Αριστοτελειο Πανεπιστημιο Θεσσαλονικης
Πολυτεχνικη Σχολη
Τμημα Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Νευρωνικα Δικτυα Ανταλλαγης Μηνυματων και Συνελικτικα Δικτυα Γραφων για Εξαγωγη Γραφου Σκηνης Εικονων

Διπλωματική Εργασία

Μιλτιάδης Κοφινάς

ΑΕΜ: 7458

Θεσσαλονίκη, Οκτώβριος 2018

Επιβλέπων Καθηγητής:

Αναστάσιος Ντελόπουλος

Περίληψη

Η ολιστική ερμηνεία εικόνων αποτελεί ένα διαχρονικά μελετημένο πρόβλημα στην υπολογιστική όραση που εκτείνεται πέρα από την ανίχνευση αντικειμένων. Οι οπτικές συσχετίσεις αποτελούν ένα μέσο προς την ερμηνεία αυτή, συλλαμβάνοντας ένα μεγάλο εύρος αλληλεπιδράσεων ανάμεσα σε ζεύγη αντικειμένων μιας εικόνας. Στην εργασία αυτή, μοντελοποιούμε τα αντικείμενα και τις συσχετίσεις τους με τη χρήση γράφων σκηνής, μία δομή γράφου που συνοψίζει το σημασιολογικό περιεχόμενο μιας εικόνας. Προτείνουμε ένα μοντέλο που εξάγει αυτές τις αναπαραστάσεις σκηνής, ενσωματώνοντας ένα σύστημα ανταλλαγής μηνυμάτων για τη διάδοση πληροφοριών συμφραζομένων ανάμεσα στα αντικείμενα και τις συσχετίσεις τους, το οποίο έχει σαν αποτέλεσμα την επαναληπτική βελτίωση των προβλέψεών του. Διεξάγουμε ένα πλήθος πειραμάτων σχετικά με τις αρχιτεκτονικές διάδοσης μηνυμάτων, συμπεριλαμβανομένης και μιας διαφοροποιημένης εκδοχής του Συνελικτικού Δικτύου Γράφων. Επιπρόσθετα, προτείνουμε ένα απλό πλην αποτελεσματικό δίκτυο διαγραφής συσχετίσεων που μαθαίνει να αναγνωρίζει και να αποκόπτει μη πιθανές συσχετίσεις. Αναφέρουμε τα αποτελέσματα στο πρόβλημα της εξαγωγής γράφου σκηνής, καθώς και άλλες βοηθητικές εργασίες αξιολόγησης, στο σύνολο δεδομένων Visual Genome, σημειώνοντας υψηλότερες αποδόσεις από ανάλογα συστήματα.

Abstract

Holistic image interpretation constitutes a long-studied problem in computer vision that extends well beyond object detection. Visual relationships comprise a means to that end, capturing a wide variety of interactions between pairs of objects in an image. In this work, we model objects and their relationships using scene graphs, a visually-grounded graphical structure of an image's semantic information. We propose an end-to-end model that generates such scene representations, by incorporating a message passing scheme that propagates contextual information between objects and their relationships to iteratively refine its predictions. We experiment on a variety of message passing propagation architectures, including a modified version of a Graph Convolutional Network. Furthermore, we propose a very simple yet effective relationship pruning network that learns to identify and dismiss unlikely relationships. We report our performance on scene graph generation and other auxiliary evaluation tasks using Visual Genome dataset, outperforming related methods.

Ευχαριστίες

Η εκπόνηση αυτής της εργασίας αποτέλεσε μία πρόκληση σε προσωπικό επίπεδο, καθώς απαιτήθηκε για πρώτη φορά ο συνδυασμός ενός μεγάλου εύρους των γνώσεων που αποκτήθηκαν κατά τη διάρκεια των σπουδών μου, αλλά και μια κοπιώδης προσπάθεια για την ολοκλήρωσή της. Η διεκπεραίωση της δε θα ήταν εφικτή δίχως τη συμβολή ορισμένων ατόμων, τα οποία και θα ήθελα να ευχαριστήσω θερμά.

Αρχικά, θα ήθελα να ευχαριστήσω τον αναπληρωτή καθηγητή κ. Αναστάσιο Ντελόπουλο για την ευκαιρία και την εμπιστοσύνη που μου έδειξε με την ανάθεση της εργασίας αυτής, αλλά και για την αγαστή συνεργασία μας στα πλαίσια της. Θα ήθελα επίσης να ευχαριστήσω τον μεταδιδακτορικό ερευνητή κ. Χρήστο Δίου για την πολύτιμη συμβολή του στην περάτωση της εργασίας αυτής μέσα από τις ιδέες και τις παρατηρήσεις του και την καθοδήγηση που μου προσέφερε.

Ιδιαίτερες ευχαριστίες οφείλω στους υπεύθυνους καθηγητές της ομάδας ρομποτικής P.A.N.D.O.R.A., τον αναπληρωτή καθηγητή κ. Λουκά Πέτρου, τον αναπληρωτή καθηγητή κ. Ανδρέα Συμεωνίδη και τον λέκτορα κ. Χαράλαμπο Δημούλα. Η ομάδα αυτή υπήρξε η ακαδημαϊκή μου μητέρα, καθώς μέσα από τη συμμετοχή της διδάχθηκα την έννοια της έρευνας και του πειράματος σε ένα πραγματικό περιβάλλον. Σαν φόρο τιμής, λοιπόν, θα ήθελα να ονομάσω άτυπα το προϊόν της εργασίας αυτής Πύρρα, κόρη της μυθολογικής Πανδώρας και μοναδική επιζήσασα του Κατακλυσμού μαζί με το σύζυγό της Δευκαλίωνα, με την ελπίδα το έργο αυτό να δημιουργήσει μια νέα αρχή για την ανθρωπότητα μέσω της τεχνητής νοημοσύνης.

Φυσικά, τίποτα από τα παραπάνω δε θα ήταν εφικτό χωρίς τη στήριξη και τη συμπαράσταση της οικογένειάς μου καθ' όλη τη διάρκεια των σπουδών μου, για τις οποίες και είμαι ευγνώμων.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου για τη βοήθειά τους και για όλα όσα μου έχουν προσφέρει στην εξέλιξή μου σαν άνθρωπος.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγικές Σημειώσεις

Η ολιστική ερμηνεία εικόνων αποτελεί το ιερό δισκοπότηρο της υπολογιστικής όρασης. Τα τελευταία χρόνια, η πρόοδος της μηχανικής μάθησης σε συνδυασμό με την ραγδαία τεχνολογική εξέλιξη του υλικού των υπολογιστών έχουν επιφέρει σημαντική ανάπτυξη στον επιστημονικό κλάδο της κατανόησης εικόνας. Αξιοσημείωτα παραδείγματα της ανάπτυξης αυτής αποτελούν η αναγνώριση εικόνων (image recognition) και η ανίχνευση αντικειμένων (object detection), τα οποία με την ενσωμάτωση της βαθιάς μηχανικής μάθησης και των συνελικτικών νευρωνικών δικτύων έχουν ανέλθει σε πολύ υψηλά επίπεδα συμπερασματολογίας, συγκρίσιμα ή και ανώτερα πολλές φορές από τα ανθρώπινα.

Καθοριστικό παράγοντα στις εξελίξεις αυτές έχει συντελέσει η συντονισμένη προσπάθεια της επιστημονική κοινότητας για τη δημιουργία ολοένα και πληρέστερων συνόλων δεδομένων, με αυξανόμενα πλήθη εικόνων και εξονυχιστικό σημασιολογικό σχολιασμό των περιεχομένων τους. Το γεγονός αυτό αποτελεί απαραίτητη προϋπόθεση για τη λειτουργία συστημάτων βαθιάς μηχανικής μάθησης, τα οποία απαιτούν μεγάλα πλήθη εικόνων για να εκπαιδευθούν, αποφεύγοντας παράλληλα την υπερπροσαρμογή. Χαρακτηριστικά παραδείγματα τέτοιων συνόλων δεδομένων αποτελούν το PASCAL Visual Object Classes (PASCAL VOC) [15], το Microsoft Common Objects in Context (MS COCO) [45] και το ImageNet [58].

Παρά την πρόοδο και την εξέλιξη, ωστόσο, τόσο των συνόλων δεδομένων όσο και των συστημάτων μηχανικής μάθησης, η κύρια εστία ενδιαφέροντος στον τομέα της υπολογιστικής όρασης παραμένει η ανίχνευση αντικειμένων και η κατάτμηση παραδειγμάτων (instance segmentation). Οι εργασίες αυτές, μολονότι υψηλού επιστημονικού ενδιαφέροντος, αποτυγχάνουν να ερμηνεύσουν ολιστικά το σημασιολογικό περιεχόμενο μίας εικόνας και εστιάζουν αντ' αυτού στην ανίχνευση αντικειμένων με παραλληλόγραμμα οριοθετικά πλαίσια (bounding boxes), ή, στην περίπτωση της κατάτμησης εικόνας, κατηγοριοποιώντας κάθε εικονοστοιχείο μιας εικόνας σε μία ή περισσότερες κλάσεις, δημιουργώντας, έτσι, τα περιγράμματα των αντικειμένων.

Μία εικόνα χαρακτηρίζεται, εντούτοις, από ένα πλήθος συσχετίσεων (relationships)

και αλληλεπιδράσεων ανάμεσα στις οντότητες που την αποτελούν, καθώς και από ένα σύνολο χαρακτηριστικών (attributes) που περιγράφουν μορφολογικά τα αντικείμενα, συνθέτοντας έναν οπτικό γράφο σκηνής (visual scene graph). Ένα σύστημα το οποίο αντιμετώπιζε ολιστικά τις εικόνες και ανίχνευε πέρα από τα αντικείμενα που τις αποτελούν και τα χαρακτηριστικά τους και τις συσχετίσεις μεταξύ τους θα μπορούσε να συλλογιστεί τον οπτικό κόσμο και να επιχειρηματολογήσει για αυτόν. Με τον τρόπο αυτό, καθίσταται εφικτή η επικοινωνία ανθρώπου-υπολογιστή σε ένα ανώτερο νοητικό επίπεδο, επιτρέποντας στον υπολογιστή να περιγράψει το περιεχόμενο μιας εικόνας σε φυσική γλώσσα, όπως επίσης και να απαντήσει σε ερωτήσεις σχετικά με αυτήν. Τέλος, η σημασιολογική ανάκτηση εικόνων με βάση προτασιακές αναζητήσεις ευνοείται σημαντικά από την αναπαράστασή τους μέσω γράφων σκηνής [33].

Σημαντική τροχοπέδη προς την κατεύθυνση αυτή αποτελούσε μέχρι και πριν λίγα χρόνια η έλλειψη συνόλων δεδομένων που εστιάζουν σε εργασίες υψηλότερου επιπέδου, όπως αυτές που αναφέρονται παραπάνω. Το σύνολο δεδομένων Visual Genome [39] αποτέλεσε την πρώτη ενορχηστρωμένη προσπάθεια ενσωμάτωσης γνωστικών εργασιών (cognitive tasks) στην κατανόηση εικόνας, ενώ ακολούθησαν, μεταξύ άλλων, τα σύνολα δεδομένων CLEVR [31] και πιο πρόσφατα το Google Open Images [38].

Η εξαγωγή γράφων σκηνής εικόνων κατέχει εξέχουσα σημασία στη σημασιολογική κατανόηση του οπτικού κόσμου και τον συλλογισμό επί αυτού, καθώς αποτελεί μια συμπαγή μαθηματική αναπαράστασή του, ενώ ταυτόχρονα εμπεριέχει πληροφορίες υψηλού εννοιολογικού επιπέδου. Ως εκ τούτου, μπορεί αφενός μεν να χρησιμοποιηθεί αυτούσια και ως κύριο πρόβλημα, όπως για παράδειγμα στο πρόβλημα της αναγνώρισης δράσης (action recognition), αφετέρου δε μπορεί να λειτουργήσει επικουρικά σε γνωστικά προβλήματα τα οποία περιλαμβάνουν την αλληλεπίδραση με τον άνθρωπο μέσω φυσικής γλώσσας, όπως είναι για παράδειγμα η προτασιακή περιγραφή εικόνων ή η απάντηση ερωτήσεων σχετικά με το περιεχόμενό τους.

1.2 Περιγραφή του Προβλήματος

Στην εργασία αυτή ερευνάται το πρόβλημα εξαγωγής γράφου σκηνής εικόνων, και πιο συγκεκριμένα το υποπρόβλημα της ανίχνευσης οπτικών συσχετίσεων (visual relationship detection). Σκοπός του προβλήματος είναι η ανίχνευση αντικειμένων, η χωροθέτηση, δηλαδή, αντικειμένων με τη χρήση οριοθετικών πλαισίων και η κατηγοριοποίησή τους, καθώς και η αναγνώριση των κατηγορημάτων (predicates) τα οποία συνδέουν ζεύγη αντικειμένων. Η ταξινόμηση τόσο των αντικειμένων όσο και των κατηγορημάτων γίνεται ανάμεσα σε ένα προδιαγεγραμμένο πλήθος κλάσεων, στις οποίες συμμετέχει και η κλάση υποβάθρου, η οποία υποδηλώνει την ανυπαρξία ενός αντικειμένου και την έλλειψη συσχέτισης ανάμεσα σε δύο αντικείμενα, αντίστοιχα.

Αναζητούνται, δηλαδή, όλες οι διατεταγμένες τριάδες της μορφής:

$$\langle αντικείμενο_1, κατηγόρημα, αντικείμενο_2 \rangle$$

που υπάρχουν σε μία εικόνα, σε συνδυασμό με την χωροθέτηση των αντικειμένων σε αυτή. Σημειώνεται στο σημείο αυτό ότι με τον όρο κατηγόρημα εννοείται η ρηματική φράση δίχως τη συμμετοχή αντικειμένων ή κατηγορουμένων, ορολογία η οποία τηρείται καθ' όλη την έκταση του παρόντος εγγράφου. Η παραπάνω διάταξη μπορεί να εκφραστεί και με την ακόλουθη γραμματική μορφή, η οποία στο εξής θα είναι και η προτιμότερη:

$$\langle υποκείμενο, κατηγόρημα, αντικείμενο \rangle$$

Όπως φανερώνεται από τα παραπάνω, στα πλαίσια της εργασίας αυτής ο όρος συσχέτιση χρησιμοποιείται αποκλειστικά για να περιγράψει την αλληλεπίδραση δύο διακριτών αντικειμένων και δε συμπεριλαμβάνει την προσδιοριστική σχέση

$$\langle υποκείμενο, είναι, κατηγορούμενο \rangle$$

η οποία εμπεριέχεται, για παράδειγμα, στο σύνολο δεδομένων Open Images και είναι πιο γνωστή στη βιβλιογραφία ως ανίχνευση χαρακτηριστικών (attribute detection).

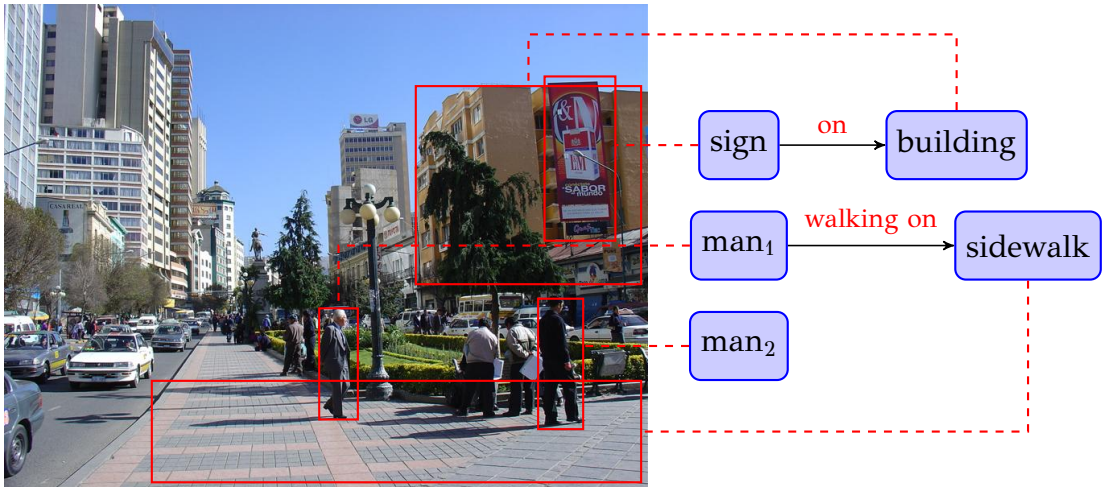
Φορμαλιστικά, ορίζουμε ως $\mathcal{I}$ την εικόνα εισόδου, $\boldsymbol{V}$ το σύνολο των περιοχών χωροθετημένων αντικειμένων στην εικόνα $\mathcal{I}$ που περιγράφει τα παραλληλόγραμμα οριοθετικά πλαίσια των αντικειμένων στο χώρο $\mathbb{R}^4$ και $\boldsymbol{E} \subseteq {}^{\boldsymbol{V}}P_2$ το σύνολο των συσχετίσεων μεταξύ των αντικειμένων, όπου ${}^{\boldsymbol{V}}P_2$ οι διατάξεις του συνόλου $\boldsymbol{V}$ ανά δύο. Επιπρόσθετα, ορίζουμε ως $\mathbf{O}$ και $\mathbf{P}$ τις κλάσεις στις οποίες ανήκουν τα αντικείμενα και τα κατηγορήματα, ανάμεσα στις προεπιλεγμένες κατηγορίες C_O και C_P, αντίστοιχα.

Σχηματίζουμε με τα στοιχεία αυτά τον κατευθυνόμενο γράφο $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathbf{O}, \mathbf{P})$, όπου $\boldsymbol{V}$ οι κορυφές και $\boldsymbol{E}$ οι ακμές του γράφου, σύμφωνα με τους παραπάνω ορισμούς των συνόλων αυτών. Στόχος μας είναι η κατά το δυνατόν επιτυχής πρόβλεψη των στοιχείων του γράφου, δεδομένης μιας εικόνας εισόδου, η μοντελοποίηση, δηλαδή, του συστήματος $P(\mathcal{G} \,|\, \mathcal{I})$.

Στο Σχήμα 1.1 απεικονίζεται διαγραμματικά ο παραγόμενος γράφος σε συνδυασμό με την φορμαλιστική μαθηματική του περιγραφή.

1.3 Διάρθρωση Εργασίας

Στα κεφάλαια που ακολουθούν αναλύεται η σχετική δουλειά στο χώρο της κατανόησης εικόνας, η μοντελοποίηση και η αρχιτεκτονική του συστήματος που υλοποιήθηκε, η πειραματική διαδικασία και η μεθοδολογία που ακολουθήθηκε, καθώς και τα αποτελέσματα των πειραμάτων και τα συμπεράσματα που εξάχθηκαν. Η εργασία αποτελείται από έξι κεφάλαια, συμπεριλαμβανομένου και του παρόντος εισαγωγικού κεφαλαίου.



(α′)

$$\mathcal{G} = \left\{ \begin{array}{l} \boldsymbol{V} = \Big\{ o_1 : box_1, o_2 : box_2, o_3 : box_3, o_4 : box_4, o_5 : box_5 \Big\} \\ \boldsymbol{E} = \Big\{ r_1 : \{o_1, o_2\}, r_2 : \{o_3, o_4\} \Big\} \\ \mathbf{O} = \Big\{ sign, building, man, sidewalk, man \Big\} \\ \mathbf{P} = \Big\{ on, walking\ on \Big\} \end{array} \right\}$$

(β′)

Σχήμα 1.1: Περιγραφή Προβλήματος

Στο Κεφάλαιο 2 αναλύεται η σχετική δουλειά στους χώρους της ανίχνευσης αντικειμένων και της εξαγωγής γράφου σκηνής και παρουσιάζεται το σύνολο δεδομένων Visual Genome.

Στο Κεφάλαιο 3 παρουσιάζεται η μοντελοποίηση και η γενική αρχιτεκτονική του συστήματος, καθώς και η λειτουργία των νευρωνικών δικτύων ανταλλαγής μηνυμάτων και των συνελικτικών δικτύων γράφων.

Στο Κεφάλαιο 4 εξηγείται η μεθοδολογία που ακολουθήθηκε για την προεπεξεργασία του συνόλου δεδομένων και η πειραματική διαδικασία της εκπαίδευσης και αξιολόγησης του συστήματος. Ακόμη, παρουσιάζονται οι παράμετροι που διαφοροποιήθηκαν κατά την εκτέλεση των πειραμάτων με σκοπό τη βελτιστοποίηση του συστήματος.

Στο Κεφάλαιο 5 αναλύονται τα αποτελέσματα αξιολόγησης που προκύπτουν από τη διεξαγωγή των πειραμάτων.

Τέλος, το Κεφάλαιο 6 εμβαθύνει στην ανάλυση των αποτελεσμάτων και προβαίνει στην εξαγωγή συμπερασμάτων για τη μεθοδολογία και τη μοντελοποίηση του συστήματος, ενώ γίνονται προτάσεις για μελλοντικές επεκτάσεις με σκοπό τη βελτίωση της απόδοσης του παρόντος συστήματος.

Κεφάλαιο 2

Βιβλιογραφική Επισκόπηση

2.1 Ταξινόμηση Εικόνας & Ανίχνευση Αντικειμένων

Η εξέλιξη των συνελικτικών νευρωνικών δικτύων και βαθιάς μηχανικής μάθησης, που ακολούθησε ως αποτέλεσμα της δημιουργίας πιο αποδοτικών τεχνικών εκπαίδευσης και της ανάπτυξης του υλικού των υπολογιστών και των καρτών γραφικών, έφερε την επανάσταση στο χώρο της υπολογιστικής όρασης, αλλά και της τεχνητής νοημοσύνης γενικότερα. Αρχής γενομένης από το AlexNet [40], ξεκίνησε μία σειρά ολοένα και πιο σύνθετων και εκλεπτυσμένων αρχιτεκτονικών και συστημάτων αναγνώρισης εικόνων. Χρονολογικά, ακολούθησαν τα δίκτυα ZF Net [73], VGGNet [62], GoogLeNet [65] και ResNet V1 [25] και ResNet V2 [27] τα οποία κατέκτησαν διακεκριμένες θέσεις στο διαγωνισμούς ILSVRC και ανέβασαν τον πήχη στην αναγνώριση εικόνων σε άνευ προηγουμένου επίπεδα.

Η εξέλιξη αυτή αποτέλεσε εφαλτήριο και για τον τομέα της ανίχνευσης αντικειμένων, στον οποίο το R-CNN (αρκτικόλεξο του όρου **R**egions with **C**onvolutional **N**eural **N**etwork features) [20] αποτέλεσε το πρώτο σύστημα το οποίο ενσωμάτωσε επιτυχώς συνελικτικά δίκτυα, εφαρμόζοντάς τα σε ένα πλήθος εξαγόμενων περιοχών ενδιαφέροντος του αλγορίθμου Selective Search [70].

Το R-CNN αποτέλεσε το πρώτο από μία μεγάλη οικογένεια συστημάτων ανίχνευσης αντικειμένων. Τη σκυτάλη έλαβαν τα συστήματα Fast R-CNN [19] και Faster R-CNN [57], τα οποία ανέβασαν τόσο την ποιότητα όσο και την ταχύτητα της ανίχνευσης. Το Faster R-CNN εισήγαγε ένα Δίκτυο Πρότασης Περιοχών (Region Proposal Network) για την εξαγωγή των περιοχών ενδιαφέροντος και έγινε με τον τρόπο αυτό το πρώτο από άκρη σε άκρη εκπαιδεύσιμο (trainable) σύστημα ανίχνευσης αντικειμένων, αποτελούμενο εξ' ολοκλήρου από συνελικτικά και πλήρως συνδεδεμένα επίπεδα.

Πέρα από την οικογένεια των δικτύων R-CNN, η οποία εστιάζει κατά κύριο λόγο στην απόδοση του συστήματος, εξέχουσα σημασία κατέχει και η οικογένεια YOLO (You Only Look Once), με τις εκδόσεις YOLO [55] και YOLO9000 [56], καθώς και το σύστημα SSD (Single Shot MultiBox Detector) [46]. Τα μοντέλα αυτά ακολουθούν μια διαφορετική προσέγγιση στην ανίχνευση αντικειμένων και ανήκουν στην κατηγορία των single-shot ανιχνευτών· η εξαγωγή, δηλαδή, των περιοχών ενδιαφέροντος δε γίνεται σαν ένα πα-

ράλληλο υποδίκτυο, όπως στην περίπτωση του Faster R-CNN, αλλά αποτελεί τμήμα του βασικού δικτύου. Τα δίκτυα αυτά απολαμβάνουν πολύ υψηλές ταχύτητες συμπερασματολογίας και επιτυγχάνουν ανίχνευση αντικειμένων σε πραγματικό χρόνο, με αρκετές δεκάδες καρέ ανά λεπτό (frames per second), ενώ δε στερούνται υψηλών επιδόσεων, αφού η απόδοσή τους είναι εφάμιλλη με αυτή των δικτύων R-CNN.

Στην κατεύθυνση της πλήρους κατανόησης εικόνας, φυσική επέκταση της ανίχνευσης αντικειμένων αποτελεί η κατάτμηση παραδειγμάτων, σκοπός της οποίας είναι η ανίχνευση και χωροθέτηση των αντικειμένων με οριοθετικά πλαίσια σε συνδυασμό με την εύρεση των εν γένει μη κυρτών και πολλαπλών περιγραμμάτων τους. Αξιοσημείωτα παραδείγματα στον τομέα αυτό αποτελούν τα [11, 12, 41] και πιο πρόσφατα το Mask R-CNN [24] το οποίο έχει σημαντικά καλύτερες επιδόσεις και αποτελεί φυσική επέκταση του Faster R-CNN στον τομέα της κατάτμησης παραδειγμάτων.

2.2 Εξαγωγή Γράφου Σκηνής & Ανίχνευση Συσχετίσεων

Η ανίχνευση αντικειμένων, αν και αποτελεί ένα κλασικό πρόβλημα υψηλού ενδιαφέροντος στον τομέα της υπολογιστικής όρασης, στερείται της ευρύτερης κατανόησης του περιεχομένου μιας εικόνας, αφού δεν ενσωματώνει τις αλληλεπιδράσεις μεταξύ αυτών και τους προσδιορισμούς που τα χαρακτηρίζουν. Τα τελευταία χρόνια, με την έλευση νέων συνόλων δεδομένων που ενσωματώνουν τα στοιχεία αυτά, έχει αναζωπυρώσει το ενδιαφέρον για την επίλυση γνωστικών προβλημάτων, όπως η εξαγωγή γράφου σκηνής, η προτασιακή περιγραφή εικόνων και η απάντηση ερωτήσεων σχετικά με το περιεχόμενο των εικόνων.

Το [47] αποτελεί μία από τις πρώτες προσπάθειες στο χώρο της ανίχνευσης συσχετίσεων, συνδυάζοντας συστήματα ανίχνευσης αντικειμένων και συνελικτικά νευρωνικά δίκτυα με γλωσσικά τμήματα τα οποία μοντελοποιούν την πιθανοφάνεια των συσχετίσεων. Τα επόμενα χρόνια προτάθηκαν αρκετές προσεγγίσεις για την ανίχνευση συσχετίσεων [52, 10, 44, 42, 71], με ολοένα και πιο εκλεπτυσμένους τρόπους επίλυσης του προβλήματος.

2.2.1 Visual Relationship Detection with Language Priors

Το σύστημα που υλοποιείται στο [47] αποτελεί μία από τις πρώτες επιτυχημένες προσπάθειες χρήσης συνελικτικών δικτύων στο χώρο της ανίχνευσης συσχετίσεων. Στο έργο αυτό, το πρόβλημα της ανίχνευσης συσχετίσεων προσεγγίζεται τόσο από την οπτική του πλευρά όσο και από τη γλωσσολογική.

Στο οπτικό μοντέλο, ένα δίκτυο R-CNN [20] εξάγει τις περιοχές ενδιαφέροντος, ενώ δύο συνελικτικά δίκτυα VGG-16 [62] κατηγοριοποιούν τα αντικείμενα και τις συσχετίσεις. Οι συσχετίσεις που εξετάζονται αποτελούν όλους τους συνδυασμούς αντικειμένων, ενώ οι περιοχές των συσχετίσεων ορίζονται μέσω της ένωσης των οριοθετικών πλαισίων των υποκειμένων και των αντικειμένων τους.

Το γλωσσικό μοντέλο χρησιμοποιεί διανύσματα λέξεων (word vectors) [48] για να προβάλλει τα αντικείμενα σε ένα διανυσματικό χώρο λέξεων. Στη συνέχεια προβάλλει τις συσχετίσεις σε ένα νέο διανυσματικό χώρο συσχετίσεων που αντιπροσωπεύει τον τρόπο με τον οποίο αλληλεπιδρούν τα αντικείμενα. Το μοντέλο αυτό εκπαιδεύεται ώστε να υποδεικνύει την πιθανοφάνεια των οπτικών συσχετίσεων, ενισχύοντας τη συνολική απόδοση του συστήματος.

2.2.2 Scene Graph Generation by Iterative Message Passing

Η προσέγγισή μας στο πρόβλημα της εξαγωγής γράφου σκηνής παρουσιάζει αρκετά κοινά με το [71]. Στο έργο αυτό, το πρόβλημα της εξαγωγής γράφου σκηνής αντιμετωπίζεται μέσω ενός συστήματος επαναλαμβανόμενης ανταλλαγής πληροφοριών ανάμεσα στα αντικείμενα και τις συσχετίσεις. Το σύστημα αυτό ορίζεται από τις ακόλουθες εξισώσεις:

$$m_i^t = \sum_{j:i\to j} \sigma(\mathbf{v}_1^T[h_i^t, h_{i\to j}^t])h_{i\to j}^t + \sum_{j:j\to i} \sigma(\mathbf{v}_2^T[h_i^t, h_{j\to i}^t])h_{j\to i}^t \tag{2.1}$$

$$m_{i\to j}^t = \sigma(\mathbf{w}_1^T[h_i^t, h_{i\to j}^t])h_i^t + \sigma(\mathbf{w}_2^T[h_j^t, h_{i\to j}^t])h_j^t \tag{2.2}$$

$$h_i^{t+1} = \mathrm{GRU}\big(m_i^t\big), \tag{2.3}$$

$$h_{i\to j}^{t+1} = \mathrm{GRU}\big(m_{i\to j}^t\big) \tag{2.4}$$

όπου h_i^t και $h_{i\to j}^t$ τα διανύσματα που κωδικοποιούν το αντικείμενο i και τη συσχέτιση $i \to j$, αντίστοιχα, $\sigma(\cdot)$ η συνάρτηση σιγμοειδούς ενεργοποίησης (sigmoid function), $[\cdot,\cdot]$ η συνένωση διανυσμάτων (vector concatenation) και $\mathrm{GRU}(\cdot)$ η επαναλαμβανόμενη μονάδα ελεγχόμενης πρόσβασης (Gated Recurrent Unit) [8].

Για την καλύτερη κατανόηση της λειτουργίας του συστήματος αυτού, ο αναγνώστης παραπέμπεται στην Ενότητα 3.3, η οποία θεμελιώνει τον ορισμό και τη λειτουργία ενός δικτύου ανταλλαγής μηνυμάτων.

2.2.3 Γράφοι Σκηνής σε Γνωστικά Προβλήματα Εικόνων

Η αναπαράσταση του περιεχομένου μιας εικόνας με τη μορφή ενός γράφου σκηνής αποτελεί μία συμπαγή αναπαράσταση του περιεχομένου της, αποφεύγοντας την αβεβαιότητα που ενυπάρχει στην αναπαράσταση κειμένου. Για το λόγο αυτό, οι γράφοι σκηνής έχουν χρησιμοποιηθεί επικουρικά σε αρκετά γνωστικά προβλήματα, όπως η απάντηση ερωτήσεων εικόνας [66], στην ανάκτηση εικόνας [33] και πιο πρόσφατα στη δημιουργία εικόνων που αποτυπώνουν το περιεχόμενο του γράφου [30].

2.3 Σύνολο Δεδομένων Visual Genome

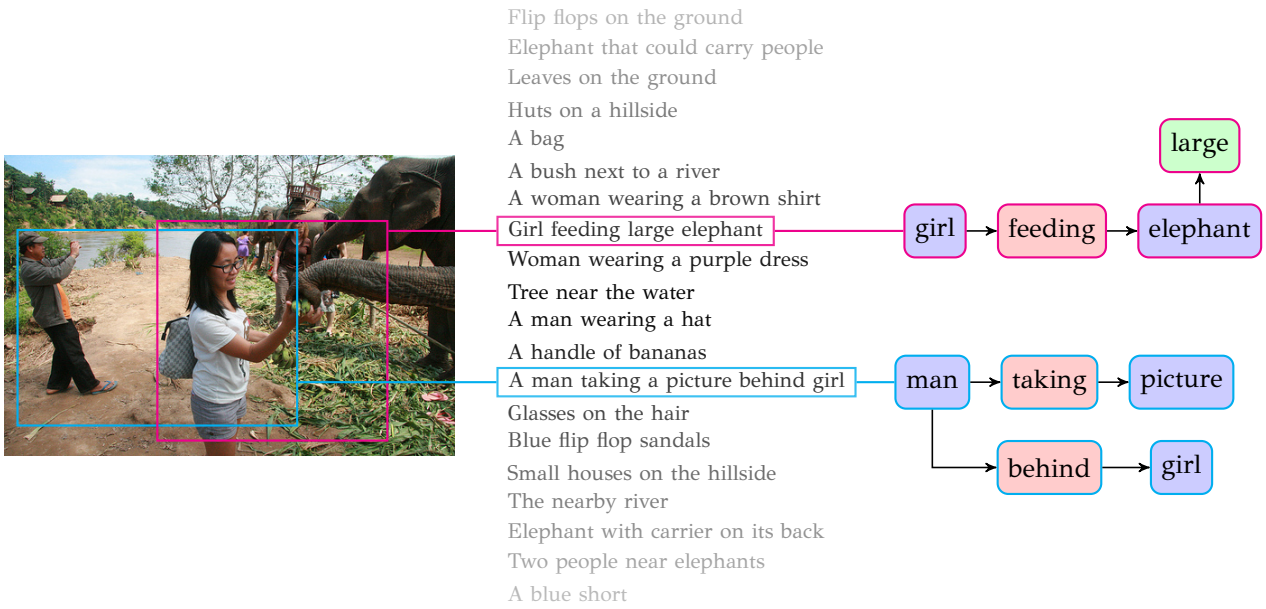
Εξέχουσα σημασία στην εξέλιξη του τομέα της εξαγωγής γράφου σκηνής είχε η δημιουργία του συνόλου δεδομένων Visual Genome [39]. Το Visual Genome δημιουργήθηκε με σκοπό να ενισχύσει την πρόοδο της τεχνητής νοημοσύνης και της υπολογιστικής όρασης πέρα από προβλήματα αντίληψης, όπως η αναγνώριση αντικειμένων, σε γνωστικά προβλήματα και προβλήματα νόησης. Η ικανότητα κατανόησης του περιεχομένου μιας εικόνας είναι απαραίτητη σε πολλά προβλήματα που απαιτούν συλλογισμό του οπτικού κόσμου, όπως είναι η αναγνώριση συσχετίσεων, η απάντηση ερωτήσεων σχετικά με το περιεχόμενο μιας εικόνας και η ανάκτηση εικόνων με βάση λεκτικές περιγραφές.

Το Visual Genome δημιουργήθηκε μέσα από τη συνένωση των συνόλων δεδομένων YFCC100M [68] και MS-COCO [45] και περιέχει περισσότερες από 100,000 εικόνες για τις οποίες έχουν συλλεχθεί εκτενείς αναφορές αντικειμένων καθώς και των ιδιοτήτων τους αλλά και των συσχετίσεων μεταξύ τους μέσω της τεχνικής του crowdsourcing. Όλες οι αναφορές είναι γραμμένες σε ελεύθερο κείμενο χωρίς περιορισμούς ενός προκαθορισμένου λεξιλογίου και κανονικοποιούνται στα σημασιολογικά τους συνώνυμα σύμφωνα με το WordNet [49]. Επιπρόσθετα, κάθε εικόνα χαρακτηρίζεται από προτασιακές περιγραφές και γράφους περιοχών ενδιαφέροντος, ζεύγη ερωτήσεων-απαντήσεων καθώς και έναν ολιστικό γράφο σκηνής.

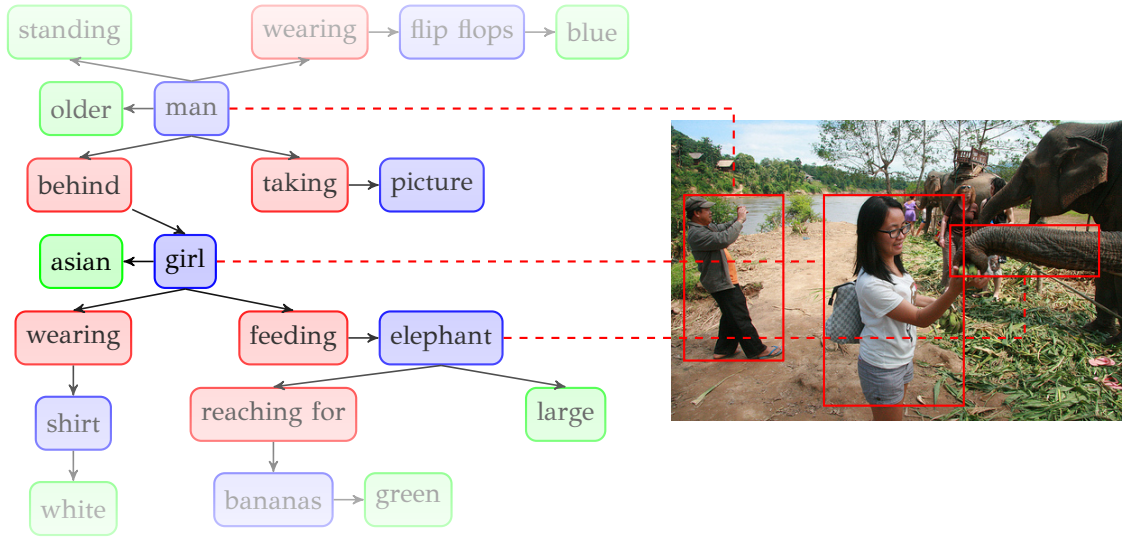
Το σύνολο δεδομένων περιέχει περισσότερες από 75,000 κατηγορίες αντικειμένων και περισσότερες από 40,000 κατηγορίες συσχετίσεων και ιδιοτήτων. Κάθε εικόνα χαρακτηρίζεται κατά μέσο όρο από 35 αντικείμενα, 26 ιδιότητες και 21 διμερείς συσχετίσεις μεταξύ των αντικειμένων. Στη συνέχεια παρατίθενται ενδεικτικά μερικές από τις εικόνες του Visual Genome σε συνδυασμό με τα μεταδεδομένα που τις χαρακτηρίζουν.

Στο Σχήμα 2.1 παρουσιάζεται ένα παράδειγμα του υψηλού επιπέδου κατανόησης εικόνων το οποίο θέτει το Visual Genome. Κάθε εικόνα περιγράφεται από ένα πλήθος αντικειμένων με οριοθετημένα πλαίσια, τις ιδιότητες των αντικειμένων και τις συσχετίσεις μεταξύ τους. Τα παραπάνω στοιχεία συνδυάζονται για να συνθέσουν γράφους περιοχών ενδιαφέροντος, σε μια δομή δεδομένων υψηλού νοητικού επιπέδου τόσο για τον άνθρωπο όσο και για τον υπολογιστή, αλλά και σε προτασιακές περιγραφές γραμμένες σε φυσική γλώσσα, σε μια μορφή απόλυτα κατανοητή στον άνθρωπο.

Στο Σχήμα 2.2 απεικονίζονται όλα τα χαρακτηριστικά με βάση τα οποία αναπαρίστανται οι εικόνες στο σύνολο δεδομένων Visual Genome. Κάθε εικόνα περιέχει τοπικούς περιγραφείς, όπως είναι οι προτασιακές περιγραφές και οι γράφοι περιοχών ενδιαφέροντος. Οι «γράφοι» αυτοί συνθέτουν έναν ολιστικό γράφο σκηνής για την εικόνα που αποτελείται από όλα τα αντικείμενα της εικόνας μαζί με το οριοθετικό πλαίσιο τους, τις ιδιότητές τους και τις συσχετίσεις μεταξύ τους. Τέλος, κάθε εικόνα χαρακτηρίζεται από ζεύγη ερωτήσεων-απαντήσεων τα οποία είναι είτε ελεύθερου τύπου, είτε συσχετίζονται με κάποια περιοχή ενδιαφέροντος της εικόνας.

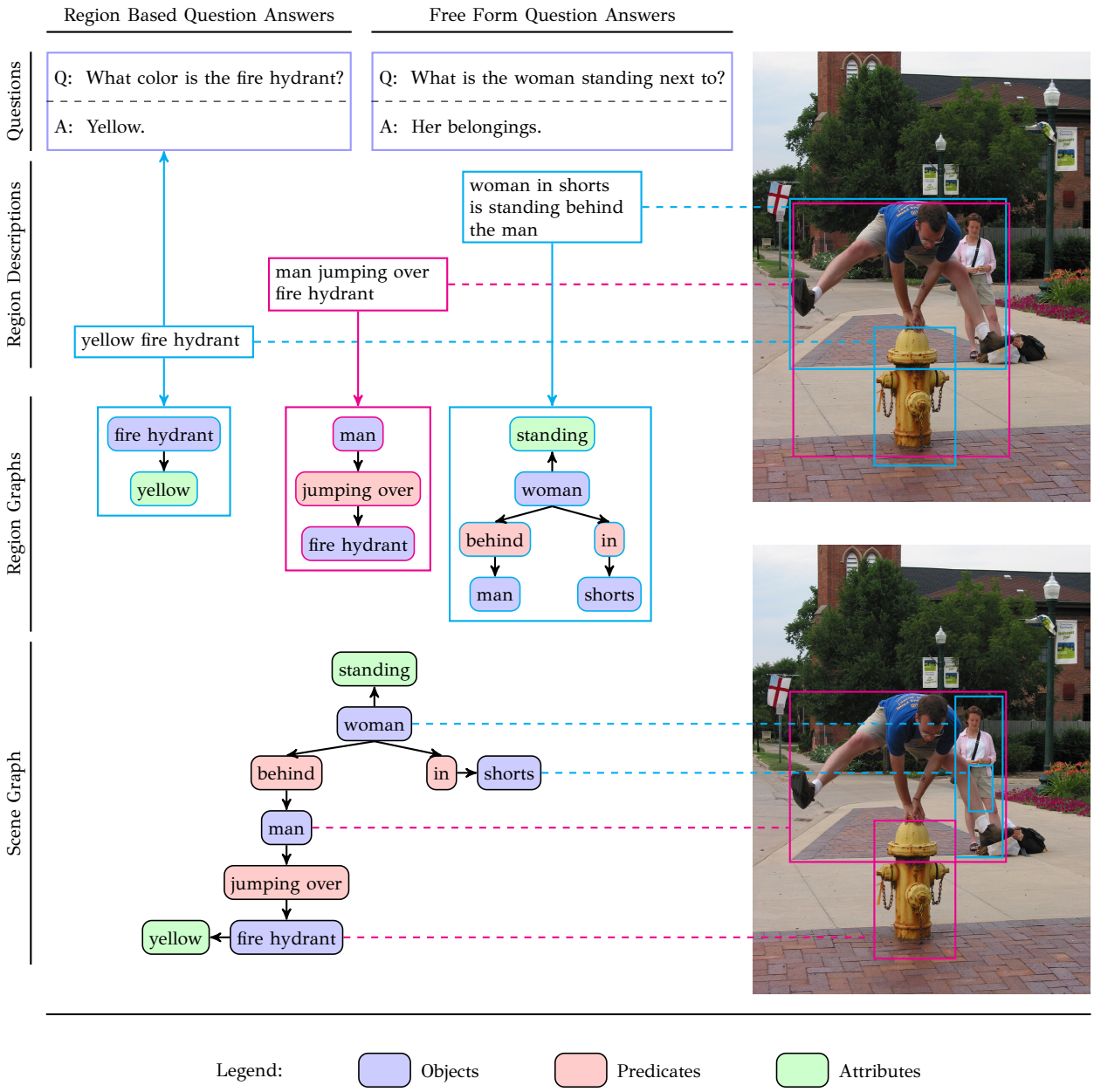


(α′)



(β′)

Σχήμα 2.1: Παραδείγματα (α′) προτασιακών περιγραφών και γράφων περιοχών εικόνων και (β′) ολιστικού γράφου σκηνής εικόνων. Προσαρμόστηκε από το [39]. *Απεικονίζεται καλύτερα έγχρωμο*



Σχήμα 2.2: Αναπαράσταση εικόνων στο σύνολο δεδομένων Visual Genome. Προσαρμόστηκε από το [39]. *Απεικονίζεται καλύτερα έγχρωμο*

Κεφάλαιο 3

Μοντελοποίηση Συστήματος

3.1 Μοντελοποίηση Συστήματος

Όπως αναφέρεται και στην Ενότητα 1.2, σκοπός του προβλήματος είναι η εξαγωγή του γράφου σκηνής εικόνων $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathbf{O}, \mathbf{P})$ δεδομένης εικόνας εισόδου $\mathcal{I}$, η μοντελοποίηση, δηλαδή, του συστήματος $P(\mathcal{G} \,|\, \mathcal{I})$.

Στην εργασία αυτή παραγοντοποιούμε το πρόβλημα της εξαγωγής γράφου σκηνής στα υποπροβλήματα της εξαγωγής περιοχών αντικειμένων, της εξαγωγής συσχετίσεων και της ταξινόμησης γράφου. Φορμαλιστικά, το πρόβλημα διατυπώνεται ως:

$$P(\mathcal{G} \,|\, \mathcal{I}) = \overbrace{P(\boldsymbol{V} \,|\, \mathcal{I})}^{\text{Εξαγωγή Περιοχών Αντικειμένων}} \underbrace{P(\boldsymbol{E} \,|\, \boldsymbol{V}, \mathcal{I})}_{\text{Εξαγωγή Συσχετίσεων}} \overbrace{P(\mathbf{O}, \mathbf{P} \,|\, \boldsymbol{V}, \boldsymbol{E}, \mathcal{I})}^{\text{Ταξινόμηση Γράφου}} \tag{3.1}$$

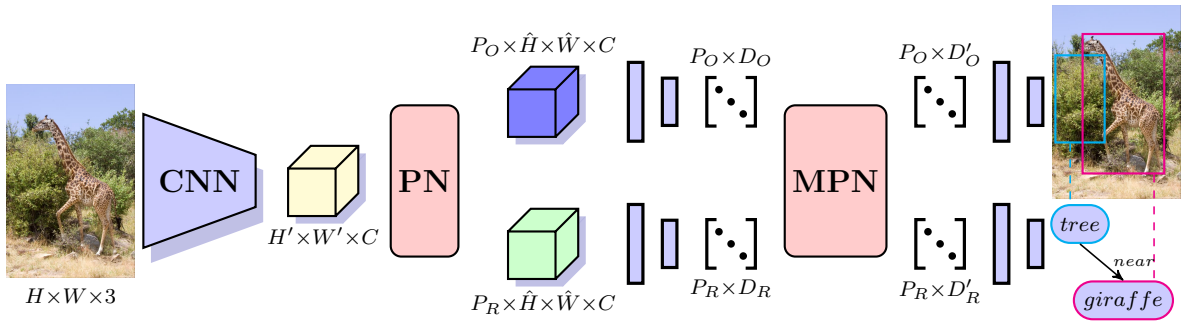
Με τρόπο αυτό διαχωρίζεται η κατασκευή του γράφου από την ταξινόμηση των ακμών και των κορυφών του. Στα πλαίσια της εργασίας αυτής δίνουμε έμφαση και εστιάζουμε στο υποπρόβλημα της ταξινόμησης γράφου. Το υποσύστημα της εξαγωγής περιοχών αντικειμένων $P(\boldsymbol{V} \,|\, \mathcal{I})$ μοντελοποιείται με τη χρήση ενός δικτύου πρότασης περιοχών ενός υπάρχοντος συστήματος ανίχνευσης αντικειμένων. Το υποσύστημα της εξαγωγής συσχετίσεων $P(\boldsymbol{E} \,|\, \boldsymbol{V}, \mathcal{I})$ μοντελοποιείται στα περισσότερα πειράματα λαμβάνοντας υπ' όψιν όλες τις πιθανές συνδέσεις και συσχετίσεις μεταξύ των αντικειμένων, ή δειγματοληπτώντας ένα τυχαίο τμήμα αυτών. Σε κάποια πειράματα, ωστόσο, δοκιμάστηκε η χρήση ενός δικτύου διαγραφής (pruning) συσχετίσεων με σκοπό τον έλεγχο της σημασίας του υποσυστήματος αυτού στη συνολική απόδοση του συστήματος. Τέλος, το υποσύστημα της ταξινόμησης γράφου $P(\mathbf{O}, \mathbf{P} \,|\, \boldsymbol{V}, \boldsymbol{E}, \mathcal{I})$ μοντελοποιείται ως ένα σύστημα επαναληπτικής ανταλλαγής πληροφοριών και ραφιναρίσματος αυτών ανάμεσα στα συνδεόμενα τμήματα του γράφου, ακολουθούμενο από συστήματα ταξινόμησης των αντικειμένων και των συσχετίσεων.

3.2 Αρχιτεκτονική Συστήματος

Σύμφωνα με τη μοντελοποίηση που προηγήθηκε, διαμορφώνεται η αρχιτεκτονική του συστήματος, η οποία παρουσιάζεται διαγραμματικά στο Σχήμα 3.1.

Η εικόνα εισόδου διαστάσεων $H \times W \times 3$ διέρχεται μέσα από ένα συνελικτικό δίκτυο από το οποίο παράγεται ο συνελικτικός χάρτης χαρακτηριστικών της (feature map), τανυστής διαστάσεων $H' \times W' \times C$. Ένα δίκτυο πρότασης περιοχών λαμβάνει ως είσοδο τον τανυστή αυτό και εξάγει P_O περιοχές αντικειμένων, οι οποίες με τη βοήθεια ενός μηχανισμού συλλογής χαρακτηριστικών περιοχών (region pooling) δημιουργούν τον τανυστή αντικειμένων $P_O \times \hat{H} \times \hat{W} \times C$. Αντίστοιχα, για κάθε διατεταγμένο συνδυασμό περιοχών αντικειμένων, εκτός από τις συνδέσεις βρόχων, υπολογίζεται η ένωση των οριοθετικών πλαισίων τους από τις οποίες δημιουργείται αντίστοιχα ο τανυστής συσχετίσεων $P_R \times \hat{H} \times \hat{W} \times C$.

Οι τανυστές αντικειμένων και συσχετίσεων διέρχονται από ένα πλήθος πλήρως συνδεδεμένων επιπέδων και μετασχηματίζονται στους πίνακες $P_O \times D_O$ και $P_R \times D_R$, αντίστοιχα. Στη συνέχεια, εισέρχονται αμφότεροι σε ένα δίκτυο ανταλλαγής μηνυμάτων στο οποίο γίνεται ελεγχόμενη ανταλλαγή πληροφοριών σύμφωνα με τις συνδέσεις των στοιχείων του γράφου. Στην έξοδο του δικτύου έχουμε τους πίνακες $P_O \times D'_O$ και $P_R \times D'_R$ που κωδικοποιούν τη συσσωρευμένη πληροφορία τους. Κάθε κωδικοποιημένο αντικείμενο και συσχέτιση διέρχεται μέσα από ένα πλήρως συνδεδεμένο επίπεδο από το οποίο και παράγεται το τελικό του σκορ και η πρόβλεψη του δικτύου.



Σχήμα 3.1: Αρχιτεκτονική Συστήματος. **CNN**: Συνελικτικό Νευρωνικό Δίκτυο, **PN**: Δίκτυο Προτάσεων, **MPN**: Δίκτυο Ανταλλαγής Μηνυμάτων

Η χρησιμότητα του δικτύου ανταλλαγής μηνυμάτων έγκειται στην αλληλεπίδραση των συνδεόμενων στοιχείων του γράφου και στην ανταλλαγή πληροφοριών ανάμεσα στις κορυφές και τις ακμές του γράφου. Η λειτουργία αυτή έρχεται σε αντιδιαστολή με ένα σύστημα που ταξινομεί τα στοιχεία του γράφου απομονωμένα από τις γειτονικές του συνδέσεις. Η χρήση του δικτύου ανταλλαγής μηνυμάτων έχει σαν αποτέλεσμα την πιο ολοκληρωμένη κατηγοριοποίηση των στοιχείων του γράφου, αφού κάθε στοιχείο καθορίζεται και από τα συμφραζόμενά του, μέσα από ένα επαναληπτικό ραφινάρισμα της συνεισφοράς τους.

3.3 Νευρωνικά Δίκτυα Ανταλλαγής Μηνυμάτων

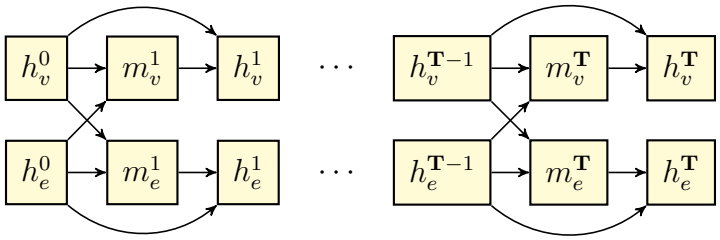
Τα νευρωνικά δίκτυα ανταλλαγής μηνυμάτων [18] αποτελούν μία προσπάθεια ομαδοποίησης κάτω από ένα κοινό μαθηματικό πλαίσιο μιας οικογένειας νευρωνικών δικτύων που εστιάζουν στην ελεγχόμενη ανταλλαγή πληροφοριών ανάμεσα σε ένα πλήθος συσχετιζόμενων κόμβων. Η ομαδοποίηση αυτή εστιάζει στην επίλυση προβλημάτων κβαντικής χημείας, μπορεί, ωστόσο, να γενικευτεί και για την επίλυση προβλημάτων διαφορετικών επιστημονικών περιοχών, όπως στην περίπτωσή μας στον κλάδο της υπολογιστικής όρασης και πιο συγκεκριμένα στο πρόβλημα της εξαγωγής γράφου σκηνής εικόνων.

Αφορμώμενοι, λοιπόν, από την εργασία αυτή, επαναορίζουμε τα νευρωνικά δίκτυα ανταλλαγής μηνυμάτων για το πρόβλημα της εξαγωγής γράφου σκηνής, σύμφωνα με τις ακόλουθες εξισώσεις ανταλλαγής μηνυμάτων:

$$
\begin{aligned}
m_{v_i}^{t+1} &= \sum_{v_j \in N_{out}(v_i)} M_{\mathcal{V}out}^t\big(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t\big) + \sum_{v_j \in N_{in}(v_i)} M_{\mathcal{V}in}^t\big(h_{v_i}^t, h_{v_j}^t, h_{e_{ji}}^t\big), \\
h_{v_i}^{t+1} &= U_{\mathcal{V}}^t\big(h_{v_i}^t, m_{v_i}^{t+1}\big), \\
m_{e_{ij}}^{t+1} &= M_{\mathcal{E}}^t\big(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t\big), \\
h_{e_{ij}}^{t+1} &= U_{\mathcal{E}}^t\big(h_{e_{ij}}^t, m_{e_{ij}}^{t+1}\big)
\end{aligned}
\tag{3.2}
$$

όπου $h_{v_i}^t$ το διάνυσμα που κωδικοποιεί την πληροφορία της κορυφής (αντικειμένου) v_i τη χρονική στιγμή t, ενώ $h_{e_{ij}}^t$ το διάνυσμα που κωδικοποιεί την πληροφορία της ακμής (συσχέτισης) ανάμεσα στις κορυφές v_i και v_j τη χρονική στιγμή t.

Τη χρονική στιγμή $t = 0$, τα διανύσματα των κορυφών και των ακμών $h_{v_i}^0$ και $h_{e_{ij}}^0$ αντίστοιχα, αποτελούν την έξοδο των πλήρως συνδεδεμένων επιπέδων που προηγούνται του δικτύου ανταλλαγής μηνυμάτων. Η ανταλλαγή μηνυμάτων γίνεται για $\mathbf{T}$ βήματα. Η διάδοση της πληροφορίας σε ένα δίκτυο ανταλλαγής μηνυμάτων αναπαρίσταται εποπτικά στο Σχήμα 3.2.



Σχήμα 3.2: Νευρωνικό Δίκτυο Ανταλλαγής Μηνυμάτων

Η δημιουργία ενός μοντέλου της οικογένειας νευρωνικών δικτύων ανταλλαγής μηνυμάτων συνίσταται στον προσδιορισμό των συναρτήσεων $M_{\mathcal{V}out}^t$, $M_{\mathcal{V}in}^t$ και $M_{\mathcal{E}}^t$ που ορίζουν τον τρόπο με τον οποίο το μοντέλο συλλέγει πληροφορίες από τις γειτονικές του περιοχές και των συναρτήσεων $U_{\mathcal{V}}^t$ και $U_{\mathcal{E}}^t$ που ορίζουν τον τρόπο με τον οποίο οι κορυφές και οι ακμές ανανεώνουν την κωδικοποιημένη πληροφορία τους. Οι συναρτήσεις αυτές είναι

στη γενική τους μορφή παραμετροποιήσιμες με παραμέτρους που μαθαίνονται κατά τη διάρκεια της εκπαίδευσης και διαφορίσιμες.

Στις υποενότητες που ακολουθούν παρουσιάζονται οι μοντελοποιήσεις που επιλέχθηκαν στα πλαίσια της εργασίας αυτής. Σημειώνεται ότι σε αρκετές περιπτώσεις οι συναρτήσεις $M^t_{\mathcal{V}out}$ και $M^t_{\mathcal{V}in}$ συμπυκνώνονται στην $M^t_{\mathcal{V}}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ij}}, h^t_{e_{ji}})$, ενώ το διάνυσμα $m^{t+1}_{v_i}$ ορίζεται ως το άθροισμα των μηνυμάτων όλων των γειτονικών περιοχών· τα μοντέλα αυτά αντιμετωπίζουν με τον ίδιο τρόπο τις εισερχόμενες και εξερχόμενες συσχετίσεις.

3.3.1 Molecular Graph Convolutions

Η μοντελοποίηση αυτή είναι βασισμένη στο [35]. Οι συναρτήσεις μηνυμάτων και ανανέωσης που την καθορίζουν είναι οι εξής:

$$\begin{aligned} M_{\mathcal{V}out}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ij}}) &= h^t_{e_{ij}}, \\ U_{\mathcal{V}}(h^t_{v_i}, m^{t+1}_{v_i}) &= \rho\Big(\mathbf{W}_1\Big[\rho\big(\mathbf{W}_0 h^t_{v_i}\big), m^{t+1}_{v_i}\Big]\Big), \\ M_{\mathcal{E}}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ij}}) &= \rho\Big(\mathbf{W}_2\Big[h^t_{v_i}, h^t_{v_j}\Big]\Big), \\ U_{\mathcal{E}}(h^t_{e_{ij}}, m^{t+1}_{e_{ij}}) &= \rho\Big(\mathbf{W}_4\Big[\rho\big(\mathbf{W}_3 h^t_{e_{ij}}\big), m^{t+1}_{e_{ij}}\Big]\Big) \end{aligned} \tag{3.3}$$

όπου $\rho(\cdot)$ η συνάρτηση ενεργοποίησης γραμμικού ανορθωτή (ReLU) και $[\cdot,\cdot]$ η συνένωση διανυσμάτων (vector concatenation).

Μία παραλλαγή της παραπάνω μοντελοποίησης που επιχειρήθηκε, με σκοπό την πληρέστερη συλλογή πληροφοριών των αντικειμένων είναι η εξής:

$$\begin{aligned} M_{\mathcal{V}out}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ij}}) &= \rho\Big(\mathbf{W}_5\Big[h^t_{v_j}, h^t_{e_{ij}}\Big]\Big), \\ M_{\mathcal{V}in}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ji}}) &= \rho\Big(\mathbf{W}_6\Big[h^t_{v_i}, h^t_{e_{ji}}\Big]\Big), \\ U_{\mathcal{V}}(h^t_{v_i}, m^{t+1}_{v_i}) &= \rho\Big(\mathbf{W}_1\Big[\rho\big(\mathbf{W}_0 h^t_{v_i}\big), m^{t+1}_{v_i}\Big]\Big), \\ M_{\mathcal{E}}(h^t_{v_i}, h^t_{v_j}, h^t_{e_{ij}}) &= \rho\Big(\mathbf{W}_2\Big[h^t_{v_i}, h^t_{v_j}\Big]\Big), \\ U_{\mathcal{E}}(h^t_{e_{ij}}, m^{t+1}_{e_{ij}}) &= \rho\Big(\mathbf{W}_4\Big[\rho\big(\mathbf{W}_3 h^t_{e_{ij}}\big), m^{t+1}_{e_{ij}}\Big]\Big) \end{aligned} \tag{3.4}$$

3.3.2 Gated Graph Neural Network

Η μοντελοποίηση αυτή βασίστηκε στο [43]. Οι συναρτήσεις που διέπουν τη λειτουργία της είναι οι εξής:

$$
\begin{aligned}
M_{\mathcal{V}}(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t, h_{e_{ji}}^t) &= \rho\Bigg(\mathbf{W}_7\Bigg[\rho\Big(\mathbf{W}_6 h_{v_i}^t\Big), \rho\Bigg(\mathbf{W}_5\Bigg[\rho\Big(\mathbf{W}_3\Big[h_{v_j}^t, h_{e_{ij}}^t\Big]\Big), \rho\Big(\mathbf{W}_4\Big[h_{v_i}^t, h_{e_{ji}}^t\Big]\Big)\Bigg]\Bigg)\Bigg]\Bigg),\\
U_{\mathcal{V}}\big(h_{v_i}^t, m_{v_i}^{t+1}\big) &= \mathrm{GRU}\big(m_{v_i}^{t+1}\big),\\
M_{\mathcal{E}}(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t) &= \rho\Bigg(\mathbf{W}_2\Bigg[\rho\Big(\mathbf{W}_0 h_{e_{ij}}^t\Big), \rho\Big(\mathbf{W}_1\Big[h_{v_i}^t, h_{v_j}^t\Big]\Big)\Bigg]\Bigg),\\
U_{\mathcal{E}}\big(h_{e_{ij}}^t, m_{e_{ij}}^{t+1}\big) &= \mathrm{GRU}\big(m_{e_{ij}}^{t+1}\big)
\end{aligned}
\tag{3.5}
$$

όπου $\mathrm{GRU}(\cdot)$ η επαναλαμβανόμενη μονάδα ελεγχόμενης πρόσβασης (Gated Recurrent Unit) [8].

Μια παραλλαγή στην παραπάνω μεθοδολογία που δοκιμάστηκε ήταν η αντικατάσταση των αρχικών οριοθετικών πλαισίων των συσχετίσεων και κατά συνέπεια της εξαγωγής χαρακτηριστικών των περιοχών αυτών με τη συνένωση των χαρακτηριστικών των υποκειμένων και των αντικειμένων των συσχετίσεων. Τα χαρακτηριστικά αυτά αποτελούν την έξοδο των πλήρως συνδεδεμένων επιπέδων και για τη μετατροπή τους σε χαρακτηριστικά των συσχετίσεων διέρχονται μέσα από ένα πλήρως συνδεδεμένο επίπεδο το οποίο τα μετασχηματίζει στις επιθυμητές για τα επόμενα επίπεδα διαστάσεις. Με βάση τη σημειογραφία των δικτύων ανταλλαγής μηνυμάτων, τα διανύσματα των συσχετίσεων $h_{e_{ij}}^0$ ορίζονται ως εξής:

$$
h_{e_{ij}}^0 = \rho\Big(\mathbf{W}\big[h_{v_i}^0, h_{v_j}^0\big]\Big) \tag{3.6}
$$

Μία δεύτερη παραλλαγή της αρχικής μεθοδολογίας, ανεξάρτητη από την παραπάνω, η οποία αποσκοπεί στην απλοποίηση του μοντέλου, χωρίς, ωστόσο, να αφαιρεί από την αποτελεσματικότητά του, είναι η εξής:

$$
\begin{aligned}
M_{\mathcal{V}}(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t, h_{e_{ji}}^t) &= \rho\Bigg(\mathbf{W}_{\mathcal{V}}\Bigg[\rho\Big(\mathbf{W}_A\Big[h_{v_i}^t, h_{e_{ij}}^t, h_{v_j}^t\Big]\Big), \rho\Big(\mathbf{W}_P\Big[h_{v_j}^t, h_{e_{ji}}^t, h_{v_i}^t\Big]\Big)\Bigg]\Bigg),\\
U_{\mathcal{V}}\big(h_{v_i}^t, m_{v_i}^{t+1}\big) &= \mathrm{GRU}\big(m_{v_i}^{t+1}\big),\\
M_{\mathcal{E}}(h_{v_i}^t, h_{v_j}^t, h_{e_{ij}}^t) &= \rho\Big(\mathbf{W}_{\mathcal{E}}\Big[h_{v_i}^t, h_{e_{ij}}^t, h_{v_j}^t\Big]\Big),\\
U_{\mathcal{E}}\big(h_{e_{ij}}^t, m_{e_{ij}}^{t+1}\big) &= \mathrm{GRU}\big(m_{e_{ij}}^{t+1}\big)
\end{aligned}
\tag{3.7}
$$

3.4 Συνελικτικά Δίκτυα Γράφων

Τα συνελικτικά δίκτυα γράφων [37] αποτελούν μία διαμόρφωση των συνελικτικών νευρωνικών δικτύων τα οποία λειτουργούν σε δεδομένα που έχουν δομή γράφου. Η διάδοση της πληροφορίας σε ένα πολυεπίπεδο συνελικτικό δίκτυο γράφων γίνεται σύμφωνα με τον ακόλουθο κανόνα:

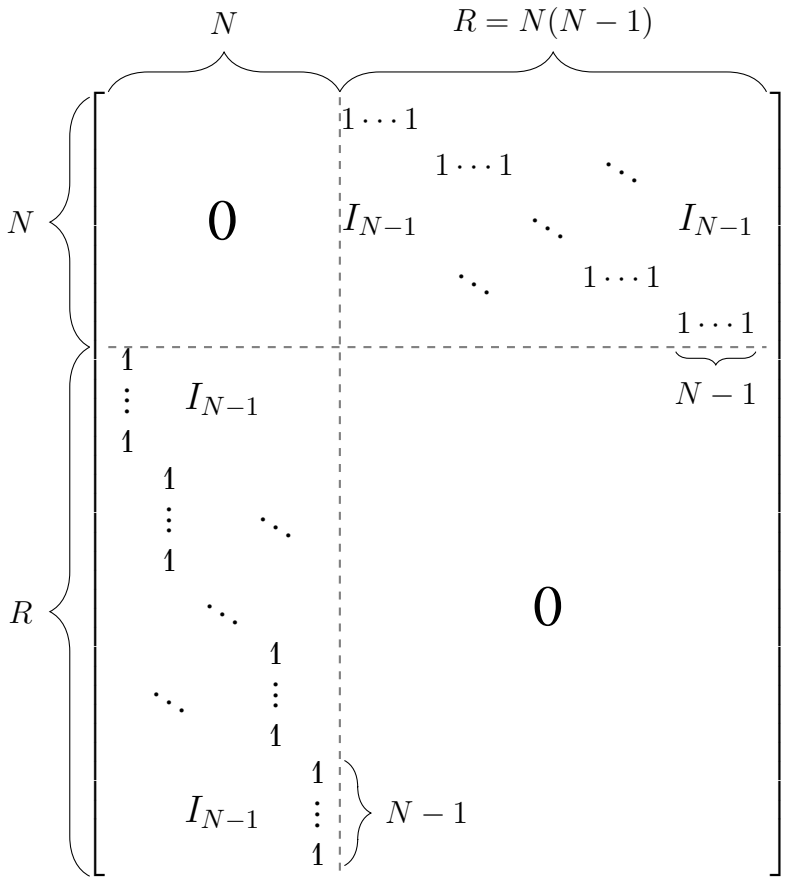
$$H^{(t+1)} = \rho\Big(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(t)}W^{(t)}\Big), \tag{3.8}$$

όπου $\tilde{A} = A + I_N$ είναι ο τετραγωνικός πίνακας γειτνίασης (adjacency matrix) του γράφου $\mathcal{G}$ στον οποίο έχουν προστεθεί βρόχοι μέσω του μοναδιαίου πίνακα I_N και $\tilde{D}$ ο διαγώνιος πίνακας βαθμών (degree matrix) του πίνακα $\tilde{A}$ με στοιχεία $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

Η παραπάνω μοντελοποίηση, αν και αποτελεσματική σε πολλές κατηγορίες προβλημάτων, δεν αντιμετωπίζει τις ακμές του γράφου ως οντότητες και έτσι δεν ενδείκνυται η άμεση εφαρμογή της στα προβλήματα της αναγνώρισης συσχετίσεων και της εξαγωγής γράφου σκηνής εικόνων. Για το λόγο αυτό, οι ακμές του γράφου μοντελοποιήθηκαν και αυτές ως κορυφές, το οποίο είχε σαν αποτέλεσμα τη δημιουργία ενός νέου διμερούς γράφου, στον οποίο υπάρχουν δύο είδη κορυφών, εκείνες που αντιστοιχούν σε αντικείμενα και οι οποίες μπορούν να συνδέονται με ένα αυθαίρετα μεγάλο πλήθος κορυφών και εκείνες που αντιστοιχούν σε συσχετίσεις και συνδέονται μόνο με το υποκείμενο και το αντικείμενο που τους αναλογεί.

Επιπρόσθετα, η αρχική μοντελοποίηση περιορίζεται σε μη κατευθυνόμενους γράφους, ενώ οι γράφοι σκηνής εικόνων είναι κατευθυνόμενοι γράφοι. Ωστόσο, αφ' ενός η κατευθυντικότητα των τριάδων *υποκείμενο* → *κατηγόρημα* → *αντικείμενο* είναι προκαθορισμένη και γνωστή εκ των προτέρων, και αφ' ετέρου η αμφικατευθυντικότητα που υπονοείται σε έναν μη κατευθυνόμενο γράφο επιτρέπει την διάδοση των μηνυμάτων σε περισσότερους και μακρινότερους κόμβους και κατ' επέκταση τη συλλογή περισσότερων πληροφοριών από τους συσχετιζόμενους κόμβους. Ως εκ τούτου, ο πίνακας γειτνίασης μοντελοποιήθηκε ως ένας συμμετρικός διμερής πίνακας γειτνίασης μη κατευθυνόμενου γράφου.

Στο Σχήμα 3.3 παρουσιάζεται ο πίνακας γειτνίασης A ενός «πλήρους» γράφου, στα πλαίσια των επιτρεπτών συνδέσεων. Ο πίνακας γειτνίασης έχει διαστάσεις $N^2 \times N^2$, αφού απαρτίζεται από N αντικείμενα και $R = N(N-1)$ συσχετίσεις οι οποίες μοντελοποιούν την αλληλεπίδραση καθενός από τα N αντικείμενα με τα υπόλοιπα $N-1$ αντικείμενα. Ο πίνακας γειτνίασης αποτελεί έναν διμερή πίνακα ο οποίος συντίθεται από μηδενικούς τετραγωνικούς πίνακες στην κύρια διαγώνιο και από μη μηδενικούς πίνακες στα τμήματα εκτός της κυρίας διαγωνίου. Τα χαρακτηριστικά αυτά οφείλονται στο ότι δεν υπάρχει άμεση επικοινωνία ανάμεσα σε δύο αντικείμενα ή δύο συσχετίσεις. Επικοινωνία υπάρχει μόνο ανάμεσα στο υποκείμενο και το κατηγόρημα και ανάμεσα στο κατηγόρημα και το αντικείμενο μιας συσχέτισης. Επιπρόσθετα, ο πίνακας αυτός είναι ένας συμμετρικός πίνακας, όπως σημειώνεται και στην προηγούμενη παράγραφο. Τέλος, τα μη μηδενικά στοιχεία του πίνακα στα μη διαγώνια τμήματα προκύπτουν από το γεγονός ότι οι συσχετίσεις ορίζονται με τη σειρά ανάμεσα στα ζεύγη αντικειμένων· οι πρώτες, δηλαδή, $N-1$ συσχετίσεις μοντελοποιούν την αλληλεπίδραση ανάμεσα στο πρώτο αντικείμενο και τα υπόλοιπα $N-1$, οι επόμενες $N-1$ ανάμεσα στο δεύτερο αντικείμενο και τα υπόλοιπα $N-1$, κ.ο.κ..



Σχήμα 3.3: Πίνακας Γειτνίασης Συνελικτικού Δικτύου Γράφων

3.4.1 Διατύπωση ως Νευρωνικό Δίκτυο Ανταλλαγής Μηνυμάτων

Τα συνελικτικά δίκτυα γράφων μπορούν να περιγραφούν και ως νευρωνικά δίκτυα ανταλλαγής μηνυμάτων. Μοντελοποιώντας τις συσχετίσεις του γράφου ως κορυφές, προκύπτει η ακόλουθη μαθηματική περιγραφή ενός συνελικτικού δικτύου γράφων, στην οποία ο όρος v_i χρησιμοποιείται για να περιγράψει τόσο τα αντικείμενα όσο και τις συσχετίσεις μιας εικόνας:

$$\begin{aligned} m_{v_i}^{t+1} &= \sum_{v_j \in N(v_i) \cup \{v_i\}} M_{\mathcal{V}}\big(h_{v_i}^t, h_{v_j}^t\big), \\ M_{\mathcal{V}}\big(h_{v_i}^t, h_{v_j}^t\big) &= \big(\deg(v_i)\deg(v_j)\big)^{-1/2} h_{v_j}^t, \\ U_{\mathcal{V}}^t\big(h_{v_i}^t, m_{v_i}^{t+1}\big) &= \rho\big(\mathbf{W}^t m_{v_i}^{t+1}\big) \end{aligned} \tag{3.9}$$

3.5 Δίκτυο Διαγραφής Συσχετίσεων

Το δίκτυο διαγραφής συσχετίσεων αποτελεί μια απλή υλοποίηση του υποσυστήματος εξαγωγής συσχετίσεων που δημιουργήθηκε με σκοπό τον έλεγχο της ποιότητας των συσχετίσεων στην απόδοση του συστήματος. Το δίκτυο αυτό δρα στο πλήρες σύνολο των συσχετίσεων μεταξύ των αντικειμένων. Η εφαρμογή του, δηλαδή, έπεται της εξαγωγής περιοχών αντικειμένων και της αρχικής εξαγωγής συσχετίσεων, ενώ η έξοδος του

δικτύου αυτού αποτελεί την αρχική είσοδο συσχετίσεων στο δίκτυο ανταλλαγής μηνυμάτων. Η χρήση του δικτύου γίνεται μία μόνο φορά και δεν εφαρμόζεται επαναληπτικά στις κωδικοποιημένες συσχετίσεις του δικτύου ανταλλαγής μηνυμάτων.

Η συνθήκη υπό την οποία η συσχέτιση $h^0_{e_{ij}}$ διαγράφεται είναι η εξής:

$$\sigma(\mathbf{w}_p^T[h^0_{v_i}, h^0_{e_{ij}}, h^0_{v_j}]) < 0.5$$

όπου $\mathbf{w}_p$ το (εκπαιδεύσιμο) διάνυσμα που βαθμολογεί τις συσχετίσεις και $\sigma(\cdot)$ η λογιστική συνάρτηση.

Κεφάλαιο 4

Πειραματική Διαδικασία

4.1 Σύνολο Δεδομένων Visual Genome

Για την εκπαίδευση και την αξιολόγηση του συστήματος που υλοποιήθηκε, επιλέχθηκε το σύνολο δεδομένων Visual Genome [39]. Η έκδοση 1.4 του Visual Genome αποτελείται από 108,077 εικόνες, κάθε μία από τις οποίες περιέχει κατά μέσο όρο 23.28 αντικείμενα και 21.43 συσχετίσεις. Ωστόσο, ένα σημαντικό ποσοστό των αντικειμένων χαρακτηρίζεται από χαμηλή ποιότητα πλαισίωσης· τα οριοθετικά πλαίσια δεν ενθυλακώνουν αυστηρά το εκάστοτε αντικείμενο αλλά περιέχουν και θόρυβο υποβάθρου ή αποκόπτουν τμήματά του. Επιπρόσθετα, παρατηρείται υψηλός βαθμός επικάλυψης σε πλαίσια που χαρακτηρίζουν το ίδιο αντικείμενο. Τέλος, η χρήση ελεύθερου κειμένου για το σχολιασμό των εικόνων παρουσιάζει το μειονέκτημα της εμφάνισης πολλαπλών ονομάτων για την περιγραφή των αντικειμένων και των συσχετίσεων, ενώ δεν απουσιάζει και η διαφορετική γραφή των περιγραφών των οντοτήτων (π.χ. η χρήση κεφαλαίων γραμμάτων και σημείων στίξης).

Για τους λόγους αυτούς πραγματοποιήθηκε προεπεξεργασία των δεδομένων, τόσο των πλαισίων των αντικειμένων όσο και των ονομάτων των οντοτήτων. Πιο συγκεκριμένα, τα ονόματα των οντοτήτων κανονικοποιήθηκαν (stemming) αφού πρώτα αφαιρέθηκαν πιθανά σημεία στίξης και μη γραμματικοί χαρακτήρες και έγινε μετατροπή σε μικρά γράμματα του λατινικού αλφαβήτου. Από τις κανονικοποιημένες οντότητες, επιλέχθηκαν τα 150 πιο συχνά αντικείμενα και οι 50 πιο συχνές συσχετίσεις με βάση τις οποίες δημιουργήθηκαν οι κατηγορίες εκπαίδευσης και αξιολόγησης των μοντέλων. Αντίστοιχα, πραγματοποιήθηκε προεπεξεργασία των πλαισίων των αντικειμένων, στην οποία συνενώθηκαν πλαίσια με υψηλό ποσοστό επικάλυψης και αφαιρέθηκαν πλαίσια με μικρό σχετικό μέγεθος. Στο τελευταίο στάδιο της προεπεξεργασίας, αφαιρέθηκαν από το σύνολο δεδομένων οι εικόνες που δεν περιείχαν ούτε μία συσχέτιση (και κατ' επέκταση είχαν λιγότερα από δύο αντικείμενα), καθώς η αξιολόγηση των μοντέλων γίνεται με βάση τις συσχετίσεις.

Το σύνολο δεδομένων που προέκυψε αποτελείται από 84,085 εικόνες, οι οποίες συνιστούν το 77.8% του αρχικού συνόλου δεδομένων. Κάθε εικόνα περιέχει κατά μέσο όρο 12.1 αντικείμενα και 4.56 συσχετίσεις. Η επιλογή των εικόνων των συνόλων εκπαίδευσης,

επικύρωσης και ελέγχου έγινε με τυχαία δειγματοληψία ώστε να αποφευχθούν πιθανές σημασιολογικές συσχετίσεις ανάμεσα σε αριθμητικά γειτονικές εικόνες. Τέλος, το σύνολο δεδομένων χωρίστηκε σε 60%/10%/30% σύνολο εκπαίδευσης/σύνολο επικύρωσης/σύνολο ελέγχου.

Το σύνολο δεδομένων εκπαίδευσης, το οποίο συνιστά το 60% του συνόλου των δεδομένων, αποτελείται από 50,397 εικόνες, ενώ το σύνολο επικύρωσης αποτελείται από 8,474 εικόνες. Για τον πολλαπλασιασμό των δεδομένων εκπαίδευσης πραγματοποιήθηκε επαύξηση των δεδομένων (data augmentation), και συγκεκριμένα προστέθηκαν στα σύνολο εκπαίδευσης και επικύρωσης οι καθρεπτισμένες εκδοχές όλων των εικόνων των συνόλων αυτών. Με τον τρόπο αυτό το σύνολο εκπαίδευσης πρακτικά διπλασιάστηκε σε μέγεθος, με 100,794 εικόνες, ενώ το άθροισμα συνόλων εκπαίδευσης και επικύρωσης απέκτησε 117,742 εικόνες.

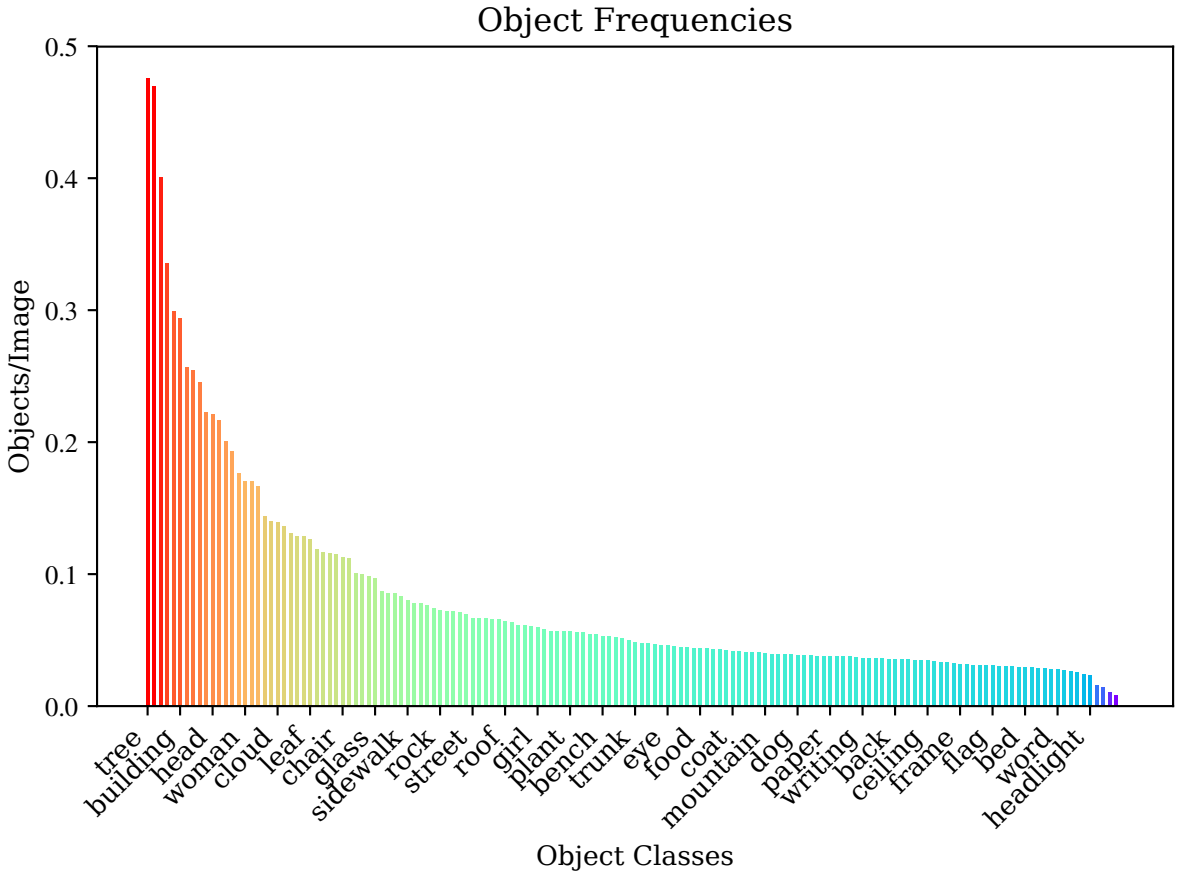
Ενδιαφέρον παρουσιάζει η συχνότητα εμφάνισης αντικειμένων και συσχετίσεων στο σύνολο δεδομένων, καθώς όπως αντικατοπτρίζεται και στα Σχήματα 4.1 και 4.2 υπάρχει μεγάλη απόκλιση στις συχνότητες των οντοτήτων. Στα Σχήματα 4.3 και 4.4 παρουσιάζεται η κατανομή των οντοτήτων στα τρία υποσύνολα δεδομένων, όπου είναι φανερή η ίση κατανομή τόσο των αντικειμένων όσο και των συσχετίσεων. Αντίθετα, στα Σχήματα 4.5 και 4.6 παρουσιάζεται η κατανομή των οντοτήτων στα υποσύνολα δεδομένων δίχως την τυχαία ανάμειξη των εικόνων, αλλά με βάση το σειριακό διαχωρισμό του συνόλου δεδομένων. Όπως είναι εμφανές, παρατηρούνται αποκλίσεις ανάμεσα στα τρία υποσύνολα σε αρκετές κλάσεις αντικειμένων και κατηγορημάτων, γεγονός που δικαιολογεί την επιλογή της ανάμειξης των δεδομένων.

4.2 Εκπαίδευση Μοντέλων

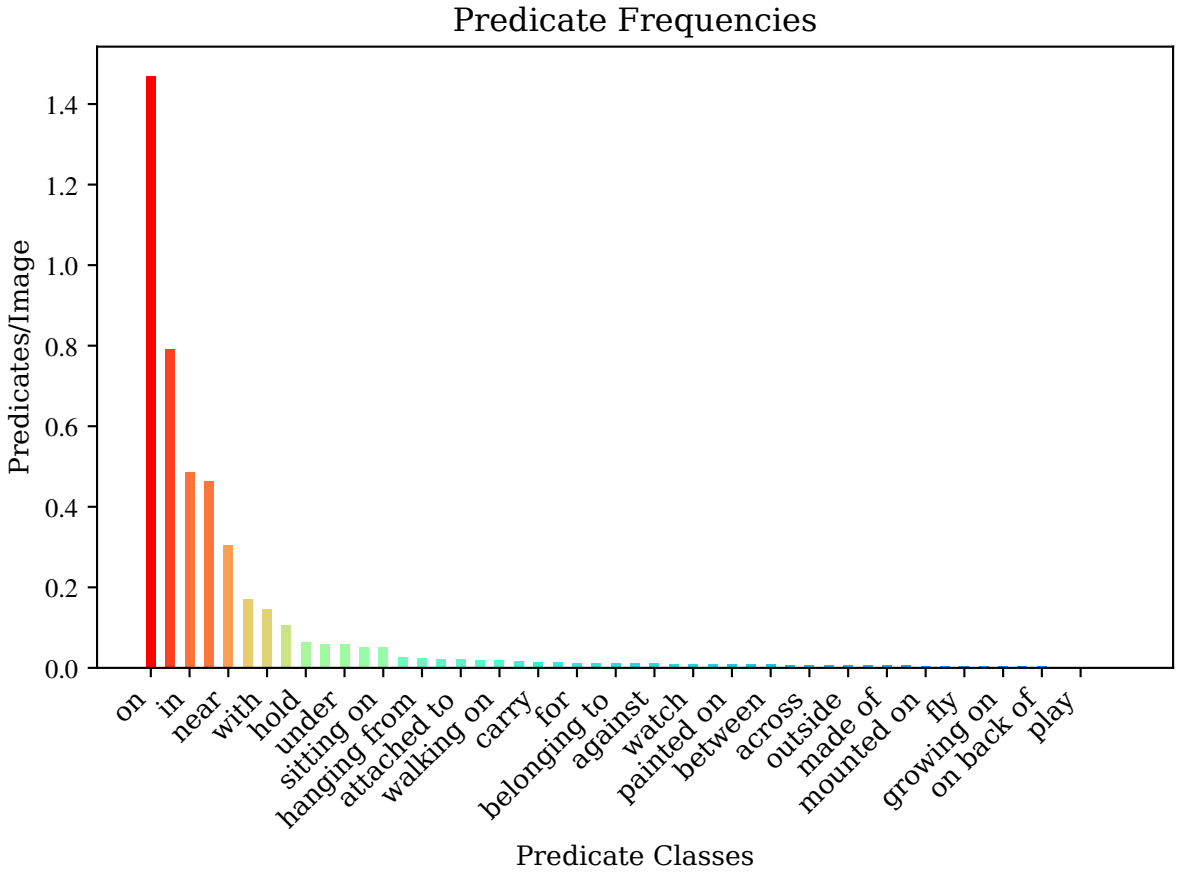
Σε κάθε επανάληψη του κύκλου εκπαίδευσης η στρατηγική δειγματοληψίας του mini-batch ακολουθεί μία εικονοκεντρική προσέγγιση, αντίστοιχη με αυτή του Faster R-CNN [57]. Συγκεκριμένα, σε κάθε mini-batch επιλέγονται 1-2 εικόνες και 256 περιοχές ενδιαφέροντος, οι οποίες αποτελούν έναν συνδυασμό των αληθών περιοχών και των προτεινόμενων περιοχών του δικτύου πρότασης περιοχών.

Η εξαγωγή οπτικών χαρακτηριστικών από τις εικόνες γίνεται από το προεκπαιδευμένο στο σύνολο δεδομένων MS COCO συνελικτικό δίκτυο VGG-16 [62] του μοντέλου ανίχνευσης αντικειμένων Faster R-CNN. Σε όλα τα πειράματα που διεξήχθησαν, εκτός και εάν αναφέρεται διαφορετικά, χρησιμοποιείται το παραπάνω μοντέλο, με σκοπό την μετέπειτα ορθότερη σύγκριση της αξιολόγησης των μοντέλων.

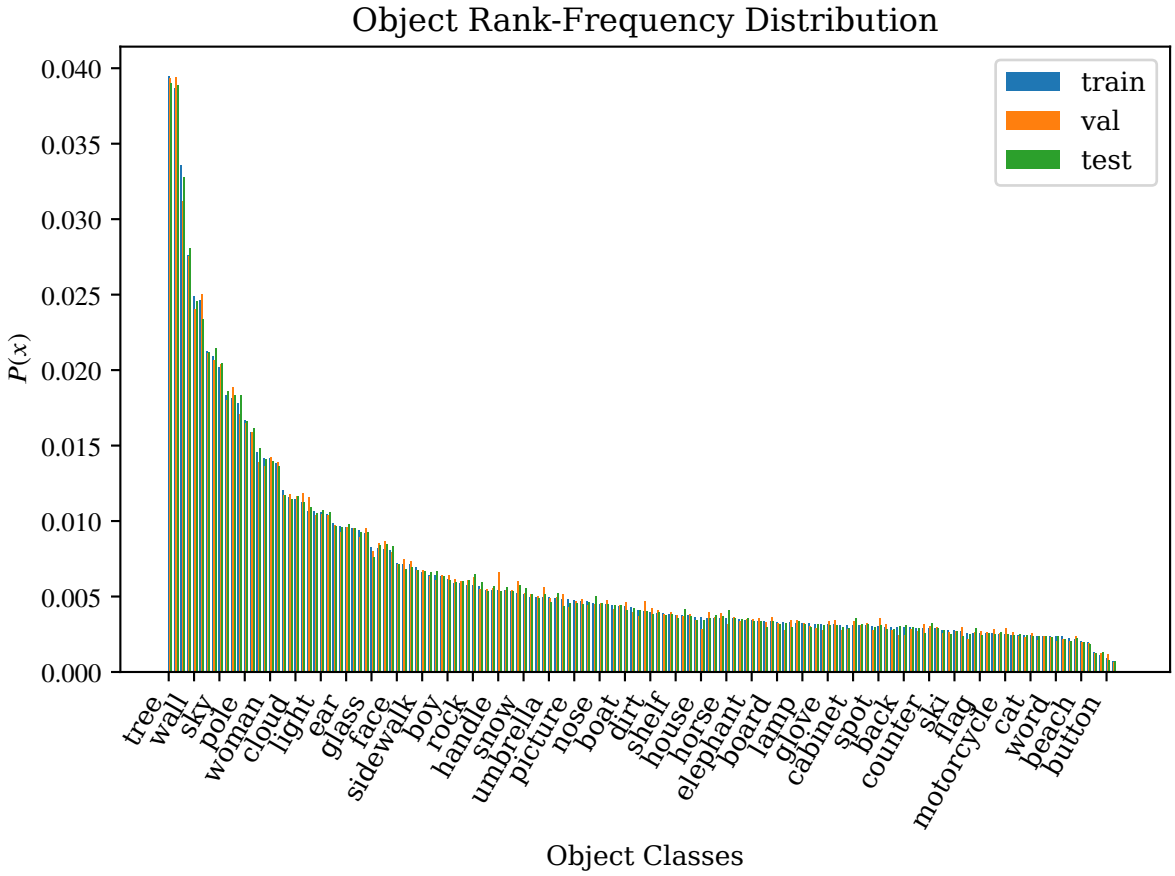
Η εξαγωγή των χαρακτηριστικών των περιοχών ενδιαφέροντος γίνεται με χρήση του αλγορίθμου RoI Align [24], με τη χρήση, δηλαδή, διγραμμικής παρεμβολής στους συνελικτικούς χάρτες χαρακτηριστικών. Η ίδια λογική ακολουθείται και στην εξαγωγή των χαρακτηριστικών των πιθανών συσχετίσεων. Ως πλαίσιο μιας συσχέτισης ορίζεται η ένωση των περιγεγραμμένων πλαισίων των αντικειμένων που την αποτελούν.



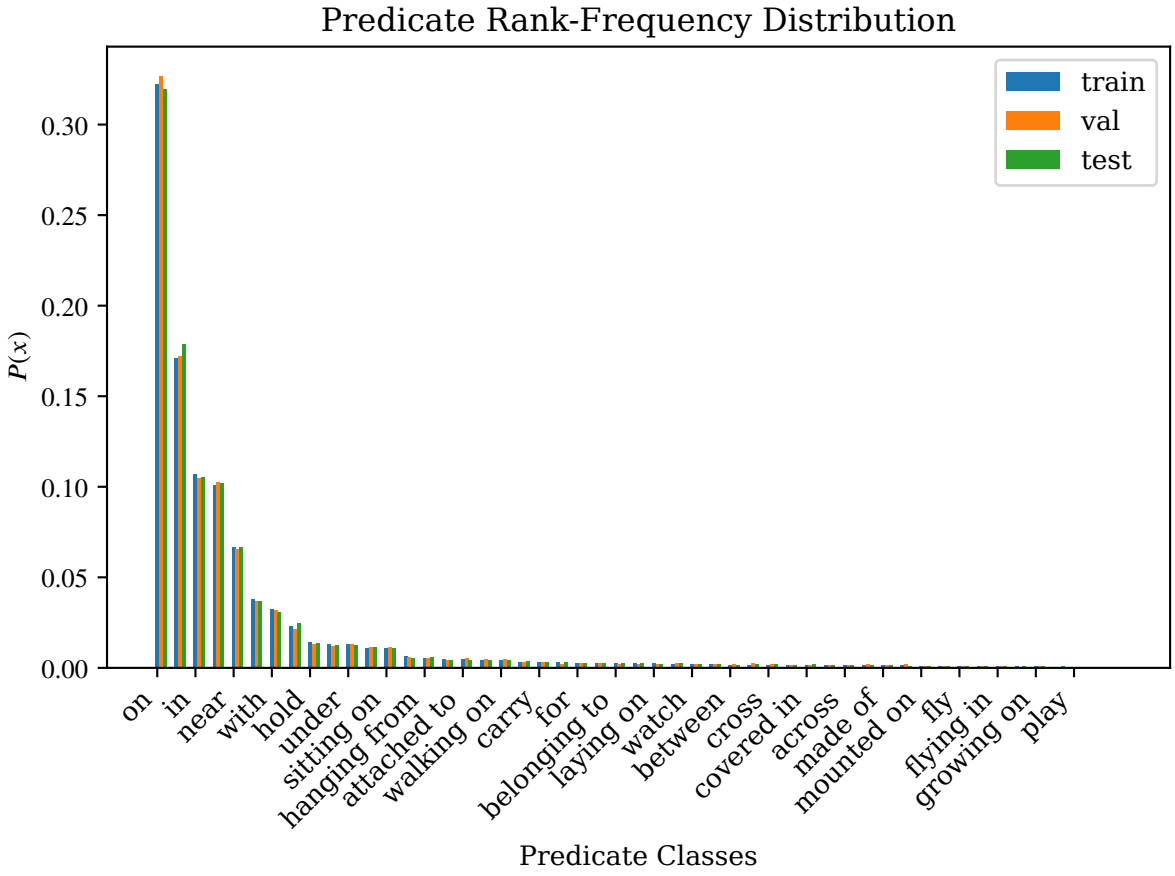
Σχήμα 4.1: Συχνότητα Εμφανίσεων Αντικειμένων



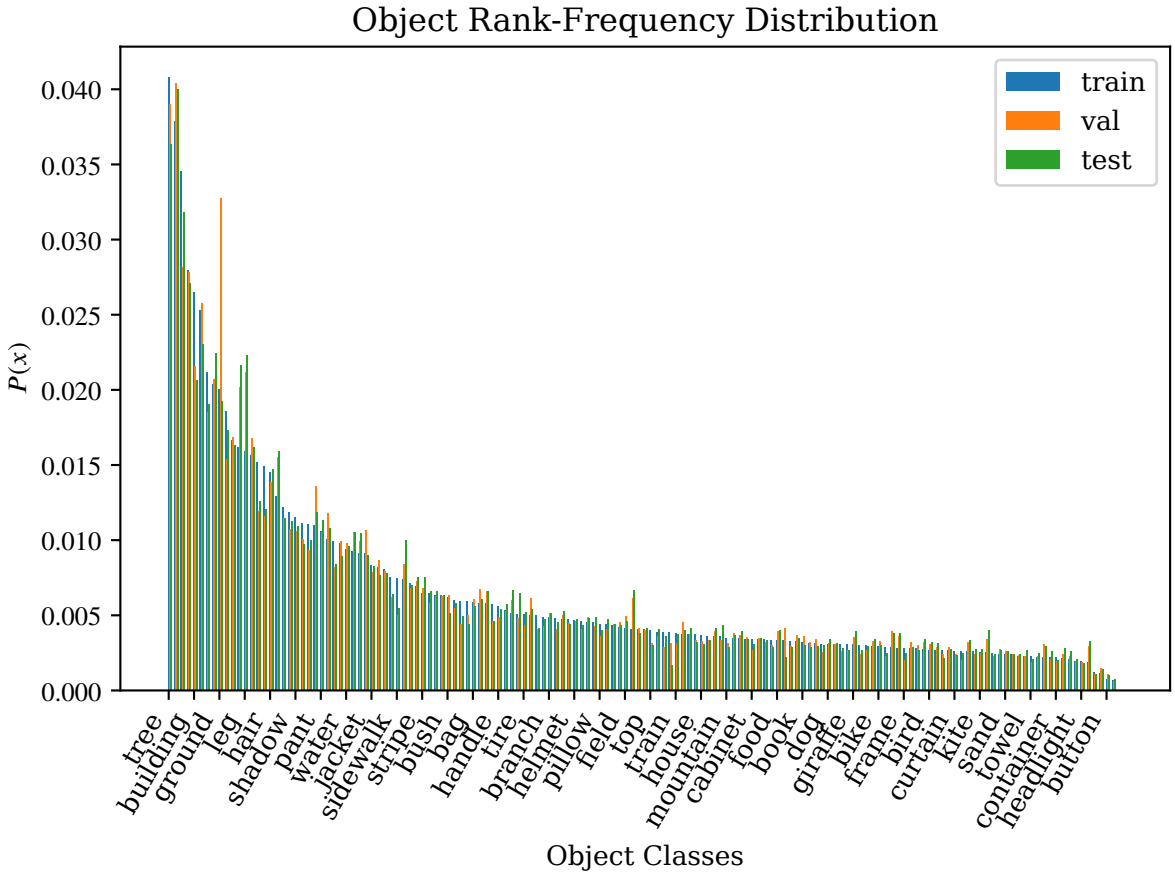
Σχήμα 4.2: Συχνότητα Εμφανίσεων Συσχετίσεων



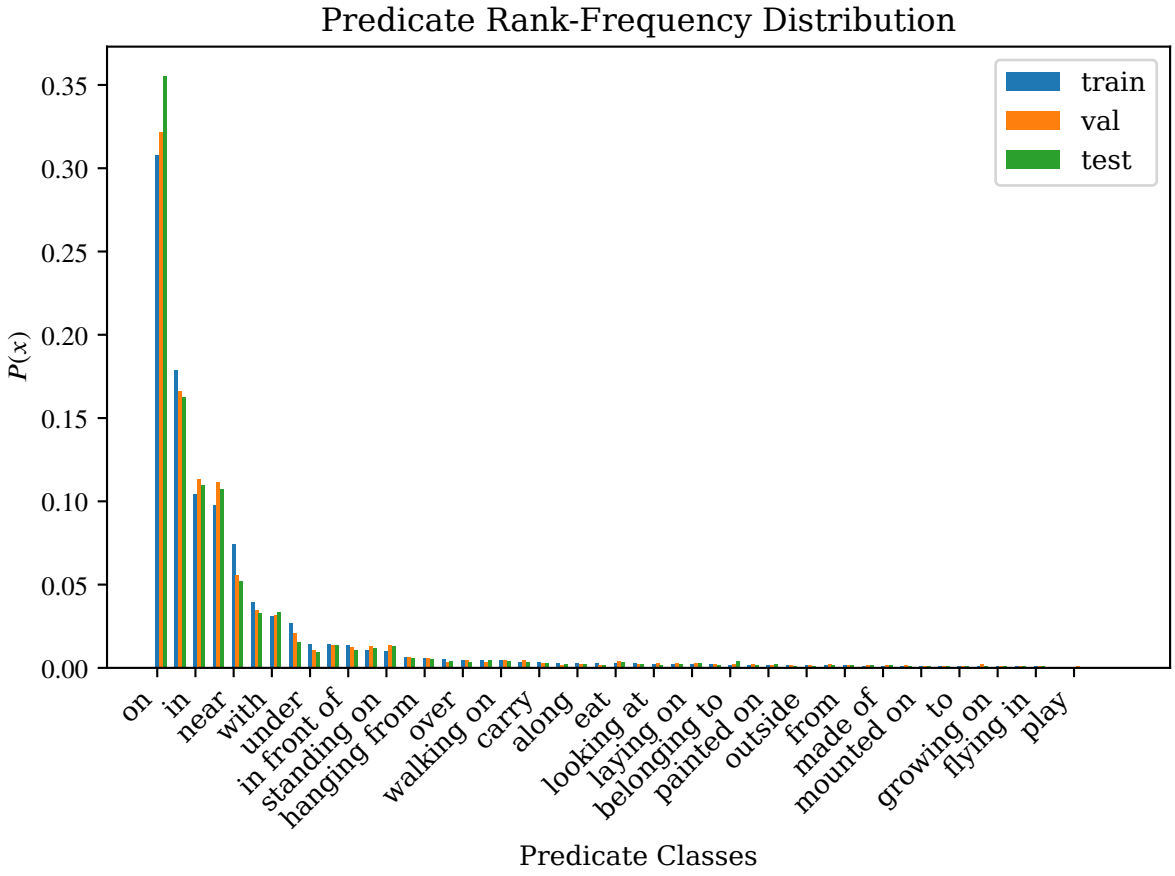
Σχήμα 4.3: Κατανομή Αντικειμένων



Σχήμα 4.4: Κατανομή Συσχετίσεων



Σχήμα 4.5: Κατανομή Αντικειμένων μη αναμεμειγμένου συνόλου δεδομένων



Σχήμα 4.6: Κατανομή Συσχετίσεων μη αναμεμειγμένου συνόλου δεδομένων

Οι επαναλήψεις στα επίπεδα ανταλλαγής μηνυμάτων τόσο των νευρωνικών δικτύων ανταλλαγής μηνυμάτων όσο και των συνελικτικών δικτύων γράφων κυμαίνονται στις 1-3, ενώ τα επίπεδα αυτά προβάλλουν τις εισόδους τους σε ένα χώρο 512 διαστάσεων, εκτός και εάν αναφέρεται διαφορετικά σε ορισμένες διατάξεις.

Οι εικόνες διέρχονται μέσα από τα επίπεδα του δικτύου, στην έξοδο των οποίων δύο επίπεδα softmax παράγουν τις τελικές προβλέψεις για τα αντικείμενα και τα κατηγορήματα, αντίστοιχα, ενώ ένα πλήρως συνδεδεμένο επίπεδο παλινδρομεί στο αντιστάθμισμα των οριοθετικών πλαισίων. Οι απώλειες δίνονται από τη συνάρτηση διεντροπίας (cross entropy) για τα αντικείμενα και τα κατηγορήματα και από τη συνάρτηση απωλειών Huber (smooth L1) για τα οριοθετικά πλαίσια.

Η εκπαίδευση των μοντέλων έγινε σύμφωνα με τον αλγόριθμο στοχαστικής καθόδου κλίσης (Stochastic Gradient Descent) και η διάδοση των σφαλμάτων στα επιμέρους βάρη του δικτύου γίνεται σύμφωνα με τον αλγόριθμο οπισθοδιάδοσης (backpropagation). Τα βάρη των επιπέδων με νευρώνες ενεργοποίησης ReLU αρχικοποιούνται σύμφωνα με την αρχικοποίηση He [26] και τα υπόλοιπα βάρη σύμφωνα με την αρχικοποίηση Glorot [21]. Τέλος, τα βάρη των συνελικτικών επιπέδων διατηρούνται σταθερά καθ' όλη τη διάρκεια της πειραματικής διαδικασίας.

4.2.1 Στρατηγική Δειγματοληψίας Mini-batch

Κατά τη διάρκεια της εκπαίδευσης εξάγονται από κάθε εικόνα 2,000 περιοχές ενδιαφέροντος από σύνολο των προτάσεων του δικτύου πρότασης περιοχών μέσω του αλγορίθμου non-maximum suppression. Οι περιοχές αυτές εξάγονται μαζικά για όλες τις εικόνες του συνόλου δεδομένων και αποθηκεύονται σε μία βάση δεδομένων με σκοπό την ταχύτερη πρόοδο στην εκπαίδευση των μοντέλων. Για κάθε προτεινόμενη περιοχή βρίσκουμε τη μέγιστη ποσοστιαία επικάλυψη με τις αληθείς περιοχές, καθώς και το αντικείμενο στο οποίο αντιστοιχεί. Κάθε περιοχή της οποίας η επικάλυψη ξεπερνάει το 0.5 χρησιμοποιείται στην παλινδρόμηση οριοθετικού πλαισίου και για το λόγο αυτό υπολογίζεται ο παρακάτω μετασχηματισμός πλαισίου, σύμφωνα με το [20]:

$$\begin{aligned} t_x &= (g_x - p_x)/p_w \\ t_y &= (g_y - p_y)/p_h \\ t_w &= \ln(g_w/p_w) \\ t_h &= \ln(g_h/p_h) \end{aligned} \tag{4.1}$$

όπου:

$\{t_x, t_y, t_w, t_h\}$ = ο στόχος της παλινδρόμησης,
$\{p_x, p_y, p_w, p_h\}$ = το οριοθετικό πλαίσιο της προτεινόμενης περιοχής,
$\{g_x, g_y, g_w, g_h\}$ = το οριοθετικό πλαίσιο της αληθούς περιοχής,

εκφρασμένα σε συντεταγμένες κέντρου (x, y) και διαστάσεων πλαισίου (w, h).

Οι πρώτες δύο εξισώσεις προσδιορίζουν έναν αναλλοίωτο στην κλιμάκωση μετασχηματισμό μετάθεσης του κέντρου του προτεινόμενου πλαισίου, ενώ οι δύο επόμενες προσδιορίζουν έναν μετασχηματισμό μετάθεσης λογαριθμικού χώρου των διαστάσεων του πλαισίου.

Στη συνέχεια επαυξάνουμε το σύνολο των συσχετίσεων, αντικαθιστώντας όλα τα αληθή αντικείμενα και υποκείμενα που τις αποτελούν με τις περιοχές ενδιαφέροντος που έχουν επικάλυψη μεγαλύτερη από 0.5, συμπεριλαμβανομένων, προφανώς, και των αρχικών οντοτήτων και δειγματοληπτούμε συσχετίσεις, μέχρις ότου συμπληρωθεί ένα προκαθορισμένο πλήθος περιοχών το οποίο ισούται καθ' όλη τη διάρκεια των πειραμάτων με 128 ή έως ότου εξαντληθούν οι αληθείς συσχετίσεις. Στην περίπτωση που εξαντληθούν οι συσχετίσεις προτού συμπληρωθεί το απαιτούμενο πλήθος περιοχών, δειγματοληπτούμε περιοχές οι οποίες δεν αποτελούν τμήμα καμίας συσχέτισης. Επιπρόσθετα, υπολογίζουμε όλες τις συσχετίσεις υποβάθρου, τις συσχετίσεις, δηλαδή, εκείνες, μεταξύ αντικειμένων που δε συσχετίζονται και δειγματοληπτούμε ένα προκαθορισμένο πλήθος από αυτές, το οποίο σε όλα τα πειράματα διατηρείται σταθερό και ισούται με 128. Τέλος, εάν το απαιτούμενο πλήθος περιοχών δεν έχει ακόμα επιτευχθεί, αυτό συμπληρώνεται με περιοχές υποβάθρου, περιοχές, δηλαδή, οι οποίες δεν έχουν επικάλυψη μεγαλύτερη από 0.5 με κανένα αληθές αντικείμενο.

Προς αποφυγήν σύγχυσης, σημειώνεται στο σημείο αυτό ότι τόσο στις κλάσεις αντικειμένων όσο και στις κλάσεις κατηγορημάτων έχει προστεθεί η κλάση υποβάθρου, η οποία αντιπροσωπεύει την έλλειψη συμμετοχής μιας οντότητας στις κατηγορίες που ορίζονται από το σύνολο δεδομένων. Η κλάση αυτή χρησιμοποιείται στην εκπαίδευση και την αξιολόγηση των μοντέλων και παρέχει σε αυτά τη δυνατότητα, μεταξύ άλλων, να εξαλείψουν περιοχές ενδιαφέροντος τις οποίες λανθασμένα προέβλεψε το δίκτυο πρότασης περιοχών αλλά και να απαλείψουν συνδυασμούς αντικειμένων τα οποία δεν συσχετίζονται.

4.2.2 Τεχνικές Βελτιστοποίησης

Η εκπαίδευση των μοντέλων έγινε σύμφωνα με τον αλγόριθμο στοχαστικής καθόδου κλίσης (Stochastic Gradient Descent). Κατά τη διάρκεια των πειραμάτων, δοκιμάστηκαν αρκετές διαφοροποιήσεις του αλγορίθμου, οι οποίες στις περισσότερες περιπτώσεις απολαμβάνουν καλύτερους ρυθμούς σύγκλισης χωρίς να είναι τόσο επιρρεπείς στην επιλογή του ρυθμού εκμάθησης και των εσωτερικών τους παραμέτρων. Συγκεκριμένα, χρησιμοποιήθηκαν οι αλγόριθμοι Momentum [53] και Nesterov Accelerated Gradient [50, 51, 64], καθώς και οι μέθοδοι προσαρμοζόμενης ρύθμισης παραμέτρων RMSProp [69] και Adam [36].

Για την καλύτερη σύγκλιση των μοντέλων χρησιμοποιήθηκε κλιμακωτή φθορά (step decay) του ρυθμού εκμάθησης. Η φθορά ήταν σταθερή σε όλα τα πειράματα και ίση με 10%, ενώ η συχνότητά της κυμάνθηκε από 0.5 έως 2 εποχές (epochs).

4.2.3 Τεχνικές Κανονικοποίησης

Για την αποφυγή της υπερπροσαρμογής των μοντέλων χρησιμοποιήθηκαν τεχνικές κανονικοποίησης (regularization). Συγκεκριμένα, χρησιμοποιήθηκε η κανονικοποίηση L_2, γνωστή και ως φθορά βαρών (weight decay), όπως επίσης και η τεχνική Dropout [63] με σταθερή πιθανότητα διατήρησης νευρώνων 0.5.

4.2.4 Συναρτήσεις Απωλειών

Οι απώλειες δίνονται από τη συνάρτηση διεντροπίας (cross entropy) για τα αντικείμενα και τα κατηγορήματα και από τη συνάρτηση απωλειών Huber (smooth L1) για τα οριοθετικά πλαίσια. Οι απώλειες των κατηγορημάτων δεν λαμβάνουν υπ' όψιν στα περισσότερα πειράματα τις αληθείς συσχετίσεις υποβάθρου. Αντίστοιχα, η παλινδρόμηση οριοθετικού πλαισίου δεν λαμβάνει υπ' όψιν τις αληθείς περιοχές υποβάθρου.

$$\mathcal{L}_{obj} = -\sum_{i=1}^{N_O}\sum_{c=0}^{C_O} y_{ic} \ln p_{ic} \tag{4.2}$$

$$\mathcal{L}_{pred} = -\sum_{i=1}^{N_P}\sum_{c=1}^{C_P} y_{ic} \ln p_{ic} \tag{4.3}$$

$$\mathcal{L}_H(x;\delta) = \begin{cases} 0.5x^2, & \text{αν } |x| \leq \delta \\ \delta(|x| - 0.5\delta), & \text{αλλιώς} \end{cases} \tag{4.4}$$

$$\mathcal{L}_{bbox} = \sum_{i=1}^{N_O}\sum_{c=1}^{C_O} y_{ic} \sum_{b\in\boldsymbol{B}} \mathcal{L}_H(t_{ib} - d_{icb}; \delta = 1) \tag{4.5}$$

όπου y_{ic} η τιμή του one-hot encoding διανύσματος $\boldsymbol{y}_i$ για το αντικείμενο/κατηγόρημα i και την κλάση c, p_{ic} η προβλεπόμενη πιθανότητα για το αντικείμενο/κατηγόρημα i και την κλάση c, t_{ib} ο στόχος της παλινδρόμησης για το αντικείμενο i, όπως ορίζεται στην Εξίσωση 4.1, d_{icb} η προβλεπόμενη απόκλιση οριοθετικού πλαισίου για το αντικείμενο i και την κλάση c, $\boldsymbol{B} = \{x, y, w, h\}$, N_O και N_P το πλήθος των αντικειμένων και των κατηγορημάτων του mini-batch, αντίστοιχα, και τέλος, C_O και C_P οι κλάσεις των αντικειμένων και των κατηγορημάτων, αντίστοιχα.

Οι συνολικές απώλειες σε κάθε κύκλο επανάληψης δίνονται από το σταθμισμένο άθροισμα των επιμέρους απωλειών δεδομένων, συναθροιζόμενες με τις απώλειες κανονικοποίησης $\mathcal{L}_{reg}$.

$$\mathcal{L}_{total} = w_{obj} \cdot \mathcal{L}_{obj} + w_{pred} \cdot \mathcal{L}_{pred} + w_{bbox} \cdot \mathcal{L}_{bbox} + \lambda \cdot \mathcal{L}_{reg} \tag{4.6}$$

Στα περισσότερα πειράματα που διεξήχθησαν οι συντελεστές των επιμέρους απωλειών δεδομένων ισούνται με 1.

Στα πειράματα που χρησιμοποιείται το δίκτυο διαγραφής συσχετίσεων προστίθεται

μία ακόμα συνάρτηση απωλειών στις συνολικές απώλειες του δικτύου. Οι απώλειες του δικτύου αυτού δίνονται από τη συνάρτηση δυαδικής διεντροπίας:

$$\mathcal{L}_{prune} = -\sum_{i=1}^{N_P} \big(y_i \ln p_i + (1 - y_i) \ln(1 - p_i)\big) \tag{4.7}$$

όπου y_i η δυαδική μεταβλητή που δηλώνει την ύπαρξη αληθούς συσχέτισης και p_i η προβλεπόμενη πιθανότητα ύπαρξης για τη συσχέτιση αυτή.

4.2.5 Βελτιστοποίηση Υπερπαραμέτρων

Η επιλογή των βέλτιστων υπερπαραμέτρων επιτεύχθηκε μέσω του χειροκίνητου συντονισμού υπερπαραμέτρων (hyperparameter tuning), σε συνδυασμό με την τυχαία αναζήτηση στο χώρο των υπερπαραμέτρων [4]. Οι κυριότερες παραμέτροι που συμμετείχαν στη διαδικασία βελτιστοποίησης είναι ο ρυθμός εκμάθησης και η συχνότητα κλιμακωτής φθοράς του, η επιλογή του αλγορίθμου βελτιστοποίησης και οι παράμετροι του, η ισχύς της κανονικοποίησης και τέλος, το πλήθος επαναλήψεων των επιπέδων ανταλλαγής μηνυμάτων και οι διαστάσεις των βαρών τους.

4.3 Αξιολόγηση Μοντέλων

Κατά την αξιολόγηση του συστήματος, επιλέχθηκαν για κάθε εικόνα οι 50 επικρατέστερες περιοχές ενδιαφέροντος από το σύνολο των προτάσεων του δικτύου πρότασης περιοχών με χρήση του αλγορίθμου NMS (non-maximum suppression), με κατώφλι λόγου IoU ίσο με 0.3. Ο αλγόριθμος NMS αποτελεί ένα άπληστο αλγόριθμο διαγραφής διπλότυπων περιοχών ενδιαφέροντος με υψηλά ποσοστά επικάλυψης, διατηρώντας μονάχα τις περιοχές με τις υψηλότερες βαθμολογίες από κάθε επικαλυπτόμενο σύνολο. Σε όλα τα πειράματα, το σύστημα έκανε προβλέψεις συσχετίσεων για όλες τις πιθανές δυάδες αντικειμένων, πέραν των αυτοσυσχετίσεων.

4.3.1 Μετρικές Αξιολόγησης

Οι συνήθεις μετρικές αξιολόγησης στην αναγνώριση αντικειμένων είναι αυτές της μέσης αντιπροσωπευτικής ακρίβειας (mean average precision ή mAP) και πιο συγκεκριμένα οι μετρικές $mAP@0.5$ και $mAP@\,[0.5 : 0.05 : 0.95]$, όπως αυτές ορίζονται από τους διαγωνισμούς PASCAL VOC 2010[15] και Microsoft COCO [45] αντίστοιχα. Ωστόσο, όπως επιχειρηματολογείται και στο [47], η μετρική αυτή, καθώς και όλες οι μετρικές που βασίζονται στην ακρίβεια (precision), αποδίδουν μια πεσσιμιστική απεικόνιση της πραγματικής ποιότητας ενός συστήματος. Πιο συγκεκριμένα, η έλλειψη εξονυχιστικής ταυτοποίησης όλων των συσχετίσεων μιας εικόνας θα είχε σαν αποτέλεσμα την σμίκρυνση της μέσης ακρίβειας στις περιπτώσεις όπου το μοντέλο προέβλεπε μη ταυτοποιημένες συσχετίσεις, ακόμα και εάν αυτές ήταν αληθείς.

Η μετρική που προτάθηκε αντ' αυτού και που αποδίδει πιο ρεαλιστικά την ποιότητα των συστημάτων είναι η μετρική ανάκλησης στα k (recall at k ή R@k) [2, 47] και πιο συγκεκριμένα η μετρική $R@\,[k, IoU = 0.5]$. Η μετρική αυτή υπολογίζει το ποσοστό εμφάνισης αληθών τριάδων συσχετίσεων στις k πιο υψηλόβαθμες προβλέψεις και στις οποίες ο λόγος της ένωσης ως προς την τομή (intersection over union ή IoU) των πλαισίων των προβλεπόμενων υποκειμένων και αντικειμένων σε σχέση με τα αληθινά είναι τουλάχιστον 0.5. Οι τιμές του k που επιλέχθηκαν για τα πειράματα είναι 20, 50 και 100 και οι μετρικές οι R@20, R@50, R@100.

4.3.2 Εργασίες Αξιολόγησης

Τα μοντέλα που εκπαιδεύθηκαν αξιολογήθηκαν σε τρεις εργασίες αυξανόμενης δυσκολίας, σύμφωνα με τα [47, 71]. Παρότι η κύρια εργασία αξιολόγησης είναι η τελευταία, η οποία και έχει τον μεγαλύτερο βαθμό ρεαλισμού και ενδιαφέροντος, οι υπόλοιπες εργασίες παρέχουν πολύ χρήσιμες πληροφορίες και βοηθούν στην περαιτέρω κατανόηση του προβλήματος, παρέχοντας επιπρόσθετα επίπεδα αφαίρεσης και απομακρύνοντας τυχόν περιορισμούς που μπορεί να θέτει η ενυπάρχουσα αρχιτεκτονική.

Η πρώτη εργασία αξιολόγησης είναι η **ταξινόμηση κατηγορημάτων** (predicate classification). Στην εργασία αυτή δίνεται ως είσοδος στο σύστημα μία εικόνα καθώς και το σύνολο των χωροθετημένων αντικειμένων που υπάρχουν σε αυτή. Στόχος της είναι η εύρεση και η ταξινόμηση των κατηγορημάτων που συνδέουν ζεύγη αντικειμένων και κατ' επέκταση των συσχετίσεών τους. Οι κανονισμοί αυτοί επιτρέπουν τη μελέτη της δυσκολίας ανίχνευσης και κατηγοριοποίησης συσχετίσεων δίχως τους επιπρόσθετους περιορισμούς που τίθενται από το σύστημα αναγνώρισης αντικειμένων.

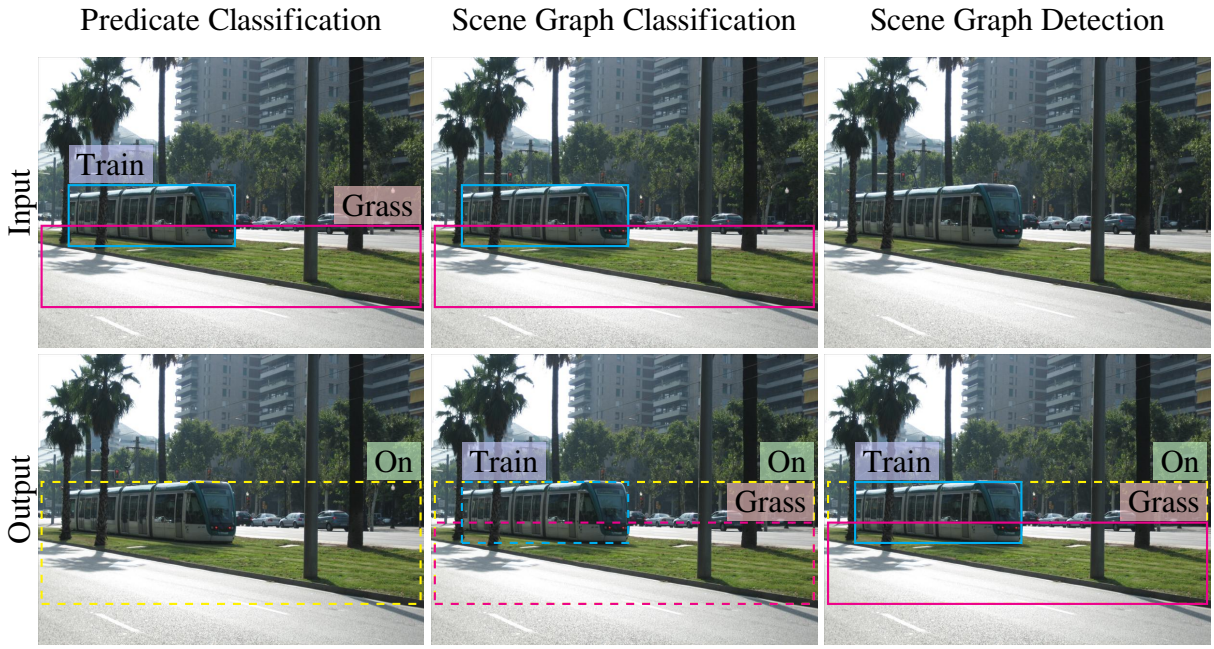
Η δεύτερη εργασία αξιολόγησης είναι η **ταξινόμηση γράφου σκηνής** (scene graph classification). Στην εργασία αυτή δίνεται ως είσοδος στο σύστημα μία εικόνα καθώς και το σύνολο των οριοθετικών πλαισίων των αντικειμένων που υπάρχουν σε αυτή, χωρίς, ωστόσο, να δίνονται οι κλάσεις των αντικειμένων αυτών. Στόχος της εργασίας είναι η ταξινόμηση των αντικειμένων, όπως και η εύρεση και η ταξινόμηση των κατηγορημάτων που συνδέουν ζεύγη διατεταγμένων αντικειμένων.

Η επιπρόσθετη δυσκολία της εργασίας αυτής σε σχέση με την πρώτη, σε συνδυασμό με την ύπαρξη ενός επιπέδου αφαίρεσης σε σχέση με την επόμενη εργασία και το οποίο αφορά την ύπαρξη οριοθετικών πλαισίων, επιτρέπουν τη μελέτη του προβλήματος της δημιουργίας γράφων σκηνής αγνοώντας τα εγγενή και μη μειονεκτήματα που ενδέχεται να χαρακτηρίζουν τα μοντέλα, τη διαδικασία εκπαίδευσης, ή ακόμα και τη διαδικασία αξιολόγησης σχετικά με την χωροθέτηση αντικειμένων.

Η τρίτη εργασία αξιολόγησης είναι η **αναγνώριση γράφου σκηνής** (scene graph detection), η οποία αποτελεί και την κύρια εργασία στην οποία αξιολογείται ένα μοντέλο. Στην εργασία αυτή δίνεται ως είσοδος στο σύστημα μία εικόνα και στόχος είναι η εύρεση όλων των πλειάδων $\langle υποκείμενο, κατηγόρημα, αντικείμενο \rangle$ αντικείμενο που υπάρχουν σε αυτή, σε συνδυασμό με την εύστοχη χωροθέτηση των αντικειμένων και των υποκειμένων

που αποτελούν τις συσχετίσεις.

Οι είσοδοι και οι έξοδοι των τριών εργασιών αξιολόγησης παρουσιάζονται σχηματικά στο Σχήμα 4.7.



Σχήμα 4.7: Σχηματική Αναπαράσταση Εργασιών Αξιολόγησης

Στο σημείο αυτό παρουσιάζει ενδιαφέρον η ανάλυση και η αξιολόγηση ενός μοντέλου τυχαίων επιλογών στις εργασίες που παρουσιάστηκαν και με βάση τις μετρικές που επιλέχθηκαν. Καθότι έχουμε 150 κατηγορίες αντικειμένων και 50 κατηγορίες κατηγορημάτων, οι πιθανοί συνδυασμοί τριάδων συσχετίσεων ισούνται με το καρτεσιανό γινόμενο $150 \times 50 \times 150$. Υπάρχουν συνολικά, δηλαδή, 1,125,000 πιθανές τριάδες συσχετίσεων. Αξιολογώντας, λοιπόν, το μοντέλο τυχαίων προβλέψεων στην εργασία ταξινόμησης γράφου σκηνής, έχουμε τα παρακάτω αποτελέσματα:

Μετρική	**Τιμή**
$R@20$	1.77e-5
$R@50$	4.44e-5
$R@100$	8.88e-5

Πίνακας 4.1: Αξιολόγηση Μοντέλου Τυχαίων Προβλέψεων στην εργασία ταξινόμησης γράφων σκηνής

4.4 Λεπτομέρειες Υλοποίησης

Η εργασία αυτή υλοποιήθηκε σε Python 2.7 με χρήση της βιβλιοθήκης TensorFlow [1], έκδοση λογισμικού 1.5. Τόσο η εκπαίδευση όσο και η αξιολόγηση των μοντέλων έγιναν

σε δύο υπολογιστικά συστήματα το πανεπιστημιακού server της Ομάδας Κατανόησης Πολυμέσων. Τα χαρακτηριστικά των συστημάτων αυτών συνοψίζονται παρακάτω:

- 1ο Σύστημα
  - CPU: Intel® Xeon® E5335 (2.00 GHz)
  - RAM: 7,972 MiB
  - GPU: Nvidia® GeForce® GTX 780 (6,144 MiB)
- 2ο Σύστημα
  - CPU: Intel® Xeon® E5-2650 v3 (2.30 GHz)
  - RAM: 64,510 MiB
  - GPU: Nvidia® Tesla® K40c (12,288 MiB)

Κεφάλαιο 5

Αποτελέσματα

5.1 Σύνοψη Αποτελεσμάτων

Όπως αναλύεται και στην Ενότητα 4.2, η βελτιστοποίηση υπερπαραμέτρων επιτεύχθηκε μέσω ενός συνδυασμού χειροκίνητου συντονισμού υπερπαραμέτρων και τυχαίας αναζήτησης υπερπαραμέτρων [4]. Τα μοντέλα εκπαιδεύθηκαν στο σύνολο εκπαίδευσης και αξιολογήθηκαν στο σύνολο επικύρωσης. Στη συνέχεια, τα μοντέλα με τις καλύτερες επιδόσεις επανεκπαιδεύθηκαν στα σύνολα εκπαίδευσης και επικύρωσης και αξιολογούνταν ανά τακτά χρονικά διαστήματα στο σύνολο επικύρωσης. Στο σημείο όπου εμφανίστηκε η βέλτιστη απόδοση για κάθε μοντέλο έγινε η τελική αξιολόγηση στο σύνολο ελέγχου. Παρά τις διακυμάνσεις στις επιδόσεις κατά τη διαφοροποίηση των υπερπαραμέτρων, παρατηρήθηκε ευρωστία των μοντέλων που εκπαιδεύθηκαν στην επιλογή των υπερπαραμέτρων, ανάμεσα, φυσικά, σε ένα εύλογο εύρος τιμών.

Το μοντέλο της οικογένειας Gated Graph Neural Network (Υποενότητα 3.3.2) που εμφάνισε την βέλτιστη απόδοση είναι αυτό που περιγράφεται από την Εξίσωση 3.7. Οι υπερπαράμετροι που οδήγησαν στη βέλτιστη απόδοση, η οποία επιτεύχθηκε μετά από μόλις 102,871 επαναλήψεις, παρουσιάζονται συνοπτικά στη Λίστα 5.1.

Οι υπερπαράμετροι που οδήγησαν στη βέλτιστη απόδοση των συνελικτικών δικτύων γράφων παρουσιάζονται συνοπτικά στη Λίστα 5.1. Το επιθυμητό τοπικό ελάχιστο της διαδικασίας βελτιστοποίησης του δικτύου GCN επιτεύχθηκε σε 235,484 επαναλήψεις, υπερδιπλάσιες από αυτές του δικτύου GCN.

Τα μοντέλα της οικογένειας Molecular Graph Convolutions (Υποενότητα 3.3.1) δεν παρουσίασαν καθόλου καλές επιδόσεις και έτσι δεν παρουσιάζονται στον πίνακα των αποτελεσμάτων.

Ο Πίνακας 5.1 συνοψίζει τα αποτελέσματα των παραπάνω μοντέλων, σε συνδυασμό με τα μοντέλα [47] και [71]. Για τα δύο αυτά μοντέλα, παρουσιάζονται τόσο τα αρχικά αποτελέσματα, όπως αυτά δημοσιεύονται στα αντίστοιχα άρθρα, καθώς οι επανεκπαιδευμένες εκδοχές τους, σύμφωνα με την πειραματική διάταξη που ορίζεται στο Κεφάλαιο 4, για τη δικαιότερη σύγκριση των αποδόσεων.

Σημειώνεται ότι για το μοντέλο [47] παρουσιάζονται τα αποτελέσματα μόνο του οπτι-

Υπερπαράμετροι Εκπαίδευσης:

Αλγόριθμος Βελτιστοποίησης: Momentum

Παράμετροι Αλγορίθμου Βελτιστοποίησης:

momentum: 0.9

Ρυθμός Εκμάθησης: 10^{-3}

Περίοδος Κλιμακωτής Φθοράς Ρυθμού Εκμάθησης: 1 εποχή (58,871 επαναλήψεις)

Υπερπαράμετροι Δικτύου:

Πλήθος Επαναλήψεων Δικτύου Ανταλλαγής Μηνυμάτων: 1

Ισχύς Κανονικοποίησης: 0.0

Παράλειψη Απωλειών Συσχετίσεων Υποβάθρου: Αληθής

Διαστάσεις διανυσμάτων κορυφών Δικτύου Ανταλλαγής Μηνυμάτων: 512

Διαστάσεις διανυσμάτων ακμών Δικτύου Ανταλλαγής Μηνυμάτων: 512

Λίστα 5.1: Υπερπαράμετροι Βέλτιστου Νευρωνικού Δικτύου Ανταλλαγής Μηνυμάτων

Υπερπαράμετροι Εκπαίδευσης:

Αλγόριθμος Βελτιστοποίησης: Momentum

Παράμετροι Αλγορίθμου Βελτιστοποίησης:

momentum: 0.9

Ρυθμός Εκμάθησης: 10^{-3}

Περίοδος Κλιμακωτής Φθοράς Ρυθμού Εκμάθησης: 2 εποχές (117,742 επαναλήψεις)

Υπερπαράμετροι Δικτύου:

Πλήθος Επαναλήψεων Συνελικτικού Δικτύου Γράφων: 2

Ισχύς Κανονικοποίησης: 0.0

Παράλειψη Απωλειών Συσχετίσεων Υποβάθρου: Ψευδής

Διαστάσεις διανυσμάτων κορυφών & ακμών Συνελικτικού Δικτύου Γράφων: 512

Λίστα 5.2: Υπερπαράμετροι Βέλτιστου Συνελικτικού Δικτύου Γράφων

κού τμήματος και όχι του γλωσσικού, καθώς, όπως επιχειρηματολογείται και στο [71], το γλωσσικό τμήμα αποτελεί μια ορθογώνια προσέγγιση στο οπτικό και μπορεί να προστεθεί ανεξάρτητα και στο δικό μας σύστημα. Το οπτικό τμήμα του [47] ισοδυναμεί με το σύστημα το οποίο υλοποιήσαμε, χωρίς το στάδιο της ανταλλαγής μηνυμάτων.

	Predicate Classification			**Scene Graph Classification**			**Scene Graph Detection**		
Model	*R*@20	*R*@50	*R*@100	*R*@20	*R*@50	*R*@100	*R*@20	*R*@50	*R*@100
VRD‡ [47]	–	1.58	1.85	–	–	–	–	7.11	7.11
IMP [71]	–	44.75	53.08	–	21.72	24.38	–	3.44	4.24
VRD‡† [47]	18.97	29.78	37.74	6.64	9.78	11.95	0.06	0.14	0.27
IMP† [71]	29.91	44.36	53.92	14.69	19.56	22.42	1.65	2.51	3.34
GGNN	32.65	**46.83**	**55.91**	**14.98**	**19.83**	**22.65**	**1.83**	**2.75**	**3.60**
GCN	**32.94**	43.18	48.23	13.06	16.12	17.72	1.76	2.57	3.21

Πίνακας 5.1: Αποτελέσματα Αξιολόγησης στο σύνολο δεδομένων Visual Genome [39]. ‡: Οπτικό Μοντέλο, †: Υλοποίηση σύμφωνα με το [71] και επανεκπαίδευση σύμφωνα με τη δική μας πειραματική διάταξη

Όπως φανερώνει ο Πίνακας 5.1, η προσθήκη ενός δικτύου ανταλλαγής μηνυμάτων επιφέρει σημαντική βελτίωση στην απόδοση του δικτύου σε όλες τις εργασίες αξιολόγησης. Τόσο το δίκτυο GCN όσο και το δίκτυο GGNN βαθμολογούνται με δεκάδες φορές καλύτερα αποτελέσματα στην εργασία της ανίχνευσης γράφου σκηνής, ενώ η απόδοσή τους στις άλλες δύο εργασίες αξιολόγησης παρατηρείται ποσοστιαία αύξηση της τάξης του 50%. Τα στοιχεία αυτά αποτελούν περίτρανη πειραματική ένδειξη για την αποτελεσματικότητα της ανταλλαγής πληροφοριών ανάμεσα στα εξαγόμενα αντικείμενα και τις συσχετίσεις.

Ο Πίνακας 5.1 φανερώνει και τη σταθερά υψηλότερη απόδοση του δικτύου GGNN σε σχέση με το δίκτυο IMP [71] σε όλες τις εργασίες αξιολόγησης. Η διαφοροποίηση είναι εντονότερη στις εργασίες ταξινόμησης κατηγορημάτων και της ανίχνευσης γράφου σκηνής, με την ποσοστιαία διαφορά να κυμαίνεται στο 3.55-8.39% στη μεν και 7.22-9.83% στη δε. Το δίκτυο GGNN φαίνεται να υπερτερεί και του δικτύου GCN με βάση όλες σχεδόν τις μετρικές αξιολόγησης.

Η σύγκριση της απόδοσης του δικτύου GCN σε σχέση με αυτή του δικτύου IMP δε φανερώνει ξεκάθαρα ποιο μοντέλο υπερτερεί, καθώς το δίκτυο GCN παρουσιάζει τμηματικά υψηλότερες αποδόσεις. Συγκεκριμένα, το δίκτυο GCN υπερτερεί στην εργασία ταξινόμησης κατηγορημάτων με βάση τη μετρική R@20 και στην εργασία ταξινόμησης γράφου σκηνής με βάση τις μετρικές R@20 και R@50. Εντούτοις, το δίκτυο IMP παρουσιάζει σταθερά υψηλότερη απόδοση στην εργασία ταξινόμησης γράφου σκηνής.

5.1.1 Αποτελέσματα προσθήκης δικτύου διαγραφής συσχετίσεων

Το δίκτυο GGNN που παρουσίασε τις υψηλότερες αποδόσεις επανεκπαιδεύθηκε με την προσθήκη του δικτύου διαγραφής συσχετίσεων που παρουσίαζεται στην Ενότητα 3.5, με σκοπό τον έλεγχο της χρησιμότητας ενός υποτμήματος εξαγωγής συσχετίσεων έναντι της χρήσης όλως των πιθανών συσχετίσεων. Η εκπαίδευση του μοντέλου αυτού έγινε δίχως την αναζήτηση υπερπαραμέτρων. Αντ' αυτού, χρησιμοποιήθηκαν ως επί το πλείστον οι υπερπαράμετροι του αρχικού δικτύου GGNN, με την εξαίρεση της μη παράλειψης των απωλειών συσχετίσεων υποβάθρου. Οι υπερπαράμετροι αυτές παρουσιάζονται συνοπτικά στη Λίστα 5.3.

Υπερπαράμετροι Εκπαίδευσης:

Αλγόριθμος Βελτιστοποίησης: Momentum

Παράμετροι Αλγορίθμου Βελτιστοποίησης:

momentum: 0.9

Ρυθμός Εκμάθησης: 10^{-3}

Περίοδος Κλιμακωτής Φθοράς Ρυθμού Εκμάθησης: 1 εποχή (58,871 επαναλήψεις)

Υπερπαράμετροι Δικτύου:

Πλήθος Επαναλήψεων Δικτύου Ανταλλαγής Μηνυμάτων: 1

Ισχύς Κανονικοποίησης: 0.0

Παράλειψη Απωλειών Συσχετίσεων Υποβάθρου: Αληθής

Διαστάσεις διανυσμάτων κορυφών Δικτύου Ανταλλαγής Μηνυμάτων: 512

Διαστάσεις διανυσμάτων ακμών Δικτύου Ανταλλαγής Μηνυμάτων: 512

Λίστα 5.3: Υπερπαράμετροι Βέλτιστου Νευρωνικού Δικτύου Ανταλλαγής Μηνυμάτων με την προσθήκη δικτύου διαγραφής συσχετίσεων

Ο Πίνακας 5.2 συνοψίζει τα αποτελέσματα αξιολόγησης του νέου δικτύου, το οποίο συμβολίζεται ως GGNN+, συγκριτικά με την αρχική εκδοχή του, όπως παρουσιάζεται παραπάνω.

	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Model	*R*@20	*R*@50	*R*@100	*R*@20	*R*@50	*R*@100	*R*@20	*R*@50	*R*@100
GGNN	32.65	46.83	**55.91**	14.98	19.83	**22.65**	1.83	2.75	3.60
GGNN+	**44.12**	**48.71**	49.56	**19.62**	**21.01**	21.22	**3.42**	**4.61**	**5.35**

Πίνακας 5.2: Αποτελέσματα Αξιολόγησης με την προσθήκη δικτύου διαγραφής συσχετίσεων

Το δίκτυο GGNN+ φαίνεται να υπερτερεί σε όλες τις εργασίες αξιολόγησης και με βάση όλες τις μετρικές πέρα από τη μετρική R@100 στις εργασίες ταξινόμησης κατηγορημάτων και ταξινόμησης γράφου σκηνής. Ιδιαίτερα στις μετρικές R@20 παρατηρείται αρκετά μεγάλη ποσοστιαία διαφορά στην απόδοση σε όλες τις εργασίες και κατ' εξοχήν στην ανίχνευση γράφου σκηνής, όπου παρατηρείται 86% αύξηση της απόδοσης.

Κεφάλαιο 6

Συμπεράσματα & Μελλοντικές Επεκτάσεις

Η αξιολόγηση των υποσυστημάτων που υλοποιήθηκαν στα πλαίσια αυτής της εργασίας παρέχουν πολύτιμα συμπεράσματα όσον αφορά τη λειτουργία και τη χρησιμότητα της ανταλλαγής μηνυμάτων στο πρόβλημα της εξαγωγής γράφου σκηνής εικόνων. Με βάση τα συμπεράσματα αυτά αλλά και τις γνώσεις που αποκτήθηκαν στην εργασία αυτή, προτείνονται ορισμένες μελλοντικές επεκτάσεις του συστήματος που σκοπό έχουν την περαιτέρω βελτίωση της ποιότητάς του.

6.1 Συμπεράσματα

Όπως παρατηρείται από τα αποτελέσματα αξιολόγησης της προηγούμενης ενότητας, η προσθήκη ενός δικτύου ανταλλαγής μηνυμάτων βελτιώνει αισθητά την απόδοση του συστήματος σε όλες τις εργασίες αξιολόγησης. Πιο συγκεκριμένα, στις εργασίες της ταξινόμησης κατηγορημάτων και της ταξινόμησης γράφου σκηνής παρατηρείται μια σχετική αύξηση της τάξης του 50 και πλέον τοις εκατό ανάμεσα στο μοντέλο VRD [47] και το μοντέλο GGNN, ενώ στην εργασία της ανίχνευσης γράφου σκηνής παρατηρείται δεκαπλάσια σχετική αύξηση. Είναι, λοιπόν, φανερή η χρησιμότητα της ανταλλαγής πληροφοριών στην αναπαράσταση των αντικειμένων και των συσχετίσεων.

Παρόμοια αύξηση της απόδοσης παρατηρείται και στα συνελικτικά δίκτυα γράφων. Τα αποτελέσματα που προκύπτουν ενέχουν ιδιαίτερο ερευνητικό ενδιαφέρον, καθώς αν και τα δίκτυα αυτά αποτελούν άμεση μοντελοποίηση του γράφου σκηνής και η διάδοση της πληροφορίας γίνεται με βάση τα ενδογενή χαρακτηριστικά των στοιχείων του γράφου, η εφαρμογή τους δεν ενδείκνυται άμεσα για το πρόβλημα της εξαγωγής γράφου σκηνής εικόνων. Παρ' όλα αυτά, η απόδοση τους είναι αρκετά υψηλή, αν και στις περισσότερες περιπτώσεις μειωμένη σε σχέση με το δίκτυο GGNN.

Η προσθήκη ενός δικτύου διαγραφής συσχετίσεων, αν και δε μελετήθηκε εκτενώς, φαίνεται να συμβάλλει ιδιαίτερα στη βελτίωση των επιδόσεων του συστήματος. Ειδικά στην εργασία της ανίχνευσης γράφου σκηνής, που αποτελεί και την κύρια εργασία αξιολόγη-

σης, παρατηρήθηκε υψηλή αύξηση των αποδόσεων. Η αύξηση αυτή αποδίδεται στην πιο εύρωστη αξιοποίηση των συλλεγόμενων πληροφοριών των κόμβων του δικτύου ανταλλαγής μηνυμάτων, αφού η δρομολόγηση των μηνυμάτων γίνεται μέσω ενός ελεγχόμενου μηχανισμού διάδοσης και όχι προς όλες τις κατευθύνσεις.

Ακόμα και χωρίς τη χρήση του δικτύου αυτού, ωστόσο, τα μοντέλα που υλοποιήθηκαν καταφέρνουν να διαχειριστούν την πληθώρα μεταδιδόμενων μηνυμάτων και οι κόμβοι επιλέγουν τις κατάλληλες πληροφορίες για να ανανεώσουν την εσωτερική τους αναπαράσταση. Το παραπάνω μπορεί να αποδοθεί στο γεγονός ότι οι αληθείς συσχετίσεις εμπεριέχουν σχεδόν κάθε συνδυασμό αντικειμένων, παρά το ότι οι ταυτοποιημένες συσχετίσεις στο σύνολο δεδομένων είναι ελλιπείς.

6.2 Μελλοντικές Επεκτάσεις

Το σύστημα που υλοποιήθηκε στα πλαίσια της εργασίας αυτής μοντελοποιεί το πρόβλημα της εξαγωγής γράφου σκηνής με τρία διακριτά υποσυστήματα, αυτά της εξαγωγής περιοχών αντικειμένων, της εξαγωγής συσχετίσεων και της ταξινόμησης του γράφου σκηνής, δίνοντας ιδιαίτερη έμφαση στη μελέτη του υποσυστήματος της κατηγοριοποίησης του γράφου μέσω ενός δικτύου ανταλλαγής μηνυμάτων. Ωστόσο, η συνολική απόδοση του συστήματος εξαρτάται σε μεγάλο βαθμό από τις επιδόσεις των δύο αρχικών υποσυστημάτων, καθώς πιθανές ελλείψεις και σφάλματα των υποσυστημάτων αυτών διαδίδονται σε όλο το σύστημα και υποβαθμίζουν την ικανότητα συμπερασματολογίας που το διέπει.

Όσον αφορά το υποσύστημα της εξαγωγής περιοχών αντικειμένων, προτείνεται η επανεκπαίδευση των συνελικτικών επιπέδων του μοντέλου ανίχνευσης αντικειμένων. Στα πλαίσια της εργασίας αυτής, χρησιμοποιήθηκαν προεκπαιδευμένα επίπεδα, τα βάρη των οποίων παρέμειναν σταθερά καθ' όλη τη διάρκεια της εκπαίδευσης. Ως εκ τούτου, η προσθήκη αυτή θεωρείται ότι μπορεί να ενισχύσει σημαντικά τις αποδόσεις του συστήματος. Η επανεκπαίδευση των συνελικτικών επιπέδων μπορεί να γίνει ενιαία με την εκπαίδευση ολόκληρου του συστήματος. Εντούτοις, η σταδιακή εκπαίδευση των υποσυστημάτων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα και προσφέρει το πλεονέκτημα της σχολαστικότερης επίβλεψης της προόδου της εκπαίδευσης. Η σταδιακή εκπαίδευση μπορεί να υλοποιηθεί με την εκπαίδευση του μοντέλου ανίχνευσης αντικειμένων στο υπάρχον σύνολο δεδομένων, ακολουθούμενη από το στάδιο της εκπαίδευσης ολόκληρου του συστήματος, στο οποίο τα συνελικτικά επίπεδα είτε «παγώνουν» (freeze), είτε γίνεται finetuning σε αυτά.

Συμπληρωματικά, προτείνεται η χρήση ενός συστήματος ανίχνευσης αντικειμένων υψηλότερων επιδόσεων. Αυτό μπορεί να γίνει αφ' ενός με τη ενσωμάτωση ενός καλύτερου συνελικτικού δικτύου στην υπάρχουσα αρχιτεκτονική και αφ' ετέρου με τη χρήση ενός τελείως διαφορετικού συστήματος ανίχνευσης αντικειμένων. Στη εργασία αυτή χρησιμοποιήθηκε το μοντέλο Faster R-CNN [57] με το συνελικτικό δίκτυο VGG-16 [62]. Η αντικατάσταση του δικτύου VGG-16 με το δίκτυο Resnet-101 [25], σε συνδυασμό, εν-

δεχομένως, με την επανεκπαίδευση του δικτύου, είναι άξια μελέτης και θεωρείται ότι μπορεί να βελτιώσει την απόδοση των ακόλουθων υποσυστημάτων και κατ' επέκταση ολόκληρου του συστήματος.

Η σημασία του συστήματος εξαγωγής συσχετίσεων μελετήθηκε σε πολύ μικρό βαθμό στην εργασία αυτή. Στα περισσότερα πειράματα αξιολόγησης χρησιμοποιήθηκαν όλες οι πιθανές συσχετίσεις ανάμεσα στα εξαγόμενα αντικείμενα. Εντούτοις, το σύστημα διαγραφής συσχετίσεων που υλοποιήθηκε αποτελεί ισχυρή ένδειξη για τη χρησιμότητά του, σύμφωνα με τα πειραματικά αποτελέσματα. Προτείνεται, λοιπόν, η υλοποίηση ενός πιο σύνθετου συστήματος εξαγωγής συσχετίσεων. Η υλοποίηση αυτή θα μπορούσε να βασίζεται είτε στη διαγραφή ενός υποσυνόλου των συσχετίσεων, αντίστοιχα με αυτή που υλοποιήθηκε, είτε στην ανεξάρτητη πρόβλεψη των πιθανών συσχετίσεων ανάμεσα στα αντικείμενα. Σημειώνεται, ωστόσο, ότι μιας και ο χώρος των πιθανών συσχετίσεων είναι τετραγωνικής πολυπλοκότητας $\mathcal{O}(n^2)$, το σύστημα που θα υλοποιηθεί θα πρέπει να λαμβάνει υπ' όψιν τον περιορισμό αυτό, είτε χρησιμοποιώντας μικρές τιμές για την παράμετρο n, είτε μηχανεύοντας μια λύση υποτετραγωνικής πολυπλοκότητας.

Τα συνελικτικά δίκτυα γράφων εικάζεται ότι θα μπορούσαν να επωφεληθούν ιδιαίτερα από την υλοποίηση ενός τέτοιου συστήματος, αφού η αρχιτεκτονική τους στοχεύει στην επίλυση προβλημάτων ημιεπιβλεπόμενης μάθησης, στα οποία ένα τμήμα των αληθών συνδέσεων του γράφου είναι γνωστό.

Η εκπαίδευση του συστήματος έγινε σύμφωνα με τον αλγόριθμο στοχαστικής καθόδου κλίσης και δοκιμάστηκαν, μεταξύ άλλων, οι αλγόριθμοι Momentum [53], RMSProp [69] και Adam [36]. Παρ' όλα αυτά, μόνο ο αλγόριθμος Momentum κατάφερε να φτάσει σε επιθυμητά επίπεδα σύγκλισης στην πλειοψηφία των πειραμάτων. Όπως επιχειρηματολογείται στο [54], η παραπάνω αποτυχία παρατηρείται συχνά σε προβλήματα με μεγάλους χώρους εξόδου και οφείλεται στον εκθετικό κινούμενο μέσο όρο των αλγορίθμων αυτών. Προτείνεται, λοιπόν, η δοκιμή του αλγορίθμου αυτού για τη βελτιστοποίηση του συστήματος, καθώς και άλλων αλγορίθμων που αντιμετωπίζουν το παραπάνω πρόβλημα.

Όπως σημειώνεται και στην Ενότητα 4.3, η μετρική που επιλέχθηκε για την αξιολόγηση των μοντέλων είναι η μετρική R@k έναντι μιας μετρικής μέσης ακρίβειας (mAP), η οποία κρίνεται ακατάλληλη λόγω των αραιών ταυτοποιήσεων των συσχετίσεων στο σύνολο δεδομένων. Μολονότι η μετρική R@k αποδίδει μια αρκετά καλή απεικόνιση της ποιότητας των μοντέλων, πάσχει από ένα βασικό μειονέκτημα. Η μετρική R@k δεν είναι μια μετρική μέση ανάκλησης (mean average recall). Αυτό σημαίνει ότι δεν υπολογίζει ανεξάρτητα για κάθε κλάση κατηγορημάτων την ανάκληση, υπολογίζοντας στη συνέχεια τη μέση τιμή αυτών. Το γεγονός αυτό, σε συνδυασμό με τη μεγάλη ουρά της κατανομής των κατηγορημάτων, επιτρέπει στο δίκτυο να έχει πολύ χαμηλές αποδόσεις στις σπάνιες κλάσεις κατηγορημάτων, χωρίς, ωστόσο να έχει συνολικά χαμηλή απόδοση εάν προβλέπει επιτυχώς την ύπαρξη συχνών κλάσεων.

Για τους λόγους αυτούς, προτείνεται η χρήση μιας τροποποιημένης μετρικής μέσης ακρίβειας, όπως αυτή που χρησιμοποιείται στο διαγωνισμό ανίχνευσης συσχετίσεων του

συνόλου δεδομένων Open Images [38]. Η μετρική αυτή αποτελεί μια τροποποίηση της μετρικής mAP@0.5 του PASCAL VOC, στην οποία διαφοροποιείται ο ορισμός της ψευδούς θετικής πρόβλεψης ώστε να αγνοεί τις περιπτώσεις που μια προβλεπόμενη συσχέτιση δεν εμφανίζεται στο σύνολο των αληθών συσχετίσεων της εικόνας. Με τον τρόπο αυτό καταφέρνει να επιλύσει το πρόβλημα της λανθασμένης ποινής ενός μοντέλου για την πρόβλεψη μη ταυτοποιημένων συσχετίσεων, χωρίς να διέπεται από τα προαναφερθέντα προβλήματα της μετρικής R@k. Οι δύο αυτές μετρικές μπορούν φυσικά να χρησιμοποιηθούν συμπληρωματικά, ως ένα σύστημα αξιολόγησης σταθμισμένου μέσου.

Παράρτημα Α΄

Τυπολόγιο

Τύπος	**Περιγραφή**
$\sigma(\cdot)$	Λογιστική συνάρτηση (Sigmoid function)
$\rho(\cdot)$	Συνάρτηση γραμμικού ανορθωτή (Rectified Linear Unit)
$\odot$	Τελεστής πολλαπλασιασμού τανυστών κατά στοιχείο (Hadamard Product)
$\left[\cdot,\cdot\right]$	Συνένωση διανυσμάτων (Vector concatenation)
$\mathbf{w}$	Διάνυσμα Παραμέτρων (Weight Vector)
$\mathbf{W}$	Πίνακας Παραμέτρων (Weight Matrix)

Πίνακας Α΄.1: Πίνακας Μαθηματικών Τύπων & Συμβάσεων Σημειογραφίας

Παράρτημα Β΄

Λεξικό Μεταφρασμένων Αγγλικών Ορολογιών

Όρος	**Μετάφραση**
Adjacency Matrix	Πίνακας Γειτνίασης
Attribute Detection	Ανίχνευση Χαρακτηριστικών
Backpropagation	Οπισθοδιάδοση
Bilinear Interpolation	Διγραμμική Παρεμβολή
Bounding Box	Οριοθετικό Πλαίσιο
Convolutional Neural Network	Συνελικτικό Νευρωνικό Δίκτυο
Cross-entropy	Διεντροπία
Data Augmentation	Επαύξηση Δεδομένων
Degree Matrix	Πίνακας Βαθμών
Feature Map	Χάρτης Χαρακτηριστικών
Fully Connected Layer	Πλήρως Συνδεδεμένο Επίπεδο
Gated Recurrent Unit	Επαναλαμβανόμενη Μονάδα Ελεγχόμενης Πρόσβασης
Hyperparameter Tuning	Συντονισμός Υπερπαραμέτρων
Image Segmentation	Κατάτμηση Εικόνας
Object Detection	Ανίχνευση Αντικειμένων
Precision	Ακρίβεια
Recall	Ανάκληση
Rectified Linear Unit	Μονάδα Γραμμικού Ανορθωτή
Recurrent Neural Network	Επαναλαμβανόμενο Νευρωνικό Δίκτυο
Regularization	Κανονικοποίηση
Relationship Detection	Ανίχνευση Συσχετίσεων
Stochastic Gradient Descent	Στοχαστική Κάθοδος Κλίσης
Test Set	Σύνολο Ελέγχου
Training Set	Σύνολο Εκπαίδευσης
Validation Set	Σύνολο Επικύρωσης

Όρος	**Μετάφραση**
Vector Concatenation	Συνένωση Διανυσμάτων
Weight Decay	Φθορά Βαρών

Πίνακας Β΄.1: Λεξικό Μεταφρασμένων Αγγλικών Ορολογιών

Βιβλιογραφία

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: *arXiv: 1603.04467 [cs]* (Mar. 14, 2016). URL: http://arxiv.org/abs/1603.04467 (cit. on p. 34).

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. "Measuring the objectness of image windows". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2189–2202 (cit. on p. 33).

[3] Yoshua Bengio. "Practical recommendations for gradient-based training of deep architectures". In: *arXiv:1206.5533 [cs]* (June 24, 2012). URL: http://arxiv.org/abs/1206.5533.

[4] James Bergstra and Yoshua Bengio. "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13 (Feb 2012), pp. 281–305. ISSN: ISSN 1533-7928. URL: http://www.jmlr.org/papers/v13/bergstra12a.html (cit. on pp. 32, 37).

[5] Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer-Verlag, 2006. ISBN: 978-0-387-31073-2. URL: //www.springer.com/us/book/9780387310732.

[6] Xinlei Chen and Abhinav Gupta. "An Implementation of Faster RCNN with Study for Region Sampling". In: *arXiv:1702.02138 [cs]* (Feb. 7, 2017). URL: http://arxiv.org/abs/1702.02138.

[7] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. "Describing Multimedia Content using Attention-based Encoder–Decoder Networks". In: *arXiv:1507.01053 [cs]* (July 3, 2015). URL: http://arxiv.org/abs/1507.01053.

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *arXiv:1406.1078 [cs, stat]* (June 3, 2014). URL: http://arxiv.org/abs/1406.1078 (visited on 10/16/2018) (cit. on pp. 9, 17).

[9] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. "Gated Feedback Recurrent Neural Networks". In: *arXiv:1502.02367 [cs, stat]* (Feb. 9, 2015). URL: http://arxiv.org/abs/1502.02367.

[10] Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting Visual Relationships with Deep Relational Networks". In: *arXiv:1704.03114 [cs]* (Apr. 10, 2017). URL: http://arxiv.org/abs/1704.03114 (cit. on p. 8).

[11] Jifeng Dai, Kaiming He, and Jian Sun. "Instance-aware Semantic Segmentation via Multi-task Network Cascades". In: *arXiv:1512.04412 [cs]* (Dec. 14, 2015). URL: http://arxiv.org/abs/1512.04412 (cit. on p. 8).

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. "R-FCN: Object Detection via Region-based Fully Convolutional Networks". In: *arXiv:1605.06409 [cs]* (May 20, 2016). URL: http://arxiv.org/abs/1605.06409 (cit. on p. 8).

[13] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. "Language Modeling with Gated Convolutional Networks". In: *arXiv:1612.08083 [cs]* (Dec. 23, 2016). URL: http://arxiv.org/abs/1612.08083.

[14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *arXiv:1606.09375 [cs, stat]* (June 30, 2016). URL: http://arxiv.org/abs/1606.09375.

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes Challenge: A Retrospective". In: *International Journal of Computer Vision* 111.1 (Jan. 2015), pp. 98–136 (cit. on pp. 1, 32).

[16] Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. "A Convolutional Encoder Model for Neural Machine Translation". In: *arXiv:1611.02344 [cs]* (Nov. 7, 2016). URL: http://arxiv.org/abs/1611.02344.

[17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. "Convolutional Sequence to Sequence Learning". In: *arXiv:1705.03122 [cs]* (May 8, 2017). URL: http://arxiv.org/abs/1705.03122.

[18] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. "Neural Message Passing for Quantum Chemistry". In: *arXiv:1704.01212 [cs]* (Apr. 4, 2017). URL: http://arxiv.org/abs/1704.01212 (cit. on p. 15).

[19] Ross Girshick. "Fast R-CNN". In: *arXiv:1504.08083 [cs]* (Apr. 30, 2015). URL: http://arxiv.org/abs/1504.08083 (cit. on p. 7).

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *arXiv:1311.2524 [cs]* (Nov. 11, 2013). URL: http://arxiv.org/abs/1311.2524 (cit. on pp. 7, 8, 29).

[21] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Mar. 31, 2010, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html (cit. on p. 29).

[22] Rafael C. Gonzalez and Richard E. Woods. *Ψηφιακή επεξεργασία εικόνας*. Ed. by Στέφανος Κόλλιας. Trans. by Αθανάσιος Ι. Μάργαρης. 3rd ed. Τζιόλα, 2011. ISBN: 978-960-418-255-8.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN". In: *arXiv:1703.06870 [cs]* (Mar. 20, 2017). URL: http://arxiv.org/abs/1703.06870 (cit. on pp. 8, 22).

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *arXiv:1512.03385 [cs]* (Dec. 10, 2015). URL: http://arxiv.org/abs/1512.03385 (cit. on pp. 7, 44).

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *arXiv:1502.01852 [cs]* (Feb. 6, 2015). URL: http://arxiv.org/abs/1502.01852 (cit. on p. 29).

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity Mappings in Deep Residual Networks". In: *arXiv:1603.05027 [cs]* (Mar. 16, 2016). URL: http://arxiv.org/abs/1603.05027 (cit. on p. 7).

[28] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. "Speed/accuracy trade-offs for modern convolutional object detectors". In: *arXiv:1611.10012 [cs]* (Nov. 30, 2016). URL: http://arxiv.org/abs/1611.10012.

[29] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *arXiv:1502.03167 [cs]* (Feb. 10, 2015). URL: http://arxiv.org/abs/1502.03167.

[30] Justin Johnson, Agrim Gupta, and Li Fei-Fei. "Image Generation from Scene Graphs". In: *arXiv:1804.01622 [cs]* (Apr. 4, 2018). URL: http://arxiv.org/abs/1804.01622 (cit. on p. 9).

[31] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: (Dec. 20, 2016). URL: https://arxiv.org/abs/1612.06890 (cit. on p. 2).

[32] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "DenseCap: Fully Convolutional Localization Networks for Dense Captioning". In: *arXiv:1511.07571 [cs]* (Nov. 24, 2015). URL: http://arxiv.org/abs/1511.07571.

[33] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. "Image retrieval using scene graphs". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3668–3678. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7298990 (cit. on pp. 2, 9).

[34] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137. URL: http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.html.

[35] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. "Molecular graph convolutions: moving beyond fingerprints". In: *Journal of computer-aided molecular design* 30.8 (2016), pp. 595–608 (cit. on p. 16).

[36] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]* (Dec. 22, 2014). URL: http://arxiv.org/abs/1412.6980 (cit. on pp. 30, 45).

[37] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *arXiv:1609.02907 [cs, stat]* (Sept. 9, 2016). URL: http://arxiv.org/abs/1609.02907 (cit. on p. 17).

[38] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, et al. "OpenImages: A public dataset for large-scale multi-label and multi-class image classification." In: (2017) (cit. on pp. 2, 46).

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *arXiv:1602.07332 [cs]* (Feb. 23, 2016). URL: http://arxiv.org/abs/1602.07332 (cit. on pp. 2, 10–12, 21, 39).

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (cit. on p. 7).

[41] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. "Fully Convolutional Instance-aware Semantic Segmentation". In: *arXiv:1611.07709 [cs]* (Nov. 23, 2016). URL: http://arxiv.org/abs/1611.07709 (cit. on p. 8).

[42] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. "Scene Graph Generation from Objects, Phrases and Caption Regions". In: *arXiv:1707.09700 [cs]* (July 30, 2017). URL: http://arxiv.org/abs/1707.09700 (cit. on p. 8).

[43] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. "Gated graph sequence neural networks". In: *arXiv preprint arXiv:1511.05493* (2015) (cit. on p. 16).

[44] Xiaodan Liang, Lisa Lee, and Eric P. Xing. "Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection". In: *arXiv:1703.03054 [cs]* (Mar. 8, 2017). URL: http://arxiv.org/abs/1703.03054 (cit. on p. 8).

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. "Microsoft COCO: Common Objects in Context". In: *arXiv:1405.0312 [cs]* (May 1, 2014). URL: http://arxiv.org/abs/1405.0312 (cit. on pp. 1, 10, 32).

[46] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "SSD: Single Shot MultiBox Detector". In: (Dec. 8, 2015). DOI: 10.1007/978-3-319-46448-0_2. URL: https://arxiv.org/abs/1512.02325 (visited on 10/11/2018) (cit. on p. 7).

[47] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. "Visual Relationship Detection with Language Priors". In: *arXiv:1608.00187 [cs]* (July 31, 2016). URL: http://arxiv.org/abs/1608.00187 (cit. on pp. 8, 32, 33, 37, 39, 43).

[48] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) (cit. on p. 9).

[49] 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: http://doi.acm.org/10.1145/219717.219748 (cit. on p. 10).

[50] Yurii Nesterov. "A Method for Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27 (1983), pp. 372–376 (cit. on p. 30).

[51] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004. ISBN: 978-1-4020-7553-7. URL: //www.springer.com/us/book/9781402075537 (cit. on p. 30).

[52] Alejandro Newell and Jia Deng. "Pixels to Graphs by Associative Embedding". In: *arXiv:1706.07365 [cs]* (June 22, 2017). URL: http://arxiv.org/abs/1706.07365 (cit. on p. 8).

[53] B. T. Polyak. "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17. ISSN: 0041-5553. DOI: https://doi.org/10.1016/0041-5553(64)90137-5. URL: http://www.sciencedirect.com/science/article/pii/0041555364901375 (cit. on pp. 30, 45).

[54] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. "On the convergence of adam and beyond". In: (2018) (cit. on p. 45).

[55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *arXiv:1506.02640 [cs]* (June 8, 2015). URL: http://arxiv.org/abs/1506.02640 (cit. on p. 7).

[56] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: *arXiv:1612.08242 [cs]* (Dec. 25, 2016). URL: http://arxiv.org/abs/1612.08242 (visited on 10/11/2018) (cit. on p. 7).

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *arXiv:1506.01497 [cs]* (June 4, 2015). URL: http://arxiv.org/abs/1506.01497 (cit. on pp. 7, 22, 44).

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cit. on p. 1).

[59] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. "Modeling Relational Data with Graph Convolutional Networks". In: *arXiv:1703.06103 [cs, stat]* (Mar. 17, 2017). URL: http://arxiv.org/abs/1703.06103.

[60] 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: *Proceedings of the Fourth Workshop on Vision and Language*. 2015, pp. 70–80. URL: http://www-nlp.stanford.edu/pubs/schuster-krishna-chang-feifei-manning-vl15.pdf.

[61] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. "Training Region-based Object Detectors with Online Hard Example Mining". In: *arXiv:1604.03540 [cs]* (Apr. 12, 2016). URL: http://arxiv.org/abs/1604.03540.

[62] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv:1409.1556 [cs]* (Sept. 4, 2014). URL: http://arxiv.org/abs/1409.1556 (cit. on pp. 7, 8, 22, 44).

[63] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://www.jmlr.org/papers/v15/srivastava14a.html (cit. on p. 31).

[64] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning". In: *International Conference on Machine Learning*. International Conference on Machine Learning. Feb. 13, 2013, pp. 1139–1147. URL: http://proceedings.mlr.press/v28/sutskever13.html (cit. on p. 30).

[65] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper with Convolutions". In: *arXiv:1409.4842 [cs]* (Sept. 16, 2014). URL: http://arxiv.org/abs/1409.4842 (cit. on p. 7).

[66] Damien Teney, Lingqiao Liu, and Anton van den Hengel. "Graph-Structured Representations for Visual Question Answering". In: *arXiv:1609.05600 [cs]* (Sept. 19, 2016). URL: http://arxiv.org/abs/1609.05600 (visited on 10/18/2018) (cit. on p. 9).

[67] Sergios Theodoridis and Konstantinos Koutroumbas. *Αναγνώριση Προτύπων*. Ed. by Άγγελος Πικράκης, Κωνσταντίνος Κουτρούμπας, and Θεόδωρος Γιαννακόπουλος. 4th ed. Π.Χ. ΠΑΣΧΑΛΙΔΗΣ, 2012. ISBN: 978-960-489-145-0.

[68] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. "YFCC100M: The New Data in Multimedia Research". In: *Communications of the ACM* 59.2 (Jan. 25, 2016), pp. 64–73. ISSN: 00010782. DOI: 10.1145/2812802. URL: http://arxiv.org/abs/1503.01817 (cit. on p. 10).

[69] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2 (2012), pp. 26–31 (cit. on pp. 30, 45).

[70] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. "Selective search for object recognition". In: *International journal of computer vision* 104.2 (2013), pp. 154–171 (cit. on p. 7).

[71] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. "Scene Graph Generation by Iterative Message Passing". In: *arXiv:1701.02426 [cs]* (Jan. 9, 2017). URL: http://arxiv.org/abs/1701.02426 (cit. on pp. 8, 9, 33, 37, 39).

[72] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. "Embedding Entities and Relations for Learning and Inference in Knowledge Bases". In: *arXiv:1412.6575 [cs]* (Dec. 19, 2014). URL: http://arxiv.org/abs/1412.6575.

[73] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *arXiv:1311.2901 [cs]* (Nov. 12, 2013). URL: http://arxiv.org/abs/1311.2901 (cit. on p. 7).