της εκπαίδευσης.

Για το λόγο αυτό, προτάθηκε η κανονικοποίηση της εισόδου των επιπέδων στο δίκτυο. Τα δεδομένα εκπαίδευσης τροφοδοτούνται στο δίκτυο σε ομάδες (batch). Η κανονικοποίηση εφαρμόζεται σε κάθε ομάδα για αυτό και ονομάστηκε κανονικοποίηση ομάδας και διορθώνει τους μέσους και τις αποκλίσεις των εισόδων των στρωμάτων.

Κάθε στρώμα νευρώνων επεξεργάζεται μία ομάδα δεδομένων κάθε φορά. Πριν από την είσοδο των δεδομένων στο στρώμα, εφαρμόζεται η κανονικοποίηση τους. Η κανονικοποίηση εφαρμόζεται στην είσοδο κάθε νευρώνα ξεχωριστά. Υπολογίζεται η μέση τιμή και η τιμή της διακύμανσης για τις εισόδους σε κάθε νευρώνα του στρώματος. Στη συνέχεια, οι τιμές εισόδου κανονικοποιούνται αφαιρώντας από κάθε τιμή την αντίστοιχη μέση τιμή κι διαιρώντας με τη διακύμανση. Κατά αυτό τον τρόπο, η είσοδος που θα λάβουν οι νευρώνες στο επίπεδο έχει μέση τιμή 0 και διακύμανση 1.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Εικόνα 8 - Αλγόριθμος κανονικοποίησης παρτίδας που εφαρμόζεται σε μία τιμή ενεργοποίησης x κατά μήκος μίας παρτίδας δεδομένων [74]

Παρόλα αυτά, η εφαρμογή της κανονικοποίησης με μέση τιμή 0 και διακύμανση 1 μπορεί να αποτρέψει το εκάστοτε επίπεδο να μάθει αποτελεσματικά την πραγματική κατανομή των δεδομένων. Για το λόγο αυτό, όπως φαίνεται και στην Εικόνα 8, η κανονικοποιημένη τιμή x_i πολλαπλασιάζεται με την παράμετρο γ και στο αποτέλεσμα προστίθεται η παράμετρος β. Οι παράμετροι αυτοί κλιμακώνουν και μετακινούν τις κανονικοποιημένες τιμές αντίστοιχα. Κατά την διάρκεια της εκπαίδευσης το μοντέλο μαθαίνει την τιμή των παραμέτρων αυτών με τέτοιο τρόπο ώστε να ελαχιστοποιήσει την συνάρτηση απώλειας (loss function). Το μοντέλο, δηλαδή μαθαίνει ποια κατανομή ταιριάζει καλύτερα στα δεδομένα. Για παράδειγμα, αν η τιμή $\gamma = sqrt(var(x))$ και $\beta = mean(x)$, τότε δεν εφαρμόζεται κανονικοποίηση στα δεδομένα αλλά διατηρείται η αρχική κατανομή που υπήρχε στα δεδομένα εισόδου.

Στο κεφάλαιο αυτό αναλύθηκε ο τρόπος λειτουργίας των συνελικτικών νευρωνικών δικτύων. Περιεγράφηκε η βασική αρχιτεκτονική των δικτύων αυτών ώστε να γίνουν κατανοητοί οι μηχανισμοί που χρησιμοποιεί για την αποτελεσματική επεξεργασία των εικόνων. Η προσοχή της επιστημονικής κοινότητας στράφηκε στην αξιοποίηση των δικτύων αυτών, μετά την επιτυχία του μοντέλου AlexNet , που πέτυχε αξιοσημείωτη απόδοση κερδίζοντας το διαγωνισμό ImageNet Large Scale Visual Recognition Competition (ILSVRC) το 2012 [23]. Πληθώρα νέων αρχιτεκτονικών συνελικτικών δικτύων προσπάθησαν τη βελτίωση του μοντέλου AlexNet προκειμένου να επιτύχουν ακόμα καλύτερα αποτελέσματα. Τα ποιο αντιπροσωπευτικά μοντέλα είναι το μοντέλο VGG (Visual Geometry Group), το μοντέλο GoogleNet, γνωστό ως InceptionNet, που χρησιμοποίει inception module και global average pooling που το επιτρέπει να μάθει χαρακτηριστικά διαφορετικής κλίμακάς ταυτόχρονα, πετυχαίνοντας αξιοσημείωτα αποτελέσματα σε εφαρμογές ταξινόμησης εικόνας. Επιπλέον, πολύ σημαντική είναι η

συνεισφορά του μοντέλου ResNet το οποίο χαρακτηρίζεται από τη χρήση υπολειπόμενων συνδέσεων (residual connections) καταφέρνοντας με αυτό τον τρόπο να αποκτήσει την ικανότητα εκπαίδευσης νευρωνικών δικτύων με πολλά στρώματα χωρίς να γίνεται υπερπροσαρμογή [68].

Τα μοντέλα αυτά διαδραματίζουν σημαντικό ρόλο σε εφαρμογές υπολογιστικής όρασης όπως η ταξινόμησης εικόνας, η ανίχνευση αντικειμένων, η κατάτμηση εικόνας και η ανάλυση βίντεο. Αυτό δείχνει ότι έχουν την ικανότητα να ταυτοποιούν την τοποθεσία ενός αντικειμένου σε μία εικόνα, να ανιχνεύουν και να αναγνωρίζουν τα διάφορα αντικείμενα, να αναλύουν οντότητες και καταστάσεις σε δεδομένα βίντεο, για παράδειγμα την ανίχνευση της κίνησης [68] .

3.2 Αναδρομικά Νευρωνικά Δίκτυα

Τα δίκτυα πρόσθιας διάδοσης διαχειρίζονται τα δεδομένα με ανεξάρτητο τρόπο. Η πρόβλεψη των τιμών δεν λαμβάνει υπόψη τα προηγούμενα δεδομένα που επεξεργάστηκε η αρχιτεκτονική. Το γεγονός αυτό δεν διευκολύνει την επεξεργασία ακολουθιακών δεδομένων, δηλαδή δεδομένων που σχετίζονται με τον χρόνο. Για παράδειγμα, η χρήση ενός απλού νευρωνικού δικτύου πρόσθιας διάδοσης σε εφαρμογή μετάφρασης μίας πρότασης, δεν θα ήταν αποδοτική. Αυτό γιατί για τη μετάφραση της πρότασης πρέπει να ληφθεί υπόψη ολόκληρη η ακολουθία εισόδου. Η αδυναμία αυτή, οδήγησε την ερευνητική κοινότητα στην αναζήτηση μοντέλων που είναι ικανά να λαμβάνουν υπόψη προηγούμενες καταστάσεις για την πρόβλεψή τους. Έτσι, προτάθηκε η αρχιτεκτονική των αναδρομικών νευρωνικών δικτύων (Recurrent Neural Networks – RNN) το 1986 [24]. Η αρχιτεκτονική αυτή συνιστά μία κατηγορία νευρωνικών δικτύων που περιέχει μηχανισμό «μνήμης», ώστε να λαμβάνει υπόψη την ακολουθιακή σχέση των δεδομένων εισόδου.

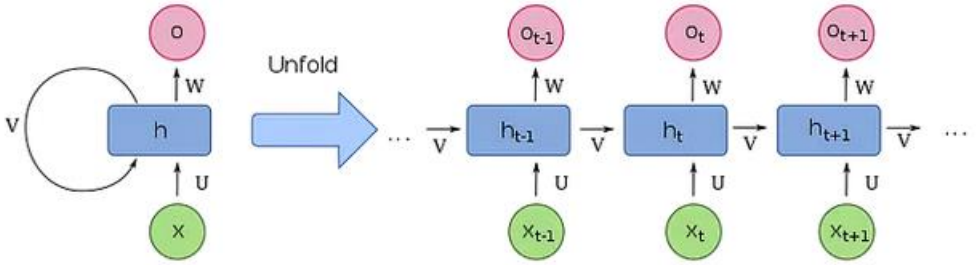
Παρόλα αυτά, τα RNNs παρουσιάζουν ένα σημαντικό πρόβλημα. Παρατηρήθηκε ότι τα μοντέλα αυτά δεν είναι αποτελεσματικά στην εκμάθηση μακροπρόθεσμων εξαρτήσεων στα δεδομένα εισόδου. Το πρόβλημα αυτό εντείνεται όσο η ακολουθία εισόδου μεγαλώνει. Αυτό συμβαίνει λόγω της εξαφάνισης της κλίσης [25]. Για το λόγο αυτό προτάθηκαν τα δίκτυα μακράς - βραχείας μνήμης (Long – Short Term Memory - LSTM). Τα δίκτυα αυτά αποτελούν εξέλιξη των δικτύων RNN και μπορούν να διαχειρίζονται αποτελεσματικά βραχυπρόθεσμες και μακροπρόθεσμες εξαρτήσεις της ακολουθίας εισόδου [26].

Οι αρχιτεκτονικές RNN και LSTM επεξεργάζονται την ακολουθία εισόδου από την αρχή προς το τέλος. Σε κάθε βήμα, δηλαδή, λαμβάνουν την πληροφορία από τα προηγούμενα βήματα. Υπάρχουν, όμως, εφαρμογές όπου η αμφίδρομη επεξεργασία ολόκληρης της ακολουθίας εισόδου πρέπει να ληφθεί υπόψη για τη καλύτερη κατανόηση της ακολουθίας εισόδου. Τα δίκτυα που αναπτύχθηκαν, για την εφαρμογή σε τέτοιου είδους προβλήματα ονομάστηκαν αμφίδρομα αναδρομικά νευρωνικά δίκτυα (Bidirectional Recurrent Neural Networks -BiRNNs) [27]. Η επιλογή ανάμεσα στην απλή αρχιτεκτονική RNN και στην αμφίδρομη BiRNN εξαρτάται από τη φύση της εκάστοτε εφαρμογής και τον τύπο των δεδομένων. Σε εφαρμογές που η πρόβλεψη στηρίζεται αποκλειστικά σε δεδομένα του παρελθόντος ή όταν δεν είναι γνωστή η ακολουθία στο μέλλον, τότε απαιτείται η χρήση του απλού μοντέλου RNN. Στην περίπτωση, όμως, που η πληροφορία διαχέεται σε όλο το μήκος της ακολουθίας εισόδου τότε η χρήση του αμφίδρομου μοντέλου BiRNN θα αποδώσει καλύτερα. Ανάλογη λογική ακολουθήθηκε και στην περίπτωση των αρχιτεκτονικών LSTM, όπου το 2005 προτάθηκε η αρχιτεκτονική του αμφίδρομου LSTM (Bidirectional LSTM-BiLSTM) [28].

Στα επόμενα κεφάλαιο πρόκειται να αναλυθεί ο τρόπος λειτουργίας των παραπάνω αρχιτεκτονικών.

3.3 Αναδρομικά Νευρωνικά Δίκτυα

Τα δίκτυα RNN διατηρούν μία κρυφή κατάσταση (hidden state), η οποία τροφοδοτείται με πληροφορίες από τις προηγούμενες καταστάσεις. Στην παρακάτω εικόνα αποτυπώνεται το μοντέλο RNN (αριστερά). Για τις ανάγκες επεξήγησης του τρόπου λειτουργίας του, δεξιά της εικόνας παρουσιάζεται η «ξεδιπλωμένη» μορφή του μοντέλου (unfold RNN). Αυτό προσφέρει καλύτερη απεικόνιση του τρόπου που επεξεργάζεται το μοντέλο μια ακολουθία εισόδου σε κάθε βήμα *t*. Η μορφή αυτή προκύπτει με την εφαρμογή σε κάθε βήμα *t*, του αντιγράφου του μοντέλου.



Εικόνα 9 - Αρχιτεκτονική RNN (Πηγή https://towardsdatascience.com/getting-started-with-recurrent-neural-network-rnns-ad1791206412)

Έστω η ακολουθία εισόδου $x = (x_0, x_1, \dots, x_t)$, όπου $x_t \in \mathbb{R}^d$, η έξοδος $O_t \in \mathbb{R}^q$ και η κρυφή κατάσταση $h_t \in \mathbb{R}^P$ στο βήμα *t*. Το μοντέλο σε κάθε βήμα *t* δέχεται σαν είσοδο το στοιχείο x_t της ακολουθίας εισόδου και την προηγούμενη κρυφή κατάσταση h_{t-1}. Στις δύο αυτές εισόδους εφαρμόζονται οι πίνακες βαρών $U \in \mathbb{R}^{p \times d}$ και $V \in \mathbb{R}^{p_x p}$ αντίστοιχα. Έτσι, ο πίνακας *U* αποτελεί τον πίνακα βαρών μεταξύ των εισόδων και των κρυφών καταστάσεων και ο πίνακας *V* αποτελεί τον πίνακα βαρών μεταξύ των κρυφών καταστάσεων. Υπάρχει ακόμα ένας πίνακας $W \in \mathbb{R}^{qxp}$ που αποτελεί τον πίνακα βαρών μεταξύ των κρυφών καταστάσεων και των εξόδων. Οι πίνακες *U*, *V*, *W* είναι κοινοί σε κάθε βήμα της ακολουθίας (parameter sharing). Έτσι, η κρυφή κατάσταση σε κάθε βήμα υπολογίζεται σύμφωνα με τη σχέση:

$$h_t = f_h(Vh_{t-1} + Ux_t + b_i), b_i \in \mathbb{R}^p \quad (1)$$

όπου η συνάρτηση $f_h(.)$ δηλώνει την συνάρτηση ενεργοποίησης κάθε κρυφής κατάστασης. Οι πιο συνήθεις συναρτήσεις είναι η υπερβολική εφαπτομένη (tanh) και η σιγμοειδής συνάρτηση (sigmoid function). Η έξοδος σε κάθε βήμα *t* υπολογίζεται με τη χρήση της κρυφής κατάστασης h_t, σύμφωνα με τη σχέση 2. Η είσοδος x_{t-1} στο βήμα $t-1$ επηρεάζει την έξοδο O_t στο βήμα *t* (και τα επόμενα βήματα) με τη χρήση των αναδρομικών συνδέσεων:

$$o_t = Vh_t + by, b_y \in \mathbb{R}^q \quad (2)$$

Στη σχέση (2) μπορεί να εφαρμοστεί οποιαδήποτε συνάρτηση ενεργοποίησης f_o. Για παράδειγμα, η συνάρτηση ενεργοποίησης f_o μπορεί να είναι η συνάρτηση SoftMax σε περίπτωση που η έξοδος χρησιμοποιηθεί ως επίπεδο ταξινόμησης. Παρόλα αυτά δεν είναι υποχρεωτική η χρήση κάθε εξόδου σε κάθε *t*. Σε πολλές εφαρμογές

επιλέγεται η χρήση μόνο της τελευταίας εξόδου στο βήμα T, η οποία περιέχει την πληροφορία ολόκληρης της ακολουθίας.

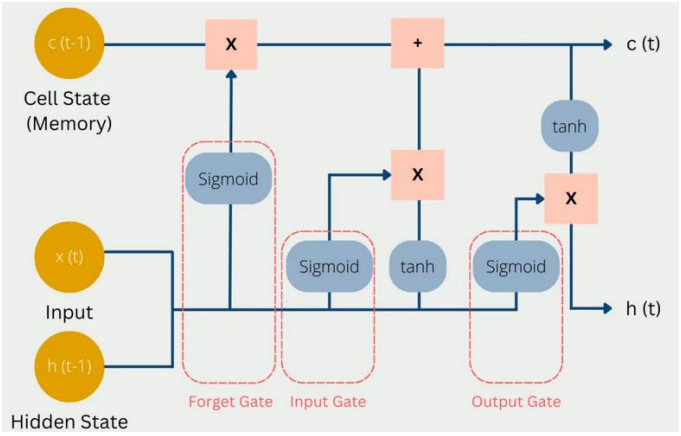
Ο υπολογισμός της κρυφής κατάστασης σε κάθε βήμα απαιτεί τον υπολογισμό όλων των προηγούμενων βημάτων. Με αυτόν τον τρόπο, κάθε κρυφή κατάσταση h_t περιέχει την πληροφορία όλων των προηγούμενων βημάτων, μέχρι το βήμα t. Έτσι, η κρυφή κατάσταση προσδίδει «μνήμη» στο δικτύου. Η έξοδος σε κάθε βήμα παράγεται σύμφωνα με την κατάσταση του δικτύου στο συγκεκριμένο βήμα.

Στην Εικόνα 9 , φαίνεται η συγκεκριμένη διάταξη δικτύου να παράγει την έξοδο σε κάθε βήμα. Ωστόσο, σε πολλές εφαρμογές επιλέγεται η αξιοποίηση μόνο της εξόδου που παράγεται στο τελευταίο βήμα O_T, η οποία περιέχει την πληροφορία που έχει εξάγει το δίκτυο από την παρατήρηση της ακολουθίας εισόδου σε ολόκληρο το μήκος της. Τα δίκτυα RNN σχεδιάστηκαν με τέτοιο τρόπο ώστε να αξιοποιούν το μηχανισμό της «μνήμης» προκειμένου να επεξεργάζονται ακολουθιακά δεδομένα. Αυτό αποτελεί και το βασικότερο χαρακτηριστικό που τα ξεχωρίζει από τα υπόλοιπα νευρωνικά δίκτυα εμπρόσθιας διάδοσης. Στην πράξη, όμως, η μνήμη αυτών των δικτύων είναι περιορισμένη. Αυτό σημαίνει ότι ειδικότερα σε μεγάλου μήκους ακολουθίες αδυνατούν να διατηρήσουν την πληροφορία από τα αρχικά βήματα της ακολουθίας. Το πρόβλημα αυτό προκαλείται καθώς τα δίκτυα αυτά, κατά τη διαδικασία της εκπαίδευσής τους, τείνουν να αντιμετωπίσουν το πρόβλημα της εξαφάνισης της κλίσης

3.4 Δίκτυα Μακράς - Βραχείας Μνήμης

Η αδυναμία των δικτύων RNN οδήγησε στην δημιουργία των δικτύων μακράς – βραχείας μνήμης (Long - Short Term Memory -LSTM) [26]. Τα δίκτυα LSTM ανήκουν στην κατηγορία των αναδρομικών νευρωνικών δικτύων και έχουν την ικανότητα να διαχειρίζονται αποτελεσματικά τις μακροπρόθεσμες εξαρτήσεις μεταξύ των στοιχείων μίας ακολουθίας. Τα δίκτυα αυτά αποτελούνται από ένα κελί μνήμης (memory cell) και από τις πύλες (gates) που τα βοηθούν να διατηρούν ή να αποβάλουν επιλεκτικά την πληροφορία που χρειάζονται. Για το σκοπό αυτό, υπάρχει η κατάσταση κελιού (cell state). Επιπλέον, διατηρούν την κρυφή κατάσταση (hidden state), η οποία αποτελεί τη βραχυπρόθεσμη μνήμη (short-term memory) και περιέχει την πληροφορία από την επεξεργασία στα προηγούμενα βήματα. Με αυτή την αρχιτεκτονική, τα δίκτυα LSTM αποτρέπουν την εξαφάνιση ή την έκρηξη της κλίσης και ανιχνεύουν τις μακροπρόθεσμες εξαρτήσεις.

Τα δίκτυα LSTM αποτελούνται από μία σειρά κελιών LSTM. Κάθε κελί αποτελείται από ένα σύνολο πυλών οι οποίες ελέγχουν τη ροή της πληροφορίας που εισέρχεται στο κελί και που εξέρχεται. Στην εικόνα που ακολουθεί αποτυπώνεται το κελί LSTM και επισημαίνεται ο τρόπος λειτουργίας των πυλών και διαφαίνεται η κατάσταση κελιού και η κρυφή κατάσταση.



Εικόνα 10 - Αρχιτεκτονική LSTM (Πηγή https://medium.com/@vanamayaswanth/naked-data-science-day-36-lstm-and-code-example-f54f3f226688)

Σε κάθε βήμα *t*, το LSTM δέχεται το στοιχείο εισόδου $x_t \in \mathbb{R}^d$ της ακολουθίας εισόδου *x*, την κρυφή κατάσταση $h_{t-1} \in \mathbb{R}^p$ και την κατάσταση κελιού $c_{t-1} \in \mathbb{R}^p$, του προηγούμενου βήματος, προκειμένου να υπολογίσει την επόμενη κατάσταση κελιού και την κρυφή κατάσταση. Πριν τον υπολογισμό των καταστάσεων είναι απαραίτητη πρώτα η επεξεργασία της πληροφορίας εισόδου από τις πύλες.

Πύλη Απόρριψης

Η πύλη απόρριψης (forget gate) είναι υπεύθυνη να αποφασίσει το μέρος της πληροφορίας από την κατάσταση κελιού που θα απορριφθεί. Στην προηγούμενη κρυφή κατάσταση h_{t-1} και τη τρέχουσα είσοδο x_t εφαρμόζεται η σιγμοειδής συνάρτηση ενεργοποίησης, η οποία λαμβάνει τιμές μεταξύ 0 και 1. Το αποτέλεσμα της συνάρτησης ενεργοποίησης εφαρμόζεται στην προηγούμενη κατάσταση κελιού, όπως θα αναλυθεί αργότερα, και υποδηλώνει ποια πληροφορία πρέπει να απορριφθεί. Με άλλα λόγια, η πύλη απόρριψης υποδηλώνει το ποσοστό της μακροπρόθεσμης μνήμης που το δίκτυο επιλέγει να θυμάται στο βήμα *t*. Η έξοδος της πύλης απόρριψης περιγράφεται από την παρακάτω σχέση (3):

$$f_t = \sigma\big(W_f[h_{t-1}, x_t] + bf\big) \quad (3)$$

Πύλη Εισόδου

Στην πύλη εισόδου (input gate) λαμβάνεται υπόψη η τρέχουσα είσοδος και η προηγούμενη κρυφή κατάσταση. Η πύλη εισόδου αποφασίζει τη νέα πληροφορία που θα προστεθεί στην κατάσταση κελιού, δηλαδή τη μακροπρόθεσμη μνήμη. Για τον σκοπό αυτό, με τη βοήθεια της συνάρτησης της υπερβολικής εφαπτομένης δημιουργείται ένα διάνυσμα $\tilde{C}_t$ (σχέση (5)) το οποίο περιέχει τη νέα υποψήφια πληροφορία. Επιπλέον, γίνεται η χρήση της σιγμοειδούς συνάρτησης η οποία θα αποφασίσει ποια πληροφορία και σε τι ποσοστό θα περάσει στην μακροπρόθεσμη μνήμη. Η έξοδος της πύλης εισόδου δίνεται από τη σχέση (4).

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \text{ (4)}$$
$$\tilde{c} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \text{ (5)}$$

Πύλη Εξόδου

Η πύλη εξόδου (output gate) δέχεται σαν είσοδο την τρέχουσα είσοδο x_t και την προηγούμενη κρυφή κατάσταση h_{t-1}. Παράγει την έξοδο στο τρέχον βήμα και ανανεώνει την κρυφή κατάσταση h_t. Για τον σκοπό αυτό λαμβάνει υπόψη, επιπλέον, την πληροφορία που έχει παραχθεί στην κατάσταση κελιού. Η κατάσταση κελιού υπολογίζεται με τη βοήθεια των δύο προηγούμενων πυλών όπως φαίνεται στην σχέση (6). Στην τρέχουσα κατάσταση κελιού εφαρμόζεται η συνάρτηση tanh και δημιουργεί την υποψήφια πληροφορία που θα αποτελέσει την νέα κρυφή κατάσταση h_t και κατ' επέκταση την έξοδο του κελιού LSTM. Η πύλη εξόδου δέχεται σαν είσοδο την τρέχουσα είσοδο x_t και την προηγούμενη κρυφή κατάσταση h_{t-1} και εφαρμόζει τη σιγμοειδή συνάρτηση ενεργοποίησης (σχέση (7)). Η επιλογή της σιγμοειδούς συνάρτησης ακολουθεί τη λογική των δύο προηγούμενων πυλών. Αποσκοπεί δηλαδή στη λήψη της απόφασης του ποσοστού της πληροφορίας που θα εξαχθεί από το κελί στο βήμα *t*.

$$C_t = i_t \cdot \tilde{c_t} + f_t \cdot C_{t-1} \text{ (6)}$$
$$o_t = \sigma(Wo \cdot [h_t - 1, x_t] + bo) \text{ (7)}$$

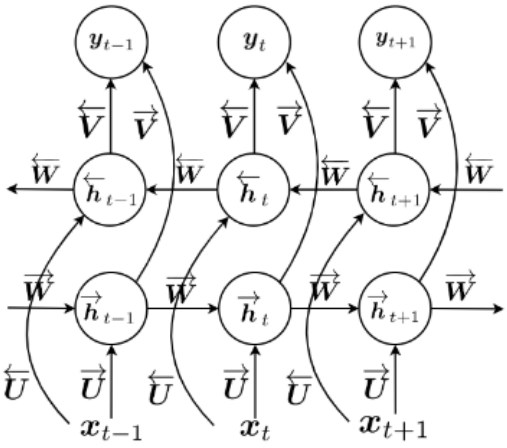
Με αυτό τον τρόπο, με τη βοήθεια των σχέσεων (6), (7) υπολογίζεται η κρυφή κατάσταση h_t (σχέση (8)) που ταυτόχρονα αποτελεί και την έξοδο του LSTM στο βήμα *t*.

$$h_t = o_t \times tanh(C_t) \quad (8)$$

Τα δίκτυα LSTM επεξεργάζονται την ακολουθία εισόδου προς μία κατεύθυνση, από την αρχή προς το τέλος. Αυτό σημαίνει ότι για την πρόβλεψη χρησιμοποιούν μόνο την πληροφορία από τα προηγούμενα στοιχεία της ακολουθίας. Σε πολλές εφαρμογές, όμως, είναι χρήσιμη η πρόσβαση στην πληροφορία που ακολουθεί. Για παράδειγμα, σε εφαρμογές μηχανικής μετάφρασης η σωστή αναπαράσταση της κάθε λέξης στην πρόταση απαιτεί τη χρήση της πληροφορίας των λέξεων που βρίσκονται πριν από αυτή και των λέξεων που βρίσκονται μετά. Για τον σκοπό αυτό, προτάθηκε το δίκτυο του αμφίδρομου RNN (Bidirectional RNN - BiRNN) [27], το οποίο επεξεργάζεται την ακολουθία εισόδου προς τις δύο κατευθύνσεις: από την αρχή προς το τέλος και αντίστροφα. Ανάλογη προσέγγιση εφαρμόστηκε και στα δίκτυα LSTM.

3.5 Αμφίδρομα Αναδρομικά Νευρωνικά Δίκτυα

Τα αμφίδρομα δίκτυα RNN (Bidirectional RNN - BiRNN) προτάθηκαν αρχικά το 1997 [27]. Χρησιμοποιούν δύο κρυφές καταστάσεις, μία για κάθε κατεύθυνση. Στην Εικόνα 11 αποτυπώνεται το δίκτυο BiRNN.



Εικόνα 11- Bidirectional RNN (BiRNN) [69]

Έστω $\vec{h}_t$, $\overleftarrow{h_t}$ οι κρυφές καταστάσεις που διατηρεί το δίκτυο από την επεξεργασία της ακολουθίας με φορά προς το τέλος και προς την αρχή αντίστοιχα. Οι δύο αυτές καταστάσεις υπολογίζονται με τις παρακάτω σχέσεις (9), (10):

$$\vec{h}_t = tanh\left(\overrightarrow{W}\overrightarrow{h_{t-1}} + \vec{U}x_t + \vec{b}_i\right) \quad (9)$$

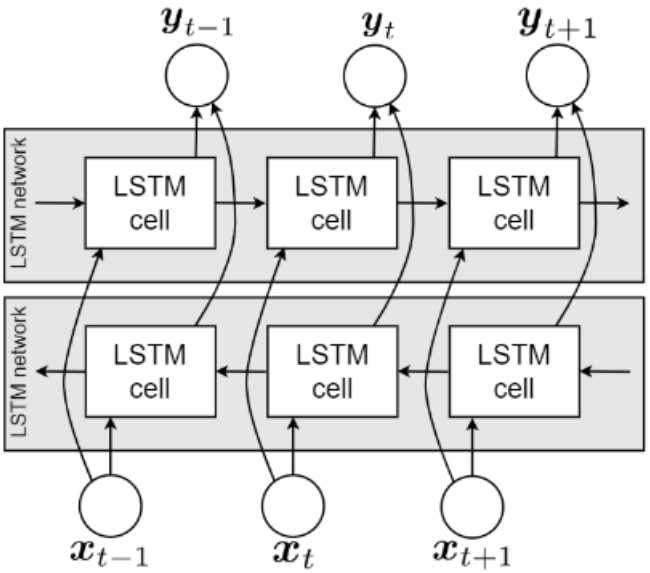
$$\overleftarrow{h_t} = tanh\left(\overleftarrow{W}\overleftarrow{h_t} + \overleftarrow{U}x_t + \overleftarrow{b}_i\right) \quad (10)$$

Ο υπολογισμός της κατάστασης $\vec{h}_t$, όπως φαίνεται και στην εικόνα, χρησιμοποιεί την τρέχουσα είσοδο και την πληροφορία από τα προηγούμενα στοιχεία της ακολουθίας μέσω της κρυφή κατάστασης. Αυτή η διαδικασία είναι η εφαρμογή ενός τυπικού δικτύου RNN. Στο BiRNN, όμως, διατηρείται ακόμη μία κρυφή κατάσταση $\vec{h}_t$, η οποία αρχίζει την επεξεργασία της ακολουθίας εισόδου από το τελευταίο στοιχείο της ακολουθίας με κατεύθυνση προς την αρχή. Έτσι, για το υπολογισμό της κατάστασης $\vec{h}_t$ χρησιμοποιείται η τρέχουσα είσοδος και η προηγούμενη κρυφή κατάσταση $\vec{h}_{t-1}$ που διατηρεί την πληροφορία από τα τελευταία στοιχεία της ακολουθίας προς τα αρχικά. Ο υπολογισμός της εξόδου σε κάθε βήμα, αξιοποιεί και τις δύο κρυφές $\vec{h}_t$, $\overleftarrow{h_t}$ του βήματος *t* και περιγράφεται στην σχέση (11).

$$y_t = \vec{V}\vec{h}_t + \vec{V}\overrightarrow{h_t} + b_y \quad (11)$$

3.6 Αμφίδρομα Δίκτυα LSTM

Αντίστοιχη εφαρμογή της αμφίδρομης επεξεργασίας της ακολουθίας εισόδου, υιοθετήθηκε και στα δίκτυα LSTM. Η πρώτη εφαρμογή αμφίδρομου δικτύου LSTM αφορούσε στην αναγνώριση ομιλίας και απέδωσε καλύτερα συγκριτικά με το κλασσικό δίκτυο LSTM [28]. Τα αμφίδρομα LSTM, όπως και τα BiRNN, αποτελούνται από δύο δίκτυα LSTM, τα οποία λαμβάνουν την ακολουθία εισόδου με αντίθετες κατευθύνσεις. Κάθε δίκτυο LSTM διατηρεί τις δικές του κρυφές καταστάσεις και κελιά μνήμης [28].



Εικόνα 12- Δίκτυο Bi-LSTM.[69]

Κατά τη διάρκεια του πρόσθιου περάσματος, όπως φαίνεται και στην Εικόνα 12, το δίκτυο LSTM επεξεργάζεται την ακολουθία εισόδου από το πρώτο βήμα προς το τελευταίο. Σε κάθε βήμα, υπολογίζει την κρυφή κατάσταση και ενημερώνει την κατάσταση κελιού με βάση την τρέχουσα είσοδο, την προηγούμενη κρυφή κατάσταση και κατάσταση κελιού. Παράλληλα, η ακολουθία εισόδου εισάγεται και στο δίκτυο LSTM, το οποία την επεξεργάζεται με αντίστροφη φορά, από το τελευταίο βήμα προς το πρώτο. Παρόμοια, το LSTM υπολογίζει την κρυφή του κατάσταση και ενημερώνει τη κατάσταση κελιού με βάση την τρέχουσα είσοδο, την προηγούμενη κρυφή κατάσταση και την κατάσταση κελιού.

Μόλις ολοκληρωθούν τα περάσματα προς τα εμπρός και προς τα πίσω, οι κρυφές καταστάσεις και από τα δύο επίπεδα LSTM συνδυάζονται σε κάθε βήμα. Το πλεονέκτημα του BiLSTM είναι ότι καταγράφει όχι μόνο την πληροφορία που λαμβάνεται πριν από ένα συγκεκριμένο χρονικό βήμα (όπως στα παραδοσιακά RNN) αλλά και την πληροφορία που ακολουθεί. Λαμβάνοντας υπόψη τόσο τις προηγούμενες όσο και τις μελλοντικές πληροφορίες, το δίκτυο BiLSTM μπορεί να καταγράψει περισσότερες εξαρτήσεις στην ακολουθία εισόδου.

3.7 Μηχανισμός Προσοχής

Ο μηχανισμός προσοχής (attention mechanism) αποτελεί μία καινοτομία που αρχικά είχε σκοπό να βελτιώσει τα δίκτυα RNNs τύπου κωδικοποιητή-αποκωδικοποιητή (encoder-decoder RNNs), που χρησιμοποιούνται σε εφαρμογές sequence-to-sequence

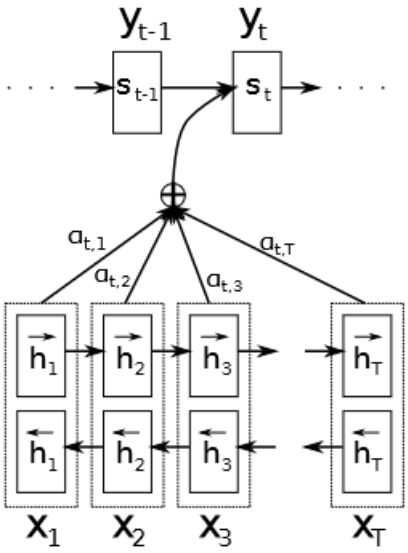
(seq2seq), όπως για παράδειγμα η μηχανική μετάφραση [32]. Η ακολουθία εισόδου κωδικοποιείται από τον κωδικοποιητή (encoder), ο οποίος τη διαβάζει, συνοψίζοντας στη συνέχεια όλη την πληροφορία της σε ένα διάνυσμα συγκεκριμένου μεγέθους, που ονομάζεται διάνυσμα συμφραζομένων (context vector). Το μέγεθος του διανύσματος συμφραζομένων είναι ανεξάρτητο από το μέγεθος της ακολουθίας εισόδου. Έτσι, για παράδειγμα, τόσο μεγάλες όσο και μικρές προτάσεις κωδικοποιούνται σε διάνυσμα ίδιου μεγέθους. Στη συνέχεια, το διάνυσμα αυτό δίνεται σαν είσοδο στον αποκωδικοποιητή (decoder), που με τη σειρά του το επεξεργάζεται ώστε τελικά να μεταφράσει την ακολουθία εισόδου. Σε περίπτωση που ο κωδικοποιητής δεν αποδώσει σωστά την ακολουθία εισόδου, ο αποκωδικοποιητής δεν θα καταφέρει να τη μεταφράσει. Συνεπώς, η αποτελεσματικότητα του αποκωδικοποιητή εξαρτάται από τη δημιουργία αντιπροσωπευτικών διανυσμάτων συμφραζομένων από τον κωδικοποιητή.

Τα πρώτα seq2seq μοντέλα βασίστηκαν στα μοντέλα των RNNs/LSTMs. Παρόλα αυτά ένα από τα κύρια μειονεκτήματα των κλασσικών μοντέλων RNN αποτελεί η αδυναμία διαχείρισης μακροπρόθεσμων εξαρτήσεων σε πολύ μεγάλες ακολουθίες δεδομένων. Αυτό συμβαίνει λόγω του προβλήματος της εξαφάνισης ή έκρηξης της κλίσης. Στην εργασία, στην οποία προτάθηκε για πρώτη φορά η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή βασιζόμενη σε μοντέλα RNNs, αποδεικνύεται πειραματικά ότι η απόδοση του μοντέλου μειώνεται με την αύξηση του μήκους της ακολουθίας εισόδου [31] Επιπλέον, παρόλο που τα μοντέλα LSTM αντιμετωπίζουν το πρόβλημα της εξαφάνισης ή έκρηξης της κλίσης, υπάρχουν περιπτώσεις που δεν είναι τόσο αποδοτικά. Επιπλέον, δεν διαθέτουν κάποιο μηχανισμό ικανό να αποδίδει μεγαλύτερη βαρύτητα στα δεδομένα της ακολουθίας εισόδου. Λύση στα παραπάνω προβλήματα προσφέρει ο μηχανισμός προσοχής, ο οποίος προτάθηκε για πρώτη φορά στην εργασία προκειμένου να ενισχύσει την απόδοση των seq2seq δικτύων.

3.7.1 Πρώτη Εφαρμογή του Μηχανισμού Προσοχής

Σύμφωνα με τον μηχανισμό αυτό ο κωδικοποιητής παράγει μία ακολουθία, ίδιου μήκους με την ακολουθία εισόδου. Στη συνέχεια ο αποκωδικοποιητής λαμβάνει, σε κάθε βήμα του, σαν είσοδο, την ακολουθία αυτή και «δίνει προσοχή» σε συγκεκριμένα σημεία της που θεωρούνται σημαντικά για τη πρόβλεψη στο συγκεκριμένο βήμα. Ο μηχανισμός προσοχής προσαρμόζεται δυναμικά. Αυτό σημαίνει ότι ο

αποκωδικοποιητής δίνει διαφορετική βαρύτητα σε διαφορετικά μέρη της ακολουθίας σε κάθε βήμα του.



Εικόνα 13 - Μοντέλο Κωδικοποιητή/Αποκωδικοποιητή με τον μηχανισμό προσοχής [33]

Στην Εικόνα 13 παρουσιάζεται το μοντέλο προσοχής όπως προτάθηκε στην εργασία [32]. Στη σχηματική αναπαράσταση αποτυπώνεται η στιγμή της πρόβλεψης της λέξης-στόχου y_t στο βήμα *t,* δεδομένης της ακολουθίας εισόδου $(x_1, x_2, \dots, x_T)$. Στη θέση του κωδικοποιητή χρησιμοποιείται το μοντέλο BiRNN. Το μοντέλο αυτό, όπως περιεγράφηκε στα κεφάλαια 4.5, έχει τη δυνατότητα επεξεργασίας της ακολουθίας εισόδου με φορά από την αρχή προς το τέλος και αντίστροφα. Το μοντέλο RNN που διαβάζει την ακολουθία εισόδου με φορά από την αρχή προς το τέλος (από τη λέξη x_1 προς τη λέξη x_T) υπολογίζει τις κρυφές καταστάσεις $(\overrightarrow{h_1}, \overrightarrow{h_{2,\dots}}, \overrightarrow{h_{Tx}})$. Αντίθετα, το RNN που την επεξεργάζεται με την αντίθετη φορά (από τη λέξη x_T προς τη λέξη x_1), υπολογίζει τις αντίστοιχες κρυφές καταστάσεις $\overleftarrow{h_1}, \dots, \overleftarrow{h_{Tx}}$. Η τελική αναπαράσταση κάθε λέξης x_j της ακολουθίας εισόδου προκύπτει από τη συνένωση της κρυφής κατάστασης του $\overrightarrow{hj}$ και του $\overleftarrow{hj}$, όπως αποτυπώνεται και στη σχέση (12):

$$h_j = \left[\overrightarrow{h_j^T}; \overleftarrow{h_j^T}\right]^T \quad (12)$$

Με αυτό τον τρόπο η αναπαράσταση κάθε λέξης της ακολουθίας επηρεάζεται τόσο από τις λέξεις πριν από αυτή όσο και από τις λέξεις που την ακολουθούν.

Το διάνυσμα συμφραζομένων προκύπτει από το σταθμισμένο άθροισμα αυτών των αναπαραστάσεων. Για την πρόβλεψη της λέξη-στόχου y_t, ο αποκωδικοποιητής δέχεται σαν είσοδο το διάνυσμα συμφραζομένων και την προηγούμενη κρυφή κατάστασή του s_{t-1}.

Ο μηχανισμός προσοχής, όπως παρουσιάστηκε στην εργασία [32] αποτελείται από τον υπολογισμό, σε κάθε βήμα, των τιμών στοίχισης (alignment scores), των βαρών και του διανύσματος συμφραζομένων.

Οι τιμές στοίχισης, e_{ij}, υπολογίζονται από το μοντέλο στοίχισης (alignment model). Για τον υπολογισμό τους, το μοντέλο δέχεται τις αναπαραστάσεις h_j και την έξοδο του προηγούμενου βήματος s_{t-1} του αποκωδικοποιητή. Κάθε τιμή στοίχισης υποδηλώνει το βαθμό συσχέτισης κάθε λέξης x_j της ακολουθίας εισόδου με την τρέχουσα έξοδο του αποκωδικοποιητή στη θέση t. Με άλλα λόγια, εκφράζει το κατά πόσο κάθε κρυφή κατάσταση της ακολουθίας εισόδου πρέπει να ληφθεί υπόψη για τον υπολογισμό της εξόδου στο βήμα *t*. Ο υπολογισμός των τιμών στοίχισης δίνεται από την παρακάτω σχέση:

$$e_{ij} = \mathrm{a}(s_{i-1}, h_j) \quad (13)$$

όπου η συνάρτηση a(.) υποδηλώνει το μοντέλο στοίχισης, το οποίο μπορεί να αποτελεί ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης (feedforward). Τα βάρη a_{ij}, υπολογίζονται με την εφαρμογή της συνάρτησης SoftMax στις τιμές στοίχισης της εξίσωσης (13):

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (14)$$

Η χρήση της συνάρτησης SoftMax αποφέρει τα εξής πλεονεκτήματα:

1. Κανονικοποιεί τα βάρη a_{ij} στο εύρος 0 και 1.
2. Το άθροισμα των βαρών ισούται με 1.

Έτσι, τα βάρη υποδηλώνουν τι ποσοστό από κάθε λέξη της ακολουθίας εισόδου πρέπει να χρησιμοποιήσει ο αποκωδικοποιητής για την πρόβλεψη του. Τέλος, τα βάρη αυτά χρησιμοποιούνται για τον υπολογισμό του διανύσματος συμφραζομένων, το οποίο αποτελεί το σταθμισμένο άθροισμα όλων των αναπαραστάσεων (κρυφές καταστάσεις) Τ, των λέξεων εισόδου του κωδικοποιητή:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \qquad (15)$$

Ο αποκωδικοποιητής λαμβάνοντας σε κάθε βήμα το διάνυσμα συμφραζομένων και την προηγούμενη κρυφή κατάστασή του, προβλέπει την τιμή-στόχος.

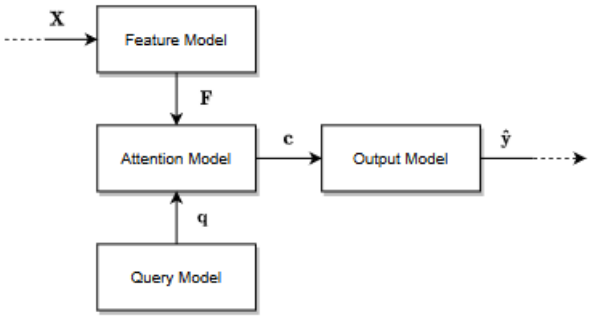
Με αυτό τον τρόπο εφαρμόζεται ο μηχανισμός προσοχής, που επιτρέπει στον αποκωδικοποιητή να αποφασίσει σε ποιες λέξεις της ακολουθίας εισόδου θα δώσει περισσότερη προσοχή. Με τη χρήση αυτού του μηχανισμού αντιμετωπίζεται ο περιορισμός του κωδικοποιητή να συνοψίζει ολόκληρη την ακολουθία εισόδου σε ένα διάνυσμα συγκεκριμένου μεγέθους. Ο μηχανισμός προσοχής επιτυγχάνει τη διάχυση της πληροφορίας ολόκληρης της ακολουθίας εισόδου, ώστε τελικά ο αποκωδικοποιητής να επιτύχει την πιο σωστή πρόβλεψη.

Παραλλαγές του μηχανισμού προσοχής, όπως παρουσιάστηκε στην εργασία [32], άρχισαν να χρησιμοποιούνται και σε άλλες εφαρμογές, εκτός της μηχανικής μετάφρασης. Παρόλα αυτά, τα βήματα που ακολουθούνται σε οποιαδήποτε μορφή του μηχανισμού προσοχής είναι ίδια και περιγράφονται από το γενικό μηχανισμό προσοχής [33,34,35]. Η περιγραφή του γενικευμένου μοντέλου προσοχής (general attention model) περιγράφεται στο επόμενο κεφάλαιο..

3.7.2 Γενικευμένο Μοντέλο Προσοχής

Η περιγραφή της γενικής μορφής του μηχανισμού προσοχής αποσκοπεί στην κατανόησή του και στον τρόπο εφαρμογής του σε άλλους τομείς, πέραν της μηχανικής μετάφρασης. Για παράδειγμα, μοντέλα που αξιοποιούν κάποιον μηχανισμό προσοχής μπορεί να έχουν στόχο την παραγωγή περίληψης ενός κειμένου εισόδου, την αντιστοίχιση ενός κειμένου με κάποιο συναίσθημα, αναλόγως του ύφους του κειμένου εισόδου. Επιπλέον, ένα τέτοιο μοντέλο μπορεί να λάβει σαν είσοδο μία εικόνα προκειμένου να παράξει ένα κείμενο που θα την περιγράφει.

Η Εικόνα 14 παρουσιάζει γραφικά τη γενική δομή του μοντέλου προσοχής.

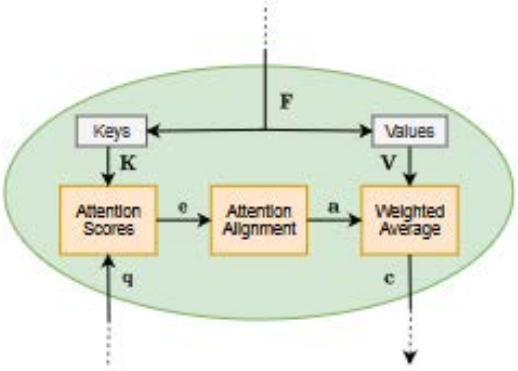


Εικόνα 14- Γενική δομή του μοντέλου προσοχής [33]

Το μοντέλο αποτελείται από τέσσερα υπό-μοντέλα, το μοντέλο χαρακτηριστικών (feature model), το μοντέλο ερωτημάτων (query model), το μοντέλο προσοχής (attention model) και το μοντέλο εξόδου (output model).

Δεδομένης της εισόδου ενός πίνακα $x \in \mathbb{R}^{dx \times n}$, όπου dx, n εκφράζουν το μέγεθος των διανυσμάτων εισόδου και το πλήθος των δεδομένων εισόδου αντίστοιχα, το μοντέλο χαρακτηριστικών παράγει n_f διανύσματα χαρακτηριστικών $f_1, \dots, f_{n_f} \in \mathbb{R}^{d_f}$ από το X , όπου d_f δηλώνει το μέγεθός των διανυσμάτων χαρακτηριστικών. Ως μοντέλο χαρακτηριστικών μπορεί να χρησιμοποιηθεί κάποια αρχιτεκτονική RNN ή CNN, ένα στρώμα που μαθαίνει τις αναπαραστάσεις (embedding layer), ένας γραμμικός μετασχηματισμός της εισόδου X. Ουσιαστικά, το μοντέλο χαρακτηριστικών αποτελείται από όλα τα βήματα που ακολουθούνται για τη μετατροπή της εισόδου X στα διανύσματα χαρακτηριστικών $f_1 \dots, f_{n_f}$. Στα διανύσματα αυτά θα «δώσει προσοχή» το μοντέλο προσοχής.

Το μοντέλο προσοχής προκειμένου να αποφασίσει τα χαρακτηριστικά διανύσματα που θα δώσει περισσότερη βαρύτητα χρειάζεται τα ερωτήματα (queries) $q \in \mathbb{R}^{dq}$, όπου dq δηλώνει το μέγεθος του διανύσματος ερωτήματος. Τα τελευταία αποτελούν αναπόσπαστο κομμάτι του μοντέλου προσοχής, αφού καθορίζουν την πληροφορία από τα διανύσματα χαρακτηριστικών που θα χρησιμοποιηθούν.



Εικόνα 15- Γενική δομή της μονάδας γενικής προσοχής [33]

Τα διανύσματα χαρακτηριστικών και τα ερωτήματα εισάγονται σαν είσοδο στο μοντέλο προσοχής. Το μοντέλο αυτό μπορεί να αποτελείται από μία ή περισσότερες γενικές μονάδες προσοχής (general attention modules). Στην Εικόνα 15 αποτυπώνεται ο εσωτερικός μηχανισμός της γενικής μονάδας προσοχής. Την είσοδο αυτής της μονάδας αποτελούν ένα ερώτημα q και ο πίνακας διανυσμάτων χαρακτηριστικών F. Από τον πίνακα χαρακτηριστικών εξάγονται δύο πίνακες. Ο πρώτος αποτελεί τον πίνακα των κλειδιών (keys) $K = \left[k_1, \dots, k_{n_f}\right] \in \mathbb{R}^{d_k \times n_f}$ και ο δεύτερος τον πίνακα των τιμών (values) $v = \left[v_1, \dots, v_{n_f}\right] \in \mathbb{R}^{d_v \times n_f}$, όπου d_k, dv δηλώνουν τη διάσταση των διανυσμάτων των κλειδιών και των τιμών αντίστοιχα. Κάθε διάνυσμα τιμών $v_1, \dots, v_{n_f}$ αντιστοιχεί σε ένα μοναδικό διάνυσμα κλειδιού $k_{1,\dots}, k_{n_f}$. Συνήθως, οι πίνακες *K, V* προέρχονται από το γραμμικό μετασχηματισμό του πίνακα χαρακτηριστικών *F*. Ο τρόπος υπολογισμού τους αποδίδεται από τις σχέσεις (16).

$$K = W_K \times F,\ V = W_v \times F \quad (16)$$

όπου $W_K \in \mathbb{R}^{d_k \times d_f},\ W_V \in \mathbb{R}^{d_v \times d_f},\ F \in \mathbb{R}^{d_f \times n_f}$.

Ανάλογα με την εφαρμογή του μοντέλου προσοχής, ο πίνακας κλειδιών *K* μπορεί να εκφράζει για παράδειγμα τα χαρακτηριστικά συγκεκριμένης περιοχής μίας εικόνας, αναπαραστάσεις λέξεων (word embeddings) κειμένου ή κρυφές καταστάσεις ενός μοντέλου RNN. Ανάγοντας τις έννοιες *Q, K, V* στο μηχανισμό προσοχής όπως αυτός προτάθηκε πρώτη φορά και περιεγράφηκε στο κεφάλαιο 4.7.1, το ερώτημα *q* αποτελεί την προηγούμενη έξοδο s_{t-1} του αποκωδικοποιητή, ενώ οι τιμές και τα κλειδιά είναι οι ανάλογες κρυφές καταστάσεις του κωδικοποιητή.

Σκοπός της μονάδας προσοχής (Εικόνα 15) αποτελεί η παραγωγή του σταθμισμένου μέσου όρου των τιμών διανυσμάτων του πίνακα V. Για το σκοπό αυτό, η μονάδα υπολογίζει τη συσχέτιση μεταξύ του ερωτήματος q και των κλειδιών με τη βοήθεια της συνάρτησης score (score function), όπως φαίνεται στη σχέση (17):

$$e = f(q, K) \qquad (17)$$

όπου $e = \left[e_1, \dots, e_{n_f}\right] \in \mathbb{R}^{n_f x n_f}$ αποτελούν τα σκορ προσοχής. Το ερώτημα q αναπαριστά κάποιο αίτημα για πληροφορία. Το σκορ προσοχής e_j υποδηλώνει το πόσο σημαντική είναι η πληροφορία που περιέχει το κλειδί k σε σχέση με το συγκεκριμένο αίτημα q. Ένα παράδειγμα της συνάρτησης *score (.)* αποτελεί η χρήση του εσωτερικού γινομένου των δύο διανυσμάτων.

Οι τιμές των σκορ προσοχής, συνήθως δεν κυμαίνονται στο διάστημα [0, 1]. Ωστόσο, επειδή ο στόχος της μονάδας προσοχής είναι η παραγωγή του σταθμισμένου αθροίσματος τιμών V , είναι απαραίτητο να κανονικοποιηθούν με τη χρήση μίας συνάρτησης κατανομής g στις τιμές των σκορ προσοχής, όπως φαίνεται στη σχέση (18):

$$a = g(e) \quad (18)$$

Κατά αυτό τον τρόπο προκύπτουν τα βάρη προσοχής $a = \left[a_1, a_2, \cdots, a_{nf}\right] \in \mathbb{R}^{n_f}$ και καθορίζουν σε ποια σημεία των δεδομένων εισόδου, το μοντέλο θα δώσει μεγαλύτερη βαρύτητα. Ως συνάρτηση g συνήθως χρησιμοποιείται η συνάρτηση SoftMax, τα πλεονεκτήματα της οποίας περιεγράφηκαν στο κεφάλαιο 4.7.1 .

Τα βάρη α χρησιμοποιούνται για τη δημιουργία του διανύσματος συμφραζομένων $c \in \mathbb{R}^{dv}$, υπολογίζοντας το σταθμισμένο άθροισμα των διανυσμάτων τιμών $v_1, \cdots, v_{nf}$, όπως φαίνεται στη σχέση (19):

$$c = \sum_{l=1}^{n_f} a_l \times v_l \quad (19)$$

Τέλος, παρατηρώντας την Εικόνα 14 το διάνυσμα συμφραζομένων c θα χρησιμοποιηθεί από το μοντέλο εξόδου για την παραγωγή της εξόδου y.

Συνοψίζοντας, ο γενικότερος μηχανισμός προσοχής μπορεί να χρησιμοποιηθεί από οποιαδήποτε αρχιτεκτονική νευρωνικών δικτύων ενισχύοντας τον τρόπο που αυτά αντιλαμβάνονται την πληροφορία για τον εκάστοτε σκοπό.

3.7.3 Αυτό-προσοχή

Ο μηχανισμός αυτό-προσοχής (self attention) αποτελεί ένα είδος μηχανισμού προσοχής που επιτρέπει κάθε στοιχείο της ακολουθίας εισόδου να «δίνει προσοχή» στα υπόλοιπα στοιχεία της ίδια ακολουθίας εισόδου. Με αυτό τον τρόπο, ένα νευρωνικό δίκτυο μπορεί να μαθαίνει μακρινές εξαρτήσεις της ακολουθίας εισόδου και να κατανοεί καλύτερα τις σχέσεις των στοιχείων της. Ο μηχανισμός αυτό-προσοχής ονομάζεται έτσι, καθώς σε αυτό τον τύπο προσοχής, οι πίνακες *Q, K* και *V* προέρχονται από την ίδια την ακολουθία εισόδου. Ένα ερώτημα q ανατίθεται σε κάθε λέξη της ακολουθίας. Στη συνέχεια, αυτό συγκρίνεται με τα κλειδιά, τα οποία επίσης προέρχονται από τις λέξεις της ακολουθίας εισόδου, προκειμένου να βρει τις πιο σχετικές πληροφορίες που αντιστοιχούν στη λέξη αυτή. Κάθε αναπαράσταση $f_1, \dots, f_{n_f}$ της ακολουθίας εισόδου μετατρέπεται σε ένα διάνυσμα q μέσω του πίνακα $W_Q \in \mathbb{R}^{dq \times d_f}$. Έτσι, ο πίνακας ερωτημάτων $Q = \left[q_1, \dots, q_{n_f}\right] \in \mathbb{R}^{dq \times n_f}$ υπολογίζεται σύμφωνα με τη σχέση (20):

$$Q = W_Q \times F \qquad (20)$$

όπου $W_Q \in \mathbb{R}^{d_q \times d_f}$. Αντίστοιχα, υπολογίζεται ο πίνακας κλειδιών *K* και ο πίνακας τιμών *V*, όπως φαίνεται στη σχέση (16) του προηγούμενου κεφαλαίου. Τόσο ο πίνακας W_Q όσο οι πίνακες W_K, W_V, περιέχουν βάρη, τα οποία μαθαίνει το νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης.

Κάθε στήλη στον πίνακα *Q* αποτελεί ένα διάνυσμα ερωτημάτων *q*, το οποίο, όπως προκύπτει από τα παραπάνω, αποτελεί μία αναπαράσταση του αντίστοιχου διανύσματος χαρακτηριστικών. Ο πίνακας *K* αποτελείται από τα διανύσματα κλειδιών για κάθε μία λέξη της ακολουθίας εισόδου. Αυτά τα διανύσματα χρησιμοποιούνται για τον υπολογισμό της ομοιότητας (similarity) ή της σχετικότητας (relevance) ανάμεσα στην εκάστοτε λέξη (χρήση του αντίστοιχου *q*) και στις υπόλοιπες λέξεις της ακολουθίας. Όσο μεγαλύτερη η τιμή της συνάρτησης score μεταξύ του ερωτήματος *q*

και του διανύσματος κλειδιού *k*, τόσο μεγαλύτερη σχέση υπάρχει μεταξύ των δύο λέξεων. Τέλος, ο πίνακας *V* περιλαμβάνει τα διανύσματα τιμών *v* για κάθε λέξη της ακολουθίας εισόδου. Τα τελευταία, περιλαμβάνουν τη συμφραζόμενη πληροφορία (contextual information) κάθε λέξης. Μετά τον υπολογισμό των τιμών της συνάρτησης score μεταξύ του ερωτήματος *q* και των διανυσμάτων *k*, υπολογίζεται το σταθμισμένο άθροισμα των διανυσμάτων *v*. Τα βάρη για κάθε διάνυσμα *v*, καθορίζονται από τις τιμές της συνάρτησης score, διασφαλίζοντας με αυτό τον τρόπο ότι η τελική αναπαράσταση της εκάστοτε λέξης έχει επηρεαστεί περισσότερο από τις σχετικότερες προς αυτή λέξεις της ακολουθίας.

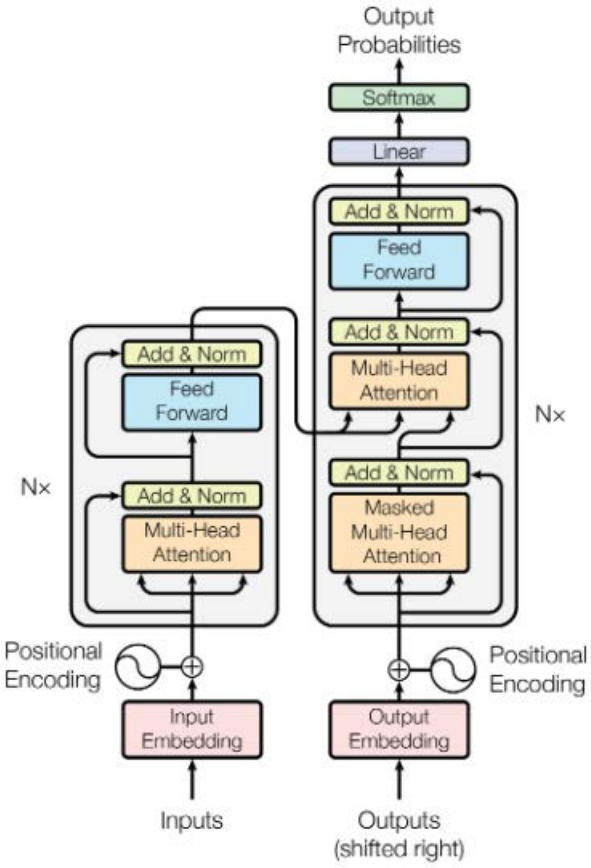
Ο μηχανισμός αυτό-προσοχής αποτελεί τον πιο διαδεδομένο τύπο μηχανισμού προσοχής, κυρίως λόγω του σημαντικού ρόλου που διαδραματίζει στην ανάπτυξη των μοντέλων των μετασχηματιστών (transformers) [36]. Οι μετασχηματιστές βασίζουν την επιτυχία τους στο μηχανισμό αυτό. Στο κεφάλαιο που ακολουθεί αναλύεται το μοντέλο των μετασχηματιστών.

3.8 Μετασχηματιστές

Τα μοντέλα των μετασχηματιστών (transformers), τα οποία παρουσιάστηκαν για πρώτη φορά το 2017 στην εργασία “Attention is ALL You Need” , αποτελούν μία κατηγορία αρχιτεκτονικών νευρωνικών δικτύων που έχουν σημειώσει σημαντική πρόοδο στον τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) [36]. Αξιοποιούνται σε πολλές εφαρμογές όπως, στη μηχανική μετάφραση κειμένου, στην ανάλυση συναισθημάτων, στην ταξινόμηση και παραγωγή κειμένου κ.α. Τα δημοφιλέστερα μοντέλα μετασχηματιστών αποτελούν τα μοντέλα BERT, RoBERTa και GPT-2 [37, 38, 39].

Η υπέροχη τους σε σχέση με τα παραδοσιακά αναδρομικά νευρωνικά δίκτυα, όπως είναι τα RNNs, τα LSTMs και τα GRUs, έγκειται στο γεγονός ότι δεν κάνουν χρήση της αναδρομής (recurrence) αλλά βασίζονται στην αξιοποίηση του μηχανισμού προσοχής. Ο αναδρομικός τρόπος λειτουργίας, καθιστά τα RNNs ακολουθιακά μοντέλα. Αυτό σημαίνει ότι η έξοδος σε κάθε βήμα εξαρτάται από την προηγούμενη κρυφή κατάσταση (hidden state) και τη τρέχουσα είσοδο. Αντίθετα, οι μετασχηματιστές αξιοποιούν τον μηχανισμό προσοχής για την κατανόηση των εξαρτήσεων μεταξύ των στοιχείων της πρότασης, δίχως να εξαρτώνται από αναδρομικές σχέσεις και κρυφές καταστάσεις. Αυτό, τα καθιστά μη ακολουθιακά

μοντέλα επιτρέποντας την παράλληλη επεξεργασία των δεδομένων εισόδου. Επιπλέον, ο μηχανισμός αυτός βοηθάει τα μοντέλα να κατανοήσουν τις σχέσεις των λέξεων μέσα στην πρόταση, ανεξαρτήτως από το μέγεθός της, συμβάλλοντας με αυτό τον τρόπο στην καλύτερη κατανόηση του νοήματος του κειμένου, επιτυγχάνοντας καλύτερη πρόβλεψη.

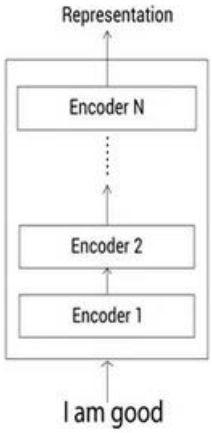


Εικόνα 16 – Αρχιτεκτονική του μοντέλου του Μετασχηματιστή [36]

Η βασική αρχιτεκτονική των μετασχηματιστών αποτελείται από τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Όπως φαίνεται και στην Εικόνα 16, το μοντέλο των μετασχηματιστών βασίζεται αποκλειστικά σε μηχανισμούς προσοχής. Ο κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου και παράγει μια συνεχή αναπαράσταση της εισόδου. Στη συνέχεια ο αποκωδικοποιητής δέχεται σαν είσοδο τις αναπαραστάσεις του κωδικοποιητή και παράγει την ακολουθία εξόδου. Ο κωδικοποιητής και ο αποκωδικοποιητής, αποτελούνται από πολλά πανομοιότυπα επίπεδα (layers). Κάθε επίπεδο περιλαμβάνει δύο βασικά υπό-επίπεδα (sublayers). Το πρώτο εκτελεί το μηχανισμό αυτό-προσοχής και το δεύτερο αποτελεί ένα νευρωνικό

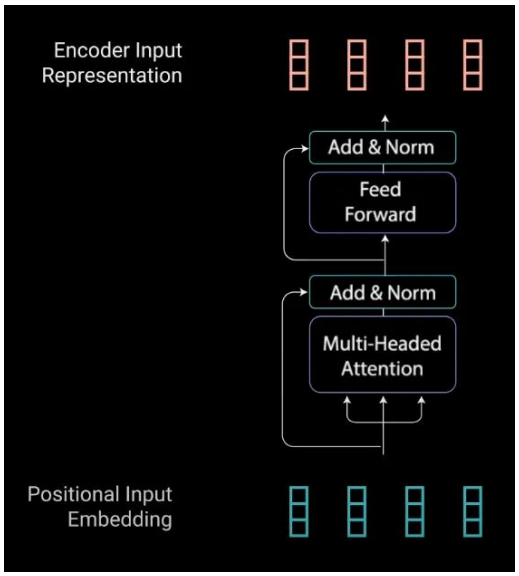
δίκτυο πρόσθιας τροφοδότησης. Στα κεφάλαια που ακολουθούν θα αναλυθεί η αρχιτεκτονική του κωδικοποιητή και του αποκωδικοποιητή.

3.8.1 Κωδικοποιητής



Εικόνα 17 - Επίπεδα Κωδικοποιητών (Encoder Layers) (Πηγή: https://www.analyticsvidhya.com/blog/2023/07/transformers-encoder-the-crux-of-the-nlp-issues/ *)*

Ο κωδικοποιητής αποτελείται από *N* πανομοιότυπα επίπεδα κωδικοποιητών. Στην εργασία που εισήγαγε το μοντέλο του μετασχηματιστή, επιλέχθηκε *N=6* [36]. Αυτό σημαίνει ότι χρησιμοποιήθηκαν έξι επίπεδα κωδικοποιητών, το ένα μετά το άλλο, προκειμένου να συγκροτηθεί το μοντέλο του κωδικοποιητή. Όπως φαίνεται και στην Εικόνα 17, η έξοδος από κάθε επίπεδο συνδέεται με την είσοδο του επόμενου επιπέδου. Το τελικό επίπεδο επιστρέφει την αναπαράσταση της ακολουθίας εισόδου.

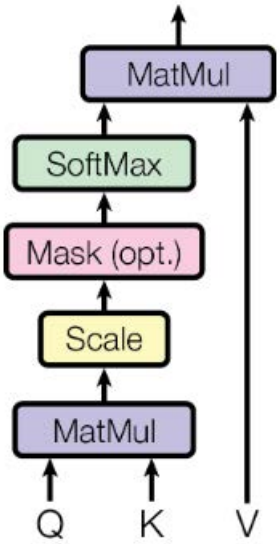


Εικόνα 18 - Υπομονάδες του επιπέδου Κωδικοποιητή (Πηγή https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0)

Κάθε επίπεδο κωδικοποιητή αποτελείται από δύο υπό-επίπεδα: το υπό-επίπεδο που εφαρμόζει την προσοχή (multi-head) και το επίπεδο που περιλαμβάνει το feed forward μοντέλο.

3.8.1.1 Μηχανισμός Προσοχής Κωδικοποιητή

Self Attention



Εικόνα 19 - Σειρά των πράξεων στο μηχανισμό Scaled Dot-Product. Το επίπεδο Mask (Opt.) χρησιμοποιείται μόνο στο μηχανισμό προσοχής της μονάδας του Κωδικοποιητή [36]

Όπως αναφέρθηκε και στο κεφάλαιο 4.7.3, ο μηχανισμός αυτό-προσοχής βοηθάει το μοντέλο στην κατανόηση των σχέσεων των λέξεων της ακολουθίας εισόδου. Στην εργασία που εισήγαγε τους μετασχηματιστές [36], ο μηχανισμός αυτός αναφέρεται ως scaled dot-product attention. Για την υλοποίησή του υπολογίζεται το εσωτερικό γινόμενο (dot product) του πίνακα ερωτημάτων Q και του ανάστροφού του, πίνακα κλειδιών K, διαιρούμενου με την ποσότητα $\sqrt{d_k}$, όπου d_k η διάσταση των διανυσμάτων των κλειδιών. Τα εσωτερικά γινόμενα που προκύπτουν είναι μη κανονικοποιημένα. Για το λόγο αυτό εφαρμόζεται η συνάρτηση SoftMax, η οποία μετατρέπει τα σκορ σε τιμές πιθανότητας μεταξύ των τιμών 0 και 1. Έτσι, προκύπτει ότι κάθε λέξη της ακολουθίας εισόδου σχετίζεται κατά ένα ποσοστό με τις υπόλοιπες λέξεις της ακολουθίας. Κατά αυτό τον τρόπο υπολογίζονται τα βάρη τα οποία στη συνέχεια πολλαπλασιάζονται με

τον πίνακα των τιμών *V*. Η παρακάτω σχέση (21) αποτυπώνει τον μηχανισμό αυτοπροσοχής, όπως αυτός παρουσιάζεται στην εργασία [36].

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (21)$$

Multi – Head Attention

Παραπάνω περιεγράφηκε ο μηχανισμός αυτό-προσοχής, όπως χρησιμοποιείται στην αρχιτεκτονική των μετασχηματιστών. Στην εργασία που εισήγαγε τους μετασχηματιστές [36], εφαρμόστηκε ο μηχανισμός αυτός πολλές φορές και παράλληλα. Συγκεκριμένα, αντί της εφαρμογής ενός μόνο μηχανισμού αυτοπροσοχής στα διανύσματα κλειδιών, ερωτημάτων και τιμών που έχουν διάσταση d_{model} (σύμφωνα με την εργασία $d_{model} = 512$), πραγματοποιήθηκε η γραμμική προβολή τους (linear projection) κατά h φορές σε διανύσματα μικρότερης διάστασης d_k, d_k και d_v αντίστοιχα, όπου

$$d_k = d_{model} / h \quad (22)$$

Η παράμετρος *h*, αποτελεί μια υπερπαράμετρο (hyperparameter) και αναφέρεται στις κεφαλές (heads) του μηχανισμού προσοχής multi-head. Εκφράζει τις φορές που θα εκτελεστεί παράλληλα ο μηχανισμός προσοχής. Στην εργασία [36], επιλέγεται η χρήση οκτώ κεφαλών ($h = 8$). Κάθε κεφαλή εκτελεί τον μηχανισμό προσοχής scaled dot-product, ανεξάρτητα από τις υπόλοιπες κεφαλές. Η λειτουργία κάθε κεφαλής συνοψίζεται στη σχέση (23):

$$\text{headi} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (23)$$

όπου $W_i^Q, W_i^K \in \mathbb{R}^{d_{model} \times d_k}$ και $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, αποτελούν τα βάρη βάσει των οποίων επιτυγχάνεται η γραμμική προβολή των ερωτημάτων, των κλειδιών και των τιμών αντίστοιχα. Τα βάρη αυτά μαθαίνονται από το μοντέλο κατά διαδικασία της εκπαίδευσης. Κάθε σύνολο W_i^Q, W_i^K, W_i^V είναι ανεξάρτητο σε κάθε κεφαλή και προβάλλει γραμμικά τα διανύσματα από την αρχική διάσταση $d_{model} = 512$ στη

διάσταση $d_k = d_v = 64$ (σύμφωνα με την σχέση (23)). Στη συνέχεια, τα αποτελέσματα των κεφαλών συνενώνονται και επεξεργάζονται από ένα γραμμικό στρώμα σύμφωνα με τη σχέση (24):

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(head_i, \dots, head_h)W^O \quad (24)$$

όπου $\text{W}^o \in \mathbb{R}^{hd_v} \times d_{modeL}$ αποτελούν τις παραμέτρους της τελευταίας γραμμικής προβολής.

Η χρήση του μηχανισμού προσοχής multi-head βοηθάει το μοντέλο στην αποτελεσματικότερη εύρεση των σχέσεων μεταξύ των λέξεων εισόδου, αφού εκτελείται παράλληλα σε διαφορετικές αναπαραστάσεις της ακολουθίας εισόδου, εστιάζοντας κάθε φορά σε διαφορετικές θέσεις. Επιπλέον, η συνεισφορά του είναι σημαντική στη βελτίωση της υπολογιστικής απόδοσης του μοντέλου, αφού παρόλο, που κάνει χρήση πολλών κεφαλών, η εκτέλεση του μηχανισμού σε κάθε κεφαλή εκτελείται παράλληλα. Επίσης, όπως αναφέρεται και στην εργασία [36] , λόγω της μείωσης της διάστασης κάθε κεφαλής, το υπολογιστικό κόστος παραμένει στα ίδια επίπεδα με την εκτέλεση του μηχανισμού, με τη χρήση μίας κεφαλής που χρησιμοποιεί όλη την αρχική διάσταση d_{model}.

Στη συνέχεια, η έξοδος από το επίπεδο προσοχής multi-head εισέρχεται σε ένα επίπεδο κανονικοποίησης και ακολούθως στο μοντέλο feed forward. Όπως φαίνεται και στην Εικόνα 18, στη μονάδα του κωδικοποιητή στο τέλος κάθε υπό-επιπέδου (multi-head attention, feed forward model) εφαρμόζονται υπολειπόμενες συνδέσεις (residual connections).

3.8.1.2 Normalization Layer, Feed Forward Model, Residuals Connection

Στην έξοδο του υπό-επιπέδου που υλοποιεί τον μηχανισμό προσοχής multi-head προστίθεται απευθείας η είσοδος του κωδικοποιητή μέσω τις ύπαρξης υπολειπόμενης σύνδεσης. Στη συνέχεια η έξοδος της υπολειπόμενης σύνδεσης εισέρχεται σε ένα στρώμα κανονικοποίησης (normalization layer). Στο σημείο αυτό το δεύτερο υπό-επίπεδο που αποτελείται ένα δίκτυο πρόσθιας τροφοδότησης ως προς τη θέση (position-wise fully connected feed-forward model) δέχεται σαν είσοδο τη κανονικοποιημένη έξοδο της υπολειπόμενης σύνδεσης.

Position-Wise Feed Forward Model

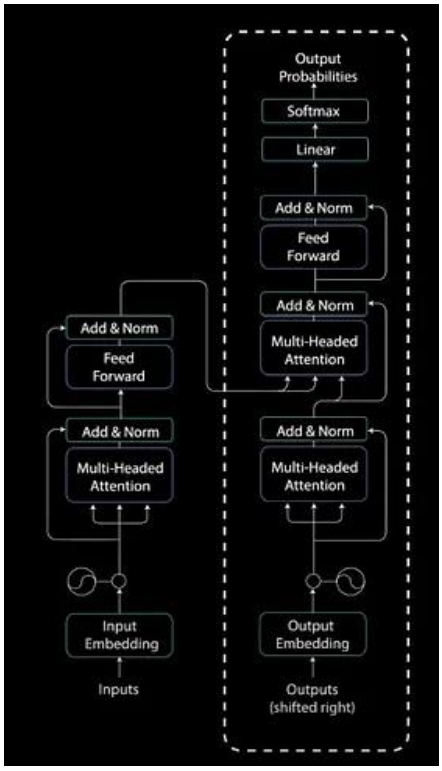
Κάθε στρώμα, τόσο στη μονάδα του κωδικοποιητή όσο και στην μονάδα του αποκωδικοποιητή, περιέχει το συγκεκριμένο υπό-επίπεδο που αποτελείται από ένα νευρωνικό δίκτυο πρόσθιας τροφοδοσίας (feed-forward network). Το δίκτυο αυτό περιλαμβάνει δύο στρώματα γραμμικών μετασχηματιστών (linear layers), όπου το πρώτο στρώμα ακολουθείται από τη συνάρτηση ενεργοποίησης ReLU, όπως αποτυπώνεται στη σχέση (25):

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (25)$$

όπου $W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$ και $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$. Η είσοδος x, όπως αναφέρθηκε αποτελεί την κανονικοποιημένη έξοδο της υπολειπόμενης σύνδεσης και περιέχει διανύσματα διάστασης $d_{model.}$ Στο πρώτο επίπεδο, η είσοδος x προβάλλεται σε μία μεγαλύτερη διάσταση d_{ff}, ($d_{ff} = 2048$ στην [36] μέσω του γραμμικού μετασχηματισμού που εφαρμόζεται με τη χρήση του W_1. Στη συνέχεια, εφαρμόζεται η συνάρτηση ενεργοποίησης ReLU προκειμένου να προσδώσει στο μοντέλο μη γραμμικότητα ώστε να μπορεί να μάθει πιο περίπλοκες σχέσεις μεταξύ των δεδομένων εισόδου. Το αποτέλεσμα της συνάρτησης ενεργοποίησης εισέρχεται στο δεύτερο επίπεδο, όπου μέσω του γραμμικού μετασχηματισμού που συνδέεται με το W_2 , μειώνεται η διάσταση εκ νέου στην αρχική $d_{model.}$ Το δίκτυο αυτό χαρακτηρίζεται ως position-wise καθώς εφαρμόζει τον ίδιο γραμμικό μετασχηματισμό σε κάθε διάνυσμα της εισόδου x. Η είσοδος x, προστίθεται απευθείας με το αποτέλεσμα του δεύτερου γραμμικού στρώματος μέσω της δεύτερης υπολειπόμενης σύνδεσης και εισέρχεται σε ένα στρώμα κανονικοποίησης, παράγοντας τις αναπαραστάσεις της εισόδου του κωδικοποιητή. Οι αναπαραστάσεις που προκύπτουν από το τελικό στρώμα του κωδικοποιητή αντιπροσωπεύουν κάθε στοιχείο της ακολουθίας εισόδου και περιλαμβάνουν όλη την απαραίτητη πληροφορία ως προς τη σημασία κάθε στοιχείου εισόδου σε σχέση με τα υπόλοιπα στοιχεία. Τις αναπαραστάσεις αυτές θα τις επεξεργαστεί το μοντέλο του αποκωδικοποιητή, η αρχιτεκτονική του οποίου πρόκειται να αναλυθεί στο επόμενο κεφάλαιο.

3.8.2 Αποκωδικοποιητής

Ο αποκωδικοποιητής (decoder) παράγει την ακολουθία εξόδου. Σε κάθε βήμα η πρόβλεψη του επόμενου στοιχείου της ακολουθίας βασίζεται στα στοιχεία που ήδη έχουν προβλεφθεί στα προηγούμενα βήματα και στην κωδικοποιημένη πληροφορία της ακολουθίας εισόδου, έτσι όπως παράγεται από τον κωδικοποιητή. Όπως και με τη μονάδα του κωδικοποιητή, έτσι και ο αποκωδικοποιητής αποτελείται από N πανομοιότυπα επίπεδα αποκωδικοποιητών. Στην εργασία [36] αναφέρεται ότι και για τον αποκωδικοποιητή ισχύει $N=6$. Αυτό σημαίνει ότι και σε αυτή τη περίπτωση η μονάδα του αποκωδικοποιητή αποτελείται από μία σειρά έξι πανομοιότυπων διασυνδεδεμένων επιπέδων αποκωδικοποιητών. Ένα επίπεδο αποκωδικοποιητή αποτελείται από παρόμοια υπό-επίπεδα, όπως στην περίπτωση του κωδικοποιητή. Συγκεκριμένα, αποτελείται από δύο υπό-επίπεδα που εφαρμόζουν τον μηχανισμό προσοχής multi-head και ένα υπό-επίπεδο που αποτελείται από ένα position-wise νευρωνικό δίκτυο πρόσθιας τροφοδοσίας. Ομοίως με τον κωδικοποιητή, εφαρμόζονται υπολειπόμενες συνδέσεις γύρω από κάθε υπό-επίπεδο. Οι συνδέσεις αυτές ακολουθούνται από ένα στρώμα κανονικοποίησης. Η τελική έξοδος της μονάδας του αποκωδικοποιητή εισέρχεται σε ένα γραμμικό στρώμα, το οποίο συμπεριφέρεται σαν επίπεδο ταξινόμησης. Τελικά, εφαρμόζεται η συνάρτηση ενεργοποίησης SoftMax στο αποτέλεσμα που προκύπτει από το γραμμικό επίπεδο, ώστε να προκύψουν οι πιθανότητες πρόβλεψης κάθε λέξης.



Εικόνα 20 - Επίπεδο Αποκωδικοποιητή (Πηγή: https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0)

Τα υπό-επίπεδα που εκτελούν τον μηχανισμό προσοχής multi-head στο επίπεδο του αποκωδικοποιητή είναι δύο και παρουσιάζουν διαφορές στον τρόπο λειτουργίας τους σε σχέση με τον αντίστοιχο μηχανισμό που εκτελείται στα επίπεδα του κωδικοποιητή. Όπως φαίνεται και στην Εικόνα 20, η μονάδα του αποκωδικοποιητή λαμβάνει δύο εισόδους. Η ακολουθία εξόδου που έχει προβλέψει ο αποκωδικοποιητής μέχρι εκείνο το βήμα αποτελεί τη μία είσοδο. Η δεύτερη είσοδος λαμβάνεται από την έξοδο του κωδικοποιητή, η οποία περιέχει κωδικοποιημένη όλη την πληροφορία της ακολουθίας εισόδου. Το πρώτο υπό-επίπεδο που εκτελεί την multi-head attention λαμβάνει σαν είσοδο την ακολουθία εξόδου που έχει προβλέψει ο αποκωδικοποιητής μέχρι εκείνο το σημείο. Παρόλα αυτά, αυτό ισχύει στη φάση της εκτίμησης (inference) του μοντέλου, όπου γίνεται η χρήση του εκπαιδευμένου μοντέλου σε δεδομένα που δεν χρησιμοποιήθηκαν κατά τη φάση της εκπαίδευσης. Με αυτό τον τρόπο, η πρόβλεψη του αποκωδικοποιητή γίνεται στοιχείο προς στοιχείο. Αυτό σημαίνει ότι για τη πρόβλεψη ενός στοιχείου της ακολουθίας εξόδου σε κάποιο βήμα, ο αποκωδικοποιητής πρέπει πρώτα να έχει προβλέψει όλα τα στοιχεία της ακολουθίας εξόδου στα προηγούμενα βήματα.

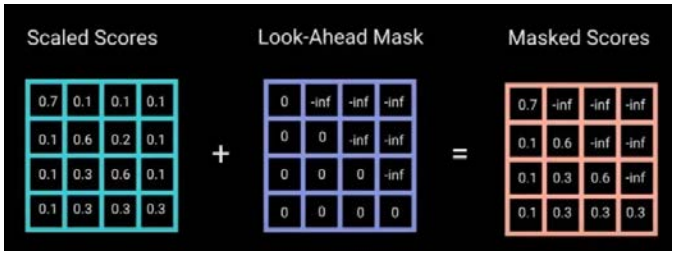
Στη φάση της εκπαίδευσης, όπου υπάρχει μεγάλος όγκος δεδομένων, αυτή η διαδικασία πρόβλεψης δεν είναι αποδοτική καθώς προσδίδει καθυστέρηση στην

εκπαίδευση του μοντέλου. Αυτό αντιμετωπίζεται με παράλληλο τρόπο και έτσι επιτυγχάνεται η βελτιστοποίηση στον τρόπο εκπαίδευσης του μοντέλου. Με άλλα λόγια, η πρόβλεψη ενός στοιχείου σε κάποιο βήμα είναι ανεξάρτητη από την πρόβλεψη των στοιχείων στα προηγούμενα βήματα. Για το σκοπό αυτό, στη διαδικασία της εκπαίδευσης, ο αποκωδικοποιητής λαμβάνει ολόκληρη την ακολουθία στόχος και με βάση αυτή προσπαθεί να κάνει τη σωστή πρόβλεψη του επόμενου βήματος.

3.8.2.1 Μηχανισμός Προσοχής Αποκωδικοποιητή

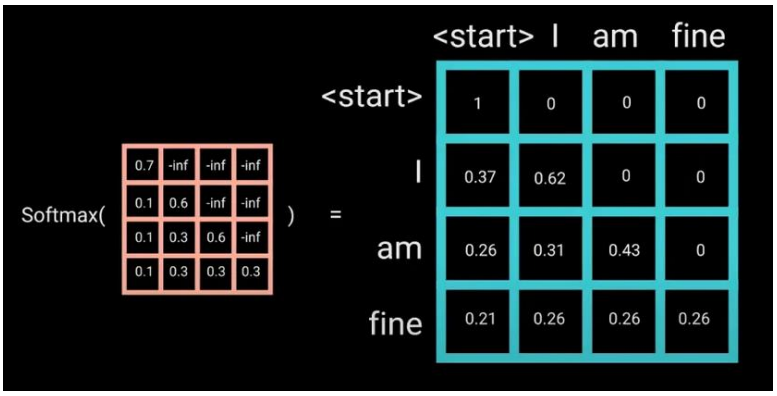
Masked Multi-Head Attention

Επισημαίνεται ότι ο όρος «ακολουθία εισόδου» σε αυτό το κεφάλαιο αναφέρεται στην ακολουθία που εισέρχεται στον αποκωδικοποιητή, ώστε να μην συγχέεται με την ακολουθία εισόδου του μοντέλου του κωδικοποιητή. Το πρώτο υπό-επίπεδο του αποκωδικοποιητή εφαρμόζει τον μηχανισμό προσοχής Masked Multi-Head στην ακολουθία που λαμβάνει σαν είσοδο (ακολουθία εισόδου), η οποία όπως αναφέρθηκε στην προηγούμενη παράγραφο αποτελεί την ακολουθία στόχος (target sequence), στη φάση της εκπαίδευσης. Στόχος της συγκεκριμένης υπομονάδας αποτελεί η κατανόηση των σχέσεων των λέξεων στην ακολουθία στόχο. Για το λόγο αυτό εφαρμόζεται ο μηχανισμός αυτό-προσοχής, όπου οι πίνακες Q, K, V σχηματίζονται από την ίδια την ακολουθία εισόδου και περιέχουν διανύσματα q,k και v, διάστασης d_{model}. Όπως περιεγράφηκε και στο κεφάλαιο 4.8.1.1, λόγω του μηχανισμού προσοχής multi-head, κάθε διάνυσμα q, k και v προβάλλεται διαμέσου γραμμικών μετασχηματιστών σε διανύσματα μικρότερης διάστασης d_k, d_k και d_v αντίστοιχα. Στα διανύσματα αυτά εφαρμόζεται παράλληλα και ανεξάρτητα ο μηχανισμός αυτό-προσοχής. Ο πίνακας προσοχής που θα προκύψει εφαρμόζοντας τον μηχανισμό αυτό-προσοχής στην ακολουθία στόχο, υπολογίζει τις σχέσεις των λέξεων μεταξύ τους. Με αυτό τον τρόπο, υπολογίζονται οι τιμές προσοχής τόσο ενός στοιχείου της ακολουθίας με τα προηγούμενα στοιχεία, όσο και με τα στοιχεία που το ακολουθούν. Αυτός ο τρόπος εκτέλεσης του μηχανισμού προσοχής δημιουργεί πρόβλημα στην πρόβλεψη του αποκωδικοποιητή καθώς αυτή δεν πρέπει να επηρεάζεται από τα στοιχεία της ακολουθίας που ακολουθούν.



Εικόνα 21 - Εφαρμογή της μάσκας στον πίνακα προσοχής (Πηγή: https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0)

Λύση σε αυτό το πρόβλημα προσφέρει η χρήση μίας μάσκας, η οποία εφαρμόζεται στον πίνακα προσοχής, πριν την εφαρμογή της συνάρτησης SoftMax. Ως μάσκα χρησιμοποιείται ένας άνω τριγωνικός πίνακας που έχει διάσταση ίση με το μέγεθος της ακολουθίας εισόδου. Τα στοιχεία κάτω από την κύρια διαγώνιο ισούνται με 0 και αυτά της άνω διαγώνιου ισούνται με ένα πολύ μεγάλο αρνητικό αριθμό ($-\infty$). Η μάσκα προστίθεται στον πίνακα προσοχής. Αυτό έχεις σαν αποτέλεσμα, τα στοιχεία κάτω από την κύρια διαγώνιο να περιλαμβάνουν τις αντίστοιχες τιμές από τις αντίστοιχες θέσεις που πίνακα προσοχής, όπως αποτυπώνεται στην Εικόνα 21. Τα στοιχεία της άνω διαγώνιου ισούνται με το $-\infty$. Η εφαρμογή της συνάρτησης SoftMax στον πίνακα αυτό έχει σαν αποτέλεσμα τον υπολογισμό των ποσοστών συσχέτισης κάθε λέξης με τις προηγούμενές της.



Εικόνα 22 - Εφαρμογή της συνάρτησης SoftMax στο αποτέλεσμα που προκύπτει μετά την εφαρμογή της μάσκας.(Πηγή https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0)

Αυτό συμβαίνει καθώς τα σκορ προσοχής των μεταγενέστερων θέσεων ισούται με έναν μεγάλο αρνητικό αριθμό με αποτέλεσμα η συνάρτηση SoftMax να μην τα λαμβάνει υπόψη για τον υπολογισμό των βαρών προσοχής, όπως διακρίνεται στην Εικόνα 22. Στη συνέχεια, ο πίνακας που προκύπτει πολλαπλασιάζεται με τον πίνακα τιμών V, που περιέχει την πληροφορία για κάθε στοιχείο της εισόδου. Μετά τον πολλαπλασιασμό, ο πίνακας που προκύπτει περιλαμβάνει τις τιμές για κάθε στοιχείο

της ακολουθίας εισόδου, που προκύπτουν λαμβάνοντας την πληροφορία συσχέτισης μόνο από προηγούμενα στοιχεία της ακολουθίας. Κατά αυτό τον τρόπο, ο αποκωδικοποιητής λαμβάνει σε κάθε βήμα αποκλειστικά την πληροφορία από τα προηγούμενα βήματα για τη πρόβλεψη του επόμενου στοιχείου της ακολουθίας.

Ο παραπάνω μηχανισμός προσοχής εφαρμόζεται παράλληλα h φορές (h = 8). Στο τέλος, οι πίνακες που προκύπτουν περιέχουν διανύσματα διάστασης $d_k=d_v$ = 64, που συνενώνονται και παράγουν πίνακα με διάσταση d_{model}. Ακολουθεί ένας γραμμικός μετασχηματισμός και στη συνέχεια η πρόσθεση της εισόδου του ακωδικοποίητή μέσω της υπολειπόμενης σύνδεσης Τέλος, εφαρμόζεται ένα στρώμα κανονικοποίησης. Το αποτέλεσμα από το στρώμα κανονικοποίησης αποτελεί τη μία από τις δύο εισόδους της δεύτερης υπομονάδας του κωδικοποιητή.

Encoder – Decoder Multi - Head Attention

Η δεύτερη υπομονάδα του αποκωδικοποιητή εκτελεί ακόμη έναν μηχανισμό Multi-Head, στον οποίο εφαρμόζεται ο μηχανισμός προσοχής cross- attention. Ο μηχανισμός αυτός επιτρέπει τον κωδικοποιητή να λάβει πληροφορία από δύο διαφορετικές ακολουθίες, την αρχική ακολουθία εισόδου του μετασχηματιστή και την ακολουθία εξόδου του αποκωδικοποιητή, μέχρι το τρέχον βήμα της πρόβλεψης. Η έξοδος από το πρώτο υπό-επίπεδο του αποκωδικοποιητή χρησιμοποιείται ως πίνακας Q και η έξοδος του κωδικοποιητή αποτελεί τους πίνακες K,V. Στη συνέχεια, εφαρμόζεται ο μηχανισμός Multi-Head, όπως περιεγράφηκε στο κεφάλαιο 4.8.1.1. Ακολουθεί μία υπολειπόμενη σύνδεση και ένα στρώμα κανονικοποίησης. Τέλος, το αποτέλεσμα εισέρχεται στην τρίτη και τελευταία υπομονάδα του αποκωδικοποιητή, όπου υπόκειται σε περαιτέρω επεξεργασία διαμέσου του μοντέλου position-wise feed forward, ίδιας αρχιτεκτονικής με του κωδικοποιητή.

Στην έξοδο της τελευταίας υπομονάδας του αποκωδικοποιητή εφαρμόζεται ένας γραμμικός μετασχηματισμός, που αποτελεί το επίπεδο ταξινόμησης. Αναλόγως το πλήθος τον κλάσεων περιέχει και το ανάλογο πλήθος παραμέτρων. Για παράδειγμα, σε περίπτωση της μηχανικής μετάφρασης το πλήθος των κλάσεων ισούται με το πλήθος του λεξιλογίου της γλώσσας στην οποία μεταφράζεται η είσοδος. Στη συνέχεια, εφαρμόζεται η συνάρτηση SoftMax για τη δημιουργία πιθανοτήτων (με τιμές μεταξύ) 0 και 1. Η μεγαλύτερη τιμή αντιστοιχεί και στην προβλεπόμενη λέξη. Τέλος, η

προβλεπόμενη τιμή εισέρχεται στη λίστα της ακολουθίας εξόδου του αποκωδικοποιητή, ώστε να χρησιμοποιηθεί στην πρόβλεψη της επόμενης λέξης.

3.8.3 Αναπαραστάσεις Εισόδου

Οι αναπαραστάσεις εισόδου (input embeddings) είναι ο τρόπος αναπαράστασης των λέξεων ως διανύσματα στο πολυδιάστατο χώρο, όπου η απόσταση και η κατεύθυνση μεταξύ δύο διανυσμάτων προσδίδουν την ομοιότητα και τη σχέση μεταξύ των δύο αντίστοιχων λέξεων. Έτσι, επιτυγχάνεται η αναπαράσταση των λέξεων με ουσιαστικό τρόπο, επιτρέποντας στο νευρωνικό δίκτυο να αναλύσει και να κατανοήσει τα δεδομένα εισόδου αποτελεσματικά.

Κατά κύριο λόγο, σαν είσοδο στο μοντέλο λαμβάνονται προτάσεις, οι οποίες αποτελούνται από λέξεις. Ένα μοντέλο, όμως, μηχανικής μάθησης δεν μπορεί να κατανοήσει τις λέξεις αυτούσιες. Αντί αυτού, οι λέξεις πρέπει να αναπαρασταθούν με αριθμούς. Μία παραδοσιακή μέθοδο αναπαράστασης λέξεων, με τέτοιο τρόπο που να μπορεί ένα μοντέλο να κατανοήσει, αποτελεί η μέθοδος one-hot-encoding. Σύμφωνα με τη μέθοδο αυτή, η αναπαράσταση κάθε λέξης αποτελεί έναν αραιό πίνακα με διάσταση ίση με το μέγεθος του λεξιλογίου. Σε αυτόν τον πίνακα μόνο ένα στοιχείο ισούται με 1 και καθορίζει την ύπαρξη της συγκεκριμένης λέξης. Αυτός ο τρόπος αναπαράστασης αν και είναι απλός στερείται σημασιολογικών πληροφοριών και δεν αποτυπώνει τις σχέσεις μεταξύ των λέξεων.

Οι αναπαραστάσεις λέξεων (word embeddings) από την άλλη μεριά, αποτελούν πυκνούς πίνακες με συνεχείς τιμές, οι οποίοι μαθαίνονται από ένα νευρωνικό δίκτυο. Στην εργασία [36], τις αναπαραστάσεις εισόδου μαθαίνει ο μετασχηματιστής κατά τη διαδικασία της εκπαίδευσης. Για το λόγο αυτό, υπάρχουν αντίστοιχα επίπεδα, πριν την είσοδο του κωδικοποιητή και του αποκωδικοποιητή, που μετατρέπουν τις ακολουθίες λέξεων σε αναπαραστάσεις λέξεων, πριν εισαχθούν στο μοντέλο. Η ακολουθία λέξεων αρχικά διαχωρίζεται σε σύμβολα (tokens). Στη συνέχεια, τα tokens εισόδου και τα tokens εξόδου μετατρέπονται σε διανύσματα διάστασης d_{model}. Τα δύο στρώματα αναπαραστάσεων (input embedding layer) και το γραμμικό στρώμα (linear layer), το οποίο βρίσκεται στο επίπεδο της ταξινόμησης, μοιράζονται τον ίδιο πίνακα βαρών. Αυτή η μέθοδος ονομάζεται Three-Way Weight Tying -TWWT. Στην εργασία [40], που προτάθηκε αυτός ο τρόπος διαχείρισης των αναπαραστάσεων, παρατηρήθηκε ότι κατά την επεξεργασία δύο διαφορετικών συνόλων δεδομένων, όπου το ένα περιείχε

μεταφράσεις από αγγλικά σε γαλλικά και το δεύτερο από αγγλικά σε γερμανικά, υπήρχε πληθώρα κοινών υπό-λέξεων. Έτσι, προτάθηκε η μέθοδος αυτή, σύμφωνα με την οποία οι αναπαραστάσεις εισόδου του αποκωδικοποιητή, οι αναπαραστάσεις εξόδου του αποκωδικοποιητή και οι αναπαραστάσεις εισόδου του κωδικοποιητή μοιράζονται τα ίδια βάρη. Σε ένα μοντέλο που επιλέγεται η χρήση αυτής της μεθόδου, το λεξιλόγιο πηγής/στόχου αυτού του μοντέλου είναι η ένωση του λεξιλογίου πηγής και του λεξιλογίου στόχου. Σε αυτό το μοντέλο, τόσο στον κωδικοποιητή όσο και στον αποκωδικοποιητή, όλες οι υπό-λέξεις είναι ενσωματωμένες στον ίδιο διαγλωσσικό χώρο.

Το επίπεδο αναπαράστασης περιέχει τις αναπαραστάσεις κάθε λέξης του λεξιλογίου ανάλογα με το εκάστοτε πρόβλημα. Κάθε αναπαράσταση αποτελεί ένα διάνυσμα διάστασης d_{model}. Κάθε ακολουθία εισόδου, όπως αναφέρθηκε και νωρίτερα, διαχωρίζεται σε σύμβολα. Κάθε σύμβολο αποτελεί μία λέξη στο λεξιλόγιο και συνεπώς εκφράζεται στο επίπεδο αναπαράστασης. Κάθε σύμβολο έχει μια συγκεκριμένη θέση στο λεξιλόγιο η οποία υποδηλώνεται από ένα συγκεκριμένο ακέραιο δείκτη. Το επίπεδο αναπαράστασης αποτελεί έναν πίνακα αναζήτησης (look-up-table) σύμφωνα με την ακέραια τιμή του δείκτη, επιλέγεται η αντίστοιχη γραμμή από τον πίνακα των αναπαραστάσεων. Με αυτό τον τρόπο κάθε ακολουθία συμβόλων μετατρέπεται στην αντίστοιχη ακολουθία από ακέραιους δείκτες, αναλόγως της θέσης τους στο λεξιλόγιο, και τελικά μετατρέπεται σε ακολουθία αναπαραστάσεων, σύμφωνα τις τιμές των δεικτών. Έτσι, επιτυγχάνεται η δημιουργία των αναπαραστάσεων εισόδου. Σε αυτές, προστίθενται τα διανύσματα θέσης προκειμένου να προσδώσουν την πληροφορία της σειράς στην ακολουθία. Στο κεφάλαιο που ακολουθεί αναλύεται ο τρόπος αναπαράστασης των θέσεων στην αρχιτεκτονική του μετασχηματιστή.

3.8.4 Αναπαραστάσεις Θέσης

Το μοντέλο του μετασχηματιστή δεν περιλαμβάνει κάποιο ακολουθιακό μηχανισμό ή κάποια συνέλιξη. Για το λόγο αυτό δεν μπορεί να λάβει υπόψη τη πληροφορία της σειράς στις ακολουθίες εισόδου. Ωστόσο, η σειρά των λέξεων σε μία πρόταση είναι σημαντική για την κατανόηση της σημασίας της. Οι αρχιτεκτονικές RNN μοντέλων, έχουν ενσωματωμένο τον μηχανισμό που λαμβάνει υπόψη τη σειρά της ακολουθίας καθώς την επεξεργάζονται ακολουθιακά και κάθε κρυφή κατάσταση περιλαμβάνει την πληροφορία όλων των προηγούμενων καταστάσεων. Στην περίπτωση των

μετασχηματιστών, η πληροφορία της σειράς ενσωματώνεται με τη χρήση των αναπαραστάσεων θέσης (positional embeddings), που περιγράφουν τη θέση κάθε στοιχείου της ακολουθίας με τέτοιο τρόπο ώστε κάθε θέση να λαμβάνει μία μοναδική αναπαράσταση.

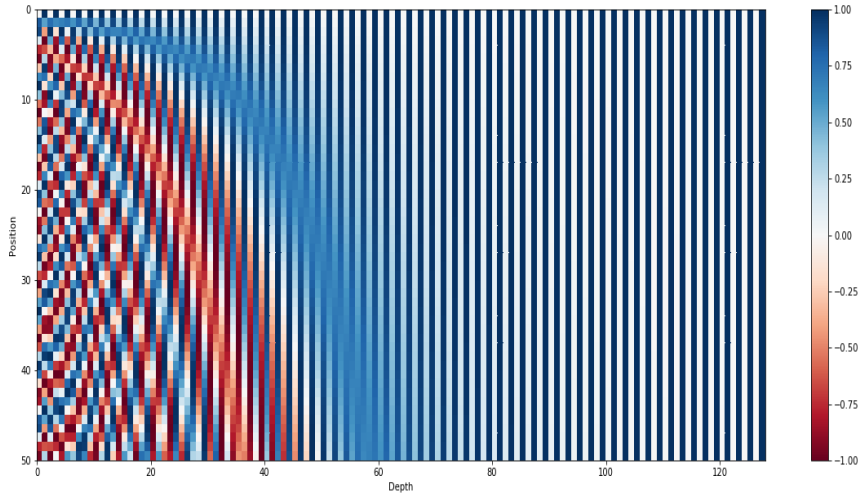
Υπάρχουν πολλοί τρόποι για τον σχηματισμό των αναπαραστάσεων αυτών. Μπορεί να αποτελούν αναπαραστάσεις που μαθαίνονται από το ίδιο το μοντέλο κατά τη διάρκεια της εκπαίδευσης ή αναπαραστάσεις που υπολογίζονται από κάποια μαθηματική συνάρτηση. Οι αναπαραστάσεις θέσης πρέπει να πληρούν ορισμένες προϋποθέσεις. Αρχικά, πρέπει να είναι μοναδικές για κάθε θέση δίχως να μεταβάλλονται με το μήκος ή τις τιμές της ακολουθίας εισόδου. Σε περίπτωση που η ακολουθία εισόδου αλλάξει τόσο σε περιεχόμενο όσο και σε μήκος, οι αναπαραστάσεις θέσης που περιγράφουν κάθε θέση πρέπει να παραμείνουν οι ίδιες. Ο δεύτερος περιορισμός έγκειται στο γεγονός ότι οι αναπαραστάσεις αυτές προστίθενται στις αναπαραστάσεις της εισόδου και δεν πρέπει να αποτελούνται από μεγάλους αριθμούς που «αλλοιώνουν » την πληροφορία των αναπαραστάσεων της εισόδου. Με άλλα λόγια, η επιλογή μεγάλων τιμών στις αναπαραστάσεις θέσης δίνει περισσότερη πληροφορία για την εκάστοτε θέση αντί για την πληροφορία και τη σημασία κάθε λέξης. Έτσι, στην εργασία [36], για τον υπολογισμό των αναπαραστάσεων θέσης χρησιμοποιείται η συνημιτονοειδής και η ημιτονοειδής συνάρτηση, σύμφωνα με τις παρακάτω σχέσεις:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (26)$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (27)$$

όπου pos εκφράζει τη θέση στην ακολουθία εισόδου και i τη διάσταση. Με αυτό τον τρόπο, οι τιμές που μπορεί να λάβει κάθε αναπαράσταση θέσης είναι οριοθετημένες μεταξύ -1 και 1. Η διάσταση των αναπαραστάσεων θέσης στην εργασία [36] ισούται με d_{model}. Όταν το i στο διάνυσμα θέσης είναι ζυγός αριθμός , τότε η τιμή της διάστασης i προκύπτει με τη χρήση του τύπου (26). Για περιττό αριθμό i στο διάνυσμα θέσης , η τιμή ακολουθεί τη σχέση (27). Με αυτό το τρόπο οι τιμές σε κάθε διάνυσμα θέσης είναι ακολουθία από ζεύγη των συναρτήσεων (26), (27), που χαρακτηρίζονται από την ίδια συχνότητα. Όσο αυξάνεται η διάσταση τόσο η μειώνεται η συχνότητα. Όπως φαίνεται

και στην Εικόνα 23 οι τιμές σε μεγαλύτερες διαστάσεις μεταβάλλονται πιο αργά σε σχέση με τις τιμές στις χαμηλότερες διαστάσεις.



Εικόνα 23 - Αναπαραστάσεις θέσης διάστασης 128 για πρόταση μεγέθους 50 στοιχείων (Κάθε γραμμή αντιπροσωπεύει μία αναπαράσταση θέσης) [36]

Η χρήση των δύο τριγωνομετρικών συναρτήσεων γίνεται για δύο λόγους. Σε περίπτωση επιλογής, για παράδειγμα μόνο της χρήσης της ημιτονοειδούς συνάρτησης, λόγω της περιοδικότητας, σε κάποιες θέσεις θα λάμβανε την ίδια τιμή. Έτσι, η ίδια αναπαράσταση θα αντιστοιχούσε σε ένα δύο ή περισσότερες θέσεις. Με τη χρήση των δύο συναρτήσεων επιτυγχάνεται η μοναδική αναπαράσταση κάθε διανύσματος θέσης. Ο δεύτερος λόγος, όπως αναφέρεται και στην εργασία [36], έγκειται στο γεγονός ότι διευκολύνει το μοντέλο να μάθει τις σχετικές θέσεις των δεδομένων εισόδου. Αυτό γιατί κάθε θέση PE_{pos+k} μπορεί να αναπαρασταθεί ως γραμμικός μετασχηματισμός της θέσης PE_{pos} . Έτσι, ο μετασχηματιστής μαθαίνει να δίνει «προσοχή» στα στοιχεία της ακολουθίας εισόδου λαμβάνοντας υπόψη τη θέσης τους.

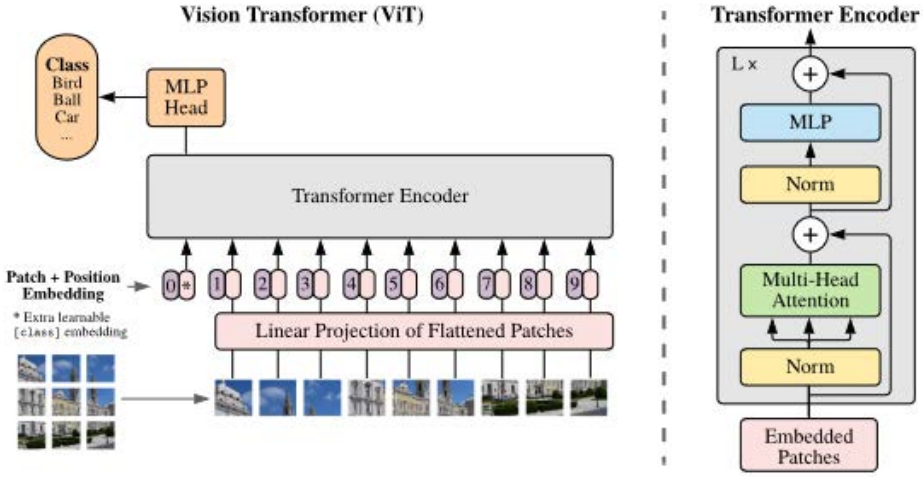
Συνεπώς, με την ενσωμάτωση της πληροφορίας της θέσης σε κάθε στοιχείο της πρότασης, το μοντέλο του μετασχηματιστή μαθαίνει, μέσω του μηχανισμού προσοχής, τις συσχετίσεις ανάμεσα στα στοιχεία, λαμβάνοντας υπόψη τη σειρά τους μέσα στην πρόταση. Η ενσωμάτωση της πληροφορίας της θέσης πραγματοποιείται με την πρόσθεση των αναπαραστάσεων θέσης στις αντίστοιχες αναπαραστάσεις κάθε στοιχείου της ακολουθίας εισόδου (input embeddings). Σημειώνεται σε αυτό το σημείο ότι και τα δύο είδη αναπαραστάσεων πρέπει να έχουν την ίδια διάσταση για την εκτέλεση της πράξης της πρόσθεσης μεταξύ τους.

3.9 Οπτικοί Μετασχηματιστές

Οι μετασχηματιστές, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, αρχικά εφαρμόστηκαν σε εφαρμογές NLP, πετυχαίνοντας σημαντικές βελτιώσεις συγκριτικά με τα αναδρομικά δίκτυα. Το 2018, ερευνητές της Google παρουσίασαν το BERT (Bidirectional Encoder Representations from Transformers), ένα μοντέλο που βασίζεται στην αρχιτεκτονική των μετασχηματιστών και έθεσε νέα πρότυπα σε 11 βασικές εφαρμογές κατανόησης της φυσικής γλώσσας (Natural Language Understanding- NLU), όπως η ανάλυση συναισθήματος, η ταξινόμηση κειμένου και η αποσαφήνιση λέξεων με πολλαπλές έννοιες [37]. Εκτός από το BERT, η επεξεργασία φυσικής γλώσσας έχει επωφεληθεί σημαντικά από την ανάπτυξη του GPT-3 (Generative Pre-trained Transformer) [41]. Αυτό το μοντέλο, βασισμένο επίσης στην αρχιτεκτονική των μετασχηματιστών, έχει ξεχωρίσει για την ικανότητά του να επιτυγχάνει κορυφαία αποτελέσματα σε διάφορες βασικές εφαρμογές NLP, όπως η μετάφραση μηχανής, η συγγραφή περιλήψεων και η απάντηση σε ερωτήσεις, χωρίς να απαιτείται fine-tuning [41].

Η μεγάλη επιτυχία της αρχιτεκτονικής των μετασχηματιστών στο χώρο της επεξεργασίας της φυσικής γλώσσας, ώθησε πολλούς ερευνητές στην αξιοποίηση τους σε εφαρμογές υπολογιστικής όρασης. Μέχρι τότε, τα συνελικτικά νευρωνικά δίκτυα αποτελούσαν την κυρίαρχη προσέγγιση για την επίτευξη υψηλής ακρίβειας σε εργασίες υπολογιστικής όρασης. Οι μετασχηματιστές πρόσφεραν μία εναλλακτική δίοδο. Το 2021 στο συνέδριο ICLR παρουσιάστηκε το μοντέλο του οπτικού μετασχηματιστή (Vision Transformer - ViT) [42]. Στόχος της εργασίας αποτέλεσε η εφαρμογή της αρχιτεκτονικής του μετασχηματιστή απευθείας σε εικόνες, διατηρώντας, όσο αυτό είναι δυνατόν, την αρχική αρχιτεκτονική των μετασχηματιστών. Για τον σκοπό αυτό, η εικόνα διαχωρίζεται σε τμήματα (patches), τα οποία μέσω του γραμμικού μετασχηματισμού μετατρέπονται σε γραμμικές αναπαραστάσεις (linear embeddings). Η ακολουθία των γραμμικών αναπαραστάσεων τροφοδοτείται στη συνέχεια στον μετασχηματιστή ο οποίος μαθαίνει να εντοπίζει τις σχέσεις και τις αλληλεπιδράσεις μεταξύ των τμημάτων [42].

3.9.1 Αρχιτεκτονική του Οπτικού Μετασχηματιστή



Εικόνα 24 – (Αριστερά) Η αρχιτεκτονική του οπτικού μετασχηματιστή, (Δεξιά) Το επίπεδο του κωδικοποιητή [42]

Η Εικόνα 24 προσφέρει μια επισκόπηση του μοντέλου του οπτικού μετασχηματιστή. Η αρχιτεκτονική αποτελείται από ένα γραμμικό επίπεδο, ένα επίπεδο κωδικοποιητή και ένα δίκτυο εμπρόσθιας τροφοδότησης πολλών επιπέδων (Multi-Layer Perceptrons-MLPs) . Στα υπό-κεφάλαια που ακολουθούν θα πραγματοποιηθεί η περιγραφή κάθε επιπέδου της αρχιτεκτονικής αυτής.

3.9.1.1 Επίπεδο Εισόδου

Η μονάδα του κωδικοποιητή στους μετασχηματιστές δέχεται σαν είσοδο μία μονοδιάστατη ακολουθία από αναπαραστάσεις (embeddings) κάθε στοιχείου της εισόδου. Για αυτούς τους λόγους, για να καταστεί δυνατή η τροφοδότηση του μοντέλου με δεδομένα εισόδου, μία εικόνα $x \in \mathbb{R}^{H \times w \times C}$ διαχωρίζεται σε N δισδιάστατα τμήματα (patches), όπου κάθε τμήμα $p \in \mathbb{R}^{P \times P \times C}$. Στη συνέχεια τα τμήματα αυτά μετατρέπονται σε διανύσματα αναπαραστάσεων $x_p \in \mathbb{R}^{N \times (P^2 * C)}$, όπου (H, W) είναι η ανάλυση της αρχικής εικόνας, (P, P) είναι η ανάλυση του κάθε τμήματος της εικόνας και $N = HW \ / \ P^2$ είναι το πλήθος των τμημάτων που προκύπτουν κατά το διαχωρισμό της εικόνας (H, W) σε τμήματα (P, P).

Τα τμήματα που προκύπτουν, απλώνονται (flatten) στα αντίστοιχα διανύσματα αναπαράστασης. Το μοντέλο του οπτικού μετασχηματιστή (όπως και του μετασχηματιστή) χρησιμοποιεί σταθερή διάσταση D σε όλα τα επίπεδά του (αντίστοιχα d_{modeL} στην αρχιτεκτονική του μετασχηματιστή). Συνεπώς, τα διανύσματα που

προκύπτουν από την διαδικασία flatten ενδέχεται να έχουν διάσταση που διαφέρει από την επιθυμητή διάσταση D. Για το σκοπό αυτό, μετατρέπονται σε γραμμικές αναπαραστάσεις διάστασης D μέσω μίας γραμμικής προβολής, της οποίας τις παραμέτρους μαθαίνει το μοντέλο κατά τη διάρκεια της εκπαίδευσης. Η γραμμική προβολή εφαρμόζεται με την ακόλουθη σχέση:

$$z_0 = x_P \times E \;\; \text{όπου } E \in \mathbb{R}^{(P^2 \cdot C) \times N} \quad (28)$$

Κάθε στήλη του πίνακα z_0 αποτελεί τη γραμμική αναπαράσταση κάθε τμήματος της αρχικής εικόνας και συνολικά οι στήλες συνθέτουν τις αναπαραστάσεις τμημάτων (patch embeddings).

Ακολουθώντας το παράδειγμα του μοντέλου BERT, γίνεται η χρήση του συμβόλου *[class]*, το οποίο αποτελεί μία αναπαράσταση την οποία μαθαίνει το μοντέλο κατά τη διάρκεια της εκπαίδευσης και την προσθέτει στην αρχή τις ακολουθίας των αναπαραστάσεων των τμημάτων ($z_0^0 = x_{class}$). Συνεπώς, η ακολουθία στη σχέση (28) μετατρέπεται ως εξής:

$$z_0 = \left[x_{class}; x_p^1 E; x_P^2 E; \ldots; x_p^N E\right] \quad (29)$$

όπου $x_{class} \in \mathbb{R}^D$. Έτσι, η ακολουθία z_0 περιλαμβάνει $N + 1$ αναπαραστάσεις.

Το σύμβολο κλάσης (class token), αρχικοποιείται με τυχαίο τρόπο και δεν περιλαμβάνει κάποια χρήσιμη πληροφορία. Ωστόσο, κατά την εκτέλεση του μηχανισμού προσοχής στη μονάδα του κωδικοποιητή συσσωρεύει την πληροφορία από τις υπόλοιπες αναπαραστάσεις της ακολουθίας. Συνεπώς, στο τελευταίο επίπεδο του κωδικοποιητή, το στοιχείο κλάσης περιλαμβάνει την απαραίτητη πληροφορία για την αναπαράσταση της εικόνας. Κατά τη διαδικασία της ταξινόμησης, το δίκτυο εμπρόσθιας τροφοδότησης πολλών επιπέδων χρησιμοποιεί μόνο το στοιχείο κλάσης από την έξοδο του κωδικοποιητή (z_L^0) προκειμένου να πραγματοποιήσει την πρόβλεψη. Με αυτή τη λειτουργία, το στοιχείο αυτό αποτελεί μία δομή που χρησιμοποιείται για την αποθήκευση της πληροφορίας που εξάγεται από τα υπόλοιπα στοιχεία της ακολουθίας. Με αυτόν τον τρόπο, ελαχιστοποιείται η επίδραση της τελικής πρόβλεψης του μοντέλου αποκλειστικά και μόνο προς ή κατά κάποιας από τις υπόλοιπες αναπαραστάσεις τμημάτων.

Όπως και στην περίπτωση του μετασχηματιστή, έτσι και στην αρχιτεκτονική του οπτικού μετασχηματιστή είναι απαραίτητη η προσθήκη των αναπαραστάσεων θέσης (positional embeddings) στις αναπαραστάσεις τμημάτων (patch embeddings). Οι μετασχηματιστές είναι μοντέλα που βασίζονται στον μηχανισμό προσοχής και δεν περιέχουν κάποιο μηχανισμό που προσδίδει την πληροφορία της θέσης (όπως αναφέρθηκε αναλυτικότερα στο κεφάλαιο 4.8.4). Στην περίπτωση του οπτικού μετασχηματιστή, επιλέχτηκε η χρήση μονοδιάστατων αναπαραστάσεων θέσης, τις οποίες μαθαίνει το μοντέλο κατά την διάρκεια της εκπαίδευσης. Οι αναπαραστάσεις αυτές προστίθενται στην αναπαραστάσεις τμημάτων και τροφοδοτούν τη μονάδα του κωδικοποιητή. Οι τελικές αναπαραστάσεις που αποτελούν την είσοδο του κωδικοποιητή αποτυπώνεται στην παρακάτω σχέση (30):

$$z_0 = \left[x_{class}; x_p^1 E; x_P^2 E; \cdots x_P^N E\right] + Epos \quad (30)$$

όπου $Epos \in \mathbb{R}^{(N+1)\times D}$ αποτελεί τις αναπαραστάσεις θέσης για κάθε θέση της ακολουθίας z_0.

3.9.1.2 Κωδικοποιητής Οπτικού Μετασχηματιστή

Ο κωδικοποιητής, όπως και στην περίπτωση του μετασχηματιστή αποτελείται από L πανομοιότυπα επίπεδα τοποθετημένα το ένα μετά το άλλο. Κάθε επίπεδο αποτελείται από ένα υπό-επίπεδο που εκτελεί τον μηχανισμό της προσοχής multi-head, όπως ακριβώς παρουσιάστηκε στην εργασία [26] και ένα υπό-επίπεδο που περιλαμβάνει ένα δίκτυο εμπρόσθιας τροφοδότησης πολλών επιπέδων (MLP) . Ένα στρώμα κανονικοποίησης εφαρμόζεται πριν από την είσοδο κάθε υπό-μονάδας. Επιπλέον, προστίθενται υπολειπόμενες συνδέσεις γύρω από κάθε υπό-επίπεδο.

Ο μηχανισμός προσοχής multi-head, παρουσιάζεται στην εργασία ως μηχανισμός αυτοπροσοχής multi-head (Multi-head Self Attention- MSA). Στην περίπτωση του οπτικού μετασχηματιστή με τον μηχανισμό προσοχής μαθαίνει το μοντέλο τη σχέση κάθε τμήματος της εικόνας με τα υπόλοιπα τμήματα. Τα τμήματα που έχουν τη μεγαλύτερη βαρύτητα είναι και αυτά που θα λάβει υπόψη το μοντέλο στην τελική πρόβλεψη. Για κάθε στοιχείο στην ακολουθία εισόδου $z \in \mathbb{R}^{N\times D}$ υπολογίζεται το σταθμισμένο άθροισμα όλων των τιμών v της ακολουθίας εισόδου. Τα διανύσματα q,

k, v υπολογίζονται από τον γραμμικό μετασχηματισμό της ακολουθίας εισόδου, όπως φαίνεται στην σχέση (31):

$$[q, v, k] = z \cdot U_{qkv}, \text{ όπου } U_{qkv} \in \mathbb{R}^{D \times 3Dh} \quad (31)$$

Τα βάρη προσοχής, τα οποία στην εργασία συμβολίζονται ως A_{ij}, βασίζονται στην τιμή της ομοιότητας μεταξύ δύο στοιχείων της ακολουθίας εισόδου και χρησιμοποιούν το αντίστοιχο διάνυσμα αιτήματος q_i και κλειδιού k_i, όπως αποτυπώνεται στη σχέση (32):

$$A = softmax\left(qk^T / \sqrt{Dh}\right), \text{ όπου } A \in \mathbb{R}^{N \times N} \quad (32)$$

Σύμφωνα, με τις παραπάνω εξισώσεις εκτελείται ο μηχανισμός αυτοπροσοχής (SA) που περιγράφεται με την παρακάτω σχέση (33):

$$SA(z) = Av \quad (33)$$

Παρόλα αυτά, όπως και στην αρχιτεκτονική του μετασχηματιστή, δεν γίνεται η χρήση ενός μηχανισμού αυτοπροσοχής που εκτελείται στην αρχική διάσταση D. Αντίθετα, επιλέγεται η χρήση του μηχανισμού αυτοπροσοχής multi-head (περιεγράφηκε αναλυτικότερα στο κεφάλαιο 4.8.1.1) που αποτελείται από h κεφαλές. Κάθε κεφαλή εκτελεί τον μηχανισμό αυτοπροσοχής σύμφωνα με την σχέση (33) και σε διανύσματα διάστασης $D_h < D$. Οι κεφαλές λειτουργούν παράλληλα και ο κάθε μηχανισμός αυτοπροσοχής εκτελείται με ανεξάρτητο τρόπο. Τα αποτελέσματα των παράλληλων μηχανισμών αυτοπροσοχής συνενώνονται και προβάλλονται γραμμικά. Προκειμένου να διατηρηθούν οι παράμετροι του μοντέλου σταθερές, η διάσταση D_h ισούται με D / h. Η λειτουργία του μηχανισμού συνοψίζεται μαθηματικά, στη σχέση (34).

$$MSA(z) = [SA_1(z); SA_2(z); \cdots; SA_h(z)]U_{msa}, \; U_{msa} \in \mathbb{R}^{(h \cdot D_h) \times D} \quad (34)$$

Τα μοντέλα MLP αποτελούν δίκτυα πρόσθιας τροφοδότησης. Οι νευρώνες οργανώνονται σε στρώματα και δεν υπάρχουν συνδέσεις μεταξύ νευρώνων που ανήκουν στο ίδιο στρώμα. Κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου στρώματος. Στην αρχιτεκτονική του οπτικού μετασχηματιστή,

η υπό-μονάδα που περιλαμβάνει το μοντέλο MLP αποτελείται από δύο κρυφά στρώματα. Σε κάθε στρώμα εφαρμόζεται η μη γραμμική συνάρτηση ενεργοποίησης GELU (Gaussian Error Linear Unit). Η χρήση της συνάρτησης GELU εισάγει μη γραμμικότητα στο δίκτυο, επιτρέποντας την εκμάθηση πιο σύνθετων μοτίβων στις αναπαραστάσεις εισόδου.

3.9.1.3 Επίπεδο Εξόδου

Ο τρόπος που συνδέονται τα υπό-επίπεδα και τα επίπεδα κάθε κωδικοποιητή, όπως αναλύθηκαν παραπάνω, μπορεί να συνοψιστεί στις παρακάτω σχέσεις:

1. Τα απλωμένα (flattened) διανύσματα που προκύπτουν από κάθε τμήμα της εικόνας μετατρέπονται σε διανύσματα διάστασης D, μέσω του γραμμικού μετασχηματισμού, ενώ σε αυτά προστίθενται οι αναπαραστάσεις θέσης. Έτσι, η ακολουθία που λαμβάνει ο κωδικοποιητής και συγκεκριμένα το πρώτο επίπεδό του, είναι αυτή που περιεγράφηκε στην σχέση (30):

$$z_0 = \left[x_{class}; x_p^1 E; x_P^2 E; \cdots x_P^N E\right] + Epo \ .$$

2. Στη συνέχεια, στην ακολουθία εισόδου κάθε επιπέδου του κωδικοποιητή εφαρμόζεται ένα στρώμα κανονικοποίησης. Η έξοδος του στρώματος κανονικοποίησης εισέρχεται στο υπό-επίπεδο που εκτελεί τον μηχανισμό αυτοπροσοχής multi-head. Στο αποτέλεσμα του μηχανισμού αυτού προστίθεται, μέσω της υπολειπόμενης σύνδεσης, η είσοδος του εκάστοτε επιπέδου κωδικοποιητή. Κατά αυτό τον τρόπο, προκύπτει η ακολουθία που περιγράφεται από την παρακάτω σχέση:

$$z_l' = MSA\big(LN(z_{l-1})\big) + z_{l-1} \ , \quad l = 1, \dots, L \qquad (35)$$

3. Η ακολουθία z_l' εισέρχεται σε ένα στρώμα κανονικοποίησης και το αποτέλεσμα τροφοδοτεί το υπό-επίπεδο που περιλαμβάνει το μοντέλο MLP. Η ακολουθία z_l', μέσω της υπολειπόμενης σύνδεσης, προστίθεται απευθείας στην έξοδο του μοντέλου MLP. Η περιγραφή του βήματος 3 συνοψίζεται στην παρακάτω σχέση:

$$z_l = MLP\big(LN(z'_l)\big) + z'_l\ , \qquad l = 1, \dots, L \qquad (36)$$

4. Η ακολουθία z_L αποτελεί την έξοδο του κωδικοποιητή και συγκεκριμένα την τελευταία κρυφή κατάσταση του επιπέδου L. Το πρώτο διάνυσμα αυτής της ακολουθίας z_L^0 αφορά την αναπαράσταση του συμβόλου κλάσης. Είναι το διάνυσμα που περιέχει όλη την πληροφορία για την αναπαράσταση y της εικόνας. Σύμφωνα με την εργασία [42], η αναπαράσταση y δίνεται από τη σχέση (37):

$$y = LN(z_L^0) \qquad (37)$$

5. Τελικό βήμα αποτελεί η διαδικασία της ταξινόμησης. Η ταξινόμηση γίνεται μέσω του επιπέδου MLP head , που περιλαμβάνει ένα μοντέλο MLP, το οποίο λαμβάνει σαν είσοδο την τελική αναπαράσταση τη εικόνας y και εκτελεί την πρόβλεψη. Στην διαδικασία της προ-εκπαίδευσης, το μοντέλο MLP στο επίπεδο της ταξινόμησης περιλαμβάνει ένα κρυφό επίπεδο ακολουθούμενο από τη μη γραμμική συνάρτηση ενεργοποίησης tanh.

3.10 Μεταφορά Μάθησης και Fine - Tuning

Οι περισσότερες εφαρμογές στον κλάδο της βαθιάς μάθησης αξιοποιούν μοντέλα τα οποία έχουν εκπαιδευτεί προηγουμένως σε μεγάλα σύνολα δεδομένων, όπως για παράδειγμα το ImageNet22k, το οποίο περιέχει περισσότερες από 14 εκατομμύρια εικόνες χωρισμένες σε 20.000 διαφορετικές κατηγορίες [23]. Η πρακτική της αξιοποίησης της γνώσης που έχει αποκτήσει ένα προεκπαιδευμένο μοντέλο βαθιάς μάθησης για την εκπαίδευσή του σε ένα πιο μικρό και διαφορετικό σύνολο δεδομένων ονομάζεται μεταφορά γνώσης (transfer learning) [43]. Η δυσκολία στην απόκτηση πραγματικών δεδομένων, ειδικότερα στο τομέα της υγείας, καθιστά την πρακτική αυτή αναγκαία για την αποδοτική εκπαίδευση των μοντέλων σε σύνολα δεδομένων που περιέχουν περιορισμένο αριθμό δειγμάτων.

Η πιο συνηθισμένη πρακτική μεταφοράς γνώσης αποτελεί το fine-tuning. Αξιοποιώντας την προηγούμενη γνώση του προεκπαιδευμένου μοντέλου, επιδιώκεται η επίλυση ενός νέου προβλήματος [44]. Για παράδειγμα στο τομέα της επεξεργασίας φυσικής γλώσσας, η τεχνική fine-tuning χρησιμοποιείται συνήθως για την προσαρμογή μεγάλων προεκπαιδευμένων μοντέλων όπως για παράδειγμα το μοντέλο του BERT,

GPT κ.α σε κάποιο συγκεκριμένο πρόβλημα όπως για παράδειγμα τη ταξινόμηση κειμένου και την ανάλυση συναισθήματος. Με τη τεχνική του fine-tuning το μοντέλο αρχικοποιείται με τα προεκπαιδευμένα βάρη. Κατά τη διάρκεια της εκπαίδευσης, τα βάρη προσαρμόζονται στο καινούριο πρόβλημα αντί να εκπαιδεύονται από την αρχή. Κατά αυτό τον τρόπο, επιτυγχάνεται η γρηγορότερη εκπαίδευσή του μοντέλου καθώς δεν απαιτείται η εκμάθηση από την αρχή. Επιπλέον, βελτιώνεται η απόδοσή του και η ικανότητά του να γενικεύεται σε ένα πρόβλημα [44].

Ωστόσο, η τεχνική του fine-tuning εμφανίζει ορισμένους περιορισμούς. Η επιλογή της εκπαίδευσης όλων των παραμέτρων του μοντέλου απαιτεί πληθώρα υπολογιστικών πόρων αφού απαιτεί τη βελτιστοποίηση όλων των παραμέτρων. Αυτό περιβάλλον με περιορισμένους διαθέσιμους υπολογιστικούς πόρους είναι προβληματικό [44]. Επιπλέον, η εκπαίδευση ενός προεκπαιδευμένου μοντέλου με περιορισμένο αριθμό δειγμάτων μπορεί να οδηγήσει σε υπερπροσαρμογή στο δεδομένα εκπαίδευσης και στην αδυναμία γενίκευσης σε νέα δεδομένα, τα οποία δεν χρησιμοποιήθηκαν κατά την εκπαίδευση. Τέλος, ένας κίνδυνος είναι η μεγάλη αλλοίωση των αρχικών παραμέτρων που έχει μάθει το μοντέλο [44]. Αυτό για παράδειγμα μπορεί να συμβεί με τη χρήση μεγάλου ρυθμού εκμάθησης (learning rate) που οδηγεί σε μεγάλες ανανεώσεις των παραμέτρων του μοντέλου, με αποτέλεσμα να διαφέρουν αρκετά από τις αρχικές. Σε αυτήν την περίπτωση, το μοντέλο ενδέχεται να «ξεχάσει» την πληροφορία που έχει μάθει στην προηγούμενη εκπαίδευσή του [44].

Οι Hu *et al.* [45], αναγνωρίζοντας τους περιορισμούς που προκύπτουν στη χρήση της τεχνικής fine-tuning, παρουσίασαν τη μέθοδο προσαρμογής χαμηλής τάξης (Low-Rank Adaptation - LoRA). Η τεχνική LoRA βασίζεται στο αποτέλεσμα της έρευνας των Li *et al.* [46] και Aghajanyan *et al.* [47], οι οποίοι εξέτασαν στις εργασίες τους το ρόλο της εγγενούς διάστασης (intrinsic dimensionality) σε μεγάλα γλωσσικά μοντέλα (Large Language Model - LLMs). Ο όρος «εγγενή διάσταση» αναφέρεται στον ελάχιστο αριθμό διαστάσεων που απαιτείται προκειμένου μεγάλα μοντέλα να προσαρμοστούν στο νέο πρόβλημα. Οι εργασίες ανέδειξαν ότι με τη χρήση μικρότερης διάστασης παραμέτρων, τα μοντέλα αυτά μπορούν να προσαρμοστούν εξίσου αποτελεσματικά όσο με τη χρήση της τεχνικής fine-tuning σε όλες τις παραμέτρους των μοντέλων. Οι Hu et al, λοιπόν, βασιζόμενοι σε αυτό το πόρισμα, υπέθεσαν ότι η αλλαγή στις παραμέτρους των προεκπαιδευμένων μοντέλων κατά τη διάρκεια της προσαρμογής τους στο καινούριο πρόβλημα απαιτεί επίσης χαμηλή εγγενή διάσταση.

Στη μέθοδό τους υποστηρίζουν ότι η αλλαγή ΔW_n στα βάρη W_{nk} για ένα επίπεδο του μοντέλου προκειμένου να προσαρμοστεί στο καινούριο πρόβλημα, είναι ένας πίνακας πολύ χαμηλότερης τάξης (low-rank matrix).

Η τάξη ή βαθμός ενός πίνακα (rank) είναι ο μέγιστος αριθμός των γραμμικά ανεξάρτητων γραμμών ή στηλών του πίνακα. Ο αριθμός των γραμμικά ανεξάρτητων γραμμών είναι πάντα ίδιος με τον αριθμό των γραμμικά ανεξάρτητων στηλών. Ο πίνακας που έχει όλες τις γραμμές ή στήλες γραμμικά ανεξάρτητες ονομάζεται πλήρους βαθμού (full rank), διαφορετικά ονομάζεται πίνακας ελλιπούς βαθμού (rank deficient). Ένας πίνακας τάξης ίση με 1 υποδηλώνει ότι όλες οι γραμμές (ή στήλες) είναι γραμμικά εξαρτημένες μεταξύ τους.

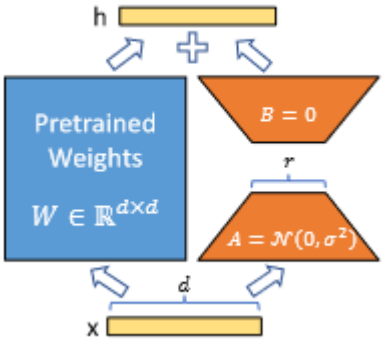
Έστω ο προεκπαιδευμένος πίνακας βαρών $W0 \in \mathbb{R}^{d\times k}$. Η ανανέωσή του ΔW μπορεί να προσεγγιστεί με τη χρήση του low rank decomposition όπως φαίνεται στην παρακάτω σχέση:

$$W0 + \Delta W = W0 + BA \quad (38)$$

όπου $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ και η τάξη $r \ll min(d,k)$. Κατά τη διάρκεια της εκπαίδευσης ο πίνακας βαρών W0 δεν μεταβάλλεται και δεν λαμβάνεται υπόψη στον υπολογισμό της κλίσης (gradient). Οι πίνακες *A, B* περιλαμβάνουν παραμέτρους που εκπαιδεύονται κατά τη διάρκεια της εκπαίδευσης του μοντέλου. Ο πίνακας *W0* και ο πίνακας ΔW, όπου $\Delta W = BA$, πολλαπλασιάζονται με την ίδια είσοδο. Τέλος, τα αντίστοιχα αποτελέσματα αθροίζονται κατά στοιχείο, όπως αποτυπώνεται στην σχέση (39):

$$h = W0x + \Delta W_x = W0x + BAx \quad (39)$$

Τα βάρη στον πίνακα *A*, αρχικοποιούνται χρησιμοποιώντας τη γκαουσιανή κατανομή ενώ στον πίνακα *B* είναι ίσα με 0 ,έτσι ώστε να ισχύει $\Delta W = BA = 0$ στην αρχή της εκπαίδευσης. Η μέθοδος LoRA αποτυπώνεται και στην Εικόνα 25.



Εικόνα 25 - Low Rank Adaptation [45]

Με την επιλογή της χρήσης των πινάκων *A, B* μικρής τάξης *r*, ο αριθμός των παραμέτρων που εκπαιδεύεται μειώνεται σημαντικά. Για παράδειγμα, έστω *W* ο πίνακας βαρών σε ένα επίπεδο του μοντέλου, διάστασης $d \times d$. Το πλήθος των παραμέτρων που απαιτούν εκπαίδευση ανέρχεται στον αριθμό d^2 . Με τη χρήση, όμως, των πινάκων *A, B* μεγέθους $d \times r$και $r \times k$, οι παράμετροι προς εκπαίδευση είναι συνολικά χαμηλότεροι αφού το $r \ll min(d, k)$, όπως έχει αναφερθεί προηγουμένως. Στα πειράματά τους, οι Hu et al έδειξαν ότι η χρήση πινάκων τάξης ίση με 4 προσδίδει εξίσου ανταγωνιστικά αποτελέσματα υποστηρίζοντας έτσι την ύπαρξη χαμηλής εγγενής διάστασης κατά την προσαρμογή των μοντέλων σε νέα δεδομένα.

Η μέθοδος εφαρμόστηκε για πρώτη φορά σε μοντέλα για την επεξεργασία φυσικής γλώσσας. Παρόλα αυτά, οι Zhu *et al.* [48] εφάρμοσαν τη μέθοδο αυτή στην εκπαίδευση του μοντέλου του οπτικού μετασχηματιστή. Συγκεκριμένα, έδειξαν στην εργασία τους, ότι διατηρώντας τα προεκπαιδευμένα βάρη αμετάβλητα και ανανεώνοντας μόνο τους πίνακες βαρών *A, B* πετυχαίνουν αξιοσημείωτα αποτελέσματα σε σχέση με τη χρήση της μεθόδου fine-tuning. Την τεχνική LoRA την αξιοποίησαν κατά την εκπαίδευση του μοντέλου σε διαφορετικά προβλήματα, κάνοντας χρήση τεσσάρων διαφορετικών συνόλων δεδομένων χρησιμοποιώντας μόνο το 0.17% των αρχικών παραμέτρων. Οι Zhu *et al.* ενσωμάτωσαν τα βάρη LoRA στα επίπεδα αυτό-προσοχής του προεκπαιδευμένου οπτικού μετασχηματιστή και συγκεκριμένα στους πίνακες *WQ* και *WV* χρησιμοποιώντας *rank = 4*. Δοκιμάζοντας διαφορετικά μεγέθη της αρχιτεκτονικής του οπτικού μετασχηματιστή στα πειράματά τους, απέδειξαν ότι η απόδοση του μοντέλου που εκπαιδεύτηκε με τη μέθοδο LoRA αποδίδει παρόμοια ή καλύτερα με την εκπαίδευσή του με τη χρήση full fine-tuning. Σημειώνεται στο σημείο αυτό ότι με τη μέθοδο LoRA εκπαιδεύονται σημαντικά λιγότερες παράμετροι μειώνοντας σημαντικά με αυτό τον τρόπο την κατανάλωση υπολογιστικών πόρων.

4. Προτεινόμενη Μεθοδολογία

Στο κεφάλαιο αυτό θα παρουσιαστεί αναλυτικά η μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία με σκοπό την ταξινόμηση εξετάσεων CTPA σε θετικές και αρνητικές αναλόγως της ύπαρξης ή μη της ασθένειας της ΠΕ αντίστοιχα. Αρχικά, θα πραγματοποιηθεί η περιγραφή του συνόλου δεδομένων που αξιοποιήθηκε καθώς και οι προκλήσεις που παρουσιάζει. Στη συνέχεια, θα ακολουθήσει αναλυτική περιγραφή της απαιτούμενης προεπεξεργασίας των δεδομένων εισόδου. Τέλος, θα γίνει αναφορά στον τρόπο εκπαίδευσης των μοντέλων. Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκε η γλώσσα Python και πιο συγκεκριμένα PyTorch framework. Ο κώδικας αναπτύχθηκε εξ ολοκλήρου στην πλατφόρμα του Kaggle, χρησιμοποιώντας Kaggle Notebooks.

4.1 Σύνολο Δεδομένων

Το σύνολο δεδομένων που αξιοποιήθηκε στην παρούσα εργασία εκδόθηκε από την ακτινολογική εταιρεία της Βόρειας Αμερικής (Radiological Society of North America - RSNA) σε συνεργασία με την Εταιρεία Ακτινολογίας Θώρακος (Society of Thoracic Radiology -STR). Αποτελείται από εξετάσεις CTPA, οι οποίες συλλέχθηκαν από διαφορετικά ιδρύματα, προερχόμενα από πέντε διαφορετικές χώρες [49]. Τα δεδομένα χαρακτηρίζονται από ποικιλομορφία όσον αφορά τους ασθενείς, τον εξοπλισμό και τα πρωτόκολλα που χρησιμοποιήθηκαν για την διενέργεια των εξετάσεων CTPA. Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία περιέχει 1790594 εικόνες από συνολικά 7279 εξετάσεις. Από τις εξετάσεις αυτές, οι 2368 είναι θετικές στην ΠΕ και οι υπόλοιπες 4911 είναι αρνητικές. Στο σύνολο των εικόνων, οι 1694054 είναι αρνητικές, δεν αποτελούν δηλαδή μέρος εμβολής. Οι θετικές εικόνες είναι συνολικά 96540.

Στην παρούσα εργασία αξιοποιήθηκε η επισημείωση για την ύπαρξη ή μη ΠΕ, τόσο σε επίπεδο εξέτασης όσο και σε επίπεδο εικόνας. Το σύνολο δεδομένων χωρίστηκε σε τρία υποσύνολα: σύνολο εκπαίδευσης (train set), σύνολο επικύρωσης (validation set) και σύνολο δοκιμής (test set) με ποσοστό 0.7, 0.2, 0.1 αντίστοιχα. Τα σύνολα εκπαίδευσης και επικύρωσης χρησιμοποιήθηκαν κατά τη διάρκεια της εκπαίδευσης του μοντέλου ενώ το σύνολο δοκιμής περιέχει δεδομένα τα οποία δεν έχουν αξιοποιηθεί κατά την εκπαίδευση του μοντέλου. Αποτελεί, δηλαδή, το σύνολο δεδομένων στο

οποίο θα πραγματοποιηθεί η αξιολόγηση της απόδοσης του μοντέλου. Στον Πίνακα 1 αναφέρεται συγκεντρωτικά το μέγεθος των υποσυνόλων.

Εξετάσεις CTPA	**5001 (70%)**	**1409 (20%)**	**712 (10%)**
Θετικές	1553	437	221
Αρνητικές	3448	972	491
Τομές CT	**1232329**	**34778**	**173805**
Θετικές	67.287	18.766	163.318
Αρνητικές	1165042	329014	10487

Πίνακας 1- Διαχωρισμός συνόλου δεδομένων

Ο διαχωρισμός στα υποσύνολα πραγματοποιήθηκε σε επίπεδο εξέτασης, έτσι ώστε να διασφαλιστεί ότι εικόνες από την ίδια εξέταση θα παραμείνουν στο ίδιο υποσύνολο. Επιπλέον, σε κάθε υποσύνολο διατηρείται η ίδια κατανομή επισημειώσεων σε επίπεδο εξέτασης με αυτή του αρχικού συνόλου δεδομένων. Τέλος, λήφθηκε υπόψη το μέγεθος των εξετάσεων, δηλαδή το πλήθος των εικόνων που τις αποτελούν. Καθώς υπάρχει σημαντική διαφορά μεταξύ των μεγεθών των εξετάσεων και προκειμένου να υπάρξει ομοιομορφία στα υποσύνολα, ο διαχωρισμός έγινε με τέτοιο τρόπο ώστε και τα τρία υποσύνολα να περιέχουν παρόμοιο πλήθος εξετάσεων από κάθε μέγεθος. Για το σκοπό αυτό, οι εξετάσεις στο αρχικό σύνολο δεδομένων διαχωρίστηκαν με βάση το μέγεθός τους σε 10 τμήματα. Στη συνέχεια, το αρχικό σύνολο δεδομένων χωρίστηκε με τέτοιο τρόπο ώστε κάθε υποσύνολο να περιέχει τον ίδιο αριθμό εξετάσεων από κάθε τμήμα. Χρησιμοποιήθηκαν 11 τμήματα, με τη μέγιστη διαφορά στο μέγεθος των εξετάσεων να κυμαίνεται στις 100 εικόνες. Για παράδειγμα, το πρώτο τμήμα περιλαμβάνει εξετάσεις με αριθμό εικόνων 0 έως 99 εικόνες, το δεύτερο 100 έως 199, το τρίτο 200 έως 299 κ.ο.κ.

Οι εικόνες CTPA είναι αποθηκευμένες με τη μορφή DICOM (Digital Imaging and Communications in Medicine) αρχείων. Τα αρχεία αυτά αποθηκεύονται σύμφωνα με το πρότυπο Digital Imaging and Communications in Medicine (DICOM), που αποτελεί διεθνές πρότυπο για την αποθήκευση και τη μετάδοση ψηφιακών ιατρικών εικόνων και της συνοδευτικής πληροφορίας [50]. Τα αρχεία DICOM οργανώνουν την πληροφορία σε δύο σύνολα. Το πρώτο σύνολο αποτελεί την κεφαλίδα και το δεύτερο το σύνολο εικόνας. Τα δύο αυτά σύνολα περιέχονται στο ίδιο αρχείο. Η κεφαλίδα ενός αρχείου DICOM περιέχει πληροφορίες για τα δημογραφικά στοιχεία του ασθενή και

πληροφορίες για την εικόνα που περιλαμβάνει, όπως για παράδειγμα τη διάσταση, το χρωματικό χώρο και κατ' επέκταση όλη την απαραίτητη πληροφορία για την ορθή προβολή της εικόνας [51]. Η πληροφορία της κεφαλίδας είναι κωδικοποιημένη προκειμένου να μην μπορεί να διαχωριστεί από τα δεδομένα εικόνας. Σε περίπτωση διαχωρισμού, ο εκάστοτε υπολογιστής αδυνατεί να προβάλει σωστά την εξέταση καθώς δεν διαθέτει την απαραίτητη πληροφορία [51].

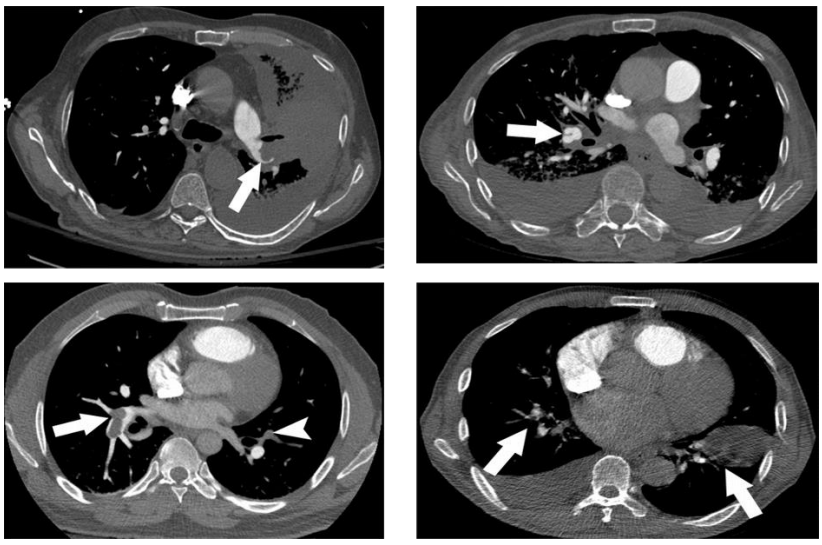
Οι εξετάσεις στο σύνολο δεδομένων είναι διαχωρισμένες σε φακέλους. Κάθε φάκελος λαμβάνει σαν όνομα το αναγνωριστικό της εξέτασης (StudyInstanceUID). Κάθε εικόνα, ομοίως, έχει ένα μοναδικό αναγνωριστικό (SOPInstanceUID). Η πρόσβαση σε κάθε εικόνα στο σύνολο δεδομένων δίνεται από τη διαδρομή: *<StudyInstanceUID>/<SeriesInstanceUID>/<SOPInstanceUID>.dcm*. Τα τρία αναγνωριστικά είναι αποθηκευμένα στις αντίστοιχες ετικέτες στην κεφαλίδα του εκάστοτε DICOM αρχείου.

Το σύνολο δεδομένων περιέχει 512 × 512 εικόνες DICOM CTPA στην αρτηριακή φάση. Οι τιμές των εικονοστοιχείων είναι σε μονάδες Hounsfield (Hounsfield Units - HU). Η κλίμακα Hounsfield αποτελεί ένα κοινό σημείο αναφοράς για τη μέτρηση των συντελεστών απορρόφησης κάθε ιστού, όταν αυτοί εκτίθεται σε ακτίνες Χ. Όσο πιο πυκνός είναι ο ιστός τόσο περισσότερη ακτινοβολία απορροφά. Ο αέρας στην κλίμακα Hounsfield λαμβάνει τιμή -1000 ενώ το νερό την τιμή 0.

Για την καλύτερη απεικόνιση του ιστού ενδιαφέροντος χρησιμοποιείται το παράθυρο (window). Το παράθυρο ορίζεται από δύο τιμές: το πλάτος (width) και το επίπεδο (level). Το πλάτος του παραθύρου αποτελεί τη μέτρηση του εύρους των μονάδων HU που περιλαμβάνει το παράθυρο. Το επίπεδο του παραθύρου αποτελεί το κέντρο του παραθύρου, δηλαδή την ενδιάμεση τιμή του εύρους των τιμών του παραθύρου.

Τα περισσότερα όργανα χρησιμοποιούν το παράθυρο του μαλακού ιστού. Το παράθυρο αυτό έχει πλάτος 400 HU και επίπεδο +40 HU. Συνεπώς, το εύρος των τιμών HU κυμαίνονται από -160 HU έως +240 HU [53]. Αυτό σημαίνει ότι οποιαδήποτε τιμή μικρότερη από -160 HU θα εμφανίζεται ως μαύρο και μεγαλύτερη από +240 HU ως άσπρο. Στην παρούσα εργασία, εφαρμόστηκε το παράθυρο για την καλύτερη ανίχνευση της ΠΕ. Το παράθυρο χαρακτηρίζεται από πλάτος 700 HU και επίπεδο +100 HU [52].

Τέλος, αξίζει να γίνει αναφορά στις ιδιαιτερότητες του συνόλου δεδομένων, το οποίο παρουσιάζει ανισορροπία μεταξύ των κλάσεων τόσο σε επίπεδο τομών, όσο και σε επίπεδο εξέτασης. Αυτό αποτυπώνεται στον Πίνακα 1. Σε επίπεδο τομών η αναλογία μεταξύ της κλάσης μειονότητας και της κλάσης πλειονότητας αντίστοιχα, είναι 1/17. Σε επίπεδο εξέτασης η αντίστοιχη αναλογία μεταξύ των κλάσεων κυμαίνεται περίπου στο 1/2. Επιπλέον, μεταξύ των διαδοχικών τομών που περιέχονται σε μία εξέταση υπάρχει μεγάλη συσχέτιση. Τέλος, το μέγεθος της ΠΕ καταλαμβάνει μικρή περιοχή τόσο σε επίπεδο 2Δ (τομές), όσο και σε 3Δ, δηλαδή στον συνολικό όγκο της εξέτασης. Στην Εικόνα 26, παρουσιάζονται ορισμένα δείγματα από θετικές τομές. Τα βέλη δείχνουν τις περιοχές της ΠΕ. Είναι εμφανής η μικρή περιοχή που καταλαμβάνουν οι ΠΕ.



Εικόνα 26 - Θετικές τομές ΠΕ [49]

4.2 Προεπεργασία Δεδομένων

Η προεπεξεργασία των εικόνων είναι απαραίτητη προκειμένου να υπάρχει συμβατότητα με το μοντέλο που πρόκειται να εκπαιδευτεί. Για τον σκοπό αυτό, οι εικόνες από DICOM αρχεία μετατράπηκαν σε JPEG RGB εικόνες. Για την ανάγνωση των DICOM αρχείων χρησιμοποιήθηκε το διαθέσιμο πακέτο της Python, *Pydicom.* Όπως αναφέρθηκε και στην προηγούμενη ενότητα, κάθε αρχείο DICOM περιέχει την κεφαλίδα και τα δεδομένα εικόνας. Οι τιμές των εικονοστοιχείων είναι αποθηκευμένες στην ετικέτα *Pixel Data.* Με τη βοήθεια της συνάρτησης που μας προσφέρει το πακέτο *Pydicom*, αποθηκεύονται τα δεδομένα των εικονοστοιχείων σε πίνακα. Στη συνέχεια

εφαρμόζεται το παράθυρο για την ΠΕ όπως αναφέρθηκε στο προηγούμενο κεφάλαιο. Οι τιμές στη συνέχεια μετατρέπονται από HU σε τιμές μεταξύ 0 και 255. Τα περισσότερα, όμως, μοντέλα βαθιάς μάθησης δέχονται σαν είσοδο RGB εικόνες. Για τον λόγο αυτό οι εικόνες αποθηκεύονται σε RGB ως JPEG εικόνες. Κατά αυτόν τον τρόπο προκύπτει ένα νέο σύνολο δεδομένων, το οποίο περιλαμβάνει τις JPEG εικόνες που προέκυψαν από τα αντίστοιχα DICOM αρχεία.

Το μοντέλο εκπαιδεύεται στο σύνολο δεδομένων που περιέχει τις JPEG εικόνες. Πριν από την τροφοδότηση του μοντέλου, οι εικόνες κανονικοποιούνται με μέση τιμή [0.485, 0.456, 0.406] και τυπική απόκλιση [0.229, 0.224, 0.225]. Οι τιμές αυτές αναφέρονται στο σύνολο δεδομένων του ImageNet και προκύπτουν λαμβάνοντας υπόψη εκατομμύρια εικόνες. Η χρήση προεκπαιδευμένου μοντέλου στο σύνολο δεδομένων ImageNet συνήθως απαιτεί τη χρήση των εν λόγω τιμών. Επιπλέον, οι εικόνες μετατρέπονται σε εικόνες 224 × 224. Τέλος, επιλέγεται η χρήση του data augmentation. Με αυτή την τεχνική, τροποποιούνται επιλεκτικά και με τυχαίο τρόπο εικόνες από το σύνολο εκπαίδευσης προκειμένου να υπάρξει μία διαφοροποίηση, προς αποφυγή της υπερπροσαρμογής στα δεδομένα εκπαίδευσης. Οι διάφορες τροποποιήσεις που επιλέχτηκαν είναι η αλλαγή στη φωτεινότητα της εικόνας και η κάθετη ή οριζόντια ανατροπή της εικόνας. Τέλος, σε ορισμένα πειράματα εφαρμόστηκε και η απαλοιφή του φόντου από τις εικόνες καθώς στις περισσότερες εικόνες CT το μαύρο φόντο πλεονάζει.

4.3 Περιγραφή Αρχιτεκτονικής

Η αρχιτεκτονική στην παρούσα εργασία περιλαμβάνει δύο επίπεδα. Το πρώτο επίπεδο έχει ως στόχο την ανίχνευση της ΠΕ σε επίπεδο εικόνας. Το δεύτερο επίπεδο χρησιμοποιεί τα χαρακτηριστικά των τομών που εξάγονται από το πρώτο επίπεδο και πραγματοποιεί ανίχνευση τόσο σε επίπεδο τομών όσο και σε επίπεδο εξέτασης.

4.3.1 Περιγραφή Πρώτου Επιπέδου

Στο πρώτο επίπεδο αξιοποιείται η αρχιτεκτονική του μετασχηματιστή ViT. Σαν είσοδο το μοντέλο λαμβάνει μεμονωμένα τις τομές, δίχως να λαμβάνεται υπόψη η σειρά ή η προέλευση των τομών στην εξέταση. Στο σημείο αυτό, τονίζεται ότι τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής δεν μοιράζονται τομές από την ίδια εξέταση. Το μοντέλο ViT, όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο,

αποτελείται εξ ολοκλήρου από μετασχηματιστές. Συνήθως, αυτά τα μοντέλα τις περισσότερες φορές απαιτούν την εκπαίδευση τους σε μεγάλα σύνολα δεδομένων προκειμένου να αποδώσουν καλύτερα από τα CNN μοντέλα. Για το λόγο αυτό στην παρούσα εργασία αξιοποιήθηκε το προεκπαιδευμένο μοντέλο ViT που προσφέρεται από το πακέτο PyTorch Image Models (timm), το οποίο δίνει πρόσβαση σε πληθώρα προεκπαιδευμένων μοντέλων. Επιλέχτηκε το βασικό μοντέλο ViT, όπως αυτό περιεγράφηκε στην αρχική δημοσιευμένη εργασία [42]. Συγκεκριμένα, το μοντέλο δέχεται σαν είσοδο την εικόνα μεγέθους 3 × 224 × 224 και τη διαχωρίζει σε τμήματα (patches) διάστασης 16 x 16. Κατά αυτό τον τρόπο, προκύπτουν 196 τμήματα τα οποία προβάλλονται γραμμικά και στη συνέχεια προστίθεται σε αυτά η πληροφορία της θέσης. Στις αναπαραστάσεις που προκύπτουν προστίθεται επιπλέον και το σύμβολο κλάσης , το οποίο χρησιμοποιείται για την τελική πρόβλεψη. Τέλος, οι 197 αναπαραστάσεις τροφοδοτούν τον κωδικοποιητή του ViT.

Το μοντέλο χρησιμοποιεί τα βάρη που προέκυψαν από την εκπαίδευσή του στο μεγαλύτερο σύνολο δεδομένων ImageNet-22k. Το τελευταίο επίπεδο ταξινόμησης αντικαταστάθηκε με ένα γραμμικό επίπεδο κλάσεις διάστασης 1. Η τελική διάσταση πρέπει να είναι συμβατή με τον αριθμό των κλάσεων που χαρακτηρίζουν το συγκεκριμένο πρόβλημα. Στην εν λόγω εργασία, στόχο αποτελεί η πρόβλεψη της επισημείωσης για την ΠΕ σε επίπεδο τομών. Η ταξινόμηση που επιδιώκει την πρόβλεψη σε μία κατηγορία και μπορεί να πάρει τιμές 1 ή 0 ονομάζεται δυαδική ταξινόμηση (binary classification).

Η ύπαρξη ανισορροπίας στο σύνολο δεδομένων οδηγεί στο να ευνοηθεί η πλειοψηφική κατηγορία, ταξινομώντας λανθασμένα τα δείγματα που προέρχονται από τη μειοψηφική κατηγορία. Αυτό έχει σαν αποτέλεσμα τη μείωση της διακριτικής ικανότητας του μοντέλου μεταξύ των δύο κλάσεων, γεγονός που δεν είναι αποδεκτό κατά την ανάπτυξη μοντέλων βαθιάς μάθησης.

Δύο τρόποι αντιμετώπισης της ανισορροπίας μεταξύ των δύο κλάσεων είναι η υπέρ-δειγματοληψία και η υπό-δειγματοληψία. Οι τεχνικές αυτές τροποποιούν την κατανομή των κλάσεων στο σύνολο δεδομένων εκπαίδευσης με στόχο τη μείωση της ανισορροπίας στα δεδομένα. Συγκεκριμένα, με τη μέθοδο της υπέρ-δειγματοληψίας, επιλέγεται με τυχαίο τρόπο η αντιγραφή δειγμάτων από τη μειοψηφική κατηγορία με στόχο την αύξηση του αριθμού τους στο σύνολο δεδομένων. Αντίθετα, με τη μέθοδο

της υπό-δειγματοληψίας επιλέγεται με τυχαίο τρόπο η απομάκρυνση ορισμένων δειγμάτων της πλειοψηφικής κατηγορίας.

Το τελικό μοντέλο αξιοποιεί την τεχνική της υπό-δειγματοληψίας επιδιώκοντας την εξισορρόπηση μεταξύ των κλάσεων, αφαιρώντας επιλεκτικά από το σύνολο εκπαίδευσης δείγματα από την πλειοψηφική κατηγορία. Σε κάθε εποχή εκπαίδευσης (epoch), χρησιμοποιούνται διαφορετικά δείγματα από τη κλάση πλειονότητας.

Για την εκπαίδευση της αρχιτεκτονικής στο στάδιο αυτό χρησιμοποιήθηκε η μέθοδος προσαρμογής LoRA, καθώς φαίνεται να οδηγεί σε αποδοτικότερη εκπαίδευση. Αυτό προκύπτει από πειράματα και συγκρίσεις που θα αναλυθούν εκτενέστερα στο επόμενο κεφάλαιο. Ακόμη, για την αντιμετώπιση της ανισορροπίας μεταξύ των κλάσεων εφαρμόστηκε υπό-δειγματοληψία της κλάσης πλειονότητας. Επιπλέον, χρησιμοποιήθηκε μέγεθος batch ίσο με 64. Τέλος, ως μέθοδος βελτιστοποίησης χρησιμοποιήθηκε η μέθοδος Adam με αρχικό ρυθμό μάθησης 0.0005, σε συνδυασμό με τη μέθοδο scheduler CosineAnnealingLR.

4.3.2 Περιγραφή Δευτέρου Επιπέδου

Το δεύτερο στάδιο περιλαμβάνει την αρχιτεκτονική που είναι υπεύθυνη για την πρόβλεψη σε επίπεδο εξέτασης. Κάθε εξέταση CTPA αποτελείται από ένα σύνολο 2Δ τομών, με το μέγεθος κάθε εξέτασης να ποικίλλει. Η διάγνωση της ΠΕ απαιτεί τη διερεύνηση των τομών της εξέτασης. Κάθε τομή είναι η συνέχεια της προηγούμενης. Συνεπώς, κάθε εξέταση CTPA αποτελεί μία ακολουθία τομών.

Τα μοντέλα βαθιάς μάθησης που είναι ικανά να επεξεργάζονται και να εκπαιδεύονται αποτελεσματικά σε ακολουθίες δεδομένων, καθώς λαμβάνουν υπόψη την πληροφορία από όλα τα στοιχεία της ακολουθίας, είναι τα RNNs. Στην παρούσα εργασία επιλέχτηκε η χρήση της αρχιτεκτονικής του LSTM και συγκεκριμένα του Bidirectional LSTM (BiLSTM). Η επιλογή του BiLSTM μπορεί να ενισχύσει την κατανόηση της ακολουθίας της εξέτασης από το μοντέλο, αφού την επεξεργάζεται και προς τις δύο κατευθύνσεις (αριστερά προς δεξιά και αντίστροφα). Τα μοντέλα LSTM δεν δέχονται απευθείας δεδομένα εικόνας, όπως για παράδειγμα τα μοντέλα CNN. Η πληροφορία της εικόνας πρέπει να κωδικοποιηθεί σε διάνυσμα προκειμένου να αξιοποιηθεί. Τα διανύσματα προκύπτουν με τη βοήθεια του μοντέλου ViT που εκπαιδεύτηκε στο πρώτο στάδιο στο πρόβλημα της ταξινόμησης των τομών σε θετικές

ή αρνητικές. Οι αναπαραστάσεις των τομών προκύπτουν από την έξοδο του τελευταίου κρυφού επιπέδου του μοντέλου ViT.

Στο επίπεδο αυτό, επιλέχτηκε η αρχιτεκτονική που περιέχει 2 επίπεδα BiLSTM. Η έξοδος από το πρώτο επίπεδο αποτελεί την είσοδο του δεύτερου επιπέδου. Και τα δυο επίπεδα LSTM έχουν την ίδια διάσταση κρυφών επιπέδων, η οποία είναι ίση με 256. Παρόλα αυτά επειδή τα LSTM είναι αμφίδρομα , η έξοδος τους έχει διάσταση 512. Οι αναπαραστάσεις των τομών από το πρώτο επίπεδο έχουν διάσταση ίση με 768. Συνεπώς και τα επίπεδα LSTM δέχονται σαν είσοδο διανύσματα της συγκεκριμένης διάστασης. Από κάθε επίπεδο LSTM αξιοποιούνται οι έξοδοι από κάθε βήμα *t* προκειμένου να χρησιμοποιηθεί η αντίστοιχη πληροφορία για κάθε τομή. Στο τέλος του δεύτερου επιπέδου LSTM εφαρμόζεται γραμμικό στρώμα , το οποίο μειώνει τη διάσταση των διανυσμάτων σε 256. Στο αποτέλεσμα εφαρμόζεται η συνάρτηση ενεργοποίησης LeakyReLU και ένα επίπεδο dropout με πιθανότητα 0.3. Στη συνέχεια στο αποτέλεσμα εφαρμόζεται attention pooling, προκειμένου να ληφθεί υπόψη η βαρύτητα της συνεισφοράς κάθε διανύσματος εικόνας στην πρόβλεψη της εξέτασης με τη χρήση του μηχανισμού προσοχής. Το αποτέλεσμα αυτού του επιπέδου αποτελεί η δημιουργία αναπαράστασης διάστασης 256 για κάθε εξέταση, σύμφωνα με τις τιμές των βαρών. Τέλος, ακολουθεί το επίπεδο ταξινόμησης για την πρόβλεψη της εξέτασης της ΠΕ σε επίπεδο εξέτασης.

Για την εκπαίδευση στο δεύτερο επίπεδο αξιοποιήθηκε το σύνολο των εξετάσεων, δίχως να γίνει η χρήση της τεχνικής padding ή περικοπή των εξετάσεων σύμφωνα με κάποιο αποδεκτό μέγεθος. Αυτό ο τρόπος εκπαίδευσης χρησιμοποιεί την μέθοδο pad packed sequence. Η μέθοδος αυτή έχει υλοποιηθεί στο framework PyTorch και χρησιμοποιείται για την αντιμετώπιση της ύπαρξης μεταβλητών μεγεθών μεταξύ των ακολουθιών εισόδου στην εκπαίδευση μοντέλων RNN. Η μέθοδος αυτή κατόπιν πειραμάτων που διενεργήθηκαν, τα οποία θα παρουσιαστούν στο επόμενο κεφάλαιο, φαίνεται να αποδίδει καλύτερα.

Για την εκπαίδευση της αρχιτεκτονικής χρησιμοποιήθηκε μέγεθος batch ίσο με 32. Ως μέθοδος βελτιστοποίησης επιλέχτηκε ο Adam με αρχικό ρυθμό μάθησης 10^{-4}. Ως scheduler χρησιμοποιήθηκε και σε αυτό το στάδιο η μέθοδος CosineAnnealingLR. Το μοντέλο εκπαιδεύτηκε για 12 εποχές εκπαίδευσης.

5. Αποτελέσματα και Δοκιμές

Στο παρόν κεφάλαιο αναλύονται τα αποτελέσματα και τα πειράματα που έχουν διεξαχθεί στην παρούσα εργασία..

5.1 Πειράματα Πρώτου Επιπέδου

Το πρώτο επίπεδο, όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, περιλαμβάνει το μοντέλο που είναι υπεύθυνο για την εξαγωγή των χαρακτηριστικών των τομών της εξέτασης. Η εκπαίδευσή του πραγματοποιήθηκε σε επίπεδο τομών. Τα πειράματα που διεξήχθησαν σε αυτό το στάδιο περιλαμβάνουν τη δοκιμή της χρήσης διαφορετικών τεχνικών δειγματοληψίας για την αντιμετώπιση της ανισορροπίας στο σύνολο δεδομένων, την εκπαίδευση του επιπέδου ταξινόμησης διατηρώντας τα υπόλοιπα βάρη του μοντέλου σταθερά και την εφαρμογή της τεχνικής LoRA, προκειμένου να εξεταστεί η επίδραση της στην απόδοση του μοντέλου. Επιπλέον, εξετάστηκε η επιρροή του μεγέθους του batch για τις διάφορες τιμές της τάξης *r* στην τεχνική LoRA.

5.1.1 Τρόποι Δειγματοληψίας

Μέθοδος υπό - δειγματοληψίας

Κατά την εκπαίδευση του μοντέλου δοκιμάστηκαν τρεις διαφορετικοί τρόποι δειγματοληψίας προκειμένου να επιτευχθεί η εξισορρόπηση του συνόλου δεδομένων εκπαίδευσης. Ο πρώτος τρόπος επιδιώκει την εξισορρόπηση των κλάσεων αφαιρώντας σε κάθε εποχή με τυχαίο τρόπο διαφορετικές τομές που ανήκουν στην κλάση πλειονότητας. Ο αριθμός των τομών που αφαιρείται είναι τέτοιος ώστε ο τελικός αριθμός τομών που περιέχονται στις δύο κλάσεις να ισούται με τον αριθμό της κλάσης μειονότητας. Αυτός ο τρόπος δειγματοληψίας είναι και ο προτεινόμενος, αφού μειώνει σημαντικά το μέγεθος του συνόλου δεδομένων επιτρέποντας τη διενέργεια και την αξιολόγηση πολλών πειραμάτων σε εύλογο χρονικό διάστημα. Με τη χρήση του συγκεκριμένου τρόπου δειγματοληψίας μία εποχή εκπαίδευσης ολοκληρώνετε σε περίπου 40 λεπτά. Η διάρκεια μίας εποχής με τη χρήση υπέρ-δειγματοληψίας που θα αναλυθεί παρακάτω απαιτεί συνολικά 3 ώρες.

Μέθοδος υπέρ – δειγματοληψίας

Με τη μέθοδο της υπέρ-δειγματοληψίας επιτυγχάνεται η εξισορρόπηση των κλάσεων με την αύξηση των δειγμάτων της κλάσης μειονότητας. Η αύξηση των δειγμάτων επιτυγχάνεται με την τυχαία επιλογή αντιγραφής ορισμένων τομών από το σύνολο των θετικών τομών. Για τον σκοπό αυτό, κάθε δείγμα στο σύνολο δεδομένων εκπαίδευσης αποκτά ένα βάρος το οποίο εξαρτάται από το πλήθος των δειγμάτων που περιέχει η κλάση στην οποία ανήκει. Συνεπώς, τα δείγματα που ανήκουν στην κλάση με τις αρνητικές τομές ως προς την ΠΕ θα λάβουν μεγαλύτερη τιμή βάρους σε σχέση με αυτές που ανήκουν στο σύνολο δεδομένων με τις θετικές τομές ως προς την ΠΕ. Κατά την εκπαίδευση του μοντέλου, στην αρχή κάθε εποχής επιλέγονται οι τομές που θα χρησιμοποιηθούν με βάση τα βάρη που τις αποδόθηκαν. Ειδικότερα, τα βάρη αυτά εκφράζουν την πιθανότητα μία εικόνα να επιλεχθεί. Μεγαλύτερη πιθανότητα επιλογής έχουν οι τομές από την κλάση μειονότητας σε σχέση με τις τομές στην κλάση πλειονότητας. Με αυτό τον τρόπο επιτυγχάνεται η αντιπροσώπευση και των δύο κλάσεων από παρόμοιο αριθμό δειγμάτων. Ο συγκεκριμένος τρόπος δειγματοληψίας είναι διαθέσιμος από το PyTorch και ονομάζεται Weighted Random Sampler.

Μέθοδος δυναμικής δειγματοληψίας

Η μέθοδος δυναμικής δειγματοληψίας (Dynamic Class Balance Sampler) αναπτύχθηκε από το framework Catalyst, που βασίζεται σε αυτό του PyTorch. Η χρήση της συγκεκριμένης μεθόδου δειγματοληψίας προτείνεται στην περίπτωση που το σύνολο δεδομένων χαρακτηρίζεται από μεγάλη ανισορροπία μεταξύ των κλάσεων. Στην αρχή της εκπαίδευσης, υπάρχει μια αρχική κατανομή των δύο κλάσεων. Κατά την εξέλιξη της εκπαίδευσης η αρχική κατανομή σταδιακά καταλήγει σε ομοιόμορφη, ως προς τις κλάσεις, κατανομή. Στο τέλος, δηλαδή, της εκπαίδευσης, το σύνολο εκπαίδευσης θα περιλαμβάνει ίσο αριθμό δειγμάτων και από τις δύο κλάσεις. Ο αριθμός αυτός ισούται με τον αριθμό των δειγμάτων στη κλάση μειονότητας.

Έστω $D_i = C_i \,/\, Cmin$ όπου C_i αποτελεί το μέγεθος της κλάσεις i ενώ $Cmin$ υποδηλώνει το μέγεθος της κλάσης μειονότητας. Η ποσότητα D_i αντιπροσωπεύει την κατανομή των κλάσεων. Σε κάθε εποχή η κατανομή D_i ανανεώνεται σύμφωνα με τον τύπο:

$D_i = D_i^{g(n\_epochs)}$, όπου $g()$ είναι μία εκθετική συνάρτηση και n_epochs ο αριθμός της εκάστοτε εποχής εκπαίδευσης.

Τα αποτελέσματα από τα πειράματα που διεξήχθησαν με βάση τη μέθοδο δειγματοληψίας αποτυπώνονται στον Πίνακα 2. Παρατηρείται ότι η απόδοση του μοντέλου σύμφωνα με την τιμή της AUC είναι παρόμοια και στις τρεις μεθοδολογίες δειγματοληψίας. Η μεγαλύτερη τιμή επιτυγχάνεται με τη μέθοδο Dynamic Balance Class Sampler. Παρόλα αυτά, ίσως η τιμή αυτή να μην είναι αντιπροσωπευτική καθώς το μοντέλο εκπαιδεύτηκε σε 22 εποχές εκπαίδευσης, όπου το σύνολο δεδομένων εκπαίδευσης θα περιέχει ακόμα ανισορροπία μεταξύ των κλάσεων. Επιπλέον, τόσο η υπέρ-δειγματοληψία όσο η υπό-δειγματοληψία συνεισφέρουν με τον ίδιο τρόπο στην απόδοση του μοντέλου. Παρόλο που με τη μέθοδο της υπό-δειγματοληψίας χρησιμοποιείται πολύ μικρότερος αριθμός δειγμάτων ανά εποχή εκπαίδευσης, το μοντέλο αποδίδει εξίσου καλά. Η μέθοδος, όμως Weight- Random Sampler χρησιμοποιεί ολόκληρο το σύνολο δεδομένων, γεγονός που αυξάνει τις ώρες εκπαίδευσης που απαιτούνται. Σημειώνεται στο σημείο αυτό ότι για τα πειράματα χρησιμοποιήθηκε το ίδιο μέγεθος batch, ίσο με 32. Επίσης, εφαρμόζεται η ίδια μέθοδος βελτιστοποίησης σε συνδυασμό με την ίδια μέθοδο learning rate scheduler.

Sampler	#Epochs	AUC	Loss
Weight- Random Sampler	**7**	0.839	0.274
Balance Class Sampler	**22**	0.839	0.343
Dynamic Balance Class Sampler	**22**	0.844	0.29

Πίνακας 2- Αποτελέσματα των διάφορων μεθόδων δειγματοληψίας

Συγκεκριμένα, ως μέθοδος βελτιστοποίησης επιλέχτηκε ο Adam με αρχικό ρυθμό μάθησης 0.0005. Σαν μέθοδος learning rate scheduler επιλέχτηκε η μέθοδος CosineAnnealingR. Τέλος, για την ταξινόμηση χρησιμοποιήθηκε το μοντέλο ViT που έχει προεκπαιδευτεί στο σύνολο δεδομένων InageNet1k [64] και περιλαμβάνει 86567656 παραμέτρους. Στο μοντέλο αντικαταστάθηκε το επίπεδο ταξινόμησης με μία σειρά από γραμμικά επίπεδα, συναρτήσεις ενεργοποίησης και dropout. Συγκεκριμένα, χρησιμοποιήθηκε ένα γραμμικό επίπεδο διάστασης 256, στις εξόδους του οποίου εφαρμόζεται η συνάρτηση ενεργοποίησης LeakyRELU. Στη συνέχεια, εφαρμόζεται dropout με πιθανότητα 0.3, προκειμένου να αποφευχθεί η

υπερπροσαρμογή στα δεδομένα εκπαίδευσης. Τελευταίο επίπεδο είναι το επίπεδο της ταξινόμησης, το οποίο περιλαμβάνει μία έξοδο. Σημειώνεται στο σημείο αυτό ότι στην έξοδο του επιπέδου ταξινόμησης δεν εφαρμόζεται κάποια συνάρτηση ενεργοποίησης, καθώς η τελευταία εφαρμόζεται στη συνάρτηση απώλειας που χρησιμοποιείται. Με αυτή την αρχιτεκτονική ρυθμίστηκαν συνολικά, 197121 παράμετροι.

5.1.2 Χρήση Μεθόδου LoRA

Στα πλαίσια των πειραμάτων διεξήχθησαν δοκιμές με τη χρήση της τεχνικής LoRA για την εκπαίδευση του μοντέλου. Αξιοποιώντας τη μέθοδο LoRA, το μοντέλο προσαρμόζεται στο εκάστοτε πρόβλημα με την ενσωμάτωση και την ανανέωση πινάκων που διαθέτουν πολύ λιγότερες παραμέτρους σε σχέση με τους αντίστοιχους πίνακες παραμέτρων του ίδιου μοντέλου στα ενδιάμεσα επίπεδα.. Στην παρούσα εργασία, υιοθετήθηκε η υλοποίηση της εργασίας [48], όπου εφαρμόστηκε η μέθοδος LoRA στα επίπεδα της αρχιτεκτονικής του ViT. Συγκεκριμένα, η μέθοδος εφαρμόστηκε στα βάρη *WQ* και *WV* των επιπέδων αυτοπροσοχής, προκειμένου να επιτευχθεί η καλύτερη προσαρμογή των προεκπαιδευμένων βαρών σε ιατρικές εικόνες. Ωστόσο, στα πειράματα της εργασίας [48] δεν χρησιμοποιήθηκαν εικόνες από εξέταση CTPA. Στην παρούσα εργασία υιοθετήθηκε η συγκεκριμένη υλοποίηση και εφαρμόστηκε σε εικόνες CTPA.

Στα πειράματα που διεξήχθησαν με τη μέθοδο LoRA, δοκιμάζεται διαφορετικός αριθμός τάξης *r*. Συγκεκριμένα, τα πειράματα διεξάγονται για τιμές *r* ίσες με 4, 12, και 16. Επιπλέον, εξετάζεται ο ρόλος στην επίδραση του μεγέθους batch στις τιμές r και το αντίκτυπο που έχουν στην απόδοση του μοντέλου. Στον Πίνακα 3 παρουσιάζονται τα αποτελέσματα των εν λόγω πειραμάτων.

BATCH SIZE = 32			
Rank (r)	**r = 4**	**r = 12**	**r = 16**
AUC	0.88	0.87	0.89
Loss	0.28	0.24	0.22
BATCH SIZE = 64			
Rank (r)	**r = 4**	**r = 12**	**r = 16**
AUC	0.891	0.882	0.884
Loss	0.25	0.27	0.22

Πίνακας 3 - Αποτελέσματα για τις διάφορες τιμές r και μέγεθος batch

Από τα παραπάνω προκύπτει ότι η μέθοδος LoRA συνεισφέρει θετικά στην απόδοση του μοντέλου συγκριτικά με την εκπαίδευση χωρίς την εφαρμογή της μεθόδου, αφού παρατηρείται ότι αυξάνει την τιμή της AUC κατά περίπου 0.3. Επιπλέον, παρατηρείται ότι με η αύξηση του βαθμού r δεν συνεπάγεται και την αύξηση της αποτελεσματικότητας του μοντέλου, παρόλο που ρυθμίζονται κατά την εκπαίδευση περισσότερες παράμετροι. Το γεγονός αυτό επιβεβαιώνει την υπόθεση της μικρής εγγενής διάστασης των παραμέτρων του μοντέλου [45]. Στα πειράματα χρησιμοποιήθηκαν ίδιες μέθοδοι βελτιστοποίησης και learning rate scheduler με τα προηγούμενα πειράματα. Επιπλέον, αξιοποιήθηκε η υπό-δειγματοληψία ως μέθοδος για την εξισορρόπηση των κλάσεων, αφού χρησιμοποιεί μικρότερο μέρος του αρχικού συνόλου δεδομένων και συνεπώς εκπαιδεύεται πιο γρήγορα, επιτρέποντας τη διεξαγωγή περισσότερων πειραμάτων, όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο.

5.2 Πειράματα Δευτέρου Επιπέδου

Κατά την εκπαίδευση του LSTM οι αναπαραστάσεις των εικόνων δίνονται με τη σειρά που έχουν στην εξέταση. Η εκπαίδευση του μοντέλου έγινε με την χρήση batches (mini batch training). Αυτό απαιτεί όλες οι εξετάσεις που εμπεριέχονται σε κάθε batch να έχουν το ίδιο μέγεθος μεταξύ τους. Παρόλα αυτά δεν είναι απαραίτητο όλα τα batches να περιέχουν ίδιο μέγεθος ακολουθιών. Ο περιορισμός αυτός, οδήγησε στη χρήση διαφορετικών τεχνικών που θα αναλυθούν στο κεφάλαιο που ακολουθεί. Επιπλέον, στο στάδιο αυτό δοκιμάστηκαν διάφορες αρχιτεκτονικές που βασίζονται στα μοντέλα LSTM, όπως επίσης και ο συνδυασμός LSTM και μηχανισμού αυτοπροσοχής. Τέλος, δοκιμάστηκε η εκπαίδευση με τη χρήση διαφορετικών αναπαραστάσεων από διαφορετικά μοντέλα του πρώτου επιπέδου.

5.2.1 Τρόποι Αντιμετώπισης Μεταβλητού Μεγέθους Εξετάσεων

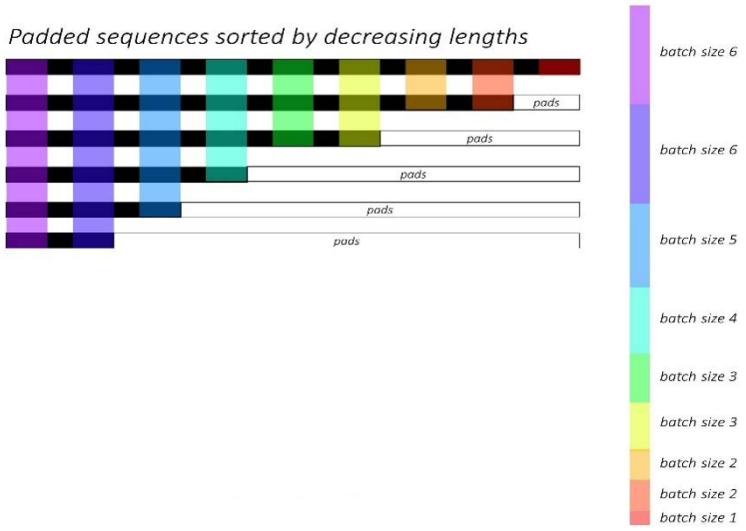
Γέμισμα με την τιμή 0

Ένας πρώτος τρόπος αντιμετώπισης του προβλήματος της ύπαρξης διαφορετικών μεγεθών μεταξύ των εξετάσεων αποτελεί το γέμισμα (padding) με την τιμή 0 για όσες θέσεις υπολείπονται, μέχρι το μέγεθος της ακολουθίας να ισούται με το μέγεθος που έχει οριστεί. Συνήθως, ως μέγεθος ορίζεται αυτό της μεγαλύτερης ακολουθίας στο

σύνολο δεδομένων εκπαίδευσης. Παρόλα αυτά η χρήση του padding ενδέχεται να αλλοιώσει την απόδοση του μοντέλου. Επιπλέον, χρησιμοποιούνται περισσότεροι πόροι για την εκπαίδευση χωρίς οι τιμές padding να προσφέρουν κάποια πληροφορία στην εκπαίδευση. Συνεπώς, η χρήση του padding πρέπει περιορίζεται σημαντικά όπου αυτό είναι εφικτό.

Μέθοδος packed pad sequence

Η μέθοδος packed pad sequence αποτελεί έναν τρόπο του PyTorch να αντιμετωπίζει ακολουθίες διαφορετικών μεγεθών κατά την εκπαίδευση των ακολουθιακών μοντέλων όπως το LSTM. Για τη χρήση της μεθόδου αυτής οι ακολουθίες σε κάθε batch ταξινομούνται από την μεγαλύτερη σε μέγεθος προς τη μικρότερη. Στη συνέχεια εφαρμόζεται padding ώστε να συμπληρωθεί το μέγεθος της μεγαλύτερης ακολουθίας. Ομοίως, ταξινομούνται και οι ετικέτες προκειμένου να υπάρχει ορθή αντιστοίχιση κατά την εκπαίδευση. Τέλος, η μέθοδος packed pad sequence θα λάβει την ταξινομημένη σειρά των ακολουθιών μαζί με τα αντίστοιχα μεγέθη και θα επιστρέψει δύο πίνακες (tensors). Ο πρώτος περιέχει τα στοιχεία των ακολουθιών ανά βήμα t. Ο δεύτερος αποτελείται από το αριθμό batch για κάθε βήμα. Στην Εικόνα 27 αποτυπώνεται ο τρόπος που η μέθοδος οργανώνει τις ακολουθίες για την εισαγωγή τους στο LSTM.



Εικόνα 27- Pack Padded Sequence

Παρατηρώντας την Εικόνα 27, στο πρώτο βήμα t το μέγεθος batch είναι 6, όπως και στο δεύτερο. Στο τρίτο βήμα, είναι 5 κ.ο.κ. Κατά αυτό τον τρόπο αποφεύγεται η χρήση του padding κατά την εκπαίδευση του LSTM.

Χρήση batches με ακολουθίες ίδιου μεγέθους

Η τελευταία τεχνική που μπορεί να χρησιμοποιηθεί για την αντιμετώπιση της διαφοράς των μεγεθών στο σύνολο δεδομένων μεταξύ των ακολουθιών αποτελεί η οργάνωση τους σε batches σύμφωνα με το μέγεθος τους. Με τον τρόπο αυτό οι ακολουθίες που έχουν το ίδιο μέγεθος θα βρίσκονται στο ίδιο batch. Στην περίπτωση του συνόλου δεδομένων που χρησιμοποιείται στην παρούσα εργασία υπάρχουν πολλές ακολουθίες οι οποίες δεν έχουν το ίδιο μέγεθος με κάποια άλλη ή είναι πολύ λίγες και δεν επαρκούν για να οργανωθούν σε batch. Παρόλα αυτά η μέθοδος αυτή δοκιμάζεται.

5.2.2 Δοκιμές Διαφορετικών Αρχιτεκτονικών LSTM

Στο δεύτερο επίπεδο, επιπλέον, δοκιμάστηκαν διαφορετικές αρχιτεκτονικές LSTM, οι οποίες θα αναλυθούν παρακάτω.

Επίπεδα LSTM

Η δοκιμή στα επίπεδα LSTM μπορούν να δώσουν την πληροφορία για τη σχέση της απόδοσης του μοντέλου με το βάθος του. Συγκεκριμένα, δοκιμάστηκε η χρήση ενός επιπέδου LSTM, όπου ένα μοντέλο LSTM διαπερνά την ακολουθία και προβλέπει τη διάγνωση. Επιπλέον, αξιολογήθηκε και η επίδραση της χρήσης ενός δεύτερου μοντέλου LSTM που λαμβάνει σαν είσοδο την ακολουθία που προκύπτει από κάθε βήμα t του πρώτου μοντέλου LSTM.

Χρήση Self Attention και Attention Pooling

Στην έξοδο που προκύπτει από κάθε βήμα *t* του μοντέλου LSTM δοκιμάστηκε η εφαρμογή του μηχανισμού self attention. Σκοπός της μεθόδου είναι η ανακάλυψη των συσχετίσεων μεταξύ των διανυσμάτων της ακολουθίας εισόδου. Για την ακρίβεια ο μηχανισμός self attention εφαρμόζεται στην έξοδο του LSTM, δηλαδή αξιοποιεί την πληροφορία που προκύπτει από την επεξεργασία της ακολουθίας από το LSTM. Οι νέες τιμές διανυσμάτων που προκύπτουν από την εφαρμογή self attention στα διανύσματα των εικόνων αξιοποιούνται για την πρόβλεψη σε επίπεδο εικόνας. Στα πειράματα αξιολογείται αν η εφαρμογή του self attention επιδρά θετικά στην κατανόηση των σχέσεων μεταξύ των διανυσμάτων των εικόνων και κατ' επέκταση στη βελτίωση της απόδοσης του μοντέλου.

Για την πρόβλεψη σε επίπεδο εξέτασης δοκιμάζεται η εφαρμογή του attention pooling στα διανύσματα των εικόνων της κάθε εξέτασης. Η υλοποίηση του attention pooling βασίζεται στην εργασία [65]. Στόχος του επιπέδου αυτού αποτελεί η δημιουργία ενός αντιπροσωπευτικού διανύσματος για κάθε εξέταση λαμβάνοντας υπόψη τα διανύσματα των εικόνων. Με τη χρήση ενός γραμμικού επιπέδου, αποδίδονται βάρη στα διανύσματα της εξέτασης. Κατά την εκπαίδευση της αρχιτεκτονικής, το μοντέλο εκπαιδεύεται στο να βρει τα διανύσματα που συνεισφέρουν περισσότερο στην πρόβλεψη προσαρμόζοντας ανάλογα τα βάρη. Στην συνέχεια στα τελευταία εφαρμόζεται η συνάρτηση SoftMax, η οποία μετατρέπει τα βάρη σε πιθανότητες, των οποίων το άθροισμα τους είναι ίσο με 1. Τελικά, για κάθε εικόνα προκύπτει το ποσοστό συνεισφοράς στην τελική αναπαράσταση της εξέτασης. Τα βάρη που προκύπτουν από τη συνάρτηση SoftMax πολλαπλασιάζονται με τα διανύσματα των εικόνων της εξέτασης. Έτσι, προκύπτουν οι νέες τιμές διανυσμάτων, οι οποίες στη συνέχεια προστίθενται κατά θέση ώστε να προκύψει το τελικό διάνυσμα της εξέτασης. Αυτό στη συνέχεια θα ληφθεί υπόψη προκειμένου να πραγματοποιηθεί η πρόβλεψη σε επίπεδο εξέτασης.

Στο σημείο αυτό να τονιστεί ότι στα πειράματα στα οποία αξιοποιείται κάποιος μηχανισμός προσοχής χρησιμοποιείται εξίσου η αντίστοιχη μάσκα για κάθε εξέταση, προκειμένου να μη ληφθούν υπόψη στις τιμές προσοχής οι τιμές όπου υπάρχει padding, δηλαδή 0.

5.2.3 Αποτελέσματα Δευτέρου Επιπέδου

Τα αποτελέσματα από τα πειράματα στο δεύτερο επίπεδο αποτυπώνονται στον Πίνακα 4. Παρατηρείται ότι στις περισσότερες δοκιμές οι τιμές AUC δεν διαφέρουν πολύ μεταξύ τους. Παρόλα αυτά παρατηρείται μία μικρή υπεροχή της μεθόδου packed pad sequence αφού κατέχει μεγαλύτερη τιμή AUC σε όλα τα πειράματα που διενεργήθηκαν σε αυτό το στάδιο. Το μοντέλο που διαθέτει την μικρότερη ικανότητα να διαχωρίζει τις εξετάσεις μεταξύ τους, είναι αυτό που εκπαιδεύτηκε με τη χρήση batch που περιέχουν εξετάσεις με το ίδιο μέγεθος . Αυτή, η μέθοδος όπως έχει ήδη αναφερθεί στο συγκεκριμένο σύνολο δεδομένων αναμένεται να παρουσιάσει μικρότερη απόδοση καθώς λόγω της ύπαρξης λίγων εξετάσεων που έχουν ίδιο μέγεθος, δεν δημιουργεί ολοκληρωμένα batch. Αυτό έχει σαν αποτέλεσμα, κατά τη διάρκεια κάθε εποχής το μοντέλο να λαμβάνει υπόψη του λίγα δείγματα για τον υπολογισμό των

κλίσεων. Αυτό μπορεί να οδηγήσει σε ασταθή εκπαίδευση. Τα δύο καλύτερα αποτελέσματα επιτυγχάνονται με τη χρήση του μοντέλου με τα δύο επίπεδα LSTM και ένα επίπεδο attention pooling και το μοντέλο που διαθέτει ένα επίπεδο LSTM, ένα επίπεδο αυτό-προσοχής και επίπεδο attention pooling. Τα δύο αυτά μοντέλα εκπαιδεύονται με τη χρήση της τεχνικής pad packed sequence.

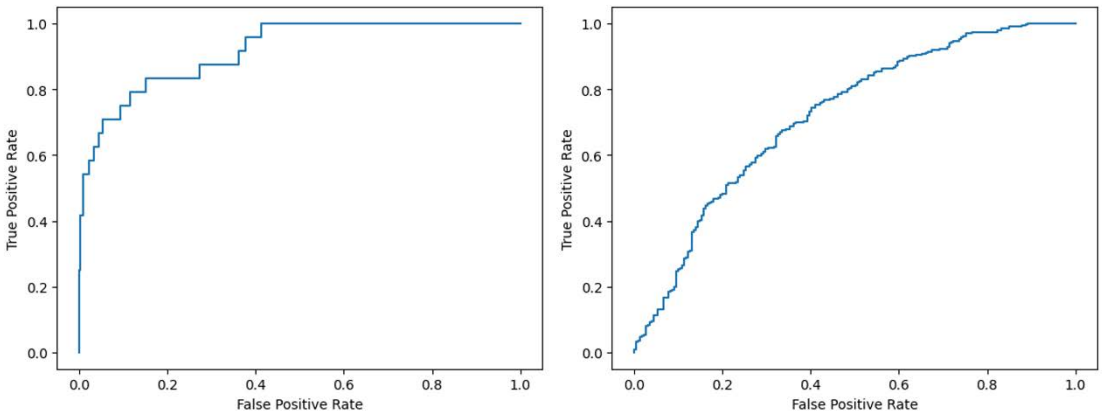
	Bucket Sampler	**Pad Packed Sequence**	**Padding**
	AUC	**AUC**	**AUC**
One Layer LSTM + Attention Pool	0.70	0.737	0.728
Two Layers LSTM + Attention Pool	x	0.738	0.721
One Layers LSTM + Self attention+ Attention Pool	x	0.735	0.716
Two Layers LSTM + Self attention+ Attention Pool	x	0.731	0.711

Πίνακας 4 - Αποτελέσματα πειραμάτων μοντέλου δεύτερου επιπέδου. Η τιμή Χ υποδηλώνει ότι δεν διενεργήθηκε πείραμα σε αυτή την κατηγορία.

Τέλος, για την εκπαίδευση των μοντέλων στο δεύτερο επίπεδο, όπως αναφέρθηκε προηγουμένως, δοκιμάστηκε η χρήση των αναπαραστάσεων εικόνων από το μοντέλο ViT που εκπαιδεύτηκε με την τεχνική της υπό-δειγματοληψίας. Το μοντέλο αυτό εκπαιδεύεται μόνο στο τελευταίο επίπεδο ταξινόμησης, επιτυγχάνοντας AUC 0.839. Κατά την εκπαίδευση του μοντέλου που περιέχει δύο επίπεδα LSTM, ένα επίπεδο αυτό-προσοχής και ένα επίπεδο attention pooling, με τη χρήση των αναπαραστάσεων του μοντέλου ViT επιτυγχάνεται AUC 0.661. Το αντίστοιχο πείραμα, με τη χρήση αναπαραστάσεων από το μοντέλο ViT που εκπαιδεύτηκε με τη μέθοδο LoRA επιτυγχάνει AUC 0.731. Επιπλέον, στο πείραμα εκπαίδευσης του μοντέλου που αποτελείται από ένα επίπεδο LSTM, ένα επίπεδο αυτοπροσοχής και ένα επίπεδο attention pooling, το μοντέλο ViT επιτυγχάνει AUC 0.65. Αντίθετα, το μοντέλο ViT που εκπαιδεύεται με την τεχνική LoRA επιτυγχάνει AUC 0.735.

6. Αξιολόγηση Αρχιτεκτονικής

Στο κεφάλαιο αυτό πραγματοποιείται η αξιολόγηση της επιλεχθείσας αρχιτεκτονικής και η σύγκριση της με αντίστοιχες εργασίες που έχουν διεξαχθεί. Η αρχιτεκτονική συνολικά σε επίπεδο διάγνωσης της ΠΕ πετυχαίνει AUC ίση με 0.72. Στο επίπεδο πρόβλεψης εικόνας επιτυγχάνεται AUC ίση με 0.92. Στην Εικόνα 28 παρουσιάζονται οι καμπύλες ROC για κάθε επίπεδο.



Εικόνα 28- (Αριστερά) Καμπύλη ROC πρώτου επιπέδου, (Δεξιά) Καμπύλη ROC δευτέρου επιπέδου

Η αξιολόγηση του μοντέλου πραγματοποιήθηκε στο σύνολο δεδομένων δοκιμής το οποίο περιέχει 712 εξετάσεις από τις οποίες οι 221 είναι θετικές στην ΠΕ. Τα αποτελέσματα έχουν ομοιότητες με τα αντίστοιχα στο επίπεδο εκπαίδευσης. Παρατηρούμε στο επίπεδο εικόνας καλή απόδοση του μοντέλου και ανεπτυγμένη ικανότητα να διαχωρίζει τις κλάσεις μεταξύ τους. Σε επίπεδο εξέτασης διαφαίνεται ότι το μοντέλο δεν διαθέτει πολύ καλή ικανότητα στην διάγνωση των περιπτώσεων ΠΕ σε επίπεδο εξέτασης,

6.1 Αξιολόγηση Αποτελεσμάτων

Οι Suman *et al.* [62] και Islam *et al.* [63] εν αντιθέσει των υπόλοιπων εργασιών που αναφέρθηκαν στο Κεφάλαιο 3, επιδιώκουν την πρόβλεψη της ΠΕ πέρα από το επίπεδο εκπαίδευσης και στο επίπεδο της εικόνας. Οι Suman *et al.* [62] επιλέγουν την χρήση του CNN μοντέλου Efficient-Net [66] για την εξαγωγή χαρακτηριστικών στο πρώτο επίπεδο. Παρόλα αυτά δεν χρησιμοποιούν μόνο την επισημείωση για την ΠΕ σε επίπεδο εικόνας, αλλά κάνουν χρήση και των υπόλοιπων επισημειώσεων σε επίπεδο εξέτασης, όπως για παράδειγμα η πρόβλεψη της χρόνιας ΠΕ. Στο πρώτο στάδιο, για την πρόβλεψη της ΠΕ η τιμή της AUC ισούται με 0.96. Σε αυτό το σημείο το μοντέλο

Efficient-Net φαίνεται να αποδίδει καλύτερα. Παρόλα αυτά δεν γίνεται κάποια αναφορά στην επίπτωση που έχει στην εκπαίδευση του μοντέλου η χρήση των επισημειώσεων σε επίπεδο εξέτασης. Επιπλέον, οι Islam *et al.* [63] είναι οι μοναδικοί που δοκίμασαν την απόδοση των μοντέλων ViT και των διαφορετικών παραλλαγών τους για τη πρόβλεψη της ΠΕ. Παρουσιάζουν στην εργασία τους τις τιμές AUC που επιτεύχθηκαν κατά την εκπαίδευση διαφορετικών μεγεθών του μοντέλου ViT και του SWIN Transformer [70]. Στον παρακάτω πίνακα φαίνονται οι τιμές AUC για τις διάφορες παραλλαγές του μοντέλου ViT, όπως επίσης η αντίστοιχη που πέτυχε το μοντέλο ViT που εκπαιδεύτηκε με τη μέθοδο LoRA στην παρούσα εργασία.

Backbone	**Image size**	**Patch size**	**Initialization**	**AUC (%)**
ViT-B - LoRA [63]	512	16	Random	83.85
ViT-B [63]	512	32	Random	82.12
ViT-B [63]	512	32	ImageNet21k	88.47
ViT-B [63]	224	32	ImageNet21k	84.56
ViT-B [63]	224	16	ImageNet21 k	88.26
ViT-B - LoRA (Παρούσα εργασία)	224	16	ImageNet21 k	**91.72**
ViT-B [63]	512	16	ImageNet21 k	90.65
ViT-B [63]	576	16	ImageNet21 k	91.79

Πίνακας 5 - Σύγκριση με αποτελέσματα των Islam et al. [63]. Με έντονο γκρι σημειώνεται η τιμή της AUC στο ViT-LoRA

Παρατηρώντας τον Πίνακα 5, προκύπτει ότι παρόλο που η επιλεχθείσα αρχιτεκτονική εκπαιδεύεται σε εικόνες μεγέθους 224 × 224 πετυχαίνει εξίσου καλή τιμή AUC με αρχιτεκτονικές που αξιοποιούν εικόνες υψηλότερη ανάλυσης. Στο σημείο αυτό να σημειωθεί ότι, οι Islam *et al.*[63] έδειξαν ότι οι η τυχαία αρχικοποίηση των βαρών στα μοντέλα και η εκπαίδευσή τους από την αρχή χωρίς την αξιοποίηση προηγούμενης μάθησης οδηγεί σε χαμηλότερες τιμές AUC. Αυτό αποτέλεσε και σημείο αναφοράς για την αξιοποίηση προεκπαιδευμένου μοντέλου στην παρούσα εργασία. Αξίζει επιπλέον να αναφερθεί, ότι οι Islam *et al.* [63] αναφέρουν στην εργασία

τους ότι οι ώρες που απαιτήθηκαν για την εκπαίδευση (fine-tuning) του μοντέλου ViT-B ξεπερνούν τις 16 με τον αριθμό των παραμέτρων να ανέρχεται στους 88 εκατομμύρια. Στην αρχιτεκτονική της παρούσας εργασίας, από την άλλη μεριά, με την αξιοποίηση της μεθόδου LoRA και την επιλογή του βαθμού ίσου με 4, οι παράμετροι που ρυθμίζονται συνολικά αποτελούν το 0.17% των συνολικών παραμέτρων του μοντέλου ViT. Η διάρκεια μίας εποχής ήταν μικρότερη από 1 ώρα. Λαμβάνοντας υπόψη όλα τα παραπάνω, προκύπτει ότι η εκπαίδευση με τη μέθοδο LoRA είναι αποτελεσματική στην προσαρμογή των παραμέτρων του προεκπαιδευμένου μοντέλου ώστε να επιτύχει μία καλή απόδοση στο συγκεκριμένο πρόβλημα που εξετάζεται, με το εν λόγω σύνολο δεδομένων δίχως να απαιτείται η ρύθμιση και των 88 εκατομμύριων παραμέτρων. Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα εργασία προσφέρει μια επιβεβαίωση της απόδοσης της μεθόδου αυτής στο πρόβλημα της ΠΕ. Τέλος, οι Islam *et al.* [63] για την εκπαίδευση των μοντέλων τους χρησιμοποίησαν το πλήρες σύνολο των δεδομένων. Στην παρούσα εργασία, αξιοποιήθηκε κατά την εκπαίδευση του μοντέλου η μέθοδος υπό-δειγματοληψίας. Ωστόσο, φαίνεται ότι η εκπαίδευση με λιγότερα είναι επαρκής για την επίτευξη συγκρίσιμων αποτελεσμάτων, παρά την απώλεια πληροφορίας από την κλάση πλειονότητας.

Οι Suman *et al.* [62] χρησιμοποιούν για την πρόβλεψη της ΠΕ σε επίπεδο εξέτασης το μοντέλο BiLSTM, επιτυγχάνοντας AUC ίση με 0.95 σε ανεξάρτητο σύνολο δεδομένων, διαφορετικό από το σύνολο που χρησιμοποιήθηκε στην εκπαίδευση. Στο σύνολο αυτό η AUC που επιτεύχθηκε είναι ίση με 0.85. Παρόλα αυτά, στην εργασία τους δεν αναφέρουν περισσότερες λεπτομέρειες σχετικά με την εκπαίδευση του μοντέλου ώστε να μελετηθούν τα σημεία στα οποία μπορεί να οφείλεται η καλύτερη απόδοσή του. Οι Islam *et al.* [63] χρησιμοποιήσαν στη θέση του LSTM το μοντέλο E-ViT, όπως το ονόμασαν, το οποίο αποτελεί ένα μοντέλο ViT που λαμβάνει τις αναπαραστάσεις των εικόνων ως είσοδο. Το μοντέλο E-ViT πετυχαίνει AUC 0.90 και αποτελείται από 105 εκατομμύρια παραμέτρους. Στην εργασία δοκιμάζουν και την απόδοση του μοντέλου BiGRU, το επιτυγχάνει AUC 0.88. Η υπεροχή των μοντέλων αυτών συγκριτικά με το μοντέλο που χρησιμοποιήθηκε στην παρούσα εργασία μπορεί να έγκειται στο ρόλο της διάστασης των δεδομένων εισόδου. Για την εκπαίδευση στο δεύτερο επίπεδο επέλεξαν να ενώσουν τις αναπαραστάσεις των εικόνων δύο γειτονικών εικόνων. Έτσι, κατέληξαν σε αναπαραστάσεις διάστασης 6144, που είναι πολύ μεγαλύτερη της διάστασης που χρησιμοποιήθηκε στα πειράματα της παρούσας

εργασίας. Ακόμη, για την εκπαίδευση στο δεύτερο επίπεδο αξιοποιήθηκαν αναπαραστάσεις που εξήχθησαν από μοντέλα που πέτυχαν μεγάλες τιμές AUC. Αυτό σημαίνει ότι οι αναπαραστάσεις που χρησιμοποιήθηκαν στο πρώτο επίπεδο είναι πολύ καλά προσαρμοσμένες στο να περιέχουν την απαραίτητη πληροφορία ως προς την ύπαρξη ή μη της ΠΕ στην εκάστοτε εικόνα. Η χρήση αντιπροσωπευτικών αναπαραστάσεων εικόνων συμβάλλουν στην απόδοση του μοντέλου στο δεύτερο επίπεδο καθώς η εκπαίδευσή του στηρίζεται σε αυτές τις αναπαραστάσεις. Μπορούμε να υποθέσουμε, έτσι, ότι μοντέλα με υψηλότερη απόδοση στο πρώτο επίπεδο δημιουργούν πιο χρήσιμες αναπαραστάσεις για την αποτελεσματικότερη εκπαίδευση του μοντέλου στο δεύτερο επίπεδο. Αυτό μπορεί να επιβεβαιωθεί και από τα αποτελέσματα αυτής της εργασίας, καθώς παραπάνω δείξαμε ότι η χρήση ενός καλύτερου μοντέλου στο πρώτο επίπεδο εκπαίδευσης επιδρά θετικά στην τιμή της AUC.

7. Συμπεράσματα και Μελλοντικές Προσεγγίσεις

Στην παρούσα εργασία , μελετήθηκε η συμβολή της αρχιτεκτονικής του ViT στο πρόβλημα της πρόβλεψης της ΠΕ με τη χρήση εικόνων CTPA. Συγκεκριμένα , αξιοποιήθηκε για την κατασκευή των αναπαραστάσεων των εικόνων των εξετάσεων της ΠΕ αφού πρώτα εκπαιδεύτηκε στο πρόβλημα της ταξινόμησής τους σε θετικές και αρνητικές. Το μοντέλο εκπαιδεύτηκε με δύο τρόπους και μελετήθηκε η συμβολή του κάθε τρόπου στην απόδοση του μοντέλου. Ο πρώτος τρόπος, περιλαμβάνει την εκπαίδευση μόνο του τελευταίου επιπέδου ταξινόμησης διατηρώντας σταθερές τις υπόλοιπες παραμέτρους του μοντέλου, που έχει ήδη μάθει από την εκπαίδευσή του σε πολύ μεγαλύτερο σύνολο δεδομένων, όπως αυτό του ImageNet. Ο δεύτερος τρόπος αφορά τη χρήση μίας καινούριας μεθόδου fine-tuning που ονομάζεται Low Rank Adaptation , η οποία τον τελευταίο καιρό αξιοποιείται όλο και περισσότερο στην εκπαίδευση μεγάλων μοντέλων που περιέχουν εκατομμύρια παραμέτρους. Παρόλα αυτά η μέθοδος αυτή δοκιμάστηκε αρχικά σε μοντέλα που αξιοποιούνται στον τομέα NPL. Τον τελευταίο καιρό, όμως, αξιοποιείται όλο και περισσότερο στον τομέα της υπολογιστικής όρασης.

Τα πειράματα που διενεργήθηκαν κατέδειξαν την υπεροχή της μεθόδου LoRA στην αποτελεσματικότερη εκπαίδευση του μοντέλου και την ενίσχυση της ακρίβειας ανίχνευσης της ΠΕ. Συγκριτικά με τη μοναδική εργασία , η οποία ομοίως με την

παρούσα εργασία μελέτησε την απόδοση μοντέλων ViT στην πρόβλεψη της ΠΕ, φαίνεται ότι η εκπαίδευση με τη χρήση LoRA μπορεί να επιτύχει παρόμοια ή και καλύτερα αποτελέσματα από τη μέθοδο fine-tuning που ρυθμίζει όλες τις παραμέτρους του μοντέλου κατά την εκπαίδευση. Επίσης, η μέθοδος LoRA επιτρέπει την αποδοτικότερη εκπαίδευση των κρυφών επιπέδων του μοντέλου, ακόμα και όταν η πρόσβαση σε διαθέσιμους υπολογιστικούς πόρους είναι περιορισμένη, όπως συνέβη στην παρούσα εργασία. Επιπλέον, η εκπαίδευση με τη χρήση της υποδειγματοληψίας (Balance Class Sampler) φαίνεται να αποδίδει καλά στο συγκεκριμένο πρόβλημα, παρά την απώλεια πληροφορίας και τη χρήση λιγότερων δειγμάτων εκπαίδευσης. Μία πιθανή εξήγηση σε αυτό, αποτελεί το γεγονός της υψηλής συσχέτισης μεταξύ των γειτονικών τομών της ίδιας εξέτασης. Συνεπώς, η χρήση λιγότερων αλλά πιο αντιπροσωπευτικών δειγμάτων αρκεί για την επαρκή εκπαίδευση του μοντέλου. Τέλος, η εκπαίδευση στο δεύτερο επίπεδο φαίνεται να εξαρτάται από τις αναπαραστάσεις στο πρώτο επίπεδο.

Παρόλα αυτά, τα αποτελέσματα και στα δύο στάδια εκπαίδευσης παραμένουν χαμηλότερα σε σύγκριση με τα αντίστοιχα αποτελέσματα της βιβλιογραφίας. Ένας πρώτος λόγος για αυτό μπορεί να σχετίζεται με τον τρόπο αντιμετώπισης της ανισορροπίας του συνόλου δεδομένων. Η μέθοδος της υποδειγματοληψίας χρησιμοποιεί πολύ λιγότερα δείγματα από την κλάση της πλειονότητας στο σύνολο δεδομένων εκπαίδευσης με αποτέλεσμα την απώλεια πληροφορίας από την κλάση αυτή. Από την άλλη μεριά, η μέθοδος της υπέρ-δειγματοληψίας έχει σαν αποτέλεσμα τη χρήση ίδιων δειγμάτων από την κλάση μειονότητας. Αυτό μπορεί να έχει σαν αποτέλεσμα την αδυναμία γενίκευσης του μοντέλου και την αδυναμία εκμάθησης χρήσιμης πληροφορίας από την κλάση μειονότητας. Η διερεύνηση ενός πιο αποδοτικού τρόπου εξισορρόπησης των κλάσεων στο σύνολο δεδομένων εκπαίδευσης μπορεί να αποτελέσει αντικείμενο μελλοντικής έρευνας. Επιπλέον, η επιλογή κατάλληλου μοντέλου στο πρώτο στάδιο εκπαίδευσης σχετίζεται σε μεγάλο βαθμό με την τελική απόδοση του μοντέλου. Η απόδοση στο δεύτερο στάδιο φαίνεται να εξαρτάται ιδιαίτερα από αυτή του πρώτου σταδίου, όπως φάνηκε και στα πειράματα στην παρούσα εργασία. Σε αυτά φάνηκε ότι η χρήση αναπαραστάσεων από μοντέλο που πετυχαίνει καλύτερη απόδοση, επιτρέπει στο μοντέλο του δεύτερου σταδίου να εκπαιδεύεται αποτελεσματικότερα. Συνεπώς, ένα επόμενο ερευνητικό βήμα θα αφορά τη δοκιμή διαφορετικών αρχιτεκτονικών που αποτελούν εξέλιξη της αρχικής

αρχιτεκτονικής του ViT , τα οποία πετυχαίνουν καλύτερα αποτελέσματα σε σχέση με τη βασική αρχιτεκτονική.. Επιπλέον, η χρήση της μεθόδου LoRA στην εκπαίδευση των μοντέλων αυτών αποτελεί μία ενδιαφέρουσα δοκιμή για την εξέταση της συνεισφοράς της μεθόδου στην συνολική απόδοση του μοντέλου. Τέλος, ένας ακόμη στόχος είναι η ερμηνευσιμότητα του μοντέλου, προκειμένου να αναδειχθούν τα σημεία που επηρεάζουν περισσότερο την τελική του πρόβλεψη, όπως επίσης και η διενέργεια μίας αναλυτικότερης ποιοτικής μελέτης για την αξιολόγηση του μοντέλου. Τέλος, η αξιοποίηση περισσότερων υπολογιστικών πόρων κρίνεται σημαντική, καθώς θα βοηθήσει στην διερεύνηση μεγαλύτερων αρχιτεκτονικών με περισσότερες παραμέτρους στο πρόβλημα της ΠΕ. Επιπλέον, θα μειώσει σημαντικά το χρόνο που απαιτείται για την υλοποίηση των πειραμάτων, γεγονός που θα επιτρέψει την αύξηση των δοκιμών και την ταχύτερη αξιολόγηση τους.

8. Βιβλιογραφία

[1] E.-O. Essien, P. Rali, and S. C. Mathai, "Pulmonary Embolism," *The Medical Clinics of North America*, vol. 103, no. 3, pp. 549–564, May 2019, doi: https://doi.org/10.1016/j.mcna.2018.12.013.

[2] G. E. Raskob *et al.*, "Thrombosis," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 34, no. 11, pp. 2363–2371, Nov. 2014, doi: 10.1161/atvbaha.114.304488.

[3] S. Konstantinides *et al.*, "2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS)," *European Heart Journal*, vol. 41, no. 4, pp. 543–603, Aug. 2019, doi: 10.1093/eurheartj/ehz405.

[4] J. Eckelt *et al.*, "Long-term mortality in patients with pulmonary embolism: results in a single-center registry," *Research and Practice in Thrombosis and Haemostasis*, vol. 7, no. 5, p. 100280, Jul. 2023, doi: 10.1016/j.rpth.2023.100280.

[5] S. Barco *et al.*, "Trends in mortality related to pulmonary embolism in the European Region, 2000–15: analysis of vital registration data from the WHO Mortality Database," ˜ *the œLancet. Respiratory Medicine*, vol. 8, no. 3, pp. 277–287, Mar. 2020, doi: 10.1016/s2213-2600(19)30354-6.

[6] D. G. Raptis, K. I. Gourgoulianis, Z. Daniil, and F. Malli, "Time trends for pulmonary embolism incidence in Greece," *Thrombosis Journal*, vol. 18, no. 1, Jan. 2020, doi: 10.1186/s12959-020-0215-7.

[7] K. Lambrini, K. Konstantinos, I. Christos, O. Petros, and T. Areti, "Pulmonary Embolism: A Literature Review," *American Journal of Nursing Science*, vol. 7, no. 3, p. 57, Nov. 2017, doi: 10.11648/j.ajns.s.2018070301.19.

[8] The Royal Australian College of general Practitioners, "Pulmonary embolism: An update," *Australian Family Physician*. https://www.racgp.org.au/afp/2017/november/pulmonary-embolism

[9] P. D. Stein *et al.*, "Multidetector Computed Tomography for Acute Pulmonary Embolism," *New England Journal of Medicine*, vol. 354, no. 22, pp. 2317–2327, Jun. 2006, doi: https://doi.org/10.1056/nejmoa052367.

[10] S. V. Konstantinides, S. Barco, M. Lankeit, and G. Meyer, "Management of pulmonary embolism," *Journal of the American College of Cardiology*, vol. 67, no. 8, pp. 976–990, Mar. 2016, doi: 10.1016/j.jacc.2015.11.061.

[11] B. D. Hutchinson, P. Navin, E. M. Marom, M. T. Truong, and J. F. Bruzzi, "Overdiagnosis of pulmonary embolism by pulmonary CT angiography," *American Journal of Roentgenology*, vol. 205, no. 2, pp. 271–277, Aug. 2015, doi: 10.2214/ajr.14.13938.

[12] M. S. Akhter, H. A. Hamali, A. A. Mobarki, H. Rashid, J. Oldenburg, and A. Biswas, "SARS-COV-2 infection: Modulator of Pulmonary Embolism Paradigm," *Journal of Clinical Medicine*, vol. 10, no. 5, p. 1064, Mar. 2021, doi: 10.3390/jcm10051064.

[13] N. A. DeMonaco, Q. Dang, W. N. Kapoor, and M. V. Ragni, "Pulmonary Embolism Incidence Is Increasing with Use of Spiral Computed Tomography," *The American Journal of Medicine*, vol. 121, no. 7, pp. 611–617, Jul. 2008, doi: https://doi.org/10.1016/j.amjmed.2008.02.035.

[14] J. Bělohlávek, V. Dytrych, and A. Linhart, "Pulmonary embolism, part I: Epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism," *PubMed Central (PMC)*, Jan. 01, 2013.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718593/

[15] S. J. Kligerman, J. W. Mitchell, J. W. Sechrist, A. K. Meeks, J. R. Galvin, and C. S. White, "Radiologist performance in the detection of pulmonary embolism," *Journal of Thoracic Imaging*, vol. 33, no. 6, pp. 350–357, Nov. 2018, doi: 10.1097/rti.0000000000000361.

[16] A. G. Bach *et al.*, "Timing of pulmonary embolism diagnosis in the emergency department," *Thrombosis Research*, vol. 137, pp. 53–57, Jan. 2016, doi: 10.1016/j.thromres.2015.11.019.

[17] R. J. M. Bruls and R. M. Kwee, "Workload for radiologists during on-call hours: dramatic increase in the past 15 years," *Insights Into Imaging*, vol. 11, no. 1, Nov. 2020, doi: 10.1186/s13244-020-00925-z.

[18] T. N. Hanna, C. Lamoureux, E. A. Krupinski, S. Weber, and J.-O. Johnson, "Effect of shift, schedule, and volume on interpretive accuracy: A retrospective analysis of 2.9 million radiologic examinations," *Radiology*, vol. 287, no. 1, pp. 205–212, Apr. 2018, doi: 10.1148/radiol.2017170555.

[19] K. Batra, Y. Xi, S. Bhagwat, A. Espino, and R. M. Peshock, "Radiologist Worklist Reprioritization using Artificial Intelligence: Impact on report turnaround Times for CTPA examinations positive for acute pulmonary embolism," *American Journal of Roentgenology*, vol. 221, no. 3, pp. 324–333, Sep. 2023, doi: 10.2214/ajr.22.28949.

[20] E. Langius-Wiffen *et al.*, “Retrospective batch analysis to evaluate the diagnostic accuracy of a clinically deployed AI algorithm for the detection of acute pulmonary embolism on CTPA,” *Insights Into Imaging*, vol. 14, no. 1, Jun. 2023, doi: 10.1186/s13244-023-01454-1.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv (Cornell University)*, Jan. 2012, doi: 10.48550/arxiv.1207.0580.

[22] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, “Regularization of Neural Networks using DropConnect,” *PMLR*, May 26, 2013. https://proceedings.mlr.press/v28/wan13.html

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[25] Y. Bengio, P. Frasconi, and P. Simard, “The problem of learning long-term dependencies in recurrent networks,” *IEEE International Conference on Neural Networks*, Dec. 2002, doi: 10.1109/icnn.1993.298725.

[26] S. Hochreiter and J. Schmidhuber, “Long Short-Term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[27] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Jan. 1997, doi: 10.1109/78.650093.

[28] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/j.neunet.2005.06.042.

[29] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” *arXiv (Cornell University)*, Jan. 2012, doi: 10.48550/arxiv.1211.5063.

[30] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation functions in deep learning: A comprehensive survey and benchmark,” *Neurocomputing*, vol. 503, pp. 92–108, Sep. 2022, doi: 10.1016/j.neucom.2022.06.111.

[31] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” *arXiv (Cornell University)*, Jan. 2014, doi: 10.48550/arxiv.1409.3215.

[32] D. Bahdanau, K. Cho, and Y. Bengio, “Published as a conference paper at ICLR 2015 NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE,” 2016. Available: https://arxiv.org/pdf/1409.0473

[33] G. Brauwers and F. Frasincar, “A General Survey on attention Mechanisms in Deep Learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023, doi: 10.1109/tkde.2021.3126456.

[34] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, doi: 10.1016/j.neucom.2021.03.091.

[35] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, “An attentive survey of attention models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 5, pp. 1–32, Oct. 2021, doi: 10.1145/3465055.

[36] A. Vaswani *et al.*, “Attention is all you need,” *arXiv (Cornell University)*, Jan. 2017, doi: 10.48550/arxiv.1706.03762.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1810.04805.

[38] Y. Liu *et al.*, “ROBERTA: A robustly optimized BERT pretraining approach,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1907.11692.

[39] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[40] O. Press and L. Wolf, “Using the output embedding to improve language models,” *arXiv (Cornell University)*, Jan. 2016, doi: 10.48550/arxiv.1608.05859.

[41] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2005.14165.

[42] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2010.11929.

[43] M. Iman, K. Rasheed, and H. R. Arabnia, “A review of deep transfer learning and recent advancements,” *arXiv (Cornell University)*, Jan. 2022, doi: 10.48550/arxiv.2201.09679.

[44] H. Zheng *et al.*, “Learn from Model Beyond Fine-Tuning: a survey,” *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2310.08184.

[45] E. J. Hu *et al.*, “LORA: Low-Rank adaptation of Large Language Models,” *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2106.09685.

[46] C. Li, H. Farkhoor, R. Liu, and J. Yosinski, “Measuring the intrinsic dimension of objective landscapes,” *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1804.08838.

[47] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, “Intrinsic dimensionality explains the effectiveness of language model Fine-Tuning,” *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2012.13255.

[48] Y. Zhu *et al.*, “MeLo: Low-rank Adaptation is Better than Fine-tuning for Medical Image Diagnosis,” *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2311.08236.

[49] E. Colak *et al.*, “The RSNA Pulmonary Embolism CT Dataset,” *Radiology. Artificial Intelligence*, vol. 3, no. 2, p. e200254, Mar. 2021, doi: 10.1148/ryai.2021200254.

[50] W. D. Bidgood, S. C. Horii, F. W. Prior, and D. E. Van Syckle, “Understanding and using DICOM, the data interchange standard for biomedical imaging,” *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 199–212, May 1997, doi: 10.1136/jamia.1997.0040199.

[51] M. Larobina, “Thirty years of the DICOM standard,” *Tomography*, vol. 9, no. 5, pp. 1829–1838, Oct. 2023, doi: 10.3390/tomography9050145.

[52] C. Wittram, “How I do it: CT pulmonary angiography,” *American Journal of Roentgenology*, vol. 188, no. 5, pp. 1255–1261, May 2007, doi: 10.2214/ajr.06.1104.

[53] K. Sahi *et al.*, “The value of ‘Liver Windows’ settings in the detection of small renal cell carcinomas on unenhanced computed tomography,” *Canadian Association of Radiologists Journal*, vol. 65, no. 1, pp. 71–76, Feb. 2014, doi: 10.1016/j.carj.2012.12.005.

[54] S. Patil, J. W. Henry, M. Rubenfire, and P. D. Stein, “Neural network in the clinical diagnosis of acute pulmonary embolism,” *Chest*, vol. 104, no. 6, pp. 1685–1689, Dec. 1993, doi: 10.1378/chest.104.6.1685.

[55] G. D. Tourassi, C. E. Floyd, H. D. Sostman, and R. E. Coleman, “Artificial neural network for diagnosis of acute pulmonary embolism: effect of case and observer selection.,” *Radiology*, vol. 194, no. 3, pp. 889–893, Mar. 1995, doi: 10.1148/radiology.194.3.7862997.

[56] J. A. Scott and E. L. Palmer, “Neural network analysis of ventilation-perfusion lung scans.,” *Radiology*, vol. 186, no. 3, pp. 661–664, Mar. 1993, doi: 10.1148/radiology.186.3.8430170.

[57] G. Serpen, D. K. Tekkedil, and M. Orra, “A knowledge-based artificial neural network classifier for pulmonary embolism diagnosis,” *Computers in Biology and Medicine*, vol. 38, no. 2, pp. 204–220, Feb. 2008, doi: 10.1016/j.compbiomed.2007.10.001.

[58] N. Tajbakhsh, M. B. Gotway, and J. Liang, “Computer-Aided pulmonary embolism detection using a novel Vessel-Aligned multi-planar image representation and convolutional neural networks,” in *Lecture notes in computer science*, 2015, pp. 62–69. doi: 10.1007/978-3-319-24571-3_8.

[59] X. Yang *et al.*, “A Two-Stage convolutional neural network for pulmonary embolism detection from CTPA images,” *IEEE Access*, vol. 7, pp. 84849–84857, Jan. 2019, doi: 10.1109/access.2019.2925210.

[60] S.-C. Huang *et al.*, “PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging,” *Npj Digital Medicine*, vol. 3, no. 1, Apr. 2020, doi: 10.1038/s41746-020-0266-y.

[61] D. Rajan, D. Beymer, S. Abedin, and E. Dehghan, “Pi-PE: A Pipeline for Pulmonary Embolism Detection using Sparsely Annotated 3D CT Images.,” *arXiv (Cornell University)*, pp. 220–232, Jan. 2019, [Online]. Available: http://arxiv.org/pdf/1910.02175.pdf

[62] S. Suman *et al.*, “Attention based CNN-LSTM Network for Pulmonary Embolism Prediction on Chest Computed Tomography Pulmonary Angiograms,” in *Lecture notes in computer science*, 2021, pp. 356–366. doi: 10.1007/978-3-030-87234-2_34.

[63] N. U. Islam, Z. Zhou, S. Gehlot, M. B. Gotway, and J. Liang, “Seeking an optimal approach for Computer-aided Diagnosis of Pulmonary Embolism,” *Medical Image Analysis*, vol. 91, p. 102988, Jan. 2024, doi: 10.1016/j.media.2023.102988.

[64] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *arXiv (Cornell University)*, Jan. 2014, doi: 10.48550/arxiv.1409.0575.

[65] P. Safari, M. India, and J. Hernando, “Self-attention encoding and pooling for speaker recognition,” *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2008.01077.

[66] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1905.11946.

[67] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, “Dying ReLU and Initialization: Theory and Numerical examples,” *Communications in Computational Physics*, vol. 28, no. 5, pp. 1671–1706, Jun. 2020, doi: 10.4208/cicp.oa-2020-0165.

[68] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv (Cornell University)*, Jan. 2015, doi: 10.48550/arxiv.1512.03385.

[69] B. Ghojogh and A. Ghodsi, “Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and survey,” *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2304.11461.

[70] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2103.14030.

[71] J. L. Carson *et al.*, “The Clinical Course of Pulmonary Embolism,” *New England Journal of Medicine*, vol. 326, no. 19, pp. 1240–1245, May 1992, doi: https://doi.org/10.1056/nejm199205073261902.

[72] S. Z. Goldhaber, “Pulmonary Embolism,” *The New England Journal of Medicine*, vol. 339, no. 2, pp. 93–104, Jul. 1998, doi: https://doi.org/10.1056/nejm199807093390207.

[73] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, “A Short Note about Kinetics-600,” *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1808.01340.

[74] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv (Cornell University)*, Jan. 2015, doi: 10.48550/arxiv.1502.03167.