Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Exploring the Interpretability of Vision Transformers: Applications in Medical Imaging

Diploma Thesis

by

Varvara - Konstantina Mangelaki

Επιβλέπων/Επιβλέπουσα: Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Exploring the Interpretability of Vision Transformers: Applications in Medical Imaging

Diploma Thesis

by

Varvara - Konstantina Mangelaki

Επιβλέπων/Επιβλέπουσα: Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1η Νοεμβρίου, 2024.

.....	.....	.....
Αθανάσιος Βουλόδημος	Γεώργιος Στάμου	Ανδρέας - Γεώργιος Σταφυλοπάτης
Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2024

.....
Βαρβαρα - Κωνσταντινα Μαγγελακη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Για αρκετά χρόνια τώρα, οι προσπάθειες επέκτασης της ψηφιακής επεξεργασίας εικόνας σε αλγορίθμους ανάλυσης και κατανόησης τους έχουν καθορίσει σε μεγάλο βαθμό την πορεία της τεχνητής νοημοσύνης. Η ανάπτυξη προηγμένων μοντέλων βαθιάς μάθησης έχει επιτρέψει την επιτυχή ανάλυση και κατανόηση πολύπλοκων εικόνων σε ποικίλες εφαρμογές, από την αυτόματη αναγνώριση αντικειμένων της καθημερινότητας μέχρι την ιατρική διάγνωση. Η χρήση της τεχνητής νοημοσύνης στην ιατρική απεικόνιση προκαλεί επανάσταση στον τομέα της υγείας, κατορθώνοντας να παρέχονται πιο ακριβείς, αποτελεσματικές και εξατομικευμένες διαγνωστικές και θεραπευτικές επιλογές στους ασθενείς. Ωστόσο, παρά τις προσπάθειες για σταδιακή ένταξη της τεχνητής νοημοσύνης στον τομέα της υγείας, η ιατρική κοινότητα δεν φαίνεται να της δείχνει απόλυτη εμπιστοσύνη. Στο πλαίσιο αυτό, η επεξηγησιμότητα (interpretability) των συστημάτων τεχνητής νοημοσύνης, όχι μόνο συμβάλλει στην ενίσχυση του κλίματος εμπιστοσύνης, αλλά έχει αποτυπωθεί και ως δικαίωμα του υποκειμένου στην επεξήγηση αποφάσεων που λαμβάνονται με αυτοματοποιημένο τρόπο. Οι Vision Transformers (ViTs) είναι μια πρόσφατη προσέγγιση στον τομέα της όρασης υπολογιστών, που έρχονται να αντικαταστήσουν τα, έως τώρα κυρίαρχα στην ανάλυση των εικόνων, Συνελικτικά Νευρωνικά Δίκτυα (CNNs), χρησιμοποιώντας μηχανισμούς προσοχής (attention mechanisms) που συναντώνται συχνά στην επεξεργασία φυσικής γλώσσας. Καθώς οι ViTs είναι πολύπλοκα μοντέλα που αντιμετωπίζουν δεδομένα υψηλής διάστασης, η ικανότητά τους να εξηγήσουν τις αποφάσεις τους είναι ζωτικής σημασίας και περιλαμβάνει την εξαγωγή χαρτών (attention, saliency, relevancy) για την επισημείωση των περιοχών της εικόνας που έπαιξαν καθοριστικό ρόλο για την πραγματοποίηση της ταξινόμησης από το μοντέλο. Στην παρούσα διπλωματική εργασία, γίνεται εφαρμογή ορισμένων Interpretable Vision Transformer δικτύων σε ιατρικά σύνολα δεδομένων διαφορετικής φύσης. Πιο συγκεκριμένα, εφαρμόζουμε το ProtoPFormer, το ViT-NeT σε τέσσερα datasets, τα οποία περιλαμβάνουν αξονικές και μαγνητικές τομογραφίες, ιστοπαθολογικές εικόνες και εικόνες από ενδοσκοπήσεις. Ακόμα, προκειμένου να αξιολογήσουμε την επίδραση των built-in μεθόδων ερμηνευσιμότητας στην ακρίβεια των μοντέλων, εφαρμόζουμε έναν απλό Transformer, τον Swin, συνδυασμένο με Grad-CAM ως post-hoc μέθοδο επεξηγησιμότητας, στα παραπάνω σύνολα δεδομένων και συγκρίνουμε τις επιδόσεις. Τα πειραματικά αποτελέσματα αποδεικνύουν ότι η προσθήκη ερμηνευσιμότητας στα δίκτυα μάλλον βελτιώνει, παρά μειώνει την ακρίβεια των ViTs.

Λέξεις-κλειδιά — Τεχνητή Νοημοσύνη, Βαθιά Μάθηση, Όραση Υπολογιστών, Interpretable Vision Transformers, Ιατρική Απεικόνιση, Επεξηγησιμότητα, Ερμηνευσιμότητα, ProtoPFormer, ViT-Ne, Swin Transformer, Grad-CAM.

Abstract

For several years now, efforts to extend digital image processing into image analysis and understanding algorithms has set the course for artificial intelligence. The developement of advanced deep learning models has enabled the successful analysis and understanding of complex images in a variety of applications, from automated recognition of everyday objects to medical diagnosis. The use of artificial intelligence in medical imaging is revolutionizing healthcare, by providing more accurate, efficient and personalized diagnostic and treatment options to patients. Nevertheless, despite efforts to gradually integrate artificial intelligence into healthcare, the medical community does not seem to fully trust it. In this context, the explainability of artificial intelligence systems, not only contributes to strengthening the climate of trust, but has also been reflected as the subject's right to explain decisions made in an automated manner. Vision Transformers (ViTs) are a recent approach in the fild of computer vision, coming to replace the hitherto dominant Convolutional Neural Networks (CNNs) in image analysis, using attention mechanisms often encountered in natural language processing. As complex models dealing with high-dimensional data, ViT's ability to explain their decisions is crucial and includes the generation of maps (attention, saliency, relevancy) to highlight the regions of the image that were definitive in making the classification by the model. In this thesis, specific Interpretable Vision Transformer networks are applied to medical imaging datasets of different nature. More specifically, we apply ProtoPFormer and ViT-NeT to four datasets, which include CT and MRI scans, histopathology images, and endoscopy images. Also, in order to evaluate the effect of interpretability methods on the accuracy of the models, we apply a simple Transformer, the Swin, combined with Grad-CAM, as a post-hoc explainability method, to the above datasets and compare the performances. Experimental results demonstrate that adding interpretability to networks rather improves, than degrades, the accuracy of ViTs.

Keywords — Artificial Intelligence, Deep Learning, Computer Vision, Interpretable Vision Transformers, Medical Imaging, Explainability, Interpretability, ProtoPFormer, ViT-NeT, Swin Transformer, Grad-CAM.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κύριο Αθανάσιο Βουλόδημο, καθώς και τον κύριο Γεώργιο Στάμου και κυρία Παρασκευή Τζούβελη για την πολύτιμη βοήθειά τους στην εκπόνηση της παρούσας διπλωματικής. Ακόμα, θα ήθελα να ευχαριστήσω τον Ιάσονα Λιάρτη, την Παρασκευή Θεοφίλου, τον Ορφέα Μένη και τον Βασίλη Καραμπίνη για την στενή συνεργασία μας και την ανεκτίμητη συνεισφορά τους καθ'όλη τη διαδικασία.

Βαρβάρα - Κωνσταντίνα Μαγγελάκη, Οκτώβριος 2024

Contents

List of Figures

Chapter 0

Εκτεταμένη Περίληψη στα Ελληνικά

0.1 Θεωρητικό Υπόβαθρο

Στη σύγχρονη τεχνολογία, λίγες εξελίξεις έχουν αναδιαμορφώσει την κοινωνία τόσο βαθιά όσο η τεχνητή νοημοσύνη (AI). Οι απαρχές της τεχνητής νοημοσύνης ανάγονται στα μέσα του 20ού αιώνα, όταν πρωτοπόροι όπως ο Alan Turing και ο John McCarthy έθεσαν τα εννοιολογικά θεμέλια για τις ευφυείς μηχανές. Έκτοτε, η AI έχει υποστεί αξιοσημείωτη εξέλιξη, ωθούμενη από την εκθετική αύξηση της υπολογιστικής ισχύος, των τεράστιων συνόλων δεδομένων και των καινοτόμων αλγορίθμων. Ένας από τους πιο σημαντικούς κλάδους της τεχνητής νοημοσύνης, η υπολογιστική όραση, επιτρέπει στις μηχανές να ερμηνεύουν και να κατανοούν τον οπτικό κόσμο [1]. Από τα συστήματα αναγνώρισης προσώπου [2] και την αυτόνομη οδήγηση [3] μέχρι τις εφαρμογές επαυξημένης πραγματικότητας [4] και την ιατρική απεικόνιση [5], η υπολογιστική όραση διαπερνά διάφορες βιομηχανίες, επαναπροσδιορίζοντας τον τρόπο με τον οποίο αντιλαμβανόμαστε και αλληλεπιδρούμε με το περιβάλλον μας.

Στον τομέα της ιατρικής απεικόνισης, παρά την πρόοδο, οι διαγνωστικές προκλήσεις παραμένουν, καθώς η τελική απόφαση εξαρτάται αποκλειστικά από την κρίση των ιατρών [6]. Η τεχνητή νοημοσύνη μπορεί να μειώσει τα ιατρικά σφάλματα μέσω αλγορίθμων που μαθαίνουν από ιατρικά δεδομένα [7–9]. Ωστόσο, η ερμηνευσιμότητα των αποφάσεων της τεχνητής νοημοσύνης είναι κρίσιμη, ειδικά στον ιατρικό τομέα, όπου οι εξηγήσεις των αλγορίθμων είναι απαραίτητες για την εμπιστοσύνη και την κατανόηση των αποτελεσμάτων [10]. Οι Vision Transformers (ViTs) είναι μια πρόσφατη εμφάνιση στον τομέα της υπολογιστικής όρασης που προσφέρει νέες δυνατότητες στην ανάλυση ιατρικών εικόνων, αλλά η ερμηνευσιμότητά τους παραμένει υπό διερεύνηση [11].

0.1.1 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (ANNs) εμφανίστηκαν τη δεκαετία του 1980 και προσομοιώνουν τα βιολογικά νευρωνικά δίκτυα του ανθρώπινου εγκεφάλου [12]. Τα ANNs αποτελούνται από διασυνδεδεμένους κόμβους ή νευρώνες, οργανωμένους σε στρώσεις. Κάθε νευρώνας λαμβάνει εισερχόμενα σήματα, τα επεξεργάζεται μέσω μιας συνάρτησης ενεργοποίησης και παράγει ένα εξερχόμενο σήμα. Η διαδικασία μάθησης σε ένα ANN περιλαμβάνει την προσαρμογή των παραμέτρων του για την ελαχιστοποίηση της απόκλισης μεταξύ των προβλεπόμενων και των πραγματικών εξόδων μέσω μιας συνάρτησης απώλειας. Παρόλο που τα ANNs είναι χρήσιμα, η ικανότητά τους να διαχειρίζονται την πολυπλοκότητα των δεδομένων εικόνας είναι περιορισμένη, γεγονός που οδηγεί σε υπερβολικές απαιτήσεις υπολογιστικής ισχύος και κίνδυνο υπερπροσαρμογής [13–15].

0.1.2 Αρχιτεκτονική CNN

Τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) είναι μια εξειδικευμένη κατηγορία Τεχνητών Νευρωνικών Δικτύων, σχεδιασμένη κυρίως για την επεξεργασία οπτικών δεδομένων [16, 17]. Οι CNNs χρησιμοποιούνται σε εφαρμογές όπως η ταξινόμηση εικόνων, η σημασιολογική τμηματοποίηση εικόνων και η ανίχνευση αντικειμένων. Οι στρώσεις των CNNs αποτελούνται από νευρώνες διατεταγμένους σε τρεις διαστάσεις (ύψος, πλάτος και βάθος), και κάθε νευρώνας συνδέεται μόνο με μια συγκεκριμένη περιοχή της προηγούμενης στρώσης. Οι βασικές στρώσεις ενός CNN περιλαμβάνουν συνελικτικές στρώσεις, στρώσεις συσσώρευσης και πλήρως συνδεδεμένες στρώσεις. Οι συνελικτικές στρώσεις εφαρμόζουν φίλτρα στα δεδομένα εισόδου, εξάγοντας χαρακτηριστικά όπως ακμές και υφές. Οι στρώσεις συσσώρευσης μειώνουν τις διαστάσεις των χαρτών χαρακτηριστικών, ενώ οι πλήρως συνδεδεμένες στρώσεις παράγουν τελικά τις ταξινομήσεις των δεδομένων. Αυτή η δομή επιτρέπει στα CNNs να μαθαίνουν πολύπλοκα ιεραρχικά χαρακτηριστικά από τις εικόνες με αποτελεσματικό τρόπο, μειώνοντας σημαντικά την υπολογιστική πολυπλοκότητα σε σύγκριση με τα παραδοσιακά ANNs [12].

0.1.3 Attention: Transformers

Σε αντίθεση με τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs), τα οποία έχουν περιορισμένη ικανότητα να καταγράφουν τις χωρικές εξαρτήσεις στα δεδομένα εισόδου, οι Transformers επιτρέπουν την αποτελεσματική εκπαίδευση σε παράλληλο επίπεδο και έχουν γίνει η κυρίαρχη αρχιτεκτονική για την επεξεργασία φυσικής γλώσσας [18].

Το μοντέλο Transformer βασίζεται αποκλειστικά σε έναν μηχανισμό αυτοπροσοχής (self-attention), ο οποίος συνδέει διάφορες θέσεις μέσα σε μια ακολουθία για να υπολογίσει την αναπαράστασή της. Στον μηχανισμό προσοχής QKV (Query, Key, Value), κάθε στοιχείο εισόδου συσχετίζεται με τρία διανύσματα: το διάνυσμα query Q, το διάνυσμα key K και το διάνυσμα value V. Ο υπολογισμός της προσοχής γίνεται με τον τύπο:

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

Το Multi-Head attention χωρίζει τα παραπάνω διανύσματα σε υπο-διανύσματα πριν από την εφαρμογή του μηχανισμού αυτοπροσοχής. Αυτό επιτρέπει στο μοντέλο να επικεντρώνεται σε διαφορετικές πτυχές της ακολουθίας εισόδου ανεξάρτητα, βελτιώνοντας την ικανότητά του να καταγράφει διάφορες σχέσεις εντός των δεδομένων. Τα αποτελέσματα από κάθε κεφαλή συνδυάζονται σε ένα ενιαίο διάνυσμα πριν περάσουν στο τελικό γραμμικό στρώμα [18–20].

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^O$$

όπου $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

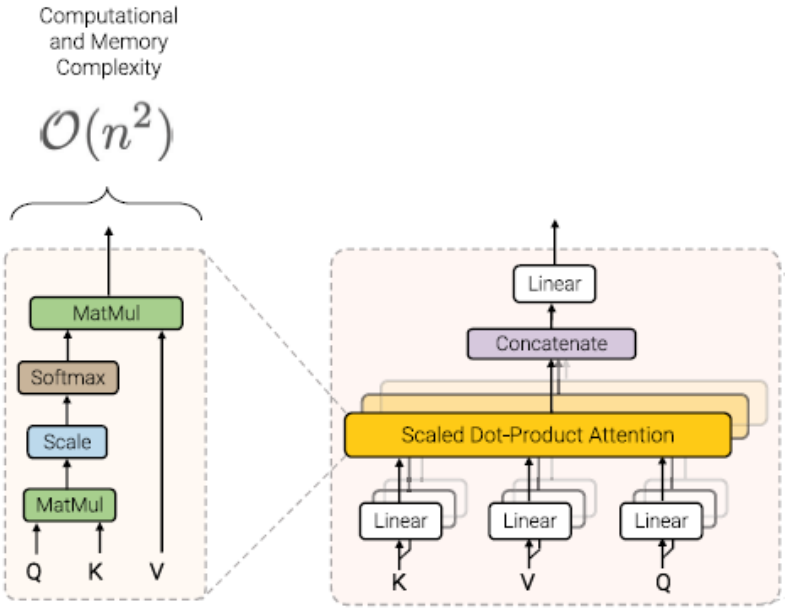


Figure 0.1.1: (αριστερά) Scaled-dot product attention. (δεξιά) Multi-head attention. "Efficient Transformers: A Survey" [20]

Το μοντέλο Transformer αποτελείται από δύο βασικά συστατικά: τον κωδικοποιητή και τον αποκωδικοποιητή.

Κωδικοποιητής: Ο κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου και εξάγει συμφραζόμενες πληροφορίες για κάθε στοιχείο. Αποτελείται από πολλαπλά στρώματα, το καθένα με έναν μηχανισμό multi-head attention και ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο με ReLU ενεργοποιήσεις [20].

$$X_A = \text{LayerNorm}(\text{MultiheadAttention}(X, X)) + X$$

$$X_B = \text{LayerNorm}(\text{PositionFFN}(X_A)) + X_A \text{ [20]}$$

Αποκωδικοποιητής: Ο αποκωδικοποιητής δημιουργεί την ακολουθία εξόδου βάσει των αναπαραστάσεων του κωδικοποιητή και των προηγούμενων παραγόμενων στοιχείων. Έχει παρόμοια αρχιτεκτονική με τον κωδικοποιητή και περιλαμβάνει έναν μηχανισμό μάσκας, έτσι ώστε, όταν το μοντέλο προσπαθεί να προβλέψει το επόμενο στοιχείο σε μια ακολουθία, να μπορεί να χρησιμοποιήσει μόνο τις πληροφορίες από τα προηγούμενα στοιχεία και όχι από τα μελλοντικά [18].

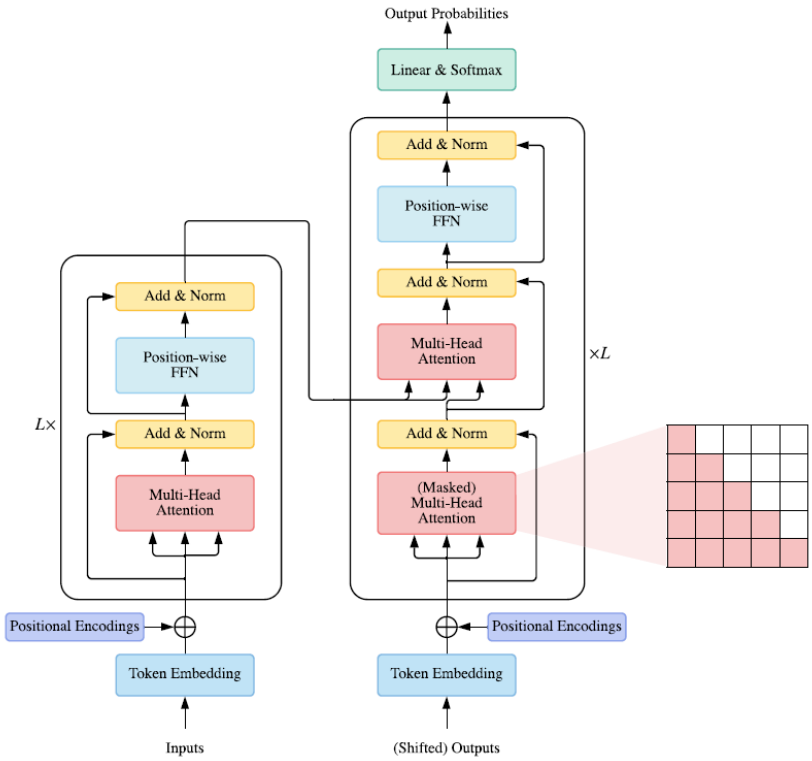


Figure 0.1.2: Απεικόνιση της αρχιτεκτονικής ενός τυπικού μοντέλου Transformer. "A survey of transformers" [19]

0.1.4 Vision Transformers

Οι Vision Transformers (ViTs) αποτελούν μια εφαρμογή της αρχιτεκτονικής των Transformers στην όραση υπολογιστών, αντικαθιστώντας τα CNNs στις εφαρμογές ανάλυσης εικόνων [11]. Οι ViTs διασπούν τις εικόνες σε tokens, τα οποία στη συνέχεια προβάλλονται γραμμικά και συνδυάζονται με κωδικοποιήσεις θέσης (positional encodings) πριν τροφοδοτηθούν σε έναν τυπικό κωδικοποιητή Transformer.

Η διαδικασία επεξεργασίας μιας εικόνας από τους ViTs περιλαμβάνει τα εξής βήματα [11]:

1. Η εικόνα διασπάται σε patches σταθερού μεγέθους.
2. Τα patches ισοπεδώνονται και προβάλλονται γραμμικά σε D διαστάσεις.
3. Προστίθεται ένα ειδικό classification token στην αρχή της ακολουθίας.
4. Κωδικοποιήσεις θέσης (positional encodings) προστίθενται στα patches.
5. Τα patches με τα positional encodings τροφοδοτούνται σε έναν κωδικοποιητή Transformer, που αποτελείται από στρώματα πολλαπλών κεφαλών προσοχής και MLP μπλοκ.

0.2 Σχετικές Αρχιτεκτονικές

0.2.1 Επεξηγησιμότητα στην όραση υπολογιστών

Επεξηγησιμότητα στην όραση υπολογιστών. Οι κύριες κατηγορίες τεχνικών για τη δημιουργία θερμικών χαρτών (heatmaps) περιλαμβάνουν τις μεθόδους κλίσης και τις μεθόδους απόδοσης. Οι μέθοδοι κλίσης, όπως οι attention rollout και GradCAM, υπολογίζουν τις κλίσεις της εξόδου του μοντέλου σε σχέση με τα εικονοστοιχεία της εισόδου για να μετρήσουν την επίδρασή τους στην πρόβλεψη. Οι μέθοδοι απόδοσης, όπως η Layer-wise Relevance Propagation μέθοδος (LRP), αναλύουν συστηματικά τη διαδικασία λήψης αποφάσεων

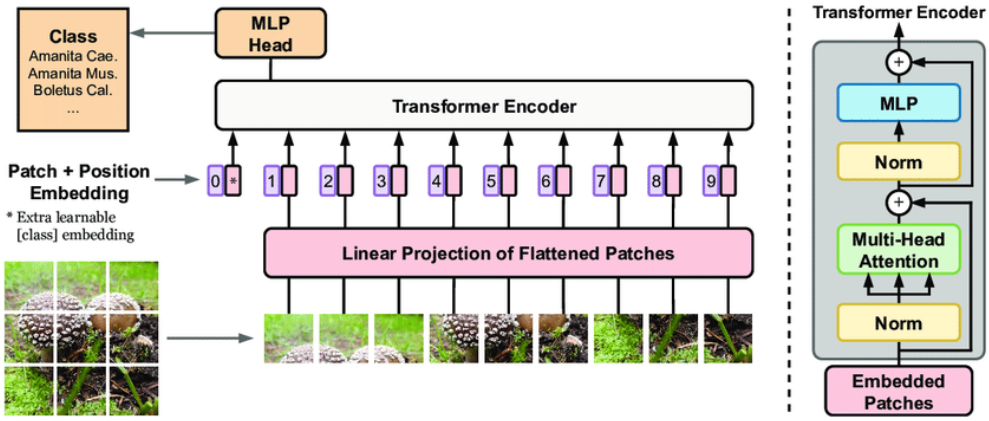


Figure 0.1.3: Επισκόπηση του ViT. Η εικόνα διασπάται σε patches σταθερού μεγέθους, που χρησιμοποιούνται ως τοκένς, τα οποία στη συνέχεια τροφοδοτούνται σε έναν τυπικό κωδικοποιητή Transformer. "Automatic fungi recognition: deep learning meets mycology" [21]

των νευρωνικών δικτύων κατανεμώντας τις συνεισφορές από προηγούμενα επίπεδα στα στοιχεία της εισόδου.

Attention rollout και Attention flow

Το **attention rollout** [22] προσφέρει έναν τρόπο για την ιχνηλάτηση της ροής πληροφορίας από το επίπεδο εισόδου έως τις ανώτερες ενσωματώσεις σε ένα μοντέλο Transformer. Για έναν Transformer με L επίπεδα, η προσοχή υπολογίζεται από όλες τις θέσεις στο επίπεδο l_i προς όλες τις θέσεις στο επίπεδο l_j, όπου $j < i$.

Ένα γράφημα προσοχής απεικονίζει τη ροή της προσοχής μέσα σε ένα νευρωνικό δίκτυο. Μια διαδρομή από τον κόμβο v στη θέση k στο επίπεδο l_i προς τον κόμβο u στη θέση m στο επίπεδο l_j αναπαριστά μια σειρά συνδέσεων. Πολλαπλασιάζοντας τα βάρη αυτών των ακμών, μπορούμε να υπολογίσουμε την ποσότητα πληροφορίας που μεταφέρεται από τον v στον u. Εάν υπάρχουν πολλαπλές διαδρομές, αθροίζουμε όλες τις διαδρομές για να βρούμε τη συνολική ροή πληροφορίας.

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}) & \text{αν } i > j \\ A(l_i) & \text{αν } i = j \end{cases}$$

όπου $\tilde{A}$ είναι η ανάπτυξη της προσοχής, A είναι η αρχική προσοχή και γίνεται πολλαπλασιασμός πινάκων.

Attention flow. Η θεώρηση του γραφήματος προσοχής ως δίκτυο ροής, με τις χωρητικότητες των ακμών ως βάρη προσοχής, επιτρέπει τον υπολογισμό της μέγιστης ροής προσοχής από οποιονδήποτε κόμβο επιπέδου στους κόμβους εισόδου με αλγορίθμους μέγιστης ροής. Αυτή η μέγιστη ροή χρησιμοποιείται ως εκτίμηση της προσοχής στους κόμβους εισόδου. Σε αντίθεση με την μέθοδο attention rollout, όπου το βάρος μιας διαδρομής είναι το γινόμενο των βαρών των ακμών, στη ροή προσοχής καθορίζεται από την ελάχιστη τιμή των βαρών των ακμών κατά μήκος της διαδρομής, λόγω πιθανής επικάλυψης διαδρομών [22].

GradCAM

Το Grad-CAM γενικεύει το CAM για να εφαρμόζεται σε οποιαδήποτε αρχιτεκτονική CNN [23]. Το CAM (Class Activation Mapping) δημιουργεί χάρτες ενεργοποίησης τάξης, επισημαίνοντας περιοχές της εικόνας που είναι σημαντικές για την πρόβλεψη μιας συγκεκριμένης τάξης [24]. Το CAM συνδυάζει τους χάρτες χαρακτηριστικών του τελευταίου συνελικτικού επιπέδου με βάρη που μαθαίνονται από ένα πλήρως συνδεδεμένο επίπεδο.

Το Grad-CAM χρησιμοποιεί τις βαθμίδες του τελικού συνελικτικού επιπέδου για να υπολογίσει τη σημασία κάθε χάρτη χαρακτηριστικών στην τελική πρόβλεψη.

$$w_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Το Grad-CAM παράγει έναν ζυγισμένο συνδυασμό των χαρτών χαρακτηριστικών χρησιμοποιώντας τα βάρη σημασίας και εφαρμόζει τη συνάρτηση ReLU για να διατηρήσει τις θετικές συσχετίσεις.

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

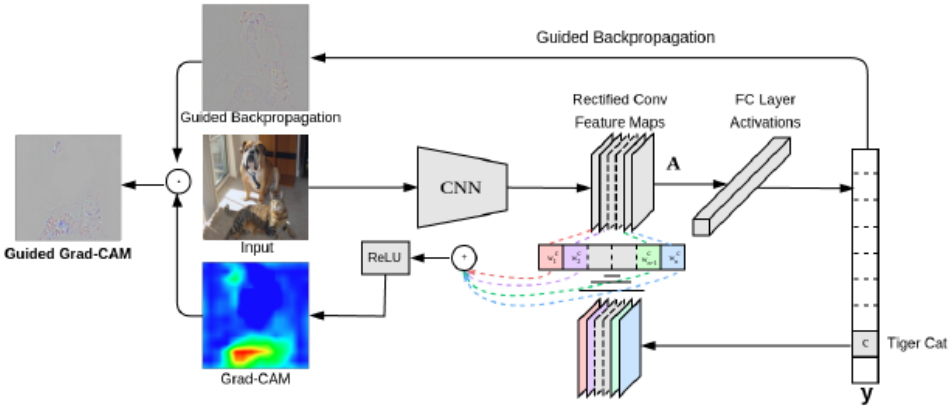


Figure 0.2.1: Το Grad-CAM ξεκινά με μια εικόνα εισόδου και μια καθορισμένη κατηγορία. Οι βαθμίδες προσαρμόζονται για να δοθεί έμφαση στην τάξη στόχο. Αυτό το σήμα προωθείται προς τα πίσω στον συνελικτικό χάρτη χαρακτηριστικών για να υπολογιστεί ο χάρτης ενεργοποίησης Grad-CAM, που αναπαρίσταται από έναν μπλε θερμικό χάρτη. Τέλος, ο θερμικός χάρτης συνδυάζεται με καθοδηγούμενη οπισθοδιάδοση, οδηγώντας σε οπτικοποιήσεις Guided Grad-CAM.
"Grad-CAM: Why did you say that?" [23]

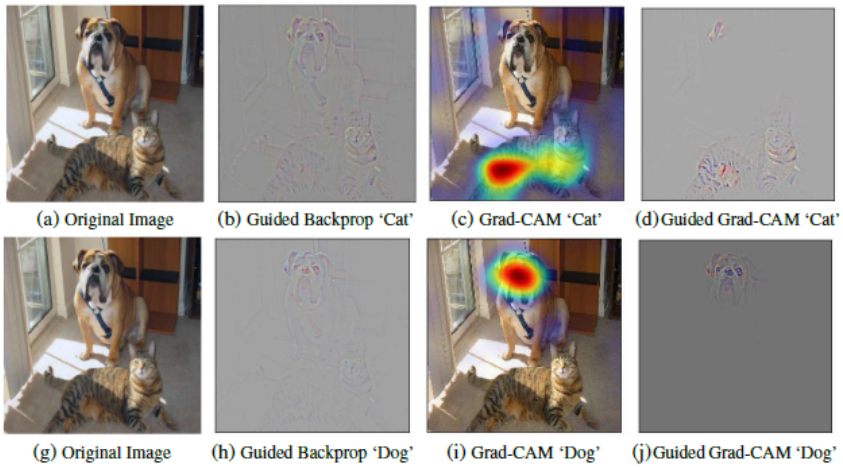


Figure 0.2.2: Σύγκριση Οπτικοποιήσεων: Αρχική εικόνα μιας γάτας και ενός σκύλου, μαζί με οπτικοποιήσεις που δημιουργήθηκαν χρησιμοποιώντας τις τεχνικές Καθοδηγούμενης Οπισθοδιάδοσης, Grad-CAM και Guided Grad-CAM, αναδεικνύοντας σαφώς τα σημαντικά χαρακτηριστικά και τις περιοχές συγκεκριμένων τάξεων.
"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" [25]

LIME

Η μέθοδος Local Interpretable Model-agnostic Explanations (LIME) [26] παρέχει εξηγήσεις για προβλέψεις πολύπλοκων μοντέλων μηχανικής μάθησης. Η LIME δημιουργεί τοπικά μοντέλα g, όπως γραμμικά μοντέλα ή δέντρα αποφάσεων, για να προσεγγίσει τη συμπεριφορά του μοντέλου f γύρω από μια συγκεκριμένη περίπτωση x. Χρησιμοποιεί ένα μέτρο εγγύτητας π_x για να επικεντρωθεί στη σχετικότητα των δειγμάτων γύρω από το x, και ελαχιστοποιεί μια συνδυασμένη συνάρτηση απώλειας και πολυπλοκότητας για την εξήγηση:

$$\hat{g}(x) = \text{argmin}_{g \in G} \left(L(f, g, \pi_x) + \Omega(g) \right)$$

Η LIME χρησιμοποιεί δειγματοληψία γύρω από το x για να προσεγγίσει την απώλεια, παραμένοντας ανθεκτική παρά τον θόρυβο δειγματοληψίας.

Layer-Wise Relevance Propagation (LRP)

Η μέθοδος Layer-Wise Relevance Propagation (LRP) [27, 28] προχωρά στην αναδρομική διάδοση της σημασίας της πρόβλεψης $f(x)$ στο νευρωνικό δίκτυο, ακολουθώντας το πλαίσιο της Deep Taylor Decomposition. Στη διαδικασία διάδοσης, που χρησιμοποιείται από τη μέθοδο LRP, η πληροφορία που δέχεται ένας νευρώνας πρέπει να αναδιανέμεται εξίσου προς τα κατώτερα επίπεδα (ιδιότητα συντήρησης). Εάν j και k αναπαριστούν νευρώνες σε διαδοχικά επίπεδα του δικτύου, ο κανόνας διάδοσης των σκορ σημασίας ($(R_k)^k$) στους νευρώνες του κατώτερου επιπέδου είναι:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k.$$

όπου z_{jk} ποσοτικοποιεί τον βαθμό επιρροής του νευρώνα j στη σημασία του νευρώνα k, και ο παρονομαστής εξασφαλίζει τη διατήρηση της ιδιότητας συντήρησης. Η διαδικασία διάδοσης ολοκληρώνεται όταν φτάσει στα χαρακτηριστικά εισόδου.

0.2.2 Ερμηνεύσιμοι Vision Transformers

ViT-NeT

Για να επιτευχθεί καλύτερη ισορροπία μεταξύ ερμηνευσιμότητας και απόδοσης, το 2022 παρουσιάστηκε το ViT-NeT [29], συνδυάζοντας τον κωδικοποιητή Swin Transformer [30] με ένα νευρωνικό αποκωδικοποιητή δενδρικής δομής. Ο Swin Transformer χρησιμοποιεί μια ιεραρχική στρατηγική κωδικοποίησης χαρακτηριστικών και δυναμική προσαρμογή των παραθύρων για να ανιχνεύσει μικρά και μεγάλα αντικείμενα με γραμμική υπολογιστική πολυπλοκότητα.

Ο Νευρωνικός Αποκωδικοποιητής Δέντρου χρησιμοποιεί κόμβους, φύλλα και ακμές για να διανείμει τη σημασία των χαρακτηριστικών σε επίπεδο εικόνας. Κάθε εσωτερικός κόμβος αντιπροσωπεύει ένα πρωτότυπο, αξιολογώντας την ομοιότητα με τμήματα εικόνας και καθοδηγώντας την δρομολόγηση. Οι τελικές προβλέψεις γίνονται με βάση τα αποτελέσματα των φύλλων και τις συγκεντρωμένες βαθμολογίες δρομολόγησης.

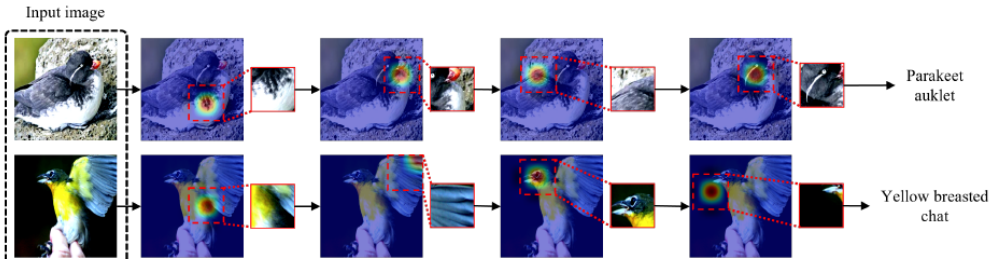


Figure 0.2.3: Τοπικές ερμηνείες που δείχνουν τις διαδικασίες απόφασης για τυχαίες εικόνες. Το NeT αναγνωρίζει συγκεκριμένα "χαρακτηριστικά πτηνού" στις εικόνες.
"ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder" [29]

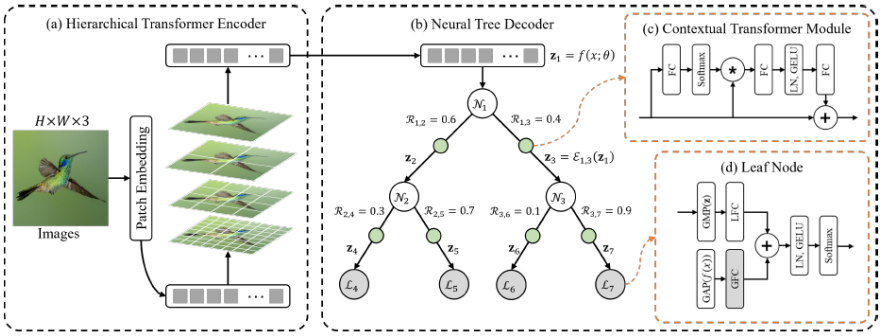


Figure 0.2.4: ViT-NeT: επισκόπηση.
"ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder" [29]

ProtoPFormer

Στη συνέχεια, οι Xue et al. πρότειναν το ProtoPFormer [31], το οποίο συνδυάζει τη μέθοδο βασισμένη σε πρωτότυπα [32] με ViTs για ερμηνεύσιμη αναγνώριση εικόνων. Το ProtoPFormer εισάγει παγκόσμια και τοπικά πρωτότυπα για να εντοπίζει και να αναδεικνύει τα χαρακτηριστικά των στόχων μέσω μιας διαδικασίας αμοιβαίας διόρθωσης και κοινής απόφασης.

Ένας κλάδος παγκόσμιου πρωτότυπου και ένας κλάδος τοπικού πρωτότυπου χρησιμοποιούνται για την ανάλυση του οπτικού σήματος. Οι τελικές προβλέψεις γίνονται με βάση το ζυγισμένο άθροισμα των αποτελεσμάτων των δύο κλάδων.

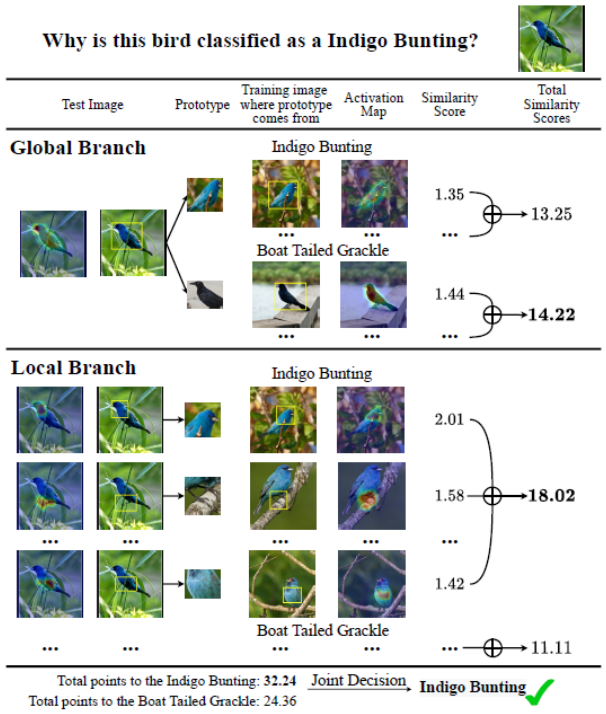


Figure 0.2.5: Διαδικασία συλλογιστικής για την ταξινόμηση μιας εικόνας πτηνού ως Indigo Bunting μέσω αμοιβαίας διόρθωσης και κοινής απόφασης των κλάδων τοπικού και παγκόσμιου πρωτότυπου.
"ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition" [31]

Το ProtoPFormer ενισχύει τη συγκέντρωση των τοπικών πρωτοτύπων σε στοιχεία πρώτου πλάνου ενώ φιλτράρει τις επιρροές του φόντου. Συγχρόνως, χρησιμοποιείται μάσκα διατήρησης πρώτου πλάνου (FP), για να

διατηρούνται μόνο τα tokens που είναι σχετικά με το πρώτο πλάνο και να εξαιρούνται εκείνα που σχετίζονται με το φόντο. Η μάσκα αυτή δημιουργείται με τη μέθοδο rollout. Επίσης, η χρήση της απώλειας συγκέντρωσης πρωτοτυπικών μερών (PPC) προάγει τη διαφοροποίηση των τοπικών πρωτοτύπων μέσα στην ίδια κατηγορία.

PaCa-ViT

Το 2023, εμφανίστηκε το PaCa-ViT [33], ένα νέο ερμηνεύσιμο ViT μοντέλο, το οποίο ξεπερνά τις προηγούμενες εκδόσεις όπως το Swin-Transformer [30] και το PVT [34, 35].

Το PaCa-ViT εισάγει τον μηχανισμό προσοχής από patch σε cluster (Patch-to-Cluster Attention, PaCa) για να διαχειριστεί την πολυπλοκότητα των υπολογισμών. Το PaCa μειώνει την πολυπλοκότητα, διατηρώντας τη σχέση $M << N$, μέσω του κατακερματισμού της ακολουθίας εισόδου $X_{N,C}$ σε "οπτικά tokens" $Z_{M,C}$.

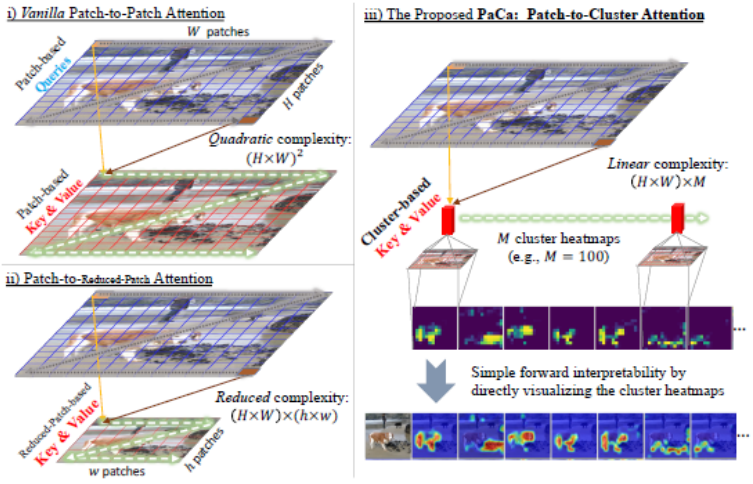


Figure 0.2.6: Η κλασική προσέγγιση patch-to-patch self-attention αντιμετωπίζει προβλήματα λόγω της τετραγωνικής πολυπλοκότητας, καθώς κάθε Query αλληλεπιδρά με κάθε Key. Μια δημοφιλής μέθοδος για τη μείωση αυτής της πολυπλοκότητας περιλαμβάνει τη χωρική μείωση μέσω τεχνικών όπως η συνέλιξη με βήμα. Το άρθρο προτείνει την προσέγγιση Patch-to-Cluster attention (PaCa), η οποία χρησιμοποιεί έναν προκαθορισμένο αριθμό clusters για τον υπολογισμό των Key και Value, επιτυγχάνοντας γραμμική πολυπλοκότητα και πιο ουσιαστικά οπτικά στοιχεία.
"PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers" [33]

Στη μέθοδο **onsite clustering**, ο υπολογισμός του $C_{N,M}$ γίνεται είτε μέσω συνελίξεων βάθους και σημείου είτε μέσω MLP. Το **external clustering** χρησιμοποιεί ένα εκπαιδευμένο CNN για να καθοδηγεί τον PaCa ViT, επιτρέποντας στο μοντέλο να μάθει από διαφορετικές πληροφοριακές πηγές.

Ερμηνευσιμότητα Δικτύου. Το PaCa ViT χρησιμοποιεί χάρτες ταξινόμησης για να εντοπίσει τα πιο σημαντικά κλάσματα μιας εικόνας. Οι χάρτες αυτοί, μετατρέπονται σε 2D χωρικούς χάρτες και χρησιμοποιούνται για να δημιουργήσουν μάσκες που εφαρμόζονται στις εισόδους, τονίζοντας τις σημαντικές περιοχές της εικόνας.

Ex-ViT

Το eX-ViT [36] λειτουργεί ως δίκτυο siamese, επεξεργάζοντας δύο διαφορετικές τυχαίες μετασχηματισμένες εκδοχές της αρχικής εικόνας. Κάθε κλάδος περιλαμβάνει έναν κωδικοποιητή με το Explainable Multi-Head Attention (E-MHA) και το Attribute-guided Explainer (AttE).

Explainable Multi-Head Attention (E-MHA): Η μονάδα E-MHA περιλαμβάνει πολλαπλές κεφαλές που μαθαίνουν επεξηγήσιμα βάρη προσοχής, βελτιώνοντας την ανθεκτικότητα στον θόρυβο και την εγγενή επεξηγήσιμότητα. Η διαδικασία ξεκινά με την προβολή των εισόδων στις μήτρες K, Q και V και τον υπολογισμό των βαρών προσοχής. Τα χαρακτηριστικά της προσοχής υπολογίζονται ως το γινόμενο της μήτρας προσοχής (A) με την μήτρα τιμών (V).

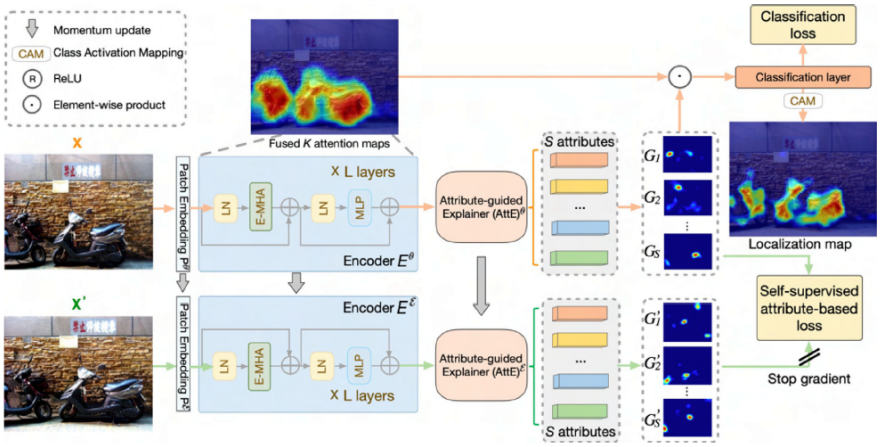


Figure 0.2.7: Γενική επισκόπηση της αρχιτεκτονικής του eX-ViT.
"eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation" [36]

Attribute-guided Explainer (AttE): Η μονάδα AttE βελτιώνει την εξηγήσιμότητα των χαρτών προσοχής, δημιουργώντας χάρτες χαρακτηριστικών που δίνουν έμφαση σε συγκεκριμένα χωρικά στοιχεία. Αυτοί οι χάρτες, χωρισμένοι σε ομάδες αντιπροσωπεύοντας διαφορετικά χαρακτηριστικά, εφαρμόζονται στους αρχικούς, για την παραγωγή αναπαραστάσεων χαρακτηριστικών. Η διαδικασία αυτή επιτρέπει στο μοντέλο να εντοπίζει ρητά τα pixels που σχετίζονται με συγκεκριμένα χαρακτηριστικά. Επιπλέον, μια απώλεια καθοδηγούμενη από τα χαρακτηριστικά χρησιμοποιείται για να ενισχύσει την αξιοπιστία του μοντέλου.

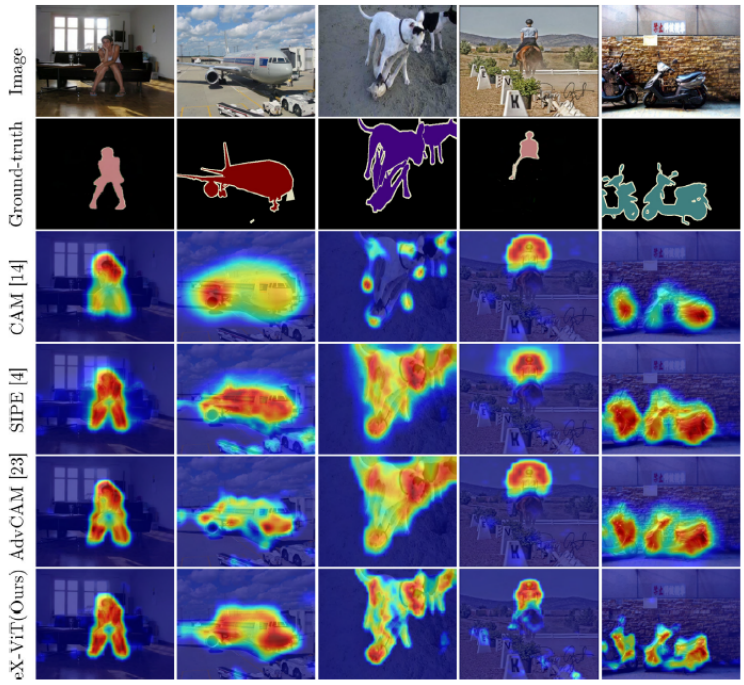


Figure 0.2.8: Σύγκριση διαφορετικών μεθόδων ερμηνείας στο PASCAL VOC 2012 Training Set.
"eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation" [36]

0.2.3 Ερμηνείες στην Ιατρική Απεικόνιση

Στον τομέα της ιατρικής απεικόνισης, η ερμηνευσιμότητα των μοντέλων τεχνητής νοημοσύνης είναι κρίσιμη για την ασφαλή και αποτελεσματική εφαρμογή τους στην κλινική πρακτική. Αυτή η ενότητα εξετάζει τις μεθόδους εξηγήσεων που έχουν αναπτυχθεί είτε ειδικά για τις ιατρικές απεικονιστικές εργασίες είτε χρησιμοποιούνται για την παραγωγή εξηγήσεων στον τομέα της ιατρικής, ξεκινώντας με την έρευνα των Komorowski et al. [37].

Η μελέτη αυτή συγκρίνει τις μεθόδους Attention Rollout, TransLRP και LIME στην ταξινόμηση ακτινογραφιών θώρακα. Τα αποτελέσματα δείχνουν ότι το TransLRP έχει ισχυρή δυναμική στην εξήγηση των προβλέψεων των ViT για τις κατηγορίες Covid, Non-Covid και Healthy στις ακτινογραφίες θώρακα. Παρόλο που το TransLRP είναι ανθεκτικό σε παραμορφώσεις ή ανωμαλίες στις ιατρικές εικόνες, μπορεί να παράγει εξηγήσεις βασισμένες σε λανθασμένες συσχετίσεις.

Το LIME παρέχει συνεπείς εξηγήσεις αλλά μπορεί να είναι ανακριβές αν τα υπερ-εικονοστοιχεία δεν εστιάζουν στις περιοχές των πνευμόνων στις ακτινογραφίες θώρακα. Οι εξηγήσεις που βασίζονται στην εξαγωγή χαρτών προσοχής είναι λιγότερο αξιόπιστες σε σύγκριση με αυτές που παράγονται από το TransLRP και το LIME. Συνολικά, τα αποτελέσματα δείχνουν ότι το TransLRP υπερέχει στις εξηγήσεις κατά την ταξινόμηση Covid σε ακτινογραφίες θώρακα, όπως φαίνεται στο Figure 3.3.1.

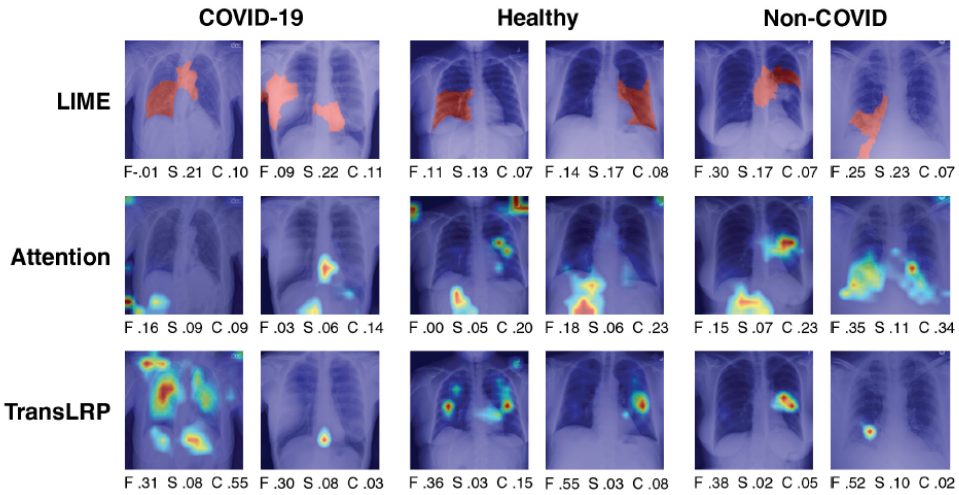


Figure 0.2.9: Οπτικές εξηγήσεις του ViT εκπαιδευμένου για την ταξινόμηση ακτινογραφιών. Δύο εικόνες εμφανίζονται για κάθε ετικέτα κλάσης και καθεμία εξηγείται χρησιμοποιώντας τρεις μεθόδους ερμηνείας. Παρέχονται μετρικές απόδοσης, συμπεριλαμβανομένων της πιστότητας (F), ευαισθησίας (S) και πολυπλοκότητας (C), με χαμηλότερα σκορ να είναι προτιμότερα για την ευαισθησία και την πολυπλοκότητα, ενώ υψηλότερα σκορ για την πιστότητα.
"Towards Evaluating Explanations of Vision Transformers for Medical Imaging" [37]

Σε μια άλλη εργασία, οι Playout et al. [38] ανέπτυξαν την τεχνική Focused Attention, που παράγει υψηλής ανάλυσης χάρτες θερμότητας με αποδοτική επεξεργασία των δεδομένων ιατρικής απεικόνισης. Η μέθοδος αυτή αντιμετωπίζει τις αυξημένες απαιτήσεις μνήμης και προτείνει την επιλογή patches μέσω επαναληπτικής διαδικασίας σε κλίμακα μειούμενων τιμών βημάτων. Τα αποτελέσματα δείχνουν ότι η μέθοδος αυτή συγκεντρώνει τον μηχανισμό προσοχής σε υποσύνολο σημαντικών τοκενών.

Σε μια άλλη εργασία, οι Demir et al. [39] εισάγουν ένα καινοτόμο μπλοκ προσοχής στην αρχιτεκτονική Convolutional Vision Transformer, που εστιάζει στη σχέση μεταξύ 'περιοχών' παρά 'εικονοστοιχείων', με ένα σύστημα βασισμένο στη μάθηση πρωτοτύπων.

Τέλος, το RadFormer [40] αντιμετωπίζει τις προκλήσεις της υπερηχογραφίας, χρησιμοποιώντας διπλή αρχιτεκτονική προσοχής για την ανίχνευση καρκίνου της χοληδόχου κύστης, συνδυάζοντας παγκόσμια και τοπικά χαρακτηριστικά για λεπτομερείς και ακριβείς ερμηνείες.

Συμπερασματικά, ενώ έχουν γίνει σημαντικά βήματα για την ανάπτυξη μεθόδων ερμηνείας στην ιατρική απεικόνιση, υπάρχει ακόμη ανάγκη για περαιτέρω έρευνα για την ανάπτυξη του τομέα και τη βελτίωση της

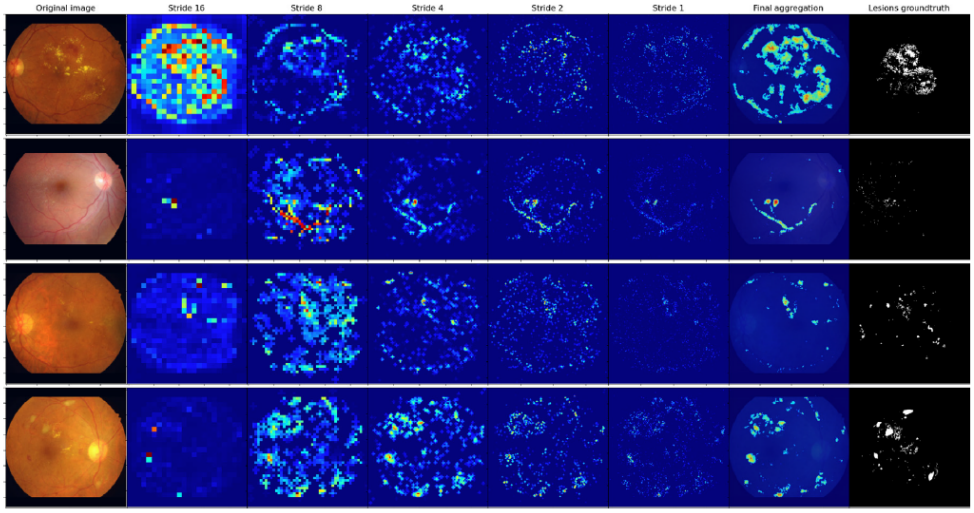


Figure 0.2.10: Οπτική αναπαράσταση της τεχνικής Focused Attention: κάθε στήλη παρουσιάζει τους χάρτες θερμότητας ανά βήμα, με τις δύο τελευταίες να δείχνουν την τελική σύνθεση και τις πραγματικές βλάβες. "Focused Attention in Transformers for interpretable classification of retinal images" [38]

φροντίδας των ασθενών.

0.3 Πειραματική Προσέγγιση

0.3.1 Σύνολα Δεδομένων

Η μεθοδολογική προσέγγιση περιλαμβάνει την προσαρμογή τριών μοντέλων ερμηνεύσιμων Vision Transformer για την αντιμετώπιση των ιδιαιτεροτήτων της ιατρικής απεικόνισης. Για το σκοπό αυτό, επιλέχθηκαν τέσσερα σύνολα δεδομένων ιατρικής απεικόνισης, διαφορετικής φύσης το καθένα. Αυτά περιλαμβάνουν μαγνητικές (MRIs) και αξονικές (CT scans) τομογραφίες, ιστοπαθολογικές εικόνες και πραγματικές εικόνες του γαστρεντερικού συστήματος από ενδοσκοπήσεις. Συγκεκριμένα, τα σύνολα δεδομένων είναι:

- Augmented Alzheimer MRI Dataset V2 [22], που περιέχει μαγνητικές τομογραφίες εγκεφάλου ασθενών με διαφορετικά στάδια της νόσου Αλτσχάιμερ.
- Large COVID-19 CT scan slice dataset [41], που χρησιμοποιείται ευρέως στη βιβλιογραφία για τη διάγνωση COVID-19.
- Gastrointestinal Cancer MSI MSS Prediction [42], που περιέχει ιστολογικές εικόνες για την ταξινόμηση MSI έναντι MSS στον γαστρεντερικό καρκίνο.
- Kvasir Dataset for Classification and Segmentation [43], που περιέχει εικόνες από το εσωτερικό του γαστρεντερικού συστήματος.

Τα βήματα προεπεξεργασίας των συνόλων δεδομένων, που περιλαμβάνουν τη λήψη, τη διαμόρφωση τς δομής και την εξισορρόπηση των δεδομένων, περιγράφονται στην ακόλουθη ενότητα:

1. Συλλογή και Λήψη Δεδομένων:
   - Αναγνώριση των σχετικών συνόλων δεδομένων ιατρικής απεικόνισης από τις πηγές τους.
   - Λήψη των συνόλων δεδομένων από αποθετήρια ή πηγές διασφαλίζοντας την ακεραιότητα και την ποιότητα των δεδομένων.
2. Διάσπαση σε Σύνολα Εκπαίδευσης-Δοκιμής:

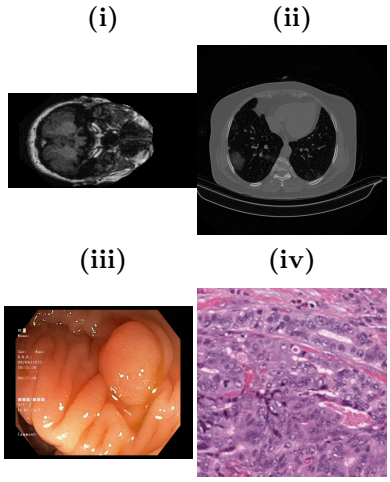


Figure 0.3.1: Τυχαίες εικόνες από το σύνολο εκπαίδευσης από: (i) το σύνολο δεδομένων ALzheimer's (ii) το σύνολο δεδομένων Covid (iii) το σύνολο δεδομένων Kvasir (iv) το σύνολο δεδομένων TCGA

- Διάσπαση του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης και δοκιμής χρησιμοποιώντας αναλογία 80-20.

3. Εξισορρόπηση Κλάσεων:
   - Αξιολόγηση της κατανομής των κλάσεων εντός του συνόλου δεδομένων για την αναγνώριση τυχόν ανισορροπιών.
   - Εάν υπάρχουν ανισορροπίες κλάσεων, εφαρμογή τυχαίας υποδειγματοληψίας (μείωση του μεγέθους των υπερεκπροσωπούμενων κλάσεων τυχαία για να ταιριάξει με το μέγεθος της μειοψηφούσας κλάσης).
4. Διαμόρφωση της Δομής του Συνόλου Δεδομένων:
   - Δημιουργία της επιθυμητής εσωτερικής δομής για την ύπαρξη επαναχρησιμοποιήσιμου κώδικα κατά την εκπαίδευση. Για παράδειγμα, για το ProtoPFormer: γονικός φάκελος/κλάση1, κλάση2, κλπ./train, test/δείγμα1.jpg, δείγμα2.jpg, κλπ.

Μετά την εφαρμογή των βημάτων προεπεξεργασίας στα τέσσερα σύνολα δεδομένων, το σύνολο δεδομένων Kvasir περιέχει 3200/800 εικόνες εκπαίδευσης/δοκιμής σε 8 κλάσεις που αναπαριστούν καταστάσεις του γαστρεντερικού συστήματος (dyed lifted polyps, dyed resection margins, esophagitis, normal cecum, normal pylorus, normal Z-line, polyps, ulcerative colitis). Το σύνολο δεδομένων Αλτσχάιμερ αποτελείται από 1564/400 εικόνες εκπαίδευσης/δοκιμής MRI χωρισμένες σε 4 κλάσεις που αναπαριστούν τα διάφορα στάδια της νόσου Alzheimer (Mild Dementia, Moderate Dementia, Non Demented, Very mild Dementia). Το σύνολο δεδομένων TCGA περιέχει 10000/2500 εικόνες ισομερώς κατανεμημένες σε δύο κλάσεις (MSS, MSIMUT) και τέλος, κάθε κλάση (Covid, Non-Covid) του συνόλου δεδομένων Covid έχει 5515/1380 εικόνες εκπαίδευσης/δοκιμής. Στον Table 4.1, περιγράφονται συνοπτικά οι διάφορες κλάσεις:

0.3.2 Εκπαιδεύοντας τα μοντέλα

Για το ProtoPFormer, η εκπαίδευση και οι οπτικοποιήσεις πραγματοποιήθηκαν στο Google Colab χρησιμοποιώντας την παρεχόμενη GPU T4, με αποθηκευτικό χώρο μέσω Google Drive. Λόγω του περιορισμένου χρόνου GPU, η διαδικασία εκπαίδευσης διήρκεσε περίπου δύο μήνες για όλα τα δεδομένα. Τα πειράματα για το ViTNet και το Swin Transformer υλοποιήθηκαν στο Kaggle, χρησιμοποιώντας δύο T4 GPUs ως επιταχυντές. Η εκπαίδευση του μοντέλου ViTNet διήρκεσε περίπου ενάμιση μήνα, ενώ το μοντέλο Swin Transformer ολοκλήρωσε την εκπαίδευσή του σε μόνο λίγες ημέρες. Τα αποτελέσματα θα μπορούσαν να βελτιστοποιηθούν περαιτέρω με τη χρήση πιο ισχυρών GPUs.

Για τους σκοπούς αυτής της εργασίας, επιλέχθηκαν τρία μοντέλα Vision Transformer για εφαρμογή σε ιατρικά δεδομένα.

Alzheimer's Dataset	
Mild Dementia	Συμπτώματα που αρχίζουν να επηρεάζουν τις καθημερινές δραστηριότητες, όπως η απώλεια μνήμης και η γνωστική εξασθένηση.
Moderate Dementia	Εντονα συμπτώματα που απαιτούν βοήθεια με τις καθημερινές δραστηριότητες, με σημαντική γνωστική εξασθένηση και προβλήματα μνήμης.
Non Demented	Άτομα που δεν εμφανίζουν σημάδια απώλειας μνήμης και χρησιμοποιούνται ως ομάδα ελέγχου.
Very mild Dementia	Πρώιμο στάδιο απώλειας μνήμης που τα συμπτώματα είναι πολύ ήπια και μπορεί να μην επηρεάζουν σημαντικά την καθημερινή ζωή.
Covid Dataset	
Covid	Εικόνες ασθενών διαγνωσμένων με COVID-19.
Non Covid	Εικόνες ασθενών που δεν έχουν διαγνωστεί με COVID-19.
Kvasir Dataset	
Dyed lifted polyps	Πολύποδες που έχουν ανυψωθεί και βαφεί για να αναδείξουν τα περίγραμμά τους.
Dyed resection margins	Περιοχές όπου έχει βαφεί ο ιστός για να σημειωθούν τα όρια εκτομής.
Esophagitis	Φλεγμονή του οισοφάγου.
Normal cecum	Υγιής ιστός στον τυφλό έντερο.
Normal pylorus	Υγιής ιστός στον πυλωρό.
Normal Z-line	Υγιής ιστός στη γαστροοισοφαγική σύνδεση.
Polyps	Ανώμαλες αναπτύξεις ιστού.
Ulcerative colitis	Φλεγμονώδης νόσος του εντέρου που προκαλεί έλκη.
TCGA Dataset	
MSIMUT	Υψηλή μικροσκοπική αστάθεια (MSI-H) υποδεικνύει υψηλό ποσοστό μεταλλάξεων.
MSS	Μικροσκοπική σταθερότητα (MSS) που υποδεικνύει χαμηλό ποσοστό μεταλλάξεων.

Table 1: Περιγραφές των κλάσεων στα σύνολα δεδομένων Alzheimer's, Covid, Kvasir και TCGA. Οι κλάσεις του συνόλου δεδομένων Αλτσχάιμερ περιγράφουν διάφορα στάδια της άνοιας. Το σύνολο δεδομένων Covid περιέχει εικόνες ασθενών με και χωρίς COVID-19. Το σύνολο δεδομένων Kvasir περιλαμβάνει διάφορες καταστάσεις του γαστρεντερικού συστήματος. Το σύνολο δεδομένων TCGA ταξινομεί τις ιστολογικές εικόνες βάσει της μικροσκοπικής σταθερότητας.

ProtoPFormer

Το μοντέλο ProtoPFormer [31] εκπαιδεύτηκε με batch size 64 και learning rate 5×10^{-4}. Χρησιμοποιήθηκε AdamW optimizer με cosine annealing scheduler. Ο συνολικός αριθμός των εποχών ήταν 60 για το dataset Alzheimer's και 100 για τα datasets Covid, Kvasir και TCGA. Οι προδιαγραφές για τα παγκόσμια και τοπικά πρωτότυπα βρίσκονται στον παρακάτω πίνακα:

Datasets	**Prototype number**	**Dimension**	**Global prototypes per class**
Alzheimer's	40	192	10
Covid	100	192	50
Kvasir	80	192	10
TCGA	100	192	50

Table 2: Προδιαγραφές Πρωτοτύπων για τα σύνολα δεδομένων Alzheimer's, Covid, Kvasir και TCGA. Ο αριθμός των παγκόσμιων πρωτοτύπων ανά κατηγορία καθορίστηκε σύμφωνα με το μέγεθος του συνόλου δεδομένων. Στα μεγαλύτερα σύνολα δεδομένων ανατέθηκαν περισσότερα πρωτότυπα ανά κατηγορία.

ViT-NeT

Το μοντέλο ViT-NeT [29] εκπαιδεύτηκε με δέντρο βάθους 4 και prototype size [1,1]. Χρησιμοποιήθηκε AdamW optimizer με αρχικό learning rate 2×10^{-5}. Ο συνολικός αριθμός των εποχών ήταν 60 για τα datasets Alzheimer's και Kvasir, 100 για το Covid, και 125 για το TCGA.

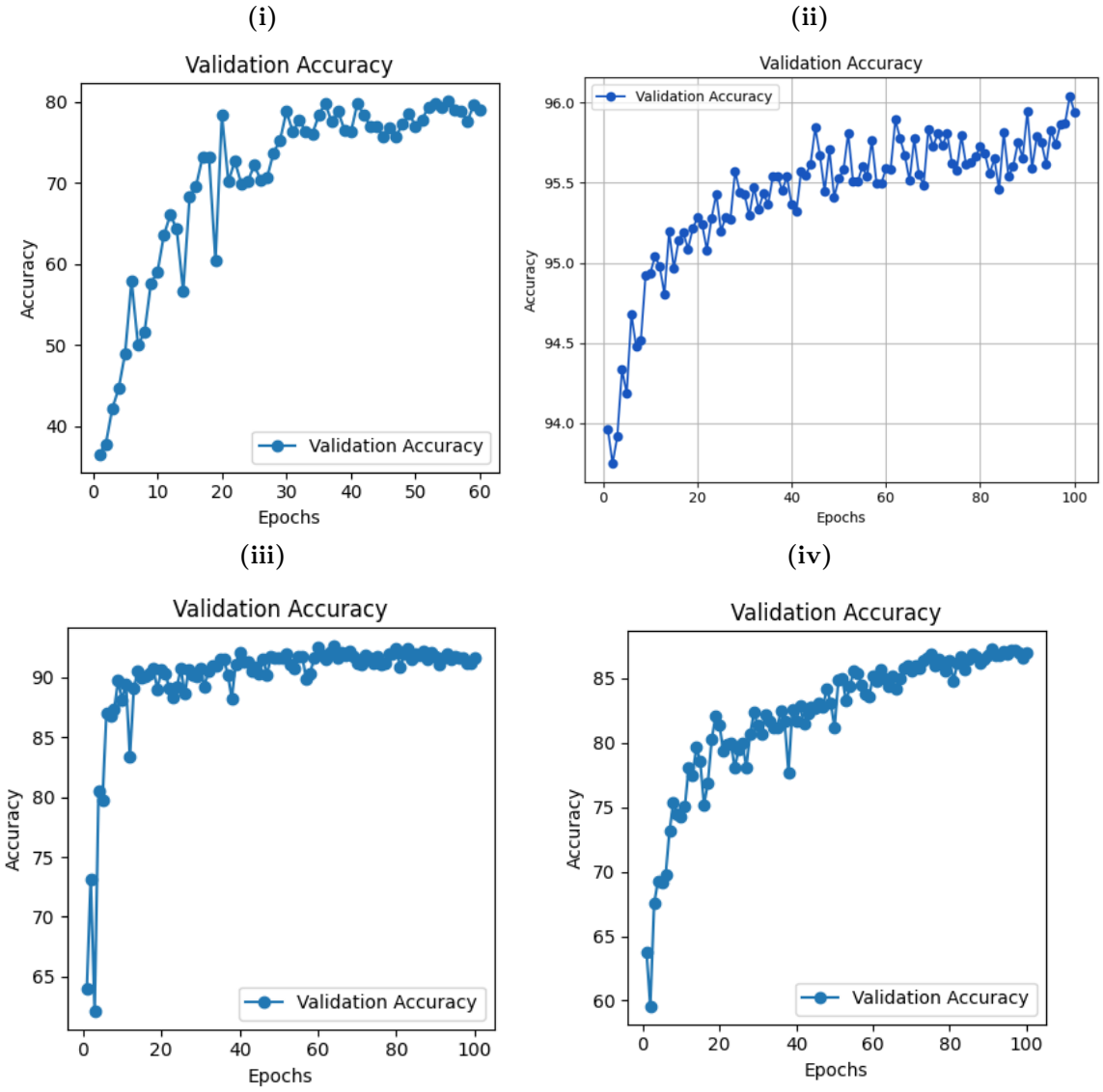


Figure 0.3.2: Καμπύλες μάθησης που δημιουργήθηκαν κατά τη διαδικασίας εκπαίδευσης του ProtoPFormer για: (i) το σύνολο δεδομένων Alzheimer's (ii) το σύνολο δεδομένων Covid (iii) το σύνολο δεδομένων Kvasir (iv) το σύνολο δεδομένων TCGA

Swin Transformer x Grad-CAM

Το μοντέλο Swin Transformer εκπαιδεύτηκε με batch size 64, χρησιμοποιώντας το variant `swin_tiny_patch4_window7_224`. Ο συνολικός αριθμός των εποχών ήταν 200 για όλα τα datasets, ενώ το μοντέλο εκπαιδεύτηκε με base learning rate 1×10^{-4}. Χρησιμοποιήθηκε Grad-CAM για οπτικοποίηση των περιοχών εστίασης του μοντέλου κατά την πρόβλεψη.

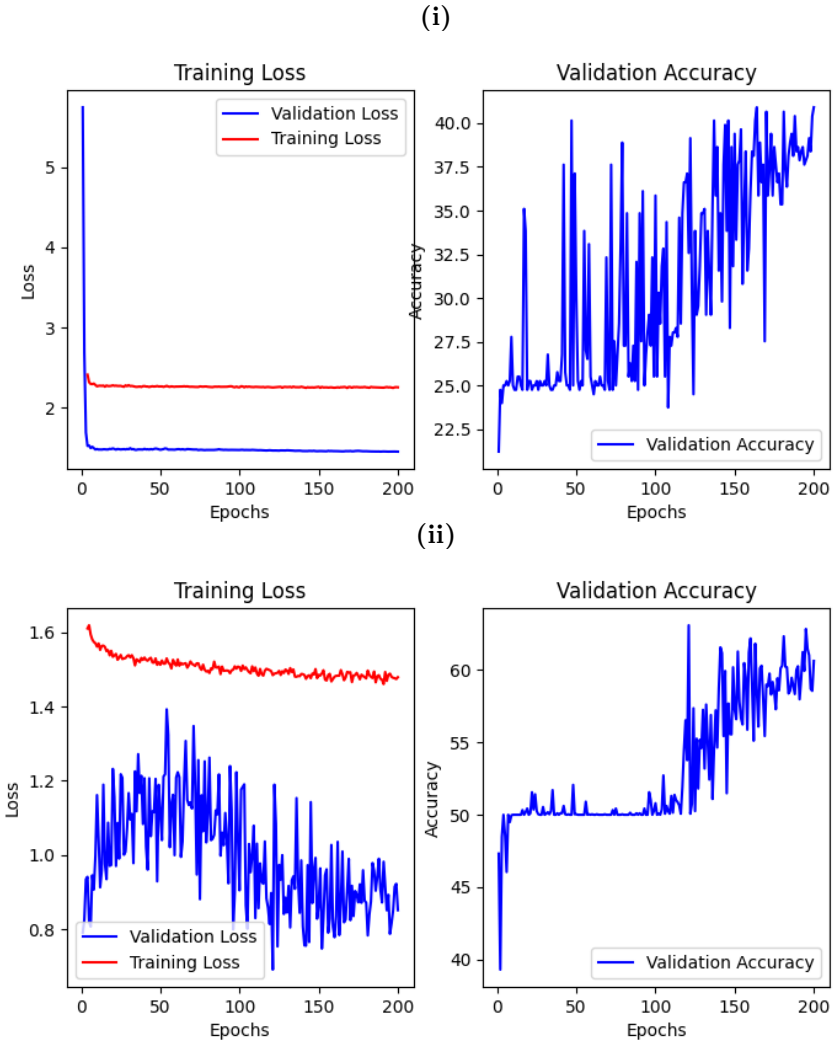
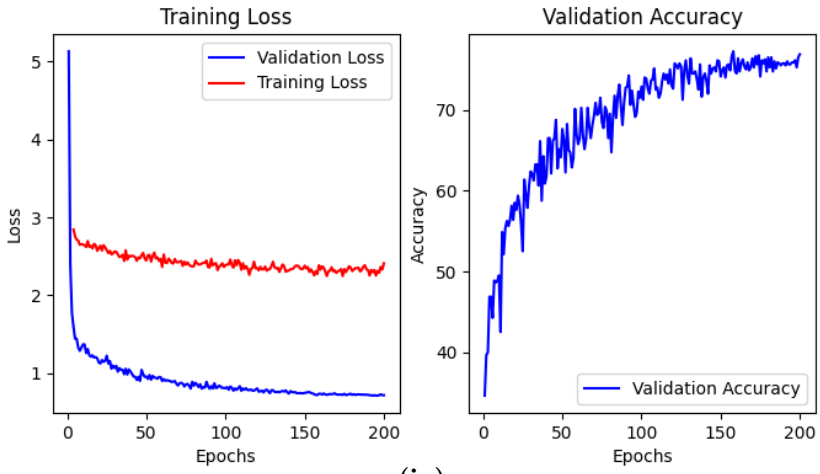


Figure 0.3.3: Καμπύλες μάθησης που δημιουργήθηκαν κατά τη διαδικασίας εκπαίδευσης του Swin Transformer για: (i) το σύνολο δεδομένων Alzheimer's (ii) το σύνολο δεδομένων Covid

(iii)



(iv)

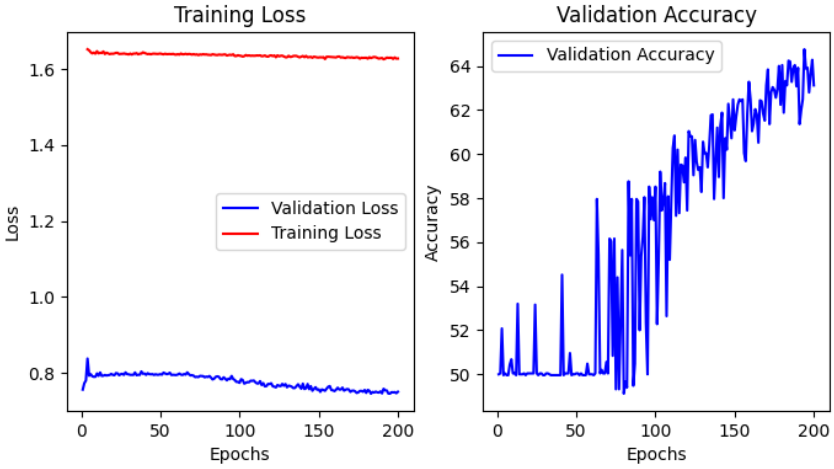


Figure 0.3.4: Καμπύλες μάθησης που δημιουργήθηκαν κατά τη διαδικασίας εκπαίδευσης του Swin Transformer για: (iii) το σύνολο δεδομένων Kvasir (iv) το σύνολο δεδομένων TCGA

0.3.3 Επίδοση και Οπτικοποιήσεις

Σε αυτή την ενότητα παρουσιάζεται η απόδοση τριών διαφορετικών μοντέλων σε τέσσερα επιλεγμένα ιατρικά σύνολα δεδομένων. Το ProtoPFormer πέτυχε την υψηλότερη ακρίβεια σε όλα τα σύνολα δεδομένων, ιδιαίτερα στα Kvasir και Covid, και όλα τα μοντέλα παρουσίασαν χαμηλότερη ακρίβεια στο σύνολο δεδομένων Alzheimer's. Στον πινακα Table 4.3, φαίνονται ανλυτικά τα ποσοστά ακρίβειας των μοντέλων.

Dataset/Model	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
Alzheimer's	**80.01**	41.27	40.91
Covid	**96.45**	67.77	63.09
Kvasir	**92.63**	86.88	77.25
TCGA	**87.28**	82.71	64.76

Table 3: Η σύγκριση της Max accuracy (%) των τριών μοντέλων στα τέσσερα επιλεγμένα ιατρικά σύνολα δεδομένων. Η μέγιστη απόδοση για κάθε σύνολο δεδομένων σημειώνεται με **έντονα** γράμματα.

Στα πλαίσια της ερμηνευσιμότητας, δημιουργήθηκαν οπτικοποιήσεις για να παρέχουν πληροφορίες σχετικά με τις περιοχές των εικόνων στις οποίες επικεντρώνεται κάθε μοντέλο για να κάνει προβλέψεις. Αυτές οι οπτικοποιήσεις είναι κρίσιμες για την κατανόηση της ερμηνευσιμότητας και της αξιοπιστίας των αποφάσεων των μοντέλων.

Οι οπτικοποιήσεις που παρήχθησαν από τα τρία μοντέλα—ProtoPFormer, ViT-NeT και Swin x Grad-CAM—εκτιμήθηκαν από μια επιτροπή ειδικών μέσω μιας αναλυτικής έρευνας, τα αποτελέσματα της οποίας θα περιγραφούν σε αυτή την ενότητα.

Οι γιατροί αξιολόγησαν αυτές τις οπτικοποιήσεις χρησιμοποιώντας ένα ερωτηματολόγιο που επικεντρωνόταν στην αποτελεσματικότητά τους, την σαφήνεια, την διαγνωστική αξία, την επισημείωση χαρακτηριστικών, την ευθυγράμμιση με τις γνώσεις των ειδικών και τη γενική χρησιμότητα στα τέσσερα ιατρικά σύνολα δεδομένων που έχουμε ήδη περιγράψει στις προηγούμενες ενότητες: Alzheimer's MRIs, Kvasir, COvid CT scans, TCGA.

Σε όλα τα αξιολογημένα σύνολα δεδομένων, το ProtoPFormer ξεχώρισε συνεχώς από τα άλλα μοντέλα, παρέχοντας οπτικοποιήσεις που ήταν ακριβείς και κλινικά χρήσιμες. Οι ειδικοί σημείωσαν ότι οι οπτικοποιήσεις του ProtoPFormer ήταν ιδιαίτερα αποτελεσματικές στην αποκάλυψη βασικών ενδείξεων και προτύπων, στην καθαρή παρουσίαση των διαγνώσεων και στην επισημείωση συγκεκριμένων χαρακτηριστικών ή περιοχών κρίσιμων για τη λήψη αποφάσεων. Οι οπτικοποιήσεις που δημιουργήθηκαν από το ProtoPFormer συμφωνούν με τις κλινικές γνώσεις και τις προσδοκίες των ειδικών, καθιστώντας το το πιο χρήσιμο εργαλείο για την ερμηνεία των προβλέψεων του μοντέλου.

Αντιθέτως, το ViTNet δημιούργησε οπτικοποιήσεις που ήταν γενικά λιγότερο ακριβείς και δεν κατάφεραν να επισημάνουν αποτελεσματικά τα σχετικά μέρη των εικόνων που είναι απαραίτητα για την τεκμηριωμένη λήψη αποφάσεων. Αυτό περιορίζει σημαντικά τη χρησιμότητά του σε όλα τα σύνολα δεδομένων.

Η μέθοδος Grad-CAM παρουσίασε επίσης αδυναμίες, καθώς τείνει να αποσπάται από μη σχετικές περιοχές των εικόνων αντί να εστιάζει στις περιοχές που είναι πιο σημαντικές για τη διάγνωση. Αυτή η έλλειψη εστίασης μείωσε την αποτελεσματικότητά της και συχνά οδήγησε σε οπτικοποιήσεις που δεν ευθυγραμμίζονταν καλά με τις κλινικές ανάγκες των ειδικών.

Συνολικά, το ProtoPFormer αναδείχθηκε ως το ανώτερο μοντέλο, παρέχοντας τις πιο αξιόπιστες και κλινικά εφαρμόσιμες οπτικοποιήσεις σε όλα τα σύνολα δεδομένων.

Οι πίνακες παρακάτω εμφανίζουν παραδείγματα εικόνων από όλα τα σύνολα δεδομένων, μαζί με τις οπτικοποιήσεις που δημιουργήθηκαν από κάθε μοντέλο.

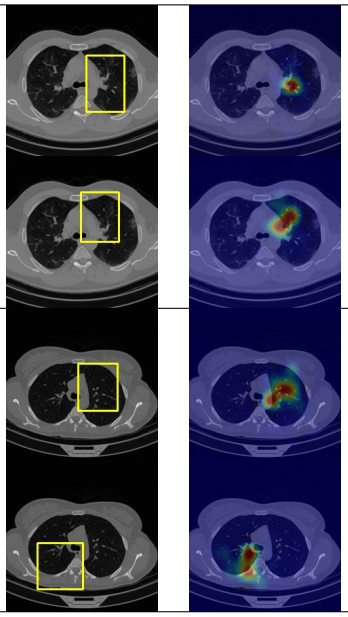
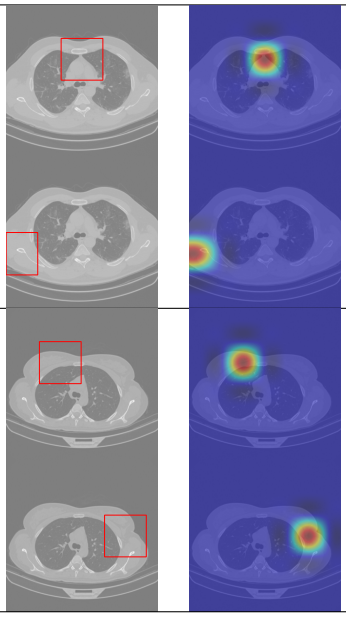
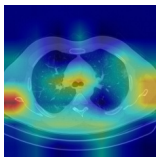
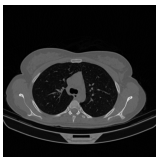
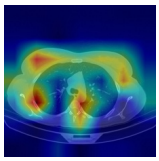
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
COVID 			
NonCOVID 			

Table 4: Οπτικοποιήσεις που δημιουργήθηκαν για το σύνολο δεδομένων Covid από τα ProtoPFormer, ViT-NeT, και Swin x Grad-CAM.

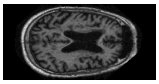
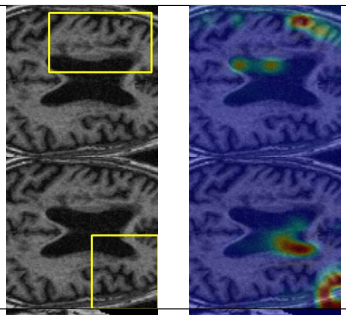
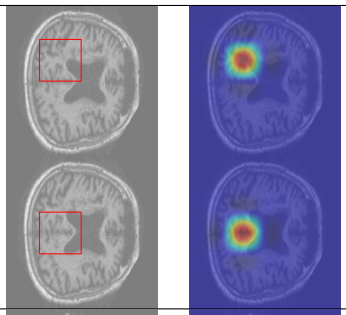
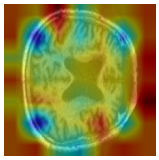
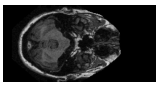
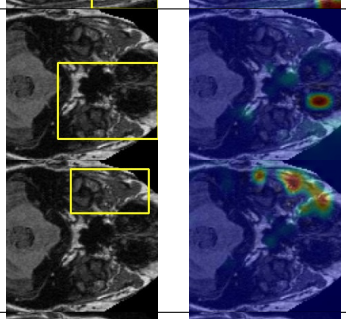
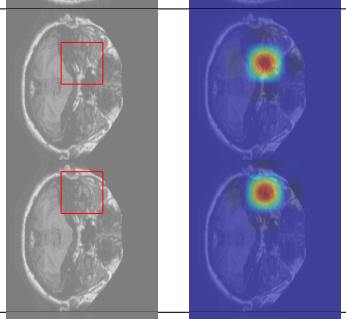
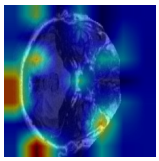
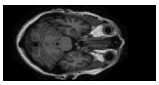
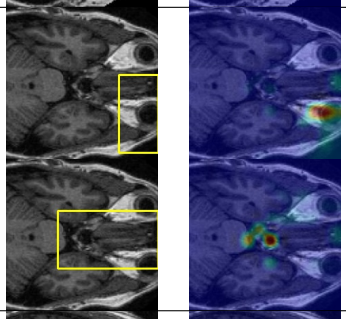
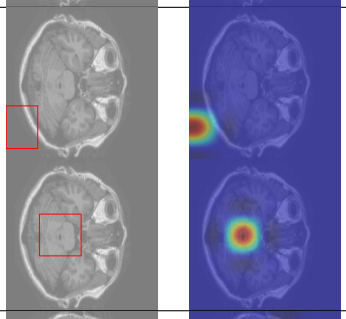
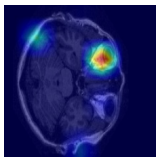
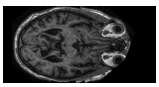
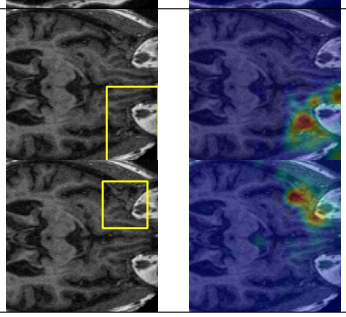
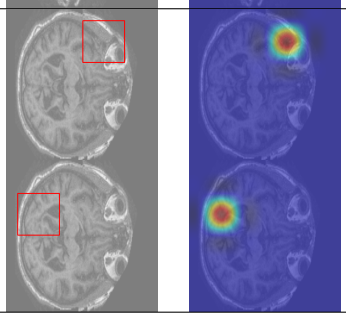
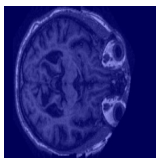
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
 Mild Dementia			
 Moderate Dementia			
 Non Demented			
 Very Mild Dementia			

Table 5: Οπτικοποιήσεις που δημιουργήθηκαν για το σύνολο δεδομένων Alzheimer's από τα ProtoPFormer, ViT-NeT, και Swin x Grad-CAM.

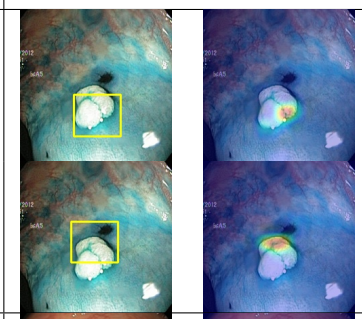
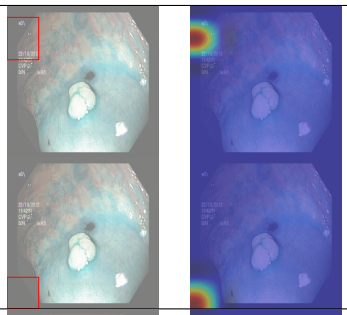
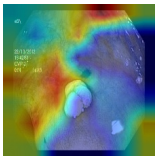
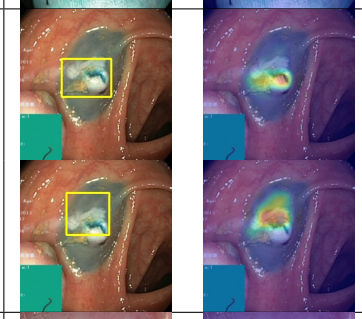
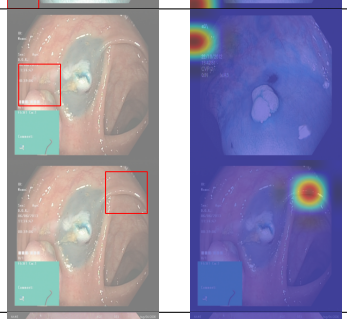
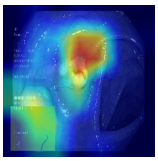
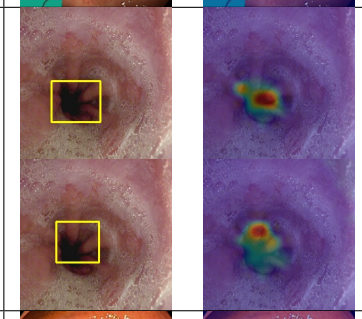
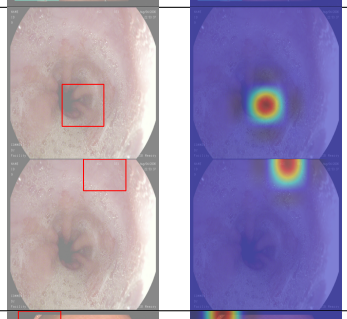
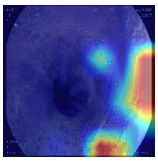
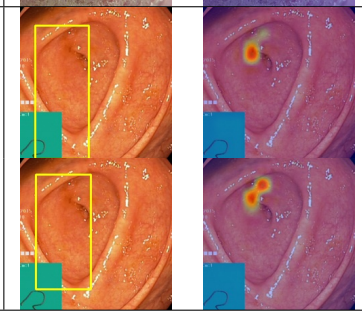
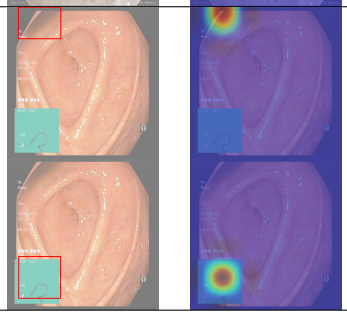
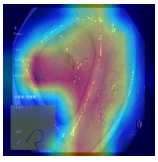
Original Image	ProtoPFormer	ViT-NeT	Swinx x Grad-CAM
Dyed Lifted Polyps			
Dyed Resection Margins			
Esophagitis			
Normal Cecum			

Table 6: Οπτικοποιήσεις που δημιουργήθηκαν για το σύνολο δεδομένων Kvasir από τα ProtoPFormer, ViT-NeT, και Swin x Grad-CAM.

Original Image	ProtoPFormer	ViT-NeT	Swinx x Grad-CAM
Normal Pylorus			
Normal Z Line			
Polyps			
Ulcerative Colitis			

Table 7: Οπτικοποιήσεις που δημιουργήθηκαν για το σύνολο δεδομένων Kvasir από τα ProtoPFormer, ViT-NeT, και Swin x Grad-CAM.

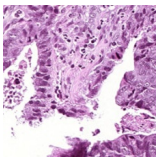
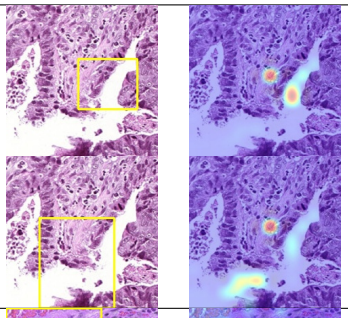
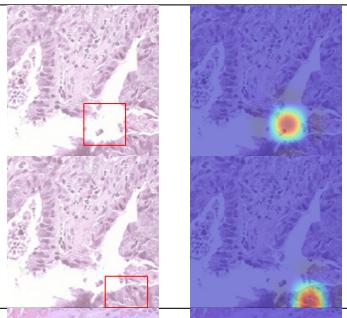
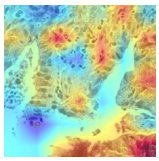
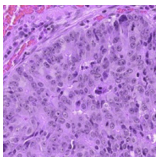
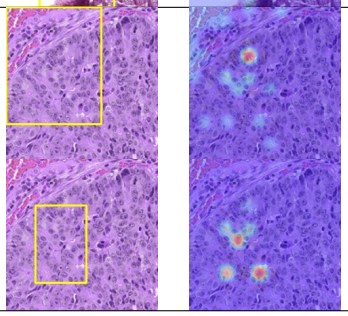
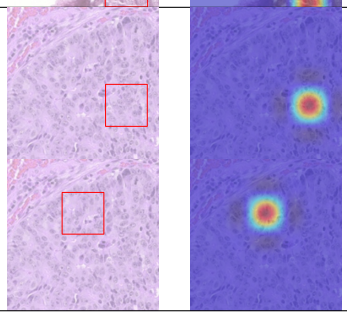
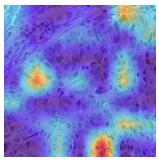
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
MSIMUT 			
MSS 			

Table 8: Οπτικοποιήσεις που δημιουργήθηκαν για το σύνολο δεδομένων TCGA από τα ProtoPFormer, ViT-NeT, και Swin x Grad-CAM.

Συμπεράσματα

Στην παρούσα διπλωματική εργασία, εξετάστηκαν τρεις αρχιτεκτονικές Vision Transformer—ProtoPFormer, ViT-NeT και Swin Transformer σε συνδυασμό με Grad-CAM—σε τέσσερα ιατρικά σύνολα δεδομένων, περιλαμβάνοντας MRI, CT, ιστοπαθολογικές εικόνες και εικόνες από ενδοσκοπήσεις. Τα αποτελέσματα έδειξαν ότι το ProtoPFormer παρουσίασε την υψηλότερη ακρίβεια σε όλα τα σύνολα δεδομένων, προσφέροντας σημαντικά πλεονεκτήματα στην ανάλυση ιατρικών εικόνων. Αντίθετα, τα ViT-NeT και Swin Transformer εμφάνισαν χαμηλότερη επίδοση, με προβλήματα στην ερμηνεία των εικόνων, γεγονός που υπογραμμίζει την ανάγκη για καλύτερες τεχνικές προεπεξεργασίας και πιθανώς εξειδικευμένες αρχιτεκτονικές.

Για το μέλλον, συνιστάται η ανάπτυξη βελτιωμένων τεχνικών προεπεξεργασίας για την καλύτερη εστίαση των μοντέλων σε κλινικά σημαντικές περιοχές, καθώς και η διερεύνηση νέων αρχιτεκτονικών Transformer και υβριδικών μοντέλων για τη βελτίωση της ακρίβειας και της ερμηνευσιμότητας. Επίσης, η ενσωμάτωση πολυτροπικών δεδομένων και η εξερεύνηση τεχνικών transfer learning μπορούν να επεκτείνουν την εφαρμογή των μοντέλων σε διάφορα ιατρικά περιβάλλοντα.

Chapter 1

Introduction

In modern technology, few advancements have reshaped society so profoundly as artificial intelligence (AI). At its core, AI empowers machines with the ability to mimic and even surpass human intelligence, enabling them to perceive, reason, learn, and act autonomously. This interdisciplinary field intersects computer science, mathematics, cognitive psychology, neuroscience, and philosophy, among other disciplines, to unlock the mysteries of intelligence and consciousness. The origins of AI trace back to the mid-20th century when pioneers such as Alan Turing [44] and John McCarthy [45] laid the conceptual groundwork for intelligent machines. Since then, AI has undergone remarkable evolution, driven by exponential growth in computational power, vast datasets, and innovative algorithms.

One of the most important branches of artificial intelligence, computer vision enables machines to interpret and understand the visual world [1]. From facial recognition systems [2] and autonomous driving [3] to augmented reality applications [4] and medical imaging [5], computer vision permeates diverse industries, revolutionizing how we perceive and interact with our environment.

1.1 Motivation

The well-being of individuals and society as a whole will forever remain a top priority. Over the years, technology has equipped the scientific community with increasingly advanced medical instruments, providing insights into the inner workings of the human body. Central to this technological evolution is medical imaging, which involves visualizing the human body using different imaging modalities, aimed at diagnosis and treatment. Every technique provides distinct details regarding the specific part of the body under examination or undergoing treatment, related to potential illnesses, injuries, or the effectiveness of medical interventions. Thus, medical imaging is frequently used for monitoring diseases already diagnosed and/or treated. There exist several medical imaging modalities, each offering unique advantages and applications to patients. Some of the most common and the ones that are going to be used for the purposes of this thesis are: Computed Tomography (CT), Magnetic Resonance Imaging (MRI) [46], histopathological images [47] and endoscopic images [48].

While these innovations have significantly improved healthcare overall, by providing more accurate visualizations of the inner topology of the human body, it is essential to acknowledge that diagnostic challenges still persist. Ultimately, diagnosis remains an exclusively human-driven decision-making process, as it depends on the medical professional's personal judgement [6]. However, inaccuracy and uncertainty should not be present in a field were mistakes cannot be forgiven.

The benefits of artificial intelligence have been discussed in detail in the medical literature [49–51]. Such benefits include applications in the three major areas of early detection and diagnosis, treatment and prediction of the course and outcome of a medical condition. By using complex algorithms that 'learn' features from a vast volume of medical data, AI can have an assisting part in the clinical practice [6]. Additionally, its self-adjusting and self-correcting capabilities can significantly improve its accuracy through feedback provided by

the model. In this way, AI systems can help minimize medical errors, which are ubiquitous and their costs, significant [7–9].

However, alongside the developments brought by AI, there arises a critical issue: interpretability. EU legislation recently required that AI algorithms utilizing user-level predictors for decision-making must offer explanations, particularly when the results have a substantial impact on the individual's outcome [52]. As a result, there has been a growing interest in the field of explainable artificial intelligence (XAI), which focuses on developing AI systems that also offer transparent explanations of their decisions, making it easier for humans to comprehend and trust them [10, 53]. Yet, this black box nature of deep learning models [54, 55] still remains unexplored and poses a challenge in the medical domain [52].

In computer vision, recent advancements in AI architecture have introduced Vision Transformers (ViTs), offering a novel approach to analyzing visual data. Unlike traditional Convolutional Neural Networks (CNNs), ViTs rely on self-attention mechanisms, allowing them to capture long-range dependencies in images more effectively [11]. While ViTs have been successfully applied to medical imaging data, showing great performance, interpretable ViTs are yet to be explored in medical image analysis.

1.2 Contribution

Through this research, we seek to bridge the gap between explainable AI techniques and clinical practice, by investigating the potential of utilizing interpretable Vision Transformers in the field of medical imaging.

To that end we:

- Apply recent developed explainable ViTs to four multimodal medical datasets.
- Apply a non interpretable ViT to the same datasets.
- Compare the results of the above ViTs and come to conclusions regarding the relationship between performance and interpretability.
- Present survey results evaluating the interpretability of the models, as assessed by medical students.

1.3 Thesis Outline

The remainder of this thesis is structured as follows. Chapters 2 and 3 are dedicated to providing the reader with the theoretical background essential for understanding our experiments. Chapter 2 covers the historical progression leading to the development of explainable Vision Transformers (ViTs), from the original transformer to ViTs and ultimately to interpretable ViTs. In chapter 3, we describe the related work, offering detail regarding specific recently developed interpretable ViT architectures. Chapter 4 describes the methodology of the experiments conducted with the selected ViT models and presents the corresponding results. In conclusion, Chapter 5 concludes this study by summarizing our discoveries and suggesting potential future avenues for developing an exclusive medical interpretable ViT.

Chapter 2

Historical Progression: From CNNs to Interpretable ViTs

In chapter 2, we describe the history behind the development of explainable Vision Transformers. We start by presenting the foundational concepts of Convolutional Neural Networks to the advancements leading to the emergence of ViTs. We make an introduction regarding diverse interpretability techniques in the field of computer vision and their integration with ViTs.

Contents

2.1 Convolutional Neural Networks

2.1.1 Artificial Neural Networks

First emerged in the 1980s, Artificial Neural Networks (ANNs) are computational models, simulating the biological neural networks of the human brain [12]. As shown in Figure 2.1.1, ANNs consist of interconnected nodes, also called neurons, arranged in layers. Each neuron receives input signals, processes them using an activation function, and produces an output signal, which may be passed to neurons in the next layer, if one exists.

The learning process in an ANN includes repeatedly adjusting its parameters to minimize the divergence between predicted and actual outputs. This divergence is calculated by the loss function, when input data is propagated through the network. Then, based on this error rate, backpropagation guides weight updates to minimize the loss. This iterative process continues over multiple epochs, gradually improving the network's ability to capture complex patterns in the data [13–15].

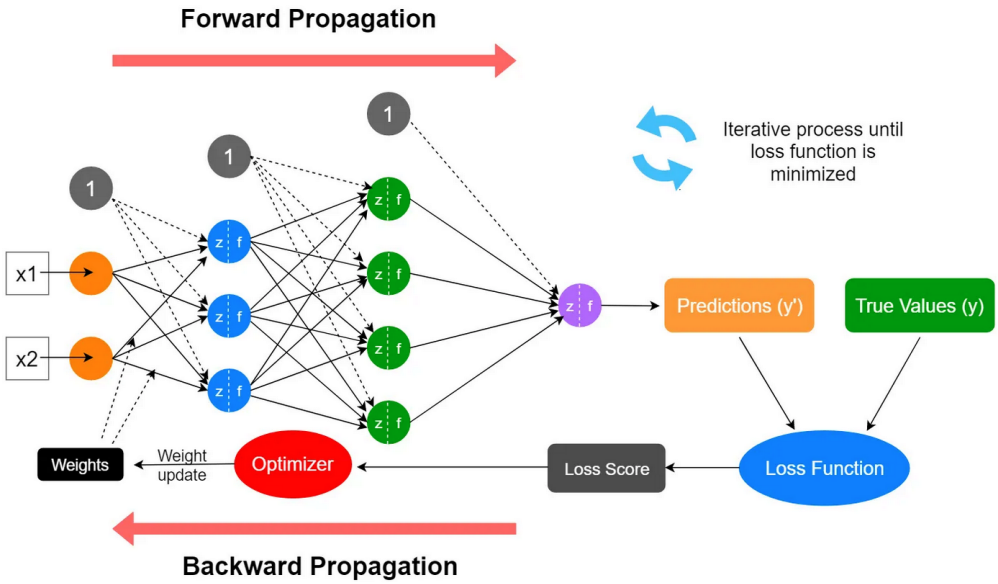


Figure 2.1.1: A four layered feedforward neural network (FNN), consisting of an input layer, two hidden layers and an output layer. This is a basic structure of a number of common ANN architectures. "Overview of a neural network's learning process" [56]

In the field of computer vision, the inability of traditional ANNs to handle the computational complexity needed to process image data is one of their biggest drawbacks. For example, when a large, 64×64 colored image input is taken into account, the number of weights on a single first-layer neuron rises significantly to 12,288. However, simply increasing the number of hidden layers in the network is not a practical solution. Doing so would exponentially increase the computational power and time required to train such a large model. It could also lead to overfitting, where the model would become too specialized to the training data, resulting in poor performance on unseen data [12].

2.1.2 CNN architecture

Convolutional Neural Networks (CNNs) represent a specialized class of Artificial Neural Networks primarily used for processing visual data [16, 17]. CNN applications include image classification, image semantic segmentation and object detection within images, etc [57]. One of the main differences between CNNs and ANNs is that the layers of the CNN are made up of neurons arranged in three dimensions: the input's height and width, and depth. The neurons in a given layer, in contrast to conventional ANNs, will only connect to a specific region of the layer before it. For the example given earlier, the dimensionality of the input would

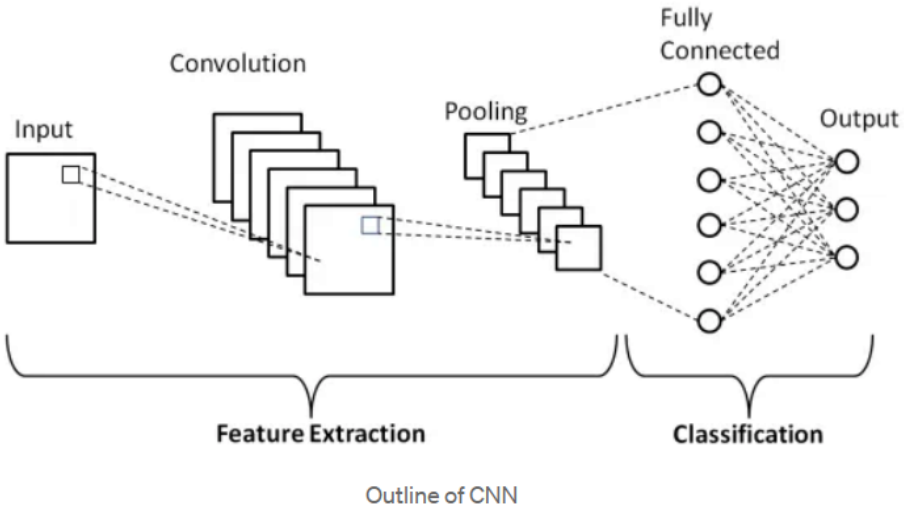


Figure 2.1.2: A simple CNN architecture, consisting of four layers.
"Binary Image classifier CNN using TensorFlow" [58]

be $64 \times 64 \times 3$ ($height \times width \times channels$) and it would result in a final output layer of dimensionality of $1 \times 1 \times n$, n being the possible number of classes. This dimensionality reduction into a smaller amount of class scores enables CNNs to effectively capture and learn complex hierarchical features from images while significantly reducing the computational complexity compared to traditional ANNs [12].

A Convolutional Neural Network (CNN) typically consists of several layers arranged sequentially. The core layers include convolutional layers, pooling layers, and fully-connected layers, as shown in Figure 2.1.2.

Convolutional Layers are the primary building blocks of a CNN. They apply a series of learnable filters to the input data, extracting features such as edges, textures, and patterns (Figure 2.1.3). Each filter slides across the input data, performing element-wise multiplication and summation to produce activation maps. Pooling layers follow convolutional layers and serve to reduce the spatial dimensions of the feature maps while retaining important information. Finally, fully-connected layers connect every neuron in one layer to every neuron in the next layer. These layers are responsible for creating a flattened vector representation of the information extracted by the previous ones, by producing class scores from the activations, to be used for classification [12].

2.2 Attention: Transformers

2.2.1 Background

However, when capturing spatial dependencies in input data, Convolutional Neural Networks have limited receptive field. These limitations when modeling long-range dependencies arise from the fixed size convolutional kernels and pooling operations, which restrict the scope of information aggregation to a local region of the input. In 2017, a novel architecture, known as Transformers [18] emerges, which, unlike CNNs, allows flexible and context-aware representation learning across the entire input sequence [59]. Overtaking other neural models, such as convolutional and recurrent neural networks, in both natural language understanding and natural language generation, the Transformer has quickly become the dominant architecture for sequence modeling tasks, as well as the building block for creating more complex extensions [60].

2.2.2 Multi-Head Attention

The Transformer model relies entirely on a self-attention mechanism, which connects various positions within a single sequence to calculate its representation. These representations are then used to predict the distri-

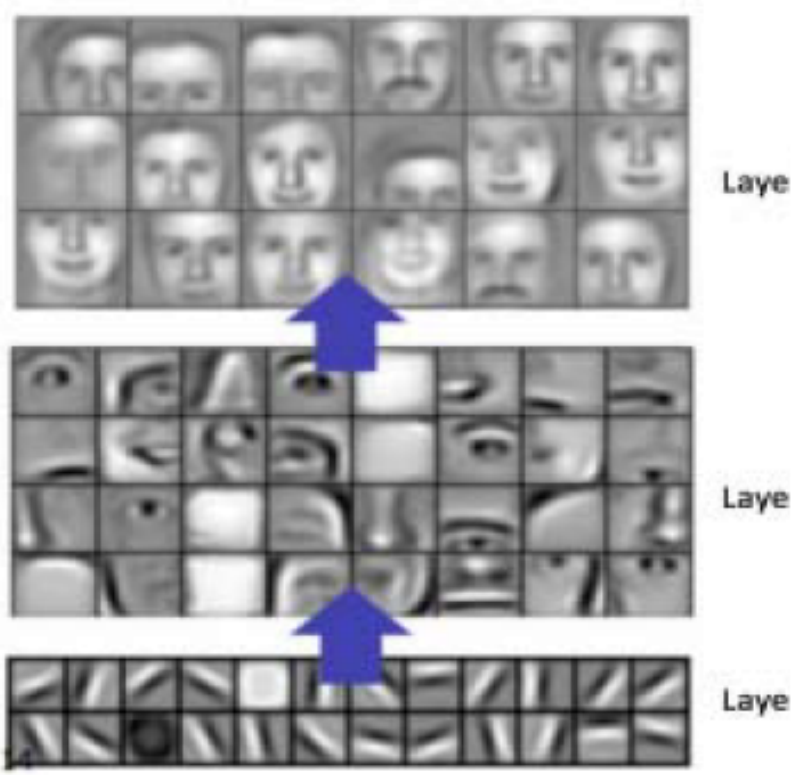


Figure 2.1.3: Learned features from different convolutional layers of a CNN.
"Understanding of a Convolutional Neural Network" [17]

bution of subsequent information as the model predicts the output sequence symbol-by-symbol, facilitating efficient parallel training [18, 59, 60]. In the QKV (Query, Key, Value) attention mechanism, each input element (or token) is associated with three learned vectors: the query vector, the key vector, and the value vector, which are obtained by linear transformations of the input embeddings. The "Scaled Dot-Product Attention", utilized by the Transformer model, is a variation of the above mechanism, wherein attention scores are computed as the dot product between the query and key vectors, divided by the square root of the dimension of the latter.More specifically, it combines queries Q, keys K, and values V as follows:

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

By doing a softmax, the highest scores get heightened and the lowest scores are depressed, which allows the model to be more confident on which words to attend to. The $\frac{1}{\sqrt{d_k}}$ is called the scaling factor. It normalizes the dot products, preventing them from becoming too large or too small, which can lead to issues like vanishing or exploding gradients during training [19].

To enable multi-head attention computation, which is the one introduced by the Transformer model, the query, key, and value vectors are split into sub-vectors before applying the self-attention mechanism. This splitting allows the model to focus on different aspects of the input sequence independently, enhancing its ability to capture diverse relationships within the data. Each split vector undergoes the same self-attention process individually, with each process referred to as a head. In essence, each head functions like a separate attention mechanism, enabling the model to attend to various parts of the input simultaneously. The outputs from each head are then concatenated into a single vector before being passed to the final linear layer. In theory, each head would learn different aspects of the input, thereby enhancing the representation power of the Transformer model [18–20].

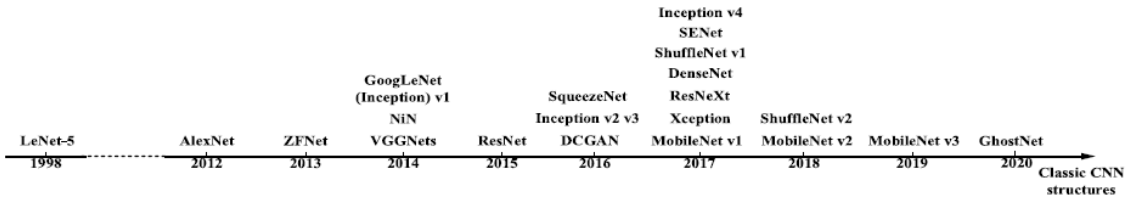


Figure 2.1.4: Classic CNN models through time.
"A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects" [16]

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

To sum it up, as shown in the left part of the Figure 2.2.1, multi-head attention is a module in a transformer network that computes the attention weights for the input and produces an output vector with encoding information on how each word should attend to all other words in a sequence.

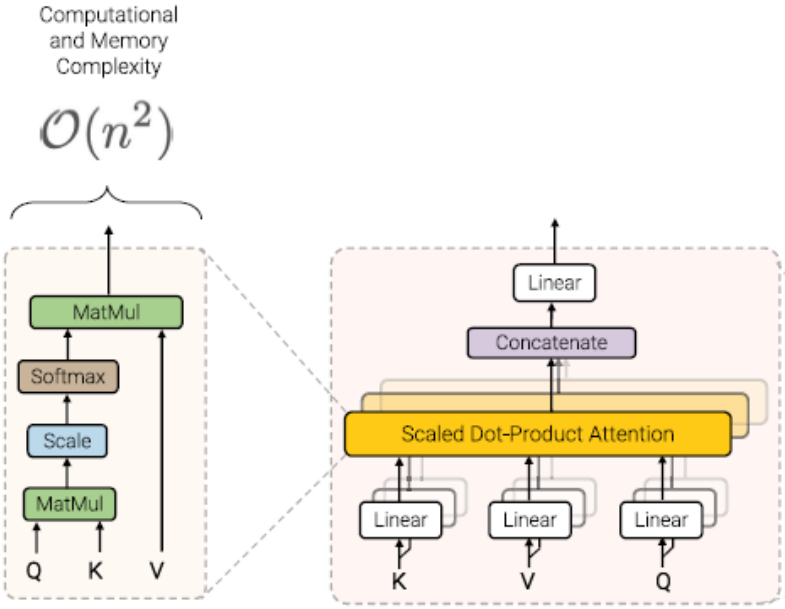


Figure 2.2.1: (left) Scaled-dot product attention. (right) Multi-head attention.
"Efficient Transformers: A Survey" [20]

2.2.3 Model Architecture

The Transformer model comprises two integral components: the encoder and the decoder, each serving essential functions in the model's architecture for sequence processing and generation.

Encoder: The encoder component of the Transformer model is responsible for processing the input sequence and extracting contextual information from each token. It consists of multiple layers, six to be exact, each containing two main sublayers: a multi-head self-attention mechanism and a position-wise fully-connected feed-forward neural network with ReLU activations [20]. In the former, each token attends to all other tokens in the input sequence, allowing the model to capture dependencies between elements at different positions [59]. Following, the position-wise feed-forward network applies a non-linear transformation to each

token's representation independently, further refining the features extracted by the self-attention mechanism. Residual connections and layer normalization are applied after each sublayer to facilitate stable training and improve gradient flow [18, 19].

$$X_A = \text{LayerNorm}(\text{MultiheadAttention}(X, X)) + X$$

$$X_B = \text{LayerNorm}(\text{PositionFFN}(X_A)) + X_A \text{ [20]}$$

Decoder: The decoder component of the Transformer model is responsible for generating the output sequence based on the representations produced by the encoder and the previously generated tokens in the output sequence. It shares a similar architecture with the encoder, consisting of multiple layers (stack of $N = 6$ identical layers [18]) with self-attention mechanisms and position-wise fully-connected feed-forward neural networks. The decoder also includes a mask to prevent tokens from attending to future tokens in the output sequence during training, ensuring that the model generates each token based only on the previously generated ones. Residual connections and layer normalization are applied after each sublayer to facilitate stable training and improve gradient flow [18].

2.2.4 Applications and Famous Transformer Models

Transformers were first introduced within the context of sequence-to-sequence machine translation in natural language processing and the majority of their early improvements still remain within the domain of language. However, their influence extends far beyond language alone, ranging from speech, to vision and reinforcement learning [20].

The T5 (Text-to-Text Transfer Transformer) [61], which was developed by Google in 2019, is among the most popular transformer encoder-decoder models. It is designed to perform numerous NLP tasks, such as question answering, summarization, language translation, by converting input text to output text.

BERT, standing for Bidirectional Encoder Representations from Transformers, [62] is another groundbreaking transformer encoder model that has significantly influenced the field of NLP. Released by Google in 2018, BERT brought about a shift by introducing bidirectionality in language representation. Unlike previous models that processed text in one direction, BERT considers both left and right context during training, enabling it to capture deeper semantic meaning and context. This bidirectional understanding revolutionized various NLP tasks, including question answering and language inference, without the need for significant tsk-specific alterations to the architecture.

GPT-3 (Generative Pre-trained Transformer 3) [63] is the third generation of the GPT series, developed by OpenAI in 2020, aimed to produce human-like text. With 175 billion parameters, GPT-3 is one of the largest and most powerful language models ever created, showing great performance in various tasks, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and more.

2.3 Vision Transformers

Inspired by the Transformer's scaling success in NLP tasks, researchers began exploring its applicability in other domains, such as computer vision. In 2017, Dosovitskiy et al. demonstrated that CNN dependency is unnecessary, as a standalone transformer applied directly to sequences of image patches can excel in image classification tasks [11]. Embracing the Transformer architecture, Vision Transformers (ViTs) revolutionized the conventional approach to image analysis. Breaking down the steps of how an input image is processed by the ViT involves several key stages that are going to be discussed in the following section.

2.3.1 Transformers in Vision

Interest in employing transformers for high/mid-level computer vision tasks has been recently increasing [64]. This growth in interest encompasses a broad spectrum of applications, including object detection [65–68], segmentation (panoptic [69], instance [70], semantic [71, 72], medical [73]), lane detection [74], and pose estimation [75].

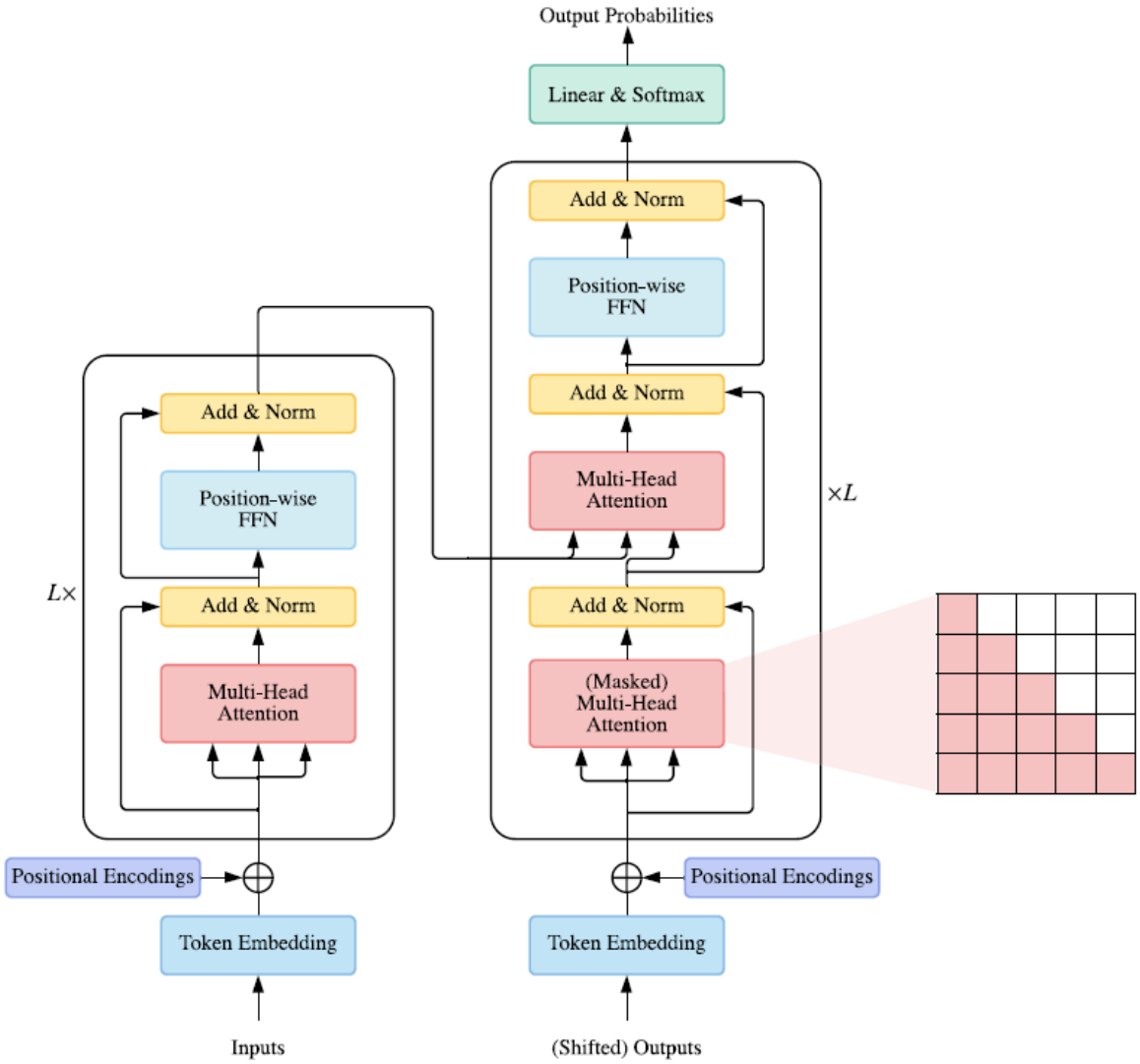


Figure 2.2.2: Illustration of the architecture of a standard Transformer model.
"A survey of transformers" [19]

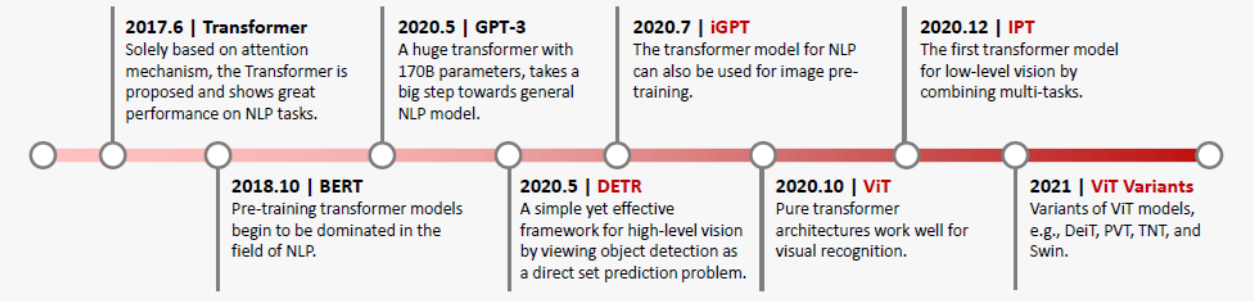


Figure 2.3.1: Significant moments in the evolution of transformer technology. The vision transformer models are highlighted in red.
"A Survey on Vision Transformer" [64]

There's a limited number of studies that employ transformers in low-level vision tasks like image super-resolution and generation. These tasks involve producing images as outputs, which poses a greater challenge compared to high-level vision task, where the outputs are labels or bounding boxes [64].

In computer vision, especially in video-related tasks, where both spatial and temporal dimensions are crucial, the Transformer architecture has also found applications [64]. These include tasks such as frame synthesis [76], action recognition [77], and video retrieval [78].

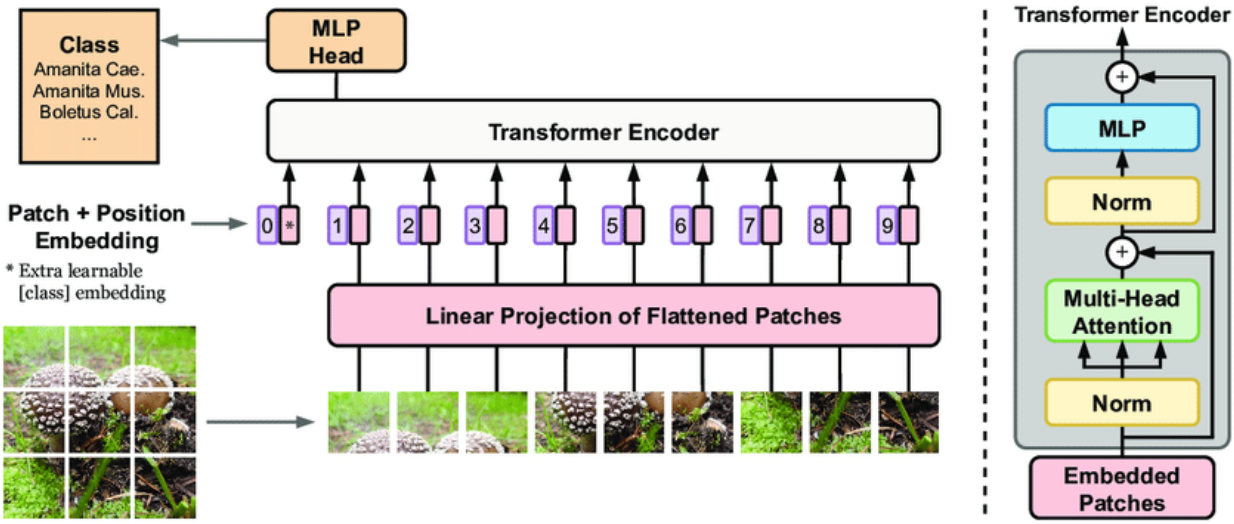


Figure 2.3.2: ViT overview. First the image is split into patches of fixed size, to be used as tokens. Then, the flattened patches are linearly projected and position embeddings are added to define the position of the patches within the image. Finally, the patches along with the embeddings are fed to a standard Transformer encoder. For classification purposes, an extra learnable class embedding is added to the sequence.
"Automatic fungi recognition: deep learning meets mycology" [21]

2.3.2 ViTs: The Method

The process of an input image, entering the Vision Transformer, as introduced by Dosovitskiy et al. [11], involves the steps described below:

1. Initially, the original 2D image, $x \in \mathbb{R}^{H \times W \times C}$ with a resolution of (H, W), is partitioned into fixed-size non-overlapping patches, effectively breaking down the visual information into manageable units. Each patch, $x_p \in \mathbb{R}^{N \times (P^2 C)}$ has a resolution of (P, P), where C is the number of channels and $N = \frac{H \times W}{P^2}$ the total number of patches, which is an efficient input sequence length for the Transformer architecture.

2. The patches are flattened and linearly projected to D dimensions, the constant size of the latent space of all the Transformer layers. This projection's output is called "patch embeddings".
3. A special learnable embedding is added to the start of the sequence $z_0^0 = x_{\text{class}}$, also known as a classification token, allowing the model to make predictions, regarding the classification, based on the aggregated information from all the patches.
4. Positional encodings are incorporated into the patch embeddings before feeding them into the Transformer layers. These encodings provide spatial information by representing the relative positions of patches within the image grid. Standard learnable 1D embeddings are used, since the 2D representation has not shown notable improvements in performance.
5. The patch embeddings, along with their positional encodings, are ready to be fed into a standard transformer encoder. The architecture of the encoder follows the pattern we have previously discussed: comprising layers of multi-head attention and MLP blocks that alternate. Layer normalization and residual connections are applied before and after each block accordingly.

As observed, the Vision Transformer displays a notably reduced image-specific inductive bias in contrast to CNNs, with only the MLP layers being both local and translationally equivariant.

Additionally, in the context of a hybrid model architecture, rather than utilizing raw image patches, feature maps extracted from a CNN can serve as the input sequence.

2.4 Interpretability in Vision Transformers

While Vision Transformers (ViTs) have shown remarkable performance across various computer vision tasks, understanding their inner workings remains a significant challenge. The primary difficulty lies in the fact that the attention generated in each layer becomes intricately intertwined with subsequent layers, posing a challenge in visually determining the proportional impact of input tokens on the final predictions [79]. Interpretability, or the ability to comprehend why a model makes certain predictions, is crucial for deploying ViTs in real-world applications, especially where transparency and trust are paramount. There exist a number of different interpretability techniques that are applied on Vision Transformers, each offering insights into how the model processes and understands visual information. These techniques, that range from attention visualization to map generation and attribution propagation [28], are going to be further discussed in the next chapter along with specific recently developed interpretable ViT models.

Chapter 3

Related Work

In Chapter 3, we delve into the domain of interpretability within Vision Transformers, focusing on three key aspects. Firstly, we explore an overview of explainability techniques, highlighting their significance in understanding complex deep learning models. Secondly, we narrow our lens to examine existing interpretability techniques specifically developed for the domain of medicine. Lastly, we dive into the latest advancements in interpretable ViT architectures, discussing innovative approaches that aim to make computer vision tasks more transparent and understandable.

Contents

3.1 Explainability Methods

Explainability encompasses a wide range of methods and techniques, including feature importance methods, rule-based explanations [80–82], counterfactual explanations [83, 84], and prototypes [85]. These approaches are applied across various modalities, such as text, images, sound, and graphs, as well as in multimodal systems that combine these data types [86]. In this thesis, however, the focus will be on explanations within the computer vision domain, specifically for images, with a particular emphasis on heatmaps as a subset of feature importance methods.

Explainability in computer vision. Before discussing specific explainability methods, it's essential to understand the broader landscape of heatmap generation techniques for identifying local relevance in images processed by CNNs. These methods typically fall into the following main categories: gradient methods and attribution methods [28].

Gradient methods, such as attention rollout and GradCAM, compute gradients of the model's output with respect to the input image pixels. They measure how changes in the input image pixels affect the output prediction. Higher gradients indicate pixels that have a stronger influence on the outcome [28].

Attribution propagation methods, such as the Layer-wise Relevance Propagation (LRP) method, rooted in the Deep Taylor Decomposition (DTD) framework, aim to systematically decompose the decision-making process of neural networks. By recursively attributing contributions from previous layers to elements of the input, these methods provide an understanding of how each layer and feature influences the network's decision [28].

Methods that do not belong in the categories described above and fall into the interpretability technique of map generation include saliency and relevance based methods, highlighting the areas of the given image, discriminative with respect to the given class. [28, 87, 88]

3.1.1 Attention rollout and Attention flow

Attention rollout [22] offers an approach to trace the flow of information from the input layer to higher-layer embeddings within a Transformer model. For a Transformer with L layers, attention is computed from all positions in layer l_i to all positions in layer l_j, where $j < i$.

An attention graph is a graphical representation used to visualize the flow of attention within a neural network, particularly in models with an attention mechanism, such as Transformers. In an attention graph, nodes represent different positions or elements in the input sequence, while edges represent the attention weights assigned between these positions. Each edge's weight indicates the strength or importance of the attention connection between the corresponding nodes. In our case, in a self-attention mechanism within a Transformer layer, each input token attends to every other token in the sequence, and the attention weights determine how much focus or influence each token has on the others.

Visualized as an attention graph, a path from node v at position k in layer l_i to node u at position m in layer l_j represents a series of connecting edges. By multiplying the weights of all the edges in a specific path, we can determine the amount of information that is propagated from v to u. This is because the weight of each edge represents the amount of information that is exchanged between two nodes. If multiple paths exist between two nodes within an attention graph, summing across all potential paths connecting these nodes is necessary to calculate the proportion of information flow between them.

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}) & \text{if } i > j \\ A(l_i) & \text{if } i = j \end{cases}$$

where $\tilde{A}$ is attention rollout, A is raw attention and a matrix multiplication is performed.

Attention flow. Viewing the attention graph as a flow network, with edge capacities serving as attention weights, enables the computation of maximum attention flow from any layer node to input nodes using standard maximum flow algorithms. This maximum-flow-value is used as an estimate of attention to input nodes. Unlike the attention rollout method where the weight of a single path is the product of edge weights,

in attention flow, it's determined by the minimum value of edge weights along the path, since there might be path overlapping [22].

3.1.2 GradCAM

In 2016, a procedure for generating class activation maps (CAM) [24] was introduced using global average pooling in CNNs [89]. Given a network mostly made up of convolutional layers, global average pooling is performed on the feature maps produced by the last convolutional layer. These pooled features are then utilized as input features for a fully-connected layer responsible for generating the desired output (classification or otherwise) [24].

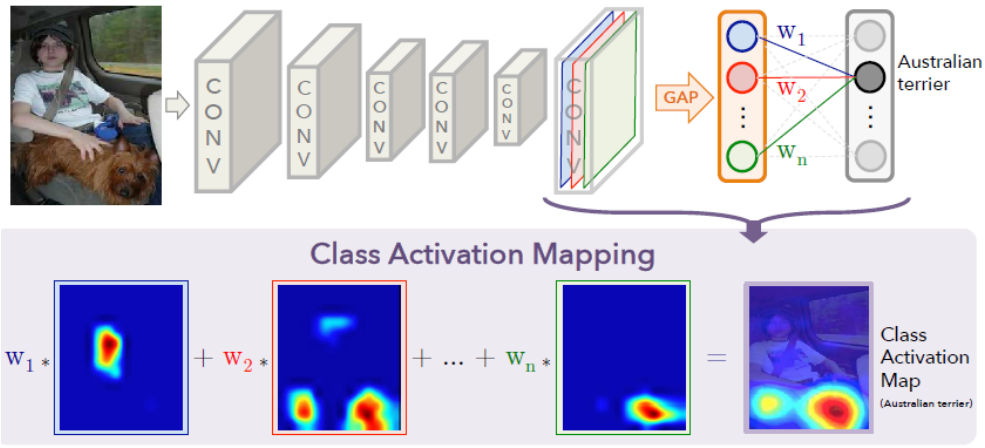


Figure 3.1.1: Class Activation Mapping (CAM): Illustrating how predicted class scores are utilized to generate class activation maps (CAMs), highlighting discriminative regions within the image. "Learning Deep Features for Discriminative Localization" [24]

If there are n filters in the last convolutional layer, corresponding to n feature maps, the activation map for a specific output class is generated by combining all n feature maps with learned weights.

To learn these weights:

Step 1: Apply global average pooling to each feature map, resulting in n scalars (GAP outputs), k_1, k_2, ..., k_n.

Global Average Pooling (GAP) [89] is a pooling technique used in convolutional neural networks (CNNs) to condense feature maps. Unlike traditional pooling methods that select the maximum value within local regions, GAP calculates the average value of each feature map across its entire spatial extent.

Step 2: Learn a linear model from these GAP outputs to the class labels. For each of the N output classes, N linear models with weights w_1, w_2, ..., w_n are learned.

Step 3: With the obtained weights for each class, weight the feature maps to generate the class activation maps. Different weighted combinations of the same feature maps produce class activation maps for different classes.

In mathematical terms, the CAM model computes the class score for an output class c as follows:

$$y_c = \frac{\sum_k w_k^c}{Z} \sum_i \sum_j A_{ij}^k$$

where A_{ij}^k represents the pixel at location (i, j) in the k-th feature map, Z is the total number of pixels in the feature map and w_k^c is the weight of the k-th feature map for class c.

CAM can only be implemented with networks that use a single fully-connected layer before the output layer. So, in a case with multiple fully-connected layers, they usually get replaced with convolutional ones and the network need to be re-trained [23].

Figure 3.1.2: The class activation maps (CAMs) depict distinctive image regions crucial for image classification, such as the animal's head for the briard class and the plates in a barbell.
"Learning Deep Features for Discriminative Localization" [24]

Grad-CAM is a generalization of CAM, developed to be applied to any CNN-based architecture [23]. It builds upon CAM by utilizing gradient information from the final convolutional layer. It computes the gradients of the predicted class score with respect to the feature maps of this layer. These gradients are then global-average-pooled to obtain importance weights for each feature map, representing the importance of each feature map in making the final prediction.

$$w_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Finally, GradCAM produces a weighted combination of the feature maps using the importance weights. To retain only the positive correlations in the final activation map, a ReLU function is applied on the weighted combination of feature maps.

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

Grad-CAM visualizations effectively localize relevant image regions based on class discrimination but lack the ability to provide fine-grained importance like pixel-space gradient visualization methods such as Guided Backpropagation and Deconvolution. To address this limitation, Guided GradCAM [23, 25], a fusion approach is proposed, combining the strengths of both techniques through pointwise multiplication. This fusion results in high-resolution visualizations that accurately identify important features specific to the predicted class while maintaining class discrimination.

3.1.3 LIME

The Local Interpretable Model-agnostic Explanations (LIME) method [26] is a technique designed to provide explanations for the predictions made by complex machine learning models. At its core, LIME defines an explanation as a model g belonging to a class G comprising models that are potentially interpretable, including linear models or decision trees. The complexity of the explanation model is quantified by a measure denoted as $\Omega(g)$, which could represent the depth of a decision tree or the number of non-zero weights in a linear model. This allows for a balance between interpretability and complexity, ensuring that the resulting explanations are understandable while still capturing the essential features of the original model's behavior.

The model being explained, denoted as $f : \mathbb{R}^d \to \mathbb{R}$, represents the function that maps input features to predictions. LIME aims to approximate the behavior of this model locally, focusing on a specific instance x for which an explanation is sought. A proximity measure $\pi_x(z)$ is used to quantify the similarity or proximity

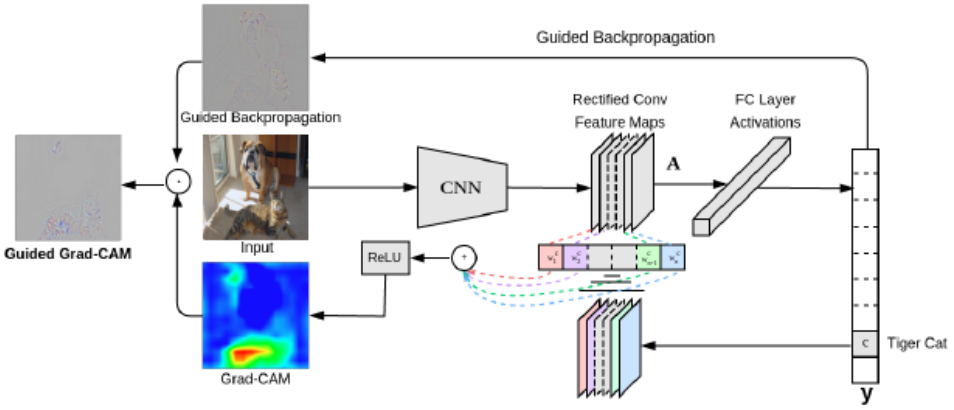


Figure 3.1.3: Guided GradCAM overview: Grad-CAM begins with an input image and a specified category, such as 'tiger cat'. The image is processed through the model to obtain raw class scores. Gradients are then adjusted to emphasize the target class while setting others to zero. This modified signal is then propagated backward to the relevant convolutional feature map, allowing computation of the coarse Grad-CAM localization represented by a blue heatmap. Lastly, the heatmap is combined with guided backpropagation through pointwise multiplication, resulting in Guided Grad-CAM visualizations known for their high resolution and ability to discriminate between classes.
"Grad-CAM: Why did you say that?" [23]

between an instance z and the instance x, defining the locality around x. This locality-aware approach allows LIME to provide explanations that are relevant to the specific context of each prediction.

The heart of LIME lies in the locality-aware loss function $L(f, g, \pi_x)$, which measures how well the explanation model g approximates the behavior of the original model f within the defined locality. The goal is to minimize this loss while simultaneously minimizing the complexity of the explanation model. This is achieved through optimization, where the explanation $\hat{g}(x)$ is obtained by minimizing the combined loss function and complexity measure:

$$\hat{g}(x) = \text{argmin}_{g \in G} \left(L(f, g, \pi_x) + \Omega(g) \right)$$

To approximate the locality-aware loss function, LIME employs a sampling approach. Samples are drawn around the instance x using the proximity measure π_x, and these samples are then used to approximate the loss function. Despite the potential presence of sampling noise, LIME remains robust due to the weighting of samples by the proximity measure.

3.1.4 Layer-Wise Relevance Propagation

Layer-Wise Relevance Propagation (LRP) [27, 28] propagates relevance of the prediction $f(x)$ backwards in the neural network, following the Deep Taylor Decomposition framework. In the propagation mechanism, utilized by the LRP, the information received by a neuron must be equally redistributed to the lower layer (conservation property). Supposing j and k represent neurons at consecutive layers of the neural network, the following rule describes the relevance scores' $((R_k)^k)$ propagation onto the neurons of the lower layer:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k.$$

where z_{jk} quantifies the degree to which neuron j has influenced the relevance of neuron k, and the denominator ensures the preservation of the conservation property, described above. The propagation process concludes when reaching the input features.

Applying the aforementioned rule to all neurons in the network allows for the straightforward expression of the layer-wise conservation principle:

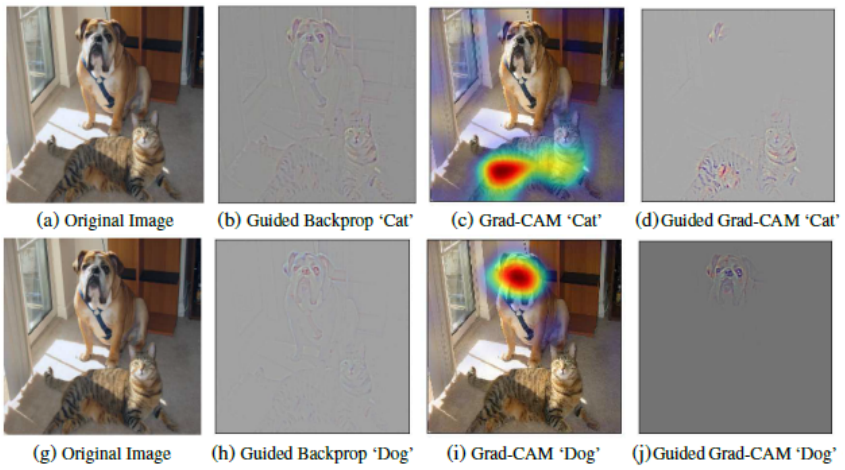


Figure 3.1.4: Comparison of Visualizations: Original image of a cat and a dog alongside visualizations generated using Guided Backpropagation, Grad-CAM, and Guided Grad-CAM techniques, showcasing distinct highlighting of salient features and class-specific regions.
"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" [25]

$$\sum_j R_j = \sum_k R_k$$

as well as the global conservation property:

$$\sum_i R_i = f(x)$$

3.2 Interpretable ViT Networks

After discussing post hoc explainability methods, which involve techniques applied externally after a model has made its predictions, we now turn our attention to reviewing specific interpretable ViT models, where interpretability is inherently built in the model architecture itself.

3.2.1 ViT-NeT

To achieve a better trade-off between model interpretability and performance, in 2022, ViT-NeT [29] was introduced, combining the Swin transformer encoder [30], as a backbone for feature representation, with a neural tree decoder. The method utilizes a shifted window approach to detect both small and large object variations at the same time, effectively managing both inductive bias and computational complexity.

The Swin Transformer Encoder. Unlike traditional transformers which rely on a single-window approach, Swin Transformer adopts a hierarchical feature encoding strategy akin to a feature pyramid found in CNNs. By dynamically adjusting window sizes and processing multiple patches with self-attention, Swin Transformer generates attention maps capable of detecting both small and large objects within images. Notably, while vanilla transformers perform quadratic computations on input images, Swin Transformer operates linearly, resulting in increased model size without significant computational overhead and improved inference speed.

The Neural Tree Decoder. The NeT comprises sets of nodes ($N()$), leaves $L()$), and edges ($E_{i,j}()$) connecting parent node i to child node j. With the use of a perfect binary tree, each internal node has exactly two child nodes: N_{2i} and N_{2i+1}. With the encoder output denoted as z1=f(x;ϑ), the proposed NeT makes the final class label predictions using a soft decision approach, meaning it selects the class label with the highest probabilities or confidence score. The NeT component of the ViT-NeT model is designed for distinguishing between objects with similar relationships across classes but differing relationships within

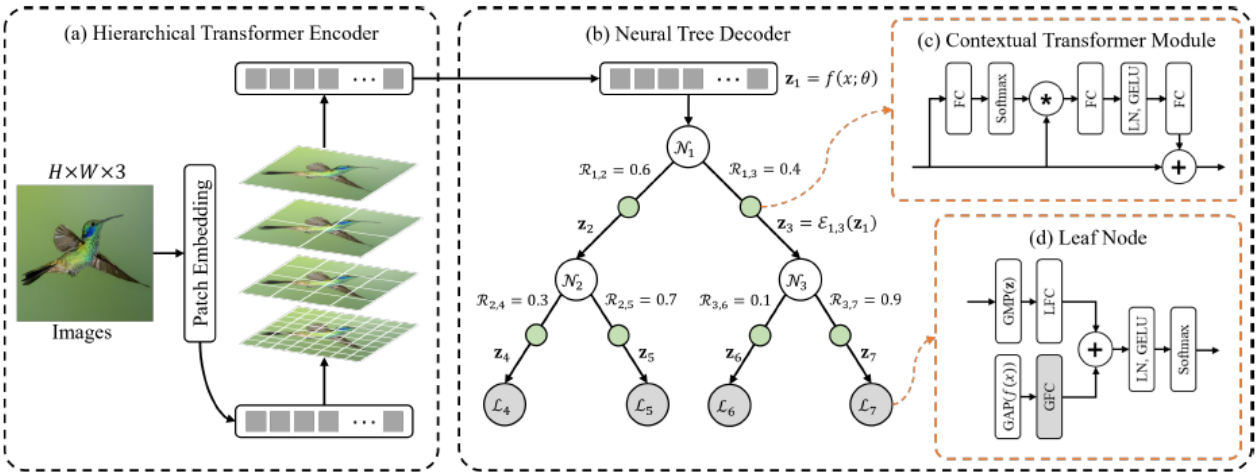


Figure 3.2.1: ViT-NeT overview.
"ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder" [29]

classes. Fine-grained classification requires identifying specific features within categories, achieved in this study by applying prototypes to condensed image patches, unlike the global classification based on general global indicators.

The NeT utilizes prototypes to detect distinctive regions within image patches, guiding routing directions through a differentiable routing module. Each internal node represents a trainable prototype, evaluating the distance between reshaped image patches and the prototype. Then, through a logarithmic similarity measure, routing scores are calculated to determine the similarity between prototypes and image patches. The contextual transformer module (CTM), used by the model, enhances the object description by integrating global context into the patches at every position. Finally, Each leaf in the tree decoder corresponds to a leaf prediction module for class probability prediction over K classes. The final prediction $\hat{y}$ is calculated as the sum of all leaf predictions multiplied by the accumulated routing scores.

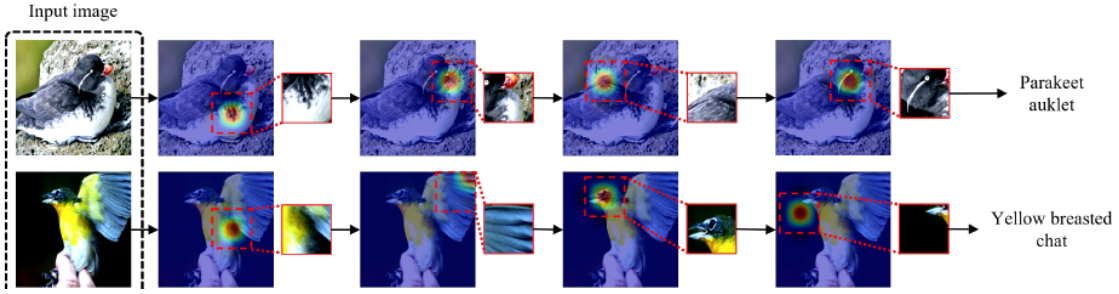


Figure 3.2.2: Illustrated local interpretations display sequential decision-making processes on randomly selected images. The proposed NeT identifies specific "bird features" within the depicted images.
"ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder" [29]

3.2.2 ProtoPFormer

In a follow-up study, Xue et al. proposed ProtoPFormer [31] for applying the prototype-based method [32] with ViTs for interpretable image recognition. Based on the architectural characteristics of ViTs, ProtoPFormer progressively solves the "prototype distraction" problem and introduces global and local prototypes to capture and highlight both the global and local features of target objects through a mutual correction and joint decision process. While ProtoPFormer and ViT-NeT, have comparable accuracies, the number of parameters added by the ProtoPFormer is significantly smaller, making it preferable to use.

The overall architectural layout is depicted in Figure 3.2.4. Within this structure, a class token t_c belonging

to $\mathbb{R}^{1\times d}$ and a feature sequence X_f from $\mathbb{R}^{(n-1)\times d}$ are extracted from the visual sequence $X = [t_c; X_f]$. These components are then used separately as inputs into a global prototype branch and a local prototype branch. The global branch encompasses m_g learnable prototypes $P_g = \{p_g^{(1)}, \ldots, p_g^{(m_g)}\}$, while the local branch contains m_l learnable prototypes $P_l = \{p_l^{(1)}, \ldots, p_l^{(m_l)}\}$, each with m_g^c and m_l^c prototypes accordingly for every class. The outputs from the FC classification layer of the global and local branches are weighted summed, ultimately classifying the input image.

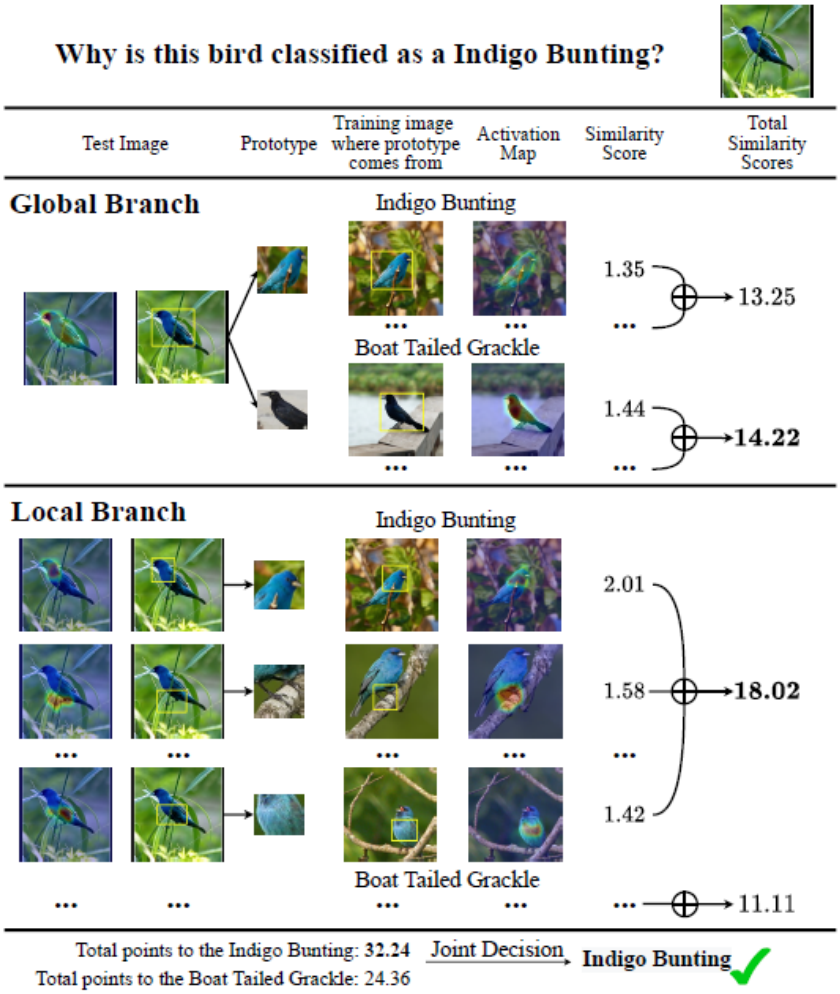


Figure 3.2.3: Reasoning process for the classification of a bird image as an Indigo Bunting through mutual correction and joint decision of the local and global branch.
"ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition" [31]

The proposed method involves several steps aimed at enhancing the concentration of local prototypes on foreground elements while filtering out background influences. Initially, a binary mask called the foreground preserving (FP) mask is applied to selectively retain foreground-related image tokens and exclude background-related ones. This mask is generated using the rollout method, which utilizes the attention rollout matrix of the class token in a Vision Transformer (ViT) model. The rollout attention values to the class token are utilized to preserve the top-K foreground tokens, which are then used to compute the subsequent encoder layers. Subsequently, a modified softmax normalization is employed to remove selected background tokens, ensuring that only foreground-related tokens contribute to further processing. This concentration of local prototypes in the foreground is achieved using the FP mask. Additionally, to focus local prototypes on diverse and centralized representative parts for each class, a prototypical part concentration (PPC) loss is introduced. This loss function minimizes the sum of eigenvalues of the dispersion parameter while encouraging diverse

center coordinates for prototypes belonging to the same class. The optimization objective of the proposed ProtoPFormer model is to minimize a combination of the conventional cross-entropy loss and the PPC loss.

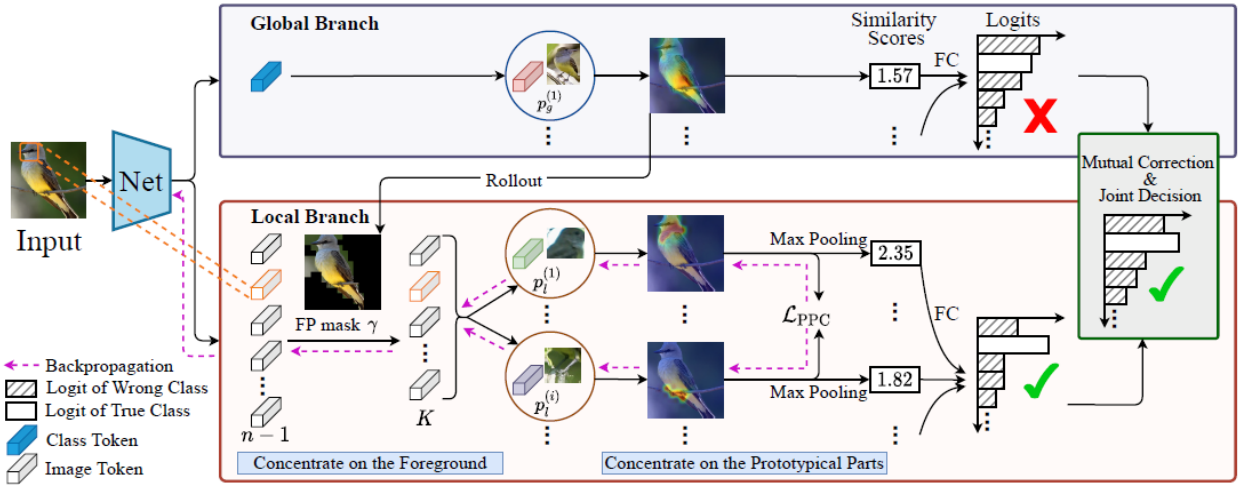


Figure 3.2.4: ProtoPFormer's depiction for interpreting image recognition, showcasing the interplay between its global and local branches. Through an approach of mutual correction and joint decision-making, they collaboratively enhance final predictions, leveraging ViTs' inherent architectures "ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition" [31]

3.2.3 PaCa ViT

Last year, a new interpretable ViT model has emerged, the PaCa-ViT [33], outperforming its ancestors, the Swin-Transformer [30] and the PVT [34, 35].

In the transition from patch-level to cluster-level attention mechanisms, an input sequence $X_{N,C}$ is processed, where N represents the number of tokens formed through patch embedding. The Transformer model's core operation involves computing scaled dot-product attention to transform $X_{N,C}$ into the output $Y_{N,C}$. This involves generating query, key, and value matrices, and applying the softmax function to obtain attention weights. In order to address the quadratic complexity problem, existing methods include spatial reduction techniques such as strided convolution or adaptive average pooling. However, although strided convolution decreases complexity according to the patch size, it does not fully eliminate quadratic complexity. Similarly, adaptive average pooling might assign equal importance to each element in a pooling window, possibly lacking adaptability and data-driven reweighing capability. So as to tackle this challenge, which arises when $M = N$, the key is to maintain a relationship where $M << N$, ideally a predetermined constant, to ensure linear complexity. This is achieved with the proposed Patch-to-Cluster attention (PaCa), as illustrated in Figure 3.2.5.

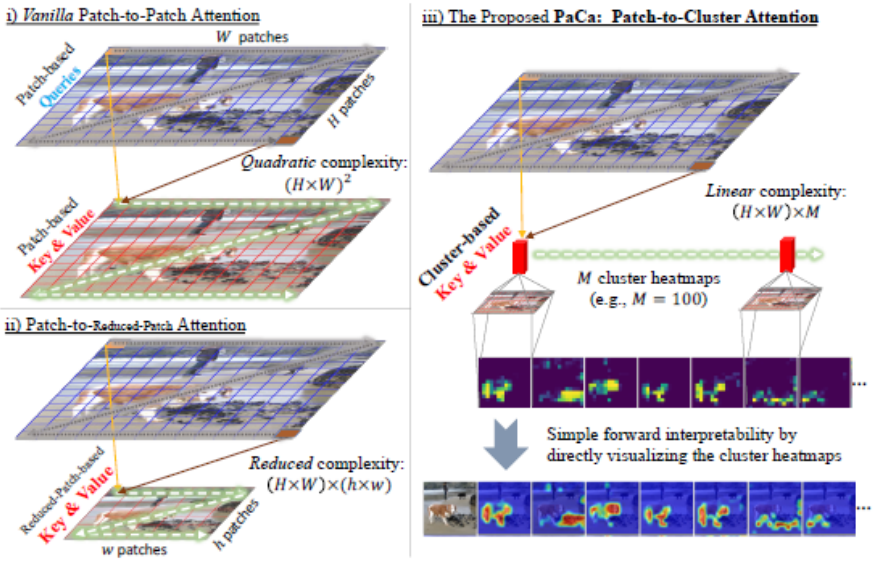


Figure 3.2.5: The vanilla patch-to-patch self-attention suffers from quadratic complexity as every query interacts with every key. A popular method to reduce this complexity involves spatial reduction through techniques like strided convolution. The paper proposes Patch-to-Cluster attention (PaCa), which uses a predefined number of cluster assignments to compute the Key and Value, achieving linear complexity and more meaningful visual tokens.
"PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers" [33]

In this approach, an input sequence $X_{N,C}$ is clustered into a set of "visual tokens" $Z_{M,C}$, with $C_{N,M}$ being the cluster assignment computed either by the onsite clustering method or by the external clustering one.

$$Z_{M,C} = \text{LayerNorm}(C_{N,M}^T \cdot X_{N,C})$$

Subsequently, the computed clusters are utilized to derive Key ($K_{M,C}$) and Value ($V_{M,C}$) components through linear transformations for attention computation.

In the **onsite clustering method**, cluster assignment $C_{N,M}$ is computed using two designs: Clustering via Convolution and Clustering via MLP.

- Clustering via Convolution employs depth-wise and point-wise convolutions followed by Softmax to generate cluster assignments.

- Clustering via MLP utilizes a Multi-Layer Perceptron for cluster assignment computation.

By utilizing Softmax across spatial dimensions, meaningful cluster creation is promoted, facilitating the visualization of $C_{N,M}$ as heatmaps to assess model interpretability during forward computation. Regarding the computation location of $C_{N,M}$ in the onsite clustering approach, it can be done either in block-wise or stage-wise fashion. Grainger et al. suggest that the latter approach is not only more computationally efficient but also yields higher accuracy.

To address the challenge of onsite clustering relying on early-stage cluster assignments rooted in low-to-middle level information, an **external clustering** teacher CNN was introduced. This approach draws insights from various research areas, such as feature pyramid networks and the slow-fast thinking paradigm, as well as empirical findings on Transformers and CNNs. The clustering teacher network, trained alongside the PaCa ViT model, computes stage-wise clustering assignments and integrates them into the model. Acting as a fast learner, it guides the slower PaCa learner, akin to a working memory mechanism that manipulates input data to facilitate post-processing via PaCa. This integration has the potential to enhance the model's ability to learn from both high and low-frequency information sources.

Network Interpretability. In the pursuit of identifying the most crucial clusters within $C_{N,M}$ for an input image in vision tasks, a direct method is employed. This involves utilizing the clustering assignment maps before undergoing the Softmax transformation, followed by a Sigmoid operation. Each cluster's heatmap is transformed into a 2D spatial heatmap, and a binary mask is generated by selecting locations with clustering scores surpassing the mean score. After upscaling this mask to match the input image's resolution, it is applied to the input image. The resulting masked image, I_m, is further categorized into positive and negative groups, depending on its ability to retain sufficient information for correct classification, providing insights into the significance of the clustered regions.

3.2.4 Ex-ViT

The eX-ViT [36] operates as a siamese network, utilizing two branches to process a pair of input images, which in reality are two different random transformations of the original input image. shown in Figure 3.2.6 Each branch includes a transformer encoder with multiple layers incorporating the Explainable Multi-Head Attention (E-MHA) module and the Attribute-guided Explainer (AttE) module positioned on top of the encoder.

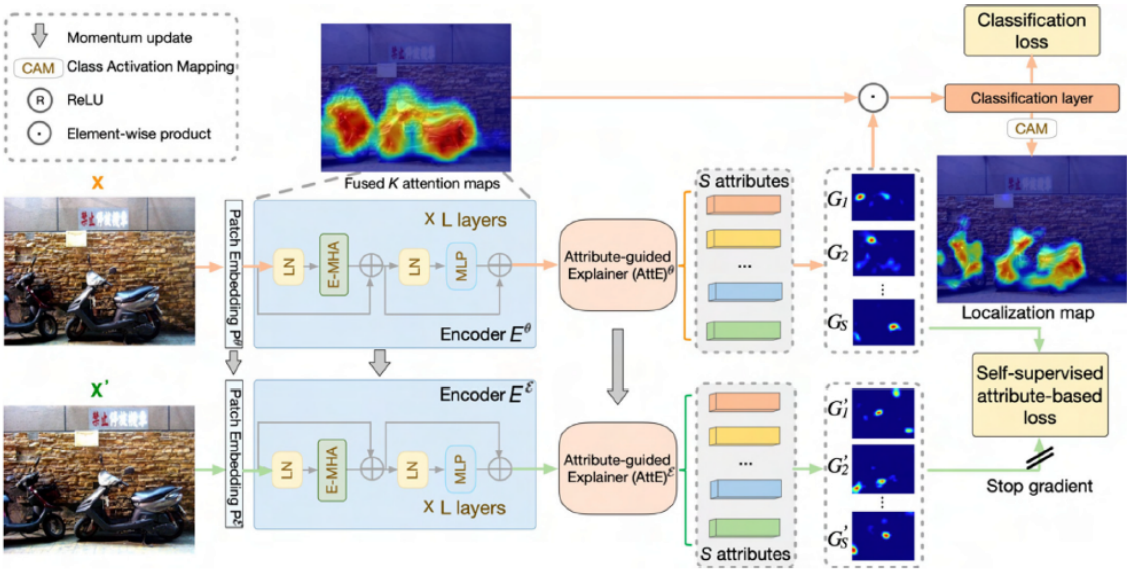


Figure 3.2.6: Overview of the eXplainable Vision Transformer architecture.
"eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation" [36]

The Explainable Multi-Head Attention. The Explainable Multi-Head Attention (E-MHA) module comprises multiple parallel heads, each holding an explainable attention weight $A_h \in \mathbb{R}^{N \times d}$ aiming to learn

interpretable features from the input feature map $X \in \mathbb{R}^{T \times d}$ (T: spatial size of X, d: feature dimension, N: spatial size of A_h). Yu et al. focus on two key attributes of the E-MHA module: noise robustness and inherent explainability. Noise robustness is achieved through dynamic alignment between input tokens and attention weight, allowing the module to focus on discriminative patterns and gradually reduce noise. Inherent explainability is attained by maximizing alignment between input tokens and attention weight, resulting in model-inherent attention weight that direct explanations for model decisions without external aids.

The computation process begins with the projection of input data X onto trainable matrices to obtain the key, query, and value matrices (K, Q, V). Self-attention is then conducted by calculating attention weights based on the dot product of the query and key matrices, scaled appropriately. These weights are then adjusted using a non-linear function and a bias term. Following this, the self-attention feature (S) is formally expressed as the product of the attention matrix (A) and the value matrix (V).

Overall, the computation in layer l is expressed as

$$S_l = E\text{-}MHA(\mathrm{LN}(F_{l-1}))$$
$$Z_l = S_l + F_{l-1}$$
$$F_l = \mathrm{MLP}(\mathrm{LN}(Z_l)) + Z_l$$

with $\mathrm{LN}(\cdot)$ being the LayerNorm layer, $\mathrm{MLP}(\cdot)$ the multi-layer perceptron layer, and F_l the output of layer l.

The Attribute-guided explainer. The Attribute-guided Explainer (AttE) module complements the E-MHA by decomposing attention maps into attribute features, aiming to enhance interpretability in Weakly Supervised Semantic Segmentation (WSSS) tasks without additional regularization. By utilizing transformer attention maps from the last layer of the eX-ViT encoder, spatial feature maps capturing relative importance are generated and normalized to emphasize or suppress specific spatial features. These maps are then sliced into groups representing different attributes, which are applied to the original feature maps to produce attribute representations. This process enables the model to explicitly identify pixels related to specific attributes, contributing to a more comprehensive understanding of object context. Additionally, an attribute-guided loss function is designed to facilitate the learning of AttE. By utilizing both global and local level losses, the introduced loss function reinforces the trustworthiness of the Transformer model and enhances the discriminative and robust qualities of the learned attribute features.

3.3 Exploring Explanations for Medical Imaging

In the domain of medical imaging, the interpretability of AI models is significantly important for ensuring their safe and effective deployment in clinical practice. To address this aspect, this section delves into the exploration of explainability methods either tailored specifically to medical imaging tasks or used for producing interpretable explanations in the field of medicine, beginning with the paper authored by Komorowski et al. [37]. This study compares the visualization results of Attention rollout, TransLRP and LIME, three methods that have already been discussed in Section 3.1, in the classification of chest X-ray images, highlighting interpretability challenges and opportunities in this domain. Based on the results provided, TransLRP demonstrates strong potential for explaining ViT's predictions in the characterization of chest X-ray images as Covid, Non-Covid and Healthy, making it a suitable choice for interpretation by experts such as radiologists. While TransLRP is resilient to distortions or abnormalities in medical images known as "imaging artifacts", there are still situations where it might produce explanations based on misleading or incorrect correlations, which could lead to inaccurate predictions from the model. On the other hand, LIME explanations tend to be similar or consistent when applied to different images. More specifically, LIME relies on superpixels, which are regions of an image used to approximate the contribution of features to the model's prediction. It is observed that if the superpixels used by LIME do not primarily focus on the lung areas in chest X-ray

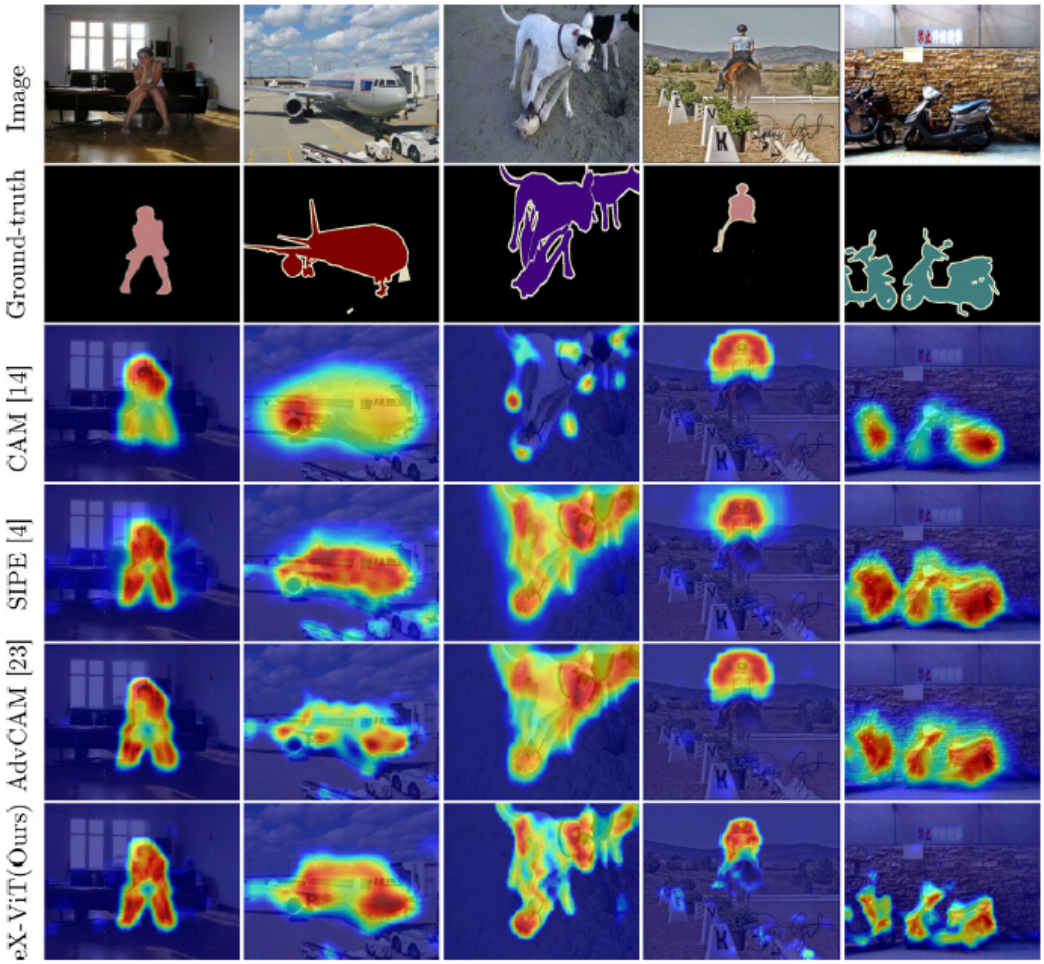
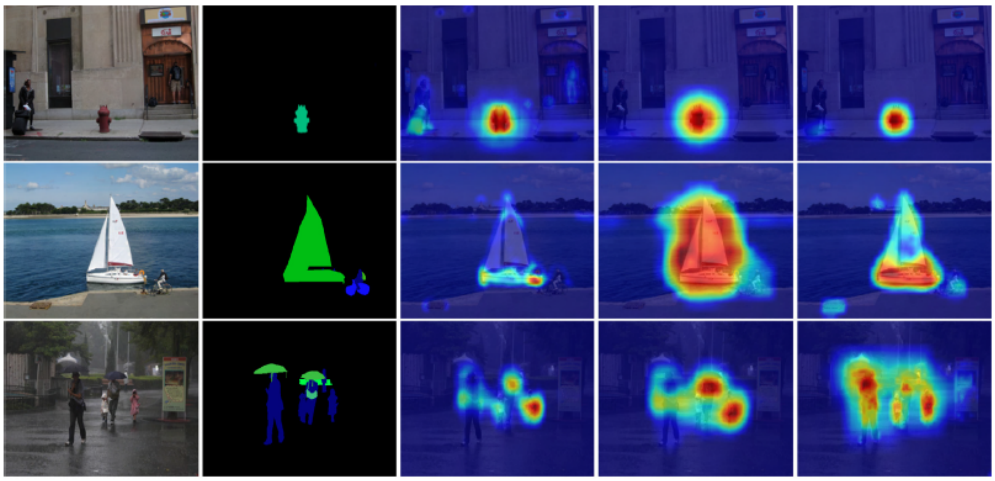


Figure 3.2.7: Visualization comparison among different interpretability methods and models on the PASCAL VOC 2012 Training Set.
"eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation" [36]



(a) Image (b) Ground-truth (c) E-MHA (d) w/o AttE (e) w/ AttE

Figure 3.2.8: Three visualization cases are showcased alongside their ground truth segmentation labels. Notably, the proposed model's attention maps, enhanced by the AttE module, demonstrate superior precision in identifying both small and large objects, exhibiting significantly improved object outlines. "eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation" [36]

images, there is a risk of inaccuracies in the explanations provided by LIME. Finally, when using attention visualization to understand why a model made a specific prediction, the explanations provided are less reliable or informative compared to those generated by TransLRP and LIME. Overall. the results, as shown in Figure 3.3.1, indicate that TransLRP overcomes the other examined explanation methods in classifying COVID in chest X-ray images.

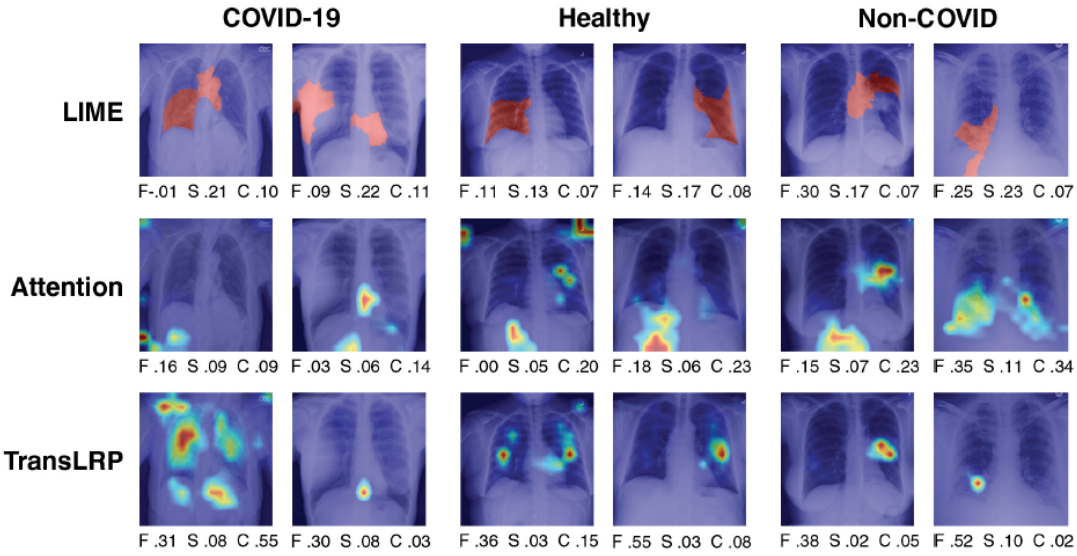


Figure 3.3.1: Visual explanations of ViT trained for X-ray classification. Two images are displayed for every class label, each explained using three interpretation methods. Performance metrics, including faithfulness (F), sensitivity (S), and complexity (C), are provided for each explanation, with lower scores being preferable for sensitivity and complexity, while higher scores for faithfulness.
"Towards Evaluating Explanations of Vision Transformers for Medical Imaging" [37]

In another work, published in 2022, Playout et al. [38] manage to produce high-resolution heatmaps by developing a technique called Focused Attention, aimed at efficiently processing medical imaging data while maintaining memory efficiency. The methodology begins by addressing the issue of increased memory requirements caused by using a small-stride projector convolution. This increase in memory usage would make attribution methods requiring backpropagation impractical due to the large number of tokens. Regarding patch selection, the proposal suggests that not all patches in an image are equally important for diagnosis, especially those centered on lesions. Therefore, the methodology proposes generating attribution maps and predictions using only a subset of patches at each scale. Thus, the process defines an iterative procedure over a scale of decreasing stride values. Strides are typically powers of two, starting from a larger value and gradually decreasing to one. For each stride, a fixed number of patches are randomly sampled from the region highlighted in the attribution heatmap constructed at the previous stride. This ensures a constant memory requirement at each scale. The approach involves repeating the sampling operation multiple times for each scale, effectively processing a total of patches with a constant memory requirement. The attribution maps generated per stride are then merged and aggregated, as shown in Figure 3.3.2, with experimentation showing that elementwise averaging provides the most consistent results. At each stride, the resolution of the attribution map is halved compared to the next stride. Bilinear interpolation is used to upsample the attribution map generated at each scale for use as a conditional sampling map for the following stride. As a whole, this method effectively concentrates the attention mechanism on a subset of meaningful tokens, iteratively refining the selection. Finally, the comparison between the Focused Attention approach and the other dominant explainability techniques used in medical imaging is shown in Figure 3.3.3.

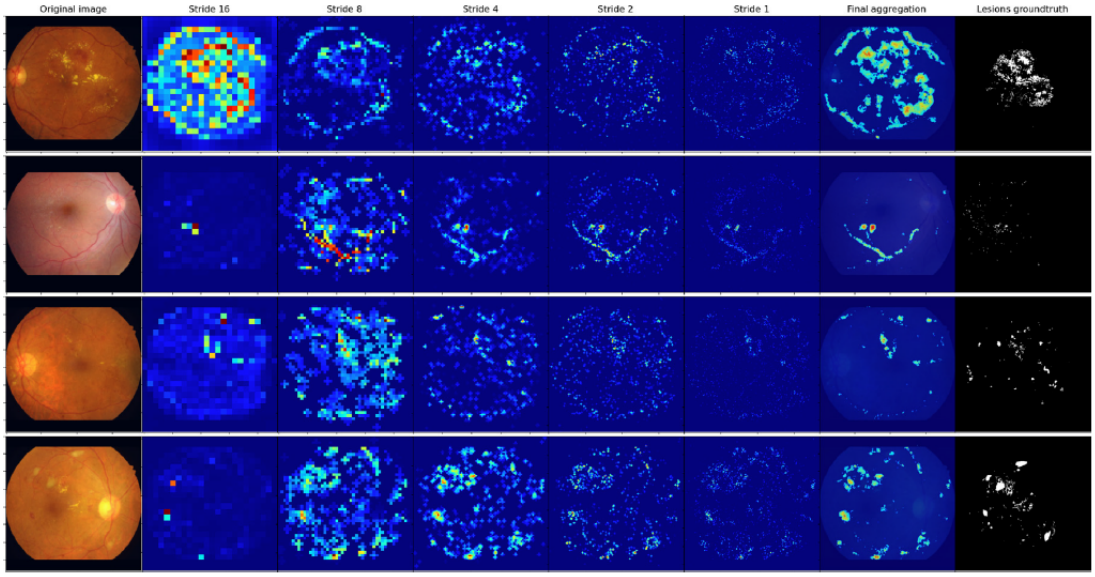


Figure 3.3.2: Visual representation of Focused Attention: each column showcases the attribution maps generated per stride, with the last two displaying the final aggregation and the ground truth lesions. "Focused Attention in Transformers for interpretable classification of retinal images" [38]

In another paper by Demir et al. [39], an innovative attention block is introduced in the Convolutional Vision Transformer architecture, emphasizing the relationship among 'regions' rather than 'pixels', along with an original system rooted in prototype learning, showcasing an advanced self-attention mechanism that surpasses traditional ad-hoc visual explanation methods by providing clear and understandable visual insights. The module comprises three primary components: (i) mask and query generation, (ii) region-to-region self-attention, and (iii) output feature reconstruction.

Mask and Query Generation. Standard transformer networks typically segment input images into patches, employing attention mechanisms to identify similarities between them. These patches contain information from the foreground as well as the background. To tackle this issue, in this novel approach, known as masking-based query generation, masks are generated to enable adaptive patch generation. This process ensures that only one mask dominates in each pixel location, enhancing the adaptability of the model. Additionally, convolutional projection is utilized to extract features from the input image, preserving essential location information necessary for subsequent processing.

Region-to-Region Self-Attention. Traditionally, attention mechanisms calculate keys based on the input features. However, in this approach, learnable key vectors are introduced to represent canonical representations of critical patterns within the data. This departure from conventional methods allows the model to store key vectors that encapsulate essential patterns, leading to more efficient and effective attention calculation. Through matrix multiplication of queries with these learnable key vectors, attention values are computed, forming an attention matrix that highlights the correlation between different regions of the input. This region-to-region self-attention mechanism enables dynamic focusing on local regions while calculating query vectors, facilitating the identification of similar patterns within the input data.

Output Feature Reconstruction. The final step of the interpretable self-attention module involves reconstructing the output features for subsequent layers in the network. This is achieved through the definition of parametric learnable prototype value vectors. These vectors, combined with the location information contained within the weighted masks generated in previous steps, enable the precise placement of similar value vectors to pixels within the same mask. By performing matrix multiplication between the prototype value vectors and the weighted masks, the output feature map is reconstructed. This process ensures that the output features retain the essential characteristics of the input data, facilitating further processing and analysis by subsequent layers in the network

The effectiveness of the proposed architecture was demonstrated on chest x-ray images, generating visual

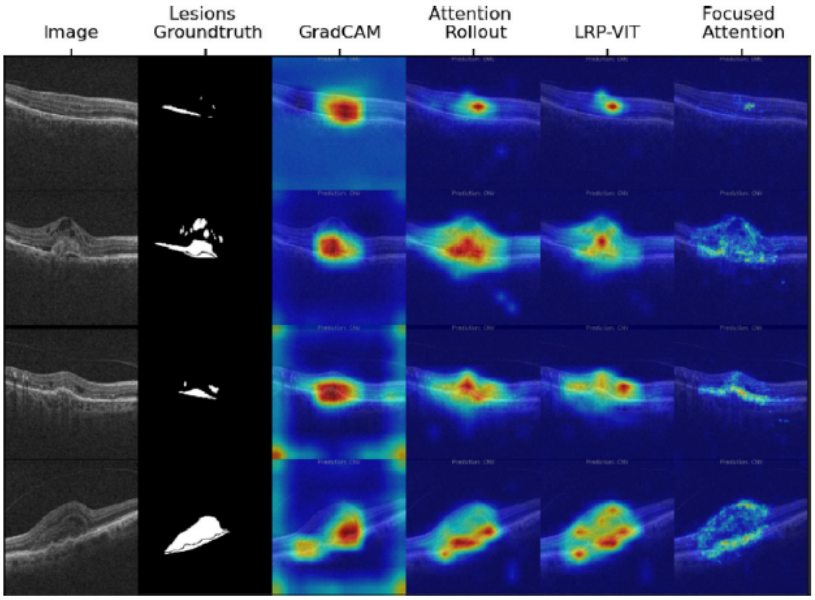


Figure 3.3.3: Comparison of different explainability methods.
"Focused Attention in Transformers for interpretable classification of retinal images" [38]

explanations at each intermediate feature layer for different resolutions, enhancing both interpretability and diagnostic performance.

Last year, the RadFormer [40] model was designed to address the challenges posed by ultrasound imaging, which include sensor noise, artifacts, and difficulties in distinguishing between normal and abnormal regions. Especially tailored for gallbladder cancer detection, it employs a dual global–local attention-based architecture to enhance interpretability and diagnostic performance. The global branch identifies the region of interest (ROI) around the gallbladder at the image level to mitigate the effects of artifacts and retain contextual information. This ROI is then used to crop the local region, which is analyzed by the local branch to identify fine-grained features using a bag-of-features (BOF) encoder. The model combines global and local features through a transformer-based classification branch, enabling both coarse and fine-grained interpretation of gallbladder conditions. The global ROI mask provides a coarse-grained visual explanation of abnormalities, while the bag-of-features offers detailed, fine-grained interpretations.

In conclusion, while significant steps have been made towards developing interpretability methods for medical imaging tasks, there remains a considerable gap in the exploration of these techniques. Further investigation needs to be carried to advance the field of medical imaging and improve patient care outcomes.

Chapter 4

Experiments

In this section, we describe the experimental process conducted to evaluate the effectiveness, as well as the interpretability of the selected Vision Transformer models. The chapter is structured to provide a clear understanding of the methodology followed, focusing on the selection of datasets, the training procedure and parameter optimization, the results' review and the comparison among the performance of the different models used.

Contents

4.1 Datasets

The methodological approach revolved around adapting three Vision Transformer model architectures, to accommodate the complexities inherent in medical imaging. For this purpose, four distinct medical imaging datasets were chosen, encompassing a variety of modalities such as MRI, CT scans, histopathological images, and real gastrointestinal images taken during endoscopies. These datasets specifically are:

- Augmented Alzheimer MRI Dataset V2 [22], showing brain MRI images of patients with different Alzheimer's disease stages.
- Large COVID-19 CT scan slice dataset [41], publicly used in COVID-19 diagnosis literature.
- Gastrointestinal Cancer MSI MSS Prediction [42], containing histological images for MSI vs MSS classification in gastrointestinal cancer.
- Kvasir Dataset for Classification and Segmentation [43], containing images from inside the gastrointestinal (GI) tract.

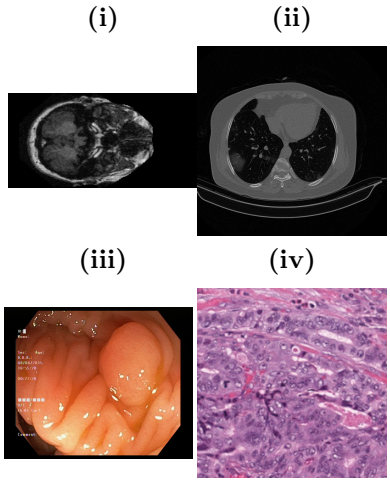


Figure 4.1.1: Training images randomly chosen from the: (i) Alzheimer's dataset (ii) Covid dataset (iii) Kvasir dataset (iv) TCGA dataset

Datasets' preprocessing steps, including downloading, structuring, and balancing the data, are described in the following section:

1. Data Collection and Downloading:
   - Identify and acquire the relevant medical imaging datasets from their sources.
   - Download the datasets from repositories or sources ensuring data integrity and quality.
2. Train-Test Split:
   - Divide the dataset into training, validation and testing subsets using an 60-20-20 split ratio.
3. Balancing Classes:
   - Assess the class distribution within the dataset to identify any imbalance.
   - If class imbalances exist, apply random undersampling (reduce the size of overrepresented classes randomly to match the size of the minority class).
4. Dataset Structuring:
   - Create the desired inner structure in order to have reusable code when training. The structure of the datasets is formed according to the model being used, in order to adjust to the needs of each respective architecture. More specifically, for the ProtoPFormer the structure looks like the following: parent folder/class1, class2, etc/train, test/sample1.jpg, sample2.jpg, etc. For the ViT-NeT,

there is a slight variation, parent folder/train, test/class1, class2, etc/sample1.jpg, sample2.jpg, etc, while for the Swin Transformer the datasets look identical to the ones used for the ViT-NeT, with 'val' being the name of the test folder, instead of 'test'.

After applying the preprocessing steps to the four datasets, the Kvasir dataset contains 3200/800 train/test images over 8 classes representing conditions of the GI tract (dyed lifted polyps, dyed resection margins, esophagitis, normal cecum, normal pylorus, normal Z-line, polyps, ulcerative colitis). The Alzheimer's dataset consists of 1564/400 train/test MRI images splitted into 4 classes representing the different dementia stages (Mild Dementia, Moderate Dementia, Non Demented, Very mild Dementia). The TCGA dataset contains 10000/2500 images equally separated into two classes (MSS, MSIMUT) and finally, each class (Covid, Non-Covid) of the Covid dataset has 5515/1380 train/test images. In Table 4.1, the different classes are briefly described:

Alzheimer's Dataset	
Mild Dementia	Noticeable symptoms that begin to interfere with daily activities, such as memory loss and cognitive decline.
Moderate Dementia	Pronounced symptoms requiring assistance with daily activities, with significant cognitive decline and memory issues.
Non Demented	Individuals showing no signs of dementia, serving as a control group.
Very mild Dementia	Early stage of dementia where symptoms are very subtle and might not significantly affect daily life.
Covid Dataset	
Covid	Images of patients diagnosed with COVID-19.
Non Covid	Images of patients not diagnosed with COVID-19.
Kvasir Dataset	
Dyed lifted polyps	Polyps that have been lifted and dyed to highlight their contours.
Dyed resection margins	Areas where tissue has been dyed to mark resection margins.
Esophagitis	Inflammation of the esophagus.
Normal cecum	Healthy tissue in the cecum.
Normal pylorus	Healthy tissue in the pylorus.
Normal Z-line	Healthy tissue at the gastroesophageal junction.
Polyps	Abnormal tissue growths.
Ulcerative colitis	Inflammatory bowel disease causing ulcers in the colon.
TCGA Dataset	
MSIMUT	Microsatellite instability-high (MSI-H) indicating a high mutation rate.
MSS	Microsatellite stability (MSS) indicating a low mutation rate.

Table 4.1: Descriptions of the classes within the Alzheimer's, Covid, Kvasir, and TCGA datasets. The Alzheimer's dataset classes describe different stages of dementia. The Covid dataset contains images of patients with and without COVID-19. The Kvasir dataset includes various conditions of the gastrointestinal tract. The TCGA dataset classifies tissue images based on microsatellite stability.

4.2 Resources

For the ProtoPFormer, the training and visualizations were conducted on Google Colab using the T4 GPU provided, with Google Drive mounted for data storage. Due to the limited GPU time available, the training process spanned approximately two months for all the data. In contrast, the experiments for ViT-NeT and Swin Transformer were executed on Kaggle, utilizing two T4 GPUs as accelerators. The training for the ViTNet model lasted about one and a half months, while the baseline Swin Transformer completed its training in only a few days. Although the training was successfully completed with these limited resources, it is important to note that the results could be further optimized with the use of more powerful GPUs.

4.3 Training the Models

For the purposes of this thesis, three Vision Transformer models were selected to be applied to the medical imaging datasets described in the previous section. More specifically, the first model, ProtoPFormer [31], was chosen for its prototype-based approach, offering a novel perspective in applying this method to medical imaging data and evaluating its response. ViT-NeT, with its unique tree-like structure, although not yet applied to the field of healthcare, would offer promising opportunities to explore its potential in the medical domain. Finally, in contrast to the previous models, a non-interpretable Transformer, the Swin Transformer was selected, combined with GradCAM, a post-hoc explainability method, in order to serve as a baseline to assess whether the built-in interpretability affects the performance.

4.3.1 ProtoPFormer

Training and Visualization Parameters

The training process was configured with several key parameters to optimize model performance. The batch size was set to 64, meaning that 64 images were processed simultaneously during each training iteration. The learning rate was initialized at 5×10^{-4}, which determines the rate at which the weights change. To gradually introduce the learning process, a warmup learning rate of 1×10^{-4} was applied for the initial 5 epochs, allowing the model to stabilize before reaching the full learning rate. The optimizer used for this task was AdamW, combined with a cosine annealing scheduler, which smoothly adjusts the learning rate following a cosine curve, preventing abrupt changes and aiding in convergence. Warmup epochs were set to 5, allowing a gradual increase in learning rate, and decay epochs were set to 10, specifying the intervals at which the learning rate decreases by a factor of 0.1. The total number of epochs for the training process was set to 60 for the Alzheimer's dataset and 100 for the Covid, Kvasir and TCGA datasets.

Prototype Configuration

The ProtoPFormer model utilizes a prototype-based method, combining both global and local prototypes to enhance interpretability and performance, through a joint decision and mutual correction process. Representative vectors called prototypes are used to capture key characteristics of different classes. The table below specifies the prototype number, as well as their dimensions, and the number of global prototypes per class for each dataset. The influence of global prototypes is balanced by the global coefficient (`global_coe`), which is set at 0.5. In contrast, local prototypes are exclusive to specific areas or characteristics found in the images; 81 tokens were set aside for these prototypes in the final layer. To further improve these prototypes, the model additionally used prototype-based contrastive loss (PPC loss), with coverage and mean thresholds set at 1.0 and 2.0, respectively. The discriminative capacity of the model is enhanced by this loss function, which aids in distinguishing between similar and dissimilar traits. The coverage (`ppc_cov_coe`) and mean (`ppc_mean_coe`) coefficients were adjusted to 0.5 and 0.1, respectively, to fine-tune their respective contributions to the total loss function.

Datasets	**Prototype number**	**Dimension**	**Global prototypes per class**
Alzheimer's	40	192	10
Covid	100	192	50
Kvasir	80	192	10
TCGA	100	192	50

Table 4.2: Prototype specifications for the Alzheimer's, Covid, Kvasir and TCGA datasets. Global prototype numbers per class were determined according to the dataset size; larger datasets were assigned a higher number of prototypes per class.

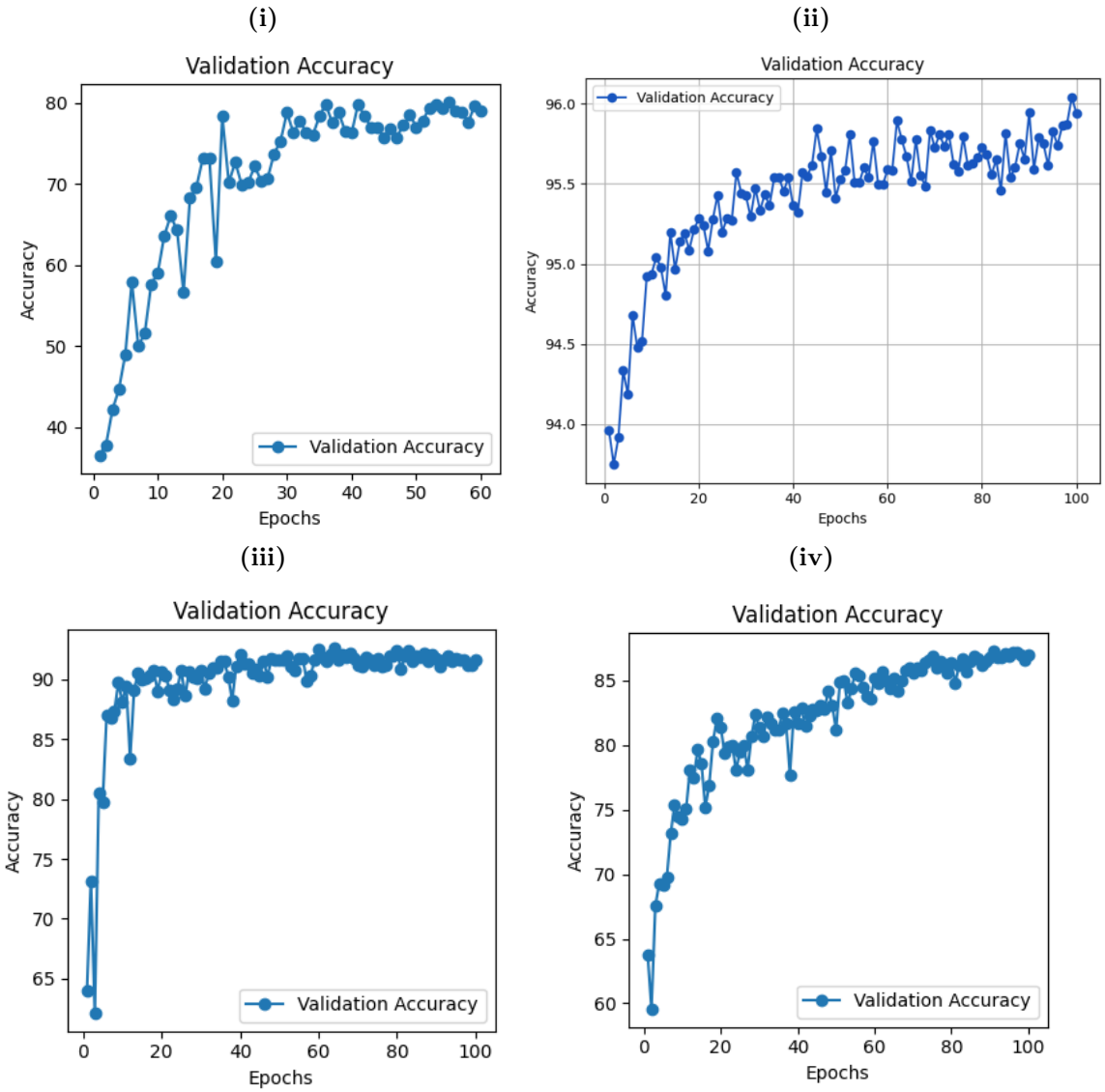


Figure 4.3.1: Learning curves generated during the training process of the ProtoPFormer for the: (i) Alzheimer's dataset (ii) Covid dataset (iii) Kvasir dataset (iv) TCGA dataset

The training and visualization code used for the purposes of this paper were originally provided by Xue et al. in their paper, "ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition" [31].

4.3.2 ViT-NeT

Training and Visualization Parameters

The ViT-NeT model [29], unique for its tree-like structure, was configured with a tree depth of 4 and a prototype size of [1,1], which facilitated the hierarchical processing of image data. The Swin Transformer parameters included an embedding dimension of 128, with depths set to [2, 2, 18, 2] and the number of heads configured as [4, 8, 16, 32]. The window size for the Swin Transformer was 14, which helped in capturing local features effectively. An initial warmup period of 1 epoch to stabilize the learning process. The weight decay was set to an extremely low value of 1×10^{-8}, ensuring minimal loss of information during weight updates. The base learning rate was configured at 2×10^{-5}, with a warmup learning rate of 1×10^{-8} and a minimum learning rate of 1×10^{-7} to maintain a steady learning progression. The optimizer used for this task was AdamW, combined with a cosine learning rate scheduler, as used in the previous model. In order to accommodate the computational demand and at the same time achieve efficient learning, the batch size for training was set to 32. The total number of epochs for the training process was set to 60 for both the Alzheimer's and the Kvasir datasets, while for the Covid one was set to 100 and for the TCGA to 125 epochs.

The training and visualization code used for ViT-NeT were originally provided by Kim et al. in their paper, "ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder" [29].

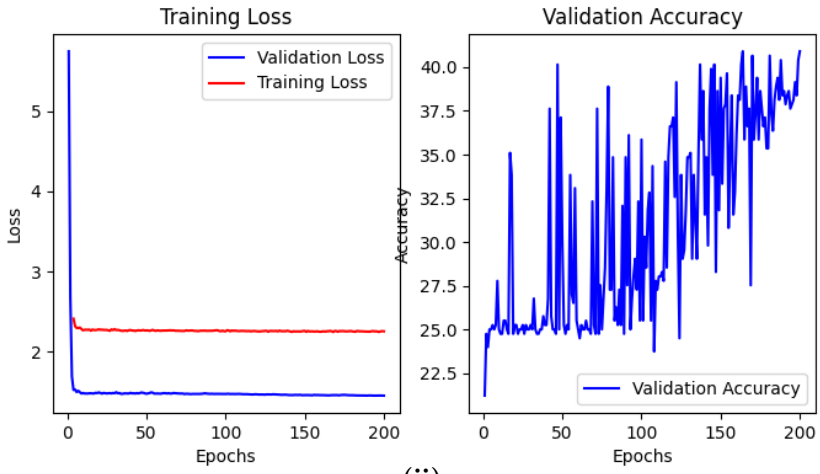
4.3.3 Swin Transformer x Grad-CAM

Training and Visualization Parameters

For the Swin Transformer model, the parameters used in this thesis were largely retained as provided by the original implementation. The specific variant `swin_tiny_patch4_window7_224` was used, which is tailored for smaller models and faster training, along with a batch size of 64. Initially, a base learning rate of 5×10^{-4} was used; however, it became apparent that the training process was not proceeding as expected. After experimenting with various learning rates, we determined that a base learning rate of 1×10^{-4} was optimal for achieving stable and effective training. The total number of epochs for the training process was set to 200 for all datasets. Finally, Grad-CAM was used to visualize and interpret the model's focus areas during prediction.

The training code, as well as the visualization code used for the Swin Transformer with the post-hoc Grad-CAM visualization method were originally provided by Liu et al. in their paper, " Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows " [30] and by Gildenblat et al. in their GitHub repository [90] accordingly.

(i)



(ii)

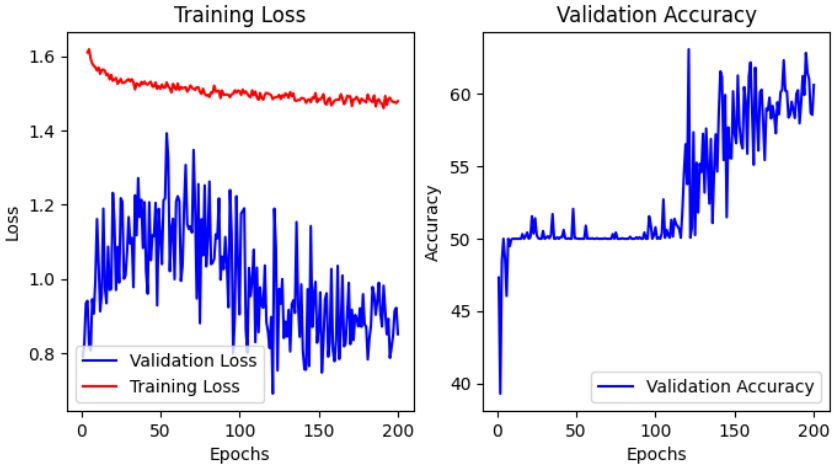
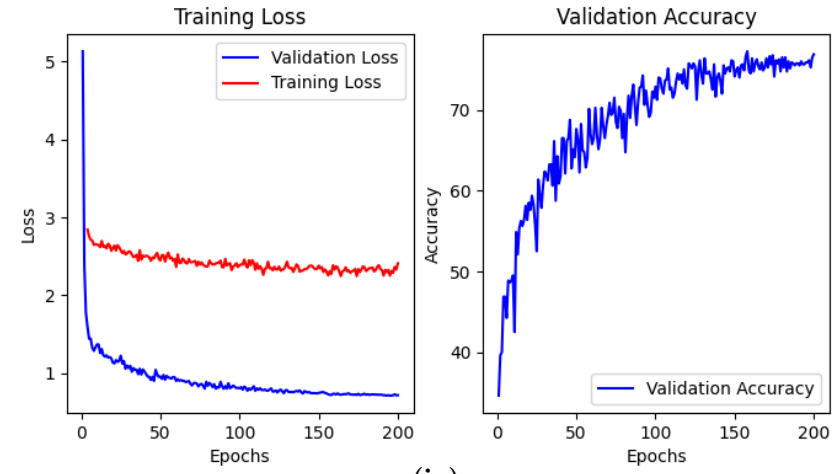


Figure 4.3.2: Learning curves generated during the training process of the Swin Transformer for the: (i) Alzheimer's dataset (ii) Covid dataset

(iii)



(iv)

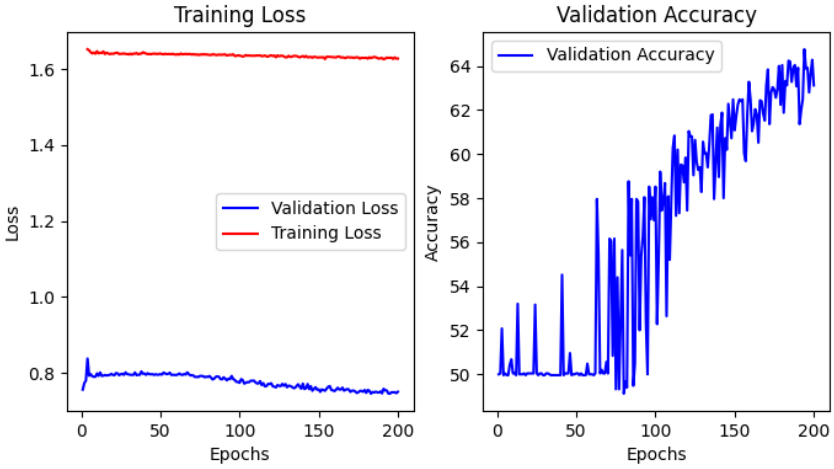


Figure 4.3.3: Learning curves generated during the training process of the Swin Transformer for the: (iii) Kvasir dataset (iv) TCGA dataset

4.4 Model Performance and Visualizations

This section presents the performance of three different models on the four selected medical datasets. The performance is evaluated in terms of accuracy and visualizations generated by each model. The analysis provides insights into the strengths and weaknesses of each model and discusses the practical implications of the findings.

4.4.1 Accuracy Analysis

Dataset/Model	ProtoPFormer	ViT-NeT	Swin Transformer
Alzheimer's	**79.07**	40.35	40.01
Covid	**95.11**	66.92	61.09
Kvasir	**91.33**	86.03	75.25
TCGA	**86.20**	81.21	64.21

Table 4.3: The Max accuracy performance comparison (%) of the three models on the four selected medical datasets. Maximum performance for each dataset is marked in **bold**.

The accuracy results for the selected medical datasets reveal significant variations in the performance of the three models. Table 4.3 shows the detailed accuracy percentages for each model on the four datasets.

ProtoPFormer achieved the highest accuracy on all datasets, indicating its superior performance. For the Covid dataset, ProtoPFormer achieved an impressive accuracy of 96.45%, significantly outperforming ViT-NeT (67.77%) and the Swin Transformer (52.03%). In the Kvasir dataset, ProtoPFormer also led with 92.63%, followed by ViT-NeT at 86.88% and the Swin at 72.0%. For the TCGA dataset, ProtoPFormer reached an accuracy of 87.28%, compared to ViT-NeT's 82.71% and Swin's 54.0%. The Alzheimer's dataset posed a significant challenge for all models. Nonetheless, ProtoPFormer achieved the highest accuracy at 80.01%, whereas ViT-NeT and the Swin Transformer reached only 41.27% and 41.67%, respectively, thus demonstrating ProtoPFormer's relatively better handling of Alzheimer's MRI data.

A general trend observed is that ProtoPFormer consistently outperforms the other models on most datasets, particularly excelling in the Kvasir and Covid CT datasets. However, all models showed relatively lower accuracy on the Alzheimer's MRI dataset, indicating potential challenges due to the complexity of this particular medical imaging task.

4.4.2 Visualizations

Visualizations were generated to provide insights into the regions of the images that each model focuses on to make predictions. These visualizations are crucial for understanding the interpretability and reliability of the models' decisions.

In this section, we describe the results of a comprehensive survey evaluating the visualizations generated by the three models: ProtoPFormer, ViT-NeT, and Swin x Grad-CAM. The evaluation was conducted by a panel of participants, including medical students and individuals not directly involved in the medical profession. This diversity provided varied perspectives on the AI systems' performance. Medical students represented a significant portion of the participants, while non-medical respondents added a layperson's point of view to the evaluation process.

The specialists assessed these visualizations using a survey focused on their effectiveness, clarity, diagnostic value, feature highlighting, alignment with expert knowledge, and overall usefulness across the four medical datasets we have already described in previous sections: Alzheimer's MRIs, Kvasir endoscopy images, COVID-19 CT scans, and TCGA histological images.

Across all the evaluated datasets ProtoPFormer consistently outperformed the other models in generating visualizations that were both accurate and clinically useful. Specialists noted that ProtoPFormer's visualizations were particularly effective in revealing key insights and patterns, clearly conveying diagnoses, and

highlighting specific features or regions crucial for decision-making. The visualizations produced by ProtoPFormer closely aligned with the specialists' clinical knowledge and expectations, making it the most useful tool for interpreting the model's predictions.

In contrast, ViTNet generated visualizations that were generally less accurate and failed to effectively mark the relevant parts of the images necessary for informed decision-making. This significantly limited its utility across all datasets.

The Grad-CAM method also exhibited consistent shortcomings, as it tended to get distracted by irrelevant parts of the images rather than focusing on the regions most important to diagnosis. This lack of focus reduced its effectiveness and often resulted in visualizations that did not align well with the clinical needs of the specialists.

Overall, ProtoPFormer emerged as the superior model, providing the most reliable and clinically actionable visualizations across all datasets.

The tables below display examples of images from all the datasets, along with the visualizations generated by each of the models.

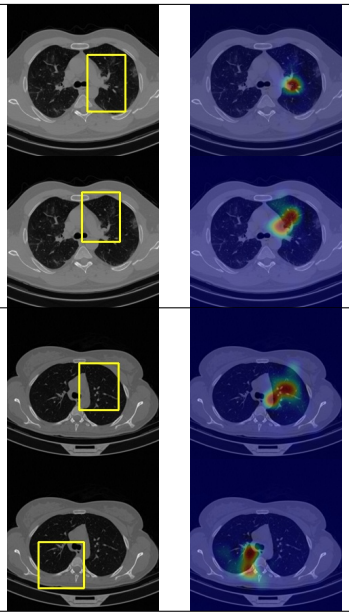
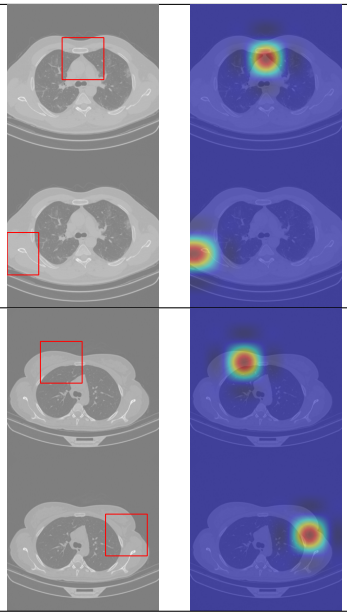
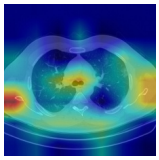
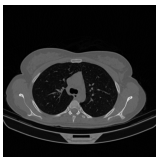
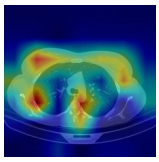
Original Image	**ProtoPFormer**	**ViT-NeT**	**Swin x Grad-CAM**
 COVID			
 NonCOVID			

Table 4.4: Visualizations generated for the Covid dataset by the ProtoPFormer, ViT-NeT, Swin x Grad-CAM.

A shown in Table 4.9, it is evident that the ViT-NeT model, rather than concentrating on areas indicating anomalies, it focuses on the annotations or labels present in the image, thus resulting in incorrect interpretations.

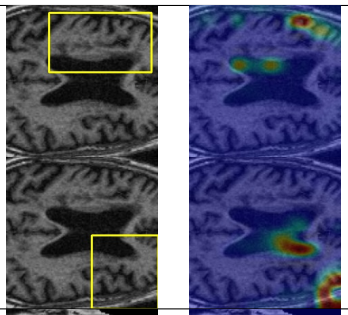
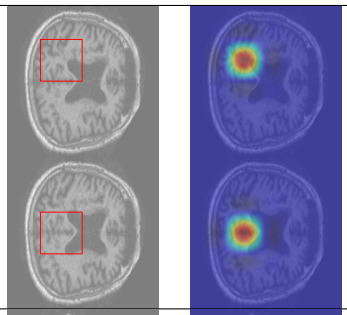
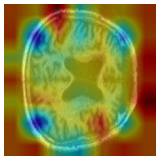
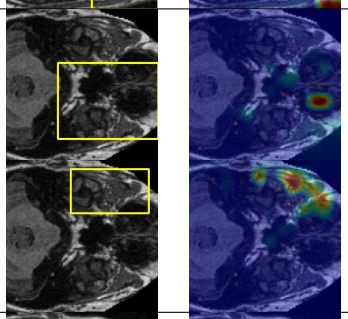
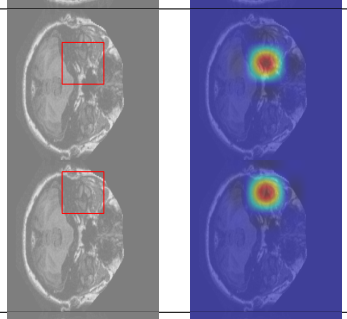
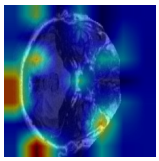
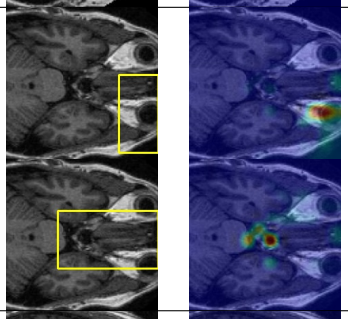
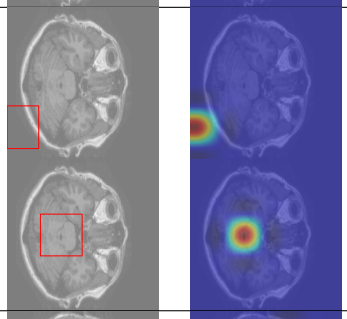
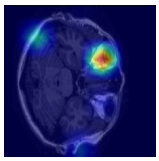
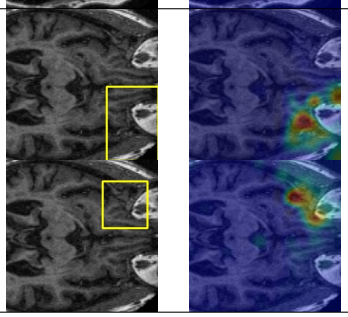
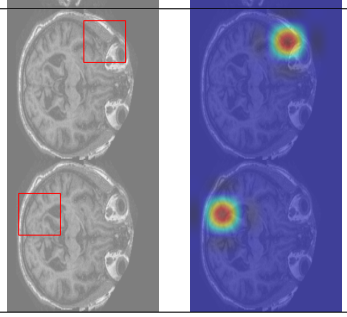
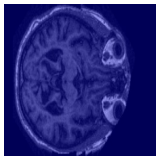
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
Mild Dementia			
Moderate Dementia			
Non Demented			
Very Mild Dementia			

Table 4.5: Visualizations generated for the Alzheimer's dataset by the ProtoPFormer, ViT-NeT, Swin x Grad-CAM.

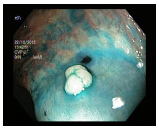
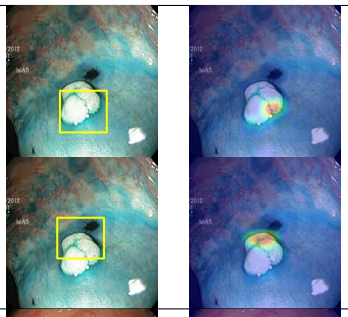
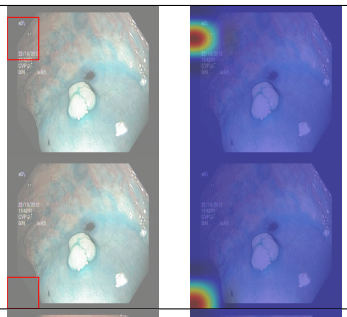
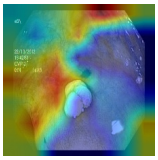
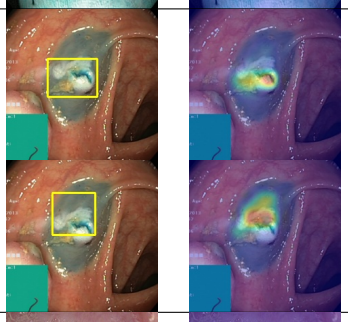
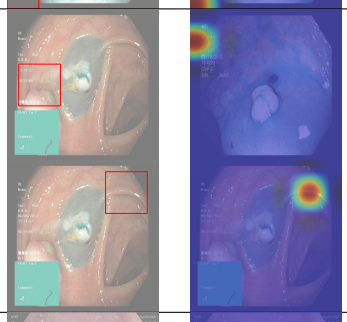
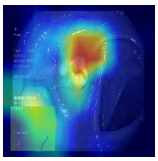
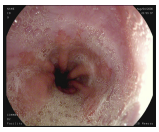
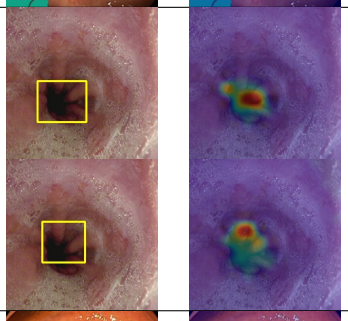
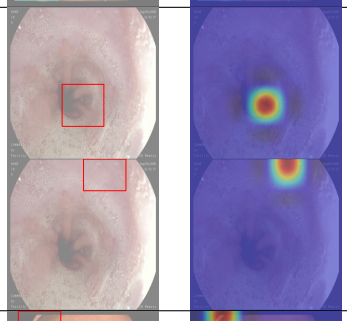
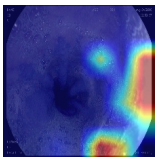
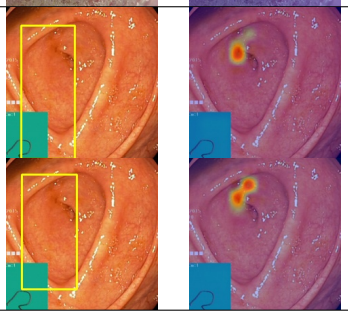
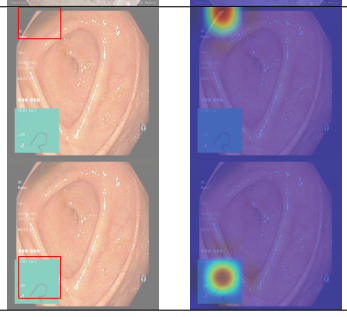
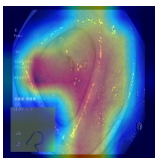
Original Image	ProtoPFormer	ViT-NeT	Swinx x Grad-CAM
 Dyed Lifted Polyps			
 Dyed Resection Margins			
 Esophagitis			
 Normal Cecum			

Table 4.6: Visualizations generated for the Kvasir dataset by the ProtoPFormer, ViT-NeT, Swin x Grad-CAM.

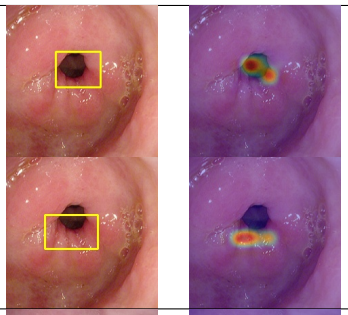
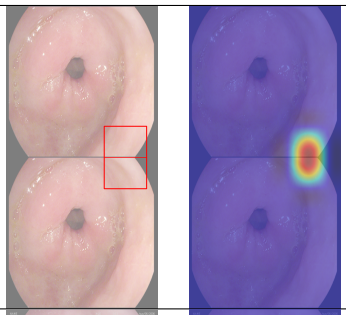
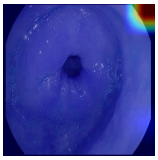
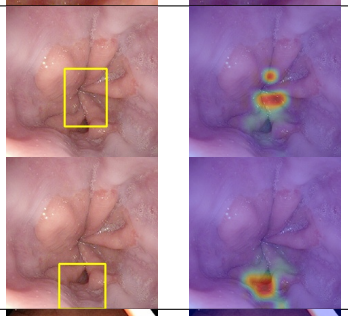
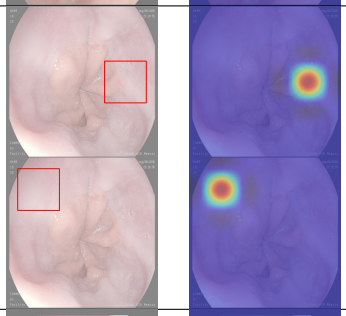
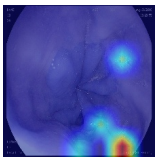
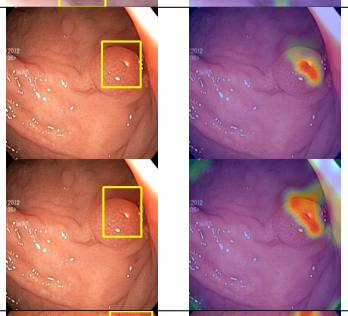
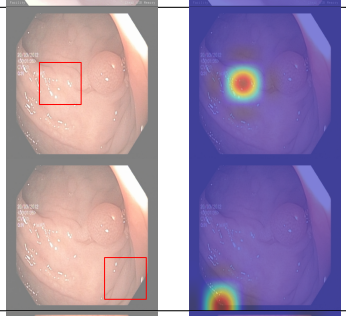
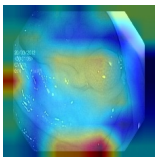
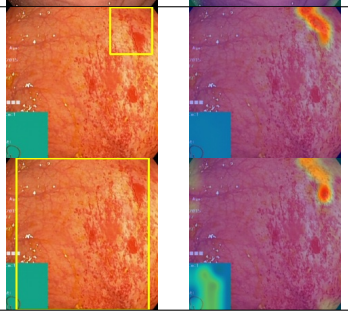
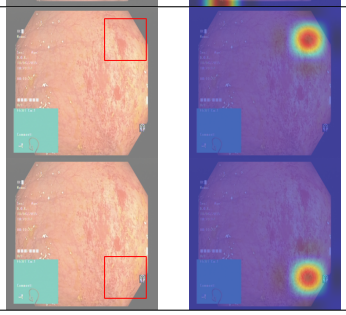
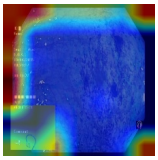
Original Image	ProtoPFormer	ViT-NeT	Swinx x Grad-CAM
Normal Pylorus			
Normal Z Line			
Polyps			
Ulcerative Colitis			

Table 4.7: Visualizations generated for the Kvasir dataset by the ProtoPFormer, ViT-NeT, Swin x Grad-CAM.

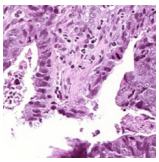
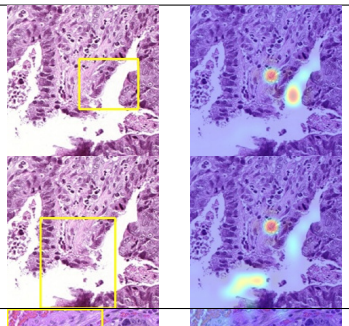
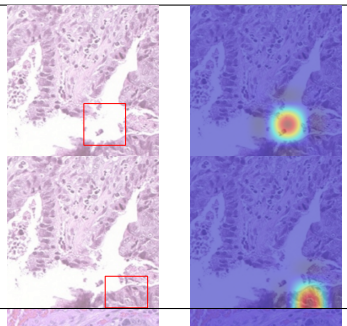
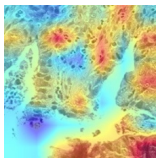
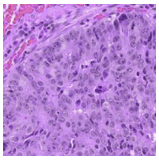
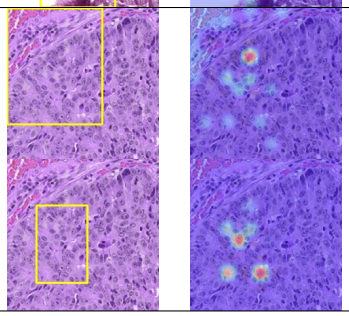
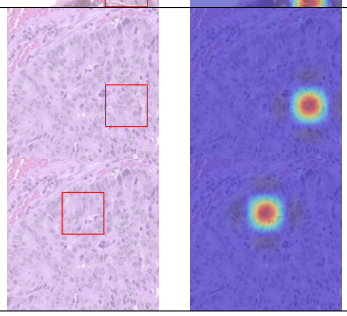
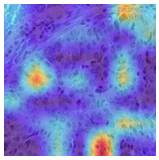
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
 MSIMUT			
 MSS			

Table 4.8: Visualizations generated for the TCGA dataset by the ProtoPFormer, ViT-NeT, Swin x Grad-CAM.

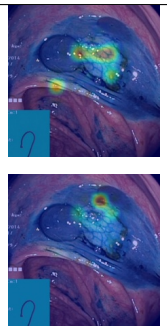
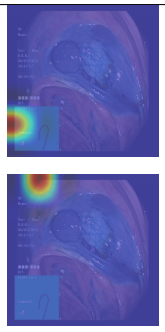
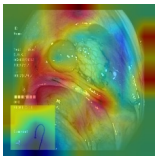
Original Image	ProtoPFormer	ViT-NeT	Swin x Grad-CAM
 Dyed Lifted Polyps			

Table 4.9: Visualizations generated for the Kvasir dataset by the ProtoPFormer, ViT-NeT, and Swin x Grad-CAM. ViT-NeT seems to get distracted by annotations in the image, rather than focusing on the abnormal tissues.

Chapter 5

Conclusion

In this thesis, we adapted and evaluated three Vision Transformer model architectures—ProtoPFormer [31], ViT-NeT [29], and Swin Transformer [30] combined with Grad-CAM [23]—on four distinct medical imaging datasets. These datasets spanned various imaging modalities, including MRI, CT scans, histopathological images, and real gastrointestinal images from endoscopies. Our experiments highlighted significant differences in performance across the models and datasets, providing valuable insights into the strengths and limitations of each approach in the context of medical imaging.

ProtoPFormer consistently outperformed ViT-NeT and Swin Transformer across all datasets, achieving the highest accuracy in every case. Its prototype-based approach demonstrated superior capability in handling the complexities of medical imaging, particularly in tasks like COVID-19 CT scan classification and GI tract image analysis. ViT-NeT, despite its innovative tree-like structure, showed potential but underperformed compared to ProtoPFormer, failing to generate insightful visualizations. The Swin Transformer, used as a baseline, provided lower accuracy, while Grad-CAM seemed to be getting lost when trying to produce visual explanations.

5.1 Discussion

The results indicate that the prototype-based method of ProtoPFormer offers significant advantages in medical imaging tasks. Its ability to use both global and local prototypes allows for a holistic understanding and successful classification of medical images. This approach seems particularly effective in the COVID-19 CT scans [41]and Kvasir GI images [43], maybe indicating that sufficient quantity of data and high variability within the dataset enhance the model's ability to generalize and accurately classify medical images.

However, the relatively lower performance of all models on the Alzheimer's MRI dataset suggests that certain medical imaging tasks may require additional preprocessing, feature extraction, or even specialized architectures to achieve higher accuracy. This dataset's complexity, involving subtle differences between dementia stages, along with its small size, posed a significant challenge, indicating that further research is needed in this area.

Both ViT-NeT and Swin Transformer struggled with interpretability, often focusing on incorrect or irrelevant parts of the images when producing visual explanations. These models did not succeed in highlighting the critical areas required for accurate medical diagnosis, further emphasizing the need for enhanced preprocessing techniques and possibly redesigned model architectures tailored specifically for medical imaging tasks.

5.2 Future Work

Future research could explore several avenues to build on the findings of this thesis. One crucial direction is the development of enhanced preprocessing techniques to address the issue of models focusing on irrelevant

parts of medical images, such as annotations or notes. Techniques such as automatic annotation detection and masking can be implemented to ensure that models concentrate on clinically significant areas.

Another important avenue is the investigation of optimized model architectures tailored specifically for medical imaging tasks. Exploring new transformer-based architectures or hybrid models can potentially address these challenges more effectively. Additionally, extensive hyperparameter tuning and experimenting with different learning rates, batch sizes, and model depths can further fine-tune model performance. Along this, enhancing the interpretability of the models could also be a key area for future work. Given the promising potential of the prototype-based method, it would be beneficial to conduct further investigations towards this direction.

Other suggestions that could contribute in a more comprehensive diagnostic tool invlolve incorporating multimodal data into model training, such as MRI and CT scans with patient history and genetic information. Moreover, exploring cross-domain adaptation techniques,by developing methods to transfer knowledge from one medical imaging domain to another, could also be a valuable direction for future research. For example, techniques that enable models trained on one type of medical image to be effective on another can significantly broaden their utility in various medical contexts.

To conclude, we believe this application of the prototype-based approach to medical imaging classification tasks to be the correct path towards new explainable state-of-the-art models in clinical practice. Despite this thesis not being the definitive attempt to solve the problem we hope to have paved the way to an optimal, exclusively medical interpretable vision Transformer architecture, to be seen in future research.

Chapter 6

Bibliography

[1] Voulodimos, A. et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[2] Jindal, V., Narayan Singh, S., and Suvra Khan, S. "Facial Recognition with Computer Vision". In: *Machine Intelligence and Data Science Applications: Proceedings of MIDAS 2021*. Springer, 2022, pp. 313–330.

[3] Janai, J. et al. "Computer vision for autonomous vehicles: Problems, datasets and state of the art". In: *Foundations and Trends® in Computer Graphics and Vision* 12.1–3 (2020), pp. 1–308.

[4] Hoff, W. A., Nguyen, K., and Lyon, T. "Computer-vision-based registration techniques for augmented reality". In: *Intelligent robots and computer vision XV: algorithms, techniques, active vision, and materials handling*. Vol. 2904. SPIE. 1996, pp. 538–548.

[5] Esteva, A. et al. "Deep learning-enabled medical computer vision". In: *NPJ digital medicine* 4.1 (2021), p. 5.

[6] Jiang, F. et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and vascular neurology* 2.4 (2017).

[7] Weingart, N. S. et al. "Epidemiology of medical error". In: *Bmj* 320.7237 (2000), pp. 774–777.

[8] Graber, M. L., Franklin, N., and Gordon, R. "Diagnostic error in internal medicine". In: *Archives of internal medicine* 165.13 (2005), pp. 1493–1499.

[9] Lee, C. S. et al. "Cognitive and system factors contributing to diagnostic errors in radiology". In: *American Journal of Roentgenology* 201.3 (2013), pp. 611–617.

[10] Gilpin, L. H. et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018, pp. 80–89. DOI: 10.1109/DSAA.2018.00018.

[11] Dosovitskiy, A. et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[12] O'shea, K. and Nash, R. "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (2015).

[13] Wu, Y.-c. and Feng, J.-w. "Development and application of artificial neural network". In: *Wireless Personal Communications* 102 (2018), pp. 1645–1656.

[14] Rojas, R. and Rojas, R. "The backpropagation algorithm". In: *Neural networks: a systematic introduction* (1996), pp. 149–182.

[15] Hecht-Nielsen, R. "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[16] Li, Z. et al. "A survey of convolutional neural networks: analysis, applications, and prospects". In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019.

[17] Albawi, S., Mohammed, T. A., and Al-Zawi, S. "Understanding of a convolutional neural network". In: *2017 international conference on engineering and technology (ICET)*. Ieee. 2017, pp. 1–6.

[18] Vaswani, A. et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[19] Lin, T. et al. "A survey of transformers". In: *AI open* 3 (2022), pp. 111–132.

[20] Tay, Y. et al. "Efficient transformers: A survey". In: *ACM Computing Surveys* 55.6 (2022), pp. 1–28.
[21] Picek, L. et al. "Automatic fungi recognition: deep learning meets mycology". In: *Sensors* 22.2 (2022), p. 633.
[22] Abnar, S. and Zuidema, W. "Quantifying attention flow in transformers". In: *arXiv preprint arXiv:2005.00928* (2020).
[23] Selvaraju, R. R. et al. "Grad-CAM: Why did you say that?" In: *arXiv preprint arXiv:1611.07450* (2016).
[24] Zhou, B. et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.
[25] Selvaraju, R. R. et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.
[26] Ribeiro, M. T., Singh, S., and Guestrin, C. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
[27] Montavon, G. et al. "Layer-wise relevance propagation: an overview". In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 193–209.
[28] Chefer, H., Gur, S., and Wolf, L. "Transformer interpretability beyond attention visualization". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 782–791.
[29] Kim, S., Nam, J., and Ko, B. C. "Vit-net: Interpretable vision transformers with neural tree decoder". In: *International conference on machine learning*. PMLR. 2022, pp. 11162–11172.
[30] Liu, Z. et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.
[31] Xue, M. et al. "Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition". In: *arXiv preprint arXiv:2208.10431* (2022).
[32] Chen, C. et al. "This looks like that: deep learning for interpretable image recognition". In: *Advances in neural information processing systems* 32 (2019).
[33] Grainger, R. et al. "PaCa-ViT: learning patch-to-cluster attention in vision transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18568–18578.
[34] Wang, W. et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 568–578.
[35] Wang, W. et al. "Pvt v2: Improved baselines with pyramid vision transformer". In: *Computational Visual Media* 8.3 (2022), pp. 415–424.
[36] Yu, L. et al. "ex-vit: A novel explainable vision transformer for weakly supervised semantic segmentation". In: *Pattern Recognition* 142 (2023), p. 109666.
[37] Komorowski, P., Baniecki, H., and Biecek, P. "Towards evaluating explanations of vision transformers for medical imaging". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 3725–3731.
[38] Playout, C. et al. "Focused attention in transformers for interpretable classification of retinal images". In: *Medical Image Analysis* 82 (2022), p. 102608.
[39] Demir, U. et al. "Explainable Transformer Prototypes for Medical Diagnoses". In: *arXiv preprint arXiv:2403.06961* (2024).
[40] Basu, S. et al. "RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection". In: *Medical Image Analysis* 83 (2023), p. 102676.
[41] Maftouni, M. et al. "A Robust Ensemble-Deep Learning Model for COVID-19 Diagnosis based on an Integrated CT Scan Images Database". In: *Proceedings of the 2021 Industrial and Systems Engineering Conference*. Virtual Conference, May 2021.
[42] Kather, J. N. *Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples*. Zenodo. Feb. 2019. DOI: 10.5281/zenodo.2530835.
[43] Pogorelov, K. et al. "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection". In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. MMSys'17. Taipei, Taiwan: ACM, 2017, pp. 164–169. ISBN: 978-1-4503-5002-0. DOI: 10.1145/3083187.3083212.
[44] Copeland, B. J. "The Church-Turing Thesis". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Spring 2024. Metaphysics Research Lab, Stanford University, 2024.

[45] Rajaraman, V. "John McCarthy — Father of artificial intelligence". In: *Resonance* 19.3 (Mar. 2014), pp. 198–207. ISSN: 0973-712X. DOI: 10.1007/s12045-014-0027-9. URL:
[46] Kasban, H., El-Bendary, M., and Salama, D. "A comparative study of medical imaging techniques". In: *International Journal of Information Science and Intelligent System* 4.2 (2015), pp. 37–58.
[47] Gurcan, M. N. et al. "Histopathological image analysis: A review". In: *IEEE reviews in biomedical engineering* 2 (2009), pp. 147–171.
[48] Gulati, S. et al. "The future of endoscopy: Advances in endoscopic image innovations". In: *Digestive Endoscopy* 32.4 (2020), pp. 512–522.
[49] Dilsizian, S. E. and Siegel, E. L. "Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment". In: *Current cardiology reports* 16 (2014), pp. 1–8.
[50] Patel, V. L. et al. "The coming of age of artificial intelligence in medicine". In: *Artificial intelligence in medicine* 46.1 (2009), pp. 5–17.
[51] Jha, S. and Topol, E. J. "Adapting to artificial intelligence: radiologists and pathologists as information specialists". In: *Jama* 316.22 (2016), pp. 2353–2354.
[52] Ghosh, A. and Kandasamy, D. "Interpretable artificial intelligence: why and when". In: *American Journal of Roentgenology* 214.5 (2020), pp. 1137–1138.
[53] MENIS-MASTROMICHALAKIS, O. "Explainable Artificial Intelligence: An STS perspective". In: ().
[54] Samek, W., Wiegand, T., and Müller, K.-R. "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *arXiv preprint arXiv:1708.08296* (2017).
[55] Fan, F.-L. et al. "On interpretability of artificial neural networks: A survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6 (2021), pp. 741–760.
[56] Pramoditha, R. "Overview of a neural network's learning process". In: *Data Science* 365 (2022).
[57] Wu, J. "Introduction to convolutional neural networks". In: *National Key Lab for Novel Software Technology. Nanjing University. China* 5.23 (2017), p. 495.
[58] Balaji, S. "Binary Image classifier CNN using TensorFlow". In: *Techiepedia* (2020).
[59] Dehghani, M. et al. "Universal transformers". In: *arXiv preprint arXiv:1807.03819* (2018).
[60] Wolf, T. et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020, pp. 38–45.
[61] Roberts, A. et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Tech. rep. Google, 2019.
[62] Devlin, J. et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).
[63] Floridi, L. and Chiriatti, M. "GPT-3: Its nature, scope, limits, and consequences". In: *Minds and Machines* 30 (2020), pp. 681–694.
[64] Han, K. et al. "A survey on vision transformer". In: *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.
[65] Carion, N. et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.
[66] Beal, J. et al. "Toward transformer-based object detection". In: *arXiv preprint arXiv:2012.09958* (2020).
[67] Fang, Y. et al. "You only look at one sequence: Rethinking transformer in vision through object detection". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26183–26197.
[68] Ma, T. et al. "Oriented object detection with transformer". In: *arXiv preprint arXiv:2106.03146* (2021).
[69] Wang, H. et al. "Max-deeplab: End-to-end panoptic segmentation with mask transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5463–5474.
[70] Wang, Y. et al. "End-to-end video instance segmentation with transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8741–8750.
[71] Strudel, R. et al. "Segmenter: Transformer for semantic segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 7262–7272.
[72] Xie, E. et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.
[73] Xiao, H. et al. "Transformers in medical image segmentation: A review". In: *Biomedical Signal Processing and Control* 84 (2023), p. 104791.

[74] Liu, R. et al. "End-to-end lane shape prediction with transformers". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 3694–3702.
[75] Shi, D. et al. "End-to-end multi-person pose estimation with transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11069–11078.
[76] Liu, Z. et al. "Convtransformer: A convolutional transformer network for video frame synthesis". In: *arXiv preprint arXiv:2011.10185* (2020).
[77] Yang, J. et al. "Recurring the transformer for video action recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14063–14073.
[78] Gabeur, V. et al. "Multi-modal transformer for video retrieval". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 214–229.
[79] Khan, S. et al. "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.
[80] Liartis, J. et al. "Semantic Queries Explaining Opaque Machine Learning Classifiers." In: *DAO-XAI*. 2021.
[81] Mastromichalakis, O. M. et al. "Rule-Based Explanations of Machine Learning Classifiers Using Knowledge Graphs". In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1. 2024, pp. 193–202.
[82] Liartis, J. et al. "Searching for explanations of black-box classifiers in the space of semantic queries". In: *Semantic Web* Preprint (2023), pp. 1–42.
[83] Mastromichalakis, O. M., Liartis, J., and Stamou, G. "Beyond One-Size-Fits-All: Adapting Counterfactual Explanations to User Objectives". In: *arXiv preprint arXiv:2404.08721* (2024).
[84] Filandrianos, G. et al. "Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors". In: *arXiv preprint arXiv:2305.17055* (2023).
[85] Menis Mastromichalakis, O. et al. "Semantic Prototypes: Enhancing Transparency without Black Boxes". In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024, pp. 1680–1688.
[86] Sotirou, T. et al. "MusicLIME: Explainable Multimodal Music Understanding". In: *arXiv preprint arXiv:2409.10496* (2024).
[87] Mahendran, A. and Vedaldi, A. "Visualizing deep convolutional neural networks using natural pre-images". In: *International Journal of Computer Vision* 120 (2016), pp. 233–255.
[88] Simonyan, K., Vedaldi, A., and Zisserman, A. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).
[89] Lin, M., Chen, Q., and Yan, S. "Network in network". In: *arXiv preprint arXiv:1312.4400* (2013).
[90] Gildenblat, J. and contributors. *PyTorch library for CAM methods*. 2021.