ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΑΛΓΟΡΙΘΜΟΙ ΑΝΑΓΝΩΡΙΣΗΣ ΕΙΚΟΝΑΣ ΚΑΙ ΑΝΑΠΤΥΞΗ ANDROID ΕΦΑΡΜΟΓΗΣ ΓΙΑ ΑΤΟΜΑ ΜΕ ΠΡΟΒΛΗΜΑΤΑ ΟΡΑΣΗΣ»

Του φοιτητή
Κωνσταντίνου Σαρανταυγά
Αρ. Μητρώου: 175097

Επιβλέπουσα
Αικατερίνη Ασδρέ
Βαθμίδα Ε.ΔΙ.Π.

Ημερομηνία 10/09/2023

Τίτλος Π.Ε. Αλγόριθμοι αναγνώρισης εικόνας και ανάπτυξη εφαρμογής Android άτομα με προβλήματα όρασης

Κωδικός Π.Ε. 22249

Ονοματεπώνυμο φοιτητή Κωσταντίνου Σαρανταυγάς

Ονοματεπώνυμο εισηγητή Αικατερίνη Ασδρέ

Ημερομηνία ανάληψης Π.Ε. 09/10/2022

Ημερομηνία περάτωσης Π.Ε. 10/09/2023

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η ταξινόμηση εικόνων έχει γίνει αναπόσπαστο μέρος πολλών εφαρμογών και βιομηχανιών, από τα αυτόνομα οχήματα και την ιατρική διάγνωση μέχρι τα μέσα κοινωνικής δικτύωσης και το ηλεκτρονικό εμπόριο. Επιπλέον, ο τομέας της ταξινόμησης εικόνων προσφέρει ένα πλούσιο και γεμάτο προκλήσεις ερευνητικό τοπίο.

Με τη ραγδαία αύξηση των δεδομένων εικόνας που είναι διαθέσιμα στο διαδίκτυο, υπάρχει επιτακτική ανάγκη για ισχυρούς και ακριβείς αλγορίθμους που μπορούν να αναλύουν και να κατανοούν αυτόματα το οπτικό περιεχόμενο. Εμβαθύνοντας σε αυτόν τον τομέα, μπορούν να διερευνηθούν διάφορες τεχνικές και μεθοδολογίες, όπως αρχιτεκτονικές βαθιάς μάθησης, μέθοδοι εξαγωγής χαρακτηριστικών και αλγόριθμοι βελτιστοποίησης, για τη βελτίωση της απόδοσης των συστημάτων ταξινόμησης εικόνων. Από τεχνικής άποψης, η ταξινόμηση εικόνων είναι ένα διεπιστημονικό πεδίο που συνδυάζει την όραση υπολογιστών, τη μηχανική μάθηση και την αναγνώριση προτύπων, με τη ραγδαία εξέλιξη των τομέων αυτών να καλεί την επιστημονική κοινότητα να συμβαδίσει με τους ρυθμούς αυτούς προκειμένου να συνεισφέρει σε χρήσιμες και ενδιαφέρουσες εφαρμογές στην καθημερινή ζωή.

Συνεπώς, η επιλογή ενός θέματος σχετικού με τη μελέτη αλγορίθμων ταξινόμησης εικόνας είναι πιο ενδιαφέρουσα από ποτέ, επιτρέποντας τη συμμετοχή επίδοξων μελλοντικών ερευνητών σε ταχέως αναπτυσσόμενους τεχνολογικούς κλάδους μέσω μιας πρώτης επαφής με τεχνολογίες αιχμής.

Περίληψη

Η ταξινόμηση εικόνων είναι μια θεμελιώδης εργασία στην όραση υπολογιστών που περιλαμβάνει την κατηγοριοποίηση εικόνων σε προκαθορισμένες κλάσεις ή ετικέτες. Έχουν αναπτυχθεί διάφορες τεχνικές για την αντιμετώπιση αυτού του προβλήματος, με στόχο την επίτευξη ακριβών και αξιόπιστων αποτελεσμάτων ταξινόμησης. Τα τελευταία χρόνια, η βαθιά μάθηση έχει φέρει επανάσταση στον τομέα της ταξινόμησης εικόνων. Τα συνελικτικά νευρωνικά δίκτυα (CNN) έχουν αναδειχθεί ως τα πιο επιτυχημένα και ευρέως χρησιμοποιούμενα μοντέλα για το έργο αυτό. Τα CNN είναι ικανά να μαθαίνουν αυτόματα ιεραρχικές αναπαραστάσεις από τα ακατέργαστα δεδομένα εικόνας, επιτρέποντάς τους να καταγράφουν περίπλοκα χαρακτηριστικά σε διαφορετικά επίπεδα αφαίρεσης.

Επιπλέον, οι εξελίξεις στις αρχιτεκτονικές των δικτύων, όπως η εισαγωγή υπολειμματικών συνδέσεων στο ResNet ή μηχανισμών προσοχής στους Vision Transformers (ViTs), έχουν βελτιώσει περαιτέρω τις δυνατότητες των μοντέλων ταξινόμησης εικόνων. Ειδικά οι ViTs έχουν φέρει επανάσταση στον τομέα της ταξινόμησης εικόνων, επιτυγχάνοντας πρωτοφανή αποτελέσματα, μετά τη μεγάλη επιτυχία του πλαισίου Transformer στον τομέα της επεξεργασίας φυσικής γλώσσας.

Επιπροσθέτως η εργασία επικεντρώνεται και στην ανάπτυξη μιας εφαρμογής για άτομα με προβλήματα όρασης, με στόχο τη βελτίωση της καθημερινότητάς τους. Η εφαρμογή επιτρέπει την αναγνώριση προτύπων και συμβόλων μέσω μιας φιλικής προς τον χρήστη διεπαφής και χρησιμοποιεί ένα μοντέλο μηχανικής μάθησης και βαθιάς μάθησης για την αναγνώριση λουλουδιών. Η εφαρμογή θα μπορούσε να αποτελέσει εργαλείο στον τομέα της τεχνολογίας υποβοηθούμενης από υπολογιστές και θα μπορούσε να βελτιώσει, αν γενικευτεί, την ποιότητα ζωής των χρηστών με προβλήματα όρασης, επιτρέποντάς τους να αναγνωρίζουν εικόνες με μεγαλύτερη ευκολία και ακρίβεια.

Συμπερασματικά, οι τεχνικές ταξινόμησης εικόνων έχουν εξελιχθεί σημαντικά, με τα μοντέλα βαθιάς μάθησης, όπως τα CNN και τα ViT, να βρίσκονται στην πρώτη γραμμή του πεδίου, δίνοντας την δυνατότητα να αναπτυχθούν τεχνολογίες που θα κάνουν την καθημερινότητα των ανθρώπων πιο εύκολη.

«Image Recognition Algorithms and Android Application Development for the Visually Impaired»

Konstantinos Sarantavgas

Abstract

Image classification is a fundamental task in computer vision that involves categorizing images into predefined classes or labels. Various techniques have been developed to tackle this problem, with the goal of achieving accurate and reliable classification results. In recent years, deep learning has revolutionized the field of image classification. Convolutional Neural Networks (CNNs) have emerged as the most successful and widely used models for this task. CNNs are capable of automatically learning hierarchical representations from raw image data, allowing them to capture intricate features at different levels of abstraction. Furthermore, advancements in network architectures, such as the introduction of residual connections in ResNet or attention mechanisms in Vision Transformers (ViTs), have further enhanced the capabilities of image classification models. Especially ViTs have revolutionarized the image classification field reaching unprecedented results, following the great success of the Transformer framework in the Natural Language Processing domain.

Additionally, the work also focuses on the development of an application for individuals with visual impairments, aiming to improve their daily lives. The application allows for the recognition of patterns and symbols through a user-friendly interface and utilizes a machine learning and deep learning model for the recognition of flowers. The application could serve as a tool in the field of computer-assisted technology and could improve the quality of life for users with visual impairments when generalized, allowing them to recognize images with greater ease and accuracy.

In conclusion, image classification techniques have evolved significantly, with deep learning models such as CNNs and ViTs leading the way, offering the potential to develop technologies that will make people's daily lives easier.

Ευχαριστίες

Θα ήθελα να εκφράσω από καρδιάς την ευγνωμοσύνη μου προς όλους όσους συνέβαλαν στην ολοκλήρωση αυτής της πτυχιακής εργασίας. Η υποστήριξή τους επιστημονικά, ηθικά ή οικονομικά, ήταν ανεκτίμητη και βοήθησε στην επίτευξη αυτού του στόχου.

Ευχαριστώ την υπεύθυνη καθηγήτρια, κυρία Ασδρέ Αικατερίνη, για την επίβλεψη και την καθοδήγηση καθ' όλη τη διάρκεια της έρευνας. Οι συμβουλές της και η εμπειρία της ήταν καθοριστικές για την ολοκλήρωση αυτής της εργασίας.

Επίσης, θέλω να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την συμβολή τους καθ' όλη τη διάρκεια αυτής της διαδρομής. Τα λόγια δεν μπορούν να εκφράσουν πόσο σημαντική ήταν η παρουσία τους.

Τέλος, ευχαριστώ όλους τους συμφοιτητές μου που μοιράστηκαν ιδέες, συζητήσεις και εμπειρίες κατά τη διάρκεια των σπουδών μας, αποκτώντας νέους στόχους και προοπτικές στον χώρο της πληροφορικής. Πρότυπα και εμπνεύσεις αποτέλεσαν τα έργα και τα παραδείγματα των διδασκόντων του τμήματος καθ’ όλα τα έτη φοίτησής μου στο Διεθνές Πανεπιστήμιο Ελλάδος.

Περιεχόμενα

Κατάλογος Πινάκων

Συντομογραφίες

CNN	Convolutional Neural Network
ViT	Vision Transformer
CapsNet	Capsule Neural Network

Εισαγωγή

Τα τελευταία χρόνια, ο τομέας της όρασης υπολογιστών έχει σημειώσει αξιοσημείωτες εξελίξεις, φέρνοντας επανάσταση στην ικανότητά του ανθρώπου να αντιλαμβάνεται και να κατανοεί τον κόσμο. Από τα αυτόνομα οχήματα που πλοηγούνται σε πολύπλοκα περιβάλλοντα μέχρι τα συστήματα αναγνώρισης προσώπου που ξεκλειδώνουν τα smartphones, η όραση υπολογιστών έχει γίνει μια απαραίτητη τεχνολογία με ευρείες εφαρμογές σε διάφορους τομείς.

Μια από τις βασικές κινητήριες δυνάμεις πίσω από τις πρόσφατες εξελίξεις στην όραση υπολογιστών είναι η έλευση της βαθιάς μάθησης. Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN), μια κατηγορία μοντέλων βαθιάς μάθησης, έχουν επιδείξει πρωτοφανή επιτυχία σε εργασίες όπως η ταξινόμηση εικόνων, η ανίχνευση αντικειμένων και η κατάτμηση. Αξιοποιώντας σύνολα δεδομένων μεγάλης κλίμακας με ετικέτες (labeled datasets) και ισχυρούς υπολογιστικούς πόρους, τα CNN έχουν επιτύχει επιδόσεις ανθρώπινου επιπέδου σε εργασίες που κάποτε θεωρούνταν πρόκληση για τις μηχανές.

Επιπλέον, η διαθεσιμότητα τεράστιων ποσοτήτων επισημασμένων δεδομένων και η αύξηση της υπολογιστικής ισχύος έχουν προωθήσει την ανάπτυξη συνόλων δεδομένων και μοντέλων οπτικής αναγνώρισης μεγάλης κλίμακας. Η προ-εκπαίδευση (pre-training) σε τεράστια σύνολα δεδομένων, όπως το ImageNet ή το COCO, ακολουθούμενη από περαιτέρω ρύθμιση παραμέτρων (fine-tuning) σε συγκεκριμένες εργασίες, έχει γίνει κοινή πρακτική στην έρευνα της όρασης υπολογιστών, οδηγώντας σε αξιοποίηση τεχνικών μεταφοράς μάθησης (transfer learning). Αυτό το παράδειγμα μεταφοράς μάθησης έχει βελτιώσει σημαντικά τη γενίκευση και την αποτελεσματικότητα των μοντέλων υπολογιστικής όρασης, καθιστώντας τα πιο προσαρμόσιμα σε νέους τομείς και μειώνοντας την ανάγκη για εκτεταμένα δεδομένα με ετικέτες σε συγκεκριμένες εφαρμογές. Για το λόγο αυτό, η μεταφορά μάθησης είναι μια ευρέως χρησιμοποιούμενη τεχνική, άμεσα συνυφασμένη με το χώρο της όρασης υπολογιστών και τις εφαρμογές της.

Ένας ακόμη σημαντικός παράγοντας στο χώρο της όρασης υπολογιστών είναι οι εξελίξεις στην τεχνολογία υλικού, όπως οι μονάδες επεξεργασίας γραφικών (GPU) και τα εξειδικευμένα τσιπ, όπως οι μονάδες επεξεργασίας αισθητήρων (TPU), οι οποίες έχουν διαδραματίσει καθοριστικό ρόλο στην επιτάχυνση της εκπαίδευσης και της ανάπτυξης μοντέλων βαθιάς μάθησης. Αυτές οι εξελίξεις στο υλικό επέτρεψαν στους ερευνητές και τους επαγγελματίες να αντιμετωπίσουν πιο σύνθετες οπτικές εργασίες, να επεξεργαστούν σύνολα δεδομένων μεγάλης κλίμακας και να αναπτύξουν συστήματα υπολογιστικής όρασης σε εφαρμογές πραγματικού χρόνου με χαμηλότερη καθυστέρηση. Στο τεχνικό μέρος ακόμη, η εμφάνιση πλαισίων (frameworks) βαθιάς μάθησης ανοικτού κώδικα, όπως το TensorFlow και το PyTorch, έχει εκδημοκρατίσει την πρόσβαση σε αλγορίθμους υπολογιστικής όρασης τελευταίας τεχνολογίας. Αυτά τα πλαίσια παρέχουν στους ερευνητές και τους προγραμματιστές ισχυρά εργαλεία και βιβλιοθήκες, απλοποιώντας την υλοποίηση και τον πειραματισμό προηγμένων μοντέλων υπολογιστικής όρασης. Η κοινότητα ανοικτού κώδικα έχει προωθήσει τη συνεργασία και την ανταλλαγή γνώσεων, επιταχύνοντας τον ρυθμό προόδου στην έρευνα της όρασης υπολογιστών.

Κοιτάζοντας τις μελλοντικές προοπτικές, η όραση υπολογιστών είναι έτοιμη να συνεχίσει την ταχεία πρόοδό της, χάρη στη συνεχιζόμενη έρευνα σε τομείς όπως η τρισδιάστατη όραση, η κατανόηση βίντεο και η εξηγήσιμη τεχνητή νοημοσύνη. Η ενσωμάτωση της όρασης υπολογιστών με αναδυόμενες τεχνολογίες όπως η επαυξημένη πραγματικότητα (AR) και η εικονική πραγματικότητα (VR) έχει τεράστιες δυνατότητες για τον μετασχηματισμό βιομηχανιών, όπως η ψυχαγωγία, η υγειονομική περίθαλψη και η εκπαίδευση. Ωστόσο, με αυτές τις εξελίξεις, είναι ζωτικής σημασίας να

αντιμετωπιστούν ηθικά ζητήματα, διασφαλίζοντας την υπεύθυνη και χωρίς αποκλεισμούς ανάπτυξη και εφαρμογή των τεχνολογιών υπολογιστικής όρασης.

Η εργασία εστιάζει στην εκτενή μελέτη αλγορίθμων ταξινόμησης εικόνας, ένα δημοφιλές και αναπτυσσόμενο πεδίο που ανήκει στον ευρύτερο χώρο της όρασης υπολογιστών. Παρουσιάζονται μια σειρά από σημαντικούς αλγορίθμους και εφαρμογές που σχετίζονται με την επεξεργασία εικόνας καθώς και μια εφαρμογή για κινητές συσκευές που απευθύνεται σε άτομα με προβλήματα όρασης.

Στο πρώτο κεφάλαιο της εργασίας, γίνεται αναφορά στον αλγόριθμο ταξινόμησης εικόνας CNN (Convolutional Neural Network) αναλύοντας τη δομή του αλγορίθμου καθώς και ο τρόπο με τον οποίο επιτυγχάνεται η ταξινόμηση εικόνων. Στο δεύτερο κεφάλαιο, παρουσιάζεται ένας ακόμα σημαντικός αλγόριθμος, ο ViT (Vision Transformer) συγκρίνοντας τον με τα CNN. Στο Κεφάλαιο 3 μελετάται η αρχιτεκτονική NAS (Neural Architecture Search) και ο αλγόριθμος Capsule Neural Network. Γίνεται αναφορά στην συμβολή τους στον τομέα της ταξινόμησης εικόνων καθώς και η σύγκριση τους με τα CNN. Στο τέταρτο κεφάλαιο αναφέρονται εφαρμογές και συσκευές που σχετίζονται με την υποβοήθηση των ατόμων με προβλήματα όρασης. Στη συνέχεια, αναλύεται η υλοποίηση μιας εφαρμογής για κινητές συσκευές, που χρησιμοποιεί τον αλγόριθμο CNN για την υποβοήθηση της προβληματικής όρασης προσφέροντας μια πρακτική προσέγγιση για την εφαρμογή των αλγορίθμων που έχουν εξεταστεί στα προηγούμενα κεφάλαια.

Κεφάλαιο 1ο: Συνελικτικά Νευρωνικά Δίκτυα σε προβλήματα όρασης υπολογιστών

1.1 Εισαγωγή

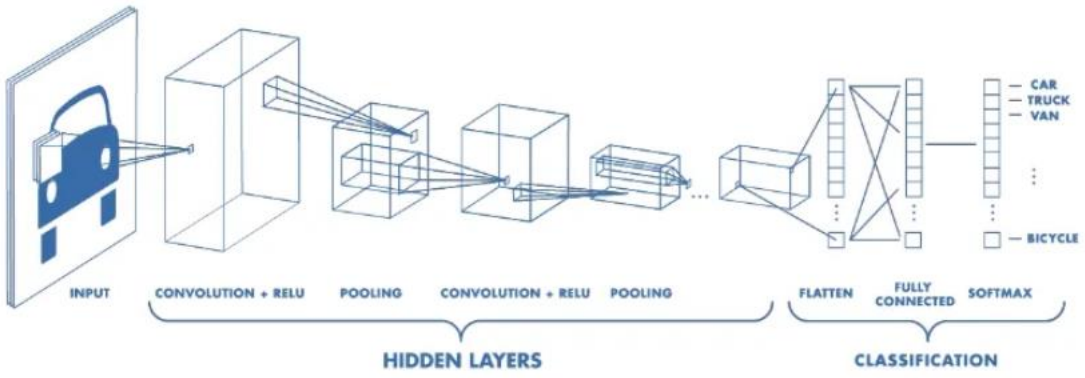
Δεδομένης της ταχείας εξέλιξης των τεχνολογιών βαθιάς μάθησης (deep learning) και των νευρωνικών δικτύων (neural networks), ο χώρος της όρασης υπολογιστών δε θα μπορούσε να μην ωφεληθεί σε μεγάλο βαθμό από τις εξελίξεις αυτές. Η δημοφιλέστερη προσέγγιση αρχιτεκτονικών νευρωνικών δικτύων που ακολουθείται στο χώρο της όρασης υπολογιστών είναι τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks – CNNs).

Πιο συγκεκριμένα, τα συνελικτικά νευρωνικά δίκτυα (CNN) έχουν σχεδιαστεί ειδικά για την επεξεργασία δεδομένων που μοιάζουν με πλέγμα, όπως οι εικόνες. Αποτελούνται από πολλαπλά στρώματα, συμπεριλαμβανομένων των επιπέδων συνελικτικής σύνδεσης (convolutional layers), των επιπέδων συγκέντρωσης (pooling layers) και των πλήρως συνδεδεμένων επιπέδων (fully connected layers). Η διάταξη αυτών των στρωμάτων επιτρέπει στα CNN να μαθαίνουν και να εξάγουν αποτελεσματικά ιεραρχικά χαρακτηριστικά από τις εικόνες εισόδου.

Διαισθητικά, η συνέλιξη βασίζεται σε τρεις θεμελιώδεις έννοιες που έχουν οδηγήσει την πρόοδο των ερευνητών της όρασης υπολογιστών: αραιή αλληλεπίδραση (sparse interaction), διαμοιρασμός παραμέτρων (parameter sharing) και ισοδύναμη αναπαράσταση (equivariant representation). Τα παραδοσιακά στρώματα νευρωνικών δικτύων χρησιμοποιούν πολλαπλασιασμό πινάκων, χρησιμοποιώντας έναν πίνακα παραμέτρων για την περιγραφή των αλληλεπιδράσεων μεταξύ των μονάδων εισόδου και εξόδου. Αυτό συνεπάγεται ότι κάθε μονάδα εξόδου αλληλεπιδρά με κάθε μονάδα εισόδου, με αποτέλεσμα ένα πυκνό μοτίβο αλληλεπίδρασης. Ωστόσο, τα συνελικτικά νευρωνικά δίκτυα (CNN) εισάγουν την έννοια της αραιής αλληλεπίδρασης. Με τη χρήση πυρήνων που είναι μικρότεροι από την είσοδο, όπως για παράδειγμα κατά την επεξεργασία μιας εικόνας με εκατομμύρια ή χιλιάδες pixels, μπορούν να συλληφθούν σημαντικές πληροφορίες μέσα σε περιοχές δεκάδων ή εκατοντάδων pixels . Αυτή η προσέγγιση όχι μόνο μειώνει τον αριθμό των παραμέτρων που πρέπει να αποθηκευτούν, μειώνοντας έτσι τις απαιτήσεις μνήμης του μοντέλου, αλλά και ενισχύει τη στατιστική αποδοτικότητα του μοντέλου [2].

Ο διαμοιρασμός παραμέτρων στα συνελικτικά νευρωνικά δίκτυα (CNNs) δημιουργεί ένα αξιοσημείωτο χαρακτηριστικό γνωστό ως ισοδιακύμανση μετάφρασης (translation equivariance) εντός των στρωμάτων του δικτύου. Αυτή η ιδιότητα δηλώνει ότι όταν η είσοδος μεταβάλλεται ή μετατοπίζεται, η έξοδος θα υποστεί αντίστοιχη μεταβολή με τον ίδιο τρόπο. Με άλλα λόγια, οποιεσδήποτε τροποποιήσεις εφαρμόζονται στην είσοδο θα αντικατοπτρίζονται με συνέπεια στην έξοδο, χάρη στις κοινές παραμέτρους στα στρώματα του δικτύου [2].

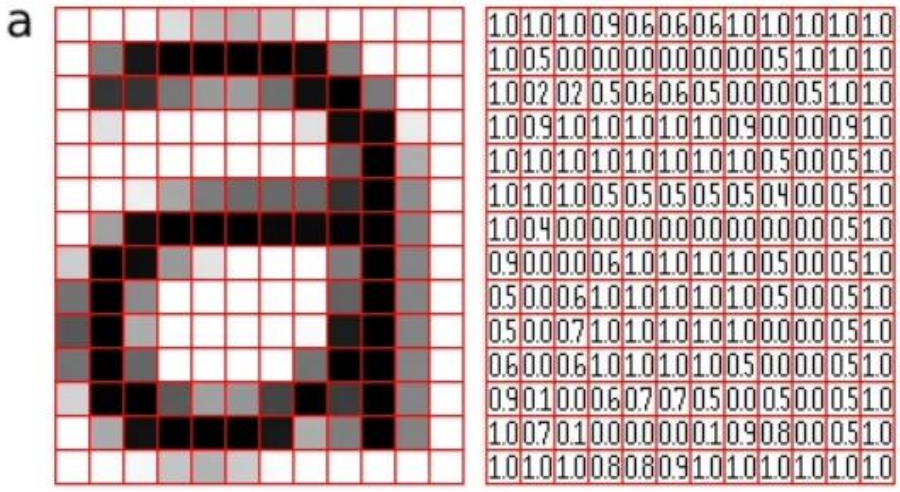
Η γενική δομή ενός CNN παρουσιάζεται στην Εικόνα 1.



Εικόνα 1. Γενική αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου (CNN) [2]

Στο παραπάνω σχήμα αναπαρίσταται αφαιρετικά πώς μια εικόνα εισέρχεται σε ένα CNN και ποια είναι η ακολουθία των σταδίων που συνεισφέρουν στην κατάλληλη αναπαράσταση της εικόνας σε μια μορφή που είναι κατανοητή από ένα νευρωνικό δίκτυο. Αυτό συμβαίνει καθώς ένα νευρωνικό δίκτυο δε μπορεί να αντιληφθεί εικόνα, παρά μόνο αριθμούς. Έτσι, τα χαρακτηριστικά των εικόνων θα πρέπει να μετατραπούν σε μια μορφή κατάλληλη για το νευρωνικό δίκτυο. Τα στάδια αυτά θα αναλυθούν στα επόμενα υποκεφάλαια.

Μια ψηφιακή εικόνα αποτελείται από pixels τα οποία περιέχουν κατάλληλη πληροφορία για στοιχεία της εικόνας (χρώμα, φωτεινότητα, σχήματα κλπ) και είναι τοποθετημένα σε μορφή πλέγματος. Αναφορικά με τα χρώματα, χρησιμοποιούνται 3 κανάλια (RGB) τα οποία διαθέτουν διαφορετικές τιμές προκειμένου να αναπαρασταθούν τα χρώματα. Στις περιπτώσεις ασπρόμαυρων εικόνων, όπως αυτή της Εικόνας 2, τα pixels σηματοδοτούν μόνο τιμές φωτεινότητας.

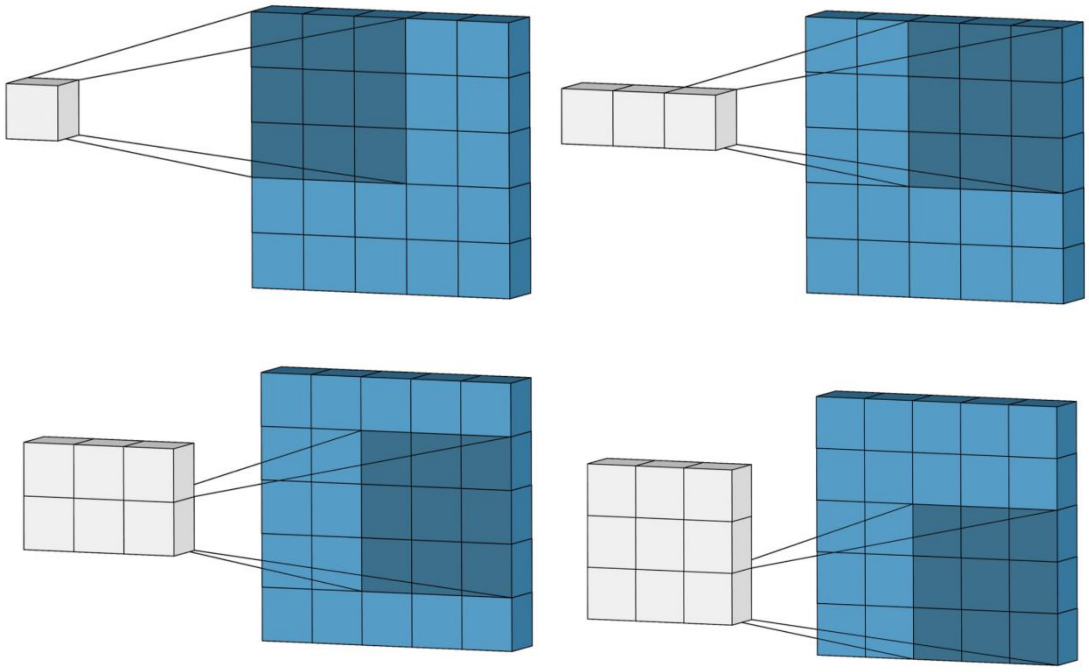


Εικόνα 2. Ασπρόμαυρη εικόνα στα αριστερά και αναπαράσταση σε pixels, τα οποία σηματοδοτούν τη φωτεινότητα σημείων της εικόνας [2].

1.2 Συνελικτικά επίπεδα (convolutional layers)

Τα βασικά δομικά στοιχεία των CNN είναι τα επίπεδα συνελίξεων (convolutional layers). Ένα συνελικτικό επίπεδο ή στρώμα εφαρμόζει ένα σύνολο μαθησιακών φίλτρων, γνωστών επίσης ως πυρήνες ή ανιχνευτές χαρακτηριστικών, στην εικόνα εισόδου ή στους χάρτες χαρακτηριστικών. Τα φίλτρα ολισθαίνουν πάνω στην είσοδο με συστηματικό τρόπο, εκτελώντας στοιχειώδη πολλαπλασιασμό μεταξύ των βαρών του φίλτρου και των αντίστοιχων τιμών εισόδου. Αυτή η διαδικασία έχει ως αποτέλεσμα έναν χάρτη χαρακτηριστικών που αντιπροσωπεύει την ενεργοποίηση κάθε φίλτρου σε διαφορετικές χωρικές θέσεις [1].

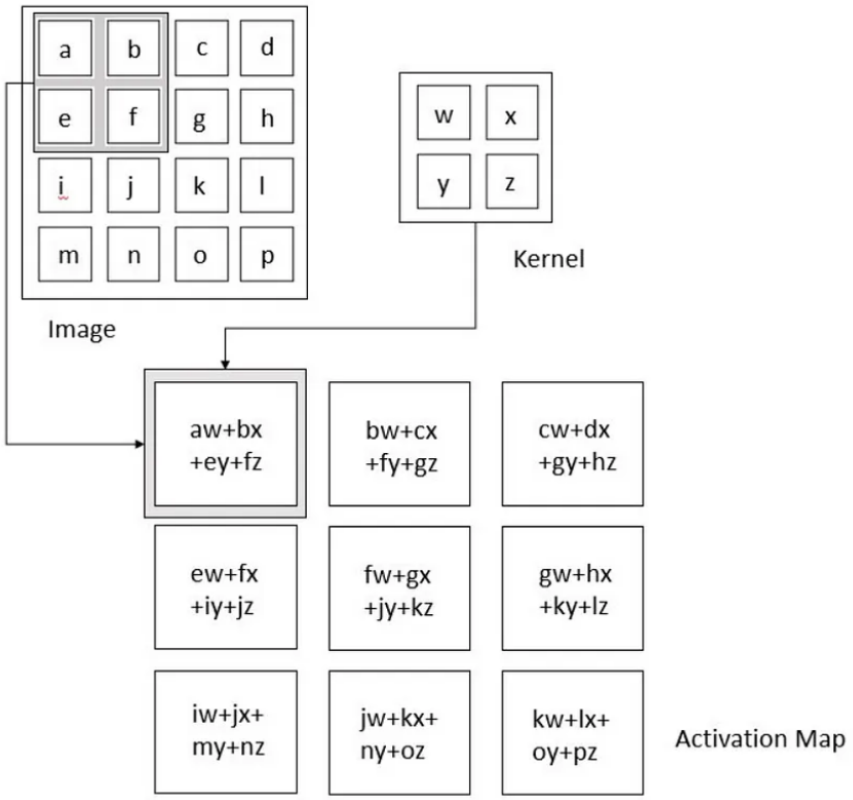
Στην ακόλουθη Εικόνα 3 παρουσιάζονται σχηματικά κάποια ενδιάμεσα βήματα της συνέλιξης, θεωρώντας το μπλε τετράγωνο ως την εικόνα χωρισμένη σε μικρότερα τμήματα, τα οποία αντιστοιχούν στο μέγεθος K ενός φίλτρου (το K είναι ένας αριθμός που επιλέγεται από το χρήστη). Κάθε υποτετράγωνο μεγέθους KxK (στην περίπτωση που παρουσιάζεται, το K=3 και τα μικρότερα τμήματα είναι 3x3), ουσιαστικά προβάλλεται σε μικρότερες διαστάσεις (1x1) μέσω της διαδικασίας της συνέλιξης. Έτσι, πρακτικά η εικόνα συμπιέζεται σε μια αναπαράσταση διαφορετικής διάστασης προκειμένου να οδηγηθεί σε περαιτέρω στάδια επεξεργασίας [1, 2].



Εικόνα 3. Εποπτική επεξήγηση της συνέλιξης και της μεταφοράς της εικόνας σε αναπαράσταση μικρότερων διαστάσεων [2].

Από τεχνικής πλευράς, η λειτουργία της συνέλιξης περιλαμβάνει την ολίσθηση των φίλτρων, τα οποία ονομάζονται kernels πάνω στην εικόνα εισόδου ή στους χάρτες χαρακτηριστικών (feature maps) και τον υπολογισμό του εσωτερικού γινομένου (dot product) μεταξύ των βαρών του φίλτρου και των αντίστοιχων τιμών εισόδου (τιμή κάθε pixel) σε κάθε θέση. Αυτή η λειτουργία συλλαμβάνει τοπικά χωρικά μοτίβα και επιτρέπει στο CNN να μαθαίνει σημαντικά χαρακτηριστικά, όπως ακμές, γωνίες και υφές. Κατά τη διαδικασία αυτή παράγεται μια πιο συμπυκνωμένη μορφή της εικόνας ονομαζόμενη

activation map. Στην Εικόνα 4 παρουσιάζεται εποπτικά η διαδικασία παραγωγής του activation map καθώς ολισθαίνει ο kernel πάνω στα pixels της αρχικής εικόνας [2, 3].



Εικόνα 4. Παραγωγή της αναπαράστασης της εικόνας (activation map) στα συνελικτικά επίπεδα μέσω ολίσθησης του φίλτρου πάνω στα pixels της εικόνας [3].

Τα φίλτρα στα συνελικτικά στρώματα λειτουργούν ως μηχανισμοί εξαγωγής χαρακτηριστικών. Κάθε φίλτρο ειδικεύεται στην ανίχνευση ενός συγκεκριμένου οπτικού μοτίβου ή χαρακτηριστικού, όπως γραμμές, καμπύλες ή κλίσεις. Κατά τη διάρκεια της εκπαίδευσης, το CNN μαθαίνει τις βέλτιστες τιμές για τα βάρη των φίλτρων μέσω οπισθοδιάδοσης (backpropagation) και καθόδου κλίσης (gradient descent), διασφαλίζοντας ότι τα φίλτρα συλλαμβάνουν σημαντικά και διακριτικά χαρακτηριστικά από την είσοδο. Η λειτουργία της συνέλιξης μπορεί να τροποποιηθεί με τη χρήση stride και padding. Το stride αναφέρεται στο μέγεθος του βήματος με το οποίο τα φίλτρα κινούνται στην είσοδο. Ένα μεγαλύτερο stride μειώνει τη χωρική ανάλυση των χαρτών χαρακτηριστικών, με αποτέλεσμα μια πιο συμπιεσμένη αναπαράσταση. Το padding, από την άλλη πλευρά, προσθέτει πρόσθετα pixel περιθωρίου γύρω από την είσοδο, επιτρέποντας στα φίλτρα να καλύψουν τις άκρες και τις γωνίες πιο αποτελεσματικά και να διατηρήσουν τη χωρική πληροφορία [1, 3].

Έχουν αναπτυχθεί διάφορες παραλλαγές του στρώματος συνελίξεων για την ενίσχυση της εκφραστικής δύναμης των CNN. Σε αυτές περιλαμβάνονται οι διευρυμένες συνελίξεις, οι οποίες αυξάνουν το δεκτικό πεδίο (receptive field) χωρίς να αυξάνουν σημαντικά τον αριθμό των παραμέτρων, και οι διαχωρίσιμες συνελίξεις κατά βάθος, οι οποίες αναλύουν την τυπική συνέλιξη σε ξεχωριστές συνελίξεις κατά βάθος και κατά σημείο για να μειώσουν την υπολογιστική πολυπλοκότητα [3, 4].

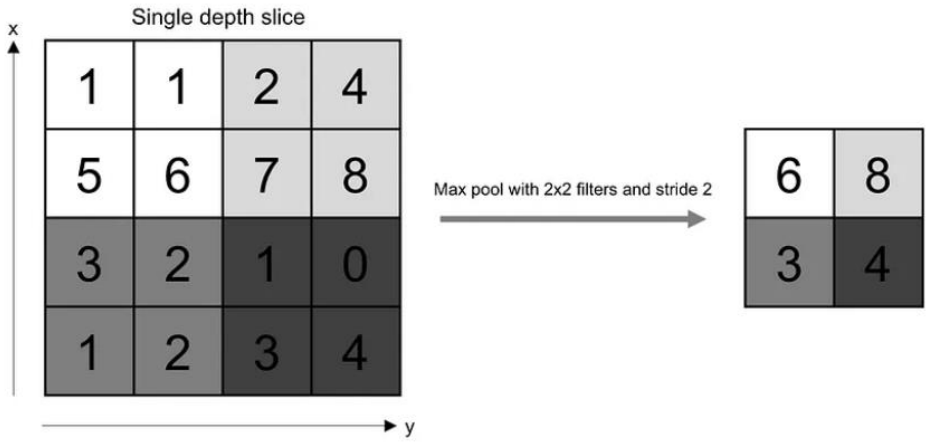
1.3 Επίπεδα συγκέντρωσης (pooling layers)

Τα στρώματα συγκέντρωσης (pooling layers) ακολουθούν τα στρώματα συνελίξεων για να μειώσουν τις χωρικές διαστάσεις των χαρτών χαρακτηριστικών, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες. Αυτή η συμπίεση βελτιώνει την υπολογιστική αποδοτικότητα μειώνοντας τον αριθμό των παραμέτρων και των λειτουργιών στα επόμενα στρώματα. Επιπλέον, η συγκέντρωση βοηθά στην αφαίρεση και τη σύλληψη των πιο σημαντικών χαρακτηριστικών, επιτρέποντας στο CNN να εστιάζει σε υψηλού επιπέδου μοτίβα και σημασιολογία στα δεδομένα εισόδου [1].

Η συγκέντρωση μπορεί να πραγματοποιηθεί με τεχνικές όπως η μέγιστη συγκέντρωση (max pooling) ή η μέση συγκέντρωση (mean pooling). Η μέγιστη συγκέντρωση είναι μια ευρέως χρησιμοποιούμενη τεχνική συγκέντρωσης στα CNN. Χωρίζει τον χάρτη χαρακτηριστικών σε μη επικαλυπτόμενες περιοχές και επιλέγει τη μέγιστη τιμή σε κάθε περιοχή. Η μέγιστη συγκέντρωση βοηθά στην υποδειγματοληψία (subsampling) των χαρτών χαρακτηριστικών, μειώνοντας τις χωρικές τους διαστάσεις, διατηρώντας παράλληλα τα πιο εξέχοντα χαρακτηριστικά. Αυτή η τεχνική βοηθά στη μεταφραστική αναλλοίωτη (translation invariance), επιτρέποντας στο CNN να εστιάζει στην παρουσία των χαρακτηριστικών και όχι στην ακριβή χωρική τους θέση. Η μέση συγκέντρωση υπολογίζει τη μέση τιμή σε κάθε περιοχή συγκέντρωσης, παρέχοντας μια ομαλότερη αναπαράσταση των χαρτών χαρακτηριστικών με μειωμένη δειγματοληψία. Παρόλο που χρησιμοποιείται λιγότερο συχνά από τη μέγιστη συγκέντρωση, η μέση συγκέντρωση μπορεί να είναι επωφελής σε ορισμένα σενάρια όπου είναι επιθυμητή η διατήρηση λεπτομερέστερων πληροφοριών σε όλες τις χωρικές διαστάσεις [3].

Υπάρχουν βέβαια και άλλες συναρτήσεις συγκέντρωσης, όπως ο μέσος όρος της ορθογώνιας γειτονιάς (rectangular neighborhood), η νόρμα L2 της ορθογώνιας γειτονιάς και ένας σταθμισμένος μέσος όρος με βάση την απόσταση από το κεντρικό pixel. Με τη μείωση της δειγματοληψίας των χαρτών χαρακτηριστικών, η συγκέντρωση μειώνει την υπολογιστική πολυπλοκότητα και παρέχει αμετάβλητη μετάφραση, καθιστώντας το δίκτυο πιο ανθεκτικό σε μικρές χωρικές μετατοπίσεις [2, 3].

Στην Εικόνα 5 δίνεται ένα οπτικό παράδειγμα της διαδικασίας συγκέντρωσης, και συγκεκριμένα της μέγιστης συγκέντρωσης.



Εικόνα 5. Χρήση max pooling (μέγιστη συγκέντρωση) για συνόψιση των χαρακτηριστικών σε μικρότερες διαστάσεις [2].

1.4 Συναρτήσεις ενεργοποίησης (activation functions)

Οι συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικότητες (non – linearities) στο CNN, επιτρέποντάς του να μοντελοποιεί πολύπλοκες σχέσεις μεταξύ εισόδου και εξόδου. Η Rectified Linear Unit (ReLU) είναι μια από τις συνήθεις συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στα CNN. Θέτει όλες τις αρνητικές τιμές στο μηδέν και διατηρεί τις θετικές τιμές αμετάβλητες, εισάγοντας ουσιαστικά αραιότητα και ενισχύοντας την ικανότητα του δικτύου να μαθαίνει πιο διακριτικά χαρακτηριστικά. Χρησιμοποιούνται επίσης και άλλες συναρτήσεις ενεργοποίησης, όπως η σιγμοειδής (sigmoid) και η υπερβολική εφαπτομένη, ανάλογα με τις ειδικές απαιτήσεις της εργασίας [3].

Η σιγμοειδής συνάρτηση είναι μια μαθηματική έκφραση που συμπιέζει έναν πραγματικό αριθμό σε ένα εύρος που εκτείνεται από το 0 έως το 1, επιτυγχάνοντας έτσι την εισαγωγή μη γραμμικότητας. Ωστόσο, η σιγμοειδής συνάρτηση παρουσιάζει ένα ανεπιθύμητο χαρακτηριστικό όπου η κλίση γίνεται εξαιρετικά μικρή όταν η ενεργοποίηση είναι κοντά σε οποιοδήποτε άκρο. Αυτή η σχεδόν μηδενική κλίση αποτελεί πρόβλημα κατά την οπισθοδιάδοση, καθώς παρεμποδίζει αποτελεσματικά τη ροή των κλίσεων. Επιπλέον, εάν τα δεδομένα εισόδου σε έναν νευρώνα είναι σταθερά θετικά, η σιγμοειδής έξοδος θα είναι είτε κατά κύριο λόγο θετική είτε κατά κύριο λόγο αρνητική. Κατά συνέπεια, αυτό οδηγεί σε μια ζιγκ-ζαγκ δυναμική στις ενημερώσεις της κλίσης για τα σχετικά βάρη [2, 3].

Η συνάρτηση της υπερβολικής εφαπτομένης (tanh) συμπιέζει έναν πραγματικό αριθμό στην περιοχή [-1, 1]. Παρόμοια με τη σιγμοειδή συνάρτηση, η ενεργοποίηση tanh μπορεί να κορεστεί, δηλαδή να φτάσει στη μέγιστη ή την ελάχιστη τιμή της. Ωστόσο, σε αντίθεση με τους σιγμοειδείς νευρώνες, η έξοδος της tanh έχει κέντρο γύρω από το μηδέν, με ίσες πιθανότητες θετικών και αρνητικών τιμών [2, 3].

1.5 Πλήρως συνελικτικά επίπεδα (fully connected layers)

Τα πλήρως συνδεδεμένα στρώματα, που ονομάζονται επίσης πυκνά στρώματα (dense layers), συνδέουν κάθε νευρώνα του προηγούμενου στρώματος με κάθε νευρώνα του επόμενου στρώματος. Αυτά τα στρώματα τοποθετούνται συνήθως στο τέλος της αρχιτεκτονικής του CNN και είναι υπεύθυνα για συλλογισμούς και αποφάσεις υψηλού επιπέδου, πατώντας πάνω σε χαρακτηριστικά που έχουν ήδη εξαχθεί σε προηγούμενα επίπεδα του CNN. Τα πλήρως συνδεδεμένα στρώματα λαμβάνουν τα μαθημένα χαρακτηριστικά από τα προηγούμενα στρώματα και τα αντιστοιχούν στις κλάσεις εξόδου ή στις προβλέψεις [3].

Σε εργασίες ταξινόμησης (classification), το στρώμα εξόδου συχνά χρησιμοποιεί τη συνάρτηση ενεργοποίησης softmax για να παράγει μια κατανομή πιθανότητας στις διάφορες κλάσεις, επιτρέποντας στο δίκτυο να κάνει προβλέψεις με βάση την κλάση με την υψηλότερη πιθανότητα [3].

Τα πλήρως συνδεδεμένα στρώματα είναι υπεύθυνα για την εκμάθηση πολύπλοκων αντιστοιχίσεων μεταξύ των εξαγόμενων χαρακτηριστικών και της επιθυμητής εξόδου. Μέσω μιας σειράς πολλαπλασιασμών πινάκων και μη γραμμικών ενεργοποιήσεων, το δίκτυο μαθαίνει να αναγνωρίζει περίπλοκα μοτίβα και να λαμβάνει τεκμηριωμένες αποφάσεις με βάση αυτές τις μαθημένες αναπαραστάσεις. Το βάθος και το μέγεθος των πλήρως συνδεδεμένων στρωμάτων επηρεάζουν άμεσα την εκφραστική δύναμη του CNN, επιτρέποντάς του να μοντελοποιεί περίπλοκες σχέσεις και να αποτυπώνει λεπτομερείς λεπτομέρειες στα δεδομένα εισόδου.

Ο αριθμός των νευρώνων στα πλήρως συνδεδεμένα στρώματα μπορεί να ποικίλλει ανάλογα με τη συγκεκριμένη αρχιτεκτονική και τις απαιτήσεις της εργασίας. Μεγαλύτερα δίκτυα με περισσότερους νευρώνες σε πλήρως συνδεδεμένα στρώματα παρέχουν στο μοντέλο αυξημένη ικανότητα εκμάθησης πολύπλοκων αναπαραστάσεων, αλλά απαιτούν επίσης περισσότερους υπολογιστικούς πόρους [3, 4].

1.6 Η συνεισφορά των CNN στην ταξινόμηση εικόνας

Η ταξινόμηση εικόνων (image classification) είναι μια από τις θεμελιώδεις εργασίες στην όραση υπολογιστών και τα συνελικτικά νευρωνικά δίκτυα (CNN) έχουν συνεισφέρει σημαντικά στον τομέα αυτό. Τα CNN έχουν φέρει επανάσταση στην ταξινόμηση εικόνων τα προηγούμενα χρόνια ξεπερνώντας τις επιδόσεις σε ανθρώπινο επίπεδο και επιτυγχάνοντας κορυφαία ακρίβεια σε διάφορα σύνολα δεδομένων αναφοράς [5].

Οι αρχιτεκτονικές CNN που έχουν σχεδιαστεί ειδικά για την ταξινόμηση εικόνων έχουν διαδραματίσει καθοριστικό ρόλο στην πρόοδο του τομέα. Αρχιτεκτονικές όπως το AlexNet, το VGGNet, το GoogLeNet (Inception) και το ResNet έχουν επιδείξει αξιοσημείωτη αύξηση επιδόσεων σε διαγωνισμούς ταξινόμησης εικόνων. Αυτές οι αρχιτεκτονικές αποτελούνται συνήθως από ένα συνδυασμό επιπέδων συνελικτικής ανάλυσης για την εξαγωγή χαρακτηριστικών, επιπέδων συγκέντρωσης για την υποδειγματοληψία και πλήρως συνδεδεμένων επιπέδων για συλλογισμό και πρόβλεψη υψηλού επιπέδου [5, 6, 7].

Τα CNN έχουν θέσει νέα σημεία αναφοράς στην ταξινόμηση εικόνων, ξεπερνώντας τις επιδόσεις των παραδοσιακών μεθόδων μηχανικής μάθησης. Ειδικότερα, ο διαγωνισμός ImageNet Large Scale Visual Recognition Challenge (ILSVRC) έχει χρησιμεύσει ως σημείο αναφοράς για την αξιολόγηση μοντέλων CNN. Τα CNN με τις καλύτερες επιδόσεις σε αυτή την πρόκληση έχουν επιδείξει πρωτοφανή ακρίβεια, μειώνοντας σημαντικά τα ποσοστά σφάλματος και προωθώντας την τελευταία λέξη της τεχνολογίας στην ταξινόμηση εικόνων [5].

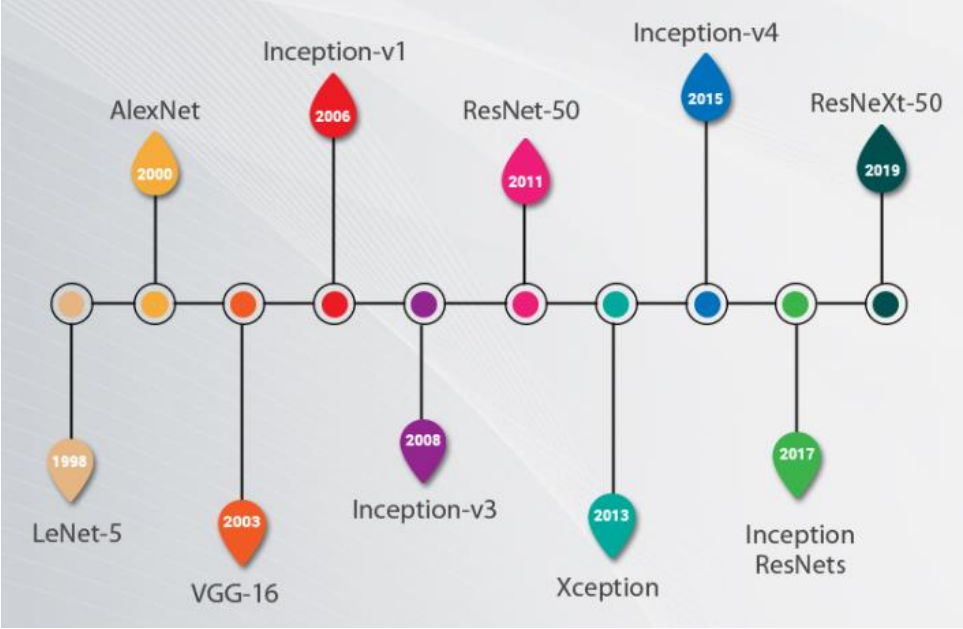
Μια από τις σημαντικότερες συνεισφορές των CNN στην ταξινόμηση εικόνων είναι η ικανότητά τους να μαθαίνουν αυτόματα διακριτικά χαρακτηριστικά. Η ιεραρχική αρχιτεκτονική των CNN επιτρέπει τη σύλληψη οπτικών μοτίβων χαμηλού επιπέδου, όπως οι ακμές και οι υφές, στα αρχικά στρώματα. Καθώς το δίκτυο βαθαίνει, τα επίπεδα συνελίξεων μαθαίνουν πιο αφηρημένες και υψηλού επιπέδου αναπαραστάσεις, επιτρέποντας τη διάκριση πολύπλοκων οπτικών εννοιών και κατηγοριών αντικειμένων. Τα CNN υπερέχουν στην εξαγωγή χαρακτηριστικών που είναι ιδιαίτερα κατατοπιστικά για την ακριβή ταξινόμηση εικόνων. Τα CNN έχουν επιδείξει αξιοσημείωτη ανθεκτικότητα στις μεταβολές των εικόνων εισόδου, όπως αλλαγές στην οπτική γωνία, στις συνθήκες φωτισμού και στις αποκρύψεις αντικειμένων (occlusions). Η προσέγγιση των κοινών παραμέτρων στα συνελικτικά στρώματα επιτρέπει στα CNN να εξάγουν και να γενικεύουν χαρακτηριστικά σε διαφορετικές περιοχές της εισόδου. Αυτή η ιδιότητα καθιστά τα CNN ανθεκτικά στις μετατοπίσεις, τις αλλαγές κλίμακας και τις παραμορφώσεις, επιτρέποντας την ακριβή ταξινόμηση ακόμη και με ποικίλες εμφανίσεις εικόνων [3, 5].

Η μεταφορά μάθησης (transfer learning), μια τεχνική που επιτρέπουν τα CNN, έχει γίνει ένα πολύτιμο εργαλείο για την ταξινόμηση εικόνων. Προεκπαιδευμένα μοντέλα CNN, εκπαιδευμένα σε σύνολα δεδομένων μεγάλης κλίμακας όπως το ImageNet [5], μπορούν να χρησιμοποιηθούν ως σημείο εκκίνησης για νέες εργασίες ταξινόμησης εικόνων. Αξιοποιώντας τις αναπαραστάσεις που έχουν μάθει από αυτά τα προ-εκπαιδευμένα μοντέλα, η μάθηση μεταφοράς επιτρέπει ταχύτερη σύγκλιση και βελτιωμένη απόδοση, ακόμη και με περιορισμένα δεδομένα με ετικέτες. Το fine - tuning επιτρέπει στο μοντέλο να προσαρμόζει τα προ-εκπαιδευμένα χαρακτηριστικά στη συγκεκριμένη εργασία, με αποτέλεσμα τη βελτίωση της ακρίβειας [5, 6].

Η συμβολή των CNN στην ταξινόμηση εικόνων εκτείνεται πέρα από τις επιδόσεις αναφοράς. Τα μοντέλα ταξινόμησης εικόνων με βάση CNN έχουν βρει πρακτικές εφαρμογές σε διάφορους τομείς, όπως η υγειονομική περίθαλψη, η αυτόνομη οδήγηση, η επιτήρηση και οι μηχανές οπτικής αναζήτησης.

Τα CNN διευκολύνουν την ακριβή και αποτελεσματική ταξινόμηση εικόνων σε πραγματικές συνθήκες, επιτρέποντας την πρόοδο σε τομείς που βασίζονται στην κατανόηση και την ανάλυση εικόνων.

Στην ακόλουθη Εικόνα 6 παρουσιάζονται οι πιο σημαντικές αρχιτεκτονικές CNN που έχουν συνεισφέρει σημαντικά σε εργασίες ταξινόμησης εικόνας.



Εικόνα 6. Τα πιο σημαντικά μοντέλα ταξινόμησης εικόνας που βασίζονται σε CNN αρχιτεκτονικές στο πέρασμα του χρόνου [8].

Στη συνέχεια, παρουσιάζονται οι αρχιτεκτονικές αυτές.

LeNet-5

Το LeNet-5, που προτάθηκε από τους LeCun et al. το 1998, ήταν μία από τις πρώτες επιτυχημένες αρχιτεκτονικές CNN. Αποτελούνταν από συνελικτικά στρώματα, στρώματα συγκέντρωσης και πλήρως συνδεδεμένα στρώματα. Το LeNet-5 έπαιξε καθοριστικό ρόλο στην επίδειξη της αποτελεσματικότητας των CNN για εργασίες αναγνώρισης χειρόγραφων ψηφίων [9].

AlexNet

Το AlexNet, που παρουσιάστηκε από τους Krizhevsky et al. το 2012, προώθησε την αναβίωση των CNN. Κέρδισε τον διαγωνισμό ImageNet Large Scale Visual Recognition Challenge (ILSVRC) το 2012 και βελτίωσε σημαντικά την ακρίβεια ταξινόμησης εικόνων. Σε αυτό οφείλεται το ότι χρησιμοποιεί μια βαθύτερη δομή δικτύου σε σύγκριση με προηγούμενα μοντέλα, επιτρέποντας την εκμάθηση πιο σύνθετων χαρακτηριστικών. Η αρχιτεκτονική του AlexNet αποτελούνταν από πολλαπλά συνελικτικά επίπεδα, όπως επίσης και pooling επίπεδα, μαζί με την καινοτόμο χρήση συναρτήσεων ενεργοποίησης ReLU (Rectified Linear Units) στη θέση των παραδοσιακά χρησιμοποιούμενων σιγμοειδών

ενεργοποιήσεων. Οι ενεργοποιήσεις ReLU επιτάχυναν την εκπαίδευση και μετρίασαν το πρόβλημα της εξαφανιζόμενης κλίσης, επιτρέποντας την αποτελεσματικότερη εκπαίδευση βαθύτερων δικτύων [5].

Η επιτυχία του AlexNet μπορεί να αποδοθεί σε διάφορους παράγοντες, συμπεριλαμβανομένης της αρχιτεκτονικής μεγάλης κλίμακας, των ενεργοποιήσεων ReLU, των τεχνικών επαύξησης δεδομένων και της χρήσης παράλληλων υπολογισμών με τη χρήση μονάδων επεξεργασίας γραφικών (GPU). Κατέδειξε σημαντική μείωση των ποσοστών σφάλματος σε σύγκριση με προηγούμενα μοντέλα τελευταίας τεχνολογίας, αναδεικνύοντας τη δύναμη της βαθιάς μάθησης για εργασίες αναγνώρισης εικόνας [5].

Η παρουσίαση του AlexNet είχε βαθύτατο αντίκτυπο στον τομέα της όρασης υπολογιστών, εμπνέοντας την επακόλουθη έρευνα και τις εξελίξεις στις αρχιτεκτονικές CNN. Η βαθιά αρχιτεκτονική του και η επιτυχία του στο ILSVRC έδωσε ώθηση σε περαιτέρω έρευνες για την ανάπτυξη βαθύτερων δικτύων, οδηγώντας στην εμφάνιση αρχιτεκτονικών όπως το VGGNet, το GoogLeNet και το ResNet.

VGGNet

Το VGGNet, που προτάθηκε από τους Simonyan και Zisserman το 2014, είναι γνωστό για την απλότητα και το βάθος του. Η αρχιτεκτονική του VGGNet επικεντρώθηκε στη χρήση πολλαπλών 3x3 συνελικτικών φίλτρων στοιβαγμένων μεταξύ τους, με αποτέλεσμα βαθύτερα δίκτυα. Πέτυχε εξαιρετικές επιδόσεις στις προκλήσεις ILSVRC και έθεσε ένα νέο σημείο αναφοράς για την ακρίβεια ταξινόμησης εικόνων [6].

Το βασικό χαρακτηριστικό του VGGNet είναι η ομοιόμορφη αρχιτεκτονική του, η οποία αποτελείται από πολλά συνελικτικά επίπεδα με μικρά φίλτρα 3x3, ακολουθούμενα από επίπεδα μέγιστης συγκέντρωσης. Διαθέτει διάφορες παραλλαγές, συμπεριλαμβανομένων των VGG16 και VGG19, οι οποίες διαφέρουν ως προς τον αριθμό των στρωμάτων. Η βαθιά δομή της αρχιτεκτονικής επιτρέπει την εκμάθηση όλο και πιο σύνθετων αναπαραστάσεων οπτικών μοτίβων. Χρησιμοποιώντας πολλαπλά στοιβαγμένα στρώματα με μικρά φίλτρα, το VGGNet συλλαμβάνει τόσο χαρακτηριστικά χαμηλού όσο και υψηλού επιπέδου με ιεραρχικό τρόπο. Η απλότητα και η ομοιογενής δομή του VGGNet καθιστούν εύκολη την κατανόηση και την εφαρμογή του. Αυτή η απλότητα έχει επίσης συμβάλει στην ευελιξία και τη μεταφερσιμότητά του. Οι ερευνητές χρησιμοποιούν συχνά το VGGNet ως πρότυπο αναφοράς και μοντέλο αναφοράς για διάφορες εργασίες υπολογιστικής όρασης. Η αρχιτεκτονική του έχει προσαρμοστεί και τελειοποιηθεί για εργασίες όπως η ανίχνευση αντικειμένων, η σημασιολογική τμηματοποίηση και η απόδοση τίτλων εικόνας. Η επιρροή του VGGNet εκτείνεται πέρα από τις επιδόσεις του σε σύνολα δεδομένων αναφοράς (benchmark datasets). Έχει εμπνεύσει μεταγενέστερη έρευνα στον σχεδιασμό της αρχιτεκτονικής και την εξαγωγή χαρακτηριστικών. Οι ερευνητές διερεύνησαν παραλλαγές και βελτιστοποιήσεις της αρχιτεκτονικής VGGNet για να βελτιώσουν την αποδοτικότητα και την επεκτασιμότητά της. Η απλή και αρθρωτή φύση του VGGNet έχει προσφέρει πολύτιμες γνώσεις σχετικά με τις αρχές σχεδιασμού των βαθιών νευρωνικών δικτύων [6].

GoogLeNet (Inception)

Το GoogLeNet (Inception Net), που αναπτύχθηκε από τους Szegedy et al. το 2014, εισήγαγε την έννοια της ενότητας Inception. Η ενότητα Inception χρησιμοποιούσε πολλαπλές παράλληλες πράξεις συνελίξεων με διαφορετικά μεγέθη πυρήνα, επιτρέποντας στο δίκτυο να συλλαμβάνει αποτελεσματικά

χαρακτηριστικά πολλαπλών κλιμάκων. Η αρχιτεκτονική του GoogLeNet μείωσε σημαντικά τον αριθμό των παραμέτρων, διατηρώντας παράλληλα υψηλή ακρίβεια [10].

Η δομή του δικτύου Inception χρησιμοποιεί παράλληλες συνελίξεις διαφορετικών μεγεθών, συμπεριλαμβανομένων των συνελίξεων 1x1, 3x3 και 5x5, για την αποτελεσματική καταγραφή πληροφοριών πολλαπλών κλιμάκων. Αυτές οι παράλληλες συνελίξεις συνδέονται στη συνέχεια κατά μήκος της διάστασης του καναλιού, επιτρέποντας στο δίκτυο να μαθαίνει ταυτόχρονα χαρακτηριστικά σε διάφορα επίπεδα αφαίρεσης. Αυτή η προσέγγιση πολλαπλών κλιμάκων βοηθά το δίκτυο να συλλάβει τόσο λεπτομερείς λεπτομέρειες όσο και σημασιολογικές πληροφορίες υψηλού επιπέδου. Η αρχιτεκτονική του GoogleNet ενισχύει περαιτέρω την αποδοτικότητα με την ενσωμάτωση στρωμάτων συμφόρησης, τα οποία χρησιμοποιούν 1x1 συνελίξεις για να μειώσουν τη διαστατικότητα των χαρτών χαρακτηριστικών πριν από την εφαρμογή πιο δαπανηρών υπολογιστικά πράξεων. Αυτό μειώνει το υπολογιστικό κόστος διατηρώντας παράλληλα την εκφραστική ισχύ. Επιπλέον, βοηθητικοί ταξινομητές (auxiliary classifiers) περιλαμβάνονται σε ενδιάμεσα στρώματα κατά τη διάρκεια της εκπαίδευσης για την αντιμετώπιση του προβλήματος της εξαφανιζόμενης κλίσης και την παροχή κανονικοποίησης [10].

Το GoogleNet σημείωσε αξιοσημείωτη επιτυχία, κερδίζοντας τον διαγωνισμό ImageNet Large Scale Visual Recognition Challenge (ILSVRC) το 2014. Απέδειξε ανώτερη ακρίβεια και υπολογιστική απόδοση σε σύγκριση με προηγούμενες αρχιτεκτονικές. Η επιτυχία του άνοιξε το δρόμο για τις επόμενες εκδόσεις, όπως οι Inception-v2, Inception-v3 και Inception-v4, οι οποίες εισήγαγαν περαιτέρω βελτιστοποιήσεις και βελτιώσεις. Σε γενικότερο επίπεδο, η εισαγωγή του GoogleNet κατέδειξε τη σημασία των παράλληλων συνελίξεων και της αποτελεσματικής συγκέντρωσης πληροφοριών στη βαθιά μάθηση. Η επιτυχία του ανέδειξε τη σημασία της μάθησης χαρακτηριστικών πολλαπλών κλιμάκων και τη δυνατότητα βελτίωσης της ακρίβειας και της υπολογιστικής απόδοσης σε αρχιτεκτονικές CNN [10].

ResNet

Το ResNet, που προτάθηκε από τους He et al. το 2015, εισήγαγε την έννοια των υπολειμματικών συνδέσεων (residual connections). Οι υπολειμματικές συνδέσεις αντιμετώπισαν το πρόβλημα της εξαφάνισης της κλίσης, επιτρέποντας στις κλίσεις να ρέουν απευθείας μέσω συντομεύσεων (skip connections) στο δίκτυο. Η αρχιτεκτονική του ResNet επέτρεψε την εκπαίδευση εξαιρετικά βαθιών δικτύων, οδηγώντας σε βελτιωμένη ακρίβεια και απόδοση [7].

Η βασική καινοτομία του ResNet είναι η ίδια έννοια της υπολειμματικής μάθησης (residual learning), η οποία αποσκοπεί στην ανακούφιση του προβλήματος υποβάθμισης (degradation) που παρατηρείται στα βαθύτερα δίκτυα. Με την εισαγωγή συνδέσεων παράλειψης, το ResNet επιτρέπει στο δίκτυο να μαθαίνει υπολειμματικές απεικονίσεις, αποτυπώνοντας τη διαφορά μεταξύ της αναπαράστασης του τρέχοντος επιπέδου και της επιθυμητής εξόδου. Αυτή η προσέγγιση διευκολύνει την εκπαίδευση πολύ βαθιών δικτύων, καθώς οι κλίσεις μπορούν να ρέουν πιο αποτελεσματικά μέσω του δικτύου και να μετριάζουν το πρόβλημα της εξαφανιζόμενης κλίσης. Η αρχιτεκτονική του ResNet αποτελείται από υπολειμματικά μπλοκ, καθένα από τα οποία περιέχει πολλαπλά επίπεδα συνελίξεων. Αυτά τα μπλοκ ενσωματώνουν αντιστοιχίσεις ταυτότητας, επιτρέποντας την άμεση διάδοση της πληροφορίας χωρίς αλλαγή των διαστάσεων. Επιπλέον, το ResNet εισάγει τη χρήση της ομαλοποίησης παρτίδων, η οποία σταθεροποιεί και επιταχύνει τη διαδικασία εκπαίδευσης [7].

Το ResNet έχει σημειώσει αξιοσημείωτη επιτυχία και έχει ξεπεράσει προηγούμενες αρχιτεκτονικές σε διάφορα κριτήρια αναφοράς για την όραση υπολογιστών. Η ικανότητά του να εκπαιδεύει εξαιρετικά

βαθιά δίκτυα με βελτιωμένη ακρίβεια το έχει καταστήσει δημοφιλή επιλογή για εργασίες όπως η ταξινόμηση εικόνων, η ανίχνευση αντικειμένων και η σημασιολογική κατάτμηση. Ο αντίκτυπος του ResNet εκτείνεται πέρα από την όραση υπολογιστών, με εφαρμογές σε άλλους τομείς, όπως η επεξεργασία φυσικής γλώσσας και η αναγνώριση ομιλίας. Οι εναπομείνασες συνδέσεις του ενέπνευσαν μεταγενέστερες έρευνες, οδηγώντας στην ανάπτυξη διαφόρων παραλλαγών του ResNet, συμπεριλαμβανομένων των ResNet-50, ResNet-101 και ResNet-152, με αυξανόμενο βάθος και βελτιωμένη απόδοση. Οι συνεισφορές του ResNet προώθησαν σημαντικά τον τομέα της βαθιάς μάθησης, αποδεικνύοντας τη σημασία των συνδέσεων παράλειψης και της υπολειμματικής μάθησης για την επιτυχή εκπαίδευση βαθιών νευρωνικών δικτύων. Οι έννοιες που εισήχθησαν στο ResNet άνοιξαν το δρόμο για περαιτέρω εξελίξεις στο σχεδιασμό της αρχιτεκτονικής, τις τεχνικές βελτιστοποίησης και την ερμηνευσιμότητα του μοντέλου [7].

DenseNet

Το DenseNet, που εισήχθη από τους Huang et al. το 2017, εισήγαγε την έννοια των πυκνά συνδεδεμένων μπλοκ. Η αρχιτεκτονική του DenseNet ενθάρρυνε την επαναχρησιμοποίηση χαρακτηριστικών και τη ροή κλίσης συνδέοντας κάθε στρώμα με κάθε άλλο στρώμα με τρόπο feed-forward. Αυτό διευκόλυνε τη διάδοση των πληροφοριών και οδήγησε σε συμπαγή δίκτυα με βελτιωμένη απόδοση παραμέτρων [11].

Το βασικό δομικό στοιχείο του DenseNet είναι το πυκνό μπλοκ, το οποίο αποτελείται από πολλαπλά στρώματα. Σε ένα πυκνό μπλοκ, κάθε στρώμα λαμβάνει χάρτες χαρακτηριστικών από όλα τα προηγούμενα στρώματα εντός του μπλοκ και μεταβιβάζει τους δικούς του χάρτες χαρακτηριστικών στα επόμενα στρώματα. Αυτή η πυκνή συνδεσιμότητα μεγιστοποιεί τη ροή πληροφοριών, διευκολύνει τη διάδοση χαρακτηριστικών και ενισχύει τις δυνατότητες αναπαράστασης του δικτύου. Η πυκνά συνδεδεμένη δομή του DenseNet προσφέρει διάφορα πλεονεκτήματα. Μειώνει τον συνολικό αριθμό των παραμέτρων σε σύγκριση με άλλες αρχιτεκτονικές, καθώς οι χάρτες χαρακτηριστικών επαναχρησιμοποιούνται εντός του δικτύου. Αυτή η αποδοτικότητα των παραμέτρων επιτρέπει βαθύτερα δίκτυα χωρίς σημαντική αύξηση της πολυπλοκότητας του μοντέλου. Το DenseNet παρουσιάζει επίσης ισχυρές ιδιότητες κανονικοποίησης, μειώνοντας την υπερπροσαρμογή και βελτιώνοντας τη γενίκευση [11].

Το DenseNet έχει επιδείξει εντυπωσιακές επιδόσεις σε διάφορες εργασίες υπολογιστικής όρασης, όπως ταξινόμηση εικόνων, ανίχνευση αντικειμένων και σημασιολογική κατάτμηση. Έχει επιτύχει κορυφαία αποτελέσματα σε σύνολα δεδομένων αναφοράς όπως το ImageNet, ξεπερνώντας τις παραδοσιακές αρχιτεκτονικές. Η πυκνή συνδεσιμότητα στο DenseNet επιτρέπει στο δίκτυο να συλλαμβάνει λεπτομερείς λεπτομέρειες, να ενισχύει την εκμάθηση χαρακτηριστικών και να βελτιώνει την ακρίβεια των προβλέψεων [11].

XceptionNet

Το XceptionNet, που προτάθηκε από τον François Chollet το 2017, είναι μια σύγχρονη αρχιτεκτονική συνελικτικού νευρωνικού δικτύου που διευρύνει τα όρια της βαθιάς μάθησης σε εργασίες όρασης υπολογιστών. Εισάγει μια καινοτόμο προσέγγιση στα στρώματα συνελικτικού δικτύου, με στόχο τη σύλληψη πολύπλοκων μοτίβων και τη βελτίωση της αποδοτικότητας των παραμέτρων. Η αρχιτεκτονική βασίζεται στην έννοια της διαχωρίσιμης κατά βάθος συνέλιξης, η οποία χωρίζει την τυπική συνέλιξη σε ξεχωριστές συνέλιξη κατά βάθος και κατά σημείο. Αυτός ο διαχωρισμός επιτρέπει στο δίκτυο να

συλλάβει πιο αποτελεσματικά τις χωρικές σχέσεις και μειώνει το υπολογιστικό κόστος μειώνοντας σημαντικά τον αριθμό των παραμέτρων. Η αρχιτεκτονική του XceptionNet του επιτρέπει να μοντελοποιεί περίπλοκα οπτικά μοτίβα και να μαθαίνει διακριτικά χαρακτηριστικά, διατηρώντας παράλληλα έναν λογικό αριθμό παραμέτρων. Αυτό το καθιστά ιδιαίτερα κατάλληλο για εργασίες αναγνώρισης εικόνων μεγάλης κλίμακας. Το XceptionNet έχει επιτύχει εξαιρετικές επιδόσεις σε διάφορα σύνολα δεδομένων αναφοράς, όπως το ImageNet, αποδεικνύοντας την αποτελεσματικότητά του στην ταξινόμηση εικόνων και σε άλλες εργασίες υπολογιστικής όρασης [12].

Η αρχιτεκτονική XceptionNet έχει επαινεθεί για την ικανότητά της να καταγράφει λεπτομερείς λεπτομέρειες και για την εξαιρετική ακρίβειά της σε σύγκριση με άλλα σύγχρονα μοντέλα. Χρησιμοποιώντας διαχωρίσιμες κατά βάθος συνελίξεις, το XceptionNet προσφέρει μια πιο αποδοτική και αποτελεσματική προσέγγιση στην εξαγωγή χαρακτηριστικών, επιτρέποντας καλύτερη γενίκευση και βελτιωμένη απόδοση. Η επιτυχία του το έχει καταστήσει δημοφιλή επιλογή για ερευνητές και επαγγελματίες που εργάζονται σε εργασίες όρασης υπολογιστών [12].

Οι συνεισφορές του XceptionNet έχουν επεκταθεί πέρα από την ταξινόμηση εικόνων. Έχει επίσης προσαρμοστεί και χρησιμοποιηθεί για εργασίες όπως η ανίχνευση αντικειμένων, η σημασιολογική κατάτμηση και η αναγνώριση προσώπων. Η ευελιξία και οι επιδόσεις του το καθιστούν πολύτιμο εργαλείο στην κοινότητα της όρασης υπολογιστών, οδηγώντας σε εξελίξεις και διευρύνοντας τα όρια της βαθιάς μάθησης.

Inception-ResNet

Το Inception-ResNet, που παρουσιάστηκε από τους Christian Szegedy et al. το 2017, είναι μια ισχυρή αρχιτεκτονική συνελικτικού νευρωνικού δικτύου που συνδυάζει τα πλεονεκτήματα τόσο της μονάδας Inception όσο και των υπολειμματικών συνδέσεων. Αυτή η αρχιτεκτονική βασίζεται στην επιτυχία της μονάδας Inception, η οποία χρησιμοποιεί παράλληλες συνελίξεις διαφορετικών μεγεθών για την αποτελεσματική σύλληψη χαρακτηριστικών πολλαπλών κλιμάκων. Με την ενσωμάτωση υπολειμματικών συνδέσεων, εμπνευσμένων από την αρχιτεκτονική ResNet, το Inception-ResNet ενισχύει τη ροή κλίσης και επιτρέπει την εκπαίδευση ακόμη βαθύτερων δικτύων [13].

Το Inception-ResNet αξιοποιεί τα πλεονεκτήματα τόσο της ενότητας Inception όσο και των υπολειπόμενων συνδέσεων για τη βελτίωση της ακρίβειας και της αποτελεσματικότητας της εκπαίδευσης. Η ενότητα Inception συλλαμβάνει πολύπλοκα μοτίβα και παραλλαγές σε διαφορετικές κλίμακες, ενώ οι υπολειμματικές συνδέσεις αντιμετωπίζουν το πρόβλημα της εξαφανιζόμενης κλίσης και διευκολύνουν τη ροή πληροφοριών μέσω του δικτύου. Αυτός ο συνδυασμός οδηγεί σε βελτιωμένη απόδοση σε διάφορες εργασίες όρασης υπολογιστών. Επίσης, το Inception-ResNet έχει επιτύχει κορυφαία αποτελέσματα σε πολυάριθμα κριτήρια αναφοράς, συμπεριλαμβανομένου του ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Η αρχιτεκτονική του επιτρέπει την αποδοτική εξαγωγή χαρακτηριστικών, μειώνοντας τον αριθμό των παραμέτρων σε σύγκριση με τα παραδοσιακά νευρωνικά δίκτυα συνελικτικού τύπου, διατηρώντας παράλληλα υψηλή ακρίβεια [13].

Η αρχιτεκτονική Inception-ResNet είχε σημαντικό αντίκτυπο στον τομέα της όρασης υπολογιστών, διευρύνοντας τα όρια των επιδόσεων της βαθιάς μάθησης. Η ενσωμάτωση των ενοτήτων Inception και των υπολειμματικών συνδέσεων παρέχει έναν ισχυρό συνδυασμό που επιτρέπει την αποτελεσματικότερη εξαγωγή χαρακτηριστικών και τη βελτιωμένη εκπαίδευση του δικτύου. Επιτυγχάνοντας κορυφαία αποτελέσματα, το Inception-ResNet έχει γίνει δημοφιλής επιλογή για την ταξινόμηση εικόνων, την ανίχνευση αντικειμένων και άλλες εργασίες όρασης υπολογιστών. Η επιτυχία

του Inception-ResNet οδήγησε σε μεταγενέστερες παραλλαγές και προσαρμογές, όπως το Inception-ResNet-v2, οι οποίες βελτιώνουν περαιτέρω την αρχιτεκτονική και βελτιώνουν τις επιδόσεις. Οι ερευνητές και οι επαγγελματίες συνεχίζουν να διερευνούν και να αξιοποιούν το πλαίσιο Inception-ResNet για την αντιμετώπιση διαφόρων προκλήσεων στην όραση υπολογιστών [13].

ResNeXt

Το ResNeXt, που παρουσιάστηκε από τους Xie et al. το 2017, είναι μια αξιοσημείωτη αρχιτεκτονική συνελικτικού νευρωνικού δικτύου που βασίζεται στην επιτυχία του μοντέλου ResNet (Residual Network). Το ResNeXt αντιμετωπίζει την πρόκληση της επεκτασιμότητας και της πολυπλοκότητας του μοντέλου με την εισαγωγή μιας νέας δομής που ονομάζεται " cardinality" (πληθάριθμος) και ενισχύει την αναπαραστατική ισχύ του δικτύου. Αυτή η αρχιτεκτονική χρησιμοποιεί μια ομαδοποιημένη συνελικτική προσέγγιση, όπου οι συνελικώσεις χωρίζονται σε πολλαπλούς κλάδους, καθένας από τους οποίους επεξεργάζεται ένα υποσύνολο των χαρακτηριστικών εισόδου. Οι κλάδοι λειτουργούν παράλληλα, επιτρέποντας στο δίκτυο να συλλαμβάνει ποικίλες και συμπληρωματικές πληροφορίες [14].

Η έννοια του πληθάριθμου (cardinality) στο δίκτυο ResNeXt αναφέρεται στον αριθμό των παράλληλων μονοπατιών ή ομάδων μέσα σε κάθε στρώμα. Αυξάνοντας τον πληθάριθμο, το ResNeXt επεκτείνει τη χωρητικότητα του δικτύου και διευκολύνει την εκμάθηση πιο λεπτομερών χαρακτηριστικών. Αυτή η προσέγγιση βοηθά το ResNeXt να επιτύχει ανώτερες επιδόσεις σε σύγκριση με τα παραδοσιακά δίκτυα, διατηρώντας παράλληλα την υπολογιστική αποδοτικότητα. Το ResNeXt έχει επιδείξει κορυφαία αποτελέσματα σε διάφορες εργασίες υπολογιστικής όρασης, συμπεριλαμβανομένης της ταξινόμησης εικόνων, της ανίχνευσης αντικειμένων και της κατάτμησης εικόνων. Έχει επιτύχει αξιοσημείωτες επιδόσεις σε σύνολα δεδομένων αναφοράς, όπως το ImageNet, ξεπερνώντας προηγούμενες αρχιτεκτονικές. Η ικανότητα της ResNeXt να μοντελοποιεί αποτελεσματικά πολύπλοκα οπτικά μοτίβα και να συλλαμβάνει ποικίλες πληροφορίες συνέβαλε στην επιτυχία της [14].

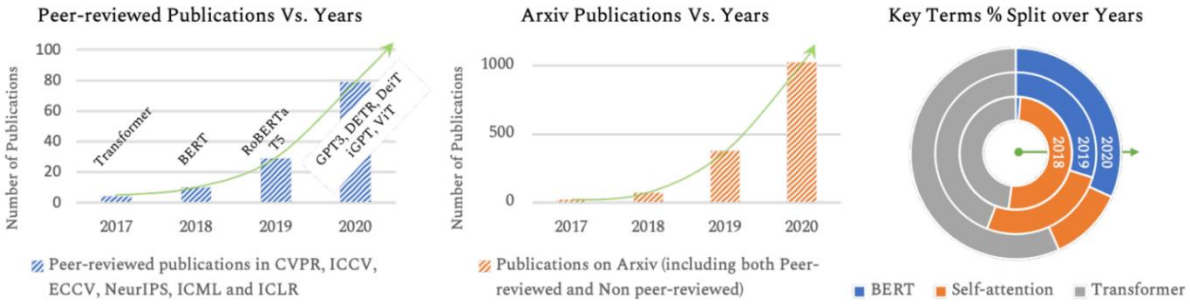
Κεφάλαιο 2ο: Δίκτυα Transformers σε προβλήματα όρασης υπολογιστών

2.1 Εισαγωγή

Αν και τα CNN έχουν αποδείξει εντυπωσιακές ιδιότητες σε διάφορες εργασίες της Όρασης Υπολογιστών, με πρώτη και κύρια την ταξινόμηση εικόνας, στην πορεία έχουν προταθεί και άλλου τύπου αρχιτεκτονικές για διάφορους λόγους.

Μια από τις κυριότερες αρχιτεκτονικές η οποία καλείται να αντικαταστήσει τα CNN είναι οι Transformers. Συγκεκριμένα, οι Transformers αποτελούν μια οικογένεια μοντέλων αρχικά σχεδιασμένων για προβλήματα επεξεργασίας φυσικής γλώσσας (NLP) και η χρήση τους για την ταξινόμηση εικόνων έχει κερδίσει σημαντική προσοχή τα τελευταία χρόνια. Οι Transformers έχουν αποδείξει την αποτελεσματικότητά τους στο χειρισμό διαδοχικών δεδομένων με εξαρτήσεις μεγάλης εμβέλειας. Αυτή η επιτυχία ώθησε τους ερευνητές να διερευνήσουν την εφαρμογή τους σε εργασίες υπολογιστικής όρασης, συμπεριλαμβανομένης της ταξινόμησης εικόνων [15, 16].

Συγκεκριμένα, η ταχεία υιοθέτηση των Transformers σε διάφορους τομείς του AI αναδεικνύεται στην ακόλουθη Εικόνα 7, όπου διαφαίνονται οι αυξητικές τάσεις αναφορικά με τις επιστημονικές δημοσιεύσεις γύρω από το αντικείμενο. Αξίζει να σημειωθεί ότι οι καταγραφόμενες δημοσιεύσεις αφορούν την αξιοποίηση των Transformer τόσο για εφαρμογές γλώσσας όσο και για εφαρμογές όρασης υπολογιστών [26].



Εικόνα 7. Αυξητική τάση δημοσιεύσεων αναφορικά με αρχιτεκτονικές Transformer και συναφείς τεχνικές όπως οι μηχανισμοί προσοχής. [26]

Οι παραπάνω εικόνες αναφέρονται τόσο σε peer – reviewed περιοδικά και συνέδρια (αριστερά), όσο και σε δημοσιεύσεις στο ArXiV (κέντρο). Επίσης, η μεταβολή αναζήτησης όρων στο πέρασμα του χρόνου αναδεικνύεται στο δεξί διάγραμμα, όπου τόσο ο όρος Transformer όσο και η εξειδικευμένη αρχιτεκτονική BERT καταλαμβάνουν μεγάλο μερίδιο στην αναζήτηση [26].

Ένα από τα κύρια κίνητρα για τη χρήση των Transformers στην ταξινόμηση εικόνων είναι η ικανότητά τους να καταγράφουν τις παγκόσμιες εξαρτήσεις και τις αλληλεπιδράσεις μεγάλης εμβέλειας μέσα σε μια εικόνα. Τα CNN, αν και επιτυχημένα στην εξαγωγή τοπικών χαρακτηριστικών μέσω συνελικτικών λειτουργιών, έχουν περιορισμένη ικανότητα μοντελοποίησης σχέσεων μεγάλης εμβέλειας σε ολόκληρη την εικόνα. Οι Transformers, από την άλλη πλευρά, χρησιμοποιούν μηχανισμούς αυτοπροσοχής (self – attention), οι οποίοι θα εξηγηθούν στη συνέχεια, οι οποίοι επιτρέπουν σε κάθε θέση εντός της εικόνας

να παρακολουθεί όλες τις άλλες θέσεις, επιτρέποντας την ενσωμάτωση παγκόσμιας πληροφορίας πλαισίου. Αυτή η ολιστική θεώρηση της εικόνας διευκολύνει την αναγνώριση σύνθετων μοτίβων και εξαρτήσεων που μπορεί να εκτείνονται σε απομακρυσμένες περιοχές. Ένα άλλο πλεονέκτημα των Transformers στην ταξινόμηση εικόνων είναι η ικανότητά τους για παραλληλισμό. Σε αντίθεση με τα CNN, τα οποία βασίζονται σε διαδοχικές λειτουργίες συνελίξεων, οι Transformers μπορούν να επεξεργάζονται ολόκληρη την εικόνα ταυτόχρονα λόγω της παράλληλης φύσης της αυτο-προσοχής. Αυτός ο παραλληλισμός οδηγεί σε βελτιωμένη υπολογιστική αποδοτικότητα και επιτρέπει τον χειρισμό μεγαλύτερων μεγεθών εικόνας, γεγονός που είναι ιδιαίτερα επωφελές όταν πρόκειται για εικόνες υψηλής ανάλυσης [16, 17, 18].

Επιπλέον, οι Transformers προσφέρουν ευελιξία όσον αφορά το μέγεθος εισόδου και τις αναλογίες διαστάσεων. Τα παραδοσιακά CNN απαιτούν την αλλαγή μεγέθους και την περικοπή των εικόνων σε σταθερό μέγεθος, με αποτέλεσμα συχνά την απώλεια πληροφοριών ή την παραμόρφωση. Οι Transformers, ωστόσο, μπορούν να επεξεργάζονται εικόνες αυθαίρετων μεγεθών χωρίς την ανάγκη αλλαγής μεγέθους ή περικοπής, διατηρώντας τις αρχικές λεπτομέρειες και αναλογίες διαστάσεων. Αυτή η ευελιξία πλεονεκτεί όταν πρόκειται για εικόνες διαφορετικών διαστάσεων ή όταν εργάζεστε με σύνολα δεδομένων που περιέχουν εικόνες με διαφορετικές κλίμακες.

Ενώ τα CNN κυριαρχούν στον τομέα της ταξινόμησης εικόνων εδώ και πολλά χρόνια, πρόσφατες μελέτες έχουν δείξει ότι οι Transformers μπορούν να επιτύχουν συγκρίσιμες ή και ανώτερες επιδόσεις σε διάφορες εργασίες ταξινόμησης εικόνων. Ο Vision Transformer (ViT), που παρουσιάστηκε από τους Dosovitskiy et al. το 2020, επέδειξε ανταγωνιστικά αποτελέσματα σε σύνολα δεδομένων αναφοράς όπως το ImageNet, ξεπερνώντας ορισμένα σύγχρονα μοντέλα CNN. Ο ViT και οι μεταγενέστερες παραλλαγές του έδειξαν τις δυνατότητες των Transformers στην ταξινόμηση εικόνων, δίνοντας ώθηση στην περαιτέρω έρευνα και εξερεύνηση σε αυτόν τον τομέα [16, 17, 18, 19, 20].

Επιπλέον, έχει επίσης διερευνηθεί η συγχώνευση των Transformers και των CNN, αξιοποιώντας τα πλεονεκτήματα και των δύο αρχιτεκτονικών. Για παράδειγμα, ο Vision Permutator που προτάθηκε από τους Zheng et al. το 2021 συνδυάζει τη δύναμη της αυτοπροσοχής των Transformers με τις δυνατότητες εξαγωγής τοπικών χαρακτηριστικών των CNNs, επιτυγχάνοντας εντυπωσιακά αποτελέσματα σε εργασίες ταξινόμησης εικόνων. Είναι σημαντικό να σημειωθεί ότι η υιοθέτηση των Transformers στην ταξινόμηση εικόνων αποτελεί ακόμη ενεργό ερευνητικό πεδίο και υπάρχουν προκλήσεις που πρέπει να ξεπεραστούν. Οι Transformers απαιτούν συνήθως μεγάλο αριθμό παραμέτρων, καθιστώντας την εκπαίδευση και την εξαγωγή συμπερασμάτων υπολογιστικά απαιτητικές. Έχουν προταθεί διάφορες στρατηγικές, όπως αποτελεσματικοί μηχανισμοί προσοχής και τεχνικές κλαδέματος δικτύου, για τον μετριασμό αυτών των προκλήσεων και τη βελτίωση της επεκτασιμότητας των μοντέλων Transformer [16, 17, 18, 19, 20].

2.2 Αρχιτεκτονική Transformer

Η αρχιτεκτονική του Transformer ενσωματώνει διάφορα βασικά στοιχεία για την αποτελεσματική επεξεργασία των ακολουθιών εισόδου. Αρχικά, η ακολουθία εισόδου υφίσταται μια διαδικασία ενσωμάτωσης όπου κάθε λέξη μετατρέπεται σε μια διανυσματική αναπαράσταση μεγέθους "d". Για την ενσωμάτωση πληροφοριών θέσης, ένα διάνυσμα κωδικοποίησης θέσης ίδιου μήκους με το διάνυσμα ενσωμάτωσης προστίθεται στο διάνυσμα ενσωμάτωσης κάθε λέξης, αποτυπώνοντας έτσι τις σχετικές θέσεις των λέξεων εντός της ακολουθίας. Μετά την επαύξηση με κωδικοποίηση θέσης, τα μετασχηματισμένα διανύσματα ενσωμάτωσης προχωρούν μέσω του μπλοκ κωδικοποιητή, το οποίο περιλαμβάνει δύο υποστρώματα. Το πρώτο υποεπίπεδο είναι ένας μηχανισμός αυτοπροσοχής που επιτρέπει στο μοντέλο να παρακολουθεί διαφορετικές λέξεις εντός της ακολουθίας εισόδου, ανεξάρτητα από τη θέση τους. Αυτή η αμφίδρομη προσοχή επιτρέπει στον κωδικοποιητή Transformer να συλλαμβάνει αποτελεσματικά τις συνολικές εξαρτήσεις και τις σχέσεις πλαισίου μεταξύ των λέξεων [15, 23].

Το δεύτερο υποεπίπεδο στο μπλοκ κωδικοποιητή είναι ένα νευρωνικό δίκτυο πρόωσης που επεξεργάζεται περαιτέρω κάθε ενισχυμένο διάνυσμα ενσωμάτωσης ξεχωριστά, βελτιώνοντας την αναπαράστασή τους και καταγράφοντας πιο σύνθετα μοτίβα και χαρακτηριστικά. Με την επαναληπτική εφαρμογή αυτών των υποστρωμάτων, το μπλοκ κωδικοποιητή καταγράφει μια πλούσια κατανόηση της ακολουθίας εισόδου, ενσωματώνοντας τόσο τοπικές όσο και παγκόσμιες πληροφορίες. Η αμφίδρομη φύση του κωδικοποιητή Transformer τον διακρίνει από τα παραδοσιακά διαδοχικά μοντέλα, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) ή τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM). Αυτά τα διαδοχικά μοντέλα επεξεργάζονται την ακολουθία εισόδου διαδοχικά, γεγονός που περιορίζει την ικανότητά τους να συλλαμβάνουν εξαρτήσεις μεγάλης εμβέλειας. Αντίθετα, ο κωδικοποιητής Transformer παρακολουθεί όλες τις λέξεις της ακολουθίας εισόδου ταυτόχρονα, επιτρέποντας την αποτελεσματικότερη μοντελοποίηση των αλληλεξαρτήσεων μεταξύ απομακρυσμένων λέξεων. Συνδυάζοντας τον μηχανισμό αυτοπροσοχής με το νευρωνικό δίκτυο τροφοδότησης προς τα εμπρός (feedforward neural network), η αρχιτεκτονική Transformer επιτυγχάνει αξιοσημείωτες επιδόσεις σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας, συμπεριλαμβανομένης της μηχανικής μετάφρασης (machine translation), της γλωσσικής μοντελοποίησης (language modelling) και της ανάλυσης συναισθήματος (sentiment analysis). Η ικανότητά του να συλλαμβάνει τόσο τοπικές όσο και σφαιρικές πληροφορίες, καθώς και η υπολογιστική του απόδοση λόγω της παράλληλης επεξεργασίας, έχουν καταστήσει τον Transformer μια αρχιτεκτονική-κλειδί στον τομέα της βαθιάς μάθησης [21, 22, 23].

Το στοιχείο αποκωδικοποιητή της αρχιτεκτονικής Transformer λειτουργεί με έναν ιδιαίτερο τρόπο. Λαμβάνει ως είσοδο την προβλεπόμενη λέξη εξόδου από το προηγούμενο χρονικό βήμα "t". Ομοίως με τον κωδικοποιητή, η είσοδος του αποκωδικοποιητή εμπλουτίζεται με κωδικοποίηση θέσης για την ενσωμάτωση πληροφοριών σχετικά με τη θέση της λέξης εντός της ακολουθίας. Αυτή η επαυξημένη είσοδος περνάει στη συνέχεια μέσα από τρία υποστρώματα εντός του μπλοκ αποκωδικοποιητή. Το πρώτο υποεπίπεδο εφαρμόζει έναν μηχανισμό κάλυψης για να διασφαλίσει ότι ο αποκωδικοποιητής δεν έχει πρόσβαση σε μελλοντικές λέξεις κατά τη διαδικασία πρόβλεψης. Με την απόκρυψη των επόμενων λέξεων, ο αποκωδικοποιητής επικεντρώνεται αποκλειστικά στις πληροφορίες που είναι διαθέσιμες μέχρι το τρέχον χρονικό βήμα, επιτρέποντάς του να παράγει ακριβείς προβλέψεις. Στο δεύτερο υποεπίπεδο, ο αποκωδικοποιητής λαμβάνει την έξοδο του κωδικοποιητή. Αυτή η σύνδεση επιτρέπει στον αποκωδικοποιητή να παρακολουθεί όλες τις λέξεις στην ακολουθία εισόδου, καταγράφοντας τις σχέσεις και τις εξαρτήσεις μεταξύ των λέξεων. Αξιοποιώντας τις πληροφορίες από τον κωδικοποιητή, ο αποκωδικοποιητής αποκτά μια ολοκληρωμένη κατανόηση της ακολουθίας εισόδου, ενισχύοντας την ικανότητά του να παράγει συνεκτική και σχετική με το πλαίσιο έξοδο. Μετά την επεξεργασία των

υποστρωμάτων, η έξοδος του αποκωδικοποιητή περνά από ένα πλήρως συνδεδεμένο στρώμα, το οποίο βελτιώνει περαιτέρω τις αναπαραστάσεις και τα χαρακτηριστικά που μαθαίνει το μοντέλο. Στη συνέχεια, εφαρμόζεται ένα στρώμα softmax για τη δημιουργία μιας κατανομής πιθανοτήτων πάνω στο λεξιλόγιο, παρέχοντας την πρόβλεψη του μοντέλου για την επόμενη λέξη στην ακολουθία εξόδου [15, 23].

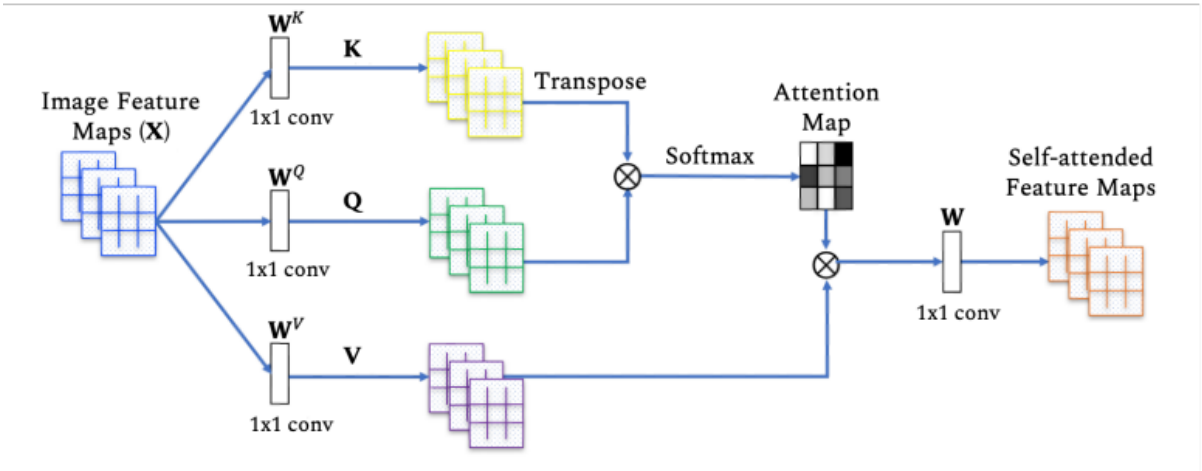
Στην αρχιτεκτονική Transformer αξιοποιούνται οι ακόλουθοι μηχανισμοί προσοχής (attention mechanisms) [23].

Self – attention (αυτοπροσοχή)

Η αυτοπροσοχή, που αναφέρεται επίσης ως ενδοπροσοχή ή προσοχή τετραγωνικού προϊόντος, είναι ένας κρίσιμος μηχανισμός που χρησιμοποιείται στα μοντέλα που βασίζονται στον Transformer για την καταγραφή των εξαρτήσεων μεταξύ διαφορετικών θέσεων μέσα σε μια ακολουθία. Σε αντίθεση με τους συμβατικούς μηχανισμούς προσοχής που εστιάζουν στο εξωτερικό περιεχόμενο, η αυτοπροσοχή επιτρέπει στο μοντέλο να παρακολουθεί διαφορετικά μέρη της ίδιας της ακολουθίας εισόδου. Στον μηχανισμό αυτοπροσοχής, κάθε στοιχείο της ακολουθίας συνδέεται με διανύσματα ερωτήματος, κλειδιού και τιμής. Οι βαθμολογίες προσοχής που υπολογίζονται μεταξύ των ερωτημάτων και των κλειδιών καθορίζουν τη συνάφεια και τη σημασία των διαφόρων στοιχείων της ακολουθίας, ενώ οι τιμές αντιπροσωπεύουν τις πληροφορίες που σχετίζονται με κάθε στοιχείο. Υπολογίζοντας ένα σταθμισμένο άθροισμα των τιμών με βάση τις βαθμολογίες προσοχής, η αυτοπροσοχή επιτρέπει στο μοντέλο να κατανέμει δυναμικά την προσοχή στα πιο συναφή στοιχεία εντός της ακολουθίας [15, 23, 24].

Η δύναμη της αυτοπροσοχής έγκειται στην ικανότητά της να συλλαμβάνει εξαρτήσεις μεγάλης εμβέλειας, επιτρέποντας στο μοντέλο να ενσωματώνει αποτελεσματικά πληροφορίες από απομακρυσμένα μέρη της ακολουθίας. Αυτή η ιδιότητα καθιστά την αυτοπροσοχή ιδιαίτερα πλεονεκτική σε εργασίες επεξεργασίας φυσικής γλώσσας, όπου η κατανόηση των σχέσεων πλαισίου και η σύλληψη των παγκόσμιων εξαρτήσεων είναι ζωτικής σημασίας. Προσέχοντας τα σχετικά στοιχεία εντός της ακολουθίας, η αυτοπροσοχή διευκολύνει την εξαγωγή σημαντικών χαρακτηριστικών και μοτίβων, βελτιώνοντας τελικά την απόδοση του μοντέλου σε εργασίες όπως η μηχανική μετάφραση, η περίληψη κειμένου και η ανάλυση συναισθήματος. Η έρευνα έχει αποδείξει την αποτελεσματικότητα της αυτοπροσοχής σε διάφορα μοντέλα που βασίζονται στον Transformer, συμπεριλαμβανομένου του αρχικού μοντέλου Transformer που εισήγαγαν οι Vaswani et al. (2017). Ο μηχανισμός αυτο-προσοχής επιτρέπει στο μοντέλο να αποτυπώνει πολύπλοκες σχέσεις και εξαρτήσεις με παραλληλοποιήσιμο τρόπο, καθιστώντας το κατάλληλο για το χειρισμό ακολουθιών μεγάλης κλίμακας. Η ικανότητά του να επεξεργάζεται ακολουθίες αποτελεσματικά και αποδοτικά έχει οδηγήσει σε σημαντικές προόδους στην επεξεργασία φυσικής γλώσσας και έχει συμβάλει στην επιτυχία των μοντέλων που βασίζονται στον Transformer σε πολυάριθμες εφαρμογές [15, 24].

Ένα παράδειγμα χρήσης μηχανισμών self-attention σε εφαρμογές όρασης υπολογιστών παρουσιάζεται στην ακόλουθη Εικόνα 8:



Εικόνα 8. Μηχανισμός αυτοπροσοχής ενσωματωμένος σε μοντέλα όρασης υπολογιστών [26].

Σύμφωνα με την παραπάνω Εικόνα 8, μετά τη λήψη της ακολουθίας χαρακτηριστικών εικόνας εισόδου, υπολογίζεται η τριπλέτα (κλειδί, ερώτημα, τιμή), η οποία στη συνέχεια χρησιμοποιείται για τον υπολογισμό των βαρών προσοχής. Αυτά τα βάρη εφαρμόζονται για την αναβάθμιση των τιμών, λαμβάνοντας υπόψη τη συνάφεια κάθε τιμής με το ερώτημα. Η διαδικασία που περιγράφεται εδώ αναπαριστά μία μόνο κεφαλή προσοχής. Για να ληφθούν χαρακτηριστικά εξόδου με την ίδια διάσταση με την είσοδο, εφαρμόζεται μια προβολή εξόδου (W). Αυτή η προβολή διασφαλίζει ότι τα χαρακτηριστικά εξόδου διατηρούν την επιθυμητή διαστατικότητα, επιτρέποντας την περαιτέρω επεξεργασία ή αξιοποίηση στη συνέχεια [26].

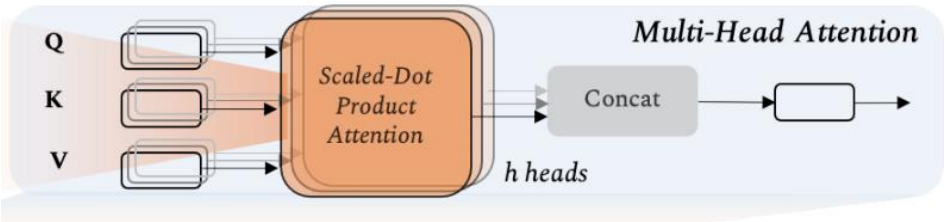
Multi-head attention (προοχοχή πολλαπλών κεφαλών)

Η προσοχή πολλαπλών κεφαλών είναι μια επέκταση του μηχανισμού αυτοπροσοχής που δίνει τη δυνατότητα στα μοντέλα που βασίζονται στον Transformer να παρακολουθούν ταυτόχρονα διαφορετικές πτυχές της ακολουθίας εισόδου. Σε αντίθεση με τον παραδοσιακό μηχανισμό αυτό-προσοχής που χρησιμοποιεί ένα ενιαίο σύνολο ερωτήσεων, κλειδιών και τιμών, η προσοχή πολλαπλών κεφαλών εισάγει πολλαπλά σύνολα αυτών των στοιχείων, καθένα από τα οποία αντιπροσωπεύει μια διαφορετική κεφαλή προσοχής. Στην προσοχή πολλαπλών κεφαλών, η ακολουθία εισόδου μετατρέπεται σε πολλαπλά σύνολα ερωτημάτων, κλειδιών και τιμών, όπου κάθε σύνολο αντιστοιχεί σε μια ξεχωριστή κεφαλή προσοχής. Κάθε κεφαλή προσοχής υπολογίζει ανεξάρτητα τις βαθμολογίες προσοχής μεταξύ του δικού της συνόλου ερωτημάτων και κλειδιών, επιτρέποντάς της να συλλάβει συγκεκριμένα μοτίβα και εξαρτήσεις εντός της ακολουθίας. Οι βαθμολογίες προσοχής καθορίζουν τη σημασία και τη συνάφεια των διαφόρων στοιχείων της ακολουθίας για κάθε κεφαλή προσοχής. Οι τιμές που σχετίζονται με κάθε κεφαλή προσοχής σταθμίζονται στη συνέχεια σύμφωνα με τις βαθμολογίες προσοχής και συνδυάζονται για να σχηματίσουν την τελική έξοδο προσοχής [15, 24].

Χρησιμοποιώντας πολλαπλές κεφαλές προσοχής, το μοντέλο μπορεί να καταγράψει διάφορους τύπους εξαρτήσεων και να εξάγει πιο διαφοροποιημένες σχέσεις εντός της ακολουθίας. Κάθε κεφαλή προσοχής ειδικεύεται στην προσοχή σε διαφορετικές πτυχές, επιτρέποντας στο μοντέλο να συλλάβει διάφορες πληροφορίες σχετικά με το πλαίσιο και να αποκτήσει μια πλουσιότερη κατανόηση της εισόδου. Αυτή η ενισχυμένη ικανότητα μοντελοποίησης της προσοχής πολλαπλών κεφαλών έχει επιδείξει αξιοσημείωτη αποτελεσματικότητα στη βελτίωση της απόδοσης των μοντέλων που βασίζονται στον

Transformer σε ένα ευρύ φάσμα εργασιών επεξεργασίας φυσικής γλώσσας. Η έρευνα έχει δείξει ότι η χρήση πολλαπλών κεφαλών προσοχής οδηγεί σε βελτιωμένη εκμάθηση αναπαράστασης και αυξημένη εκφραστικότητα του μοντέλου. Επιτρέπει μια πιο ολοκληρωμένη και λεπτομερή ανάλυση της ακολουθίας εισόδου, επιτρέποντας στο μοντέλο να συλλαμβάνει σύνθετα μοτίβα και εξαρτήσεις. Η αποτελεσματικότητα της προσοχής πολλαπλών κεφαλών έχει αποδειχθεί σε διάφορα μοντέλα που βασίζονται στον Transformer, συμπεριλαμβανομένου του αρχικού μοντέλου Transformer και μεταγενέστερων παραλλαγών [15, 24].

Στην παρακάτω Εικόνα 9, δίνεται σχηματικά η διαδικασία υπολογισμού της προσοχής πολλαπλών κεφαλών για έναν αριθμό h κεφαλών.



Εικόνα 9. Σχηματική παρουσίαση του υπολογισμού προσοχής πολλαπλών κεφαλών [26].

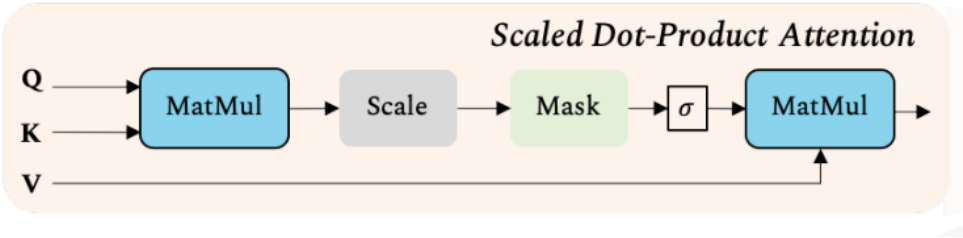
Scaled dot-product attention (Κλιμακωτή προσοχή εσωτερικού γινομένου)

Η κλιμακωτή προσοχή εσωτερικού γινομένου είναι ένα κρίσιμο δομικό στοιχείο των μηχανισμών αυτοπροσοχής που χρησιμοποιούνται στα μοντέλα που βασίζονται στον Transformer. Στην κλιμακωτή προσοχή εσωτερικού γινομένου, η ομοιότητα μεταξύ ερωτημάτων και κλειδιών υπολογίζεται με τη λήψη του σωτερικού γινομένου εμεταξύ τους, το οποίο μετρά τη συνάφεια και τη συγγένεια μεταξύ διαφορετικών στοιχείων εντός της ακολουθίας. Για να διασφαλιστεί ότι οι βαθμολογίες προσοχής είναι κατάλληλα κλιμακωτές, τα γινόμενα τελείας διαιρούνται με την τετραγωνική ρίζα της διάστασης των διανυσμάτων κλειδιών. Αυτός ο παράγοντας κλιμάκωσης αποτρέπει τις βαθμολογίες προσοχής από το να γίνουν πολύ μεγάλες ή πολύ μικρές, διατηρώντας τη σταθερότητα και διευκολύνοντας την αποτελεσματική μάθηση. Οι προκύπτουσες βαθμολογίες προσοχής κανονικοποιούνται στη συνέχεια χρησιμοποιώντας τη συνάρτηση softmax, η οποία τις μετατρέπει σε βάρη που αντιπροσωπεύουν τη σχετική σημασία κάθε κλειδιού σε σχέση με ένα δεδομένο ερώτημα. Αυτά τα βάρη προσοχής αντικατοπτρίζουν τον βαθμό προσοχής ή εστίασης που πρέπει να ανατεθεί σε κάθε κλειδί κατά τη δημιουργία της εξόδου. Το τελικό βήμα περιλαμβάνει τον πολλαπλασιασμό των βαρών προσοχής με τις αντίστοιχες τιμές που σχετίζονται με κάθε κλειδί και το άθροισμά τους. Αυτό το σταθμισμένο άθροισμα τιμών χρησιμεύει ως έξοδος προσοχής, αποτυπώνοντας τις συνδυασμένες πληροφορίες από τα σχετικά στοιχεία της ακολουθίας [15, 24].

Η κλιμακωτή προσοχή εσωτερικού γινομένου έχει υιοθετηθεί ευρέως και έχει αποδειχθεί αποτελεσματική σε διάφορους τομείς, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας και των εργασιών επεξεργασίας εικόνας. Επιτρέπει στο μοντέλο να συλλάβει εξαρτήσεις και σχέσεις εντός της ακολουθίας, επιτρέποντας αναπαραστάσεις με βάση το πλαίσιο και διευκολύνοντας ποικίλες εργασίες, μεταξύ των οποίων είναι και η ταξινόμηση εικόνας. Η αξιοποίηση της κλιμακωτής προσοχής εσωτερικού γινομένου σε μοντέλα που βασίζονται σε Transformers έχει φέρει επανάσταση στον τομέα της βαθιάς μάθησης, παρέχοντας έναν αποδοτικό και αποτελεσματικό μηχανισμό για τη σύλληψη των πολύπλοκων εξαρτήσεων και αλληλεπιδράσεων εντός μιας ακολουθίας. Η απλότητα και η επεκτασιμότητά της την έχουν καταστήσει θεμελιώδες συστατικό πολλών μοντέλων τελευταίας

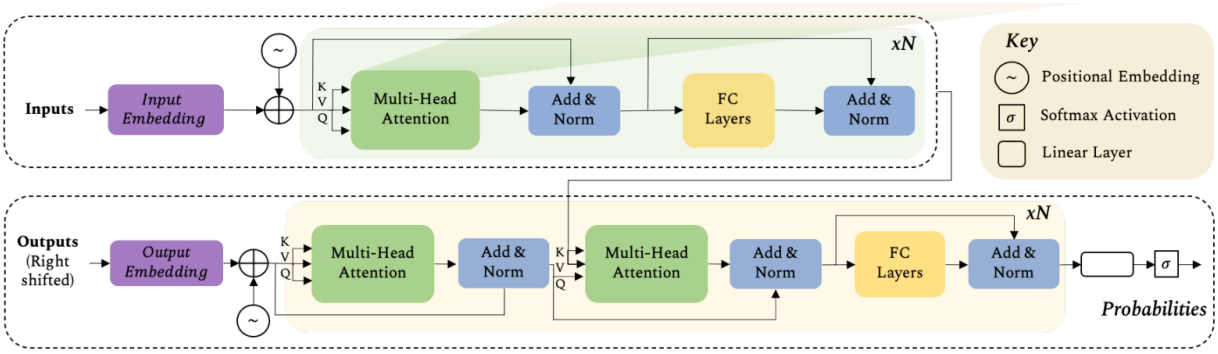
τεχνολογίας, επιτρέποντας σημαντικές προόδους σε διάφορους τομείς της έρευνας και των πρακτικών εφαρμογών [15, 24].

Ένα παράδειγμα της ροής πληροφορίας για τον υπολογισμό της κλιμακωτής προσοχής εσωτερικού γινομένου παρουσιάζεται στην ακόλουθη Εικόνα 10:



Εικόνα 10. Αφαιρετική περιγραφή για τον υπολογισμό κλιμακωτής προσοχής εσωτερικού γινομένου. [26].

Στη συνέχεια, δίνεται μια σχηματική αναπαράσταση της ροής πληροφορίας σε έναν Transformer από την είσοδο έως την έξοδο στην Εικόνα 11. Απεικονίζονται οι μηχανισμοί προσοχής που συνεισφέρουν στην εκμάθηση των κατάλληλων σχέσεων, όπως επίσης και άλλα στοιχεία που θα περιγραφούν στη συνέχεια.



Εικόνα 11. Σχηματική αναπαράσταση της αρχιτεκτονικής Transformer [26].

Ακολούθως θα αναλυθούν τα επιμέρους modules που συναποτελούν έναν Transformer.

Κωδικοποιητής (Encoder)

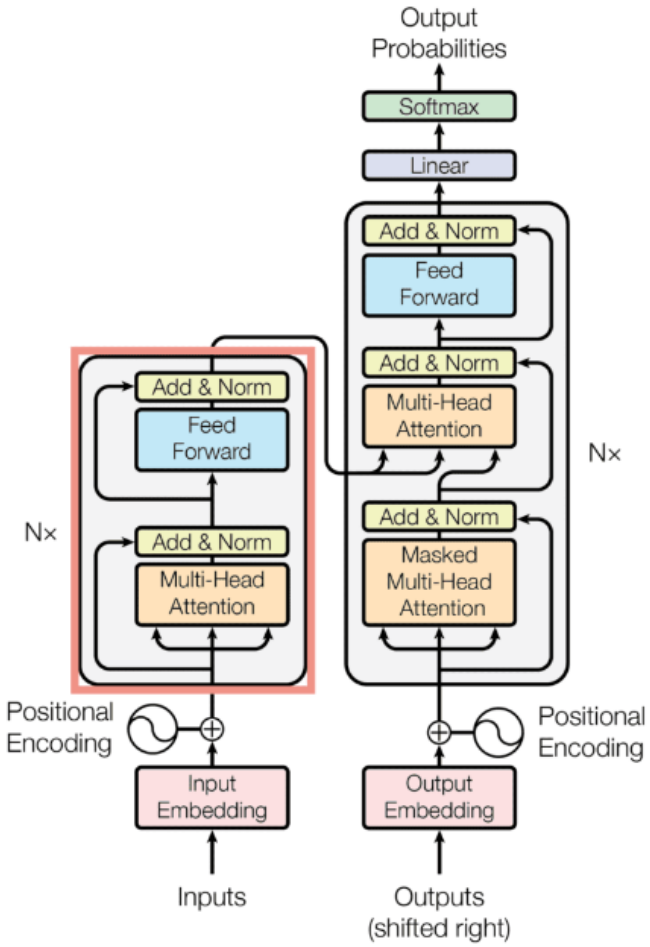
Κάθε μονάδα κωδικοποιητή (encoder) στην αρχιτεκτονική Transformer αποτελείται από δύο βασικά στοιχεία: έναν μηχανισμό αυτοπροσοχής και ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (feed forward neural network). Ο μηχανισμός αυτοπροσοχής αποτελεί θεμελιώδη πτυχή της διαδικασίας κωδικοποίησης, καθώς αξιολογεί τη συνάφεια και τις εξαρτήσεις μεταξύ των κωδικοποιήσεων εισόδου από τον προηγούμενο κωδικοποιητή, με αποτέλεσμα τη δημιουργία κωδικοποιήσεων εξόδου. Αυτός ο μηχανισμός αυτοπροσοχής αποδίδει βάρη στις κωδικοποιήσεις εισόδου με βάση τη σημασία και τις σχέσεις τους, επιτρέποντας στο μοντέλο να εστιάζει στις πιο σχετικές πληροφορίες για κάθε κωδικοποίηση εξόδου. Το νευρωνικό δίκτυο τροφοδότησης επεξεργάζεται περαιτέρω κάθε κωδικοποίηση εξόδου ξεχωριστά, εφαρμόζοντας μη γραμμικούς μετασχηματισμούς για να συλλάβει αναπαραστάσεις υψηλότερου επιπέδου και να εξάγει πιο σύνθετα χαρακτηριστικά. Στο μοντέλο

Transformer, η πρώτη μονάδα κωδικοποιητή χειρίζεται ειδικά την ακολουθία εισόδου ενσωματώνοντας τόσο πληροφορίες θέσης όσο και ενσωματώσεις. Η συμπερίληψη των κωδικοποιήσεων θέσης είναι ζωτικής σημασίας για τον Transformer ώστε να αξιοποιήσει τη διαδοχική σειρά της ακολουθίας. Σε αντίθεση με άλλα στοιχεία του μοντέλου που δεν λαμβάνουν ρητά υπόψη τη σειρά των στοιχείων, οι κωδικοποιήσεις θέσης επιτρέπουν στον Transformer να κατανοήσει τις σχετικές θέσεις των διαφόρων στοιχείων στην ακολουθία. Με την ενσωμάτωση αυτής της πληροφορίας θέσης, το μοντέλο αποκτά καλύτερη κατανόηση της διαδοχικής φύσης της εισόδου, επιτρέποντάς του να συλλάβει και να αξιοποιήσει τις εγγενείς εξαρτήσεις και τις χρονικές σχέσεις μεταξύ των στοιχείων [15, 25].

Αυτός ο συνδυασμός του μηχανισμού αυτοπροσοχής και των κωδικοποιήσεων θέσης εντός των μονάδων κωδικοποιητή δίνει στο μοντέλο Transformer την ικανότητα να επεξεργάζεται αποτελεσματικά διαδοχικά δεδομένα, όπως προτάσεις φυσικής γλώσσας ή δεδομένα χρονοσειρών. Ο μηχανισμός αυτοπροσοχής επιτρέπει στο μοντέλο να παρακολουθεί δυναμικά διάφορα μέρη της ακολουθίας εισόδου, καταγράφοντας τόσο τοπικές όσο και παγκόσμιες εξαρτήσεις, ενώ οι κωδικοποιήσεις θέσης παρέχουν το απαραίτητο πλαίσιο για την κατανόηση της σειράς των στοιχείων. Αξίζει να σημειωθεί ότι ο κωδικοποιητής στην αρχιτεκτονική Transformer λειτουργεί αμφίδρομα, πράγμα που σημαίνει ότι κατά την επεξεργασία κάθε συμβόλου λαμβάνει υπόψη τόσο τα προηγούμενα όσο και τα επόμενα συμβολίσματα. Αυτός ο αμφίδρομος χαρακτήρας επιτρέπει στον κωδικοποιητή να καταγράφει μια ολιστική άποψη του πλαισίου που περιβάλλει κάθε λέξη, λαμβάνοντας υπόψη τα σημεία που προηγούνται και έπονται αυτής. Με την ενσωμάτωση πληροφοριών και από τις δύο κατευθύνσεις, ο κωδικοποιητής μπορεί να κατανοήσει αποτελεσματικά τις εξαρτήσεις και τις σχέσεις εντός της ακολουθίας. Αυτή η προσέγγιση είναι ιδιαίτερα επωφελής όταν χρησιμοποιούνται tokens αντί για μεμονωμένες λέξεις, καθώς λαμβάνει υπόψη το φαινόμενο της πολυσημίας, όπου μια λέξη μπορεί να έχει πολλαπλές σημασίες ή ερμηνείες. Αναλυτικότερα, οι ενότητες κωδικοποιητή στην αρχιτεκτονική Transformer αποτελούνται από δύο βασικά στοιχεία: μηχανισμούς αυτοπροσοχής και νευρωνικά δίκτυα τροφοδότησης προς τα εμπρός. Ο μηχανισμός αυτοπροσοχής αξιολογεί τη συνάφεια και τη σημασία των κωδικοποιήσεων εισόδου, επιτρέποντας στο μοντέλο να εστιάζει στα πιο κατατοπιστικά στοιχεία εντός της ακολουθίας. Στη συνέχεια, το νευρωνικό δίκτυο προώθησης επεξεργάζεται τις προκύπτουσες κωδικοποιήσεις εξόδου, εφαρμόζοντας περαιτέρω μετασχηματισμούς για τη σύλληψη αναπαραστάσεων υψηλότερου επιπέδου και την εξαγωγή σημαντικών χαρακτηριστικών [15, 24, 25].

Στη συγκεκριμένη περίπτωση της πρώτης μονάδας κωδικοποιητή, ενσωματώνει τόσο πληροφορίες θέσης όσο και ενσωματώσεις για να βελτιώσει την κατανόηση της ακολουθίας εισόδου. Με την ενσωμάτωση πληροφοριών θέσης, το μοντέλο αποκτά επίγνωση της σειράς και της ακολουθίας των tokens, επιτρέποντάς του να διακρίνει τις σχετικές θέσεις των διαφόρων στοιχείων. Αυτή η εξέταση της διαδοχικής φύσης της εισόδου είναι ζωτικής σημασίας για την ακριβή αποτύπωση του νοήματος και του πλαισίου της ακολουθίας. Συνολικά, ο αμφίδρομος κωδικοποιητής, μαζί με τους μηχανισμούς αυτοπροσοχής και τις κωδικοποιήσεις θέσης, δίνει τη δυνατότητα στο μοντέλο Transformer να αναλύει και να κατανοεί πλήρως το πλαίσιο της ακολουθίας εισόδου. Προσέχοντας τόσο τα προηγούμενα όσο και τα επόμενα tokens, το μοντέλο καταγράφει μια ολοκληρωμένη κατανόηση του περιβάλλοντος πλαισίου, επιτρέποντάς του να επεξεργάζεται αποτελεσματικά και να παράγει ουσιαστικές αναπαραστάσεις των δεδομένων εισόδου [15, 24].

Ο κωδικοποιητής σηματοδοτείται με κόκκινο πλαίσιο στην παρακάτω Εικόνα 12.



Εικόνα 12. Ο κωδικοποιητής (encoder) ενός Transformer. [25]

Αποκωδικοποιητής (Decoder)

Κάθε module αποκωδικοποιητή στην αρχιτεκτονική Transformer αποτελείται από τρία βασικά στοιχεία: αυτοπροσοχή, προσοχή κωδικοποιητή-αποκωδικοποιητή και ένα νευρωνικό δίκτυο τροφοδότησης προς τα εμπρός (feedforward neural network). Παρόμοια με τον κωδικοποιητή, ο αποκωδικοποιητής χρησιμοποιεί την αυτοπροσοχή για να συλλάβει τις εξαρτήσεις και τις σχέσεις μεταξύ των tokens μέσα στην ακολουθία εξόδου που παράγει. Αυτός ο μηχανισμός αυτοπροσοχής επιτρέπει στον αποκωδικοποιητή να λαμβάνει υπόψη του τα δικά του σήματα εξόδου που έχουν προηγουμένως παραχθεί και να σταθμίζει τη σημασία τους κατά τη δημιουργία του επόμενου συμβόλου. Εξετάζοντας τις εξαρτήσεις εντός της δικής του εξόδου, ο αποκωδικοποιητής μπορεί να διασφαλίσει τη συνοχή και τη συνέπεια στην παραγόμενη ακολουθία. Εκτός από την αυτοπροσοχή, ο αποκωδικοποιητής ενσωματώνει έναν μηχανισμό προσοχής ειδικά σχεδιασμένο για να παρακολουθεί τις κωδικοποιήσεις που παράγονται από τις μονάδες κωδικοποιητή. Αυτός ο μηχανισμός προσοχής κωδικοποιητή-αποκωδικοποιητή (encoder – decoder) επιτρέπει στον αποκωδικοποιητή να συλλέγει σχετικές πληροφορίες από τις κωδικοποιήσεις, ευθυγραμμίζοντας την κατανόησή του με το πλαίσιο (context) που παρέχει ο κωδικοποιητής, συνεισφέροντας έτσι σε αναπαραστάσεις που ακολουθούν το πλαίσιο (contextualized representations). Με τη χρήση μηχανισμού προσοχής στις κωδικοποιήσεις του κωδικοποιητή, ο αποκωδικοποιητής μπορεί να έχει πρόσβαση σε αναπαραστάσεις υψηλότερου επιπέδου της ακολουθίας εισόδου, επιτρέποντάς του να λαμβάνει τεκμηριωμένες αποφάσεις κατά τη διαδικασία αποκωδικοποίησης [15, 24, 25].

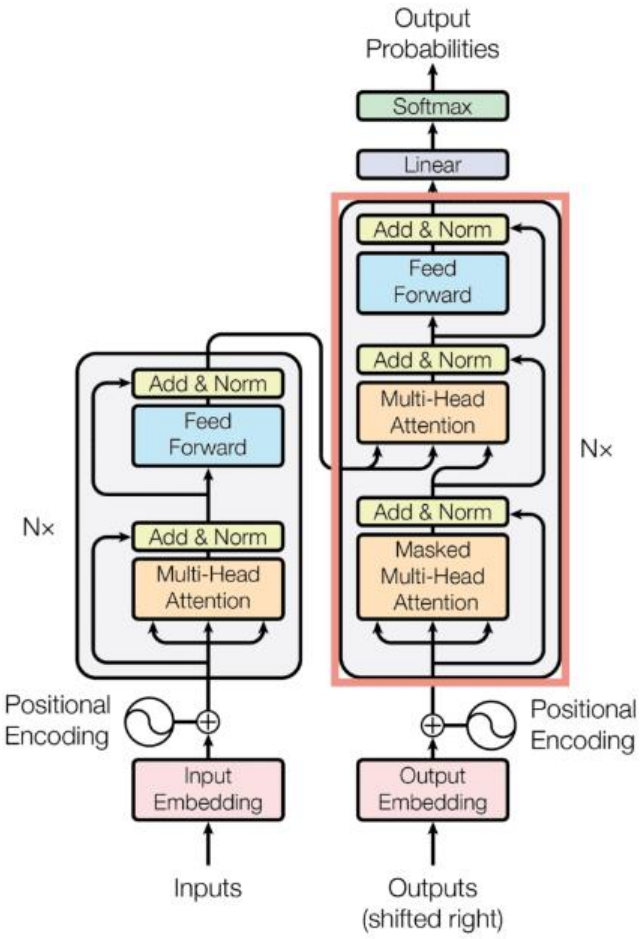
Συνδυάζοντας την αυτοπροσοχή και την προσοχή του κωδικοποιητή-αποκωδικοποιητή, ο αποκωδικοποιητής στην αρχιτεκτονική Transformer μπορεί να παράγει αποτελεσματικά ακριβείς και σχετικές με τα συμφραζόμενα ακολουθίες εξόδου. Ο μηχανισμός αυτοπροσοχής επιτρέπει στον αποκωδικοποιητή να καταγράφει τις εξαρτήσεις εντός της ακολουθίας, ενώ ο μηχανισμός προσοχής κωδικοποιητή-αποκωδικοποιητή διευκολύνει την ενσωμάτωση των πληροφοριών από τις κωδικοποιήσεις του κωδικοποιητή, εξασφαλίζοντας μια ολοκληρωμένη κατανόηση της ακολουθίας εισόδου. Επιπλέον, η μονάδα αποκωδικοποιητή περιλαμβάνει επίσης ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης που επεξεργάζεται περαιτέρω την έξοδο των μηχανισμών προσοχής. Αυτό το δίκτυο εφαρμόζει μετασχηματισμούς και μη γραμμικότητες (για παράδειγμα μέσω συναρτήσεων ενεργοποίησης - activation functions όπως αυτές που περιγράφηκαν στην ενότητα των CNNs) στις αναπαραστάσεις του αποκωδικοποιητή, ενισχύοντας την ικανότητά του να παράγει ακολουθίες εξόδου υψηλής ποιότητας [15, 24, 25].

Παρόμοια με την αρχική μονάδα κωδικοποιητή, η πρώτη μονάδα αποκωδικοποιητή της αρχιτεκτονικής Transformer λαμβάνει επίσης υπόψη τις πληροφορίες θέσης (positional information) και τις ενσωματώσεις της ακολουθίας εξόδου (output sequence embeddings), αντί να βασίζεται σε προ-υπολογισμένες κωδικοποιήσεις. Με την ενσωμάτωση πληροφοριών θέσης, ο αποκωδικοποιητής είναι σε θέση να συλλάβει τη διαδοχική σειρά της ακολουθίας εξόδου, η οποία είναι ζωτικής σημασίας για τη δημιουργία συνεκτικού και ουσιαστικού κειμένου. Αυτό εξασφαλίζει ότι ο αποκωδικοποιητής κατανοεί το πλαίσιο και τη σχέση μεταξύ των διαφόρων λέξεων στην παραγόμενη ακολουθία. Για να διατηρηθεί ο αυτοπαλίνδρομος (autoregressive) χαρακτήρας της διαδικασίας παραγωγής κειμένου, δηλαδή η πρόβλεψη μελλοντικών τιμών βασισμένων σε παρελθοντικές τιμές, η ακολουθία εξόδου καλύπτεται εν μέρει. Αυτό σημαίνει ότι ο αποκωδικοποιητής δεν έχει πρόσβαση σε πληροφορίες σχετικά με μελλοντικές μάρκες κατά τη διάρκεια της διαδικασίας παραγωγής. Αυτή η απόκρυψη εξασφαλίζει ότι το μοντέλο προβλέπει κάθε λέξη με βάση αποκλειστικά τις προηγούμενες λέξεις, διατηρώντας την αιτιότητα και αποτρέποντας τη διαρροή πληροφοριών από μελλοντικές μάρκες. Ακολουθώντας αυτή την προσέγγιση, ο αποκωδικοποιητής παρουσιάζει αυτοπαλινδρόμηση,

παράγοντας κείμενο που συνάδει με την αυτοπαλινδρόμηση του μοντέλου. Μέσα στον αποκωδικοποιητή, οι κεφαλές προσοχής περιορίζονται από το να προσέχουν τα σημεία που ακολουθούν το τρέχον σημείο. Αυτός ο περιορισμός είναι απαραίτητος για την επιβολή της αυτοπαλινδρομικής διαδικασίας παραγωγής, όπου κάθε λέξη παράγεται μόνο με βάση τις προηγούμενες λέξεις. Αποτρέποντας την προσοχή σε μελλοντικά tokens, ο αποκωδικοποιητής παραμένει ευθυγραμμισμένος με το αυτοπαλινδρομικό πλαίσιο και αποφεύγει τη διαρροή πληροφοριών από tokens που δεν έχουν ακόμη παραχθεί (βρίσκονται αργότερα στην ακολουθία). Τέλος, η τελευταία ενότητα αποκωδικοποιητή ακολουθείται από ένα στρώμα γραμμικού μετασχηματισμού και ένα στρώμα softmax. Ο γραμμικός μετασχηματισμός εφαρμόζει μια γραμμική απεικόνιση στις αναπαραστάσεις εξόδου του αποκωδικοποιητή, ενώ το στρώμα softmax μετατρέπει αυτές τις αναπαραστάσεις σε πιθανότητες πάνω στο λεξιλόγιο. Αυτό επιτρέπει στο μοντέλο να παράγει το πιο πιθανό επόμενο token για την ακολουθία εξόδου αποδίδοντας υψηλότερες πιθανότητες σε λέξεις που είναι πιο πιθανές στο συγκεκριμένο πλαίσιο [15, 24, 25].

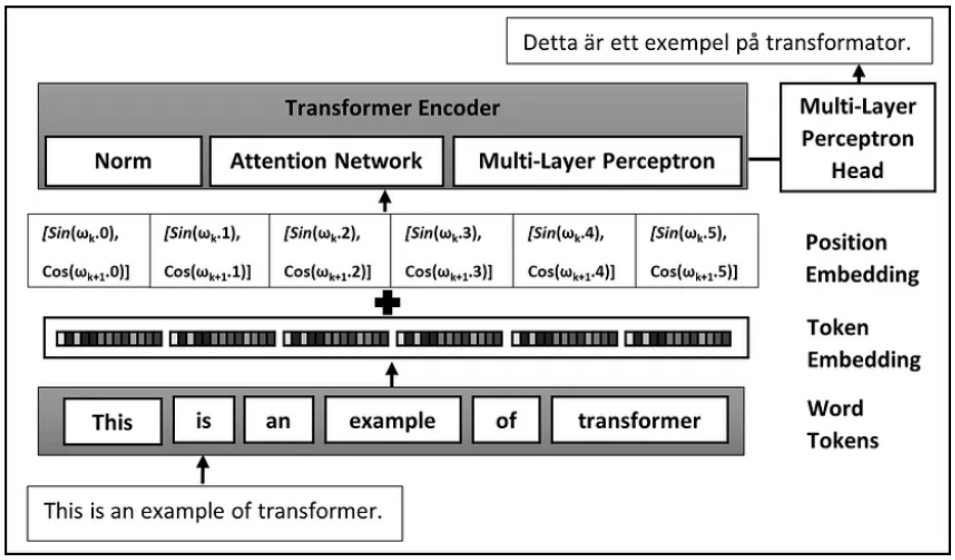
Συνοπτικά, οι μονάδες αποκωδικοποιητή στην αρχιτεκτονική Transformer διασφαλίζουν ότι η παραγόμενη ακολουθία εξόδου ακολουθεί μια διαδοχική σειρά και διατηρεί την αυτοπαλινδρομική φύση της διαδικασίας παραγωγής κειμένου. Με την ενσωμάτωση πληροφοριών θέσης, την απόκρυψη μελλοντικών σημείων και τον περιορισμό της προσοχής σε προηγούμενα σημεία, ο αποκωδικοποιητής παράγει συνεκτικό και σχετικό με το πλαίσιο κείμενο.

Ο αποκωδικοποιητής σηματοδοτείται με κόκκινο πλαίσιο στην παρακάτω Εικόνα 13.



Εικόνα 13. Ο αποκωδικοποιητής (decoder) ενός Transformer. [25]

Ένα διαισθητικό παράδειγμα χρήσης Transformer σε λειτουργίες φυσικής γλώσσας, και συγκεκριμένα μηχανικής μετάφρασης (machine translation) απεικονίζεται στην ακόλουθη Εικόνα 14. Η διαδικασία της πραγματοποίησης της μηχανικής μετάφρασης είναι η ακόλουθη: Αρχικά, το κείμενο διαιρείται σε μια συλλογή λέξεων που είναι γνωστές ως tokens. Φαντάζοντας κάθε token ως ένα σημασιολογικά επαρκές κομμάτι μιας λέξης. Κάθε token υφίσταται μια διαδικασία μετατροπής για να αναπαρασταθεί σε ένα κωδικοποιημένο ή ενσωματωμένο διάνυσμα (embedding), με τεχνικές όπως το word2vec, η οποία αντιστοιχίζει κάθε λέξη σε ένα σημείο του διανυσματικού χώρου. Η θέση κάθε λέξης στην ακολουθία αναπαρίσταται μέσω της ενσωμάτωσης θέσης (ositional encoding), η οποία συνδυάζεται με την ενσωμάτωση λέξης για την αναπαράσταση πληροφοριών θέσης. Αυτές οι ενσωματώσεις τροφοδοτούνται στη συνέχεια στον κωδικοποιητή Transformer. Ο κωδικοποιητής περιέχει ένα πολυεπίπεδο δίκτυο αυτοπροσοχής (MSP) που αποδίδει βάρη στα tokens με βάση τη σχετική τους σημασία εντός της πρότασης, αποτυπώνοντας έτσι τις σχέσεις πλαισίου (contextual relationships). Μετά το MSP, ένα δίκτυο πολλαπλών επιπέδων Perceptron (MLP) χρησιμοποιείται για την επεξεργασία της εξόδου από το δίκτυο προσοχής. Ο κωδικοποιητής, ο οποίος όπως περιγράψαμε παραπάνω αποτελείται από πολλαπλά μπλοκ MSP και MLP, μαζί με στρώματα κανονικοποίησης, λαμβάνει την πληροφορία από το MLP. Επιπλέον, ένα MLP-Head Layer βρίσκεται εκτός του δικτύου κωδικοποιητή, το οποίο παράγει τα logits. Τα logits μπορούν να μετατραπούν περαιτέρω σε πιθανότητες με την εφαρμογή ενός στρώματος ενεργοποίησης, όπως το softmax. Τέλος, η softmax αποδίδει την έξοδο με τη μεγαλύτερη πιθανότητα, η οποία στην περίπτωση της μηχανικής μετάφρασης αντιστοιχεί στην πιο πιθανή επόμενη λέξη της ακολουθίας εξόδου. [41]



Εικόνα 14. Παράδειγμα μηχανικής μετάφρασης με χρήση αρχιτεκτονικής Transformer [41].

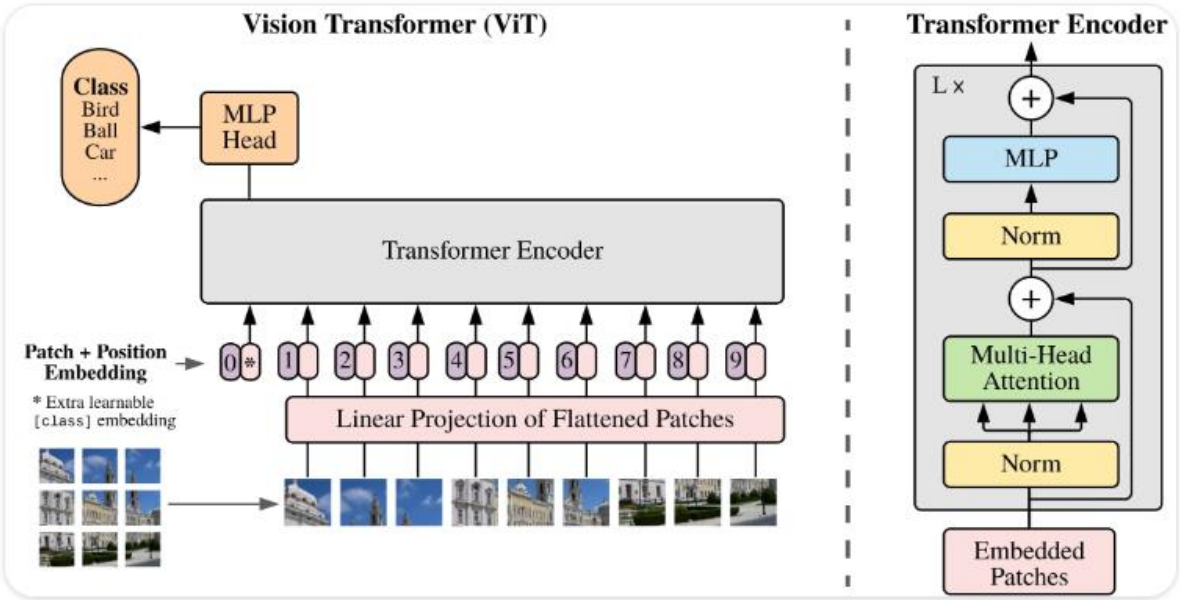
Σε μια παρόμοια λογική κινούνται και οι Vision Transformers που θα περιγραφούν στη συνέχεια. Στους Vision Transformers αντί να δίνονται tokens λέξεων στην είσοδο. Θα δίνονται τμήματα εικόνων, τα οποία συχνά αποτελούνται από patches της εικόνας εισόδου.

2.3 Οπτικοί Transformers (Vision Transformers)

Αρκετές εμπειρικές μελέτες έχουν καταδείξει την αποτελεσματικότητα της προ-εκπαίδευσης στους Vision Transformers. Οι Dosovitskiy et al. [16] παρουσίασαν την αρχιτεκτονική ViT και παρουσίασαν την ισχυρή απόδοσή της σε διάφορα σημεία αναφοράς ταξινόμησης εικόνων. Ακολουθώντας τις βασικές αρχές σχεδιασμού της αρχικής Transformer αρχιτεκτονικής, οι οπτικοί Transformers (Vision Transformers - ViTs) ακολουθούν δύο στάδια εκπαίδευσης. Το πρώτο είναι η προεκπαίδευση (pre – training) η οποία πραγματοποιείται μέσω μιας διαδικασίας αυτοεπιβλεπόμενης μάθησης (self – supervised learning). Σύμφωνα με αυτό το σχήμα μάθησης, ένα μοντέλο καλείται να μάθει συσχετίσεις πραγματοποιώντας κάποιες ενέργειες πάνω στα δεδομένα εισόδου, χωρίς να λαμβάνει κάποιο επιπλέον σήμα για τη μάθηση, όπως θα ήταν για παράδειγμα η χρήση ετικετών (κάτι το οποίο ακολουθείται στο σχήμα της επιβλεπόμενης μάθησης). Κατά την προεκπαίδευση λοιπόν ένα μοντέλο εκπαιδεύεται πάνω σε μεγάλο όγκο μη επισημειωμένων δεδομένων. Η πληθώρα δεδομένων είναι αναγκαία προκειμένου το μοντέλο να μάθει γενικές συσχετίσεις της εικόνας (ή της γλώσσας στην περίπτωση της παραδοσιακής Transformer αρχιτεκτονικής για NLP), αφού θα ‘δει’ πολλές φορές τα υπάρχοντα μοτίβα, δημιουργώντας έτσι συσχετίσεις αυτών [16, 26].

Το δεύτερο στάδιο εκπαίδευσης είναι το fine – tuning, όπου αξιοποιούνται μικρότερα σύνολα δεδομένων, τα οποία είναι επισημειωμένα με ετικέτες. Στη φάση αυτή, το μέγεθος του dataset επιτρέπει μια τέτοια πολυτέλεια, καθώς το μοντέλο έχει ήδη αποκτήσει μια γνώση του κόσμου (οπτική ή λεκτική, ανάλογα το πεδίο εφαρμογής και τα αντίστοιχα δεδομένα εκπαίδευσης), και το μόνο που απομένει είναι να εφαρμόσει τη γνώση αυτή πάνω σε μια συγκεκριμένη εργασία. Στην περίπτωση της όρασης υπολογιστών, μια σχετική εργασία θα ήταν η ταξινόμηση εικόνας [16, 26].

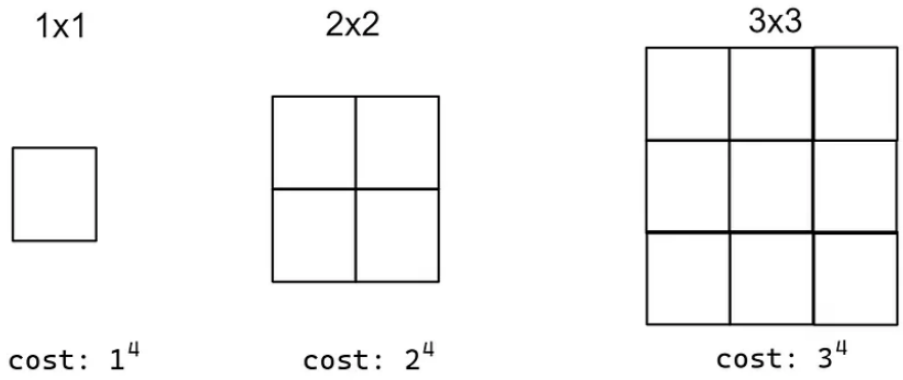
Οι ViT χαρακτηρίζονται από την αρχιτεκτονική που παρουσιάζεται στην Εικόνα 15, ως παραλλαγή της κλασικής αρχιτεκτονικής Transformer, όπως αυτή παρουσιάστηκε αρχικά στο “Attention is all you need”.



Εικόνα 15. Βασική αρχιτεκτονική ViT για ταξινόμηση εικόνων. [30]

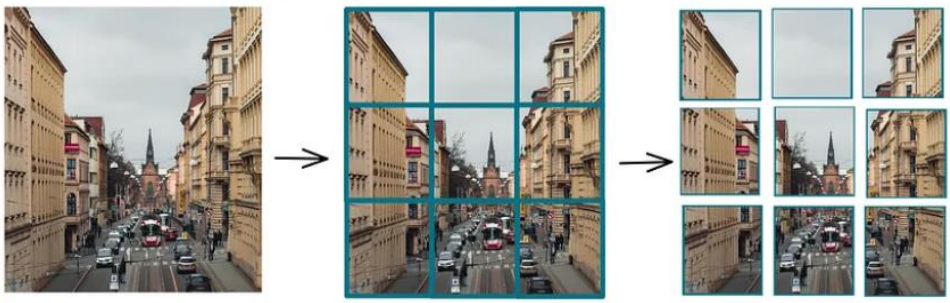
Αν και η αρχή λειτουργίας των Vision Transformers εμπνέεται και βασίζεται στους γλωσσικούς Transformers, παρόλα αυτά η αυτοπροσοχή αποτελεί πρόκληση λόγω της υπολογιστικής πολυπλοκότητάς της. Αυτό οφείλεται στο γεγονός ότι απαιτεί από κάθε token να παρακολουθεί και να συγκρίνει κάθε άλλο token μέσα σε μια ακολουθία. Η εφαρμογή του μηχανισμού αυτοπροσοχής σε δεδομένα εικόνας εισάγει ένα ακόμη εμπόδιο. Σε αυτό το σενάριο, κάθε pixel σε μια εικόνα θα πρέπει να παρακολουθεί και να συγκρίνεται με κάθε άλλο pixel. Ωστόσο, το πρόβλημα προκύπτει όταν εξεταστεί η εκθετική αύξηση του υπολογιστικού κόστους εάν η ανάλυση της εικόνας είναι σημαντικά μεγάλη. Ακόμη και μια μικρή αύξηση του αριθμού των pixels θα είχε ως αποτέλεσμα μια τετραγωνική αύξηση των υπολογιστικών απαιτήσεων, καθιστώντας το μη πρακτικό για την επεξεργασία εικόνων με λογικά υψηλή ανάλυση. [42]

Ένα παράδειγμα που αναδεικνύει το εν λόγω πρόβλημα παρουσιάζεται στην Εικόνα 16:



Εικόνα 16. Υπολογιστική πολυπλοκότητα της αυτοπροσοχής στους Vision Transformer. Εικόνες μεγαλύτερης ανάλυσης επιβαρύνουν σημαντικά το υπολογιστικό κόστος [42].

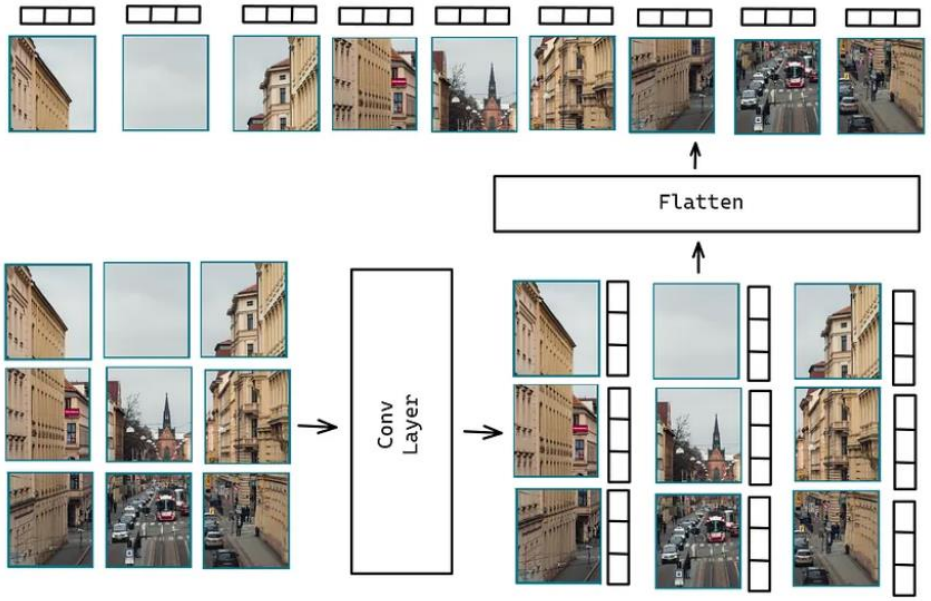
Για την αντιμετώπιση αυτής της πρόκλησης, η αρχιτεκτονική Vision Transformer (ViT) παρουσιάζει μια λύση με την εισαγωγή μιας προσέγγισης βασισμένης σε επιθέματα. Η εικόνα εισόδου διαιρείται σε μικρότερα patches, καθένα από τα οποία αποτελείται από ένα σταθερό μέγεθος, όπως 16 x 16 pixels. Για παράδειγμα, μια εικόνα με διαστάσεις 48 x 48 pixels , θα διαιρεθεί σε patches ως εξής (Εικόνα 17): [16, 42]



Εικόνα 17. Διαίρεση της εικόνας σε patches για μείωση της υπολογιστικής πολυπλοκότητας που εισάγει ο μηχανισμός αυτοπροσοχής [42].

Αυτή η αναπαράσταση με βάση τα patches επιτρέπει στο μοντέλο ViT να επεξεργάζεται αποτελεσματικά τις εικόνες, λειτουργώντας σε μικρότερες και πιο εύχρηστες μονάδες. Με την αποσύνθεση της εικόνας σε patches, η υπολογιστική πολυπλοκότητα μειώνεται σημαντικά σε σύγκριση με την επεξεργασία σε επίπεδο pixel. Επιπλέον, τα patches διατηρούν χωρικές πληροφορίες, επιτρέποντας στο μοντέλο να συλλαμβάνει τοπικά μοτίβα και σχέσεις εντός της εικόνας. Αυτή η προσέγγιση με βάση τα patches αποτελεί το θεμέλιο της αρχιτεκτονικής ViT, επιτρέποντας την αποτελεσματική κατανόηση εικόνων και εργασίες ταξινόμησης [42].

Μετά τον διαχωρισμό της εικόνας εισόδου σε patches, το επόμενο βήμα στο μοντέλο ViT περιλαμβάνει την επιπεδοποίηση (flattening) των patches και την τακτοποίησή τους με διαδοχικό τρόπο. Αυτή η διαδικασία επιπεδοποίησης μετατρέπει τη δισδιάστατη δομή των επιφανειών σε μονοδιάστατη ακολουθία, όπως απεικονίζεται στην παρακάτω Εικόνα 18.



Εικόνα 18. Μετατροπή των patches εισόδου σε μονοδιάστατη αναπαράσταση μέσω της διαδικασίας επιπεδοποίησης (flattening). [42]

Με την επιπεδοποίηση των patches και τη διαδοχική τους διάταξη, το μοντέλο ViT μετατρέπει τα δεδομένα της εικόνας σε μορφή που μπορεί να επεξεργαστεί αποτελεσματικά από τα επόμενα στοιχεία της αρχιτεκτονικής. Αυτή η διαδοχική αναπαράσταση επιτρέπει στο μοντέλο να καταγράφει τις χωρικές σχέσεις και εξαρτήσεις μεταξύ των patches με συνεκτικό τρόπο. Κάθε patch στην ακολουθία διατηρεί τις πληροφορίες θέσης του, οι οποίες είναι ζωτικής σημασίας για την κατανόηση του χωρικού πλαισίου από το μοντέλο και τη διατήρηση της εγγενούς δομής της εικόνας. Αυτή η διαδοχική διάταξη των πεπλατυσμένων patches χρησιμεύει ως είσοδος στα επόμενα στρώματα του μοντέλου ViT, επιτρέποντάς του να εκτελεί ισχυρές εργασίες οπτικής αναπαράστασης και ανάλυσης [42].

2.3.1 Pre – training στους οπτικούς Transformers

Η διαδικασία προ-εκπαίδευσης στους Vision Transformers (ViT) περιλαμβάνει διάφορα κρίσιμα στοιχεία και στρατηγικές. Αρχικά, χρησιμοποιούνται σύνολα δεδομένων εικόνων μεγάλης κλίμακας χωρίς ετικέτες για να εκτεθεί το μοντέλο σε ένα ευρύ φάσμα οπτικών μοτίβων. Η αρχιτεκτονική του Transformer, με τον μηχανισμό αυτοπροσοχής και το νευρωνικό δίκτυο τροφοδότησης προς τα εμπρός, χρησιμοποιείται για τη σύλληψη των χωρικών σχέσεων και των ιεραρχικών αναπαραστάσεων εντός των εικόνων. Διάφοροι στόχοι προ-εκπαίδευσης, όπως η αφαίρεση χρωμάτων εικόνας, η πρόβλεψη περιστροφής και η αντιθετική μάθηση, καθοδηγούν τη διαδικασία μάθησης. Οι προσεγγίσεις αυτοεπιβλεπόμενης μάθησης έχουν αποδειχθεί αποτελεσματικές στην εκπαίδευση των ViT για την εκμάθηση χρήσιμων οπτικών αναπαραστάσεων χωρίς την ανάγκη χειροκίνητων σχολίων. Για παράδειγμα, οι ViTs μπορούν να εκπαιδευτούν για να προβλέπουν τις σχετικές θέσεις των patches εικόνας ή να λύνουν παζλ που αποτελούνται από ανακατεμένα patches εικόνας. Η αυτοεπιβλεπόμενη μάθηση ενθαρρύνει τους ViTs να καταγράφουν πλούσιες χωρικές σχέσεις και εξαρτήσεις από το πλαίσιο εντός των εικόνων, βελτιώνοντας την ικανότητά τους να κατανοούν σύνθετες οπτικές σκηνές [16, 26].

Η προ-εκπαίδευση προσφέρει σημαντικά οφέλη στους Vision Transformers, ενισχύοντας τις ικανότητές τους σε πολλαπλές πτυχές της κατανόησης εικόνας. Μία από αυτές είναι η βελτιωμένη εξαγωγή χαρακτηριστικών: Η προ-εκπαίδευση επιτρέπει στους ViTs να εξάγουν πλούσια και ουσιαστικά οπτικά χαρακτηριστικά από εικόνες, επιτρέποντάς τους να συλλαμβάνουν σύνθετα μοτίβα και σημασιολογία. Επιπλέον, μέσω της προ-εκπαίδευσης επιτυγχάνεται βελτιωμένη μεταφορά μάθησης (transfer learning). Πιο συγκεκριμένα, η προ-εκπαίδευση διευκολύνει τη μάθηση μεταφοράς, επιτρέποντας στους ViTs να αξιοποιούν τη γνώση από τα προ-εκπαιδευμένα μοντέλα για να βελτιώνουν την απόδοση σε επόμενες εργασίες. Ένα ακόμη πλεονέκτημα που προσφέρει η προεκπαίδευση είναι η γενίκευση και ανθεκτικότητα των ViT. Αυτό οφείλεται στο γεγονός ότι οι ViTs που εκπαιδεύονται με προ-εκπαίδευση παρουσιάζουν βελτιωμένες δυνατότητες γενίκευσης, επιτρέποντάς τους να αποδίδουν καλά σε διαφορετικά σύνολα δεδομένων και τομείς. Επιπλέον, η προεκπαίδευση προσφέρει προσαρμογή σε περιορισμένα δεδομένα, εφόσον επιτρέπει την αποτελεσματική μάθηση ακόμη και σε σενάρια με περιορισμένα επισημασμένα δεδομένα, καθώς τα μοντέλα έχουν ήδη μάθει πολύτιμες οπτικές αναπαραστάσεις από μη επισημασμένα δεδομένα. Ενώ η προ-εκπαίδευση έχει επιδείξει αξιοσημείωτη επιτυχία στους Vision Transformers, πρέπει να αντιμετωπιστούν διάφορες προκλήσεις και μελλοντικές ερευνητικές κατευθύνσεις. Αυτές περιλαμβάνουν τη διερεύνηση βελτιωμένων στόχων προ-εκπαίδευσης (pre – training objectives), το σχεδιασμό αποτελεσματικών τεχνικών επαύξησης δεδομένων (data augumentation) και τη βελτιστοποίηση της αρχιτεκτονικής (model optimization) του μοντέλου για καλύτερες επιδόσεις σε ποικίλες εργασίες [16, 26].

Στόχοι προεκπαίδευσης

Οι στόχοι της προ-εκπαίδευσης παίζουν καθοριστικό ρόλο στη διαμόρφωση της διαδικασίας μάθησης των ViTs και στην καθοδήγησή τους προς την καταγραφή ουσιαστικών οπτικών αναπαραστάσεων. Έχουν διερευνηθεί διάφοροι στόχοι προ-εκπαίδευσης για την ενίσχυση της αποτελεσματικότητας των ViTs σε εργασίες κατανόησης εικόνων. Όλες αυτές οι τεχνικές που θα αναπτυχθούν στη συνέχεια ακολουθούν το μοντέλο της αυτοεπιβλεπόμενης μάθησης. Συνδυάζοντας αυτούς τους στόχους προ-εκπαίδευσης ή εξερευνώντας νέους στόχους, οι ερευνητές συνεχίζουν να διευρύνουν τα όρια των δυνατοτήτων μάθησης αναπαράστασης των ViTs. Η επιλογή ενός κατάλληλου στόχου προ-εκπαίδευσης εξαρτάται από τη συγκεκριμένη εργασία και το σύνολο δεδομένων και οι εμπειρικές μελέτες είναι απαραίτητες για την αξιολόγηση της αποτελεσματικότητάς τους στη βελτίωση της απόδοσης των ViTs.

Συμπλήρωση τμημάτων εικόνας (image inpainting): Η συμπλήρωση τμημάτων εικόνας περιλαμβάνει την πρόβλεψη των τμημάτων μιας εικόνας που λείπουν ή που έχουν αποκλειστεί. Τα ViT μπορούν να εκπαιδευτούν για να συμπληρώνουν τις περιοχές που λείπουν με βάση το πλαίσιο που παρέχουν τα γύρω pixels. Αξιοποιώντας την αφαίρεση εικόνας ως στόχο προ-εκπαίδευσης, οι ViTs μαθαίνουν να κατανοούν τις χωρικές σχέσεις και τις πληροφορίες πλαισίου εντός των εικόνων, επιτρέποντάς τους να συλλαμβάνουν λεπτομερείς λεπτομέρειες και να δημιουργούν συνεκτικές οπτικές αναπαραστάσεις. [16, 20]

Προβλέψεις περιστροφής (rotation prediction): Ένας άλλος ευρέως υιοθετημένος στόχος προ-εκπαίδευσης είναι η πρόβλεψη περιστροφής. Οι ViTs εκπαιδεύονται για να προβλέπουν τη γωνία περιστροφής που εφαρμόζεται σε μια εικόνα. Αναγκάζοντας το μοντέλο να σκεφτεί για διαφορετικούς προσανατολισμούς, η πρόβλεψη περιστροφής προωθεί την ισχυρή μάθηση χαρακτηριστικών και ενθαρρύνει τα ViTs να κωδικοποιήσουν γεωμετρικούς μετασχηματισμούς στις αναπαραστάσεις τους. Αυτός ο στόχος προ-εκπαίδευσης βοηθά τα ViTs να συλλάβουν την αναλλοίωτη οπτική γωνία και βελτιώνει την ικανότητά τους να χειρίζονται μεταβολές στους προσανατολισμούς των αντικειμένων [16, 26].

Αντιθετική μάθηση (contrastive learning): Η αντιθετική μάθηση είναι μια ισχυρή προσέγγιση προ-εκπαίδευσης που στοχεύει στην εκμάθηση σημασιολογικά σημαντικών αναπαραστάσεων μεγιστοποιώντας τη συμφωνία μεταξύ παρόμοιων ζευγών και ελαχιστοποιώντας την για ανόμοια ζεύγη. Στο πλαίσιο των ViTs, η αντιθετική μάθηση περιλαμβάνει την εκπαίδευση του μοντέλου για τη διάκριση μεταξύ θετικών (παρόμοιων) και αρνητικών (ανόμοιων) ζευγών τμημάτων εικόνας ή επαυξημένων όψεων της ίδιας εικόνας. Με την αντιπαραβολή παρόμοιων και ανόμοιων αναπαραστάσεων, τα ViTs μπορούν να μάθουν να συλλαμβάνουν υψηλού επιπέδου οπτική σημασιολογία και να διαχωρίζουν τους υποκείμενους παράγοντες της διακύμανσης [16, 26].

Γενετική μοντελοποίηση (generative modelling): Η προ-εκπαίδευση των ViTs με τη χρήση τεχνικών γεννητικής μοντελοποίησης, όπως η αυτοπαλινδρομική μοντελοποίηση ή τα γεννητικά αντιφατικά δίκτυα, έχει επίσης δείξει ότι είναι πολλά υποσχόμενη. Οι εργασίες δημιουργικής μοντελοποίησης, όπως η πρόβλεψη των pixelsπου λείπουν ή η παραγωγή ρεαλιστικών δειγμάτων εικόνας, ενθαρρύνουν τα ViTs να μάθουν την υποκείμενη κατανομή των οπτικών δεδομένων και να καταγράψουν λεπτομερή δομή της εικόνας. Αυτός ο στόχος προεκπαίδευσης ενισχύει την ικανότητα των ViTs να παράγουν οπτικά συνεκτικές και ποικίλες οπτικές αναπαραστάσεις [16, 20, 26].

Επαύξηση δεδομένων προεκπαίδευσης

Η επαύξηση των δεδομένων είναι μια κρίσιμη τεχνική στην εκπαίδευση των Transformers όρασης (ViTs) για τη βελτίωση της ευρωστίας, της γενίκευσης και της ικανότητάς τους να χειρίζονται διακυμάνσεις στα οπτικά δεδομένα. Η επαύξηση δεδομένων περιλαμβάνει την εφαρμογή ποικίλων μετασχηματισμών ή τροποποιήσεων στις εικόνες εισόδου για τη δημιουργία πρόσθετων δειγμάτων εκπαίδευσης. Η επιλογή και ο συνδυασμός των τεχνικών επαύξησης των δεδομένων στις ViT εξαρτώνται από τα χαρακτηριστικά του συνόλου δεδομένων και του συγκεκριμένου έργου. Είναι ζωτικής σημασίας να βρεθεί μια ισορροπία μεταξύ της εισαγωγής παραλλαγών που ενισχύουν τη γενίκευση του μοντέλου και της αποφυγής υπερβολικών παραμορφώσεων που μπορεί να εισάγουν μη ρεαλιστικά πρότυπα. Απαιτείται προσεκτικός πειραματισμός και επικύρωση για τον προσδιορισμό του πιο αποτελεσματικού συνόλου τεχνικών επαύξησης για ένα δεδομένο σενάριο. Στις επόμενες παραγράφους γίνεται αναφορά σε ορισμένες δημοφιλείς τεχνικές επαύξησης δεδομένων που χρησιμοποιούνται στους ViTs [27, 28].

Τυχαία περικοπή και αλλαγή μεγέθους (Random Cropping and Resizing): Η τυχαία περικοπή περιλαμβάνει την τυχαία επιλογή μιας μικρότερης περιοχής από την εικόνα εισόδου, διατηρώντας παράλληλα την αναλογία διαστάσεων της. Αυτή η τεχνική επαύξησης βοηθά τα ViTs να μάθουν να εστιάζουν σε σχετικές περιοχές της εικόνας και βελτιώνει την ικανότητά τους να χειρίζονται διαφορετικές κλίμακες και θέσεις αντικειμένων. Η αλλαγή μεγέθους, από την άλλη πλευρά, περιλαμβάνει την αλλαγή του μεγέθους της εικόνας διατηρώντας την αναλογία διαστάσεων της. Η αλλαγή μεγέθους μπορεί να πραγματοποιηθεί με αναβάθμιση ή μείωση της δειγματοληψίας της εικόνας, επιτρέποντας στα ViTs να μάθουν αναπαραστάσεις αναλλοίωτες σε μεταβολές κλίμακας. [27, 28]

Αναδίπλωση και περιστροφή (Flipping and Rotation): Η αναδίπλωση μιας εικόνας οριζόντια ή κάθετα εισάγει διαφοροποιήσεις στον προσανατολισμό των αντικειμένων και στις οπτικές γωνίες. Αυτή η τεχνική επαύξησης ενθαρρύνει τους ViTs να μάθουν αναλλοίωτες αναπαραστάσεις για τις αναποδογυρισμένες εικόνες και ενισχύει την ικανότητά τους να χειρίζονται την κατοπτρική συμμετρία. Η επαύξηση περιστροφής περιλαμβάνει την περιστροφή της εικόνας κατά μια συγκεκριμένη γωνία, όπως 90 μοίρες ή 180 μοίρες. Με την εκπαίδευση των ViTs σε περιστρεφόμενες εικόνες, γίνονται πιο ανθεκτικοί σε διαφορετικούς προσανατολισμούς αντικειμένων και μπορούν να αποτυπώσουν καλύτερα την αναλλοίωτη περιστροφή [27, 28].

Διαταραχές χρώματος (Color Jittering): Το χρωματικό ‘τρεμόπαιγμα’ περιλαμβάνει την εφαρμογή τυχαίων διαταραχών στα χρωματικά κανάλια μιας εικόνας, όπως η αλλαγή της φωτεινότητας, της αντίθεσης, του κορεσμού ή της απόχρωσης. Αυτή η τεχνική επαύξησης βοηθά τα ViTs να μάθουν εύρωστες αναπαραστάσεις που είναι λιγότερο ευαίσθητες στις μεταβολές των συνθηκών φωτισμού, στις μετατοπίσεις των χρωμάτων ή στα επίπεδα αντίθεσης. Η χρωματική τρεμοποίηση επιτρέπει στα ViTs να γενικεύουν καλά σε διαφορετικές συνθήκες φωτισμού και κατανομές χρωμάτων [27, 28].

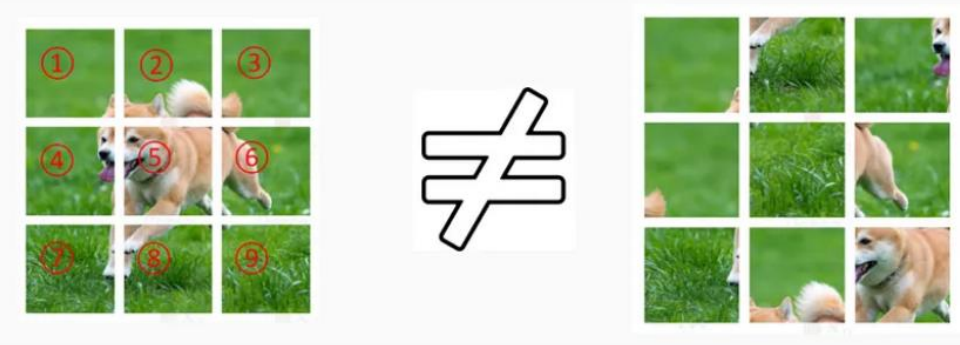
Γκαουσιανός θόρυβος και θόλωση (Gaussian Noise and Blur): Η προσθήκη γκαουσιανού θορύβου σε μια εικόνα εισάγει τυχαίες διαταραχές σε επίπεδο pixel, προσομοιώνοντας το θόρυβο ή τις ατέλειες σε εικόνες του πραγματικού κόσμου. Αυτή η τεχνική επαύξησης βοηθά τα ViTs να μάθουν να είναι πιο ανθεκτικά στο θόρυβο της εικόνας και ενισχύει την ικανότητά τους να συλλαμβάνουν σημαντικά οπτικά χαρακτηριστικά παρουσία διαταραχών. Παρομοίως, η εφαρμογή θολότητας σε μια εικόνα εξομαλύνει τις λεπτές λεπτομέρειες και ενθαρρύνει τα ViTs να επικεντρωθούν σε πιο ισχυρά και υψηλού επιπέδου χαρακτηριστικά [27, 28].

Ελαστική παραμόρφωση (Elastic Deformation): Η ελαστική παραμόρφωση περιλαμβάνει την παραμόρφωση μιας εικόνας χρησιμοποιώντας τοπικές μετατοπίσεις pixels . Αυτή η τεχνική επαύξησης προσομοιώνει τοπικές παραμορφώσεις ή παραμορφώσεις που προκαλούνται από διαφορετικές οπτικές γωνίες ή θέσεις αντικειμένων. Με την εκπαίδευση των ViTs σε ελαστικά παραμορφωμένες εικόνες, γίνονται πιο ανθεκτικοί σε χωρικούς μετασχηματισμούς και μπορούν να συλλάβουν τις χωρικές σχέσεις μεταξύ διαφορετικών περιοχών της εικόνας πιο αποτελεσματικά [27, 28].

Κωδικοποίηση θέσης (positional encoding)

Η κωδικοποίηση θέσης είναι ένα κρίσιμο συστατικό στους Transformers όρασης (ViTs), καθώς επιτρέπει στο μοντέλο να κωδικοποιεί χωρική πληροφορία και να κατανοεί τις σχετικές θέσεις των διαφορετικών σημείων μέσα σε μια εικόνα. Στις εργασίες υπολογιστικής όρασης, η κατανόηση των χωρικών σχέσεων μεταξύ αντικειμένων, η σύλληψη του πλαισίου της σκηνής και η διατήρηση της χωρικής διάταξης της εικόνας είναι απαραίτητες για την ακριβή οπτική αναγνώριση και κατανόηση.

Στα ViTs, η απουσία συνελικτικών λειτουργιών περιορίζει την ικανότητα του μοντέλου να συλλαμβάνει ρητά χωρικές πληροφορίες. Σε αντίθεση με τα Συνελικτικά Νευρωνικά Δίκτυα (CNN) που εκμεταλλεύονται φυσικά τα τοπικά δεκτικά πεδία των συνελικτικών φίλτρων, τα ViT επεξεργάζονται την εικόνα με συμβολικό τρόπο, αντιμετωπίζοντας κάθε εικόνα ως ανεξάρτητο συμβολικό πεδίο. Ωστόσο, χωρίς κωδικοποίηση θέσης, το μοντέλο στερείται άμεσης γνώσης της χωρικής διάταξης αυτών των tokens. Για να ξεπεραστεί αυτός ο περιορισμός, εισάγεται η κωδικοποίηση θέσης.Παρέχει στο μοντέλο ViT έναν μηχανισμό για την κωδικοποίηση των χωρικών σχέσεων μεταξύ των tokens. Με την ενσωμάτωση της πληροφορίας θέσης, τα ViT μπορούν να διακρίνουν μεταξύ των tokens με βάση τη θέση τους και να αποτυπώνουν το χωρικό πλαίσιο της εικόνας. Αυτό διαφαίνεται στην ακόλουθη Εικόνα 19.



Εικόνα 19. Ένα παράδειγμα της σημαντικότητας χρήσης positional encoding. [29]

Το σύστημα κωδικοποίησης θέσης ημιτόνου και συνημιτόνου χρησιμοποιείται συνήθως στα ViTs. Αυτό το σύστημα κωδικοποίησης αποδίδει μια μοναδική κωδικοποίηση σε κάθε σύμβολο με βάση τη θέση του στο πλέγμα της εικόνας.Με την εφαρμογή συναρτήσεων ημιτόνου και συνημιτόνου διαφορετικών συχνοτήτων, η κωδικοποίηση θέσης παρέχει έναν τρόπο αναπαράστασης των σχετικών θέσεων των tokens. Η συχνότητα και το πλάτος αυτών των συναρτήσεων κωδικοποιούν πληροφορίες σχετικά με τη θέση του συμβόλου μέσα στην εικόνα. Με τη συμπερίληψη της κωδικοποίησης θέσης, οι ViTs μπορούν να συλλάβουν αποτελεσματικά τόσο τοπικές όσο και παγκόσμιες χωρικές πληροφορίες. Το μοντέλο μαθαίνει να παρακολουθεί διαφορετικές περιοχές της εικόνας και να συλλαμβάνει εξαρτήσεις μεγάλης εμβέλειας, διευκολύνοντας την ολιστική κατανόηση του περιεχομένου της εικόνας. Με την ενσωμάτωση πληροφοριών θέσης, τα ViTs μπορούν να κωδικοποιήσουν όχι μόνο την εμφάνιση αλλά και τη χωρική διάταξη των τμημάτων (pathces) της εικόνας. Η παρουσία της κωδικοποίησης θέσης επιτρέπει στα ViTs να υπερέχουν σε διάφορες εργασίες όρασης υπολογιστών. Επιτρέπει στο μοντέλο να συλλογίζεται τις σχέσεις των αντικειμένων, να κατανοεί σύνθετες σκηνές και να εκτελεί χωρικούς συλλογισμούς. Επιπλέον, η κωδικοποίηση θέσης παρέχει στο μοντέλο ViT την ικανότητα να χειρίζεται εικόνες διαφορετικών μεγεθών και να προσαρμόζεται σε διαφορετικές χωρικές δομές [16, 29].

Διάνυσμα εισόδου (Input embedding)

Το διάνυσμα εισόδου σε έναν Vision Transformer(ViT) αναφέρεται στη διαδικασία μετατροπής της ακατέργαστης εικόνας εισόδου σε ένα σύνολο ενσωματωμένων αναπαραστάσεων που μπορούν να υποβληθούν σε επεξεργασία από το μοντέλο του Transformer. Το βήμα της ενσωμάτωσης εισόδου είναι ζωτικής σημασίας, καθώς επιτρέπει στον ViT να συλλάβει σημαντικά χαρακτηριστικά της εικόνας και να τα κωδικοποιήσει σε μορφή που μπορεί να υποστεί επεξεργασία από τα επόμενα στρώματα Transformers [16, 29].

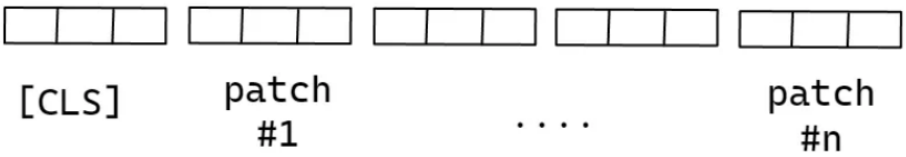
Στα ViTs, η ενσωμάτωση εισόδου επιτυγχάνεται συνήθως μέσω δύο κύριων βημάτων: του διαχωρισμού σε τμήματα (patching) και της γραμμικής προβολής (linear projection).

Το πρώτο βήμα, η επιδιόρθωση, περιλαμβάνει τη διαίρεση της εικόνας εισόδου σε ένα πλέγμα μη επικαλυπτόμενων επιφανειών. Κάθε patches αντιπροσωπεύει μια τοπική περιοχή της εικόνας και χρησιμεύει ως σύμβολο στο μοντέλο του Transformer. Η επιλογή του μεγέθους των patches καθορίζει την ανάλυση (granularity) της αναπαράστασης της εικόνας. Τα μικρότερα μεγέθη patches αποτυπώνουν λεπτότερες λεπτομέρειες, αλλά μπορεί να οδηγήσουν σε μεγαλύτερο αριθμό patches, οδηγώντας σε αυξημένη υπολογιστική πολυπλοκότητα. Αντίθετα, τα μεγαλύτερα μεγέθη patches καταγράφουν περισσότερες σφαιρικές πληροφορίες, αλλά μπορεί να χάσουν ορισμένες λεπτομέρεις. Τα κοινά χρησιμοποιούμενα μεγέθη patch στα ViTs κυμαίνονται από 8x8 έως 16x16 pixels. Αφού η εικόνα διαιρεθεί σε patches, κάθε patch μετασχηματίζεται γραμμικά σε ένα διάνυσμα ενσωμάτωσης. Αυτό το βήμα γραμμικής προβολής απεικονίζει τις τιμές των pixelsκάθε patch σε έναν χώρο ενσωμάτωσης υψηλότερων διαστάσεων, επιτρέποντας στο μοντέλο να μάθει ουσιαστικές αναπαραστάσεις του περιεχομένου της εικόνας. Η γραμμική προβολή υλοποιείται συνήθως με τη χρήση ενός πλήρως συνδεδεμένου στρώματος ή ενός συνελικτικού στρώματος που ακολουθείται από τη συγκέντρωση του παγκόσμιου μέσου όρου. Οι προκύπτουσες ενσωματωμένες αναπαραστάσεις συλλαμβάνουν τόσο τα τοπικά όσο και τα παγκόσμια χαρακτηριστικά της εικόνας και χρησιμεύουν ως είσοδος στα επόμενα στρώματα μετασχηματισμού [16, 29].

Η επιλογή της διάστασης ενσωμάτωσης είναι ένα σημαντικό ζήτημα στις ViTs. Ένας χώρος ενσωμάτωσης μεγαλύτερης διάστασης επιτρέπει στο μοντέλο να συλλάβει πιο σύνθετα μοτίβα και λεπτομερείς λεπτομέρειες, αλλά αυξάνει επίσης τις υπολογιστικές απαιτήσεις. Αντίθετα, ένας χώρος ενσωμάτωσης χαμηλότερης διάστασης μπορεί να θυσιάσει κάποιες λεπτές λεπτομέρειες, αλλά μπορεί να είναι πιο αποδοτικός από υπολογιστική άποψη. Οι διαστάσεις ενσωμάτωσης που χρησιμοποιούνται συνήθως στα ViTs κυμαίνονται από μερικές εκατοντάδες έως μερικές χιλιάδες. Η διαδικασία ενσωμάτωσης της εισόδου στις ViTs διαδραματίζει κρίσιμο ρόλο στην καταγραφή ουσιαστικών αναπαραστάσεων εικόνας. Χωρίζοντας την εικόνα σε τμήματα και προβάλλοντάς τα σε έναν χώρο ενσωμάτωσης, τα ViTs μπορούν να κωδικοποιήσουν αποτελεσματικά οπτικές πληροφορίες και να επιτρέψουν στα επόμενα στρώματα μετασχηματισμού να συλλογιστούν για το περιεχόμενο της εικόνας. Οι ενσωματωμένες αναπαραστάσεις χρησιμεύουν ως γέφυρα μεταξύ των ακατέργαστων δεδομένων εικόνας και του μηχανισμού αυτοπροσοχής του Transformer, επιτρέποντας στο μοντέλο να συλλαμβάνει τόσο τα τοπικά όσο και τα παγκόσμια χαρακτηριστικά της εικόνας και να κάνει προβλέψεις με βάση τις μαθημένες αναπαραστάσεις [16, 29].

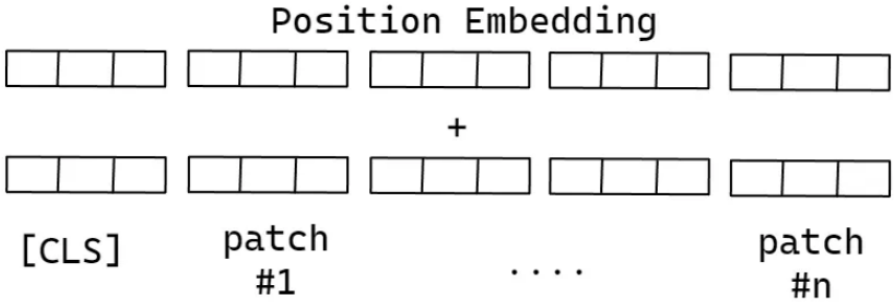
Συνοπτικά, η ενσωμάτωση της εισόδου στις ViTs περιλαμβάνει τη διαίρεση της εικόνας εισόδου σε μικρότερες περιοχές και την προβολή κάθε διαίρεσης σε ένα χώρο ενσωμάτωσης. Αυτή η διαδικασία επιτρέπει στο μοντέλο να κωδικοποιήσει σημαντικά χαρακτηριστικά της εικόνας και επιτρέπει στα επόμενα στρώματα Transformers να αιτιολογήσουν το περιεχόμενο της εικόνας. Η επιλογή του μεγέθους των patches, της διάστασης ενσωμάτωσης και του συγκεκριμένου μηχανισμού προβολής είναι σημαντικά στοιχεία για το σχεδιασμό μιας αποτελεσματικής αρχιτεκτονικής ViT.

Στην ακόλουθη Εικόνα 20 παρουσιάζεται η μορφή ενός input embedding για ταξινόμηση εικόνας βασισμένου στα patches στα οποία έχει χωριστεί η εικόνα εισόδου. Το ειδικό token [CLS] τοποθετείται στην αρχή σηματοδοτώντας την έναρξη της ακολουθίας του embedding.



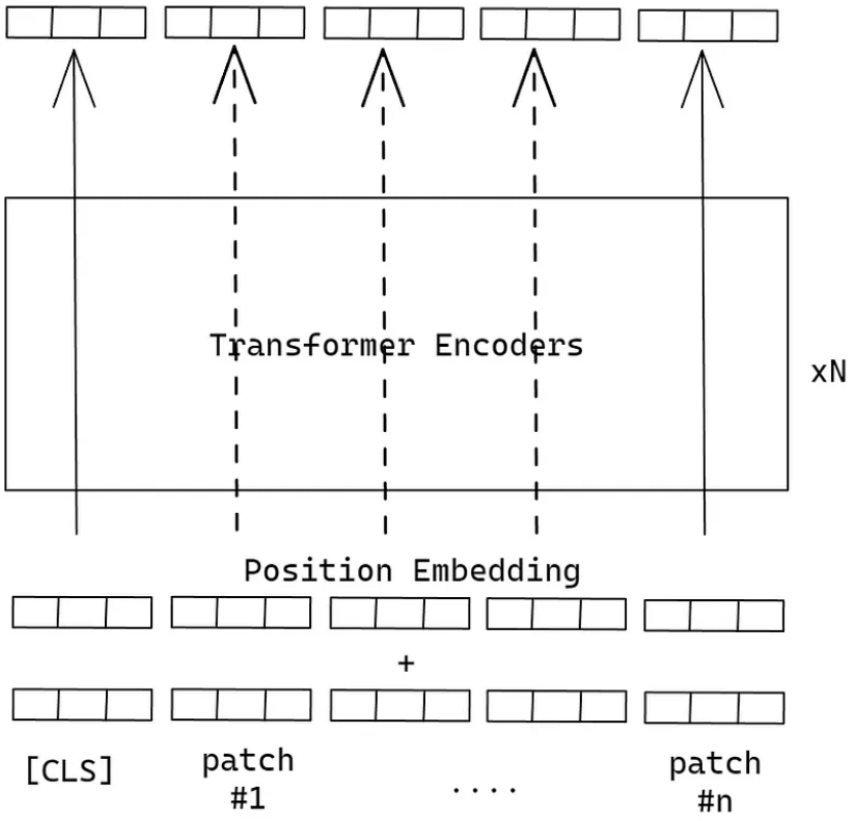
Εικόνα 20. Διάνυσμα εισόδου με βάση τα patches μιας εικόνας. [42]

Με τη συμμετοχή των διανυσμάτων θέσης που περιγράφηκαν παραπάνω, το διάνυσμα εισόδου μετατρέπεται στην ακόλουθη μορφή (Εικόνα 21):



Εικόνα 21. Διάνυσμα εισόδου μαζί με τα διανύσματα θέσης για κάθε patch. Τα διανύσματα θέσης μαθαίνονται κατά τη διαδικασία της εκπαίδευσης του Vision Transformer. [42]

Τελικά, η ροή πληροφορίας στον κωδικοποιητή του Vision Transformer με τα δοθέντα embeddings εισόδου θα μοιάζει με την ακόλουθη Εικόνα 22:



Εικόνα 22. Είσοδος των διανυσμάτων εισόδου στον κωδικοποιητή του Vision Transformer. [42]

Εξαγωγή χαρακτηριστικών (Feature extraction)

Η διαδικασία εξαγωγής χαρακτηριστικών στους Transformers όρασης (ViTs) διαφέρει από εκείνη των CNNs σε αρκετές βασικές πτυχές. Στα CNN, η εξαγωγή χαρακτηριστικών πραγματοποιείται από μια σειρά επιπέδων συνελικτικής και συγκέντρωσης. Αυτά τα στρώματα χρησιμοποιούν τοπικά δεκτικά πεδία για να συλλάβουν τις χωρικές σχέσεις στην εικόνα εισόδου. Τα συνελικτικά φίλτρα ολισθαίνουν στην εικόνα, υπολογίζοντας πολλαπλασιασμούς ανά στοιχείο και αθροίζοντας τα αποτελέσματα για την παραγωγή χαρτών χαρακτηριστικών. Αυτοί οι χάρτες χαρακτηριστικών αντιπροσωπεύουν διαφορετικά επίπεδα αφαίρεσης και αποτυπώνουν ιεραρχικά μοτίβα στην εικόνα εισόδου. Στη συνέχεια, τα στρώματα συγκέντρωσης μειώνουν τη δειγματοληψία των χαρτών χαρακτηριστικών, μειώνοντας τις χωρικές τους διαστάσεις, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες [2, 16].

Αντίθετα, οι ViTs χρησιμοποιούν μηχανισμούς αυτοπροσοχής για την εξαγωγή χαρακτηριστικών. Αντί να βασίζονται σε φίλτρα συνελίξεων, οι ViT επεξεργάζονται ολόκληρη την εικόνα ως μια ακολουθία πεπλατυσμένων επιφανειών. Αυτά τα patches προβάλλονται γραμμικά για τη δημιουργία διανυσμάτων ενσωμάτωσης. Οι κωδικοποιήσεις θέσης προστίθενται στις ενσωματώσεις για τη σύλληψη χωρικών πληροφοριών. Η ακολουθία ενσωμάτωσης τροφοδοτείται στη συνέχεια σε μια στοίβα στρωμάτων κωδικοποιητή μετασχηματισμού. Μέσα σε κάθε στρώμα, μηχανισμοί αυτοπροσοχής παρακολουθούν τις ενσωματώσεις, επιτρέποντας σε κάθε έμπλαστρο να συλλέγει πληροφορίες από όλα τα άλλα patches. Αυτός ο μηχανισμός προσοχής επιτρέπει στους ViTs να συλλαμβάνουν τις παγκόσμιες εξαρτήσεις και τις αλληλεπιδράσεις μεγάλης εμβέλειας σε ολόκληρη την εικόνα. Μετά το βήμα της αυτοπροσοχής, χρησιμοποιούνται νευρωνικά δίκτυα πρόωσης (feedforward neural networks) για την περαιτέρω επεξεργασία των ενσωματώσεων και τη σύλληψη αναπαραστάσεων υψηλότερου επιπέδου. Μια σημαντική διαφορά μεταξύ των ViTs και των CNNs είναι η απουσία πράξεων συνέλιξης (convolution) και συγκέντρωσης (pooling) στα ViTs. Αυτό εξαλείφει την ανάγκη για χειροποίητες χωρικές ιεραρχίες και επιτρέπει στους ViTs να συλλαμβάνουν πληροφορίες συνολικού πλαισίου. Ωστόσο, η εξάρτηση από την αυτοπροσοχή εισάγει επίσης υψηλότερο υπολογιστικό κόστος σε σύγκριση με τα CNN. Ως εκ τούτου, οι ViTs συχνά προ-εκπαιδεύονται σε σύνολα δεδομένων μεγάλης κλίμακας και τελειοποιούνται σε συγκεκριμένες μεταγενέστερες εργασίες για να αξιοποιήσουν την ικανότητά τους να εξάγουν πλούσια και πλαισιωμένα χαρακτηριστικά [16].

Συνολικά, ενώ τα CNN υπερέχουν στη σύλληψη τοπικών χωρικών μοτίβων, τα ViT αξιοποιούν την αυτοπροσοχή για τη σύλληψη παγκόσμιων σχέσεων και έχουν επιδείξει αξιοσημείωτες επιδόσεις σε εργασίες αναγνώρισης εικόνων, ιδίως όταν εκπαιδεύονται σε σύνολα δεδομένων μεγάλης κλίμακας.

Μέγεθος δεδομένων προεκπαίδευσης

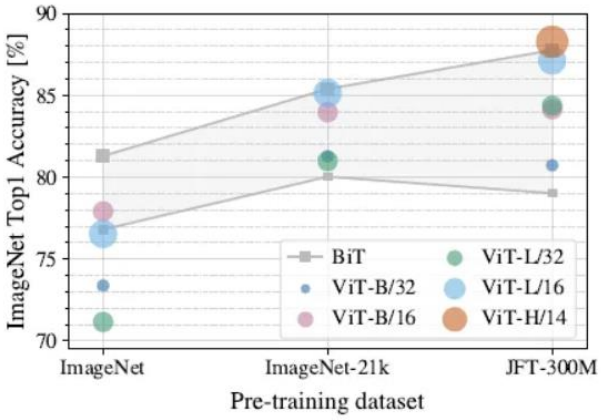
Ένα σημαντικό σύνολο προεκπαίδευσης που χρησιμοποιείται παραδοσιακά στην όραση υπολογιστών είναι το ImageNet. Το ImageNet είναι ένα ευρέως αναγνωρισμένο σύνολο δεδομένων μεγάλης κλίμακας που έχει διαδραματίσει καθοριστικό ρόλο στην πρόοδο του τομέα της όρασης υπολογιστών. Αποτελείται από πάνω από ένα εκατομμύριο επισημειωμένες εικόνες που καλύπτουν χιλιάδες διαφορετικές κατηγορίες αντικειμένων, καθιστώντας το πολύτιμο πόρο για την εκπαίδευση και την αξιολόγηση μοντέλων ταξινόμησης εικόνων. Το σύνολο δεδομένων ImageNet έχει συμβάλει σημαντικά στην ανάπτυξη της βαθιάς μάθησης και των συνελικτικών νευρωνικών δικτύων (CNN), καθώς έχει χρησιμοποιηθεί ως σημείο αναφοράς για διάφορες εργασίες υπολογιστικής όρασης, όπως η ταξινόμηση εικόνων, η ανίχνευση αντικειμένων και η κατάτμηση εικόνων. Τα μοντέλα που εκπαιδεύονται στο ImageNet έχουν επιτύχει πρωτοποριακά αποτελέσματα και έχουν γίνει η βάση για πολλές σύγχρονες προσεγγίσεις στην όραση υπολογιστών. Η διαθεσιμότητα του συνόλου δεδομένων ImageNet έχει προωθήσει την ανάπτυξη ισχυρότερων και ακριβέστερων μοντέλων, επιτρέποντας εξελίξεις στην κατανόηση και την ερμηνεία της οπτικής πληροφορίας. Η ευρεία υιοθέτηση και η επιρροή του έχουν καταστήσει το ImageNet βασικό πόρο για τους ερευνητές και τους επαγγελματίες στον τομέα της όρασης υπολογιστών [31].

Στην περίπτωση χρήσης ViT για εργασίες εικόνας, με κυριότερη την ταξινόμηση εικόνων, το μέγεθος του συνόλου προεκπαίδευσης παίζει σημαντικό ρόλο στις τελικές επιδόσεις του μοντέλου. Μάλιστα, το μέγεθος του συνόλου δεδομένων προ-εκπαίδευσης επηρεάζει περισσότερο την απόδοση των Vision Transformers (ViTs) σε σύγκριση με αυτή των Convolutional Neural Networks (CNNs), αν θεωρηθεί ότι και οι δύο αρχιτεκτονικές εκπαιδεύονται ακριβώς πάνω στο ίδιο σύνολο δεδομένων. Γενικά, τα μεγαλύτερα σύνολα δεδομένων προ-εκπαίδευσης μπορούν να παρέχουν πιο ποικίλα και αντιπροσωπευτικά παραδείγματα, επιτρέποντας στα μοντέλα να μαθαίνουν πλουσιότερες και πιο γενικευμένες οπτικές αναπαραστάσεις [29].

Τα CNN παραδοσιακά επωφελούνται από σύνολα δεδομένων μεγάλης κλίμακας για προ-εκπαίδευση, όπως το ImageNet, το οποίο περιέχει εκατομμύρια εικόνες με ετικέτες. Το μεγάλο μέγεθος του συνόλου δεδομένων βοηθά τα CNN να συλλάβουν ένα ευρύ φάσμα οπτικών μοτίβων και τους επιτρέπει να γενικεύουν καλά σε διάφορες μεταγενέστερες εργασίες. Αυτό οφείλεται στο γεγονός ότι τα CNN μαθαίνουν ιεραρχικές αναπαραστάσεις αξιοποιώντας τα τοπικά δεκτικά πεδία, τα στρώματα συγκέντρωσης και τον καταμερισμό βαρών, τα οποία είναι αποτελεσματικά για τη σύλληψη χωρικά τοπικών μοτίβων στις εικόνες. Τα CNN έχουν σημειώσει μεγάλη επιτυχία σε εργασίες όρασης υπολογιστών λόγω της ικανότητάς τους να εκμεταλλεύονται τις τοπικές χωρικές εξαρτήσεις και το αμετάβλητο στη μετάφραση. Αντίθετα, τα ViT βασίζονται σε μηχανισμούς αυτοπροστασίας για να μοντελοποιήσουν τις παγκόσμιες αλληλεπιδράσεις μεταξύ των patches εικόνας και να συλλάβουν εξαρτήσεις μεγάλης εμβέλειας. Ενώ η αυτοπροσοχή επιτρέπει στους ViTs να συλλογίζονται ολιστικά για το περιεχόμενο της εικόνας, απαιτεί έκθεση σε ποικίλα και αντιπροσωπευτικά οπτικά πρότυπα κατά την προ-εκπαίδευση. Με ένα περιορισμένο σύνολο δεδομένων προ-εκπαίδευσης, οι ViTs μπορεί να δυσκολευτούν να μάθουν ισχυρές οπτικές αναπαραστάσεις και μπορεί να παρουσιάσουν υπερβολική προσαρμογή στα συγκεκριμένα παραδείγματα που είδαν κατά την προ-εκπαίδευση [29].

Εμπειρικές μελέτες έχουν δείξει ότι οι ViT επωφελούνται από μεγαλύτερα σύνολα δεδομένων προ-εκπαίδευσης, παρόμοια με τα CNN. Ειδικότερα, οι ViTs έχουν παρουσιάσει σημαντική βελτίωση των επιδόσεων όταν εκπαιδεύονται σε σύνολα δεδομένων μεγάλης κλίμακας, όπως το σύνολο δεδομένων JFT-300M, το οποίο περιέχει 300 εκατομμύρια εικόνες. Το μεγαλύτερο μέγεθος του συνόλου δεδομένων επιτρέπει στους ViTs να μαθαίνουν πιο γενικεύσιμα χαρακτηριστικά και να επιτυγχάνουν καλύτερες επιδόσεις σε επόμενες εργασίες. Είναι σημαντικό να σημειωθεί ότι τα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων, όπως η ποικιλομορφία των εικόνων, η κατανομή των οπτικών εννοιών και η ποιότητα των σχολίων, επηρεάζουν επίσης την απόδοση των ViTs. Ενώ τα μεγαλύτερα σύνολα δεδομένων παρέχουν περισσότερες ευκαιρίες στο μοντέλο να μάθει, η ποιότητα και η ποικιλομορφία των δεδομένων είναι εξίσου σημαντικές. Η εξασφάλιση ενός ποικιλόμορφου και αντιπροσωπευτικού συνόλου δεδομένων για την προ-εκπαίδευση των ViTs μπορεί να βοηθήσει στην αντιμετώπιση πιθανών προκαταλήψεων και στη βελτίωση της γενίκευσης σε διαφορετικά οπτικά σενάρια [29].

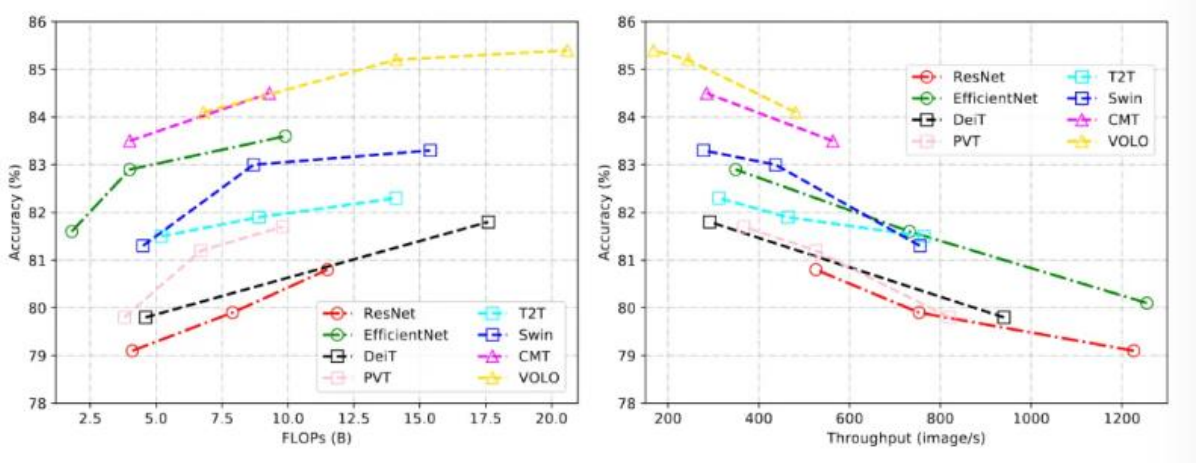
Στην ακόλουθη Εικόνα 23 παρουσιάζεται ένα διάγραμμα που αποδεικνύει τη σημασία του μεγέθους του συνόλου προεκπαίδευσης ενός οπτικού Transformer (εν προκειμένω του ViT σε 5 διαφορετικά μεγέθη με 14, 16 και 32 κεφαλές, και με διαφορετικά μεγέθη με το B να σηματοδοτεί το μικρότερο και το H το μεγαλύτερο. Σημειώνεται ότι το μέγεθος αφορά στον αριθμό εκπαιδεύσιμων παραμέτρων του κάθε μοντέλου). Παράλληλα, το μοντέλο BiT αναφέρεται σε CNN αρχιτεκτονική.



Εικόνα 23. Ακρίβεια στο σύνολο δεδομένων ελέγχου (test set) του ImageNet για μοντέλα που έχουν εκπαιδευτεί σε σύνολα προεκπαίδευσης διαφορετικού μεγέθους. [29]

Με βάση το παραπάνω διάγραμμα, γίνεται αντιληπτό ότι η απόδοση των ViT μοντέλων σε σύγκριση με τα αντίστοιχα CNN, όπως το Big Transfer (BiT), ποικίλλει ανάλογα με το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση. Όταν εκπαιδεύονται στο σύνολο δεδομένων ImageNet, το οποίο αποτελείται από 1 εκατομμύριο εικόνες, οι ViT τείνουν να έχουν χειρότερες επιδόσεις από τους BiT. Ωστόσο, καθώς αυξάνεται το μέγεθος του συνόλου δεδομένων, η απόδοση της ViT βελτιώνεται σε σχέση με την BiT. Για παράδειγμα, στο μεγαλύτερο σύνολο δεδομένων ImageNet-21k, το οποίο περιέχει 14 εκατομμύρια εικόνες, το ViT επιδεικνύει συγκρίσιμες επιδόσεις με το BiT. Επιπλέον, όταν εκπαιδεύεται στο σύνολο δεδομένων JFT, το οποίο περιλαμβάνει 300 εκατομμύρια εικόνες, το ViT ξεπερνά το BiT από άποψη επιδόσεων. Τα ευρήματα αυτά δείχνουν ότι η σχετική απόδοση του ViT και του BiT επηρεάζεται από την κλίμακα του συνόλου δεδομένων εκπαίδευσης, με το ViT να παρουσιάζει τα πλεονεκτήματά του σε σύνολα δεδομένων με μεγαλύτερες συλλογές εικόνων.

Επιπλέον, στην Εικόνα 24 παρουσιάζεται ο αριθμός των FLOPS και ο χρόνος διεκπεραίωσης (throughput) για διαφορετικές αρχιτεκτονικές. Υψηλές επιδόσεις απαιτούν περισσότερα FLOPS και μειωμένο throughput, αναδεικνύοντας έτσι τις θυσίες που πρέπει να γίνουν για βελτίωση της μετρικής της επίδοσης για την ταξινόμηση εικόνας.



Εικόνα 24. Σύγκριση ποικίλων μοντέλων ταξινόμησης εικόνας, τόσο CNN όσο και Vision Transformer αναφορικά με τις υπολογιστικές τους απαιτήσεις προκειμένου να πετύχουν υψηλή ακρίβεια ταξινόμησης [43].

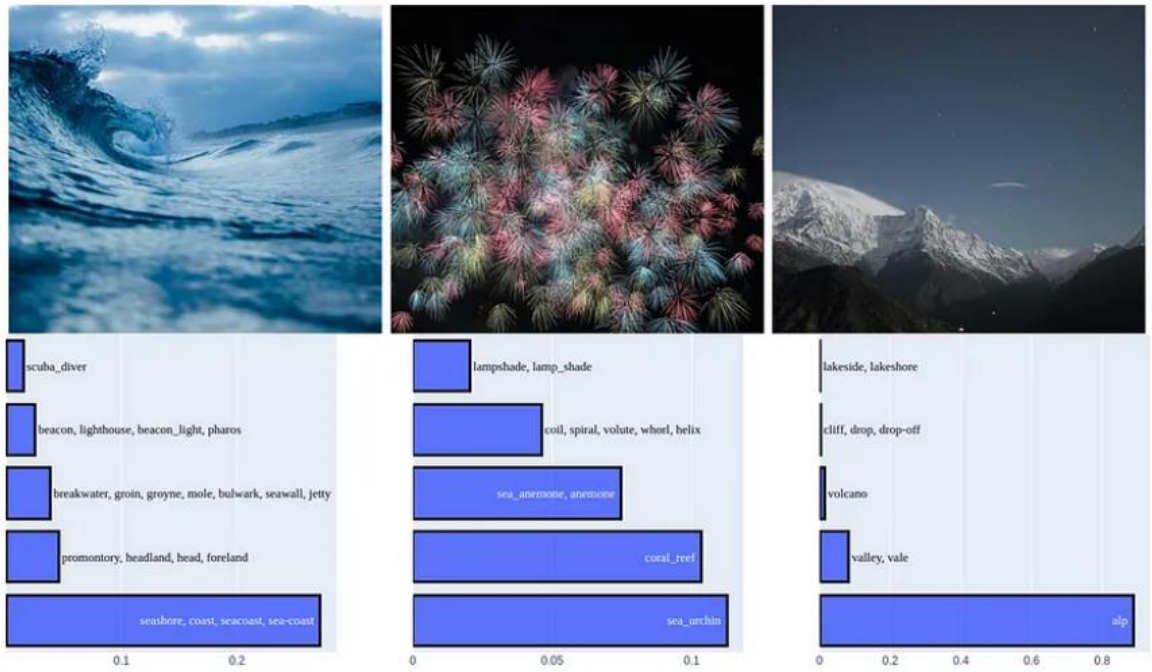
2.3.2 Fine – tuning στους οπτικούς Transformers

Η διαδικασία fine - tuning των ViTs διαδραματίζει κρίσιμο ρόλο στη μεγιστοποίηση της απόδοσής τους σε συγκεκριμένες μεταγενέστερες εργασίες. Ενώ η προ-εκπαίδευση των ViTs σε σύνολα δεδομένων μεγάλης κλίμακας, όπως το ImageNet, παρέχει ισχυρά θεμέλια για την κατανόηση οπτικών εννοιών, το fine - tuning επιτρέπει στο μοντέλο να προσαρμόζεται και να εξειδικεύεται για συγκεκριμένες εργασίες με περιορισμένα επισημασμένα δεδομένα. Το fine - tuning περιλαμβάνει την εκπαίδευση του προ-εκπαιδευμένου ViT σε ένα σύνολο δεδομένων συγκεκριμένης εργασίας, συνήθως με μικρότερο αριθμό επισημασμένων δειγμάτων. Η διαδικασία του fine – tuning των ViT ξεκινά με την αρχικοποίηση των παραμέτρων του μοντέλου χρησιμοποιώντας τα προ-εκπαιδευμένα βάρη που λαμβάνονται από τη φάση της προ-εκπαίδευσης. Ωστόσο, λόγω των διαφορών μεταξύ των συνόλων δεδομένων προ-εκπαίδευσης και στόχου, το fine - tuning απαιτεί προσεκτική εξέταση διαφόρων βασικών πτυχών [16, 26].

Πρώτον, η διαδικασία τελειοποίησης περιλαμβάνει την προσαρμογή του τελικού στρώματος ταξινόμησης του μοντέλου ώστε να ταιριάζει με τον αριθμό των κλάσεων στην εργασία-στόχο. Αυτό διασφαλίζει ότι το ViT παράγει προβλέψεις συμβατές με το χώρο ετικετών του συνόλου δεδομένων-στόχου. Εκτός από το τελικό στρώμα ταξινόμησης, μπορούν να προστεθούν ή να τροποποιηθούν και άλλα ειδικά για την εργασία στρώματα, όπως ανιχνευτές αντικειμένων ή σημασιολογικοί τμηματοποιητές, ώστε να ανταποκρίνονται στις ειδικές απαιτήσεις της εργασίας-στόχου. Δεύτερον, είναι σημαντικό να καθοριστεί ο βέλτιστος ρυθμός μάθησης και το βέλτιστο χρονοδιάγραμμα για το fine - tuning των ViTs. Σε ορισμένες περιπτώσεις, προτιμάται ένας χαμηλότερος ρυθμός μάθησης για να επιτραπεί στο μοντέλο να συγκλίνει σταδιακά και να αποφευχθεί η καταστροφική λήθη. Εναλλακτικά, μπορούν να χρησιμοποιηθούν στρατηγικές προθέρμανσης του ρυθμού μάθησης για να εξασφαλιστεί ταχύτερη αρχική μάθηση και καλύτερη προσαρμογή στο σύνολο δεδομένων-στόχου [16, 26].

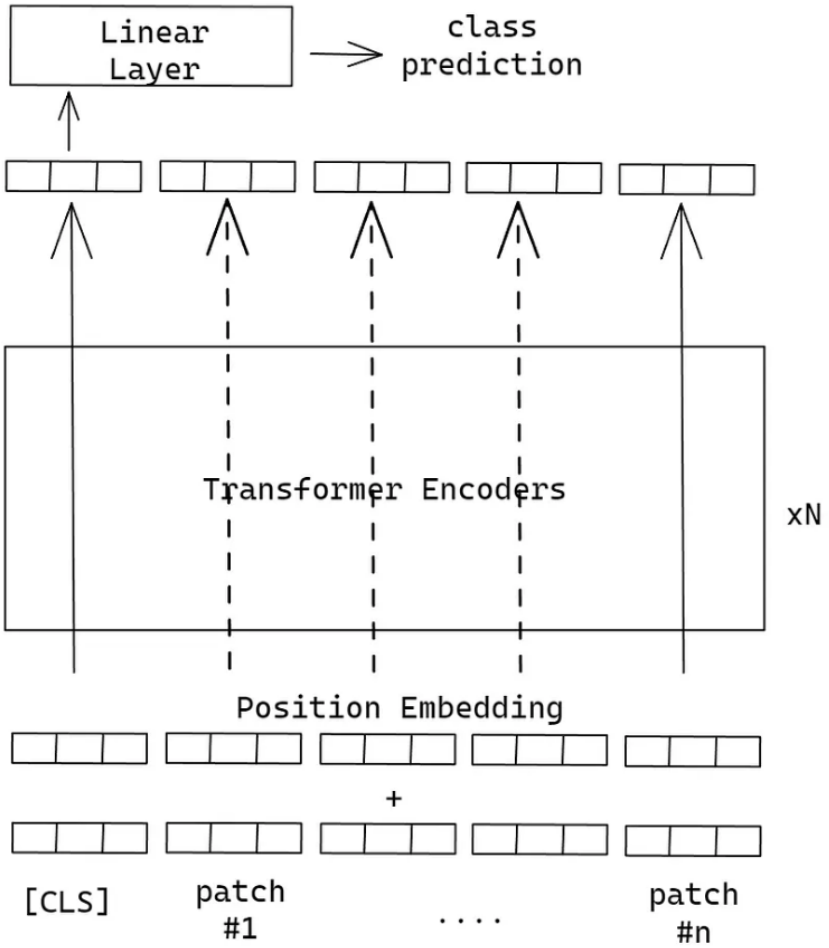
Ένα άλλο σημαντικό στοιχείο για το fine - tuning των ViT είναι η επιλογή των τεχνικών επαύξησης (data augumentation), οι οποίες περιγράφηκαν και παραπάνω αναφορικά με το στάδιο της προεκπαίδευσης. Η επαύξηση των δεδομένων διαδραματίζει καθοριστικό ρόλο στη βελτίωση της ικανότητας γενίκευσης του μοντέλου και στη μείωση της υπερπροσαρμογής (overfitting). Οι συνήθεις τεχνικές επαύξησης για δεδομένα εικόνας περιλαμβάνουν τυχαία περικοπή, αναστροφή, περιστροφή και χρωματικό ‘τρεμόπαιγμα’ (jittering). Με την εφαρμογή αυτών των επαυξήσεων, το εκλεπτυσμένο ViT γίνεται πιο ανθεκτικό στις μεταβολές του συνόλου δεδομένων-στόχου, βελτιώνοντας την απόδοσή του σε αθέατα δεδομένα. Επιπλέον, η μάθηση μεταφοράς μπορεί να χρησιμοποιηθεί στη διαδικασία τελειοποίησης. Αντί να εκπαιδεύεται το ViT από την αρχή στην εργασία-στόχο, μπορεί να αξιοποιηθεί η μεταφορά γνώσης από συναφείς εργασίες ή τομείς. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να επωφεληθεί από τα προ-εκπαιδευμένα βάρη που λαμβάνονται από παρόμοιες εργασίες, μειώνοντας τον όγκο των επισημασμένων δεδομένων που απαιτούνται για το fine - tuning. Τέλος, το fine - tuning των ViTs περιλαμβάνει συχνά μια ισορροπία μεταξύ της αξιοποίησης της προ-εκπαιδευμένης γνώσης και της προσαρμογής στην εργασία-στόχο. Η υπερβολικά επιθετική λεπτομερής ρύθμιση μπορεί να έχει ως αποτέλεσμα τη λήθη των οπτικών εννοιών που έχουν διδαχθεί προηγουμένως, ενώ η ανεπαρκής λεπτομερής ρύθμιση μπορεί να περιορίσει την ικανότητα του μοντέλου να προσαρμοστεί στο σύνολο δεδομένων-στόχου. Η εύρεση της σωστής ισορροπίας μέσω εμπειρικού πειραματισμού είναι απαραίτητη [16, 26].

Λαμβάνοντας σαν παράδειγμα χρήσης την ταξινόμηση εικόνας, ένας Vision Transformer θα γίνει fine – tuned σε ένα ανάλογο σύνολο δεδομένων με labels στις εικόνες. Στην περίπτωση αυτή, η έξοδος του fine – tuned Vision Transformer θα μοιάζει με την ακόλουθη της Εικόνας 25:



Εικόνα 25. Έξοδος ενός fine – tuned Vision Transformer. Οι πιθανότητες κάθε κατηγορίας στο επίπεδο softmax καθορίζουν και την τελική κατηγορία που προβλέπεται για κάθε εικόνα εκ μέρους του Vision Transformer [41].

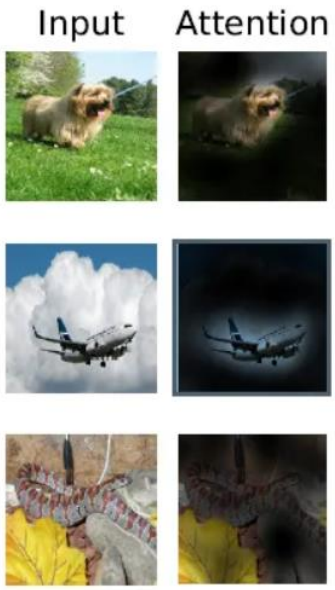
Τέλος, λαμβάνοντας ολόκληρη τη ροή πληροφορίας ήδη από τα input embeddings, στο ακόλουθο διάγραμμα της Εικόνας 26 παρουσιάζεται η πορεία που ακολουθεί μια εικόνα από τη μετατροπή της σε patches μέχρι να ληφθούν οι προβλέψεις στην έξοδο του Vision Transformer.



Εικόνα 26. Συνολική ροή πληροφορίας από την είσοδο (input embedding της εικόνας σε μορφή patch) ως την πρόβλεψη εξόδου [42].

2.3.3 Ευρωστία και γενικευσιμότητα (Robustness and Generalization)

Όσον αφορά τις δυνατότητες γενίκευσης και την ανθεκτικότητας σε μεταβολές της εισόδου, οι ViTs και τα CNNs παρουσιάζουν κάποιες ομοιότητες και διαφορές. Τα CNNs έχουν αποδεδειγμένη γενίκευση και ευρωστία σε διάφορες εργασίες όρασης υπολογιστών. Υπερέχουν στη σύλληψη τοπικών χωρικών μοτίβων και έχουν επιτύχει σε εργασίες αναγνώρισης εικόνων, ακόμη και με περιορισμένα δεδομένα εκπαίδευσης. Τα CNN αξιοποιούν τον διαμοιρασμό βαρών και τις χωρικές ιεραρχίες μέσω των λειτουργιών συνένωσης και συγκέντρωσης, επιτρέποντάς τους να μαθαίνουν ιεραρχικές αναπαραστάσεις που είναι αναλλοίωτες στη μετάφραση, την περιστροφή και την κλίμακα. Αυτή η ιδιότητα επιτρέπει στα CNNs να γενικεύουν καλά σε αόρατα δεδομένα και να παρουσιάζουν ανθεκτικότητα έναντι μικρών παραλλαγών στις εικόνες εισόδου. Επιπροσθέτως, τεχνικές όπως η αύξηση των δεδομένων, η διακοπή και η κανονικοποίηση ενισχύουν περαιτέρω τις ικανότητες γενίκευσής τους. Από την άλλη πλευρά, οι ViT παρουσιάζουν ορισμένα μοναδικά χαρακτηριστικά όσον αφορά τη γενίκευση και την ευρωστία. Λόγω της εξάρτησής τους από τους μηχανισμούς αυτοπροσοχής, οι ViTs μπορούν να συλλάβουν συνολικές σχέσεις και εξαρτήσεις μεγάλης εμβέλειας στην εικόνα εισόδου. Αυτή η ικανότητα επιτρέπει στους ViTs να λαμβάνουν υπόψη τους πληροφορίες πλαισίου σε ολόκληρη την εικόνα, επιτρέποντάς τους να συλλαμβάνουν πλουσιότερες σημασιολογικές πληροφορίες. Ένα τέτοιο παράδειγμα δίνεται στην Εικόνα 27: οι μηχανισμοί προσοχής πάνω στους οποίους έχουν χτιστεί τα ViT επιτρέπει την εστίαση στην πιο σημαντική πληροφορία της εικόνας. Δηλαδή, σε παρόμοια σημεία θα εστίαζε και ο ίδιος ο άνθρωπος προκειμένου να κατατάξει μια εικόνα σε κατάλληλη κατηγορία [16, 29].



Εικόνα 27. Χάρτες προσοχής που προκύπτουν κατά την ταξινόμηση εικόνων από ένα ViT. [29]

Ωστόσο, οι ViTs είναι πιο απαιτητικοί υπολογιστικά από τους CNNs, γεγονός που μπορεί να περιορίσει την επεκτασιμότητά τους και τη γενίκευσή τους σε σύνολα δεδομένων μεγάλης κλίμακας. Για τον μετριασμό αυτού του φαινομένου, η προ-εκπαίδευση σε σύνολα δεδομένων μεγάλης κλίμακας και το fine - tuning σε συγκεκριμένες εργασίες χρησιμοποιούνται συνήθως για τους ViTs [16].

Ενώ τα CNN έχουν μελετηθεί και βελτιστοποιηθεί εκτενώς για γενίκευση και ανθεκτικότητα, τα ViT εξακολουθούν να αποτελούν σχετικά νέο πεδίο έρευνας. Ως αποτέλεσμα, η κατανόηση των χαρακτηριστικών γενίκευσης και ευρωστίας τους εξακολουθεί να εξελίσσεται. Αρκετές μελέτες έχουν δείξει ότι τα ViTs μπορούν να επιτύχουν ανταγωνιστικές επιδόσεις με τα CNNs σε σύνολα δεδομένων αναφοράς όπως το ImageNet-21k και το JFT, γεγονός που υποδηλώνει τις δυνατότητές τους για γενίκευση. Ωστόσο, απαιτείται περαιτέρω έρευνα για τη διερεύνηση τεχνικών όπως η κανονικοποίηση, η αύξηση των δεδομένων και ο έλεγχος ευρωστίας ειδικά για τα ViTs [16].

2.3.4 Δημοφιλή μοντέλα Vision Transformer

Παρακάτω δίνονται κάποιες δημοφιλείς αρχιτεκτονικές Vision Transformers οι οποίες έχουν επηρεάσει σημαντικά το χώρο της ταξινόμησης εικόνας.

ViT (Vision Transformer)

Η αρχική αρχιτεκτονική ViT που παρουσιάστηκε από τους Dosovitskiy et al. (2021) έχει απλή δομή που αποτελείται από στοιβαγμένα στρώματα κωδικοποιητή Transformer. Χωρίζει την εικόνα εισόδου σε patches, τα οποία στη συνέχεια ενσωματώνονται γραμμικά και υποβάλλονται σε επεξεργασία από τα στρώματα Transformer. Παρά την απλότητά του, το ViT πέτυχε αξιοσημείωτες επιδόσεις σε διάφορους δείκτες αναφοράς ταξινόμησης εικόνων [16].

DeiT (Data-efficient Image Transformers)

Το μοντέλο αυτό εισήχθη το 2020 και επικεντρώνεται στη βελτίωση της αποδοτικότητας των δεδομένων των ViTs. Εισάγει τεχνικές απόσταξης (distillation), όπου ένα μικρότερο ViT εκπαιδεύεται με γνώση που μεταφέρεται από ένα μεγαλύτερο μοντέλο. Το DeiT επιτυγχάνει ανταγωνιστικές επιδόσεις με το ViT, ενώ απαιτεί σημαντικά λιγότερα δεδομένα εκπαίδευσης. Με την απόσταξη της γνώσης από το μοντέλο του δασκάλου (teacher model), το DeiT είναι σε θέση να συλλάβει σημαντικές οπτικές αναπαραστάσεις, ενώ παράλληλα επωφελείται από την αποτελεσματικότητα και την επεκτασιμότητα των Transformers. Τα πειραματικά αποτελέσματα δείχνουν ότι το DeiT επιτυγχάνει αξιοσημείωτες επιδόσεις σε διάφορα σύνολα δεδομένων αναφοράς, όπως το ImageNet, με ακρίβεια συγκρίσιμη με τα παραδοσιακά δίκτυα συνελικτικής ανάλυσης ή ακόμη και καλύτερη από αυτά. Το DeiT καταδεικνύει τις δυνατότητες της απόσταξης γνώσης που επιτρέπει την αποδοτική εκπαίδευση Transformers όρασης, καθιστώντας τους προσιτούς σε εφαρμογές όπου σύνολα δεδομένων μεγάλης κλίμακας με ετικέτες δεν είναι άμεσα διαθέσιμα [18].

TNT (Transformer in Transformer)

Η αρχιτεκτονική TNT ενισχύει τον μηχανισμό αυτοπροσοχής στα ViTs. Εισάγει ένα μπλοκ Transformer μέσα σε κάθε μπλοκ αυτοπροσοχής, επιτρέποντας την ιεραρχική μοντελοποίηση των χαρακτηριστικών της εικόνας, κι επιτυγχάνοντας έτσι λεπτομερή (fine – grained) ταξινόμηση. Το TNT επιδεικνύει βελτιωμένη απόδοση σε διάφορες εργασίες αναγνώρισης εικόνων. Ενώ οι Transformers όρασης υπερέχουν στη μοντελοποίηση του συνολικού πλαισίου με τη χρήση μηχανισμών αυτοπροσοχής, μπορεί να δυσκολεύονται με τη σύλληψη τοπικών λεπτομερειών λόγω των patches σταθερού μεγέθους που επεξεργάζονται. Το TNT αντιμετωπίζει αυτό το ζήτημα εισάγοντας μια ιεραρχική αρχιτεκτονική που ενσωματώνει μια ενότητα "Transformer in Transformer". Η ενότητα " Transformer in Transformer " ενεργεί ως ένθετος Transformer εντός του κύριου δικτύου Transformers. Λειτουργεί σε τοπικά τεμάχια της εικόνας εισόδου, επιτρέποντας τη λεπτομερή μοντελοποίηση των χωρικών εξαρτήσεων εντός κάθε τεμαχίου. Με την ενσωμάτωση μηχανισμών αυτο-προσοχής τόσο σε επίπεδο patch όσο και σε επίπεδο token, το TNT συλλαμβάνει αποτελεσματικά τόσο τοπικές όσο και παγκόσμιες πληροφορίες πλαισίου. Ο εσωτερικός Transformer παρακολουθεί τις τοπικές αναπαραστάσεις των επιφανειών, επιτρέποντας την ακριβή μοντελοποίηση των λεπτών λεπτομερειών, ενώ ο εξωτερικός Transformer παρακολουθεί το παγκόσμιο πλαίσιο, καταγράφοντας ευρύτερες σχέσεις μεταξύ των επιφανειών. Αυτός ο ιεραρχικός σχεδιασμός του TNT του επιτρέπει να μοντελοποιεί αποτελεσματικά τόσο τις τοπικές όσο και τις παγκόσμιες αλληλεπιδράσεις, με αποτέλεσμα βελτιωμένη απόδοση σε διάφορες εργασίες όρασης υπολογιστών. Τα πειραματικά αποτελέσματα καταδεικνύουν ότι το TNT επιτυγχάνει κορυφαίες επιδόσεις σε δείκτες αναφοράς ταξινόμησης εικόνων, όπως το ImageNet, ξεπερνώντας τους παραδοσιακούς vision transformers. Επιπλέον, το TNT παρουσιάζει ισχυρές δυνατότητες γενίκευσης, ακόμη και όταν εκπαιδεύεται σε περιορισμένα δεδομένα με ετικέτες, καθιστώντας το μια πολλά υποσχόμενη αρχιτεκτονική για σενάρια με περιορισμένη διαθεσιμότητα δεδομένων με ετικέτες [32].

PiT (pooling in Tranformer)Το μοντέλο PiT συνδυάζει τα πλεονεκτήματα των CNN και των ViT εισάγοντας λειτουργίες συγκέντρωσης μέσα στην αρχιτεκτονική Transformer. Αντικαθιστά τη σήμανση με βάση το patch με σήμανση με βάση τη συγκέντρωση, η οποία επιτρέπει στο μοντέλο να συλλαμβάνει αποτελεσματικά τόσο τις τοπικές όσο και τις παγκόσμιες πληροφορίες. Σε αντίθεση με τους παραδοσιακούς vision transformers που λειτουργούν σε patches εικόνας σταθερού μεγέθους, το PiT εισάγει μια υβριδική προσέγγιση όπου η εικόνα εισόδου επεξεργάζεται από ένα CNN κορμού για την εξαγωγή τοπικών χαρακτηριστικών, τα οποία στη συνέχεια τροφοδοτούνται σε ένα δίκτυο Transformer για την καταγραφή των παγκόσμιων εξαρτήσεων. Αυτός ο υβριδικός σχεδιασμός επιτρέπει στο PiT να μοντελοποιεί αποτελεσματικά τόσο τις τοπικές όσο και τις παγκόσμιες πληροφορίες, αξιοποιώντας τη χωρική ιεραρχία που αποτυπώνεται από τα CNN και τον μηχανισμό προσοχής των Transformers. Επιπλέον, το PiT ενσωματώνει έναν μηχανισμό επισήμανσης συμβολικών ετικετών κατά την προ-εκπαίδευση, όπου ένα υποσύνολο των patches εικόνας καλύπτεται και το μοντέλο εκπαιδεύεται για να προβλέψει τις ετικέτες αυτών των patches. Αυτός ο στόχος επισήμανσης συμβόλων ενισχύει την ικανότητα του μοντέλου να συλλαμβάνει λεπτομερείς λεπτομέρειες και βελτιώνει την ανθεκτικότητά του σε αποκρύψεις και μεταβολές στην είσοδο. Τα πειραματικά αποτελέσματα καταδεικνύουν ότι το PiT επιτυγχάνει ανταγωνιστικές επιδόσεις σε διάφορα σύνολα δεδομένων αναφοράς, αναδεικνύοντας την αποτελεσματικότητά του στον συνδυασμό των πλεονεκτημάτων των CNN και των Transformers για εργασίες αναγνώρισης εικόνων. Το PiT επιτυγχάνει ανταγωνιστικές επιδόσεις με το ViT, ενώ είναι πιο αποδοτικό από υπολογιστική άποψη [33].

CaiT (Class-Attention in Image Transformers)

Το CaiT επεκτείνει τον ViT με μηχανισμούς διασταυρούμενης προσοχής μεταξύ των patches και των οπτικών σημείων. Βελτιώνει την ικανότητα του μοντέλου να παρακολουθεί τα σχετικά patches και τις αλληλεπιδράσεις τους. Το CaiT επιδεικνύει βελτιωμένη απόδοση σε εργασίες ανίχνευσης αντικειμένων [34].

CoaT (Co-Scale Conv-Attentional Image Transformers)

Ο CoaT που προτάθηκε από τους Yuan et al. (2021), συνδυάζει μηχανισμούς συνελίξεων και προσοχής σε μια υβριδική αρχιτεκτονική. Εισάγει μια μονάδα σύντηξης πολλαπλών κλιμάκων που ενσωματώνει τα χαρακτηριστικά συνελικτικής ανάλυσης με τα χαρακτηριστικά προσοχής. Το CoaT επιτυγχάνει ανταγωνιστικές επιδόσεις με το ViT, ενώ είναι πιο αποδοτικό από υπολογιστική άποψη [35].

LinViT (Linear Vision Transformers)

Το LinViT (2021), επικεντρώνεται στη βελτίωση της γραμμικής προβολής των patches εικόνας σε ViTs. Αντικαθιστά τη συμβατική γραμμική προβολή με μια λειτουργία γραμμικοποιημένης συνέλιξης, η οποία ενισχύει την ικανότητα μοντελοποίησης των τοπικών χαρακτηριστικών της εικόνας. Το LinViT επιδεικνύει βελτιωμένη απόδοση σε διάφορες εργασίες αναγνώρισης εικόνων [36].

Swin Transformer

Ο Transformer Swin είναι μια πρόσφατη αρχιτεκτονική Vision Transformer που προτάθηκε από τους Liu et al. (2021) και εισάγει την έννοια των ιεραρχικών παραθύρων μετατόπισης. Σε αντίθεση με τους συμβατικούς Transformers όρασης που επεξεργάζονται εικόνες με patches σταθερού μεγέθους, ο Swin Transformer διαιρεί την εικόνα σε μικρότερα παράθυρα και εκτελεί τους υπολογισμούς ιεραρχικά σε πολλαπλά επίπεδα. Αυτός ο μηχανισμός ιεραρχικής μετατόπισης επιτρέπει στο μοντέλο να συλλαμβάνει αποτελεσματικά τόσο τοπικές όσο και παγκόσμιες πληροφορίες πλαισίου. Ο Swin Transformer ενσωματώνει επίσης έναν μηχανισμό προσοχής μετατοπισμένων παραθύρων, ο οποίος μειώνει την υπολογιστική πολυπλοκότητα περιορίζοντας τον υπολογισμό της προσοχής σε τοπικά παράθυρα αντί να παρακολουθεί ολόκληρη την εικόνα. Αυτός ο μηχανισμός προσοχής βοηθά στον αποτελεσματικό χειρισμό εικόνων μεγάλης κλίμακας. Ο Swin Transformer έχει επιδείξει ισχυρές επιδόσεις σε διάφορες εργασίες όρασης υπολογιστών, συμπεριλαμβανομένης της ταξινόμησης εικόνων και της ανίχνευσης αντικειμένων, ξεπερνώντας προηγούμενα μοντέλα τελευταίας τεχνολογίας. Η ικανότητα της αρχιτεκτονικής να μοντελοποιεί εξαρτήσεις μεγάλης εμβέλειας και η υπολογιστική της αποδοτικότητα καθιστούν τον Swin Transformer μια πολλά υποσχόμενη προσέγγιση για την προώθηση του πεδίου των Transformers όρασης [37].

Μια παραλλαγή του Swin είναι ο CSWin Transformer [43].

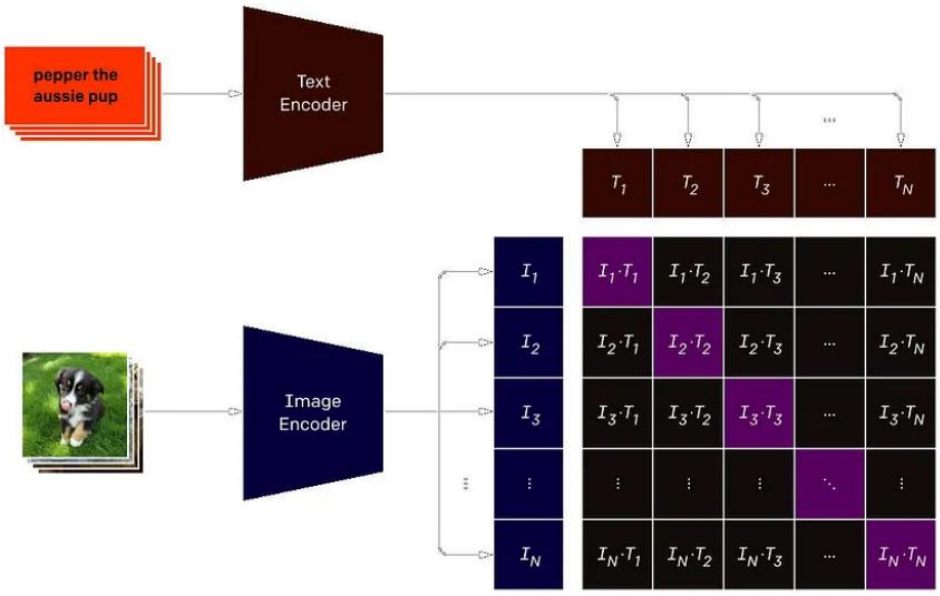
BEiT (Bidirectional Encoder representation from Image Transformers)

Το BeiT (Bidirectional Encoder representation from Image Transformers) είναι ένα μοντέλο αναπαράστασης όρασης με αυτοεπιτήρηση που εισήχθη από τους Bao et al. (2021). Εμπνευσμένο από την επιτυχία του BERT στην επεξεργασία φυσικής γλώσσας, το BeiT εφαρμόζει την εργασία μοντελοποίησης καλυμμένων εικόνων για την προεκπαίδευση Vision Transformers. Σε αυτό το στάδιο προεκπαίδευσης, κάθε εικόνα αναπαρίσταται από δύο όψεις: patches εικόνας και οπτικά σημεία. Η αρχική εικόνα συμβολίζεται σε οπτικά tokens και εφαρμόζεται τυχαία μάσκα σε ορισμένα patches εικόνας. Ο στόχος της προ-εκπαίδευσης είναι η ανάκτηση των αρχικών οπτικών tokens χρησιμοποιώντας τα αλλοιωμένα patches εικόνας. Μετά την προ-εκπαίδευση, οι παράμετροι του μοντέλου BeiT ρυθμίζονται λεπτομερώς στις επόμενες εργασίες προσθέτοντας στρώματα ειδικά για τις εργασίες στον προ-εκπαιδευμένο κωδικοποιητή. Τα πειραματικά αποτελέσματα καταδεικνύουν την αποτελεσματικότητα του BeiT σε εργασίες ταξινόμησης εικόνων και σημασιολογικής κατάτμησης. Για παράδειγμα, το βασικό μέγεθος BeiT επιτυγχάνει κορυφαία-1 ακρίβεια 83,2% στο ImageNet-1K, ξεπερνώντας την απόδοση της εκπαίδευσης από την αρχή με την ίδια ρύθμιση. Επιπλέον, το BeiT μεγάλου μεγέθους επιτυγχάνει εντυπωσιακή ακρίβεια 86,3% στο ImageNet-1K, ξεπερνώντας ακόμη και το ViT-L με επιβλεπόμενη προ-εκπαίδευση στο ImageNet-22K (85,2%). Το BeiT αναδεικνύει τη δυνατότητα αξιοποίησης της προ-εκπαίδευσης με μοντελοποίηση καλυμμένων εικόνων για την επίτευξη ανταγωνιστικών αποτελεσμάτων σε εργασίες όρασης [38].

CLIP (Contrastive Language-Image Pretraining)

Το CLIP είναι ένα μοντέλο μετασχηματισμού όρασης που εισήχθη από τους Radford et al. (2021) και αξιοποιεί τη δύναμη τόσο της φυσικής γλώσσας όσο και των δεδομένων εικόνας για την ταξινόμηση εικόνων. Σε αντίθεση με τις παραδοσιακές προσεγγίσεις που βασίζονται αποκλειστικά σε επισημασμένα δεδομένα εικόνας, το CLIP συνδυάζει έναν vision Transformer με έναν Transformer γλώσσας για να μάθει κοινές αναπαραστάσεις εικόνων και των σχετικών κειμενικών περιγραφών τους. Το μοντέλο εκπαιδεύεται με αντιθετικό (contrastive) τρόπο, όπου μαθαίνει να συσχετίζει παρόμοιες εικόνες και κείμενα ενώ διαχωρίζει ανόμοια ζεύγη κειμένων - εικόνων. Αυτό επιτρέπει στο CLIP να κατανοεί τις σημασιολογικές σχέσεις μεταξύ εικόνων και των αντίστοιχων κειμενικών περιγραφών τους, επιτρέποντάς του να εκτελεί εργασίες ταξινόμησης εικόνων με βάση prompts φυσικής γλώσσας. Τα πειραματικά αποτελέσματα δείχνουν ότι το CLIP επιτυγχάνει εντυπωσιακές επιδόσεις σε ένα ευρύ φάσμα σημείων αναφοράς ταξινόμησης εικόνων, ξεπερνώντας τις προηγούμενες σύγχρονες μεθόδους. Η ικανότητα του CLIP να κατανοεί εικόνες στο πλαίσιο της γλώσσας όχι μόνο αποδεικνύει την ευελιξία του, αλλά και ανοίγει νέες δυνατότητες για διατροπική (cross – modal) κατανόηση και συλλογιστική σε συστήματα τεχνητής νοημοσύνης [39].

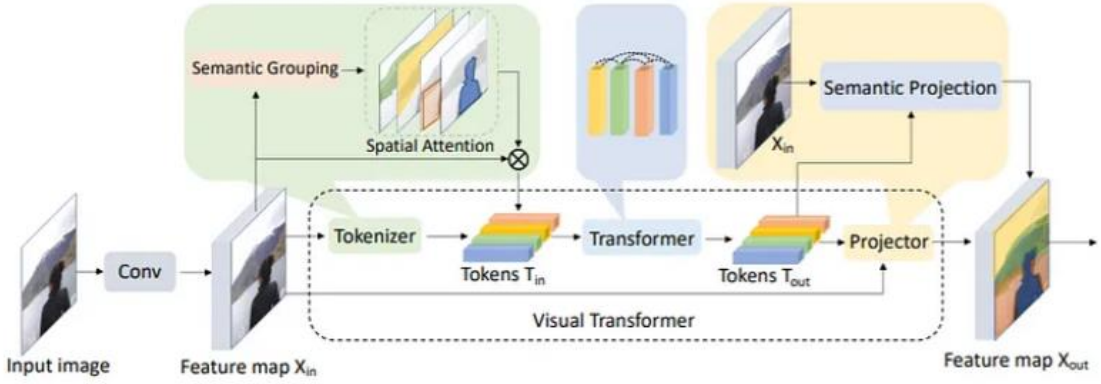
Η αναπαράσταση μέσω contrastive learning είναι πολύ θεμελιώδης στο χώρο της όρασης υπολογιστών. Αξίζει να σημειώθει ότι κατά τη χρήση contrastive learning, οι εικόνες και τα κείμενα τοποθετούνται στον ίδιο χώρο, με τρόπο παρόμοιο με αυτό που παρουσιάζεται στην ακόλουθη Εικόνα 28:



Εικόνα 28. Αναπαράσταση εικόνων και κειμένων σε έναν ενιαίο διανυσματικό χώρο με χρήση contrastive learning. [41]

2.4 Συνδυασμοί Visual Transformers και CNNs

Έχουν επίσης προταθεί αρχιτεκτονικές οι οποίες συνδυάζουν τα οφέλη των Vision transformers και των CNN. Ένα παράδειγμα μιας τέτοιας αρχιτεκτονικής δίνεται στην Εικόνα 29.



Εικόνα 29. Αρχιτεκτονική που ενσωματώνει στοιχεία Vision Transformer και CNN [40].

Οι συγγραφείς της συγκεκριμένης υλοποίησης υποστηρίζουν ότι, ενώ οι συμβατικές μέθοδοι αναπαριστούν τις εικόνες ως ομοιόμορφα τοποθετημένες συστοιχίες pixels και εφαρμόζουν συνελίξεις για να συλλάβουν έντονα εντοπισμένα χαρακτηριστικά, οι προσεγγίσεις αυτές έχουν περιορισμούς όσον αφορά την ισότιμη αντιμετώπιση όλων των pixels , τη ρητή μοντελοποίηση όλων των εννοιών σε όλες τις εικόνες και τη συσχέτιση χωρικά απομακρυσμένων εννοιών. Αντίθετα, το προτεινόμενο μοντέλο αναπαριστά τις εικόνες ως σημασιολογικές οπτικές μάρκες και χρησιμοποιεί Transformers για την πυκνή μοντελοποίηση των σχέσεων μεταξύ αυτών των μαρκών. Ο visual transformer λειτουργεί σε ένα σημασιολογικό χώρο συμβόλων και επιλεκτικά παρακολουθεί διαφορετικά μέρη της εικόνας με βάση πληροφορίες σχετικά με το πλαίσιο. Αυτή η προσέγγιση επιτρέπει στον visual transformer να εστιάζει σε σημαντικές περιοχές και να κωδικοποιεί σημασιολογικές έννοιες σε ένα συμπαγές σύνολο οπτικών συμβόλων χρησιμοποιώντας μηχανισμούς αυτοπροσοχής. Αυτές οι οπτικές μάρκες μπορούν να αξιοποιηθούν για εργασίες όπως η ταξινόμηση εικόνων ή να προβάλλονται στο χάρτη χαρακτηριστικών για σημασιολογική κατάτμηση. [40]

Η εργασία παρουσιάζει πειραματικά αποτελέσματα που αποδεικνύουν την υπεροχή της προσέγγισης με βάση το VT σε σύγκριση με τις παραδοσιακές αντίστοιχες συνελικτικές μεθόδους. Για παράδειγμα, στο σύνολο δεδομένων ImageNet, η VT επιτυγχάνει σημαντική βελτίωση της ακρίβειας του ResNet κατά 4,6 έως 7 μονάδες, ενώ χρησιμοποιεί λιγότερες πράξεις κινητής υποδιαστολής (FLOP) και παραμέτρους. Επιπλέον, σε εργασίες σημασιολογικής τμηματοποίησης στα σύνολα δεδομένων LIP και COCO-stuff, τα δίκτυα πυραμίδας χαρακτηριστικών (FPN) με βάση το VT επιτυγχάνουν υψηλότερη μέση διατομή πάνω από την ένωση (mIoU) κατά 0,35 μονάδες, ενώ μειώνουν τις FLOPs της μονάδας FPN κατά 6,5 φορές. Συνολικά, το προτεινόμενο μοντέλο αντιμετωπίζει με επιτυχία τους περιορισμούς των συμβατικών προσεγγίσεων με την ενσωμάτωση των CNN και των ViT. Ο συνδυασμός της εξαγωγής χαρακτηριστικών χαμηλού επιπέδου από τα CNN και της μοντελοποίησης εννοιών υψηλού επιπέδου από τα ViT οδηγεί σε ανώτερες επιδόσεις σε εργασίες ανάλυσης εικόνας, προσφέροντας βελτιωμένη ακρίβεια και αποδοτικότητα σε σύγκριση με τις παραδοσιακές συνελικτικές μεθόδους. Τα ευρήματα της παρούσας εργασίας συμβάλλουν στην πρόοδο των τεχνικών υπολογιστικής όρασης και ανοίγουν νέους δρόμους για περαιτέρω έρευνα στον τομέα αυτό [40].

2.5 Επεξηγησιμότητα και ερμηνευσιμότητα των Vision Transformers

Η ερμηνευσιμότητα (interpretability) και η επεξηγησιμότητα (explainability) είναι δύο κρίσιμες πτυχές των μοντέλων βαθιάς μάθησης, συμπεριλαμβανομένων των Vision Transformers (ViTs), καθώς μας επιτρέπει να αποκτηθούν πληροφορίες για τη διαδικασία λήψης αποφάσεων του μοντέλου και να κατανοηθούν οι παράγοντες που συμβάλλουν στις προβλέψεις του. Ενώ οι ViTs έχουν επιτύχει αξιοσημείωτες επιδόσεις σε διάφορες εργασίες όρασης υπολογιστών, η εγγενής πολυπλοκότητά τους καθιστά συχνά δύσκολη την ερμηνεία της εσωτερικής τους λειτουργίας. Η ερμηνευσιμότητα των ViTs μπορεί να προσεγγιστεί από διαφορετικές οπτικές γωνίες. Μια πτυχή είναι η ερμηνεία των μεμονωμένων προβλέψεων, η κατανόηση του γιατί το μοντέλο αποδίδει μια συγκεκριμένη ετικέτα κλάσης σε μια εικόνα. Έχουν προταθεί διάφορες τεχνικές για την οπτικοποίηση των χαρτών προσοχής που παράγονται από τον μηχανισμό αυτοπροσοχής στα ViTs. Οι χάρτες προσοχής αναδεικνύουν τις περιοχές της εικόνας εισόδου στις οποίες το μοντέλο δίνει προσοχή κατά την πραγματοποίηση προβλέψεων. Με την οπτικοποίηση αυτών των χαρτών προσοχής, μπορούν να αποκτηθούν πληροφορίες για τις περιοχές της εικόνας που θεωρούνται σημαντικές για την απόφαση του μοντέλου. Μια άλλη πτυχή της ερμηνευσιμότητας των ViTs είναι η ανάλυση των μαθημένων αναπαραστάσεων. Τα ViTs, όπως και άλλα μοντέλα βαθιάς μάθησης, μαθαίνουν να εξάγουν ιεραρχικές αναπαραστάσεις των δεδομένων εισόδου. Η κατανόηση της φύσης αυτών των αναπαραστάσεων μπορεί να προσφέρει πολύτιμες πληροφορίες σχετικά με το ποιες οπτικές έννοιες και χαρακτηριστικά έχει μάθει το μοντέλο. Τεχνικές όπως η οπτικοποίηση χαρακτηριστικών και η μεγιστοποίηση της ενεργοποίησης μπορούν να βοηθήσουν στην οπτικοποίηση των μοτίβων που ενεργοποιούν συγκεκριμένους νευρώνες ή στρώματα στο ViT, ρίχνοντας φως στις αναπαραστάσεις που έχουν μάθει. Επιπλέον, οι μέθοδοι απόδοσης μπορούν να χρησιμοποιηθούν για την κατανόηση της συμβολής διαφορετικών περιοχών ή σημείων στην εικόνα εισόδου στις προβλέψεις του μοντέλου. Αυτές οι μέθοδοι αποδίδουν βαθμολογίες σπουδαιότητας σε μεμονωμένα pixels ή μάρκες, υποδεικνύοντας τον αντίκτυπό τους στην τελική πρόβλεψη. Αποδίδοντας σημασία σε συγκεκριμένες περιοχές, γίνεται αντιληπτό ότι καλύτερα ποια μέρη της εικόνας επηρέασαν περισσότερο την απόφαση του μοντέλου [26].

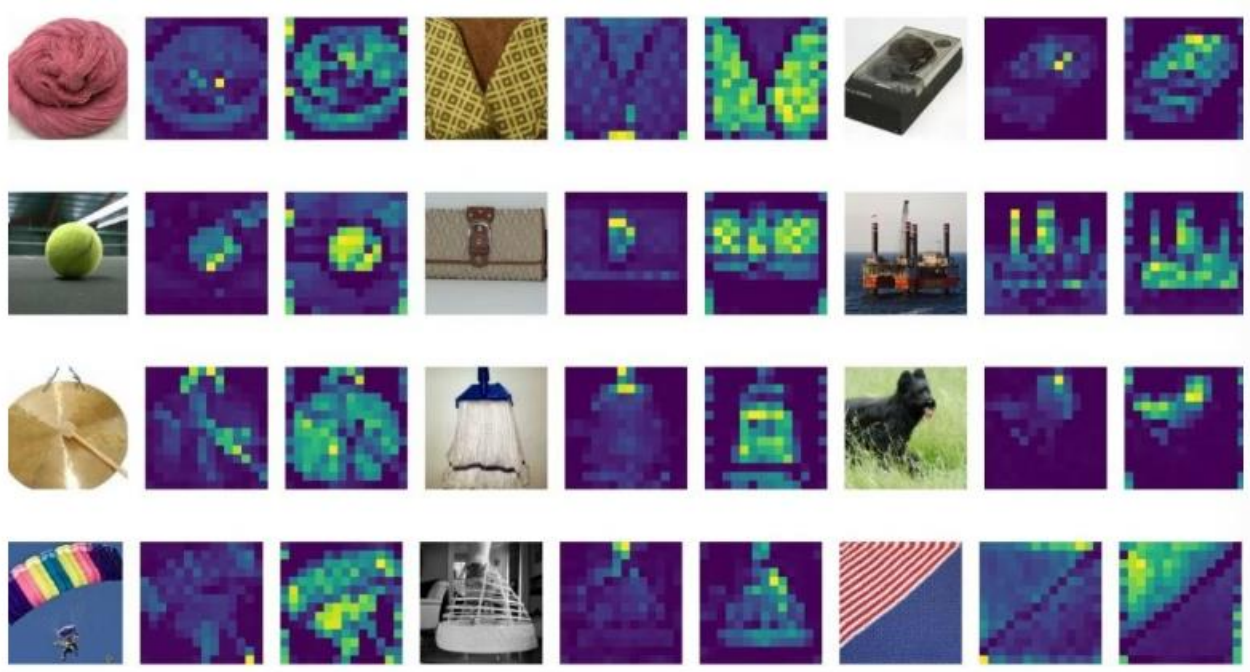
Επιπλέον, οι τεχνικές διερεύνησης (probing) μπορούν να χρησιμοποιηθούν για την ανίχνευση και την εύρεση των σημασιολογικών πληροφοριών που καταγράφονται από διαφορετικά στρώματα ή κεφαλές στο ViT. Η διερεύνηση περιλαμβάνει το σχεδιασμό συγκεκριμένων εργασιών για τον έλεγχο της κατανόησης από το μοντέλο διαφόρων οπτικών εννοιών ή σχέσεων. Αναλύοντας τις επιδόσεις του ViT σε αυτές τις δοκιμαστικές εργασίες, μπορούν να αποτυπωθούν πληροφορίες σχετικά με το επίπεδο αφαίρεσης και σημασιολογικής κατανόησης που επιτυγχάνεται από τα διάφορα συστατικά του μοντέλου. Ενώ η ερμηνευσιμότητα των ViTs αποτελεί ενεργό πεδίο έρευνας, καταβάλλονται προσπάθειες για την ανάπτυξη πιο ισχυρών και αξιόπιστων μεθόδων για την κατανόηση αυτών των μοντέλων. Η ερμηνευσιμότητα και η επεξηγησιμότητα όχι μόνο συμβάλλει στην οικοδόμηση εμπιστοσύνης στα ViTs, αλλά παρέχει επίσης δυνατότητες βελτίωσης των μοντέλων και ανάλυσης σφαλμάτων. Με την αποκάλυψη της διαδικασίας λήψης αποφάσεων των ViTs, μπορεί να ενισχυθεί η αξιοπιστία τους, να αντιμετωπιστούν πιθανές προκαταλήψεις και να εντοπιστούν τομείς για περαιτέρω έρευνα. Εν κατακλείδι, η ερμηνευσιμότητα είναι μια κρίσιμη πτυχή των Vision Transformers, η οποία επιτρέπει την κατανόηση και επεξήγηση στις προβλέψεις και τις μαθησιακές αναπαραστάσεις του μοντέλου. Τεχνικές όπως η οπτικοποίηση της προσοχής, η οπτικοποίηση των χαρακτηριστικών, οι μέθοδοι απόδοσης και οι δοκιμαστικές εργασίες συμβάλλουν στην κατανόηση του τρόπου με τον οποίο οι ViTs επεξεργάζονται και ερμηνεύουν τις οπτικές πληροφορίες. Η συνεχής έρευνα στην ερμηνευσιμότητα θα προωθήσει τη διαφάνεια, την εμπιστοσύνη και την ανάπτυξη πιο ισχυρών και εξηγήσιμων μοντέλων ViT [26].

Η επεξηγησιμότητα στο χώρο των Vision Transformers είναι ιδιαιτέρως σημαντική. Στην Εικόνα 30 παρουσιάζεται ένα παράδειγμα όπου η ταξινόμηση θα πρέπει να είναι εντελώς αξιόπιστη. Τα περισσότερα μοντέλα δίνουν απλά μια πιθανότητα στην έξοδο, από την οποία συμπεραίνεται η ταμπέλα της εικόνας. Από την άλλη, ένας γιατρός δε θα μπορούσε εύκολα να εμπιστευτεί στα τυφλά ένα τέτοιο μοντέλο για την αξιολόγηση του ασθενούς, καθώς πιθανές μεταβολές στην εικόνα (χρωματικές αλλαγές, θόρυβος κλπ) θα μπορούσαν να επηρεάσουν τον τρόπο λήψης αποφάσεων και ανάθεσης πιθανοτήτων στα labels εκ μέρους του Vision Transformer. Έτσι λοιπόν, είναι σημαντικό να αναπτυχθούν τεχνικές επεξηγησιμότητας, οι οποίες θα δίνουν μια οπτική ή άλλου τύπου εξήγηση σχετικά με την ανάθεση των συγκεκριμένων πιθανοτήτων σε κάθε κατηγορία.

Εικόνα 30. Ταξινόμηση ιατρικής εικόνας αναφορικά με την πιθανότητα μια δερματική ανωμαλία να είναι αθώα ή επικίνδυνη. [43]

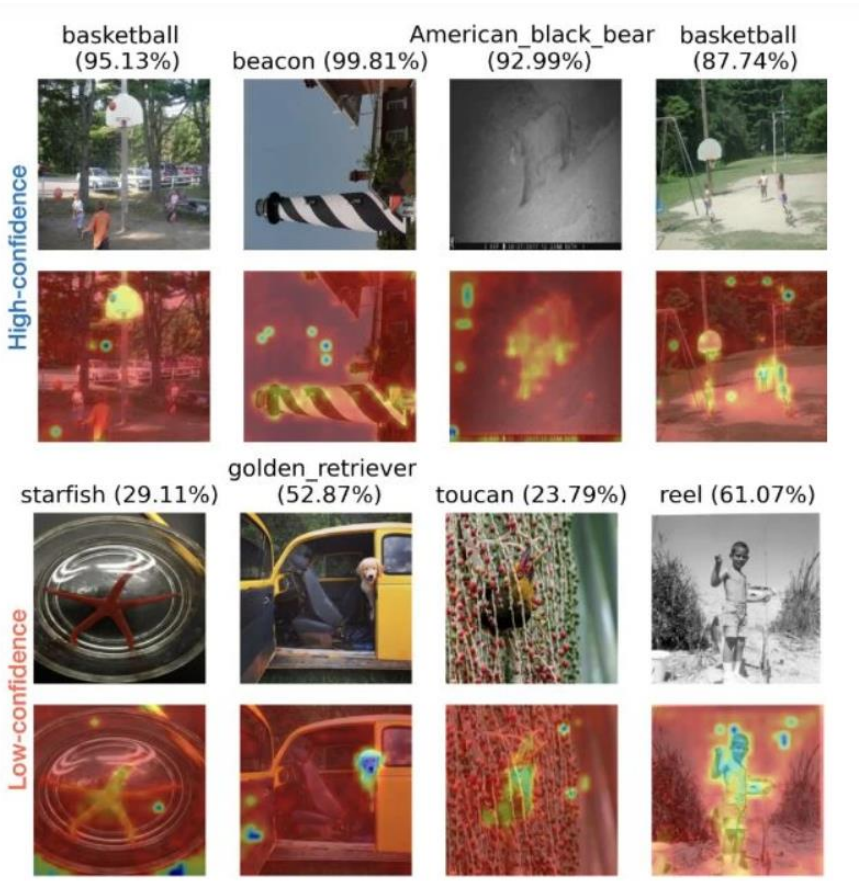
Στο σημείο αυτό, οι χάρτες προσοχής (attention maps) θα μπορούσαν να δώσουν ένα βαθμό οπτικής επεξηγησιμότητας στις αποφάσεις, ‘χρωματίζοντας’ περιοχές της εικόνας οι οποίες είναι σημαντικές για την ταξινόμηση. Πιο συγκεκριμένα, οι χάρτες προσοχής χρησιμεύουν ως πίνακες που απεικονίζουν τη σημασία των διαφόρων περιοχών μιας εικόνας εισόδου σε σχέση με διάφορα στοιχεία των μαθημένων αναπαραστάσεων του μοντέλου. Στην αρχιτεκτονική του ViT, η εικόνα εισόδου διαιρείται αρχικά σε μη επικαλυπτόμενα τμήματα, τα οποία στη συνέχεια ισοπεδώνονται και υποβάλλονται σε επεξεργασία από τον κωδικοποιητή Transformer. Αυτοί οι χάρτες προσοχής απεικονίζουν τα βάρη προσοχής που υπολογίζονται μεταξύ κάθε συμβόλου (ή patch) στην εικόνα και όλων των άλλων συμβόλων. Αξιοποιώντας έναν μηχανισμό αυτοπροσοχής, κάθε token προσέχει όλα τα άλλα tokens, με αποτέλεσμα ένα σταθμισμένο άθροισμα των αντίστοιχων αναπαραστάσεών τους. Η οπτικοποίηση των χαρτών προσοχής ως πλέγμα θερμικών χαρτών επιτρέπει την παρατήρηση των βαρών προσοχής μεταξύ συγκεκριμένων σημείων. Τα φωτεινότερα χρώματα στον χάρτη θερμότητας υποδηλώνουν υψηλότερα βάρη προσοχής, υποδηλώνοντας τη σημασία των αντίστοιχων tokens. Η ανάλυση των χαρτών προσοχής διευκολύνει την κατανόηση των περιοχών της εικόνας που έχουν τη μεγαλύτερη σημασία για το συγκεκριμένο έργο ταξινόμησης, παρέχοντας πολύτιμες πληροφορίες για τη διαδικασία λήψης αποφάσεων του μοντέλου [43].

Ένα σχετικό παράδειγμα παρουσιάζεται στην Εικόνα 31. Με κίτρινο χρώμα απεικονίζονται οι πιο σημαντικές για την ταξινόμηση περιοχές της εικόνας, σε αντίθεση με το σκούρο μπλε. Οι χάρτες προσοχής της εικόνας έχουν παραχθεί με χρήση του μοντέλου ViT-S/16 model [43].



Εικόνα 31. Χάρτες προσοχής για διάφορες εικόνες εισόδου. [43]

Συνεπώς, με χρήση των χαρτών προσοχής σαν συνοδεία της απόφασης ταξινόμησης του Vision Transformer, παρέχοντας μια επεξηγησιμότητα των αποφάσεών του. Για παράδειγμα, στην Εικόνα 32, αναπαρίσταται χάρτες προσοχής για ταξινομήσεις μεγάλης βεβαιότητας (high – confidence), εν αντιθέσει με ταξινομήσεις χαμηλης βεβαιότητας (low – confidence) [43].



Εικόνα 32. Χάρτες προσοχής για ταξινομήσεις υψηλής και χαμηλής βεβαιότητας από μοντέλο Vision Transformer [43].

Σύμφωνα με την Εικόνα 32, παρατηρείτε ότι στις περιπτώσεις ταξινόμησης υψηλής βεβαιότητας το αντικείμενο που παίζει σημαντικό ρόλο λαμβάνει υψηλό attention, με χρώματα πιο κοντά στο κόκκινο. Από την άλλη, στις ταξινομήσεις χαμηλής βεβαιότητας παρατηρούνται αρκετές μπλε περιοχές, ειδικά αν παρατηρηθεί η πιο δεξιά εικόνα, όπου το κύριο στοιχείο της συγκεκριμένης εικόνας (το παιδί) έχει χαμηλό attention, το οποίο σηματοδοτείται με μπλε χρώμα.

Φυσικά, η διάκριση μέσω των χαρτών προσοχής δεν είναι τόσο προφανής. Για το λόγο αυτό είναι σημαντικό να βελτιωθούν και να επεκταθούν οι παρούσες τεχνικές επεξηγησιμότητας των ViT για ταξινόμηση εικόνας, ιδιαίτερα όταν εμπλέκονται σε κρίσιμες εφαρμογές, όπως για παράδειγμα η ταξινόμηση ιατρικής εικόνας.

Κεφάλαιο 3ο: Άλλες τεχνικές ταξινόμησης εικόνας

Εκτός των δημοφιλέστερων αρχιτεκτονικών CNN και Vision Transformers, υπάρχουν κάποιες πιο εξειδικευμένες τεχνικές που έχουν συνεισφέρει στην ταξινόμηση εικόνας, όπως επίσης και σε άλλες συναφείς εργασίες της όρασης υπολογιστών. Στην ενότητα αυτή θα αναλύονται τέτοιες τεχνικές.

3.1 Capsule Neural Networks

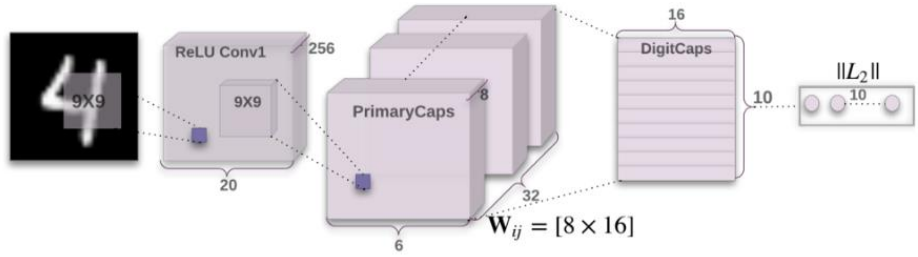
Τα Νευρωνικά Δίκτυα Κάψουλας (Capsule Neural Networks - CapsNets) είναι μια εναλλακτική αρχιτεκτονική βαθιάς μάθησης που αποσκοπεί στην αντιμετώπιση ορισμένων περιορισμών των παραδοσιακών CNN σε εργασίες ταξινόμησης εικόνων. Τα CapsNets προτάθηκαν από τον Geoffrey Hinton και τους συνεργάτες του το 2011 ως μια νέα προσέγγιση για την αποτύπωση ιεραρχικών σχέσεων μεταξύ των συστατικών της εικόνας. Ένα από τα πρωταρχικά κίνητρα εισαγωγής των CapsNets ήταν ο εντοπισμός της εγγενούς αδυναμίας του max pooling, το οποίο χρησιμοποιείται ευρέως σε αρχιτεκτονικές CNN. Πιο συγκεκριμένα, η λειτουργία max pooling στα παραδοσιακά νευρωνικά δίκτυα μπορεί να οδηγήσει σε απώλεια σημαντικών χωρικών πληροφοριών λόγω της επιλεκτικής φύσης της επιλογής μόνο των πιο ενεργών νευρώνων. Ένα τέτοιο παράδειγμα παρουσιάζεται στην Εικόνα 33. Αυτός ο περιορισμός παρακίνησε τον Geoffrey Hinton να εισαγάγει μια έννοια που ονομάζεται "δρομολόγηση με συμφωνία" (routing by agreement) για την αντιμετώπιση αυτού του ζητήματος. Αντί να απορρίπτονται πολύτιμες πληροφορίες, η διαδικασία αυτή διασφαλίζει ότι χαρακτηριστικά χαμηλότερου επιπέδου, όπως τα δάχτυλα, τα μάτια και το στόμα, δρομολογούνται επιλεκτικά σε στρώματα υψηλότερου επιπέδου που ταιριάζουν με το περιεχόμενό τους [44, 47].

Εικόνα 33. Η περιστροφή και το zoom της εικόνας του σκύλου επηρεάζει δραματικά την απόφαση ταξινόμησης ενός CNN. Όμως κάτι τέτοιο δε θα επηρέαζε την ταξινόμηση ενός CapsNet. [48]

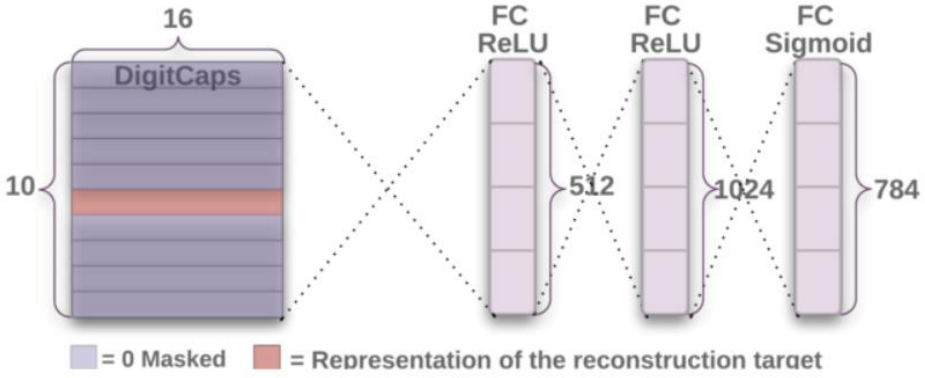
Στα CapsNets, το βασικό δομικό στοιχείο είναι μια "κάψουλα", η οποία είναι μια ομάδα νευρώνων που κωδικοποιεί διάφορες ιδιότητες μιας συγκεκριμένης οντότητας σε μια εικόνα, όπως η παρουσία της, η πόζα της και άλλα σχετικά χαρακτηριστικά. Οι κάψουλες στοχεύουν στην αναπαράσταση πιο δομημένων και κατατοπιστικών αναπαραστάσεων σε σύγκριση με τους μεμονωμένους νευρώνες στα CNNs. Ένα βασικό χαρακτηριστικό των CapsNets είναι η δυναμική δρομολόγηση (dynamic routing), η οποία επιτρέπει στις κάψουλες σε ένα επίπεδο να επικοινωνούν με κάψουλες στο επόμενο επίπεδο μέσω της διαβίβασης μηνυμάτων (message passing). Αυτός ο μηχανισμός δρομολόγησης επιτρέπει στις κάψουλες να καταλήξουν σε συναίνεση σχετικά με την έξοδό τους, συνδυάζοντας πληροφορίες από διαφορετικά μέρη της εικόνας. Δημαδή, οι κάψουλες σε χαμηλότερα επίπεδα είναι υπεύθυνες για την ενθυλάκωση συγκεκριμένων χαρακτηριστικών και ιδιοτήτων των στοιχείων εικόνας. Αυτές οι

κάψουλες επικοινωνούν με κάψουλες σε στρώματα υψηλότερου επιπέδου μέσω επαναληπτικών βημάτων συμφωνίας. Κατά τη διάρκεια αυτής της διαδικασίας, οι κάψουλες στα χαμηλότερα επίπεδα στέλνουν τις εξόδους τους σε κάψουλες υψηλότερου επιπέδου που συμφωνούν με το περιεχόμενό τους. Για παράδειγμα, εάν τα χαρακτηριστικά χαμηλότερου επιπέδου μοιάζουν με εκείνα ενός ματιού ή ενός στόματος, δρομολογούνται σε μια κάψουλα υψηλότερου επιπέδου που αντιπροσωπεύει ένα "πρόσωπο". Ομοίως, εάν τα χαρακτηριστικά περιέχουν δάχτυλα και παλάμη, κατευθύνονται σε μια κάψουλα υψηλότερου επιπέδου που αντιπροσωπεύει ένα "χέρι". Επίσης, βοηθά στην καταγραφή των χωρικών σχέσεων μεταξύ διαφορετικών οντοτήτων (entities), ενισχύοντας την ικανότητα του δικτύου να χειρίζεται τις μεταβολές της οπτικής γωνίας και τις αποκρύψεις [44, 45, 46, 47].

Η αρχιτεκτονική του CapsNet περιλαμβάνει δύο κύρια στοιχεία: έναν κωδικοποιητή (encoder) και έναν αποκωδικοποιητή (decoder), καθένα από τα οποία αποτελείται από τρία επίπεδα. Ο κωδικοποιητής παροθυσιάζεται στην Εικόνα 34, ενώ ο αποκωδικοποιητής στην Εικόνα 35 [48].



Εικόνα 34. Κωδικοποιητής ενός CapsNet [48]



Εικόνα 35. Αποκωδικοποιητής ενός CapsNet. [48]

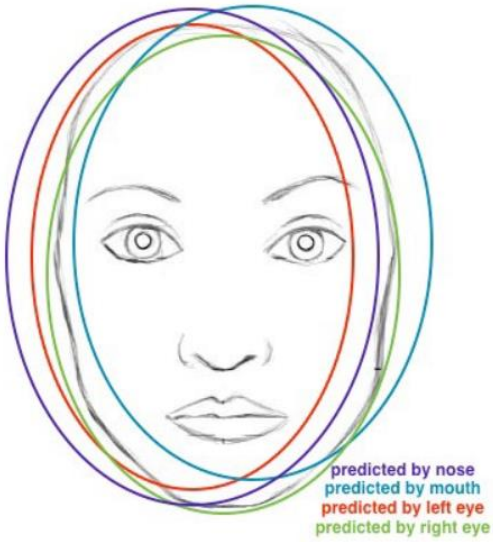
Ο κωδικοποιητής αποτελείται από ένα συνελικτικό στρώμα, ένα στρώμα PrimaryCaps και ένα στρώμα DigitCaps, ενώ ο αποκωδικοποιητής αποτελείται από τρία πλήρως συνδεδεμένα στρώματα. Στον κωδικοποιητή, η εικόνα εισόδου υποβάλλεται σε πράξεις συνελίξεων για την εξαγωγή ουσιαστικών χαρακτηριστικών. Το στρώμα συνελικτικής επεξεργασίας εφαρμόζει φίλτρα για να συλλάβει τοπικά μοτίβα και χωρικές σχέσεις εντός της εικόνας. Η έξοδος του συνελικτικού στρώματος περνά στη συνέχεια από το στρώμα PrimaryCaps, το οποίο εισάγει κάψουλες ως πρωταρχικές μονάδες αναπαράστασης πληροφοριών. Κάθε κάψουλα σε αυτό το στρώμα κωδικοποιεί την ύπαρξη και τις ιδιότητες μιας συγκεκριμένης οπτικής οντότητας. Το στρώμα DigitCaps, το τελευταίο στρώμα του κωδικοποιητή, ενσωματώνει την έννοια της δυναμικής δρομολόγησης. Επικεντρώνεται στον καθορισμό των παραμέτρων ενστάλαξης των καψουλών υψηλότερου επιπέδου με βάση τις εξόδους των καψουλών

χαμηλότερου επιπέδου. Αυτή η διαδικασία δρομολόγησης επιτρέπει το σχηματισμό ιεραρχικών σχέσεων και επιτρέπει στις κάψουλες να καταλήξουν σε συναίνεση σχετικά με τις εξόδους τους [44, 45, 48].

Το στοιχείο αποκωδικοποιητή του CapsNet λαμβάνει την έξοδο του στρώματος DigitCaps και ανακατασκευάζει την εικόνα εισόδου. Αποτελείται από πλήρως συνδεδεμένα στρώματα που αντιστοιχίζουν τις αναπαραστάσεις των κάψουλων πίσω στο χώρο των pixels , με στόχο την ανακατασκευή της αρχικής εικόνας από τα κωδικοποιημένα χαρακτηριστικά. Αυτή η διαδικασία ανακατασκευής συμβάλλει στην ενίσχυση της ευρωστίας και της ερμηνευσιμότητας του δικτύου. Η αρχιτεκτονική CapsNet, με τη δομή κωδικοποιητή-αποκωδικοποιητή και την ενσωμάτωση των καψουλών, προσφέρει μια νέα προσέγγιση στην αναπαράσταση και κατανόηση εικόνων. Αξιοποιώντας τη δυναμική δρομολόγηση και τη ρητή μοντελοποίηση των σχέσεων μεταξύ των καψουλών, το CapsNet στοχεύει στη βελτίωση των δυνατοτήτων των παραδοσιακών νευρωνικών δικτύων συνελικτικού τύπου στο χειρισμό πολύπλοκων οπτικών μοτίβων και στη διατήρηση της χωρικής πληροφορίας [44, 45, 48].

Τα νευρωνικά δίκτυα κάψουλας προσφέρουν μια ενδιαφέρουσα προσέγγιση για τη σύλληψη ιεραρχικών σχέσεων και χωρικών πληροφοριών σε εργασίες ταξινόμησης εικόνων. Αν και απαιτείται περαιτέρω έρευνα και εξερεύνηση για την πλήρη κατανόηση των δυνατοτήτων τους και την αντιμετώπιση των προκλήσεών τους, τα CapsNets παρέχουν μια εναλλακτική αρχιτεκτονική που υπόσχεται τη βελτίωση της ερμηνευσιμότητας και της ευρωστίας των συστημάτων ταξινόμησης εικόνων. Επιπλέον, έχουν δείξει ότι υπόσχονται πολλά σε διάφορες εργασίες ταξινόμησης εικόνων. Συγκεκριμένα, έχουν επιδείξει καλύτερες επιδόσεις σε σενάρια όπου οι χωρικές σχέσεις και οι μεταβολές της οπτικής γωνίας είναι ζωτικής σημασίας, όπως η αναγνώριση αντικειμένων με διαφορετικό προσανατολισμό. Τα CapsNets έχουν επίσης διερευνηθεί για εργασίες όπως η ανίχνευση αντικειμένων (object detection) και η εκτίμηση στάσης (pose estimation). Ωστόσο, τα CapsNets εξακολουθούν να αποτελούν ενεργό πεδίο έρευνας και υπάρχουν προκλήσεις που πρέπει να ξεπεραστούν. Η εκπαίδευση των CapsNets μπορεί να είναι υπολογιστικά δαπανηρή και ο αποτελεσματικός σχεδιασμός της διαδικασίας δρομολόγησης παραμένει ένα ανοιχτό και συνεχές ερευνητικό θέμα [44, 45, 46].

Τα CapsNets είναι ιδιαίτερα αποτελεσματικά στο να προβλέπουν αντικείμενα χωρίς να επηρεάζονται από μη σημασιολογικές ιδιότητες αυτών, όπως η θέση, η πόζα ή η απόκρυψη λεπτομερειών. Για παράδειγμα, στην ακόλουθη Εικόνα 36 παρουσιάζεται ένα πρόσωπο, το οποίο μπορεί να αναγνωριστεί επιτυχώς από ένα CapsNet παρά τις πιθανές παραλλαγές στις οποίες απεικονίζεται. Αυτό συμβαίνει καθώς διαφορετικές κάψουλες ανιχνεύουν σημασιολογικά χαρακτηριστικά που συνδέονται με ένα πρόσωπο, όπως για παράδειγμα η μύτη και τα μάτια [47] .



Εικόνα 36. Παραλλαγές ενός προσώπου οι οποίες δεν επηρεάζουν τις ικανότητες ταξινόμησης ενός CapsNet [48].

Κάποια σημαντικά δίκτυα CapsNets ήταν τα ακόλουθα:

CapsNet: Η αρχική αρχιτεκτονική CapsNet που προτάθηκε από τον Geoffrey Hinton στην εργασία του "Dynamic Routing Between Capsules" εισήγαγε την έννοια των καψουλών και του μηχανισμού δρομολόγησης με συμφωνία [44].

CapsNet with EM Routing: Η αρχιτεκτονική "Matrix Capsules with EM Routing" που παρουσιάστηκε από τους Sara Sabour, Nicholas Frosst και Geoffrey Hinton βελτίωσε τον μηχανισμό δρομολόγησης ενσωματώνοντας τον αλγόριθμο Expectation-Maximization (EM), ο οποίος βελτίωσε τη διαδικασία συμφωνίας και ανάθεσης μεταξύ των capsules [51].

Attention-based Capsule Networks: Τα δίκτυα κάψουλας με βάση την προσοχή, όπως προτάθηκαν από τους Zhang et al., εισήγαγαν μηχανισμούς προσοχής στις κάψουλες, επιτρέποντάς τους να εστιάζουν σε συγκεκριμένες περιοχές ενδιαφέροντος στην είσοδο. Αυτός ο μηχανισμός προσοχής βελτίωσε την ικανότητα του μοντέλου να χειρίζεται πολύπλοκες και ακατάστατες σκηνές [52].

3.1.1 Πλεονεκτήματα των CapsNets έναντι των CNNs

Παραπάνω, αναφέρθηκε το κίνητρο μετάβασης σε CapsNets αρχιτεκτονικές σε σχέση με τα κλασικά CNN. Στο σημείο αυτό, συγκεντρώνονται τα πλεονεκτήματα των CapsNets ως ακολούθως: [44, 45, 46]

Αναλλοίωτη οπτική γωνία (viewpoint invariance): Τα CapsNets αξιοποιούν τους πίνακες πόζας για να επιτρέπουν την αναγνώριση αντικειμένων ανεξάρτητα από το σημείο θέασης ή την προοπτική από την οποία παρατηρούνται. Αυτή η ικανότητα επιτρέπει στα CapsNets να χειρίζονται τις μεταβολές στον προσανατολισμό και την προβολή των αντικειμένων. Το γεγονός αυτό οφείλεται στο ότι κάθε κάψουλα είναι υπεύθυνη για την κωδικοποίηση της παρουσίας, της στάσης και άλλων ιδιοτήτων ενός αντικειμένου. Με τη χρήση πινάκων στάσης, τα CapsNets μπορούν να καταγράψουν τις χωρικές σχέσεις μεταξύ διαφορετικών τμημάτων ενός αντικειμένου και να συμπεράνουν τη συνολική δομή του, ανεξάρτητα από τον προσανατολισμό ή την οπτική του γωνία. Μέσω της διαδικασίας της "δρομολόγησης με συμφωνία", οι κάψουλες σε ανώτερα επίπεδα του δικτύου λαμβάνουν δεδομένα από κάψουλες σε χαμηλότερα επίπεδα που ταιριάζουν με το περιεχόμενό τους. Αυτή η δρομολόγηση βάσει συμφωνίας διασφαλίζει ότι χαρακτηριστικά χαμηλότερου επιπέδου, όπως δάχτυλα, μάτια ή άλλα συστατικά αντικειμένων, κατευθύνονται σε κάψουλες υψηλότερου επιπέδου που αναπαριστούν πιο σύνθετα αντικείμενα, όπως ένα πρόσωπο ή ένα χέρι. Η χρήση πινάκων πόζας επιτρέπει στα CapsNets να κωδικοποιούν πληροφορίες σχετικά με τη θέση, τον προσανατολισμό, την κλίμακα και άλλους μετασχηματισμούς ενός αντικειμένου. Λαμβάνοντας υπόψη αυτούς τους παράγοντες, τα CapsNets μπορούν να αναγνωρίζουν αντικείμενα από διάφορες οπτικές γωνίες χωρίς να βασίζονται αποκλειστικά σε συγκεκριμένες μαθαμένες οπτικές γωνίες κατά την εκπαίδευση. Αυτή η αναλλοίωτη οπτική γωνία επιτρέπει στα CapsNets να χειρίζονται περιστροφές, μεταφράσεις και άλλους μετασχηματισμούς, παρέχοντας ισχυρή αναγνώριση αντικειμένων σε διαφορετικούς προσανατολισμούς και οπτικές γωνίες.

Αποδοτικότητα παραμέτρων (parameter efficiency): Τα CapsNets απαιτούν λιγότερες παραμέτρους σε σύγκριση με τα CNN, επειδή οι κάψουλες ομαδοποιούν τους νευρώνες, με αποτέλεσμα τη μείωση των συνδέσεων μεταξύ των επιπέδων. Αυτή η αποδοτικότητα των παραμέτρων καθιστά τα CapsNets πιο αποδοτικά από υπολογιστική άποψη και λιγότερο επιρρεπή στην υπερπροσαρμογή.

Γενίκευση σε νέες οπτικές γωνίες (Generalization to New Viewpoints): Τα CNN, όταν εκπαιδεύονται στην κατανόηση περιστροφών, συχνά δυσκολεύονται να γενικεύσουν σε αθέατες οπτικές γωνίες. Αντίθετα, τα CapsNets υπερέχουν στο χειρισμό νέων σημείων θέασης λόγω της χρήσης πινάκων πόζας, οι οποίοι αποτυπώνουν τα χαρακτηριστικά των αντικειμένων ως γραμμικούς μετασχηματισμούς. Αυτή η βελτιωμένη γενίκευση επιτρέπει στα CapsNets να αναγνωρίζουν αντικείμενα από διάφορες γωνίες και προσανατολισμούς.

Ανθεκτικότητα έναντι επιθέσεων (Robustness Against Adversarial Attacks): Τα CapsNets παρουσιάζουν μεγαλύτερη ανθεκτικότητα σε αντιπολιτευτικές επιθέσεις λευκού κουτιού (white box) σε σύγκριση με τα CNN. Τεχνικές όπως η μέθοδος Fast Gradient Sign Method (FGSM) μπορούν να μειώσουν σημαντικά την ακρίβεια των CNNs, αλλά τα CapsNets διατηρούν υψηλότερα επίπεδα ακρίβειας, συχνά πάνω από 70%, ακόμη και κάτω από τέτοιες επιθέσεις. Αυτή η ανθεκτικότητα και η ευρωστία μπορεί να αποδοθεί στην ιδιαίτερη αρχιτεκτονική και τις αρχές λειτουργίας των CapsNets. Χρησιμοποιώντας αναπαραστάσεις με βάση τις κάψουλες, τα CapsNets κωδικοποιούν πληροφορίες όχι μόνο για την παρουσία αντικειμένων αλλά και για την πόζα, τον προσανατολισμό και άλλες σχετικές ιδιότητες. Αυτή η ολιστική αναπαράσταση καταγράφει πιο ‘αποχρωματισμένες’ λεπτομέρειες και σχέσεις μεταξύ των διαφόρων συστατικών ενός αντικειμένου. Η χρήση των πινάκων πόζας στα CapsNets τους επιτρέπει να μοντελοποιούν αποτελεσματικά τις χωρικές σχέσεις και να αποτυπώνουν εξαρτήσεις ανώτερης τάξης μεταξύ των διαφόρων τμημάτων ενός αντικειμένου. Αυτό καθιστά

δυσκολότερη τη διατάραξη της εσωτερικής συνέπειας και των σχέσεων που κωδικοποιούνται μέσα στα CapsNets από αντίπαλες διαταραχές. Ως αποτέλεσμα, ακόμη και αν εισαχθούν μικρές διαταραχές στην εικόνα εισόδου, οι κάψουλες του δικτύου μπορούν να συλλάβουν την εγγενή δομή και τα χαρακτηριστικά του αντικειμένου, οδηγώντας σε πιο εύρωστες και ακριβείς προβλέψεις. Η αυξημένη ανθεκτικότητα των CapsNets έναντι αντιπολιτευτικών επιθέσεων αποτελεί σημαντικό πλεονέκτημα σε εφαρμογές όπου η ασφάλεια και η αξιοπιστία είναι υψίστης σημασίας. Διατηρώντας υψηλότερη ακρίβεια και μειώνοντας την ευπάθεια σε αντίπαλες διαταραχές, τα CapsNets προσφέρουν βελτιωμένη ευρωστία και αξιοπιστία, καθιστώντας τα ελκυστικά για εργασίες όπου η ακεραιότητα των προβλέψεων του μοντέλου είναι κρίσιμη.

3.2 Neural Architecture Search - NAS

Η αναζήτηση νευρωνικής αρχιτεκτονικής (Neural Architecture Search - NAS) είναι μια τεχνική που αποσκοπεί στην αυτοματοποίηση της διαδικασίας σχεδιασμού και βελτιστοποίησης αρχιτεκτονικών βαθιάς μάθησης για την ταξινόμηση εικόνων. Περιλαμβάνει τη χρήση αλγορίθμων αναζήτησης, όπως η ενισχυτική μάθηση ή οι εξελικτικοί αλγόριθμοι, για τη διερεύνηση ενός μεγάλου χώρου αναζήτησης πιθανών αρχιτεκτονικών και την εύρεση των πιο αποτελεσματικών. Ένα από τα βασικά πλεονεκτήματα της NAS είναι ότι απαλλάσσει τους ανθρώπινους εμπειρογνώμονες από το βάρος του χειροκίνητου σχεδιασμού αρχιτεκτονικών, ο οποίος μπορεί να είναι χρονοβόρος και επιρρεπής σε σφάλματα. Με την αυτοματοποίηση της διαδικασίας, η NAS επιτρέπει την ανακάλυψη νέων και δυνητικά πιο αποτελεσματικών αρχιτεκτονικών που μπορεί να μην είχαν ληφθεί υπόψη από τους ανθρώπινους σχεδιαστές. Οι αλγόριθμοι NAS συνήθως λειτουργούν ορίζοντας έναν χώρο αναζήτησης που περιλαμβάνει διάφορες αρχιτεκτονικές επιλογές, όπως ο αριθμός των στρωμάτων, οι τύποι των στρωμάτων (συνελικτικά, pooling, κ.λπ.), τα πρότυπα συνδεσιμότητας και οι υπερπαράμετροι. Στη συνέχεια, ο αλγόριθμος αναζήτησης αξιολογεί διάφορες αρχιτεκτονικές με βάση την απόδοσή τους σε ένα σύνολο επικύρωσης και χρησιμοποιεί αυτή την ανατροφοδότηση για να κατευθύνει την αναζήτηση προς πιο υποσχόμενες αρχιτεκτονικές [49, 50].

Μια αξιοσημείωτη προσέγγιση στη NAS είναι η χρήση της ενισχυτικής μάθησης (reinforcement learning), όπου ένας πράκτορας μαθαίνει να επιλέγει διαδοχικά αρχιτεκτονικές αποφάσεις με βάση ένα σήμα ανταμοιβής που αντανακλά την απόδοση του μοντέλου που προκύπτει. Μια άλλη προσέγγιση είναι οι εξελικτικοί αλγόριθμοι (genetic algorithms), οι οποίοι εξελίσσουν επαναληπτικά έναν πληθυσμό αρχιτεκτονικών επιλέγοντας, μεταλλάσσοντας και ανασυνδυάζοντας υποψήφιες λύσεις. Η NAS έχει δείξει πολλά υποσχόμενα αποτελέσματα στην αυτόματη ανακάλυψη αρχιτεκτονικών που ξεπερνούν τις χειροκίνητα σχεδιασμένες όσον αφορά την ακρίβεια, την αποδοτικότητα ή και τα δύο. Έχει χρησιμοποιηθεί για την ανάπτυξη μοντέλων τελευταίας τεχνολογίας σε διάφορα κριτήρια αναφοράς ταξινόμησης εικόνων, συμπεριλαμβανομένου του ImageNet. Ωστόσο, η NAS μπορεί να είναι υπολογιστικά δαπανηρή λόγω του μεγάλου χώρου αναζήτησης και της ανάγκης για εκτεταμένη εκπαίδευση και αξιολόγηση πολλαπλών αρχιτεκτονικών. Μια πρόκληση στη NAS είναι η εξεύρεση ισορροπίας μεταξύ της εξερεύνησης του χώρου αναζήτησης και των απαιτούμενων υπολογιστικών πόρων. Τεχνικές όπως ο διαμοιρασμός παραμέτρων (parameter sharing), ο διαμοιρασμός βαρών (weight sharing) και οι μέθοδοι που βασίζονται στην κλίση (gradient-based methods) έχουν προταθεί για την αντιμετώπιση αυτής της πρόκλησης και την αποδοτικότερη λειτουργία της NAS. Συνολικά, η NAS προσφέρει μια συναρπαστική διέξοδο για την προώθηση της ταξινόμησης εικόνων με την αυτοματοποίηση της διαδικασίας σχεδιασμού και την ανακάλυψη αρχιτεκτονικών που ξεπερνούν τα όρια της απόδοσης και της αποδοτικότητας. Καθώς ο τομέας εξελίσσεται, οι περαιτέρω εξελίξεις στις τεχνικές NAS και στους υπολογιστικούς πόρους αναμένεται να οδηγήσουν σε ακόμη πιο σημαντικές ανακαλύψεις στην ταξινόμηση εικόνων [49, 50].

Κεφάλαιο 4ο: Εφαρμογές για άτομα με προβλήματα όρασης

Η παροχή βοήθειας σε άτομα με προβλήματα όρασης εξελίσσονταν αργά τις προηγούμενες δεκαετίες. Πριν από τον μετασχηματισμό της υγειονομικής περίθαλψης από την τεχνολογία, το κύριο βοηθητικό εργαλείο των ατόμων με προβλήματα όρασης ήταν οι μεγεθυντικοί φακοί [53]. Σήμερα, η βοήθεια σε αυτά τα άτομα περιλαμβάνει πέρα από καλύτερες συσκευές μεγέθυνσης από αυτές που χρησιμοποιούνταν στο παρελθόν, περιλαμβάνει χρήση εφαρμογών και άλλων προϊόντων που χρησιμοποιούν ηχητική ή απτική ανάδραση, αντί εκείνων που χρησιμοποιούν οπτική ανατροφοδότηση. Μέσα από αυτές τις εξελίξεις, κατέστη δυνατό τα άτομα με προβλήματα όρασης να είναι πιο ανεξάρτητα καθώς σταμάτησαν να εγκαταλείπουν δραστηριότητες όπως π.χ το διάβασμα. Πλέον με τη χρήση των πρόσφατα αναπτυγμένων εφαρμογών, όπως(Seeing AI) μπορούν να διαβάσουν [54].

4.1 Συσκευές

Διάφορες τεχνολογίες έχουν υλοποιηθεί κατά καιρούς για άτομα με προβλήματα όρασης. Μερικές από αυτές περιλαμβάνουν βελτιώσεις σε ηλεκτρονικές συσκευές, όπως tablet και Kindle, οι οποίες είναι ευκολά προσβάσιμες καθώς και αρκετά οικονομικές [54]. Αυτές οι συσκευές έχουν ενσωματωμένα χαρακτηριστικά, απαραίτητα για την παροχή βοήθειας σε άτομα με προβλήματα όρασης, συμπεριλαμβανομένης της γραμματοσειριακής διεύρυνσης, η οποία επίσης όχι μόνο βοηθά τα άτομα με προβλήματα όρασης αλλά και αυτά που αναπτύσσουν την κοινή πρεσβυωπία που συνοδεύεται με τη γήρανση. Επιπρόσθετα, η τεχνολογία υπαγόρευσης που διαθέτουν πολλές συσκευές όπως και τα smartphones με την χρήση της οποίας οι χρήστες μπορούν να χρησιμοποιήσουν πληθώρα λειτουργιών χωρίς να χρειάζεται η απτική επαφή με την οθόνη του κινητού τους τηλέφωνου.

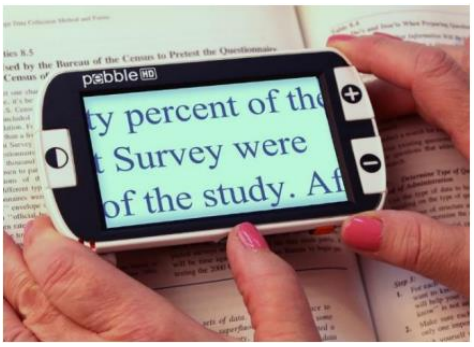
Μια άλλη εξέλιξη που βοηθά τα άτομα με προβλήματα όρασης είναι το Victor reader stream. Μια φορητή ψηφιακή συσκευή αναπαραγωγής πολυμέσων, η οποία είναι ειδικά σχεδιασμένη για τυφλούς άτομα ή άτομα με χαμηλή όραση [55]. Αυτή η συσκευή χρησιμεύει ως συσκευή ανάγνωσης ακουστικών βιβλίων, συσκευή αναπαραγωγής μουσικής και ψηφιακή συσκευή εγγραφής σε ένα. Ακούει βιβλία, παίζει μουσική και ακούει οποιαδήποτε άλλη μορφή πολυμέσων, όπως το DAISY.

Εικόνα 37: Συσκευή Victor Reader Stream [63]

Είναι επίσης χρήσιμο στην παραγωγή υψηλών ποιοτικά ηχογραφήσεων συσκέψεων ή ομιλιων, μέσω της χρήσης του ενσωματωμένου μικροφώνου τους. Μπορεί να ακούσει τις ηχογραφήσεις ή να τις αναπαράγει χρησιμοποιώντας το ενσωματωμένο ηχείο ή ένα ζευγάρι ακουστικά [55]. Το Victor Reader Stream μπορεί να συνδεθεί σε ένα υπολογιστή, μέσω του οποίου μπορεί να αναπαράγει ένα ευρύ φάσμα μορφών πολυμέσων όπως MP3, MP4 και EPUB και άλλα, και εάν συνδεθεί σε Wi-Fi, ένα άτομο μπορεί

να έχει πρόσβαση σε podcast και σε διαδικτυακούς ήχους και να κατεβάσει ασύρματα οτιδήποτε άλλο χρειάζεται μέσα σε λίγα λεπτά.

Ακόμα έχουν αναπτυχθεί μεγεθυντικοί φακοί βίντεο, οι οποίοι επιτρέπουν σε άτομα με προβλήματα όρασης να τοποθετούν έντυπα υλικά σε σαρωτή και να τα βλέπουν μεγεθυσμένα στην οθόνη[56]. Οι μεγεθυντικοί αυτοί φακοί βίντεο, μπορεί να είναι φορητοί ή μπορούν να τοποθετημένοι σε βάση, κάτι που εξυπηρετεί τα άτομα που διαβάζουν για μεγάλο χρονικό διάστημα. Οι μεγεθυντικοί φακοί που είναι τοποθετημένοι σε βάση έχουν ως επί το πλείστων λειτουργία κειμένου σε ομιλία, απαραίτητη για την ανάγνωση κειμένου δυνατά, μια δυνατότητα που ονομάζεται οπτική αναγνώριση χαρακτήρων (OCR) [56]. Μερικοί από αυτούς τους μεγεθυντικούς φακούς μπορούν επίσης να μετατρέψουν μαύρο κείμενο σε λευκό φόντο σε λευκό κείμενο μαύρο, διευκολύνοντας ορισμένους να διαβάζουν. Το δημοφιλές μοντέλο αυτών των μεγεθυντικών φακών περιλαμβάνει το pebble HD, πάντα στο χέρι, και το Merlin elite, το οποίο χρησιμοποιείται σε επιτραπέζιο υπολογιστή. Το Pebble HD αποτελείται από μια νέα κάμερα HD, που παρέχει μια καθαρή, πολύχρωμη και υψηλής ευκρίνειας εικόνα [56]. Αποτελείται από ένα νέο εργονομικό ελαφρύ compact σχεδιασμό, που το καθιστά τέλειο σύντροφο για άτομα με προβλήματα όρασης. Μπορεί να το μεταφερθεί στις τσάντες τους, στην τσέπη τους ή να το κουμπώσουν στη ζώνη τους. Μέσω του pebble HD, οι χρήστες μπορούν να διατηρήσουν την ανεξαρτησία τους και να το εκτελέσουν αποτελεσματικά τις καθημερινές τους εργασίες, όπως η ανάγνωση επιστολών, η υπογραφή εντύπων και η ανάγνωση συνταγών. Στα χαρακτηριστικά του περιλαμβάνεται η μεγέθυνση έως και 20x, ρυθμιζόμενη φωτεινότητα, πολλαπλή θέση λαβή, ένα ρολόι σε πραγματικό χρόνο και εύχρηστα μεγάλα απτικά κουμπιά [56].



Εικόνα 38: Pebble HD σε χρήση [64]

4.2 Εφαρμογές κινητών συσκευών

Παρακάτω περιγράφονται μερικές αξιόλογες εφαρμογές για κινητά που έχουν βοηθήσει άτομα με προβλήματα όρασης.

4.3 WayAround

Μια άλλη ανεπτυγμένη τεχνολογία είναι η WayAround, που έχει δημιουργηθεί για να βοηθήσει τα άτομα με προβλήματα όρασης να εκτελούν ευκολότερα τις καθημερινές τους εργασίες [57]. Αυτή είναι μια υποβοηθητική τεχνολογία, μια εφαρμογή για έξυπνες συσκευές που παρέχουν κατ' απαίτηση λεπτομέρειες για καθημερινά πράγματα. Χρησιμοποιεί μια απλή προσέγγιση tag-and-scan, η οποία επιτρέπει σε άτομα αναγνωρίζουν γρήγορα και εύκολα τα πράγματα γύρω τους. Παρέχει επίσης επιπλέον λεπτομέρειες, όπως η γνώση πώς λειτουργεί κάτι καθώς και πότε κάτι λήγει. Το WayAround κάνει ένα τα καθημερινά αντικείμενα του ατόμου να είναι μόνιμα προσβάσιμα κάθε φορά μέσω της ετικέτας και του συστήματος σάρωσης [57]. Οι χρήστες μπορούν να τοποθετήσουν τα smartphone τους σε οποιοδήποτε είδος ετικέτας και να ακούσουν οδηγίες για αυτό. Επί του παρόντος, αναπτύσσονται ετικέτες σε δημόσιους χώρους, κάτι που θα επιτρέψει άτομα με προβλήματα όρασης π.χ να πάρουν τη διάταξη των δημόσιων τουαλετών για να τους επιτρέψουν να πλοηγηθούν σε αυτές εύκολα.

4.4 Be My Eyes

Μια άλλη τεχνολογία που είναι χρήσιμη στους τυφλούς ή σε άτομα με χαμηλή όραση είναι το Be my eyes. Αυτό είναι μια δωρεάν εφαρμογή μέσω της οποίας έχουν πρόσβαση σε έναν βλέποντα εθελοντή [58]. Καθημερινά, εθελοντές με όραση δανείζουν τα μάτια τους για να λύσουν μεγάλες και μικρές εργασίες, βοηθώντας έτσι τα άτομα με προβλήματα όρασης να ζουν πιο ανεξάρτητα. Αυτά τα άτομα μπορούν να ζητήσουν βοήθεια, όπου μπορούν να πραγματοποιούν βιντεοκλήσεις και μπορούν να επικοινωνούν με τον εθελοντή για να λύσει το πρόβλημα τους [58]. Ένας βλέποντας εθελοντής μπορεί να βοηθήσει ένα άτομο μόνο εάν εγκαταστήσει την εφαρμογή Be My Eyes.

4.5 Moovit

Μια άλλη εφαρμογή είναι το Moovit, το οποίο παρέχει διαφορετικές δυνατότητες προσβασιμότητας. Η οθόνη του έχει βελτιστοποιηθεί με τεχνολογίες VoiceOver και Talkback σε συσκευές IOS και Android, οι οποίες είναι μια βελτιωμένη ενσωμάτωση προσβασιμότητας [59]. Μέσω αυτού, οι χρήστες μπορούν να χρησιμοποιούν χειρονομίες για να περιηγηθούν στα στοιχεία της οθόνης καθώς και να αλληλοεπιδράσουν με αυτά. Αφού προηγηθούν σε ένα στοιχείο, το VoiceOver διαβάζει το κείμενο που εμφανίζεται σε αυτό. Οι χρήστες λαμβάνουν βήμα προς βήμα καθοδήγηση στυλ GPS για το ταξίδι τους, ειδικά όταν χρησιμοποιούν τη λειτουργία «ζωντανές οδηγίες» του Moovit [59]. Οι χρήστες μπορούν επίσης να ειδοποιηθούν μέσω ειδοποιήσεων όταν καταφθάνει ένα μέσω μαζικής μεταφοράς για επιβίβαση καθώς και ειδοποιήσεις αποβίβασης για προετοιμασία όταν πλησιάζουν στον προορισμού τους. Η εφαρμογή Be MY Eyes που συζητήθηκε νωρίτερα εντάχθηκε στο Moovit, ενισχύοντας την προσβασιμότητα σε τυφλούς και άτομα με χαμηλά επίπεδα όρασης καθιστώντας τη δημόσια συγκοινωνία πιο προσιτή σε αυτούς όταν ταξιδεύουν [59]. Το Be My Eyes είναι προσβάσιμο απευθείας από την εφαρμογή Moovit για αυτά τα άτομα, διασφαλίζοντας την ομαλή διαδρομή όταν γίνεται χρήση των μέσων μαζικής μεταφοράς.

4.6 Seeing AI

Μια άλλη τεχνολογία που βοηθά άτομα με προβλήματα όρασης είναι η εφαρμογή τεχνητής νοημοσύνης της Microsoft, η οποία είναι μια δωρεάν εφαρμογή που διαβάζει δυνατά μικρά κομμάτια κειμένου [60]. Παρέχει στους ανθρώπους με προβλήματα όρασης έναν ευκολότερο τρόπο κατανόησης του κόσμου γύρω τους, χρησιμοποιώντας την κάμερα των smartphone τους. Αυτή η εφαρμογή κυκλοφόρησε το 2017 και από τότε οι τυφλοί και τα άτομα με χαμηλή όραση έχουν επιτύχει να ολοκληρώνουν τις καθημερινές τους εργασίες όπως ποτέ άλλοτε. Βοηθά τους χρήστες της να διαβάζουν έντυπα κείμενα, πινακίδες και χειρόγραφες σημειώσεις [60]. Τους δίνει επίσης τη δυνατότητα να αναγνωρίζουν τα τραπεζογραμμάτια και τα προϊόντα τους χρησιμοποιώντας τον γραμμωτό κώδικα(barcode) τους. Βελτιώνει τις εμπειρίες των χρηστών της με τις διάφορες δυνατότητες που περιέχει. Ένα από αυτά τα χαρακτηριστικά είναι η εξερεύνηση φωτογραφιών με την αφή, όπου αξιοποιείται η τεχνολογία από το Azure Cognitive Services, όπως η υπηρεσία Custom vision που σε συνδυασμό με άλλες, επιτρέπουν στους χρήστες να πατήσουν με τα δάχτυλά τους σε μια εικόνα στην οθόνη αφής, δίνοντάς τους έτσι τη δυνατότητα να ακούν μια περιγραφή της εικόνας [60].

4.7 VoiceOver

Μια άλλη εφαρμογή είναι το VoiceOver, το οποίο είναι ένα πρόγραμμα ανάγνωσης οθόνης που έχει ενσωματωθεί απευθείας σε κινητές συσκευές Iphone [61]. Η κύρια χρήση αυτής της εφαρμογής είναι η εκφώνηση μηνυμάτων ηλεκτρονικού ταχυδρομείου ή μηνύματα κειμένου. Χρησιμοποιεί επίσης τεχνητή νοημοσύνη για την περιγραφή εικονιδίων εφαρμογών, επίπεδο μπαταριών, και εν μέρει εικόνες. Το Apple VoiceOver περιλαμβάνει επιλογές για μεγέθυνση και έλεγχο πληκτρολογίου και λεκτικές περιγραφές στα αγγλικά, που περιγράφουν τι συμβαίνει στην οθόνη κάποιου [61]. Το πρόγραμμα ανάγνωσης οθόνης διαβάζει επίσης δυνατά το περιεχόμενο των αρχείων, για παράδειγμα, ιστοσελίδες, μηνύματα email και αρχεία επεξεργασίας κειμένου, καθώς και παρέχουν ακριβής αφήγηση του χώρου εργασίας του χρήστη [61].

Αντίστοιχα εφαρμογή είναι το TalkBack, η οποία έχει δημιουργηθεί με αντίστοιχες λειτουργίες για κινητές συσκευές Android [62].

4.8 Εικονικοί Βοηθοί

Ένας εικονικός βοηθός είναι ένα πρόγραμμα που σχεδιάστηκε για να παρέχει βοήθεια σε χρήστες σε διάφορες δραστηριότητες. Οι εικονικοί βοηθοί μπορούν να απαντήσουν σε ερωτήσεις, να παρέχουν οδηγίες, να εκτελούν εργασίες και να παρέχουν πληροφορίες στους χρήστες.

Για άτομα με προβλήματα όρασης, ένας εικονικός βοηθός μπορεί να παρέχει βοήθεια σε πληροφορίες και υπηρεσίες σε μια συσκευή με τη χρήση της φωνής. Ο εικονικός βοηθός μπορεί να εκτελέσει εντολές που δίνονται μέσω της φωνής και να παρέχει πληροφορίες για το περιεχόμενο. Ο εικονικός βοηθός μπορεί επίσης να διαβάσει τα περιεχόμενα στην οθόνη, να παρέχει οδηγίες για την πλοήγηση στον υπολογιστή και να εκτελέσει άλλες λειτουργίες που μπορούν να βοηθήσουν το άτομο με προβλήματα όρασης να χρησιμοποιήσει τον υπολογιστή με μεγαλύτερη άνεση. Επιπλέον, οι εικονικοί βοηθοί μπορούν να χρησιμοποιηθούν για την παροχή βοήθειας στην περιήγηση στο διαδίκτυο, την ανάγνωση ηλεκτρονικών μηνυμάτων, την αποστολή μηνυμάτων ηλεκτρονικού ταχυδρομείου και τη χρήση εφαρμογών. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για άτομα με προβλήματα όρασης που δυσκολεύονται να διαβάσουν κείμενο σε μια οθόνη ή να χρησιμοποιήσουν το ποντίκι και το πληκτρολόγιο.

Μερικά παραδείγματα εικονικών βοηθών που μπορούν να βοηθήσουν άτομα με προβλήματα όρασης είναι οι εικονικοί βοηθοί των λειτουργικών συστημάτων όπως ο Siri της Apple, ο Cortana της Microsoft και ο Google Assistant της Google.

Κεφάλαιο 5ο: Υλοποίηση εφαρμογής για άτομα με προβλήματα όρασης

Σε αυτό το κεφάλαιο θα περιγραφεί η διαδικασία υλοποίησης. Αρχικά, θα παρουσιαστούν τα δύο μέρη του έργου. Αυτά είναι η ενότητα προεπεξεργασίας και εκπαίδευσης του μοντέλου και η εφαρμογή για κινητά τηλέφωνα (android) που εισάγει το μοντέλο για τη ταξινόμηση των εικόνων. Όπως έχει αναφερθεί, οι πιο διαδεδομένοι αλγόριθμοι αναφορικά με την ταξινόμηση εικόνας είναι οι: CNN, VIT, CapsNets, NAS. Στην παρούσα πτυχιακή εργασία επελέγη ο αλγόριθμος CNN, καθώς αποτελεί την πιο διαδεδομένη επιλογή στον κλάδο, η υπάρχουσα βιβλιογραφία και πρακτική πληροφορία πάνω στο αντικείμενο είναι παραπάνω από επαρκείς και η περαιτέρω ανάπτυξη του θέματος θεωρείται πως θα συμβάλει στην κατανόηση εις βάθος και επακόλουθη ανάπτυξη στον τομέα της ταξινόμησης εικόνων.

5.1 Επεξήγηση δεδομένων

Το ανεπτυγμένο μοντέλο πρόβλεψης είναι ένα μοντέλο πολλαπλών ταξινομήσεων, επομένως τα δεδομένα πρέπει να έχουν τουλάχιστον δύο κατηγορίες. Όλες οι εικόνες σε αυτό το σύνολο δεδομένων έχουν άδεια χρήσης υπό την άδεια Creative Commons με αναφορά στον δημιουργό (By-Attribution License). Τα δεδομένα που χρησιμοποιήθηκαν για το έργο έχουν 5 κατηγορίες. Οι κατηγορίες είναι ‘Μαργαρίτα’, ‘Πικραλίδα’, ‘Τριαντάφυλλα’, ‘Ηλιοτρόπια’ και ‘Τουλίπες’. Παρακάτω ο Πίνακας 1, περιγράφει τον αριθμό φωτογραφιών για κάθε κατηγορία.

Πίνακας 1 : Αριθμητικά δεδομένα

Κατηγορίες	Αριθμός Εικόνων	Αριθμός Εικόνων (σε ποσοστό %)
Τριαντάφυλλα	641	15.96%
Μαργαρίτα	633	15.71%
Πικραλίδα	898	22.29%
Ηλιοτρόπιο	699	17.33%
Τουλίπες	799	19.70%

5.2 Προεπεξεργασία δεδομένων

Οι εικόνες υποβάλλονται σε προεπεξεργασία για να γίνουν έτοιμα σύνολα δεδομένων για το εκπαιδευόμενο μοντέλο. Αρχικά αποδίδονται ετικέτες για κάθε κατηγορία π.χ. ‘Μαργαρίτα’ και έπειτα υποβάλλονται σε επεξεργασία αλλαγής μεγέθους. Οι εικόνες στη συνέχεια αλλάζουν μέγεθος σε 200 pixel για διαστάσεις x και y. Αυτό είναι ένα σημαντικό βήμα για να μειώσει τον υπολογιστικό φόρτο εργασίας καθώς και να μεγιστοποιήσει την αποδοτικότητα της μάθησης. Στη συνέχεια, τα δεδομένα

ανακατεύονται, ώστε ο αλγόριθμος να μάθει να αναγνωρίζει κάθε τάξη σε ένα άρτιο ποσοστό. Αυτό είναι ένα σημαντικό βήμα γιατί αν το μοντέλο μάθαινε κάθε τάξη μία προς μία, την στιγμή που το μοντέλο θα έφτανε στην τελευταία κατηγορία, θα «ξεχνούσε» πολύτιμα συμπεράσματα. Τα δεδομένα διαχωρίζονται σε δύο μέρη "X" και "Y". Για κάθε "X" περιέχεται μια εικόνα από το σύνολο των εικόνων και το «Y» αποτελεί όλες τις αντίστοιχες ετικέτες για τις τάξεις τους. Το μέρος "X" περιέχει όλες τις εικόνες με βάση την τιμή pixel οι οποίες διαιρούνται με το 255 ώστε να κανονικοποιηθούν και να βρίσκονται εντός της τιμής 0 και 1. Χρησιμοποιώντας τη βιβλιοθήκη της γλώσσας προγραμματισμού Python 'Numpy', οι πολυδιάστατοι πίνακες μορφοποιούνται στο σωστό σχήμα, ώστε να είναι έτοιμοι για την εισαγωγή τους στο μοντέλο μάθησης.

5.3 Περιγραφή του μοντέλου μάθησης

Η υλοποίηση του μοντέλου (ο κώδικας του μοντέλου παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ Α) ενσωματώνει διάφορες αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων για την ταξινόμηση εικόνων. Οι αρχιτεκτονικές αυτές προσπαθούν να αξιολογήσουν την επίδραση του αριθμού και του μεγέθους των στρωμάτων σε σχέση με την απόδοση του μοντέλου. Η βασική αρχιτεκτονική υλοποίησης του μοντέλου χρησιμοποιεί Sequential δομή.

Το Sequential μοντέλο είναι ένας τρόπος για τη δημιουργία ενός νευρωνικού δικτύου σε Keras, και είναι ο πιο απλός τρόπος να οριστεί μια στοίβα συνελικτικών και πλήρως συνδεδεμένων στρωμάτων για ένα νευρωνικό δίκτυο. Σε ένα Sequential μοντέλο, τα στρώματα προστίθενται με τη σειρά, ένα-ένα, από την είσοδο στην έξοδο του δικτύου. Αυτό σημαίνει ότι τα δεδομένα ρέουν μέσα από το δίκτυο σε μια γραμμική σειρά. Αυτό το μοντέλο είναι κατάλληλο για πολύ απλές αρχιτεκτονικές, όπου δεν υπάρχει ανάγκη για πολλαπλά διακλαδωτικά σημεία ή κοινή χρήση στρωμάτων.

Τα είδη στρωμάτων για την υλοποίηση του μοντέλου είναι:

Συνελικτικό Στρώμα (Conv2D): Είναι το βασικό στρώμα για την ανίχνευση χαρακτηριστικών σε εικόνες. Χρησιμοποιείται για τη σάρωση της εικόνας με μια σειρά μικρών φίλτρων για την ανίχνευση συγκεκριμένων χαρακτηριστικών (όπως ακμές, γωνίες, χρώματα). Τα μεγέθη των φίλτρων και ο αριθμός των συνελικτικών φίλτρων καθορίζονται από τις παραμέτρους size_of_layer και convultional. Τα επίπεδα Conv2D συνήθως ακολουθούνται από επίπεδα ενεργοποίησης ReLU για την εισαγωγή μη γραμμικότητας.

Επίπεδο Μείωσης των Διαστάσεων (MaxPooling2D): Χρησιμοποιείται για τη μείωση των διαστάσεων των χαρακτηριστικών χωρίς απώλεια σημασίας. Τα μεγέθη των παραθύρων μείωσης των διαστάσεων καθορίζονται από την παράμετρο pool_size.

Επίπεδο Επίπεδης Καταστολής (Flatten): Το Flatten είναι ένα είδος στρώματος σε ένα νευρωνικό δίκτυο που χρησιμοποιείται για να μετατρέψει έναν πολυδιάστατο πίνακα σε έναν πυκνό (1D) πίνακα. Στην συγκεκριμένη περίπτωση, αν το είσοδο είναι ένας τρισδιάστατος πίνακας, όπως ένας πίνακας εικόνας με πλάτος, ύψος και βάθος (π.χ., πλάτος x ύψος x 3 για έναν RGB εικονοσκοπικό πίνακα), το Flatten στρώμα τον μετατρέπει σε έναν πυκνό πίνακα, όπου όλες οι τιμές του πίνακα είναι συνεχόμενες χωρίς καμία διάσταση. Το Flatten στρώμα χρησιμοποιείται συνήθως μετά από ένα σύνολο συνελικτικών και MaxPooling στρωμάτων σε ένα συνελικτικό νευρωνικό δίκτυο για να προετοιμάσει τα δεδομένα για την είσοδο σε ένα πλήρως συνδεδεμένο (Fully Connected) στρώμα. Αυτό είναι απαραίτητο, καθώς τα πλήρως συνδεδεμένα στρώματα δέχονται ως είσοδο έναν πυκνό πίνακα.

Πλήρως Συνδεδεμένο Στρώμα (Dense): Είναι πλήρως συνδεδεμένα επίπεδα, όπου κάθε νευρώνας συνδέεται με όλους τους νευρώνες του προηγούμενου στρώματος. Χρησιμοποιούνται για την εκμάθηση σχέσεων μεταξύ των χαρακτηριστικών. Το μέγεθος τους καθορίζεται από την παράμετρο size_of_layer.

Οι λόγοι για τους οποίους έγινε η χρήση των συγκεκριμένων τιμών στην δομή της αρχιτεκτονικής είναι οι εξής:

Αποδοτικότητα υπολογισμού: Στον υπολογισμό σε μηχανική μάθηση, η χρήση δυνάμεων του 2 και πολλαπλασιαστών του 8 μπορεί να είναι πολύ πιο αποδοτική από πλευράς υπολογιστικής απόδοσης. Πολλές GPU και επεξεργαστές είναι βελτιστοποιημένοι για τις συγκεκριμένες τιμές, κάτι που καθιστά την εκπαίδευση του μοντέλου πιο γρήγορη.

Βέλτιστη γεωμετρία στο νευρωνικό δίκτυο: Οι συγκεκριμένες τιμές επιλέγονται επίσης λαμβάνοντας υπόψη τη γεωμετρία του νευρωνικού δικτύου. Η χρήση δυνάμεων του 2 για τον αριθμό των νευρώνων και των φίλτρων στα συνελικτικά στρώματα είναι πιο συμβατή με την αρχιτεκτονική των συστημάτων όπου οι διαστάσεις των δεδομένων συχνά μειώνονται κατά τον υπολογισμό. Αυτό βοηθά στη διατήρηση της πληροφορίας και της αποδοτικότητας του υπολογισμού.

Συνολικά, αυτές οι τιμές είναι συχνά προτιμώμενες λόγω της βελτιστοποίησης των υπολογιστικών πόρων και της γεωμετρίας του δικτύου.

5.4 Εκπαίδευση του μοντέλου μάθησης

Κατά την προετοιμασία της εκπαίδευσης του αλγορίθμου θα οριστεί η απώλεια (loss) που θα χρησιμοποιηθεί, ο βελτιστοποιητής (optimizer) και οι μετρικές (metrics) θα χρησιμοποιηθούν κατά την εκπαίδευση του μοντέλου. Πιο αναλυτικά:

Συνάρτηση Απώλειας (Categorical_crossentropy): Αυτή είναι η συνάρτηση απώλειας που χρησιμοποιείται κατά την εκπαίδευση του μοντέλου. Σε αυτή την περίπτωση, χρησιμοποιείται η κατηγορική απώλεια (categorical crossentropy) γιατί πρόκειται για ένα πρόβλημα ταξινόμησης πολλαπλών κατηγοριών.

Βελτιστοποιητής (Adam): Ο βελτιστοποιητής που χρησιμοποιείται κατά την εκπαίδευση του μοντέλου είναι ο Adam. Ο Adam είναι ένας αποδοτικός αλγόριθμος βελτιστοποίησης που χρησιμοποιείται για την προσαρμογή των βαρών του μοντέλου με σκοπό τη μείωση της απώλειας κατά τη διάρκεια της εκπαίδευσης.

Μετρικό (Accuracy): Εδώ καθορίζονται οι μετρικές που θα χρησιμοποιηθούν για να αξιολογηθεί η απόδοση του μοντέλου κατά την εκπαίδευση. Σε αυτή την περίπτωση, χρησιμοποιείται μόνο η μετρική "ακρίβεια" (accuracy), που μετρά το ποσοστό των σωστών προβλέψεων στα δεδομένα εκπαίδευσης.

Κατά την εκπαίδευση τελικά θα οριστούν οι αναγκαίες τιμές:

X_features: Είναι το σύνολο των δεδομένων εισόδου που χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Αυτό περιλαμβάνει τις εικόνες που χρησιμοποιούνται ως χαρακτηριστικά.

Y_labels: Είναι οι ετικέτες των δεδομένων εκπαίδευσης. Κάθε ετικέτα αντιστοιχεί σε μια κατηγορία.

Αριθμός Δειγμάτων (batch_size): Αυτό το όριο καθορίζει πόσα δείγματα θα χρησιμοποιηθούν σε κάθε επανάληψη της εκπαίδευσης. Η τιμή 32 σημαίνει ότι κάθε επανάληψη θα χρησιμοποιεί ένα παρτίδα 32 δειγμάτων. Οι παρτίδες χρησιμοποιούνται για να βελτιστοποιήσουν την εκπαίδευση του μοντέλου.

Εποχές (epochs): Αυτό το όριο καθορίζει πόσες φορές θα εκπαιδευτεί το μοντέλο σε ολόκληρο το σύνολο των δεδομένων εκπαίδευσης. Σε αυτή την περίπτωση, το μοντέλο θα εκπαιδευτεί για 10 εποχές, προς τα εμπρός και προς τα πίσω μέσα στο σύνολο των δεδομένων.

Διαχωρισμός εκπαίδευσης – επαλήθευσης (validation_split): Αυτό το όριο καθορίζει το ποσοστό των δεδομένων εκπαίδευσης που θα χρησιμοποιηθεί για επαλήθευση (validation). Σε αυτή την περίπτωση, το 20% των δεδομένων εκπαίδευσης θα χρησιμοποιηθεί για επαλήθευση της απόδοσης του μοντέλου κατά την εκπαίδευση.

Κατά την εκπαίδευση, το μοντέλο προσπαθεί να μάθει τις σχέσεις μεταξύ των χαρακτηριστικών (εικόνες) και των ετικετών (κατηγορίες λουλουδιών) ώστε να μπορεί να κάνει ακριβείς προβλέψεις για νέα δεδομένα. Η διαδικασία εκπαίδευσης επαναλαμβάνεται για τον αριθμό των εποχών που καθόριζα, με στόχο τη βελτίωση της απόδοσης του μοντέλου με κάθε επανάληψη. Κατά την διάρκεια της εκπαίδευσης, το μοντέλο προσπαθεί να προσαρμοστεί στα δεδομένα εισόδου ώστε να μπορεί να ταξινομεί σωστά τις εικόνες σε μια από τις πέντε κατηγορίες λουλουδιών που έχουν οριστεί στο πρόβλημα. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο παράγει μετρικές αξιολόγησης όπως η ακρίβεια (accuracy), που μετράει πόσες από τις προβλέψεις του μοντέλου είναι σωστές σε σχέση με τις πραγματικές ετικέτες. Ο στόχος είναι να εκπαιδευτεί το μοντέλο ώστε να επιτυγχάνει υψηλή ακρίβεια στα δεδομένα εκπαίδευσης και ταυτόχρονα να γενικεύει καλά σε νέα, μη έχοντα δει πριν, δεδομένα. Μετά την εκπαίδευση, το μοντέλο μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις για νέες εικόνες λουλουδιών με βάση την εκπαιδευτική εμπειρία που απέκτησε.

5.5 Η Android εφαρμογή

Στο αυτή την ενότητα, θα εξεταστεί η υλοποίηση μιας εφαρμογής Android (ο κώδικας της εφαρμογής παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ Β) που απευθύνεται σε άτομα με προβλήματα όρασης. Η εφαρμογή ενσωματώνει το μοντέλο ταξινόμησης εικόνας που αναπτύχθηκε παραπάνω. Η εφαρμογή αποτελεί ένα χρήσιμο εργαλείο για τους χρήστες με προβλήματα όρασης, προσφέροντας λειτουργίες που τους επιτρέπουν να αναγνωρίζουν και να διαχειρίζονται εικόνες με άνεση και αυτονομία.

5.6 Δομή της εφαρμογής

Η κεντρική υλοποίηση της Android εφαρμογής βρίσκεται στο αρχείο MainActivity.java, όπου εκτελούνται όλες οι σημαντικές λειτουργίες. Σε αυτό το αρχείο συνδέονται τα στοιχεία της διεπαφής χρήστη, όπως τα TextViews, τα Buttons και τα ImageViews. Τα TextViews χρησιμοποιούνται για να εμφανίζουν πληροφορίες σχετικά με τις προβλέψεις που δημιουργούνται, όπως τις κατηγορίες και το ποσοστό πιθανότητας της πρόβλεψης. Το κουμπί χρησιμοποιείται για να τραβήξετε μια φωτογραφία χρησιμοποιώντας την κάμερα της συσκευής. Το ImageView χρησιμοποιείται για να προβάλει τη φωτογραφία που έχει ληφθεί, επιτρέποντας στον χρήστη να τη συγκρίνει με τα αποτελέσματα που έλαβε.

5.7 Εισαγωγή μοντέλου και προεπεξεργασία εικόνας στην εφαρμογή

Το κύριο μέρος της εφαρμογής επικεντρώνεται στην εισαγωγή του μοντέλου χρησιμοποιώντας τη βιβλιοθήκη TensorFlowLite για την εισαγωγή του ταξινομητή. Είναι κρίσιμο να καθοριστεί το μέγεθος της εικόνας που απαιτείται από το μοντέλο, διότι η εφαρμογή πρέπει να προεπεξεργαστεί την εικόνα που τράβηξε ο χρήστης για να ταιριάζει με την αναμενόμενη μορφή του μοντέλου. Για να ξεκινήσει, η δραστηριότητα αρχικοποιεί έναν μονοδιάστατο πίνακα ικανό να αποθηκεύσει 200x200 τιμές. Αυτός ο πίνακας λειτουργεί ως αποθηκευτικός χώρος για όλες τις πληροφορίες που αφορούν τα pixel της

εικόνας. Η διαδικασία περιλαμβάνει την εξαγωγή των τιμών RGB (Κόκκινο, Πράσινο, Μπλε) από την εικόνα και την προσθήκη αυτών των δεδομένων στον πίνακα.

5.8 Μετατροπή κειμένου σε ομιλία

Ένα σημαντικό κομμάτι της εφαρμογής αφορά την ομάδα εστίασης. Οι χρήστες με προβλήματα όρασης ενδέχεται να μην μπορούν πάντα να αναγνωρίσουν το περιεχόμενο που εμφανίζεται στην οθόνη. Για αυτόν τον λόγο, απαιτήθηκε η υλοποίηση ενός άλλου τρόπου επικοινωνίας των αποτελεσμάτων με τον χρήστη. Εδώ είναι όπου έρχεται να συμβάλει η λειτουργία Text-To-Speech της Google. Η εφαρμογή χρησιμοποιεί μια βιβλιοθήκη που μπορεί να δεχτεί ως είσοδο μια τιμή "κειμένου" και να την αναπαράγει στο ηχείο της κινητής συσκευής. Με αυτόν τον τρόπο, ο χρήστης μπορεί να ακούσει το αποτέλεσμα και να ενημερωθεί άμεσα για το περιβάλλον του.

5.9 Μελλοντικές επεκτάσεις της εφαρμογής

Μερικές μελλοντικές επεκτάσεις που θα μπορούσαν να προστεθούν σε αυτήν την εφαρμογή για να την καταστήσουν ακόμη πιο χρήσιμη και ευέλικτη είναι οι παρακάτω:

1. Αυξημένη Βάση Δεδομένων Φυτών: Αυξάνοντας την βάση δεδομένων των λουλουδιών που μπορεί να αναγνωρίσει η εφαρμογή οι χρήστες θα έχουν την δυνατότητα να αναγνωρίζουν μεγαλύτερη ποικιλία φυτών και λουλουδιών.
2. Επιπλέον Πληροφορίες: Η εφαρμογή θα μπορούσε να παρέχει πληροφορίες για κάθε λουλούδι που αναγνωρίζεται, όπως το όνομα του, την προέλευσή του, τον τρόπο φροντίδας και άλλα χρήσιμα στοιχεία.
3. Σύνδεση με Κοινότητες: Δημιουργία μιας κοινότητας χρηστών όπου μπορούν να ανταλλάσσουν εμπειρίες και συμβουλές σχετικά με την αναγνώριση και τη φροντίδα των λουλουδιών.
4. Φωνητικές Εντολές: Ενσωμάτωση φωνητικών εντολών ώστε οι χρήστες να μπορούν να χρησιμοποιούν την εφαρμογή χωρίς να χρειάζεται να αγγίζουν την οθόνη.
5. Επιπλέον Λειτουργίες Εκπαίδευσης: Δημιουργία λειτουργίες εκπαίδευσης που θα επιτρέπουν στους χρήστες να βελτιώσουν τις δεξιότητές τους στην αναγνώριση λουλουδιών.
6. Επικοινωνία με Άλλες Συσκευές: Με την σύνδεση της εφαρμογής με άλλες συσκευές, όπως φορητές κάμερες ή γυαλιά επαυξημένης πραγματικότητας, για ακόμη πιο ακριβείς αναγνώριση λουλουδιών.
7. Προσαρμογή στις Προτιμήσεις: Οι χρήστες θα πρέπει να μπορούν να προσαρμόζουν τις ρυθμίσεις της εφαρμογής στις προτιμήσεις τους, όπως την φωτεινότητα, τον ήχο και τη γραμματοσειρά.
8. Ανίχνευση Κατάστασης του Φυτού: Λειτουργίες που θα επιτρέπουν στους χρήστες να ανιχνεύουν την κατάσταση του φυτού, όπως αν χρειάζεται να το ποτίσουν ή να αλλάξουν τη θέση του.

Αυτές οι επεκτάσεις θα μπορούσαν να κάνουν την εφαρμογή ακόμη πιο ευέλικτη και χρήσιμη για τα άτομα με προβλήματα όρασης, βοηθώντας τους να απολαύσουν την φύση και την ομορφιά των λουλουδιών.

Κεφάλαιο 6ο: Συμπεράσματα και μελλοντικές επεκτάσεις

Η ταξινόμηση εικόνων έχει σημειώσει αξιοσημείωτες εξελίξεις τα τελευταία χρόνια, χάρη στην εμφάνιση ισχυρών μοντέλων βαθιάς μάθησης, όπως τα νευρωνικά δίκτυα συνελικτικής μάθησης (Convolutional Neural Networks - CNN) και οι μετασχηματιστές όρασης (Vision Transformers - ViT). Τόσο τα CNN όσο και οι ViT έχουν επιδείξει εξαιρετικές ικανότητες στην εξαγωγή ουσιαστικών χαρακτηριστικών από εικόνες και στην ακριβή ταξινόμησή τους σε διάφορες κατηγορίες.

Τα CNN, με την ιεραρχική δομή τους και τις συνελικτικές λειτουργίες τους, αποτελούν την κυρίαρχη προσέγγιση στην ταξινόμηση εικόνων εδώ και πολύ καιρό. Ξεχωρίζουν για την καταγραφή τοπικών χωρικών μοτίβων και την αξιοποίηση κοινών βαρών για την εξαγωγή σχετικών χαρακτηριστικών σε ολόκληρη την εικόνα. Τα CNN έχουν επιτύχει πρωτοποριακά αποτελέσματα σε ένα ευρύ φάσμα εργασιών αναγνώρισης εικόνας, συμπεριλαμβανομένης της αναγνώρισης αντικειμένων, της κατανόησης σκηνών και της οπτικής σημασιολογικής κατάτμησης. Η ικανότητά τους να μαθαίνουν ιεραρχικές αναπαραστάσεις και να προσαρμόζονται σε διαφορετικές οπτικές πολυπλοκότητες τα έχει καταστήσει δημοφιλή επιλογή στην όραση υπολογιστών.

Ωστόσο, η πρόσφατη εμφάνιση των Vision Transformers έφερε αλλαγή στην ταξινόμηση εικόνων. Οι ViTs αξιοποιούν τη δύναμη των μηχανισμών αυτο-προσοχής για να συλλάβουν σφαιρικές εξαρτήσεις και σχέσεις εντός μιας εικόνας. Αντιμετωπίζοντας τις εικόνες ως αλληλουχίες συμβόλων και χρησιμοποιώντας αρχιτεκτονικές μετασχηματιστών, οι ViTs έχουν επιδείξει εντυπωσιακές επιδόσεις σε εργασίες αναγνώρισης εικόνων μεγάλης κλίμακας. Μπορούν να μοντελοποιήσουν αποτελεσματικά εξαρτήσεις μεγάλης εμβέλειας, να συλλάβουν το παγκόσμιο πλαίσιο και να επιτύχουν αξιοσημείωτα αποτελέσματα σε διάφορα σύνολα δεδομένων αναφοράς. Οι ViTs προσφέρουν μια πιο ευέλικτη και κλιμακούμενη προσέγγιση στην ταξινόμηση εικόνων, με τη δυνατότητα να γενικευτούν καλά σε διάφορους οπτικούς τομείς.

Ενώ τα CNN έχουν ισχυρή βάση στην ταξινόμηση εικόνων και υπερέχουν σε σενάρια όπου η χωρική εντοπιότητα είναι ζωτικής σημασίας, τα ViT παρέχουν μια πολλά υποσχόμενη εναλλακτική λύση που δίνει έμφαση στην ολιστική κατανόηση και στις αναπαραστάσεις με επίγνωση του πλαισίου. Τα πλεονεκτήματα των CNN και των ViT είναι συμπληρωματικά και οι ερευνητές συνεχίζουν να διερευνούν τρόπους συνδυασμού των πλεονεκτημάτων τους σε υβριδικά μοντέλα για την επίτευξη ακόμη υψηλότερων επιδόσεων σε εργασίες ταξινόμησης εικόνων.

Καθώς η ταξινόμηση εικόνων συνεχίζει να εξελίσσεται, είναι προφανές ότι τα CNNs και τα ViTs αντιπροσωπεύουν δύο κυρίαρχες προσεγγίσεις που έχουν φέρει επανάσταση στον τομέα. Τα CNN έχουν θέσει τα θεμέλια και έχουν χρησιμεύσει ως σημείο αναφοράς για την ταξινόμηση εικόνων, ενώ τα ViT έχουν αναδειχθεί σε μετασχηματιστική δύναμη, διευρύνοντας τα όρια του εφικτού στην οπτική αναγνώριση. Και τα δύο μοντέλα έχουν τα μοναδικά τους πλεονεκτήματα και περιορισμούς και η καταλληλότητά τους εξαρτάται από τις συγκεκριμένες απαιτήσεις της εργασίας, τα χαρακτηριστικά του συνόλου δεδομένων και τους υπολογιστικούς περιορισμούς.

Το μέλλον της ταξινόμησης εικόνων έγκειται στην περαιτέρω ανάπτυξη των δυνατοτήτων των CNN και των ViT, καθώς και στη διερεύνηση νέων αρχιτεκτονικών που συνδυάζουν τα πλεονεκτήματά τους. Νεότερες αρχιτεκτονικές όπως τα Νευρωνικά Δίκτυα Κάψουλας (CapsNets) αναμένεται να τραβήξουν περισσότερο ερευνητικό ενδιαφέρον, καθώς επιλύουν θεμελιώδη προβλήματα συνδεδεμένα με τα δημοφιλέστερα CNN μοντέλα. Ωστόσο, υπάρχει μακρύς δρόμος στη σχετική έρευνα προκειμένου να εγκαθιδρυθεί η χρήση τους και να μπορέσουν να ενσωματωθούν σε πραγματικές εφαρμογές. Η συνεχής

έρευνα και ανάπτυξη σε αυτούς τους τομείς θα ξεκλειδώσει νέες δυνατότητες στην όραση υπολογιστών και θα οδηγήσει σε καινοτομίες στην κατανόηση εικόνων, στην αναγνώριση αντικειμένων και όχι μόνο. Η επιτυχία της ταξινόμησης εικόνων εξαρτάται από την ικανότητα αξιοποίησης των δυνατοτήτων των CNNs, των ViTs αλλά και άλλων αρχιτεκτονικών για τη δημιουργία μοντέλων που όχι μόνο επιτυγχάνουν υψηλή ακρίβεια αλλά και επιδεικνύουν ευρωστία, ερμηνευσιμότητα και προσαρμοστικότητα σε πραγματικές συνθήκες.

Τέλος η εφαρμογή που αναπτύχθηκε θα μπορούσε είτε να αποτελέσει έναν οδηγό διαχείρισης, αναγνώρισης και φροντίδας των λουλουδιών ενσωματώνοντας τις επεκτάσεις που αναφέρθηκαν στο Κεφάλαιο 5 ή θα μπορούσε να εμπλουτιστεί με μεγαλύτερα σύνολα δεδομένων με διάφορες εικόνες, πολλών κατηγοριών και να αποτελέσει ένα μεγαλύτερου βελτηνεκούς οδηγό για τα άτομα με προβλήματα όρασης. Αυτές οι εφαρμογές δεν βελτιώνουν μόνο την καθημερινή ζωή των χρηστών τους, αλλά και προάγουν την πρόσβαση και τη συμμετοχή σε όλες τις πτυχές της κοινωνίας. Έτσι με τη σωστή σχεδίαση και ανάπτυξη, αυτές οι εφαρμογές θα μπορούν να εξαλείψουν το εμπόδιο της ανεπαρκούς πρόσβασης σε πληροφορίες και υπηρεσίες, προσφέροντας ίσες ευκαιρίες και βελτιώνοντας την ποιότητα ζωής για όλους αλλά και θα δημιουργήσουν μια κοινωνία ισότητας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] O'Shea, K. & Nash, R. (2015). An Introduction to Convolutional Neural Networks.. CoRR, abs/1511.08458.

[2] Medium, "Convolutional Neural Networks, Explained" [Online]. Available: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

[3] Bengio, Y. (2016). Deep Learning. MIT Press.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[8] AI Smartz, "CNN Architectures Over a Timeline (1998-2019)" [Online]. Available: https://www.aismartz.com/blog/cnn-architectures/

[9] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan D., Vanhoucke V. & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[11] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[12] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[13] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[14] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Łukasz. Kaiser, and I. Polosukhin (2017). Attention is all you need. Advances in Neural Information Processing Systems , page 5998--6008.

[16] Dosovitskiy, A., et al. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[17] Zheng, H., et al. (2021). Vision Permutator: A Permutable MLP-Like Architecture for Visual Recognition. arXiv preprint arXiv:2103.16302.

[18] Touvron, H., et al. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

[19] Carion, N., et al. (2020). End-to-End Object Detection with Transformers. arXiv preprint arXiv:2005.12872.

[20] Wu, Z., et al. (2021). Training Vision Transformers with ImageNet Labels. arXiv preprint arXiv:2103.15358.

[21] Rumelhart, David E; Hinton, Geoffrey E, and Williams, Ronald J (Sept. 1985). Learning internal representations by error propagation. Tech. rep. ICS 8504. San Diego, California: Institute for Cognitive Science, University of California.

[22] Hochreiter, S., & Schmidhuber, Jurgen. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[23] Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[25] Machine Learning Mastery "The Transformer Model". [Online]. Available: https://machinelearningmastery.com/the-transformer-model/

[26] Khan S., Naseer M., Hayat M., Waqas Zamir S., Shahbaz Khan F., and Shah M. (2021). Transformers in Vision: A Survey. arXiv:2006.03677v4 https://arxiv.org/pdf/2101.01169.pdf

[27] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 163–172).

[28] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0

[29] Medium. ViT: Vision Transformer . [Online]. Available: https://medium.com/machine-intelligence-and-deep-learning-lab/vit-vision-transformer-cc56c8071a20

[30] Huggingface. Vision Transformer (ViT) [Online]. Available: https://huggingface.co/docs/transformers/model_doc/vit

[31] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255). IEEE.

[32] Han, K., Wang, J., & Han, J. (2021). Transformer in transformer. arXiv preprint arXiv:2103.00112.

[33] Xiangxiang Chu, , Zhi Tian, , Bo Zhang, Xinlong Wang, , Chunhua Shen. (2021) Conditional Positional encodings for Vision Transformers

[34] Touvron, H., Vedaldi, A., Douze, M., Cord, M., & Auffray, C. (2021). Going deeper with image transformers. arXiv preprint arXiv:2103.17239.

[35] Xu W., Xu Y., Chang T., Tu Z. (2021) Co-Scale Conv-Attentional Image Transformers arXiv:2104.06399 https://arxiv.org/abs/2104.06399

[36] Jiang, Y., Hou, Q., Wu, M., Xie, L., & Liang, X. (2021). Linear vision transformers. arXiv preprint arXiv:2106.00699.

[37] Liu Z., LinY., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 https://arxiv.org/abs/2103.14030

[38] Bao H., Dong L., Piao S., Wei F. (2021). BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 https://arxiv.org/abs/2106.08254

[39] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

[40] Wu B., Xu C., Dai X., Wan A., Zhang P., Yan Z., Tomizuka M., Gonzalez J., Keutzer K., Vajda P. (2020) Visual Transformers: Token-based Image Representation and Processing for Computer Vision. arXiv:2006.03677v4 https://arxiv.org/pdf/2006.03677.pdf

[41] Medium. What Are Vision Transformers And How Are They Important For General Purpose Learning? [Online]. Available: https://towardsdatascience.com/what-are-vision-transformers-and-how-are-they-important-for-general-purpose-learning-edd008545e9e

[42] Medium. Image Classification with Vision Transformer. [Online]. Available: https://towardsdatascience.com/image-classification-with-vision-transformer-8bfde8e541d4

[43] Viso.ai. Vision Transformers (ViT) in Image Recognition – 2023 Guide. [Online]. Available: https://viso.ai/deep-learning/vision-transformer-vit/

[44] Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in Neural Information Processing Systems (pp. 3856-3866).

[45] Hinton, G. E., & Salakhutdinov, R. R. (2011). Transforming auto-encoders. In International Conference on Artificial Neural Networks (pp. 44-51).

[46] Zhao, W., Wang, Y., Zhao, Y., Liu, R., & Zhang, S. (2018). Capsule network performance on complex data. Neural Computing and Applications, 30(9), 2723-2735.

[47] Medium. Capsule Networks: The New Deep Learning Network [Online]. Available: https://towardsdatascience.com/capsule-networks-the-new-deep-learning-network-bd917e6818e8

[48] Paperspace. Capsule Networks: A Quick Primer [Online]. Available: https://blog.paperspace.com/capsule-networks/

[49] Zoph, B., & Le, Q. V. (2017). Neural Architecture Search with Reinforcement Learning. In Proceedings of the 34th International Conference on Machine Learning (ICML) (Vol. 70, pp. 2973–2982). http://proceedings.mlr.press/v70/zoph17a.html

[50] Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized Evolution for Image Classifier Architecture Search. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)(pp.405–415). https://proceedings.neurips.cc/paper/2019/file/88e55c6c3d31f8af85d25bb4a70f740b-Paper.pdf

[51] Google Blog. Matrix capsules with EM routing. [Online]. Available: https://research.google/pubs/pub46653/

[52] Zhang N., Deng S., Sun Z., Xi X., Zhang W., Chen H. (2018) Attention-Based Capsule Networks with Dynamic Routing for Relation Extraction. arXiv:1812.1132 https://arxiv.org/abs/1812.11321

[53] Senjam, S. S., Foster, A., & Bascaran, C. (2021). Assistive technology for visual impairment and trainers at schools for the blind in Delhi. Assistive Technology, 1-5. https://doi.org/10.1080/10400435.2020.1839144

[54] Bhowmick, A., & Hazarika, S. M. (2017). An insight into assistive technology for the visually impaired and blind people: State-of-the-art and future trends. Journal on Multimodal User Interfaces, 11(2), 149-172. https://doi.org/10.1007/s12193-016-0235-6

[55] Yadav, A. B., Bindal, L., Namhakumar, V. U., Namitha, K., & Harsha, H. (2016). Design and development of smart assistive device for visually impaired people. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). https://doi.org/10.1109/rteict.2016.7808083

[56] Li, Z., Pundlik, S., & Luo, G. (2013). Stabilization of magnified videos on a mobile device for visually impaired. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/cvprw.2013.15

[57] Sayal, R. (2020). Mobile app accessibility for visually impaired. International Journal ofAdvanced Trends in Computer Science and Engineering, 9(1), 182-185. https://doi.org/10.30534/ijatcse/2020/27912020

[58] Griffin-Shirley, N., Banda, D. R., Ajuwon, P. M., Cheon, J., Lee, J., Park, H. R., & Lyngdoh, S. N. (2017). A survey on the use of mobile applications for people who are visually impaired. Journal of Visual Impairment & Blindness, 111(4), 307-323. https://doi.org/10.1177/0145482x1711100402

[59] Chottin, M. (2017). Mobility technologies for visually impaired people through the prism of classic theories of perception. Mobility of Visually Impaired People, 77-108. https://doi.org/10.1007/978-3-319-54446-5_3

[60] Granquist, C., Sun, S. Y., Montezuma, S. R., Tran, T. M., Gage, R., & Legge, G. E. (2021). Evaluation and comparison of artificial intelligence vision aids: Orcam MyEye 1 and seeing AI. Journal of Visual Impairment & Blindness, 115(4), 277-285. https://doi.org/10.1177/0145482x211027492

[61] Smaradottir, B. F., Håland, J. A., & Martinez, S. G. (2018). User evaluation of the smartphone screen reader VoiceOver with visually disabled participants. Mobile Information Systems, 2018, 1-9. https://doi.org/10.1155/2018/6941631

[62] Venugopal, G. (2015). A review of popular applications on Google play-do they cater to visually impaired users? Indian Journal of Science and Technology, 8(S4), 221. https://doi.org/10.17485/ijst/2015/v8is4/61436

[63] https://store.humanware.com/hus/victor-reader-stream-handheld-media-player.html

[64] https://www.palmervision.com/product/pebble-hd-4-3-digital-handheld-magnifier/

ΠΑΡΑΡΤΗΜΑ Α : ΚΩΔΙΚΑΣ ΜΟΝΤΕΛΟΥ ΤΑΞΙΝΟΜΗΣΗΣ ΕΙΚΟΝΑΣ

```
import numpy as np # πολυσδιάστατοι πίνακες
import matplotlib.pyplot as plt # σχεδίαση δεδομένων και αποτελεσμάτων
import os # χειρισμός αρχείου - διαδρομής
import cv2 # μετατροπή εικόνων σε δεδομένα
import random # τυχαιοποίηση συστοιχίας δεδομένων
from pathlib import Path # χειραγώγηση διαδρομής
import seaborn as sns # οπτικοποίηση δεδομένων
from tensorflow.keras.models import Sequential # είδος δημιουργίας μοντέλου
from tensorflow.keras.layers import Dense, Dropout, Activation, Flatten, Conv2D, MaxPooling2D # διαφορετικά στρώματα
from tensorflow.keras.callbacks import TensorBoard
from tensorflow.keras.utils import to_categorical
import tensorflow as tf # tensorflow βιβλιοθήκη

path_to_data = str(Path('').parent.resolve())+"\\Data\\flower_photos"

flowerTypes = ["daisy", "dandelion","roses","sunflowers","tulips"]
image_size = 200
data_for_training = []
for flowerType in flowerTypes:
    path = os.path.join(path_to_data, flowerType)
    flower_label = flowerTypes.index(flowerType)
    for img in os.listdir(path):
        try:
            image_data = cv2.imread(os.path.join(path,img))
            resized_image = cv2.resize(image_data, (image_size, image_size))
            data_for_training.append([resized_image,flower_label])
        except Exception as e:
            pass
random.shuffle(data_for_training)
X_features=[]
```

```
Y_labels=[]
for features, label in data_for_training:
    X_features.append(features)
    Y_labels.append(label)
X_features = np.array(X_features).reshape(-1, image_size, image_size, 3)
X_features = X_features/255.0
X_features = np.array(X_features)
Y_labels = np.array(Y_labels)
Y_labels = to_categorical(Y_labels,5)

for dense in [0, 1, 2]:
    for size_of_layer in [32, 64, 128]:
        for convultional in [1, 2, 3]:
            deep_learning_model = Sequential()
            deep_learning_model.add(Conv2D(size_of_layer, (3, 3), input_shape=X_features.shape[1:]))
            deep_learning_model.add(Activation('relu'))
            deep_learning_model.add(MaxPooling2D(pool_size=(2, 2)))
            for l in range(convultional-1):
                deep_learning_model.add(Conv2D(size_of_layer, (3, 3)))
                deep_learning_model.add(Activation('relu'))
                deep_learning_model.add(MaxPooling2D(pool_size=(2, 2)))
            deep_learning_model.add(Flatten())
            for _ in range(dense):
                deep_learning_model.add(Dense(size_of_layer))
                deep_learning_model.add(Activation('relu'))
            deep_learning_model.add(Dense(5))
            deep_learning_model.add(Activation('softmax'))
deep_learning_model.compile(loss='categorical_crossentropy',optimizer='adam',metrics=['accuracy'])
deep_learning_model.fit(X_features, Y_labels, batch_size=32, epochs=10, validation_split=0.2)
deep_learning_model.save('deep_learning_modelWithBatchSize32AndEpoch10')
converter =
tf.lite.TFLiteConverter.from_saved_model('deep_learning_modelWithBatchSize32AndEpoch10')  #
path to the SavedModel directory
tflite_model = converter.convert()
```

```
with open('deep_learning_modelWithBatchSize32AndEpoch10.tflite', 'wb') as f:
    f.write(tflite_model)
```

ΠΑΡΑΡΤΗΜΑ Β : ΚΩΔΙΚΑΣ ΕΦΑΡΜΟΓΗΣ ANDROID

```
ΠΑΡΑΡΤΗΜΑ C /*
 * Created by Kostas Sarantavgas
 *
 * The following code is responsible for initiating the application and
 * creating the necessary tools to be able to take a picture, show it
 * and using an imported model, validate it to produce predictions
 * on what the image is showing.
*/

package app;


/**
 * Αυτές είναι οι απαραίτητες βιβλιοθήκες και τα εργαλεία που χρειάζονται
για την
 * υλοποίηση σχεδίου.
 */

// Εργαλεία Android X που βοηθούν στη δημιουργία και την αλληλεπίδραση
διεπαφής χρήστη
import androidx.annotation.Nullable;
import androidx.appcompat.app.AppCompatActivity;
import android.Manifest;
import android.content.Intent;
import android.content.pm.PackageManager;
import android.graphics.Bitmap;
import android.media.ThumbnailUtils;
import android.os.Bundle;
import android.provider.MediaStore;
import android.speech.tts.TextToSpeech;
import android.view.View;
import android.widget.Button;
import android.widget.ImageView;
import android.widget.TextView;

// Βιβλιοθήκη Tensorflow που θα χρησιμοποιηθεί για την εισαγωγή και τη
χρήση του μοντέλου για την πραγματοποίηση προβλέψεων
import org.tensorflow.lite.DataType;
import org.tensorflow.lite.support.tensorbuffer.TensorBuffer;

// Βιβλιοθήκες Java για χειρισμό αρχείων
import java.io.IOException;
import java.nio.ByteBuffer;
import java.nio.ByteOrder;
import java.util.Locale;

// Εισαγάγετε το μοντέλο tflite που παράγεται από τον αλγόριθμο Python
import app.ml.Model;

/**
 * Κύρια κλάση του έργου που είναι υπεύθυνη για τη ρύθμιση του UI,
 * η λήψη εικόνας και η εισαγωγή του μοντέλου σε
 * Κάντε ακριβείς προβλέψεις με βάση την εικόνα που τραβήξατε.
 */
public class MainActivity extends AppCompatActivity {

    // UI components objects
    TextView final_result, list_of_confidences; // πεδία κειμένου για
```

```
εμφάνιση αποτελεσμάτων
    ImageView picture; // η φωτογραφία που τραβήχτηκε
    Button takePictureBtn; //κουμπί για να ξεκινήσει ο μηχανισμός λήψης
φωτογραφιών
    TextToSpeech speaker;
    TextToSpeech speaker2;


    int imageSize = 200;

    @Override
    protected void onCreate(Bundle savedInstanceState) {
        super.onCreate(savedInstanceState);
        setContentView(R.layout.activity_main);

        // init των στοιχείων διεπαφής χρήστη, συνδεθείτε με τους
κατάλληλους πόρους
        final_result = findViewById(R.id.final_result);
        list_of_confidences = findViewById(R.id.list_of_confidences);
        picture = findViewById(R.id.picture);
        takePictureBtn = findViewById(R.id.takePictureBtn);

        // αρχικοποίηση ομιλητή κειμένου σε ομιλία
        speaker=new TextToSpeech(getApplicationContext(), new
TextToSpeech.OnInitListener() {
            @Override
            public void onInit(int status) {
                if(status != TextToSpeech.ERROR) {
                    speaker.setLanguage(Locale.UK);
                }
            }
        });

        speaker2=new TextToSpeech(getApplicationContext(), new
TextToSpeech.OnInitListener() {
            @Override
            public void onInit(int status) {
                if(status != TextToSpeech.ERROR) {
                    speaker2.setLanguage(Locale.UK);
                }
            }
        });


        // ορίστε τη λειτουργικότητα του κουμπιού "Take Picture".
        takePictureBtn.setOnClickListener(new View.OnClickListener() {
            @Override
            public void onClick(View view) {
                // Η κάμερα θα πρέπει να εκκινείται μόνο εάν παραχωρηθεί
άδεια πρόσβασης σε αυτήν
                if ( checkSelfPermission(Manifest.permission.CAMERA) ==
PackageManager.PERMISSION_GRANTED ) {
                    // create intent to be used for initializing picture-
taking-screen
                    Intent cameraIntent = new
Intent(MediaStore.ACTION_IMAGE_CAPTURE);
                    startActivityForResult(cameraIntent, 1);
                } else {
                    // εάν όχι, τότε ζητήστε την άδεια
                    requestPermissions(new
String[]{Manifest.permission.CAMERA}, 100);
```

```
                }
            }
        });
    }
    /**
     * Μέθοδος υπεύθυνη για την εισαγωγή του μοντέλου tensorflow και τη δημιουργία ενός buffer για την είσοδο εικόνας που είναι η φωτογραφία που έχει ληφθεί.
     * Επιπλέον, θα δημιουργήσει μια σειρά από pixel που αντιπροσωπεύουν την εικόνα που τραβήχτηκε.
     * Το μοντέλο θα λάβει στη συνέχεια ως είσοδο τη διάταξη των pixel και θα παράγει προβλέψεις ταξινόμησης.
     * Εικόνα @param
     */
    public void classifyImage(Bitmap image){
        try {
            //Δημιουργία αντικειμένου μοντέλου με βάση το μοντέλο tensorflow lite
            Model model = Model.newInstance(getApplicationContext());

            //Δημιουργεί εισόδους για αναφορά.
            TensorBuffer inputFeature0 = TensorBuffer.createFixedSize(new int[]{1, 200, 200, 3}, DataType.FLOAT32);
            ByteBuffer byteBuffer = ByteBuffer.allocateDirect(4 * imageSize * imageSize * 3);
            byteBuffer.order(ByteOrder.nativeOrder());

            // λάβετε 1D πίνακα 224 * 224 pixel στην εικόνα
            int [] intValues = new int[imageSize * imageSize];
            image.getPixels(intValues, 0, image.getWidth(), 0, 0, image.getWidth(), image.getHeight());

            // επανάληψη σε εικονοστοιχεία και εξαγωγή τιμών R, G και B. Προσθήκη στο bytebuffer.
            int pixel = 0;
            for(int i = 0; i < imageSize; i++){
                for(int j = 0; j < imageSize; j++){
                    int val = intValues[pixel++]; // RGB
                    byteBuffer.putFloat(((val >> 16) & 0xFF) * (1.f / 255.f));
                    byteBuffer.putFloat(((val >> 8) & 0xFF) * (1.f / 255.f));
                    byteBuffer.putFloat((val & 0xFF) * (1.f / 255.f));
                }
            }

            // φορτώστε το buffer με τα byte
            inputFeature0.loadBuffer(byteBuffer);

            // Εκτελεί το συμπέρασμα μοντέλου και λαμβάνει αποτέλεσμα.
            Model.Outputs outputs = model.process(inputFeature0);
            TensorBuffer outputFeature0 = outputs.getOutputFeature0AsTensorBuffer();

            // η λίστα των παραγόμενων εμπιστευτικών στοιχείων για κάθε τάξη
            float[] confidences = outputFeature0.getFloatArray();
            // βρείτε το ευρετήριο της τάξης με τη μεγαλύτερη εμπιστοσύνη.
            int maxPos = 0;
            float maxConfidence = 0;
            for(int i = 0; i < confidences.length; i++){
```

```
                if(confidences[i] > maxConfidence){
                    maxConfidence = confidences[i];
                    maxPos = i;
                }
            }

            // Τύποι λουλουδιών - ετικέτες που θα χρησιμοποιηθούν για την
ταξινόμηση της ετικέτας
            String[] classes = {"DAISY", "DANDELION",
"ROSES","SUNFLOWERS","TULIPS"};

            // Πάρτε την πρόβλεψη με την υψηλότερη σιγουριά και εμφανίστε
την στην οθόνη
            final_result.setText(classes[maxPos]);
            // καθώς και να το παίξετε
            speaker.speak("The flower is:" + classes[maxPos],
TextToSpeech.QUEUE_FLUSH, null);

            // για τις υπόλοιπες αξιοπιστίες, παράγετε %
            String s = "";
            for(int i = 0; i < classes.length; i++){
                s += String.format("%s: %.1f%%\n", classes[i],
confidences[i] * 100);
                speaker2.speak( s , TextToSpeech.QUEUE_FLUSH,null);
            }
            // προσθέστε τις υπόλοιπες εμπιστεύσεις από κάτω
            list_of_confidences.setText(s);


            // Απελευθερώνει πόρους μοντέλου εάν δεν χρησιμοποιούνται
πλέον.
            model.close();
        } catch (IOException e) {
            // TODO Handle the exception
        }
    }


    /**
     * Μέθοδος υπεύθυνη για τα επακόλουθα της εκτέλεσης του μηχανισμού
λήψης φωτογραφιών για τη λήψη της φωτογραφίας,
     * Προεπεξεργαστείτε το για το μοντέλο νευρωνικού δικτύου και, στη
συνέχεια, δώστε το στο μοντέλο για να παράγει αποτελέσματα.
     * Η μικρογραφία δημιουργείται για να εμφανίζεται στον χρήστη ποια
εικόνα χρησιμοποιείται.
     *
     * @param requestCode
     * @param resultCode
     * Δεδομένα @param
     */
    @Override
    public void onActivityResult(int requestCode, int resultCode, @Nullable
Intent data) {
        if (requestCode == 1 && resultCode == RESULT_OK) {
            Bitmap image = (Bitmap) data.getExtras().get("data");
            int dimension = Math.min(image.getWidth(), image.getHeight());
            image = ThumbnailUtils.extractThumbnail(image, dimension,
dimension);
            picture.setImageBitmap(image);

            image = Bitmap.createScaledBitmap(image, imageSize, imageSize,
```

```
false);
            classifyImage(image);
        }
        super.onActivityResult(requestCode, resultCode, data);
    }
}
```