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

Αυτο-επιβλεπόμενη Μάθηση για Οπτική Αναγνώριση Συσχετίσεων

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΖΑΧΑΡΙΑ Ν. ΑΝΑΣΤΑΣΑΚΗ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής
Συνεπιβλέπων: Γεώργιος Αλεξανδρίδης
Εργαστηριακό Διδακτικό Προσωπικό

Αθήνα, Μάρτιος 2023

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

Αυτο-επιβλεπόμενη Μάθηση για Οπτική Αναγνώριση Συσχετίσεων

Μελέτη και υλοποίηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΖΑΧΑΡΙΑ Ν. ΑΝΑΣΤΑΣΑΚΗ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής
Συνεπιβλέπων: Γεώργιος Αλεξανδρίδης
Εργαστηριακό Διδακτικό Προσωπικό

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15/03/2023.

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

Αθήνα, Μάρτιος 2023

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

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

(Υπογραφή)

.....
Ζαχαρίας Ν. Αναστασάκης

15/03/2023

Περίληψη

Το πρόβλημα της δημιουργίας γράφου σκηνής στην όραση υπολογιστών περιλαμβάνει τη δημιουργία μιας αναπαράστασης μιας εικόνας με βάση ένα γράφο. Ο γράφος αποτελείται από αντικείμενα (κόμβους) με τις σχέσεις (ακμές) μεταξύ τους, αναπαριστώντας τη σκηνή και τα στοιχεία της με δομημένο τρόπο. Στόχος είναι η εξαγωγή σημασιολογικών πληροφοριών από μια εικόνα και η αναπαράστασή τους σε μορφή που να μπορεί εύκολα να αναλυθεί και να γίνει κατανοητή.

Παρατηρώντας τη συμπεριφορά σύγχρονων μοντέλων στη βιβλιογραφία, καθίσταται σαφές πως η πλειονότητα των σημερινών μεθόδων βασίζεται στην επιβλεπόμενη μάθηση, όπου το μοντέλο εκπαιδεύεται σε ένα μεγάλο όγκο επισημειωμένων δεδομένων εικόνων. Παρά την επιτυχία της, η μάθηση με επίβλεψη είναι ακριβή και χρονοβόρα, καθώς απαιτεί μεγάλες ποσότητες επισημειωμένων δεδομένων και ετικετών. Η έλλειψη μάθησης με αυτο-επίβλεψη στη δημιουργία γράφων σκηνής είναι εύκολο να παρατηρηθεί. Ένας από τους κύριους λόγους που μπορεί να συμβαίνει αυτό είναι ότι η δημιουργία γράφων σκηνής είναι μια σύνθετη εργασία που απαιτεί την εξαγωγή πληροφοριών από εικόνες, η οποία μπορεί να είναι πρόκληση για τις μεθόδους χωρίς επίβλεψη.

Η συνεισφορά αυτής της διπλωματικής εργασίας αφορά την εισαγωγή μίας νέας αρχιτεκτονικής μοντέλου αυτο-επιβλεπόμενης μάθησης, το οποίο προ-εκπαιδεύεται σε μη επισημειωμένα δεδομένα και καταφέρνει να πετύχει έως και 7% σχετική βελτίωση συγκριτικά με επανυλοποιήσεις μεθόδων της βιβλιογραφίας όταν εκπαιδεύεται με λίγα επισημειωμένα δεδομένα (few-shot learning) τόσο στο VRD όσο και στο VG200 σύνολο δεδομένων, δύο από τα δημοφιλέστερα σύνολα δεδομένων του προβλήματος.

Λέξεις Κλειδιά

Οπτική Αναγνώριση Συσχετίσεων, Παραγωγή Γράφων Σκηνής, Αυτο-επιβλεπόμενη Μάθηση, Εκπαίδευση με λίγα δείγματα, Μετασχηματιστές, VRD, VG200

Abstract

The task of Scene Graph Generation (SGG) in computer vision involves creating a graph-based representation of an image. The graph consists of objects (nodes) with relationships (edges) between them, representing the scene and its elements in a structured way. The goal is to extract semantic information from an image and represent it in a form that can be easily analyzed and understood.

Looking at the behavior of modern, state-of-the-art models in the literature, it becomes clear that the majority of current methods are based on supervised learning, where the model is trained on a large amount of labeled image data. Despite its success, supervised learning is expensive and time-consuming, as it requires large amounts of data and labels. The lack of self-supervised learning in scene graph generation can be easily be observed. One of the main reasons why this may be the case is that scene graph generation is a complex task that requires extracting information from images, which can be challenging for unsupervised methods.

The contribution of this thesis is the introduction of a new self-supervised learning model architecture, which is pre-trained on unlabeled data and manages to achieve up to 7% relative improvement compared to reimplementations of methods in the literature when trained with a few labeled data (few-shot learning) on both VRD and VG200 datasets, two of the most popular datasets in problem.

Keywords

Visual Relationship Detection, Scene Graph Generation (SGG), Self Supervised Learning, Few-shot Learning, Transformers, VRD, VG200

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον κ. Κόλλια Στέφανο, Καθηγητή Ε.Μ.Π., για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης (AILS). Επίσης θα ήθελα να ευχαριστήσω τους κ. κ. Στάμου Γεώργιο, Καθηγητή Ε.Μ.Π., και Βουλόδημο Αθανάσιο, Επίκουρο Καθηγητή Ε.Μ.Π., για την τιμή που μου έκαναν να είναι μέλη της τριμελούς εξεταστικής επιτροπής. Ακόμα ευχαριστώ ιδιαίτερα τον Δρ. Γεώργιο Αλεξανδρίδη για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε, καθώς και τους Βασίλη Πιτσικάλη, Μάρκο Διοματάρη και Δημήτρη Μαλλή που συνεπίβλεψαν την διπλωματική μου στα πλαίσια συνεργασίας του εργαστηρίου AILS και της Deeplab. Ο ένας χρόνος συνεργασίας μαζί τους, μου έμαθε πολλά πράγματα και θα αποτελέσει γερό θεμέλιο για την εξέλιξη της μετέπειτα πορείας μου.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την καθοδήγηση που μου προσέφεραν όλα αυτά τα χρόνια, τους φίλους μου για την υποστήριξη τους, και την κοπέλα μου η οποία μου στάθηκε όλο αυτό τον καιρό, πίστευε σε εμένα και της οποίας η βοήθεια ήταν ιδιαίτερα σημαντική καθόλη την διάρκεια εκπόνησης της διπλωματικής εργασίας.

Αθήνα, Μάρτιος 2023

Ζαχαρίας Ν. Αναστασάκης

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Ο τομέας της όρασης υπολογιστών χρησιμοποιεί τη μηχανική μάθηση με στόχο την ερμηνεία και την κατανόηση οπτικών πληροφοριών από τον κόσμο, όπως εικόνες και βίντεο. Είναι ένα πολυεπιστημονικό πεδίο που συνδυάζει έννοιες από την επεξεργασία εικόνας και τη μηχανική μάθηση για να επιτρέψει στους υπολογιστές να κατανοούν και να λαμβάνουν αποφάσεις με βάση οπτικά δεδομένα. Τα τελευταία χρόνια, έχει δοθεί μεγάλη έμφαση στην ταξινόμηση εικόνων και στην ανίχνευση αντικειμένων, που περιλαμβάνουν αλγόριθμους για την αναγνώριση και ταξινόμηση αντικειμένων μέσα σε μια εικόνα. Η υπολογιστική όραση είναι ένας ενεργός και ταχέως εξελισσόμενος τομέας, με νέες τεχνικές και εφαρμογές να αναπτύσονται συνεχώς. Η εξέλιξή του είναι συναρπαστική, με τεράστιες δυνατότητες και θα συνεχίζει να διαδραματίζει σημαντικό ρόλο στη διαμόρφωση του μέλλοντος της τεχνολογίας και της αλληλεπίδρασης ανθρώπου-υπολογιστή.

Ωστόσο, η αναγνώριση αντικειμένων από μόνη της δεν αρκεί για την κατανόηση των σκηνών σε μια εικόνα. Για να κατανοήσουμε τις σκηνές και τα αντικείμενα στις εικόνες με πιο λεπτομερή τρόπο, προέκυψε το πρόβλημα της ανίχνευσης οπτικών συσχετίσεων (Visual Relationship Detection - VRD). Το πρόβλημα της ανίχνευσης οπτικών συσχετίσεων περιλαμβάνει τον εντοπισμό και την περιγραφή των σχέσεων μεταξύ των αντικειμένων σε μια εικόνα. Για παράδειγμα, ένας αλγόριθμος VRD μπορεί να προσδιορίσει ότι υπάρχει ένα άτομο που ιππεύει ένα άλογο σε μια εικόνα ή ότι ένα βιβλίο βρίσκεται επάνω σε ένα τραπέζι.

Το πρόβλημα της αναγνώρισης συσχετίσεων είναι αρκετά απαιτητικό λόγω της μεγάλης ποικιλίας σχέσεων που μπορεί να υπάρχουν μεταξύ των αντικειμένων. Μπορεί να υπάρχουν χιλιάδες διαφορετικές σχέσεις και κάθε εικόνα μπορεί να περιέχει πολλαπλές σχέσεις. Μια άλλη πρόκληση του προβλήματος είναι η αντιμετώπιση της μεγάλης μεταβλητότητας στον τρόπο με τον οποίο οι σχέσεις μπορούν να απεικονιστούν σε μια εικόνα. Για παράδειγμα, η ίδια σχέση, όπως "άτομο πάνω από ένα άλογο", μπορεί να απεικονιστεί με πολλούς διαφορετικούς τρόπους. Το άτομο μπορεί να ιππεύει το άλογο, να κάθεται πάνω του ή ακόμα και να ξαπλώνει πάνω του. Το άλογο μπορεί να στέκεται ακίνητο, να καλπάζει ή να πηδά. Για να αναγνωρίσει και να περιγράψει σωστά αυτές τις σχέσεις, ο αλγόριθμος πρέπει να μπορεί να χειριστεί αυτή τη μεταβλητότητα.

Πρόσφατα, μια πληθώρα τεχνικών βαθιάς μάθησης έχουν εφαρμοστεί για την επίλυση αυτού του προβλήματος, μερικές από τις οποίες έχουν επιτύχει εξαιρετικές επιδόσεις χωρίς όμως να τελικά να καταφέρνουν πάντα να κωδικοποιούν την εννοιολογική σημασία των σχέσεων που εμπεριέχονται μέσα σε μία εικόνα.

1.1 Περιγραφή Προβλήματος

Ως σχέση μεταξύ δύο οντοτήτων ορίζουμε μία τριπλέτα της μορφής <υποκείμενο-κατηγόρημα-αντικείμενο> (subject-predicate-object) που υποδηλώνει πως το υποκείμενο σχετίζεται μέσω του κατηγορήματος με το αντικείμενο. Η δημιουργία γράφων σκηνής (Scene Graph Generation - SGG) αποσκοπεί στην κατανόηση των σχέσεων μεταξύ των οντοτήτων που βρίσκονται μέσα σε μία εικόνα και στην αναπαράσταση τους σε μορφή ενός κατευθυνόμενου γράφου. Οι κόμβοι του κατευθυνόμενου γράφου αποτελούν τα αντικείμενα που εντοπίζονται στην εικόνα ενώ οι ακμές αντιπροσωπεύουν τις σχέσεις με τις οποίες συνδέονται με κατεύθυνση από το υποκείμενο προς το αντικείμενο.

Για την καλύτερη δυνατή αξιολόγηση της απόδοσης ενός μοντέλου SGG χωρίζουμε το πρόβλημα μας σε επιμέρους υποπροβλήματα ανάλογα με το αν γνωρίζουμε τις συντεταγμένες των περιγραμμάτων (Bounding Boxes) των εντοπισμένων αντικειμένων μιας εικόνας, τις κατηγορίες των εντοπισμένων αντικειμένων αλλά και αν σχετίζονται μεταξύ τους τα αντικείμενα. Πιο συγκεκριμένα τα υποπροβλήματα είναι 5 και είναι τα εξής:

1. **Ανίχνευση σχέσης** (Predicate Detection - PredDet): Δοθέντων των περιγραμμάτων, των κατηγοριών των αντικειμένων και των ζευγών που αλληλεπιδρούν προβλέπουμε τις σχέσεις μεταξύ τους.

2. **Ταξινόμηση σχέσης** (Predicate Classification - PredCls): Δοθέντων των περιγραμμάτων και των κατηγοριών των αντικειμένων αποφασίζουμε ποια ζεύγη αλληλεπιδρούν και για αυτά προβλέπουμε σχέσεις.

3. **Ταξινόμηση γράφων σκηνής** (Scene Graph Classification - SGCls): Δοθέντων των περιγραμμάτων των αντικειμένων, τα κατηγοριοποιούμε, βρίσκουμε ποια αλληλεπιδρούν και προβλέπουμε σχέσεις.

4. **Δημιουργία γράφων σκηνής** (SGG): Τίποτα δεν είναι γνωστό. Εντοπίζουμε και κατηγοριοποιούμε τα αντικείμενα, βρίσκουμε ποια αλληλεπιδρούν και προβλέπουμε σχέσεις.

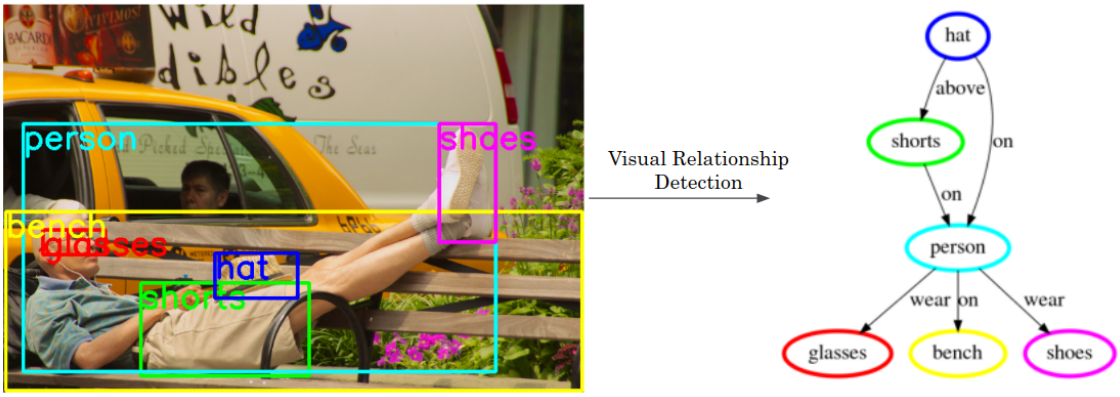
5. **Ανίχνευση φράσης** (Phrase Detection - PhrDet): Ομοίως με SGG μόνο που αξιολογεί την επικάλυψη του περιγράμματος της ένωσης του υποκειμένου και του αντικειμένου (να είναι δηλαδή < 0.5) αντί το ξεχωριστό τους γινόμενο.

Ο Πίνακας 1.1 παρακάτω παρουσιάζει τα παραπάνω υποπροβλήματα κατηγοριοποιημένα.

Στο υπόλοιπο της παρούσας διπλωματικής εργασίας, θα προχωρήσουμε με την υπόθεση ότι έχουμε πρόσβαση και στις τρεις πληροφορίες, δηλαδή στις συντεταγμένες των περιγραμμάτων, στις κατηγορίες των αντικειμένων αλλά και στην πληροφορία για το ποια αντικείμενα αλληλεπιδρούν μεταξύ τους και ποια όχι. Αυτό μας επιτρέπει να επικεντρωθούμε σε μια συγκεκριμένη παραλλαγή του προβλήματος που είναι η ανίχνευση σχέσεων (PredDet). Ένα παράδειγμα ανίχνευσης σχέσεων παρουσιάζεται στο Σχήμα 1.1. Αυτή η επιλογή της προσέγγισης γίνεται προκειμένου να απομονώσουμε σαφώς το πρόβλημα που στοχεύουμε να

Υποπρόβλημα	Συντεταγμένες Περιγραμμάτων	Κατηγορίες Αντικειμένων	Αλληλεπίδραση
PredDet	ναι	ναι	ναι
PredCls	ναι	ναι	οχι
SGCls	ναι	οχι	οχι
SGGen	οχι	οχι	οχι
PhrDet	οχι	οχι	οχι

Πίνακας 1.1: *Απαρίθμηση των παραλλαγών της ανίχνευσης οπτικών σχέσεων.* ***ναι*** *σημαίνει πως η συγκεκριμένη παραλλαγή χρησιμοποιεί την αντίστοιχη πληροφορία ενώ σε αντίθετη περίπτωση σημειώνουμε* ***οχι****.*



Σχήμα 1.1: *Εντοπισμός οπτικών σχέσεων (Visual Relationship Detection). Για είσοδο μία εικόνα, τις ανιχνευμένες οντότητες και τις κατηγορίες κάθε ανιχνευμένης οντότητας υπολογίζουμε έναν κατευθυνόμενο γράφο όπου κάθε κόμβος αντιστοιχεί σε μία οντότητα και κάθε ακμή συνδέει το υποκείμενο με το αντικείμενο σύμφωνα με τη σχέση τους. Ο παραπάνω κατευθυνόμενος γράφος αναφέρεται σε ένα εκπαιδευμένο δίκτυο ATR-Net [5].*

διερευνήσουμε από τις πολυπλοκότητες και τους περιορισμούς ενός δικτύου ανίχνευσης και αναγνώρισης αντικειμένων (Object Detector).

Για την αξιολόγηση του προβλήματος της ανίχνευσης οπτικών σχέσεων χρησιμοποιείται η μετρική της Ανάκλησης@x (Recall@x - R@x) η οποία μετρά το ποσοστό των σωστών σχέσεων που περιλαμβάνονται στις πρώτες x προβλέψεις αφού τις κατατάξουμε σύμφωνα με την πιθανότητα πρόβλεψης σε φθίνουσα σειρά. Μία άλλη παράμετρος k μετρά τον μέγιστο αριθμό προβλέψεων που επιτρέπουμε ανά ακμή. Για k = 1 θεωρούμε σωστή την πρόβλεψη για μία ακμή όταν η πρώτη σχέση (αυτή με τη μεγαλύτερη πιθανότητα) ταυτίζεται με την πραγματική. Επειδή πολλές φορές κάποιες ακμές είναι επισημειωμένες με πάνω από μία σχέση, έχει νόημα να αυξήσουμε το k μεταβαίνοντας έτσι σε ένα πρόβλημα πολλαπλών-κλάσεων και πολλαπλών-επισημειώσεων (multi-class multi-label classification). Έτσι στο [5] ορίζουν τη μετρική Rk@x όπου για n ζευγάρια αντικειμένων σε μία εικόνα κρατάει τις x πιο πιθανές προβλέψεις από συνολικά nk για να μετρήσει το Recall. Στην παρούσα διπλωματική, οποιαδήποτε αναφορά σε Recall υπονοεί k = 1.

1.2 Εφαρμογές

Η οπτική αναγνώριση σχέσεων (VRD) είναι μια ισχυρή τεχνική στην όραση υπολογιστών που έχει πολυάριθμες εφαρμογές σε διάφορους τομείς. Πιο συγκεκριμένα:

1. **Κατανόηση σκηνών**: Η οπτική αναγνώριση σχέσεων μπορεί να προσδιορίσει τις σχέσεις μεταξύ των αντικειμένων σε μια εικόνα για να βελτιώσει την κατανόηση σκηνών. Ανιχνεύοντας τη σχέση μεταξύ των αντικειμένων, μπορεί να προσδιοριστεί ο τρόπος με τον οποίο αλληλεπιδρούν τα αντικείμενα και να παράξει μεγαλύτερο νόημα στη σκηνή. Αυτό μπορεί να είναι χρήσιμο για εργασίες όπως η επιτήρηση (surveillance), όπου ο εντοπισμός αλληλεπιδράσεων μεταξύ αντικειμένων μπορεί να είναι ζωτικής σημασίας για την ανίχνευση ύποπτης δραστηριότητας.

2. **Αναζήτηση εικόνων και βίντεο**: Η οπτική αναγνώριση σχέσεων μπορεί να βελτιώσει την αναζήτηση εικόνων και βίντεο, επιτρέποντας στους χρήστες να αναζητούν εικόνες και βίντεο με βάση τις σχέσεις μεταξύ των αντικειμένων σε αυτά. Για παράδειγμα, ένας χρήστης θα μπορούσε να αναζητήσει εικόνες που περιέχουν ένα άτομο και έναν σκύλο που αλληλεπιδρούν μεταξύ τους, ή βίντεο που δείχνουν ένα αυτοκίνητο να πλησιάζει σε μια πινακίδα στοπ. Αυτό μπορεί να είναι χρήσιμο για ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της ανάκτησης εικόνων και βίντεο βάσει περιεχομένου, της παρακολούθησης αντικειμένων και της ανίχνευσης συμβάντων.

3. **Ευφυής επεξεργασία εικόνας**: Η οπτική αναγνώριση σχέσεων μπορεί να ανιχνεύει αυτόματα τις σχέσεις μεταξύ αντικειμένων σε μια εικόνα, επιτρέποντας την πιο έξυπνη επεξεργασία εικόνας. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί για την αυτόματη περικοπή μιας εικόνας με βάση τα αντικείμενα και τις σχέσεις που ανιχνεύονται σε αυτήν, ή για την αφαίρεση ή αντικατάσταση αντικειμένων διατηρώντας τις σχέσεις μεταξύ τους. Αυτό μπορεί να είναι χρήσιμο για εργασίες όπως η επεξεργασία εικόνας, η δημιουργία περιεχομένου και η επεξεργασία ψηφιακών μέσων.

4. **Αυτόνομη οδήγηση**: Η οπτική αναγνώριση σχέσεων μπορεί να βοηθήσει τα αυτο-οδηγούμενα αυτοκίνητα να κατανοήσουν τις σχέσεις μεταξύ αντικειμένων στο δρόμο, όπως αυτοκίνητα, πεζούς και φανάρια. Ανιχνεύοντας τις σχέσεις μεταξύ των αντικειμένων, τα αυτο-οδηγούμενα αυτοκίνητα μπορούν να προβλέπουν καλύτερα τις κινήσεις άλλων αντικειμένων και να λαμβάνουν πιο τεκμηριωμένες αποφάσεις σχετικά με τον τρόπο πλοήγησης στο περιβάλλον. Αυτό μπορεί να βελτιώσει την ασφάλεια και την αποτελεσματικότητα στο δρόμο, καθιστώντας τα αυτο-οδηγούμενα αυτοκίνητα πιο πρακτικά για καθημερινή χρήση.

5. **Υγειονομική περίθαλψη**: Η οπτική αναγνώριση σχέσεων μπορεί να χρησιμοποιηθεί για τον εντοπισμό σχέσεων μεταξύ διαφορετικών ιατρικών εικόνων, βοηθώντας τους γιατρούς να εντοπίζουν ανωμαλίες και να διαγιγνώσκουν ασθένειες. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί για την ανίχνευση των σχέσεων μεταξύ διαφορετικών τύπων ιστών σε μια μαγνητική τομογραφία, επιτρέποντας στους γιατρούς να εντοπίζουν με μεγαλύτερη ακρίβεια τις ανωμαλίες. Αυτό μπορεί να είναι χρήσιμο για ένα ευρύ φάσμα ιατρικών εφαρμογών, συμπεριλαμβανομένης της ανίχνευσης καρκίνου.

1.3 Προκλήσεις

Η παραγωγή γράφου σκηνής για μια εικόνα αποτελεί ένα ιδιαίτερα δύσκολο πρόβλημα το οποίο συνδυάζει πολλαπλές προκλήσεις που πρέπει να αντιμετωπιστούν για τη λύση του. Μία από τις κύριες προκλήσεις είναι η πολυπλοκότητα των σχέσεων, οι οποίες μπορεί να είναι εξαιρετικά ποικίλες και να περιλαμβάνουν πολλαπλά αντικείμενα και τύπους σχέσεων. Για παράδειγμα, ένας γράφος σκηνής μπορεί να περιλαμβάνει σχέσεις όπως "άτομο που κρατάει μια μπάλα" ή "καρέκλα δίπλα σε τραπέζι", καθιστώντας δύσκολη την ακριβή πρόβλεψή τους από το μοντέλο. Μια άλλη πρόκληση είναι οι ποικίλες κατηγορίες αντικειμένων που υπάρχουν στα γραφήματα σκηνών. Οι γράφοι σκηνής συχνά περιλαμβάνουν μεγάλο αριθμό κατηγοριών αντικειμένων, γεγονός που μπορεί να καταστήσει δύσκολο για το μοντέλο να προβλέψει με ακρίβεια όλες αυτές τις κατηγορίες σε μια εικόνα. Αυτό ισχύει ιδιαίτερα για τις λεπτόκοκκες κατηγορίες, όπου τα αντικείμενα μπορεί να μοιάζουν πολύ στην εμφάνιση αλλά να ανήκουν σε διαφορετικές κατηγορίες. Επιπλέον, οι αποκρύψεις αντικειμένων μπορεί επίσης να δυσχεράνουν τον εντοπισμό και την πρόβλεψη των σχέσεων σε μια εικόνα από το μοντέλο. Στις εικόνες του πραγματικού κόσμου, τα αντικείμενα μπορεί να είναι μερικώς ή πλήρως καλυμμένα, καθιστώντας δύσκολη την ακριβή πρόβλεψη των σχέσεων μεταξύ των αντικειμένων από το μοντέλο. Τέλος, η επισημείωση των δεδομένων είναι μια εργασία που συχνά εκτελείται με ασυνεπή τρόπο, με αποτέλεσμα να αποδίδεται διαφορετικός βαθμός σημαντικότητας στις ίδιες πληροφορίες για διαφορετικές εικόνες. Ενώ μια εικόνα μπορεί να θεωρεί ορισμένες πληροφορίες σημαντικές και να έχουν επισημειωθεί ανάλογα, οι ίδιες πληροφορίες σε μια άλλη εικόνα μπορεί να αγνοούνται και να μην έχουν επισημειωθεί καθόλου. Επιπλέον, δεν είναι ασυνήθιστο το φαινόμενο να αγνοούνται εντελώς σημαντικές οντότητες ή σχέσεις που υπάρχουν σε μια εικόνα.

1.4 Κίνητρο και Συνεισφορά

1.4.1 Κίνητρο

Βασικό κίνητρο αυτής της διπλωματικής αποτελεί η παρατήρηση πως τα τελευταία χρόνια η επιβλεπόμενη μάθηση (supervised learning) είναι η κυρίαρχη προσέγγιση για τη δημιουργία γράφων σκηνής, όπου τα μοντέλα ανεξαρτήτως της αρχιτεκτονικής τους εκπαιδεύονται σε μεγάλα σύνολα δεδομένων επισημειωμένων εικόνων για την πρόβλεψη κατηγοριών αντικειμένων και σχέσεων μεταξύ αντικειμένων.

Η αυτοεπιβλεπόμενη μάθηση (self-supervised learning), από την άλλη πλευρά, περιλαμβάνει την εκπαίδευση μοντέλων με τη χρήση μιας προαπαιτούμενης εργασίας που παρέχει επίβλεψη χωρίς να απαιτούνται επιγεγραμμένα δεδομένα (labelled data). Ένα από τα κύρια κίνητρα για τη χρήση της αυτοεπιβλεπόμενης μάθησης στη δημιουργία γράφων σκηνής είναι να ξεπεραστεί η περιορισμένη διαθεσιμότητα επιγεγραμμένων δεδομένων. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο σε τομείς όπου είναι δύσκολο να ληφθούν επιγεγραμμένα δεδομένα ή όπου η επισημείωση δεδομένων είναι χρονοβόρα και ακριβή. Επιπλέον, η αυτοεπίβλεπόμενη μάθηση μπορεί επίσης να χρησιμοποιηθεί για την εκμάθηση ουσιαστικών αναπαραστάσεων εικόνων που μπορούν να χρησιμοποιηθούν ως είσοδος σε άλλες εργασίες, συμπεριλαμβα-

νομένης της δημιουργίας γράφων σκηνών. Για παράδειγμα, η αυτοεπιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για την εκμάθηση οπτικών αναπαραστάσεων αντικειμένων και σχέσεων που μπορούν να χρησιμοποιηθούν για τη βελτίωση της απόδοσης των μοντέλων δημιουργίας γράφων σκηνών.

Η αυτοεπιβλεπόμενη μάθηση έχει τη δυνατότητα να αποτελέσει ένα πολύτιμο εργαλείο για τη δημιουργία γράφων σκηνών, ιδίως σε περιπτώσεις όπου τα επιγεγραμμένα δεδομένα είναι περιορισμένα. Ωστόσο, απαιτείται περισσότερη έρευνα για την πλήρη κατανόηση του αντίκτυπου της αυτοεπιβλεπόμενης μάθησης στην απόδοση της δημιουργίας γράφων σκηνής και για τον προσδιορισμό των καλύτερων τρόπων ενσωμάτωσης της αυτοεπιβλεπόμενης μάθησης στα υπάρχοντα μοντέλα δημιουργίας γράφων σκηνής.

1.4.2 Συνεισφορά

Με βάση τις παρατηρήσεις που κάναμε στην Ενότητα 1.4.1, με την παρούσα διπλωματική εργασία συμβάλλουμε στην αντιμετώπιση των προβλημάτων που δημιουργεί η διαδικασία της επισημείωσης δεδομένων, προτείνοντας ένα νέο μοντέλο αυτο-επιβλεπόμενης μάθησης το οποίο είναι ικανό να μάθει καλύτερα τις σχέσεις μεταξύ των οντοτήτων μίας εικόνας, συγκριτικά με υπάρχουσες δουλειές στην βιβλιογραφία, όταν αυτά εκπαιδεύονται με λίγα δείγματα. Πιο συγκεκριμένα:

- Προτείνουμε μια νέα αρχιτεκτονική μοντέλου (Vision Decoder - ViD) το οποίο προεκπαιδεύεται σε ένα μεγάλο όγκο δεδομένων χωρίς την ανάγκη ύπαρξης ετικετών σχέσεων, το οποίο στη συνέχεια χρησιμοποιείται σαν εξαγωγέας οπτικών χαρακτηριστικών των οντοτήτων, ικανών να εκπαιδεύσουν έναν κατηγοριοποιητή σχέσεων με λίγα δείγματα (few shot learning).

- Πραγματοποιούμε ποσοτική αλλά και ποιοτική σύγκριση μεταξύ της μεθόδου που προτείνουμε αλλά και με τη σχετική βιβλιογραφία στα VRD [1] και VG200 [9], τα δύο δημοφιλέστερα σύνολα δεδομένων του προβλήματος, όπου πετυχαίνουμε 7% και 5% μέγιστη σχετική βελτίωση αντίστοιχα, χρησιμοποιώντας 5 επισημειωμένα δείγματα ανά σχέση στην εκπαίδευση.

1.5 Δομή Διπλωματικής Εργασίας

Το υπόλοιπο της παρούσας διπλωματικής εργασίας οργανώνεται ως εξής. Στο Κεφάλαιο 2 γίνεται μια γενική ανασκόπηση της βιβλιογραφίας στον χώρο της οπτικής αναγνώρισης συσχετίσεων, στο Κεφάλαιο 3 περιγράφεται αναλυτικά η αρχιτεκτονική του δικτύου των μετασχηματιστών καθώς χρησιμοποιούμε μία παραλλαγή του δικτύου αυτού στην προτεινόμενη αρχιτεκτονική μας και στο Κεφάλαιο 4 περιγράφουμε την αρχιτεκτονική που προτείνουμε για την εκπαίδευση δικτύων με λίγα δείγματα. Αμέσως μετά, στο Κεφάλαιο 5 παρουσιάζουμε και αναλύουμε τα αποτελέσματα που προέκυψαν από τα πειράματα μας και τέλος κλείνουμε με το Κεφάλαιο 6, όπου περιγράφουμε τα συμπεράσματα μας και προτείνουμε πιθανές μελλοντικές κατευθύνσεις.

Κεφάλαιο 2

Ανασκόπηση Βιβλιογραφίας

2.1 Βασική Αρχιτεκτονική Δικτύων Παραγωγής Γράφου Σκηνής

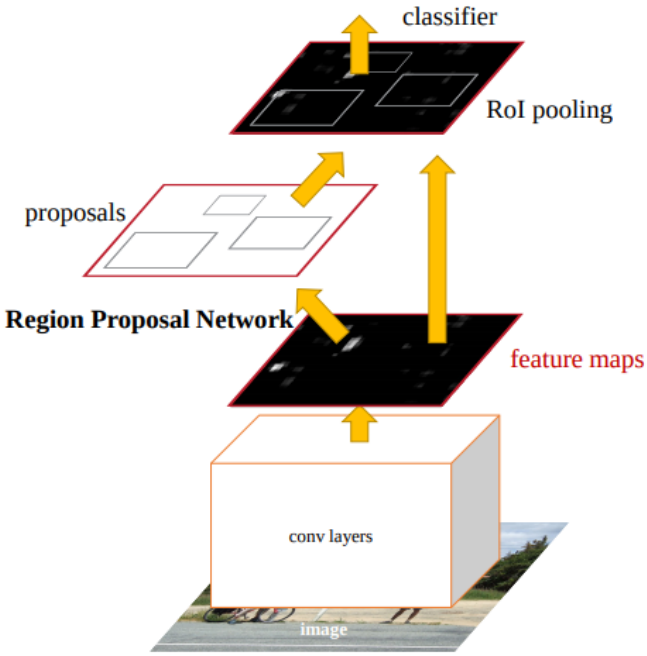
Όπως αναφέρουμε και παραπάνω, σε αυτήν την διπλωματική εργασία θα ασχοληθούμε μια συγκεκριμένη παραλλαγή του προβλήματος της παραγωγής γράφων σκηνής (SGG) που είναι η ανίχνευση σχέσεων έχοντας δεδομένες τις πληροφορίες των περιγραμμάτων, τις κατηγορίες των αντικειμένων αλλά και πληροφορία για το αν τα εντοπισμένα αντικείμενα αλληλεπιδρούν ή όχι. Επομένως, δεν χρειάζεται να χρησιμοποιήσουμε κάποιο δίκτυο ανίχνευσης αντικειμένων κατά την διάρκεια της εκπαίδευσης ή και της αξιολόγησης του μοντέλου μας εφόσον παίρνουμε τις αντίστοιχες πληροφορίες για τις συντεταγμένες των περιγραμμάτων των αντικειμένων και τις κατηγορίες τους (classes) από τα σύνολα δεδομένων (Datasets).

Παρόλ'αυτά, έχουν γίνει αρκετές μελέτες στα προβλήματα ανίχνευσης φράσης, όπου χρειάζεται η επανεκπαίδευση ενός ανιχνευτή αντικειμένων παράλληλα με την εκπαίδευση του ανιχνευτή σχέσεων. Στις περισσότερες ερευνητικές μελέτες χρησιμοποιείται το δίκτυο Faster R-CNN [10] (Σχήμα 2.1) με ραχοκοκαλιά το VGG-16 [11] ή το ResNet [12] για τον εντοπισμό των αντικειμένων σε συνδυασμό με κάποιο επιπρόσθετο δίκτυο για κατηγοριοποίηση των εντοπισμένων αντικειμένων [13, 14, 15, 16, 17, 18]. Ακόμη, αρκετές σύγχρονες μελέτες στην βιβλιογραφία [19] χρησιμοποιούν το δίκτυο DETR [20] (Σχήμα 2.2), με ραχοκοκαλιά ένα συνελικτικό δίκτυο (Convolutional Neural Network - CNN), ως δίκτυο ανίχνευσης αντικειμένων, το οποίο αποτελείται από δίκτυα μετασχηματιστών (Transformers) [8], με επιδόσεις πολύ κοντά στο Faster R-CNN, αλλά με μία αρχιτεκτονική αρκετά πιο απλή και κατανοητή.

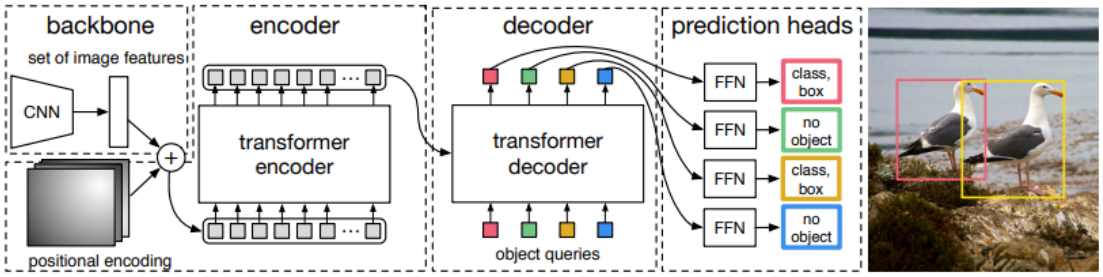
Επομένως ανάλογα με το ποιο υποπρόβλημα προσπαθούν να λύσουν (Πίνακας 1.1), είτε εκπαιδεύουν από κοινού τον ανιχνευτή αντικειμένων με τον ανιχνευτή σχέσεων (Multi-task learning) [5], είτε παγώνουν τα βάρη (weights) του ανιχνευτή αντικειμένων και εκπαιδεύουν αποκλειστικά τον ανιχνευτή σχέσεων [21].

2.2 Είδη Πληροφορίας

Προτού ξεκινήσουμε να περιγράφουμε τη βιβλιογραφική ανασκόπηση γύρω από το θέμα της Παραγωγής Γράφου Σκηνής, πρέπει πρώτα να εξηγήσουμε τι είναι και από που προέρχεται η πληροφορία που χρησιμοποιούν οι περισσότερες (εάν όχι όλες) επιστημονικές έρευνες για την κατηγοριοποίηση των σχέσεων. Τα τρία βασικά είδη πληροφορίας είναι τα εξής:



Σχήμα 2.1: *Η αρχιτεκτονική του μοντέλου ανιχνευτή αντικειμένων Faster R-CNN [10] η οποία αποτελείται από 2 κύρια υποδίκτυα. Το πρώτο υποδίκτυο ονομάζεται δίκτυο προτεινόμενων περιοχών αντικειμένων (Region Proposal Network - RPN) το οποίο παίρνει σαν είσοδο μια εικόνα και βγάζει σαν έξοδο πιθανές περιοχές της εικόνας στις οποίες βρίσκονται αντικείμενα. Το δεύτερο υποδίκτυο είναι το Fast R-CNN [22] το οποίο λαμβάνει σαν είσοδο τις προτεινόμενες περιοχές αντικειμένων του προηγούμενου υποδικτύου και εξάγει τις τελικές προβλέψεις (Πηγή [10]).*



Σχήμα 2.2: *Το DETR χρησιμοποιεί ως ραχοκοκαλιά ένα συνελικτικό δίκτυο (CNN) για να μάθει μια 2D αναπαράσταση μιας εικόνας εισόδου. Το μοντέλο προσθέτει σε αυτή την αναπαράσταση μια κωδικοποίηση θέσης (positional encoding) προτού τη διοχετεύσει σε έναν κωδικοποιητή μετασχηματιστή (Transformer Encoder). Στη συνέχεια, ένας αποκωδικοποιητής μετασχηματιστή (Transformer Decoder) λαμβάνει ως είσοδο ένα σταθερό αριθμό μαθημένων χαρακτηριστικών θέσης (object queries) ενώ παράλληλα δίνει προσοχή στην έξοδο του κωδικοποιητή. Επομένως το δίκτυο είτε προβλέπει ένα αντικείμενο (κλάση και περίγραμμα) είτε μια κλάση που υποδεικνύει ότι δεν εντοπίσθηκε αντικείμενο.(Πηγή [20]).*

1. **Οπτικά Χαρακτηριστικά** (Visual features): Χρησιμοποιείται ένα προεκπαιδευμένο συνελικτικό δίκτυο (CNN) ως ραχοκοκαλιά (backbone) για την εξαγωγή χαρακτηριστικών αναπαράστασης οπτικής πληροφορίας (feature maps). Συνήθως γίνεται χρήση δικτύων όπως το ResNet [12] ή το VGG-16 [11]. Μέσω μιας προεκπαιδευμένης κεφαλής εντοπισμού αντικειμένων λαμβάνουμε τα χαρακτηριστικά που εκφράζουν τις οπτικές πληροφορίες για κάθε εντοπισμένο αντικείμενο.

2. **Χωρικά Χαρακτηριστικά** (Spatial features): Χρησιμοποιώντας τις συντεταγμένες των περιγραμμάτων των εντοπισμένων οντοτήτων, μπορούμε να εξάγουμε πληροφορίες που να αντιπροσωπεύουν την χωρική διάταξη μεταξύ όλων των ζευγών υποκειμένου - αντικειμένου που εμπεριέχονται σε μία εικόνα.

3. **Γλωσσικά Χαρακτηριστικά** (Linguistic features): Οι κατηγορίες του υποκειμένου και του αντικειμένου κωδικοποιούνται μέσω νευρωνικών γλωσσικών μοντέλων (π.χ. word2vec [23]) σε διανύσματα ενός σημασιολογικού χώρου όπου οι γεωμετρικές σχέσεις αντιστοιχούν σε νοηματικές.

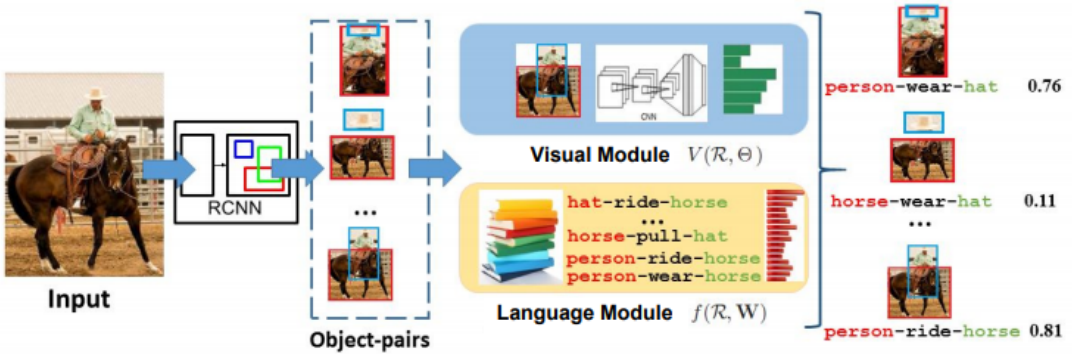
2.3 Γενική Βιβλιογραφία

2.3.1 Visual Relationship Detection with Language Priors [1]

Η πρώτη επιστημονική έρευνα κοντά στο πρόβλημα της αναγνώρισης οπτικών σχέσεων πραγματοποιήθηκε στην εργασία [24], όπου ορίστηκε το πρόβλημα της αναγνώρισης οπτικών φράσεων. Η προσέγγιση των συγγραφέων περιλάμβανε την εκπαίδευση ενός ξεχωριστού ταξινομητή για κάθε διαφορετική τριπλέτα που εμπεριέχεται στο σύνολο δεδομένων. Για παράδειγμα, ένας ταξινομητής εκπαιδευόταν για να αναγνωρίζει μόνο την τριπλέτα <άνθρωπος, κάθεται, στην καρέκλα> και ένας διαφορετικός ταξινομητής για την τριπλέτα <άνθρωπος, κάθεται, στον καναπέ> κ.ο.κ. Όπως είναι λογικό, αυτή η μέθοδος δεν μπορούσε να λειτουργήσει όταν έχουμε σύνολα δεδομένων μεγάλης έκτασης καθώς θα έπρεπε να εκπαιδευτούν N^2K διαφορετικοί ταξινομητές, όπου N είναι ο αριθμός των οντοτήτων και K ο αριθμός των κατηγορημάτων, ανεβάζοντας με αυτόν τον τρόπο σημαντικά την πολυπλοκότητα του δικτύου. Για παράδειγμα, στο σύνολο δεδομένων VG200 [9] εμπεριέχονται 150 κατηγορίες οντοτήτων και 50 κατηγορίες κατηγορημάτων. Έτσι θα έπρεπε να εκπαιδευτούν $150^2 \cdot 50 = 1.125.000$ διαφορετικοί ταξινομητές, πράγμα που είναι αδύνατο.

Η λύση λοιπόν στο παραπάνω πρόβλημα καθώς και η εισαγωγή ενός νέου συνόλου δεδομένων (VRD) έγινε στην εργασία [1], όπου εκπαιδεύονται ταξινομητές για την αναγνώριση των εντοπισμένων οντοτήτων αλλά και για την κατηγοριοποίηση των κατηγορημάτων, μειώνοντας με αυτόν την τρόπο την πολυπλοκότητα του προβλήματος σε $O(N + K)$ αντί του $O(N^2K)$ [24]. Πιο συγκεκριμένα, το προτεινόμενο δίκτυο μπορεί να χωριστεί σε 2 υπο-δίκτυα, το οπτικό και το γλωσσικό, συνδυάζοντας την οπτική αλλά και την γλωσσική πληροφορία που εμπεριέχεται μέσα σε μια εικόνα (Σχήμα 2.3).

Το οπτικό υπο-δίκτυο αποτελείται από 2 συνελικτικά δίκτυα (VGG [11]), ένα δίκτυο για την κατηγοριοποίηση των εντοπισμένων οντοτήτων (υποκείμενο και αντικείμενο) της εικόνας και ένα δίκτυο για την κατηγοριοποίηση των κατηγορημάτων. Ταυτόχρονα με το οπτικό υπο-

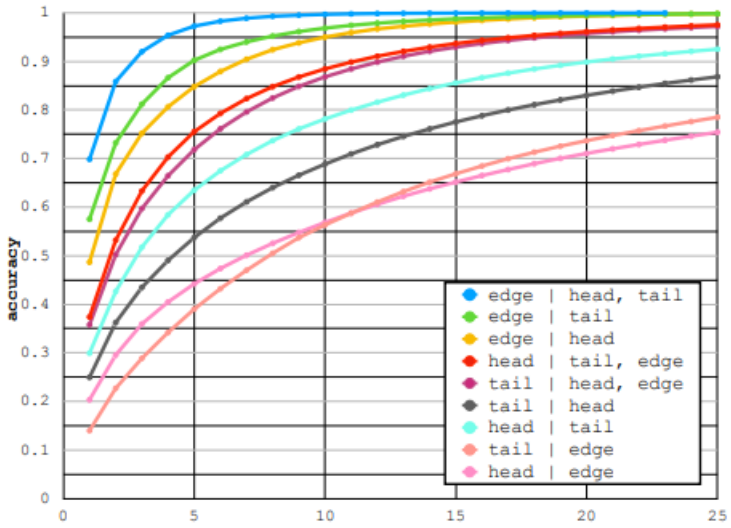


Σχήμα 2.3: *Επισκόπηση της αρχιτεκτονικής VRD [1]. Δεδομένης μιας εικόνας εισόδου, το RCNN [25] παράγει ένα σύνολο προτάσεων οντοτήτων. Κάθε ζεύγος οντοτήτων (υποκείμενο-αντικείμενο) στη συνέχεια βαθμολογείται με τη χρήση του οπτικούκαι του γλωσσικού υποδικτύου. Αυτές οι βαθμολογίες στη συνέχεια περνάνε από μια συνάρτηση κατωφλίου και παράγουν τις τελικές σχέσεις των οντοτήτων. (Πηγή [1]).*

δίκτυο, εκπαιδεύεται και το γλωσσικό υπο-δίκτυο το οποίο έχει δύο στόχους. Πρώτον, να κατηγοριοποιήσει τα κατηγορήματα χρησιμοποιώντας μόνο τα γλωσσικά χαρακτηριστικά των υποκειμένων και αντικειμένων και δεύτερον, να προβάλλει τις σχέσεις σε ένα διανυσματικό χώρο, όπου νοηματικά παρόμοιες σχέσεις θα βρίσκονται αρκετά κοντά μεταξύ τους. Τέλος, το αποτέλεσμα του συνολικού δικτύου θα είναι το γινόμενο του οπτικού και του γλωσσικού υποδικτύου.

2.3.2 Neural Motifs: Scene Graph Parsing with Global Context [2]

Μια διαφορετική και αρκετά ενδιαφέρουσα προσέγγιση γίνεται στην εργασία [2], όσον αφορά το πρόβλημα της παραγωγής γράφου σκηνής (SGG). Οι συγγραφείς έκαναν μια στατιστική ανάλυση στις εικόνες που εμπεριέχονται στο σύνολο δεδομένων εκπαίδευσης και έβγαλαν το συμπέρασμα πως ορισμένα αντικείμενα συνδέονται σε πολύ μεγάλο βαθμό με κάποια άλλα και πως εάν έχουμε δύο αντικείμενα τότε υπάρχει μεγάλη πιθανότητα να γνωρίζουμε την σχέση που υπάρχει μεταξύ τους (Σχήμα 2.4). Στη βάση αυτή προτείνουν μία νέα αρχιτεκτονική μοντέλου, που την ονομάζουν MotifsNet, η οποία εκτός του τοπικού πλαισίου (local context) μαθαίνει να εκμεταλλεύεται και το ολική πλαίσιο που μπορεί να υπάρχει μέσα στην εικόνα. Για την κατηγοριοποίηση, δηλαδή, της σχέσης μεταξύ 2 αντικειμένων θα χρησιμοποιηθεί πληροφορία και από τα υπόλοιπα αντικείμενα που περιέχονται μέσα στην φωτογραφία. Η δομή της αρχιτεκτονικής τους στηρίζεται σε μια σειρά από κελιά μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory - LSTM) ώστε να έχουν την δυνατότητα να εκμεταλλευτούν το ολικό πλαίσιο της εικόνας, ενώ ταυτόχρονα να εξακολουθούν να έχουν το τοπικό πλαίσιο. Αρχικά όλες οι οντότητες εξάγονται με την χρήση ενός ανιχνευτή αντικειμένων (Faster R-CNN [10]) και αμέσως μετά εξάγεται το πλαίσιο των αντικειμένων (object context) χρησιμοποιώντας LSTMs δύο κατευθύνσεων για να ληφθεί υπόψιν η πιθανή σχέση μεταξύ των αντικειμένων. Στην συνέχεια, χρησιμοποιούνται LSTMs μονής κατεύθυνσης για την εξαγωγή των ετικετών των εντοπισμένων αντικειμένων. Ακόμη, το πλαίσιο των σχέσεων (edge context) εξάγεται μέσω LSTMs δύο κατευθύνσεων με είσοδο την πληροφορία



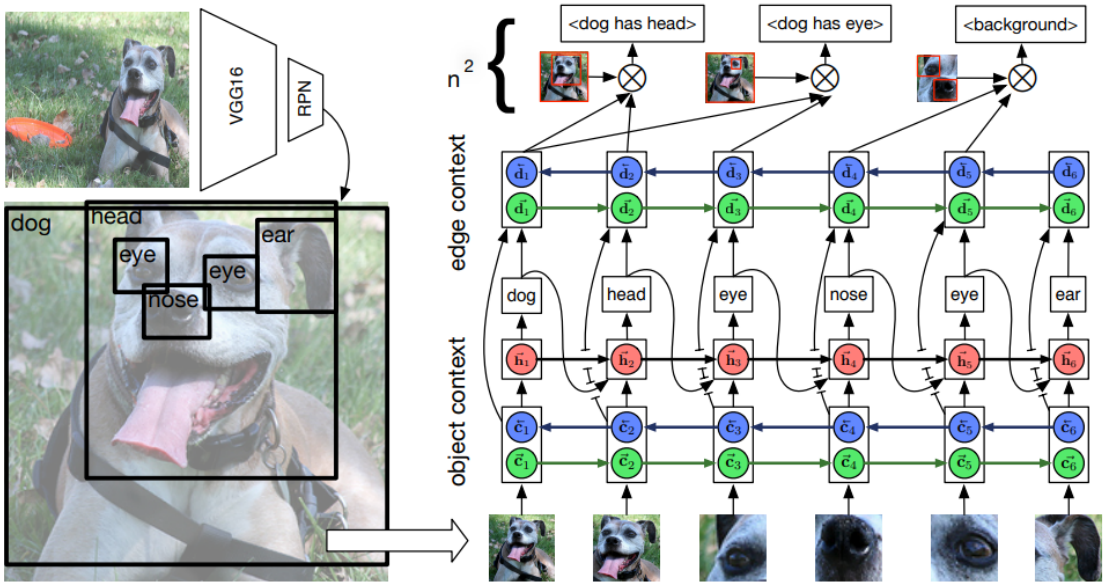
Σχήμα 2.4: *Στην εργασία [2] εξετάζεται πόσο σημαντική είναι η γνώση των ετικετών των υποκειμένων (head), κατηγορημάτων (edge) και των αντικειμένων (tail) στην παραγωγή γράφου σκηνής. Συγκεκριμένα, υπολογίζεται πόσες προσπάθειες προβλέψεων απαιτούνται για τον προσδιορισμό των ετικετών του υποκειμένου, του κατηγορήματος και του αντικειμένου δεδομένων των ετικετών των άλλων στοιχείων, χρησιμοποιώντας μόνο στατιστικά στοιχεία ετικετών. Υψηλότερες καμπύλες υποδηλώνουν ότι το στοιχείο είναι σε μεγάλο βαθμό καθορισμένο, δεδομένων των άλλων τιμών. Οι ετικέτες των κατηγορημάτων που εμπλέκονται σε μια σχέση δεν είναι ιδιαίτερα πληροφοριακές για τα υπόλοιπα στοιχεία της (υποκείμενο, αντικείμενο), ενώ οι ετικέτες του υποκειμένου ή του αντικειμένου παρέχουν σημαντικές πληροφορίες, τόσο μεταξύ τους όσο και στις ετικέτες των κατηγορημάτων. (Πηγή [2]).*

των αντικειμένων και τις ετικέτες του κάθε αντικειμένου και τέλος υπολογίζεται η πιθανότητα για την ετικέτα του της σχέσης του κάθε πιθανού ζευγαριού αντικειμένων. Η παραπάνω αρχιτεκτονική περιγράφεται και στο Σχήμα 2.5.

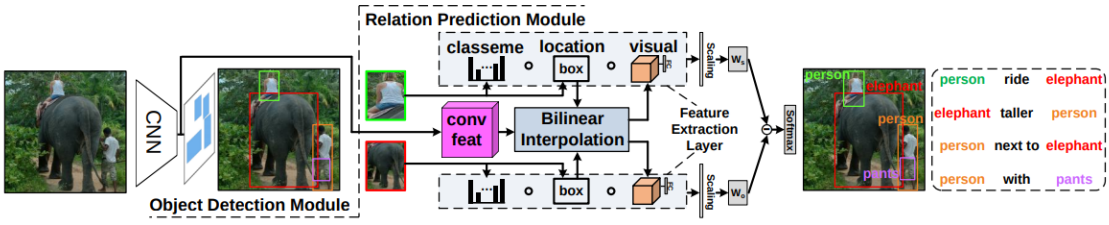
2.3.3 Visual Translation Embedding Network for Visual Relation Detection [3]

Σημαντική ήταν και η συνεισφορά της εργασίας [3] στο πρόβλημα της οπτικής αναγνώρισης σχέσεων, η οποία εμπνευσμένη από την εργασία [26], προβάλλει σε ένα διανυσματικό χώρο χαμηλών διαστάσεων τα υποκείμενα, κατηγορήματα και αντικείμενα και στην περίπτωση που το ζεύγος υποκείμενο - αντικείμενο αλληλεπιδρά, απαιτούν τα διανύσματα χαρακτηριστικών τους να ικανοποιούν την σχέση υποκείμενο + κατηγόρημα ≈ αντικείμενο. Έστω λοιπόν, **S, P, O** τα διανύσματα χαρακτηριστικών του υποκείμενου, κατηγορήματος και αντικειμένου αντίστοιχα στο διανυσματικό χώρο χαμηλών διαστάσεων τότε πρέπει **S** + **P** ≈ **O** ή αλλιώς **S** − **O** ≈ **P**. Για παράδειγμα, άνθρωπος + οδηγάει ≈ αμάξι.

Ομοίως με την εργασία [2], χρησιμοποιείται ένας ανιχνευτής οντοτήτων (Faster R-CNN [10]) με ραχοκοκαλιά το VGG-16 [11], από τον οποίο χρησιμοποιείται ο οπτικός χάρτης χαρακτηριστικών με στόχο να εξαχθούν τα οπτικά χαρακτηριστικά των εντοπισμένων οντοτήτων. Σαν τελικά χαρακτηριστικά (features) των οντοτήτων ορίζονται η συνένωση των οπτικών με τα γλωσσικά και τα χωρικά χαρακτηριστικά. Το Σχήμα 2.6 παρουσιάζει την εν λόγω αρχιτεκτονική πιο αναλυτικά [3].



Σχήμα 2.5: *Η αρχιτεκτονική του MotifNet. Το μοντέλο χωρίζει την ανάλυση του γράφου σκηνής σε στάδια πρόβλεψης των περιγραμμάτων, των ετικετών για κάθε περίγραμμα (δηλ. για κάθε εντοπισμένη οντότητα) και, στη συνέχεια, των ετικετών των σχέσεων. Μεταξύ κάθε σταδίου, το ολικό πλαίσιο υπολογίζεται με τη χρήση αμφίδρομων LSTMs και στη συνέχεια χρησιμοποιείται για τα επόμενα στάδια. (Πηγή [2]).*



Σχήμα 2.6: *Αρχιτεκτονική του VTransE [3]. Τα οπτικά, γλωσσικά και χωρικά χαρακτηριστικά των οντοτήτων συνενώνονται πριν προβληθούν στον διανυσματικό χώρο χαμηλών διαστάσεων. Το σύμβολο ∘ συμβολίζει τη συνένωση και το ⊖ την αφαίρεση κατά στοιχείο (element-wise subtraction). (Πηγή [3]).*

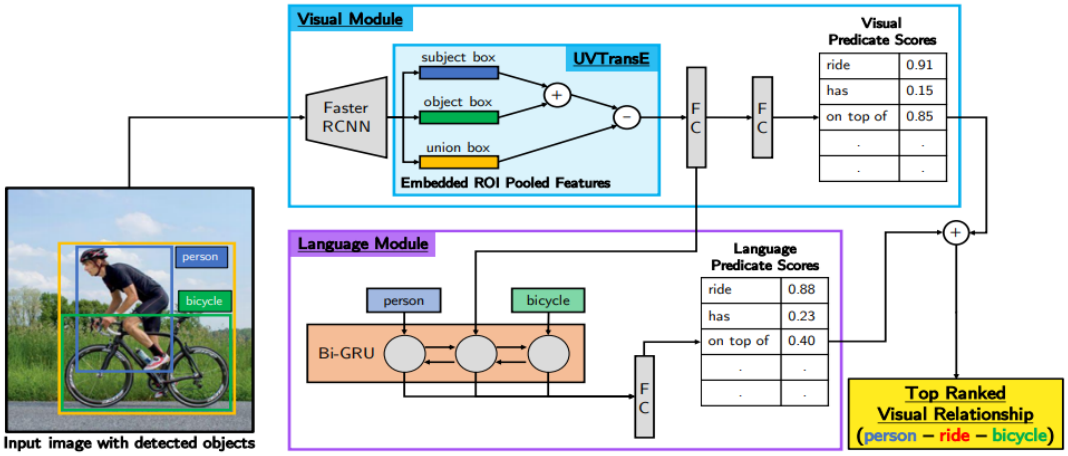
2.3.4 Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation [4]

Το UVTransE [4] αποτελεί προέκταση του VTransE μιας και το δεύτερο αποδίδει καλά σε τριπλέτες που έχει ξαναδεί, αλλά όχι τόσο καλά σε τριπλέτες που δεν έχει ξαναδεί. Αυτό οφείλεται στο γεγονός ότι το VTransE υπολογίζει το διάνυσμα του αντικειμένου μόνο από τα διανύσματα του υποκειμένου και του κατηγορήματος και επομένως το αντικείμενο **O** καθορίζεται ολοκληρωτικά από το υποκείμενο και το κατηγόρημα ως **S** + **P**. Για παράδειγμα, ας θεωρήσουμε μία όχι τόσο συχνή σχέση όπως <άνθρωπος, κάθεται στο, αμάξι>. Το γεγονός ότι είναι αρκετά σπάνια σχέση στο σύνολο δεδομένων εκπαίδευσης καθιστά αδύνατο στο VTransE να κατασκευάσει το αντικείμενο <αμάξι> με βάση το υποκείμενο <άνθρωπος> και το κατηγόρημα <κάθεται στο>.

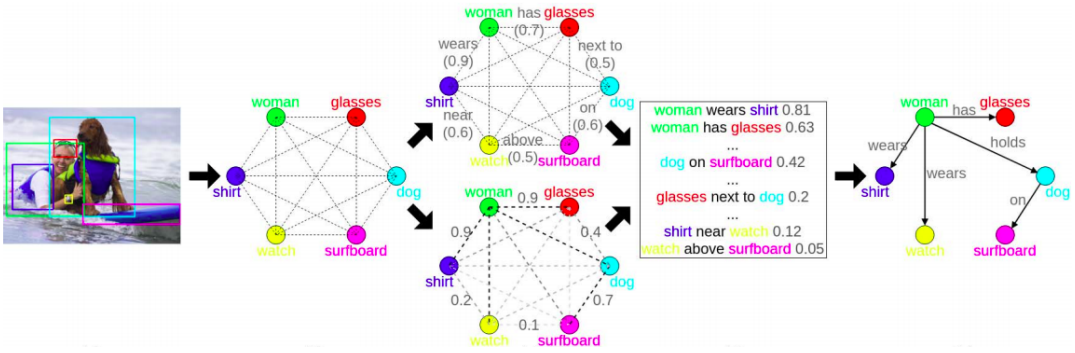
Η κύρια ιδέα του UVTransE, και η οποία προτείνεται ως λύση στο παραπάνω πρόβλημα, είναι πως αφαιρώντας τα διανύσματα οπτικών χαρακτηριστικών του υποκειμένου και του αντικειμένου από το διάνυσμα οπτικων χαρακτηριστικών της περιοχής της εικόνας η οποία διαμορφώνεται από την ένωση των περιγραμμάτων του ζεύγους υποκείμενο - αντικείμενο, τότε θα προκύψει ένα διάνυσμα το οποίο αντιστοιχεί στο κατηγόρημα τους. Επομένως, αν **S, P, O** και **U** τα διανύσματα οπτικών χαρακτηριστικών του υποκείμενου, κατηγορήματος, αντικειμένου και της ένωσης τους αντίστοιχα στο διανυσματικό χώρο χαμηλών διαστάσεων τότε θα πρέπει να είναι **U** − **(S** + **O)** ≈ **P**. Παράλληλα με το UVTransE, προτείνουν και ένα γλωσσικό δίκτυο το οποίο είναι ένα Bi-GRU (Bidirectional Gated Reccurrent Unit) και λαμβάνει σαν είσοδο τα γλωσσικά χαρακτηριστικά του υποκείμενου, αντικείμενου αλλά και το διάνυσμα χαρακτηριστικών του κατηγορήματος από το UVTransE και κατηγοριοποιεί το κατηγόρημα. Τέλος, η τελική πρόβλεψη του δικτύου για την ετικέτα του κατηγορήματος διαμορφώνεται από το άθροισμα των πιθανοτήτων του UVTransE και του γλωσσικού δικτύου. Η αρχιτεκτονική του UVTransE παρουσιάζεται και στο Σχήμα 2.7.

2.3.5 Attention-Translation-Relation Network for Scalable Scene Graph Generation [5]

Μία καινοτόμα λύση στο πρόβλημα της παραγωγής γράφου σκηνής εισάγεται στην εργασία [5], όπου προτείνεται το ATR-Net (Attention-Translation-Relation Network) το οποίο εμφανίζει πολύ καλά αποτελέσματα. Πιο συγκεκριμένα, κάθε εικόνα εισόδου τροφοδοτεί-ται σε ένα δίκτυο ανίχνευσης αντικειμένων (Faster-RCNN [10]) το οποίο εξάγει τις πιθανότητες κλάσεων για κάθε εντοπισμένη οντότητα $P(S)$ όταν η οντότητα αναφέρεται σε υποκείμενο και $P(O)$ όταν αναφέρεται σε αντικείμενο. Στη συνέχεια, υπολογίζει την πιθανότητα $P(P)$ του κατηγορήματος για κάθε πιθανό ζεύγος υποκειμένου-αντικειμένου, η οποία ισούται με $P(P) = P(P|related)P(related)$, όπου $P(P|related)$ είναι η πιθανότητα κατηγοριοποίησης της σχέσης μεταξύ ενός ζεύγους υποκειμένου-αντικειμένου δεδομένου πως αλληλεπιδρούν μεταξύ τους, και $P(related)$ η πιθανότητα αλληλεπίδρασης τους. Η τελική πρόβλεψη του δικτύου ATR-Net είναι ίση με $P(S, P, O) = P(S)P(P)P(O)$.



Σχήμα 2.7: *Επισκόπηση του μουτέλου ανίχνευσης οπτικών σχέσεων UVTransE. Δεδομένης μιας εικόνας, το Faster R-CNN [10] χρησιμοποιείται για την ανίχνευση οντοτήτων. Για κάθε ζεύγος εντοπισμένων οντοτήτων, εξάγονται οπτικά και χωρικά χαρακτηριστικά και τροφοδοτούνται στην οπτική μονάδα, η οποία υπολογίζει το διάνυσμα χαρακτηριστικών UVTransE: U − (S + O). Το διάνυσμα χαρακτηριστικών του κατηγορήματος που εξάγεται από το UVTransE μπορεί προαιρετικά να σταλεί σε ένα γλωσσικό μουτέλο Bi-GRU (γλωσσική μονάδα). Τέλος, αθροίζονται τα αποτελέσματα από την οπτική και την γλωσσική μονάδα και διαμορφώνονται έτσι οι τελικές προβλέψεις του δικτύου για τις ετικέτες των κατηγορημάτων. (Πηγή [3]).*



Σχήμα 2.8: *Το ATR-Net λύνει δύο ξεχωριστά προβλήματα: το πρώτο είναι η κατηγοριοποίηση την σχέσης μεταξύ υποκειμένου και αντικειμένου και το δεύτερο είναι η κατηγοριοποίηση της συσχέτισης των αντικειμένων δηλαδή αν αλληλεπιδρούν δύο οντότητες ή όχι. Η τελική έξοδος του δικτύου είναι το γινόμενο της πιθανότητας της κλάσης ενός ζεύγους οντοτήτων με την πιθανότητα αλληλεπίδρασής τους. (Πηγή [5]).*

2.4 Παρόμοιες Εργασίες - Αυτο-Επιβλεπόμενη Μάθηση

Στην παρούσα ενότητα, περιγράφουμε 2 βασικές βιβλιογραφικές έρευνες που έχουν γίνει, οι οποίες παρόλο που δεν έχουν σχέση με τον χώρο της ανίχνευσης οπτικών σχέσεων, αποτέλεσαν πηγές έμπνευσης για την ανάπτυξη της προτεινόμενης αρχιτεκτονικής καθώς κάνοντας χρήση αυτο-επιβλεπόμενης μάθησης καταφέρνουν να έχουν εξαιρετικές επιδόσεις.

2.4.1 BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding [6]

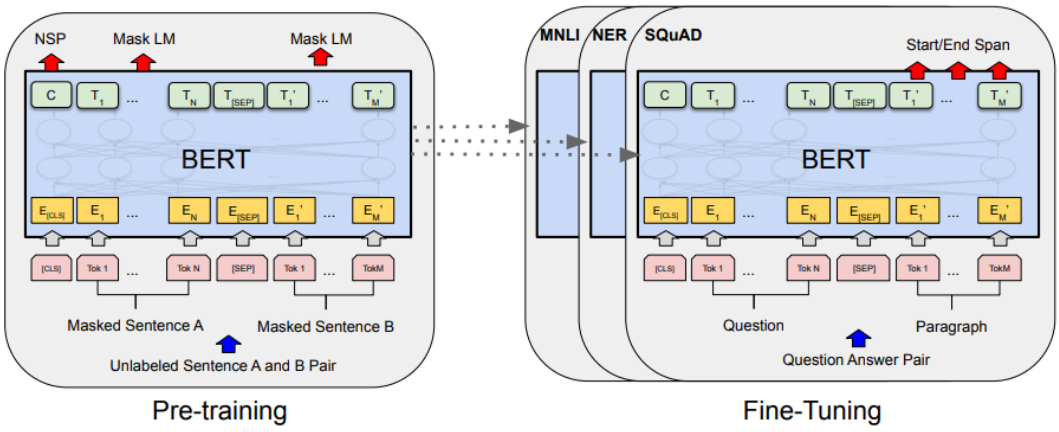
Το BERT (Bidirectional Encoder Representations from Transformers) [6] είναι μια επιστημονική έρευνα που δημοσιεύθηκε από ερευνητές της Google AI Language οποία παρουσίασε πολύ καλά αποτελέσματα σε μια ευρεία ποικιλία εργασιών επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP), όπως η Απάντηση Ερωτήσεων (SQuAD), η Συμπερασματολογία Φυσικής Γλώσσας (MNLI) και άλλες. Η βασική καινοτομία του BERT είναι η εφαρμογή της αμφίδρομης εκπαίδευσης του δικτύου μετασχηματιστών [8], ενός δημοφιλούς μοντέλου προσοχής, στη μοντελοποίηση γλώσσας. Αυτό έρχεται σε αντίθεση με προηγούμενες προσπάθειες που εξέταζαν μια ακολουθία κειμένου είτε από αριστερά προς τα δεξιά, είτε με συνδυασμένη εκπαίδευση από αριστερά προς τα δεξιά και από δεξιά προς τα αριστερά. Τα αποτελέσματα της επιστημονικής έρευνας δείχνουν ότι ένα γλωσσικό μοντέλο που εκπαιδεύεται αμφίδρομα μπορεί να έχει βαθύτερη κατανόηση του γλωσσικού πλαισίου και της ροής από τα γλωσσικά μοντέλα μίας κατεύθυνσης. Οι ερευνητές περιγράφουν λεπτομερώς μια νέα τεχνική με την ονομασία *μοντελοποίηση γλώσσας με κάλυψη* (Masked Language Modeling - MLM), η οποία επιτρέπει την αμφίδρομη εκπαίδευση σε μοντέλα στα οποία ήταν προηγουμένως αδύνατη.

Γενικότερα, κατά την εκπαίδευση γλωσσικών μοντέλων, υπάρχει η πρόκληση του καθορισμού ενός στόχου πρόβλεψης. Πολλά μοντέλα προβλέπουν την επόμενη λέξη σε μια ακολουθία, μια κατευθυνόμενη προσέγγιση που περιορίζει εγγενώς την εκμάθηση συμφραζομένων. Για να ξεπεραστεί αυτή η πρόκληση, το BERT χρησιμοποιεί δύο στρατηγικές εκπαίδευσης:

1. Πριν από την τροφοδοσία ακολουθιών λέξεων στο BERT, το 15% των λέξεων σε κάθε ακολουθία εισόδου αντικαθίσταται με ένα σύμβολο [MASK]. Στη συνέχεια, το μοντέλο προσπαθεί να προβλέψει την αρχική τιμή των κρυμμένων λέξεων, με βάση το πλαίσιο που παρέχουν οι άλλες, μη κρυμμένες, λέξεις στην ακολουθία (αυτο-επιβλεπόμενη μάθηση).

2. Στη διαδικασία εκπαίδευσης το μοντέλο λαμβάνει ζεύγη προτάσεων ως είσοδο και μαθαίνει να προβλέπει αν η δεύτερη πρόταση στο ζεύγος είναι η επόμενη πρόταση στο αρχικό έγγραφο. Κατά τη διάρκεια της εκπαίδευσης, το 50% των εισόδων είναι ένα ζεύγος στο οποίο η δεύτερη πρόταση είναι η επόμενη πρόταση του αρχικού έγγραφου, ενώ στο υπόλοιπο 50% επιλέγεται ως δεύτερη πρόταση μια τυχαία πρόταση από το σώμα κειμένων. Η υπόθεση είναι ότι η τυχαία πρόταση θα είναι αποσυνδεδεμένη από την πρώτη πρόταση.

Υπάρχουν δύο στάδια εκπαίδευσης (Σχήμα 2.9), η προ-εκπαίδευση (pre-training) και η



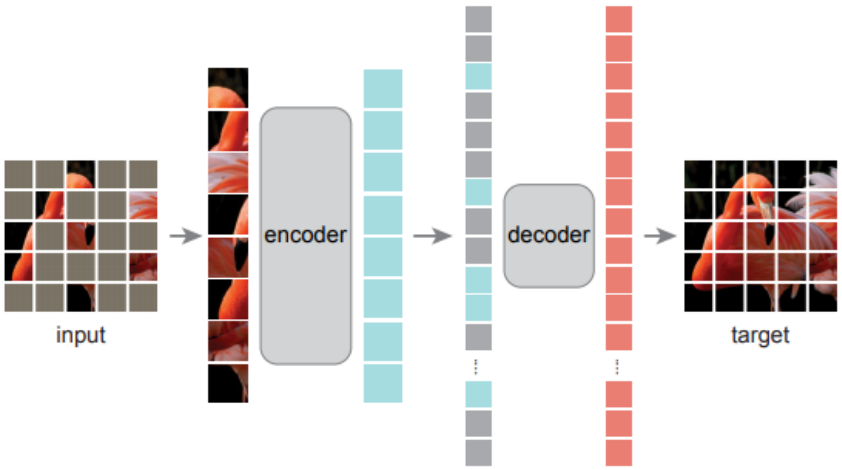
Σχήμα 2.9: *Διαδικασία προ-εκπαίδευσης και βελτιστοποίησης του BERT. Οι ίδιες αρχιτεκτονικές μουτέλων χρησιμοποιούνται τόσο στην προ-εκπαίδευση (pre-training) όσο και στην βελτιστοποίηση (fine-tuning). Οι ίδιες προ-εκπαιδευμένες παράμετροι χρησιμοποιούνται για την αρχικοποίηση μοντέλων για διαφορετικές εργασίες (downstream tasks). (Πηγή [6]).*

βελτιστοποίηση (fine-tuning). Κατά την προ-εκπαίδευση, το μοντέλο εκπαιδεύεται σε μη επισημειωμένα δεδομένα σε διάφορες εργασίες προ-εκπαίδευσης. Για την βελτιστοποίηση, το μοντέλο BERT αρχικοποιείται πρώτα με τις προ-εκπαιδευμένες παραμέτρους και όλες οι παράμετροι βελτιστοποιούνται, χρησιμοποιώντας επισημειωμένα δεδομένα από τις επιμέρους εργασίες (downstream tasks). Κάθε εργασία έχει ξεχωριστά finetuned μοντέλα, παρόλο που αρχικοποιούνται με τις ίδιες προ-εκπαιδευμένες παραμέτρους.

2.4.2 Masked Autoencoders Are Scalable Vision Learners [7]

Ένα σύστημα που κωδικοποιεί δεδομένα για να τα αποκωδικοποιήσει ξανά, όχι μόνο έχει το πλεονέκτημα της αυτοεπίβλεψης (self supervision), καθώς δεν χρειάζεται δεδομένα με ετικέτες, επειδή απλά παίρνει την είσοδο και την ανακατασκευάζει, αλλά μαθαίνει επίσης γενικότερες αναπαραστάσεις της συγκεκριμένης μορφής. Αυτή την ιδιότητα των αυτο-κωδικοποιητών (autoencoders) εκμεταλλεύονται στην εργασία [7], με την εισαγωγή των Masked AutoEncoders (MAE). Η βασική καινοτομία αυτής της αρχιτεκτονικής περιλαμβάνεται ήδη στον τίτλο και είναι η απόκρυψη (masking) της εικόνας. Η βασική ιδέα της ερευνητικής αυτής εργασίας είναι να αφαιρεθούν εικονοστοιχεία (pixels) από την εικόνα και επομένως να τροφοδοτηθεί το μοντέλο με μια ελλιπή εικόνα. Ο στόχος του μοντέλου είναι να μάθει πώς έμοιαζε η πλήρης, αρχική εικόνα. Οι συγγραφείς διαπίστωσαν ότι ένα αρκετά υψηλό ποσοστό απόκρυψης της εικόνας είναι πιο αποτελεσματικό καλύπτοντας το 75% της εικόνας.

Πιο συγκεκριμένα, όπως παρουσιάζεται και στο Σχήμα 2.10, αρχικά η εικόνα χωρίζεται σε τμήματα (patches), στα οποία αποδίδονται κωδικοποιήσεις θέσης (positional encoding). Στη συνέχεια, καλύπτεται τυχαία το 75% των κομματιών και τροφοδοτείται σαν είσοδο το υπόλοιπο 25% σε έναν μετασχηματιστή εικόνας (Vision Transformer [27]). Η έξοδος του μετασχηματιστή είναι μια διανυσματική αναπαράσταση χαμηλών διαστάσεων (latent representation) των τροφοδοτούμενων κομματιών της εικόνας εισόδου. Στη συνέχεια, εισάγονται τα επικαλυπτόμενα τμήματα, καθώς το επόμενο βήμα είναι ο αποκωδικοποιητής να ανα-



Σχήμα 2.10: *Κατά τη διάρκεια της προ-εκπαίδευσης, ένα μεγάλο τυχαίο υποσύνολο τμημάτων εικόνας (75%) επικαλύπτονται. Στην είσοδο του κωδικοποιητή εισέρχεται το μικρό υποσύνολο των ορατών τμημάτων της εικόνας (25%). Τα επικαλυπτόμενα τμήματα εισάγονται μετά τον κωδικοποιητή και το πλήρες σύνολο των κωδικοποιημένων και των επικαλυπτόμενων τμημάτων (patches) επεξεργάζεται από έναν αποκωδικοποιητή, που ανακατασκευάζει την αρχική εικόνα σε πιξελς. Μετά την προ-εκπαίδευση, ο αποκωδικοποιητής απορρίπτεται και ο κωδικοποιητής εφαρμόζεται σε ολόκληρες εικόνες για εργασίες αναγνώρισης. (Πηγή [7]).*

κατασκευάσει την αρχική εικόνα. Οι κωδικοποιήσεις θέσης εφαρμόζονται και πάλι ώστε ο αποκωδικοποιητής να αντιληφθεί πού βρίσκονται τα μεμονωμένα τμήματα στην αρχική εικόνα. Ο αποκωδικοποιητής λαμβάνει τη διανυσματική αναπαράσταση από τον κωδικοποιητή μαζί με τα επικαλυπτόμενα τμήματα ως είσοδο και εξάγει τις τιμές εικονοστοιχείων για κάθε ένα από αυτά, συμπεριλαμβανομένων και των επικαλυπτόμενων. Από αυτές τις πληροφορίες, η αρχική εικόνα μπορεί να ανακατασκευαστεί για να σχηματιστεί η προβλεπόμενη έκδοση της πλήρους εικόνας εισόδου. Αφού ανακατασκευαστεί η εικόνα-στόχος, μετράται η διαφορά της από την αρχική εικόνα εισόδου και χρησιμοποιείται ως κόστος (loss). Μετά την εκπαίδευση του μοντέλου, ο αποκωδικοποιητής απορρίπτεται και μόνο ο κωδικοποιητής, δηλαδή ο οπτικός μετασχηματιστής [27], διατηρείται για περαιτέρω χρήση και είναι πλέον ικανός να υπολογίζει διανυσματικές αναπαραστάσεις εικόνων για περαιτέρω επεξεργασία.

Κεφάλαιο 3

Δίκτυα Μετασχηματιστών

Καθώς η προτεινόμενη αρχιτεκτονική του οπτικού μετασχηματιστή (ViD) χρησιμοποιεί έναν αποκωδικοποιητή μετασχηματιστή (Transformer Decoder), σε αυτό το κεφάλαιο γίνεται μια ανάλυση των δικτύων μετασχηματιστών, αναλύοντας ξεχωριστά κάθε παράγοντα της αρχιτεκτονικής τους, επεξηγώντας παράλληλα και τον ρόλο και την επίδραση του στην συνολική απόδοση του δικτύου. Τέλος, αναλύουμε πως τα η συγκεκριμένη αρχιτεκτονική των μετασχηματιστών προσαρμόστηκε στον χώρο της εικόνας, περιγράφοντας την αντίστοιχη επιστημονική έρευνα.

3.1 Μηχανισμός Προσοχής (Attention Mechanism)

Η ανθρώπινη οπτική προσοχή μας επιτρέπει να εστιάζουμε σε μια συγκεκριμένη περιοχή με "ὑψηλή ανάλυση" (για παράδειγμα το μυτερό αυτί στο Σχήμα 3.1), ενώ αντιλαμβανόμαστε την περιβάλλουσα εικόνα με "χαμηλή ανάλυση" (για παράδειγμα τον δρόμο, το χορτάρι ή και την γάτα) και στη συνέχεια να προσαρμόζουμε το σημείο εστίασης ή να εξάγουμε συμπεράσματα με βάση αυτή την περιοχή. Δεδομένου ενός μικρού τμήματος μιας εικόνας, τα εικονοστοιχεία στα υπόλοιπα μέρη της εικόνας παρέχουν ενδείξεις για το τι πρέπει να εμφανίζεται εκεί. Περιμένουμε να δούμε ένα μυτερό αυτί στο λευκό πλαίσιο επειδή έχουμε δει τη μύτη ενός σκύλου, ένα άλλο μυτερό αυτί στα αριστερά. Ωστόσο, το χορτάρι και ο δρόμος δεν θα ήταν τόσο χρήσιμα όσο αυτά τα χαρακτηριστικά του σκύλου. Έτσι λοιπόν, ο μηχανισμός προσοχής στα νευρωνικά δίκτυα είναι μια τεχνική που επιτρέπει σε ένα μοντέλο να εστιάζει σε συγκεκριμένα τμήματα μιας εισόδου όταν κάνει προβλέψεις ή παίρνει αποφάσεις. Αυτό επιτρέπει στο μοντέλο να κάνει πιο ακριβείς προβλέψεις, αποδίδοντας διαφορετικά επίπεδα σπουδαιότητας ή "προσοχής" σε διαφορετικά μέρη της εισόδου. Με λίγα λόγια, η προσοχή (attention) στη βαθιά μάθηση μπορεί να ερμηνευτεί ως ένα διάνυσμα βαρών σημαντικότητας. Προκειμένου να προβλέψουμε ένα στοιχείο, όπως ένα εικονοστοιχείο σε μια εικόνα ή μια λέξη σε μια πρόταση, εκτιμούμε, χρησιμοποιώντας το διάνυσμα προσοχής, πόσο έντονα συσχετίζεται με τα υπόλοιπα στοιχεία και λαμβάνουμε το άθροισμα των τιμών τους, σταθμισμένο με το διάνυσμα προσοχής ως προσέγγιση του στοιχείου. Παρόλο που ο μηχανισμός αυτός χρησιμοποιείται πλέον σε διάφορα προβλήματα, σχεδιάστηκε αρχικά στο πλαίσιο της Νευρωνικής Μηχανικής Μετάφρασης με χρήση μοντέλων ακολουθίας σε ακολουθία (Seq2Seq).

Ένα μοντέλο Seq2Seq είναι ένα μοντέλο που λαμβάνει μια ακολουθία στοιχείων (λέξεις,

Σχήμα 3.1: *Εικόνα από το σύνολο δεδομένων VRD. (Πηγή [1]).*

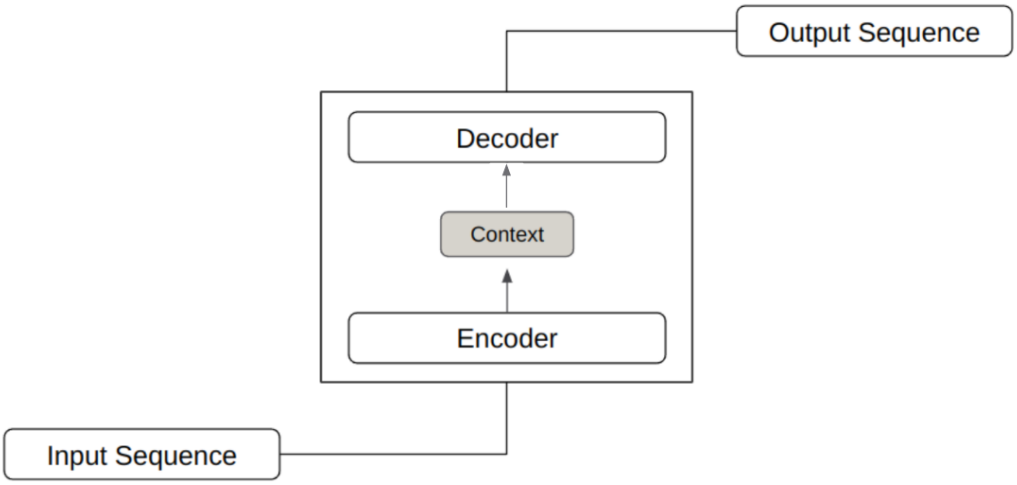
γράμματα, χρονοσειρές κ.λπ.) και εξάγει μια άλλη ακολουθία στοιχείων. Στην περίπτωση της νευρωνικής μηχανικής μετάφρασης, η είσοδος είναι μια σειρά λέξεων και η έξοδος είναι η μεταφρασμένη σειρά λέξεων. Ένα μοντέλο Seq2Seq αποτελείται από έναν κωδικοποιητή (κωδικοποιητής) και έναν αποκωδικοποιητή (αποκωδικοποιητής) (Σχήμα 3.2). Ο κωδικοποιητής συλλέγει το περιεχόμενο της ακολουθίας εισόδου με τη μορφή ενός κρυφού διανύσματος καταστάσεων (context vector) και το στέλνει στον αποκωδικοποιητή, ο οποίος στη συνέχεια παράγει την ακολουθία εξόδου. Δεδομένου ότι η εργασία του μοντέλου βασίζεται στην ακολουθία, τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής τείνουν να χρησιμοποιούν κάποια μορφή επαναληπτικών δικτύων όπως RNN (Recursive Neural Networks), LSTM, GRU κ.λπ. Το διάνυσμα κρυμμένης κατάστασης μπορεί να έχει οποιοδήποτε μέγεθος, αν και στις περισσότερες περιπτώσεις, λαμβάνεται ως δύναμη του 2 και είναι ένας μεγάλος αριθμός (256, 512, 1024), ο οποίος μπορεί κατά κάποιο τρόπο να αντιπροσωπεύει την πολυπλοκότητα της πλήρους ακολουθίας.

Ο μηχανισμός προσοχής εισήχθη στην εργασία [28] για να αντιμετωπίσει το πρόβλημα που προκύπτει με τη χρήση ενός σταθερού μήκους διανύσματος περιεχομένου, όπου ο αποκωδικοποιητής έχει περιορισμένη πρόσβαση στις πληροφορίες που παρέχονται από την είσοδο. Αυτό θεωρείται ότι καθίσταται ιδιαίτερα προβληματικό για μεγάλες ή/και πολύπλοκες ακολουθίες, όπου η διαστατικότητα της αναπαράστασής τους θα ήταν αναγκαστικά η ίδια με εκείνη για μικρότερες ή απλούστερες ακολουθίες. Ο προτεινόμενος μηχανισμός προσοχής [28] μπορεί να διαιρεθεί σε 3 επιμέρους στάδια υπολογισμού:

1. **Υπολογισμός των Βαθμολογιών Ευθυγράμμισης** (Alignment scores): Το μοντέλο ευθυγράμμισης (alignment model) λαμβάνει τις κωδικοποιημένες κρυφές καταστάσεις h_i (encoder hidden states) και την προηγούμενη έξοδο του αποκωδικοποιητή, S_{t-1}, για να υπολογίσει μια βαθμολογία, $e_{t,i}$, που δείχνει πόσο καλά τα στοιχεία της ακολουθίας εισόδου ευθυγραμμίζονται με την τρέχουσα έξοδο στη θέση t. Το μοντέλο ευθυγράμμισης αναπαρίσταται από μια συνάρτηση, $a(\cdot)$, η οποία μπορεί να υλοποιηθεί από ένα νευρωνικό δίκτυο:

$$e_{t,i} = a(S_{t-1}, h_i) \tag{3.1}$$

2. **Υπολογισμός Βαρών Προσοχής** (attention weights): Τα βάρη υπολογίζονται με την



Σχήμα 3.2: *Αρχιτεκτονική ενός μοντέλου Seq2Seq σε υψηλό επίπεδο.*

εφαρμογή της συνάρτησης Softmax στις βαθμολογίες ευθυγράμμισης που υπολογίστηκαν παραπάνω:

$$a_{t,i} = Softmax(e_{t,i}) \tag{3.2}$$

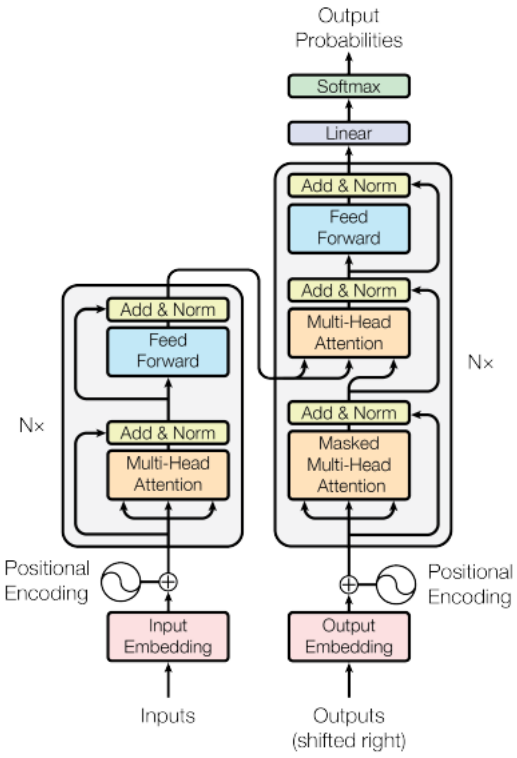
3. **Υπολογισμός διανύσματος περιεχομένου** (context vector): Το διάνυσμα περιεχομένου τροφοδοτείται στον αποκωδικοποιητή, το οποίο υπολογίζεται από ένα σταθμισμένο άθροισμα όλων των T κρυφών καταστάσεων του κωδικοποιητή (encoder hidden states):

$$c_t = \sum_{n=1}^{T} a_{t,i} \cdot h_i \tag{3.3}$$

3.2 Δίκτυα Μετασχηματιστών (Transformers)

Ο παραπάνω μηχανισμός υλοποιήθηκε με δίκτυα RNNs τόσο για τον κωδικοποιητή (encoder) όσο και για τον αποκωδικοποιητή (decoder) [28]. Ωστόσο, τα RNNs, δυσκολεύονται να συλλάβουν εξαρτήσεις μεγάλης εμβέλειας επειδή επεξεργάζονται την είσοδο διαδοχικά. Αυτό σημαίνει ότι καθώς η είσοδος γίνεται μεγαλύτερη, οι πληροφορίες από τα πρώτα τμήματα της εισόδου μπορεί να χαθούν ή να καταστούν λιγότερο σημαντικές μέχρι τη στιγμή που το μοντέλο φτάσει στο τέλος της. Σε αυτό το πρόβλημα δίνουν λύση οι μετασχηματιστές [8] (transformers).

Οι μετασχηματιστές [8] είναι ένας τύπος αρχιτεκτονικής νευρωνικών δικτύων που έγινε πρόσφατα δημοφιλής λόγω της ικανότητάς του να αποδίδει καλά σε ένα ευρύ φάσμα εργασιών επεξεργασίας φυσικής γλώσσας (Σχήμα 3.3). Ένας από τους κύριους λόγους για τους οποίους οι μετασχηματιστές θεωρούνται σημαντικότεροι από τα αναδρομικά νευρωνικά δίκτυα είναι η ικανότητά τους να συλλαμβάνουν αποτελεσματικά εξαρτήσεις και σχέσεις



Σχήμα 3.3: *Αρχιτεκτονική του δικτύου Μετασχηματιστή. (Πηγή [8]).*

μεγάλης εμβέλειας μεταξύ διαφορετικών τμημάτων της εισόδου. Οι μετασχηματιστές χρησιμοποιούν μηχανισμούς αυτοπροσοχής (self-attention) για να παρακολουθούν διαφορετικά μέρη της εισόδου, ανεξάρτητα από τη θέση τους στην ακολουθία εισόδου. Αυτό επιτρέπει στο μοντέλο να καταγράφει αποτελεσματικά εξαρτήσεις και σχέσεις μεγάλης εμβέλειας μεταξύ διαφορετικών τμημάτων της εισόδου, ακόμη και όταν η είσοδος είναι πολύ μεγάλη. Ένας ακόμη αρκετά σημαντικός λόγος για τον οποίο οι μετασχηματιστές προτιμώνται περισσότερο από τα αναδρομικά νευρωνικά δίκτυα, είναι η ικανότητά τους να επεξεργάζονται την ακολουθία εισόδου παράλληλα. Στα RNNs η πληροφορία ρέει με διαδοχικό τρόπο, οπότε κατά την εκπαίδευση και την εξαγωγή συμπερασμάτων η είσοδος πρέπει να επεξεργάζεται κατά ένα βήμα τη φορά. Στους μετασχηματιστές, ο μηχανισμός αυτοπροσοχής (self-attention) επιτρέπει στο μοντέλο να παρακολουθεί ταυτόχρονα διαφορετικά μέρη της εισόδου, γεγονός που καθιστά δυνατή την παραλληλία των υπολογισμών. Αυτό επιτρέπει στο μοντέλο να εκπαιδευτεί πολύ πιο γρήγορα από τα RNN, ειδικά όταν η είσοδος είναι μεγάλη.

Η βασική αρχιτεκτονική ενός μετασχηματιστή αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή. Ο πρώτος δέχεται σαν είσοδο την ακολουθία εισόδου υπό μορφή συμβόλων (tokens). Κάθε σύμβολο αποτελεί ένα γλωσσικό διάνυσμα (word embedding) (εάν αναφερόμαστε σε πρόβλημα επεξεργασίας φυσικής γλώσσας) τα οποία στην συνέχεια κωδικοποιούνται κατάλληλα ώστε το δίκτυο να γνωρίζει σε ποια θέση της πρότασης ανήκει το κάθε ένα (περισσότερες πληροφορίες στην Ενότητα 3.2.1). Έχοντας λοιπόν την ακολουθία εισόδου με επίγνωση θέσης, την διοχετεύουμε στην συνέχεια στο δίκτυο του κωδικοποιητή το οποίο αποτελείται από 4 βασικά στοιχεία:

1. **Στρώμα προσοχής πολλαπλών κεφαλών** (Multi-head attention layer), το οποίο πε-

ριγράφεται στην Ενότητα 3.2.2.

2. **Γραμμικά στρώματα** (Linear Layers).

3. **Υπολειπόμενες συνδέσεις** (residual connections): Οι λόγοι ύπαρξης των υπολειπόμενων συνδέσεων είναι κυρίως δύο, για την διατήρηση της γνώσης κατά την διάρκεια της εκπαίδευσης αλλά για την καταπολέμηση του προβλήματος των εξαφανιζούμενων κλίσεων (vanishing gradient problem).

4. **Στρώματα πρόσθεσης και κανονικοποίησης** (Add & Norm layers): Αρχικά προστίθεται η είσοδος του δικτύου με την έξοδο του στρώματος προσοχής πολλαπλών κεφαλών και στην συνέχεια γίνεται κανονικοποίηση του αθροίσματος αυτού. Η κανονικοποίηση γίνεται με την διαίρεση της κάθε τιμής του προηγούμενου αθροίσματος με τη μέση τιμή του άξονα των διανυσμάτων και διαίρεση με την τυπική απόκλιση

Επομένως με αυτόν τρόπο παίρνουμε την έξοδο του κωδικοποιητή την οποία παρέχουμε σαν είσοδο στο δίκτυο του αποκωδικοποιητή, μαζί με την ακολουθία στόχο (target sequence) ώστε να κάνουμε εκπαίδευση του δικτύου μας. Η αρχιτεκτονική του αποκωδικοποιητή είναι αρκετά "κοντά" με αυτή του κωδικοποιητή, με την διαφορά ότι περιέχει ένα επιπλέον στρώμα προσοχής πολλαπλών κεφαλών το οποίο ονομάζεται Masked Multi-head Attention. Η ανάλυση της αρχιτεκτονικής του αποκωδικοποιητή γίνεται στην Ενότητα 3.2.3.

Παρακάτω παρουσιάζονται αναλυτικά τα στοιχεία της αρχιτεκτονικής του δικτύου των μετασχηματιστών και η λειτουργία τους.

3.2.1 Κωδικοποίηση Θέσης (Positional Encoding)

Μπορούμε να καταλάβουμε την αναγκαιότητα των διανυσμάτων θέσης (positional embeddings) αν σκεφτούμε το εξής: αν ένα δίκτυο LSTM αναλάμβανε την ακολουθία εισόδου, θα το έκανε διαδοχικά, ένα στοιχείο της ακολουθίας τη φορά, γι' αυτό και είναι τόσο αργό δίκτυο. Υπάρχει όμως και μια θετική πλευρά σε αυτό, αφού τα LSTMs παίρνουν τα διανύσματα διαδοχικά με την καθορισμένη σειρά τους, γνωρίζουν ποια λέξη ήρθε πρώτη, ποια λέξη ήρθε δεύτερη κ.ο.κ. Από την άλλη πλευρά, οι μετασχηματιστές λαμβάνουν όλα τα στοιχεία της ακολουθίας εισόδου ταυτόχρονα. Παρόλο που αυτό είναι ένα τεράστιο πλεονέκτημα και κάνει τους μετασχηματιστές πολύ πιο γρήγορους, το μειονέκτημα είναι ότι χάνουν τις κρίσιμες πληροφορίες που σχετίζονται με τη σειρά των λέξεων. Με απλά λόγια, δεν γνωρίζουν ποια λέξη ήρθε πρώτη στην ακολουθία και ποια τελευταία. Επομένως η χρήση της πληροφορίας θέσης είναι αρκετά σημαντική. Μπορούμε να σκεφτούμε για παράδειγμα την εξής πρόταση "Παρόλο που δεν κέρδισε το βραβείο, ήταν ικανοποιημένη" και μία διαφορετική πρόταση η οποία περιέχει τις ίδιες λέξεις αλλά με ελαφρώς διαφορετική σειρά: "Παρόλο που κέρδισε το βραβείο, δεν ήταν ικανοποιημένη". Παρατηρούμε πώς η θέση μιας και μόνο λέξης αλλάζει όχι μόνο το συναίσθημα αλλά και το νόημα αυτής της πρότασης.

Αυτό που μπορούμε να κάνουμε λοιπόν για να επαναφέρουμε την πληροφορία της σειράς των λέξεων στους μετασχηματιστές χωρίς να τους έχουμε επαναλαμβανόμενους όπως τα δίκτυα LSTMs είναι να εισάγουμε ένα νέο σύνολο διανυσμάτων που περιέχουν την πληροφορία της θέσης, τα διανύσματα θέσης. Προσθέτοντας τα γλωσσικά διανύσματα στα αντίστοιχα

διανύσματα θέσης δημιουργούμε νέα γλωσσικά διανύσματα με επίγνωση της σειράς εμφάνισης τους στην πρόταση. Μία αρχική σκέψη για το τι θα μπορούσαν να περιέχουν τα διανύσματα θέσης είναι οι αριθμοί θέσεων των λέξεων. Έτσι λοιπόν, το διάνυσμα θέσης που αντιστοιχεί στο πρώτο στοιχείο της ακολουθίας εισόδου θα περιέχει μόνο μηδενικά, το διάνυσμα θέσης που αντιστοιχεί στο δεύτερο στοιχείο της ακολουθίας εισόδου μόνο μονάδες κ.ο.κ. Ωστόσο η προσθήκη της πληροφορίας θέσης με αυτόν τον τρόπο μπορεί να παραμορφώσει τα γλωσσικά διανύσματα, ιδίως εκείνα των θέσεων που εμφανίζονται αργότερα στο κείμενο. Για παράδειγμα, αν το κείμενο έχει 30 λέξεις το τελευταίο γλωσσικό διάνυσμα θα προστεθεί στον τεράστιο αριθμό 30. Μια επόμενη σκέψη θα μπορούσε να ήταν να προσθέταμε κλάσματα. Επομένως, αν ένα κείμενο αποτελείται από τέσσερις λέξεις τα διανύσματα θέσης μπορούν απλά να αναπαριστούν τη θέση της λέξης ως κλάσματα του συνολικού μήκους. με αποτέλεσμα η μέγιστη τιμή του διανύσματος θέσης να μην ξεπερνά ποτέ την μονάδα. Ωστόσο και αυτή η τεχνική έχει περιορισμένα αποτελέσματα, διότι το να μετατρέψουμε τα διανύσματα θέσης συναρτήσει του συνολικού μήκους του κειμένου θα σήμαινε ότι αν οι προτάσεις διαφέρουν σε μήκος, πράγμα που συμβαίνει συχνά, θα διέθεταν και διαφορετικά διανύσματα θέσης για την ίδια θέση. Αυτό μπορεί να μπερδέψει το μοντέλο μας και ούτε αυτό το θέλουμε. Ιδανικά οι τιμές των διανυσμάτων θέσης σε μια δεδομένη θέση θα πρέπει να παραμένουν ίδιες, ανεξάρτητα από το συνολικό μήκος του κειμένου ή οποιονδήποτε άλλο παράγοντα.

Επομένως χρειαζόμαστε μια συνάρτηση, η οποία να παίρνει τιμές σε ένα συγκεκριμένο εύρος τιμών αλλά και να μην εξαρτάται από το μέγεθος της ακολουθίας. Μια τέτοια συνάρτηση αποτελεί το ημίτονο και συνημίτονο, οι οποίες περιοδικά επιστρέφουν τιμές μεταξύ του -1 και του 1 αλλά και ταυτόχρονα έχουν το θετικό πως ορίζονται μέχρι το άπειρο, οπότε ακόμα και με τεράστιες εκτάσεις ακολουθιών θα είχαμε τιμές μεταξύ -1 και 1 στο διάνυσμα θέσης. Θα μπορούσε κανείς να σκεφτεί και την σιγμοειδή συνάρτηση (sigmoid) η οποία επίσης έχει τιμές περιορισμένες, παρόλα αυτά όμως προτιμάμε τις συναρτήσεις των ημίτονων και συνημίτονων διότι δίνουν μεγάλη μεταβλητότητα στις τιμές ακόμα και σε τεράστους αριθμούς. Έστω λοιπόν ότι διαλέγουμε την συνάρτηση του ημίτονου με περίοδο p για να μοντελοποιήσουμε την κωδικοποίηση θέσης. Μπορούμε εύκολα να παρατηρήσουμε πως με μία μόνο περιοδική συνάρτηση, το ίδιο αποτέλεσμα θα επαναλαμβανόταν για διαφορετικές θέσεις (για παράδειγμα στο πρώτο στοιχείο της ακολουθίας εισόδου και στο p-οστό στοιχείο της ακολουθίας). Εάν όμως χρησιμοποιήσουμε την συνάρτηση του ημιτόνου για μια διάσταση του διανύσματος θέσης και την συνάρτηση του συνημιτόνου αλλά με διαφορετικά συχνότητα για μια άλλη διαφορετική διάσταση, παρατηρούμε ότι οι τιμές είναι μοναδικές και διαφοροποιούν το ένα στοιχείο της ακολουθίας εισόδου από το άλλο. Αν λοιπόν το κάνουμε αυτό για κάθε διάσταση, δηλαδή εναλλάσσουμε τις συναρτήσεις των ημίτονων και των συνημίτονων με αυξανόμενη συχνότητα, δίνουμε αρκετές πληροφορίες ώστε να διασφαλίσουμε ότι ο μετασχηματιστής δεν μπορεί να χάσει αυτή τη σειρά της ακολουθίας εισόδου.

Έτσι λοιπόν στην εργασία [8] χρησιμοποιούνται συχνότητες κυμάτων για να καταγράψουν τις πληροφορίες θέσης. Ας υποθέσουμε ότι έχουμε μια ακολουθία εισόδου μήκους L και χρειαζόμαστε τη θέση του pos^{th} αντικειμένου μέσα σε αυτή. Η κωδικοποίηση θέσης δίνεται από ημιτονοειδείς και συνημιτονοειδείς συναρτήσεις διαφορετικών συχνοτήτων:

$$PE(pos, 2i) = \sin\Big(\frac{pos}{10000^{2i/d_{model}}}\Big) \tag{3.4}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.5}$$

όπου pos είναι η θέση του στοιχείου στην ακολουθία εισόδου με εύρος τιμών $0 \leq pos < L$, d_{model} είναι οι διαστάσεις του διανύσματος της ακολουθίας εισόδου / εξόδου (άρα και του διανύσματος θέσης), $PE(pos, j)$ είναι η συνάρτηση για την αντιστοίχηση μιας θέσης pos της ακολουθίας εισόδου στο δείκτη (pos, j) του διανύσματος θέσης, και i διάσταση του διανύσματος με $0 \leq i < \frac{d_{model}}{2}$. Ένα αναλυτικό παράδειγμα του υπολογισμού του διανύσματος θέσης παρουσιάζεται στον Πίνακα 3.1, όπου υπολογίζεται το διάνυσμα θέσης των γλωσσικών διανυσμάτων της φράσης "I just graduated", με διάσταση $d_{model} = 4$ για λόγους απλότητας.

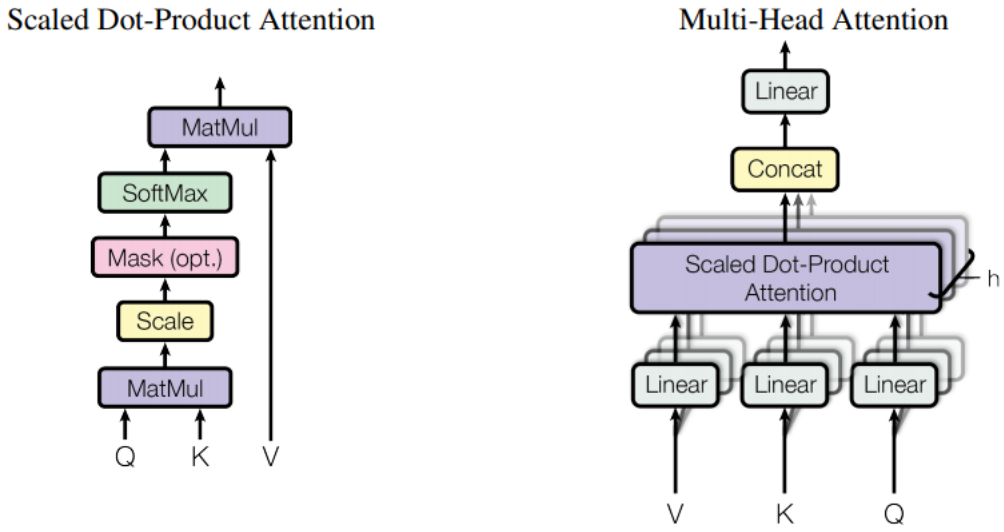
Input sequence	Index of token	Positional Encoding			
		i = 0	i = 0	i = 1	i = 1
I	0	$PE_{0,0} = \sin\left(\frac{0}{10000^{2\cdot 0/4}}\right)$	$PE_{0,1} = \cos\left(\frac{0}{10000^{2\cdot 0/4}}\right)$	$PE_{0,2} = \sin\left(\frac{0}{10000^{2\cdot 1/4}}\right)$	$PE_{0,3} = \cos\left(\frac{0}{10000^{2\cdot 1/4}}\right)$
just	1	$PE_{1,0} = \sin\left(\frac{1}{10000^{2\cdot 0/4}}\right)$	$PE_{1,1} = \cos\left(\frac{1}{10000^{2\cdot 0/4}}\right)$	$PE_{1,2} = \sin\left(\frac{1}{10000^{2\cdot 1/4}}\right)$	$PE_{1,3} = \cos\left(\frac{1}{10000^{2\cdot 1/4}}\right)$
graduated	2	$PE_{2,0} = \sin\left(\frac{2}{10000^{2\cdot 0/4}}\right)$	$PE_{2,1} = \cos\left(\frac{2}{10000^{2\cdot 0/4}}\right)$	$PE_{2,2} = \sin\left(\frac{2}{10000^{2\cdot 1/4}}\right)$	$PE_{2,3} = \cos\left(\frac{2}{10000^{2\cdot 1/4}}\right)$

Πίνακας 3.1: *Υπολογισμός του διανύσματος θέσης της ακολουθίας "I just graduated". Για λόγους απλότητας και ευκολίας στην κατανόηση θεωρούμε πως η διάσταση των διανυσμάτων ισούται με $d_{model} = 4$. Αξίζει να αναφέρουμε ότι στην εργασία [8], όπου παρουσιάστηκε ο συγκεκριμένος μηχανισμό, τα διανύσματα είχαν $d_{model} = 512$.*

3.2.2 Αυτο-προσοχή και Προσοχή πολλαπλών κεφαλών (Multihead & Self-attention)

Ο μηχανισμός προσοχής βοηθά το μοντέλο να εστιάζει σε σημαντικές λέξεις σε μια δεδομένη ακολουθία εισόδου. Οι μετασχηματιστές δεν χρησιμοποιούν τον μηχανισμό προσοχής που αναφέρθηκε παραπάνω [28], αλλά εισάγουν έναν νέο μηχανισμό προσοχής που ονομάζεται αυτο-προσοχή (self-attention). Η κύρια διαφορά μεταξύ της απλής προσοχής [28] και της αυτο-προσοχής είναι ότι η απλή προσοχή εστιάζει επιλεκτικά στις λέξεις σε σχέση με κάποιο εξωτερικό ερώτημα. Όσο πιο σημαντική είναι μια λέξη για τον προσδιορισμό της απάντησης σε αυτό το ερώτημα, τόσο μεγαλύτερη εστίαση της δίνεται. Η αυτοπροσοχή, από την άλλη πλευρά, λαμβάνει υπόψη τη σχέση μεταξύ των λέξεων εντός της ίδιας πρότασης. Στην Σχήμα 3.4 φαίνονται όλοι οι απαραίτητοι μηχανισμοί για τον υπολογισμό της προσοχής.

Πιο συγκεκριμένα, το πρώτο στοιχείο στην αρχιτεκτονική της προσοχής πολλών-κεφαλών (multi-head attention) είναι τρία γραμμικά στρώματα. Κάθε ένα από αυτά τα στρώματα έχει μια ειδική λειτουργία, τα οποία ονομάζουμε Queries, Keys και Values. Ας κατανοήσουμε αυτούς τους όρους με ένα παράδειγμα. Αν λοιπόν θέλαμε να αναζητήσουμε κάτι στο Youtube ή στο Google, το κείμενο που πληκτρολογούμε στο πλαίσιο αναζήτησης είναι το Query. Τα αποτελέσματα που εμφανίζονται ως τίτλος βίντεο ή άρθρου είναι τα Keys και το περιεχόμενο μέσα σε αυτά είναι τα Values. Για να βρούμε τις καλύτερες αντιστοιχίες, πρέπει να βρούμε την ομοιότητα μεταξύ του Query και των αντίστοιχων Keys. Μόλις βρεθεί το πιο παρόμοιο



Σχήμα 3.4: *Αρχιτεκτονική του δικτύου προσοχής πολλών-κεφαλών. (Πηγή [8]).*

Key, επιστρέφει το βίντεο ή το άρθρο που συνδέεται με αυτό το Key, δηλαδή το Value. Η ομοιότητα αυτή, μπορεί να θεωρηθεί ως ένα υποκατάστατο της προσοχής. Αυτό συμβαίνει επειδή το μοντέλο επιστρέφει το καλύτερο βίντεο/άρθρο μόνο δίνοντας προσοχή στον πιο παρόμοιο τίτλο βίντεο/άρθρου σε σύγκριση με το Query. Ένας καλός τρόπος υπολογισμού της ομοιότητας μεταξύ 2 διανυσμάτων είναι η ομοιότητα συνημίτονου, που κυμαίνεται σε ένα εύρος από -1 έως +1, όπου +1 (μέγιστη ομοιότητα) σημαίνει ότι 2 διανύσματα δείχνουν προς την ίδια ακριβώς κατεύθυνση επειδή το συνημίτονο της γωνίας μεταξύ τους είναι 1 ($\cos(0) = 1$) και -1 (μέγιστη ανομοιότητα) σημαίνει ότι δείχνουν προς την αντίθετη κατεύθυνση καθώς το συνημίτονο της γωνίας μεταξύ τους είναι -1 ($\cos(\pi) = -1$). Το μέτρο της ομοιότητας συνημιτόνου μεταξύ 2 διανυσμάτων υπολογίζεται ως εξής:

$$CosSim(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{3.6}$$

Μπορούμε να ξαναγράψουμε την Εξίσωση 3.6 αλλά με την διαφορά ότι θα αναφέρεται σε πίνακες και όχι σε διανύσματα:

$$CosSim(A, B) = \frac{A \cdot B^T}{scaling} \tag{3.7}$$

Επομένως εφόσον ψάχνουμε την ομοιότητα των Queries και των Keys, η Εξίσωση 3.7 ξαναγράφεται ως εξής:

$$Similarity(Q, K) = \frac{Q \cdot K^T}{scaling} \tag{3.8}$$

Στα τρία αυτά λοιπόν γραμμικά στρώματα τροφοδοτούμε τα γλωσσικά διανύσματα της ακολουθίας εισόδου (αφου πρώτα έχουμε προσθέσει τα διανύσματα θέσης) και κάθε γραμμικό στρώμα εξάγει 3 πίνακες, τα queries, keys και values. Επομένως, υπολογίζουμε το γινόμενο μεταξύ του query και του ανάστροφου πίνακα key. Η έξοδος αυτού του γινομένου μπορεί να ονομαστεί φίλτρο προσοχής (attention filter). Στην αρχή της διαδικασίας της εκ-

παίδευσης τα περιεχόμενα του φίλτρου προσοχής είναι τυχαίοι αριθμοί, αλλά μόλις αυτή ολοκληρωθεί παίρνουν αρκετά πιο σημαντικές τιμές. Οι βαθμολογίες στο εσωτερικό αυτού του πίνακα είναι στην πραγματικότητα βαθμολογίες αυτο-προσοχής (self-attention scores). Για παράδειγμα (Πίνακας 3.2), ας θεωρήσουμε τη γραμμή που αντιστοιχεί στη λέξη "βροχή". Η υψηλότερη προσοχή που ρυθμίζει μια λέξη είναι συνήθως στον εαυτό της, αφού είναι η πιο παρόμοια με τον εαυτό της. Τέλος, κλιμακώνουμε τις βαθμολογίες προσοχής, διαιρώντας τις με τη διάσταση του διανύσματος key (στο παράδειγμά μας είναι 4) και τις απεικονίζουμε στο διάστημα [0, 1] χρησιμοποιώντας μια συνάρτηση softmax, υπολογίζοντας κατ' αυτόν τον τρόπο τις τελικές βαθμολογίες αυτο-προσοχής.

Επομένως έχουμε το φίλτρο προσοχής, το οποίο περιέχει βαθμολογίες προσοχής υπολογισμένες από το γινόμενο των πινάκων query και key και τον αρχικό πίνακα value, ο οποίος στην ουσία αντιπροσωπεύει όλη την πληροφορία της ακολουθίας εισόδου. Για να καταλάβουμε όμως περισσότερο τον ρόλο των βαθμολογιών προσοχής στην απόδοση ενός μοντέλου, καλύτερα να καταφύγουμε σε ένα παράδειγμα όπου η ακολουθία εισόδου είναι μία εικόνα (Σχήμα 3.5). Αντί να παίρνουμε όλη την πληροφορία της εικόνας, πολλαπλασιάζουμε τις βαθμολογίες αυτο-προσοχής με την αρχική εικόνα με αποτέλεσμα οι πρώτες να φιλτράρουν όλη την περιττή πληροφορία υποβάθρου, δίνοντας σημασία μόνο στην πληροφορία που πραγματικά πρέπει, ώστε να αντλήσουμε την μεγαλύτερη δυνατή πληροφορία του αντικειμένου που απεικονίζεται. Με τον ίδιο τρόπο, όταν πολλαπλασιάζουμε το φίλτρο προσοχής με τον πίνακα value λαμβάνουμε έναν φιλτραρισμένο πίνακα value ο οποίος αποδίδει υψηλή τιμή εστίασης στα χαρακτηριστικά που είναι πιο σημαντικά (Εξίσωση 3.9):

$$Attention(Q, K, V) = \frac{Q \cdot K^T}{\sqrt{d_k}} \cdot V \tag{3.9}$$

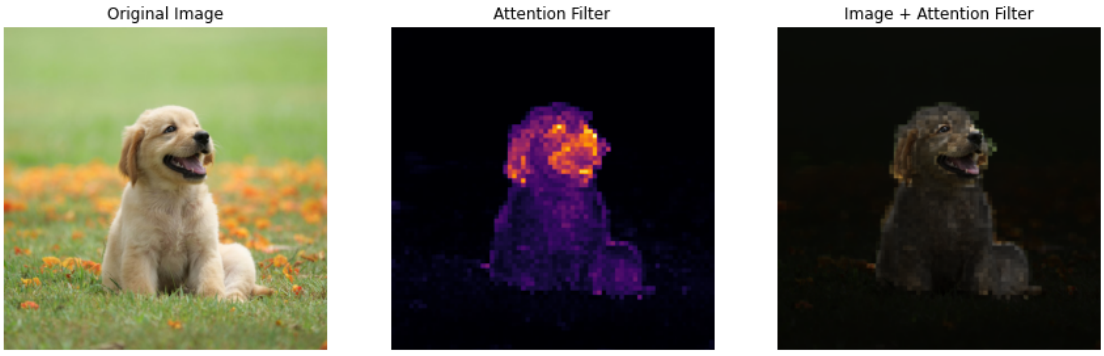
Όμως αυτός ο φιλτραρισμένος πίνακας value είναι η έξοδος ενός στρώματος προσοχής μιας κεφαλής. Το πρώτο φίλτρο μας βοήθησε να εστιάσουμε στο ποιος ή τι ήταν στην εικόνα, αλλά ίσως πρέπει να εστιάσουμε και σε άλλα μέρη της εικόνας, προκειμένου να δούμε αν υπάρχει κάποια πληροφορία που μπορεί να μας βοηθήσει. Οι μετασχηματιστές δεν μαθαίνουν ένα φίλτρο προσοχής, μαθαίνουν πολλαπλά· το καθένα εστιάζοντας σε ένα διαφορετικό γλωσσικό φαινόμενο ή στο παράδειγμά μας σε διαφορετικά μέρη της εικόνας. Επομένως, η τελική έξοδος του στρώματος προσοχής πολλαπλών κεφαλών αποτελείται από τη συνένωση (concatenation) όλων των φιλτραρισμένων πινάκων των values (Εξίσωση 3.11 και Σχήμα 3.6).

$$Multihead(Q, K, V) = Concat(head_1, head_2, ..., head_h) \cdot W^O, \tag{3.10}$$

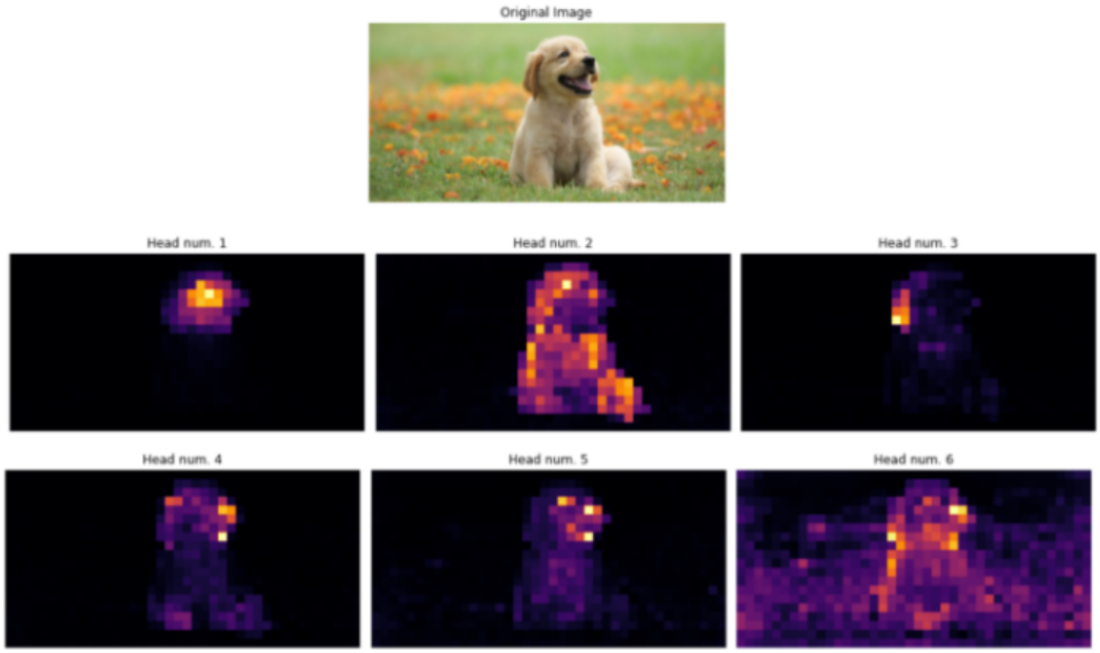
$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3.11}$$

3.2.3 Καλυπτόμενη προσοχή πολλαπλών κεφαλών (Masked Multi-head Attention)

Γενικά, ο κωδικοποιητής λαμβάνει την ακολουθία εισόδου και την μετατρέπει σε διάνυσμα. Στη συνέχεια, ο αποκωδικοποιητής λαμβάνει αυτό το διάνυσμα και το μετατρέπει



Σχήμα 3.5: *Παράδειγμα αυτο-προσοχής σε μία εικόνα που απεικονίζεται ένας σκύλος. Στην αριστερή εικόνα απεικονίζεται ο σκύλος, η μεσαία εικόνα είναι οι βαθμολογίες αυτο-προσοχής και στην δεξιά εικόνα απεικονίζεται το γινόμενο της αρχικής εικόνας με τις βαθμολογίες αυτο-προσοχής απομονώνοντας με αυτόν του τρόπο την απαραίτητη πληροφορία της εικόνας από το υπόβαθρο.*



Σχήμα 3.6: *Παράδειγμα προσοχής πολλαπλών κεφαλών στην εικόνα με του σκύλο. Κάθε "κεφαλή" εστιάζει σε διαφορετικό σημείο της εικόνας.*

	Today	is	raining
Today	89	20	41
is	22	90	81
raining	81	41	95

	Today	is	raining
Today	22	5	10
is	5	22	20
raining	20	10	23

	Today	is	raining
Today	0.98	0.005	0.015
is	0.01	0.88	0.11
raining	0.047	0.003	0.95

Πίνακας 3.2: *Οι βαθμολογίες του φίλτρου προσοχής όταν η ακολουθία εισόδου είναι η "Today is raining". Στον επάνω αριστερά πίνακα απεικονίζονται οι βαθμολογίες μετά τον προσδιορισμό του εσωτερικού γινομένου, στον επάνω αριστερά οι βαθμολογίες μετά την κλιμάκωση (διαίρεση με την διάσταση του Key, $d = 4$) και τέλος ο πίνακας κάτω στο κέντρο δείχνει τις τελικές βαθμολογίες αυτο-προσοχής μετά το Softmax.*

σε μια νέα ακολουθία ανάλογα με το πρόβλημα το οποίο προσπαθεί να λύσει το μοντέλο. Μια σημαντική διαφορά μεταξύ του κωδικοποιητή και του αποκωδικοποιητή είναι ότι ενώ ο πρώτος δέχεται μόνο μία είσοδο, δηλαδή την αρχική ακολουθία εισόδου, ο δεύτερος δέχεται δύο, την έξοδο του κωδικοποιητή και την ακολουθία-στόχο της εξόδου. Όσον αφορά την πρώτη είσοδο, αφού πάρουμε την έξοδο του κωδικοποιητή τη χωρίζουμε σε δύο αντίγραφα. Το ένα αντίγραφο αποτελεί τα queries και το δεύτερο αντίγραφο αποτελεί τα keys. Αυτά τα δύο αντίγραφα θα είναι τα queries και τα keys εισόδου του δεύτερου στρώματος προσοχής πολλαπλών κεφαλών του αποκωδικοποιητή. Όσον αφορά το πρώτο στρώμα καλυπτόμενης προσοχής πολλαπλών κεφαλών (masked multi-head attention layer) του αποκωδικοποιητή, η μόνη διαφορά με το παραπάνω στρώμα προσοχής πολλαπλών κεφαλών είναι μια πρόσθετη λειτουργία κάλυψης (masking). Όπως αναφέρθηκε και προηγουμένως, όταν εκπαιδεύεται το μοντέλο περνάμε την ακολουθία-στόχο στον αποκωδικοποιητή. Έτσι, κατά τη διάρκεια της λειτουργίας προσοχής πολλαπλών κεφαλών, λίγο πριν περάσουμε τα αποτελέσματα στο softmax εκτελούμε την λειτουργία κάλυψης (masking operation) προσθέτοντας στο φίλτρο κάλυψης (masking filter) τον πίνακα με τις βαθμολογίες κάλυψης. Οι βαθμολογίες κάλυψης είναι ένας πίνακας που όλες οι μελλοντικές λέξεις σε σχέση με κάθε λέξη-στόχο έχουν βαθμολογία $-\infty$. Για παράδειγμα (Πίνακας 3.3), αν βρισκόμαστε στο χρονικό βήμα όπου μόλις καταλήξαμε στη δημιουργία μιας ακολουθίας μέχρι τη λέξη "I" και θα θέλαμε να δημιουργήσουμε την επόμενη λέξη στην ακολουθία, το φίλτρο μάσκας θα προσθέσει αρνητικό άπειρο σε όλες τις λέξεις μετά το τέλος της λέξης ως βαθμολογία. Μόλις περάσουμε το φίλτρο προσοχής μάσκας από το στρώμα softmax, όλα τα $-\infty$ θα πάρουν τιμή ίση με το μηδέν και επομένως κατά την πρόβλεψη της λέξης μετά το "I" το μοντέλο δίνει προσοχή μόνο σε όλες τις λέξεις πριν το "I" και δίνει μηδενική προσοχή στις λέξεις που ακολουθούν.

Στην συνέχεια, η έξοδος του στρώματος καλυπτόμενης προσοχής πολλαπλών κεφαλών πηγαίνει στο στρώμα add & norm. Αυτό μας δίνει σαν αποτέλεσμα έναν πίνακα τον οποίο θα χρησιμοποιήσουμε ως πίνακα value στη δεύτερη μονάδα προσοχής πολλαπλών κεφαλών του αποκωδικοποιητή. Το στρώμα προσοχής πολλαπλών κεφαλών του αποκωδικοποιητή λαμβάνει στην πραγματικότητα τρεις εισόδους. Οι πρώτες δύο είσοδοι είναι οι πίνακες queries και keys που προήλθαν από τον κωδικοποιητή και η τρίτη είσοδος είναι ο πίνακας value που προέρχεται από την προηγουμένως παραγόμενη ακολουθία κειμένου. Η μονάδα προσοχής

	Yes	,	I	am
Yes	0	$-\infty$	$-\infty$	$-\infty$
,	0	0	$-\infty$	$-\infty$
I	0	0	0	$-\infty$
am	0	0	0	0

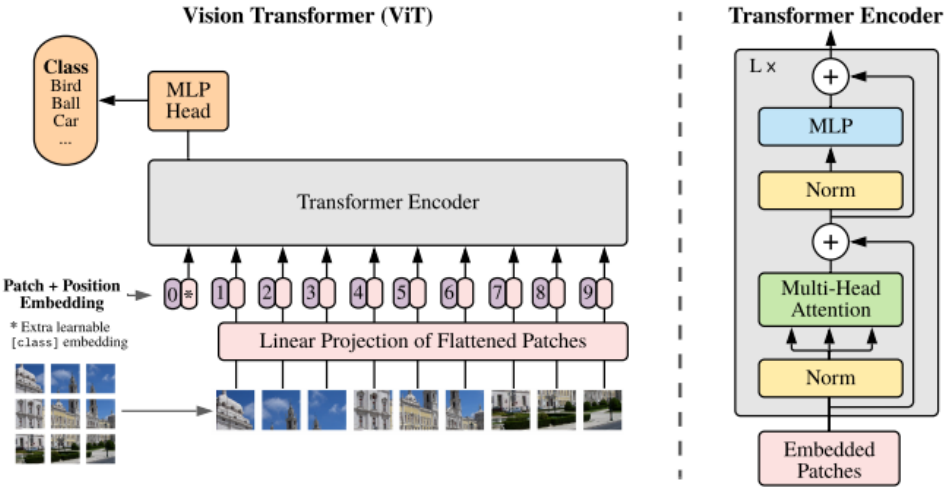
Πίνακας 3.3: *Παράδειγμα βαθμολογιών κάλυψης όταν η ακολουθία εξόδου είναι η φράση "Yes, I am".*

πολλαπλών κεφαλών λαμβάνει αυτές τις τρεις εισόδους και στη συνέχεια λειτουργεί ακριβώς με τον ίδιο τρόπο που περιγράψαμε στην Ενότητα 3.2.2. Τέλος, η έξοδος του επιπέδου προσοχής πολλαπλών κεφαλών ρέει προς τα άλλα στρώματα, όπως έχουμε ήδη περιγράψει στα παραπάνω κεφάλαια.

3.3 Οπτικοί Μετασχηματιστές (Vision Transformers - ViT)

Ένας μετασχηματιστής ξεπερνά σε απόδοση και τα καλύτερα μοντέλα στην αναγνώριση εικόνων [27], ένα πεδίο όπου μέχρι πρόσφατα κυριαρχούσαν τα συνελικτικά νευρωνικά δίκτυα. Ένα CNN χρησιμοποιεί πυρήνες (kernels) για να συγκεντρώσει τις πολύ τοπικές πληροφορίες σε κάθε στρώμα, οι οποίες στη συνέχεια περνούν στο επόμενο στρώμα που συγκεντρώνει και πάλι τοπικές πληροφορίες, αλλά αυτή τη φορά με μεγαλύτερο οπτικό πεδίο επειδή εξετάζει πληροφορίες που είχαν ήδη συγκεντρωθεί από το πρώτο στρώμα. Δηλαδή τα CNNs αρχίζουν να εξετάζουν πολύ τοπικά την εικόνα εισόδου και το πεδίο προσοχής τους γίνεται πιο σφαιρικό σε κάθε στρώμα αλλά μετά από αρκετό χρόνο εκπαίδευσης.

Αντίθετα, ο μετασχηματιστής εξετάζει την εικόνα παίρνοντας την εικόνα εισόδου και χωρίζοντάς την σε τμήματα των 16 επί 16 εικονοστοιχείων, τα οποία στη συνέχεια ισοπεδώνονται (flatten) με έναν γραμμικό μετασχηματισμό (έναν πίνακα), έτσι ώστε να γίνουν διανύσματα. Στη συνέχεια, η λειτουργία προσομοιάζει με τους κλασσικούς μετασχηματιστές (Ενότητα 3.2.2)· το διάνυσμα κάθε τμήματος λαμβάνει το διάνυσμα θέσης (Ενότητα 3.2.1) και στη συνέχεια χρησιμοποιείται ένας συνηθισμένος κωδικοποιητής μετασχηματιστή για να προβλέψει τι μπορεί να αναπαριστά η εικόνα χρησιμοποιώντας την μέθοδο προσοχής πολλαπλών κεφαλών. Στην εργασία [27] εκτελείται κατηγοριοποίηση εικόνων χρησιμοποιώντας μια ειδική είσοδο που ονομάζεται σύμβολο ταξινόμησης (classification token - CLS), το οποίο προστίθεται στην ακολουθία εισόδου του μετασχηματιστή κωδικοποίησης, μαζί με την κωδικοποίηση θέσης. Το CLS είναι είναι ένα εκπαιδεύσιμο διάνυσμα και η έξοδος του από τον μετασχηματιστή κωδικοποίησης χρησιμοποιείται προκειμένου να ταξινομηθεί η εικόνα. Η αρχιτεκτονική του οπτικού μετασχηματιστή παρουσιάζεται στο Σχήμα 3.7.



Σχήμα 3.7: *Επισκόπηση αρχιτεκτονικής μουτέλου οπτικού μετασχηματιστή. Αρχικά, η εικόνα χωρίζεται σε τμήματα σταθερού μεγέθους, ισοπεδώνονται γραμμικά καθένα από αυτά, και αφού προστεθούν κωδικοποιήσεις θέσης τροφοδοτείται η προκύπτουσα ακολουθία διανυσμάτων σε έναν τυπικό μετασχηματιστή κωδικοποίησης. Προκειμένου να πραγματοποιηθεί η ταξινόμηση της εικόνας, χρησιμοποιείται η προσθήκη ενός επιπλέον "συμβόλου ταξινόμησης" στην ακολουθία εισόδου. Η απεικόνιση του μετασχηματιστή κωδικοποίησης είναι εμπνευσμένη από [8]. (Πηγή [27]).*

Κεφάλαιο 4

Αυτο-επιβλεπόμενη μάθηση οπτικών συσχετίσεων

Στην παρούσα ενότητα περιγράφουμε αρχικά τα σύνολα δεδομένων που έχουμε στην διάθεση μας και αμέσως μετά αναφέρουμε τι είναι η αυτο-επιβλεπόμενη μάθηση. Τέλος, παρουσιάζουμε την προτεινόμενη αρχιτεκτονική του δικτύου ViD και εξηγούμε τα επιμέρους στοιχεία του.

4.1 Σύνολα Δεδομένων

Υπάρχει πληθώρα συνόλων δεδομένων για το πρόβλημα του εντοπισμού σχέσεων [1, 29, 30, 3, 9, 31]. Στον Πίνακα 4.1 αναγράφονται τα βασικά στατιστικά τους όπως το πλήθος των εικόνων στα σύνολα εκπαίδευσης και επαλήθευσης, τον αριθμό των κλάσεων των κατηγορημάτων αλλά και τον αριθμό των κλάσεων των οντοτήτων. Τα σύνολα δεδομένων που θα χρησιμοποιήσουμε για την διεξαγωγή των πειραμάτων μας είναι τα VRD [1], VG200 [9], δύο από τα πιο διαδεδομένα σύνολα δεδομένων στην βιβλιογραφία.

Dataset	Train/Test Images	Predicates	Objects
VRD [1]	4k/1k	70	100
VG-MSDN [29]	46.2k/10k	50	150
VG-VTE [3]	73.8k/25.8k	100	200
sVG [30]	64.7k/8.7k	24	399
VG200 [9]	75.6k/32.4k	50	150
VG80K [31]	99.9k/4.8k	29086	53304

Πίνακας 4.1: *Απαρίθμηση των συνόλων δεδομένων της βιβλιογραφίας με τις χαρακτηριστικές στατιστικές πληροφορίες τους.*

Πιο συγκεκριμένα, το σύνολο δεδομένων Visual Relationships Detection (VRD) [1] είναι ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων αναφοράς για τη δημιουργία γράφου σκηνής. Περιέχει 4.000 εικόνες εκπαίδευσης και 1.000 εικόνες επαλήθευσης με 70 κατηγορίες σχέσεων και 100 κατηγορίες αντικειμένων. Ο συνολικός αριθμός των επισημειωμένων τριπλετών που περιέχει είναι 203.284 στα δεδομένα εκπαίδευσης και 7.624 στα δεδομένα επαλήθευσης.

Επιπλέον, χρησιμοποιούμε το σύνολο δεδομένων Visual Genome (VG) [9]. Το VG περιέχει 108.077 εικόνες, 3.8 εκατομμύρια επισημειωμένα αντικείμενα και 2.3 εκατομμύρια επισημειωμένες τριπλέτες. Ακολουθούμε το φιλτράρισμα των [9] που χρησιμοποιείται συνήθως για εκπαίδευση SGG μοντέλων. Αυτό χωρίζεται σε 75.651 εικόνες εκπαίδευσης και 32.422 εικόνες επαλήθευσης και περιλαμβάνει 150 κατηγορίες αντικειμένων και 50 κατηγορίες σχέσεων. Στα Σχήματα 4.1 και 4.2 παρουσιάζουμε σε λογαριθμική κλίμακα τον αριθμό των δειγμάτων ανά κλάση στο σύνολο δεδομένων εκπαίδευσης.

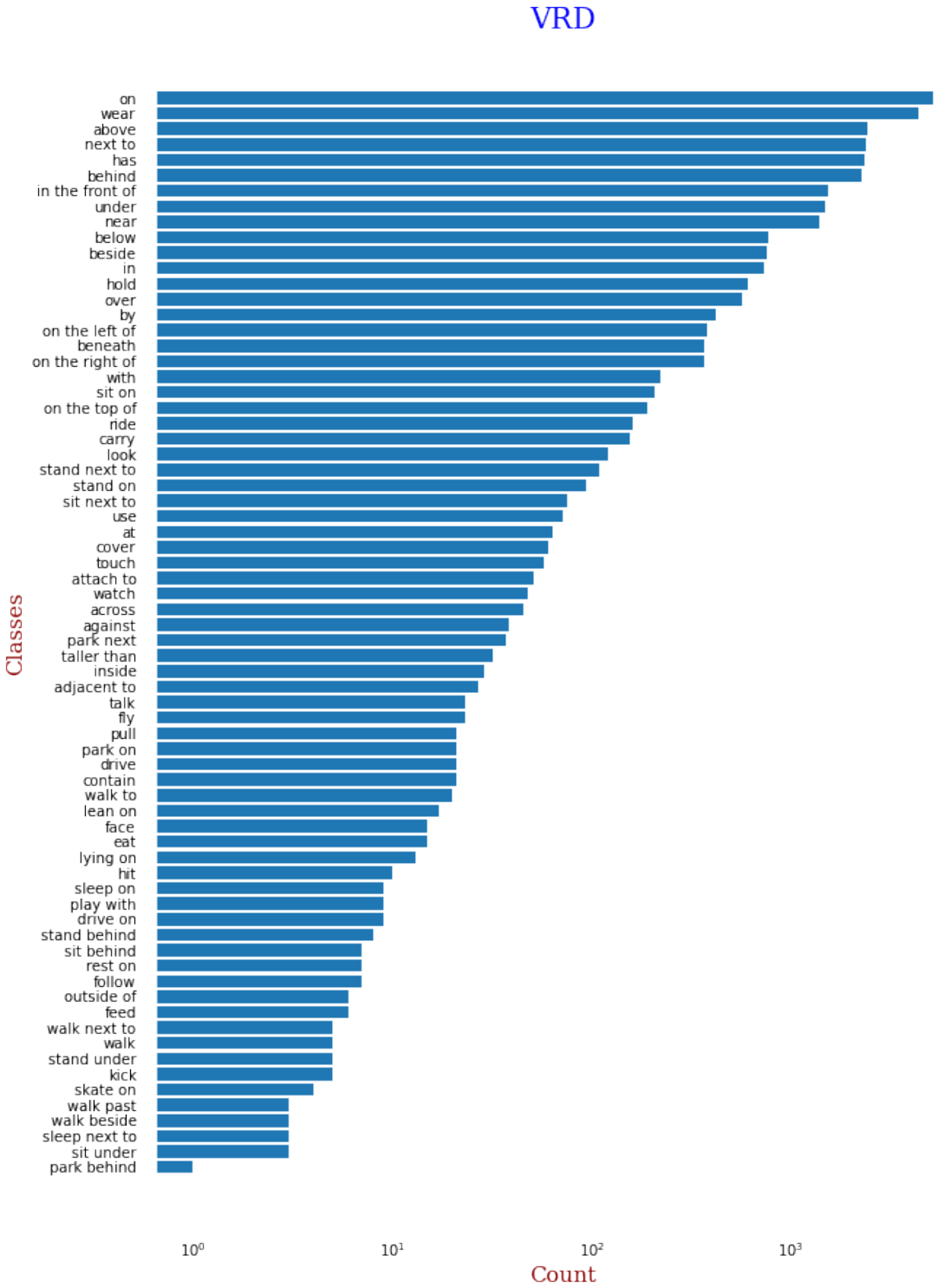
4.2 Αυτο-Επιβλεπόμενη Μάθηση

Η αυτοεπιβλεπόμενη μάθηση είναι ένας τύπος μηχανικής μάθησης που εκπαιδεύει μοντέλα τεχνητής νοημοσύνης χρησιμοποιώντας αυτοδημιουργούμενες ετικέτες αντί για χειροκίνητα δημιουργούμενες. Αυτή η προσέγγιση είναι χρήσιμη σε περιπτώσεις όπου είναι δύσκολο, ακριβό ή και χρονοβόρο να αποκτηθούν δεδομένα με ετικέτες ή όταν υπάρχει μεγάλος όγκος δεδομένων χωρίς ετικέτες που μπορούν να χρησιμοποιηθούν για εκπαίδευση.

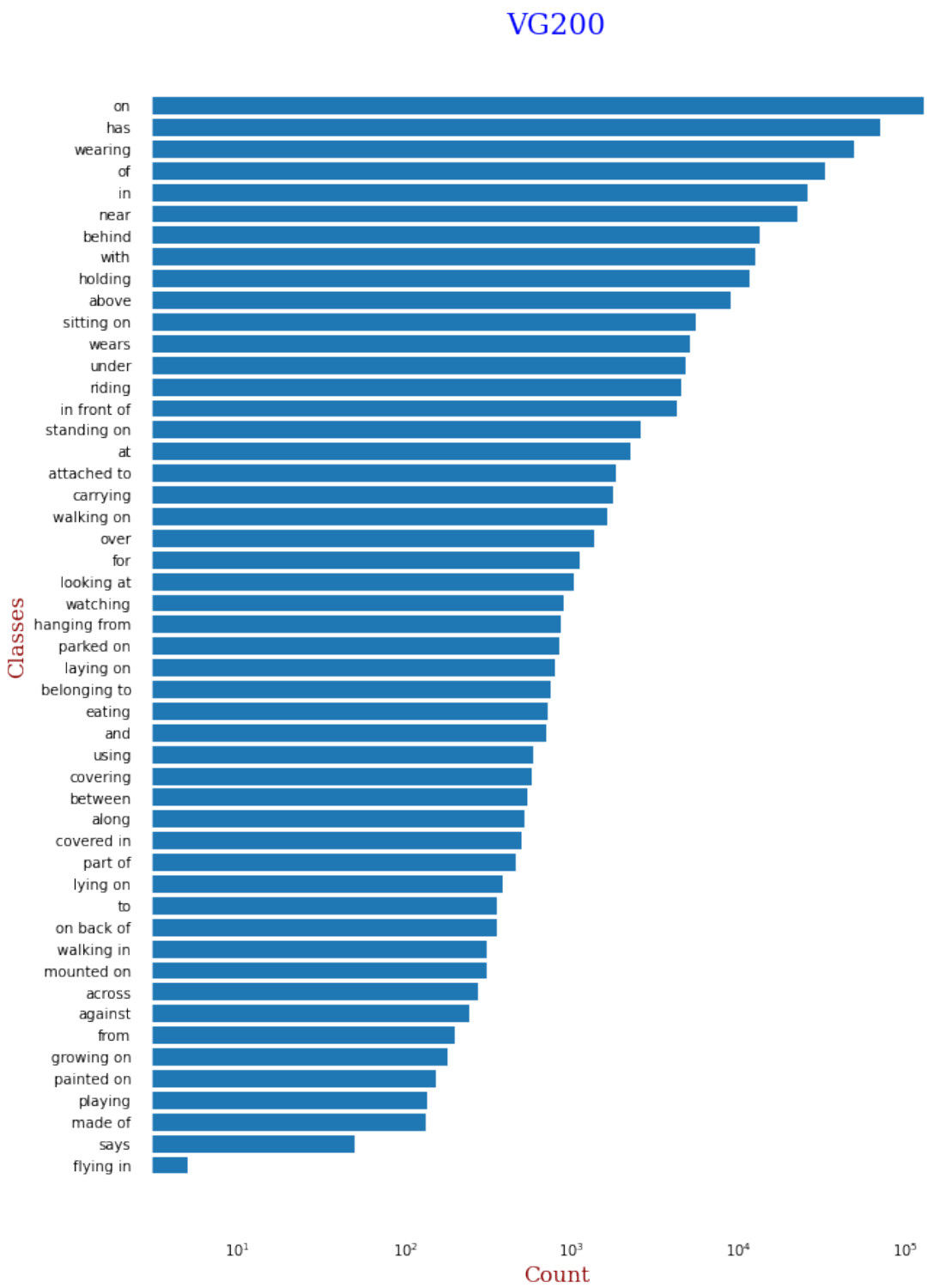
Στην αυτοεπιβλεπόμενη μάθηση, αντικαθιστούμε τις ανθρώπινες επισημειώσεις εκμεταλλευόμενοι δημιουργικά κάποια ιδιότητα των δεδομένων για τη δημιουργία μιας ψευδοεποπτευόμενης εργασίας. Για παράδειγμα, αντί να επισημάνουμε εικόνες ως γάτα/σκύλος, θα μπορούσαμε να τις περιστρέψουμε κατά 0/90/180/270 μοίρες και να εκπαιδεύσουμε ένα μοντέλο για την πρόβλεψη της περιστροφής. Μπορούμε να δημιουργήσουμε πρακτικά απεριόριστα δεδομένα εκπαίδευσης από εκατομμύρια εικόνες που διατίθενται ελεύθερα στο Διαδίκτυο. Το μοντέλο παράγει τις δικές του ετικέτες καθώς εκπαιδεύεται στα δεδομένα και ο στόχος είναι να μάθει χρήσιμες αναπαραστάσεις των δεδομένων που μπορούν να χρησιμοποιηθούν αργότερα για άλλες εργασίες με επίβλεψη. Η μάθηση με αυτοεπίβλεψη έχει υιοθετηθεί ευρέως στους τομείς της όρασης υπολογιστών και της επεξεργασίας φυσικής γλώσσας, όπου έχει αποδειχθεί ότι παράγει ισχυρές αναπαραστάσεις που μπορούν να φτάσουν ή και να ξεπεράσουν τις επιδόσεις των μοντέλων που εκπαιδεύονται σε δεδομένα με επίβλεψη. Αυτό ήταν ιδιαίτερα χρήσιμο για την προ-εκπαίδευση μοντέλων σε μεγάλες ποσότητες δεδομένων χωρίς ετικέτες, τα οποία μπορούν στη συνέχεια να τελειοποιηθούν (finetune) σε μικρότερες ποσότητες δεδομένων με ετικέτες για να επιτύχουν κορυφαία αποτελέσματα σε διάφορες εργασίες.

Υπάρχουν διάφορες προσεγγίσεις που έχουν προτείνει οι ερευνητές για την αξιοποίηση των ιδιοτήτων της εικόνας και την εφαρμογή αυτοεπιβλεπόμενης μάθησης για τη εκμάθηση αναπαραστάσεων. Παρακάτω παρουσιάζονται ενδεικτικά ορισμένες προσεγγίσεις:

1. **Ανακατασκευή** (recontruction)
    - Χρωματισμός Εικόνας (Image Colorization) [32]
    - Υπερ-ανάλυση εικόνας (Image Superresolution) [33]
    - Ανοικοδόμηση εικόνας (Image Inpainting) [34]
2. **Common Sense Tasks**
    - Jigsaw Puzzle Εικόνας [35]



Σχήμα 4.1: *Αριθμός δειγμάτων ανά κλάση σε λογαριθμική κλίμακα στο σύνολο δεδομένων εκπαίδευσης για το VRD [1].*



Σχήμα 4.2: *Αριθμός δειγμάτων ανά κλάση σε λογαριθμική κλίμακα στο σύνολο δεδομένων εκπαίδευσης για το VG200 [9].*

- Πρόβλεψη συμφραζομένων (Context Prediction) [36]
- Αναγνώριση γεωμετρικών μετασχηματισμών (Geometric Transformation Recognition)

Στην παρούσα διπλωματική εργασία θα κάνουμε ανακατασκευή εικόνας και πιο συγκεκριμένα, ανακατασκευή πολλών διαφορετικών περιοχών της εικόνας. Περισσότερες λεπτομέρειες παρουσιάζονται στην Ενότητα 4.3

4.3 Vision Decoder (ViD)

Στην παρούσα ενότητα παρουσιάζουμε το Vision Decoder - ViD και τη λεπτομερή υλοποίησή του. Υπάρχουν δύο στάδια στην μέθοδο που προτείνουμε: η προ-εκπαίδευση (pre-training) και η τελειοποίηση (fine-tuning) σε λίγα δείγματα. Οι αρχιτεκτονικές των ViD και δικτύων τελειοποίησης παρουσιάζονται αναλυτικά στα Σχήματα 4.3 και 4.4. Κατά την προ-εκπαίδευση, το μοντέλο εκπαιδεύεται σε μη επισημειωμένα δεδομένα. Για την τελειοποίηση με λίγα δείγματα, χρησιμοποιούμε ένα απλό πλήρως διασυνδεδεμένο νευρωνικό δίκτυο (Multi-Layer Perceptron - MLP) το οποίο μαθαίνει να κατηγοριοποιεί τις σχέσεις μεταξύ των αντικειμένων γνωρίζοντας ποια είναι τα ζευγάρια των αντικειμένων που αλληλεπιδρούν (PredDet).

Αρχιτεκτονική Μοντέλου

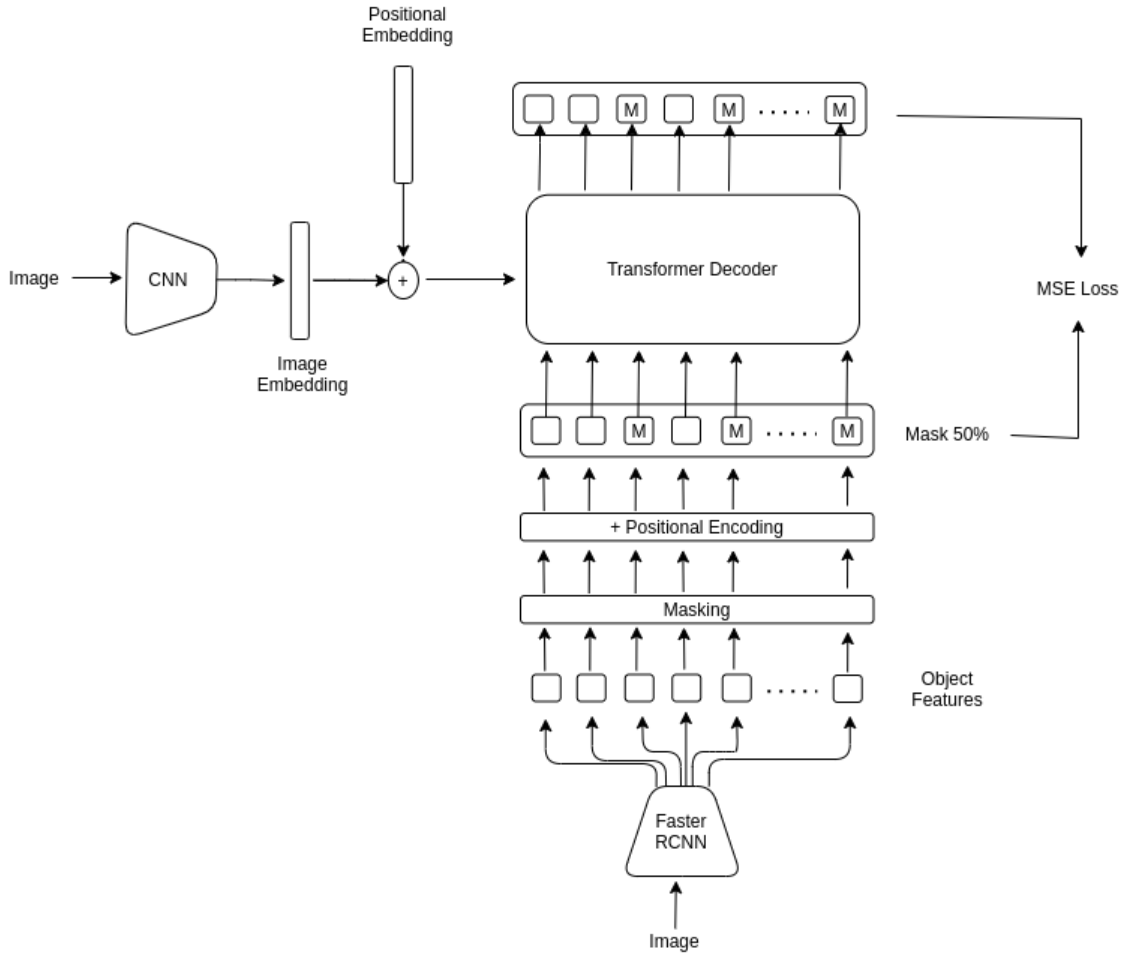
Τροποποιούμε το αρχικό μοντέλο μετασχηματιστή [8] κρατώντας μόνο τον αποκωδικοποιητή μετασχηματιστή (σε αντίθεση με τους οπτικούς μετασχηματιστές [27]) προσθέτοντας νέα στοιχεία για την προσαρμογή του οπτικού περιεχομένου. Πιο συγκεκριμένα, το μοντέλο μας αποτελείται από έναν αποκωδικοποιητή μετασχηματιστή [8] πολλαπλών επιπέδων έχοντας δύο εισόδους στο σύνολο, επιτρέποντας τη μοντελοποίηση εξαρτήσεων μεταξύ όλων των στοιχείων της πρώτης εισόδου ενώ παράλληλα λαμβάνονται υπόψιν και εξαρτήσεις της πρώτης εισόδου και της δεύτερης. Σε αντίθεση με προηγούμενες δουλειές στην βιβλιογραφία που χρησιμοποιούν αποκλειστικά ολόκληρη την εικόνα, το ViD δέχεται ως είσοδο

1. οπτικά στοιχεία, τα οποία αποτελούνται από οπτικά χαρακτηριστικά που ορίζονται από περιοχές ενδιαφέροντος (RoIs) στις εικόνες και
2. ολόκληρη την εικόνα.

Τα RoIs αποτελούν τα περιγράμματα που παράγονται από ανιχνευτές αντικειμένων (Faster-RCNN [10]).

Αναπαράσταση Εισόδου

Για την διαμόρφωση των διανυσμάτων της ακολουθίας εισόδου συνδυάζουμε 2 ειδών πληροφορίας. Τα οπτικά χαρακτηριστικά και την ενσωμάτωση θέσης. Όσον αφορά τα οπτικά χαρακτηριστικά χρησιμοποιούμε την οπτική πληροφορία που αντιστοιχεί στις περιοχές ενδιαφέροντος (RoIs) οι οποίες έχουν υπολογιστεί από τον ανιχνευτή αντικειμένων Faster-RCNN [10], δηλαδή το διάνυσμα χαρακτηριστικών που προκύπτει πριν από το στρώμα εξόδου κάθε



Σχήμα 4.3: *Προτεινόμενη αρχιτεκτονική προ-εκπαιδευμένου δικτύου ViD.*

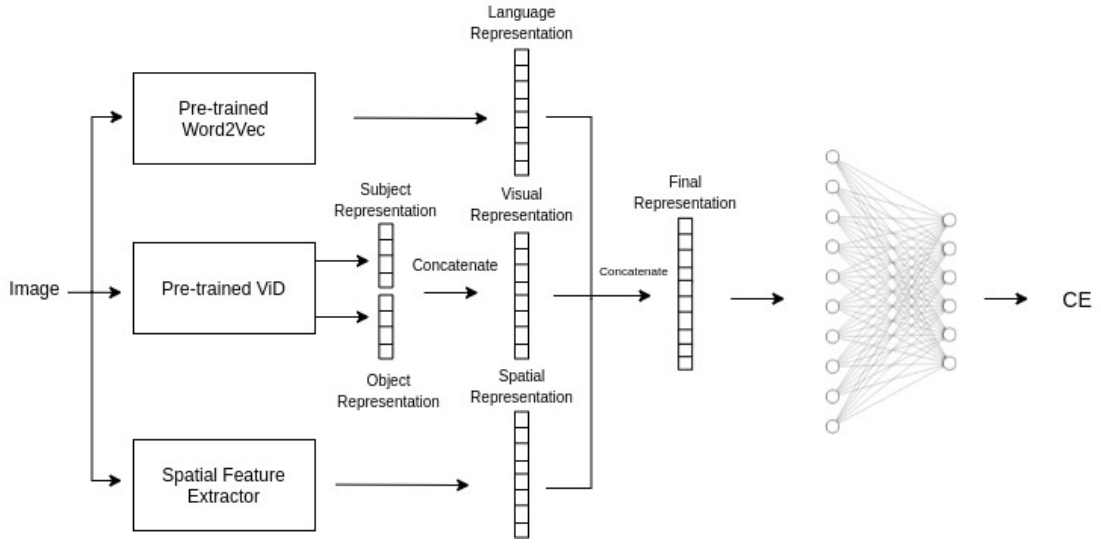
RoI στο Faster-RCNN. Στις περιπτώσεις που θέλουμε να εξάγουμε οπτικά χαρακτηριστικά για ολόκληρη την εικόνα εφαρμόζουμε πάλι το Faster-RCNN σε μια περιοχή ενδιαφέροντος ίση με τις διαστάσεις της εικόνας.

Χάρη στη φύση της μη ταξινομημένης αναπαράστασης της προσοχής του μετασχηματιστή, δηλαδή η θέση ενός στοιχείου της ακολουθίας εισόδου δεν είναι γνωστή στο δίκτυο, χρησιμοποιούμε τις ενσωματώσεις θέσεων για να διαμορφώσουμε την τελική αναπαράσταση της εισόδου. Στις περιπτώσεις που χρησιμοποιούμε ολόκληρη την εικόνα κάνουμε χρήση της κλασσικής ενσωμάτωσης θέσης [8], την οποία έχουμε περιγράψει αναλυτικά στην Ενότητα 3.2.1. Στην περίπτωση που η ακολουθία εισόδου του ViD είναι οι περιοχές ενδιαφέροντος ακολουθούμε μια διαφορετική προσέγγιση. Εφόσον οι περιοχές ενδιαφέροντος (δηλ. τα αντικείμενα) της εικόνας δεν ακολουθούν κάποια σχετική κατάταξη στην εικόνα αλλά και ούτε σχετίζονται μεταξύ τους με κάποιο σταθερό τρόπο, η κλασσική ενσωμάτωση θέσης δεν έχει νόημα να εφαρμοστεί. Ακολουθώντας την εργασία [37], η ενσωμάτωση θέσης του κάθε αντικειμένου αποκτά γεωμετρικό χαρακτήρα καθώς το δίκτυο θα ενημερώνεται για την γεωμετρική θέση κάθε αντικειμένου της εισόδου στην εικόνα. Πιο συγκεκριμένα, κάθε περιοχή ενδιαφέροντος (αντικείμενο) θα χαρακτηρίζεται από ένα διάνυσμα τεσσάρων διαστάσεων ως $(\frac{X_{LT}}{W}, \frac{Y_{LT}}{H}, \frac{X_{RB}}{W}, \frac{Y_{RB}}{H})$, όπου (X_{LT}, Y_{LT}) και (X_{RB}, Y_{RB}) οι συντεταγμένες του επάνω αριστερά και κάτω δεξιά σημείων του περιγράμματος των αντικειμένων, και W, H είναι το πλάτος και το ύψος της εικόνας. Στην συνέχεια, το διάνυσμα αυτό προβάλλεται σε ένα διανυσματικό χώρο υψηλών διαστάσεων υπολογίζοντας συναρτήσεις ημίτονου και συνημιτόνου διαφορετικών μηκών κύματος. Η τελική αναπαράσταση της εισόδου των περιοχών ενδιαφέροντος του ViD αποτελείται από την έξοδο ενός πλήρως συνδεδεμένου στρώματος που λαμβάνει ως είσοδο την συνένωση των οπτικών χαρακτηριστικών και των ενσωματώσεων γεωμετρικής θέσης.

Προ-εκπαίδευση (Pre-training)

Η αρχιτεκτονική του ViD μας επιτρέπει να το εκπαιδεύσουμε σε θεωρητικά άπειρο αριθμό εικόνων καθώς δεν υπάρχει η ανάγκη για επισημειωμένες σχέσεις. Στην παρούσα διπλωματική εργασία, προ-εκπαιδεύουμε το ViD στο VG200 σύνολο δεδομένων καθώς πειραματικά αποδείχθηκε πως ο αριθμός των εικόνων (4.000) στο VRD σύνολο δεν επαρκεί ώστε το δίκτυο μας να μάθει μία τέτοια απαιτητική εργασία όπως η ανακατασκευή αντικειμένων μιας εικόνας.

Πιο συγκεκριμένα, κατα την διάρκεια της εκπαίδευσης του ViD για κάθε εικόνα χρησιμοποιούμε το Faster-RCNN [10] ως ανιχνευτή αντικειμένων με παγωμένα βάρη, ώστε να λάβουμε τα οπτικά χαρακτηριστικά που αντιστοιχούν σε κάθε περιοχή ενδιαφέροντος (δηλ. αντικείμενα της εικόνας) και καλύπτουμε το 50% αυτών. Η προκύπτουσα ακολουθία οπτικών χαρακτηριστικών κάθε περιοχής ενδιαφέροντος αλλά και ολόκληρη η εικόνα μαζί με τις αντίστοιχες ενσωματώσεις θέσεων (όπως περιγράφουμε παραπάνω) αποτελούν τις δύο εισόδους του ViD. Το πρώτο στρώμα αυτο-προσοχής του ViD επιτρέπει τις αλληλεξαρτήσεις μεταξύ των αντικειμένων της ακολουθίας εισόδου μεταφέροντας με αυτόν τον τρόπο οπτική πληροφορία από το ένα αντικείμενο σε οποιοδήποτε άλλο της ακολουθίας. Στη συνέχεια, κάθε αντικείμενο της ακολουθίας κάνει διασταυρούμενη προσοχή (cross-attention) σε ολόκληρη την εικόνα σχηματίζοντας την τελική ακολουθία εξόδου. Τέλος, υπολογίζουμε το



Σχήμα 4.4: *Προτεινόμενη αρχιτεκτονική του δικτύου τελειοποίησης.*

μέσο τετραγωνικό σφάλμα (Mean Squared Error - MSE) ανάμεσα στην ακολουθία εξόδου και την ακολουθία εισόδου καθώς στόχος της προ-εκπαίδευσης είναι η ανακατασκευή των καλυπτομένων αντικειμένων. Οι αναπαραστάσεις που προκύπτουν από το ViD για τα αντικείμενα κάθε εικόνας, είναι ικανές να βοηθήσουν ένα δίκτυο λίγων μόνο παραμέτρων να μάθει πιο γρήγορα αλλά και πιο εύκολα τις σχέσεις μεταξύ τους.

Τελειοποίηση (Fine-tuning)

Αφού εκπαιδεύσουμε το ViD, το χρησιμοποιούμε με παγωμένα βάρη ως εξαγωγέα οπτικών αναπαραστάσεων. Πιο συγκεκριμένα, χρησιμοποιούμε ένα MLP 2 στρωμάτων το οποίο αρχικοποιείται με τυχαίες παραμέτρους και δέχεται σαν είσοδο γλωσσικά, χωρικά και οπτικά χαρακτηριστικά με τα τελευταία να εξάγονται από τον ViD με παγωμένα βάρη. Οι παραμέτροι του δικτύου τελειοποίησης (2-layer MLP) τελειοποιούνται χρησιμοποιώντας λίγα επισημειωμένα δεδομένα (1, 2 & 5 ανά κατηγορία σχέσης) για την το πρόβλημα του PredDet. Αποδεικνύουμε πως οι αναπαραστάσεις που εξάγει το ViD βοηθάνε ένα πολύ απλό δίκτυο να ξεπεράσει σε σημαντικό βαθμό τα υπάρχοντα μοντέλα της βιβλιογραφίας όταν αυτά εκπαιδεύονται σε λίγα επισημειωμένα δεδομένα ενώ πιάνει αποδόσεις σχεδόν ίδιες ή και καλύτερες από αυτά όταν εκπαιδεύονται σε ολόκληρα τα σύνολα δεδομένων εκπαίδευσης.

Κεφάλαιο 5

Πειραματική διαδικασία

Σε αυτό το κεφάλαιο, παρέχουμε μια λεπτομερή αναφορά στα εργαλεία και τις παραμέτρους που χρησιμοποιήθηκαν κατά την εκπαίδευση όλων των μεθόδων που αναφέρθηκαν παραπάνω. Στη συνέχεια, θα πραγματοποιήσουμε μια ολοκληρωμένη ανάλυση των μοντέλων από την υπάρχουσα βιβλιογραφία που έχουμε επανυλοποιήσει, συγκρίνοντας τόσο τις ποσοτικές όσο και τις ποιοτικές πτυχές τους.

5.1 Εκπαίδευση

Υλισμικό/Λογισμικό Όλα τα μοντέλα εκπαιδεύτηκαν σε NVIDIA 1080 Ti GPU σε σύστημα με 64GB RAM και Ubuntu 16.04. Όσον αφορά την προ-εκπαίδευση, μία εποχή εκπαίδευσης του ViD διαρκεί κατά μέσο όρο 85 λεπτά στο VG200. Στην περίπτωση της τελειοποίησης, μία εποχή εκπαίδευσης των μοντέλων σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης διαρκεί περίπου 5 λεπτά για τα μοντέλα της βιβλιογραφίας και 7 λεπτά για το ViD στο VRD και 70,93 λεπτά αντίστοιχα στο VG200. Ο χρόνος συμπερασμού (inference time) του ViD παρουσιάζει μία αύξηση της τάξεως των 80 δευτερολέπτων για το VRD και 10 λεπτών για το VG200. Κατά μέσο όρο οι εποχές εκπαίδευσης είναι 25 για το ViD στην προ-εκπαίδευση και 15 στην τελειοποίηση (ανάλογα με την αρχιτεκτονική του μοντέλου). Τα αποτελέσματα που θα παρουσιάσουμε αποτελούν την μέση τιμή μαζί με την τυπική απόκλιση από εκπαίδευση που πραγματοποιήθηκε 5 φορές.

Επανυλοποιήσεις Για την υλοποίηση των μοντέλων χρησιμοποιήσαμε τη βιβλιοθήκη PyTorch [38]. Επιλέξαμε μοντέλα διαφορετικών αρχιτεκτονικών και δυνατοτήτων προκειμένου να εξετάσουμε την συμπεριφορά τους σε εκπαίδευση με λίγα δεδομένα και για την βελτιστοποίηση χρησιμοποιήσαμε τον αλγόριθμο Adam [39] με weight decay ίσο με 5×10^{-4} για το VRD και 5×10^{-5} για το VG200. Δεδομένου του ότι ερευνούμε την απόδοση των μοντέλων σε εκπαίδευση με λίγα δεδομένα λογικό είναι να συγκρίνουμε μεταξύ της ίδιας υλοποίησης των δικτύων προκειμένου να είναι συγκρίσιμα τα αποτελέσματα. Για αυτό τον λόγο υλοποιήσαμε τα Motifs-Net [2], VTransE [3], UVTransE [4] και ATR-Net [5]. Παρόλα αυτά, παραθέτουμε στον Πίνακα 5.1 σύγκριση μεταξύ των δικών μας αποτελεσμάτων και εκείνων που αναφέρουν οι συγγραφείς τους. Έτσι βεβαιώνουμε ότι οι υλοποιήσεις μας είναι πολύ κοντά σε αυτές των συγγραφέων και μάλιστα σε πολλές περιπτώσεις ξεπερνούν.

Model	Original	Ours	Dataset
Motifs-Net [2]	65.2	64.2	VG200
VTransE [3]	44.76	53.04	VRD
UVTransE [4]	55.5	53.96	VRD
ART-Net [5]	58.4	58.04	VRD

Πίνακας 5.1: *Σύγκριση του R@20 που αναφέρεται επίσημα από την βιβλιογραφία με τις επανυλοποιήσεις μας.*

5.2 Ποσοτικά αποτελέσματα

5.2.1 Επιβεβαίωση λειτουργικότητας (Proof of Concept)

Πριν περάσουμε σε οποιαδήποτε διεξαγωγή πειραμάτων πρέπει να εξασφαλίσουμε την λειτουργικότητα της προτεινόμενης αρχιτεκτονικής μοντέλου. Η επιβεβαίωση της λειτουργικότητας είναι μια αρχική απόδειξη ότι μια συγκεκριμένη ιδέα ή προσέγγιση είναι εφικτή και στο πλαίσιο των νευρωνικών δικτύων είναι ένα ουσιαστικό βήμα πριν προχωρήσουμε σε περαιτέρω πειράματα ή ανάπτυξη. Τα νευρωνικά δίκτυα είναι πολύπλοκα συστήματα που απαιτούν σημαντικό χρόνο και υπολογιστικούς πόρους για την εκπαίδευσή τους και την αξιολόγησή τους. Ως εκ τούτου, είναι ζωτικής σημασίας να διασφαλιστεί ότι η επιλεγμένη προσέγγιση ή αρχιτεκτονική είναι κατάλληλη για το συγκεκριμένο πρόβλημα και μπορεί να παράγει ικανοποιητικά αποτελέσματα. Επιπλέον, μπορεί να βοηθήσει στον εντοπισμό πιθανών προβλημάτων και περιορισμών της προσέγγισης με αποτέλεσμα την αποφυγή της σπατάλης χρόνου και πόρων σε περαιτέρω πειράματα που είναι απίθανο να επιτύχουν.

Για την απόδειξη της λειτουργικότητας της αρχιτεκτονική μας αρχικά προ-εκπαιδεύουμε το ViD, παίρνοντας σαν είσοδο τα οπτικά χαρακτηριστικά ολόκληρης της εικόνας και τα οπτικά χαρακτηριστικά των περιοχών ενδιαφέροντος (αντικείμενα) της εικόνας, στα VRD και VG200 καλύπτοντας τυχαία ένα αντικείμενο σε κάθε εικόνα της δέσμης (batch) και τελειοποιούμε το δίκτυο που περιγράψαμε στην Ενότητα 4.3 σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης του VRD. Τα αποτελέσματα φαίνονται παρακάτω στον Πίνακα 5.2.

Model	Pre-trained	*R*@20
Motifs-Net [2]	—	53.7
VTransE [3]	—	53.04
UVTransE [4]	—	53.96
ART-Net [5]	—	58.4
ViD (ours)	VRD	23.67
ViD (ours)	VG200	49.89

Πίνακας 5.2: *Αρχική σύγκριση των μοντέλων της βιβλιογραφίας με το ViD με βάση την μετρική Recall@20 σε ολόκληρο το test-set του VRD.*

Όπως παρατηρούμε και από τον Πίνακα 5.2, τα αποτελέσματα του παραπάνω πειράμα-

Model	Masking ratio during pre-training	*R*@20
Visual Net	—	46.03
ViD (fine-tuned)	10%	40.1
	20%	40.6
	30%	46.1
	40%	46.27
	50%	**47.5**
	60%	<u>47.2</u>
	70%	43.32

Πίνακας 5.3: *Σύγκριση της μετρικής Recall@20 όταν το δίκτυο τελειοποίησης εκπαιδεύεται μόνο με τις αναπαραστάσεις που εξάγει ο ViD όταν αυτός έχει εκπαιδευτεί με διαφορετικές τιμές του ποσοστού επικάλυψης των αντικειμένων μιας εικόνας. Το Visual Net δέχεται ως είσοδο οπτικές αναπαραστάσεις από ένα ResNet 50 [12].*

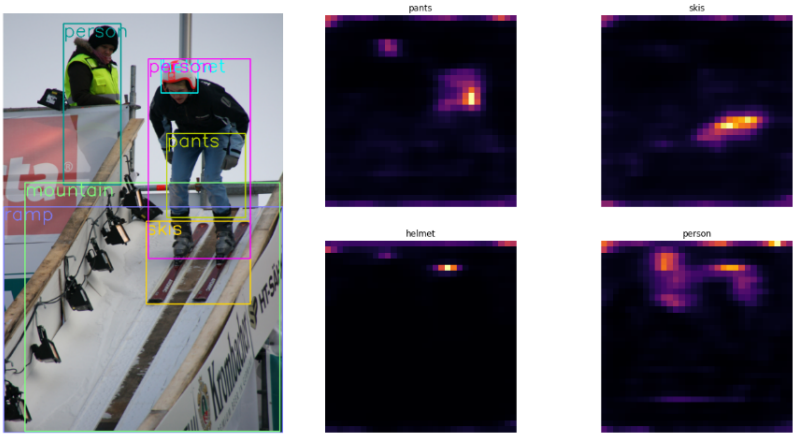
τος ήταν αρκετά ενθαρρυντικά καθώς προ-εκπαιδεύοντας το ViD στο VG200 παρατηρούμε πολύ μικρή μείωση της μετρικής *R*@20, της τάξεως του 3% – 4% οδηγώντας μας στο συμπέρασμα πως η αρχιτεκτονική του μοντέλου μας έχει σταθερές βάσεις ενώ παράλληλα όμως υπάρχει αρκετός χώρος για βελτίωση. Αξίζει σε αυτό το σημείο να αναφέρουμε πως η πτώση της μετρικής *R*@20 όταν το ViD έχει προ-εκπαιδευτεί στο σύνολο δεδομένων VRD είναι δικαιολογημένη, καθώς η ανακατασκευή περιοχών της εικόνας είναι μία δύσκολη εργασία και χρειάζεται αρκετά δείγματα εικόνων για να γίνει. Επομένως το VRD δεν επαρκεί καθώς περιέχει μόνο 4.000 εικόνες στο σύνολο δεδομένων εκπαίδευσης έναντι του VG200 που περιέχει 75.600 εικόνες.

5.2.2 Ablation Studies

Ποσοστό επικάλυψης (Masking Ratio)

Προκειμένου να μάθουμε το καλύτερο ποσοστό επικάλυψης των περιοχών ενδιαφέροντος κατά την διάρκεια της εκπαίδευσης του ViD διεξάγουμε πειράματα μετρώντας το *Recall*@20. Σε αυτά τα πειράματα όμως κατα την διάρκεια εκπαίδευσης του δικτύου τελειοποίησης θα δίνουμε σαν είσοδο μόνο τα οπτικά χαρακτηριστικά που εξάγει ο ViD εξαιρώντας τα γλωσσικά και χωρικά χαρακτηριστικά. Επιλέγουμε την συγκεκριμένη προσέγγιση διότι τα γλωσσικά και τα χωρικά χαρακτηριστικά βοηθάνε αρκετά την απόδοση του μοντέλου (της τάξεως του 10%) κάνοντας πιο δύσκολη την κατανόηση της αλλαγής της συμπεριφοράς του ViD στις διαφορετικές παραμέτρους όπως το ποσοστό επικάλυψης. Επομένως απομονώνοντας τα οπτικά χαρακτηριστικά του ViD είμαστε σε θέση να καταλάβουμε σε έναν αρκετά καλύτερο βαθμό το αντίκτυπο του ποσοστού επικάλυψης στην απόδοση του μοντέλου.

Στον Πίνακα 5.3 φαίνονται τα αποτελέσματα. Πειραματιζόμαστε με τις τιμές του ποσοστού επικάλυψης από 10% – 70%. Επιπλέον, συγκρίνουμε με το δίκτυο Visual Net το οποίο λειτουργεί σαν μοντέλο αναφοράς για το συγκεκριμένο πείραμα. Το Visual Net αποτελεί ένα δίκτυο ίδιας αρχιτεκτονικής με το δικό μας δίκτυο τελειοποίησης (2-layer MLP) το οποίο δέχεται σαν είσοδο τα οπτικά χαρακτηριστικά από τον Faster-RCNN [10] (όπως εξηγούμε στην Ενότητα 4.3) των αντικειμένων της εικόνας. Παρατηρούμε πως το 50% αποτελεί το καλύτερο ποσοστό επικάλυψης για το ViD κατά την διάρκεια της προ-εκπαίδευσης του κα-



Σχήμα 5.1: *Παράδειγμα εξαγωγής διασταυρούμενης προσοχής για τα αντικείμενα "άνθρωπος", "σκι", "παντελόνι" και "κράνος". Αριστερά απεικονίζεται η εικόνα με τα επισημειωμένα αντικείμενο από το σύνολο δεδομένων και δεξιά αυτής απεικονίζονται οι χάρτες διασταυρούμενης προσοχής. Παρατηρούμε πως κάθε αντικείμενο εστιάζει στην σωστή περιοχή της εικόνας.*

θώς το αντίστοιχο δίκτυο τελειοποίησης παρουσιάζει την καλύτερη επίδοση με τιμή ίση με 47.5 στην μετρική Recall@20 ενώ παράλληλα ξεπερνάει το Visual Net δίκτυο κατά 1.47%. Αξίζει να σημειωθεί σε αυτό το σημείο πως η πτώση της απόδοσης του μοντέλου σε χαμηλά (10%) αλλά και σε πολύ υψηλά (70%) ποσοστά επικάλυψης είναι δικαιολογημένη καθώς στην πρώτη περίπτωση η διαδικασία της ανακατασκευής περιοχών της εικόνας είναι σχετικά εύκολη επομένως το ViD λύνει εύκολα το πρόβλημα με αποτέλεσμα να μην μαθαίνει κάτι ουσιαστικό για τις αναπαραστάσεις των αντικειμένων, ενώ στην περίπτωση του υψηλού ποσοστού επικάλυψης το πρόβλημα γίνεται αρκετά δύσκολο για το ViD με αποτέλεσμα να μην μπορεί να το λύσει.

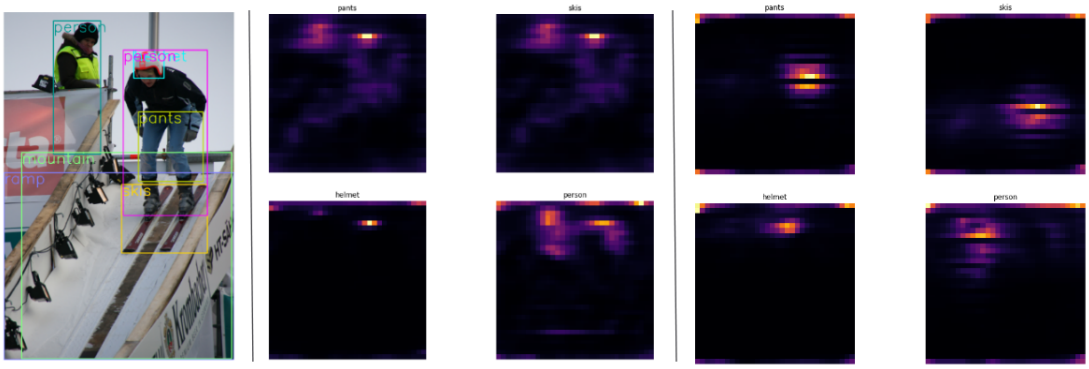
Επιπλέον,για να μελετήσουμε καλύτερα την συμπεριφορά του ViD εξάγουμε τους χάρτες διασταυρούμενης προσοχής (cross-attention maps) του κάθε αντικειμένου προς ολόκληρη την εικόνα και τους οπτικοποιούμε ώστε να καταλάβουμε που δίνει προσοχή το δίκτυο μας ανάλογα με το αντικείμενο και συνεπώς τι έχει μάθει. Όπως φαίνεται και στο Σχήμα 5.1 κάθε αντικείμενο της εικόνας εστιάζει στην σωστή περιοχή της εικόνας και επομένως το ViD φαίνεται να καταλαβαίνει και να αντιλαμβάνεται τα αντικείμενα εισόδου του καθώς μπορεί επιτυχώς να τα διαχωρίζει στην εικόνα.

Επομένως, με βάση τα παραπάνω μπορούμε να συμπεράνουμε πως οι αναπαραστάσεις των αντικειμένων που εξάγει το ViD δίκτυο είναι καλύτερες από αυτές ενός ResNet [12] καθώς είναι πιο ικανές να μάθουν ένα δίκτυο μικρής χωρητικότητας τις σχέσεις μεταξύ τους.

Κωδικοποίηση Θέσης (Positional Encoding)

Όπως έχουμε αναφέρει παραπάνω, τα δίκτυα μετασχηματιστών δεν διαθέτουν πληροφορίες σχετικά με τη θέση κάθε στοιχείου (αντικειμένου) στην ακολουθία εισόδου και επομένως δεν μπορεί να διακρίνει μεταξύ των διαφορετικών μετασχηματισμών της ίδιας ακολουθίας.

Πριν όμως περάσουμε στην μελέτη της κωδικοποίησης θέσης στην ακολουθία αντικειμένων εισόδου του ViD ας καταλάβουμε πρώτα καλύτερα τι πρόβλημα αντιμετωπίζουμε. Σκοπός του ViD αποτελεί η ανακατασκευή περιοχών μιας εικόνας. Όταν καλύπτουμε 2



Σχήμα 5.2: *Παράδειγμα εξαγωγής διασταυρούμενης προσοχής για τα αντικείμενα "άνθρωπος", "σκι", "παντελόνι" και "κράνος" όταν έχουμε επικαλύψει τα αντικείμενα "σκι" και "παντελόνι". Παρατηρούμε πως όταν ο ViD δεν διαθέτει κάποια πληροφορία σχετική με την θέση των αντικειμένων οι αναπαραστάσεις των επικαλυπτώμενων αντικειμένων είναι οι ίδιες (δεύτερη στήλη), ενώ όταν χρησιμοποιήσουμε κωδικοποίηση θέσης στην ακολουθία αντικειμένων τότε το ViD καταφέρνει και βρίσκει και τα δύο επιτυχώς.*

ή και περισσότερα αντικείμενα της ακολουθίας αντικειμένων, εφόσον αυτά δεν διαθέτουν κάποια πληροφορία για την σχετική τους θέση πάνω στην εικόνα, τότε για το ViD τα επικαλυπτόμενα αντικείμενα είναι ακριβώς τα ίδια. Επομένως το μοντέλο μας δεν μπορεί να ξεχωρίσει τα επικαλυπτόμενα αντικείμενα μεταξύ τους και με αυτό τον τρόπο το ViD μαθαίνει μια κατανομή πάνω στην εικόνα, η οποία είναι ίδια για κάθε επικαλυπτόμενο αντικείμενο, με βάση την οποία ελαχιστοποιείται η συνάρτηση κόστους. Επομένως οι αναπαραστάσεις των επικαλυπτόμενων αντικειμένων στην έξοδο του ViD θα είναι οι ίδιες.

Ο παραπάνω συλλογισμός μπορεί να γίνει καλύτερα αντιληπτός διεξάγοντας ένα πείραμα. Έχοντας εκπαιδεύσει το ViD θα εφαρμόσουμε συμπερασματολογία με παγωμένα βάρη αλλά επικαλύπτοντας 2 από τα αντικείμενα της εικόνας. Χρησιμοποιώντας την ίδια εικόνα με το Σχήμα 5.1, καλύπτουμε τα αντικείμενα "σκι" και "παντελόνι" και στη συνέχεια εξάγουμε τους χάρτες διασταυρούμενης περιοχής και τους οπτικοποιούμε. Τα αποτελέσματα φαίνονται στο Σχήμα 5.2. Όπως περιμέναμε, οι χάρτες διασταυρούμενης προσοχής και των δύο επικαλυπτόμενων αντικειμένων είναι ίδιοι για τον λόγο που περιγράψαμε παραπάνω· το ViD δεν διαθέτει κάποια πληροφορία σχετική με την θέση των αντικειμένων, ενώ όταν χρησιμοποιήσουμε κωδικοποίηση θέσης στην ακολουθία αντικειμένων τότε το ViD καταφέρνει και βρίσκει και τα δύο επιτυχώς ενώ ταυτόχρονα το μοντέλο είναι και πιο ακριβές όσον αφορά τον εντοπισμό όλων των αντικειμένων.

Όμως για να κατανοήσουμε και ποσοτικά καλύτερα την σημαντικότητα και το αντίκτυπο της κωδικοποίησης θέσης των αντικειμένων εισόδου του ViD στην απόδοση του μοντέλου μας διεξάγουμε πειράματα. Πιο συγκεκριμένα, στον Πίνακα 5.4 φαίνονται τα αποτελέσματα των πειραμάτων. Ομοίως όπως και παραπάνω για την καλύτερη αξιολόγηση της σημαντικότητας της κωδικοποίηση θέσης στην απόδοση του μοντέλου, απομονώνουμε τα οπτικά χαρακτηριστικά από τα γλωσσικά και τα χωρικά και εκπαιδεύουμε το δίκτυο τελειοποίησης μόνο με αυτά. Παρατηρούμε αρκετά μεγάλη αύξηση στην απόδοση του μοντέλου μας με τιμή ίση με 1.1% και επομένως η αύξηση του *Recall*@20 σε σχέση με το Visual Net ανεβαίνει στο 2.57%.

Model	Positional Encoding	*R*@20
Visual Net	—	46.03
ViD (fine-tuned)	no	47.5
	yes	**48.6**

Πίνακας 5.4: *Σύγκριση της μετρικής Recall@20 όταν το δίκτυο τελειοποίησης εκπαιδεύεται μόνο με τις οπτικές αναπαραστάσεις που εξάγει το Vid όταν αυτός έχει προ-εκπαιδευτεί με κωδικοποίηση θέσης και χωρίς. Το Visual Net δέχεται ως είσοδο οπτικές αναπαραστάσεις από ένα ResNet 50 [12].*

5.2.3 Εκπαίδευση με λίγα δείγματα (Few-shot Learning)

Όπως έχουμε αναφέρει, σε αυτή την διπλωματική εργασία εξετάζουμε το κατα πόσο ένα προ-εκπαιδευμένο δίκτυο αυτο-επιβλεπόμενης μάθησης (ViD) βοηθάει ένα δίκτυο μικρής χωρητικότητας να μάθει τις σχέσεις μεταξύ των οντοτήτων μιας εικόνας όταν αυτό εκπαιδεύεται με λίγα δείγματα ανά κλάση σχέσης. Ωστόσο αξίζει να επισημάνουμε πως και στα δύο σύνολα δεδομένων που ασχολούμαστε, στο VRD [1] και VG200 [9], παρατηρείται πληθώρα επισημειωμένων σχέσεων οι οποίες περιέχουν αρκετό θόρυβο, όπως για παράδειγμα η επισημειωμένη σχέση <Γυναίκα, φοράει, Γυναίκα>. Τέτοια δείγματα σχέσεων θα προκαλούσαν μεγάλη σύγχυση στο μοντέλο τελειοποίησης κατά την διάρκεια εκπαίδευσης του με αποτέλεσμα να μαθαίνει λάθος τις σχέσεις ή και να μην μαθαίνει τίποτα.

Για να αντιμετωπίσουμε το παραπάνω πρόβλημα, πηγαίνουμε εμείς οι ίδιοι χειροκίνητα και στο VRD αλλά και στο VG200 και εντοπίζουμε k σχέσεις ανά κλάση στα δεδομένα εκπαίδευσης οι οποίες είναι νοηματικά σωστές, όπου k είναι ο αριθμός λίγων δειγμάτων εκπαίδευσης. Πειραματιζόμαστε με 1, 2 & 5 δείγματα ανά κλάση. Τα αποτελέσματα της εκπαίδευσης με λίγα δείγματα αναγράφονται στον Πίνακα 5.5 για το VRD και στον Πίνακα 5.6 για το VG200 σύνολο δεδομένων. Όπως μπορούμε να παρατηρήσουμε, το μοντέλο μας ξεπερνάει σε όλα τα πειράματα όλα τα δίκτυα αναφοράς και στα 2 σύνολα δεδομένων. Η διαφορά απόδοσης είναι αρκετά πιο εμφανής όταν εκπαιδεύουμε τα δίκτυα με 5 δείγματα ανά κλάση, όπου το μοντέλο τελειοποίησης πετυχαίνει απόδοση με μέση τιμή ίση με **18.57** στο VRD ενώ η αμέσως καλύτερη απόδοση είναι του UVTransE [4] με μέση τιμή ίση με **11.2**, δηλαδή με διαφορά ίση με **7.37**. Όσον αφορά το VG200 σύνολο δεδομένων, το δίκτυο τελειοποίησης μας πετυχαίνει απόδοση με μέση τιμή ίση με **13.06** όταν εκπαιδεύεται με 5 δείγματα ανά κλάση ενώ η αμέσως καλύτερη επίδοση είναι του UVTransE με μέση τιμή ίση με **7.93**, δηλαδή με διαφορά ίση με **5.13**.

Ωστόσο ο εντοπισμός των νοηματικά σωστών σχέσεων από εμάς για κάθε σύνολο δεδομένων αποδείχτηκε μία αρκετά χρονοβόρα διαδικασία με αποτέλεσμα η εκπαίδευση των δικτύων σε περισσότερα δεδομένα από 5 ανά σχέση να είναι διαδικασία αρκετά απαιτητική. Για αυτό τον λόγο, επιλέγουμε να εκπαιδεύσουμε τα δίκτυα σε δεδομένα τα οποία έχουν επιλεγεί τυχαία από το σύνολο δεδομένων με στόχο την μελέτη της απόδοσης των δικτύων σε περισσότερα από 5 δεδομένα ανά σχέση. Στον Πίνακα 5.7 παρουσιάζονται τα αποτελέσματα για τις επιδόσεις των δικτύων της βιβλιογραφίας σε σύγκριση με το δικό μας δίκτυο όταν αυτά έχουν εκπαιδευτεί με 1, 5, 10, 20, 50 & 100 δείγματα ανά κλάση στο VRD σύνολο δεδομένων. Παρατηρούμε ότι το δίκτυο τελειοποίησης με το προ-εκπαιδευμένο ViD πετυχα-

VRD Dataset	R@20		
Model	1-shot	2-shot	5-shot
Motifs Net [2]	$0.1_{\pm 0.1}$	$3.62_{\pm 1.91}$	$1.17_{\pm 2.34}$
VTransE Net [3]	$4.06_{\pm 2.62}$	$4.63_{\pm 2.62}$	$10.7_{\pm 1.34}$
UVTransE Net [4]	$3.07_{\pm 1.93}$	$4.92_{\pm 1.99}$	$11.2_{\pm 0.63}$
ATR Net [5]	$1.43_{\pm 1.3}$	$1.17_{\pm 0.62}$	$2.67_{\pm 0.84}$
ViD (ours)	$\mathbf{6.73}_{\pm 2.92}$	$\mathbf{7.86}_{\pm 2.54}$	$\mathbf{18.57}_{\pm 1.54}$

Πίνακας 5.5: *Σύγκριση της μετρικής Recall@20 όταν το δίκτυο τελειοποίησης εκπαιδεύεται με 1, 2, & 5 δείγματα ανά κλάση στο VRD. Παρατηρούμε ότι το μοντέλο μας ξεπερνάει όλα τα μοντέλα αναφοράς.*

VG200 Dataset	R@20		
Model	1-shot	2-shot	5-shot
Motifs Net [2]	$1.66_{\pm 2.07}$	$2.58_{\pm 4.51}$	$0.6_{\pm 0.72}$
VTransE Net [3]	$4.16_{\pm 0.86}$	$4.13_{\pm 2.05}$	$7.1_{\pm 2.73}$
UVTransE Net [4]	$5.42_{\pm 3.18}$	$6.53_{\pm 3.57}$	$7.93_{\pm 2.81}$
ATR Net [5]	$0.99_{\pm 1.56}$	$1.96_{\pm 3.5}$	$1.56_{\pm 1.61}$
ViD (ours)	$\mathbf{5.67}_{\pm 3.1}$	$\mathbf{10.28}_{\pm 3.09}$	$\mathbf{13.06}_{\pm 3.01}$

Πίνακας 5.6: *Σύγκριση της μετρικής Recall@20 όταν το δίκτυο τελειοποίησης εκπαιδεύεται με 1, 2, & 5 δείγματα ανά κλάση στο VG200. Παρατηρούμε ότι το μοντέλο μας ξεπερνάει όλα τα μοντέλα αναφοράς.*

VRD Dataset	R@20					
Model	1-shot	5-shot	10-shot	20-shot	50-shot	100-shot
Motifs Net [2]	$0.17_{\pm 0.12}$	$0.18_{\pm 0.2}$	$2.48_{\pm 3.28}$	$2.88_{\pm 6.1}$	$3.56_{\pm 1.98}$	$4.84_{\pm 2.51}$
VTransE Net [3]	$2.65_{\pm 1.96}$	$5.13_{\pm 4.32}$	$9.75_{\pm 2.55}$	$14.66_{\pm 3.54}$	$19.14_{\pm 1.3}$	$23.3_{\pm 0.88}$
UVTransE Net [4]	$1.75_{\pm 1.39}$	$5.37_{\pm 2.13}$	$10.41_{\pm 3.29}$	$14.98_{\pm 0.82}$	$20.25_{\pm 2.58}$	$24.81_{\pm 1.37}$
ATR Net [5]	$0.64_{\pm 0.68}$	$1.1_{\pm 1.49}$	$2.05_{\pm 0.58}$	$15.68_{\pm 0.82}$	$22.83_{\pm 1.29}$	$27.2_{\pm 0.66}$
ViD (ours)	$\mathbf{2.65}_{\pm 1.51}$	$\mathbf{9.09}_{\pm 2.47}$	$\mathbf{17.57}_{\pm 1.27}$	$\mathbf{19.17}_{\pm 1.51}$	$\mathbf{24.5}_{\pm 1.29}$	$\mathbf{29.57}_{\pm 0.85}$

Πίνακας 5.7: *Σύγκριση της μετρικής Recall@20 όταν το δίκτυο τελειοποίησης εκπαιδεύεται με περισσότερα δείγματα ανά κλάση στο VRD.*

ίνει καλύτερες επιδόσεις σε όλες τις περιπτώσεις εκπαίδευσης. Όπως περιμέναμε, η επίδοση των 1 και 5 δειγμάτων ανά σχέση εκπαίδευσης παρουσιάζουν πτώση σε σχέση με αυτές του Πίνακα 5.5 σε όλα τα δίκτυα. Η διαφορά μεταξύ των επιδόσεων του δικού μας δικτύου σε σχέση με τα δίκτυα της βιβλιογραφίας είναι κυρίως αισθητή με τις εκπαιδεύσεις με 5, 10 και 20 δείγματα ανά σχέση, με μεγαλύτερη να είναι διαφορά με εκπαίδευση με 10 δείγματα ανά σχέση και με τιμή ίση με **7.16**. Όσο αυξάνονται τα δείγματα εκπαίδευσης ανά σχέση τόσο και μικρότερη είναι η διαφορά στην τιμή του Recall@20.

5.2.4 Εκπαίδευση με ολόκληρα δεδομένα εκπαίδευσης

Εκτός από την επίδοση του δικτύου μας στην εκπαίδευση με λίγα δεδομένα εξετάζουμε και την επίδοση του όταν αυτό εκπαιδεύεται με ολόκληρα δεδομένα. Οι Πίνακες 5.8 και 5.9 παρουσιάζουν τα αποτελέσματα στα VRD και VG200 σύνολα δεδομένων αντίστοιχα. Παρατηρούμε πως και στα δύο σύνολα δεδομένων το δίκτυο μας πετυχαίνει την δεύτερη καλύτερη επίδοση με την διαφορά από το μοντέλο με την καλύτερη επίδοση (ATR-Net) να έχει τιμή ίση με 1.6 και 0.71 στα VRD και VG200 αντίστοιχα.

Model	R@20
Motifs-Net [2]	53.7
VTransE [3]	53.04
UVTransE [4]	53.96
ATR-Net [5]	**58.04**
ViD (ours)	<u>56.8</u>

Πίνακας 5.8: *Σύγκριση του R@20 που αναφέρεται σε εκπαιδεύσεις χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων εκπαίδευσης του VRD.*

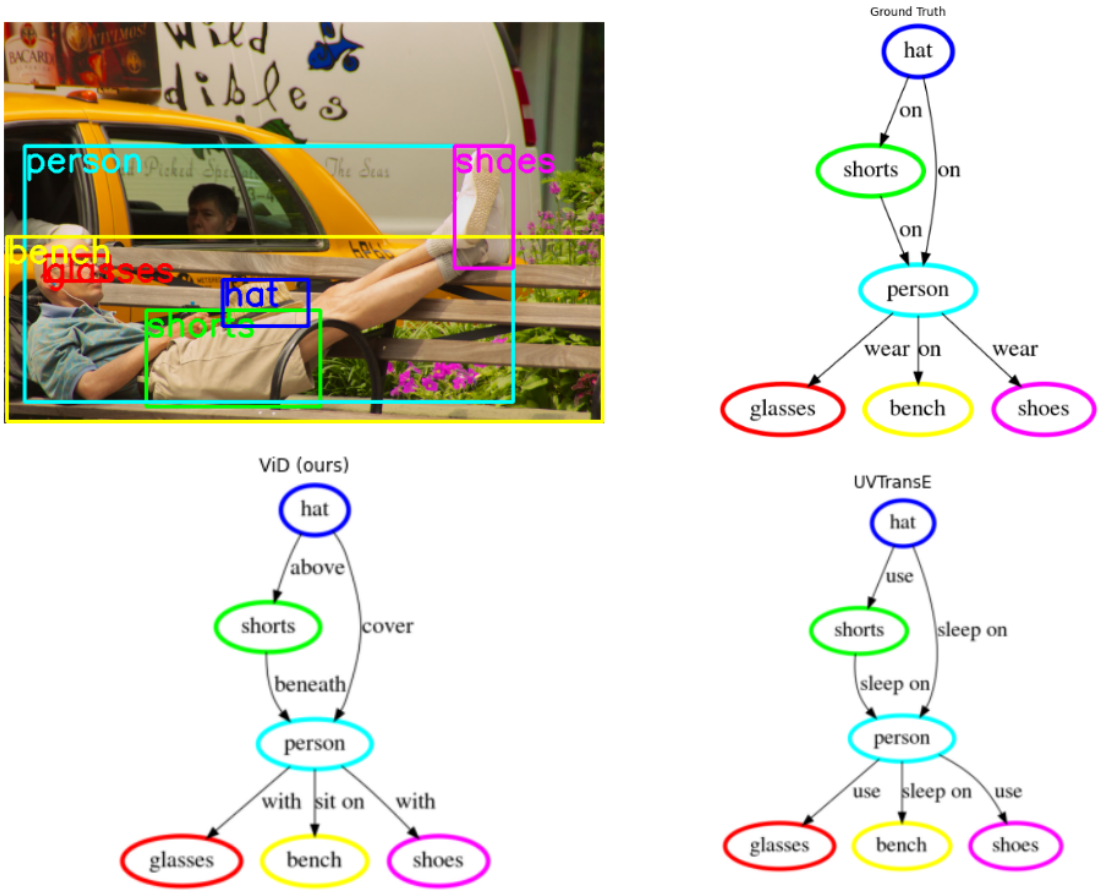
5.3 Ποιοτικά αποτελέσματα

Στα Σχήματα 5.3 έως 5.6 βλέπουμε ορισμένα ποιοτικά παραδείγματα που αναδεικνύουν την βελτίωση που προκύπτει από την δική μας μέθοδο σε σύγκριση με μεθόδους της βι-

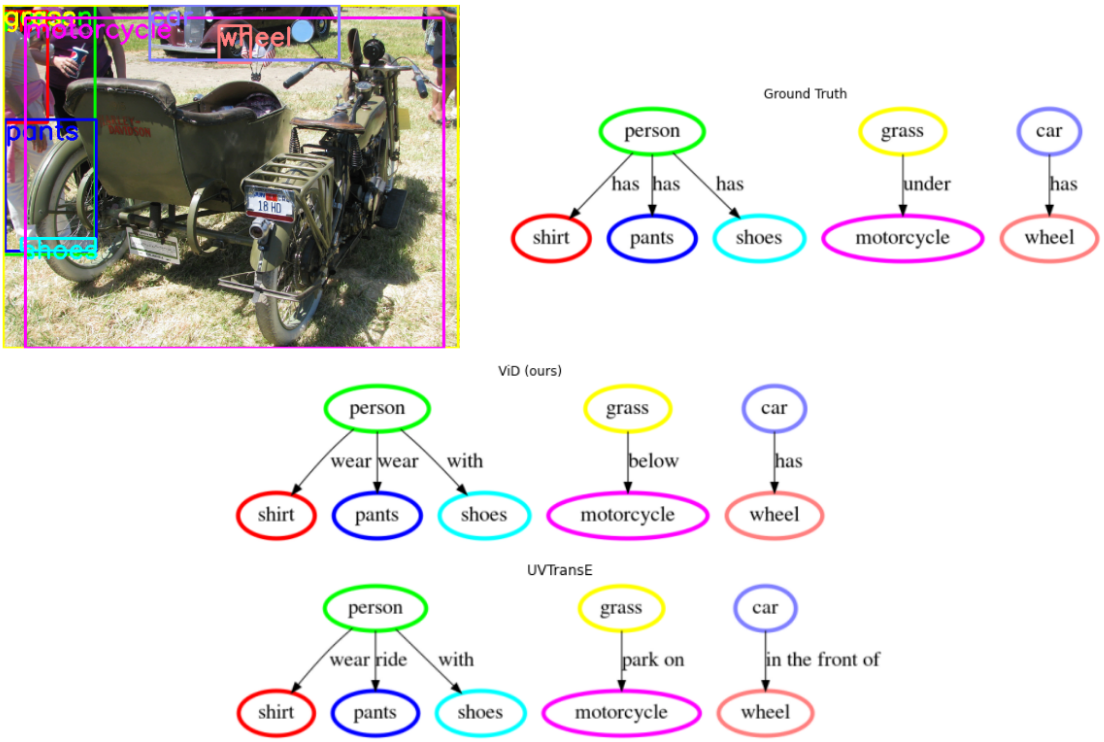
Model	R@20
Motifs-Net [2]	67.46
VTransE [3]	67.42
UVTransE [4]	64.76
ATR-Net [5]	**68.61**
ViD (ours)	<u>67.9</u>

Πίνακας 5.9: *Σύγκριση του R@20 που αναφέρεται σε εκπαιδεύσεις χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων εκπαίδευσης του VG200.*

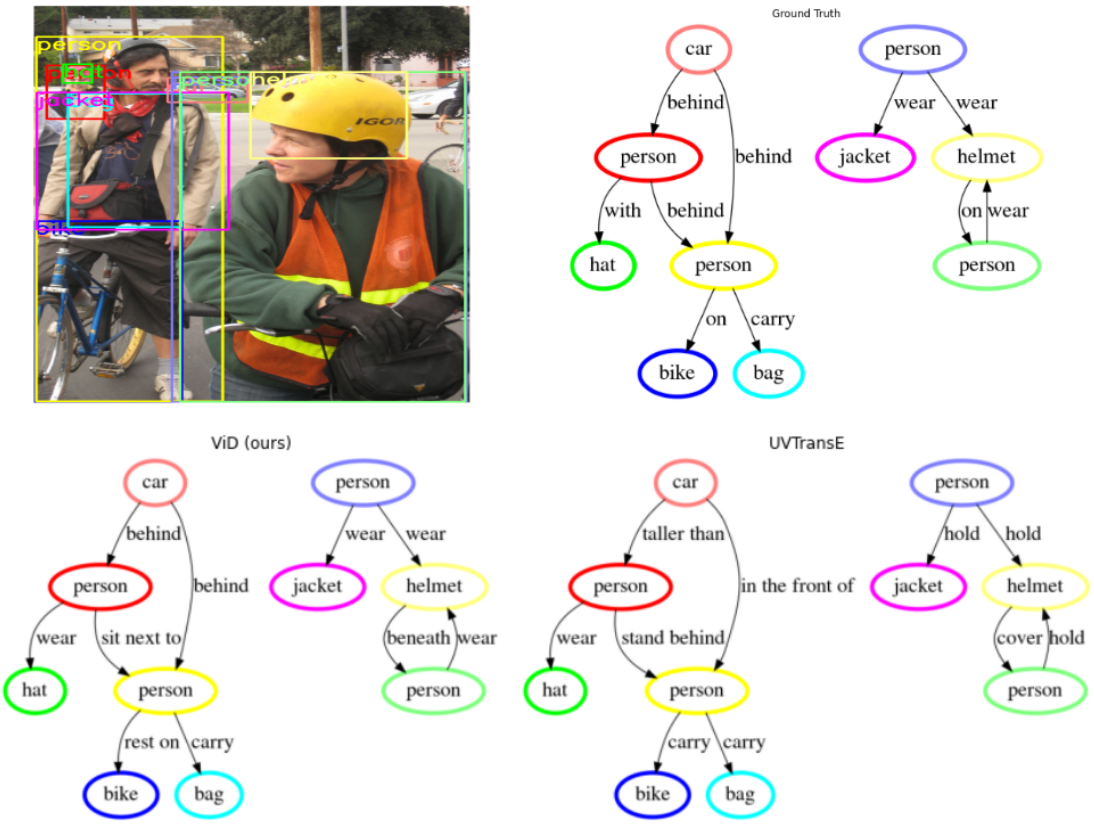
βλιογραφίας. Πιο συγκεκριμένα παρουσιάζουμε τους παραγόμενους γράφους σκηνής που δημιουργούνται από το δικό μας δίκτυο τελειοποίησης σε σύγκριση με το καλύτερο μοντέλο της βιβλιογραφίας δηλαδή το UVTransE [4], όταν αυτά έχουν εκπαιδευτεί με 5 δείγματα ανά σχέση. Κάθε εικόνα περιέχει πρώτα την εικόνα πάνω στην οποία θα κάνουμε οπτική αναγνώριση σχέσεων, μαζί με τα περιγράμματα των αντικειμένων και τις κατηγορίες τους, αμέσως δίπλα απεικονίζεται ο γράφος σκηνής που προκύπτει με βάσει τις επισημειώσεις από το σύνολο δεδομένων και τέλος από κάτω απεικονίζονται άλλοι 2 γράφοι, από τους οποίους ο πρώτος προκύπτει από το δικό μας μοντέλο και ο τελευταίος προκύπτει από το UVTransE. Όπως μπορούμε να παρατηρήσουμε, ο παραγόμενος γράφος του ViD δεν πετυχαίνει ακριβώς τις επισημειωμένες κατηγορίες του συνόλου δεδομένων παρ' όλα αυτά, οι προβλέψεις του είναι σημασιολογικά κοντά σε αυτές ή και πολλές φορές είναι σωστές, σε αντίθεση με τις προβλέψεις του UVTransE που απέχουν σημασιολογικά αρκετά από τις πραγματικές σχέσεις.



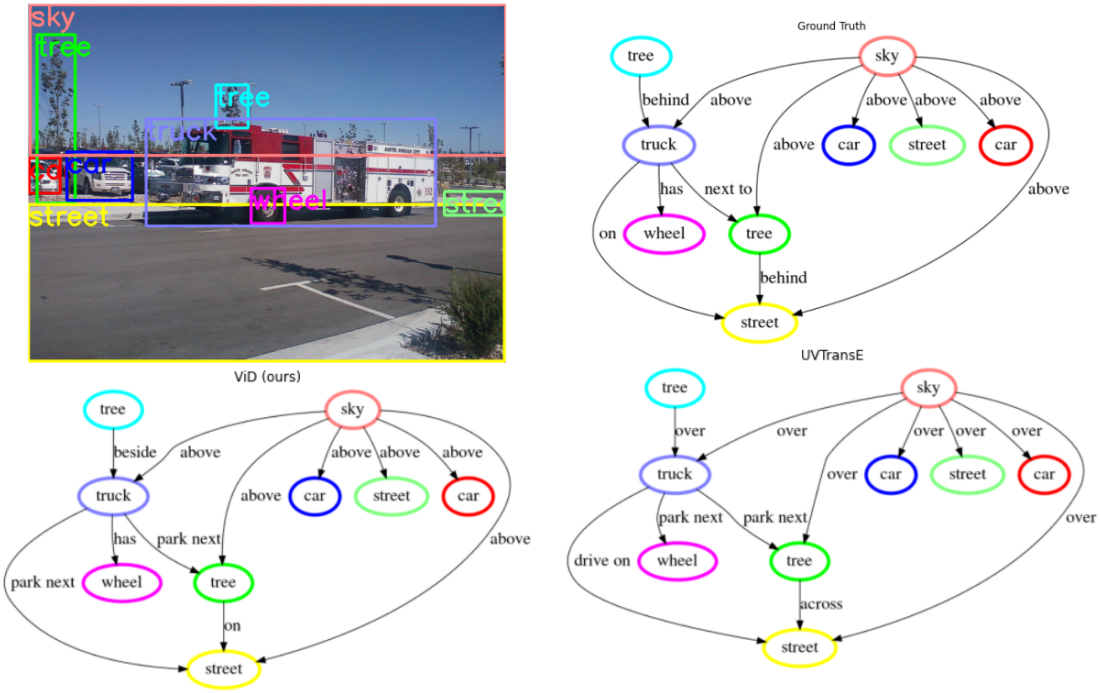
Σχήμα 5.3: *Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας έναντι του UVTransE. Η μέθοδος μας καταφέρνει να εντοπίσει το κυρίως γεγονός <person-sit on-bench> σε αντίθεση με το UVTransE που προβλέπει <person-sleep on-bench> το οποίο είναι λάθος. Ακόμη παρατηρούμε πως το δίκτυο μας έχει προβλέψει διαφορετικές σχέσεις συγκριτικά με τις επισημειωμένες όμως σημασιολογικά είναι σωστές αρκετές από αυτές, όπως <person-sit on-bench>, <hat-above-shorts>, <person-with-glasses> κ.α. ενώ οι προβλέψεις του UVTransE είναι όλες λάθος από κάθε άποψη.*



Σχήμα 5.4: *Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας έναντι του UVTransE. Η μέθοδος μας προβλέπει μόνο 1 σχέση σωστά με βάση τις επισημειωμένες σχέσεις αλλά παρατηρούμε πως οι υπόλοιπες σχέσεις είναι σημασιολογικά σωστές. Για παράδειγμα, το ViD προβλέπει <person-wear-shirt>, <person-wear-pants> ενώ οι αντίστοιχες επισημειωμένες σχέσεις είναι <person-has-shirt>, <person-has-pants>. Επίσης η επισημειωμένη σχέση για τις οντότητες grass, motorcycle είναι <grass-under-motorcycle> ενώ το δίκτυο μας πρόβλεψε <grass-below-motorcycle> που είναι σημασιολογικά σωστή. Από την άλλη μεριά, το UVTransE δεν προέβλεψε καμία σχέση σωστά ενώ παράλληλα πολλές από τις προβλεπόμενες σχέσεις δεν έχουν νόημα όπως <person-ride-pants>, <grass-park on-motorcycle>, <car-in the front of-wheel>*



Σχήμα 5.5: *Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας έναντι του UVTransE. Η μέθοδος μας προβλέπει 6 από τις 10 σχέσεις σωστά σε σχέση με τις επισημειωμένες σχέσεις. Και σε αυτή την περίπτωση οι περισσότερες από τις υπολειπόμενες σχέσεις είναι σημασιολογικά σωστές όπως οι <person-wear-hat>, <person-rest on-bike> παρόλο που έχουν επισημειωθεί ως <person-with-hat>, <person-on-bike>. Ακόμη το UVTransE προβλέπει μόνο 1 από τις 10 σχέσεις σωστά, με τις περισσότερες από αυτές να μην έχουν κανένα απολύτως νόημα όπως <car-in the front of-person>, <person-carry-bike>.*



Σχήμα 5.6: *Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας έναντι του UVTransE. Η μέθοδος μας προβλέπει σωστά 7 από τις 11 σχέσεις σε σύγκριση με τις επισημειωμένες από το σύνολο δεδομένων. Οι περισσότερες από τις υπόλοιπες 4 σχέσεις είναι σημασιολογικά σωστές όπως <truck-park next-tree>. Το UVTransE δεν προβλέπει καμία σχέση σωστά σε σύγκριση με τις επισημειωμένες και για άλλη μια φορά προβλέπει σχέσεις με κανένα νόημα όπως <truck-park next-wheel>.*

Κεφάλαιο 6

Επίλογος και μελλοντικές κατευθύνσεις

6.1 Επίλογος

Η ανίχνευση οπτικών σχέσεων είναι ένα δύσκολο πρόβλημα στην όραση υπολογιστών που απαιτεί την ανίχνευση αντικειμένων και των σχέσεών τους μέσα σε μια εικόνα. Μια από τις βασικές προκλήσεις σε αυτόν τον τομέα είναι η έλλειψη δεδομένων με ετικέτες, η οποία περιορίζει την αποτελεσματικότητα των προσεγγίσεων μάθησης με επίβλεψη. Για να ξεπεραστεί αυτή η πρόκληση, οι ερευνητές στράφηκαν πρόσφατα στην αυτοεπιβλεπόμενη μάθηση, η οποία χρησιμοποιεί μη επιβλεπόμενες τεχνικές για την εκμάθηση χαρακτηριστικών από μη επισημειωμένα δεδομένα. Ωστόσο, παρά την επιτυχία της αυτοεπιβλεπόμενης μάθησης σε άλλες εργασίες όρασης υπολογιστών, δεν έχει διερευνηθεί ευρέως στον τομέα της ανίχνευσης οπτικών σχέσεων.

Στην παρούσα διπλωματική εργασία, αναπτύξαμε ένα δίκτυο αυτοεπιβλεπόμενης μάθησης για την ανίχνευση οπτικών σχέσεων, το οποίο εκπαιδεύτηκε σε μη επισημειωμένα δεδομένα για να μάθει χαρακτηριστικά και στη συνέχεια τελειοποιήθηκε σε ελάχιστα επισημειωμένα δεδομένα. Τα αποτελέσματα των πειραμάτων μας δείχνουν ότι αυτή η προσέγγιση υπερτερεί των υφιστάμενων βασικών λύσεων τόσο στο σύνολο δεδομένων VRD όσο και στο σύνολο δεδομένων VG200 όταν αυτά εκπαιδεύονται με λίγα δείγματα ανά σχέση. Τα αποτελέσματα αυτά υποδηλώνουν ότι η αυτοεπιβλεπόμενη μάθηση είναι μια πολλά υποσχόμενη τεχνική για την ανίχνευση οπτικών σχέσεων και ότι έχει τη δυνατότητα να βελτιώσει τις επιδόσεις των ήδη υπαρχόντων προσεγγίσεων στην βιβλιογραφία αλλά και να ορίσει νέες επιδόσεις σε αυτόν τον τομέα. Η επιτυχία της προσέγγισής μας στην αυτο-επιβλεπόμενη μάθηση για την οπτική αναγνώριση συσχετίσεων οφείλεται πιθανότατα σε διάφορους παράγοντες. Αρχικά, η αυτοεπιβλεπόμενη μάθηση επέτρεψε στο μοντέλο μας να μάθει χαρακτηριστικά από ένα μεγάλο όγκο μη επισημειωμένων δεδομένων, τα οποία του παρείχαν μια πιο ολοκληρωμένη κατανόηση των οπτικών σχέσεων μεταξύ των αντικειμένων. Επίσης, η λεπτομερής τελειοποίηση σε λίγα επισημειωμένα δεδομένα επέτρεψε στο μοντέλο μας να προσαρμόσει τα χαρακτηριστικά που έμαθε κατά την διάρκεια της προ-εκπαίδευσης, στο συγκεκριμένο πρόβλημα της ανίχνευσης οπτικών σχέσεων με αποτέλεσμα την βαθύτερη και καλύτερη κατανόηση αυτών των σχέσεων.

Συμπερασματικά, η έρευνά μας δείχνει ότι η αυτο-επιβλεπόμενη μάθηση είναι μια πολλά υποσχόμενη τεχνική για την ανίχνευση οπτικών σχέσεων και ότι μπορεί να βελτιώσει σημαντικά την απόδοση στην εκπαίδευση με λίγα δείγματα τόσο στο VRD όσο και στο VG200 σύνολο

δεδομένων. Αν και υπάρχουν ακόμη πολλά που πρέπει να διερευνηθούν σε αυτόν τον τομέα, η επιτυχία της προσέγγισής μας υποδηλώνει ότι η αυτο-επιβλεπόμενη μάθηση μπορεί να είναι μια αποτελεσματική λύση για την αντιμετώπιση της πρόκλησης των περιορισμένων επισημειωμένων δεδομένων στην ανίχνευση οπτικών σχέσεων.

6.2 Μελλοντικές Επεκτάσεις

Ενώ η επιτυχία της αυτοεπιβλεπόμενης μάθησης στην ανίχνευση οπτικών σχέσεων είναι πολλά υποσχόμενη, υπάρχουν αρκετοί τομείς στους οποίους μπορεί να γίνει περαιτέρω έρευνα. Μία πιθανή κατεύθυνση για μελλοντική έρευνα είναι διερεύνηση διαφορετικών τεχνικών αυτο-επιβλεπόμενης μάθησης. Υπάρχουν πολλές διαφορετικές τεχνικές αυτο-επιβλεπόμενης μάθησης που μπορούν να χρησιμοποιηθούν για την εκμάθηση χαρακτηριστικών από μη επισημειωμένα δεδομένα όπως για παράδειγμα η κατηγοριοποίηση των εντοπισμένων αντικειμένων μέσα σε μία εικόνα αντί της ανακατασκευής τους που γίνεται στην παρούσα εργασία και θα ήταν ενδιαφέρον να διερευνήσουμε πώς αποδίδει η συγκεκριμένη τεχνική στο πλαίσιο της ανίχνευσης οπτικών σχέσεων και αν είναι πιο αποτελεσματική για αυτό το πρόβλημα.

Μία επιπλέον πιθανή μελλοντική κατεύθυνση αποτελεί η ενσωμάτωση περισσότερης σημασιολογίας στην ανίχνευση οπτικών σχέσεων. Παρόλο που το μοντέλο αυτο-επιβλεπόμενης μάθησης είναι σε θέση να μαθαίνει χαρακτηριστικά από μη επισημειωμένα δεδομένα, εξακολουθεί να βασίζεται σε μεγάλο βαθμό σε χαρακτηριστικά σε επίπεδο αντικειμένου. Η ενσωμάτωση περισσότερων σημασιολογικών πληροφοριών της εικόνας, όπως χαρακτηριστικά σε επίπεδο σκηνής, θα μπορούσε ενδεχομένως να βελτιώσει ακόμη περισσότερο την απόδοση.

Βιβλιογραφία

[1] Cewu Lu, Ranjay Krishna, Michael Bernstein and Li Fei-Fei. *Visual Relationship Detection with Language Priors. European Conference on Computer Vision*, 2016.

[2] Rowan Zellers, Mark Yatskar, Sam Thomson and Yejin Choi. *Neural Motifs: Scene Graph Parsing with Global Context. Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Hanwang Zhang, Zawlin Kyaw, Shih Fu Chang and Tat Seng Chua. *Visual Translation Embedding Network for Visual Relation Detection.* 2017.

[4] Zih Siou Hung, Arun Mallya and Svetlana Lazebnik. *Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence*, ΠΠ, 2020.

[5] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras and Petros Maragos. *Attention-Translation-Relation Network for Scalable Scene Graph Generation. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1754–1764, 2019.

[6] Jacob Devlin, Ming Wei Chang, Kenton Lee and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv*, 1810.04805, 2019.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar and Ross B. Girshick. *Masked Autoencoders Are Scalable Vision Learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021.

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

[9] Danfei Xu, Yuke Zhu, Christopher B. Choy and Li Fei-Fei. *Scene Graph Generation by Iterative Message Passing. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, 2017.

[10] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems (NIPS)*, 2015.

[11] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep Residual Learning for Image Recognition*. pages 770–778, 2016.

[13] Zhen Cui, Chunyan Xu, Wenming Zheng and Jian Yang. *Context-Dependent Diffusion Network for Visual Relationship Detection. Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[14] Kongming Liang, Yuhong Guo, Hong Chang and Xilin Chen. *Visual Relationship Detection with Deep Structural Ranking*. 2018.

[15] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao and Chen Change Loy. *Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition*. 2018.

[16] Ji Zhang, Kevin J. Shih, Andrew Tao, Bryan Catanzaro and A. Elgammal. *An Interpretable Model for Scene Graph Generation. ArXiv*, 1811.09543, 2018.

[17] Yaohui Zhu, Shuqiang Jiang and Xiangyang Li. *Visual relationship detection with object spatial distribution. 2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 379–384, 2017.

[18] Bohan Zhuang, Lingqiao Liu, Chunhua Shen and Ian Reid. *Towards Context-Aware Interaction Recognition for Visual Relationship Detection. 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 589–598, 2017.

[19] Rongjie Li, Songyang Zhang and Xuming He. *SGTR: End-to-end Scene Graph Generation with Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19486–19496, 2022.

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov and Sergey Zagoruyko. *End-to-End Object Detection with Transformers. Computer Vision - ECCV 2020*Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan Michael Frahm, , pages 213–229, Cham, 2020. Springer International Publishing.

[21] Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis and Petros Maragos. *Grounding Consistency: Distilling Spatial Common Sense for Precise Visual Relationship Detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15891–15900, 2021.

[22] Ross Girshick. *Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[23] Tomas Mikolov, Kai Chen, G.s Corrado and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR*, 2013, 2013.

[24] Mohammad Amin Sadeghi and Ali Farhadi. *Recognition using visual phrases*. *CVPR 2011*, pages 1745–1752, 2011.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*, 2013.

[26] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston and Oksana Yakhnenko. *Translating Embeddings for Modeling Multi-relational Data*. *Advances in Neural Information Processing Systems*C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, , Vol. 26. Curran Associates, Inc., 2013.

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

[28] Dzmitry Bahdanau, Kyunghyun Cho and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. *ArXiv*, 1409, 2014.

[29] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang and Xiaogang Wang. *Scene Graph Generation from Objects, Phrases and Region Captions*. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1270–1279, 2017.

[30] Bo Dai, Yuqi Zhang and Dahua Lin. *Detecting Visual Relationships with Deep Relational Networks*. pages 3298–3308, 2017.

[31] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal and Mohamed Elhoseiny. *Large-Scale Visual Relationship Understanding*. 2018.

[32] Richard Zhang, Phillip Isola and Alexei A. Efros. *Colorful Image Colorization*. *European Conference on Computer Vision*, 2016.

[33] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang and Wenzhe Shi. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. pages 105–114, 2017.

[34] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell and Alexei A. Efros. *Context Encoders: Feature Learning by Inpainting*. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

[35] Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*. *European Conference on Computer Vision*, 2016.

[36] Carl Doersch, Abhinav Kumar Gupta and Alexei A. Efros. *Unsupervised Visual Representation Learning by Context Prediction*. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei and Jifeng Dai. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. *ArXiv*, 1908.08530, 2019.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[39] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. *Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015*, 2014.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

Αγγλικοί Όροι

Bi-GRU	Bidirectional Gated Recccurrent Unit
CNNs	Convolutions Neural Networks
LSTM	Long Short Term Memory
MLP	Multi-layer Perceptron
MSE	Mean Squeared Error
PE	Positional Embedding
PhrDet	Phrase Detection
PredCls	Predicate Classification
PredDet	Predicate Detection
RNNs	Reccurrent Neural Networks
Seq2Seq	Sequence to sequence
SGCls	Scene Graph Classification
SGG	Scene Graph Generation
SOTA	State of the art
std	Standard Deviation
VG	Visual Genome
ViT	Vision Transformer
ViD	Vision Decoder
VRD	Visual Relationship Detection

Ελληνικοί Όροι

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
κ.α.	και άλλα
δηλ.	δηλαδή

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
Προσοχή	Attention
Αυτοκωδικοποιητής	Autoencoder
Περιγράμματα Περιορισμού	Bounding Boxes
Κατηγορίες	Classes
Διάνυσμα Περιεχομένου	Context Vector
Συνελικτικά Νευρωνικά δίκτυα	Convolutional Neural Network (CNN)
Σύνολο δεδομένων	Dataset
Αποκωδικοποιητής	Decoder
Επιμέρους Εργασία	Downstream Task
Κωδικοποιητής	Encoder
Εκπαίδευση με λίγα δείγματα	Few-shot learning
Τελειοποίηση	Fine-tuning
Παραγωγή Περιγραφών Εικόνων	Image Captioning
Αναπαράσταση χαμηλών διαστάσεων	Latent Representation
Γλωσσικά Χαρακτηριστικά	Linguistic Features
Μεγάλη Βραχυπρόθεσμη μνήμη	Long Short-Term Memory
Επικάλυψη	Masking
Ταξινόμηση πολλαπλών κατηγοριών	Multi-class multi-label classification
Στρώμα Προσοχής Πολλαπλών Κεφαλών	Multi-head Attention Layer
Φυσική Γλώσσα	Natural language
Ανιχνευτής Αντικειμένων	Object Detector
Περιοχές εικόνας χαμηλών διαστάσεων	Patches
Εικονοστοιχεία	Pixels
Χαρακτηριστικά Θέσης	Positional Embedding
Κωδικοποίηση Θέσης	Positional Encoding
Προ-εκπαίδευση	Pre-Training
Ανίχνευση Σχέσεων	Predicate Detection
Αναδρομικά Νευρωνικά δίκτυα	Recurrent Neural Networks (RNN)
Επιδόσεις Αιχμής	SOTA
Παραγωγή Γράφου Σκηνής	Scene Graph Generation
Αυτο-προσοχή	Self-attention
Μοντέλα με είσοδο και έξοδο σε μορφή ακολουθίας	Seq2Seq
Χωρικά Χαρακτηριστικά	Spatial Features
Μετασχηματιστές	Transformers

Οπτικός Μετασχηματιστής	Vision Transformer
Οπτικά Χαρακτηριστικά	Visual Features
Οπτική Απάντηση Ερωτήσεων	Visual Question Answering
Οπτική Αναγνώριση Συσχετίσεων	Visual Relationship Detection