ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

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

Παραγωγή Περιγραφικών Γράφων Σκηνής Χρησιμοποιώντας Ασθενή Επίβλεψη σε Περιγραφές Εικόνων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μπενετάτου Αλέξανδρου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και
Επεξεργασίας Σημάτων

Παραγωγή Περιγραφικών Γράφων Σκηνής Χρησιμοποιώντας Ασθενή Επίβλεψη σε Περιγραφές Εικόνων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μπενετάτου Αλέξανδρου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Ιουλίου 2022.

.....	.....	.....
Πέτρος Μαραγκός	Αθανασιος Ροντογιαννης	Γεράσιμος Ποταμιάνος
Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΠΘ

Αθήνα, Ιούλιος 2022.

.....
Αλέξανδρος Μπενετάτος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Το πρόβλημα παραγωγής γράφων σκηνής (scene graph generation) του τομέα της όρασης υπολογιστών αφορά την εξαγωγή κατευθυνόμενων γράφων ως αναπαράσταση των σχέσεων (ακμές) μεταξύ των αντικειμένων (κόμβοι) σε μία εικόνα.

Παρατηρώντας τη συμπεριφορά σύγχρονων μοντέλων στη βιβλιογραφία σε εικόνες με επισημειωμένα δείγματα, γίνεται σαφές πως τα μοντέλα που εκπαιδεύουμε δυσκολεύονται να ξεχωρίσουν ποιες από τις πιθανές σχέσεις είναι πιο σημαντικές για την περιγραφή της εικόνας. Μάλιστα, αυτό δεν οφείλεται σε κάποιο πρόβλημα εκπαίδευσης καθώς, πολύ συχνά, τα μοντέλα θα προβλέψουν τις σχέσεις που είναι επισημειωμένες, ωστόσο ακόμα και αυτές δεν θα παρέχουν σημαντική πληροφορία για την εικόνα. Θα αναφερόμαστε στην ικανότητα των μοντέλων να εντοπίσουν ποιες από τις πιθανές σχέσεις είναι πιο σημαντικές για την περιγραφή της εικόνας ως saliency και, από όσο γνωρίζουμε, είμαστε οι πρώτοι που αναφερόμαστε σε αυτό το χαρακτηριστικό.

Η συνεισφορά αυτής της διπλωματικής αφορά τόσο τη μέτρηση του saliency ενός Scene Graph Generation (SGG) μοντέλου όσο και τη παραγωγή πιο salient γράφων σκηνής σύμφωνα με ποιοτικά και ποσοτικά αποτελέσματα που εξάγουμε. Συγκεκριμένα:

- Εισάγουμε μια γενικευμένη μέθοδο εκπαίδευσης SGG μοντέλων με ασθενή επίβλεψη χρησιμοποιώντας περιγραφές εικόνων.
- Εισάγουμε δύο παραλλαγές της μέτρησης του Recall@N όπου, με χρήση των περιγραφών εικόνων, μπορούμε να εξάγουμε μετρήσεις για το saliency SGG μοντέλων.
- Πραγματοποιούμε τόσο ποσοτική όσο και ποιοτική σύγκριση μεταξύ των μεθόδων που προτείνουμε και με τη σχετική βιβλιογραφία στο VG200, το δημοφιλέστερο σύνολο δεδομένων του προβλήματος όπου πετυχαίνουμε 35% μέγιστη σχετική βελτίωση συγκριτικά με επαναϋλοποίηση της SOTA μεθόδου.

Θεμελιώνουμε, λοιπόν, την αιτία έλλειψης saliency στους γράφους σκηνής, προτείνουμε μετρικές για την αξιολόγηση του saliency ενός μοντέλου και τέλος σχεδιάζουμε μια μέθοδο εκπαίδευσης μοντέλων ώστε αυτά να αντιλαμβάνονται καλύτερα την έννοια του saliency και να παράγουν πιο ουσιώδεις γράφους σκηνής.

Τα παραπάνω τονίζουν την ανάγκη παραγωγής περιγραφικών γράφων σκηνής και αναδεικνύουν την ανάγκη αλλαγής προσανατολισμού στην αντιμετώπιση του προβλήματος. Η χρήση πλήρως επιβλεπόμενων μεθόδων, δυστυχώς, δεν κλιμακώνονται καλά σε αυξημένο αριθμό από κατηγορίες αντικειμένων ή σχέσεων. Αλλά ακόμα και σε μικρότερα λεξιλόγια, εξαιτίας της αραιής μη-περιγραφικής επισημείωσης, οδηγούμαστε σε μεροληπτικά μοντέλα που δεν κατανοούν την εικόνα και αδυνατούν να εντοπίσουν τη σημαντική πληροφορία σε αυτή.

Λέξεις Κλειδιά Παραγωγή Γράφων Σκηνής, Ασθενής Επίβλεψη, Περιγραφικότητα, Παραγωγή Γράφων Σκηνής από Περιγραφές Εικόνων, COCO, VG200, Open Images

Abstract

Scene graph generation is a computer vision task regarding the generation of a directed graph as a representation of relations (edges) and object entities (nodes) in an image

Observing the behavior of state-of-the-art models in images with labeled data, it is clear that the models we train find it difficult to separate which of the possible relations are more important for the description of the image. In fact, this is not due to some problem in training since, very often, the models will predict relations that are labeled, though even those do not provide important information for the image. We will refer to the ability of the models to identify which of the possible relations are more important for the description of the image as saliency and, as far as we know, we are the first to study this characteristic of the SGG models.

This thesis contributes to the quantification of the saliency of a Scene Graph Generation (SGG) model as well as the generation of more salient scene graphs, as shown from qualitative and quantitative results we gather. Specifically we:

- Introduce a generalized method for training SGG models with weak supervision using image captions.
- Introduce two variations of the common Recall@N metric with which, using image captions, we can calculate measurements regarding the saliency of SGG models
- Perform quantitative and qualitative comparison between the methods we propose and the relative literature in VG200, the most common dataset for SGG where we achieve 35% maximum relative improvement compared to the re-implementation of the SOTA method for weakly supervised training with image captions.

So, we establish the reason for lack of saliency in scene graphs, we introduce metrics to evaluate the saliency of a model and lastly we propose a generalized method for training SGG models that better incorporate the concept of saliency and generate more descriptive scene graphs.

The above emphasizes the need to produce descriptive scene graphs and highlights the need to change the way we deal with the problem. The use of fully supervised methods, unfortunately, does not scale well into an increasing number of object categories or relationships. But even in smaller vocabularies, due to the sparse, non-salient annotations, we end up with biased models that do not understand the image and are unable to locate important relationships.

Keywords Scene Graph Generation (SGG), Weak Supervision, Saliency, Scene Graph Generation from Image Captions, COCO, VG200, Open Images

Ευχαριστίες

Με το πέρας της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά τον κύριο Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική αυτή υπό την επίβλεψη και καθοδήγησή του. Τα μαθήματα του κυρίου Μαραγκού στη σχολή αποτέλεσαν τη πρώτη μου επαφή με τον κόσμο της όρασης υπολογιστών αλλά και γενικότερα της μηχανικής μάθησης, δίνοντάς μου έτσι την έμπνευση και ταυτόχρονα καλλιεργώντας την αγάπη μου για τον εντυπωσιακό αυτό τομέα.

Ένα ισάξιο ευχαριστώ πρέπει να πάει στον φίλο και συνεργάτη μου Μάρκο Διοματάρη καθώς και στον Βασίλη Πιτσικάλη που συνεπίβλεψαν τη διπλωματική μου στο πλαίσιο της συνεργασίας του εργαστηρίου CVSP και της Deeplab. Η αλληλεπίδραση μαζί τους, αλλά και γενικότερα με τους υπόλοιπους ανθρώπους στον χώρο της Deeplab ήταν και είναι μια συνεχής πηγή έμπνευσης και γνώσης για εμένα και έχει αποτελέσει το εφαλτήριο για μια βαθύτερη κατανόηση και την πιο κριτική ματιά στον τομέα. Ο ένας χρόνος που περάσαμε μαζί θα αποτελέσει αναπόσπαστο κομμάτι στην μετέπειτα μου πορεία. Μάρκο, σε ευχαριστώ για την ανοχή που έδειξες στην ξεροκεφαλιά, τη γκρίνια μου και την αργοπορία μου καθώς και στις δεκάδες ώρες που αφιέρωσες για τη διπλωματική μου.

Τέλος, ευχαριστώ την οικογένειά μου, τους φίλους μου και την κοπέλα μου που πίστεψαν σε μένα και με υποστήριξαν όλα αυτά τα χρόνια στη σχολή και ειδικά τα τελευταία χρόνια που ο συνδυασμός πανδημίας, καραντίνας και διπλωματικής θα ήταν ανυπόφορος χωρίς τη βοήθειά τους και τις αναμνήσεις που δημιουργήσαμε μαζί.

Contents

List of Figures

List of Tables

List of Algorithms

1 Εισαγωγή

Εδώ και αρκετά χρόνια παρατηρείται μια αλματώδης ανάπτυξη του τομέα της όρασης υπολογιστών με χρήση μηχανικής μάθησης και νευρωνικών δικτύων, με εκατοντάδες χιλιάδες δημοσιεύσεις σχετικές με αναγνώριση εικόνων, εντοπισμό αντικειμένων σε εικόνες καθώς και πολλών άλλων προβλημάτων που φάνταζαν εντελώς απίθανα πριν μερικά, μόλις, χρόνια. Πολλά, όμως, από αυτά τα προβλήματα τα οποία λύνουμε αφορούν περισσότερο την ανάκτηση πληροφορίας από εικόνες και όχι γνώσης. Ένας άνθρωπος, ένα αδιάβροχο, μια μηχανή και ένα φορτηγό που βρίσκονται σε κάποιες περιοχές της εικόνας αποτελούν σημαντικές πληροφορίες, ωστόσο δεν συγκρίνονται με τη γνώση ότι αυτός ο άνθρωπος φοράει το αδιάβροχο, οδηγάει τη μηχανή είναι δίπλα από το φορτηγό.

Και αυτό γιατί η ύπαρξη αυτών των αντικειμένων στην εικόνα δημιουργεί εκθετικά πολλά σενάρια για το τι μπορεί να συμβαίνει στην εικόνα, ο άνθρωπος μπορεί να οδηγάει το φορτηγό, να είναι μπροστά τους, να κρατάει τη μηχανή κλπ, ενώ η γνώση που περιγράψαμε είναι πολύ πιο σαφής, ξεκάθαρη και χρήσιμη. Μπορεί να αναπαραστήσει πολύ καλύτερα, σε κάποιον που δεν έχει δει την εικόνα, πως αυτή θα μπορούσε να είναι, μπορεί να απαντήσει ερωτήσεις όπως "τι κάνει ο άνθρωπος;" και "τι είναι δίπλα στο φορτηγό;" αλλά και να βοηθήσει στην περαιτέρω εξαγωγή της γνώσης οτι "όταν τραβήχτηκε η φωτογραφία, μάλλον έβρεχε"!

Βασιζόμενοι σε αυτή την ιδέα, οι ερευνητές της όρασης υπολογιστών εισήγαγαν το πρόβλημα της αναγνώρισης οπτικών σχέσεων, που αφορά τον εντοπισμό των σχέσεων μεταξύ των αντικειμένων που εμφανίζονται σε μια εικόνα. Μάλιστα, αν για μια εικόνα εντοπίσουμε όλα τα αντικείμενα και τα βάλουμε σε έναν γράφο σαν κόμβους, ενώ σημειώσουμε με ακμές στον γράφο τις σχέσεις μεταξύ των αντικειμένων αυτό, τότε λαμβάνουμε τον λεγόμενο γράφο σκηνής της εικόνα. Παρά, όμως, την μεγάλη εξέλιξη στον εντοπισμό και την αναγνώριση αντικειμένων σε εικόνες, χάρη στην ανάπτυξη βαθιών δικτύων μηχανικής μάθησης, η πρόοδος στο πρόβλημα της παραγωγής γράφου σκηνής είναι πιο μετριασμένη, και ακόμα και τα state-of-the-art (SOTA) μοντέλα παράγουν γράφους που συχνά δεν κωδικοποιούν κάποια σημαντική γνώση όπως θα δούμε παρακάτω.

1.1 Περιγραφή προβλήματος

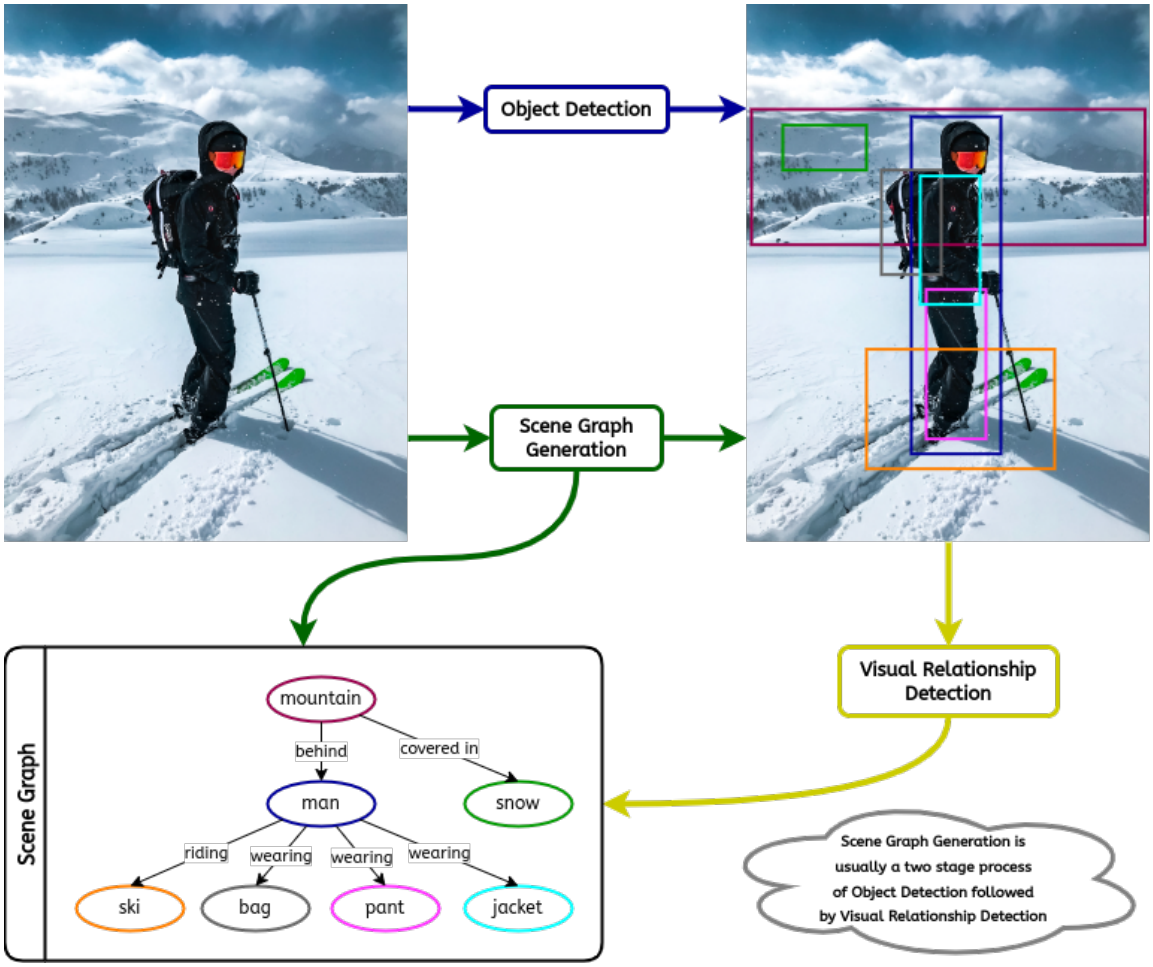
Για τον ορισμό του προβλήματος Παραγωγής Γράφου Σκηνής (Scene Graph Generation - SGG) πρέπει πρώτα να ορίσουμε την έννοια της σχέσης. Συγκεκριμένα, μια τριπλέτα της μορφής `<subject - predicate - object>` (υποκείμενο - κατηγόρημα - αντικείμενο) που περιγράφει πως το υποκείμενο σχετίζεται με τον αντικείμενο μέσω του κατηγορήματος, θα λέμε ότι αποτελεί μια σχέση. Έτσι, το SGG περιγράφει τη διαδικασία δημιουργίας ενός κατευθυνόμενου γράφου (βλ. εικόνα 1) με τους κόμβους να περιγράφουν τις οντότητες που εντοπίζονται στην εικόνα και τις ακμές να περιγράφουν τις σχέσεις που συνδέουν τους κόμβους (με κατεύθυνση από το υποκείμενο στο αντικείμενο).

Στα δεδομένα εκπαίδευσης, κάθε εικόνα αντιστοιχείται με

- ένα σύνολο από N επισημειωμένες οντότητες μαζί με τα κουτιά περιορισμών τους καθώς και τις κατηγορίες τους
- ένα σύνολο από T επισημειωμένες τριπλέτες της μορφής `(subject id, predicate, object id)` όπου το subject/object id αναφέρονται σε συγκεκριμένες οντότητες από αυτές που έχουν εντοπιστεί στην εικόνα (grounded οντότητες)

Μάλιστα, οι τριπλέτες είναι αραιώς επισημειωμένες με τις περισσότερες από τις N^2 πιθανές σχέσεις να μην είναι επισημειωμένες. Συγκεκριμένα για τα VRD [15] και VG200 [28], δύο από τα πιο συνηθισμένα σύνολα δεδομένων για το πρόβλημα, επισημειώνεται μόλις το 13% και 3% αντίστοιχα όλων των πιθανών σχέσεων. Για περισσότερες πληροφορίες για τα σύνολα δεδομένων παραπέμπουμε στο παράρτημα A.

Με βάση τα παραπάνω ορίζονται και διάφορες παραλλαγές του προβλήματος. Συγκεκριμένα, η γνώση των κουτιά περιορισμού (bounding boxes) των οντοτήτων, των κατηγοριών τους ή/και των επισημειωμένων ζευγαριών (χωρίς την επισημειωμένη σχέση μεταξύ τους) μπορούν να θεωρούνται



Σχήμα 1: Για το πρόβλημα της Παραγωγή Γράφου Σκηνής (Scene Graph Generation - SGG), δεδομένης μιας εικόνας, ανιχνεύουμε τις οντότητες στην εικόνα (object detection) και υπολογίζουμε έναν κατευθυνόμενο γράφο όπου οι κόμβοι αντιστοιχούν στις οντότητες και οι ακμές αποτελούν τη σχέση μια οντότητας υποκειμένου με μια άλλη οντότητα αντικειμένου, αν αυτή δεν θεωρείται background σχέση (visual relationship detection).

δεδομένα ή οχι ορίζοντας τις παραλλαγές Phrase Detection (PhrDet), Scene Graph Generation (SGGen), Scene Graph Classification (SGCls), Predicate Classification (PredCls) και Predicate Detection (PredDet). Ο λόγος χρήσης παραλλαγών όπως το PredCls είναι πως με αυτόν τον τρόπο μπορούμε να αποπλέξουμε το πρόβλημα της παραγωγής γράφου σκηνής από τις παραμέτρους του object detector. Για αναλυτικές πληροφορίες των παραλλαγών του προβλήματος παραπέμπουμε στο παράρτημα B.

Η πιο συνηθισμένη μετρική για την μέτρηση της επίδοσης των μοντέλων SGG αποτελεί το Recall@K (R@K). Έτσι, αν θεωρήσουμε ότι προβλέπουμε μία σχέση για κάθε πιθανό ζευγάρι αντικειμένων ($N \times (N-1)$ προβλέψεις) και τις 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

Το πρόβλημα του εντοπισμού σχέσεων σε εικόνες αποτελεί ένα ενδιάμεσου επιπέδου πρόβλημα για τον τομέα της όρασης υπολογιστών. Αποτελεί ανωτέρου επιπέδου πρόβλημα συγκριτικά με τα προβλήματα εντοπισμού ή κατηγοριοποίησης αντικειμένων (object detection ή classification) αλλά

χαμηλότερου επιπέδου πρόβλημα σε σύγκριση άλλα όπως η παραγωγή περιγραφών εικόνων (image captioning), απάντηση ερωτήσεων σχετικά με μια εικόνα (visual question answering), απάντηση και λογική αιτιολόγηση σχετικά με οπτικές ερωτήσεις (visual commonsense reasoning). Αυτό, τοποθετή το πρόβλημά μας σε μια θέση όπου μπορεί να εξάγει υψηλότερου επιπέδου γνώση από μια εικόνα, έναντι απλής πληροφορίας όπως προβλήματα χαμηλότερου επιπέδου, αλλά ταυτόχρονα οι γράφοι που παράγονται μπορούν να χρησιμοποιηθούν σαν τη βασική αναπαράσταση για τη λύση προβλημάτων υψηλότερου επιπέδου. Για παράδειγμα, η κατανόηση μια σκηνής σε μορφή γράφου μπορεί εύκολα να απαντήσει σε ερωτήσεις όπως "τι κάνει ο άνθρωπος A στην εικόνα;" αν για τον "άνθρωπο Α" ξέρουμε πως <άνθρωπος A - οδηγάει - μηχανή>.

1.3 Προκλήσεις

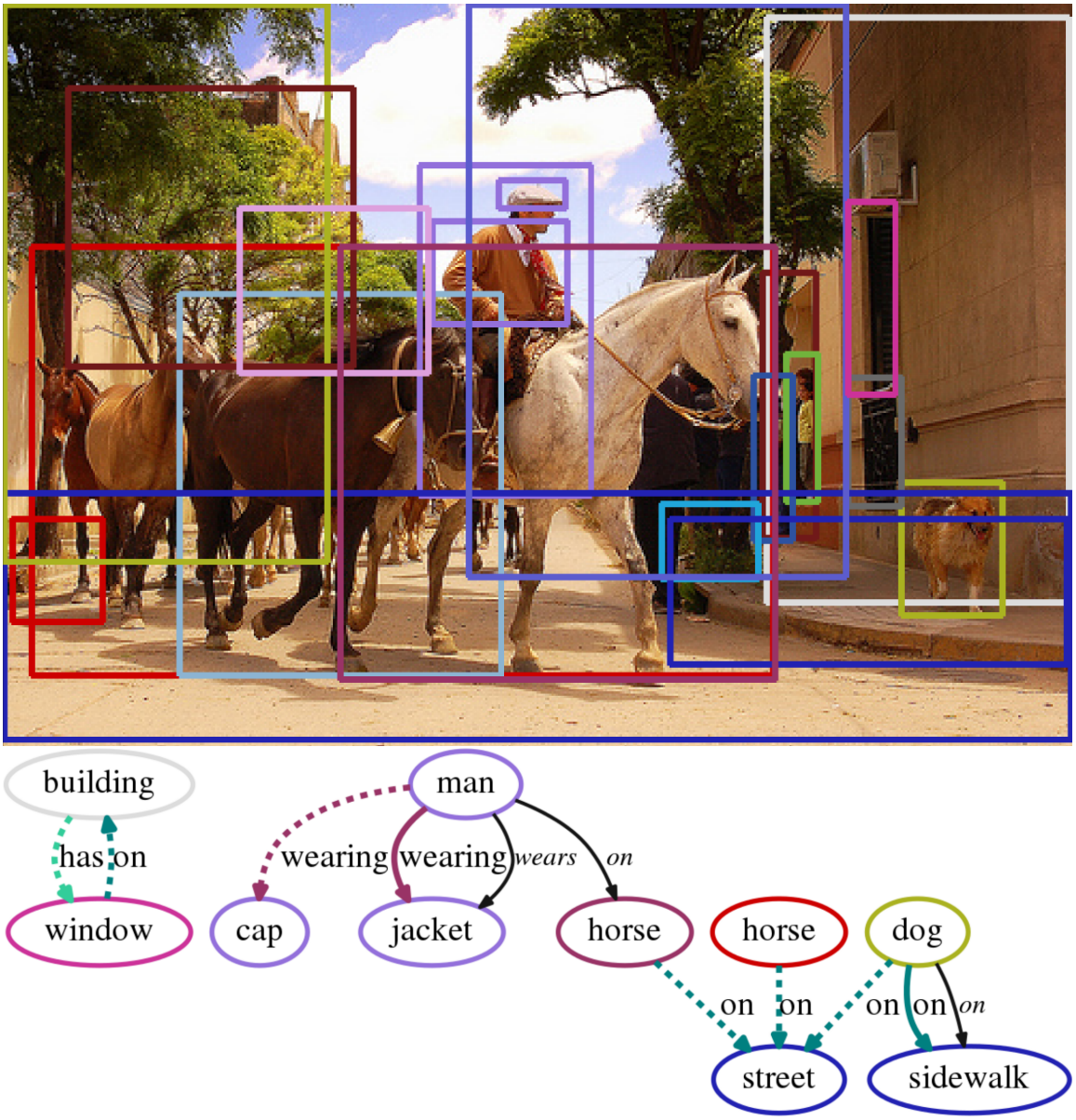
Η παραγωγή γράφου σκηνής για μια εικόνα αποτελεί ένα ιδιαίτερα δύσκολο πρόβλημα το οποίο συνδυάζει πολλαπλές προκλήσεις που πρέπει να αντιμετωπιστούν για τη λύση του. Ξεκινώντας, βασική δυσκολία αποτελεί η **συνδυαστική φύση** του προβλήματος όπου για ένα δοσμένο λεξιλόγιο O αντικειμένων και P κατηγορημάτων, πρέπει να μπορούμε να εντοπίζουμε PO^2 διαφορετικές τριπλέτες, καταλήγοντας στο VG200 να έχουμε 1.125.000 πιθανές τριπλέτες (με τις περισσότερες μη επισημειωμένες) και στο σύνολο δοκιμής του, το 4% των τριπλετών να μην υπάρχουν στο σύνολο εκπαίδευσης. Ακόμα, η χρήση γλώσσα εισάγει προβλήματα **συνωνυμίας** και **πολυσημίας** όπου, τα κατηγορήματα "next to" και "near" μπορούν σε κάποιο περιβάλλον να μπορού να χρησιμοποιηθούν ως συνώνημα, αλλα θα πρέπει να επιλέξουμε μονο ενα απο τα δύο σαν σωστά. Αντίστοιχα, κατηγορήματα όπως το "on" μπορεί να έχουν πολλές διαφορετικές σημασίες, από το χωρικό <person - on - car> μέχρι έννοιες "ride" στο <person - on - bike> ή "wearing" στο <person - on - skis>. Επιπλέον, δεν πρέπει να αγνοήσουμε τη δυσκολία εντοπισμού σχέσεων που καθορίζονται από **μικρές οπτικές λεπτομέρειες** σε εικόνα ή αφορούν αρκετά μικρές οντότητες στην εικόνα, για παράδειγμα, θα ήταν ιδιάιτερα δύσκολο να καταλάβουμε εάν κάποιος άνθρωπος κρατάει ένα κινητό ή αν πληκτρολογεί σε αυτό. Τέλος, σαν να μην έφταναν τα παραπάνω, η επισημείωση των δεδομένων γίνεται με **ασυνεπή** τρόπο όπου η ίδια πληροφορία μπορεί να θεωρηθεί σημαντική και να επισημειωθεί σε μια εικόνα, ενώ σε μια άλλη να αγνοηθεί, ενώ δεν είναι και λίγες οι φορές όπου σημαντικές οντότητες ή σχέσεις στην εικόνα δεν είναι επισημειωμένες.

1.4 Κίνητρο & Συνεισφορές

1.4.1 Κίνητρο

Βασικό κίνητρο αυτής της ερευνητικής προσπάθειας αποτελεί η παρατήρηση ότι όλα τα μοντέλα, ανεξάρτητα από την αρχιτεκτονική τους, δεν έχουν την ικανότητα να εντοπίσουν με ακρίβεια ποιες από τις σχέσεις είναι σημαντικές και πρέπει να προβλεφθούν και ποιες όχι. Σύμφωνα με τα μοντέλα που έχουμε, συχνά σε μια εικόνα θεωρούνται πιο σημαντικές σχέσεις όπως <ουρανός - πάνω από - βουνο> ή <παράθυρο - σε - κτήριο> έναντι των <άνθρωπος - φοράει - σκι> ή <γυναίκα - κρατάει - ρακέτα>. Για παράδειγμα, στην εικόνα 2 φαίνεται ο γράφος που προβλέπεται από το [16] που συνεχίζει να πετυχαίνει μια από τις καλύτερες επιδόσεις σε Recall στη βιβλιογραφία. Βλέπουμε πως κατατάσεται σαν πιο σημαντικό πως <window - on - building> έναντι του <man - riding - horse> που κωδικοποιούν την ουσία αυτού που απεικονίζεται. Μάλιστα, αξίζει να παρατηρήσουμε πως το μοντέλο όντως προβλέπει επισημειωμένα δεδομένα (συνεχόμενη γραμμή) και πως οι σημαντικές σχέσεις στην εικόνα δεν είναι επισημειωμένες (μαύρη γραμμή). Και αυτός είναι και ο λόγος που δεν υπάρχει αναφορά σε αυτό το πρόβλημα ως τώρα. Αφού τα μοντέλα προβλέπουν τις σχέσεις που υπάρχουν στα επισημειωμένα δεδομένα δεν είναι εμφανής η ύπαρξη του προβλήματος.

Αιτία του προβλήματος αυτού, σε μεγάλο βαθμό, είναι τα ίδια τα επισημειωμένα δεδομένα καθώς και ο τρόπος επισημείωσης και φιλτραρίσματός τους. Για παράδειγμα, εάν δοθούν σε έναν άνθρωπο δύο περιοχές στην εικόνα με τα αντικείμενα ουρανός και βουνό, συχνά θα τα επισημειώσει με τη σχέση <ουρανός - πάνω από - βουνό> παρόλο που αυτή δεν δίνει καμία πληροφορία για την εικόνα.



Σχήμα 2: Ακόμα και με τις πιο σύγχρονες μεθόδους ([16]) παρατηρούμε πως επιλέγονται ως πιο σημαντικές σχέσεις όπως `<window - on - building` έναντι του `<man - riding - horse>`. Αυτό υποδηλώνει πως υπάρχει κάποιο σημαντικό πρόβλημα στον τρόπο εκπαίδευσης των υπαρχόντων μοντέλων που τα εμποδίζει από το να μάθουν ποιες σχέσεις είναι πιο σημαντικές για μια εικόνα. Τα διακεκομμένα βέλη αντιπροσωπεύουν ζευγάρια οντοτήτων για τα οποία δεν υπάρχει επισημειωμένη σχέση, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα.

Ιδιαίτερο ενδιαφέρον, μάλιστα, έχουν εικόνες όπου όλες οι επισημειωμένες σχέσεις είναι μηδαμινού ενδιαφέροντος ή/και εικόνες όπου δεν υπάρχουν καν επισημειωμένα τα αντικείμενα που βρίσκονται στο επίκεντρό τους (βλ. εικόνα 3).

Η δεύτερη πολύ σημαντική παρατήρηση που συνοδεύει το κίνητρό μας είναι πως οι περιγραφές εικόνων περιλαμβάνουν, εξ' ορισμού, πληροφορία σχετικά με το ποιες σχέσεις στην εικόνα είναι σημαντικές. Συγκεκριμένα, για να συνδεθούν δύο οντότητες νοηματικά σε μια περιγραφή εικόνας από κάποιον επισημειωτή, σημαίνει πως η σχέση τους είναι σημαντική μιας και η περιγραφή της εικόνα αποτελεί μια σύνοψη του τι συμβαίνει σε αυτή. Συνεπώς, το bottleneck που παρουσιάζεται στην επισημείωση εικόνων για image captioning, που αποτελεί ο περιορισμός του μεγέθους της

Σχήμα 3: Στην παρούσα εικόνα από το VG200 δεν είναι επισημειωμένη η ρακέτα, που αποτελεί ένα κεντρικό αντικείμενο για την εικόνα. Αλλά ακόμα και από τα επισημειωμένα αντικείμενα, η μόνη επισημειωμένη σχέση για τη παρούσα εικόνα είναι το `<sign - on - board>`.

περιγραφής (δεν μπορούμε να γράφουμε δύο παραγράφους για να περιγράψουμε την εικόνα) οδηγεί στην κωδικοποίηση της πολύ σημαντικής πληροφορίας του saliency στην εικόνα που μας ενδιαφέρει.

1.4.2 Συνεισφορές

Βασισμένοι στις παρατηρήσεις από το 1.4.1, με τη παρούσα έρευνα συμβάλλουμε τόσο στη μέτρηση του saliency ενός Scene Graph Generation (SGG) μοντέλου όσο και στη παραγωγή πιο salient γράφων σκηνής σύμφωνα με ποιοτικά και ποσοτικά αποτελέσματα που εξάγουμε. Συγκεκριμένα:

- Εισάγουμε μια γενικευμένη μέθοδο εκπαίδευσης SGG μοντέλων με ασθενή επίβλεψη χρησιμοποιώντας περιγραφές εικόνων.
- Εισάγουμε δύο παραλλαγές της μέτρησης του Recall@N όπου, με χρήση των περιγραφών εικόνων, μπορούμε να εξάγουμε μετρήσεις για το saliency SGG μοντέλων.
- Πραγματοποιούμε τόσο ποσοτική όσο και ποιοτική σύγκριση μεταξύ των μεθόδων που προτείνουμε και με τη σχετική βιβλιογραφία στο VG200, το δημοφιλέστερο σύνολο δεδομένων του προβλήματος όπου πετυχαίνουμε 35% μέγιστη σχετική βελτίωση συγκριτικά με επανυλοποίηση της SOTA μεθόδου.

Θεμελιώνουμε, λοιπόν, την αιτία έλλειψης saliency στους γράφους σκηνής, προτείνουμε μετρικές για την αξιολόγηση του saliency ενός μοντέλου και τέλος σχεδιάζουμε μια μέθοδο εκπαίδευσης μοντέλων ώστε αυτά να αντιλαμβάνονται καλύτερα την έννοια του saliency και να παράγουν πιο ουσιώδεις γράφους σκηνής.

2 Ανασκόπηση Βιβλιογραφίας

Λόγω των πολλαπλών προκλήσεων που παρουσιάζει το πρόβλημα της παραγωγής γράφου σκηνής, κάποιες από τις οποίες παρουσιάσαμε στην υποενότητα 1.3, υπάρχει αρκετή βιβλιογραφία που επικεντρώνεται σε διαφορετικές πτυχές του. Για παράδειγμα στα [25, 33] επικεντρώνονται στο πρόβλημα της long-tail κατανομής, οι [4, 7, 26] ασχολούνται με το πρόβλημα των zero-shot ή few-shot προβλέψεων, στα [25, 3] ορίζεται και αντιμετωπίζεται το πρόβλημα του context bias, ενώ δημοσιεύσεις όπως [12] ασχολούνται με αρχιτεκτονικές εκπαίδευσης απο άκρη σε άκρη (end-to-end). Στη παρούσα εργασία θα παρουσιάσουμε μία γενική ανασκόπηση μέρους της βιβλιογραφίας που αφορά ορισμένες σημαντικές κατευθύνσεις στον χώρο της μοντελοποίησης των δικτύων πρόβλεψης ενώ στη συνέχεια θα γίνει μια πιο αναλυτική παρουσίαση βιβλιογραφίας σχετικής με εκπαίδευση μοντέλων παραγωγής γράφων σκηνής με χρήση ασθενούς επίβλεψης.

2.1 Είδη πληροφορίας

Ξεκινώντας, είναι ιδιαίτερα σημαντικό να αναφερθούμε στις ροές πληροφορίας και χαρακτηριστικών που χρησιμοποιούνται από τα μοντέλα για την κατηγοριοποίηση των σχέσεων. Οι τρεις βασικές αυτές ροές είναι:

- Οπτική (**V**isual): Χρησιμοποιείται κάποιο προεκπαιδευμένο δίκτυο ως backbone (συνήθως VGG-16 [23] ή ResNet [24]) για την εξαγωγή feature maps. Μέσω μιας προεκπαιδευμένης κεφαλής εντοπισμού αντικειμένων λαμβάνουμε feature vectors για τα προβλεπόμενα αντικείμενα, τα οποία χρησιμοποιούνται από το δίκτυο.

- Χωρική (**S**patial): μέσω των συντεταγμένων των 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 [5].

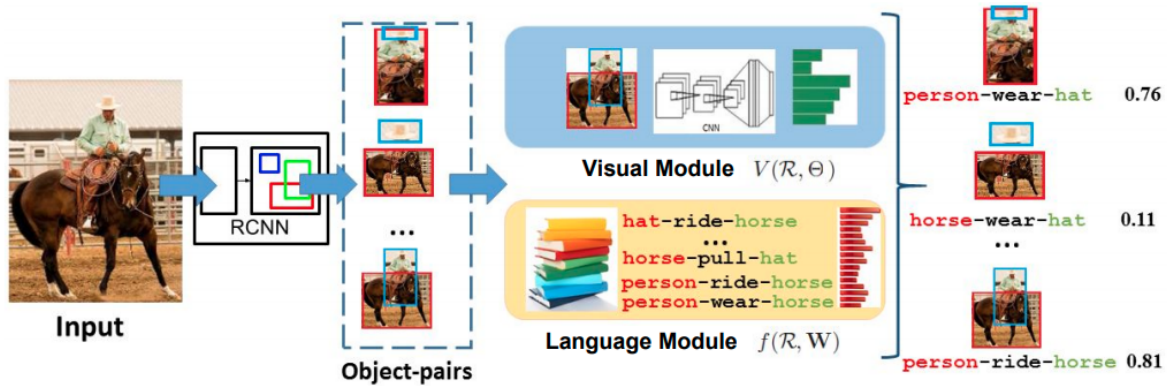
- Γλωσσική (**L**inguistic): Οι κατηγορίες του υποκειμένου και του αντικειμένου κωδικοποιούνται μέσω *word2vec* σε διανύσματα ενός σημασιολογικού χώρου όπου γεωμετρικές σχέσεις αντιστοιχούν σε νοηματικές.

Από τα παραπάνω, λοιπόν, λαμβάνουμε πληροφορίες για τα οπτικά και σημασιολογικά χαρακτηριστικά αντικειμένου και υποκειμένου καθώς και για τη χωρική τους σχέση στην εικόνα.

2.2 Γενική Βιβλιογραφία

Η ουσιαστική εισαγωγή του προβλήματος γίνεται με τους [15] και τη ταυτόχρονη θεσμοθέτηση του συνόλου δεδομένων VRD. Σε αντίθεση με προηγούμενες προσεγγίσεις, με την εισαγωγή του VRD μέθοδοι που χρησιμοποιούσαν έναν ταξινομητή για κάθε τριπλέτα σχέσεων (δηλαδή, διαφορετικός ταξινομητής για το `<άνθρωπος - καβαλάει - άλογο>` και για το `<άνθρωπος - καβαλάει - ποδήλατο>`) δεν είναι εφικτές λόγο του μεγάλου αριθμού σχέσεων και αντικειμένων, και άρα πιθανών τριπλετών. Έτσι, αντιμετωπίζουν το πρόβλημα με τη χρήση ενός ταξινομητή για κάθε είδος σχέσης καθώς και ταξινομητών για οντότητες (αντικείμενο, υποκείμενο) της εικόνας μειώνοντας τη πολυπλοκότητα από $\mathcal{O}(N^2K)$ σε $\mathcal{O}(N+K)$, όπου K ο αριθμός των σχέσεων και N ο αριθμός των οντοτήτων (βλ. Σχήμα 4).

Έπειτα, οι [8] με το UVTransE θέτουν ένα ισχυρό baseline για το πρόβλημα με μια αρχιτεκτονική δικτύου που, σε μεγάλο βαθμό, ακολουθούν πολλά σύγχρονα μοντέλα. Συγκεκριμένα, στηριζόμενοι και στους [31] προβάλλουν τα διανύσματα χαρακτηριστικών του αντικειμένου, υποκειμένου, της ένωσής τους (union) και του κατηγορουμένου σε έναν κοινό χώρο και απαιτούν στην εκπαίδευση $U - S - O \approx p$ όπου U, S, O, P το διανύσματα χαρακτηριστικών της ένωσης, του υποκειμένου, του αντικειμένου και του κατηγορήματος αντίστοιχα. Η ιδέα στηρίζεται στους [31] όπου παρόμοια "απαιτούν" $S + P \approx O$, το οποίο με τη σειρά του έχει ρίζες στα word embeddings όπου, αν



Σχήμα 4: Μια σύνοψη του μοντέλου για εντοπισμό οπτικών σχέσεων που παρουσιάζεται στο [15]. Δεδομένης μιας εικόνας, ένας object detection δίκτυο παράγει ένα σύνολο από προτάσεις αντικειμένων. Κάθε ζευγάρι αντικειμένων λαμβάνει μια βαθμολογία με τη χρήση (i) της οπτικής μονάδας και της (ii) γλωσσικής μονάδας. Αυτές οι βαθμολογίες περνάνε από μια συνάρτηση κατωφλίου και παράγουν τελικά σχέσεις αντικειμένων (π.χ. `<person - ride - horse>`. Εικόνα από [15]

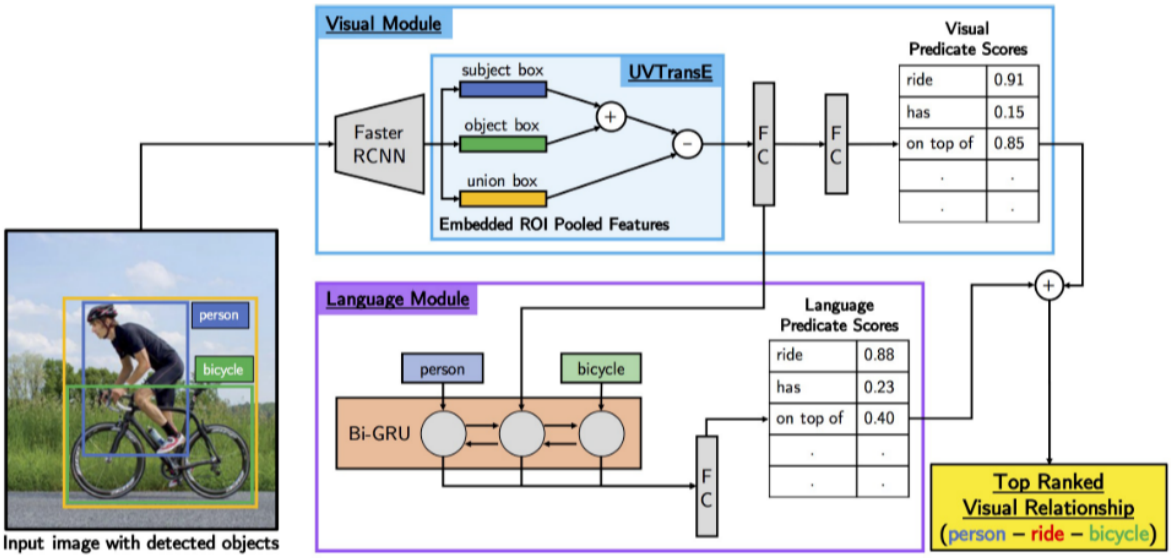
σκεφτούμε το P σαν ένα διάνυσμα στον χώρο, θα περιμέναμε από το διάνυσμα του S να πηγαίνουμε στο O μέσω του P, πχ. άνθρωπος + φοράει ≈ μπλούζα. Μάλιστα, επιπλέον αυτού, στο UVTransE προσθέτουν και ένα γλωσσικό μοντέλο (ένα απλό Bidirectional Gated Recurrent Units (Bi-GRU) δίκτυο) το οποίο, βασισμένο στα word embeddings υποκειμένου και αντικειμένου, καθώς και στην αναπαράσταση του κατηγορήματος που έχει προκύψει από το προηγούμενο κομμάτι του δικτύου, παράγει ορισμένες πιθανότητες εξόδου για κάθε κατηγόρημα. Τελικά, από το άθροισμα των πιθανοτήτων της καθαρά "οπτική μονάδας" και της "γλωσσικής μονάδας" λαμβάνουμε της τελικές πιθανότητες για τα κατηγορήματα (βλ. Σχήμα 5). Στην ίδια κατηγορία, οι [6] με το MATransE (Multimodal Attentional Translation Embeddings) χρησιμοποιούν μια αρχιτεκτονική δύο κλάδων (βλ. εικόνα 6) όπου σε έναν latent χώρο ο ένας κλαδος υπολογίζει ενα διανύσματα χαρακτηριστικών για το $S-O$ ενώ ο άλλος ένα διανύσματα χαρακτηριστικών, από τα όλη τη περιοχή της σχέσης, για το P. Συνδιάζοντας τα διανύσματα αυτά εφαρμόζεται μια συνολική επίβλεψη στο μοντέλο, ωστόσο εφαρμόζεται και deep supervision στον κάθε κλάδο.

Ιδιαίτερου ενδιαφέροντος είναι και ο σχεδιασμός μοντέλων με μηχανισμούς προσοχής (attention mechanisms) που χρησιμοποιούνται ευρέως σε πλήθος πεδίων με πιο χαρακτηριστικά παραδείγματα το image captioning και τα προβλήματα επεξεργασίας φυσικής γλώσσας. Η βασική ιδέα είναι η αλλαγή της συμβολής που έχει ένας δίαυλος πληροφορίας, με βάση την περίσταση και το γενικότερο πλαίσιο, μέσα από κάποια εκπαιδεύσιμα βάρη, για τον εντοπισμό της χρήσιμης πληροφορίας κάθε φορά. Στο δικό μας πεδίο, στο [6] μπορούμε να δούμε μηχανισμούς προσοχής στα οπτικά χαρακτηριστικά οδηγούμενους από συνδυασμό σημασιολογικής πληροφορίας, μέσα από τα γλωσσικά χαρακτηριστικά που προκύπτουν από κάθε κλάση αντικειμένου, και χωρικών πληροφοριών, μέσα από τις δυαδικές μάσκες αντικειμένων (βλ. εικόνα 6). Παρόμοιος μηχανισμός χρησιμοποιείται και στο [5], με τη διαφορά πως κάθε κλάση έχει διαφορετικές παραμέτρους για τον μηχανισμό προσοχής οι οποίες εκπαιδεύονται ανεξάρτητα (multi-head attention) (βλ. εικόνα 7).

2.3 Κοντά σε εμάς - Προηγούμενες Δουλειές

2.3.1 Weakly-supervised learning of visual relations [19]

Στη παρούσα δουλειά οι συγγραφείς ασχολούνται με το πρόβλημα της ασθενούς εκπαίδευσης SGG μοντέλου από επισημειώσεις τριπλετών σε επίπεδο εικόνας. Συγκεκριμένα, χρησιμοποιούν μόνο τις τριπλέτες από το πλήρες επισημειωμένο VRD και τα αντικείμενα που προβλέπει ένας προεκπαιδευμένος Object Detector, finetuned όμως στο VRD. Συνεπώς, για την εκπαίδευση του μοντέλου, χρησιμοποιούνται τόσο τα επισημειωμένα objects του VRD όσο και οι unlocalized τριπλέτες



Σχήμα 5: Σύνοψη του μοντέλου για εντοπισμό οπτικών σχέσεων που παρουσιάζεται στο [8]. Για κάθε ζευγάρι αντικειμένων, οπτικά και χωρικά χαρακτηριστικά συνδυάζονται και προωθούνται στην "οπτική μονάδα" που υπολογίζει το UVTransE διάνυσμα αλληλεπίδρασης: ένωση - (υποκείμενο + αντικείμενο). Το παραγόμενο διάνυσμα του κατηγορήματος μπορεί να σταλεί στη "γλωσσική μονάδα" (Bi-GRU). Τελικά, αθροίζονται οι πιθανότητες των δύο μονάδων και οι τριπλέτες ταξινομούνται από αυτή με την υψηλότερη βαθμολογία προς αυτή με τη μικρότερη. Εικόνα από [8]

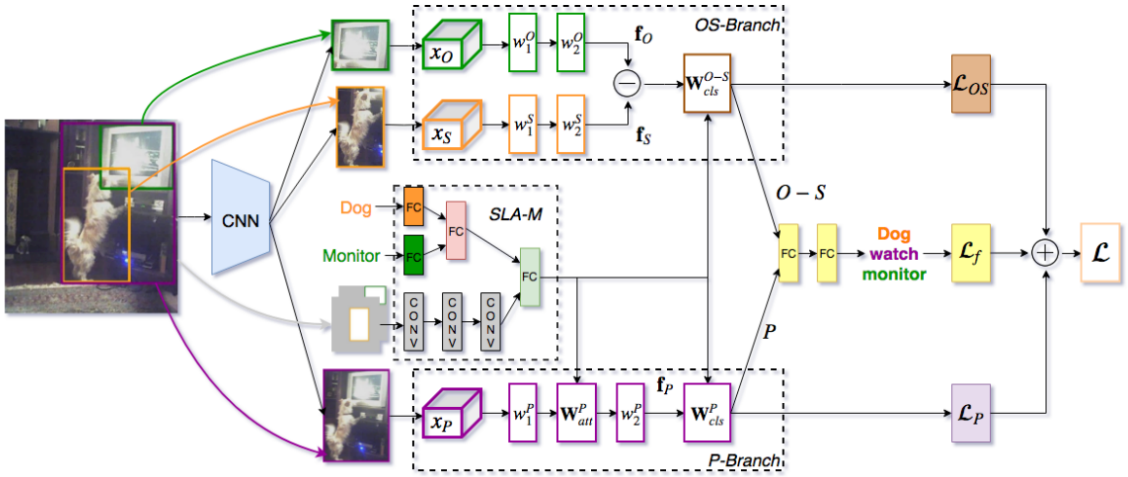
του (το μόνο που λείπει είναι η αντιστοίχιση των οντοτήτων στις τριπλέτες με τα αντικείμενα στην εικόνα). Με αυτά τα δεδομένα, διατυπώνουν το πρόβλημα του ταιριάσματος σημασιολογικών οντοτήτων και οπτικών αντικειμένων ως ένα constrained convex optimization πρόβλημα το οποίο λύνουν με τη χρήση του Frank–Wolfe αλγορίθμου κατά τη διάρκεια της εκπαίδευσης.

Επιπλέον των παραπάνω, και λιγότερο σχετικά με τη δική μας δουλειά, κωδικοποιούν και χωρικά features μαζί με τα οπτικά για χρήση στην πρόβλεψη του predicate, ενώ εισάγουν και το UnRel (Unusual Relations), ένα σύνολο δεδομένων από ασυνήθιστες σχέσεις, το οποίο είναι πλήρως και εξαντλητικά επισημειωμένο, και μπορεί να χρησιμοποιηθεί για το αξιολόγηση μοντέλων SGG που έχουν εκπαιδευτεί στο VRD.

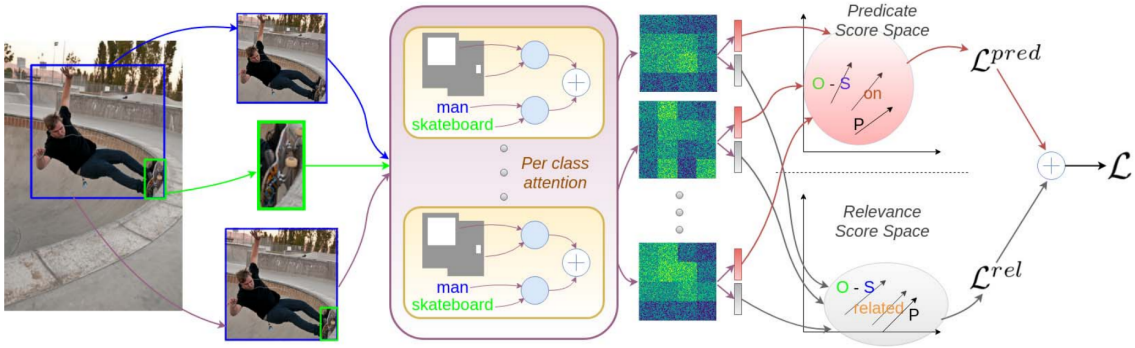
2.3.2 PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN [32]

Στη παρούσα δουλειά οι συγγραφείς ασχολούνται, όπως και πάνω, με το πρόβλημα της ασθενούς εκπαίδευσης SGG μοντέλου από επισημειώσεις τριπλετών σε επίπεδο εικόνας. Μάλιστα, το μοντέλο αξιολογείται τόσο στο VRD όσο και στο Visual Genome (VG). Ωστόσο, εδώ πέρα, αντί για τη χρήση των επισημειωμένων κουτιών περιορισμού αντικειμένων του VRD ή του VG για το finetune ενός προεκπαιδευμένου Object Detector όπως παραπάνω, δεν χρησιμοποιείται καθόλου αυτή η πληροφορία. Δηλαδή, χρησιμοποιούνται μόνο οι σημασιολογικές τριπλέτες για την εκπαίδευση και όχι και η χωρική τοποθεσία αντικειμένων στην εικόνα. Συνεπώς, το μοντέλο πρέπει να μάθει συνδυαστικά και να κάνει Object Detection αλλά και Predicate Classification με ασθενή τρόπο, το οποίο είναι αρκετά πιο δύσκολο task από αυτό της προηγούμενης δημοσίευσης όσο και της επόμενης. Δυστυχώς, όμως, η μέθοδος δεν φαίνεται να είναι model agnostic.

Αυτή η προσέγγιση είναι αρκετά πιο σημαντική από αυτή στη προηγούμενη δουλειά καθώς, δεδομένου ότι έχουν επισημειωθεί τόσο τα αντικείμενα στην εικόνα όσο και οι σημασιολογικές τριπλέτες, η επισημείωση του grounding είναι σχετικά αμελητέα οπότε δεν υπάρχει πολύ σοβαρός λόγος να ενδιαφερόμαστε για την εκπαίδευση μοντέλων αγνοώντας μόνο αυτό το δεδομένο. Αντί-



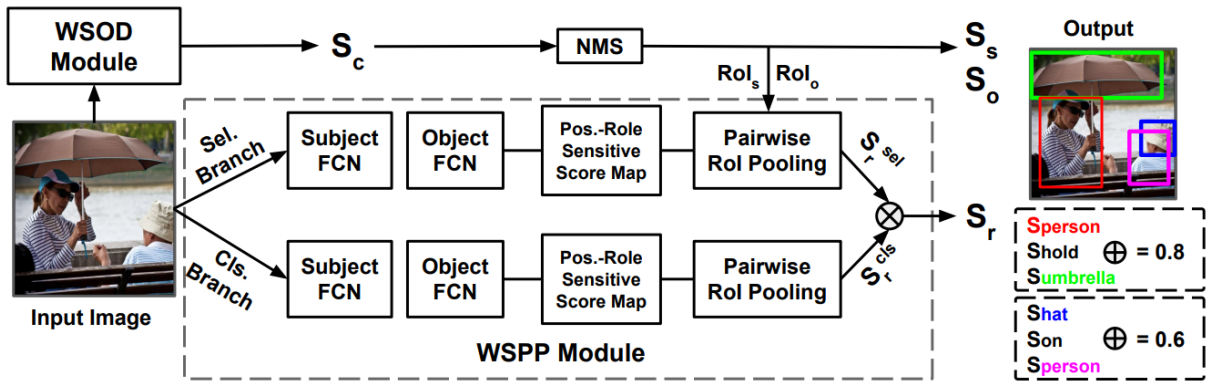
Σχήμα 6: Αρχιτεκτονική δύο κλάδων του MATransE. Ο ένας κλάδος μαθαίνει τη διανυσματική μετατόπιση των χαρακτηριστικών ανάμεσα στο υποκείμενο και το αντικείμενο ενώ ο άλλος προσπαθεί να προβλέψει τη σχέση από τα χαρακτηριστικά του κουτιού ένωσης (union). Ακόμα, βλέπουμε τον μηχανισμό προσοχής στην οπτική πληροφορία καθοδηγούμενο από τον σημασιολογικό και χωρικό δίαυλο πληροφορίας. Εικόνα από [6]



Σχήμα 7: Βλέπουμε πως κάθε κλάση έχει τις δικές της παραμέτρους που μαθαίνουν ανεξάρτητα την προσοχή που χρειάζεται (multi-head attention). Ακόμα, λύνονται χωριστά τα προβλήματα της κατηγοριοποίησης σχέσης και συσχέτισης δύο αντικειμένων. Το γινόμενο των πιθανοτήτων αυτών υπολογίζει τη τελική πιθανότητα μιας κλάσης. Εικόνα από [5]

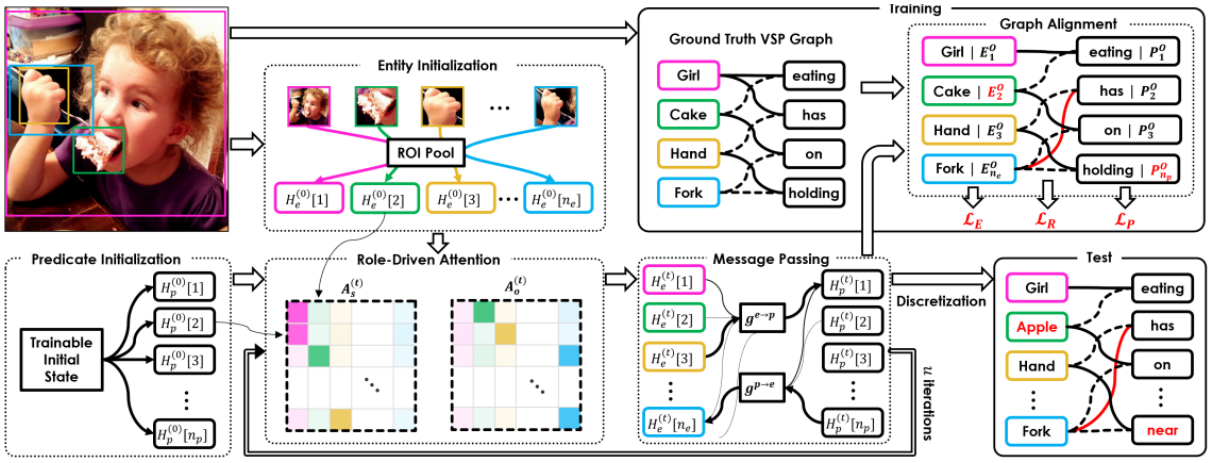
θετα, οι σημασιολογικές τριπλέτες είναι αρκετά πιο εύκολο και γρήγορο να επισημειωθούν (έναντι της επισημείωσης και των αντικειμένων και του grounding) και μάλιστα μπορούν συχνά να παραχθούν με αυτοματοποιημένο τρόπο από image captions, όπως θα δούμε αργότερα, που είναι πολύ εύκολο να επισημειώσουμε και για τα οποία έχουμε τεράστια σύνολα δεδομένων.

Ιδιαίτερο ενδιαφέρον, μάλιστα, αποτελεί και το γεγονός πως δεν εφαρμόζεται κατά την εκπαίδευση κάποια ξεκάθαρη αντιστοίχιση επισημειωμένης τριπλέτας με μια από τις τριπλέτες πρόβλεψης για την εφαρμογή Cross Entropy σε αυτή τη πρόβλεψη. Αντίθετα, εφαρμόζεται ένα Loss μόνο στο πλαίσιο της εικόνας όπου, αν s, o είναι δύο κατηγορίες αντικειμένων και C_s, C_o είναι όλες οι περιοχές στην εικόνα που αντιστοιχούν σε αυτά τα αντικείμενα, τότε συμβολίζουμε με S_r το συνολικό score για το predicate r που είναι το άθροισμα $\sum_{i \in C_s, j \in C_o} S_r(P_i, P_j)$ και εφαρμόζει Binary Cross Entropy για κάθε επισημειωμένη τριπλέτα. Ο λόγος είναι πως ξέρουμε την συνδιασμένη πιθανότητα να υπάρχει μια τριπλέτα στην εικόνα $P(s, r, o) = \sum_{i \in C_s, j \in C_o} P(i, r, j)$, αλλα όχι χωριστά την πιθανότητα δύο περιοχές στην εικόνα $i \in C_s$, $j \in C_o$ με αντικείμενα τύπου s, o, αντίστοιχα, να συνδέονται με το κατηγόρημα r, $P(C_s, r, C_o)$.



Σχήμα 8: Μέθοδος από το [32]. Βλέπουμε πως το WSPP Module έχει δύο χωριστές ροές για τη πρόβλεψη του predicate χωριστά από τη πρόβλεψη του αν το ζευγάρι είναι σχέση προσκηνίου. Εικόνα από [32]

2.3.3 Weakly Supervised Visual Semantic Parsing [30]



Σχήμα 9: Μια σύνοψη του μοντέλου για εντοπισμό οπτικών σχέσεων που παρουσιάζεται στο [30]. Δεδομένων προτάσεων αντικειμένων, ένας γράφος σκηνής παράγεται από μια επαναληπτική διαδικασία που περιλαμβάνει μια μονάδα προσοχής πολλαπλών κεφαλών που συμπεραίνει ακμές μεταξύ οντοτήτων και κατηγορημάτων, και μια μονάδα αποστολής μηνυμάτων για την διάδοση πληροφορίας μεταξύ κόμβων και την ανανέωση της κατάστασής τους. Για τον ορισμό μιας συνάρτησης κόστους για κάθε κόμβο και ακμή, ο επισημειωμένος γράφος σκηνής ευθυγραμμίζεται με τον γράφο εξόδου μέσω ενος αλγορίθμου ασθενής επίβλεψης. Εικόνα από [30]

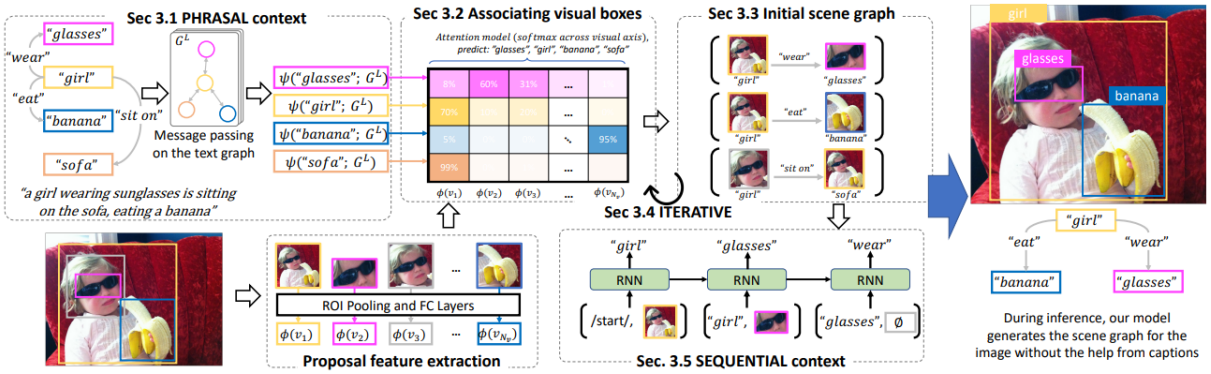
Όπως και με τις προηγούμενες δημοσιεύσεις, έτσι και εδώ οι συγγραφείς ασχολούνται, με το πρόβλημα της ασθενούς εκπαίδευσης SGG μοντέλου από επισημειώσεις τριπλετών σε επίπεδο εικόνας. Στη διαδικασία ασθενούς επίβλεψης ενός μοντέλου SGG, οι [30], ασχολούνται με το ελαφρώς γενικότερο πρόβλημα του "Visual Semantic Parsing" όπου τόσο τα κατηγορήματα όσο και οι οντότητες παρουσιάζονται σαν κόμβοι ενός γράφου και ενώνονται μεταξύ τους με ακμές που καθορίζουν το είδος συσχέτισης (υποκείμενο, αντικείμενο, εργαλείο). Η ασθενής επίβλεψη προκύπτει από το γεγονός πως δεν χρησιμοποιούν τα bounding boxes από τα επισημειωμένα σύνολα δεδομένων για την εκπαίδευση. Στη συγκεκριμένη περίπτωση χρησιμοποιείται ένας off-the-shelf (OTS) Object Detector εκπαιδευμένος στο Open Images [18] σύνολο δεδομένων για την εξαγωγή των bounding boxes και των features των αντικειμένων στην εικόνα.

Η αρχιτεκτονική του μοντέλου στηρίζεται σε πέρασμα μηνυμάτων από κόμβο σε κόμβο που επηρεάζει το διάνυσμα χαρακτηριστικών του καθενός, καθώς και από πίνακες προσοχής που ανανεώνονται και αυτοί επαναληπτικά μαζί με τους κόμβους για ένα σταθερό αριθμό επαναλήψεων. Τελικά, για την πρόβλεψη των κόμβων (οντότητες και κατηγορήματα) εφαρμόζεται μια διακριτοποίηση (επιλέγεται το κοντινότερο διάνυσμα οντότητας ή κατηγορήματος αντίστοιχα από το λεξικό με όλα τα διαθέσιμα) και από τους πίνακες προσοχής λαμβάνουμε το είδος της συσχέτισης. Το ενδιαφέρον κομμάτι, που ταιριάζει και με το δικό μας πρόβλημα, αποτελεί η επαναληπτική διαδικασία με την οποία ταιριάζουν τους γράφους. Συγκεκριμένα, (i) θεωρούν κάποιο δοσμένο ταίριασμα μεταξύ των οντοτήτων των δύο γράφων και βρίσκουν το βέλτιστο ταίριασμα κατηγορημάτων με τη βοήθεια του Kuhn–Munkres αλγορίθμου (ή Hungarian Algorithm) [10] (ii) με δοσμένο το προηγούμενο ταίριασμα κατηγορημάτων βρίσκουν αντίστοιχα ένα βέλτιστο ταίριασμα οντοτήτων (iii) επαναλαμβάνεται το βήμα (i) (βλ. Σχήμα 9). Δυστυχώς, ούτε αυτή η μέθοδος φαίνεται να είναι model agnostic.

2.4 Κοντά σε εμάς - Παράλληλες Δουλειές

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

2.4.1 Linguistic Structures as Weak Supervision for Visual Scene Graph Generation [29]



Σχήμα 10: Μέθοδος από [29]. Βλέπουμε πως πρόκειται για μια περίπλοκη μέθοδο με πολλά βήματα και η οποία δεν μπορεί να γενικευθεί για άλλα SGG μοντέλα. Μέσω ενός message passing παράγονται contextual language χαρακτηριστικά τα οποία προβάλλονται σε έναν latent χώρο. Αντίστοιχα προβάλλονται οπτικά χαρακτηριστικά στον ιδιο latent χώρο. Συγκρίνοντας αυτά δημιουργείται ένα ταίριασμα και μάλιστα μπορούμε, από το ταίριασμα αυτό να παράξουμε καινούργιο με επαναληπτικό τρόπο. Τέλος, χρησιμοποιείται και ένα LSTM για βελτίωση της επίδοσης. Εικόνα από το [29]

Το αρχικό κίνητρο που παρουσιάζεται στη παρούσα δουλειά είναι η χρήση των captions για εκπαίδευση SGG μοντέλων λόγω του χαμηλότερου κόστους δημιουργίας τους, καθώς και της ύπαρξης μεγάλων συνόλων δεδομένων για image captioning. Για τη μέθοδο εκπαίδευσης από captions, ξεκινάει, όπως και εμείς, με την εξαγωγή τριπλετών επίβλεψης από τα captions με χρήση ενός OTS Scene Graph Parser [20], τα οποία χρησιμοποιεί αργότερα σαν ασθενή σήματα επίβλεψης κατά την εκπαίδευση του μοντέλου.

Σχετικά με το μοντέλο, αρχικά χρησιμοποιείται ένα message passing δίκτυο για την δημιουργία semantic contextual object features σε έναν latent χώρο. Παράλληλα, μετασχηματίζει τα visual object features στον ίδιο latent και τελικά, χρησιμοποιεί attention για να υπολογίσει μια

πιθανότητα ταιριάσματος των visual με τα contextual semantic object features. Έτσι, αντιμετωπίζει το πρόβλημα σαν ένα grounding πρόβλημα όπου ταιριάζει αντικείμενα στην εικόνα με την αντίστοιχη σημασιολογική οντότητα.

Χρησιμοποιώντας αυτά, δημιουργεί τα σήματα επίβλεψης για την πρόβλεψη του predicate μεταξύ ζευγαριών αντικειμένων στην εικόνα. Συγκεκριμένα, υπολογίζονται οι πιθανότητες $P_X[i,j]$ όπου $X \in \{det, relsub, relobj\}$ με το:

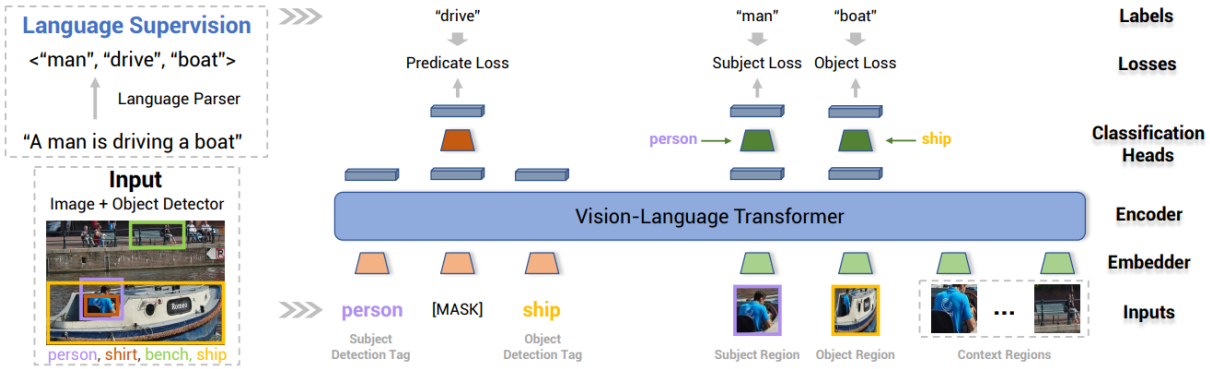
- $\mathbf{P_{det}[i,j]}$ να δείχνει τη πιθανότητα το visual region v_i να σχετίζεται με τη σημασιολογική οντότητα j
- $\mathbf{P_{relsub}[i,j]}$ να δείχνει τη πιθανότητα το visual region v_i ενός subject να λαμβάνει μέρος σε σχέση μέσω του jοστού predicate με κάποιο object
- $\mathbf{P_{relobj}[i,j]}$ να δείχνει τη πιθανότητα το visual region v_i ενός object να λαμβάνει μέρος σε σχέση μέσω του jοστού predicate με κάποιο subject

Τελικά, μπορούμε να εντοπίσουμε τριπλέτες της μορφής `<subj - pred - obj>` μέσα από το P_{rel} το οποίο προκύπτει ως:

$$P_{rel}\left[subj, pred, obj\right] = \min\left(P_{relsub}\left[subj, pred\right], P_{relobj}\left[obj, pred\right]\right)$$

Επιπλέον, με τη βοήθεια του $P_{det}[i,j]$ μπορεί να δημιουργηθεί ένας νέος πίνακας με πιθανότητες ταιριάσματος και από αυτόν να ξανά παραχθούν νέα $P_X[i,j]$, $X \in \{det, relsub, relobj\}$ επαναληπτικά σε μια διαδικασία μείωσης του θορύβου από τα image captions. Τελικά, χρησιμοποιείται και ένα Long Short-Term Memory (LSTM) δίκτυο για την βελτίωση των αποτελεσμάτων, το οποίο στη πραγματικότητα ενσωματώνει ένα γλωσσικό bias στο μοντέλο.

2.4.2 Learning to Generate Scene Graph from Natural Language Supervision [34]



Σχήμα 11: Παρατηρούμε πως τα classification heads που βρίσκονται πάνω δεξια στην εικόνα χρησιμοποιούνται για τη μετάφραση από το λεξιλόγιο του Object Detector στο λεξιλόγιο του Scene Graph Generation. Ο Vision-Language Transformer κάνει ταυτόχρονα και predicate classification και object classification. Εικόνα από το [34]

Και σε αυτή τη δουλειά, ο συγγραφείς χρησιμοποιούν σαν πρώτο βήμα, στην αντιμετώπιση του προβλήματος εκπαίδευσης από από περιγραφές εικόνων, την παραγωγή σημασιολογικών γράφων από τα image captions με χρήση OTS Scene Graph Parsers με κάποια post-processing βήματα. Κάτι σημαντικό, ωστόσο, στη συγκεκριμένη δουλειά, είναι πως το ταίριασμα των σημασιολογικών τριπλετών με τις τριπλέτες μεταξύ των αντικειμένων στην εικόνα δεν γίνεται κατά τη διάρκεια

της εκπαίδευσης. Αντίθετα, οι συγγραφείς επέλεξαν να ταιριάξουν τυχαία κάθε οντότητα που εμφανίζεται στις σημασιολογικές τριπλέτες με ένα αντικείμενο (ίδιας κατηγορίας) από την εικόνα.

Έτσι, offline δημιουργούν ένα fully supervised dataset με soft labels που προέκυψαν από τις τριπλέτες των captions και το τυχαίο ταίριασμα καθώς και από τα αντικείμενα που εντόπισε στην εικόνα ένας OTS Object Detector εκπαιδευμένος στο σύνολο δεδομένων Open Images [18].

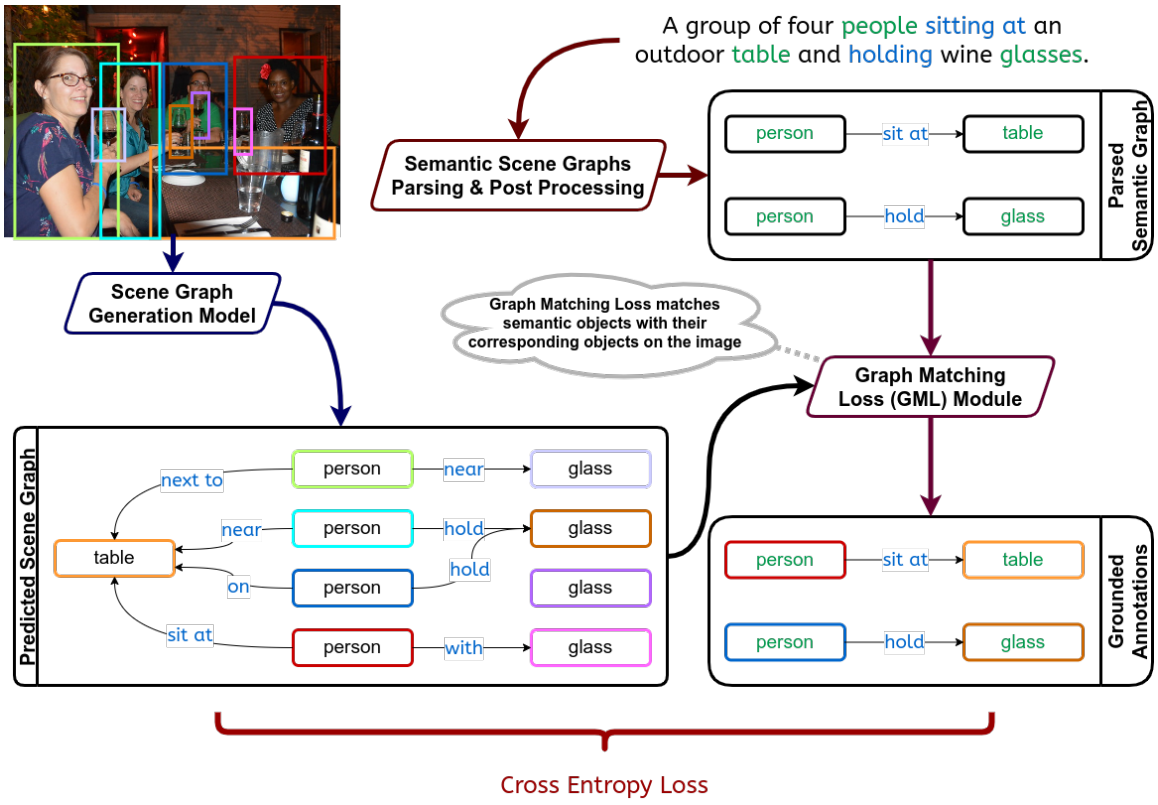
Συνεπώς, χρησιμοποιούν αυτό το fully supervised dataset για να εκπαιδεύσουν ένα SGG μοντέλο, το οποίο στη περίπτωσή τους αποτελείται από έναν Vision-Language Transformer. Κάτι που έχει ενδιαφέρον να παρατηρήσουμε είναι πως τα classification heads που βρίσκονται πάνω δεξια στην εικόνα 11 χρησιμοποιούνται για τη μετάφραση από το λεξιλόγιο του Object Detector (Open Images σε αυτή τη περίπτωση) στο λεξιλόγιο του Scene Graph Generation (VG200 κατά το evaluation). Βλέπουμε άρα τον Vision-Language Transformer να κάνει ταυτόχρονα και predicate classification και object classification.

Σημαντική, επίσης, για την βελτίωση της επίδοσης του δικτύου, είναι η χρήση ενός "Weighted Loss" κατά την εκπαίδευση. Συγκεκριμένα, χρησιμοποιείται η κατανομή των predicates του evaluation dataset VG200 για να δοθεί ανάλογη βαρύτητα σε κάθε κλάση κατά το Cross Entropy, βελτιώνοντας τα τελικά σκορ κατά 50% ή περισσότερο. Ωστόσο, αυτό αποτελεί ξεκάθαρη διαρροή δεδομένων από το test dataset μέσω της χρήση των priors του. Άλλωστε, χρησιμοποιώντας αποκλειστικά και μόνο αυτές τις prior πιθανότητες, μπορούμε να πετύχουμε ακόμα μεγαλύτερη βελτίωση των μετρικών από αυτή τη δημοσίευση, όπως θα δούμε στη συνέχεια στην ενότητα 4 (Πίνακας 5).

3 Μεθοδολογία εκπαίδευσης με δείγματα εξορυγμένα από περιγραφές εικόνων

Όπως παρουσιάσαμε προηγουμένως, τα σύνολα δεδομένων για SGG δεν εμπεριέχουν στα επισημειωμένα τους δείγματά την έννοια του saliency, με τριπλέτες συχνά να αναφέρονται σε εντελώς προφανείς ή μη σημαντικές, για το σύνολο της εικόνας, σχέσεις. Έτσι, στραφήκαμε στα σύνολα περιγραφών εικόνων που, εξ' ορισμού, ενσωματώνουν την έννοια του saliency.

Συγκεκριμένα, όταν ζητάμε από έναν annotator να περιγράψει μια εικόνα, αναγκαστικά θα πρέπει να εντοπίσει και να αναφερθεί μόνο σε αλληλεπιδράσεις που είναι σημαντικές και συμπεριλαμβάνουν την ουσία της εικόνας. Για παράδειγμα, αν έχουμε έναν άνθρωπο να κάνει σκι στις Άλπεις, δεν θα γίνεται αναφορά στη σχέση μεταξύ ανθρώπου και ουρανού, αλλά η πληροφορία θα επικεντρώνεται στον άνθρωπο, τα σκι και το χιόνι στη πλαγιά. Μάλιστα, συχνά θα χρησιμοποιήσουν αρκετά περιγραφικές και ειδικές σχέσεις έναντι γενικών, για παράδειγμα, αν ένας άνθρωπος κάνει σκι, δεν θα γίνει αναφορά στο ότι ο άνθρωπος είναι κοντά σε εξοπλισμό σκι, αλλά ότι είναι πάνω τους ή τα φοράει.



Σχήμα 12: Σύντομη περιγραφή της ιδέας. Χρησιμοποιούμε έναν Scene Graph Parser για να εξάγουμε τριπλέτες ασθενούς επίβλεψης από τα captions (ενότητα 3.2). Παράλληλα εντοπίζουμε τα αντικείμενα της εικόνας με έναν object detector (ενότητα 3.3). Τέλος, ταιριάζουμε τις προβλέψεις με τις τριπλέτες επίβλεψης, κατά την εκπαίδευση, με χρήση του Hungarian αλγορίθμου για την ελαχιστοποίηση ενός Graph Matching Loss (GML) (ενότητα 3.4).

Ωστόσο, για την εκπαίδευση μοντέλων SGG με χρήση captions χρειάζονται ορισμένες τροποποιήσεις του pipeline και του τρόπου εκπαίδευσης. Συγκεκριμένα, για κάθε δείγμα εκπαίδευσης που περιλαμβάνει εικόνα και caption, επιλέξαμε να εξορύξουμε τριπλέτες εκπαίδευσης από το captions (πάνω δεξιά στο σχήμα 12 και ενότητα 3.2) καθώς και να εντοπίσουμε τις οντότητες αντικειμένων στην εικόνα (πάνω αριστερά στο σχήμα 12 και ενότητα 3.3). Έτσι, έχοντας ζευγάρια απο τριπλέτες και αντικείμενα στην εικόνα, κατά την εκπαίδευση χρειάζεται να τα ταιριάξουμε ώστε να μπορέσουμε

να εφαρμόσουμε σήματα επίβλεψης στο εκάστοτε μοντέλο. Το ταίριασμασ αυτό γίνεται μέσω του "Graph Matching Loss Module" που υπολογίζει τη βέλτιση ανάθεση τριπλέτας στην εικόνα με τριπλέτα στο σύνολο δεδομένων (κάτω δεξιά στο σχήμα 12 και ενότητα 3.4) ώστε να μπορέσουμε να εφαρμόσουμε Cross Entropy και να εκπαιδευτεί το μοντέλο.

Έτσι, σε αυτό το κεφάλαιο θα αναλύουμε τη μέθοδο εξαγωγής τριπλετών επίβλεψης από περιγραφές εικόνων, ορισμένα προβλήματα σχετικά με το object detection, καθώς και τον αλγόριθμο ταιριάσματος των σημασιολογικών τριπλετών με τις τριπλέτες που παράγει το μοντέλο μας για κάθε ζευγάρι αντικειμένων στην εικόνα.

Πριν από αυτό όμως, θα κάνουμε μια μικρή αναφορά στα χαρακτηριστικά που κάνουν τη μέθοδό μας να αποτελεί weakly supervised μέθοδο έναντι fully supervised.

3.1 Weakly Supervised SGG

Supervision	SGG Dataset	Objects Dataset	Objects Grounding
Full	VG200	VG200	✓
Weak	VG200	VG200	✗
Weaker	COCO	VG200	✗
Weakest	COCO	Open Images	✗
Full w/ soft labels [34]	COCO	Open Images	Random Assignment

Πίνακας 1: Μπορούμε να έχουμε διάφορες παραλλαγές για εκπαίδευση με ασθενή επίβλεψη, ανάλογα με τη πληροφορία που θεωρούμε ως δεδομένη ή οχι. Η τελευταία γραμμή αναφέρεται στη μέθοδο του [34] που έχουμε αναλύσει και στην ενότητα 2.4.2.

Για να είναι πιο σαφής η διαδικασία που ακολουθούμε, καθώς ίσως δημιουργηθεί κάποια σύγχυση με τα διάφορα δεδομένα, πρέπει να ξεκαθαρίσουμε πως όταν αναφερόμαστε σε weak supervision, μπορούμε να έχουμε διαφορετικές βαθμίδες ανάλογα με το πόσο ασθενή επίβλεψη έχουμε (weak, weaker, weakest στον Πίνακα 1). Ακόμα, από εδώ και στο εξής, όταν αναφερόμαστε σε grounding, εννοούμε τη γνώση ότι μια οντότητα που αναφέρεται σε μια τριπλέτα (π.χ. άνθρωπος) σχετίζεται με μια συγκεκριμένη περιοχή στην εικόνα (ROI, bounding box). Συγκεκριμένα, για εκπαίδευση SGG μοντέλων μπορούμε να έχουμε:

- **full supervision**: χρησιμοποιούμε το fully supervised VG200 dataset (ή άλλο fully supervised SGG dataset), όπου για κάθε τριπλέτα γνωρίζουμε το grounding μεταξύ μιας οντότητας και της περιοχής στην εικόνα στην οποία αναφέρεται, για παράδειγμα, για μια τριπλέτα `<άνθρωπος - οδηγάει - ποδήλατο>` γνωρίζουμε τη θέση των ποδηλάτων και των ανθρώπων στην εικόνα αλλά και σε ποιόν άνθρωπο και ποδήλατο αναφέρεται η τριπλέτα μας.

- **weak supervision**: χρησιμοποιούμε το fully supervised VG200 dataset (ή άλλο fully supervised SGG dataset) αγνοώντας το grounding μεταξύ επισημειωμένων γράφων και εικόνας, για παράδειγμα, για μια τριπλέτα `<άνθρωπος - οδηγάει - ποδήλατο>` να γνωρίζουμε τη θέση των ποδηλάτων και των ανθρώπων στην εικόνα αλλά να μη ξέρουμε σε ποιον άνθρωπο και ποδήλατο αναφέρεται η τριπλέτα μας (εμφανίζεται στη βιβλιογραφία και ως unlocalized graphs setting)

- **weaker supervision**: χρησιμοποιούμε το dataset από περιγραφές εικόνων COCO [14] (ή άλλο caption dataset) από το οποίο θα κάνουμε extract ungrounded τριπλέτες ίδιας μορφής με πριν με χρήση ενός Semantic Scene Graph Parser. Ωστόσο τώρα δεν γνωρίζουμε τη θέση κανενός αντικειμένου στην εικόνα. Δηλαδη, στο προηγούμενο παράδειγμα, ξέρουμε ότι κάποιος άνθρωπος οδηγάει κάποιο ποδήλατο αλλά ούτε ξέρουμε που είναι αυτά στην εικόνα ούτε έχουμε εντοπίσει κάποιον άνθρωπο ή ποδήλατο στην εικόνα γενικά. Εδώ, για να

μπορέσουμε να εκπαιδεύσουμε το μοντέλο μας χρησιμοποιούμε έναν object detector για να εντοπίσουμε τα αντικείμενα στις εικόνες του dataset μας. Ανάλογα τώρα τον object detector μπορεί να αλλάξει το πόσο ασθενή επίβλεψη έχουμε. Αν χρησιμοποιήσουμε object detector εκπαιδευμένο στο VG200 (ή γενικά στο dataset στο οποίο θέλουμε να κάνουμε evaluation) τότε θα αναφερόμαστε σε weaker supervision.

- **weakest supervision**: Χρησιμοποιούμε τις extracted ungrounded τριπλέτες από το COCO dataset όπως παραπάνω, ωστόσο ο object detector είναι εκπαιδευμένος σε dataset διαφορετικό του evaluation dataset (στη περίπτωσή μας στο Open Images). Έτσι έχουμε ακόμα πιο ασθενή επίβλεψη μιας και στη προηγούμενη περίπτωση ο object detector έχει εκπαιδευτεί στην κατανομή του evaluation dataset και έχει μάθει όλα τα objects που έχει αυτό.

Στην ακόλουθη εργασία θα μας απασχολήσει κυρίως το weakest supervision από τα παραπάνω, μιας και είναι αυτό που θεωρούμε πως έχει τη μεγαλύτερη σημασία. Ο λόγος ειναι πως, για να κάνουμε weak supervision θα πρέπει να έχουμε επισημειωμένα κουτιά στις εικόνες καθώς και τριπλέτες, χωρίς όμως το μεταξύ τους grounding. Τέτοιο dataset ούτε υπάρχει και ούτε θα είχε νόημα να δημιουργηθεί. Για να έχουμε weaker εκπαίδευση, πάλι χρησιμοποιούμε τα δεδομένα από το fully supervised dataset που έχουμε για evaluation ώστε να εκπαιδεύσουμε τον detector οπότε ούτε αυτό θεωρούμε ότι είναι ένα ρεαλιστικό σενάριο. Αντίθετα, τα captioning datasets που χρειάζονται για την weakest εκπαίδευση υπάρχουν σε πολύ μεγάλες ποσότητες, είναι εύκολο να πάρουμε νέα δείγματα και, πιστεύουμε, περιέχουν τελικά όλη τη πληροφορία που χρειαζόμαστε, καταναλώνοντας τον ελάχιστο δυνατό χρόνο επισημείωσης.

3.2 Εξόρυξη τριπλετών από περιγραφές εικόνων

Πρώτο βήμα της διαδικασίας αποτελεί η εξαγωγή σημασιολογικών τριπλετών από περιγραφές εικόνας. Τη διαδικασία αυτή και τις δυσκολίες της συνοψίζει το παρακάτω παράδειγμα:

Man wearing a red shirt playing with a ball.
A tall man that wears a shirt plays with a yellow ball.
A tall man that plays with a ball wears a red shirt.

Όλες οι παραπάνω περιγραφές θα πρέπει να αντιστοιχούν στον ίδιο σημασιολογικό γράφο όπως φαίνεται παρακάτω

man	wear	shirt
man	playing with	ball

Η δυσκολία έγκειται σε πολλούς παράγοντες, μερικοί εκ των οποίων αποτελούν:

i. η πληθώρα τρόπων με την οποία μπορεί να εμφανίζεται μια σχέση (wear, wore, is wearing)

ii. γνωρίσματα μιας οντότητας (tall man, short man αντιστοιχίζονται όλα στο man)

iii. αντωνυμίες (a man that plays with a ball – το that αναφέρεται στον man)

iv. αναφορές που υπονοούν αντικείμενα (a man and another – υπονοείται another man)

v. πληθυντικός (men – αντιστοιχεί σε περισσότερους από έναν man αλλά δεν γνωρίζουμε πόσους)

Για την εξαγωγή σημασιολογικών γράφων, χρησιμοποιήσαμε υπάρχοντες αναλυτές περιγραφών σε γράφους σκηνής με τόσο θετικά όσο και αρνητικά αποτελέσματα, όπως φαίνεται στον Πίνακα 2. Συγκεκριμένα, χρησιμοποιήσαμε:

Caption	Subject	Predicate	Object	Entities	Source
A woman sits on a bench.	woman	sit	bench	woman, bench	[21]
	woman	sits on	bench	woman (a), bench (a)	[27]
	woman	sits on	bench	woman, bench	[1]
A woman that sits on a bench.	-	-	-	-	[21]
	that	sits on	bench	woman (a), that, bench (a)	[27]
	-	-	-	woman, bench	[1]
A woman holds a glass of wine next to a man.	woman woman glass	hold hold of	glass man wine	woman, glass, wine, man	[21]
	woman woman glass	holds next to of	glass man wine	woman (a), glass (a), wine, man (a)	[27]
	woman man glass	hold next to of	glass wine wine	woman, glass, wine, man	[1]
People riding their bikes.	people people people' people'	have ride have ride	bike bike bike bike	people, people', bike	[21]
	people	riding	bikes	people, bikes	[27]
	people people	ride have	bikes bikes	people, bikes	[1]
A man sitting on a bench eating a sandwich.	man bench	sit eat	bench sandwich	man, bench, sandwich	[21]
	man	sitting on	bench	man (a), bench (a), sandwich(a)	[27]
	man bench	sit on eat	bench sandwich	man, bench, sandwich	[1]
A woman is playing the piano in the room.	woman piano	play in	piano room	woman, piano, room	[21]
	woman woman	playing in	piano room	woman (a), piano (the), room (the)	[27]
	woman piano	play in	piano room	woman, piano, room	[1]
A piano is played by a woman in the room.	woman woman	play in	piano room	woman, piano, room	[21]
	woman woman	played in	piano room	woman (a), piano (the), room (the)	[27]
	woman woman	play in	piano room	woman, piano, room	[1]
A giraffe grazing a tree in the wildness with other wildlife.	tree wildness	in with	wildness wildlife	wildness, tree, wildlife (other)	[21]
	giraffe giraffe giraffe	grazing in with	tree wildness wildlife	giraffe (a), wildness (the), tree (a), wildlife (other)	[27]
	giraffe giraffe giraffe	graze graze in graze with	tree wildness wildlife	giraffe, wildness, tree, wildlife (other)	[1]

Πίνακας 2: Ανάλυση περιγραφών σε σημασιολογικούς γράφους από τρεις διαφορετικούς αναλυτές περιγραφών σε γράφους. Για κάθε caption, η πρώτη σειρά δείχνει τα αποτελέσματα από το [21], η μεσαία από το [27] ενώ η τελευταία από το [1].

i. τον rule-based Stanford Scene Graph Parser [21], έναν dependency parser Πιθανοτικής Context-Free Γραμματικής (PCFG) που σε συνδυασμό με ορισμένα post-processing βήματα και κάποιους γλωσσικούς κανόνες τελικά οδηγούν στη δημιουργία ενός Scene Graph Parser

ii. έναν "Scene Graph Parser" βασισμένο σε αυτόν του Stanford αλλά γραμμένο από την αρχή

σε python με ορισμένες αλλαγές στο πλαίσιο του [27], καθώς και

iii. τον parser από το SPICE [1] το οποίο αποτελεί μια μετρική εκτίμησης της επίδοσης μοντέλων που παράγουν περιγραφές εικόνων. Ωστόσο, για τη λειτουργία του παράγει σημασιολογικούς γράφους από τα caption, και συγκρίνει τους δύο γράφους μεταξύ τους. Το πρώτο κομμάτι της λειτουργίας είναι και αυτό που μας ενδιαφέρει αφού για αυτό χρησιμοποιείται μια παραλλαγή του rule-based Stanford Scene Graph Parser.

Παρατηρούμε πως, ο ένας από τους αναλυτές μπορεί και εντοπίζει πως μια οντότητα ειναι σε πληθυντικό αριθμό και οπότε διπλασιάζει τους κόμβους που αντιστοιχούν σε αυτήν (π.χ. people, people', Πίνακας 2). Σχετικά με τα επίθετα και τις διαφορετικές εκφάνσεις του ίδιου κατηγορήματος σε γενικές γραμμές δεν εμφανίζεται ιδιαίτερο πρόβλημα κατά τη παραγωγή του γράφου. Δυστυχώς, προτάσεις που περιέχουν αντωνυμίες (π.χ. a woman that sits on a bench, Πίνακας 2) ή πιο μεγάλες προτάσεις όπου μια λέξη προς το τέλος της πρότασης υποδηλώνει σύνδεση με μια λέξη πιο κοντά στην αρχή της (π.χ. a man sitting on a bench eating a sandwich, Πίνακας 2) δεν αναλύονται με καλή ακρίβεια από τους αναλυτές. Τελικά, για την εξαγωγή γράφων σκηνής από captions, χρησιμοποιούμε τον parser από το SPICE.

3.2.1 Scene Graph Parsing Post-Processing

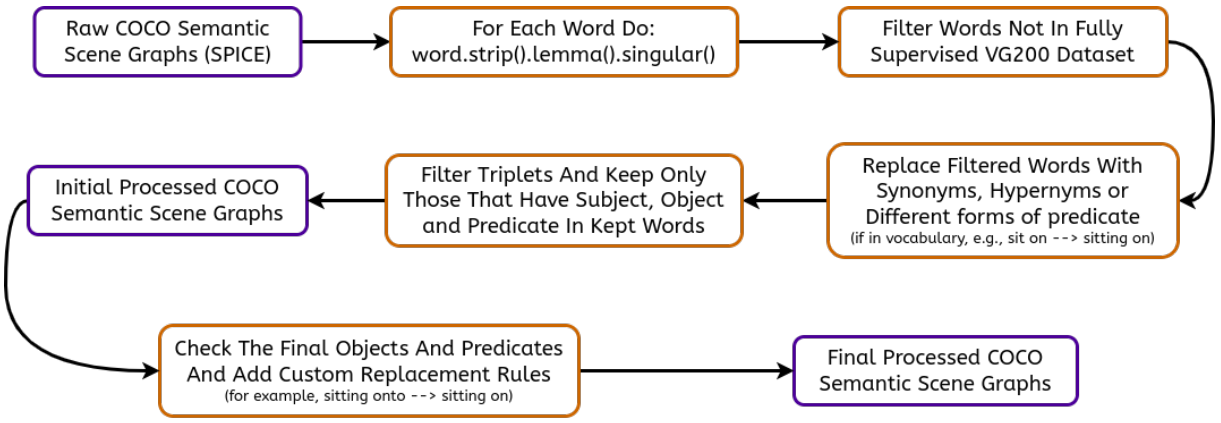
Post Processing	total		relations per image		objects per image	
	images	relations	average	max	average	max
none	58460	105277	1.8	12	2.6	10
+ syno/hyper-nym replacement rules	78445	178559	2.3	21	3.1	18
+ predicate replacement rules	93600	275296	2.9	21	3.5	18

Πίνακας 3: Στατιστικά του training dataset από την εξόρυξη τριπλετών των captions με εφαρμογή διαφορετικών βημάτων post-processing. Οι συνολικές εικόνες από το COCO dataset είναι 118287.

Εμείς με τη σειρά μας, στους γράφους που παρήγαγε το SPICE, αρχικά εφαρμόσαμε κάποια λημματοποίηση και στη συνέχεια φιλτράραμε όσες λέξεις δεν περιέχονταν στο λεξιλόγιο του VG200 όπου και θέλαμε να κάνουμε evaluation (πάνω σειρά στο σχήμα 13). Έτσι, οι τριπλέτες που είχαν το υποκείμενο, κατηγόρημα και αντικείμενό τους στο λεξιλόγιο κρατήθηκαν και καταλήξαμε με 58460 εικόνες και 105277 τριπλέτες. Ωστόσο, οι αρχικές εικόνες του dataset ήταν 118287, δηλαδή δεν αξιοποιούμε το 51% των εικόνων.

Εξαιτίας της δραματικής μείωσης του αριθμού των εικόνων του dataset που αξιοποιούμε λόγω του φιλτραρίσματος, θέλαμε να ελέγξουμε αν υπάρχουν τριπλέτες με αντικείμενα των οποίο το συνώνυμο ή το υπερώνυμο ανήκει στο λεξιλόγιο. Για να το κάνουμε αυτό, χρησιμοποιήσαμε τα synsets από τη βάση γνωσης του WordNet [17]. Έτσι, αν για μια λέξη που έχει φιλτραριστεί από το προηγούμενο βήμα, υπάρχει συνώνυμο ή υπερώνυμό της στο λεξιλόγιο, την αντικαθιστούμε με αυτό (μεσαία σειρά στο σχήμα 13). Αυτός ο κανόνας πρόσθεσε επιπλέον 19985 εικόνες και 45490 τριπλέτες στο dataset μας.

Επιπλέον, όμοια με το λεξιλόγιο των αντικειμένων, εφαρμόζουμε και ορισμένους κανόνες συγχώνευσης και αντικατάστασης για τα κατηγορήματα. Για παράδειγμα, στο λεξιλόγιο του VG200 έχουμε το κατηγόρημα "belonging to" το οποίο, λέξη προς λέξη, είναι διαφορετικό από το "belong to" παρόλο που έχουν την ίδια ακριβώς σημασία. Συνεπώς εφαρμόζουμε και μια λημματοποίηση στα κατηγορήματα του λεξιλογίου για το dataset αξιολόγησης το οποίο και συγκρίνουμε με τα λημματοποιημένα κατηγορήματα που προέκυψαν από το SPICE και ξανά αντικαθιστούμε ό,τι είναι ίδιο. Την ίδια στιγμή, προσθέσαμε κανόνες για να διορθώσουμε το προβλήματα όπως το "attached on" και το "attached onto" που λαμβάνονταν ως διαφορετικά (κάτω σειρά στο σχήμα 13). Τελικά, καταλήξαμε με 93600 εικόνες, αξιοποιώντας το 79% των εικόνων του αρχικού dataset.



Σχήμα 13: Περίληψη της διαδικασίας μετα-επεξεργασίας των τριπλετών που παράγονται από τον Εξαγωγέα Σημασιολογικών Γράφων. Έχοντας παράξει τους γράφους αυτούς, επεξεργαζόμαστε κάθε λέξη για να τη φέρουμε στην απλούστερη της μορφή και φιλτράρουμε όσες λέξεις δεν υπάρχουν στο σύνολο δεδομένων VG200. Εξαιτίας της μείωσης του αριθμού των εικόνων του dataset που αξιοποιούμε λόγω του φιλτραρίσματος, ελέγχουμε αν υπάρχουν τριπλέτες με αντικείμενα των οποίο το συνώνυμο ή το υπερώνυμο ανήκει στο λεξιλόγιο και τα αντικαθιστούμε με αυτό. Τέλος, εφαρμόζουμε και ορισμένους κανόνες συγχώνευσης και αντικατάστασης για τα κατηγορήματα και λαμβάνουμε το τελικό επεξεργασμένο dataset με ungrounded τριπλέτες το οποίο χρησιμοποιούμε για την εκπαίδευση των μοντέλων μας.

Όλα αυτά συνοψίζονται στα παρακάτω, ενώ στον πίνακα 3 βλέπουμε την επίδραση βασικών βημάτων του post-processing:

- συνώνυμο: αν μια λέξη δεν βρίσκεται στο λεξιλόγιο του VG200 αλλά ένα συνώνυμό της βρίσκεται, αντικαθιστούμε τη μια λέξη με την άλλη
- υπερώνυμο: αν μια λέξη δεν βρίσκεται στο λεξιλόγιο του VG200 αλλά ένα υπερώνυμό της βρίσκεται, αντικαθιστούμε τη μια λέξη με την άλλη
- γερούνδια, χρόνοι, αριθμοί, φωνές : αν μια λέξη (αντικείμενο ή κατηγόρημα) δεν βρίσκεται στο λεξιλόγιο του VG200 αλλά αλλάζοντας τον χρόνο, το γένος, τον αριθμό ή τη φωνή της, ταριάζει στο λεξιλόγιο, τότε κάνω την αλλαγή αυτή
- "on": αν η αντικατάσταση μίας από τις λέξεις "on", "onto", "upon" με κάποια άλλη οδηγεί στο ταίριασμα ενός κατηγορήματος με ένα κατηγόρημα στο λεξιλόγιο του VG200, τότε κάνουμε την αντικατάσταση
- "in": αν η αντικατάσταση μίας από τις λέξεις "in", "into", "inside" με κάποια άλλη οδηγεί στο ταίριασμα ενός κατηγορήματος με ένα κατηγόρημα στο λεξιλόγιο του VG200, τότε κάνουμε την αντικατάσταση

3.3 Εντοπισμός Αντικειμένων στις Εικόνες

Έχοντας πλέον έναν τρόπο εξαγωγής σημασιολογικών τριπλετών από περιγραφές εικόνων, χρειάζεται να εντοπίσουμε τα αντικείμενα μεταξύ των οποίων θα προβλέψουμε σχέσεις, και θα πρέπει αργότερα να ταιριάξουμε με κάποια από τις σημασιολογικές τριπλέτες που εξάγαμε στη προηγούμενη ενότητα.

Ο πιο εύκολος τρόπος, και ταυτόχρονα ο λιγότερο γενικεύσιμος, θα ήταν να χρησιμοποιήσουμε έναν object detector εκπαιδευμένο στο evaluation dataset, που στη περίπτωσή μας είναι το VG200, και να χρησιμοποιήσουμε τα αντικείμενα που προβλέπει αυτός. Ωστόσο, σε αυτή τη περίπτωση

έχουμε χρησιμοποιήσει δεδομένα από το dataset το οποίο είπαμε εξ' αρχής να μη χρησιμοποιήσουμε και το οποίο περιέχει επισημειωμένες τριπλέτες.

Μια άλλη ιδέα θα ήταν να χρησιμοποιήσουμε τα ήδη επισημειωμένα αντικείμενα που περιέχει το COCO dataset από το οποίο λαμβάνουμε τις περιγραφές εικόνων, με τη βοήθεια ενός λεξικού μετάφρασης των αντικειμένων του COCO σε αντικείμενα του VG200. Ωστόσο, καθώς το COCO περιέχει μόνο 80 αντικείμενα έναντι των 150 του VG200, εκ των οποίων ακόμα μικρότερο ποσοστό ταιριάζει με κάποιο αντικείμενο από το VG200, θα έπρεπε ξανά να χρησιμοποιήσουμε και κάποιον pretrained detector για τις κλάσεις αντικειμένων του VG200 που δεν υπάρχουν στο COCO. Ωστόσο, και σε αυτή τη περίπτωση, θα χρησιμοποιούσαμε τη πληροφορία των επισημειωμένων αντικειμένων στις εικόνες που δεν βρίσκεται εξ' ορισμού σε image captioning dataset, κάνοντας τη μέθοδό μας λιγότερο γενική.

Τελικά, βλέπουμε την ανάγκη χρήσης ενός object detector για τον εντοπισμό των αντικειμένων στις εικόνες του captioning dataset ο οποίος όμως να μην σχετίζεται με το Scene Graph Generation dataset μας ή με το captioning dataset. Έτσι, λοιπόν, καταλήξαμε στη χρήση ενός detector εκπαιδευμένου στο Open Images dataset το οποίο περιέχει 600 κλάσεις αντικειμένων που τελικά έχουν και μεγάλη επικάλυψη με αυτά του τελικού evaluation dataset.

3.3.1 Object Detection Post-Processing

Ωστόσο, στην περίπτωση όπου χρησιμοποιούμε έναν object detector εκπαιδευμένο στο Open Images, ή και οποιοδήποτε άλλο dataset με διαφορετικό λεξιλόγιο αντικειμένων από το evaluation dataset, θα πρέπει να δημιουργήσουμε ένα επίπεδο μετάφρασης από τις κλάσεις αντικειμένων του Open Images στις κλάσεις αντικειμένων του VG200 στο οποίο και θα κάνουμε το evaluation. Για αυτό, επανερχόμαστε στη χρήση του WordNet από τη προηγούμενη ενότητα. Κάθε κλάση του Open Images την ταιριάζουμε, αν υπάρχει, με (i) την ίδια (ii) συνώνυμη ή (iii) υπερώνυμη κλάση στο VG200.

Συγκεκριμένα, για την εκπαίδευση, κάνουμε ένα inference του object detector στην εικόνα και φιλτράρουμε τα αντικείμενα που προβλέπει. Κρατάμε, για αρχή, μόνο όσα αντικείμενα εμφανίζονται και στις τριπλέτες ασθενούς επίβλεψης, καθώς κανένα από τα άλλα αντικείμενα δεν λαμβάνει μέρος σε σχέση. Στη συνέχεια, για να μειώσουμε το φαινόμενο όπου το ίδιο αντικείμενο προβλέπεται περισσότερες από μια φορές με ελαφρώς διαφορετικό κουτί περιορισμού (bounding box), φιλτράρουμε, ανά είδος αντικειμένου, όλα εκείνα με το ίδιο είδος των οποίων τα bounding boxes έχουν μεγαλύτερη επικάλυψη από μια τιμή κατωφλίου. Τέλος, για καθένα από τα είδη αντικειμένων κρατάμε τα 4 αντικείμενα με το υψηλότερο σκορ, ενώ στο σύνολο των αντικειμένων, κρατάμε τα 12 με το κορυφαίο σκορ.

3.4 Αλγόριθμος Ταιριάσματος Γράφων (GML)

Έχοντας, λοιπόν, εξάγει γράφους από τις περιγραφές εικόνας, παρατηρούμε ότι αυτοί δεν είναι "grounded" υπό την έννοια πως δεν ξέρουμε κάθε οντότητα τους (υποκείμενο - αντικείμενο) σε ποιά οντότητα/bounding box της εικόνας αναφέρονται. Συνεπώς, θα πρέπει να υπολογίσουμε ένα βέλτιστο ταίριασμα μεταξύ του γράφου σκηνής του μοντέλου μας και του σημασιολογικού γράφου από τις περιγραφές σκηνής. Σε αυτό το ταίριασμα, άλλωστε, έγκειται και το γεγονός πως η μέθοδός μας επιβάλει ασθενή επίβλεψη (weak supervision) κατά την εκπαίδευση.

Συγκεκριμένα, αν G_i είναι ο γράφος σκηνής που παράγει το μοντέλο μας απο την εικόνα και G_s είναι ο σημασιολογικός γράφος που έχει παραχθεί από τις περιγραφές εικόνας και M είναι κάποιο ταίριασμα ανάμεσα σε γράφους (συνδέει κάθε κόμβο του ενός με έναν από τον άλλο ή και κάθε σχέση του ενός με μια του άλλου) και $\mathcal{L}$ είναι μια συνάρτηση κόστους, θέλουμε να επιλέξουμε εκείνο το ταίριασμα που ελαχιστοποιεί το $\mathcal{L}$. Συγκεκριμένα:

$$M^* = \arg\min_M \left[\mathcal{L}\left(G_i, G_s, M\right)\right]$$

Και τελικά, αν ϕ είναι οι παράμετροι του μοντέλου, όλη η διαδικασία εκπαίδευσης και εύρεσης των βέλτιστων παραμέτρων συνοψίζεται από το ακόλουθο:

$$\phi^* = \arg\min_{\phi} \mathbb{E}\left[\min_{M} \mathcal{L}\left(G_i, G_s, M\right)\right]$$

Για να βρούμε το βέλτιστο ταίριασμα χρησιμοποιούμε τον Kuhn–Munkres αλγόριθμο (ή Hungarian Algorithm) [10] όπου, δωσμένου του κόστους ταιριάσματος κάθε στοιχείου με κάθε άλλο χωριστά (του ενός γράφου με στοιχεία του άλλου), βρίσκει το συνολικό ταίριασμα ελαχίστου κόστους. Ωστόσο, θα πρέπει να καθορίσουμε και ποιο θα είναι το κόστος το οποίο θα πρεπει να ελαχιστοποιήσει ο αλγόριθμος. Το κόστος ταιριάσματος των τριπλετών θα πρεπει να λαμβάνει υπόψη τόσο το κόστος ταιριάσματος των οντοτήτων όσο και των σχέσεων μεταξύ τους. Οι τριπλέτες `<man - in - car>` και `<dog - eat - pizza>` δεν θα μπορούσαν να ταιριάζουν, αλλά `<man - in - car>` και `<man - near - car>` αποτελεί ένα πιθανό ταίριασμα για τον αλγόριθμο.

Προηγούμενες δουλειές [30] έχουν δοκιμάσει μάλιστα να χρησιμοποιήσουν ταιριάσματα 2ου βαθμού, προσεγγίζοντας με επαναληπτικές μεθόδους την ελαχιστοποίηση του κόστους τόσο για το ταιριασμα των οντοτήτων όσο και των σχέσεων, εφαρμόζοντας επαναληπτικά τον Hungarian αλγόριθμο. Ωστόσο, με βάση τα αποτελέσματά τους, αλλά και με βάση το [22], αυτή η μέθοδος δεν φαίνεται να αποδίδει τόσο καλά πρακτικά, παρά τη θεωρητική της βάση.

3.4.1 Κόστος Ταιριάσματος Τριπλετών

Τελικά, καταλήγουμε να χρησιμοποιούμε ένα ταίριασμα όπου το κόστος ταιριάσματος ανάμεσα σε δύο τριπλέτες είναι:

i το κόστος ταιριάσματος των σχέσεων, αν το υποκείμενο και το αντικείμενο των δύο τριπλετών ταιριάζει

ii ∞, αλλιώς

Συνεπώς, εφαρμόζουμε τον Hungarian αλγόριθμο μόνο μεταξύ των τριπλετών που έχουν κοινό υποκείμενο-αντικείμενο σύμφωνα με το κόστος ταιριάσματος των σχέσεων που τα ενώνει. Συγκεκριμένα, το κόστος αυτό αποτελεί το Cross Entropy Loss μεταξύ της κατανομής της πρόβλεψης του δικτύου και της επισημείωσης. Δηλαδή, αν για μια τριπλέτα i η κατανομή μεταξύ των κλάσεων που προβλέπει το δίκτυο μας είναι $\mathbf{F}_{\mathbf{pred_i}}$ και l_j είναι ένα από τα labels σχέσεων, που έχουν ίδιο όμως υποκείμενο και αντικείμενο τότε:

$$\mathcal{L}_{i,j} = CE\left(\mathbf{F}_{\mathbf{pred_i}}, l_j\right)$$

Έτσι, ο Hungarian αλγόριθμος, δοσμένου του κόστους $\mathcal{L}_{i,j} \forall i \in n_p, \forall j \in n_l$, όπου n_p, n_l ο αριθμών των τριπλετών που προβλέψαμε και των labels αντίστοιχα, βρίσκει το κατάλληλο ταίριασμα $M^* : \{0, 1, \dots n_p\} \longrightarrow \{0, 1, \dots n_l\}$ που ελαχιστοποιεί το συνολικό κόστος:

$$\mathcal{L} = \sum_{i=0}^{n_p} CE\left(\mathbf{F}_{\mathbf{pred_i}}, l_{M^*(i)}\right)$$

3.4.2 Απεξαρτητοποιώντας το Κατηγόρημα από το Saliency

Σε αυτό το σημείο οφείλουμε να αναφέρουμε και μια ακόμα λεπτομέρεια υλοποίησης. Ακολουθώντας τους [5], [32] χωρίζουμε τη πρόβλεψη μιας οπτικής σχέσης P για ένα ζευγάρι αντικειμένων S, O σε δύο ανεξάρτητα προβλήματα:

i πρόβλεψη ενός foreground score το οποίο περιγράφει πόσο foreground ή background είναι η αλληλεπίδραση των δύο αντικειμένων. Χαμηλό foreground score σημαίνει πως το ζευγάρι δεν αλληλεπιδρά με κάποια salient σχέση ενώ υψηλό score σημαίνει πως έχουμε salient αλληλεπίδραση - $Pr(foreground)$

ii πρόβλεψη ενός predicate score για κάθε αλληλεπίδραση στο λεξιλόγιο. Ανεξάρτητα με το η αλληλεπίδραση είναι foreground σύμφωνα με το παραπάνω score, μπορούν να αλληλεπιδρούν με μη salient σχέσεις. - $Pr(P)$

Τελικά, η πιθανότητα δύο αντικείμενα S, O να συνδέονται με μια σχέση P είναι:

$$Pr(P, foreground|S, O) = Pr(P|S, O)Pr(foreground|S, O)$$

Μάλιστα, ενώ στα [5], [32] το $Pr(foreground)$ περιγράφει τη πιθανότητα S, O να συσχετίζονται (άρα είναι πιο πολύ $Pr(related)$), στη περίπτωσή μας το συγκεκριμένο score περιγράφει τη πιθανότητα η συσχέτιση αυτή να είναι και salient για την εικόνα.

Επομένως, έχουμε δύο Cross Entropy Scores (CE) να υπολογίσουμε, το CE για το foreground score και το CE για το predicate score. Μάλιστα, αυτά τα δύο CE είναι και ελαφρώς διαφορετικά δίνοντας αποτελέσματα σε διαφορετικές κλίμακες. Αυτό δημιουργεί ορισμένα προβλήματα, καθώς κατά το ταίριασμα, το score λόγω του predicate είναι τουλάχιστον μια τάξη μεγέθους μεγαλύτερο από αυτό του foreground score με αποτέλεσμα το δεύτερο να μην επηρεάζει πολύ το ταίριασμα.

Λαμβάνοντας υπόψη τα παραπάνω, δοκιμάσαμε διάφορες τεχνικές. Η βασική ιδέα σχετίζεται με τη χρήση ενός βεβαρημένου άθροισμα των δύο Cross Entropy Losses σαν το Loss ταιριάσματος τριπλέτας με επισημειωμένο label ώστε να μπορούμε να δώσουμε τη κατάλληλη βαρύτητα σε κάθε όρο. Ακόμα, καθώς τα annotations ανά εικόνα είναι πολύ λιγότερα συγκριτικά με τις προβλεπόμενες σχέσεις (λόγω του αριθμού των αντικειμένων που προβλέπει ο Object Detector), και καθώς αγνοούμε για το predicate Cross Entropy όσα ζευγάρια δεν είναι foreground related, πολλαπλασιάζουμε το predicate CE score με ένα δυναμικό παράγοντα που μεγαλώνει όσο μεγαλώνει ο λόγος ανάμεσα στον αριθμό των προβλέψεων προς τον αριθμό των foreground labels.

Τελικά, μετά από αρκετά πειράματα με τη χρήση ή όχι των παραπάνω και με διάφορες τιμές για τη βαρύτητα μεταξύ κάθε όρου, καταλήξαμε πως ο δυναμικός παράγοντας για το predicate CE score ως επί το πλείστον βοηθάει λίγο το ταίριασμα, ενώ η βαρύτητα μεταξύ των δύο όρων τελικά δεν φαίνεται να είναι ιδιαίτερα σημαντική όσο κρατάμε τους πολλαπλασιαστικούς παράγοντες σε λογικά πλαίσια. Αν χρησιμοποιήσουμε έναν παράγοντα μόνο για το predicate score, τότε τιμές από 0.5 έως και 3 φαίνεται να δίνουν καλά αποτελέσματα. Δυστυχώς όμως, η βέλτιστη επιλογή αυτών των υπερπαραμέτρων φαίνεται να αλλάζει με την αλλαγή του νευρωνικού δικτύου, οπότε, μετά από αρκετές δοκιμές, θα κρατήσουμε σε όλα τα δίκτυα έναν παράγοντα 0.5 για τον πολλαπλασιασμό με το predicate score. Μάλιστα, χρησιμοποιούμε αυτόν τον παράγοντα τόσο κατά το υπολογισμό του score για το ταίριασμα των γράφων, όσο και για την υπολογισμό της συνάρτηση ελαχιστοποίησης κατά την εκπαίδευση.

3.4.3 Αλγόριθμος Graph Matching Loss Module

3.4.3.1 Αρχική αντιμετώπιση Αρχικά, δοκιμάσαμε να εφαρμόσουμε τον αλγόριθμο ταιριάσματος μεταξύ όλων των τριπλετών εκπαίδευσης και πρόβλεψης. Δηλαδή, να υπολογίσουμε το κόστος μεταξύ του ταιριάσματος κάθε πιθανού κατηγορήματος στις επισημειώσεις με κάθε κατηγόρημα στις προβλέψεις και να κάνουμε την ανάθεση που ελαχιστοποιεί το συνολικό κόστος με χρήση του Hungarian αλγορίθμου. Αυτό ωστόσο, λόγο του πολύ υψηλού θορύβου και αβεβαιότητας που εισήγαγε στην εκπαίδευση του μοντέλου μας, δεν δούλεψε. Μια άλλη σκέψη ήταν να εφαρμόσουμε ταίριασμα δύο σταδίων, όπου στο αρχικό στάδιο θα βρίσκαμε τη βέλτιστη ανάθεση για το ποιες σχέσεις είναι salient, και στο δεύτερο στάδιο θα αναθέταμε σε κάθε μία από αυτές από ένα κατηγόρημα. Ωστόσο, λόγω της αρκετά υψηλής αρχικής αβεβαιότητας για το saliency κάθε τριπλέτας, ούτε αυτή η ιδέα δούλεψε.

3.4.3.2 Υλοποίηση Τελικά, στην υλοποίησή μας εφαρμόσαμε χωριστά Hungarian Algorithms για κάθε ζευγάρι από μοναδικά (με βάσει το είδος) ζευγάρια οντοτήτων. Για παράδειγμα, αν έχουμε σαν επισημειωμένες τριπλέτες `<man - in - car>` και `<man - eat - pizza>` θα εφαρμόσουμε χωριστά Hungarian Algorithm για όλα τα ζευγάρια "man" - "car" και χωριστά για όλα τα

ζευγάρια “man” - “pizza”. Για τον υπολογισμό του ταιριάσματος, υπολογίζουμε το κόστος ανάθεσης του κατηγορήματος της επισημειωμένης τριπλέτας (“’in” και “eat” στο παράδειγμα) με κάθε κατηγόρημα που έχει προβλεφθεί από το μοντέλο για τα αντίστοιχα ζευγάρια. Το κόστος αυτό είναι ένα βεβαρημένο άθροισμα ανάμεσα στο (i) Cross Entropy κόστος μεταξύ της κατανομής του κατηγορήματος που προβλέφθηκε και του επισημειωμένου κατηγορήματος, (ii) Binary Cross Entropy κόστος ανάμεσα στην κατανομή του saliency που προβλέφθηκε και στο foreground saliency (στο αν είναι αυτή τη τριπλέτα salient και όχι background δηλαδή). Τα παραπάνω φαίνονται στον αλγόριθμο 1.

Algorithm 1 Ο ψευδοκώδικας για τον αλγόριθμο ταιριάσματος που εφαρμόζουμε χρησιμοποιώντας τον Hungarian Algorithm

Require: $0 \leq \alpha \leq 1$
  $DetTrplts \leftarrow$ detected triplets
  $AnnoTrplts \leftarrow$ weakly annotated triplets
  $APs_u \leftarrow unique(pairs(AnnoTrplts))$
  $Assignments \leftarrow [\,]$
  for all $AP_u \in APs_u$ **do**
    $DetPredsDist_{cur} \leftarrow [\,]$
    $DetSalDist_{cur} \leftarrow [\,]$
    $AnnoPreds_{cur} \leftarrow [\,]$
    for all $DetTrplt \in DetTrplts$ **do**
      if $pair(DetTrplt) == AP_u$ **then**
        $DetPredsDist_{cur}.append(distribution(predicate(DetTrplt)))$
        $DetSalDist_{cur}.append(distribution(saliency(DetTrplt)))$
      end if
    end for
    for all $AnnoTrplt \in AnnoTrplts$ **do**
      if $pair(AnnoTrplt) = AP_u$ **then**
        $AnnoPreds_{cur}.append(predicate(AnnoTrplt))$
      end if
    end for
    $PredScores_{cur} \leftarrow CEloss(allCombinations(AnnoPreds_{cur},\ DetPredsDist_{cur}))$
    $SalScores_{cur} \leftarrow BCEloss(allCombinations(AnnoPreds_{cur},\ DetSalDist_{cur}))$
    $TotalScores_{cur} \leftarrow \alpha \cdot PredScores_{cur} + (1-\alpha) \cdot SalScores_{cur}$
    $Assignments.append(HungarianAlgorithm(TotalScores_{cur}))$
  end for

3.4.4 Unlocalized Graphs Setting

Τέλος, ένα ακόμα θετικό αυτής της μεθόδου ταιριάσματος είναι πως μπορεί να χρησιμοποιηθεί τόσο στο weakest supervised setting με το οποίο ασχολούμαστε, όσο και σε λιγότερο weak supervised εκπαίδευση που αναφέρουμε στο 3.1 (weaker και weak). Συνεπώς, με το ίδιο Graph Matching Module μπορούμε να εκπαιδεύσουμε και μοντέλα στο unlocalized graphs setting.

3.5 Συνολική Περίληψη Μεθόδου

Τελικά, η μέθοδός μας ξεκινάει με την εξόρυξη τριπλετών εκπαίδευσης από τα captions του COCO με τη βοήθεια του Scene Graph Parser που χρησιμοποιείται στη μετρική του SPICE (ενότητα 3.2). Αυτές τις αρχικές τριπλέτες επεξεργαζόμαστε με γλωσσικούς κανόνες, φιλτράροντας όσες λέξεις δεν υπάρχουν στο λεξιλόγιο αξιολόγησης από το dataset VG200, και αντικαθιστώντας όσες

λέξεις φιλτράραμε με κάποιο συνώνυμο ή υπερώνυμό τους, εάν αυτό ανήκει στο λεξιλόγιο (ενότητα 3.2.1). Παράλληλα, χρησιμοποιούμε έναν Object Detector προεκπαιδευμένο στο Open Images dataset για να εντοπίσουμε τις οντότητες αντικειμένων στην εικόνα του συνόλου δεδομένων COCO (ενότητα 3.3).

Έχοντας, λοιπόν, ζευγάρια από ungrounded τριπλέτες και αντικείμενα στην εικόνα, κατά την εκπαίδευση χρειάζεται να τα ταιριάξουμε ώστε να μπορέσουμε να εφαρμόσουμε σήματα επίβλεψης στο εκάστοτε μοντέλο. Το ταίριασμα αυτό γίνεται μέσω του "Graph Matching Loss" που υπολογίζει το κόστος ταιριάσματος μεταξύ τριπλετών στην εικόνα και τριπλετών στο σύνολο δεδομένων και με χρήση του Kuhn–Munkres αλγορίθμου (ή Hungarian Algorithm) που, δεδομένου αυτού του κόστους, υπολογίζει τη βέλτιστη ανάθεση τριπλέτας στην εικόνα με τριπλέτα στο σύνολο δεδομένων που ελαχιστοποιεί το συνολικό κόστος ταιριάσματος (ενότητα 3.4).

4 Πειράματα, Αποτελέσματα και Συγκρίσεις

Σε αυτό το κεφάλαιο αρχικά θα ορίσουμε δύο νέες παραλλαγές της κλασικής μετρικής Recall@N που εκμεταλλεύονται ασθενώς επισημειωμένα δείγματα τριπλετών από περιγραφές εικόνων για την εκτίμηση του saliency ενός SGG μοντέλου. Στη συνέχεια, θα πραγματοποιήσουμε τόσο ποσοτικές όσο και ποιοτικές συγκρίσεις σε μοντέλα της βιβλιογραφίας που επανυλοποιήσαμε. Τέλος θα περιγράψουμε τα εργαλεία και τις παραμέτρους με τα οποία εκπαιδεύσαμε όλες τις παραπάνω μεθόδους που παραθέσαμε.

4.1 Εισαγωγή νέων μετρικών

Οι παραδοσιακές μετρικές που χρησιμοποιούνται για τη μέτρηση της επίδοσης SGG μοντέλων αδυνατούν να μετρήσουν το πόσο salient γράφους παράγουν τα μοντέλα αυτά. Ο λόγος είναι πως όλες εφαρμόζονται πάνω στα επισημειωμένα δεδομένα, για τα οποία δεν υπάρχει καμία εγγύηση ότι περιγράφουν salient σχέσεις όπως είδαμε στο κεφάλαιο 1.4.1. Συνεπώς, η βασική ιδέα είναι να χρησιμοποιήσουμε τους ασθενώς επισημειωμένους σημασιολογικούς γράφους που εξάγαμε από τις περιγραφές εικόνων, για να εκτιμήσουμε το saliency των μοντέλων μας. Ωστόσο, οι υπάρχουσες μετρικές για εκτίμηση της επίδοσης SGG μοντέλων χρειάζονται το grounding μεταξύ των αντικειμένων των επισημειωμένων γράφων και αυτών στην εικόνα.

Συνεπώς, εισάγουμε δύο νέες μετρικές που ονομάζουμε wR@N-bpsl (weak Recall@N - background predicate saliency) και wR@N-bsl (weak Recall@N - background saliency). Η πρώτη εκτιμάει πόσο καλά το μοντέλο μπορεί να εντοπίσει τριπλέτες (υποκείμενο - κατηγόρημα - αντικείμενο) που είναι σημαντικές, ενώ η δεύτερη μετράει πόσο καλά το μοντέλο μας μπορεί να εντοπίσει ποιά ζευγάρια αντικειμένων είναι σημαντικά και ποιά όχι (background). Έτσι, μετράμε τον αριθμό των τριπλετών (bpsl) ή ζευγαριών (bsl), που εντοπίζονται στις κορυφαίες N προβλέψεις, που ταιριάζουν σημασιολογικά/λεξικολογικά με τριπλέτες ή ζευγάρια που υπάρχουν στα ασθενώς επισημειωμένα δεδομένα.

Και οι δύο μετρικές είναι weak καθώς αποτελούν αναγκαία, αλλά όχι ικανή συνθήκη για τους γράφους ενός μοντέλου να είναι salient. Για παράδειγμα, εάν από τα επισημειωμένα δεδομένα γνωρίζω πως `<man - ride - horse>`, τότε για την bsl παραλλαγή της μετρικής με ενδιαφέρουν όλες οι προβλέψεις της μορφής `<man - [predicate] - horse>`, ανεξάρτητα με τα κουτιά περιορισμού που περιγράφουν κάθε μία από τις οντότητες. Αντίστοιχα, για την bpsl παραλλαγή της μετρικής με ενδιαφέρουν όλες οι προβλέψεις της μορφής `<man - ride - horse>`, ξανά ανεξάρτητα με τα κουτιά περιορισμού που περιγράφουν κάθε μία από τις οντότητες.

Σύμφωνα με αυτές τις μετρικές, μάλιστα, μπορούμε να δώσουμε και έναν **ασθενή ορισμό του saliency στο πλαίσιο της Παραγωγής Γράφων Σκηνής**. Θα λέμε πως για να είναι ένα μοντέλο salient θα πρέπει να μπορεί να εντοπίσει ποια ζευγάρια σημασιολογικών οντοτήτων είναι σημαντικά (η wR@N-bsl μετρική) καθώς και συνολικά ποιες σημασιολογικές τριπλέτες από `<υποκείμενο - κατηγόρημα - αντικείμενο>` είναι πιο περιγραφικές (η wR@N-bpsl μετρική), όπως προκύπτουν σύμφωνα με τις περιγραφές εικόνων. Αξίζει, επίσης, να σημειωθεί πως παρόλο που ένα μοντέλο, αν είναι salient θα πρέπει να ικανοποιεί αυτόν τον οριμό, ένα μοντέλο που τον ικανοποιεί δεν σημαίνει πως είναι υποχρεωτικά salient, δηλαδή, η συνθήκη αυτή είναι αναγκαία αλλά όχι ικανή. Ο λόγος είναι πως το μοντέλο μπορεί να έχει αναθέσει υψηλή βαθμολογία σε κάποιο ζευγάρι ή τριπλέτα η οποία σημασιολογικά φαίνεται να είναι περιγραφική, ωστόσο να έχει λανθασμένο grounding στην εικόνα, π.χ. μπορεί να λέει ορθά πως `<man - eating - pizza>` αλλά να αναφέρεται στον λάθος άνθρωπο ή στην λάθος πίτσα στην εικόνα.

4.2 Σύνολα Δεδομένων και Παραλλαγές Παραμέτρων Εκπαίδευσης

Στο σύνολο της διπλωματικής εργασίας, χρησιμοποιούμε το σύνολο δεδομένων Visual Genome (VG) που περιλαμβάνει πλήρως επισημειωμένουν γράφου σκηνής για πλήρη επίβλεψη, καθώς και το Common Objects in Context (COCO) σύνολο δεδομένων από περιγραφές εικόνων. Το VG έχει 108,077 εικόνες 3.8 εκατομμύρια επισημειωμένα αντικείμενα και 2.3 εκατομμύρια επισημειωμένες τριπλέτες. Ακολουθούμε το φιλτράρισμα των Xu et al. [28] που χρησιμοποιείται συνήθως για εκπαίδευση SGG μοντέλων με πλήρη επίβλεψη. Αυτό χωρίζεται σε 75,651/32,422 εικόνες για training/testing και περιλαμβάνει 150 κατηγορίες αντικειμένων και 50 κατηγορίες σχέσεων (predicates). Για το COCO, χρησιμοποιούμε το training split του 2017 με 118,287 εικόνες. Τέλος, αφού θα εκπαιδεύσουμε στο COCO αλλά θα κάνουμε evaluate στο test set του VG, αφαιρούμε από το COCO όσες εικόνες βρίσκονται και στο test set του VG και καταλήγουμε με 50,267 εικόνες για training.

Έτσι, ονομάζουμε COCOSG το σύνολο δεδομένων που περιλαμβάνει τις τριπλέτες που έχουμε εξάγει από το COCO με τη διαδικασία που περιγράφεται στο 3.1. Ακόμα, ως SGGfromNLS ονομάζουμε το ίδιο σύνολο δεδομένων, που όμως έχει προκύψει από το preprocessing που γίνεται στο paper [34]. Ακόμα, για καθεμία από αυτές τις παραλλαγές, θα δοκιμάσουμε τρεις κανόνες ταιριάσματος τριπλετών με αντικείμενα στην εικόνα λαμβάνοντας τις έξι παραλλαγές που φαίνονται στον πίνακα 4:

- "full random" αφορά το τυχαίο offline (πριν το training) ταίριασμα των σημασιολογικών τριπλετών από τα caption με ζευγάρια αντικειμένων (κατάλληλου είδους) που έχουν εντοπιστεί στην εικόνα από τον object detector - σύμφωνα με τη μέθοδο του [34]. Έτσι δημιουργούμε ένα fully supervised σύνολο δεδομένων με soft labels.

- "full best" είναι ίδιο με το παραπάνω, απλά ταιριάζουμε με τη χρήση μιας ευρεστικής όπου κάθε σημασιολογική τριπλέτα ταιριάζεται με το ζευγάρι αντικειμένων που έχει το μεγαλύτερο γινόμενο από objectness scores και ταυτόχρονα δεν έχει ήδη ταιριαστεί

- "weak" χρησιμοποιούμε τη μέθοδό μας για online (κατά τη διάρκεια του training) ταίριασμα των σημασιολογικών τριπλετών με κάποιο ζευγάρι αντικειμένων

Setting Variations	Caption Preprocessing	Triplet Matching	Supervision
COCOSG weak	ours	HA (ours)	weak
SGGfromNLS weak	[34]		
COCOSG full best	ours	Heuristic (ours)	full w/ soft labels
SGGfromNLS full best	[34]		
COCOSG full random	ours	random [34]	full w/ soft labels
SGGfromNLS full random	[34]		

Πίνακας 4: Παρουσίαση όλων το παραλλαγών για εκπαίδευση SGG μοντέλων από περιγραφές εικόνων που μελετάμε στη παρούσα δουλειά.

4.3 Priors

Στον πίνακα 5 μπορούμε να δούμε την επίδοση ενός απλοϊκού μοντέλου το οποίο κάνει προβλέψεις κάθε φορά με χρήση προϋπολογισμένων a priori πιθανοτήτων για κάθε ζευγάρι υποκειμένου αντικειμένου. Συγκεκριμένα, το μοντέλο υπολογίζει τη πιθανότητα $P(predicate \mid subject, object)$ $\forall subject, object, predicate \in Vocabulary$ και κατά το inference, δοσμένου του *subject* και του

Prior Probabilities	Object Detector Dataset	PredCls			SGGen		
		R@20	R@50	R@100	R@20	R@50	R@100
VG200	VG200	41.940	56.861	63.586	8.664	12.202	14.912
	Open Images				5.107	7.460	9.284
COCOSG	VG200	13.313	19.229	22.648	3.249	4.791	6.055
	Open Images				2.010	3.027	4.043
SGGFromNLS [34]	VG200	10.755	15.683	18.490	2.855	4.149	5, 241
	Open Images				1.898	2.776	3.538

Πίνακας 5: Παρουσίαση των αποτελεσμάτων ενός στοιχειώδους baseline μοντέλου που κάνει προβλέψεις χρησιμοποιώντας προ-υπολογισμένες a priori πιθανότητες για κάθε πιθανό ζευγάρι υποκειμένου αντικειμένου που βλέπει. Παρατηρούμε πως οι δικές μας τριπλέτες (COCOSG) φαίνεται να είναι καλύτερης ποιότητας από αυτές του [34] (SGGFromNLS) συγκρίνοντας την 4η με την 6η γραμμή. Ακόμα, συγκρίνοντας 4η με 5η γραμμή παρατηρούμε την ευαισθησία της μεθόδου σε χρήση διαφορετικού μοντέλου εντοπισμού αντικειμένων.

object από τον Object Detector, το μοντέλο προβλέπει σαν *predicate* αυτό με τη μεγαλύτερη πιθανότητα σύμφωνα με το ακόλουθο:

$$predicate_{predicted} = \arg\max_{predicate} P(predicate \mid subject, object)$$

Αξίζει να παρατηρήσουμε εδώ πως, συγκρίνοντας την 4η με την 6η γραμμή του Πίνακα 5, η μέθοδος που προτείνουμε για την εξαγωγή των τελικών τριπλετών εκπαίδευσης (COCOSG) φαίνεται να είναι καλύτερη από αυτή που χρησιμοποιείται στη μέθοδο του [34] (SGGFromNLS). Αντίστοιχα, από την 4η και 5η γραμμή του ίδιου πίνακα γίνεται ήδη ξεκάθαρη η ευαισθησία της μεθόδου αυτής στη χρήση διαφορετικού Object Detector. Μάλιστα αυτός είναι και ο λόγος που δεν έχει πολύ νόημα να συγκρίνουμε μεθόδους που δεν γνωρίζουμε πως χρησιμοποιούν τον ίδιο Object Detector εκπαιδευμένο στα ίδια δεδομένα και με ίδια μετα-επεξεργασία των αποτελεσμάτων του.

4.4 Ποσοτικά αποτελέσματα

4.4.1 Proof of Concept

Ξεκινώντας τα πειράματα, το βασικό κομμάτι αφορούσε την επιβεβαίωση της λειτουργικότητας της ιδέας μας που αφορούσε το ταίριασμα ungrounded τριπλετών με τριπλέτες στην εικόνα με χρήση του Hungarian Algorithm στο Cross Entropy κόστος ανάμεσα στο επισημειωμένο κατηγόρημα και στην κατανομή πρόβλεψης για το κατηγόρημα από το Scene Graph Generation μοντέλο. Έτσι, για την μελέτη αυτής της ιδέας μόνο αφαιρέσαμε από το VG200 τη πληροφορία που αφορούσε το grounding ανάμεσα στις τριπλέτες εκπαίδευσης και την εικόνα. Συγκεκριμένα, στο VG200, όπως και σε κάθε άλλο πλήρες επισημειωμένο σύνολο δεδομένων για Scene Graph Generation, υπάρχει η πληροφορία που αφορά την περιοχή της εικόνας (την οντότητα στην εικόνα) στην οποία αναφέρεται κάθε σημασιολογική οντότητα στις τριπλέτες. Αφαιρώντας αυτή τη πληροφορία μετατρέψαμε το πρόβλημα εκπαίδευσης σε Weak Supervision (βλ. ενότητα 3.1) που απαιτεί τη χρήση ενός αλγορίθμου για ταίριασμα των σημασιολογικών τριπλετών με τις grounded στην εικόνα τριπλέτες που προβλέπει το μοντέλο μας. Αυτό το ταίριασμα ανέλαβε να πραγματοποιήσει το GML module που περιγράψαμε στην ενότητα 3.4).

Όντως, στο αρχικό στάδιο των πειραμάτων μας, είναι θεμιτό να πραγματοποιούμε γρήγορα πειράματα για να δοκιμάσουμε τις ιδεές μας. Έτσι, το παρόν πείραμα πραγματοποιήθηκε με ένα μοντέλο που χρησιμοποιεί μόνο γλωσσική και χωρική πληροφορία (LangSpat μοντέλο), χωρίς να βλέπει τα pixels της εικόνας καθόλου, επιταχύνοντας κατά πολύ τους χρόνους εκπαίδευσης. Τα αποτελέσματα του πειράματος αυτού ήταν πολύ ενθαρυντικά, καθώς είδαμε μείωση των μετρικών Recall@N κατά

1.3 – 2.5% στο υποπρόβλημα του Predicate Classification και κατά 0.7 – 0.9% στο υποπρόβλημα του Scene Graph Generation (πίνακας 6). Αυτό το πείραμα μας οδηγεί στο συμπέρασμα πως η τεχνική ανάθεσης τριπλετών που χρησιμοποιούμε για την εκπαίδευση έχει σταθερές βάσεις μιας και προκαλεί μόνο μικρή μείωση στην επίδοση του μοντέλου μας, παρόλο που εκπαιδεύουμε με ασθενή επίβλεψη.

Supervision	PredCls			SGGen		
	R@20	R@50	R@100	R@20	R@50	R@100
Full	50.493	63.046	66.915	10.189	14.021	16.575
Weak	47.981	61.558	65.586	9.309	13.058	15.808
Metric Reduction	2.511	1.488	1.328	0.879	0.962	0.767

Πίνακας 6: Σύγκριση του Recall@N στο VG200 εκπαίδευση με πλήρη επίβλεψη (Full Supervision) ή με ασθενή επίβλεψη (Weak Supervisin) που δεν χρησιμοποιεί την grounding πληροφορία για τις σημασιολογικές οντότητες στις τριπλέτες (βλ. ενότητα 3.1). Βλέπουμε

4.4.2 Ablation Studies

Προκειμένου να συγκρίνουμε τις μεθόδους εξαγωγής γράφων από captions, τις διαφορετικές μεθόδους ταιριάσματος όπως παρουσιάστηκαν στο 4.2 καθώς και την ευαισθησία στη χρήση διαφορετικού Object Detector, διεξάγουμε ablation studies που παρουσιάζονται παρακάτω. Αξίζει να αναφέρουμε πως όλα τα πειράματα σε αυτή την ενότητα έγιναν με τη χρήση του μοντέλου UVTransE [8].

Dataset Variation	PredCls			SGGen		
	R@20	R@50	R@100	R@20	R@50	R@100
COCOSG weak (proposed)	**16.715**	**23.757**	**28.900**	**2.501**	**3.698**	**4.647**
SGGfromNLS [34] weak	10.008	15.008	19.066	2.003	2.790	3.446
COCOSG full best (ours)	13.705	20.378	25.174	2.210	3.315	4.220
SGGfromNLS [34] full best	11.274	17.253	21.820	2.094	2.904	3.669
COCOSG full random	10.135	15.846	20.085	1.656	2.635	3.438
SGGfromNLS full random [34]	8.970	14.625	19.320	1.860	2.675	3.359

Πίνακας 7: Σύγκριση του Recall@N στο VG200 για τις διαφορετικές παραλλαγές εκπαίδευσης με συνδυασμό διαφορετικών μεθόδων εξαγωγής γράφων από captions και ταιριάσματος γράφων όπως παρουσιάστηκαν στο 4.2.

4.4.2.1 Εξαγωγή γράφων από captions Είναι ξεκάθαρο, παρατηρώντας τον πίνακα 7 ότι η διαδικασία εξαγωγής τριπλετών από τα caption που χρησιμοποιούμε (COCOSG) οδηγεί σε μοντέλα που καταφέρνουν καλύτερα evaluation scores στο VG200 συγκριτικά με το προηγούμενο SOTA (SGGFromNLS [34]). Οι λόγοι για αυτή τη βελτίωση μπορεί να είναι διάφοροι. Χρήση διαφορετικού Semantic Parser, διαφορετικό post-processing, διαφορετικοί κανόνες κλπ. Συγκεκριμένα, στο [34] αναφέρεται η χρήση τόσο του OTS parser που χρησιμοποιούμε και εμείς όσο και του wordnet για την εύρεση συνωνύμων, υπερωνύμων και υπωνύμων. Ωστόσο, από ανασκόπηση αποτελεσμάτων, φαίνεται να μη χρησιμοποιούν αντίστοιχους κανόνες με τους δικούς μας για τη συγχώνευση λέξεων καθώς και για την αντικατάσταση ορισμένων predicates.

Όσον αφορά τις saliency μετρικές, παρατηρούμε ότι για την bpsl μετρική, που εκτιμάει το saliency στα triplets, παρουσιάζει ιδιαίτερη βελτίωση με τη χρήση της δικής μας παραλλαγής του

Dataset Variation	SGGen					
	wR@5-bsl	wR@10-bsl	wR@20-bsl	wR@5-bpsl	wR@10-bpsl	wR@20-bpsl
COCOSG weak (proposed)	28.031	38.735	48.292	**15.425**	**21.522**	**28.215**
SGGfromNLS [34] weak	27.550	35.686	44.167	13.546	18.428	23.699
COCOSG full best (ours)	29.154	**39.033**	**49.370**	15.081	20.032	25.762
SGGfromNLS [34] full best	**30.117**	38.872	48.247	12.973	16.709	20.720
COCOSG full random	27.229	35.870	45.817	13.202	17.236	21.614
SGGfromNLS full random [34]	27.412	35.205	43.502	12.331	15.723	19.390

Πίνακας 8: Σύγκριση των weak saliency μετρικών για τις διαφορετικές παραλλαγές εκπαίδευσης με συνδυασμό διαφορετικών μεθόδων εξαγωγής γράφων από captions και ταιριάσματος γράφων όπως παρουσιάστηκαν στο 4.2. Οι πρώτες τρεις στήλες (wR@N-bsl) ποσοτικοποιούν το saliency μεταξύ `subject-object`, ενώ οι επόμενες τρεις (wR@N-bpsl) το saliency ολόκληρης της τριπλέτας `<subject - predicate - object>`

triplet dataset έναντι του [34] για όλες τις παραλλαγές ταιριάσματων τριπλετών. Ωστόσο, για την bsl μετρική, που αφορά αποκλειστικά το saliency των ζευγαριών, βλέπουμε μια λιγότερο ξεκάθαρη βελτίωση, με την δική μας παραλλαγή για εξαγωγή γράφων να οδηγεί σε καλύτερα αποτελέσματα σε κάθε περίπτωση, πέρα από τη μετρική wR@5-bsl όπου η μέθοδός μας έχει το δεύτερο καλύτερο σκορ.

Για να εντοπίσουμε μια πιθανή αιτία για αυτό αρκεί να θυμηθούμε πως πρόκειται για μια μετρική που αγνοεί το predicate στις τριπλέτες. Έτσι, μπορούμε να καταλάβουμε πως, ανεξάρτητα από το post-processing των τριπλετών που εξάγονται από τα captions, η πληροφορία σχετικά με το ποιες κατηγορίες αντικειμένων αλληλεπιδρούν βρίσκεται ήδη εκεί κωδικοποιημένη στο dataset αυτό.

4.4.2.2 Αλγόριθμοι ταιριάσματος

Και για τον αλγόριθμο ταιριάσματος, παρατηρούμε μια ξεκάθαρη άνοδο του Recall στον πίνακα 7, για το δικό μας dataset, με χρήση του GML module έναντι του τυχαίου ταιριάσματος ή του "best" offline ταιριάσματος όπως περιγράψαμε στην ενότητα 4.2. Ωστόσο, για το dataset με τις τριπλέτες από το [34] βλέπουμε η δική μας μέθοδος ταιριάσματος να τα πηγαίνει χειρότερα από την "best". Πιθανολογούμε πως ο λόγος για αυτό σχετίζεται με τον αυξημένο θόρυβο στα labels του συγκεκριμένου dataset καθώς και στην αυξημένη τυχαιότητα στο training λόγω έλλειψης hyperparameter tuning και σχετικής αστοχίας στη σύγκλιση των μοντέλων.

Όσον αφορά τις saliency μετρικές, παρατηρούμε ξανά βελτίωση της bpsl μετρικής με τη χρήση του GML module έναντι των heuristic ταιριασμάτων πριν το training και στις δύο παραλλαγές του dataset. Ωστόσο, παρατηρούμε ότι για το saliency των ζευγαριών, τα ταιριάσματα με την "best" ευρεστική φαίνεται να τα καταφέρνουν καλύτερα από το GML module που προτείνουμε. Αυτό συμβαίνει καθώς η χρήση της ευρεστικής δεν επιβαρύνει τη μετρική του bsl, την διευκολύνει μάλιστα. Συγκεκριμένα, με την ευρεστική, παρόλο που μπορεί να ταιριάξουμε λάθος αντικείμενα στην εικόνα με λάθος σημασιολογικές οντότητες, θα ταιριάξουμε και πάλι ίδια είδη αντικειμένων, που είναι και αυτό που θα ελέγξει η bsl μετρική, αν δηλαδή στα κορυφαία N predictions υπάρχει επικάλυψη των κατηγοριών των αντικειμένων ενός ζευγαριού με αυτά ενός σημασιολογικού επισημειωμένου ζευγαριού. Μάλιστα, καθώς με τη χρήση του ταιριάσματος πριν την εκπαίδευση έχουμε πολύ μικρότερο θόρυβο κατά την εκπαίδευση, είναι λογικό αυτό τελικά να ευνοεί το "best" ταίριασμα. Ο λόγος που αυτή η βελτίωση δεν φαίνεται και στο "random" ταίριασμα είναι καθώς εκεί εισάγουμε έναν σταθερό θόρυβο στο μοντέλο μας πριν την εκπαίδευση όπου μπορεί και συχνά ταιριάζουμε αντικείμενα με πολύ χαμηλό objectness score με σημασιολογικές οντότητες, και μπορεί αυτά τα αντικείμενα να είναι εντελώς λάθος.

Τέλος, να σημειώσουμε πως για τις δύο τελευταίες σειρές του πίνακα 7, που αντιπροσωπεύουν τη μέθοδο ταιριάσματος του [34], τα μοντέλα δυσκολεύονται τόσο πολύ με την εκπαίδευση, που έχουν χειρότερη επίδοση στις μετρικές από ένα μοντέλο που χρησιμοποιεί μόνο a priori γνώση και δεν κοιτάει καν την εικόνα, όπως φαίνεται στον πίνακα 5.

Object Detector Finetuning Dataset		PredCls			SGGen		
for Training	for Inference (SGGen)	R@20	R@50	R@100	R@20	R@50	R@100
Open Images	Open Images	16.715	23.757	28.900	2.501	3.698	4, 647
Open Images	VG200				3.330	5.070	6.621

Πίνακας 9: Σύγκριση του Recall@N στο VG200 με χρήση διαφορετικών object detectors για την εξαγωγή των bounding boxes και labels για χρήση στη πρόβλεψη γράφων. Παρατηρούμε πως η χρήση διαφορετικής πηγής δεδομένων για την εκπαίδευση του object detector που χρησιμοποιείται για την πρόβλεψη των γράφων μπορεί να επιφέρει πολύ μεγάλες διαφοροποιήσεις στην επίδοση.

Object Detector Finetuning Dataset		SGGen					
for Training	for Inference	wR@5-bsl	wR@10-bsl	wR@20-bsl	wR@5-bpsl	wR@10-bpsl	wR@20-bpsl
Open Images	Open Images	28.031	38.735	48.292	15.425	21.522	28.215
Open Images	VG200	19.574	27.252	34.999	11.59	16.800	22.026

Πίνακας 10: Σύγκριση των weak saliency μετρικών με χρήση διαφορετικών object detectors για την εξαγωγή των bounding boxes και labels για χρήση στη πρόβλεψη γράφων. Παρατηρούμε πως η χρήση διαφορετικής πηγής δεδομένων για την εκπαίδευση του object detector που χρησιμοποιείται για την πρόβλεψη των γράφων μπορεί να επιφέρει πολύ μεγάλες διαφοροποιήσεις στην επίδοση.

4.4.2.3 Object Detector Μπορούμε εύκολα από τους πίνακες 9 και 10 να αντιληφθούμε πως η επιλογή του Object Detector έχει αρκετή επίδραση στην τελική επίδοση του μοντέλου μας. Όπου αναφέρεται "Open Images", έχουμε χρησιμοποιήσει τα bounding boxes και τα labels που προέκυψαν από έναν προεκπαιδευμένο object detector στο dataset "Open Images" και τα οποία επεξεργαστήκαμε, όπως παρουσιάσαμε στην ενότητα 3.3. Αντίστοιχα, όπου αναφέρεται "VG200", χρησιμοποιήσαμε έναν object detector εκπαιδευμένο στο VG200 για την εξαγωγή των bounding boxes και labels.

Παρατηρούμε πως στον πίνακα 9, η χρήση προεκπαιδευμένου object detector στο VG200 βοηθάει την επίδοση του μοντέλου, μιας και το αξιολογούμε στο VG200 σύνολο δεδομένων. Ωστόσο, Στον πίνακα 10 όπου αξιολογούμε το saliency του κάθε μοντέλου, παρατηρούμε πως η χρήση detector προεκπαιδευμένου στο "Open Images" βελτιώνει τα αποτελέσματα. Ο λόγος για αυτή τη συμπεριφορά είναι πως ο detector, προεκπαιδευμένος στο VG200, έχει μάθει να εντοπίζει τα αντικείμενα που λαμβάνουν μέρος σε αλληλεπιδράσεις, σύμφωνα με τους επισημειωτές, και άρα έχει μάθει μια biased κατανομή αντικειμένων που ταιριάζει στο fully supervised σύνολο δεδομένων VG200. Αυτή όμως, όπως είδαμε στην ενότητα 1.4.1, δεν είναι ιδανική για την παραγωγή salient γράφων. Έτσι, ο object detector εκπαιδευμένος στο πιο γενικό (και πιο μεγάλο) Open Images σύνολο δεδομένων οδηγεί σε καλύτερα αποτελέσματα.

Αξίζει να αναφέρουμε εδώ πως, παρόλο που η υλοποίηση στη τελευταία γραμμή του πίνακα 7 που αντιπροσωπεύει την μέθοδο του [34] βγάζει πολύ χαμηλά αποτελέσματα, στη δημοσίευση τους οι [34] αναφέρουν για τις μετρικές του Recall στο SGGen ως R@50 το 3.8% (έναντι του 2.7% στην επαναϋλοποίησή μας) και ως R@100 το 4.5% (έναντι του 3.4% στην επαναϋλοποίησή μας). Ωστόσο, όπως είναι σαφές από την παρούσα ενότητα και τον πίνακα 9, δεν μπορούμε εύκολα να συγκρίνουμε μετρικές αν έχει χρησιμοποιηθεί διαφορετικός object detector ή ακόμα και διαφορετική μετα-επεξεργασία των δεδομένων με φιλτράρισμα bounding boxes κ.α.

4.4.3 Ποσοτικά αποτελέσματα GML module σε πολλαπλά μοντέλα

Στους πίνακες 11 και 12 παρουσιάζουμε την επίδραση που έχει η ασθενής επίβλεψη με χρήση του GML σε διαφορετικά επανυλοποιημένα μοντέλα από τη βιβλιογραφία. Στον πίνακα 11 παρατηρούμε μια σημαντική αλλά αναμενόμενη μείωση του Recall για το evaluation στο VG200 dataset καθώς το μοντέλο μας έχει εκπαιδευτεί σε ένα εντελώς διαφορετικό dataset με διαφορετικές κατανομές από επισημειώσεις και εικόνες. Από τον πίνακα 12 ωστόσο είναι σαφής η βελτίωση στο saliency των

Model	PredCls			SGGen		
	R@20	R@50	R@100	R@20	R@50	R@100
UVTranE [8]	50.646	62.179	65.596	6.611	9.059	10.692
VTransE [31]	48.749	60.641	64.879	6.237	8.558	10.141
ATR-Net [5]	51.491	63.140	66.497	6.469	8.944	10.625
UVTranE + GML	16.715	23.757	28.900	2.501	3.698	4,647
VTransE + GML	15.460	22.393	27.151	2.274	3.338	4.206
ATR-Net + GML	15.171	21.559	26.500	2.584	3.789	4.720

Πίνακας 11: Αποτελέσματα από τα 4 μοντέλα που επανυλοποιήσαμε με και χωρίς GML για το VG200 στο πρόβλημα του PredCls και SGGen. Είναι σαφές πως η εκπαίδευση με weakly supervised τρόπο σε ένα διαφορετικό dataset από αυτό στο οποίο κάνουμε evaluate μειώνει αισθητά τις μετρικές του Recall.

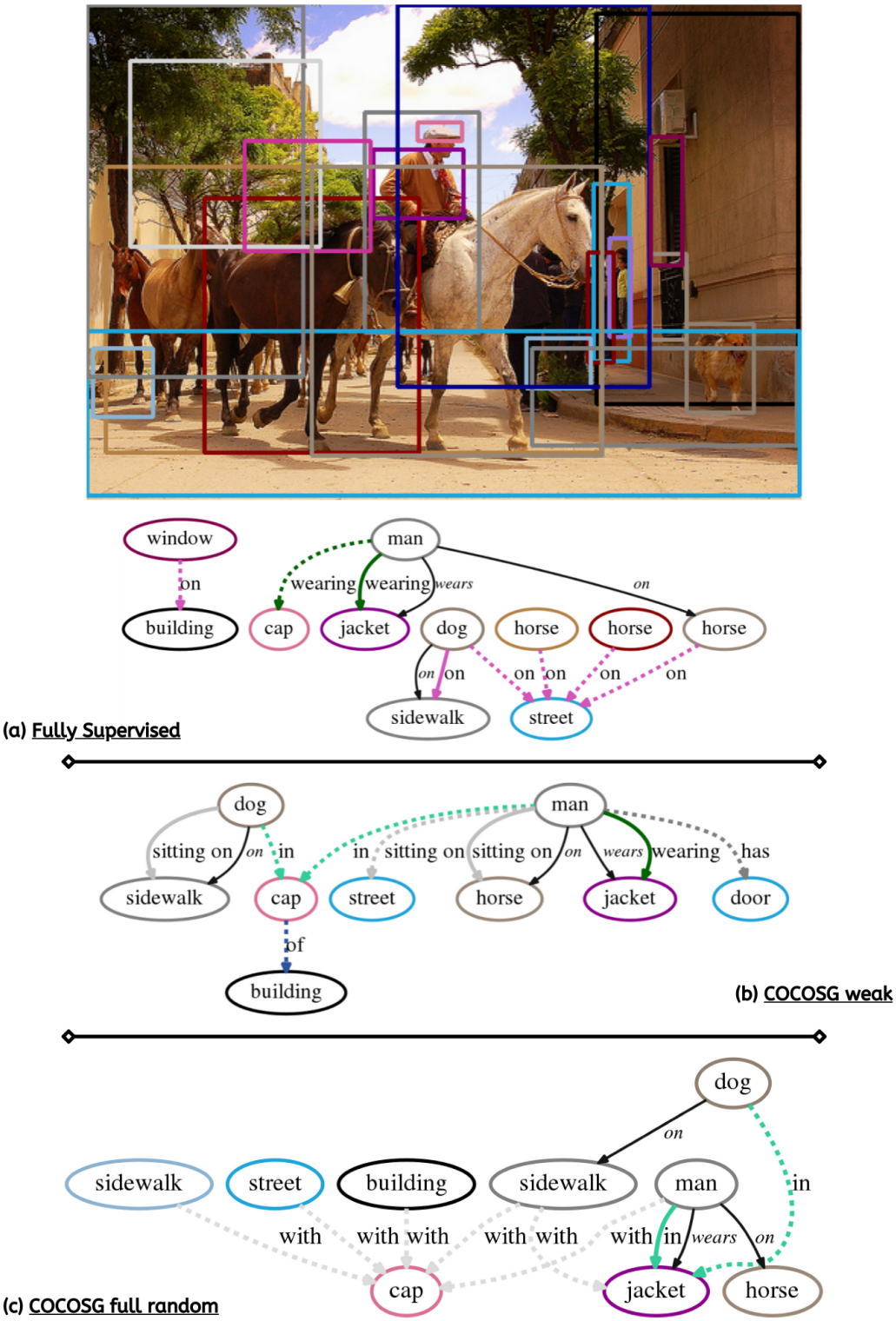
Model	SGGen					
	wR@5-bsl	wR@10-bsl	wR@20-bsl	wR@5-bpsl	wR@10-bpsl	wR@20-bpsl
UVTransE [8]	21.843	29.383	38.574	9.328	11.850	14.852
VTransE [31]	22,393	29.911	37.772	9.351	11.850	14.852
ATR-Net [5]	20.238	26.266	34.426	8.664	10.497	13.133
UVTransE + GML	28.031	38.735	48.292	15.425	21.522	28.215
VTransE + GML	29.154	39.033	49.370	15.081	20.032	25.762
ATR-Net + GML	28.833	37.703	47.376	15.425	22.072	29.177

Πίνακας 12: Αποτελέσματα για το saliency από τα 4 μοντέλα που επανυλοποιήσαμε με και χωρίς GML. Είναι σαφές πως η εκπαίδευση με weakly supervised τρόπο βελτιώνει αισθητά την επίδοση των μοντέλων στις weak saliency μετρικές (εως και 90% βελτίωση).

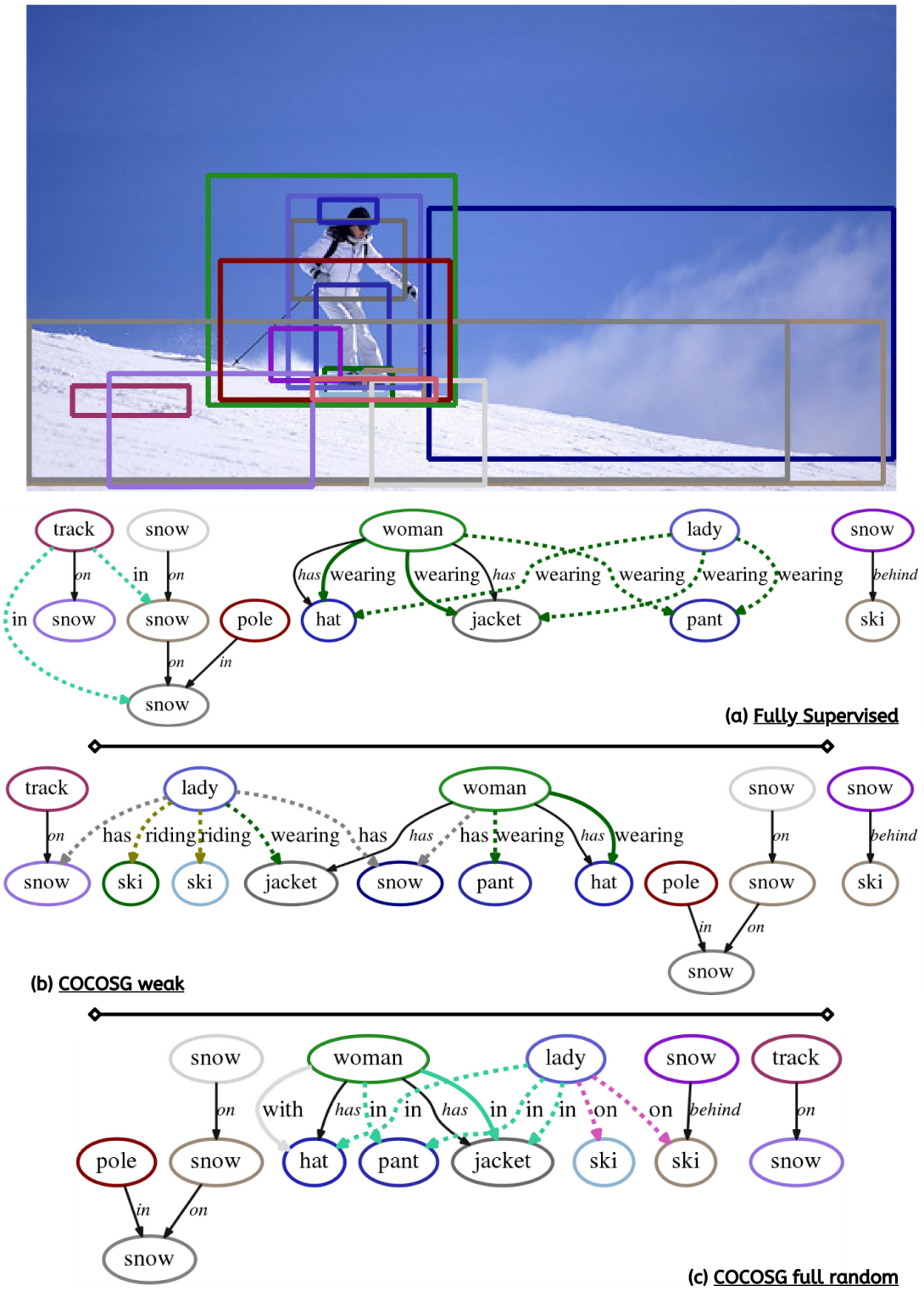
μοντέλων που επιφέρεται από τη χρήση weak supervision. Παρατηρούμε πως στην wR@20 μετρική υπάρχει βελτίωση μέχρι και 90% συγκριτικά με την εκπαίδευση με πλήρη επίβλεψη στο VG200. Αυτό επιβεβαιώνει και την αρχική μας παρατήρηση πως πλήρη επισημειωμένα σύνολα δεδομένων όπως το VG200 δεν ενσωματώνουν την έννοια του saliency στης επισημειώσεις τους και συνεπώς δεν είναι κάτι το οποίο μπορούν να μάθουν τα μοντέλα που εκπαιδεύονται σε αυτό.

4.5 Ποιοτικά αποτελέσματα

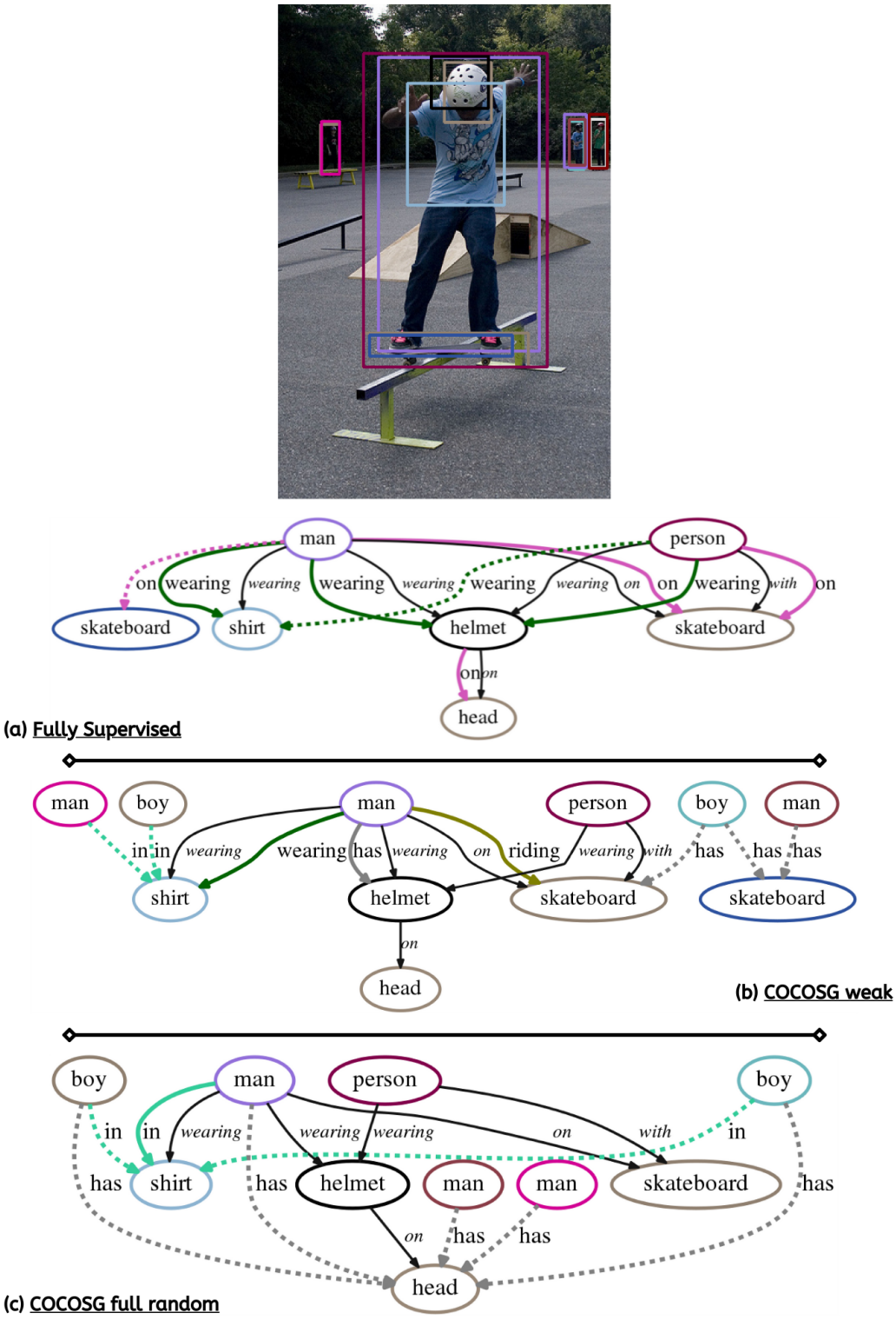
Στα σχήματα 14 έως 17 βλέπουμε ορισμένα ποιοτικά παραδείγματα που αναδεικνύουν τη βελτίωση που προκύπτει χάρη στη μέθοδό μας. Σε κάθε σχήμα υπάρχει μια εικόνα με εντοπισμένα αντικείμενα (κουτιά περιορισμού) σχεδιασμένα πάνω. Από κάτω, τρεις διαφορετικοί γράφοι σκηνής (α) ένας που έχει παραχθεί από μοντέλο με πλήρη επίβλεψη (fully supervised), (β) ένας που έχει παραχθεί με τη μέθοδό μας (COCOSG weak) στη μέση και (γ) ένας που έχει παραχθεί με τις δικές μας τριπλέτες (που σύμφωνα με την ενότητα 4 φαίνεται να είναι καλύτερης ποιότητας) αλλά τη μέθοδο από το [34] (COCOSG full random). Σε κάθε γράφο, με διακεκομμένη γραμμή σχεδιάζουμε σχέσεις που το μοντέλο προέβλεψε αλλά δεν υπήρχαν στα επισημειωμένα δεδομένα, με συνεχόμενη χρωματιστή γραμμή σημειώνουμε σχέσεις που το μοντέλο προέβλεψε και υπήρχαν (με το ίδιο ή με άλλο κατηγόρημα) στα επισημειωμένα δεδομένα, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα. Σε κάθε περίπτωση έχουμε κρατήσει μόνο τις κορυφαίες οχτώ προβλέψεις κάθε δικτύου για να καταλάβουμε ποιες σχέσεις θεωρεί πιο σημαντικές/salient κάθε φορά, ενώ για κάθε μια από τις μεθόδους χρησιμοποιήθηκε το μοντέλο UVTransE.



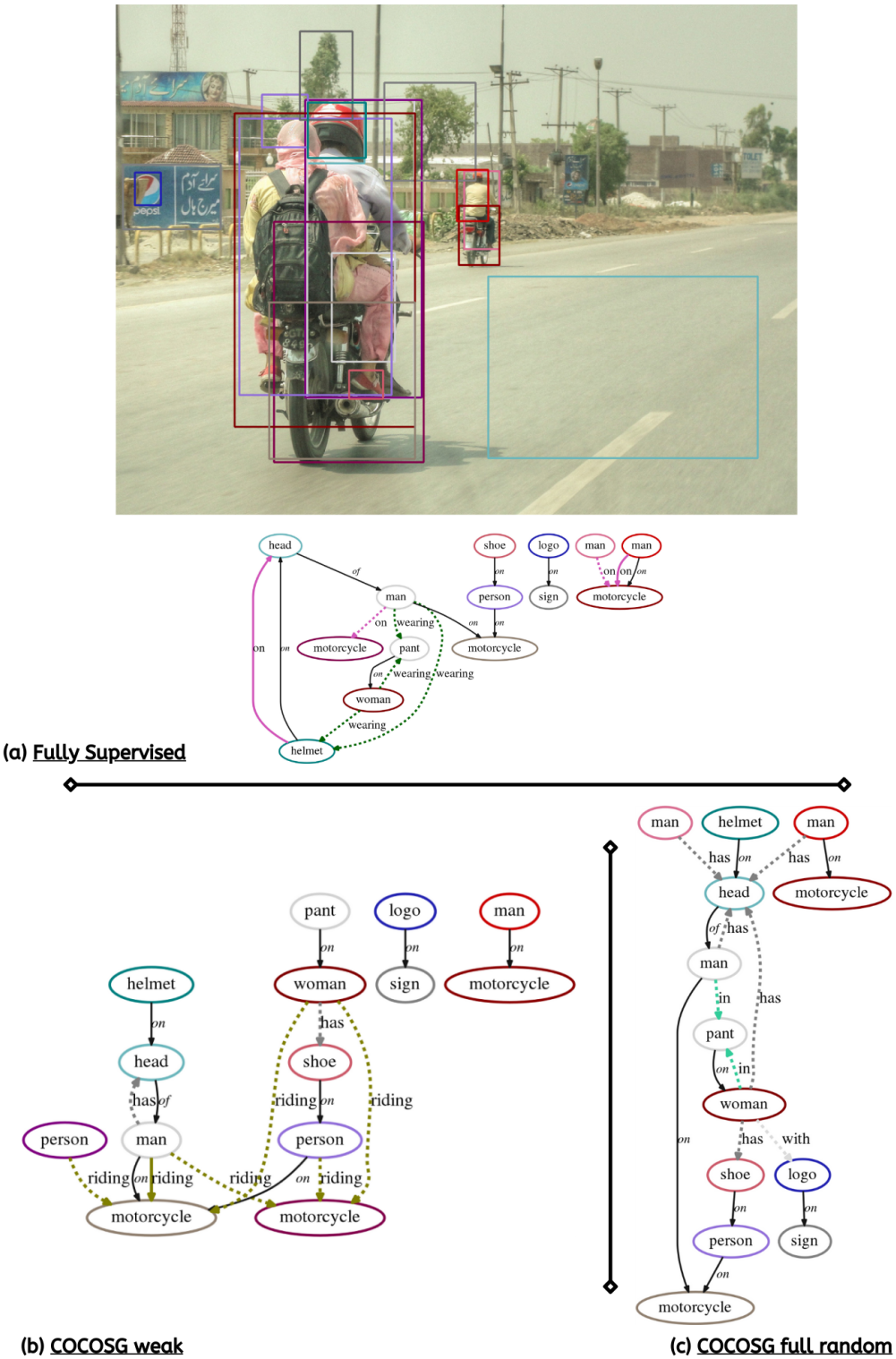
Σχήμα 14: Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας (b) έναντι της μεθόδου στο [34] (c) καθώς και εκπαίδευσης με πλήρη επίβλεψη (a). Με διακεκομμένη γραμμή σχεδιάζουμε σχέσεις που το μοντέλο προέβλεψε αλλά δεν υπήρχαν στα επισημειωμένα δεδομένα, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα. Η μέθοδός μας καταφέρνει αμέσως να εντοπίσει το κυρίως γεγονός `<man-sitting on-horse>` και μάλιστα να επιλέξει πιο salient κατηγόρημα από το "on" που υπάρχει ως επισημείωση σε αντίθεση με τις άλλες δύο μεθόδους όπου δεν εντοπίζουν καθόλου την αλληλεπίδραση αυτών των δύο οντοτήτων.



Σχήμα 15: Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας (b) έναντι της μεθόδου στο [34] (c) καθώς και εκπαίδευσης με πλήρη επίβλεψη (a). Με διακεκομμένη γραμμή σχεδιάζουμε σχέσεις που το μοντέλο προέβλεψε αλλά δεν υπήρχαν στα επισημειωμένα δεδομένα, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα. Η μέθοδος πλήρους επίβλεψης αγνοεί εντελώς το κυρίως γεγονός της εικόνας `<woman-riding-ski>` το οποίο καταφέρνει να εντοπίσει η μέθοδός μας. Μάλιστα, η αλληλεπίδραση μεταξύ `woman` και `ski` δεν είναι καν επισημειωμένη στο σύνολο δεδομένων VG200, αντίθετα επιλέχθηκε η επισημείωση της σχέσης `<snow-on-snow>`!



Σχήμα 16: Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας (b) έναντι της μεθόδου στο [34] (c) καθώς και εκπαίδευσης με πλήρη επίβλεψη (a). Με διακεκομμένη γραμμή σχεδιάζουμε σχέσεις που το μοντέλο προέβλεψε αλλά δεν υπήρχαν στα επισημειωμένα δεδομένα, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα. Η μέθοδός μας επιλέγει πιο περιγραφικό κατηγόρημα για να περιγράψει πως `<man-riding-skateboard>` έναντι του `<man-on-skateboard>`. Αντίθετα, η μέθοδος από το [34] δεν καταφέρνει να εντοπίσει πως το `skateboard>` αποτελεί σημαντικό κομμάτι της σκηνής. Δυστυχώς, αυτές οι επιτυχίες της μεθόδου μας τιμωρούνται από την Recall@N μετρική κάτι που αναδεικνύει ακόμα περισσότερο την ανάγκη ορισμού των νέων μετρικών στην ενότητα 4.1.



Σχήμα 17: Ποιοτικά παραδείγματα της αποτελεσματικότητας της μεθόδου μας (b) έναντι της μεθόδου στο [34] (c) καθώς και εκπαίδευσης με πλήρη επίβλεψη (a). Με διακεκομμένη γραμμή σχεδιάζουμε σχέσεις που το μοντέλο προέβλεψε αλλά δεν υπήρχαν στα επισημειωμένα δεδομένα, ενώ με μαύρη λεπτή γραμμή και μικρότερης γραμματοσειράς πλάγια γράμματα συμβολίζουμε τις σχέσεις που βρίσκονται στα επισημειωμένα δεδομένα. Η μέθοδός μας επιλέγει πιο περιγραφικό κατηγόρημα για να περιγράψει πως `<man-riding-motorcycle>` έναντι του `<man-on-motorcycle>`. Αντίθετα, η μέθοδος από το [34] δεν καταφέρνει να εντοπίσει πως το `motorcycle>` αποτελεί σημαντικό κομμάτι της σκηνής.

4.5.1 Πιο Salient αποτελέσματα με χρήση της μέθοδου μας

Μπορούμε να παρατηρήσουμε πως στην εικόνα 14 η μέθοδός μας (b) καταφέρνει αμέσως να εντοπίσει το κυρίως γεγονός `<man-sitting on-horse>` και μάλιστα να επιλέξει πιο salient κατηγόρημα από το "on" που υπάρχει ως επισημείωση. Αυτή η πρόβλεψη λαμβάνεται ως λάθος από τη κλασική μετρική Recall@N παρόλο που όπως τη παρατηρούμε είναι πιο σωστή και ενδιαφέρουσα από αυτή που το μοντέλο καλείται να εντοπίσει. Βλέπουμε ότι με την "random" μέθοδο του [34] (b) το μοντέλο δεν καταλαβαίνει σχεδόν καθόλου τι συμβαίνει στη σκηνή, ενώ με τη πλήρη επίβλεψη (a), δυσκολεύεται να εντοπίσει την σημαντική αλληλεπίδραση μεταξύ `man` και `horse`.

Αντίστοιχα, στην εικόνα 15 μπορούμε να δούμε πως το μοντέλο με πλήρη επίβλεψη αγνοεί εντελώς το κυρίως γεγονός της εικόνας `<woman-riding-ski>` το οποίο καταφέρνει να εντοπίσει η μέθοδός μας. Μάλιστα, εδώ αξίζει να τονίσουμε πως η αλληλεπίδραση μεταξύ `woman` και `ski` δεν είναι καν επισημειωμένη στο σύνολο δεδομένων VG200, αντίθετα επιλέχθηκε η επισημείωση της σχέσης `<snow-on-snow>`!

Στις εικόνες 16 και 17 φαίνεται για ακόμα μια φορά η μέθοδός μας να επιλέγει πιο περιγραφικό κατηγόρημα για να περιγράψει πως `<man-riding-skateboard/motorcycle>` έναντι του `<man-on-skateboard/motorcycle>` (που είναι και επισημειωμένο). Αντίθετα, η μέθοδος από το [34] δεν καταφέρνει να εντοπίσει πως το `skateboard/motorcycle>` αποτελεί σημαντικό κομμάτι της σκηνής. Δυστυχώς, αυτές οι επιτυχίες της μεθόδου μας τιμωρούνται από την Recall@N μετρική κάτι που αναδεικνύει ακόμα περισσότερο την ανάγκη ορισμού των νέων μετρικών στην ενότητα 4.1.

4.5.2 Επεξήγηση αστοχιών της μεθόδου

Οφείλουμε να εξετάσουμε τις περιπτώσεις όπως το `<cap-of-building>` στην εικόνα 14, όπου με τη μέθοδό μας προβλέπονται λανθασμένες τριπλέτες. Ο λόγος για το παραπάνω είναι η υψηλή συχνότητα με την οποία εντοπίζεται το κατηγόρημα `of` τόσο μόνο του, $Pr(P)$ όσο και σε ορισμένο περιβάλλον οντοτήτων, $Pr(P|O)$, $Pr(P|S)$, $Pr(P|S,O)$ όπου P, S, O το κατηγόρημα, υποκείμενο και αντικείμενο αντίστοιχα. Για παράδειγμα, στη περίπτωσή μας, το μοντέλο κατά την εκπαίδευση έχει δει πολλές φορές το `<window/door-of-building>` καθώς και το `<cap-of-woman/man/person>` οπότε το κατηγόρημα "of" καταλήγει να έχει ιδιαίτερα υψηλή πιθανότητα. Έτσι, παρόλο που η πιθανότητα συσχέτισης `cap` και `building` δεν είναι υψηλή (κάτι που φαίνεται και από τη χαμηλή πιθανότητα foreground που αναθέτει το μοντέλο στη σχέση), η τριπλέτα καταλήγει στις κορυφαίες οχτώ λόγω της υψηλής πιθανότητας που προέρχεται από τη πρόβλεψη του κατηγορήματος. Θεωρούμε πως πιο ποικιλόμορφα λεξιλόγια για οντότητες/κατηγορήματα ή και πιο ποικιλόμορφα caption datasets θα οδηγήσουν σε σαφή μείωση αυτού του προβλήματος και εύκολα ενσωματώνονται από τη μέθοδό μας. Ωστόσο, λόγω της χρήσης του VG200 για την αξιολόγηση των μοντέλων μας, η επιλογή του λεξιλογίου ήταν ήδη καθορισμένη για τη παρούσα δουλειά.

4.5.3 Σύνοψη

Συνολικά, είναι σαφές πως με χρήση της μεθόδου μας μπορούμε να παράξουμε γράφους που κωδικοποιούν καλύτερα την σημαντική/salient πληροφορία στην εικόνα, τόσο όσο αφορά την επιλογή περιγραφικών κατηγορημάτων, όσο και την ορθή επιλογή των ζευγαριών αντικειμένων που είναι στο προσκήνιο με βάση τα συμφραζόμενα της εικόνα. Τα σχήματα 15 και 14 αναδεικνύουν πολύ καλά την ανωτερότητα της μεθόδου μας, μιας και παρατηρούμε πως το μοντέλο που έχει εκπαιδευτεί με πλήρη επίβλεψη δεν καταφέρνει να εντοπίσει τη βασική δράση που συμβαίνει στην εκάστοτε εικόνας. Βέβαια, η βελτίωση αυτή του saliency δεν έρχεται εντελώς χωρίς προβλήματα μιας και βλέπουμε ορισμένες λάθος τριπλέτες, όπως το `<cap-of-building>` στην εικόνα 14, να εμφανίζονται στους γράφους λόγω της υψηλής πιθανότητας πρόβλεψης του υποκειμένου με το κατηγόρημα ή/και του κατηγορήματος με το αντικείμενο που προκύπτει μέσα από την παραγωγή του dataset από τα captions.

4.6 Εκπαίδευση

Υλισμικό/Λογισμικό Τα μοντέλα εκπαιδεύτηκαν σε NVIDIA 2080Ti GPU και 1080Ti GPU σε σύστημα με 64GB RAM και Ubuntu 16.04. Μία εποχή εκπαίδευσης με χρήση του Graph Matching Loss (GML) στο VG200 διαρκεί κατά μέσο όρο (ανάλογα με την αρχιτεκτονική του μοντέλου) 75 λεπτά στις 2080Ti GPU και 88 λεπτά στις 1080Ti GPU. Χωρίς GML, οι χρόνοι μεταβαίνουν σε 55 και 75 λεπτά για τις 2080Ti και 1080Ti αντίστοιχα. Αυτή η χρονική αύξηση που επιφέρει το GML είναι δικαιολογημένη καθώς χρειάζεται να τρέξουμε πολλαπλούς Hungarian Algorithms (HA) για το ταίριασμα των προβλεπόμενων τριπλετών με τα επισημειωμένα δεδομένα. Ωστόσο ο χρόνος συμπερασμού (inference time) δεν αλλάζει μιας και το GML δεν χρησιμοποιείται εκεί. Κατά μέσο όρο οι εποχές εκπαίδευσης είναι 14. Κάθε μετρική που θα παρουσιάσουμε είναι το καλύτερο αποτέλεσμα από εκπαίδευση που πραγματοποιήθηκε 2 φορές.

Επανυλοποιήσεις Για την υλοποίηση των μοντέλων χρησιμοποιήσαμε Pytorch μέσω του Pytorch Lightning. 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 [9] με weight decay ίσο με 5×10^{-5}. Δεδομένου του ότι ερευνούμε την αποτελεσματικότητα ενός module για εκπαίδευση με ασθενή επίβλεψη, λογικό είναι να συγκρίνουμε μεταξύ της ίδιας υλοποίησης των δικτύων προκειμένου να είναι συγκρίσιμα τα αποτελέσματα. Για αυτόν το λόγο υλοποιήσαμε τα VTransE [31], ATR-Net [5] και UVTransE [8]. Παρόλα αυτά, παραθέτουμε στον πίνακα 13 σύγκριση μεταξύ των δικών μας αποτελεσμάτων και εκείνων που αναφέρουν οι συγγραφείς τους. Έτσι βεβαιώνουμε ότι οι υλοποιήσεις μας είναι κοντά σε αυτές των συγγραφέων. Αποκλίσεις αφορούν περισσότερο την επιλογή υπερπαραμέτρων, παρά την υλοποίηση αυτή καθ' αυτή.

Model	Original	Ours
ATR-Net [5]	65.87	63.14
UVTransE [8]	65.3	62.18
VTransE [31]	62.63	60.64

Πίνακας 13: Σύγκριση του R@50 για το πρόβλημα του PredCls που αναφέρεται από τους συγγραφείς των μοντέλων με τις επανυλοποιήσεις μας στο VG200.

5 Επίλογος και μελλοντικές επεκτάσεις

5.1 Επίλογος

Στη παρούσα διπλωματική εντοπίσαμε και αναλύσαμε το σημαντικό πρόβλημα της έλλειψης saliency στα Scene Graph Generation μοντέλα, ενώ ταυτόχρονα προτείναμε τρόπους μέτρησης και επίλυσής του. Μέσω παρατήρησης είδαμε πως οι υπάρχουσες μετρικές δεν αρκούν για την αξιολόγηση του saliency των μοντέλων SGG. Μάλιστα, όχι μόνο αυτό αλλά και τα ίδια τα επισημειωμένα δεδομένα εκπαίδευσης δεν περιλαμβάνουν την έννοια του saliency και συχνά παρατηρούνται περιπτώσεις όπου γίνονται annotate εντελώς ανούσιες σχέσεις ή ακόμα και περιπτώσεις όπου η κεντρική οντότητα στην εικόνα δεν έχει επισημειωθεί καν.

Ακόμα, η σημαντική παρατήρηση, που θεμελίωσε και την ενασχόληση με το πρόβλημα, υπήρξε το γεγονός πως οι περιγραφές εικόνων περιλαμβάνουν εξορισμού τη πληροφορία του ποιες σχέσεις είναι σημαντικές και περιγραφικές για την εικόνα και ποιες όχι.

Έτσι για την αντιμετώπιση του προβλήματος, θέσαμε έναν ξεκάθαρο τρόπο παραγωγής σημασιολογικών γράφων σκηνής από περιγραφές εικόνων με χρήση off-the-shelf Scene Graph Parsers και πολύ συγκεκριμένων βημάτων post processing, ανάλογα το λεξιλόγιο του συνόλου δεδομένων στο οποίο θέλουμε να αξιολογήσουμε το μοντέλο. Έπειτα, εντοπίσουμε τα αντικείμενα που εμφανίζονται στην εικόνα.

Ιδιαίτερα σημαντική συνεισφορά αποτελεί η χρήση First Order Matching με τον Hungarian Algorithm για το ταίριασμα των γράφων, χρησιμοποιώντας, ωστόσο, μόνο το Cross Entropy μεταξύ της προβλεπόμενης κατανομής των predicates και του πιθανού label.

Τέλος για την αξιολόγηση και σύγκριση όλων των ιδεών μας δημιουργήσαμε δύο νέες παραλλαγές της κλασικής Recall@N, wR@N-bsl και wR@N-bpsl. Αυτές, χρησιμοποιώντας τα ασθενή σήματα επίβλεψης που εξορύξαμε από τα captions αντικατοπτρίζουν καλύτερα εάν ένα μοντέλο μπορεί να εντοπίσει τις οντότητες που είναι foreground related καθώς και το predicate που τις συσχετίζει. Μάλιστα, αυτές οι μετρικές, όντως weak, αποτελούν μια αναγκαία συνθήκη για ένα μοντέλο να παράγει salient γράφους, αλλα όχι και ικανή, καθώς δεν μετράμε κατά πόσο η σημασιολογικές οντότητες που συνδέονται είναι και σωστά grounded στην εικόνα.

Συνολικά, με χρήση της μεθόδου μας μπορούμε να παράξουμε γράφους που κωδικοποιούν καλύτερα την σημαντική/salient πληροφορία στην εικόνα, τόσο από θέμα επιλογής περιγραφικών κατηγορημάτων, όσο από ορθής επιλογής των ζευγαριών αντικειμένων που είναι στο προσκήνιο με βάση τα συμφραζόμενα της εικόνας. Ωστόσο, η βελτίωση αυτή του saliency δεν έρχεται εντελώς χωρίς προβλήματα μιας και εμφανίζονται ορισμένες λάθος τριπλέτες στους γράφους λόγω της υψηλής πιθανότητας πρόβλεψης του υποκειμένου με το κατηγόρημα ή/και του κατηγορήματος με το αντικείμενο, όπως προκύπτει μέσα από την παραγωγή του συνόλου δεδομένων από τις περιγραφές εικόνων.

5.2 Μελλοντικές επεκτάσεις

Πιστεύουμε ότι η χρήση των περιγραφών εικόνων είναι όντως κομβική για την παραγωγή γράφων σκηνής με ενδιαφέρον, όπως φαίνεται άλλωστε και από την βελτίωση των saliency μετρικών που εισάγαμε. Ωστόσο, η διαδικασία παραγωγής ασθενών γράφων από τα caption με τη χρήση υπαρχόντων εργαλείων και στη συνέχεια ο εντοπισμός αντικειμένων με προεκπαιδευμένους object detectors δεν είμαστε τελικά σίγουροι ότι αποτελεί τη βέλτιστη λύση του προβλήματος.

Ο λόγος είναι πως υπάρχουν πολλά πιθανά σημεία αστοχίας, τα οποία εντοπίζονται στα ποιοτικά αποτελέσματα. Αστοχίες στον object detector που μπορεί να μην εντοπίσει ένα σημαντικό αντικείμενο της εικόνας, αστοχίες στον semantic Scene Graph Parser που δυσκολεύεται να εντοπίσει ορισμένες σχέσεις ή εντοπίζει λανθασμένες τριπλέτες, αστοχίες, ακόμα, και στα post processing στάδια που εφαρμόζουμε στις εξόδους του object detector ή του Scene Graph Parser. Η μέθοδος αυτή βασίζεται σε πολλά hardcoded κομμάτια και σε ορισμένες θεωρήσεις που δεν έχουμε τη βεβαίωση

ότι ισχύουν πέρα από τα πειραματικά αποτελέσματα που εξάγαμε. Και παρόλο που η μέθοδος ταιριάσματος με χρήση First Order Matching πάνω στα Cross Entropies δίνει πολύ καλά αποτελέσματα, μελλοντική δουλειά θα επικεντρωθεί σε διαφορετικό τρόπο εκπαίδευσης από περιγραφές εικόνων.

Σκοπός μας είναι να αφαιρέσουμε εντελώς το στάδιο παραγωγής σημασιολογικών γράφων από τα captions, και να επικεντρωθούμε στην παραγωγή των γράφων με κάποιο attention distillation από τα captions κατά τη διαδικασία της εκπαίδευσης. Το attention αυτό θα αφορά τόσο τις οντότητες που εμφανίζονται στο caption όσο και των σχέσεων που τις συνδέει. Η ιδέα είναι να εκπαιδεύσουμε από άκρη σε άκρη (end-to-end) ένα δίκτυο που να εντοπίζει σημαντικές τριπλέτες χωρίς χρήση ξεχωριστού object detector. Για την επίβλεψη και εκπαίδευσή του, ο γράφος που παράγει το δίκτυό μας θα περνάει από ένα Graph Neural Network (GNN) το οποίο θα παράγει ένα caption από τον γράφο και στη συνέχεια θα κάνουμε επίβλεψη σε όλο το δίκτυο συνολικά. Άλλωστε, ξέρουμε ήδη από το [11] πως από αναπαραστάσεις γράφων μπορούμε να εξάγουμε περιγραφές εικόνων. Έτσι, θα μπορούσαμε να εξαφανίσουμε όλα τα post-processing στάδια, αλλά και την εξάρτηση από off-the-shelf Scene Graph Parsers ή Object Detectors που φαίνεται να αποτελούν το κύριο πρόβλημα της διαδικασίας.

Μια ακόμα κατεύθυνση που θα θέλαμε να μελετήσουμε στο μέλλον είναι η εκπαίδευση μοντέλων SGG με unsupervised ή self-supervised τρόπο. Σε αυτή τη περίπτωση, θα μετασχηματίζαμε το πρόβλημα ως το reconstruction ενός μέρους μιας εικόνας (ή έστω του feature map της περιοχής) με χρήση του γράφου σκηνής που θα προέβλεπε ένα SGG μοντέλο. Για να μπορέσει, ωστόσο, αυτή η μέθοδος να συμπεριλάβει και την έννοια του saliency, θα χρειαζόταν κάποιος συνδυασμός από weak και self supervised εκπαίδευση.

Σε κάθε περίπτωση, θα πρέπει να έχει γίνει σαφές μέχρι τώρα πως τα επισημειωμένα σύνολα δεδομένων είναι σε ένα βαθμό θεμελιωδώς ελαττωματικά, συνεπώς ο σκοπός της έρευνας, πια, στο SGG δεν θα πρέπει να είναι η εκπαίδευση μοντέλων που να μαθαίνουν καλύτερα το VG200 για παράδειγμα, αλλά η εκπαίδευση μοντέλων που χρησιμοποιούν όλο και λιγότερο τέτοια επισημειωμένα σύνολα δεδομένων.

Απόδοση ξενόγλωσσων όρων

Table 14: Απόδοση ξενόγλωσσων όρων

Ξενόγλωσσος Όρος	Απόδοση
<subject - predicate - object>	<υποκείμενο - κατηγόρημα - αντικείμενο>
Attention (mechanisms)	Μηχανισμοί Προσοχής
Backbone	Ραχοκοκαλιά
Bias	Προκατάληψη/Προδιάθεση
Bidirectional Gated Recurrent Units	Αμφίδρομη Αναδρομική Μονάδα με Πύλες
Binary Cross Entropy	Δυαδική Διασταυρούμενη Εντροπία
Bottleneck	Σημείο Συμφόρησης
Bounding boxes	Κουτιά Περιορισμού
Constrained convex optimization	Περιορισμένη Κυρτή Βελτιστοποίηση
Convolutional Neural Network	Συνελικτικό Νευρωνικό Δίκτυο
Classification Heads	κεφαλές κατηγοριοποίησης
Cross Entropy	Διασταυρούμενη Εντροπία
Dataset	Σύνολο Δεδομένων
Deep Supervision	Βαθιά Επίβλεψη
End-to-end (training)	Εκπαίδευσης από άκρη σε άκρη
Evaluation	Αξιολόγηση
Features	Χαρακτηριστικά
Feature maps	Χάρτες Αναπαράστασης
Feature vector	Διάνυσμα Χαρακτηριστικών
Few-shot	Με λίγα δείγματα εκπαίδευσης
Finetuned	Προεκπαιδευμένα μοντέλα που εκπαιδεύουμε ένα μέρος τους σε ένα νέο σύνολο δεδομένων
First Order Matching	Ταίριασμα Πρώτου Βαθμού
Fully Supervised	με/για Πλήρη Επίβλεψη
Graph Neural Network (GNN)	Νευρωνικό Δίκτυο Γράφου
Grounding	το ταίριασμα μιας σημασιολογικής οντότητας με μια περιοχή στην εικόνα
Hardcoded	Ενσωματωμένο στον κώδικα από τον προγραμματιστή
Image captioning	το πρόβλημα της παραγωγής περιγραφής για μια εικόνα
Image captions	Περιγραφές Εικόνων
Latent	Λανθάνον
Linguistic	Γλωσσικός/Σημασιολογικός
Long Short-Term Memory	Μεγάλη Βραχυπρόθεσμη μνήμη

Continued on next page

Table 14: Απόδοση ξενόγλωσσων όρων (Continued)

Message Passing	Πέρασμα Μηνυμάτων
Model agnostic	Ανεξάρτητο του Μοντέλου
Object detector	δικτύου εντοπισμού οντοτήτων
Off-the-shelf	Ανεπτυγμένο και διαθέσιμο για χρήση
Offline	όχι κατά τη διάρκεια της εκπαίδευσης
Post-processing	Μετα-επεξεργασία
Predicate Classification	Κατηγοριοποίηση Σχέσεων
Priors	εκ των προτέρων πιθανότητες υπολογισμένες από την κατανομή που έχει το σύνολο δεδομένων, P(predicate | subject, object)
Saliency	Σημαντικότητα/Περιγραφικότητα
Scene Graph Generation (SGG)	Παραγωγή Γράφου Σκηνής
Scene Graph Parser	Εξαγωγέας Γράφου Σκηνής από περιγραφές εικόνας
Self Supervised	Αυτοεπιβλεπόμενο
Semantic Contextual Object Features	Σημασιολογικά με Συμφραζόμενα Χαρακτηριστικά Αντικειμένου
Soft Labels	Επισημειωμένα δείγματα που δεν έχουν προκύψει με ανθρώπινη επισημείωση
State-of-the-art	Τελευταίας Τεχνολογίας
Test set	Σύνολο Δεδομένων Δοκιμής
Train set	Σύνολο Δεδομένων Εκπαίδευσης
Transformer	ένα μοντέλο που χρησιμοποιεί μονάδες μηχανισμών προσοχής σε πολλαπλά επίπεδα
Unlocalized	Σημασιολογικές οντότητες που δεν γνωρίζουμε το ταίριασμά τους με κάποια περιοχή στην εικόνα
Unsupervised	Χωρίς επίβλεψη
Visual	Οπτικός
Visual Region	Οπτική Περιοχή
Visual Semantic Parsing	Οπτική Σημασιολογική Ανάλυση
Weight Decay	Περιοριστής Τιμών Βαρών του Δικτύου
Weighted Loss	Συνάρτηση κόστους που δίνει διαφορετική βαρύτητα σε διαφορετικούς όρους/κατηγορίες
Word embeddings	Λεξιλογικά Διανύσματα
Zero-shot	Χωρίς δείγματα εκπαίδευσης

Παραρτήματα

A Σύνολα δεδομένων

Υπάρχει πληθώρα συνόλων δεδομένων για το πρόβλημα του εντοπισμού σχέσεων [15, 13, 31, 2, 28, 33]. Στον πίνακα 15 παρουσιάζουμε τα βασικά στατιστικά τους. Το βασικό σύνολα δεδομένων

Dataset	Train/Test images	Predicates	Objects
VRD[15]	4k/1k	70	100
VG-MSDN[13]	46.2k/10k	50	150
VG-VTE[31]	73.8k/25.8k	100	200
sVG[2]	64.7k/8.7k	24	399
VG200[28]	75.6k/32.4k	50	150
VG80K[33]	99.9k/4.8k	29086	53304

Πίνακας 15: Απαρίθμηση των συνόλων δεδομένων της βιβλιογραφίας με τις χαρακτηριστικές στατιστικές πληροφορίες τους.

που χρησιμοποιούμε για τη διεξαγωγή των πειραμάτων είναι το VG200, ένα από τα πιο διαδεδομένα στη βιβλιογραφία. Στην εικόνα 18 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 16. 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

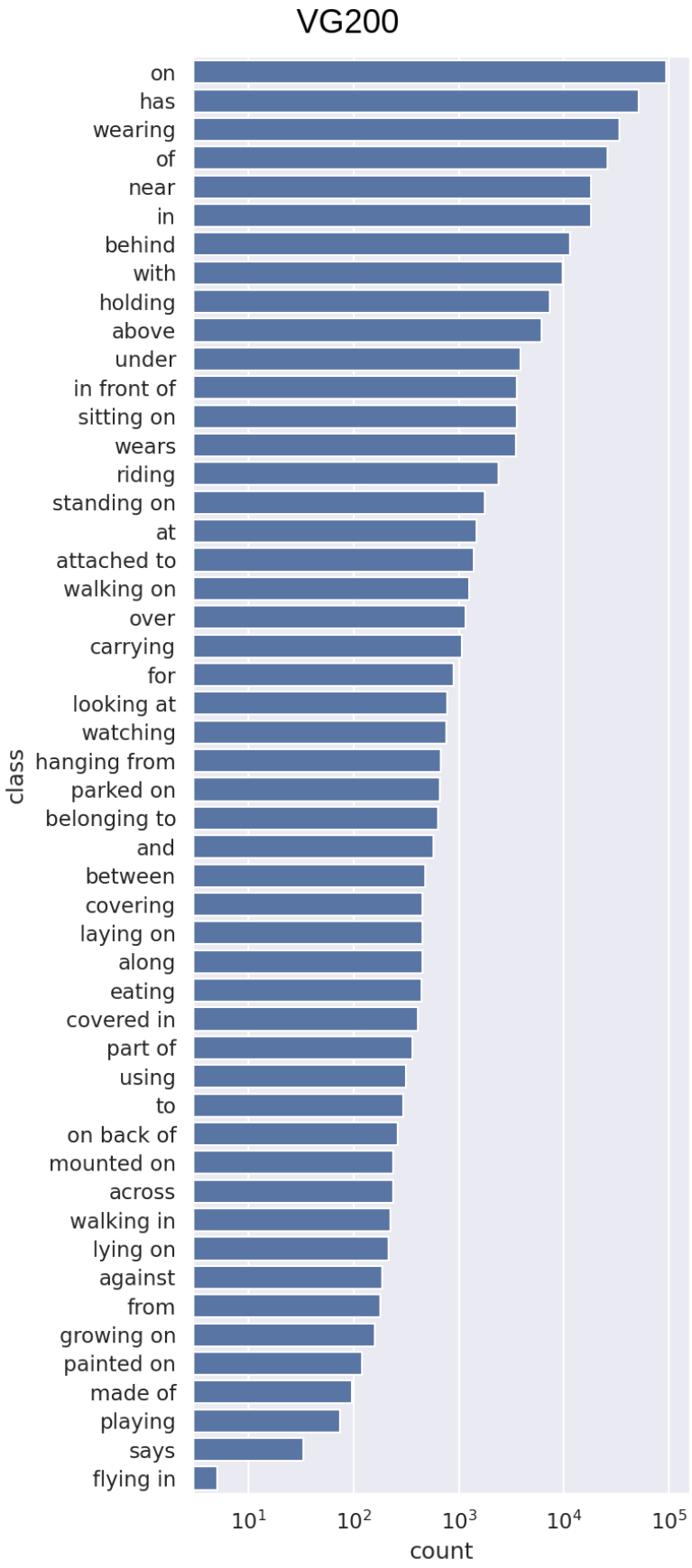
Πίνακας 16: Απαρίθμηση των παραλλαγών της ανίχνευσης οπτικών σχέσεων. `yes` σημαίνει πως η συγκεκριμένη παραλλαγή χρησιμοποιεί την αντίστοιχη πληροφορία ενώ σε αντίθετη περίπτωση σημειώνουμε `no`.

- 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 [5] ορίζουν τη μετρική Rk@x όπου για n ζευγάρια αντικειμένων σε μία εικόνα κρατάει τις x πιο πιθανές προβλέψεις από συνολικά nk για να μετρήσει το Recall. Στην παρούσα διπλωματική, οποιαδήποτε αναφορά σε Recall υπονοεί $k = 1$.



Σχήμα 18: Αριθμός δειγμάτων ανά κλάση σε λογαριθμική κλίμακα στο σύνολο δεδομένων εκπαίδευσης για το VG200.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. "SPICE: Semantic Propositional Image Caption Evaluation". In: *ECCV*. 2016.

[2] Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting Visual Relationships with Deep Relational Networks". In: *Proc. CVPR*. 2017.

[3] Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. "Grounding Consistency: Distilling Spatial Common Sense for Precise Visual Relationship Detection". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 15891–15900. DOI: 10.1109/ICCV48922.2021.01561.

[4] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. "Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction". In: *Proc. ICCV Workshops*. 2019.

[5] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. "Attention-Translation-Relation Network for Scalable Scene Graph Generation". In: *Proc. ICCV Workshops*. 2019.

[6] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, Athanasia Zlatintsi, and Petros Maragos. "Deeply Supervised Multimodal Attentional Translation Embeddings for Visual Relationship Detection". In: *Proc. ICIP*. 2019.

[7] Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. "From Saturation to Zero-Shot Visual Relationship Detection Using Local Context". In: *Proc. BMVC*. 2020.

[8] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. "Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation". In: *PAMI* (2020).

[9] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014).

[10] H.W. Kuhn. "The Hungarian Method for the Assignment Problem". In: *Naval Res. Logist. Quart.* 2 (Jan. 1955), pp. 83–98. DOI: 10.1002/nav.20053.

[11] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. "Baby talk: Understanding and generating simple image descriptions". In: *CVPR 2011*. 2011, pp. 1601–1608. DOI: 10.1109/CVPR.2011.5995466.

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

[13] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. "Scene Graph Generation from Objects, Phrases and Region Captions". In: *Proc. ICCV*. 2017.

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[15] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. "Visual Relationship Detection with Language Priors". In: *Proc. ECCV*. 2016.

[16] Li Mi and Zhenzhong Chen. "Hierarchical Graph Attention Network for Visual Relationship Detection". In: *Proc. CVPR*. 2020.

[17] George A. Miller. "WordNet: A Lexical Database for English". In: *Commun. ACM* 38.11 (Nov. 1995), pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: https://doi.org/10.1145/219717.219748.

[18] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. "We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 854–863.

[19] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. "Weakly-Supervised Learning of Visual Relations". In: *Proc. ICCV*. 2017.

[20] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. "Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval". In: *Workshop on Vision and Language (VL15)*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015.

[21] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. "Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval". In: *Workshop on Vision and Language (VL15)*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015.

[22] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. "A Simple Baseline for Weakly-Supervised Scene Graph Generation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 16373–16382. DOI: 10.1109/ICCV48922.2021.01608.

[23] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[24] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *CoRR* abs/1602.07261 (2016).

[25] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. "Unbiased Scene Graph Generation from Biased Training". In: *Proc. CVPR*. 2020.

[26] W. Wang, Meng Wang, Sen Wang, Guodong Long, L. Yao, G. Qi, and Y. A. Chen. "One-Shot Learning for Long-Tail Visual Relation Detection". In: *Proc. AAAI*. 2020.

[27] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. "Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 6602–6611.

[28] Danfei Xu, Yuke Zhu, Christopher Bongsoo Choy, and Li Fei-Fei. "Scene Graph Generation by Iterative Message Passing". In: *Proc. CVPR*. 2017.

[29] Keren Ye and Adriana Kovashka. "Linguistic Structures as Weak Supervision for Visual Scene Graph Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021.

[30] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. "Weakly Supervised Visual Semantic Parsing". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[31] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. "Visual Translation Embedding Network for Visual Relation Detection". In: *Proc. CVPR*. 2017.

[32] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. "PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN". In: *Proc. ICCV*. 2017.

[33] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed M. Elgammal, and Mohamed Elhoseiny. "Large-Scale Visual Relationship Understanding". In: *Proc. AAAI*. 2019.

[34] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. "Learning to Generate Scene Graph from Natural Language Supervision". In: *ICCV*. 2021.