ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Ανάλυση και επίλυση της σημασιολογικής μεροληψίας στον εντοπισμό οπτικών σχέσεων μέσω ημι-επιβλεπόμενων τεχνικών μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Διοματάρη Μάρκου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Diomataris Markos

Αθήνα, Νοέμβριος 2020

Εθνικο Μετσοβιο Πολυτεχνειο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και
Επεξεργασίας Σημάτων

Ανάλυση και επίλυση της σημασιολογικής μεροληψίας στον εντοπισμό οπτικών σχέσεων μέσω ημι-επιβλεπόμενων τεχνικών μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Διοματάρη Μάρκου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3η Νοεμβρίου 2020.

.....	.....	.....
Πέτρος Μαραγκός	Κωνσταντίνος Τζαφέστας	Γεράσιμος Ποταμιάνος
Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Παν/μιο Θεσσαλίας

Αθήνα, Νοέμβριος 2020.

.....

Μάρκος Διοματάρης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Το πρόβλημα αναγνώρισης οπτικών σχέσεων (visual relationship detection) της όρασης υπολογιστών αφορά τη εξαγωγή κατευθυνόμενων γράφων ως αναπαράσταση των σχέσεων (ακμές) μεταξύ των αντικειμένων (κόμβοι) σε μία εικόνα. Παρατηρώντας τη συμπεριφορά των σύγχρονων μοντέλων της βιβλιογραφίας στα μη επισημειωμένα δείγματα αποκαλύπτεται πως υπάρχουν πολλές περιπτώσεις όπου το περιεχόμενο της εικόνας αγνοείται πλήρως και χρησιμοποιείται μόνο η σημασιολογική πληροφορία των αντικειμένων για την πρόβλεψη των σχέσεων. Ονομάζουμε αυτό το πρόβλημα **context bias** και από όσο γνωρίζουμε είμαστε οι πρώτοι που το εντοπίζουν.

Η συνεισφορά αυτής της διπλωματικής αφορά τόσο την ανάλυση και πρόταση μεθόδων επίλυσης του context bias, όσο και την εισαγωγή νέων μετρικών οι οποίες, σε αντίθεση με τις υιοθετούμενες, είναι ικανές να το αναδείξουν. Συγκεκριμένα:

- Εισάγουμε το πείραμα του κυλιόμενου κουτιού (**sliding box experiment**) με το οποίο διερευνούμε ποιοτικά την επίδραση του context bias στα μοντέλα.
- Δημιουργούμε μία μέθοδο εντοπισμού κλάσεων που προκαλούν context bias μετρώντας την εντροπία της κατανομής τους στο σύνολο δεδομένων (**entropy ranking**).
- Παρουσιάζουμε ένα σύνολο κανόνων εξόρυξης αρνητικών μη επισημειωμένων δειγμάτων που ονομάζουμε αρνητική συμπλήρωση γράφου (**Negative Graph Completion** ή **NGC**).
- Η ανάλυσή μας, μας επιτρέπει να σχεδιάσουμε τις εξής τρεις μεθόδους επίλυσης του context bias:
  - Αρνητικής Εντροπίας (**NCE**): συνάρτηση κόστους αρνητικής εντροπίας που εφαρμόζεται στα αρνητικά δείγματα που παράγονται από την NGC.
  - Κατάταξης Αρνητικότητας (**NR**): αντικατάσταση της NGC με ένα δίκτυο που μαθαίνει να αξιολογεί την αρνητικότητα δειγμάτων προκαθορισμένων κλάσεων.
  - Συνέπειας Grounding (**GCL**): χρησιμοποιώντας το αντίστροφο πρόβλημα της πρόβλεψης οπτικών σχέσεων (grounding) επιβάλουμε συνέπεια ανάμεσα στην πρόβλεψη σχέσεων και την επαναπροβολή τους πίσω στην εικόνα μέσω του grounding με πλήρως ημι-επιβλεπόμενο τρόπο.
- Εισάγουμε δύο παραλλαγές μέτρησης του Precision που, χρησιμοποιώντας τα αρνητικά δείγματα που παράγονται από την NGC, είναι ικανές να αναδείξουν το πρόβλημα του context bias.
- Πραγματοποιούμε τόσο ποσοτική όσο και ποιοτική σύγκριση μεταξύ των μεθόδων που προτείνουμε αλλά και με τη σχετική βιβλιογραφία στα VRD και VG200, τα δύο δημοφιλέστερα σύνολα δεδομένων του προβλήματος όπου πετυχαίνουμε 42.2% και 54% μέγιστη σχετική βελτίωση αντίστοιχα.

Όλα τα παραπάνω αναδεικνύουν την ανάγκη χρήσης ημι-επιβλεπόμενων μεθόδων καθώς και επαναπροσδιορίζουν την μετρική του Precision ως μία αναπόσπαστη πτυχή του προβλήματος ανίχνευσης οπτικών σχέσεων, συμβάλλοντας έτσι στην περαιτέρω εμβάθυνση της μέχρι τώρα κατανόησής του.

Μεγάλο μέρος των συνεισφορών υποβλήθηκαν στο Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) με συγγραφείς τους Μάρκος Διοματάρης, Νικόλαος Γκανάτσιος, Βασίλης Πιτσικάλης και Πέτρος Μαραγκός.

Λέξεις Κλειδιά— Αναγνώριση οπτικών σχέσεων, Κοινή λογική του χώρου, Βαθιά νευρωνικά δίκτυα, ημι-επιβλεπόμενη μάθηση, Μεροληψία, Συμφραζόμενα

Abstract

Visual relationship detection is a task of computer vision that concerns extracting directed graphs as a representation of relationships (edges) between entities (nodes) in an image. Performing inference with state of the art models on unlabeled samples reveals plenty of cases where the image is completely neglected and predictions are based solely on the semantic information of the entities. We name this problem **context bias** and to our knowledge we are the first to discover it.

This thesis contributes not only to analyzing and solving context bias but also introducing new metrics which, in contrast to the most adopted ones, are able to reflect it. Specifically we:

- Introduce the **sliding box experiment** to qualitatively investigate the effect of context bias on models.
- Create a method that detects classes which generate context bias by measuring their sample distribution entropy called **entropy ranking**.
- Present a set of rules for mining negative samples that we call **Negative Graph Completion** (**NGC**).
- Our analysis enables us to design the three following methods for solving context bias:
  - Negative Cross Entropy (**NCE**): apply a negative cross entropy cost function on negative samples generated from NGC.
  - Negativity Ranking (**NR**): replace NGC with a network trained to assess the negativity of samples for a set of prespecified classes.
  - Grounding Consistency Loss (**GCL**): by using the inverse problem of visual relationship detection (grounding) we impose consistency between the predicted relation and its back-projection to the image through grounding in a fully semi-supervised manner.
- Introduce two Precision variations that, by using NGC's generated negative samples, are able to reflect the problem of context bias.
- Perform extensive quantitative and qualitative comparisons between our proposed methods and the relative literature on VRD and VG200, the two most popular datasets of the task, where we achieve 42.2% and 54% maximum relative improvement respectively.

All the aforementioned highlight the effectiveness of adopting semi-supervised solutions as well as redefine Precision as a core aspect of visual relationship detection taking a step towards improving our understanding of the task.

A big portion of our contributions was submitted to the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) with the authors being Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis and Petros Maragos.

Keywords— Visual relationship detection, Spatial common sense, Deep neural networks, semi-supervised learning, Bias, Context

Ευχαριστίες

Πρώτο από όλους θα ήθελα να ευχαριστήσω τον κύριο Πέτρο Μαραγκό, όχι μόνο για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική μου υπό την επίβλεψή του αλλά και για τον τρόπο με τον οποίο με έκανε να γνωρίσω και να αγαπήσω την όραση υπολογιστών μέσα από τα μαθήματά του.

Ισάξια, ένα μεγάλο ευχαριστώ πηγαίνει στους Βασίλη Πιτσικάλη και Νικόλαο Γκανάτσιο οι οποίοι στα πλαίσια της συνεργασίας CVSP-DeepLab συνεπίβλεψαν τη διπλωματική μου. Η συναναστροφή μαζί τους ήταν από μόνη της μία πλούσια και όμορφη εμπειρία και σίγουρα η καθοδήγησή και οι συμβουλές τους θα αποτελέσουν πολύτιμα εφόδια για τη μετέπειτα πορεία μου. Νίκο, ευχαριστώ για τις ατελείωτες ώρες που πέρασες προσφέροντας μου βοήθεια για θέματα που αφορούσαν τη διπλωματική.

Ευχαριστώ τους γονείς μου και την αδερφή μου, που με υποστήριξαν με πίστεψαν με συμβούλεψαν και με βοήθησαν σε κάθε στιγμή αυτής της προσπάθειας.

Τέλος ένα μεγάλο ευχαριστώ στους φίλους μου και την κοπέλα μου για όλες τις ωραίες αναμνήσεις που δημιουργήσαμε και θα κάνουν τα χρόνια της φοίτησής μου στη σχολή αξέχαστα.

Contents

List of Figures

List of Tables

Κεφάλαιο 1

Εισαγωγή

Άρρηκτα συνυφασμένη με την ανθρώπινη νοημοσύνη είναι η ικανότητα κωδικοποίησης της οπτικής πληροφορίας (visual information encoding), η επεξεργασία και μετατροπή δηλαδή οπτικών ερεθισμάτων σε μία πυκνή αναπαράσταση σημασιολογικά πλούσια σε πληροφορία. Ένα πρώτο βήμα αφαίρεσης αποτελεί η εξαγωγή του συνόλου των κατηγοριών των αντικειμένων που υπάρχουν στην εικόνα. Αυτό ωστόσο απέχει αρκετά από την ανθρώπινη κωδικοποίηση καθώς εικόνες που περιέχουν ίδια σύνολα αντικειμένων, ακόμη και αν οπτικά είναι όμοιες, μπορούν να απέχουν αρκετά σε νοηματικό επίπεδο μεταξύ τους. Προκειμένου να διασαφηνιστεί ο τρόπος κωδικοποίησης της οπτικής πληροφορίας, στο [42] δοκίμασαν να δείχνουν ένα σύνολο από φωτογραφίες, κάθε μία για ένα δευτερόλεπτο, σε ανθρώπους. Έπειτα, σε μια προσπάθεια να καταλάβουν τι είναι αυτό που αποτυπωνόταν στη μνήμη τους βλέποντας για τόσο λίγο χρόνο τις εικόνες, τους παρουσίασαν ένα άλλο σύνολο από εικόνες όπου μέσα υπήρχαν κάποιες που τους είχαν ήδη δείξει αλλά και καινούριες. Παρατήρησαν πως οι άνθρωποι έχουν μία εξαιρετική ικανότητα να θυμούνται αν έχουν δει μία σκηνή ή όχι και συμπέραναν πως πέρα από τα αντικείμενα των εικόνων χρησιμοποιούσαν και τη συσχέτιση αυτών ως βασική πληροφορία κωδικοποίησης.

Βασιζόμενο σε αυτές τις παρατηρήσεις, το πρόβλημα της αναγνώρισης οπτικών σχέσεων της όρασης υπολογιστών προσπαθεί να αναγνωρίσει τις σχέσεις μεταξύ των αντικειμένων που εντοπίζονται σε μία εικόνα. Τα τελευταία χρόνια, ο εντοπισμός και η αναγνώριση αντικειμένων έχουν γνωρίσει μεγάλη εξέλιξη, ειδικά με την πρόοδο των βαθιών μεθόδων και αρχιτεκτονικών μάθησης, καθιστώντας τα ικανά να υποστηρίξουν πραγματικές εφαρμογές. Δυστυχώς, όπως θα δείξουμε και στην παρούσα διπλωματική εργασία, το ίδιο δεν ισχύει για την αναγνώριση οπτικών σχέσεων η οποία ακόμη βρίσκεται σε πρώιμο στάδιο.

1.1 Περιγραφή προβλήματος

Ως σχέση ορίζεται μία τριπλέτα της μορφής `<subject-predicate-object>` (υποκείμενο-κατηγόρημα-αντικείμενο) όπου υποδηλώνει πως το υποκείμενο σχετίζεται μέσω του κατηγορήματος με το αντικείμενο. Συγκεκριμένα, για είσοδο μία εικόνα η έξοδος αντιστοιχεί σε ένα κατευθυνόμενο γράφο (βλ. εικόνα 1.1) όπου κόμβοι είναι οι οντότητες που εντοπίστηκαν ενώ οι ακμές αντιπροσωπεύουν σχέσεις που συνδέουν τις οντότητες με κατεύθυνση από το υποκείμενο στο αντικείμενο.

Τα διαθέσιμα δεδομένα εκπαίδευσης για κάθε εικόνα είναι μία λίστα με τα N οντότητες που έχουν επισημειωθεί μαζί με τις συντεταγμένες των κουτιών περιορισμού και τις κατηγορίες τους καθώς και μία λίστα από T τούπλες της μορφής `(subject id, predicate id, object id)` που ορίζουν τις ακμές του γράφου σχέσεων. Οι επισημειωμένες σχέσεις δεν καλύπτουν κάθε πιθανή ακμή του γράφου και μάλιστα ισχύει ότι $T << N^2$. Συγκεκριμένα για τα VRD [28] και VG200 [43], δύο σύνολα δεδομένων που θα χρησιμοποιήσουμε, επισημειώνεται μόλις το 13% και 3% αντίστοιχα όλων των πιθανών σχέσεων. Για περισσότερες πληροφορίες για τα σύνολα δεδομένων παραπέμπουμε στο παράρτημα A.

Figure 1.1: Εντοπισμός οπτικών σχέσεων (visual relationship detection). Για είσοδο μία εικόνα και τις ανιχνευμένες οντότητες 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 (αν αυτή θεωρείται αρκετά πιθανή).

Αναλόγως με το αν γνωρίζουμε τα κουτιά περιορισμού (bounding boxes) των οντοτήτων που έχουν εντοπιστεί, την κατηγορία τους ή τα ζευγάρια που έχουν επισημειωθεί να αλληλεπιδρούν μπορούν να οριστούν διαφορετικές παραλλαγές του προβλήματος της πρόβλεψης σχέσεων. Για το υπόλοιπο αυτής της διπλωματικής θα θεωρούμε και τις τρεις αυτές πληροφορίες γνωστές λύνοντας έτσι την παραλλαγή που ονομάζεται ανίχνευση σχέσεων (Predicate Detection ή PredDet). Πρακτικά δηλαδή, θεωρούμε δεδομένες τις επισημειωμένες οντότητες (κουτί και κατηγορία) και ασχολούμαστε μόνο με την αναγνώριση της μεταξύ τους σχέσης. Η επιλογή αυτή γίνεται προκειμένου να αποπλέξουμε το πρόβλημα που θα διερευνηθεί στη συνέχεια από τις παραμέτρους και τους περιορισμούς ενός δικτύου εντοπισμού και αναγνώρισης οντοτήτων (object detector). Για αναλυτικές πληροφορίες των παραλλαγών του προβλήματος παραπέμπουμε στο παράρτημα B.

Συνήθης μετρική στην κατηγοριοποίηση σχέσεων αποτελεί το Recall@K (R@K). Έστω λοιπόν ότι προβλέπουμε μία σχέση για κάθε πιθανό ζευγάρι αντικειμένων ($N \times (N-1)$ προβλέψεις) και τα κατατάσσουμε σε φθίνουσα σειρά με βάση την πιθανότητα πρόβλεψής τους. Το R@K προκύπτει μετρώντας το ποσοστό των επισημειωμένων σχέσεων που έχουν προβλεφθεί σωστά και βρίσκονται στις K καλύτερες σχέσεις. Παραπέμπουμε στο παράρτημα C για επιπλέον πληροφορίες σχετικά με τις μετρικές.

1.2 Εφαρμογές

Ο εντοπισμός σχέσεων σε εικόνες αποτελεί ένα πρόβλημα μέσης αφαίρεσης όπως δείχνουμε και στην εικόνα 1.2. Αυτό σημαίνει πως από τη μία εξάγει πληροφορία υψηλότερου επιπέδου από βασικά προβλήματα της όρασης υπολογιστών όπως ο εντοπισμός αντικειμένων (object detection) η κατηγοριοποίηση αντικειμένων (object classification) ή η κατηγοριοποίηση δράσεων (action recognition). Από την άλλη οι γράφοι που παράγονται μπορούν να χρησιμοποιηθούν ως ενδιάμεση αναπαράσταση της εικόνας και να αποτελέσουν είσοδο σε προβλήματα υψηλότερου επιπέδου όπως η περιγραφή εικόνας (captioning) ή η απάντηση οπτικών ερωτήσεων (visual question answering). Αυτή η ενδιάμεση θέση όσον αφορά το βαθμό αφαίρεσης της πληροφορίας καθιστά το πρόβλημα ιδιαίτερα χρήσιμο για πολλές εφαρμογές όπου απαιτείται κατανόηση μιας σκηνής όπως για παράδειγμα η σημασιολογική αναζήτηση ή συμπίεση εικόνων, εντοπισμός δράσεων σε βίντεο, αυτόνομη πλοήγηση κ.ά..

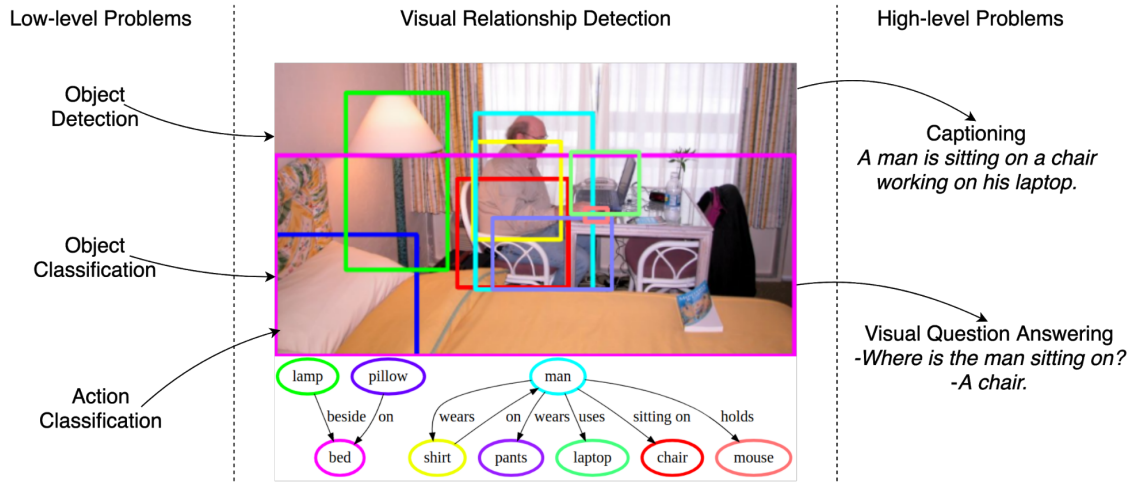

Figure 1.2: Θέση του προβλήματος εντοπισμού οπτικών σχέσεων σε σχέση με άλλα προβλήματα. Ενώ δεν παρέχει πληροφορία χαμηλού αφαιρετικού επιπέδου, μπορεί να αποτελέσει ενδιάμεση αναπαράσταση μίας εικόνας για την επίλυση προβλημάτων υψηλότερης αφαίρεσης. Αυτό το καθιστά ένα από τα πλέον σημαντικά προβλήματα της όρασης υπολογιστών.

1.3 Προκλήσεις

Όπως τα περισσότερα προβλήματα της όρασης υπολογιστών, έτσι και η αναγνώριση σχέσεων σε εικόνες μας φέρνει αντιμέτωπους με ένα σύνολο από προκλήσεις στη προσπάθεια σχεδιασμού μίας καλής λύσης.

Συνδυαστική φύση Μία σημαντική πρόκληση, εγγενής στη συνδυαστική φύση του προβλήματος, είναι η εκθετική σχέση των πιθανών σχέσεων και των οντοτήτων. Για O κλάσεις οντοτήτων και P κλάσεις σχέσεων, ο αριθμός όλων των πιθανών συνδυασμών της μορφής `<subject-predicate-object>` είναι PO^2. Δηλαδή σε ένα σύνολο δεδομένων όπως το VRD όπου $O = 100$ και $P = 70$ έχουμε $700,000$ πιθανές τριπλέτες εκ των οποίων μόνο $6,672$ έχουν επισημειώσεις. Αυτό σημαίνει πως από τη μία αν αντιμετωπίσουμε ως κλάση κάθε τριπλέτα θα έχουμε έναν μεγάλο αριθμό κλάσεων με πολύ λίγα (ή και καθόλου) δείγματα η κάθε μία, ενώ αν κάθε κλάση αντιστοιχεί σε μία σχέση η κατανομή κάθε κλάσης θα γίνει αρκετά περίπλοκη καθώς τριπλέτες με τελείως διαφορετικά χαρακτηριστικά θα ανήκουν στην ίδια κλάση (π.χ. `<person-ride-bike>` και `<dog-ride-horse>`). Προκειμένου οι λύσεις να είναι κλιμακώσιμες, στη πλειοψηφία της βιβλιογραφίας (καθώς και στην παρούσα διπλωματική) υιοθετείται η δεύτερη προσέγγιση. Αυτό σημαίνει πως 15% των δειγμάτων στο σύνολο δοκιμής (test set) του VRD και 4% του VG200 αποτελούν τριπλέτες που δεν υπάρχουν στο σύνολο εκπαίδευσης (train set).

Οπτική λεπτομέρεια Ακόμη μία δυσκολία είναι η εξάρτηση των σχέσεων από μικρές οπτικές λεπτομέρειες. Πολύ μικρές αλλαγές της διάταξης ή των χαρακτηριστικών των αντικειμένων μπορεί να επιφέρουν μεγάλη αλλαγή στον γράφο αναπαράστασης. Για παράδειγμα στην εικόνα 1.3 για να καταλάβουμε ότι στην αριστερή φωτογραφία δεν είναι ο αστυνομικός που κρατάει την κάμερα, αλλά ο επισημειωμένος άνθρωπος, θα πρέπει να δώσουμε σημασία σε οπτικές λεπτομέρειες (π.χ. την πόζα των δύο ανθρώπων ή το λουρί της κάμερας στον επισημειωμένο άνθρωπο). Στην δεξιά εικόνα επίσης κατανοούμε ότι η γυναίκα κρατάει μια κάμερα και πως μάλλον τη χρησιμοποιεί από τη στάση του σώματός της αλλά και από το γεγονός ότι φαίνεται να βγάζει φωτογραφία μία άλλη γυναίκα.

Πολυσημία Ίσως από τις σημαντικότερες προκλήσεις αποτελεί η ύπαρξη σχέσεων με πολλαπλά νοήματα. Για παράδειγμα η σχέση "has" θα μπορούσε να χρησιμοποιηθεί ως `<person-has-shoes>` με την έννοια δηλαδή του ότι φοράω κάτι ή στην τριπλέτα `<tower-has-clock>` με την έννοια του ότι κάτι είναι προσκολλημένο κάπου. Ακόμη ένα παράδειγμα φαίνεται στην εικόνα 1.4. Η κλάση "on" έχει πάνω από ένα νοήματα και μάλιστα όχι μόνο χωρικά αφού για παράδειγμα χρησιμοποιείται και στην τριπλέτα `<jeans-on-person>` όταν ένα ρούχο φοριέται από κάποιον. Αυτή η πολυσημία αυξάνει

Figure 1.3: Πολύ μικρές οπτικές λεπτομέρειες κρίνουν το ποια αντικείμενα και με ποιον τρόπο αλληλεπιδρούν.

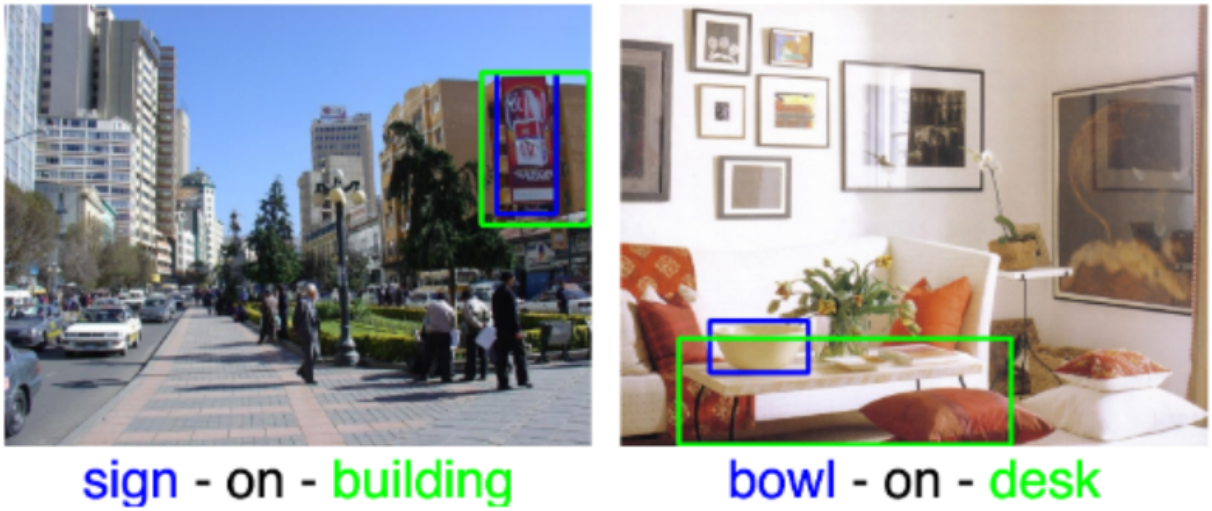

Figure 1.4: Μία από τις δυσκολίες του εντοπισμού οπτικών σχέσεων είναι η αμφισημία των κατηγορημάτων. Για παράδειγμα, παρουσιάζουμε δύο από τα νοήματα της σχέσης "on". Αριστερά χρησιμοποιείται με την έννοια της προσκόλλησης ενώ δεξιά με την έννοια ενός υποκειμένου που βρίσκεται πάνω στο αντικείμενο.

ακόμη περισσότερο την ποικιλία των δεδομένων κάνοντας δυσκολότερη την κατηγοριοποίηση των σχέσεων.

Συνωνυμία Ενώ τα προηγούμενα προβλήματα προέρχονται από την ίδια τη φύση του προβλήματος και της φυσικής γλώσσας, υπάρχει ένα πρόβλημα που έχει να κάνει με τον τρόπο κατασκευής των συνόλων δεδομένων και πρόκειται για την ύπαρξη συνώνυμων κλάσεων. Υπάρχουν διαφορετικές κλάσεις που μπορεί να σημαίνουν ακριβώς το ίδιο όπως για παράδειγμα οι "next to" και "near" ή "below" και "under". Ακόμη μπορεί μία κλάση να εμπεριέχει νοηματικά κάποια άλλη όπως για παράδειγμα το "next to" το "on the right of". Καθώς σχεδόν ποτέ δεν επισημειώνονται για ένα ζευγάρι οντοτήτων όλες οι νοηματικά έγγυρες σχέσεις, το πρόβλημα του εντοπισμού οπτικών σχέσεων δεν αντιμετωπίζεται ως κατηγοριοποίηση πολλαπλών κλάσεων (multi-class & multi-label classification) και έτσι δημιουργούνται συνθήκες ανταγωνισμού μεταξύ συνωνύμων οι οποίες εξορισμού δεν γίνεται να λυθούν. Εντύπωση προκαλεί στο [14] η διαπίστωση πως αν ενώσουμε τις συνώνυμες τριπλέτες η μετρική του R@50 πηγαίνει από 58.48% στο 82.76% στο VRD.

Μεροληπτική επισημείωση Τέλος, κάτι το οποίο θα μας απασχολήσει ιδιαίτερα στη συνέχεια είναι το πρόβλημα της μεροληπτικής επισημείωσης των δεδομένων. Σαν να μην ήταν αρκετή η έλλειψη

δεδομένων, οι επισημειώσεις που υπάρχουν ακολουθούν την ανθρώπινη κοινή λογική η οποία καθορίζει τη σημαντική από την ασήμαντη πληροφορία. Αυτό οδηγεί σε μοντέλα με φτωχή κατανόηση των σχέσεων που απλώς προσπαθούν να απομνημονεύσουν τα στατιστικά των δεδομένων χωρίς καμία ικανότητα γενίκευσης.

1.4 Κίνητρο & Συνεισφορές

1.4.1 Κίνητρο

Κινητήριος δύναμη αυτής της ερευνητικής προσπάθειας είναι η παρατήρηση ότι σε όλα τα μοντέλα, ανεξαρτήτως αρχιτεκτονικής και επίδοσης, παρατηρείται ότι σε αρκετές περιπτώσεις κάνουν εντελώς παράλογες προβλέψεις όπως για παράδειγμα ότι ένας άνθρωπος "φοράει" όλα τα παπούτσια που υπάρχουν σε μία εικόνα ή ένα φορτηγό "έχει" τη ρόδα ενός άλλου φορτηγού. Το ενδιαφέρον είναι ότι αυτή η συμπεριφορά μπορεί να παρατηρηθεί μόνο όταν κάνουμε προβλέψεις σε μη επισημειωμένα δείγματα και μάλιστα ακόμη και για κλάσεις με πολύ μεγάλο πλήθος δειγμάτων. Για

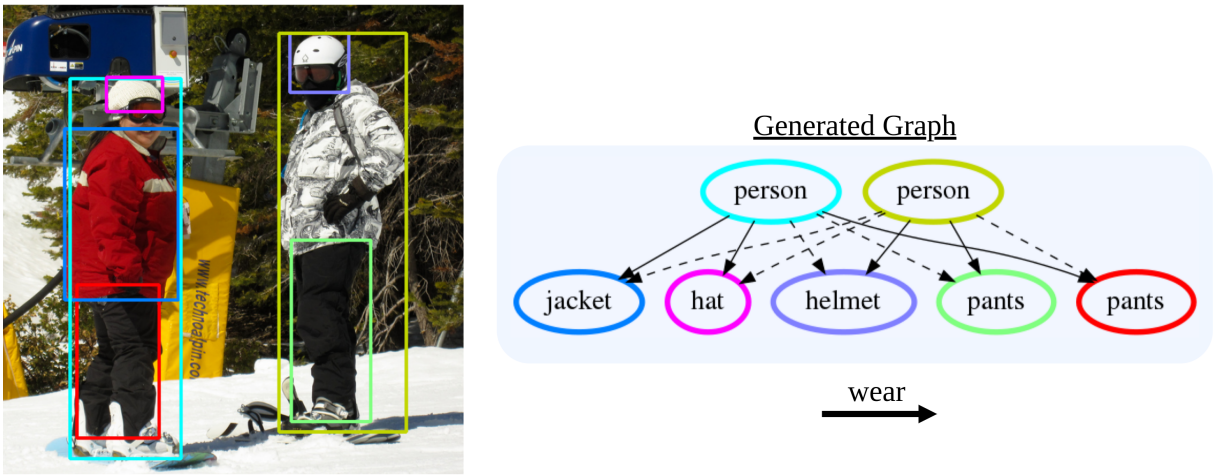

Figure 1.5: Ακόμη και τα πιο σύγχρονα μοντέλα ([29]) προβλέπουν πως ένας άνθρωπος "φοράει" την μπλούζα, το παντελόνι και το καπέλο του διπλανού του. Αυτή η συμπεριφορά υποδηλώνει πως υπάρχει σημαντικό πρόβλημα στους μέχρι τώρα τρόπους εκπαίδευσης μοντέλων στη βιβλιογραφία. Τα διακεκομμένα βέλη αντιπροσωπεύουν ζευγάρια οντοτήτων για τα οποία δεν υπάρχει επισημειωμένη σχέση.

παράδειγμα, στην εικόνα 1.5 δείχνουμε τον γράφο που προβλέπει το μοντέλο του [29] το οποίο ,τη στιγμή που γράφεται αυτή η διπλωματική, πετυχαίνει την καλύτερη επίδοση σε Recall στη βιβλιογραφία. Παρατηρούμε πως υπάρχουν αρκετές παράλογες προβλέψεις, όπως ότι ο ένας άνθρωπος φοράει τα ρούχα ή το καπέλο του άλλου. Ενδιαφέρον έχει το γεγονός ότι αν αγνοήσουμε τελείως τα μη επισημειωμένα ζευγάρια (διακεκομμένα βέλη) θα παρατηρήσουμε πως ο γράφος φαίνεται απόλυτα φυσιολογικός. Λόγω αυτού, μέχρι στιγμής αυτό το πρόβλημα λάνθανε της προσοχής των ερευνητών καθώς η πιο διαδεδομένη μετρική του Recall δεν δίνει ποινή για λάθη σε μη επισημειωμένα δείγματα. Αιτία αυτού του προβλήματος όπως θα δούμε και στην ανάλυση που θα ακολουθήσει μετέπειτα στο 3 είναι η μεροληψία της διαδικασίας επισημείωσης των δεδομένων ως προς το είδος των αντικειμένων. Για παράδειγμα υπάρχουν ελάχιστες επισημειώσεις ανάμεσα στις οντότητες `person-jacket` που να μην είναι "wear" με αποτέλεσμα τα μοντέλα να νομίζουν πως ένα ζευγάρι `person-jacket` θα συνδέεται πάντα με τη σχέση "wear" (βλ. εικόνα 1.5). Για αυτόν τον λόγο ονομάζουμε το συγκεκριμένο πρόβλημα `context bias` και από όσο γνωρίζουμε είμαστε οι πρώτοι που το εντοπίζουν. Παρά λοιπόν το γεγονός ότι η κλάση "wear" έχει πολύ μεγάλο αριθμό δειγμάτων, τα μοντέλα απέχουν πολύ από το να κατανοήσουν τη σημασία της.

1.4.2 Συνεισφορές

Ορμώμενοι από την παρατήρηση που αναφέραμε στο 1.4.1, με αυτήν την διπλωματική εργασία συμβάλλουμε με σημαντικό τρόπο τόσο στην ανάλυση όσο και στην αντιμετώπιση του προβλήματος του context bias υποστηρίζοντας τα συμπεράσματά μας τόσο ποσοτικά όσο και ποιοτικά με τις παρακάτω συνεισφορές:

- Εισάγουμε το πείραμα του κυλιόμενου κουτιού (**sliding box experiment**) με το οποίο διερευνούμε ποιοτικά την επίδραση του context bias στα μοντέλα.
- Δημιουργούμε μία μέθοδο εντοπισμού κλάσεων που προκαλούν context bias μετρώντας την εντροπία της κατανομής τους στο σύνολο δεδομένων (**entropy ranking**).
- Παρουσιάζουμε ένα σύνολο κανόνων εξόρυξης αρνητικών μη επισημειωμένων δειγμάτων που ονομάζουμε αρνητική συμπλήρωση γράφου (**Negative Graph Completion** ή **NGC**).
- Η ανάλυσή μας, μας επιτρέπει να σχεδιάσουμε τις εξής τρεις μεθόδους επίλυσης του context bias:
  - Αρνητικής Εντροπίας (**NCE**): συνάρτηση κόστους αρνητικής εντροπίας που εφαρμόζεται στα αρνητικά δείγματα που παράγονται από την NGC.
  - Κατάταξης Αρνητικότητας (**NR**): αντικατάσταση της NGC με ένα δίκτυο που μαθαίνει να αξιολογεί την αρνητικότητα δειγμάτων προκαθορισμένων κλάσεων.
  - Συνέπειας Grounding (**GCL**): χρησιμοποιώντας το αντίστροφο πρόβλημα της πρόβλεψης οπτικών σχέσεων (grounding) επιβάλουμε συνέπεια ανάμεσα στην πρόβλεψη σχέσεων και την επαναπροβολή τους πίσω στην εικόνα μέσω του grounding με πλήρως ημι-επιβλεπόμενο τρόπο.
- Εισάγουμε δύο παραλλαγές μέτρησης του Precision που, χρησιμοποιώντας τα αρνητικά δείγματα που παράγονται από την NGC, είναι ικανές να αναδείξουν το πρόβλημα του context bias.
- Πραγματοποιούμε τόσο ποσοτική όσο και ποιοτική σύγκριση μεταξύ των μεθόδων που προτείνουμε αλλά και με τη σχετική βιβλιογραφία στα VRD και VG200, τα δύο δημοφιλέστερα σύνολα δεδομένων του προβλήματος όπου πετυχαίνουμε 42.2% και 54% μέγιστη σχετική βελτίωση αντίστοιχα.

Καταφέρνουμε λοιπόν να θεμελιώσουμε τα αίτια και την επίδραση του προβλήματος αλλά και μέσω αυτής της ανάλυσης να σχεδιάσουμε λύσεις και μετρικές που επαναπροσδιορίζουν τις σημαντικές πτυχές του προβλήματος πρόβλεψης σχέσεων τόσο στον τρόπο επίλυσής του αλλά και στον τρόπο αξιολόγησης των μοντέλων.

Κεφάλαιο 2

Θεματική ανασκόπηση της βιβλιογραφίας

2.1 Εντοπισμός και αναγνώριση αντικειμένων

Καθώς αναφερόμαστε στην αναγνώριση οπτικών σχέσεων μεταξύ αντικειμένων προϋπόθεση φυσικά αποτελεί ο εντοπισμός και η αναγνώριση της κατηγορίας τους. Όπως αναφέραμε και πιο πάνω, καθώς στην παρούσα διπλωματική θα ασχοληθούμε κυρίως με την παραλλαγή του προβλήματος που ονομάζεται Predicate Detection (PredDet) δεν θα χρειαστεί να συμπεριλάβουμε ως αρχικό στάδιο των μοντέλων μας κάποιον object detector. Ωστόσο αξίζει να αναφέρουμε πως στην περίπτωση άλλων παραλλαγών του προβλήματος όπου ένας object detector είναι απαραίτητος, συνήθως χρησιμοποιείται Faster R-CNN [35] με ραχοκοκαλιά VGG-16 [36] ή ResNet [37] για τον εντοπισμό καθώς και κάποια απλή δομή (π.χ. αλληλουχία γραμμικών επιπέδων με παρεμβαλλόμενες Relu) για την κατηγοριοποίηση αντικειμένων. Κατά κανόνα υιοθετούνται οι εξής δύο προσεγγίσεις: από κοινού εκπαίδευση του object classifier με το δίκτυο ανίχνευσης σχέσεων ([27]) ή πάγωμα των παραμέτρων του object classifier και εκπαίδευση μόνο του δικτύου ανίχνευσης σχέσεων ([55, 12]). Η από κοινού εκπαίδευση προσπαθεί να συνδυάσει τα προβλήματα της αναγνώρισης αντικειμένων και σχέσεων ώστε το ένα να επωφεληθεί από το άλλο (multi-task learning) και σύμφωνα με τα [8, 48] κάτι τέτοιο πράγματι συμβαίνει. Ωστόσο για λόγους απλοποίησης και απομόνωσης του προβλήματος τις περισσότερες φορές ο object-classifier παραμένει παγωμένος κατά τη διάρκεια της εκπαίδευσης.

2.2 Είδη πληροφορίας

Πριν περάσουμε στην θεματική κατηγοριοποίηση της βιβλιογραφίας αξίζει να αναφερθούμε στον κοινό παράγοντα των περισσότερων μοντέλων που θα παρατεθούν παρακάτω, ο οποίος είναι η πληροφορία που χρησιμοποιούν για την κατηγοριοποίηση σχέσεων. Οι τρεις βασικές πηγές πληροφορίας διαθέσιμες είναι:

- Οπτικές (**V**isual): Χρησιμοποιείται κάποιο δίκτυο ως ραχοκοκαλιά (VGG-16 [36] ή ResNet [37]) για την εξαγωγή χαρτών αναπαράστασης (feature maps). Μέσω μίας προεκπαιδευμένης κεφαλής εντοπισμού αντικειμένων λαμβάνουμε το διάνυσμα που παράγεται ένα στάδιο πριν την πρόβλεψη του αντικειμένου.

- Χωρικές (**S**patial): μέσω των κουτιών περιορισμού (bounding boxes) εξάγουμε κανονικοποιημένες μετρικές που αφορούν τη χωρική σχέση και διάταξη υποκειμένου-αντικειμένου. Μπορούν ακόμη να χρησιμοποιηθούν οι δυαδικές τους μάσκες όπως για παράδειγμα στο [12].

- Γλωσσικές (**L**inguistic): Οι κατηγορίες του υποκειμένου και του αντικειμένου κωδικοποιούνται μέσω *word2vec* σε διανύσματα ενός σημασιολογικού χώρου όπου γεωμετρικές σχέσεις

αντιστοιχούν σε νοηματικές.

Έτσι έχουμε πληροφορία για τα οπτικά χαρακτηριστικά, τη χωρική διάταξη και τη σημασιολογία του υποκειμένου και του αντικειμένου.

Στη συνέχεια θα παρουσιάσουμε μία ανασκόπηση ενός μεγάλου εύρους της βιβλιογραφίας κατηγοριοποιημένη ανά πεδίο.

2.3 Αρχιτεκτονικές μετατόπισης

Από τις πρώτες αξιοσημείωτες αρχιτεκτονικές στο πρόβλημα του εντοπισμού οπτικών σχέσεων σε εικόνες αποτελεί το [52] εισάγοντας το μοντέλο VTransE (Visual Translation Embeddings). Εμπνεόμενοι από το [3] όπου επιχειρούν πως διαισθητικά, σε ένα δέντρο αναπαράστασης ιεραρχικής γνώσης οι διάφορες σχέσεις αντιστοιχούν σε κατακόρυφες και οριζόντιες μετατοπίσεις, εφαρμόζουν αυτή τη λογική για οπτικές σχέσεις αντικειμένων όμως σε περισσότερες διαστάσεις (βλ. εικόνα 2.1. Αν δηλαδή S, O είναι οι αναπαραστάσεις του υποκειμένου και του αντικειμένου αντίστοιχα, τότε θα

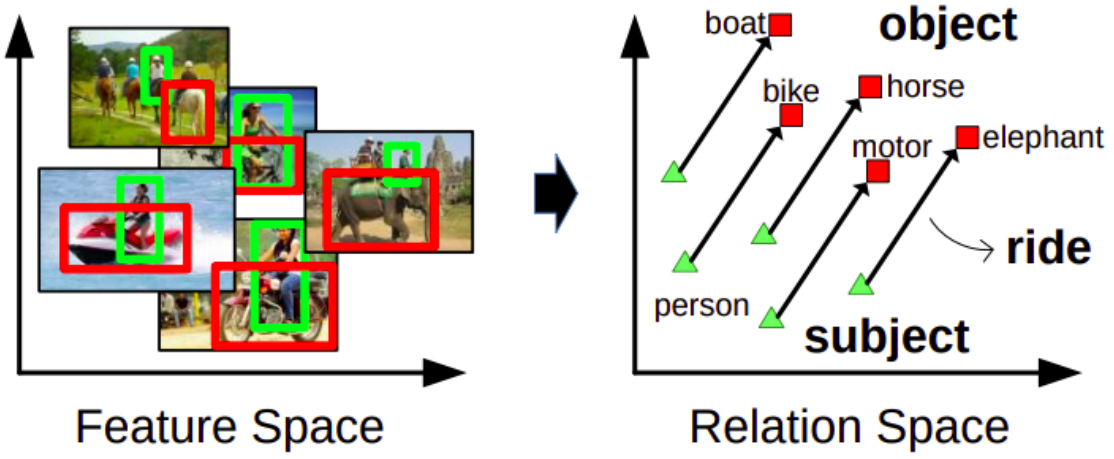

Figure 2.1: Σχέσεις αναπαρίστανται ως μετατοπίσεις σε έναν χώρο προβολής των αντικειμένων. Εικόνα από [52].

πρέπει να ισχύει $S+P \approx O$ ή ισοδύναμα $S-O \approx P$. Σχεδιάζουν λοιπόν ένα δίκτυο όπου προσπαθεί να μάθει να προβάλλει τα S, O με τρόπου που για ίδιες σχέσεις, θα έχουν ίδια διανυσματική απόσταση. Ως απάντηση στο [52] έρχεται το [19] το οποίο προσπαθεί να αυξήσει την ικανότητα γενίκευσης του VTransE εισάγοντας μια αρχιτεκτονική που ονομάζουν UVTransE (Union VTransE). Ισχυρίζονται πως το $S+P$, δηλαδή τα οπτικά χαρακτηριστικά του υποκειμένου και η σχέση του με το αντικείμενο, καθορίζουν μονοσήμαντα τα χαρακτηριστικά του αντικειμένου O. Προτείνουν λοιπόν τη χρήση των χαρακτηριστικών U του κουτιού που περικλείει το υποκείμενο και το αντικείμενο (union box). Έτσι, μία σχέση καθώς πρέπει να είναι ανεξάρτητη της κατηγορίας των αντικειμένων που περιλαμβάνει θα πρέπει να είναι $P \approx U-(S+O)$. Προτείνουν επίσης τη χρήσης μίας μονάδας GRU η οποία θα δέχεται ως είσοδο γλωσσική και οπτική πληροφορία για τα αντικείμενα και θα συμβάλει ανεξάρτητα στην τελική κατηγοριοποίηση. Η αρχιτεκτονική τους φαίνεται στην εικόνα 2.2. Στην ίδια οικογένεια ανήκει και το [13] με το μοντέλο MATransE (Multimodal Attentional Translation Embeddings). Εδώ υιοθετείται μία αρχιτεκτονική δύο κλάδων (βλ. εικόνα 2.3) όπου το δίκτυο από τη μία προβάλει τη διαφορά $S-O$ και από την άλλη τα χαρακτηριστικά όλης της περιοχής της σχέσης P. Επιπλέον της τελικής επίβλεψης, υλοποιούν και βαθιά επίβλεψη (deep supervision) στα $S-O$ και P παίρνοντας καλά αποτελέσματα.

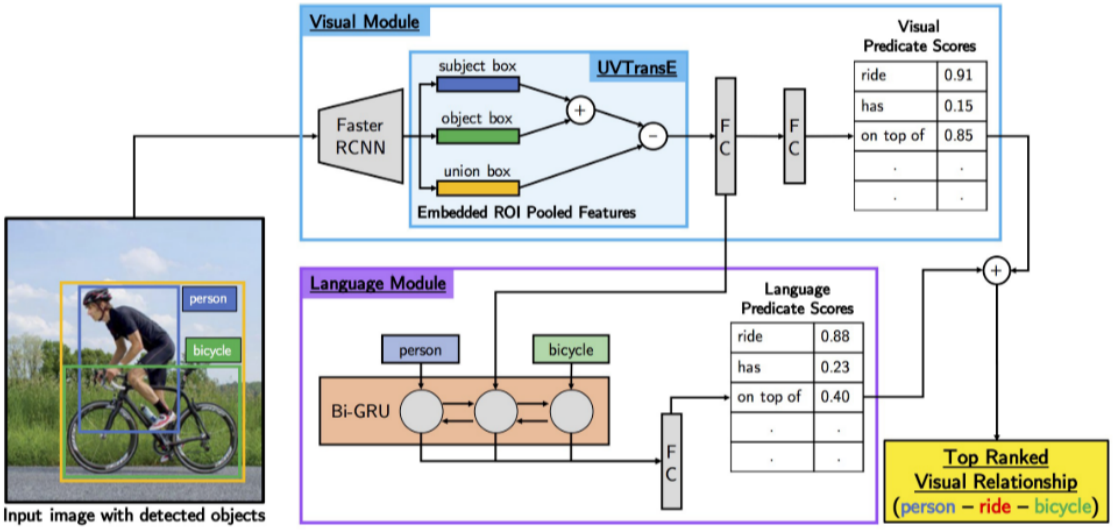

Figure 2.2: Συνολική αρχιτεκτονική του UVTransE. Τα χαρακτηριστικά των αντικειμένων αφαιρούνται από τη συνολική εμφάνιση της σχέσης ώστε να απομείνει μόνο πληροφορία που αφορά την αλληλεπίδρασή τους. Μία μονάδα GRU συνδυάζει γλωσσική με οπτική πληροφορία για να συμβάλει ανεξάρτητα στο τελικό αποτέλεσμα. Εικόνα από [19].

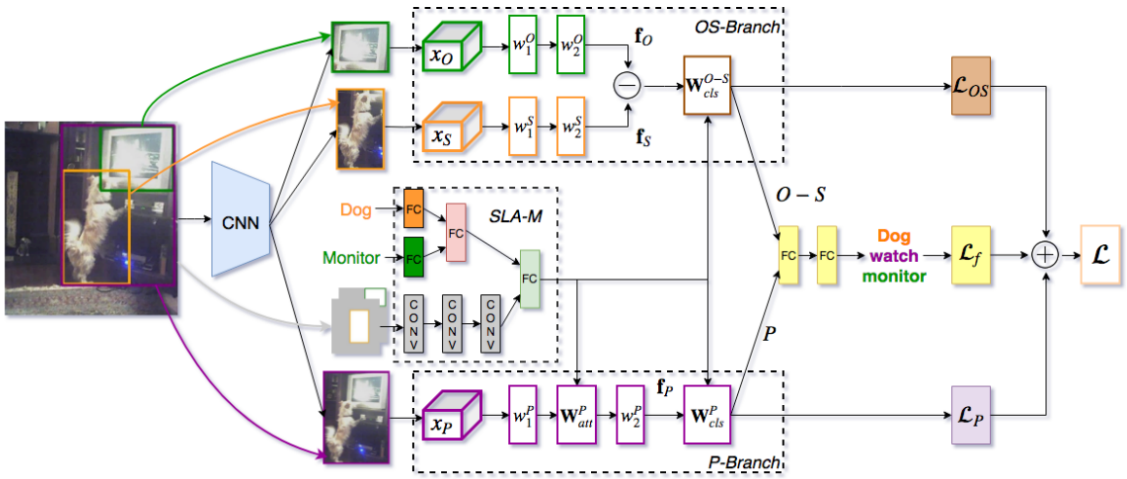

Figure 2.3: Αρχιτεκτονική δύο κλάδων του MATransE. Ο ένας κλάδος μαθαίνει τη διανυσματική μετατόπιση των χαρακτηριστικών ανάμεσα στο υποκείμενο και το αντικείμενο ενώ ο άλλος προσπαθεί να προβλέψει τη σχέση από τα χαρακτηριστικά του κουτιού ένωσης (union). Εικόνα από [13].

2.4 Μηχανισμοί προσοχής

Σημαντική προσοχή έχει τραβήξει επίσης ο σχεδιασμός μεθόδων προσοχής (attention mechanisms). Γενικά, προσοχή είναι ένας ευρέως χρησιμοποιούμενος μηχανισμός σε μία πληθώρα πεδίων και εφαρμογών όπως για παράδειγμα η περιγραφή εικόνων (captioning) στο [44]. Σκοπός του είναι, δεδομένου ότι υπό διαφορετικές περιπτώσεις η σημασία και η συμβολή της πληροφορίας αλλάζει, να μπορέσει ένα δίκτυο να μάθει να εστιάζει μέσω κάποιων βαρών στη χρήσιμη πληροφορία. Στο πρόβλημα της ανίχνευσης οπτικών σχέσεων κάποιοι υλοποιούν προσοχή οδηγούμενη από γλωσσικά χαρακτηριστικά [56], συνδυασμό σημασιολογικών και χωρικών [13] ή και προσοχής πολλαπλών κεφαλών (multi-head attention) [12] όπου για διαφορετικές κλάσεις μαθαίνεται διαφορετική προσοχή. Συγκεκριμένα, στο [56] εφαρμόζουν προσοχή στον οπτικό χάρτη πληροφορίας (feature map) που εξάγουν από το VGG-16. Η προσοχή αυτή δημιουργείται με βάση τη γλωσσική πληροφορία του υποκειμένου και του αντικειμένου. Στο [13] χρησιμοποιούν συνδυασμό πληροφορίας από τις δυαδικές μάσκες των αντικειμένων και της γλωσσικής πληροφορίας τους για να οδηγήσουν την προσοχή στα οπτικά χαρακτηριστικά (βλ. εικόνα 2.3). Όμοια είναι η μέθοδος του [12] με την εξαίρεση ότι κάθε κλάση έχει τις δικές της παραμέτρους που μαθαίνουν ανεξάρτητα την προσοχή που χρειάζεται (multi-head attention).

2.5 Γλωσσικά οδηγούμενοι ταξινομητές

Η χρήση γλωσσικής πληροφορίας σε συνδυασμό με την οπτική έχει αποτελέσει κοινή καθώς και προσοδοφόρα πρακτική σε πολλά προβλήματα της όρασης υπολογιστών. Καθώς η γλώσσα μπορεί να εκφράσει σε ένα βαθμό τη δομή της ανθρώπινης σκέψης, η αντιστοίχιση και συσχέτισή της με την οπτική πληροφορία μπορεί να συνεισφέρει σε μία σημασιολογικά πλούσια αναπαράσταση των αντικειμένων. Για παράδειγμα στο [4] χρησιμοποιούν τη συμπληρωματική φύση μεταξύ βίντεο και κειμένου για την αναγνώριση δράσεων και προσώπων.

Παρά το γεγονός ότι στο πρόβλημά μας ασχολούμαστε με στατικές εικόνες, αυτή η δομή της γλώσσας εξακολουθεί να υπάρχει. Ενδιαφέρον παρουσιάζουν αρχιτεκτονικές όπου οι τελικοί ταξινομητές των σχέσεων δεν είναι σταθεροί αλλά κατασκευάζονται με δυναμικό τρόπο. Στο [56]

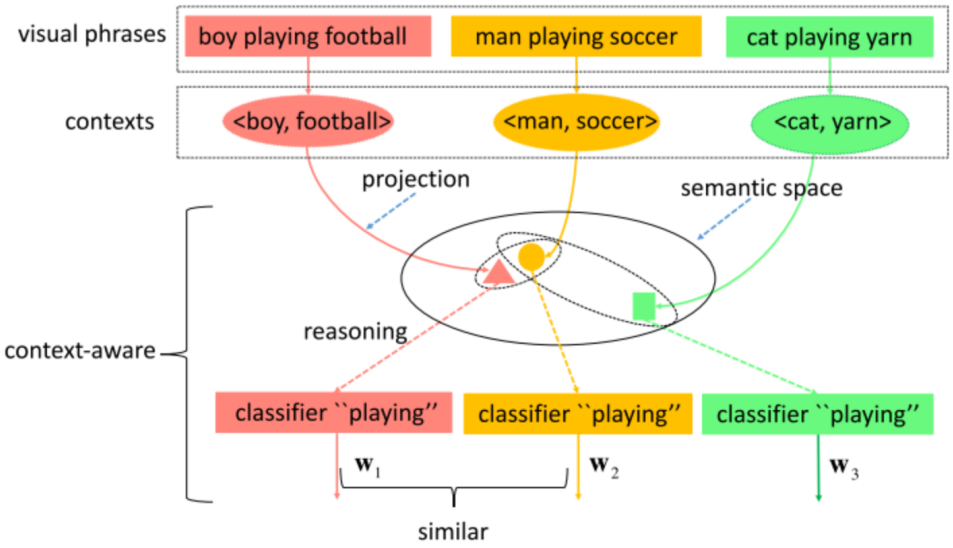

Figure 2.4: Ο ταξινομητής μίας κλάσης δημιουργείται από συνδυασμό της γλωσσικής πληροφορίας των αντικειμένων. Εικόνα από [56].

για πρώτη φορά, οι ταξινομητές προκύπτουν ως άθροισμα ενός όρου ανεξάρτητου της γλωσσικής πληροφορίας, και ενός εξαρτημένου από το υποκείμενο και το αντικείμενο (βλ. εικόνα 2.4). Έτσι ο

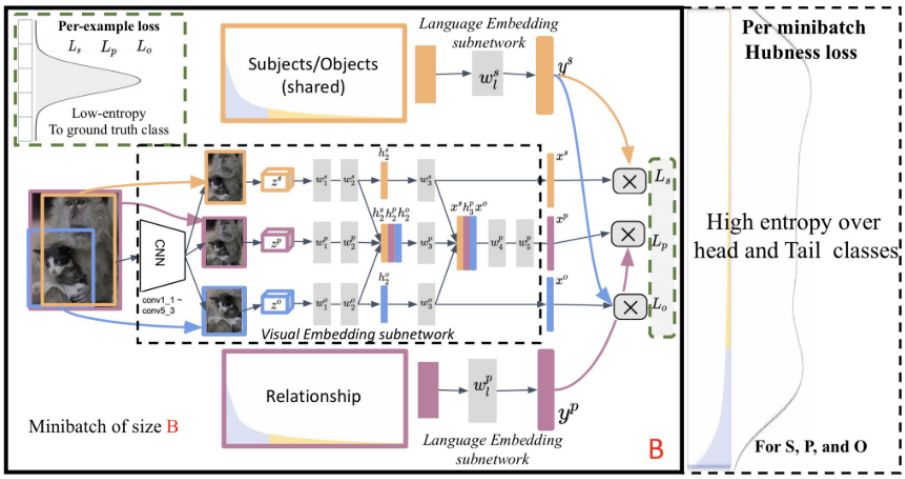

Figure 2.5: Η αρχιτεκτονική του [54] μαζί με τη μέθοδο ελαχιστοποίησης εντροπίας του [1].

ταξινομητής κάθε σχέσης συνδιαμορφώνεται από κάποια σταθερή συνιστώσα που είναι ανεξάρτητη των αντικειμένων και μία δυναμική που εξαρτάται από το υποκείμενο και το αντικείμενο. Στο [54] το υποκείμενο, το κατηγόρημα και το αντικείμενο καθώς και οι αντίστοιχες σημασιολογικές τους αναπαραστάσεις, προβάλλονται στον ίδιο χώρο και απαιτείται να έχουν κοντινή απόσταση (με την έννοια της απόστασης συνημιτόνου). Για να γίνει πρόβλεψη σχέσης, προβάλλονται σημασιολογικά όλες οι σχέσεις και επιλέγεται ο κοντινότερος γείτονας των αναπαραστάσεων των οπτικών διανυσμάτων. Τέλος, στο [14] οι ταξινομητές δημιουργούνται από μία μονάδα GRU όπου ως είσοδος της δίνεται η τριπλέτα `<subject-predicate-object>` και προσπαθεί να ευθυγραμμίσει τα οπτικά χαρακτηριστικά, με τον ταξινομητή.

2.6 Long-tail κατανομή

Μία άλλη μεγάλη συστάδα ερευνών, επικεντρώνεται στο πρόβλημα της ανισορροπίας των δεδομένων (long-tail distribution) εφαρμόζοντας τεχνικές μάθησης με λίγα δείγματα (low-shot learning) [6, 40, 31, 10] ή κατασκευής εικόνων [41, 9, 22]. Όπως έχει φανεί, υπάρχει μεγάλη μεροληψία στον τρόπο επισημείωσης των σχέσεων το οποίο εν μέρει οφείλεται στη συνδυαστική φύση του προβλήματος. Στο [1] προσπαθούν να αυξήσουν την ικανότητα γενίκευσης προκειμένου κλάσεις με λίγα δείγματα να μαθαίνονται καλύτερα (βλ. εικόνα 2.5). Παίρνοντας την αρχιτεκτονική του [54] επιβάλουν η κατανομή πρόβλεψης να έχει υψηλή εντροπία. Ακόμη, στο [27] ακολουθώντας παρόμοιο σκεπτικό χρησιμοποιούν τα στατιστικά του συνόλου δεδομένων ώστε να εξομαλύνουν τον σταθερό όρο του τελικού ταξινομητή (bias). Στο ίδιο πνεύμα, στο [38] κάνουν το διαχωρισμό του bias σε "καλό" και "κακό". Αναφέρουν πως το πρώτο είδος βοηθά στην γενίκευση ενώ το δεύτερο οδηγεί σε απομνημόνευση. Υποστηρίζουν πως οι κλάσεις με λίγα δείγματα αποτελούν ειδικότερες έννοιες των κυρίαρχων κλάσεων όπως για παράδειγμα τα "in the front of", "behind" είναι ειδικές περιπτώσεις του "near" (βλ. εικόνα 2.6).

2.7 Συνολικά συμφραζόμενα & Διαβίβαση μηνυμάτων (Global context & Message passing)

Ακόμη μία σημαντική ιδέα για το πρόβλημα της ανίχνευσης σχέσεων είναι η χρήση παγκόσμιας πληροφορίας (global contex). Σύμφωνα με αυτή, για την πρόβλεψη μίας σχέσης ενός ζευγαριού αντικειμένων συμμετέχουν όλα τα αντικείμενα της εικόνας. Για παράδειγμα, ο εντοπισμός ενός ανθρώπου που κάνει ποδήλατο ίσως βοηθηθεί από την ύπαρξη δρόμου ή αυτοκινήτων στην εικόνα.

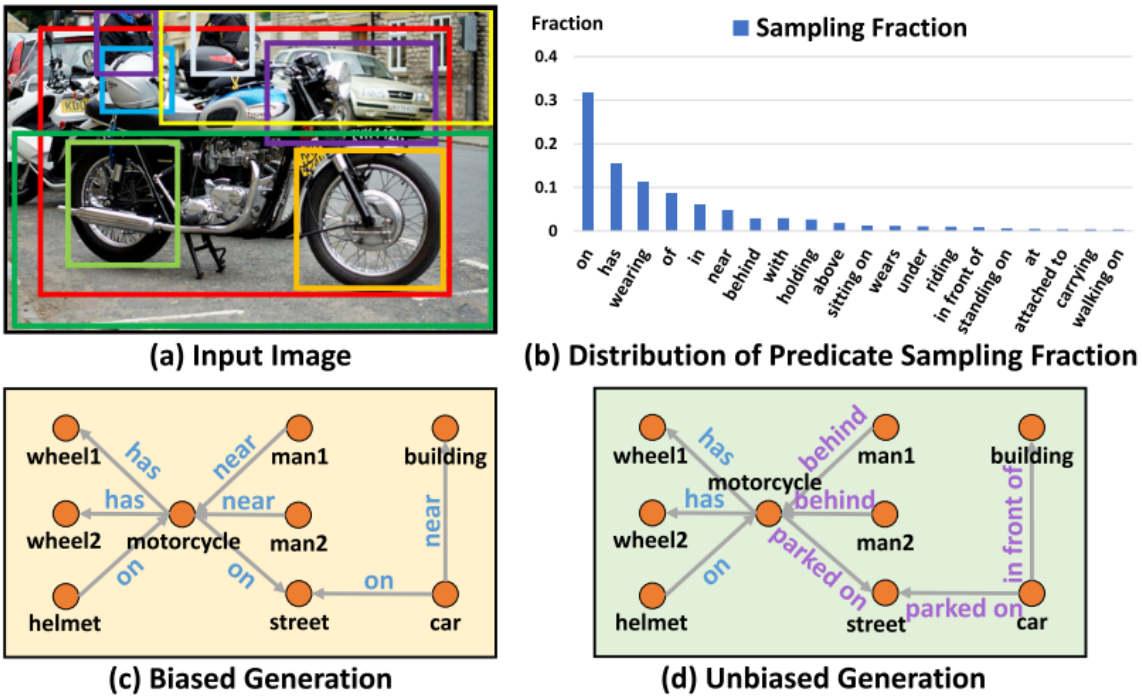

Figure 2.6: Εικόνα του [38] δείχνει τη μετάβαση από ένα γράφο με γενικές σχέσεις, σε σχέσεις με ειδικότερα νοήματα οι οποίες όμως έχουν πολύ λιγότερα δείγματα.

Σε αυτή τη λογική κινούνται δουλειές όπως οι [51, 27, 29, 25, 48]. Στο [51] χρησιμοποιούν μία σειρά από LSTM δύο κατευθύνσεων τα οποία δέχονται ως είσοδο τα αντικείμενα της εικόνας και συνδυάζουν πληροφορίες από όλη την εικόνα για την κατηγοριοποίηση κάθε σχέσης όπως φαίνεται στην εικόνα 2.7. Επίσης στα [27, 29, 25, 48] υιοθετούν μία μέθοδο μεταβίβασης μηνυμάτων (message passing) όπου θεωρείται κάθε αντικείμενο ως κόμβος σε ένα γράφο και επαναληπτικά μηνύματα με τη μορφή διανυσμάτων ανταλλάσσονται μεταξύ των κόμβων ώστε να δημιουργηθούν αναπαραστάσεις των αντικειμένων που συνδυάζουν πληροφορία από όλη την εικόνα.

2.8 Χρησιμοποιώντας μη επισημειωμένη πληροφορία

Κατά την εκπαίδευση ενός μοντέλου, σημαντική απόφαση αποτελεί το πως θα χρησιμοποιηθούν τα μη επισημειωμένα δείγματα, τα οποία αποτελούν και τη συντριπτική πλειοψηφία των δειγμάτων. Η πιο διαδεδομένη πρακτική είναι όλα τα δείγματα που δεν έχουν επισημείωση να αντιμετωπίζονται ως αρνητικά και να μην ανήκουν σε καμία κλάση [19, 56, 13]. Όμως, το ότι ένα δείγμα δεν έχει επισημείωση δε συνεπάγεται ότι δεν ανήκει σε καμία κλάση με αποτέλεσμα θεωρώντας το ως αρνητικό να δημιουργήσει σύγχυση στο μοντέλο. Με αυτό το σκεπτικό στα [8, 46] μαθαίνουν μέσω χωρικών και σημασιολογικών πληροφοριών να φιλτράρουν τα ασυσχέτιστα από τα συσχετιζόμενα ζευγάρια. Άλλη προσέγγιση είναι να θεωρήσουμε μία επιπλέον κλάση που θα ονομάζεται "background" και θα αντιπροσωπεύει όλα τα μη επισημειωμένα ζευγάρια ([43, 33]) η οποία έχει επίσης το πρόβλημα που αναφέραμε παραπάνω. Στα [5, 51] θεωρούν ως μη συσχετιζόμενα αντικείμενα εκείνα που τα κουτιά περιορισμού τους έχουν μηδενική τομή. Μια συνολικότερη λύση προτείνεται στο [12] (βλ. εικόνα 2.8) όπου για κάθε ακμή του γράφου λύνουν δύο προβλήματα. Το ένα είναι η κατηγοριοποίησή της (χωρίς να υπάρχει κλάση ασυσχέτιστων αντικειμένων) ενώ το άλλο είναι ένα δυαδικό πρόβλημα που αποφασίζει το αν το συγκεκριμένο ζευγάρι έχει συσχέτιση. Στο τελικό στάδιο πολλαπλασιάζουν αυτές τις δύο πιθανότητες και προκύπτει η τελική κατάταξη των προβλέψεων.

Καθώς όπως αναφέραμε τα μη επισημειωμένα δεδομένα περιέχουν χρήσιμη πληροφορία, έχουν

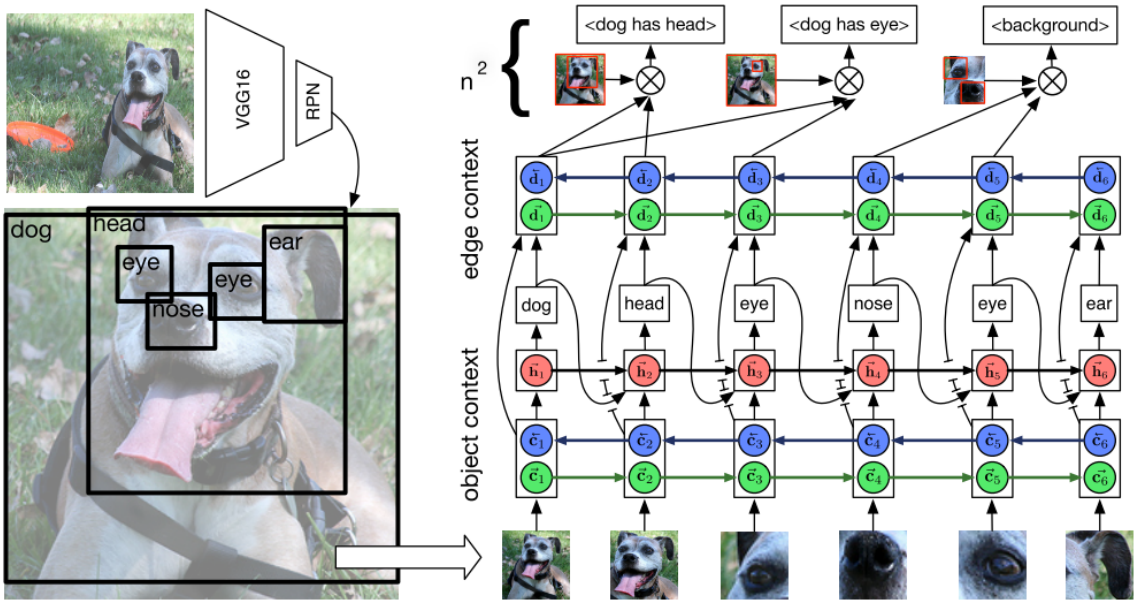

Figure 2.7: Εικόνα του [51]. Μία σειρά από bi-directional LSTMs συνδυάζουν πληροφορία από αντικείμενα όλης της εικόνας για την περιγραφή κάθε σχέσης.

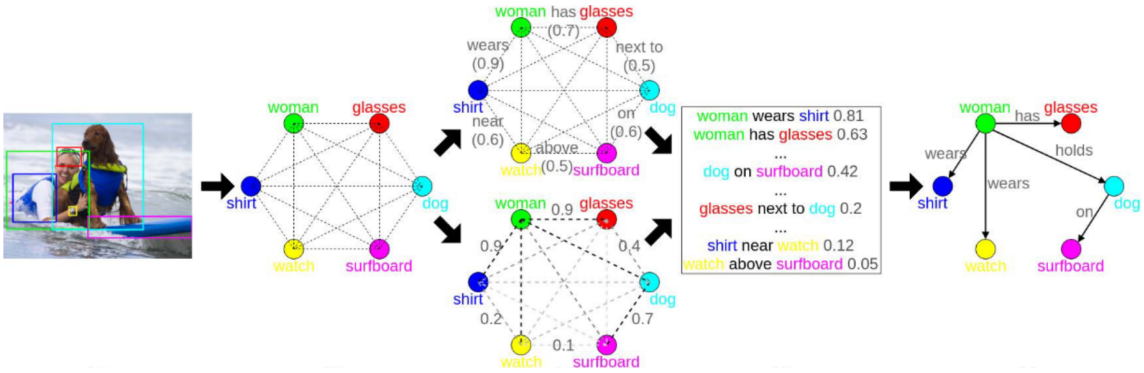

Figure 2.8: Εικόνα του [12]. Τα προβλήματα της κατηγοριοποίησης σχέσης και συσχέτισης δύο αντικειμένων λύνονται ξεχωριστά για κάθε ακμή. Η τελική πιθανότητα μίας κλάσης είναι το γινόμενο της πιθανότητάς της με την πιθανότητα συσχέτισης των αντικειμένων.

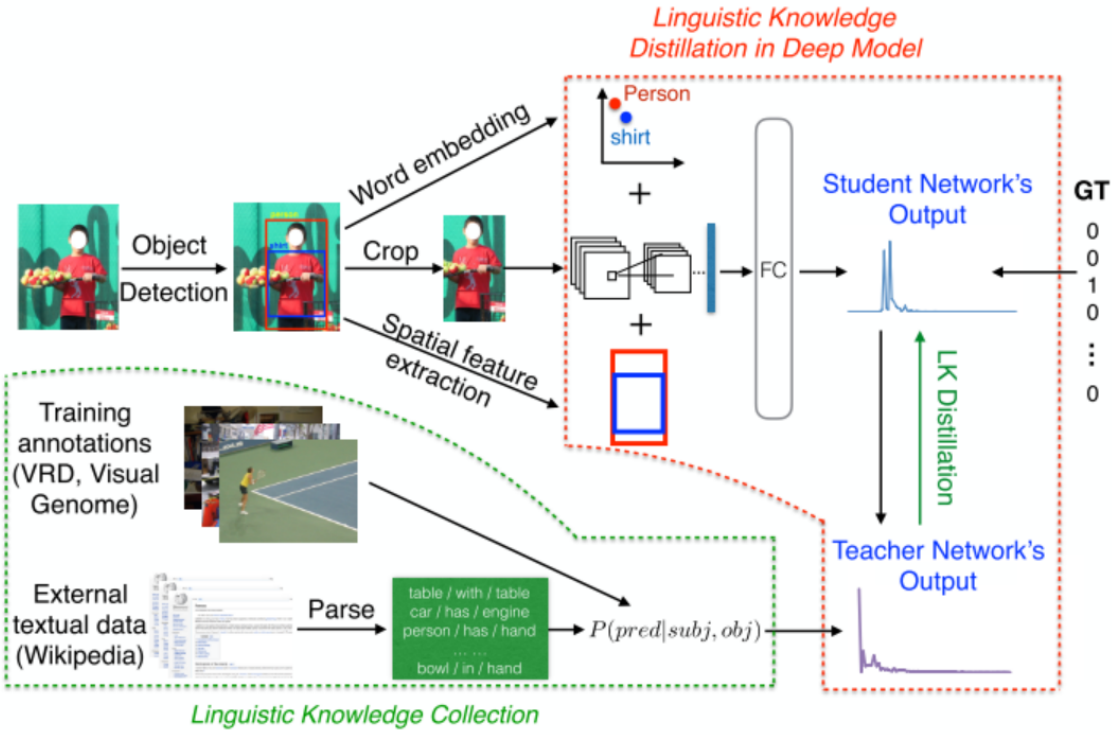

Figure 2.9: Με χρήση εξωτερικής γνώσης από κείμενο το δίκτυο-μαθητής κανονικοποιείται μέσω της KL απόκλισης ενώ παράλληλα προσπαθεί να προβλέψει τις σωστές κλάσεις. Εικόνα από [49].

πολλές φορές υιοθετηθεί ημι-επιβλεπόμενες μέθοδοι μάθησης. Η ημι-επιβλεπόμενη μάθηση συνδυάζει την εκμετάλλευση ενός σχετικά μικρού συνόλου επισημειωμένων και ενός μεγάλου συνόλου μη επισημειωμένων δεδομένων για την εκπαίδευση ενός μοντέλου. Στον εντοπισμό οπτικών σχέσεων έχουν γίνει προσπάθειες για την καλύτερη εκμετάλλευσή τους [30, 53, 11, 50, 2]. Το [6] αποτελεί μέθοδο ημι-επιβλεπόμενης (semi-supervised) μάθησης όπου μέσω ευριστικών μετρικών δημιουργούν επισημειώσεις για τα μη επισημειωμένα δείγματα. Σε παρόμοια λογική, τα [49, 32] εκτιμούν μέσω του εκπαιδευόμενου μοντέλου ή πρότερων σημασιολογικών στατιστικών επισημειώσεις που χρησιμοποιούνται κατά την εκπαίδευση συμβουλευτικά.

2.9 Μεταβίβαση γνώσης

Προκειμένου να αυξηθεί η ικανότητα κατανόησης και γενίκευσης των μοντέλων μία διαδεδομένη μέθοδος που εφαρμόζεται και στον εντοπισμό οπτικών σχέσεων είναι η μεταβίβαση γνώσης (knowledge distillation). Η γνώση μπορεί να μεταβιβάζεται από μία βάση πληροφοριών ή ακόμη και από ένα άλλο μοντέλο το οποίο ονομάζεται δάσκαλος. Όπως αναφέρεται και στο [16], το εκπαιδευόμενο δίκτυο-μαθητής προσπαθεί ταυτόχρονα να μειώσει το κόστος του προβλήματος που λύνει αλλά και να δίνει ως έξοδο κατανομές που πλησιάζουν (π.χ. με την έννοια της KL απόκλισης) αυτές του δασκάλου. Ακόμη, στο [18] περιγράφουν μία μέθοδο όπου γνώση από κανόνες σε συνδυασμό με μη επισημειωμένα δείγματα χρησιμοποιούνται ως κανονικοποιητές του εκπαιδευόμενου δικτύου. Τα κέρδη που μπορούμε να έχουμε σε τέτοιες περιπτώσεις είναι γρηγορότερη σύγκλιση του δικτύου-μαθητή, χρήση λιγότερων παραμέτρων και καλύτερη ικανότητα γενίκευσης.

Στο πρόβλημά μας χρησιμοποιείται μεταβίβαση γνώσης κυρίως με σκοπό την αύξηση της ικανότητας γενίκευσης των μοντέλων. Στα [32, 49] χρησιμοποιούν ως δάσκαλο εξωτερική γνώση από κείμενο σε μία προσπάθεια να εκτιμήσουν την $Pr(predicate|subject, object)$ (βλ. εικόνα 2.9) και μέσω της KL απόκλισης κανονικοποιούν την έξοδο πρόβλεψης του δικτύου εντοπισμού σχέσεων. Ενδιαφέρον έχει η παρατήρηση στο [49] όπου ισχυρίζεται πως είναι πιο προσοδοφόρα η ενσωμάτωση

εξωτερικής γνώσης μέσω ενός δασκάλου παρά απλώς η επαύξηση των δεδομένων εκπαίδευσης. Ακόμη, στο [14] προκειμένου να μάθει το δίκτυο να κάνει προβλέψεις όπου οι προβλέψεις με υψηλή πιθανότητα για ένα δείγμα θα είναι συνώνυμες, συνδυάζουν στατιστικά του συνόλου δεδομένων με ένα γλωσσικό μοντέλα και τα χρησιμοποιούν ως δάσκαλο.

2.10 Κοντά σε εμάς

Επιπλέον των παραπάνω κατηγοριοποιήσεων της βιβλιογραφίας, θεωρούμε χρήσιμο να αναφερθούμε πιο αναλυτικά σε κάποιες δουλειές που είτε έρχονται θεματικά κοντά στην παρούσα διπλωματική, είτε κάποια από τα συμπεράσματά τους θα φανούν χρήσιμα στη μετέπειτα ανάλυση.

2.10.1 Γλωσσική μεροληψία

Όπως διαισθητικά μπορούμε να καταλάβουμε, κάθε μία από αυτές τις πηγές πληροφορίας που αναφέραμε στο 2.2 συμβάλει σε διαφορετικό βαθμό στην αναγνώριση διαφορετικών σχέσεων. Στο [14] οι συγγραφείς διερευνούν την επίδοση μοντέλων που μεμονωμένα ή συνδυαστικά χρησιμοποιούν τις S, V, L πηγές πληροφορίας. Συγκεκριμένα, κατασκευάζουν τα εξής τέσσερα πειραματικά μοντέλα:

- Spatial (S): Μοντέλο που χρησιμοποιεί μόνο χωρική πληροφορία.
- Visual (V): Μοντέλο που χρησιμοποιεί μόνο οπτική πληροφορία.
- Language (L): Μοντέλο που χρησιμοποιεί μόνο γλωσσική πληροφορία.
- Visual-Spatial (V-S): Μοντέλο που χρησιμοποιεί συνδυασμό οπτικής και χωρικής πληροφορίας.
- Language-Spatial (L-S): Μοντέλο που χρησιμοποιεί συνδυασμό γλωσσικής και χωρικής πληροφορίας.

Ακόμη τα συγκρίνουν με state of the art μοντέλα όπως το ATR-Net [12] στο VRD και το VG200. Στον πίνακα 2.1 παραθέτουμε ένα μέρος των αποτελεσμάτων τους.

Model	VRD	VG200
S	47.58	52.6
V	52.59	64.55
L	54.28	68.64
V-S	54.55	-
L-S	<u>57.56</u>	<u>69.67</u>
ATR-Net	**58.48**	**70.18**

Table 2.1: Αποτελέσματα από [14] για R@50 σε τέσσερα πειραματικά μοντέλα στα VRD και VG200.

Έκπληξη προκαλεί το γεγονός ότι ένα μοντέλο που δεν "βλέπει" την εικόνα πετυχαίνει επίδοση πολύ κοντά σε ένα state of the art μοντέλο. Η κατανομή των συνόλων δεδομένων φαίνεται να είναι τέτοια που μόνο και μόνο με χρήση γλωσσικής και χωρικής πληροφορίας επιτυγχάνεται πολύ καλή επίδοση. Επιπλέον αυτού, το μοντέλο L καταφέρνει απρόσμενα υψηλό R@50. Συγκεκριμένα, το L πετυχαίνει 71% R@50 στην κλάση "above", 64% στην "under" και 100% στην κλάση "wear". Αυτές οι κλάσεις (ειδικά οι δύο πρώτες) έχουν μία αναπόσπαστη χωρική νοηματική συνιστώσα η οποία φαίνεται να μην είναι πλήρως απαραίτητη για υψηλά επίπεδα R@50. Φανερώνεται λοιπόν ότι ορισμένες κλάσεις έχουν κάποιο μοτίβο επισημείωσης το οποίο εξαρτάται από τα αντικείμενα που

εμπλέκονται. Η γλωσσική πληροφορία είναι χρήσιμη επειδή λόγω του ότι σημασιολογικά κοντινές κλάσεις αντικειμένων έχουν και κοντινές *word2vec* αναπαραστάσεις ένα μοντέλο μπορεί να γενικεύσει τη γνώση για την τριπλέτα `<person-ride-horse>` στην `<man-ride-horse>`. Από την άλλη όμως συμπεραίνουν πως αν και απαραίτητη μπορεί να οδηγήσει σε στείρα απομνημόνευση των δεδομένων αποτρέποντας τα μοντέλα από το να γενικεύουν σε νέους συνδυασμούς αντικειμένων και σχέσεων που δεν έχουν ξαναδεί. Το γιατί, πότε και κατά πόσο ένα μοντέλο στηρίζεται στη γλωσσική πληροφορία είναι κάτι που θα μας απασχολήσει έντονα στα επόμενα κεφάλαια.

Figure 2.10: Λόγω εγγύτητας το μοντέλο δεν μπορεί να ξεχωρίσει ποιος άνθρωπος κρατάει την κάμερα οπότε προβλέπει πως και οι δύο την κρατούν.Αριστερά βλέπουμε τη λανθασμένη πρόβλεψη ενώ δεξιά τη πρόβλεψη εφαρμόζοντας τη μέθοδο των [55]. Εικόνα από [55].

2.10.2 Αποσαφήνιση γράφου

Ένα πρόβλημα παρεμφερές με αυτό που θα εξετάσουμε στο κεφάλαιο 3 αντιμετωπίζουν και στο [55]. Παρατήρησαν πως πολλές φορές τα μοντέλα, όταν υπάρχουν σε εγγύτητα δύο ή παραπάνω αντικείμενα της ίδιας κλάσης, αδυνατούν να ξεχωρίσουν ποιο ακριβώς είναι αυτό που συμμετέχει στη σχέση που προβλέπεται. Για παράδειγμα στην εικόνα 2.10 φαίνεται αριστερά πως το μοντέλο δεν χρησιμοποιεί την οπτική, ή ακόμα και τη χωρική, πληροφορία που του παρέχεται για να διαχωρίσει το υποκείμενο που κρατάει την κάμερα. Προσπαθώντας να λύσουν αυτό το πρόβλημα, δημιουργούν τρεις κανόνες με τους οποίους εντοπίζουν παραδείγματα τα οποία δημιουργούν αυτή τη σύγχυση και τα χρησιμοποιούν ως αρνητικά. Αυτό όμως στο οποίο υστερούν είναι η παραδοχή ότι δείγματα χωρίς επισημείωση δεν ανήκουν σε καμία κλάση κάτι που σε καμία περίπτωση δεν ισχύει. Όπως θα φανεί και στη συνέχεια η ανάλυση αυτής της διπλωματικής διαφοροποιείται αρκετά από το [55] καθώς εμβαθύνουμε περισσότερο στο πρόβλημα αποκαλύπτοντας τη γενικότερη αιτία του αλλά και το προσεγγίζουμε με αποτελεσματικότερο τρόπο όπως θα δείξουμε στη σύγκριση στο κεφάλαιο 5.3.1.

Κεφάλαιο 3

Context Bias

Οι προτιμώμενες μετρικές (Recall) καθώς και όπως είδαμε στο κεφάλαιο 2.8 οι περισσότεροι τρόποι εκπαίδευσης αντιμετωπίζουν τα μη επισημειωμένα δείγματα ως ασυσχέτιστα ζεύγη αντικειμένων. Αγνοούν τις προβλέψεις του δικτύου για αυτά τα δείγματα και το μόνο που απαιτούν είναι να κατατάσσονται χαμηλότερα, με βάση την πιθανοφάνειά τους, από τα επισημειωμένα. Έτσι, η συμπεριφορά των μοντέλων παραμένει ανεξερεύνητη για ένα πολύ μεγάλο ποσοστό δειγμάτων (87% στο VRD και 97% στο VG). Φυσικά, η τάση αποφυγής διερεύνησης των μη επισημειωμένων δειγμάτων είναι εν μέρει δικαιολογημένη καθώς δεν υπάρχει καμία πληροφορία για αυτά. Παρόλα αυτά μπορούμε διαισθητικά να υποθέσουμε ότι είναι πολύ λίγες οι περιπτώσεις όπου δύο αντικείμενα είναι πραγματικά ασυσχέτιστα και πως στην πλειοψηφία των περιπτώσεων ακόμη και μία γενική σχέση (π.χ. "near") θα μπορούσε να περιγράψει τον τρόπο αλληλεπίδρασής τους. Ακολουθώντας αυτήν την αντιμετώπιση, η μόνη γνώση που μπορεί να αποκτήσει ένα δίκτυο από τα μη επισημειωμένα δείγματα είναι η κατανομή που εκφράζει το κατά πόσο δύο αντικείμενα αλληλεπιδρούν, συμφωνα πάντα με τους επισημειωτές (reporting bias [39]).

Βάζοντας ακόμη και σύγχρονα (state of the art) μοντέλα να κάνουν προβλέψεις σε μη επισημειωμένα δείγματα μπορούμε αμέσως να καταλάβουμε ότι η μετρική του Recall δημιουργεί την "ψευδαίσθηση" πως αναπτύσσεται υψηλός βαθμός κατανόησης για κλάσεις με πολλά δείγματα, κάτι το οποίο δεν ισχύει. Όπως παρατηρούμε και στην εικόνα 3.1 ακόμη και ένα s.o.t.a. μοντέλο

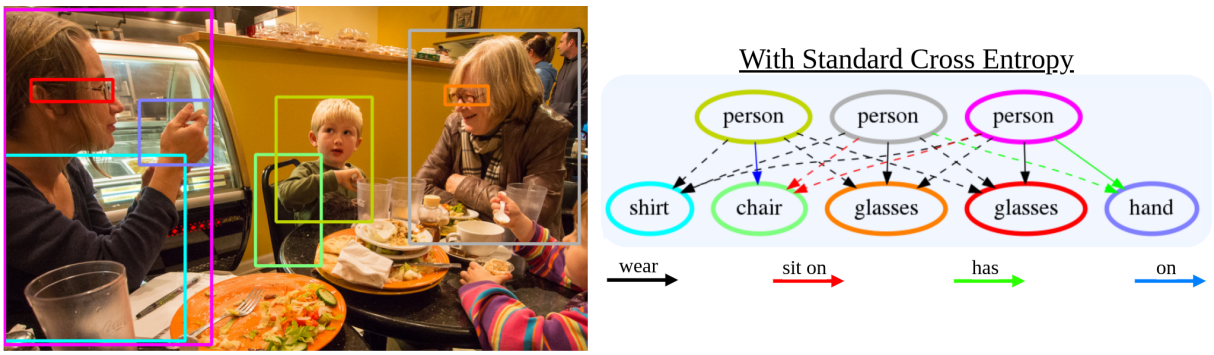

Figure 3.1: Προβλέψεις του [19] εκπαιδευμένο στα επισημειωμένα δείγματα με cross entropy ως συνάρτηση κόστους. Οι μη επισημειωμένες σχέσεις συμβολίζονται με διακεκομμένο βέλος.

προβλέπει πως όλοι οι άνθρωποι φορούν όλα τα γυαλιά καθώς και όλες τις μπλούζες, κάθονται στην ίδια καρέκλα και έχουν το ίδιο χέρι. Σημειωτέον ότι η κλάση "wear" είναι η δεύτερη πολυπληθέστερη σχέση στο VRD και τα περισσότερα μοντέλα επιτυγχάνουν Recall κοντά στο 100% για τη συγκεκριμένη κλάση. Ακόμη, αξιοσημείωτο είναι πως απομονώνοντας τις προβλέψεις για τα επισημειωμένα δείγματα (συνεχή βέλη) αποκρύπτεται ολοκληρωτικά το γεγονός ότι το μοντέλο είναι πολύ μακρυά από το να κατανοήσει θεμελιωδώς τα νοήματα των κλάσεων "wear", "sit on" και "has". Συμπεραίνουμε λοιπόν πως δεν αρκεί να απαιτούμε απλά χαμηλή πιθανοφάνεια για τα

μη επισημειωμένα δείγματα αλλά θα πρέπει να γίνεται έλεγχος και για την κλάση που προβλέπεται για αυτά. Επειδή οι προβλέψεις του δικτύου φαίνεται να χρησιμοποιούν τυφλά μόνο την πληροφορία για τα αντικείμενα που εξετάζουν, δηλαδή τα συμφραζόμενα (context [56]), π.χ. ένας άνθρωπος πάντα φοράει μια μπλούζα ανεξαρτήτως του περιεχομένου της εικόνας, ονομάζουμε αυτό το πρόβλημα ***context bias***. Συνήθως στην βιβλιογραφία ο όρος context χρησιμοποιείται εννοώντας την κατηγορία του υποκειμένου και του αντικειμένου, εμείς όμως διευρύνουμε την έννοια του όρου συμπεριλαμβάνοντας οποιαδήποτε σημασιολογική τους πληροφορία. Έτσι το context μίας σχέσης είναι οποιαδήποτε σημασιολογική πληροφορία υποκειμένου και αντικειμένου η οποία βρίσκεται στα γλωσσικά (*word2vec*) διανύσματα των κατηγοριών τους ή και στα οπτικά χαρακτηριστικά τους. Σε αυτό το κεφάλαιο θα διερευνήσουμε τα αίτια και τα αποτελέσματα του context bias ώστε στη συνέχεια να προτείνουμε λύσεις για την αντιμετώπισή του.

3.1 Το πείραμα του κυλιόμενου παραθύρου

Για να εξετάσουμε πειραματικά το κατά πόσο ένα μοντέλο βασίζεται στο context για να κάνει μία πρόβλεψη δημιουργήσαμε το πείραμα του κυλιόμενου παραθύρου (**sliding box experiment**). Σε αυτό το πείραμα, κρατώντας το κουτί του υποκειμένου ακίνητο αρχίζουμε να μετακινούμε το κέντρο του κουτιού του αντικειμένου σε όλη την έκταση της εικόνας κάνοντας προβλέψεις. Θεωρώντας 0 όποτε η πρόβλεψη είναι λανθασμένη και 1 σωστή, δημιουργούμε ένα δυαδικό χάρτη έντασης (heatmap) το οποίο υπερθέτουμε στην αρχική εικόνα.

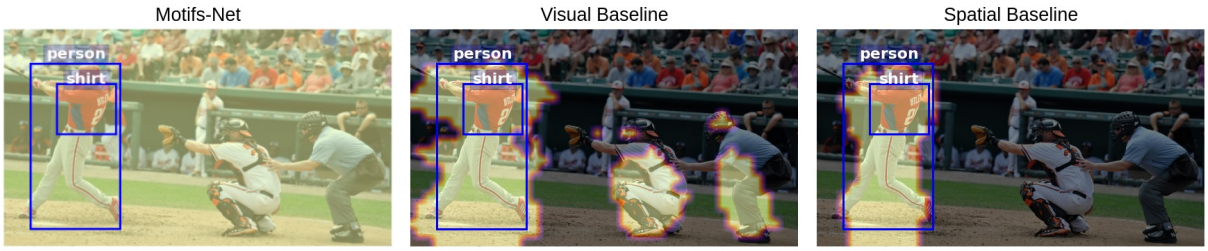

Figure 3.2: Sliding box experiment για τρία μοντέλα. Το bounding box του `person` παραμένει σταθερό καθώς εκείνο του `shirt` μετακινείται και το κάθε σημείο του heatmap εκφράζει το αν, δεδομένου ότι το κέντρο του bounding box του `shirt` βρίσκεται εκεί, η πρόβλεψη είναι "wear". Αριστερά: μοντέλο του [51]. Μέση: μοντέλο που χρησιμοποιεί μόνο visual πληροφορία. Δεξιά: μοντέλο που χρησιμοποιεί μόνο χωρική πληροφορία

Στην εικόνα 3.2 παρατηρούμε το sliding box experiment για τρία μοντέλα. Αριστερά το Motifs-Net, ένα από τα πιο διαδεδομένα μοντέλα, προβλέπει πως οπουδήποτε και αν βρίσκεται ένα bounding box επισημειωμένο ως `shirt` ο άνθρωπος θα το φοράει, αγνοώντας πλήρως οποιαδήποτε χωρική ή οπτική πληροφορία. Στη μέση, ένα δίκτυο που χρησιμοποιεί μόνο οπτική πληροφορία θεωρεί πως από τη στιγμή που μέσα στο κουτί του object εικονίζεται κάποιο μέλος ενός ανθρώπου, ακόμη και αν δεν ανήκει στον άνθρωπο που έχει τον ρόλο του subject, τότε θα το φοράει. Δεξιά, ένα δίκτυο που χρησιμοποιεί μόνο χωρική πληροφορία έρχεται πιο κοντά από όλα τα προηγούμενα καθώς κατανοεί πως για να φοράει ένα υποκείμενο κάποιο αντικείμενο, το αντικείμενο θα πρέπει να βρίσκεται μέσα στο κουτί του υποκειμένου. Από το πείραμα αυτό μπορούμε να εξάγουμε τα παρακάτω ποιοτικά συμπεράσματα. Υπό την παρουσία σημασιολογικής πληροφορίας για την κατηγορία των αντικειμένων ακόμη και ένα σύνθετο δίκτυο (Motifs-Net), που χρησιμοποιεί γλωσσικά, χωρικά και οπτικά χαρακτηριστικά, τείνει να "βραχυκυκλώνει" την πρόβλεψή του με τα γλωσσικά χαρακτηριστικά για συγκεκριμένες κλάσεις όπως το "wear" στην περίπτωσή μας. Ακόμη, παρόμοια συμπεριφορά μπορεί να προκύψει και μόνο με τα οπτικά χαρακτηριστικά καθώς εμπεριέχουν την σημασιολογική πληροφορία της κατηγορίας των αντικειμένων που απεικονίζονται. Τέλος, η χωρική πληροφορία

αν και ικανή να περιορίσει το δίκτυο σε προβλέψεις που ακολουθούν την κοινή λογική του χώρου ορισμένων κλάσεων (spatial common sense) φαίνεται να αγνοείται πλήρως.

3.2 Στατιστικές ενδείξεις του context bias

Για να εντοπίσουμε τα αίτια της παραπάνω συμπεριφοράς δεν έχουμε παρά να εξετάσουμε την κατανομή των κλάσεων δεδομένου ενός context (`subject-object` ζευγάρι) που παρατηρούμε να είναι προβληματικό. Στην εικόνα 3.3 φαίνεται η κατανομή των επισημειώσεων για ορισμένα contexts

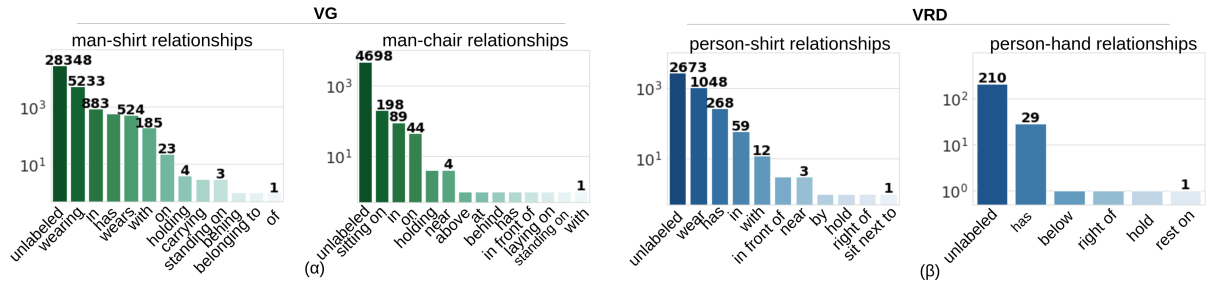

Figure 3.3: Κατανομή επισημειώσεων δεδομένου ενός context (`subject-object` pair) για το VG2 και το VRD. Τα ραβδογράμματα είναι σε λογαριθμική κλίματα και όσες κλάσεις έχουν μηδενικό αριθμό δειγμάτων παραλείπονται.

για τα οποία παρατηρήθηκαν biased προβλέψεις. Είναι φανερό ότι η συντριπτική πλειοψηφία των δειγμάτων είναι είτε μη επισημειωμένα (unlabeled) είτε ανήκουν σε μία συγκεκριμένη κλάση και κάποια συνώνυμά της, όπου ο όρος συνώνυμα στο πλαίσιο του εντοπισμού σχέσεων χρησιμοποιείται με την έννοια που αναλύεται και στο [14]. Οι επισημειωτές δηλαδή, σύμφωνα με την ανθρώπινη κοινή λογική, για συγκεκριμένα ζεύγη υποκειμένου-αντικειμένου επισημειώνουν μόνο μία συγκεκριμένη κλάση θεωρώντας πως οποιαδήποτε άλλη περίπτωση δεν έχει χρήσιμη πληροφορία. Για παράδειγμα το υποκείμενο `person` με το αντικείμενο `shirt` επισημειώνονται κυρίως όταν συνδέονται με τη σχέση `wear` (εικόνα 3.3β αριστερά) καθώς οι επισημειωτές κρίνουν πως για παράδειγμα η τριπλέτα `<person-next to-shirt>` δεν προσφέρει ουσιαστική πληροφορία. Έτσι, κατά τη διάρκεια της εκπαίδευσης τα μοντέλα δέχονται δείγματα όπου ένας άνθρωπος πάντα θα φοράει ("wear") μία μπλούζα με αποτέλεσμα να αρκεί μόνο και μόνο η σημασιολογική πληροφορία (`person-shirt`) ή αλλιώς το σημασιολογικό context για την κατηγοριοποίηση ενός τέτοιου δείγματος. Αυτό το πρόβλημα είναι ανεξάρτητο αρχιτεκτονικής καθώς πρόκειται για στατιστική ανισορροπία του συνόλου δεδομένων. Μέχρι στιγμής, το context bias δεν έχει αντιμετωπιστεί στη βιβλιογραφία και οι κύριοι λόγοι είναι δύο. Πρώτον, καθώς ως κλάσεις αντιμετωπίζονται μεμονωμένα οι σχέσεις και όχι οι τριπλέτες σχέσεων δημιουργείται η εντύπωση πως μόνο οι κλάσεις με μικρό αριθμό δειγμάτων είναι προβληματικές, κάτι το οποίο όπως δείξαμε δεν ισχύει (το "wear" αποτελεί τη δεύτερη κλάση με τα περισσότερα δείγματα στο VRD). Δεύτερον, το Recall αγνοεί την ορθότητα των προβλέψεων για τα μη επισημειωμένα δείγματα. Η μετρική που θα μπορούσε να αναδείξει το πρόβλημα είναι το Precision αλλά αποφεύγεται συστηματικά στη βιβλιογραφία καθώς η εφαρμογή του συνεπάγεται τη θεώρηση όλων των μη επισημειωμένων δειγμάτων ως αρνητικά, κάτι το οποίο θα ήταν αφελές και παραπλανητικό.

3.3 Κατάταξη με βάση την εντροπία

Έχοντας εκθέσει το context bias καθώς και τα αίτια δημιουργίας του, το επόμενο βήμα είναι ο εντοπισμός , ή καλύτερα ο σχεδιασμός ενός μηχανισμού εντοπισμού, των κλάσεων που πάσχουν από το συγκεκριμένο πρόβλημα. Οδηγούμενοι από την εικόνα 3.3 και παρατηρώντας ότι οι πάσχουσες κλάσεις (π.χ. "wear", "has") συγκεντρώνουν την πλειοψηφία των δειγμάτων δεδομένων ορισμένων

context, επιλέγουμε να κατατάξουμε τις κλάσεις μετρώντας για κάθε μία τη μέση εντροπία των κατανομών στις οποίες είναι η πολυπληθέστερη, για όλα τα πιθανά contexts. Αυτόν τον τρόπο κατάταξης τον ονομάζουμε **entropy ranking**. Συγκεκριμένα ορίζουμε την κατανομή δεδομένου ενός context ως:

$$p_{ij}(k) = Pr(predicate = k | subj = i, obj = j)$$

όπου η κλάση "background" (unlabeled) δεν συμπεριλαμβάνεται. Επιπλέον, για να μειώσουμε τον θόρυβο αγνοούμε ορισμένες τριπλέτες οι οποίες είναι εξ ορισμού biased όπως για παράδειγμα **<sky-above-person>** ενώ φιλτράρουμε contexts με πολύ μικρό αριθμό δειγμάτων. Αναπαριστούμε κάθε context με την κλάση με την οποία έχει τα περισσότερα δείγματα:

$$\hat{k}(i,j) = \underset{k}{argmax}\ p_{ij}(k)$$

και το σύνολο των contexts που αντιπροσωπεύει η κλάση k είναι

$$\mathcal{C}_k = \{i, j : \hat{k}(i,j) = k\}$$

Τέλος υπολογίζουμε τη μέση εντροπία για κάθε κλάση

$$e(k) = \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} E(p_{ij})$$

όπου $E(\cdot)$ είναι η συνάρτηση εντροπίας και κατατάσσουμε σε αύξουσα σειρά με βάση το $e(k)$.

Ως γνωστόν, η εντροπία μίας κατανομής μπορεί να θεωρηθεί ως το μέτρο της αβεβαιότητάς της. Δηλαδή, κατανομές με μικρή εντροπία (κοντά στο μηδέν) εκφράζουν μεγαλύτερη βεβαιότητα, και συνεπώς συγκεντρώνονται περισσότερο σε ορισμένες κλάσεις, ενώ εκείνες με μεγαλύτερη εντροπία είναι πιο ομοιόμορφα κατανεμημένες.

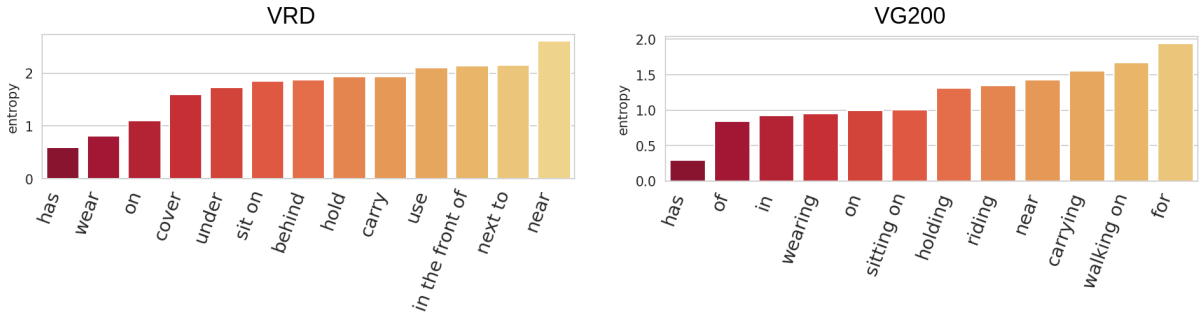

Figure 3.4: Κλάσεις με τη μικρότερη μέση εντροπία στα VRD και VG200 υπολογισμένη σύμφωνα με το entropy ranking.

Στην εικόνα 3.4 βλέπουμε τις 13 κλάσεις με την μικρότερη εντροπία στο VRD. Παρατηρούμε πως πράγματι σε κλάσεις που έχουν context bias σημειώνεται χαμηλότερη μέση εντροπία από τις υπόλοιπες. Ακόμη, σχέσεις που περιγράφουν μόνο τη χωρική διάταξη δύο αντικειμένων π.χ. "next to", "near" κλπ τείνουν να έχουν μεγαλύτερη εντροπία από σχέσεις που δηλώνουν κάποια αλληλεπίδραση εγγύτητας που όμως απαιτεί παραπάνω πληροφορία πέραν της χωρικής. Για παράδειγμα οι επισημειωτές θα αποφύγουν να επισημειώσουν τη σχέση μεταξύ ενός ανθρώπου και μίας καρέκλας στην οποία δεν κάθεται καθώς το συγκεκριμένο context (**person-chair**) εμπεριέχει σημαντικότερη (κατά την ανθρώπινη κοινή λογική) πληροφορία υπό τη σχέση "sit on" παρά "next to". Έτσι τα μοντέλα δεν έχουν την ευκαιρία να κατανοήσουν την κοινή λογική που αφορά την χωρική διάταξη (spatial common sense [7]) που απαιτείται για να ισχύουν ορισμένες σχέσεις, όπως για παράδειγμα ότι δεν γίνεται ένας άνθρωπος να φοράει μία μπλούζα και τα bounding boxes τους να είναι τελείως ξένα μεταξύ τους (εικόνα 3.1). Τις σχέσεις αυτές που για να ισχύουν προϋποθέτουν χωρική εγγύτητα αλλά εννοιολογικά δεν είναι αμιγώς χωρικές τις ονομάζουμε σχέσεις **εγγύτητας** (**proximal** relations).

Κεφάλαιο 4

Εξόρυξη και εκπαίδευση με μη επισημειωμένα δείγματα

Όπως δείξαμε στο κεφάλαιο 3 υπάρχει έλλειψη επισημειώσεων για συγκεκριμένα contexts όμως η μη επισημειωμένη πληροφορία εμπεριέχει πολλά ζεύγη αντικειμένων που θα μπορούσαν να αποτελέσουν αρνητικά δείγματα (negative samples) για τις σχέσεις εγγύτητας (proximals). Αν με κάποιο τρόπο καταφέρναμε να χρησιμοποιήσουμε τη μη επισημειωμένη πληροφορία και να αναγκάσουμε το μοντέλο να μάθει και από αυτή, οι κατανομές θα εξομαλύνονταν και το context bias θα αμβλυνόταν. Πώς όμως μπορούμε να εξορύξουμε με έγκυρο, ακριβή και αυτόματο τρόπο αρνητικά παραδείγματα και κυρίως πώς θα τα χρησιμοποιήσουμε κατά τη διάρκεια της εκπαίδευσης;

Σε αυτό το κεφάλαιο, αρχικά θα προτείνουμε μία μέθοδο εξόρυξης αρνητικών παραδειγμάτων για σχέσεις εγγύτητας τον οποίο ονομάζουμε **αρνητική συμπλήρωση γράφου** (**negative graph completion**). Έπειτα θα προτείνουμε τρεις μεθόδους ενσωμάτωσης και αξιοποίησης των αρνητικών μη επισημειωμένων δειγμάτων οι οποίες είναι ανεξάρτητες της αρχιτεκτονικής των μοντέλων:

- Negative Cross Entropy (NCE): χρήση των αρνητικών δειγμάτων που έχουν εξορυχθεί με τη μέθοδο αρνητικής συμπλήρωσης γράφου. (απαιτείται πρότερη γνώση και των προβληματικών κλάσεων αλλά και των αρνητικών δειγμάτων στο σύνολο δεδομένων εκπαίδευσης)
- Negativity Ranking (NR): χρήση προεκπαιδεύμενου δικτύου που έχει μάθει να ταξινομεί την "αρνητικότητα" δειγμάτων ορισμένων κλάσεων. (απαιτείται πρότερη γνώση μόνο των προβληματικών κλάσεων)
- Grounding Consistency Loss (GCL): επιβολή συνέπειας στην επαναπροβολή της προβλεφθείσας σχέσης πάνω στην εικόνα. (δεν απαιτείται καμία πρότερη γνώση)

4.1 Αρνητική συμπλήρωση γράφου

Για την εξόρυξη αρνητικών δειγμάτων για τις κλάσεις εγγύτητας με αξιόπιστο τρόπο και τη συμπλήρωση των γράφων με αρνητικά δείγματα στηριχθήκαμε στο γεγονός ότι οι σχέσεις εγγύτητας μπορούν να χωριστούν σε δύο κατηγορίες ανάλογα με το πως δεσμεύουν το υποκείμενο ή το αντικείμενο που συνδέουν. Συγκεκριμένα, ονομάζουμε την πρώτη κατηγορία σχέσεων **κτητικές** (**possessive**) ενώ τη δεύτερη **ιδιοκτησίας** (**belonging**) (βλ. εικόνα 4.1). Η πρώτη δηλώνει πως το υποκείμενο "δεσμεύει" το αντικείμενο και δεν μπορεί να υπάρξει άλλο υποκείμενο που να συνδέεται με το αντικείμενο με αυτή την σχέση. Για παράδειγμα η σχέση **<person-wear-shirt>** επιβάλλει πως το αντικείμενο **shirt** δεν μπορεί να έχει άλλο υποκείμενο που να συνδέεται με αυτό με τη σχέση "wear". Η δεύτερη, συμμετρικά, δηλώνει πως το αντικείμενο "δεσμεύεται" από το υποκείμενο. Για παράδειγμα η σχέση **<shirt-on-person>** επιβάλλει πως η συγκεκριμένη μπλούζα δεν μπορεί να είναι πάνω ("on") σε άλλον άνθρωπο. Παρεκτείνοντας με βάση την λογική αυτών

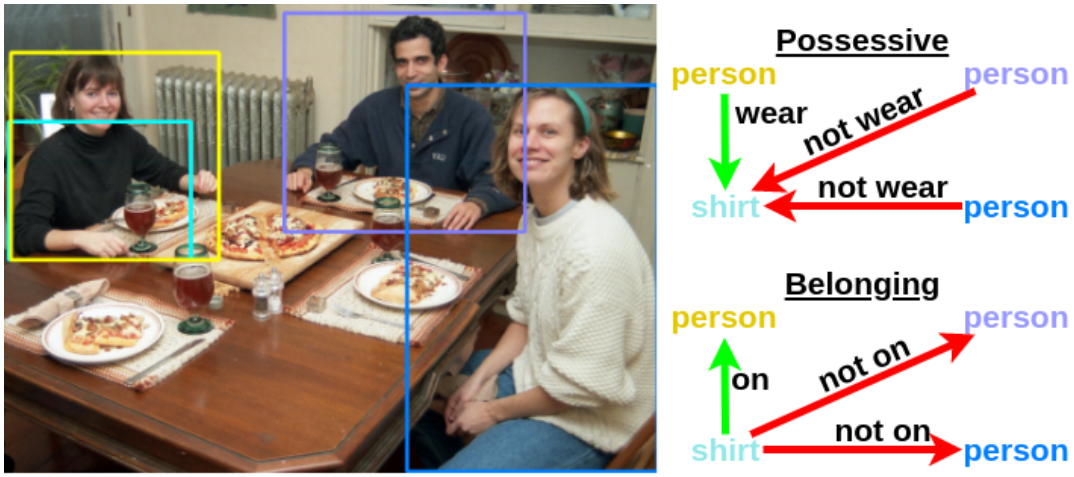

Figure 4.1: Οι επισημειωμένες σχέσεις (πράσινο) χρησιμοποιούνται για την παραγωγή αρνητικών σχέσεων (κόκκινο). `Possessive`: Ένας άνθρωπος που φοράει μία μπλούζα συνεπάγεται πως κανείς άλλος δεν μπορεί να τη φοράει. `Belonging`: Μία μπλούζα μπορεί να είναι μόνο πάνω σε έναν άνθρωπο.

των δύο συνόλων σχέσεων σε όλες τις σχέσεις των VRD και VG200 εκθέτουμε στον πίνακα 4.1 πως κατηγοριοποιούνται όλες οι σχέσεις εγγύτητας σε κτητικές και ιδιοκτησίας. Πιο συγκεκριμένα, ορίζουμε το σύνολο των σχέσεων κτήσης και ιδιοκτησίας ως $\mathcal{R}_p$, $\mathcal{R}_b$ αντίστοιχα και $r(s, o)$ τη σχέση ενός υποκειμένου s με ένα αντικείμενο o με το κατηγόρημα r. Η αρνητική συμπλήρωση γράφου λοιπόν πραγματοποιείται από δύο κανόνες παραγωγής αρνητικών δειγμάτων οι οποίοι ορίζονται ως:

- Κτητικοί: $\forall r \in \mathcal{R}_p, \forall s, o, s' : r(s, o) \implies \neg r(s', o)$
- Ιδιοκτησίας: $\forall r \in \mathcal{R}_b, \forall s, o, o' : r(s, o) \implies \neg r(s, o')$

Rules	VRD	VG200
Possessive	carry, contain, cover, drive, eat, feed, fly, has, hit, hold, kick, play with, pull, ride, touch, use, wear, with	carrying, eating, has, holding, playing, riding, using, wearing, wears, with
Belonging	at, drive on, in, inside, lean on, lying on, on, park on,rest on, sit on, skate on, sleep on, stand on	at, attached to, belonging to, flying in,for,from, growing on, hanging from, in, laying on, looking at, lying on, made of,mounted on, of, on, painted on, parked on, part of, says, sitting on, standing on, to,walking in, walking on, watching

Table 4.1: Τα σύνολα των κτητικών και ιδιοκτησίας σχέσεων στις οποίες εφαρμόζεται ο κάθε κανόνας για το VRD και το VG200.

4.2 Αρνητική εντροπία

Έχοντας πλέον έναν τρόπο εξαγωγής αρνητικών δειγμάτων μία απλή λύση θα ήταν να εφαρμόσουμε τους κανόνες που περιγράψαμε στην ενότητα 4.1 στο σύνολο δεδομένων εκπαίδευσης και να χρησιμοποιήσουμε τα αρνητικά δείγματα που θα εξάγουμε. Έτσι, πέρα από τον όρο του cross entropy στη συνάρτηση σφάλματος $\mathcal{L}_{CE}$ θα προστεθεί ένας όρος αρνητικής εντροπίας $\mathcal{L}_{NCE} = -log(1-p)$ που θα ωθεί την πιθανότητα των αρνητικών δειγμάτων στο μηδέν όπως στο [20] όπου εδώ p είναι η πιθανότητα της κλάσης από την οποία προέκυψε το αρνητικό δείγμα. Η συνολική συνάρτηση σφάλματος λοιπόν θα είναι $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{NCE}$. Λόγω του ότι το p παράγεται από το τελικό επίπεδο softmax του δικτύου η ώθησή του στο μηδέν ταυτόχρονα ισοδυναμεί και με αύξηση μίας ή περισσότερων άλλων κλάσεων. Ακόμη λοιπόν και αν φαίνεται πως επηρεάζουμε την έξοδο του δικτύου μόνο για μία κλάση (την αρνητική) στην πραγματικότητα όλες οι έξοδοι συμμετέχουν για να συνδιαμορφώσουν χαμηλή πιθανότητα στην έξοδο p.

4.3 Κατάταξη αρνητικότητας

Η μέθοδος του NCE που περιγράψαμε στο 4.2 για να λειτουργήσει απαιτεί την πρότερη γνώση δύο πληροφοριών. Πρώτα, των κλάσεων που έχουν context bias (σχέσεις εγγύτητας) καθώς και τον διαχωρισμό τους στις δύο υποκατηγορίες και έπειτα την εφαρμογή των δύο κανόνων για την εξαγωγή αρνητικών επισημειώσεων στο σύνολο δεδομένων εκπαίδευσης (negative graph completion).

Προσπαθώντας να μειώσουμε όσο το δυνατόν περισσότερο γίνεται τη χρήση πρότερης γνώσης, προτείνουμε ως επόμενο βήμα μία μέθοδο η οποία δεν απαιτεί την εύρεση αρνητικών δειγμάτων μέσω των κανόνων που προτείναμε στο 4.1. Αντ' αυτού, προεκπαιδεύουμε ένα δίκτυο που ονομάζουμε **ταξινομητή** (`ranker`) και συμβολίζουμε ως $\mathcal{R}$ το οποίο μαθαίνει να προβλέπει την "αρνητικότητα" κάποιου δείγματος.Για παράδειγμα, στην εικόνα 4.3 θα θέλαμε δίνοντας στον ταξινομητή για υποκείμενο έναν άνθρωπο και για αντικείμενο ένα ζευγάρι γυαλιά που δεν τα φοράει, η έξοδος (logit) του να είναι αρνητική για την κλάση "wear" και θετική στην περίπτωση όπου τα γυαλιά πράγματι φοριούνται από τον άνθρωπο. Έτσι, κατά τη διάρκεια της εκπαίδευσης κάποιου μοντέλου, έχοντας παγωμένες τις παραμέτρους του ταξινομητή, τον χρησιμοποιούμε ως συνάρτηση βαθμολόγησης της "αρνητικότητας" των μη επισημειωμένων δειγμάτων. Μένουν λοιπόν να απαντηθούν δύο ερωτήματα: με ποιον τρόπο διαλέγουμε τα δείγματα που θα χρησιμοποιηθούν ως αρνητικά κατά την εκπαίδευση και το πώς εκπαιδεύουμε το δίκτυο του ταξινομητή.

Για να απαντήσουμε στο πρώτο ερώτημα πρέπει να συλλογιστούμε πως δεν έχουν όλα τα αρνητικά δείγματα μίας κλάσης την ίδια συνεισφορά στο βαθμό κατανόησής της. Για παράδειγμα, ένας άνθρωπος που στέκεται δίπλα σε ένα κράνος και ένας άνθρωπος που καβαλάει ένα ποδήλατο είναι και τα δύο αρνητικά δείγματα της τριπλέτας `<person-wear-helmet>`. Όμως, μόνο ο άνθρωπος δίπλα στο κράνος θα βοηθήσει ουσιαστικά στην κατανόηση του τι σημαίνει να μην φοράω κάτι καθώς δεν είναι πιθανοφανές ένας άνθρωπος να φοράει ένα ποδήλατο. Φιλτράρουμε λοιπόν τα δείγματα με βάση αυτήν την πιθανοφάνεια η οποία υπολογίζεται από το δίκτυο του ταξινομητή το οποίο συμβολίζουμε ως $\mathcal{R}$. Αναπαριστούμε λοιπόν την πιθανότητα που προβλέπει για το δείγμα (s, o) (υποκείμενο, αντικείμενο) να ανήκει στην κλάση r ως $\sigma(\mathcal{R}(s,o))_r$, όπου $\sigma(\cdot)$ η συνάρτηση softmax. Άρα δείγματα με υψηλό $\sigma(\mathcal{R}(s,o))_r$ αποτελούν *πιθανά* δείγματα ενδιαφέροντος (είτε θετικά είτε αρνητικά) για την κλάση r. Στην πράξη ορίζουμε κάποιο κατώφλι, συνήθως 0.5, και έτσι δεδομένου ότι $\sigma(\mathcal{R}(s,o))_r > 0.5$ και $\mathcal{R}(s,o)_r < 0$ για το (s,o), το θεωρούμε ως αρνητικό της κλάσης r και εφαρμόζουμε τη συνάρτηση κόστους που περιγράφηκε στο κεφάλαιο 4.2. Ισοδύναμα, ακολουθώντας τον συμβολισμό του κεφαλαίου 4.1:

$$\sigma(\mathcal{R}(s,o))_r \times sgn(\mathcal{R}(s,o)_r) < -0.5 \implies \neg r(s,o) \tag{4.1}$$

όπου $sgn(\cdot)$ η συνάρτηση προσήμου.

Σκοπός του ταξινομητή (ranker) είναι να καταφέρει να διαχωρίσει τα θετικά από τα αρνητικά δείγματα μίας κλάσης, έχοντας στη διάθεσή του μόνο θετικά και μη επισημειωμένα δείγματα ([45]).

Για να το πετύχουμε αυτό, εκπαιδεύουμε τον ταξινομητή όπως οποιοδήποτε άλλο δίκτυο με τη συνηθισμένη συνάρτηση cross entropy στο πρόβλημα της κατηγοριοποίησης σχέσεων, προσθέτοντας όμως έναν επιπλέον όρο $\mathcal{L}_{NEG}$ ο οποίος προσπαθεί να ωθήσει τις εξόδους (logits) των επισημειωμένων δειγμάτων να γίνουν θετικές ενώ των μη επισημειωμένων αρνητικές. Επειδή όμως, όπως

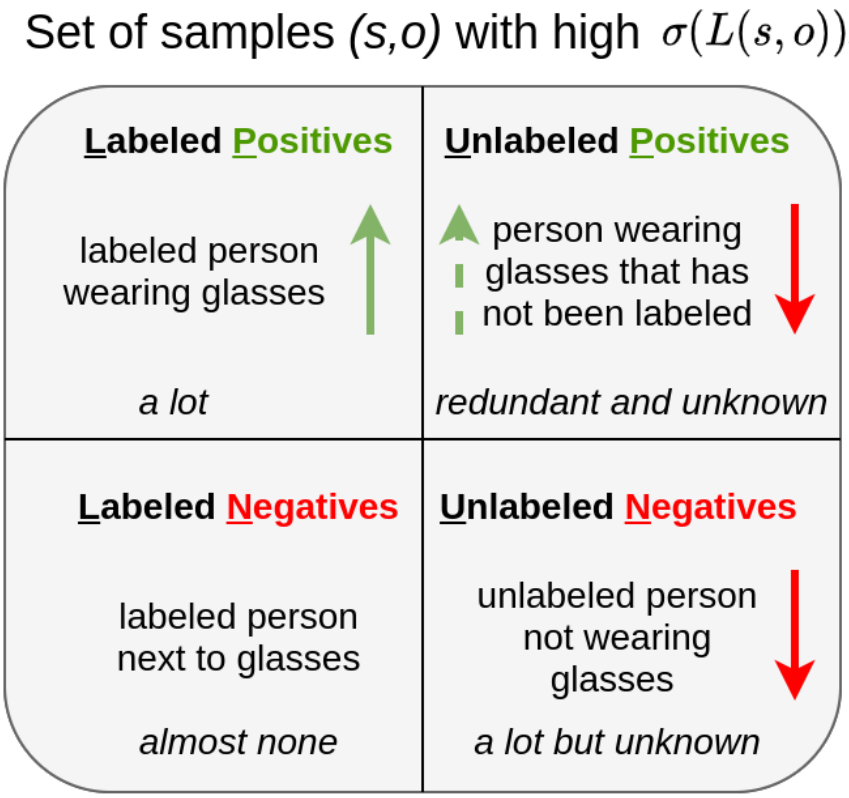

Figure 4.2: Διαχωρισμός δειγμάτων ενδιαφέροντος για μία κλάση r ανάλογα με την ύπαρξη επισημείωσης και το αν είναι θετικά η αρνητικά. Με τα βέλη συμβολίζουμε την επίδραση του $\mathcal{L}_{NEG}$ στις εξόδους της κλάσης r του $\mathcal{R}$.

αναφέραμε και πριν, δεν είναι όλα τα μη επισημειωμένα δείγματα χρήσιμα για να χρησιμοποιηθούν ως αρνητικά, εκπαιδεύουμε ένα δίκτυο L (βλ. κεφ 2.10) το οποίο χρησιμοποιεί μόνο γλωσσική πληροφορία και, δεδομένης μίας προβληματικής κλάσης r, εφαρμόζουμε το $\mathcal{L}_{NEG}$ μόνο σε δείγματα όπου $\sigma(L(s,o))_r > 0.5$.

Έτσι:

$$\mathcal{L}_{NEG}(s,o) = \begin{cases} -log(sig(\mathcal{R}(s,o)_r)) & \text{if } labeled \\ -log(1 - sig(\mathcal{R}(s,o)_r)) & \text{if } unlabeled \end{cases} \tag{4.2}$$

όπου $sig(\cdot)$ είναι η σιγμοειδής συνάρτηση.

Για να εξηγήσουμε τη λειτουργία του $\mathcal{L}_{NEG}$ πρέπει πρώτα να παρατηρήσουμε ότι τα δείγματα ενδιαφέροντος για μία κλάση r, δηλαδή εκείνα με υψηλό $\sigma(L(s,o))_r$, μπορούν να χωριστούν σε τέσσερις κατηγορίες ανάλογα με το αν έχουν επισημείωση και το αν είναι θετικά ή αρνητικά (βλ. σχήμα 4.2). Σύμφωνα με τη λογική που περιγράψαμε, λόγω του $\mathcal{L}_{NEG}$ τα επισημειωμένα και θετικά δείγματα (LP) αποκτούν θετικές εξόδους ενώ τα μη επισημειωμένα αρνητικά (UN) αρνητικές. Το ενδιαφέρον είναι στην περίπτωση των μη επισημειωμένων θετικών δειγμάτων (UP). Εκεί, οι έξοδοι από τη μία ωθούνται στα αρνητικά (συνεχές κόκκινο βέλος), από την άλλη όμως λόγω του ότι είναι θετικά δείγματα έμμεσα θα ωθούνται και στα θετικά (διακεκομμένο πράσινο βέλος) και επειδή η κατανομή των UP είναι πιο κοντά στα LP θα καταλήξουν να είναι θετικά. Για παράδειγμα, στον πίνακα 4.2 βλέπουμε πως ο ταξινομητής αρνητικότητας $\mathcal{R}$ μπορεί ακόμη και στα μη επισημειωμένα δείγματα να μας δώσει πληροφορία για το αν είναι αρνητικά της κλάσης “wear”.

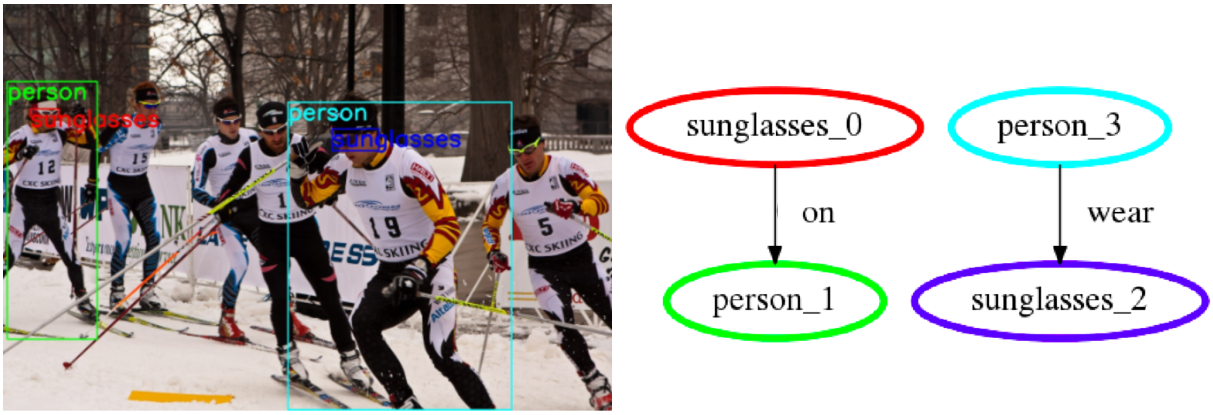

Figure 4.3: Εικόνα δύο επισημειωμένων ανθρώπων και οι επισημειώσεις που είναι διαθέσιμες κατά τη διάρκεια της εκπαίδευσης. Σημειωτέον το γεγονός ότι δεν υπάρχει πληροφορία για τη σχέση μεταξύ ενός ανθρώπου με τα γυαλιά ενός άλλου (reporting bias).

subject	object	"wear" logit	labeled
sunglasses_0	person_1	1.31	yes
sunglasses_2	person_3	1.21	no
sunglasses_0	person_3	-3.18	no
sunglasses_2	person_1	-3.89	no

Table 4.2: Έξοδοι του ταξινομητή ($\mathcal{R}$) για τα ζευγάρια της εικόνας 4.3. Παρατηρούμε πως ανεξάρτητα από το αν είναι επισημειωμένα ή όχι καταφέρνει να διαχωρίσει με το πρόσημο της εξόδου της κλάσης "wear" τα θετικά από τα αρνητικά δείγματα.

Συνοψίζοντας, η διαδικασία εκπαίδευσης με ταξινόμηση αρνητικών δειγμάτων (NR) που περιγράψαμε έχει τα εξής βήματα:

- Εκπαίδευση μοντέλου L που χρησιμοποιεί μόνο γλωσσική πληροφορία για να χρησιμοποιηθεί ως πρότερη γλωσσική πιθανότητα (a priori).
- Εκπαίδευση ενός μοντέλου εντοπισμού σχέσεων $\mathcal{R}$ με τον πρόσθετο όρο $\mathcal{L}_{NEG}$ όπως ορίζεται στην 4.2 να εφαρμόζεται στις σχέσεις εγγύτητας που έχουμε επιλέξει.
- Εκπαίδευση οποιουδήποτε δικτύου το οποίο χρησιμοποιεί το $\mathcal{L}_{NCE}$ σύμφωνα με την σχέση 4.1.

Το δεύτερο βήμα προαπαιτεί να ορίσουμε σε ποιες σχέσεις θα εφαρμοστεί ο όρος $\mathcal{L}_{NEG}$ καθώς ενδέχεται για παράδειγμα να μην έχει αποτέλεσμα σε κλάσεις με πολύ λίγα δείγματα.

4.4 Κοινή λογική του χώρου και grounding συνέπεια

Οι δύο μέθοδοι που παρουσιάσαμε παραπάνω αν και είναι αποτελεσματικές, όπως θα δούμε και στο κεφάλαιο 5.3.1, απαιτούν την πρότερη γνώση των προβληματικών κλάσεων, εκείνων δηλαδή που επηρεάζονται αρνητικά από το context bias. Επίσης, στην περίπτωση της μεθόδου αρνητικής εντροπίας που περιγράψαμε στο κεφάλαιο 4.2 απαιτείται και η γνώση των δύο κανόνων που δημιουργούν τα αρνητικά δείγματα. Αυτοί οι περιορισμοί καθιστούν αυτές τις μεθόδους δύσκολα κλιμακώσιμες καθώς η μεταφορά τους σε νέα σύνολα δεδομένων απαιτεί την προσεκτική προεπεξεργασία των κλάσεων και τον επαναπροσδιορισμό των προβληματικών περιπτώσεων. Πώς λοιπόν μπορούμε να άρουμε όλους αυτούς τους περιορισμούς και να δημιουργήσουμε μία μέθοδο που δε θα χρειάζεται καμία ανθρώπινη παρέμβαση και θα μπορεί να εφαρμοστεί ανεξαρτήτως αρχιτεκτονικής μοντέλου αλλά και κυρίως συνόλου δεδομένων;

Η απάντηση σε αυτήν την ερώτηση βρίσκεται στον ορισμό της κοινής λογικής του χώρου (spatial common sense) που αναφέραμε στο κεφάλαιο 3.1 και στο τι ακριβώς σημαίνει για ένα μοντέλο να αντιλαμβάνεται τη λογική της χωρικής διάταξης δύο αντικειμένων δεδομένης της μεταξύ τους σχέσης. Ας πάρουμε για παράδειγμα την κλάση "wear". Η συγκεκριμένη κλάση έχει κάποια σημασιολογική λογική αλλά και μία χωρική λογική. Σύμφωνα με την πρώτη ένας άνθρωπος δεν μπορεί να φοράει ένα κτήριο ενώ σύμφωνα με τη δεύτερη ένας άνθρωπος δεν μπορεί να φοράει μία μπλούζα που δεν είναι "πάνω" του. Πιο συγκεκριμένα, μπορούμε να πούμε ότι δεδομένου του ανθρώπου αλλά και του ότι φοράει μία μπλούζα, η κοινή λογική του χώρου επιτρέπει την ύπαρξη της μπλούζας μόνο "πάνω" στον συγκεκριμένο άνθρωπο. Αντιστρόφως, δεδομένης της ύπαρξης μίας μπλούζας αλλά και του ότι την φοράει ένας άνθρωπος, ο τελευταίος υποχρεούται να βρίσκεται σε ένα συγκεκριμένο εύρος του χώρου όπου θα μπορούσε η μπλούζα να φορεθεί από αυτόν. Όσον αφορά τη σημασιολογική κοινή λογική, καλύπτεται (και ίσως περισσότερο από όσο θα θέλαμε) από τα γλωσσικά και οπτικά χαρακτηριστικά εισόδου. Όμως, όπως δείξαμε πειραματικά στο κεφάλαιο 3.1 τα χωρικά χαρακτηριστικά των κουτιών οριοθέτησης εμπεριέχουν την κοινή λογική του χώρου των σχέσεων αλλά αγνοούνται πλήρως από τα μοντέλα.

4.4.1 Grounding συνέπεια (Grounding consistency)

Θέλουμε λοιπόν να αναγκάσουμε τα μοντέλα να μάθουν να απαντούν στις ερωτήσεις: "που πρέπει να είναι μία μπλούζα δεδομένου ότι ένας άνθρωπος τη φοράει;" και "που πρέπει να είναι ένας άνθρωπος δεδομένου ότι μία μπλούζα φοριέται από αυτόν;". Ουσιαστικά, αυτές οι ερωτήσεις περιγράφουν το αντίστροφο πρόβλημα της πρόβλεψης σχέσεων που ονομάζεται `grounding` στο οποίο δεδομένης μίας τριπλέτας `<subject-predicate-object>` πρέπει να εντοπίσουμε τα αντικείμενα στην εικόνα. Προς το παρόν θα θεωρήσουμε ότι έχουμε μία συνάρτηση g που μπορεί να λύνει αυτό το αντίστροφο πρόβλημα και για είσοδο μία εικόνα και μία τριπλέτα, δίνει ως έξοδο δύο διδιάστατους χάρτες έντασης (heatmaps) για τον εντοπισμό του υποκειμένου και του αντικειμένου. Η αρχιτεκτονική και ο σχεδιασμός του g θα περιγραφεί αναλυτικά στο κεφάλαιο 4.4.4. Δεδομένου λοιπόν του g και ενός οποιουδήποτε δικτύου πρόβλεψης σχέσεων f μπορούμε ουσιαστικά να δημιουργήσουμε έναν κλειστό βρόχο όπου για ένα ζεύγος αντικειμένων προβλέπεται μία σχέση και έπειτα απαιτούμε η σχέση που προβλέφθηκε δεδομένης της εικόνας να μπορεί να οδηγήσει μέσω του

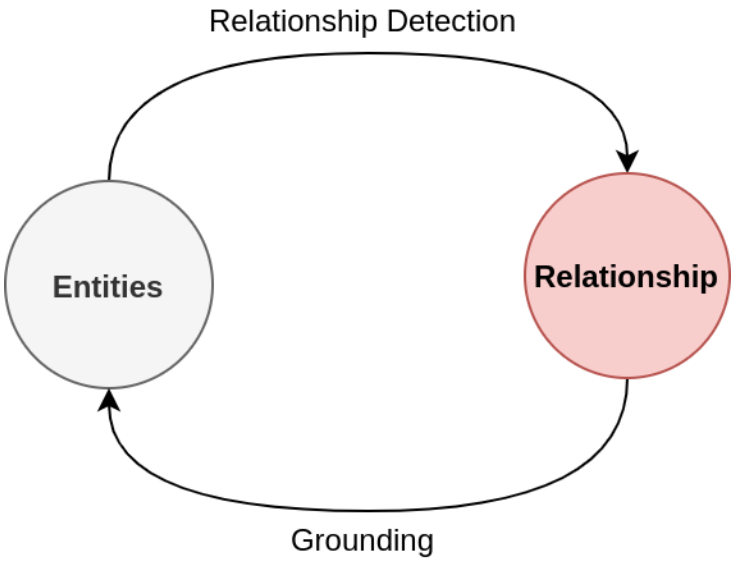

Figure 4.4: Η κοινή λογική του χώρου επιβάλλεται κλείνοντας τον βρόχο πρόβλεψης μίας σχέσης μέσω του αντίστροφου προβλήματος (grounding). Έτσι η προβλεπόμενη σχέση πρέπει να μπορεί να εξηγήσει και χωρικά τη διάταξη των αντικειμένων.

g πίσω στα αρχικά αντικείμενα (βλ. εικόνα 4.4). Πολύ σημαντικό χαρακτηριστικό της διάταξης της εικόνας 4.4 είναι ότι σε κανένα σημείο της διαδικασίας δε χρειαζόμαστε επισημειώσεις. Μπορούμε δηλαδή να εφαρμόσουμε, κάποια συνάρτηση κόστους που θα επιβάλλει αυτού του είδους τη συνέπεια (consistency) σε όλα τα δείγματα του συνόλου δεδομένων. Ωστόσο εμείς θα το κάνουμε μόνο στα μη επισημειωμένα δείγματα καθώς για τα υπόλοιπα έχουμε πληροφορία που επιβλέπει την εκμάθησή τους.

Ορίζουμε $f(S, O) \rightarrow p \in \mathcal{P}$ ένα οποιοδήποτε δίκτυο εντοπισμού σχέσεων το οποίο δέχεται ως είσοδο το υποκείμενο $S = (s_v, s_{sem}, s_{sp})$ και το αντικείμενο $O = (o_v, o_{sem}, o_{sp})$ ως πληροφορία της μορφής (οπτική, γλωσσική, χωρική) και προβλέπει μία σχέση p του συνόλου $\mathcal{P}$ των σχέσεων. Ακόμη, ορίζουμε τη συνάρτηση:

$$g(s_{sem}, p, o_{sem}) \rightarrow (\mathbf{h}_s \in \mathbb{R}^{H \times W}, \mathbf{h}_o \in \mathbb{R}^{H \times W}) \tag{4.3}$$

η οποία λύνει το αντίστροφο πρόβλημα από την f και παράγει τους χάρτες εντοπισμού $\mathbf{h}_s, \mathbf{h}_o$ για το υποκείμενο και το αντικείμενο αντίστοιχα. Κάθε στοιχείο των $\mathbf{h}_s, \mathbf{h}_o$ ανήκει στο εύρος $[0, 1]$ και αντιπροσωπεύει το κατά πόσο το κέντρο του υποκειμένου (ή αντικειμένου) βρίσκεται σε εκείνο το σημείο της εικόνας. Θα παρουσιάσουμε δύο τρόπους χρήσης της g στα κεφάλαια 4.4.2 και 4.4.3 και όπως θα εξηγήσουμε πειραματικά στο κεφάλαιο 5 ο δεύτερος τρόπος υπερτερεί.

4.4.2 Προσέγγιση συνέπειας χρησιμοποιώντας τον grounder ως ταξινομητή

Εκ πρώτης όψεως η 4.3 ορίζει πως η g παράγει ένα ζευγάρι από χάρτες έντασης (heatmaps) για τη δοθείσα σχέση. Όμως εμείς μπορούμε να υπολογίσουμε τα $g(s_{sem}, r, o_{sem})$ για οποιαδήποτε σχέση $r \in \mathcal{P}$ παίρνοντας έτσι $|\mathcal{P}|$ ζεύγη από χάρτες έντασης. Κάθε ένα από αυτά τα ζευγάρια εκφράζει το που, σύμφωνα με την αντίστοιχη σχέση, είναι αναμενόμενο να βρεθεί το υποκείμενο και το αντικείμενο. Μπορούμε λοιπόν όπως φαίνεται και στην εικόνα 4.5 να υπολογίσουμε την KL απόκλιση κάθε χάρτη έντασης που προβλέψαμε με τον πραγματικό δυαδικό χάρτη (ground truth binary mask) και ύστερα να περάσουμε τα αποτελέσματα από ένα επίπεδο softmin ώστε να πάρουμε μία κατανομή $P^{sub}_{ground} \in \mathbb{R}^{|\mathcal{P}|}$ (και P^{obj}_{ground} αντίστοιχα για το αντικείμενο) όπου κλάσεις που "εξηγούν" καλά τη χωρική διάταξη των αντικειμένων θα έχουν υψηλή πιθανότητα και αντιστρόφως. Τέλος υπολογίζουμε και αθροίζουμε την KL απόκλιση μεταξύ των $P^{sub}_{ground}, P^{obj}_{ground}$ και της εξόδου $P_{pred} \in \mathbb{R}^{|\mathcal{P}|}$ του μοντέλου πρόβλεψης σχέσεων όπου:

$$KL(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{4.4}$$

Ως τελική συνάρτηση κόστους παίρνουμε την:

$$\mathcal{L} = \mathcal{L}_{CE} + \frac{KL(P_{pred} || P^{sub}_{ground}) + KL(P_{pred} || P^{obj}_{ground})}{2}$$

Σημαντική λεπτομέρεια αποτελεί η επιλογή της κατανομής στόχου ($P(x)$ στην εξίσωση 4.4) στην KL απόκλιση. Στην περίπτωσή μας η κατανομή P_{pred} είναι διαφορίσιμη ενώ οι $P^{sub/obj}_{ground}$ όχι, αφού το δίκτυο g (grounder) έχει παγωμένες παραμέτρους. Έτσι, (βλ. εικόνα 4.6) θέτοντας το P_{pred} ως στόχο η KL απόκλιση θα λειτουργεί με `mode-seeking` τρόπο, που σημαίνει πως το δίκτυο θα προσπαθήσει οι κλάσεις με υψηλή πιθανότητα στο P_{pred} να αντιστοιχούν σε υψηλή πιθανότητα και στα $P^{sub/obj}_{ground}$. Δεν θα δημιουργήσουν μεγάλο σφάλμα όμως ασυμφωνίες σε κλάσεις όπου έχουν χαμηλή πιθανότητα στο P_{pred}. Σε αντίθετη περίπτωση όπου για στόχο έχουμε τα $P^{sub/obj}_{ground}$ η KL απόκλιση θα λειτουργήσει ως `mean-seeking` και το δίκτυο θα προσπαθήσει να αυξήσει την πιθανότητα σε κλάσεις όπου έχουν υψηλή πιθανότητα στο $P^{sub/obj}_{ground}$ ενώ δεν θα επηρεάσει η ασυμφωνία σε κλάσεις με χαμηλή πιθανότητα στο $P^{sub/obj}_{ground}$.

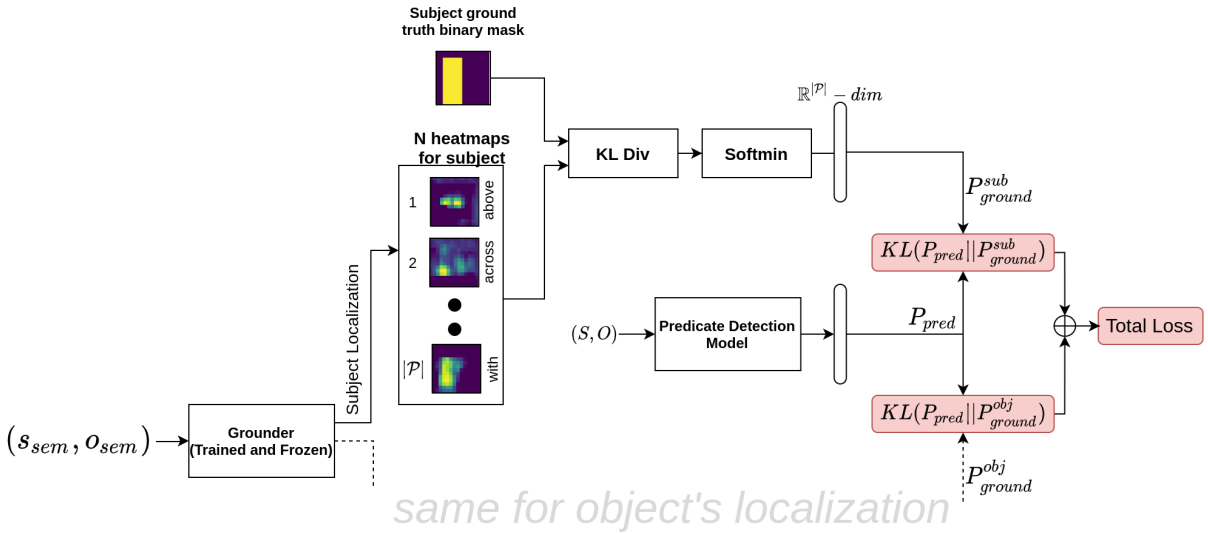

Figure 4.5: Εκπαίδευση δικτύου χρησιμοποιώντας τον grounder για να εξάγουμε την πιθανοφάνεια κάθε κλάσης. Κλάσεις που παράγουν χάρτες χαμηλής KL απόκλισης από τον πραγματικό έχουν υψηλή πιθανότητα και αντίστροφα. Συνολική συνάρτηση σφάλματος είναι η KL απόκλιση των κατανομών $P^{sub/obj}_{ground}$ με την κατανομή P_{pred} που προβλέπει το μοντέλο προς εκπαίδευση.

Η επιλογή μας εξαρτάται από τα χαρακτηριστικά των κατανομών $P^{sub/obj}_{ground}$ οι οποίες λόγω της φύσης του αντίστροφου προβλήματος (grounding) έχουν καλή ικανότητα στο να δίνουν χαμηλή πιθανότητα σε σχέσεις που δεν μπορούν να εξηγήσουν χωρικά, όμως αδυνατούν να ξεχωρίσουν ανάμεσα σε σχέσεις που μπορούν εξίσου καλά να εξηγήσουν χωρικά τη διάταξη των αντικειμένων. Για παράδειγμα, για έναν άνθρωπο που δεν κάθεται σε μία καρέκλα, η πιθανότητα της κλάσης "sit on" θα είναι χαμηλή όμως στην περίπτωση που κάθεται οι πιθανότητες των "sit on", "on" δε θα είναι τέτοιες ώστε να ευνοηθεί το bias του συνόλου δεδομένων. Είναι λογικό λοιπόν να επιλέξουμε την P_{pred} ως κατανομή στόχο και να χρησιμοποιήσουμε την KL απόκλιση ως `mode-seeking`.

4.4.3 Προσέγγιση συνέπειας χρησιμοποιώντας τον grounder ως παλινδρομητή

Πέρα από τη μέθοδο που περιγράψαμε στο κεφάλαιο 4.4.2 υπάρχει και ένας ακόμη τρόπος προσέγγισης της χρήσης της συνάρτησης g ο οποίος, όπως θα δούμε και στο κεφάλαιο 5 υπερτερεί. Αντί να υπολογίζουμε την έξοδο της συνάρτησης g για όλες τις σχέσεις, το κάνουμε μόνο για εκείνη που έχει προβλέψει το δίκτυο f που εκπαιδεύουμε. Θέλουμε η σχέση που προβλέπεται να μπορεί να εντοπίσει "επιτυχώς" τα αντικείμενα από τα οποία προβλέφθηκε. Για να συγκεκριμενοποιήσουμε το "επιτυχώς" ορίζουμε έναν αριθμό q τον οποίο αποκαλούμε ποιότητα (quality) εντοπισμού αντικειμένων και ορίζουμε ως:

$$q \triangleq \frac{max(\mathbf{h}_s \odot \mathbf{m}_s) + max(\mathbf{h}_o \odot \mathbf{m}_o)}{2} \tag{4.5}$$

όπου $\mathbf{m}_s, \mathbf{m}_o$ είναι $H \times W$ δυαδικοί χάρτες έντασης με τα πραγματικά κουτιά περιορισμού του υποκειμένου και του αντικειμένου αντίστοιχα (ground truth binary masks) και με $\odot$ συμβολίζουμε το hadamard γινόμενο. Το q υπολογίζει την κατά μέσο όρο ικανότητα του g να προβλέψει την πραγματική θέση του υποκειμένου και του αντικειμένου. Από τον ορισμό 4.5 μπορούμε να συμπεράνουμε ότι $q \in [0, 1]$. Συνεπώς, όσο μεγαλύτερο το q τόσο καλύτερα η σχέση που προβλέφθηκε μπορεί να εξηγήσει τη χωρική διάταξη των αντικειμένων και αντίστροφα.

Συμβολίζουμε με $\hat{p}$ την πιθανότητα της πρόβλεψης της κλάσης p και ελαχιστοποιούμε την εντροπία μεταξύ του $\hat{p}$ και του q. Ονομάζουμε αυτόν τον όρο σφάλματος GCL (Grounding Consis-

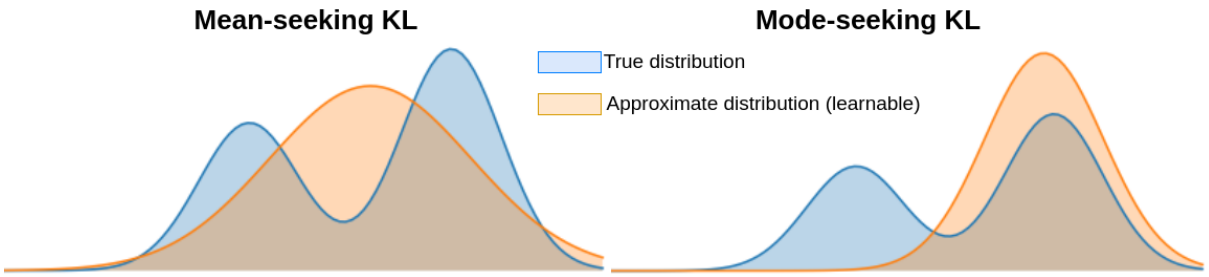

Figure 4.6: Έστω ότι χρησιμοποιούμε μία γκαουσιανή κατανομή(πορτοκαλί) για να μάθουμε ένα μείγμα δύο γκαουσιανών (μπλε) ελαχιστοποιώντας την KL απόκλισή τους. Ανάλογα με το αν θα επιλέξουμε η κατανομή που μαθαίνουμε να είναι η κατανομή στόχος ή όχι, μπορούμε να έχουμε δύο διαφορετικές συμπεριφορές της KL απόκλισης ως συνάρτησης σφάλματος. `Mean-seeking`: η κατανομή που μαθαίνουμε προσπαθεί να επεκτείνει τη μάζα της σε όλα τα τμήματα της αληθινής κατανομής (μπλε) που έχουν υψηλή πιθανότητα. `Mode-seeking`: η κατανομή που μαθαίνουμε προσπαθεί να συγκεντρώσει τη μάζα της σε τμήμα της πραγματικής κατανομής (μπλε) με υψηλή πιθανότητα.

tency Loss) και ορίζουμε:

$$\mathcal{L}_{GCL} = -[q\ log\ \hat{p} + (1-q)\ log\ (1-\hat{p})] \tag{4.6}$$

και επομένως η συνολική συνάρτηση κόστους γίνεται $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{GCL}$.

Διαισθητικά το $\mathcal{L}_{GCL}$ αναγκάζει το δίκτυο όταν προβλέπει κάποια κλάση p με μεγάλη βεβαιότητα (υψηλό $\hat{p}$), εκείνη να έχει την ικανότητα να εντοπίσει μέσω του g πάλι τα αντικείμενα από τα οποία προβλέφθηκε. Ομοίως, για προβλέψεις που έχουν χαμηλό q και δεν μπορούν να εξηγήσουν τη χωρική διάταξη των αντικειμένων το δίκτυο πρόβλεψης σχέσεων θα αναγκάζεται να μειώνει την πιθανότητα $\hat{p}$ και επομένως κατά τη διάρκεια της εκπαίδευσης άλλες κλάσεις θα αρχίσουν να προβλέπονται. Ακόμη υπενθυμίζουμε ότι ο όρος $\mathcal{L}_{GCL}$ μπορεί να εφαρμοστεί σε οποιοδήποτε δείγμα ανεξάρτητα από το αν έχει επισημείωση και μάλιστα χωρίς την απαίτηση κάποιας πρότερης γνώσης.

Σχηματικά η διαδικασία φαίνεται στην εικόνα 4.7 όπου δίνεται το παράδειγμα για ένα ζευγάρι ενός ανθρώπου και μίας μπλούζας που δεν φοράει. Αρχικά, το δίκτυο πρόβλεψης σχέσης (Rel. Detection) λόγω του context bias προβλέπει πως ο άνθρωπος φοράει ("wear") την μπλούζα. Έπειτα, το δίκτυο g (Grounding) προβλέπει πως ο άνθρωπος (Subject) θα έπρεπε να βρίσκεται πάνω στην μπλούζα για να τη φοράει και η μπλούζα (Object) πάνω στον άνθρωπο για να φοριέται από αυτόν. Όμως αυτοί οι χάρτες έντασης βλέπουμε πως είναι σε διαφωνία με τις πραγματικές θέσεις των αντικειμένων (GT Subject/Object). Έτσι δημιουργείται μεγάλο σφάλμα το οποίο διαδίδεται (backpropagates) για να ανανεώσει τα βάρη του δικτύου πρόβλεψης σχέσης.

4.4.4 Δίκτυο Grounding

Έπειτα από την εισαγωγή των μεθόδων που επιβάλουν συνέπεια μέσω grounding (grounding consistency) μένει να περιγράψουμε το δίκτυο g στο οποίο θα αναφερόμαστε ως grounder. Η αρχιτεκτονική του grounder αν και ανεξάρτητη της συνολικής λύσης επηρεάζει σε μεγάλο βαθμό την αποτελεσματικότητα του όρου $\mathcal{L}_{GCL}$. Ακόμη, υπάρχουν κάποιες προδιαγραφές που θέλουμε να πληρεί προκειμένου να είναι δυνατή η εφαρμογή του GCL σε οποιοδήποτε δείγμα.

Ας πάρουμε για παράδειγμα την εικόνα 4.8, δεδομένης της σχέσης `<person-on-street>` υπάρχουν δύο ζευγάρια `person-street` που μπορούν να την ικανοποιήσουν αφού και οι δύο άνθρωποι της εικόνας είναι "πάνω" στον δρόμο. Το να εντοπίσουμε το αντικείμενο `street` δεδομένης της τριπλέτας `<person-on-street>` και ενός ανθρώπου μπορεί να γίνει μόνο με έναν τρόπο αφού υπάρχει ένας μόνο δρόμος που θα είναι κάτω από αυτόν τον άνθρωπο. Το αντίστροφο όμως, δηλαδή να εντοπίσουμε τον άνθρωπο που είναι πάνω στον δρόμο δεδομένου του δρόμου, θα μπορούσε να έχει

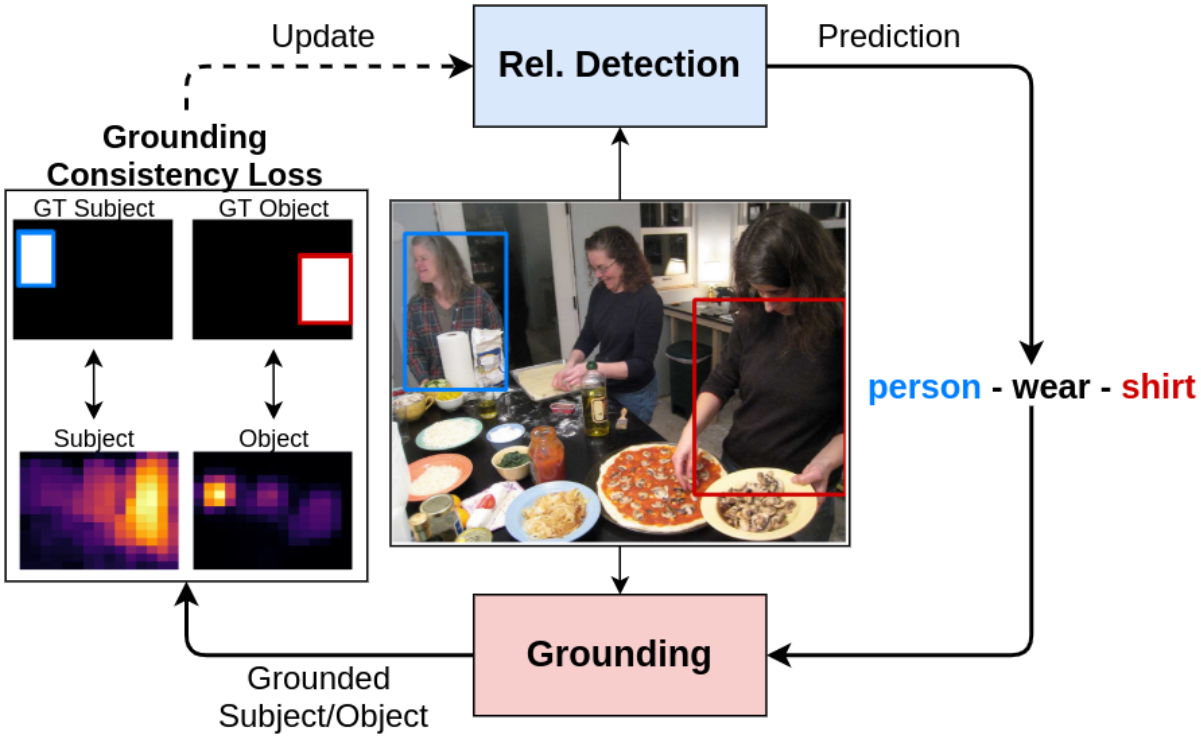

Figure 4.7: Ένα ζευγάρι ενός ανθρώπου και μίας μπλούζας που δε φοράει δίνονται ως είσοδο στο δίκτυο πρόβλεψης σχέσης. Σε περίπτωση biased πρόβλεψης όπως το "wear" ο grounder περιμένει την μπλούζα να βρίσκεται "πάνω" στον άνθρωπο (μπλε κουτί) και τον άνθρωπο "πάνω" στη μπλούζα (κόκκινο κουτί). Καθώς κάτι τέτοιο δεν ισχύει δημιουργείται μεγάλο σφάλμα, το οποίο ανανεώνει τα βάρη του δικτύου πρόβλεψης σχέσεων.

δύο πιθανές εκβάσεις, μία για κάθε άνθρωπο. Κατανοούμε λοιπόν πως η συνάρτηση g θα πρέπει να μπορεί να εντοπίσει και τους δύο ανθρώπους δεδομένης της τριπλέτας `<person-on-street>` ώστε όποιο ζευγάρι `person-street` και να διαλέξουμε το GCL να μπορεί να είναι αποτελεσματικό. Ο τρόπος με τον οποίο το πετυχαίνουμε αυτό είναι ανεξαρτητοποιώντας τον εντοπισμό του υποκειμένου και του αντικειμένου δεσμεύοντας την πρόβλεψη του υποκειμένου στο αντικείμενο και αντίστροφα. Αυτός ο διαχωρισμός έρχεται σε συμφωνία και με τις δύο συμπληρωματικές ερωτήσεις που αναφέραμε προηγουμένως: "που είναι το υποκείμενο (αντικείμενο) δεδομένης της σχέσης και της θέσης του αντικειμένου (υποκειμένου);".

4.4.4.1 Αρχιτεκτονική

Υπάρχουν αρκετές περιπτώσεις στη βιβλιογραφία που ασχολούνται με το πρόβλημα του grounding ([23, 47, 34, 17]). Στο [23] δεσμεύουν επίσης την πρόβλεψη του υποκειμένου στο αντικείμενο και αντίστροφα όμως η λύση τους δεν έχει τη δυνατότητα εντοπισμού πολλαπλών αντικειμένων. Αυτό καθιστά ιδιαίτερη την περίπτωσή μας και επιβάλλει τη δημιουργία ενός νέου μοντέλου. Παρακάτω θα περιγράψουμε τη διαδικασία εντοπισμού του αντικειμένου, δεδομένου του υποκειμένου και ομοίως ισχύει για το αντίστροφο. Η συνολική αρχιτεκτονική φαίνεται στην εικόνα 4.9.

Αναπαριστούμε τις κανονικοποιημένες συντεταγμένες του κέντρου του κουτιού περιορισμού του υποκειμένου ως $s_c = [s_{c,x}, s_{c,y}] \in \mathbb{R}^2$ και το κανονικοποιημένο πλάτος και ύψος ως $s_b = [s_{b,w}, s_{b,h}] \in \mathbb{R}^2$. Η γλωσσική πληροφορία του υποκειμένου και του αντικειμένου s_{sem}, o_{sem} μαζί με τη σχέση p που προβλέφθηκε από το δίκτυο πρόβλεψης σχέσης συνθέτουν την προβλεφθείσα τριπλέτα $t =< s_{sem}, p, o_{sem} >$. Η κατανομή $Pr(o_b, o_c)$ των διαστάσεων και των συντεταγμένων

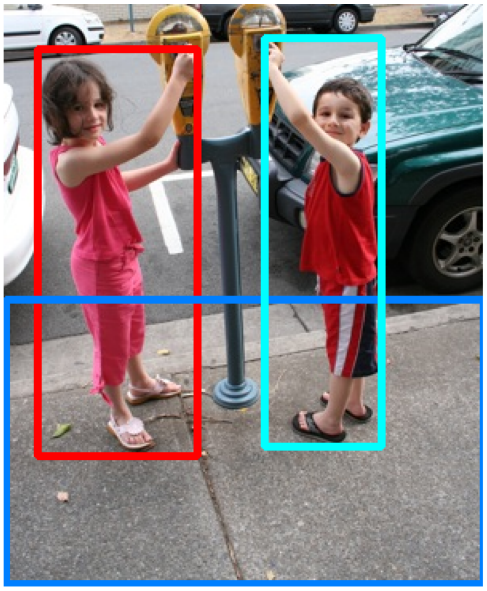

Figure 4.8: Περίπτωση όπου η τριπλέτα `<person-on-street>` θα μπορούσε να ικανοποιείται για διαφορετικούς ανθρώπους στην εικόνα. Σε αυτές τις περιπτώσεις ο grounder θα πρέπει να μπορεί να εντοπίζει όλα τα αντικείμενα που ικανοποιούν την τριπλέτα αναφοράς.

κέντρου του κουτιού του αντικειμένου μπορούν να μοντελοποιηθούν ως:

$$\begin{aligned} Pr(o_b, o_c) &= Pr(o_b, o_c|t, s_c, s_b) \\ &= Pr(o_c|t, s_c, s_b, o_b)Pr(o_b|t, s_c, s_b) \end{aligned}$$

και υποθέτουμε ότι οι διαστάσεις του αντικειμένου o_b είναι ανεξάρτητες από τη θέση του υποκειμένου s_c, το οποίο μας οδηγεί στην:

$$Pr(o_b, o_c) = Pr(o_c|t, s_c, s_b, o_b)Pr(o_b|t, s_b)$$

Εμπνεόμενοι από το [7], μοντελοποιούμε την $Pr(o_b|t, s_b)$ ως μία αλληλουχία γραμμικών επιπέδων (Linear Layers) με Relu για συνάρτηση ενεργοποίησης όπου δεδομένης της t και s_b, παλινδρομεί (regresses) πάνω στο κανονικοποιημένο πλάτος και ύψος του κουτιού του αντικειμένου δίνοντας ως έξοδο το $\hat{o_b}$. Αυτό το στάδιο επιβλέπεται με συνάρτηση κόστους ελαχίστων τετραγώνων (MSE Loss).

Για να μοντελοποιήσουμε την $Pr(o_c|t, s_c, s_b, \hat{o_b})$, παλινδρομούμε σε ένα χάρτη έντασης (heatmap regression) όπου κάθε στοιχείο εκφράζει την βεβαιότητα το κέντρο του αντικειμένου να βρίσκεται εκεί. Πρώτα κωδικοποιούμε την εικόνα μέσω του ResNet-50 [15] σε ένα χάρτη οπτικής αναπαράστασης (feature map) διαστάσεων $H \times W \times D$ το οποίο συμβολίζουμε ως $\mathbf{F}$, και H, W είναι οι οι χωρικές διαστάσεις ύψους και πλάτους αντίστοιχα ενώ D η διάσταση των διανυσμάτων αναπαράστασης. Έπειτα, υπολογίζουμε μία μάσκα προσοχής (attention mask) A_{att} διαστάσεων $H \times W$, οδηγούμενη από τη γλωσσική πληροφορία του αντικειμένου. Η προσοχή, ακολουθώντας σε γενικές γραμμές το [23], είναι το εσωτερικό γινόμενο του $\mathbf{F}$ με ένα διάνυσμα D διαστάσεων το οποίο μαθαίνει το δίκτυο ως προβολή της γλωσσικής πληροφορίας o_{sem} του αντικειμένου. Έπειτα αυτή η μάσκα πολλαπλασιάζεται με τον χάρτη αναπαράστασης ώστε να πάρουμε $\mathbf{F}_{att} = \mathbf{F} \star A_{att}$.

Προκειμένου να μπορέσουμε να αξιολογήσουμε όλες τις πιθανές θέσεις που θα μπορούσε να βρίσκεται ένα αντικείμενο που ικανοποιεί τη σχέση t, συνελίσουμε κάθε κανάλι (channel) του $\mathbf{F}_{att}$ με μία δυαδική μάσκα στις διαστάσεις $\hat{o_b}$ που προβλέφθηκαν. Αποτέλεσμα αυτής της πράξης είναι ένας χάρτης αναπαράστασης $\mathbf{F}_{gath}$ διαστάσεων $H \times W \times D$. Έτσι, είναι σαν κάθε διάνυσμα D διαστάσεων να συλλέγει (gathers) την οπτική πληροφορία που θα υπήρχε μέσα στο κουτί του αντικειμένου (έχοντας $\hat{o_b}$ διαστάσεις) αν το κέντρο του βρισκόταν σε εκείνη τη θέση του χάρτη αναπαράστασης.

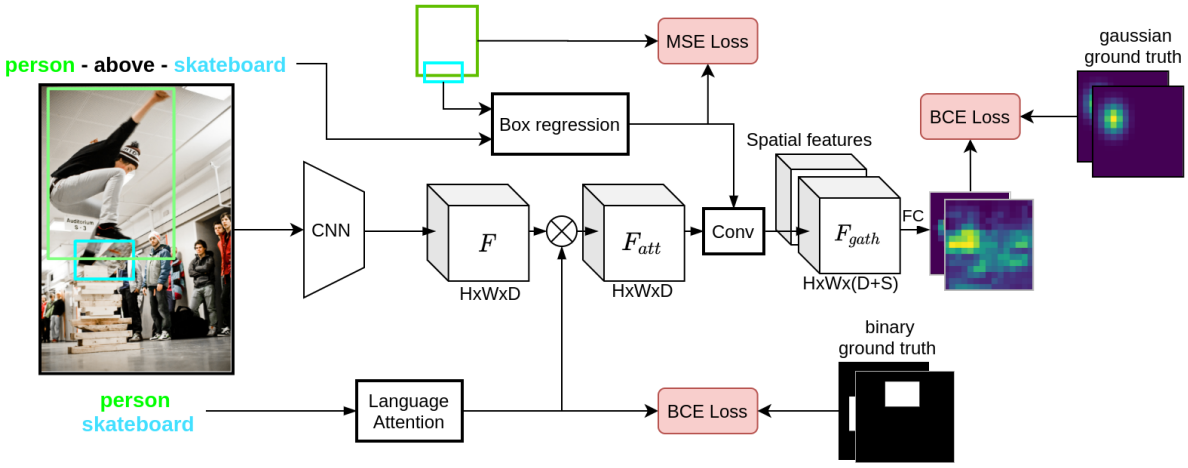

Figure 4.9: Η αρχιτεκτονική του grounding δικτύου. Κάνοντας αρχικά μία εκτίμηση των διαστάσεων του αντικειμένου που ψάχνουμε συνδυάζουμε οπτική με χωρική πληροφορία αξιολογώντας σε όλη την εικόνα την πιθανότητα ύπαρξης ενός ή περισσότερων αντικειμένων που ικανοποιούν τη δοθείσα σχέση.

Επειδή η χωρική πληροφορία είναι απαραίτητη για κάθε σχέση, συνενώνουμε το $\mathbf{F}_{gath}$ με ένα σύνολο χωρικών χαρακτηριστικών (spatial features) που χρησιμοποιούνται και στο [12] και αντιστοιχούν στη χωρική διάταξη του υποκειμένου (σταθερό) με την αντίστοιχη θέση του αντικειμένου σε κάθε πιθανή θέση. Αυτά τα χαρακτηριστικά έχουν διάσταση S. Ως αποτέλεσμα λοιπόν έχουμε ένα χάρτη χωρικής και σημασιολογικά οδηγούμενης οπτικής αναπαράστασης διαστάσεων $H \times W \times (D + S)$. Τέλος, μέσω γραμμικών επιπέδων με συναρτήσεις Relu προβάλουμε κάθε διάνυσμα $D+S$ διαστάσεων σε μία τιμή εύρους $[0, 1]$ που αξιολογεί το κατά πόσο το κέντρο του αντικειμένου βρίσκεται εκεί. Αυτός ο χάρτης έντασης της εξόδου επιβλέπεται μέσω της συνάρτησης δυαδικής εντροπίας (binary cross entropy) έχοντας ως πραγματικό χάρτη γκαουσιανές κατανομές με κέντρο το πραγματικό κέντρο του αντικειμένου και διασπορά ανάλογη των διαστάσεών του.

4.4.4.2 Ταυτόχρονη εκπαίδευση g και f

Μία απόφαση που πήραμε ήταν να προεκπαιδεύσουμε το δίκτυο g και αφού παγώσουμε τις παραμέτρους του να το χρησιμοποιήσουμε για να εκπαιδευτεί το δίκτυο εντοπισμού σχέσεων f. Ο λόγος είναι ότι η από κοινού εκπαίδευση δεν πρόκειται να ευνοήσει κανένα από τα δύο δίκτυα. Καθώς σκοπός του g είναι να δημιουργεί σφάλμα σε προβλέψεις που δεν εξηγούν τη χωρική διάταξη των αντικειμένων και το f πρέπει να μάθει να κάνει προβλέψεις που θα έχουν το ελάχιστο δυνατό σφάλμα είναι πιθανό να καταλήξουν σε κάποιο από κοινού ελάχιστο στο οποίο η συνάρτηση g θα προσπαθεί να διευκολύνει την f και αντίστροφα.

Κεφάλαιο 5

Πειράματα, Αποτελέσματα και Συγκρίσεις

Σε αυτό το κεφάλαιο αρχικά θα περιγράψουμε τα εργαλεία και τις παραμέτρους με τα οποία εκπαιδεύσαμε όλες τις παραπάνω μεθόδους που παραθέσαμε. Έπειτα, θα ορίσουμε δύο νέες παραλλαγές της κλασικής μέσης ακρίβειας (mean Precision) που εκμεταλλεύονται επιπλέον δείγματα αυξημένης δυσκολίας που εξάγουμε μέσω της μεθόδου αρνητικής συμπλήρωσης γράφου που εισηγάγαμε στο κεφάλαιο 4.1. Στη συνέχεια, θα πραγματοποιήσουμε τόσο ποσοτικές όσο και ποιοτικές συγκρίσεις σε μοντέλα της βιβλιογραφίας που επανυλοποιήσαμε αποδεικνύοντας την υπεροχή της μεθόδου GCL του κεφαλαίου 4.4.3 έναντι άλλων εναλλακτικών.

5.1 Εκπαίδευση

Υλισμικό/Λογισμικό Όλα τα μοντέλα εκπαιδεύτηκαν σε NVIDIA 2080Ti GPU σε σύστημα με 64GB RAM και Ubuntu 16.04. Μία εποχή εκπαίδευσης με χρήση του GCL διαρκεί κατά μέσο όρο (ανάλογα με την αρχιτεκτονική του μοντέλου) 9 λεπτά στο VRD και 75 λεπτά στο VG200 χρησιμοποιώντας 50% του ποσοστού των μη επισημειωμένων δεδομένων. Χωρίς GCL, οι χρόνοι μεταβαίνουν σε 3 και 55 λεπτά για τα VRD και VG200 αντίστοιχα. Αυτή η χρονική αύξηση που επιφέρει το GCL είναι δικαιολογημένη καθώς από τη μία αυξάνεται το σύνολο δεδομένων εκπαίδευσης λόγω χρήσης και μη επισημειωμένων δειγμάτων και από την άλλη χρησιμοποιείται επιπλέον και το grounding δίκτυο. Ωστόσο ο χρόνος συμπερασμού (inference time) δεν αλλάζει αφού ο grounder δεν είναι απαραίτητος εκεί. Κατά μέσο όρο οι εποχές εκπαίδευσης είναι 16. Κάθε μετρική που θα παρουσιάσουμε είναι το μέσο αποτέλεσμα από εκπαίδευση που πραγματοποιήθηκε 5 φορές. Οι μέγιστες διασπορές για το VRD είναι: R@50 0.007, mP 0.01 mP^+ 0.11, f-mP^+ 0.45. Για το VG200: R@50 0.002, mP 0.0005, mP^+ 0.005, f-mP^+ 0.15.

Επανυλοποιήσεις Για την υλοποίηση των μοντέλων χρησιμοποιήσαμε Pytorch. Επιλέξαμε μοντέλα διαφορετικών αρχιτεκτονικών και δυνατοτήτων προκειμένου να εξετάσουμε το κατά πόσο οι μέθοδοι είναι ανεξάρτητες των μοντέλων και για βελτιστοποίηση χρησιμοποιήσαμε τον αλγόριθμο Adam [21] με weight decay ίσο με 5×10^{-4} για το VRD και 5×10^{-5} για το VG200. Δεδομένου του ότι ερευνούμε την αποτελεσματικότητα συναρτήσεων σφάλματος λογικό είναι να συγκρίνουμε μεταξύ της ίδιας υλοποίησης των δικτύων προκειμένου να είναι συγκρίσιμα τα αποτελέσματα. Για αυτόν το λόγο υλοποιήσαμε τα VTransE [52], Motis-Net [51], RelDN [55] , ATR-Net [12], UVTransE [19], HGAT-Net [29]. 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
VTransE [52]	44.76	53.17	VRD
Motis-Net [51]	65.20	64.02	VG200
RelDN [55]	68.3	64.65	VG200
ATR-Net [12]	58.40	58.06	VRD
UVTransE [19]	55.5	56.88	VRD
HGAT-Net [29]	59.54	57.00	VRD

Table 5.1: Σύγκριση του R@50 που αναφέρεται από τους συγγραφείς των μοντέλων με τις επανυλοποιήσεις μας.

5.2 Εισαγωγή νέων μετρικών

Όπως αναφέραμε και στο κεφάλαιο 3 η προτιμώμενη μετρική του Recall είναι ανίκανη να αναδείξει το πρόβλημα του context bias αφού δεν λαμβάνει υπόψιν προβλέψεις για μη επισημειωμένα δείγματα. Από την άλλη το Precision αγνοεί τα μη επισημειωμένα δείγματα τα οποία δεν μπορούμε να θεωρήσουμε ως απλώς αρνητικά για όλες τις κλάσεις. Απαραίτητο λοιπόν για την αξιολόγηση και σύγκριση των προτεινόμενων μεθόδων είναι η εισαγωγή μίας νέας μετρικής. Εφόσον έχουμε ήδη σχεδιάσει τους δύο κανόνες στο κεφάλαιο 4.1 τους χρησιμοποιούμε για να συλλέξουμε αρνητικά δείγματα και να τα συμπεριλάβουμε στο σύνολο δεδομένων δοκιμής (test set). Αυτά τα δείγματα αποτελούν περιπτώσεις αυξημένης δυσκολίας που επιφέρουν μεγάλη ποινή για δίκτυα που αποκτούν context bias. Συμβολίζουμε τη μέτρηση της μέσης ακρίβειας στο επαυξημένο σύνολο δοκιμής mP^+. Ακόμη για να εστιάσουμε στις προβληματικές κλάσεις εγγύτητας περιορίζουμε τη μέτρηση του mP μόνο σε αυτές (βλ. πίνακας 4.1) και συμβολίζουμε με f-mP^+ (focused mP^+). Επίσης, συμπεριλαμβάνουμε και τον αρμονικό μέσο (HarMean) των R@50 και f-mP^+ ώς μίας συνολικής μετρικής αξιολόγησης των μοντέλων. Στην εικόνα 5.1 συγκρίνουμε τον αριθμό των θετικών και αρνητικών κλάσεων για κάποιες από τις πάσχουσες κλάσεις αφού επαυξήσουμε το σύνολο δοκιμών με αρνητικά δείγματα. Βλέπουμε πως υπάρχει μεγάλο πλήθος αρνητικών δειγμάτων σε σημείο που

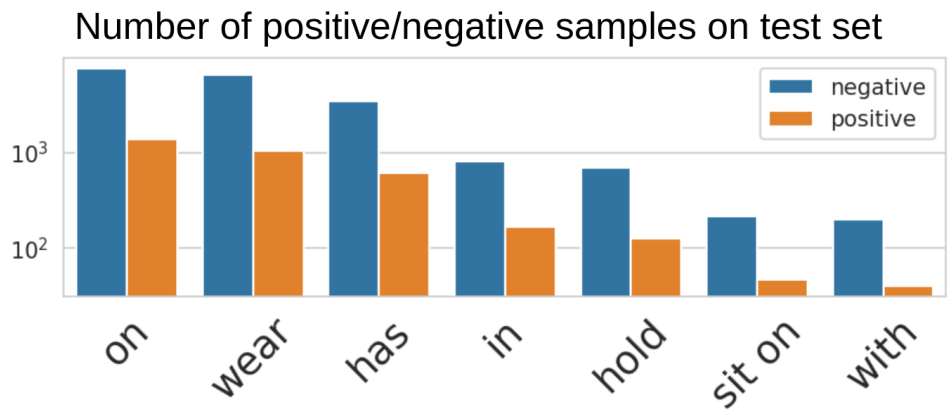

Figure 5.1: Σύγκριση αριθμού θετικών και αρνητικών δειγμάτων ορισμένων κλάσεων στο επαυξημένο σύνολο δεδομένων δοκιμής (test set).

ξεπερνούν ακόμη και τα θετικά. Στην εικόνα 5.2 παρατηρούμε για όλες τις κλάσεις και για τις σχέσεις εγγύτητας (σύμφωνα με τον πίνακα 4.1) ξεχωριστά την κατανομή του λόγου του εμβαδού της τομής υποκειμένου-αντικειμένου με το εμβαδόν του μικρότερου από τα δύο κουτιού. Αυτό μας δείχνει πως πράγματι οι σχέσεις εγγύτητας απομονώνουν ζευγάρια όπου το υποκείμενο είναι πολύ κοντά στο αντικείμενο και σχεδόν το καλύπτει.

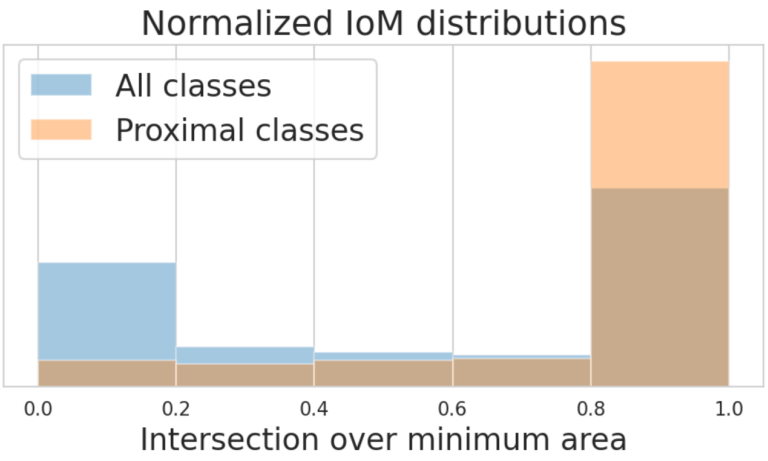

Figure 5.2: Κατανομή του λόγου του εμβαδού της τομής υποκειμένου-αντικειμένου με το εμβαδόν του μικρότερου από τα δύο κουτιού στο VRD. Παρατηρούμε πως για τις σχέσεις εγγύτητας που ορίσαμε το υποκείμενο με το αντικείμενο σχεδόν πάντα βρίσκονται το ένα μέσα στο άλλο.

5.3 Ποσοτικά αποτελέσματα

5.3.1 Ποσοτική σύγκριση όμοιων μεθόδων

Αρχικά θα κάνουμε μία ποσοτική σύγκριση όλων των μεθόδων που προτείναμε και αποσκοπούν στην αντιμετώπιση του context bias, καθώς και τη μόνη κοντινή [55] (βλ. κεφάλαιο 2.10.2). Οι μέθοδοι που συγκρίνονται στον πίνακα 5.2 είναι:

- SpatDistill: Προκειμένου να γίνει πληρέστερη η σύγκριση υλοποιήσαμε μέσω μεταβίβασης γνώσης [16] (knowledge distillation) μία συνάρτηση κόστους η οποία χρησιμοποιεί ένα δίκτυο εκπαιδευμένο μόνο με χωρικά χαρακτηριστικά (βλ. κεφάλαιο 2.10) ως δάσκαλο. Το σκεπτικό πίσω από αυτό είναι ότι δεδομένου του ότι τα μοντέλα αγνοούν τα χωρικά χαρακτηριστικά (βλ. κεφάλαιο 3.1), το να χρησιμοποιήσει κάποιος ένα δίκτυο που έχει κοινή λογική του χώρου ως δάσκαλο ώστε να τη μεταβιβάσει στο μαθητή φαίνεται μία λογική προσέγγιση.
- GraphL: μέδοδος του [55].
- NegRank: μέθοδος κεφαλαίου 4.3 με το δίκτυο του ranker $\mathcal{R}$ εκπαιδευμένο για όλες τις σχέσεις εγγύτητας.
- GCL^{KL}: μέθοδος κεφαλαίου 4.4.2.
- GCL: μέθοδος κεφαλαίου 4.4.3.
- oracle with NCE: μέθοδος κεφαλαίου 4.2.

Στον πίνακα 5.2 παρατηρούμε ότι το GCL υπερτερεί σε όλες τις μετρικές ενώ το NegRank ακολουθεί. Είναι φανερός ο συμβιβασμός που αναγκάζονται να κάνουν τα μοντέλα μεταξύ των Recall και Precision κάτι φυσικά αναμενόμενο καθώς αναγκάζονται να γίνουν πιο "συντηρητικά" και βασίζονται περισσότερο σε χωρικά χαρακτηριστικά για να κάνουν προβλέψεις. Συνολικά όμως είναι φανερό το μεγάλο κέρδος (26% σχετικό) στο f-mP^{+} σε σύγκριση με την πολύ μικρή μείωση του R@50 (-1% σχετικό) κάτι που επιβεβαιώνει και η αύξηση του HarMean.

Ιδιαίτερο ενδιαφέρον αποτελεί η διαφορά μεταξύ των GCL και GCL^{KL} καθώς πρόκειται για δύο πολύ κοντινές μεθόδους. Ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι η GCL^{KL} χρησιμοποιεί τον grounder για να κατατάξει τις κλάσεις με βάση το ποια μπορεί να κάνει καλή εκτίμηση της χωρικής διάταξης. Αυτομάτως λοιπόν, κλάσεις που χωρικά "μοιάζουν", δηλαδή ικανοποιούνται από

Model	R@50	mP$^+$	f-mP$^+$	HarMean
Spatial Baseline	47.08	20.09	32.87	38.71
ATR-Net	58.06	24.29	38.78	46.50
SpatDistill	<u>56.88</u>	27.02	44.58	49.98
GraphL	56.68	22.72	36.93	44.72
NegRank* (Ours)	54.44	<u>28.67</u>	<u>47.64</u>	<u>50.81</u>
GCL^{KL} (Ours)	53.69	25.6	42.27	47.30
GCL (Ours)	**57.21**	**29.43**	**48.97**	**52.77**
oracle with NCE* (Ours)	56.65	33.66	56.83	56.73

Table 5.2: Σύγκριση των μεθόδων που περιγράφηκαν και της βιβλιογραφίας στο VRD. Με * συμβολίζουμε τις μεθόδους που χρησιμοποιούν πρότερη γνώση πέραν του συνόλου δεδομένου εκπαίδευσης. Όλες οι συναρτήσεις κόστους έχουν εφαρμοστεί στο ATR-Net. Η NCE δεν συμμετέχει στη σύγκριση καθώς έχει πρότερη γνώση των αρνητικών δειγμάτων, σε αντίθεση με όλες τις άλλες. Μαζί παρουσιάζουμε και την επίδοση των Spatial Baseline και ATR-Net εκπαιδευμένα χωρίς κάποια επιπλέον συνάρτηση κόστους.

παρόμοιες χωρικές διατάξεις, θα πρέπει να ανταγωνιστούν. Όπως είναι αναμενόμενο, ο grounder δεν έχει μάθει το σημασιολογικό bias του συνόλου δεδομένων με αποτέλεσμα να μην μπορεί να κατατάξει σωστά κλάσεις που μοιάζουν ως προς τη χωρική διάταξη. Έτσι, οι κατανομές $P_{ground}^{sub/obj}$ που θα προκύψουν θα επηρεάσουν αρνητικά την κατανομή της εξόδου P_{pred} του δικτύου πρόβλεψης σχέσεων και θα το αποτρέψουν από το να μάθει το σημασιολογικό bias του συνόλου δεδομένων που όπως είδαμε στο κεφάλαιο 2.10 είναι ιδιαίτερα χρήσιμο για την αύξηση του Recall. Για τον ίδιο λόγο και η SpatDistill δεν καταφέρνει μεγάλα κέρδη. Ένα δίκτυο εκπαιδευμένο μόνο με χωρικά χαρακτηριστικά, όπως είδαμε και στο κεφάλαιο 2.10, αδυνατεί να μάθει χρήσιμα biases των σημασιολογικών δεδομένων και επομένως ως δάσκαλος αποτρέπει και το δίκτυο-μαθητή να τα μάθει μέσω της KL απόκλισης. Σε αντίθεση με το GCL^{KL} όμως, το GCL εκτιμά μεμονωμένα την συμβατότητα της σχέσης που προβλέφθηκε με τη διάταξη των αντικειμένων οπότε αποφεύγει αυτό το πρόβλημα.

Η oracle-NCE μέθοδος αδιαμφισβήτητα δίνει το μεγαλύτερο κέρδος όμως δεν πρέπει να αγνοούμε ότι χρησιμοποιεί αρνητικά δείγματα που εξορύχτηκαν από κανόνες που κατασκευάσαμε μέσω της ανθρώπινης διαίσθησης, κάτι που σε καμία περίπτωση δεν είναι κλιμακώσιμο. Σημειώνουμε επίσης ότι την καλύτερη επίδοση την πετύχαμε με μία πλήρως ημι-επιβλεπόμενη (semi-supervised) μέθοδο που δεν απαιτεί καμία απολύτως πρότερη γνώση ή ανθρώπινη επέμβαση.

5.3.2 Ποσοτικά αποτελέσματα του GCL

Στον πίνακα 5.3 παρουσιάζουμε την επίδραση του GCL σε ένα σύνολο επανυλοποιημένων μοντέλων. Το μέγιστο σχετικό κέρδος στο f-mP$^+$ είναι 42.2% στο VRD και 54% στο VG200. Παρατηρούμε ότι σε όλες τις περιπτώσεις το HarMean καθώς και τα mP$^+$, mP$^+$ αυξάνονται ενώ υπάρχει σχετικά μικρή πτώση στο Recall. Αξιοσημείωτο είναι το ATR-Net το οποίο καταφέρνει σε κάθε μετρική να βρίσκεται στις πρώτες δύο θέσης της κατάταξης. Σε αυτό το σημείο, παρατηρώντας τη στήλη του mP επιβεβαιώνουμε το γεγονός ότι οι παραδοσιακές μετρικές του Precision δεν αντικατοπτρίζουν το πρόβλημα του context bias ούτε την ποιοτικά επιβεβαιωμένη βελτίωση που επιφέρει το GCL.

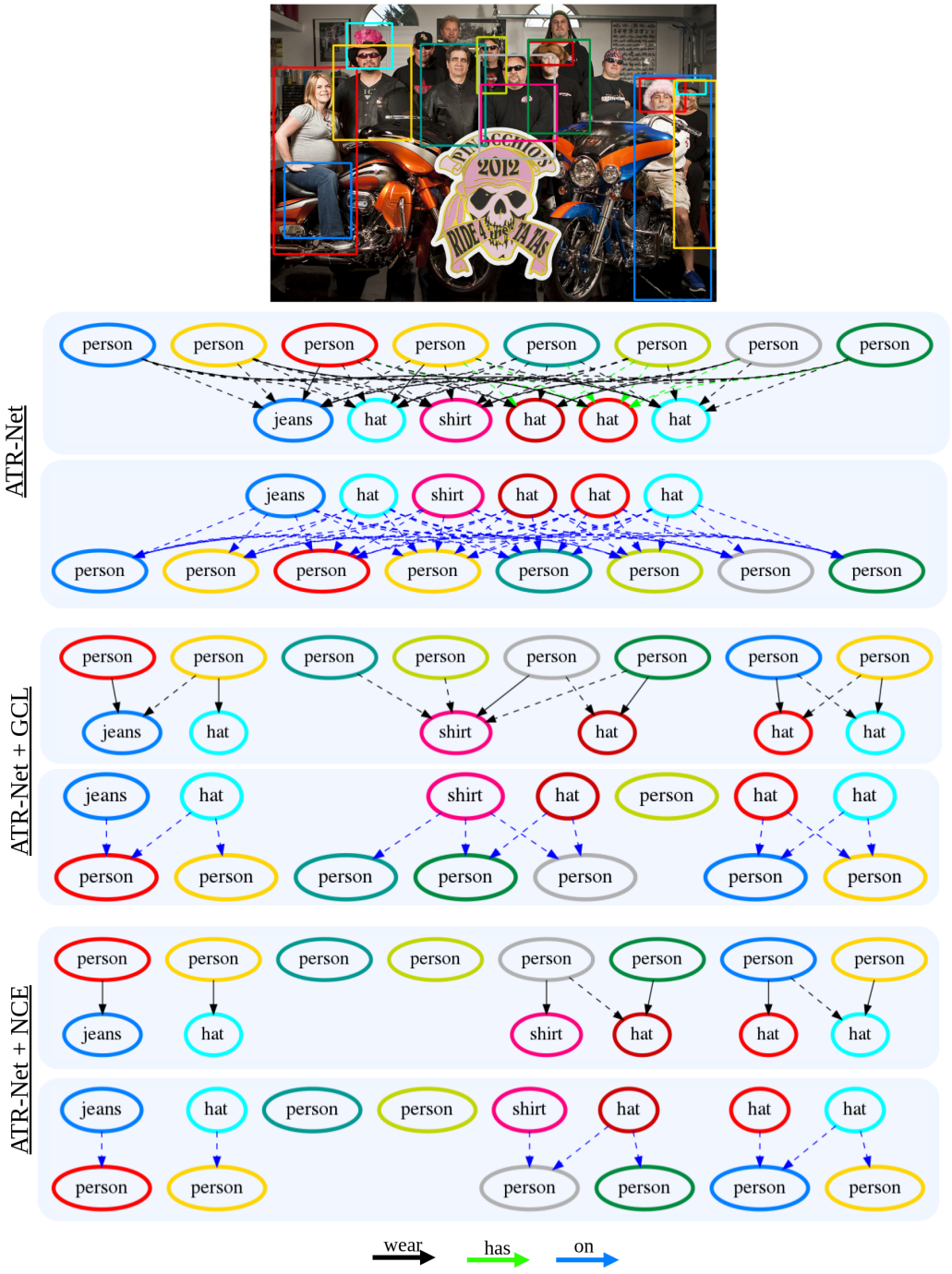

Figure 5.3: Ποιοτική σύγκριση εκπαίδευσης του ATR-Net μόνο με cross entropy, GCL και NCE. Χωρίς επιπλέον επίβλεψη το δίκτυο γίνεται πλήρως biased στο context (πάνω γράφος). Η GCL αραιώνει κατά πολύ τον γράφο δείχνοντας βελτιωμένη κατανόηση της σχέσης "wear" και "on". Η NCE λόγω του ότι χρησιμοποιεί εξωτερική γνώση καταφέρνει να ξεχωρίσει λίγο καλύτερα κοντινές περιπτώσεις από την GCL. Για λόγους ευκρίνειας παρουσιάζουμε τα ζευγάρια που ο άνθρωπος έχει μόνο τον ρόλο υποκειμένου και φιλτράρουμε τις χωρικές σχέσεις. Με διακεκομμένο βέλος είναι οι μη επισημειωμένες σχέσεις.

Models	VRD					VG200				
	R@50	mP	mP$^+$	f-mP$^+$	HarMean	R@50	mP	mP$^+$	f-mP$^+$	HarMean
VTransE [52]	53.17	13.11	17.42	26.95	35.77	68.98	6.8	9.13	16.08	26.08
Motis-Net [51]	55.06	13.31	20.67	32.38	40.78	69.4	6.83	10.19	17.98	28.56
RelDN [55]	55.02	13.66	22.94	36.63	43.98	64.65	6.3	9.42	16.6	26.42
ATR-Net [12]	**58.06**	**13.99**	24.29	38.78	46.50	**70.27**	**6.93**	10.51	18.53	29.33
UVTransE [19]	56.88	13.46	21.63	34.69	43.10	69.31	6.87	9.57	16.82	27.07
HGAT-Net [29]	57.00	13.84	22.46	36.26	44.32	70.20	6.92	10.29	18.14	28.83
VTransE + GCL	51.04	12.29	17.43	27.54	35.78	61.57	6.05	11.29	20.05	30.25
Motifs-Net + GCL	53.85	12.86	24.72	40.89	46.48	63.28	5.97	13.27	19.31	29.59
RelDN + GCL	54.11	13.29	26.96	44.55	48.87	61.37	5.97	12.93	23.00	33.46
ATR-Net + GCL	57.21	13.98	29.43	48.97	**52.77**	66.78	6.52	12.15	21.55	32.58
UVTransE + GCL	54.98	13.43	**29.47**	**49.33**	52.00	63.97	6.16	**14.53**	**25.91**	**36.88**
HGAT-Net + GCL	55.99	13.56	23.27	37.94	45.23	64.00	6.09	11.17	19.87	30.33

Table 5.3: Αποτελέσματα από τα έξι μοντέλα που επανυλοποιήσαμε με και χωρίς GCL για το VRD και το VG200. Αναφέρουμε Recall@50 (R@50), micro Precision (mP), mP$^+$, f-mP$^+$ και τον αρμονικό μέσο των R@50 and f-mP$^+$, με $^+$ συμβολίζουμε το επαυξημένο test set μέσω της αρνητικής συμπλήρωσης γράφου, f- τον υπολογισμό μόνο για τις σχέσεις εγγύτητας.

5.4 Ποιοτικά αποτελέσματα

5.4.1 Ποιοτική σύγκριση GCL με NCE

Στην εικόνα 5.3 μπορούμε να δούμε τα ποιοτικά αποτελέσματα των GCL και NCE. Καταρχήν είναι προφανής η μεγάλη βελτίωση στον γράφο σκηνής (scene graph) που προβλέπεται από το μοντέλο και στις δύο μεθόδους. Ο γράφος γίνεται πιο αραιός και πλέον φαίνεται πως κατανοείται σε καλύτερο επίπεδο η χωρική λογική των κλάσεων "wear" και "on". Οι περιπτώσεις αποτυχίας φαίνεται να είναι εκείνες όπου οι άνθρωποι είναι πολύ κοντά στα αντικείμενα. Όπως θα δούμε και στο κεφάλαιο 5.4.2 η διακριτική ικανότητα του grounder για πολύ κοντινά αντικείμενα δεν είναι μεγάλη. Έτσι λοιπόν τέτοιες περιπτώσεις δεν δημιουργούν αρκετά μεγάλο σφάλμα και το δίκτυο δεν αναγκάζεται να τις μάθει. Στην περίπτωση του NCE όμως, επειδή τα αρνητικά δείγματα προέρχονται από κανόνες θα δημιουργηθεί σφάλμα και για εκείνες τις κοντινές περιπτώσεις.

5.4.2 Ποιοτικά παραδείγματα για Grounding

Στην εικόνα 5.4 παρατηρούμε ποιοτικά αποτελέσματα της αρχιτεκτονικής που περιγράψαμε. Στην 5.4α βλέπουμε πως παρά το γεγονός ότι υπάρχουν δύο άνθρωποι που φορούν `jeans` (καθώς και δύο διαφορετικά `jeans` που φοριούνται από ανθρώπους) το δίκτυο καταφέρνει και εντοπίζει το ζευγάρι που πρέπει. Αυτό το πετυχαίνει μέσω της δέσμευσης του εντοπισμού της μίας οντότητας στην άλλη όπως περιγράψαμε παραπάνω. Στην 5.4γ παρατηρούμε το αποτέλεσμα για την εικόνα 4.8 όπου και οι δύο άνθρωποι εντοπίζονται. Αυτό είναι επιθυμητό καθώς θα μπορούσαμε να εφαρμόσουμε το GCL και στα δύο ζευγάρια `person-street`. Ακόμη βλέπουμε στην 5.4β πως υπάρχει δυσκολία στον εντοπισμό του ανθρώπου και των πέδιλων του σκι λόγω της εγγύτητας των αντικειμένων. Αυτές οι περιπτώσεις δεν αντιμετωπίζονται πάντα αποτελεσματικά με αποτέλεσμα να έχουμε προβλέψεις που μπορεί να παραπλανήσουν το GCL και να μειώσουν την αποδοτικότητα της μεθόδου όπως θα δούμε και στο κεφάλαιο 5.4.3.

Στην εικόνα 5.5 παραθέτουμε τυχαία παραδείγματα όπου υπερθέτουμε εικόνες με το αντίστοιχο A_{att} που παράγει εσωτερικά ο grounder. Αξίζει να σχολιάσουμε ότι παρά το γεγονός ότι η κλίμακα κάθε στοιχείου του χάρτη είναι πολύ μικρότερη της κλίμακας του αντικειμένου που θέλουμε να εντοπίσουμε, η προσοχή που μαθαίνει το δίκτυο έχει την ικανότητα να εντοπίσει ακόμη και μικρά τμήματα του συνόλου. Για παράδειγμα ψάχνοντας το αντικείμενο `dog` μπορεί να εντοπίσει για ένα μικρό τμήμα του ποδιού του ότι πρόκειται για μέρος ενός σκύλου. Αυτό είναι αποτέλεσμα της μεθόδου FPN (Feature Pyramid Network [26]) που υλοποιείται για το Faster R-CNN [35]. Ένα

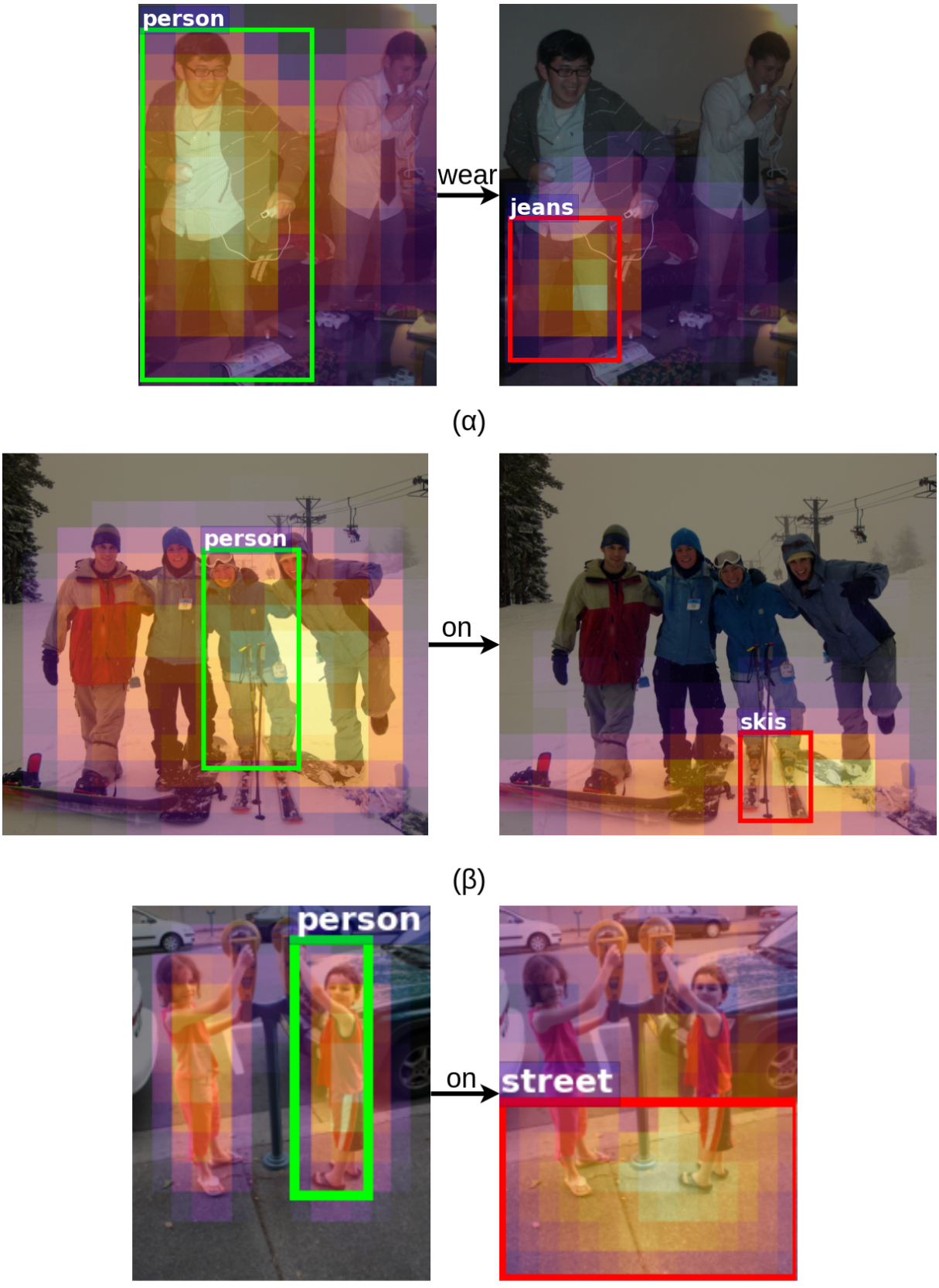

Figure 5.4: Ποιοτικά αποτελέσματα του grounder. Αριστερά τα υποκείμενα, δεξιά τα αντικείμενα με επισημειωμένα τα κουτιά της αναφερόμενης σχέσης.

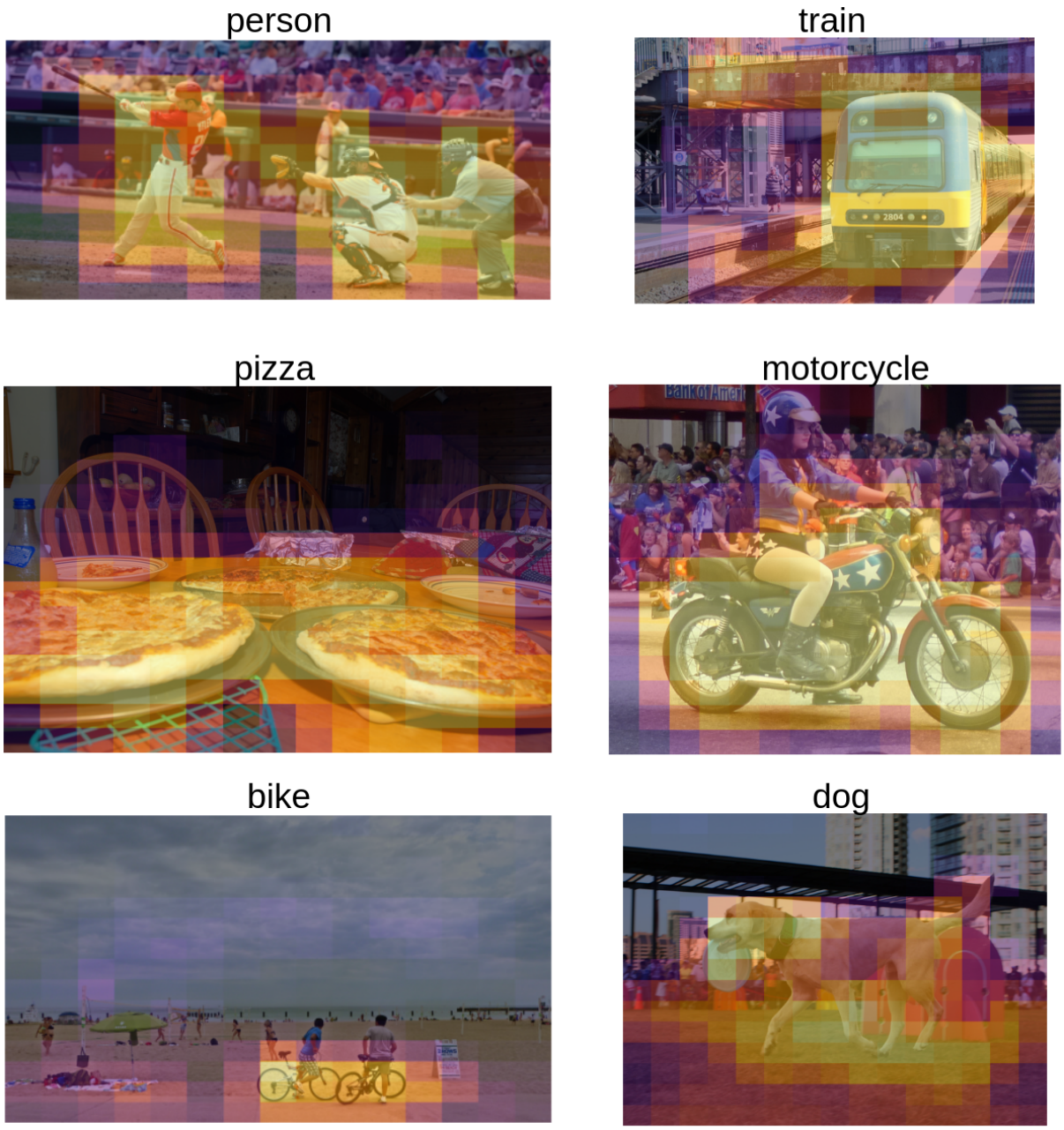

Figure 5.5: Παραδείγματα της γλωσσικά οδηγούμενης προσοχής (attention) που περιγράψαμε στο 4.4.4.1.

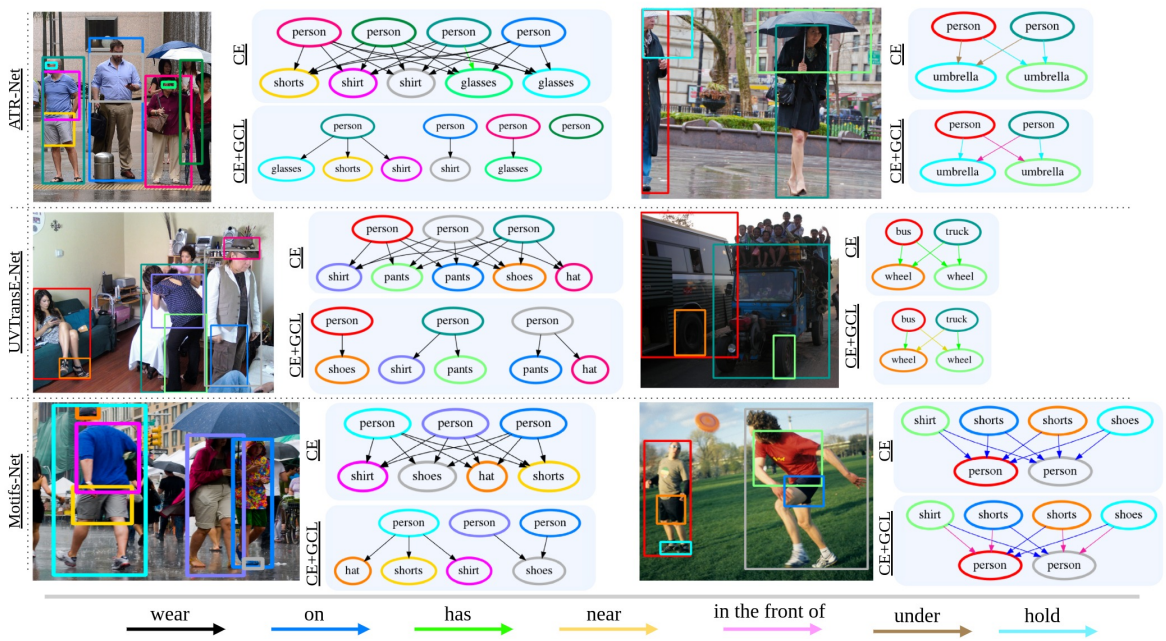

Figure 5.6: Ποιοτικά παραδείγματα της επίδρασης του GCL. Στην αριστερή στήλη για λόγους ευκρίνειας φιλτράρουμε τις ακμές χωρικών σχέσεων ενώ στη δεξιά στήλη τις επιτρέπουμε.

FPN καταφέρνει να αναπαραστήσει διανύσματα πλούσια σε σημασιολογική πληροφορία η οποία δεν χάνεται όσο μειώνεται η κλίμακα.

5.4.3 Ποιοτικά αποτελέσματα GCL

Στην εικόνα 5.6 παρουσιάζουμε μερικά ακόμη ποιοτικά αποτελέσματα του GCL για τα ATR-Net, UVTransE-Net και Motifs-Net. Βλέπουμε (αριστερή στήλη) πως οι γράφοι σχηματίζουν συνεκτικές συνιστώσες που αποτελούνται από τους ανθρώπους και τα αντικείμενα με τα οποία αλληλεπιδρούν. Ακόμη παρατηρούμε (δεξιά στήλη) πως έπειτα της εφαρμογής του GCL οι biased προβλέψεις αντικαθίστανται από χωρικές και όχι από συνώνυμες π.χ. το `<person-wear-shirt>` δεν γίνεται `<person-has-shirt>`.

Ακόμη, στην εικόνα 5.7 παρουσιάζουμε μέσω του sliding box experiment την επίδραση του GCL στη λογική του χώρου που αναπτύσσουν τα μοντέλα. Πλέον υπάρχει μία συγκεντρωμένη περιοχή πάνω στον άνθρωπο που προβλέπεται ως "wear" σε αντίθεση με πριν όπου το πως συνδέεται ο άνθρωπος με την μπλούζα χωρικά ήταν αδιάφορο για το μοντέλο.

Μία ακόμη σημαντική παρατήρηση αφορά την επίδραση του GCL σε ένα δίκτυο που χρησιμοποιεί μόνο οπτική πληροφορία (βλ. κεφάλαιο 2.10). Όπως βλέπουμε και στην εικόνα 5.8 ένα τέτοιο δίκτυο καταφέρνει να εξάγει έμμεσα τη χωρική πληροφορία από τα οπτικά χαρακτηριστικά. Συμπεραίνουμε λοιπόν πως με τα κατάλληλα δεδομένα και τη σωστή επίβλεψη, τα μοντέλα μπορούν να αξιοποιήσουν ακόμη μεγαλύτερη πληροφορία. Το Visual Baseline προφανώς και πριν λάμβανε την ίδια πληροφορία όμως ποτέ δεν "αναγκάστηκε" να τη χρησιμοποιήσει.

5.4.4 Ποιοτική εξήγηση της πτώσης του Recall

Θέλουμε να δώσουμε μία ποιοτική εξήγηση για τη μικρή πτώση του Recall που επιφέρει το GCL. Επιλέγουμε την κλάση "next to" του VRD η οποία δέχεται τη μεγαλύτερη μείωση Recall έπειτα την εφαρμογή του GCL και παρουσιάζουμε στην εικόνα 5.9 την κατανομή των προβλέψεων επισημειωμένων ως "next to" στο test set. Ιδανικά θα θέλαμε να μην υπάρχει πτώση της καμπύλης στο "next to". Παρόλα αυτά βλέπουμε πως αρχίζουν να προβλέπονται περισσότερες κλάσεις όπως

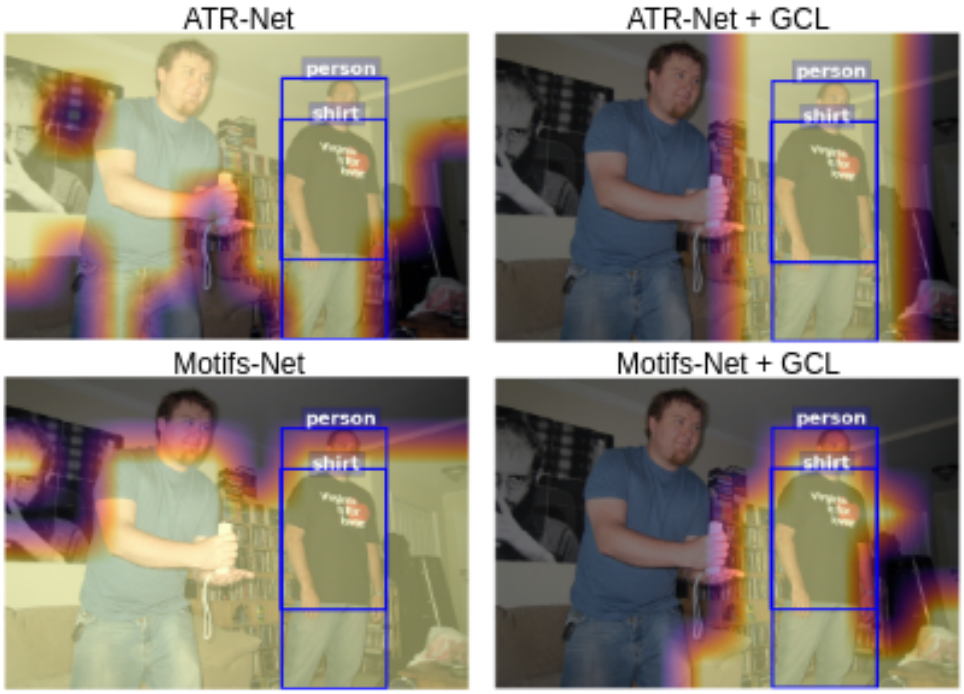

Figure 5.7: Το πείραμα του sliding box για τα ATR-Net και Motifs-Net με και χωρίς GCL. Μετακινούμε το κέντρο του κουτιού του `shirt` και σημειώνουμε σε ποια σημεία προβλέπεται η κλάση "wear" . Ενώ στην αρχή (αριστερά στήλη) δεν υπάρχει χωρικός περιορισμός με το GCL κατανοείται πολύ καλύτερα ότι η σχέση "wear" προϋποθέτει εγγύτητα.

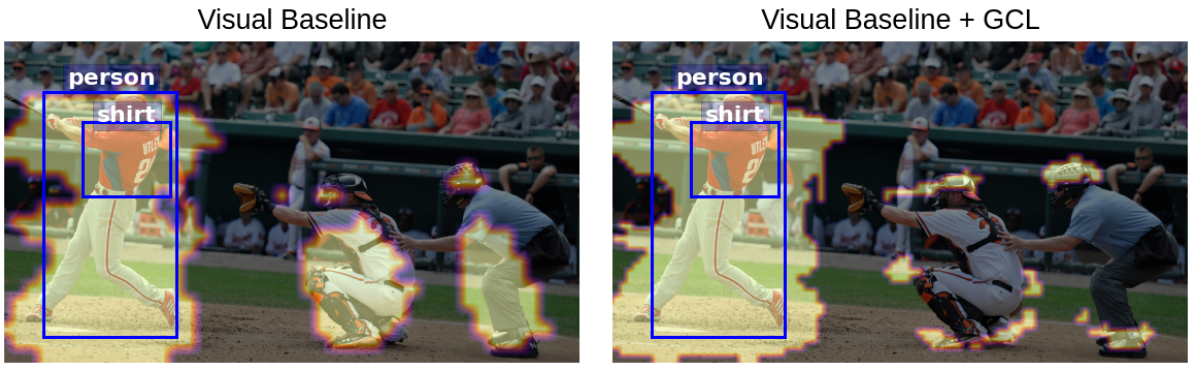

Figure 5.8: Το πείραμα του sliding box το Visual Baseline με και χωρίς GCL. Μετακινούμε το κέντρο του κουτιού του `shirt` και σημειώνουμε σε ποια σημεία προβλέπεται η κλάση "wear". Παρατηρούμε πως με χρήση του GCL ακόμη και ένα δίκτυο χωρίς χωρική πληροφορία καταφέρνει να μάθει μέσω της οπτικής πληροφορίας να κατανοεί τη χωρική διάταξη των οντοτήτων.

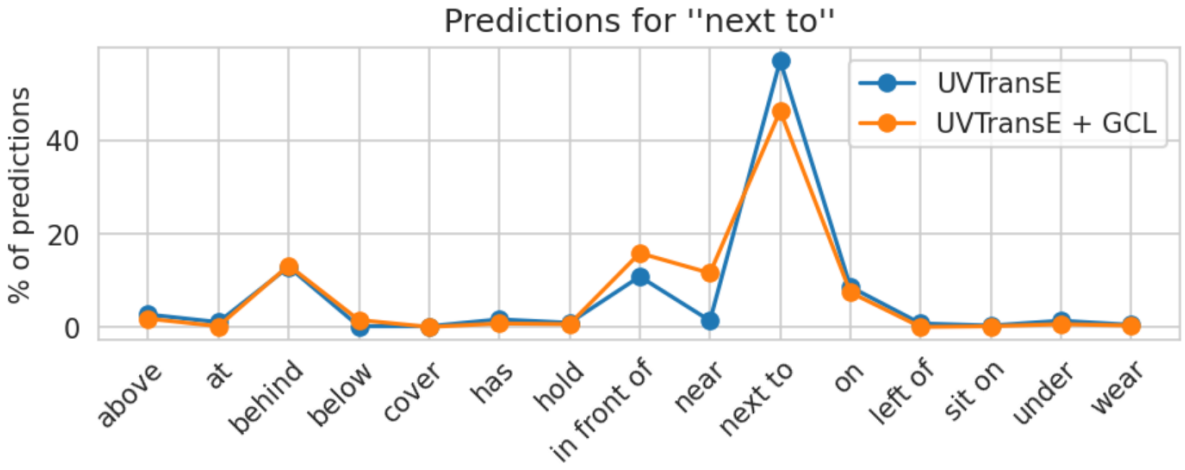

Figure 5.9: Κατανομή προβλέψεων για τα δείγματα που είναι επισημειωμένα ως "next to" στο test set για το UVTransE με και χωρίς GCL. Τα δείγματα που δεν προβλέπονται ως "next to" μειώνονται όμως αντικαθίστανται από νοηματικά κοντινές κλάσεις όπως "near", "in the front of".

"near", "in the front of" οι οποίες είναι πολύ κοντά νοηματικά. Ποιοτικά λοιπόν φαίνεται η πτώση του Recall να μην έχει ουσιαστικό αντίκτυπο στην ποιότητα του μοντέλου.

5.5 Ποσοστό μη επισημειωμένων δειγμάτων στην εκπαίδευση

Εφόσον αναφερόμαστε σε ημι-επιβλεπόμενη μέθοδο, μία σημαντική παράμετρος είναι το ποσοστό των δεδομένων εκπαίδευσης που είναι μη επισημειωμένα. Εκπαιδεύουμε το ATR-Net και ξεκινώντας από 0% ποσοστό μη επισημειωμένων κινούμαστε ανοδικά. Στην εικόνα 5.10 παρατηρούμε πως όσο περισσότερα δείγματα χωρίς επισημείωση χρησιμοποιούμε, τόσο αυξάνεται το f-mP$^+$ με μία μικρή μείωση του R@50. Ο λόγος είναι ότι σιγά σιγά εξισορροπείται το context bias και μάλιστα όσο ανεβάζουμε το ποσοστό όλο και μειώνεται ο ρυθμός που επωφελείται το f-mP$^+$ καθώς τα νέα δείγματα δεν έχουν να προσφέρουν αρκετά καινούρια πληροφορία. Αφού χρησιμοποιήσουμε όλα τα δεδομένα που έχουμε στη διάθεσή μας, για να συνεχίσει να αυξάνεται το ποσοστό των μη επισημειωμένων πρέπει να αρχίσουμε να μειώνουμε τα επισημειωμένα. Σε αυτό το σημείο (διακεκομμένη γραμμή) βλέπουμε πως απότομα ξεκινά να μειώνεται το R@50 σε αντίθεση με το f-mP$^+$ το οποίο παραμένει υψηλό. Ο λόγος πίσω από αυτή τη συμπεριφορά είναι ότι η μη επισημειωμένη πληροφορία έχει τη δυνατότητα να αναγκάσει το δίκτυο να κάνει προβλέψεις που τουλάχιστον έχουν χωρική λογική. Πρακτικά όσο διατρέχουμε το γράφημα από αριστερά προς τα δεξιά, πηγαίνουμε από μία ημι-επιβλεπόμενη μέθοδο σε μία μη-επιβλεπόμενη.

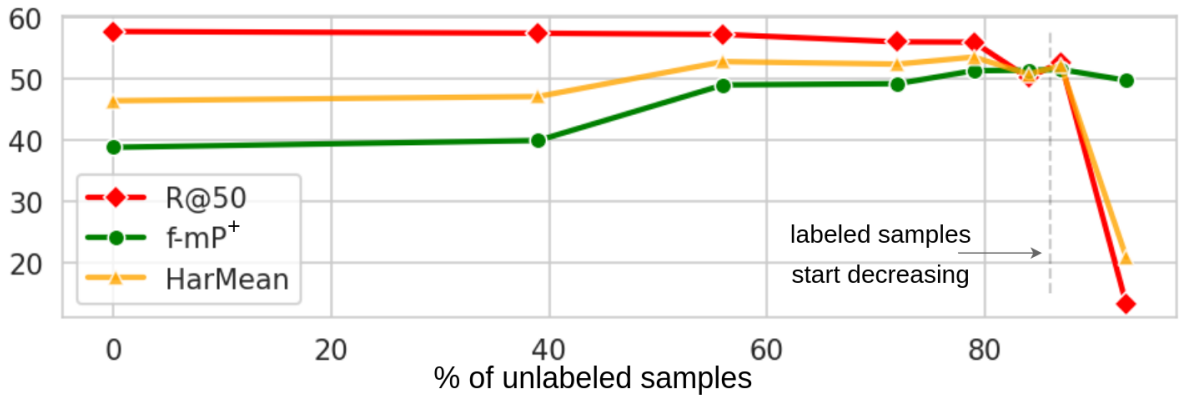

Figure 5.10: Παράθεση R@50, f-mP$^+$ και HarMean για διαφορετικά ποσοστά χρήσης μη επισημειωμένων δειγμάτων. Όσο αυξάνονται τα μη επισημειωμένα δείγματα, υπάρχει και αύξηση του f-mP$^+$ ενώ ελάχιστη είναι η πτώση του R@50. Σημειώνουμε με διακεκομμένη γραμμή το σημείο στο οποίο ξεκινούν να μειώνονται τα επισημειωμένα δείγματα που χρησιμοποιούνται.

Κεφάλαιο 6

Επίλογος και μελλοντικές επεκτάσεις

6.1 Επίλογος

Σε αυτή τη διπλωματική προσπαθήσαμε να συμβάλουμε στο πρόβλημα της αναγνώρισης οπτικών σχέσεων μέσω της έκθεσης, ανάλυσης και επίλυσης ενός σημαντικού προβλήματος που μέχρι στιγμής διέφευγε στην κοινότητα ρίχνοντας έτσι φως σε μία ως τώρα ανεξερεύνητη πτυχή του προβλήματος.

Μέσω παρατήρησης αποκαλύψαμε πως η μετρική του Recall δεν αρκεί για την πλήρη αξιολόγηση των μοντέλων καθώς το bias που κρύβουν τα μοντέλα σε ορισμένες κλάσεις δεν αναδεικνύεται. Ακόμη, ανατρέψαμε την πεποίθηση ότι μόνο οι κλάσεις με πολύ λίγα δείγματα (tail) είναι αυτές που πάσχουν από έλλειψη δεδομένων και αποδείξαμε πειραματικά ότι ακόμη και οι κλάσεις με τα περισσότερα δείγματα είναι πολύ μακρυά από το να κατανοηθούν από τα μοντέλα. Δείξαμε λοιπόν έτσι τη σημασία διερεύνησης των μη επισημειωμένων δειγμάτων.

Αρχικά, για τη θεμελίωση του **context bias** εισηγάγαμε το πείραμα του κυλιόμενου παραθύρου (**sliding box experiment**) με το οποίο μελετήσαμε ποιοτικά την επίδραση κάθε είδους πληροφορίας αποκαλύπτοντας πως τα μοντέλα μέχρι σήμερα για πολλές σχέσεις δεν "κοιτούσαν" την εικόνα. Αυτό το στηρίξαμε με στατιστικά καθώς και με μία αυτοματοποιημένη μέθοδο εντοπισμού των σχέσεων που πάσχουν βασισμένη στον υπολογισμό της εντροπίας (**entropy ranking**) των κατανομών ανά ζεύγος υποκειμένου-αντικειμένου.

Για την αντιμετώπιση του προβλήματος του context bias στραφήκαμε στη χρήση **αρνητικών δειγμάτων** τα οποία όπως δείξαμε υπάρχουν σε αφθονία στο σύνολο των μη επισημειωμένων ζευγαριών. Με βάση την μέθοδο της εντροπίας εντοπίσαμε το σύνολο των σχέσεων που πάσχουν και αφού τις χωρίσαμε σε δύο κατηγορίες σχεδιάσαμε δύο κανόνες εξαγωγής αρνητικών δειγμάτων (**negative graph completion**). Έπειτα ως πρώτη προσέγγιση, δημιουργήσαμε μία μέθοδο που εφαρμόσαμε αρνητική εντροπία (**NCE**) στα αρνητικά δείγματα παίρνοντας πολύ καλά αποτελέσματα προς την επίλυση του προβλήματος. Έπειτα, και επειδή σκοπός μας ήταν μία μέθοδος που δεν απαιτεί πρότερη γνώση και ανθρώπινη επέμβαση, δημιουργήσαμε τη μέθοδο κατάταξης αρνητικότητας (**NR**). Εκεί, χωρίς πλέον να χρησιμοποιούμε τους κανόνες, εκπαιδεύσαμε ένα δίκτυο στο να μάθει πότε ένα δείγμα είναι αρνητικό για μία κλάση που ορίσαμε.

Πιο σημαντική όμως ήταν η εισαγωγή της ιδέας της grounding συνέπειας (**Grounding consistency**) με σκοπό τη δημιουργία μίας μεθόδου όπου καμία πληροφορία σχετικά με το πρόβλημα δε θα ενσωματωνόταν στη διαδικασία της εκπαίδευσης. Πράγματι, η συνάρτηση κόστους **GCL** κατάφερε καλύτερα από όλες τις προηγούμενες να βελτιώσει κατά πολύ την ακρίβεια των μοντέλων και χωρίς καθόλου χρήση πρότερης γνώσης για το πρόβλημα. Κλειδί στο σχεδιασμό της ήταν το να επιτρέψουμε στο δίκτυο να μάθει το χρήσιμο bias που χρειάζεται για την επίτευξη υψηλού Recall. Για να το πετύχουμε αυτό επιβάλαμε μεμονωμένα στην πιθανότητα κάθε πρόβλεψης του δικτύου εντοπισμού σχέσεων να είναι συνεπής με τη δυνατότητα επαναπροδιορισμού των αντικειμένων στην εικόνα. Η

αρχιτεκτονική κλειστού βρόχου της λύσης μας, επέτρεψε την εφαρμογή της συνάρτησης κόστους στα μη επισημειωμένα δείγματα κάτι το οποίο ήταν κομβικής σημασίας για την αποτελεσματικότητά της.

Για την εφαρμογή του GCL, αναγκαστήκαμε να λύσουμε και το αντίστροφο πρόβλημα του εντοπισμού σχέσεων (grounding). Σχεδιάσαμε μία νέα αρχιτεκτονική που από μία τριπλέτα `<subject-predicate-object>` εντοπίζει στην εικόνα τα συσχετιζόμενα αντικείμενα.

Τέλος για την αξιολόγηση και σύγκριση όλων των ιδεών μας δημιουργήσαμε δύο νέες παραλλαγές της κλασικής μέσης ακρίβειας mP^+ και f-mP^+. Αυτές με χρήση των αρνητικών δειγμάτων που εξορύξαμε κατάφεραν να αντικατοπτρίσουν καλύτερα το πρόβλημα του context bias αλλά και της βελτίωσης που επέφεραν οι μέθοδοί μας. Έτσι πραγματοποιήσαμε εκτενείς ποσοτικές και ποιοτικές συγκρίσεις σε πληθώρα μοντέλων που επανυλοποιήσαμε για να καταλήξουμε στο ότι πράγματι οι προτεινόμενες μέθοδοι αποτελούν μία ισχυρή προσπάθεια επίλυσης των προβλημάτων που εκθέσαμε.

6.2 Μελλοντικές επεκτάσεις

Παρά την αποτελεσματικότητα των μεθόδων μας οι μετρικές που εισηγάγαμε δείχνουν πως υπάρχει ακόμη μεγάλο περιθώριο για βελτίωση. Η αύξηση του f-mP^+ από 38.78% σε 56.83% είναι σίγουρα μία καλή αρχή όμως τι κρύβεται στο υπόλοιπο 43.17% των περιπτώσεων που αποτυγχάνουμε; Στην ανάλυσή μας αναφέραμε πως μία βασική πηγή λαθών είναι περιπτώσεις όπου οι οντότητες βρίσκονται σε μεγάλη εγγύτητα μεταξύ τους. Αυτό όμως θα πρέπει να διερευνηθεί καθώς ίσως υπάρχουν και άλλα αίτια αστοχίας.

Κάτι ακόμη που θα μπορούσε να αποτελέσει αντικείμενο μελλοντικής έρευνας είναι η προσπάθεια βελτίωσης των χαρτών έντασης που παράγονται από το πείραμα του κυλιόμενου κουτιού. Όπως είδαμε μία μπλούζα μπορεί να φορεθεί σε όλη την έκταση ενός ανθρώπου. Αν και σε σχέση με πριν αυτό αποτελεί μεγάλη πρόοδο, ιδανικά θα θέλαμε να περιορίσουμε ακόμη περισσότερο αυτήν την περιοχή σε σημεία όπου πράγματι υπάρχει μία μπλούζα που την φοράει ο άνθρωπος.

Μία άλλη κατεύθυνση βελτίωσης αποτελεί το δίκτυο του grounder. Κατά την ανάλυσή του αναφέραμε πως αν και αποτελεσματικός, σε καμία περίπτωση δεν είναι τέλειος. Κάποια διαφορετική αρχιτεκτονική ίσως κατάφερνε να αποδώσει μεγαλύτερη χωρική ακρίβεια λύνοντας έτσι τα προβλήματα εγγύτητας που αναφέραμε.

Επιπλέον, εφαρμογή των μεθόδων και σε καινούρια σύνολα δεδομένων θα ήταν χρήσιμη καθώς ίσως αποκάλυπτε νέα προβλήματα και λύσεις. Ιδέες όπως η χρήση δικτύων για παραγωγή νέων δειγμάτων ίσως να καταφέρουν να βελτιώσουν το πολύ μικρό ποσοστό επισημειωμένων δειγμάτων που έχουμε στη διάθεσή μας.

Κάτι που διαπιστώσαμε κατά τη διάρκεια ενασχόλησης με το πρόβλημα είναι πως μέθοδοι με μικρό βαθμό επίβλεψης έχουν καλύτερη προοπτική επιτυχίας. Ο λόγος είναι ότι από τη φύση του προβλήματος μπορούμε να έχουμε πολλά δεδομένα με λίγες όμως επισημειώσεις. Οπότε τρόποι εκμετάλλευσης των μη επισημειωμένων σχέσεων θα ήταν προτιμώμενοι.

Παραρτήματα

A Σύνολα δεδομένων

Υπάρχει πληθώρα συνόλων δεδομένων για το πρόβλημα του εντοπισμού σχέσεων [28, 24, 52, 8, 43, 54]. Στον πίνακα 1 παρουσιάζουμε τα βασικά στατιστικά τους. Τα σύνολα δεδομένων τα

Dataset	Train/Test images	Predicates	Objects
VRD[28]	4k/1k	70	100
VG-MSDN[24]	46.2k/10k	50	150
VG-VTE[52]	73.8k/25.8k	100	200
sVG[8]	64.7k/8.7k	24	399
VG200[43]	75.6k/32.4k	50	150
VG80K[54]	99.9k/4.8k	29086	53304

Table 1: Απαρίθμηση των συνόλων δεδομένων της βιβλιογραφίας με τις χαρακτηριστικές στατιστικές πληροφορίες τους.

οποία χρησιμοποιούμε για τη διεξαγωγή των πειραμάτων είναι τα VRD και VG200, δύο από τα πιο διαδεδομένα στη βιβλιογραφία. Στις εικόνες 1 και 2 παρουσιάζουμε σε λογαριθμική κλίμακα τον αριθμό των δειγμάτων ανά κλάση στο σύνολο δεδομένων εκπαίδευσης.

B Παραλλαγές προβλήματος

Υπάρχουν πέντε παραλλαγές του προβλήματος εντοπισμού οπτικών σχέσεων τις οποίες και παραθέτουμε στον πίνακα 2. Οι τρεις πηγές πληροφορίας που τις καθορίζουν είναι: οι συντεταγμένες των κουτιών περιορισμού (Object boxes), οι κατηγορίες των αντικειμένων τους (Object categories) και ποιες ποια ζευγάρια σχέσεων έχουν επισημειωθεί (Interactions). Συγκεκριμένα:

Problem	Object boxes	Object categories	Interactions
PredDet	yes	yes	yes
PredCls	yes	yes	no
SGCls	yes	no	no
SGGen	no	no	no
PhrDet	no	no	no

Table 2: Απαρίθμηση των παραλλαγών της ανίχνευσης οπτικών σχέσεων. `yes` σημαίνει πως η συγκεκριμένη παραλλαγή χρησιμοποιεί την αντίστοιχη πληροφορία ενώ σε αντίθετη περίπτωση σημειώνουμε `no`.

- Predicate Detection (PredDet): Δοθέντων των κουτιών, των κατηγοριών των αντικειμένων και των ζευγαριών που αλληλεπιδρούν προβλέπουμε τις σχέσεις μεταξύ τους.

- Predicate Classification (PredCls): Δοθέντων των κουτιών, των κατηγοριών των αντικειμένων αποφασίζουμε ποια ζευγάρια αλληλεπιδρούν και για αυτά προβλέπουμε σχέσεις.
- Scene Graph Classification (SGCls): Δοθέντων των κουτιών των αντικειμένων, τα κατηγοριοποιούμε, βρίσκουμε ποια αλληλεπιδρούν και προβλέπουμε σχέσεις.
- Scene Graph Generation (SGGen): Τίποτα δεν είναι γνωστό. Εντοπίζουμε και κατηγοριοποιούμε τα αντικείμενα, βρίσκουμε ποια αλληλεπιδρούν και προβλέπουμε σχέσεις.
- Phrase Detection (PhrDet): Ομοίως με SGGen μόνο που αξιολογεί την επικάλυψη του κουτιού της ένωσης του υποκειμένου και του αντικειμένου (να είναι δηλαδή > 0.5) αντί το ξεχωριστό τους γινόμενο.

C Μετρικές

Η πιο συνηθισμένη μετρική είναι το 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 [12] 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 $k = 1$.

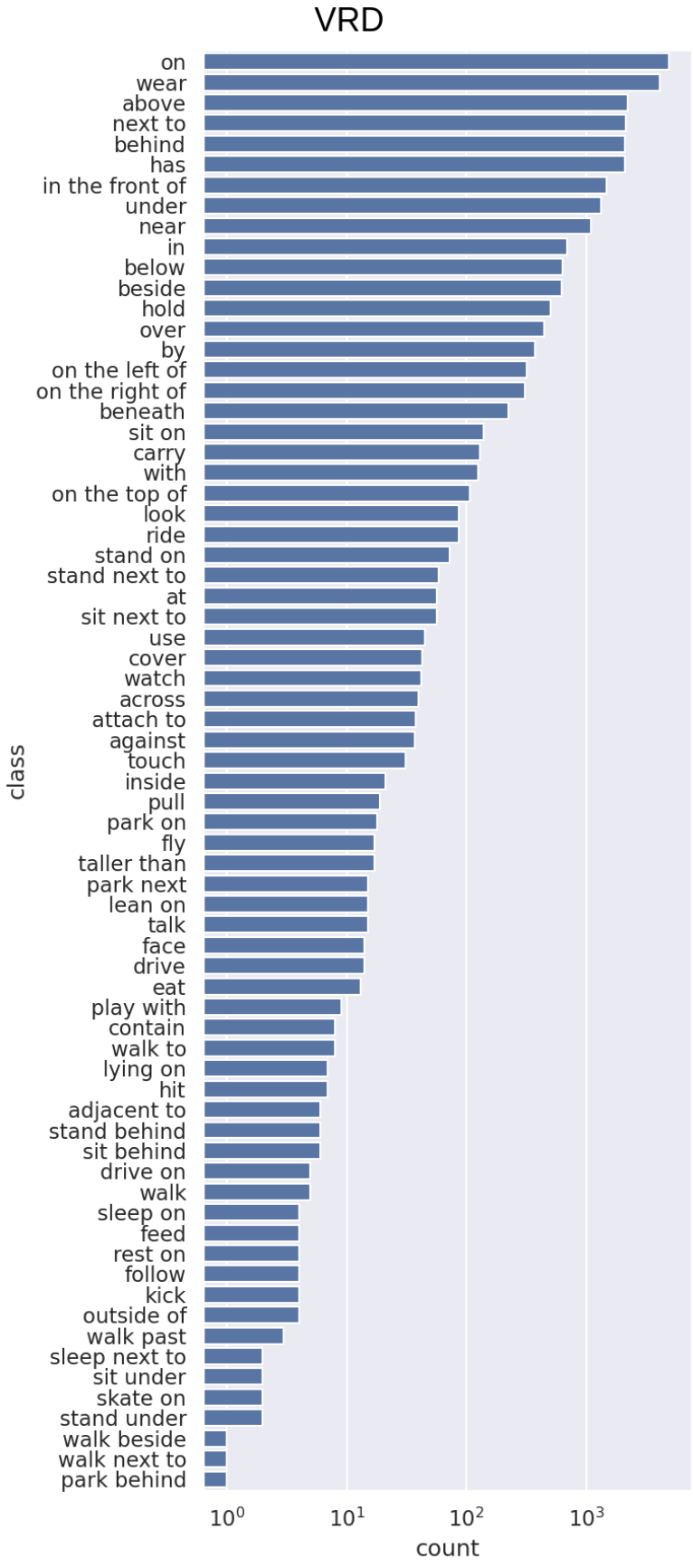

Figure 1: Αριθμός δειγμάτων ανά κλάση σε λογαριθμική κλίμακα στο σύνολο δεδομένων εκπαίδευσης για το VRD.

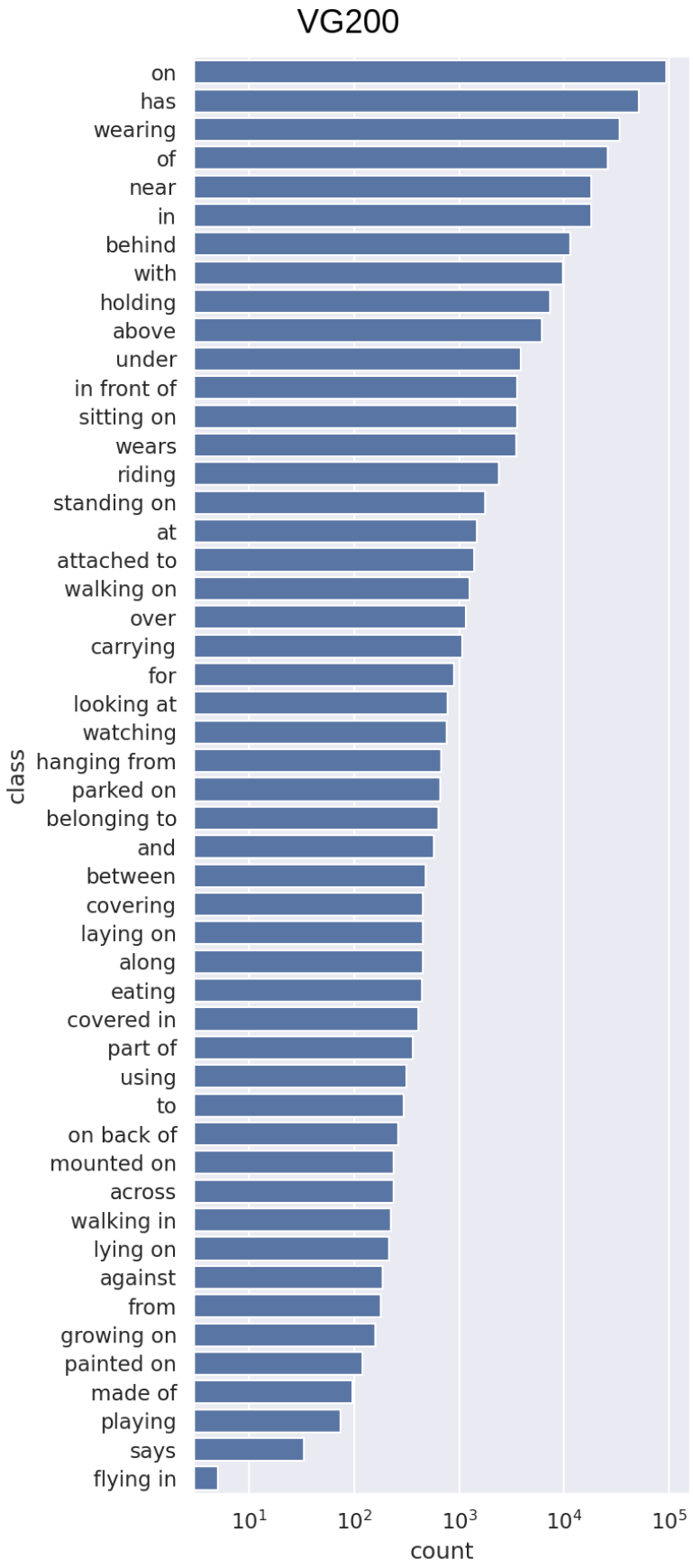

Figure 2: Αριθμός δειγμάτων ανά κλάση σε λογαριθμική κλίμακα στο σύνολο δεδομένων εκπαίδευσης για το VG200.

Bibliography

[1] S. Abdelkarim, P. Achlioptas, J. Huang, B. Li, K. W. Church, and M. Elhoseiny, "Long-tail Visual Relationship Recognition with a Visiolinguistic Hubless Loss," *ArXiv*, vol. abs/2004.00436, 2020.

[2] F. Baldassarre, K. Smith, J. Sullivan, and H. Azizpour, "Explanation-based Weakly-supervised Learning of Visual Relations with Graph Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[3] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," in *Proc. NIPS*, 2013.

[4] G. Bouritsas, P. Koutras, A. Zlatintsi, and P. Maragos, "Multimodal visual concept learning with weakly supervised techniques," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-Embedded Routing Network for Scene Graph Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] V. S. Chen, P. Varma, R. Krishna, M. Bernstein, C. Ré, and L. Fei-Fei, "Scene Graph Prediction With Limited Labels," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] G. Collell, L. Van Gool, and M.-F. Moens, "Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[8] B. Dai, Y. Zhang, and D. Lin, "Detecting Visual Relationships with Deep Relational Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic Image Manipulation Using Scene Graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] A. Dornadula, A. Narcomey, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop on Scene Graph Representation and Learning*, 2019.

[11] S. Garg, J. R. A. Moniz, A. Aviral, and P. Bollimpalli, "Learning to Relate from Captions and Bounding Boxes," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[12] N. Gkanatsios, V. Pitsikalis, P. Koutras, and P. Maragos, "Attention-Translation-Relation Network for Scalable Scene Graph Generation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop on Scene Graph Representation and Learning*, 2019.

[13] N. Gkanatsios, V. Pitsikalis, P. Koutras, A. Zlatintsi, and P. Maragos, "Deeply Supervised Multimodal Attentional Translation Embeddings for Visual Relationship Detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[14] N. Gkanatsios, V. Pitsikalis, and P. Maragos, "From Saturation to Zero-Shot Visual Relationship Detection Using Local Context," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[16] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *ArXiv*, vol. abs/1503.02531, 2015.

[17] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 10 294–10 303.

[18] Z. Hu, X. Ma, Z. Liu, E. H. Hovy, and E. P. Xing, "Harnessing Deep Neural Networks with Logic Rules," *ArXiv*, vol. abs/1603.06318, 2016.

[19] Z. Hung, A. Mallya, and S. Lazebnik, "Contextual translation embedding for visual relationship detection and scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1–1, 2020. DOI: 10.1109/TPAMI.2020.2992222.

[20] Y. Kim, J. Yim, J. Yun, and J. Kim, "NLNL: Negative Learning for Noisy Labels," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[22] M. Klawonn and E. Heim, "Generating Triples With Adversarial Networks for Scene Graph Construction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[23] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei, "Referring Relationships," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene Graph Generation from Objects, Phrases and Region Captions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[25] X. Liang, L. Lee, and E. P. Xing, "Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph Property Sensing Network for Scene Graph Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual Relationship Detection with Language Priors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[29] L. Mi and Z. Chen, "Hierarchical Graph Attention Network for Visual Relationship Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Weakly-Supervised Learning of Visual Relations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[31] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Detecting Unseen Visual Relations Using Analogies," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[32] F. Plesse, A. Ginsca, B. Delezoide, and F. J. Prêteux, "Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[33] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive Relational Networks for Mapping Images to Scene Graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] M. Raboh, R. Herzig, G. Chechik, J. Berant, and A. Globerson, "Differentiable Scene Graphs," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[37] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. arXiv: 1602.07261. [Online]. Available: http://arxiv.org/abs/1602.07261.

[38] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased Scene Graph Generation from Biased Training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] R. Vedantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh, "Learning Common Sense through Visual Abstraction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[40] W. Wang, M. Wang, S. Wang, G. Long, L. Yao, G. Qi, and Y. A. Chen, "One-Shot Learning for Long-Tail Visual Relation Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[41] X. Wang, Q. Sun, T.-S. Chua, and M. H. Ang, "Generating Expensive Relationship Features from Cheap Objects," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[42] J. M. Wolfe, "Visual Memory: What Do You Know About What You Saw?" *Current Biology*, vol. 8, R303–R304, 1998.

[43] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene Graph Generation by Iterative Message Passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[44] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[45] Y. Xu, C. Xu, C. Xu, and D. Tao, "Multi-positive and unlabeled learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[46] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for Scene Graph Generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[47] S. Yang, G. Li, and Y. Yu, "Cross-Modal Relationship Inference for Grounding Referring Expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[48] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. C. Loy, "Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[49] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[50] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging Knowledge Graphs to Generate Scene Graphs," in *Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[51] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural Motifs: Scene Graph Parsing with Global Context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[52] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual Translation Embedding Network for Visual Relation Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[53] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, "PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[54] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. M. Elgammal, and M. Elhoseiny, "Large-Scale Visual Relationship Understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[55] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical Contrastive Losses for Scene Graph Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[56] B. Zhuang, L. Liu, C. Shen, and I. D. Reid, "Towards Context-Aware Interaction Recognition for Visual Relationship Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.