Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας σηματων, ελεγχου και ρομποτικης

Out of Distribution generalization methods for Visual Question Answering.

Διπλωματικη Εργασια

του

ΖΕΡΒΑ Δ. ΝΙΚΟΛΑΟΥ

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας σηματων, ελεγχου και ρομποτικης

Out of Distribution generalization methods for Visual Question Answering.

Διπλωματικη Εργασια

του

ΖΕΡΒΑ Δ. ΝΙΚΟΛΑΟΥ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13 Φεβρουαρίου 2024.

.....	.....	.....
Αλέξανδρος Ποταμιάνος	Κωνσταντίνος Τζαφέστας	Στέφανος Κόλλιας
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας σηματων, ελεγχου και ρομποτικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

.....
Ζέρβας Δ. Νικόλαος
Διπλωματούχος
Ηλεκτρολόγος Μηχανικός
και Μηχανικός
Υπολογιστών Ε.Μ.Π.

10 Οκτωβρίου 2022

Περίληψη

Η Απάντηση σε Οπτικές Ερωτήσεις (VQA) βρίσκεται στο προσκήνιο της προαγωγής της Γενικής Τεχνητής Νοημοσύνης (AGI), συνδυάζοντας τον τομέα της υπολογιστικής όρασης με την επεξεργασία φυσικής γλώσσας.

Τα τρέχοντα μοντέλα στο VQA επιτυγχάνουν υψηλές επιδόσεις σε κλασικά σύνολα δεδομένων, αλλά συχνά περιορίζονται από την εξάρτησή τους στις συσχετίσεις της γλώσσας στα δεδομένα εκπαίδευσης. Συχνά απαντούν χωρίς να λαμβάνουν υπόψη τις εικόνες, με αποτέλεσμα να αποτυγχάνουν σε ποικίλα περιβάλλοντα δοκιμών. Αυτή η διατριβή αντιμετωπίζει αυτές τις προκλήσεις, εστιάζοντας στη γενίκευση στη VQA, ιδιαίτερα σε σενάρια εκτός κατανομής.

Η διπλωματική εργασία ξεκινά από θεμελιώδεις έννοιες της μηχανικής μάθησης και στη συνέχεια, διεξάγει μια σφαιρική βιβλιογραφική ανασκόπηση του τομέα γενίκευσης στο αντικείμενο του VQA, με στόχο την κατανόηση των διάφορων μεθόδων γενίκευσης σε δεδομένα εκτός κατανομής και επανεκτελέσεις καινοτόμων μεθόδων. Αναφέρουμε ορισμένα ευρήματα και συμπεράσματα βασισμένα στα αποτελέσματα των μεθόδων στα σύνολα δεδομένων GQA OOD και VQA-CPv2.

Ακολουθούν, αρχικά πειράματα στη δημιουργία οπτικών ερωτήσεων ως τεχνική επαύξησης δεδομένων και άναλυση των αποτελεσμάτων.

Το κύριο αντικείμενο αυτής της εργασίας είναι η ανάπτυξη μιας νέας μεθοδολογίας μάσκας αντικειμένων εικόνας, που διαφέρει από τις παραδοσιακές προσεγγίσεις. Οι προσαρμοσμένες μέθοδοι μας βασίζονται στον εντοπισμό σημαντικών αντικειμένων μέσω καλύψεων και στη δημιουργία θετικών και αρνητικών τριάδων Εικόνας-Ερώτησης. Χρησιμοποιείται μια συνάρτηση κόστους τριπλών απωλειών, η οποία πλησιάζει τις πολυδιάστατες αναπαραστάσεις των πραγματικών δειγμάτων πιο κοντά στα θετικά δείγματα και μακριά από τα αρνητικά. Επιπλέον, χρησιμοποιήσαμε μια συνάρτηση κόστους επαύξησης δεδομένων με θετικά δείγματα. Τέλος, πειραματιστήκαμε με μια τυχαία μέθοδο μάσκας που έδειξε σημαντικές βελτιώσεις στην απόδοση, σε συνδυασμό με την αρχική μας μεθοδολογία.

Τα προτεινόμενα μοντέλα μας συνδυάζοντας τις αναφερθείσες μεθοδολογίες οδηγούν σε σημαντικές βελτιώσεις σε συνθήκες εντός και εκτός κατανομής στο σύνολο δεδομένων GQA OOD.

Συνοψίζοντας, αυτή η διατριβή περιλαμβάνει τις νέες συνεισφορές μας στον τομέα του VQA, αναλύοντας τα κύρια ευρήματά μας και προτείνοντας κατευθύνσεις για μελλοντική έρευνα για να βελτιώσουν περαιτέρω τις δυνατότητες γενίκευσης των μοντέλων VQA.

Λέξεις Κλειδιά

Απάντηση σε Οπτικές Ερωτήσεις, Κάλυψη οπτικών αντικειμένων, Γενίκευση, Δεδομένα εκτός κατανομής, Συνάρτηση κόστους τριπλών απωλειών, Τεχνικές επαύξησης δεδομένων

Abstract

Visual Question Answering (VQA) stands at the forefront of advancing General AI, blending visual perception with linguistic analysis. VQA requires a deep understanding of visual content and natural language queries, demanding an advanced level of object, scene, and activity recognition and contextual understanding.

Despite the progress in VQA, current models often struggle with out-of-distribution conditions, relying heavily on spurious correlations and language biases. This thesis addresses these challenges by focusing on generalization in VQA, particularly in out-of-distribution scenarios.

The thesis is structured to gradually build the reader's of fundamental machine learning concepts and afterwards, we conduct a comprehensive survey of the existing methodologies of generalization in VQA and reimplement several key approaches on the out-of-distribution datasets VQACPv2 GQA OOD. We report certain findings and conclusions based on our survey the entire field and our reimplementations.

Our initial experiments include an augmentation strategy using question generation based on image and answer features. Our experiments in visual question generation showcase that trying to perturbate the question without changing the answers resulted in suboptimal performance and we propose an alternative augmentation strategy for constructing new question-answer pairs.

The core contribution of this thesis is developing a novel image object masking methodology that diverges from traditional approaches. Our custom masking methods are based on identifying important objects by leveraging annotations in our dataset and using masking to construct positive and negative Image-Question tuples. It leverages a triplet contrastive loss function responsible for pulling the multimodal representations of the real samples closer to the positive samples and away from the negative ones. Additionally, we leveraged an augmentation loss using only the positive samples. Lastly, we experimented with a random masking approach that showcased significant performance improvements paired with our initial methodology. Our proposed models combining the mentioned methodologies lead to significant performance improvements under out-of-distribution conditions in the GQA OOD dataset.

In summary, this thesis encapsulates our novel contributions to the VQA field, detailing our primary findings and proposing directions for future research to improve the generalization capabilities of VQA models further.

Keywords

Visual Question Answering (VQA), Masking of visual objects, Out-of-distribution data, Triplet loss function, Data Augmentation

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Αλέξανδρο Ποταμιάνο για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο SLP-NTUA.

Επίσης ευχαριστώ ιδιαίτερα τον Υποψήφιο Διδάκτορα Γιώργο Παρασκευόπουλο για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε.

Ευχαριστώ επιπλέον τους καθηγητές κ. Τζαφέστα και κ. Κόλλια που ανταποκρίθηκαν με προθυμία να συμμετάσχουν στην τριμελή επιτροπή.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την καθοδήγηση και την στήριξη που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Ιούνιος 2024

Ζέρβας Δ. Νικόλαος

Περιεχόμενα

Κεφάλαιο 0

Εκτεταμένη Ελληνική Περίληψη

0.1 Εισαγωγή

0.1.1 Κίνητρο

Η Απάντηση Ερωτήσεων για Εικόνες (VQA), δηλαδή η απάντηση ερωτήσεων σχετικά με το περιεχόμενο μιας εικόνας, είναι ένας από τους σημαντικότερους τομείς επεξεργασίας εικόνας-γλώσσας και μία κεντρική μέθοδος προς την πορεία για τη Γενική Τεχνητή Νοημοσύνη. Παρόλο που τα σημαντικότερα μοντέλα μπορούν να επιτύχουν καλά αποτελέσματα στα VQA benchmarks όπως το VQA v2, συχνά βασίζονται σε συγκεκριμένα biases του συνόλου δεδομένων τους και κατά συνέπεια υπεραποδίδουν στον τομέα I.D, αλλά 'υποαποδίδουν" σε διαφορετικές συνθήκες εξέτασης.

Τα παραπάνω ενισχύονται από τις γλωσσικές προκαταλήψεις που υπάρχουν σε διάφορα VQA σύνολα δεδομένων και την έλλειψη κατάλληλων μετρικών για τη μέτρηση της απόδοσης. Επειδή η γλώσσα είναι ευκολότερη στην επεξεργασία και συχνά πιο σημαντική για τις δοκιμασίες ερωτήσεων και απαντήσεων, οι περισσότερες κορυφαίες μέθοδοι VQA τείνουν να εξαρτώνται υπερβολικά από αυτές τις γλωσσικές προκαταλήψεις και να εκμεταλλεύονται συντομίες για να επιτύχουν καλύτερη απόδοση, με αποτέλεσμα την ανεπαρκή οπτική αντίληψη. Για παράδειγμα, δείχνοντας μια εικόνα από πράσινες άγουρες μπανάνες, στην ερώτηση "Τι χρώμα είναι οι μπανάνες·",τα περισσότερα μοντέλα θα απαντήσουν "κίτρινο" χωρίς να εστιάζουν στις κατάλληλες περιοχές της εικόνας, επειδή είναι η πιο συχνή απάντηση σχετικά με αυτή την συγκεκριμένη ερώτηση στο σετ εκπαίδευσης. Κατά συνέπεια, το μοντέλο θεωρητικά θα επιτυγχάνει πολύ υψηλή ακρίβεια παρά την κακή οπτική κατανόηση του μοντέλου.

Σχήμα 1: *Κοινή λάθος απάντηση μοντέλων υπερβολικά εξαρτημένων από τη γλώσσα.*

Με αυτό το κίνητρο, αποφασίσαμε να εμβαθύνουμε στις μεθόδους γενίκευσης που χρησιμοποιούνται στο VQA και να προτείνουμε τις δικές μας μεθοδολογίες για τη βελτίωση της απόδοσης σε συνθήκες εκτός κατανομής (out-of-distribution).

0.1.2 Συνεισφορές

Αυτή η διπλωματική εργασία προσφέρει σημαντικές συνεισφορές στον τομέα της Οπτικής Απάντησης Ερωτήσεων (VQA). Αρχικά, διεξήχθη μια εκτενής έρευνα της σχετικής βιβλιογραφίας, επικεντρωμένη στις υπάρχουσες μεθοδολογίες γενίκευσης για VQA.

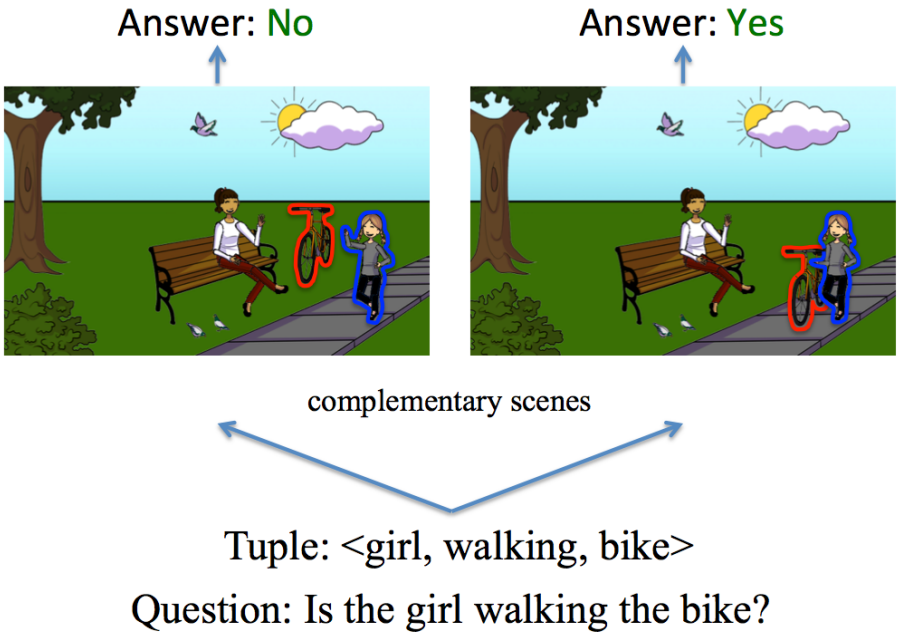
Πιο συγκεκριμένα, πραγματοποιήσαμε μια συνολική ανάλυση της σχετικής βιβλιογραφίας και επανεκτελέσαμε αρκετές σημαντικές μεθόδους που αναφέρονται σε αυτήν, παρέχοντας ένα σταθερό θεμέλιο για την πειραματική μας ανάλυση, αναλύοντας, παράλληλα, τα πλεονεκτήματα και τα μειονεκτήματα των κυρίαρχων τεχνικών γενίκευσης στην απάντηση οπτικών ερωτήσεων.

Αυτό αποτέλεσε τη βάση για την ανάπτυξη μιας καινοτόμου μεθοδολογίας ¨απόκρυψης¨ (masking) αντικειμένων εικόνας. Αυτή η νέα προσέγγιση αποκλίνει από τις παραδοσιακές μεθόδους, ενσωματώνοντας ευφυή και τυχαία στοιχεία, ενισχύοντας την ανθεκτικότητα και τις δυνατότητες γενίκευσης του μοντέλου. Η πειραματική μας εργασία με αυτή τη μεθοδολογία επέδειξε σημαντικές βελτιώσεις στην απόδοση του μοντέλου σε συνθήκες εκτός κατανομής στο GQA OOD dataset. Ένα σημαντικό στοιχείο αυτής της μεθοδολογίας είναι η χρήση της προσέγγισης triplet loss, η οποία κατασκευάζει ένα τριπλέτο πραγματικών, θετικών και αρνητικών δειγμάτων. Με την μεγιστοποίηση/ελαχιστοποίηση της αμοιβαίας πληροφορίας μεταξύ των πραγματικών και θετικών/αρνητικών δειγμάτων αντίστοιχα, παρέχουμε ένα σήμα καθοδηγησης του μοντέλου στην ανεύρεση σημαντικών οπτικών πληροφοριών. Επιπλέον, χρησιμοποιείται μια απώλεια ενίσχυσης, βελτιώνοντας περαιτέρω την απόδοση του μοντέλου. Αυτές οι συνεισφορές συνολικά προάγουν την κατανόηση και την αποτελεσματικότητα των συστημάτων VQA, ιδιαίτερα την γενίκευση τους σε περιβάλλοντα εκτός κατανομής.

0.2 Απάντηση Ερωτήσεων Πάνω σε Εικόνες.

Η Απάντηση Ερωτήσεων Πάνω σε Εικόνες (VQA) αποτελεί έναν τομέα της τεχνητής νοημοσύνης (AI), που ο στόχος είναι η ανάπτυξη συστημάτων που είναι ικανά να απαντούν σε ερωτήσεις σχετικά με το περιεχόμενο εικόνων. Πρόκειται για ένα περίπλοκο task που συνδυάζει την υπολογιστική όραση και την επεξεργασία φυσικής γλώσσας, απαιτώντας από τις μηχανές να αναγνωρίζουν όχι μόνο τα στοιχεία εντός μιας εικόνας αλλά και να κατανοούν τα χαρακτηριστικά τους και τις επιμέρους σχέσεις μεταξύ τους.

Για παράδειγμα, στο Σχήμα 2 μπορούμε να δούμε διαφορετικές απαντήσεις να δίνονται για την ίδια ερώτηση και για παρόμοιες αλλά διαφορετικές εικόνες. Για να μπορέσει το μοντέλο να συνάγει την σωστή απάντηση και στις δύο περιπτώσεις, πρέπει να μην αναγνωρίζει μόνο με ακρίβεια τα αντικείμενα αλλά και να συνάγει, μέσω των σχετικών θέσεών τους και της τοποθέτησης του χεριού του κοριτσιού, ότι η σχέση τους είναι πραγματικά ‘walking¨, το οποίο μεταφορικά χρησιμοποιείται για την ολίσθηση του ποδηλάτου στο δρόμο.



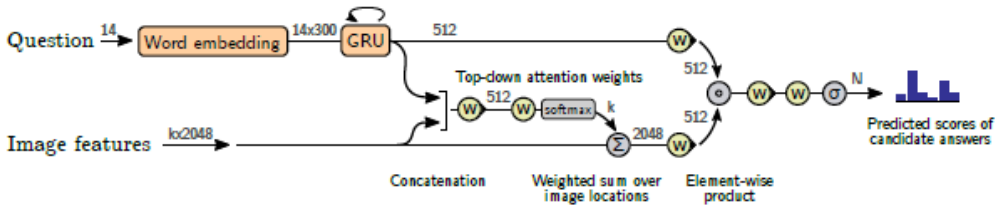
Σχήμα 2: *Ένα παράδειγμα οπτικής απάντησης ερωτήσεων χρησιμοποιώντας την ίδια ερώτηση για διαφορετικές εικόνες. Το μοντέλο αξιοποιεί τη σχετική θέση του ποδηλάτου σε σύγκριση με το κορίτσι για να συνάγει δύο διαφορετικές απαντήσεις. Προσαρμοσμένο από* https://παπερσωιτησοδε.σομ/τασκ/ισυαλ-χυεστιον-ανσωερινγ.

0.3 Βασικό Μοντέλο - Bottom Up Top Down Attention

Ένα από τα σημαντικότερα μοντέλα του πεδίου του VQA είναι το Bottom Up Top Down Attention το οποίο βασίζεται στην ιδέα του bottom-up attention και top-down attention .

Το μοντέλο UpDn αξιοποιεί τους μηχανισμούς bottom-up και top-down για να ενσωματώσει αποτελεσματικά οπτικά και κειμενικά χαρακτηριστικά. Ο μηχανισμός bottom-up επικεντρώνεται στην κατάτμηση της εικόνας σε οπτικά αντικείμενα-περιοχές και στην εξαγωγή σημαντικών χαρακτηριστικών για αυτά, χρησιμοποιώντας το μοντέλο Faster-RCNN , ενώ ο top-down μηχανισμός προσοχής συνδυάζει τα περιεχόμενα των οπτικών αντικειμένων με την αναπαράσταση της εικόνας, εξαγόμενη από ένα RNN για να καταλήξει στην τελική απάντηση.

Το τελικό αποτέλεσμα του δικτύου, φαίνεται στο Σχήμα 3, Επιλέξαμε το παραπάνω μοντέλο λόγω της ικανότητας του να γενικεύει ευρέως σε task εικόνας και γλώσσας.



Σχήμα 3: *Bottom Up- Top Down attention architecture.*

0.4 Σύνολα Δεδομένων γενίκευσης.

Στο πλαίσιο της διπλωματικής εργασίας θα ασχοληθούμε με τα δύο κυριότερα σύνολα δεδομένων για τεχνικές γενίκευσης στο VQA.

Το σύνολο δεδομένων **VQA-CPv2** αποτελεί μια παραλλαγή του VQA v2 dataset, που δημιουργήθηκε για να αντιμετωπίσει την υπερβολική εξάρτηση από τη γλώσσα που παρατηρήθηκε στο αρχικό VQA dataset. Σε αυτό το σύνολο δεδομένων, κάθε ερώτηση συνδέεται με ένα ζευγάρι παρόμοιων εικόνων που οδηγούν σε διαφορετικές απαντήσεις σε train και test set. Οι βασικές μετρικές είναι το accuracy σε 3 βασικές κατηγορίες ερωτήσεων ‘Yes/No¨, ‘Number¨, ‘Other¨. Περιέχει εύρος ερωτήσεων για πολλαπλές κατηγορίες και σχέσεις μεταξύ αντικειμένων.

Το παραπάνω σύνολο δεδομένων έχει δεχθεί κριτική στην βιβλιογραφία [1, 2, 3] και οι επανυλοποιήσεις που πραγματοποιήσαμε σε παρόμοια συμπεράσματα. Επομένως, για την κύρια μεθοδολογία μας βασιστήκαμε στο παρακάτω σύνολο δεδομένων.

Το σύνολο δεδομένων **GQA OOD** είναι μια εξέλιξη του συνόλου δεδομένων GQA, σχεδιασμένο για την αξιολόγηση οπτικής θεμελίωσης και συλλογιστικής σε πραγματικά σενάρια. Αυτό το σύνολο δεδομένων δοκιμάζει τα μοντέλα VQA σε συνθήκες εντός και εκτός κατανομής (ID και OOD), περιλαμβάνοντας διαχωρισμούς για επικύρωση σε αμφότερες συνθήκες. Περιλαμβάνει επίσης, επιπρόσθετες πληροφορίες για τις εικόνες (scene graphs) και οι ερωτήσεις είναι γραμμένες και σε λογική προγραμμάτων.

Η κύρια κατανομή του συνόλου δεδομένων διακρίνεται σε δύο τμήματα:

Κεφαλή της Κατανομής: Αναφέρεται σε δειγμάτα συχνών απαντήσεων και αντιπροσωπεύουν τις συνήθεις περιπτώσεις (ID).

Ουρά της Κατανομής: Περιέχει τις σπανιότερες απαντήσεις, αντιπροσωπεύοντας ασυνήθιστες ή εκτός κατανομής περιπτώσεις (OOD).

Εισάγονται νέες μετρήσεις απόδοσης για την ακρίβεια σε συνθήκες ID (Acc-head) και OOD (Acc-tail), καθώς και μια μετρική (Δ) που δείχνει τη διαφορά απόδοσης μεταξύ αυτών των δύο σεναρίων. Συγκεκριμένα, $\Delta = \frac{Acc_{Head} - Acc_{Tail}}{Acc_{head}}$.

Η συνάρτηση κόστους που χρησιμοιείται ευρέως στη βιβλιογραφία για τα σύνολα δεδομένων αυτά, ονομάζεται Binary Cross Enropy Loss και αντιπροσωπεύει την κατανομή πιθανότητας πάνω στις πιθανές απαντήσεις και ανατίθεται σε μια συνάρτηση απώλειας δυαδικής διασταυρωμένης εντροπίας. Η συνάρτηση απώλειας διαμορφώνεται ως εξής:

$$L_{\text{BCE}} = -\sum_{i=1}^{C} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{0.1}$$

όπου C είναι ο αριθμός των τάξεων απάντησης, y_i είναι η πραγματική ετικέτα για την τάξη i, και p_i είναι η προβλεπόμενη πιθανότητα για την τάξη i από το μοντέλο. Αυτή η συνάρτηση απώλειας υπολογίζει τη διασταυρωμένη εντροπία μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών ετικετών, τιμωρώντας αποτελεσματικά τις προβλέψεις που αποκλίνουν από τις πραγματικές ετικέτες.

0.5 Σχετική βιβλιογραφία

Η βιβλιογραφία σε σχέση με την γενίκευση σε VQA μπορεί να κατανεμηθεί σε 4 κύριες κατηγορίες . Στο πλαίσιο της διπλωματικής μου θα ασχοληθούμε κατά κύριο λόγο με τις παρακάτω:

Language debiasing ensemble based methods: Οι μέθοδοι της κατηγορίας αυτής βασίζονται στην ιδέα ότι μέσω ενός μοντέλου το οποίο βλέπει μόνο γλώσσα, ωθούμε τα VQA μοντέλα που παίρνουν σαν είσοδο γλώσσα και εικόνα να προβλέπει διαφορετικές απαντήσεις με την ελπίδα να αποσυμπλεχθεί το μοντέλο από την υπερβολική εξάρτηση στην γλώσσα (ερώτηση). Οι σημαντικότερες μέθοδοι αυτής της κατηγορίας είναι LMH, RuBi, AdvReg, Οι μέθοδοι αυτές οδηγούν σε πολύ υψηλή αύξηση της ακρίβειας στο dataset VQA-CPv2 αλλά έχουν δεχθεί σφοδρή κριτική [2, 1, 4, 3].

Μέθοδοι επάυξησης δεδομένων (augmentation): Οι μέθοδοι αυτές ποικίλουν. Κάποιες εξ αυτών βασίζονται σε τεχνικές για perturbations, οι οποίες όμως φαίνεται να οδηγούν σε καλύτερα αποτελέσματα σε συγκεκριμένες συνθήκες [5, 6, 7], ενώ άλλες που θα αναλυθούν στην συνέχεια, βασίζονται στην κατασκευή αντιθετικών (counterfactual) δειγμάτων [8, 3, 9]. Τέλος, η μέθοδος του Mutant [10] βασίζεται στην παραλλαγή των υπάρχοντων δεδομένων με ποικίλες τεχνικές οι οποίες, όμως, είναι συγκεκριμένες αναφορικά με το dataset VQACPv2, και άρα είναι δύσκολο να εφαρμοστούν σε άλλα σύνολα δεδομένων.

Answer Reranking: Οι μέθοδοι αυτές [11, 12] ασχολούνται κατά κύριο λόγο με το συνδυασμό της εικόνας και απάντησης σε μια συνολική γλωσσική περιγραφή. Η αναδιαμόρφωση της γλωσσικής πληροφορίας οδηγεί σε καλύτερα αποτελέσματα, αλλά τα μοντέλα αυτά είναι περιορισμένα λόγω του αυξημένου υπολογιστικού κόστους.

0.6 Επανεκτελέσεις μεθόδων και Σχολιασμός

Το κεφάλαιο περιλαμβάνει τα πρώτα πειράματα μας σχετικά με την επανυλοποίηση διάφορων ερευνητικών εργασιών για τη βελτίωση της γενίκευσης στο σύνολο δεδομένων VQACPv2. Έχοντας δεχθεί έντονη κριτική για την εκμετάλλευση προκαταλήψεων στο σύνολο δεδομένων VQACPv2, αυτές οι μέθοδοι θα επαναυλοποιήθηκαν στο σύνολο δεδομένων GQA-OOD για την αξιολόγηση της απόδοσής τους σε διαφορετικές συνθήκες εκτός κατανομής. Οι μέθοδοι που επανυλοποίηθηκαν μπορούν να συνοψιστούν ως εξής:

- Τα Rubi [13] και LMH [14] είναι τα σημαντικότερα μοντέλα της πρώτης προαναφερθείσας κατηγορίας (Language debiasing ensemble based methods).
- Το CSS [9] βασίζεται στην κατασκευή counterfactual δειγμάτων μέσω του μασκαρίσματος συγκεκριμένων περιοχών της εικόνας ή λέξεων της ερώτησης και της χρήσης ψευδοετικετών για επίβλεψη.
- Το SSL [8] βασίζεται στην κατασκευή και στην απαίτηση να μην προβλεφθει καμία σωστή απάντηση μέσω self-supervision.

Τα μοντέλα οδήγησαν σε παρόμοια αποτελέσματα με τα δημοσιευμένα, στο δυνολο δεδομένων VQACP-v2 Το μοντέλο [8] χρησιμοιείται έπειτα από pre-training με το κλασικό B"E

Λοσς, και δεν προσέφερε βελτίωση στο validation loss με αποτέλεσμα να μην συμπεριληφθεί στον πίνακα. Το μοντέλο SAR [12] επανεξετάστηκε, αλλά δεν καταφέραμε να ολοκληρώσουμε την εκπαίδευση λόγω του μεγάλου υπολογιστικού κόστους, όπως αναφέρθηκε παραπάνω.

Model	Baseline	GQA-OOD test results			
		Acc tail	Acc head	Acc all	Delta
Baselines					
UpDn	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 4.1}$
Ensemble based methods					
Rubi	UpDn	$30.78_{\pm 2.3}$	$39.52_{\pm 2.4}$	$36.2_{\pm 3.1}$	$22.10_{\pm 4.7}$
LMH	UpDn	$27.621_{\pm 2.5}$	$38.41_{\pm 1.9}$	$34.31_{\pm 2.7}$	$28.09_{\pm 3.9}$
Data augmentation methods					
CSS	UpDn	$41.75_{\pm 1.6}$	$49.11_{\pm 1.9}$	$46.31_{\pm 1.3}$	$14.95_{\pm 2.8}$
LMH+CSS	UpDn	$29.19_{\pm 3.1}$	$37.67_{\pm 2.2}$	$34.45_{\pm 2.3}$	$22.49_{\pm 3.7}$

Οι μέθοδοι αυτές δεν οδήγησαν σε βελτίωση των αποτελεσμάτων στο GQA OOD dataset δείχνοντας ότι η χρήση language biased ensemble based models αλλά και η χρήση τεχνικών για αντιθετικά παραδείγματα δεν γενικεύουν σε όλες τις περιπτώσεις εκτός κατανομής, το οποίο συνάδει με την αντίστοιχη βιβλιογραφική κριτική [2, 1].

0.7 Παραγωγή Νέων Ερωτήσεων μέσω ζεύγους εικόνας και απάντησης.

Τα αρχικά μας πειράματα περιλαμβάνουν την μέθοδο παραγωγής οπτικών ερωτήσεων από εικόνα και απάντηση με σκοπό την δημιουργία καινούργιων παρεμφερών δειγμάτων για ενίσχυση δεδομένων data augmentation.

0.7.1 Συνολο δεδομένων και Baseline Μοντέλο

Για την παραγωγή ερωτήσεων χρησιμοποιήσαμε ένα απλοικό αρχικά μοντέλο διαρθρωμένο ως εξής:

- **Language Encoder:** Μετατρέπει τις απαντήσεις σε σειρά από Glove embeddings και τα περνάει από ένα LSTM για να κατασκευάσει embeddings απάντησης στον κοινό χώρο χαρακτηριστικών .
- **Image Encoder:** Εξαγωγή χαρακτηριστικών από Faster-RCNN για τα αντικείμενα της εικόνας και πέρασμα του μέσου όρου τους από ένα μη γραμμικό layer.
- **Fusion Layer:** Ένα υπολογιστικό μέρος από MLP για την δημιουργία κοινού διανύσματος χαρακτηριστικών για εικόνα, απάντηση.
- **Generator LSTM Network:** Ένα generator LSTM που χρησιμοποιεί το κοινό διάνυσμα χαρακτηριστικών για να προβλέψει την παραφρασμένη ερώτηση.

Για το VQA task χρησιμοποιήσαμε το μοντέλο Bottom Up Top Down Attention που αναφέρθηκε παραπάνω.

0.7.2 Μεθοδολογία

Η μεθοδολογία αυτής της έρευνας περιλαμβάνει αρκετά σημαντικά στοιχεία:

1. **Προσεγγίσεις Εκπαίδευσης:** Το μοντέλο VQG εκπαιδεύτηκε χρησιμοποιώντας διάφορες μεθόδους: με τη χρήση δασκάλου (teacher forcing), χωρίς τη χρήση δασκάλου (no-teacher forcing) και μέσω διδακτικής εξέλιξης (curriculum learning). Χωρίς την χρήση δασκάλου, φάνηκε ότι τα μοντέλα δυσκολεύονταν να κατασκευάσουν σημασιολογικά ή γραμματικά ορθές ερωτήσεις .

2. **Ανάλυση Επιπτώσεων στους Τύπους Ερωτήσεων:** Έπειτα από εκπαίδευση του UpDn μοντέλου με τα παραγώμενα δεδομένα παρατηρήσαμε σημαντική μείωση της αποδοτικότητας του μοντέλου. Πραγματοποιήθηκε μια εκτενής ανάλυση για την αξιολόγηση των επιδράσεων του VQG σε διάφορους τύπους ερωτήσεων μέσα στο σετ δεδομένων. Αυτή η εξέταση αποκάλυψε ότι συγκεκριμένα είδη ερωτήσεων δημιουργούσαν την πτώση απόδοση αυτή.

3. **Εφαρμογή της Μεθόδου Beam Search:** Για να αυξηθεί περαιτέρω η ποικιλία των παραγόμενων ερωτήσεων καθώς τα μοντέλα μας αδυνατούσαν να παράξουν παραπάνω από ένα νέο δείγμα, υλοποιήθηκε η μέθοδος beam search η οποία παρέχει την δυνατότητα παραγωγής περισσότερων διαφορετικών δειγμάτων. Παρότι καταφέραμε να παράξουμε μεγαλύτερη ποικιλία ερώτησεων απαντήσεων, οδήγησε τελικώς σε χειρότερα τελικά αποτελέσματα, πιθανών λόγω της παραγωγής ναι μεν διαφορετικών αλλά χειρότερης ποιότητας δειγμάτων.

Κάποια παραδείγμα την παραγωγής ερωτήσεων μπορούν να φανούν παρακάτω στο Σχήμα4:

0.7.3 Αποτελέσματα

Model	All	Yes/No	Num	Other
UpDn (Baseline)	41.53	43.45	13.64	48.18
Updn + base_VQG (all questions)	38.7	42.34	9.4	43.21
UpDn + base_VQG	41.21	42.73	14.08	48.01
UpDn + 3-beam	40.64	42.60	13.32	47.12

0.7.4 Σχολιασμός

Τα αποτελέσματα υποδεικνύουν ότι η αναδιατύπωση ερωτήσεων μέσω της VQG δεν βελτιώνει σημαντικά, και μπορεί ακόμη και να επιδεινώσει, την απόδοση της VQA σε συνθήκες OOD. Η εργασία καταλήγει στο συμπέρασμα ότι για καλύτερα αποτελέσματα OOD, πρέπει να εφαρμοστεί μια στρατηγική που δημιουργεί νέα ζεύγη ερωτήσεων-απαντήσεων, όπως στην περίπτωση του [10], καθώς η δημιουργία ερωτήσεων με την ίδια αναμενόμενη απάντηση περιορίζει σημαντικά τη διαδικασία γεννήτριας.

Σχήμα 4: *Σε αυτό το σχήμα, μπορούμε να δούμε ότι το πρώτο παράδειγμα είναι μια παραφρασμένη έκδοση της αρχικής ερώτησης. Στο δεύτερο παράδειγμα για την απάντηση 'μπάλα', η δημιουργημένη ερώτηση αναφέρεται στο αγόρι αντί για τον άντρα (το οποίο είναι επιθυμητό), και στην τελευταία ερώτηση, το μοντέλο λανθασμένα αναφέρεται στο κολιέ του σκύλου αντί για το σχέδιο στην άμμο.*

0.8 Προτεινόμενη Μέθοδος: Γενίκευση μέσω απόκρυψης οπτικών αντικειμένων.

Έπειται από την προσεκτική επανυλοποίηση μεθόδων για language debiasing και συγκεκριμένων μεθόδων που αφορούν την παραγωγή παραλλαγμένων δειγμάτων augmented samples, αποφασίσαμε να επικεντρωθούμε σε μια μέθοδο παραλλάγής εικόνων με σκοπό την χρήση θετικών και αρνητικών δειγμάτων μέσω διάφορων τεχνικών για απόκρυψη masking σημαντικών αντικειμένων. Η μέθοδος αυτή υλοποιήθηκε στο dataset ΓΧΑ ΟΟΔ.

0.8.1 Σύνολο Δεδομένων

Για το σύνολο δεδομένων μας, χρησιμοποιούμε το GQA-OOD, το οποίο διαμορφώνεται ως μια εργασία ταξινόμησης μονής ετικέτας VQA. Περιλαμβάνει επιπλέον επεξηγήσεις που περιλαμβάνουν ερωτήσεις ως σημασιολογικά προγράμματα που αναφέρονται σε αντικείμενα που περιλαμβάνονται στις επεξηγήσεις του γράφου σκηνής της εικόνας.

0.8.2 Κατασκευή θετικών και αρνητικών παραδειγμάτων

Εμπνευσμένοι από την εξέταση των τεχνικών [9, 15], δημιουργούμε παρόμοια και αντιφατικά δείγματα για κάθε εικόνα βάσει σημαντικών περιοχών εικόνων.

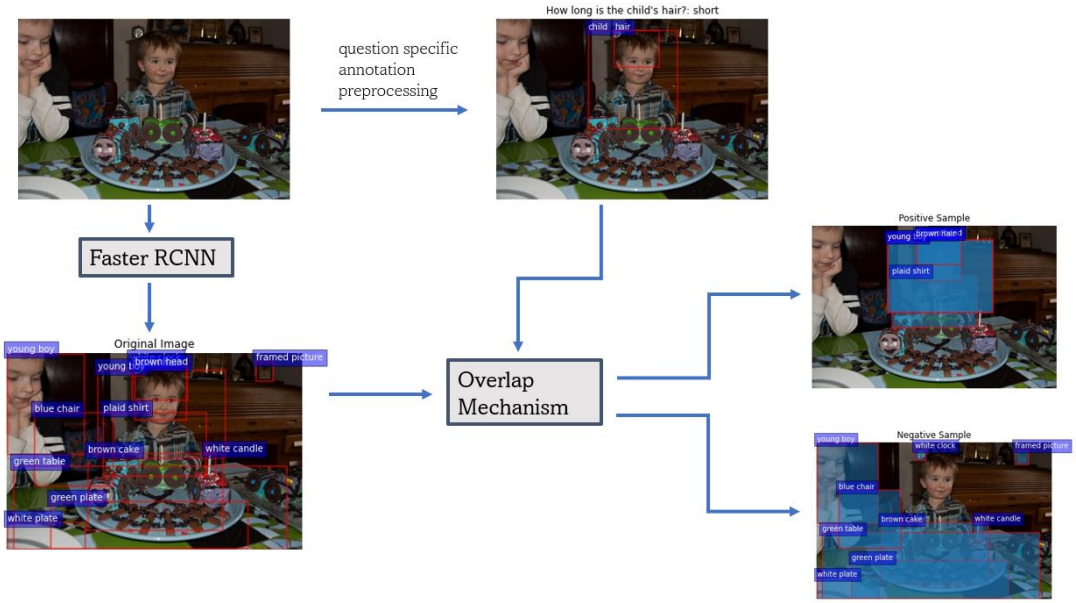
Εάν έχουμε ισχυρότερες σημειώσεις με κάποια μορφή οπτικής εξήγησης, μπορούμε να εντοπίσουμε άμεσα τις σημαντικές περιοχές. Το σετ δεδομένων GQA παρέχει τα βήματα συλλογισμού (προγράμματα) για κάθε ερώτημα και τα επιλεγμένα αντικείμενα μετά από κάθε

βήμα. Χρησιμοποιούμε αυτά τα βήματα συλλογισμού για να εξάγουμε όλα τα σχετικά και μη σχετικά αντικείμενα για το ερώτημα. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε ένα μηχανισμό επικάλυψης πλαισίων αναφοράς για να ταιριάξουμε τα αντικείμενα του Faster-RCNN με τα πραγματικά. Πολλές μεθοδολογίες στο σετ δεδομένων GQA χρησιμοποιούν τον δείκτη IoU παρομοίως ως επιπλέον πληροφορία για τις αντίστοιχες μεθοδολογίες τους [16, 7]. Ωστόσο, ο δείκτης IoU δεν είναι απαραίτητα ο κατάλληλος μηχανισμός επικάλυψης για τη μέθοδό μας. Σε αντίθεση με το [16], δεν προσπαθούμε να εκτελέσουμε μια 1-1 αντιστοίχιση μεταξύ πραγματικών και εξαγόμενων αντικειμένων, αλλά θέλουμε να επιλέξουμε όλα τα αντικείμενα με τουλάχιστον κάποιο βαθμό επικάλυψης με τα πραγματικά και να τα κατηγοριοποιήσουμε ως σημαντικά. Η μετρική που χρησιμοποιήσαμε είναι :

$$Overlap = \frac{ObjectArea \cap ExtractedObjectArea}{ObjectArea}$$

Όπως φαίνεται στην εικόνα 5 , αποκρύπτουμε τα σημαντικά αντικείμενα δημιουργώντας ένα αρνητικό δείγμα εικόνας-απάντησης και αντίθετα αποκρύπτοντας τα μη σημαντικά αντικείμενα δημιουργούμε ένα θετικό δείγμα.

Σχήμα 5: *Is the hair brown and thin?: yes*



0.8.3 Συναρτήσεις Απώλειας για Regularization

Χρησιμοποιώντας διάφορες συναρτήσεις απώλειας, εξερευνούμε τέσσερις κρίσιμες πτυχές της γενίκευσης στην απάντηση οπτικών ερωτήσεων (VQA).

Αυτές οι πτυχές περιλαμβάνουν την αξιολόγηση της αντιθετικότητας των αρνητικών παραδειγμάτων, τη δυνατότητα χρήσης θετικών δειγμάτων για ενίσχυση, την εγκυρότητα των υποθέσεων που υποκρύπτονται στην απώλεια τριάδας, και τη ρύθμιση της δυνατότητας των τυχαίων μηχανισμών μάσκας. Ας εξετάσουμε κάθε πτυχή αναλυτικά :

Counterfactual Losses

Χρησιμοποιούμε τα εξαγόμενα αρνητικά δείγματα ως αντιθετικά (counterfactual) σε τρεις διακριτές περιπτώσεις.

Self-Supervised Loss Η παρακάτω συνάρτηση κόστους υποδηλώνει ότι για ένα αντιθετικό ζευγάρι θα πρέπει να να προβλέπει μηδενική κατανομή για το σύνολο ων απαντήσεων και χρησιμοποιήθηκε στο SSL μοντέλο [8].

$$L_{qd} = \frac{1}{N}\sum_{i=1}^{N} P(A_i|Q_i, \hat{I}_{i0})$$

όπου Q_i είναι η ερώτηση, $\hat{I}_{i0}$ η αντιθετική απάντηση και $P(A_i)$ η έξοδος του δικτύου για την απάντηση A_i .

Gradient Supervision Η απώλεια Εποπτείας Κλίσης (GS), η οποία χρησιμοποιείται σε συνδυασμό με δείγματα αντιθετικών παραδειγμάτων, είναι μια μαθηματική μέθοδος που χρησιμοποιεί αντιθετικά παραδείγματα για να οδηγήσει η κλίση του δικτύου σε κάθε σημείο εισόδου ευθυγραμμίζεται με ένα ground-truth διάνυσμα κλίσης.

$$L_{\mathrm{GS}}(g_i, \hat{g}_i) = 1 - \frac{g_i \cdot \hat{g}_i}{||g_i||||\hat{g}_i||} \tag{0.2}$$

όπου $\hat{g}$ είναι η αναπαράσταση του αντιθετικού παραδείγματος και g είναι η αναπαράσταση του πραγματικού.

Τέλος, στην αναφορά [9], χρησιμοποιούν **supervised loss** για τα counterfactual δείγματα. Τα counterfactual labels είναι το αντίστροφο των καλύτερων k προβλέψεων του μοντέλου από την επεξεργασία του θετικού δείγματος. Στο πλαίσιο του GQA, όπου η ταξινόμηση πολλαπλών ετικετών δεν είναι εφαρμόσιμη,το loss function που αναφέρεται στις προηγούμενες ενότητες απλοποιείται. Για ένα ζεύγος VQA, αναθέτουμε $a = 1$ εάν η σωστή απάντηση δεν προβλέπεται σωστά. Αντιστρόφως, εάν η απάντηση προβλέπεται σωστά, αναθέτουμε $a = 0$. Αυτό αντιπροσωπεύει μια προσέγγιση **αντίστροφης επισήμανσης**, όπου $a = 0$ δηλώνει σωστές προβλέψεις και $a = 1$ δηλώνει λανθασμένες προβλέψεις,

Θετικά Δείγματα ως τεχνική augmentation

Προπονούμε το μοντέλο με τα θετικά παραδείγματα, μετατρέποντας την επιβλεπόμενη μέθοδο supervised BCE loss σε μία τεχνική ενίσχυσης.

Triplet Loss

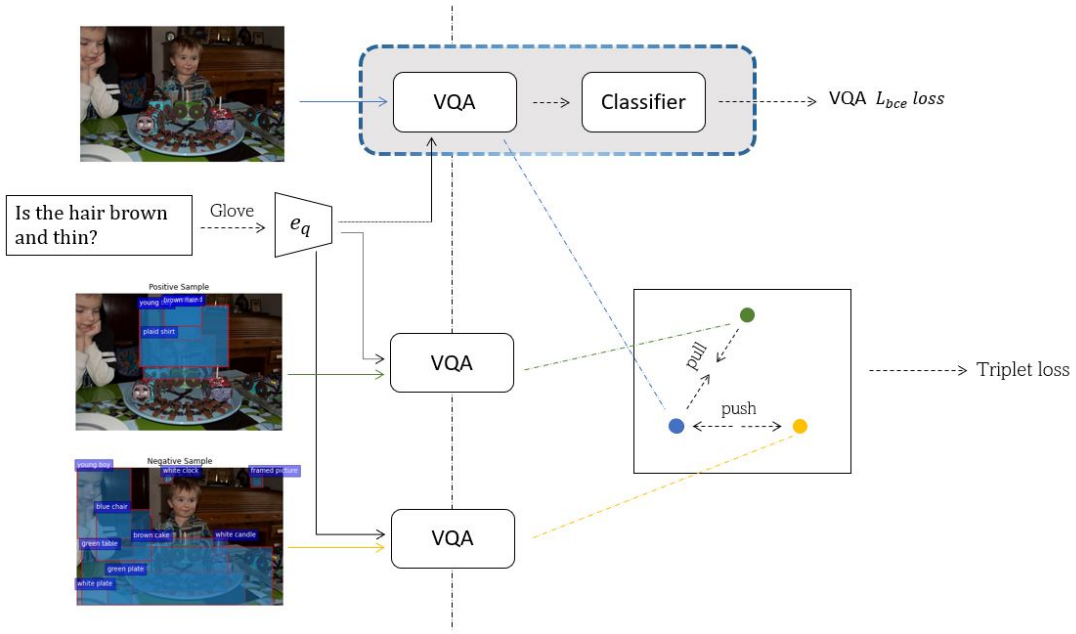
Το Triplet loss που χρησιμοποιήσαμε, αποτελεί μια πιο σταθερή αναδιαμόρφωση του κλασικού triplet loss [17] χρησιμοποιώντας τον μαθηματικό μετασχηματισμό του cosine similarity στον πολυδιάστατο χώρο πριν τον classifier και διατυπώνεται ως εξής:

$$L_c = \mathbb{E}_{p,n,a}\left[-\log\left(\frac{e^{s(a,p)}}{e^{s(a,p)} + e^{s(a,n)}}\right)\right]$$

Η ιδέα του βασίζεται στην υπόθεση ότι τα θετικά δείγματα θα έπρεπε να βρίσκονται πιο

κοντά στον πολυδιάστατο χώρο χαρακτηριστικών στα πραγματικά δείγματα σε σχέση με τα αρνητικά δείγματα και αποτυπώνεται γραφικά στην εικόνα 6.

Σχήμα 6: *Is the hair brown and thin?: yes*



Training Objective

Για όλες τις μεθόδους μας, η συνάρτηση απώλειας διαμορφώνεται ως το άθροισμα της απώλειας BCE με την επιβαρυμένη μας απώλεια κανονικοποίησης από μία παράμετρο λ_{reg}:

$$L = L_{\text{vqa}} + \lambda_{\text{reg}} \times L_{\text{reg}}$$

0.8.4 Πειράματα

Τα πειράματα μας διαρθρώθηκαν ως εξής:

Αντιθετική Μάθηση Η αντιθετική μάθηση συμπεριλάμβανε πειράματα με αντιπαραδειγματικά δείγματα και τα αντίστοιχα counterfactual losses, μεταβάλλοντας την υπερπαράμετρο βάρους (λ). Υψηλότερες τιμές του λ οδήγησαν σε **χαμηλότερη απόδοση**, ενώ χαμηλότερες τιμές έδειξαν παρόμοια απόδοση με το βασικό μοντέλο.

Πειράματα Triplet Loss: Παρατηρήσαμε μια σαφή τάση σχετικά με τη ρύθμιση της υπερπαραμέτρου λ, η οποία ελέγχει τη συνεισφορά της απώλειας τριπλέτας στη συνολική συνάρτηση απώλειας. Χαμηλότερες τιμές του λ (0.1, 0.2) έδειξαν αξιοσημείωτες βελτιώσεις σε όλες τις μετρήσεις.

Πειράματα με επιβλεπόμενη μάθησης χρησιμοποιώντας τεχνικές επαύξησης: Σε συνδυασμό με τα αρχικά δεδομένα, χρησιμοποιούμε ένα BCE supervised loss για τα θετικα παραδείγματα. Η μέθοδος αυτή μπορεί να θεωρηθεί μέθοδος επαύξησης βελτιώνοντας τα αποτελέσματα σε όλες τις μετρικές.

Πειράματα τυχαίας μάσκας: Για να δοκιμάσουμε την επίδοση του custom masking μας αποφασίσαμε να χρησιμοποιήσουμε τυχαίες μάσκες στα αντικείμενα της εικόνας με

βάση κάποια πιθανότητα. Η χρήση τυχαίων μάσκών τόσο για augmentation και triplet loss οδήγησε σε καλύτερα αποτελέσματα. Πειραματιζόμενοι με το ποσοστό των τυχαίων μασκών παρατηρήσαμε ότι τα καλύτερα αποτελέσματα είχαμε για πιθανότητα 0.82 που είναι αντίστοιχη με την μέσο ποσοστό σημαντικών αντικειμένων σε σχέση με τα συνολικά αντικείμενα στο δατασετ μας.

Τελικά Μοντέλα

Τα τελικά μοντέλα, με τα καλύτερα αποτελέσματα, που προέκυψαν από την πειραματατική μελέτη μας, διαφαίνονται στον παρακάτω πίνακα και είνα:

1. Χρήση BCE loss με augmented θετικά δειγματα για τυχαίες μάσκες με μασκαρίσματος πιθανότητα 0.82.

2. Χρήση BCE loss με augmented θετικά δειγματα με τυχαίες μάσκες και χρήση triplet loss με custom μάσκες.

3. Χρήση BCE loss με augmented θετικά δειγματα με συστομ μάσκες και χρήση triplet loss με custom μάσκες.

GQA OOD					
Loss	Model	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Augm Rand	UpDn	$44.803_{\pm 0.7}$	$\mathbf{52.352}_{\pm 0.8}$	$49.482_{\pm 0.4}$	$16.898_{\pm 3.3}$
BCE+ Augm Rand +Triplet with Heur.	UpDn	$\mathbf{45.2}_{\pm 1.3}$	$52.284_{\pm 0.5}$	$\mathbf{49.704}_{\pm 0.5}$	$15.369_{\pm 4.0}$
BCE+ Augm + Triplet with Heur.	UpDn	$44.73_{\pm 0.47}$	$52.26_{\pm 0.65}$	$49.41_{\pm 0.42}$	$\mathbf{14.4}_{\pm 2.06}$

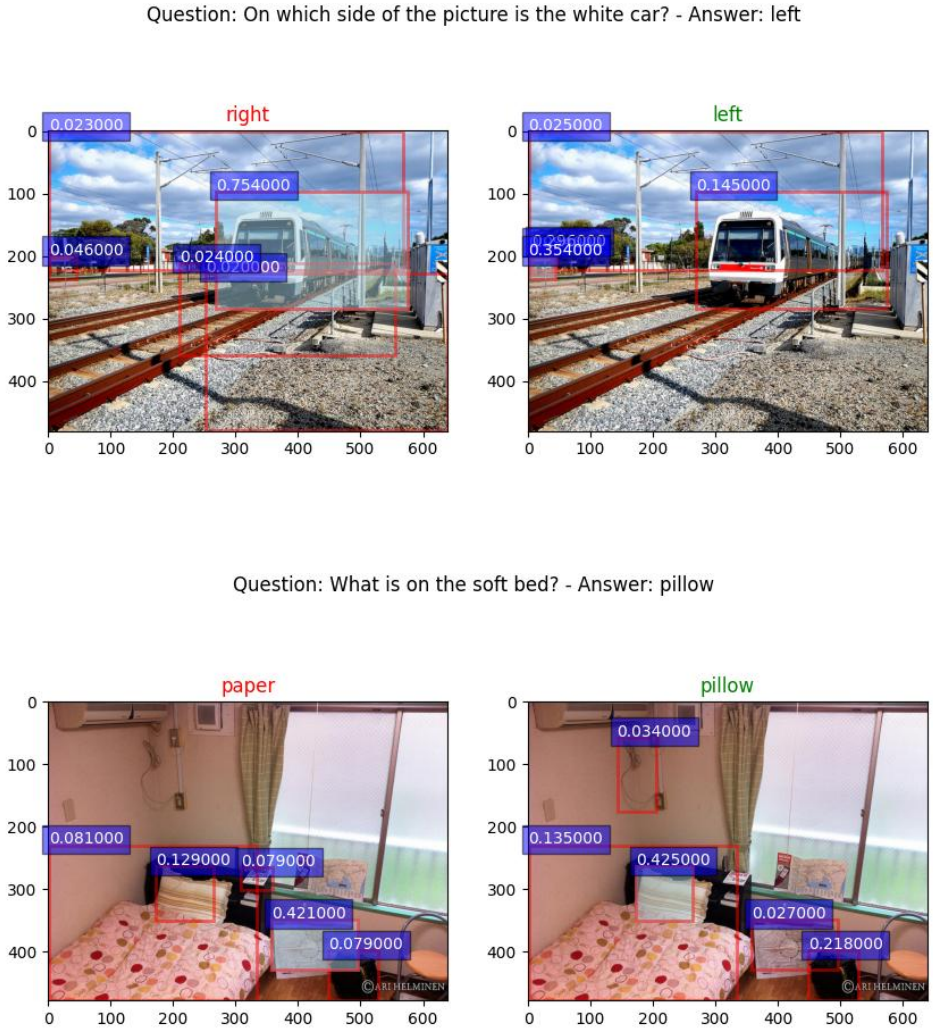
Στο Σχήμα 7 μπορούμε να δούμε κάποια παραδείγματα του μοντέλου 3 σε σχέση με το baseline και την ικανότητα του να επικεντρώνει καλύτερα σε σημαντικότερες περιοχές της εικόνας.

0.9 Συμπεράσματα

Επισκόπηση Βιβλιογραφίας και Αναπαραγωγής Υλοποιημένων Μεθόδων: Η έρευνά μας τονίζει την κρισιμότητα της προσεκτικής αντιμετώπισης των γλωσσικών προκαταλήψεων στα μοντέλα οπτικής απάντησης ερωτήσεων (VQA). Αυτό το θέμα είναι ιδιαίτερα έντονο στα μοντέλα συνόλου που βασίζονται σε αντιστροφή αποτελεσμάτων με μοντέλα προκατειλημμένα στη γλώσσα, τα οποία δείχνουν μειωμένη αποτελεσματικότητα σε σενάρια GQA-OOD και είναι περιορισμένα σε ευρύτερες εφαρμογές.

Σημαντικό, επίσης, είναι να μην γνωρίζουμε την κατανομή του test set σε σχέση με το train set εξαρχής, καθώς τεχνικές εξαρτώμενες σε αυτή την συγκεκριμένη αλλαγή ενδέχεται να οδηγήσει σε πλασματική αύξηση επιδόσεων που δεν γενικεύει σε άλλα σύνολα δεδομένων.

Η επισκόπησή μας αναγνωρίζει το δυναμικό των μεθόδων βασιζόμενων σε data augmentation στο VQA. Ωστόσο, όπως τεκμηριώνεται από την βιβλιογραφία, σημαντικές βελτιώσεις απόδοσης περιορίζονται σε συγκεκριμένες στρατηγικές αύξησης και δεν είναι γενικά εφαρμόσιμες σε όλες τις εργασίες VQA. Τέλος, οι μέθοδοι για answer reranking έχουν περιθώρια



Σχήμα 7: *Ο χάρτης προσοχής του τελικού μοντέλου 3 δείχνει ότι απαντά σωστά γιατί επικεντρώνει καλύτερα σε σημαντικότερες περιοχές της εικόνας.*

βελτίωσης αποτελεσμάτων αλλά χρειάζεται αναπροσαρμογή τους ώστε να καταπολεμηθεί το μεγάλο βάρος χρονικής και τοπικής πολυπλοκότητας (hardware requirements and training time) που επιφέρουν στα μοντέλα.

Παραγωγή Οπτικών Ερωτήσεων: Τα πειράματά μας στην παραγωγή οπτικών ερωτήσεων δείχνουν ότι η προσπάθεια για διαταραχή της ερώτησης χωρίς αλλαγή των απαντήσεων οδήγησε σε υποβέλτιστη απόδοση. Επομένως, θα πρέπει να υλοποιηθούν μεθοδολογίες που δημιουργούν νέα παρόμοια ζεύγη ερώτησης-απάντησης από τα αρχικά μας δείγματα ώστε να εκμετταλευτούμε πλήρως την επιπλέον πληροφορία που ενδεχομένων να προσέφερε ένα ζεύγος εικόνας-ερώτησης-απάντησης.

Συμπεράσματα για κύρια μεθοδολογία: Η εκτεταμένη μας πειραματική εργασία στον τομέα των οπτικών ερωτήσεων και απαντήσεων έχει οδηγήσει σε αρκετά σημαντικά ευρήματα. Πιο συγκεκριμένα, τα πειράματα δείχνουν ότι τα αρνητικά δείγματα που χρησιμοποιήθηκαν στη μελέτη μας δεν μπορούν να θεωρηθούν εντελώς αντιθετικά, καθώς οι μέθοδοι

βασιζόμενες σε συναρτήσεις κόστους αντιθετικών παραδειγμάτων δεν οδηγούν σε βελτιώσεις απτελεσμάτων. Αυτό οφείλεται, ίσως, στο γεγονός ότι ακόμη και με μασκαρισμένα τα σημαντικά αντικείμενα διατηρούν σημαντικές πληροφορίες εικόνας που εξάγονται μέσω μοντέλων CNN. Αυτή η παρατήρηση υποδηλώνει μια πιθανή περιοριστική διάσταση στην προσέγγισή μας για τη δημιουργία αρνητικών δειγμάτων, η οποία μπορεί να προκαλεί υπερβολική διαρροή πληροφοριών.

Οι custom και τυχαίες τεχνικές μασκαρίσματος οδήγησαν σε βελτίωση αποτελεσμάτων. Λειτουργούν ως τεχνικές κανονικοποίησης, ενισχύοντας την ανθεκτικότητα και την απόδοση του μοντέλου. Η ενισχυμένη απώλεια BCE (augmentation loss), ιδιαίτερα όταν συνδυάζεται με τυχαίο μασκάρισμα, λειτουργεί παρόμοια με μια εκλεπτυσμένη μέθοδο dropout και αυξάνει σημαντικά τις γενικευτικές δυνατότητες του μοντέλου. Επίσης, η custom τεχνική μασκαρίσματος φαίνεται να βοηθά το μοντέλο να εστιάσει περισσότερο σε σημαντικές περιοχές της εικόνας, βελτιώνοντας έτσι τις ικανότητες προσοχής του.

Επιπλέον, η μεθοδολογία triplet loss μας βελτιώνει τα συνολικά αποτελέσματα και μπορεί να συνδυαστεί αποτελεσματικά με την προσέγγιση ενίσχυσης. Αν και ως μοναδική μέθοδος κανονικοποίησης, παρουσιάζει χειρότερα αποτελέσματα από τις μέθοδους βασιζόμενες σε ενίσχυση, αποδεικνύει τη μεγάλη βελτίωση στην τιμή Delta, η οποία είναι κρίσιμη για την επίτευξη παρόμοιας απόδοσης σε σενάρια εκτός κατανομής (OOD) και εντός κατανομής (ID). Ωστόσο, τα πειράματά μας με απώλειες αντιφατικών περιστάσεων υποδηλώνουν ότι μπορεί να μην αξιοποιούμε πλήρως το triplet loss λόγω της πιθανότητας χρήσης υπο-βέλτιστων αρνητικών παραδειγμάτων.

0.10 Μελλοντικές προεκτάσεις

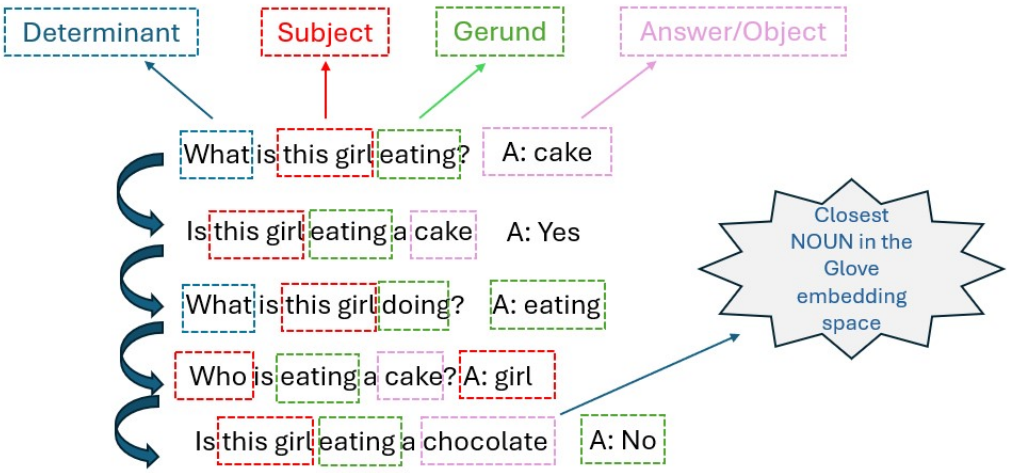
Στο μέλλον, οι επεκτάσεις της εργασίας μας μπορούν να περιλαμβάνουν:

- **Βελτιώσεις στη Δημιουργία Αρνητικών Δειγμάτων:** Η εκπαίδευση με αντιθετικές μεθόδους για τα αρνητικά δείγματα δεν βελτίωσε τα συνολικά αποτελέσματα, δείχνοντας ότι η απώλεια τριπλέτας (triplet loss) δεν χρησιμοποιήθηκε πλήρως. Οι πειραματισμοί μας θα μπορούσαν να περιλαμβάνουν πιο λεπτομερείς αρνητικές εικόνες ή ερωτήσεις για τα τριπλέτα μας. Θα μπορούσαμε επίσης, να πειραματιστούμε με τη μάσκα σημαντικών λεκτικών στοιχείων εκτός από τη μάσκα του περιεχομένου της εικόνας ή/και τη μάσκα αντικειμένων που έχουν την ίδια κλάση με τα σημαντικά μας αντικείμενα. Για παράδειγμα, σχετικά με το δείγμα VQA στο Σχήμα 6, στην πρώτη περίπτωση θα μασκαριστούν και οι περιοχές των αντικειμένων των αγοριών και στη δεύτερη περίπτωση η μάσκα της λέξης ¨μαλλιά¨ θα δημιουργούσε ένα σημασιολογικά διαφορετικό δείγμα ερώτησης-εικόνας (Q-I).

- **Αντικατάσταση του UpDn με καλύτερα μοντέλα:** Θα μπορούσαμε να πειραματιστούμε με τα μοντέλα διμερούς προσοχής [18], τα οποία επιτυγχάνουν υψηλότερη απόδοση στα περισσότερα καθήκοντα χα [18, 19, 2] ή παρόμοια μοντέλα βασισμένα σε μετασχηματιστές [20, 21] εφόσον ξεπεράσουμε το ζήτημα της διαρροής δεδομένων δοκιμής στην προεκπαίδευσή τους. Τα μοντέλα διμερούς προσοχής έχουν παρουσιάσει σημαντικές βελτιώσεις σε σύγκριση με το αρχικό μοντέλο UpDn όπως φαίνεται

στο [21, 18, 20] καθώς ο μηχανισμός προσοχής τους *NXM* επιτρέπει πιο λεπτομερή αλληλεπίδραση λεκτικών στοιχείων-αντικειμένου. Με την προϋπόθεση ότι διαθέτουμε επαρκείς υπολογιστικούς πόρους, η δοκιμή των μεθοδολογιών μας σε άλλα αγνωστικά οπτικο-γλωσσικά μοντέλα θα ήταν μια εξαιρετική μορφή επικύρωσης της architecture agnostic μεθοδολογίας μας για γενίκευση σε VQA.

- **Δημιουργία νέων ενισχυμένων ζευγών ερώτησης-απάντησης μέσω σημασιολογικής αναδιατύπωσης:** Εμπνευσμένοι από τις μεθόδους επαναταξινόμησης στην επιστημονική βιβλιογραφία [11, 12] και από τα απογοητευτικά αποτελέσματά μας στη μέθοδο VQG, ένα ζεύγος ερώτησης-απάντησης θα μπορούσε να περιέχει επιπλέον σημασιολογικές πληροφορίες αν ανακασκευαστεί κατάλληλα. Πιο συγκεκριμένα, η πλειοψηφία των ζευγών ερώτησης-απάντησης στο VQA περιλαμβάνονται σε συγκεκριμένο τύπο ερώτησης (π.χ. "What type") και μπορούν να αποδομηθούν χρησιμοποιώντας τον ετικετοποιητή θέσης spacy [22] και να αναδιατυπωθούν με τρόπο ώστε να δημιουργούμε σημασιολογικά παρόμοια αλλά διαφορετικά ζεύγη ερώτησης-απάντησης. Αυτή η διαδικασία θα μπορούσε να επιτευχθεί τόσο με τη χρήση LLMs όπως το GPT-3 όσο και με ένα σύστημα βασισμένο σε κανόνες παρόμοιο με τη δημιουργία αρκετών συνόλων δεδομένων VQA [19], όπως παρουσιάζεται στο Σχήμα 6.1. Με την ενίσχυση του συνόλου δεδομένων μας με παρόμοια δείγματα που παρουσιάζουν σημασιολογικές διαφορές και διαφορετικές απαντήσεις όπως φαίνεται στην εικόνα 8, θα μπορούσαμε να ενισχύσουμε τη λογική και τη σημασιολογική κατανόηση των μοντέλων VQA μας και να αποτρέψουμε την εξάρτησή τους από γλωσσικές προκαταλήψεις ή από ανομοιόμορφες κατανομές απαντήσεων.



Σχήμα 8: *Αυτοματοποιημένη παραγωγή νέων σημασιολογικά κοντινών ζευγαριών ερώτησης απάντησης μέσω συνταντικής αποδόμησης.*

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
μασκάρισμα	masking
απώλεια	loss
σύνολο δεδομένων	dataset
ενισχυμένη απώλεια	augmented loss
αντιθετικός	counterfactua
υποβέλτιστη απόδοση	suboptimal performance
αρνητικά δείγματα	negative samples
τεχνικές κανονικοποίησης	regularization methods
κανότητες προσοχής	attention mechanisms
επαυξημένα	augmented
αντίστροφη επισήμανση	inverted label
γράφοω σκηνής	scene graph
οπτικές ερωτήσεις	Visual Questions

Chapter 1

Introduction

1.1 Motivation

Visual Question Answering (VQA), i.e., answering natural language questions about the content of an image, is one of the most important visual-linguistic tasks and an important method towards the path to General AI. The potential applications of VQA are vast in numerous fields, such as healthcare diagnostics, autonomous vehicle navigation, robotics, educational tools, smart home devices, and interactive digital assistants.

These models need to accurately recognize objects, scenes, and activities in images and understand the context and nuances of natural language queries. This dual understanding enables them to provide precise and relevant answers to various questions about an image.

Developing robust VQA models involves challenges like understanding the interplay between visual elements and textual descriptions, dealing with ambiguity in both visual and textual inputs, and handling a wide variety of question types. As research in VQA and related fields progresses, these models are expected to become more sophisticated and capable. Although the state-of-the-art[23, 18, 21] can achieve good results on the VQA benchmarks such as VQA v2[24], they tend to rely on spurious correlations and consequently overperform in the I.D domain, but underperform in different testing conditions.

The above are amplified by the language biases in various VQA datasets and the lack of proper performance metrics. Because language is easier to process and is often more important for QA tasks, most state-of-the-art VQA models tend to over-rely on those language biases and exploit shortcuts to achieve better performance, resulting in poor visual grounding. For example, in the question “What color is the banana?", most models will answer “yellow" without attending to proper image regions because it is the most common answer relative to that specific question in the training set. Consequently, the model would theoretically achieve high accuracy, despite the poor visual understanding of the model.

With this motivation, we decided to delve into the out-of-distribution datasets and generalization methods utilized in VQA and propose our methodologies for improving performance in out-of-distribution conditions.

Figure 1.1: *Common false answer of a VQA model hyper-dependent on language.*

1.2 Contributions

This thesis constitutes a significant advancement in Visual Question Answering (VQA), with a particular emphasis on improving model generalization. An extensive review of existing approaches was conducted to understand the current landscape of VQA methodologies better, concentrating on their generalization aspects. We successfully re-implemented several pivotal methods reported in the literature, applying them to out-of-distribution datasets such as VQACPv2 [25] and GQA OOD [2]. This process laid the foundation for our experimental analysis, allowing for a detailed assessment of the strengths and weaknesses of prevalent generalization techniques in VQA. Furthermore, our research also involved exploring question generation for data augmentation in VQA, utilizing image content and answer features.

The core contribution of this thesis is developing a novel image object masking methodology that diverges from traditional approaches. Our custom masking methods are based on identifying important objects by leveraging annotations in our dataset and using masking to construct positive and negative Image-Question-Answer triplets. It leverages a triple contrastive loss function responsible for pulling the multimodal representations of the real samples closer to the positive samples and away from the negative ones. Additionally, we leveraged an augmentation loss using only the positive samples. Lastly, we experimented with a random masking approach that showcased significant performance improvements paired with our initial methodology. Our proposed models combining the mentioned methodologies lead to significant performance improvements under out-of-distribution conditions in the GQA OOD.

1.3 Thesis outline

The thesis is structured into several sections, with Chapter 1 providing fundamental knowledge about machine learning. This chapter serves as a basis for comprehending more advanced concepts relevant to the thesis, including various aspects of neural networks, such as recurrent neural networks.

Chapter 3 is dedicated to visual question answering, starting with an introduction

to the field, the baseline system typically used in the literature, and our thesis. It also includes a comprehensive literature review that covers most generalization methods and the datasets commonly used in this domain.

Chapter 4 focuses on re-implementing specific methodologies presented in the literature, and we conduct initial experiments related to visual question generation as an augmentation task, highlighting key findings and insights.

Chapter 5 presents our proposed methodology, which includes a custom visual object masking methodology, augmentation strategies, and contrastive learning. We outline the extensive experiments we conducted and discuss the results obtained, aiming to clearly understand how our approach contributes to the field and the implications of our findings.

In Chapter 6, we summarize our work and our main findings, and we provide ideas for future work.

Chapter 2

Machine Learning

Machine learning (ML), a crucial subfield of Artificial Intelligence (AI), focuses on developing algorithms capable of autonomously learning from data to accomplish specific tasks. Unlike traditional rule-based systems or classical algorithms explicitly programmed for specific problem-solving, machine learning models are designed model statistical correlations in the training data, in order to make predictions or decisions in new, unseen situations.

The potency of machine learning lies in its applicability to complex problems where crafting explicit, rule-based solutions is either highly challenging or virtually impossible. Tasks such as sentiment analysis and autonomous driving serve as prime examples. While humans can navigate these tasks relatively easily, developing high-performance algorithms based solely on predefined rules proves more feasible.

Machine learning methods are applied to increasing domains and are prevalent in tasks such as computer vision, computational biology, speech and music recognition, recommendation systems, and robotics. The availability of big data and the computational power afforded by graphical processing units (GPUs) have paved the way for large-scale machine learning models, often called deep learning. Models designed with the appropriate architectures and trained on ample data have the potential to outperform traditional rule-based systems and simpler machine learning algorithms in terms of generalizing to new, unseen data.

In summary, machine learning offers a powerful alternative to traditional algorithms, especially regarding complex tasks that cannot be solved computationally or analytically. It leverages data to train models capable of generalization, thereby serving as an essential tool in various scientific and technological applications. Advances in computational hardware and data availability have catalyzed the evolution of large-scale deep-learning models, further enhancing the ability of machine learning to tackle intricate problems.

2.1 Supervision Types

Regarding the topic of learning, there exist different types, but our focus will be primarily on Supervised, Unsupervised, Self-Supervised learning, and Contrastive Learning.

2.1.1 Supervised Learning

In supervised learning, the dataset comprises labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i represents the feature vector and y_i is the corresponding label or supervisory

signal. The objective is to learn a mapping function $g : X \rightarrow Y$, where X and Y denote the input and output spaces, respectively. It is assumed that an unknown function $y = g(x)$ maps input features x_i to output labels y_i. Machine learning models aim to approximate g with $\hat{g}$, leveraging the information available in the training data [26] .

Supervised learning primarily focuses on two types of tasks: **classification** and **regression**. Classification involves categorizing input data into predefined classes. For example, in image classification, a model might be trained to distinguish between pictures of dogs and cats. Regression, conversely, seeks to model the relationship between dependent and independent variables, producing a continuous output.

Figure 2.1 illustrates the difference between classification and regression within the context of supervised learning.

On the left side, we see a classification task, where the data points are divided into two classes, represented by 'X's and 'O's. The decision boundary, shown as a line, separates the two classes. The goal of a classification model is to categorize new observations correctly into one of these classes based on their features.

On the right side, the regression task is depicted, showing a scatter plot of data points. The goal of regression is to fit a function (curve) that best represents the relationship between the independent variables (on the horizontal axis) and the dependent variable (on the vertical axis). The fitted curve is the model's prediction for the dependent variable's value, given new input data. This curve can be used to predict numeric values for new, unseen inputs, which is exemplified by the dotted line extending from a new input point to the fitted curve, indicating the predicted output.

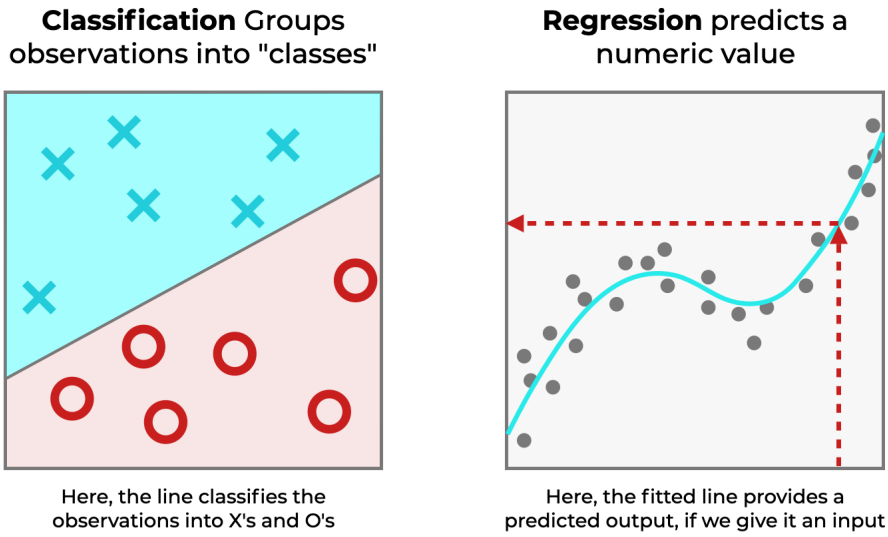


Figure 2.1: *Classification vs Regression. Adjusted from* https://r-craft.org/the-roc-curve-explained/

2.1.2 Unsupervised Learning

Unsupervised learning algorithms can primarily be categorized into three essential tasks: **Clustering**, **Association**, and **Dimensionality Reduction**, each serving a unique function in the realm of data analysis and interpretation.

1. **Clustering Algorithms**, such as k-means, meticulously partition data into distinct clusters based on the similarity of features as seen in 2.2, enabling insightful comprehension of inherent data patterns and structures [27].

2. **Association Rule Learning** is crucial for uncovering interesting relationships between variables. It is fundamental for deciphering underlying structures and relations in datasets, notably in the field of word embeddings, where it elucidates associations between words within a corpus.

3. **Dimensionality Reduction Techniques**, like Principal Component Analysis (PCA), are paramount for transforming high-dimensional datasets into more tractable, lower-dimensional forms, preserving the essential relationships and characteristics inherent to the original data.

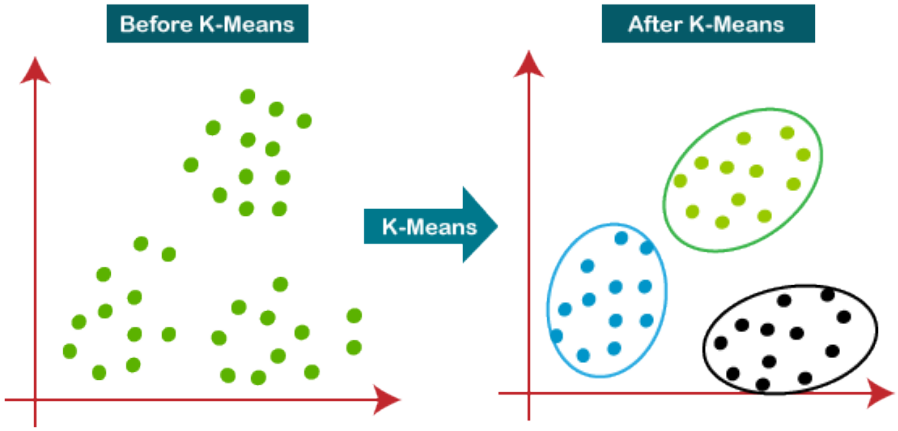


Figure 2.2: *K means clustering on unsupervised data.*

There are also other types of unsupervised methods as we seen in GloVe (Global Vectors for Word Representation)[28]. The learning algorithm leverages world co-occurency probabilities, which can be seen as an association based method to discern relationships between words in a corpus and could be seen as performing dimensionality reduction in the Bag-of-Words embedding space.

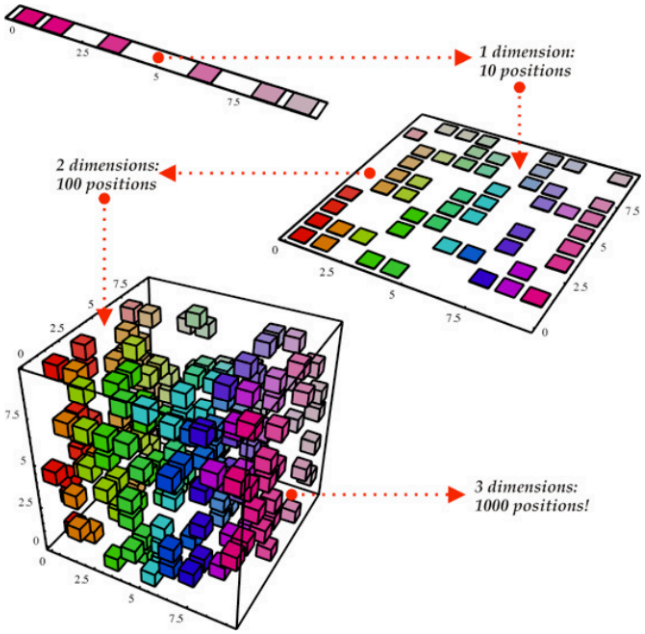


Figure 2.3: *Dimensionality reduction of 3d data to 2d and 1d. Adjusted from* https://teamraft.com/2019/01/04/dimensionality-reduction.html

2.1.3 Adversarial Learning

Often associated with Generative Adversarial Networks (GANs) [29], adversarial learning involves training two models simultaneously: a generator and a discriminator. The generator creates data instances that are intended to be indistinguishable from real data, while the discriminator tries to distinguish between real and generated data.

Adversarial methods can also be utilized as regularizers for a dsicriminator model in cases where the adversary is the output of a biased model [30] or an adversarily perturbated input [6, 31] as seen in 2.4.

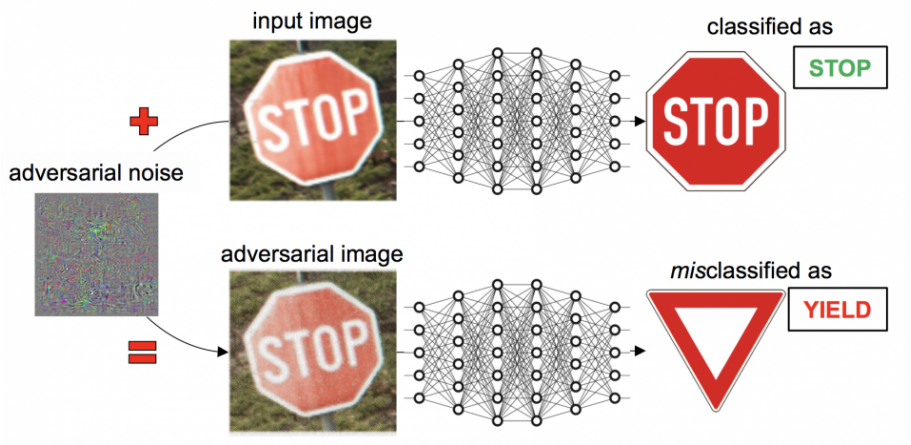


Figure 2.4: *Missclassification in case of adversarial perturbated inputs.*

2.1.4 Self-Supervised Learning

Self-supervised learning (SSL) is a paradigm where the model is trained to predict part of the input data from other parts of the same data. This approach effectively leverages the data's inherent structure to automatically generate supervisory signals. Consequently, SSL can be seen as an intermediate form between supervised and unsupervised learning,

Different strategies are employed within SSL, include:

- **Autoencoders**: Autoencoders are neural networks designed to replicate their input at the output layer. During this process, leverages an encoding and a decoding model. The network compresses the data in a lower-dimensional latent space using the encoder and decompresses the data using the decoder to re-obtain a reconstruction of the original input. In Figure 2.5 we can observe the basic architecture. Autoencoders are used primarily for feature extraction and dimensionality reduction but, in some instances, can also be utilized in [32] for generative models that can create new data points or reconstruct partial or noisy inputs as seen in Figure 2.6.

- **Self-supervised Pretraining**: Models are pretrained using automatically generated labels. For instance, in multimodal NLP-Vision transformers like LXMERT, pretraining involves matching pairs of image-text data collected from the web [21]. Another potential self-supervised learning method is similar to autoencoding where certain parts of the input signals are masked and we try to predict them [33].

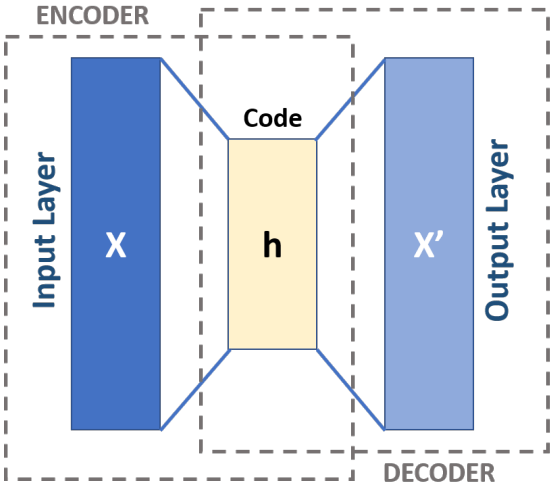


Figure 2.5: *Visualization of an encoder, decoder architecture and the hidden embedding space*

.

2.1.5 Multitask Learning

Multitask learning involves leveraging different strategies of learning at the same time (basically different loss functions). Multitask learning is either used for trying to solve

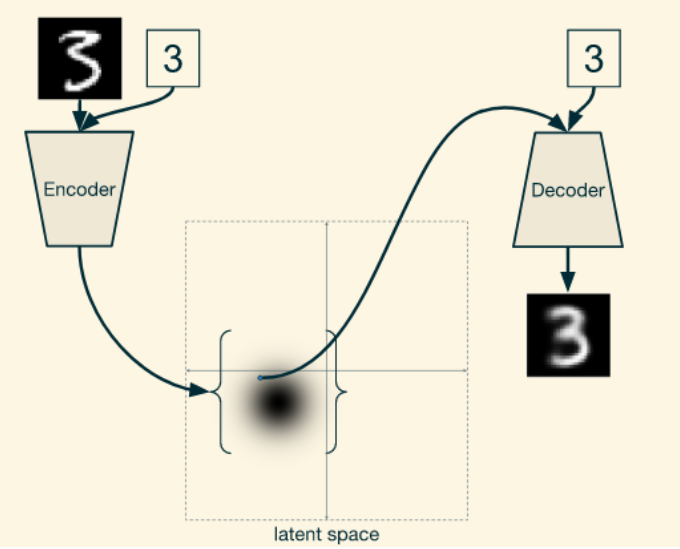


Figure 2.6: *Reconstruction of a digit using Conditional Variational autoencoders. Adjusted from* hhttps://towardsdatascience.com/understanding-conditional-variational-autoencoders-cd62b4f57bf8

.

different aspects of a similar problem at the same time as in [34] where we try to predict the objects of the image the classes and their bounding boxes, or using extra supervisory signals of similar task that introduces a bias that helps to produce better results.

2.1.6 Contrastive Learning

This approach differentiates between similar and dissimilar data points through a specialized loss function [17]. We usually construct the dissimilar pairs (or triplets) [17, 14] by leveraging specific data patterns or by constructing pseudo-labels [14]. In the large multimodal pre trained transformer Clip [17], they train the model considering the N image-text pairs of the batch of the model as positive pairs and then $N \cdot (N-1)$ pairs as negative pairs. We can visualize the positive and negative pairs in Figure 2.7

2.2 Understanding Generalization in Machine Learning

Generalization in machine learning refers to a model's ability to effectively perform on new, unseen inputs that are similar in distribution to the training data. It is a fundamental aspect in developing models that are not just tailored to their training dataset but are also capable of accurately predicting on data they have not previously encountered.

In practice, the available data for a machine learning model is typically split into two parts: a training set for building the model and a test set for evaluating its performance. It is crucial to differentiate between training error and generalization error. The training error is the error computed on the training dataset and is minimized through optimization techniques. The generalization error, in contrast, represents the expected error if the model were applied to an infinite number of unseen inputs. In practical terms, this is approximated by the error on the test set.

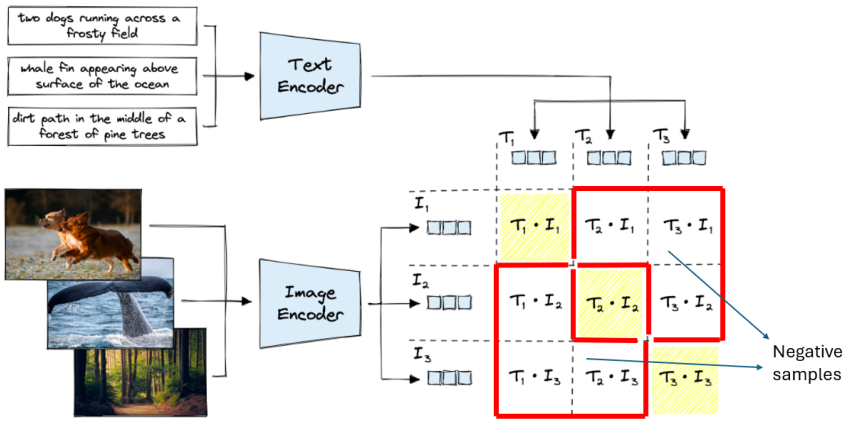


Figure 2.7: *CLIP contrastive pretraining using batch matching of image pairs. The T_i, I_j pairs where $i \neq j$ are considered negative pairs. Adjusted from* https://www.pinecone.io/learn/series/image-search/clip/

.

The ideal scenario is for a model to have both low training and generalization errors. The errors in machine learning predictions consist of bias and variance components, and minimizing the sum of these errors is key to effective model performance. However, there is a trade-off between minimizing bias and variance.

The Bias-Variance Trade-Off

The following analysis is adjusted from https://brc-deep-analytics.medium.com/bias-variance-tradeoff-855e5116a5e2. Consider a training dataset $D = \{(x_n, y_n), n = 1, \ldots, N\}$, generated by a function f, such that $y = f(x) + \epsilon$, where ϵ is normally distributed noise with zero mean: $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. The learning process aims to find an approximating model $\hat{f}(x)$ that best replicates $f(x)$. The expected squared prediction error at a point x is given by:

$$\text{Err}[x] = \mathbb{E}[(\hat{f}(x) - f(x))^2] \tag{2.1}$$

This error can be decomposed into three components:

$$\text{Err}[x] = (\mathbb{E}[\hat{f}(x)] - f(x))^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma_\epsilon^2 \tag{2.2}$$

$$\text{Err}[x] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \tag{2.3}$$

The above components can be analyzed as follows:

1. The bias is the difference between the average model prediction and the correct value. A model with overly high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

2. The variance term corresponds the variance of the approximating function f^over all

the training data D. It represents the model sensitivity to the choice of the training data D.

3. The irreducible error is produced by the noise in the data and cannot be reduced regardless of the learning algorithm.

The goal for creating generalizable machine learning algorithms is to find the bias-variance trade-off that performs optimally on unseen data.

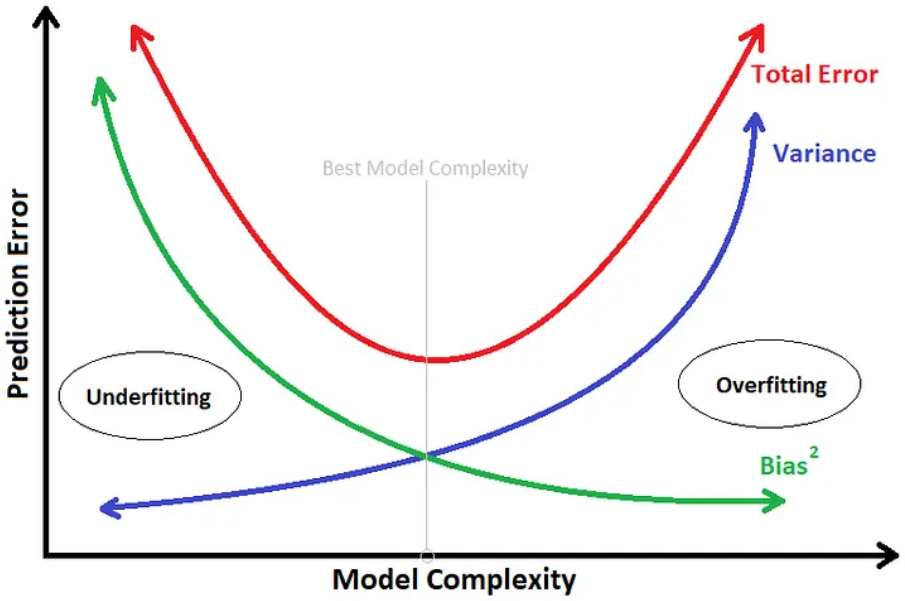


Figure 2.8: *Bias-variance trade-off in respect to model complexity and total error. Adjusted from* `https://brc-deep-analytics.medium.com/bias-variance-tradeoff-855e5116a5e2`.

Overfitting vs Underfitting

The difference between the errors during training and generalization is called the generalization gap. To minimize these errors, ML algorithms must avoid underfitting and overfitting as shown in Figure 2.9. The occurrence of these phenomena depends on the model's complexity (Figure 2.8) and the amount of training data available.

Underfitting happens when the model cannot reduce the training error. In such cases, the model fails to capture the relationship between the inputs and target outputs, resulting in low variance - high bias errors.

On the other hand, overfitting occurs when the model can reduce the training error, but the generalization gap is significant. This happens because the model is complex enough to adapt to the noise in the training data, leading to high variance-low bias errors. To combat overfitting, a common strategy is to hold back a subset of the training examples as the validation set. This set can evaluate different models to determine the appropriate model complexity.

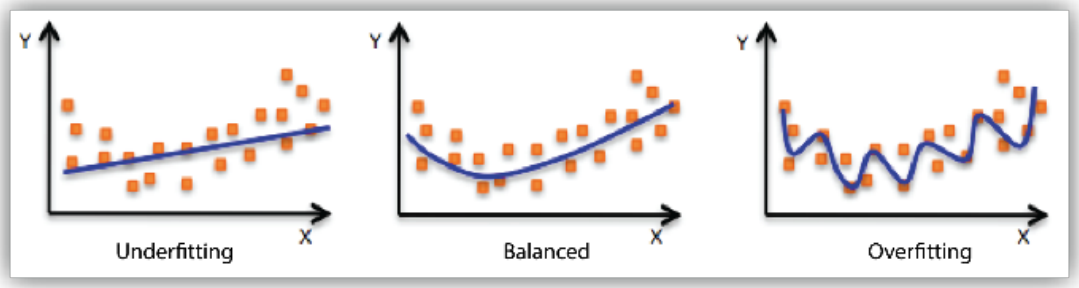


Figure 2.9: *Underfitting vs Robust Fitting vs Overfitting. Adjusted from* https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

2.3 Neural Networks

Neural networks are the foundational computational framework for many machine learning algorithms, prominently within deep learning. The distinguishing attribute of neural networks, as contrasted with conventional machine learning methodologies, is their capability to theoretically approximate any continuous function with an arbitrary degree of accuracy when at least one hidden layer is included, a proposition posited by the Universal Approximation Theorem. These networks exhibit exceptional versatility regarding the type of supervision they can receive; they can be architected to resolve any task that can be rendered as a differentiable problem, necessitating the implementation of backpropagation for optimizing the model parameters. We can observe the convergence to our function approximation through gradient descent in Figure 2.10.

Given an adequate volume of data and meticulous architectural design, neural networks thus hold the theoretical potential to solve an extensive range of problems. This characteristic enables them to address tasks across a diverse array of domains, providing solutions that can adapt and generalize effectively to different data distributions and structures, thereby presenting unparalleled flexibility and adaptability in problem-solving.

2.3.1 Architecture and Activation Functions.

A typical neural network is composed of several layers, each fulfilling unique functions:

- Input Layer: Serves as the gateway for raw data, typically represented as a one-dimensional array. Each element in this array corresponds to a specific feature in the dataset.

- Hidden Layers: These crucial intermediate layers perform most computational processing and feature transformation. Their architecture and activation functions are meticulously designed to meet the specific needs of the task.

- Output Layer: Delivers the network's final output. Depending on the task, different activation functions are employed, such as softmax for classification tasks. The

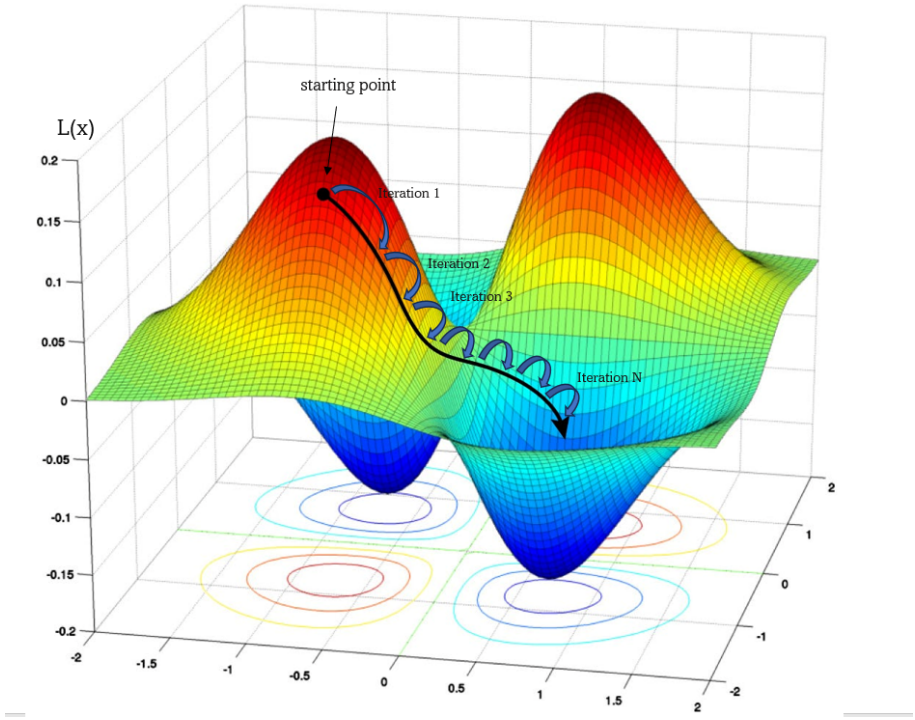


Figure 2.10: *Convergence of a loss function to a local optimum in a 3d hidden space. Adjusted from* https://medium.com/@tejovk311/optimization-challenges-in-deep-learning-a4b085d529b6.

ability to select an appropriate output layer makes neural networks highly versatile for various applications.

The hidden layers comprise linear layers and nonlinear activation functions. Linear layers are vital computational units that correlate hidden dimensions and facilitate the creation of complex functions. Without nonlinearity, models would be limited to mere linear input representations. Incorporating nonlinear functions, such as sigmoid, allows for more intricate representations. Furthermore, certain nonlinear functions possess unique properties. For instance, ReLU introduces feature selection and sparsity into the network, while tanh and sigmoid are crucial for gating mechanisms in Recurrent Neural Networks, as discussed in the following section. A hidden layer can be formulated mathematically as follows and can be visualized in Figure 2.11:

$$\mathbf{h} = f(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{2.4}$$

- $\mathbf{h}$ represents the output of the hidden layer.
- f is the activation function (like ReLU, Sigmoid, or Tanh).
- $\mathbf{W}$ is the weight matrix associated with the layer.
- $\mathbf{x}$ is the input vector to the hidden layer.
- $\mathbf{b}$ is the bias vector.

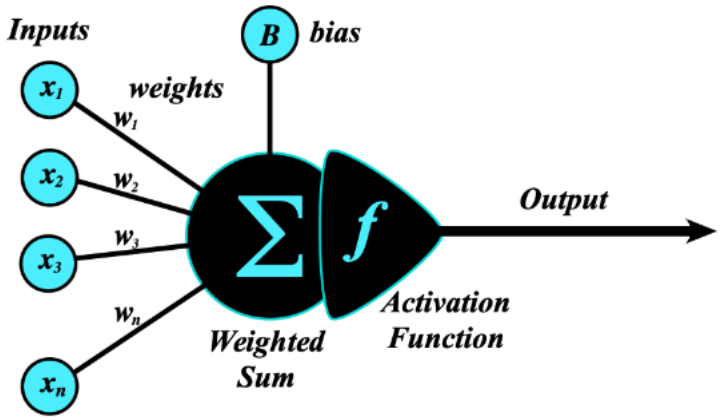


Figure 2.11: *An overview of a single neural network cell. Adjusted from* https://mriquestions.com/what-is-a-neural-network.html

2.3.2 Learning through backpropagation

During forward propagation, the input data passes through the network until a prediction is made. This prediction is then compared to the target to determine the loss value. The next step, backward propagation, optimizes the network weights to minimize this loss using gradient descent techniques. Backpropagation is a dynamic programming algorithm that efficiently calculates the gradient of the cost function concerning each weight in feedforward neural networks. This algorithm uses the chain rule, starting from the end of the network, to avoid redundant computations by leveraging intermediate computed terms. Instead of computing partial derivatives independently, backpropagation begins at the output layer and works backward to find the gradient concerning the weighted input of each layer. These terms are calculated recursively and are used to find the partial derivatives. For an input-output pair $(x_i, g(x_i))$ with a cost function of $C(y_i, g(x_i))$ where y_i is the ground-truth label, we want to compute the partial derivatives $\partial C/\partial w^l_{jk}$ concerning the weights. A detailed algorithm explanation can be found in [29]. We can visualize the propagation of derivatives through the network during training in Figure 2.12.

Optimization Strategies

Stochastic Gradient Descent, or SGD for short, is an optimization algorithm that minimizes the cost function in neural networks. This algorithm updates the network weights in small batches of training data rather than all at once, making it more efficient for large datasets. Each batch is chosen randomly, which makes it stochastic.

Adam, on the other hand, is a more advanced optimization algorithm that combines the benefits of additional extensions of stochastic gradient descent, like Adagrad and RMSProp [35], to provide a more efficient and effective way to update the weights of a neural network.

Adam optimizer uses adaptive learning rates, which means it adjusts the learning rate for each weight based on the gradients' history. This helps to address the issue of sparse

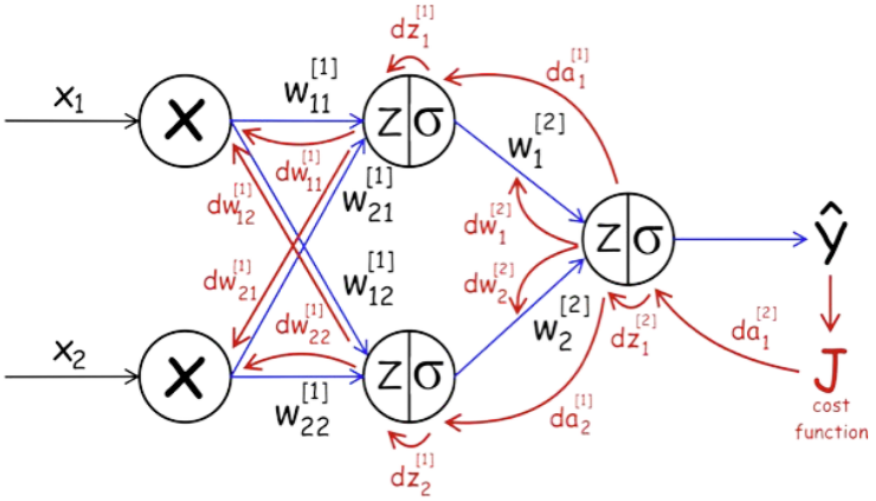


Figure 2.12: *The partial derivatives are calculated recursively through the network.*

gradients, which can be problematic for other optimization algorithms.

In summary, while SGD and Adam are optimization algorithms used for minimizing the cost function in neural networks, Adam is more advanced and adept at addressing sparse gradients, making it a more efficient and effective option for convergence.

Evaluation and High Parameter Tuning.

High parameter tuning, often referred to as hyperparameter tuning, involves adjusting various neural network parameters that are not directly learned from the data. These include learning rate, number of layers, number of neurons in each layer, and choice of activation functions. The objective is to find the optimal set of hyperparameters that yield the best performance on the validation set. Techniques such as grid search, random search, and Bayesian optimization are commonly employed.

By meticulously tuning these parameters and rigorously evaluating the model across these three data segments, we can significantly enhance its effectiveness and ensure its robustness in practical applications.

2.4 Regularization Methods

Previously, we discussed underfitting and overfitting. Neural networks, intense neural networks, rarely underfit because they often have a lot of parameters. However, overfitting is common. Techniques that help ML models avoid overfitting and generalize better are called regularization techniques. This section will describe the most widely used regularization techniques in (deep) neural networks.

- **L1 Regularization**. L1 regularization aims to penalize large values for the weights

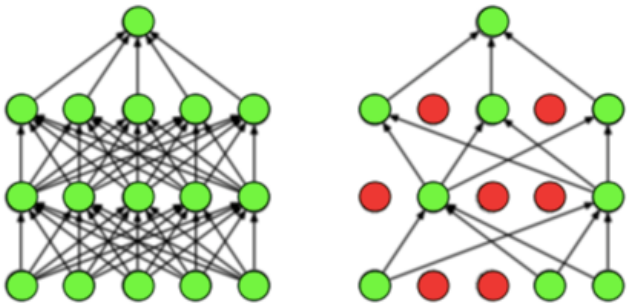
$\mathbf{w}$ of the network by adding the term $||w||$ to the loss function:

$$L' = L + \lambda ||\mathbf{w}|| \tag{2.5}$$

- **L2 Regularization or Weight Decay**. L2 regularization works similarly to L1 regularization but adds the term $||\mathbf{w}||^2$ to the loss function:

$$L' = L + \lambda ||\mathbf{w}||^2 \tag{2.6}$$

- **Dropout**. Dropout [36] chooses a random subset of neurons during each training iteration and removes it. Because this random dropout of neurons is only performed during training, this method can be seen as an efficient averaging (ensemble) of different neural networks, greatly improving generalization by forcing the neurons to learn representations independently of other neurons.



(a) Standard Neural Net (b) After applying dropout.

Figure 2.13: *A neural network with two hidden layers before (a) and after (b) applying dropout. Adjusted from* https://medium.com/analytics-vidhya/a-simple-introduction-to-dropout-regularization-with-code-5279489dda1e.

- **Early Stopping**. In early stopping, we keep one small part of the training set (development or validation set), which is not fed to the model but is rather periodically used to estimate the generalization ability of the model as it is being trained. When we observe that the performance on the validation set starts getting worse, we stop training, as this may be an indicator of overfitting.

- **Data Augmentation** Data augmentation artificially increases the training set by creating modified copies of a dataset using existing data. It includes multiple different methodologies based on our desired goal. Augmentation can be used to balance imbalanced datasets [10], to protect models against specific attacks [5], and to aid generalization using self-supervised training [15]. We often resort to small perturbation changes to the dataset like noise injection (Figure 2.4), using generative models like seen in (4.2) or even generating pseudo-labels for our new points (Figure 3.9).

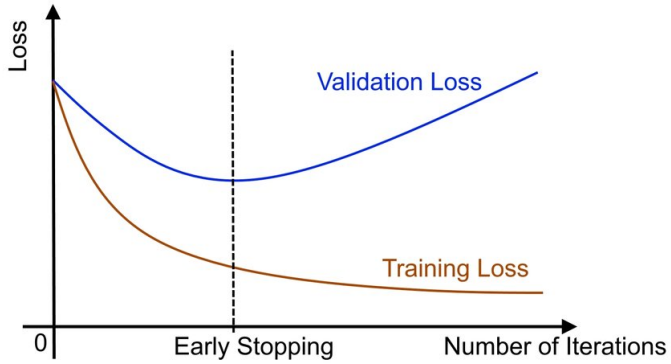


Figure 2.14: *An example of early stopping. Training stops when the validation set error starts increasing, which indicates overfitting.*

2.5 Loss Functions

Loss functions play a crucial role in the training of neural networks, acting as guides by quantifying the difference between the predicted outputs and the actual values. Among the most prevalent loss functions are Mean Square Error (MSE), Cross-Entropy Loss, and Binary Cross-Entropy Loss.

Mean Square Error (MSE) is often utilized in regression tasks. It is mathematically expressed as:

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.7}$$

where Y_i represents the actual value, $\hat{Y}_i$ is the predicted value, and n is the number of observations. MSE penalizes larger errors more, hence ensuring sensitivity to outliers and stability in gradient descent.

Cross-Entropy Loss measures the performance of a classification model outputting probability values. It is defined as:

$$\text{Loss} = -\sum_{i=1}^{n} Y_i \log(\hat{Y}_i) + (1 - Y_i)\log(1 - \hat{Y}_i) \tag{2.8}$$

where Y_i is the true label, and $\hat{Y}_i$ is the predicted probability.

Binary Cross-Entropy Loss is a variant of Cross-Entropy Loss used for binary classification tasks. It is given by:

$$\text{Loss} = -[Y \log(\hat{Y}) + (1 - Y)\log(1 - \hat{Y})] \tag{2.9}$$

where Y is the binary true label, and $\hat{Y}$ is the predicted probability.

Beyond these supervised loss functions, unsupervised, self-supervised, or semi-supervised losses can be employed. As mentioned in Section 1.5 **L1 regularization**is an unsupervised loss that introduces sparsity to the network.

In multitask learning, multiple losses can be combined for solving related tasks. Depending on the assigned weights, the model optimizes its performance by learning to

balance the trade-offs between different tasks. This approach is beneficial when tasks share underlying similarities or when learning one task can provide useful insights for another:

$$\text{Loss} = a\text{Loss}_1 + \beta\text{Loss}_2 \tag{2.10}$$

where a and β are weights assigned to the respective task losses, Loss_1 and Loss_2.

In the context of my thesis complementary non-supervised loss functions we utilized for better regularization.

2.6 Recurrent Neural Networks

In this thesis, we utilize Recurrent Neural Networks (RNNs) for diverse applications. This includes the use of encoding RNNs to transform sequences into single outputs, as well as employing decoding RNNs that take specific inputs to create sequences. Therefore, delving into RNNs is crucial for better understanding the models and experiments used.

2.6.1 Introduction to Recurrent Neural Architectures

Traditional neural architectures, primarily feedforward designs, offer limited capabilities in handling variable-length sequences. In contrast, Recurrent Neural Networks (RNNs) shine in this realm, adeptly handling non-fixed input lengths by maintaining an internal memory-like representation, the hidden state. This state evolves with each input in the sequence, ensuring a dynamic response irrespective of sequence length.

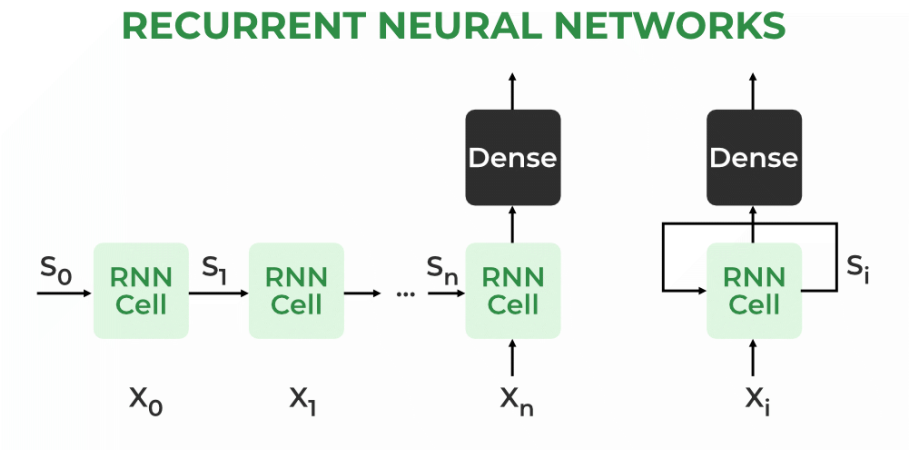


Figure 2.15: *Depiction of an RNN and its unfolded representation. Adjusted from: [37]*

Wherein, σ_h and σ_y represent activation functions, with the trainable parameters denoted by W_h, U_h, W_y, b_h, and b_y. One imperative feature is the consistent usage of these parameters across all time steps, facilitating handling sequences of any length with constant model size. The internal state, h_{t-1}, effectively acts as a conduit for historical information, encapsulating past sequence insights.

For my thesis, we will elaborate on LSTMs, a special type of RNN addressing the and their respective. They were introduced to overcome the limitations of conventional RNNs, especially the problem of vanishing and exploding gradients that arise through

their sequential structure. This problem significantly deteriorates the ability of RNNs to learn and retain long-term dependencies in data sequences. However, LSTMs include long-term and short-term memory mechanisms, enabling them to maintain information over extended periods and process it more effectively.

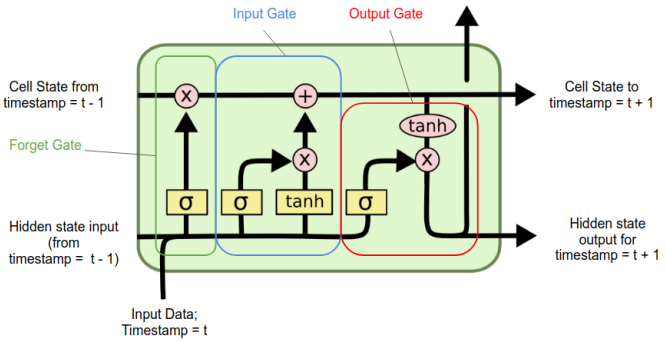


Figure 2.16: *Schematic of an LSTM cell. Adjusted from:* https://ai.stackexchange.com/questions/14326/structure-discrepancy-of-an-lstm.

The LSTM cell operates through Equations 2.11 to 2.16, wherein gate layers dictate information flow using activation functions. Given:

- Input vector at time t: x_t
- Hidden state from the previous time step: h_{t-1}
- Cell state from the previous time step: C_{t-1}

The LSTM updates are defined as:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2.11}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{2.12}$$

$$\tilde{C}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.13}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2.14}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{2.15}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{2.16}$$

Where:

- $\sigma(\cdot)$ is the sigmoid activation function.

- W_{xy} represents weight matrices, with x being the input type (x or h) and y being the gate type (i, f, c, or o).
- b_y is the bias for gate y.

Lastly, acknowledging tasks where both past and future data offer value, Bidirectional RNNs (Bi-RNNs) emerged, as outlined by Schuster and Paliwal in 1997. They encapsulate comprehensive sequence knowledge by concurrently processing data in forward and backward directions. This dual-direction processing gives the network historical and upcoming context insights, enhancing its predictive accuracy, especially in sequence prediction tasks like language modeling and speech recognition.

This bidirectional mechanism extends to more advanced recurrent neural network architectures like Long Short-Term Memory (LSTMs) and Gated Recurrent Units (GRUs), resulting in Bi-LSTMs and Bi-GRUs. These variants combine the benefits of LSTMs and GRUs—such as handling long-term dependencies and avoiding vanishing gradient problems—with the bidirectional context awareness of Bi-RNNs.

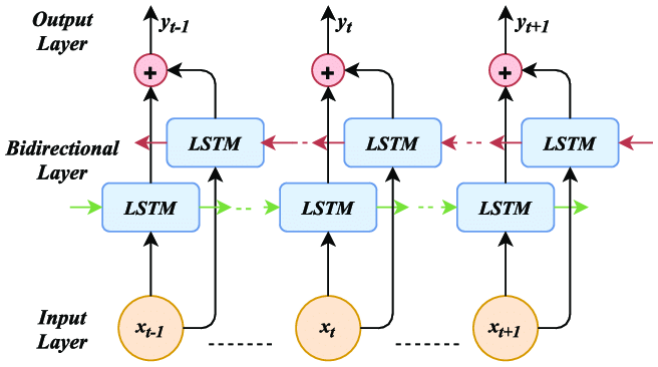


Figure 2.17: *Schematic of a bidirectional LSTM.*

2.6.2 RNNs in Decoding and Generation

Encoding and Decoding in RNNs Recurrent Neural Networks (RNNs) are exceptionally versatile in handling sequence data, playing a dual role in encoding and decoding processes. In encoding, RNNs convert a sequence of inputs into a singular, compact representation, capturing the essence of the sequence in its internal state. This ability is crucial in tasks where understanding the context of an entire sequence is necessary before making a prediction or decision.

In contrast, during decoding, RNNs perform the inverse operation. Starting from an initial state or input, they generate a sequence of outputs over time. This capability is integral to generative modeling tasks, which aim to produce coherent and contextually relevant data sequences. For example, in natural language processing, RNNs can generate sentences or translate text by decoding a condensed representation into a sequence of words.

The effectiveness of RNNs in these roles stems from their internal memory, which captures information about previous elements in a sequence, allowing the network to

make informed predictions about future elements.

Teacher Forcing

Teacher forcing is a training strategy used to speed up the convergence and improve the performance of RNNs, especially in sequence generation tasks. As shown in Figure 2.18, during training, the actual output from the previous time step is provided as input to the next step rather than the predicted output from the model. This approach guides the model with the correct sequence during the early training phases, helping it learn the dependencies between sequence elements more effectively.

However, while teacher forcing can lead to faster convergence, it may also cause a discrepancy between training and inference phases, known as exposure bias. During inference, the model only has access to its predictions, not the ground truth. This difference can lead to compounding errors in generated sequences. Strategies such as scheduled sampling can mitigate this, gradually transitioning from teacher forcing to a more autonomous generation as training progresses.

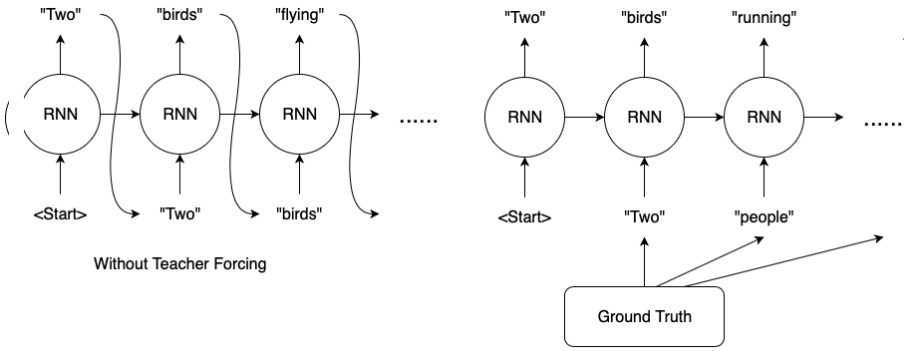


Figure 2.18: *Teacher forcing usage on sentence generation. Adjusted from* https://towardsdatascience.com/what-is-teacher-forcing-3da6217fed1c

Beam Search in Sequence Generation

Beam search is a heuristic search algorithm widely used in sequence generation tasks with RNNs. Unlike greedy search, which selects the most probable next element at each step, beam search keeps track of a fixed number of the most promising sequences at each time step. This number, known as the beam width, balances the breadth and depth of the search.

In the greedy approach for generating a sequence $X = (x_1, x_2, ..., x_N)$, at each step t, the selected token is:

$$x_t = \arg\max_{x_t} P(x_t | x_1, ..., x_{t-1})$$

Where x_t is the sequence's token at position t. This process is repeated for each token in the sequence until completion.

In contrast, given a sequence $X = (x_1, x_2, ..., x_N)$, the **beam search** algorithm selects the top k candidates at each step based on the conditional probability:

$$P(x_t|x_1, ..., x_{t-1}) = \arg\max_{x_t} P(x_t|x_1, ..., x_{t-1})$$

Where:

- x_t is the token at position t in the sequence.
- k is the beam width.

The algorithm continues this process at each step, keeping only the top k sequences based on the cumulative probability until the end of the sequence is reached.

In the examples shown, we can see that in Figure 2.19, we can see that the greedy algorithm generates the 1st sequence even though it has less total probability compared to the 2nd.

Beam search helps generate more accurate and coherent sequences than greedy search, as it explores a broader range of possibilities before making a decision. It also produces a more diverse generation of samples.

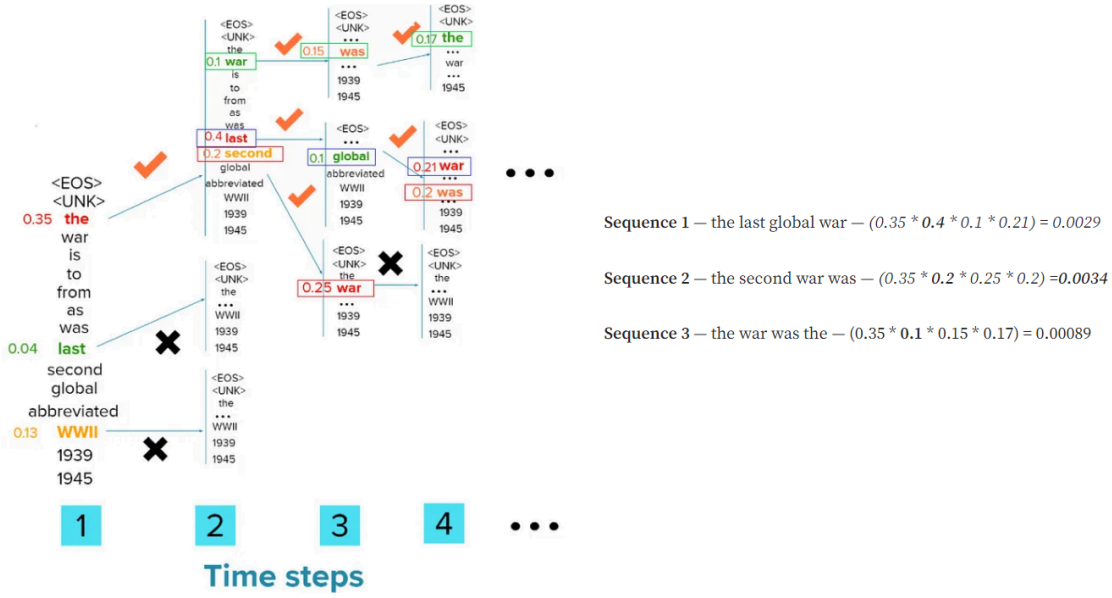


Figure 2.19: *The top 3 generated sequences using k=3 beam search in a LLM.*

2.7 Convolutional Neural Networks

Although the primary focus of this thesis is not Convolutional Neural Networks (CNNs), it is important to briefly discuss their relevance as feature extractors, especially in computer vision and object detection. CNNs have become the state-of-the-art solution for various image processing tasks due to their topological inductive bias, resulting in an innate capability to understand and automatically extract spatial hierarchies of features from images. This thesis will specifically focus on the Faster RCNN model for feature extraction.

2.7.1 General CNN Overview

A CNN comprises layers designed to automatically and adaptively learn spatial hierarchies from images, as seen in Figure 2.21. Key components include:

- Convolutional Layer: Utilizes filters to scan the input data (like images) to learn features, such as edges, textures, and other patterns.
- Pooling Layer: Helps reduce the spatial dimensions, retain only significant information, and reduce computational overhead and overfitting.
- Fully-Connected Layer: This is where neurons from the previous layers connect to decide on the image's content based on the learned features.

The strength of CNNs lies in their ability to learn filters that detect patterns, making manual feature extraction obsolete. In Figure 2.20, we can see that the features of the shallow layers detect lower-level features while the deepest layers detect more fine-grained ones. The feature extraction capabilities of the CNNs have been widely used in numerous applications, from image and video recognition to some aspects of natural language processing. They are of paramount importance for our thesis.

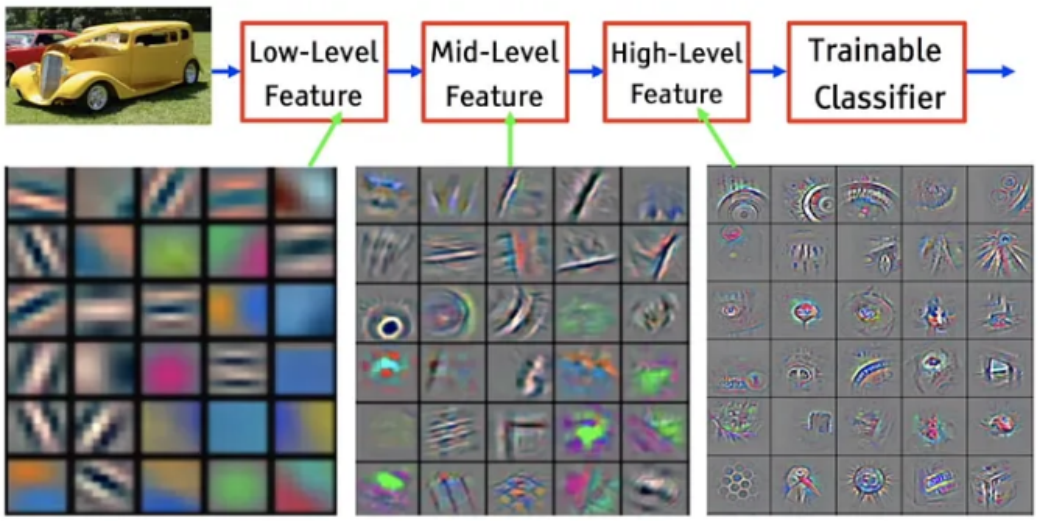


Figure 2.20: *: The features extracted from the first, second, and third CNN layers are used in image classification.* https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb

2.7.2 CNNs for Object Detection

CNNs have shown remarkable performance in tasks like image classification and recognition. In object detection, CNNs are crucial in identifying and categorizing objects within an image. Starting from basic architectures, the field has evolved toward more sophisticated and practical models. This progression includes notable models like

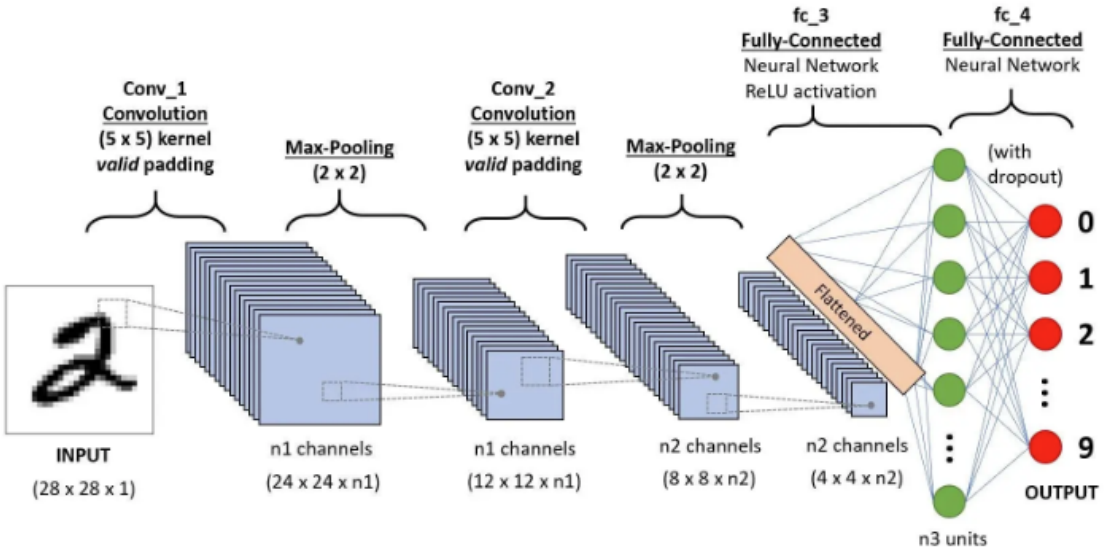


Figure 2.21: *Schematic of a 4-layer CNN model followed by a fully connected layer for digit classification. Adapted from: Source:* `https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/`

R-CNN [34], Fast R-CNN[38], and ultimately, the Faster R-CNN [39], each improving upon the last in terms of speed and accuracy. **Faster-RCNN**

The Faster R-CNN model aligns remarkably well with the key objectives outlined in our thesis. It effectively addresses the following essential conditions:

1. **Object Segmentation:** Faster R-CNN adeptly segments the image into distinct objects, ensuring focused analysis on elements of interest.

2. **Feature Extraction for Each Object:** It employs advanced feature extraction techniques for each identified object, harnessing the power of CNNs.

3. **Object Localization:** The model accurately predicts the spatial location of each object within the image, represented by precise bounding boxes.

4. **Object Classification:** Beyond localization, Faster R-CNN efficiently classifies the segmented objects into their respective categories.

5. **Prediction of Relationships and Attributes:** Modifications to the Faster R-CNN framework enable it to predict the objects and their attributes and interrelationships, as discussed in [31].

Faster-RCNN overview

The Faster R-CNN architecture [39] shown in Figure 2.22 is as follows :

1. **Backbone Network**: Utilizes a Convolutional Neural Network, often a pre-trained model such as VGGNet, to process the input image.

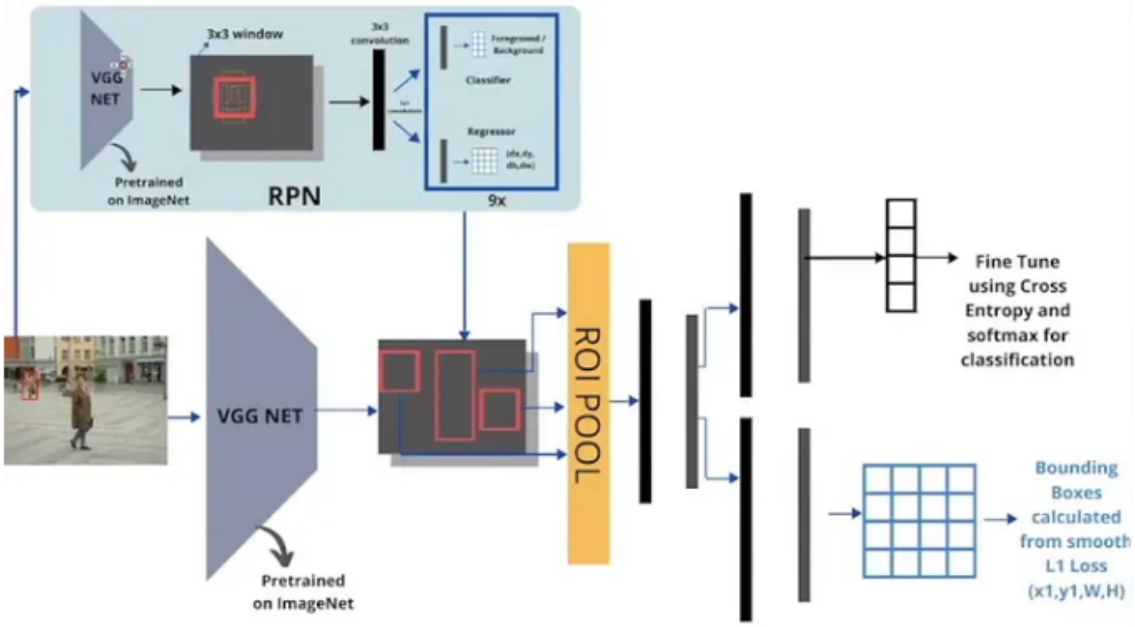


Figure 2.22: *Faster RCNN architecture. Adapted from* https://medium.com/@2003priyanshusingh/evolution-of-object-detection-rcnn-fast-rcnn-and-faster-rcnn-90cc872e6dae.

2. **Region Proposal Network (RPN)**: This network component generates region proposals based on the feature maps obtained from the backbone network.

3. **Region of Interest (RoI) Pooling**: This step aligns the features extracted by the backbone network with the region proposals, akin to the process in Fast R-CNN.

4. **Classification and Regression**: The final stages involve classifying objects within the proposed regions and fine-tuning the bounding boxes, like the processes in Fast R-CNN.

2.8 Attention Mechanisms

The utilization of attention mechanisms has been transformative in deep learning, especially since its inception in Neural Machine Translation (NMT). This paradigm allows a sequence-based model to adaptively highlight the most critical input data segments, resulting in better prediction models and a higher explainability, as seen in Figure 2.23. Attention mechanisms are beneficial when combining different inputs split into explainable parts, as seen in [20]. The widespread adoption of attention across various methods, such as recurrent networks for language processing tasks in [40], has led to the innovation of the Transformer architecture [41], which is predicated entirely on the principles of self-attention and cross-attention.

Attention is essentially a method of dynamically weighting the significance of input components of a sequence, for example, regions in an image or words in a textual sequence. The attention process can be a *Nx*1 (one-to-many), *NxM* (cross-attention), or

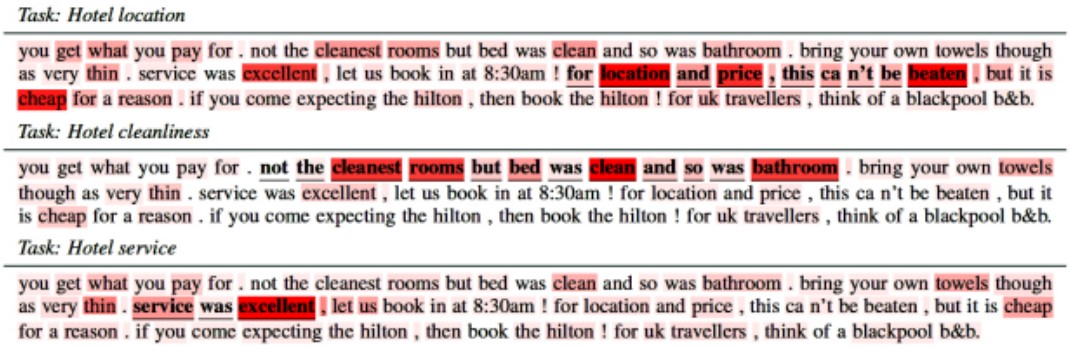


Figure 2.23: *Highlighted words concerning the task. The bolder the red color, the closer the attention values are to 1.*

NXN (self-attention). The weights are derived from the computation of alignment scores between the input components and a query vector. We can formalize the mechanism of attention and its computation and visualize it as follows:

$$s_i = \text{align}(q, x_i) \tag{2.37}$$

$$a_i = \text{softmax}(s_i) = \frac{e^{s_i}}{\sum_j e^{s_j}} \tag{2.38}$$

$$\hat{x} = \sum_i a_i x_i \tag{2.39}$$

The score s_i for each vector x_i is a scalar computed with an alignment score function.

The alignment method varies in the literature and can be implemented in multiple ways:

Name	**Alignment Score Function**
General [50]	$s^\top W_a x_i$
Additive [8]	$v_a^\top \tanh(W_a[s; x_i])$
Dot-Product [50]	$s^\top x_i$
Scaled dot-product [69]	$\frac{s^\top x_i}{\sqrt{n}}$
Content-based [28]	$\cos(s, x_i)$

Table 2.1: *Types of alignment score functions that take as input the representations $x_i \in \mathbb{R}^n$ and the query $q \in \mathbb{R}^d$.*

2.9 Glove Embeddings

The processing of word tokens and sentences is an integral part of my thesis; a vital aspect of this work involves glove embeddings as my initial feature vectors of each word. These embeddings are a powerful tool in natural language processing that allows us to represent words as numerical vectors, capturing their meaning and relationships with other words in a given text corpus. By training on global word-word co-occurrence counts,

the GloVe algorithm generates these embeddings, which can be utilized in various NLP tasks. As such, understanding how these embeddings are extracted is crucial.

In Figure 2.24, the glove embeddings are projected to a 3d space through a dimensionality reduction method and clustered into different colors. The red cluster refers to geographical terms as seen by words like borders, Maldives, and Swaziland.

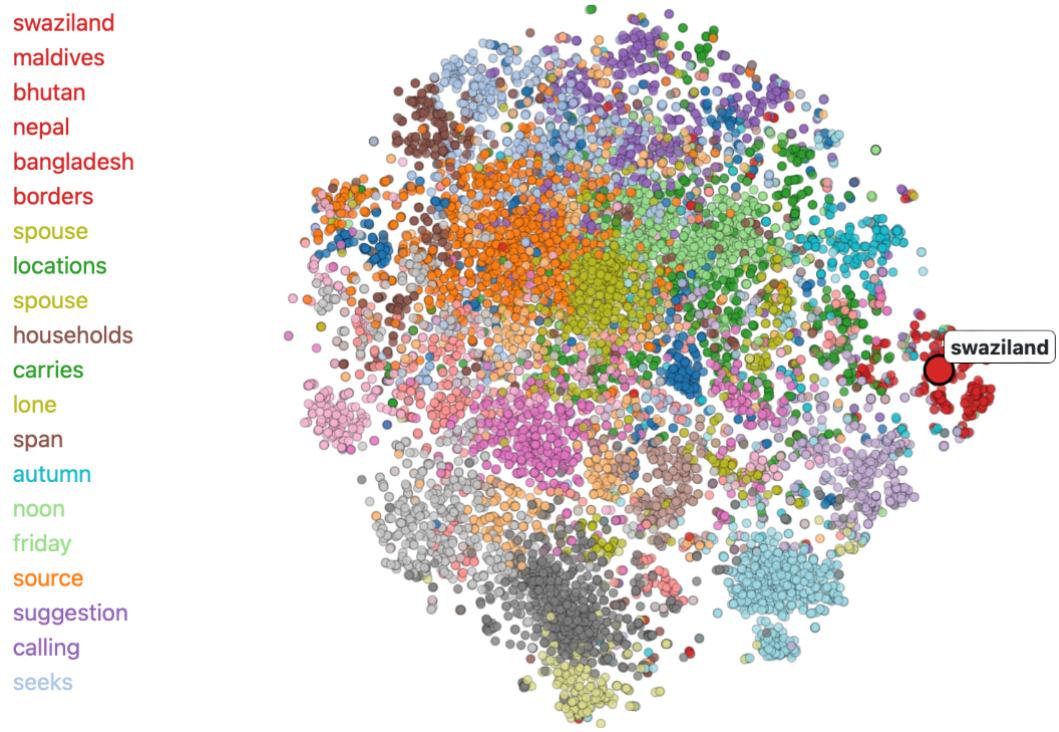


Figure 2.24: *Glove embeddings clustered in a 3d space. Adjusted from* https://blog.echen.me/2022/02/11/a-visual-tool-for-exploring-word-embeddings/.

Word co-occurences.

GloVe is essentially a log-bilinear model that uses weighted least squares for its objectives. It's based on the idea that the ratios of how often words appear together (word-word co-occurrence probabilities) are essential for representing them in a joint space.

In a large corpus, 'ice' is seen more with 'solid' than 'gas,' and 'steam' is more with 'gas' than 'solid.' 'ice' and 'steam' often appear with 'water' and rarely with 'fashion.' When looking at the ratio of these probabilities, the background noise from words like 'water' and 'fashion' is eliminated. This helps to identify characteristics unique to 'ice' or 'steam.' The GloVe model aims to create word vectors where their dot product is the logarithm of the probability that the words will co-occur. Since the logarithm of a ratio is the difference of logarithms, the model effectively links the logarithm of probability ratios with differences in word vector space. This way, it encodes meanings into the vector differences.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figure 2.25: *Word co-occurrence probability and ratio of words solid, gas, water, fashion with ice, steam.*

Loss function

The Glove objective function is mathematically formulated as follows:

$$J = \sum_{i,j} w_{ij}(f(X_{ij}) - \log(X_{ij}))^2 \tag{2.17}$$

Where w_{ij} is a weighting function that assigns more weight to rare word co-occurrences, X_{ij} is the entry in the co-occurrence matrix corresponding to the co-occurrence of words i and j, and f is a function that maps the dot product of two-word embeddings to a log space:

$$f(W_i^T W_j) = \frac{\log(X_{ij})}{\log Xi} \tag{2.18}$$

The weighting function w_{ij} is designed to down-weight the importance of frequent word co-occurrences, which tend to be less informative about the meaning of the words.

The GloVe algorithm minimizes the loss function J using stochastic gradient descent (SGD) with a fixed learning rate. The gradient of the loss concerning the word embeddings is given by:

$$\nabla_{W_i} J = 2W_i \sum_j w_{ij}(f(W_i^T W_j) - \log(X_{ij})) \tag{2.19}$$

The gradient is computed for each pair of words (i,j) in the co-occurrence matrix, and the word embeddings are updated accordingly.

Chapter 3

Visual Question Answering.

3.1 Introduction

Visual Question Answering (VQA) is a field in artificial intelligence (AI) where the goal is to develop systems capable of answering questions about the content of images. It's a complex task that merges computer vision and natural language processing (NLP), requiring machines not only to recognize elements within an image but also to understand and respond to questions posed in natural language.

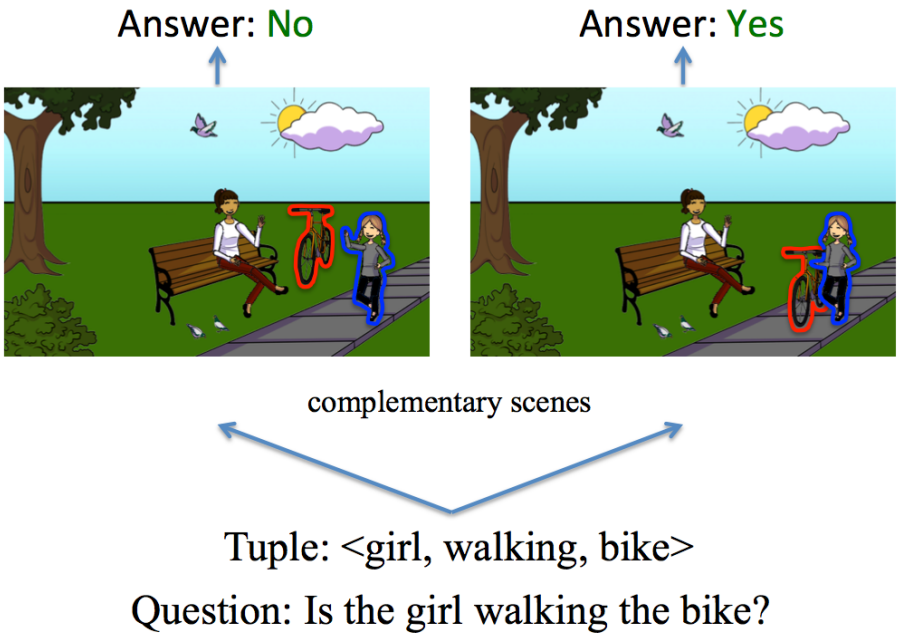


Figure 3.1: *An example of visual question answering using the same question for different images. The model leverages the relative position of the bike compared to the girl to infer two different answers. Adjusted from* https://paperswithcode.com/task/visual-question-answering.

In other words, VQA systems aim to enable machines to interpret and articulate the visual and linguistic world, somewhat akin to human perception and reasoning. The challenge VQA presents is an essential step towards achieving Artificial General Intelligence (AGI), which is the hypothetical ability of an AI system to understand, learn, and apply knowledge in an autonomous, flexible manner across a breadth of tasks similar to human cognitive capabilities.

To achieve VQA, AI systems must not only "see" or "read" but also "understand" and

"reason". This requires the AI to draw inferences that are not explicitly stated and often rely on common-sense knowledge or contextual subtleties that are second nature to humans but exceedingly complex for machines to mimic. For example, in Figure 3.1 we can see different answers being given for the same question and similar but different images. In order for the model to infer the same answer in both cases, it must not only accurately classify the objects but also, by their relative positions and the girlds hand placement, infer that their relationship is actually “walking", which is metaphorically being used for sliding the bike on the road.

Traditional VQA approaches were initially focused on generalized solutions , extracting generalized embeddings from images and questions from CNN and RNN models as seen in 3.2. Recent advancements have seen the integration of object detection frameworks, such

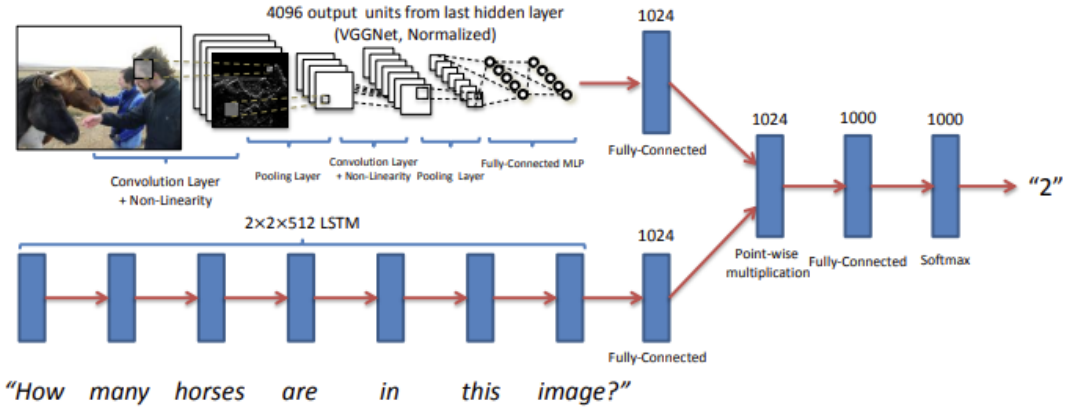


Figure 3.2: *A schematic of a simple VQA model using VGGNnet for the image, LSTM for the question, and simple fusion using point multiplication.*

as Faster R-CNN, which pinpoint and classify objects within an image, providing a rich, detailed understanding of the visual content. Concurrently, incorporating GloVe (Global Vectors for Word Representation) embeddings enriches the system’s grasp of semantic language nuances, enabling a more sophisticated interpretation of the questions posed. The initial implementation of such methods is the bottom-up, top-down attention model [23] and has been used in several other Transformer [41] based models like [18, 20, 42].

For our thesis, we will examine the bottom-up and top-down attention model and use it as a baseline for our implementations. Notably, the UpDn model and subsequent models were initially developed as a versatile framework for various visual-linguistic tasks, including image captioning, visual dialogue, and image-text retrieval.

This inherent versatility renders the UpDn model an ideal candidate for evaluation within Visual Question Answering. Its capacity to process and interpret visual-linguistic information makes it a potent tool for this investigation, particularly in terms of its generalization capabilities to data it has not previously encountered.

The selection of the UpDn architecture over other comparable but more efficient transformer-based frameworks, such as those documented in [21, 20, 42], is justified by several factors:

- The Bottom Up Attention model was developed as a task-neutral visuolinguistic architecture, offering a level of generality comparable to its transformer-based peers.
- The generalization datasets we use are GQA-OOD [2] and VQA-CPv2 [25] . Architectures like Vilbert [20], VisualBert [42], and LXMERT [21] have undergone pre-training on the COCO captioning dataset [43] and the GQA dataset[19]. The VQA dataset and its variants, including VQA-CPv2, contain images mostly from the COCO dataset. Similarly, GQA-OOD is part of the GQA dataset. Consequently, as indicated in [2], there exists a dataleak of test data in the pre-training phase for these models, which could potentially skew the results regarding their generalization capabilities.
- The computational demands of these models exceeded the capabilities of the available hardware resources.

3.2 Bottom-Up and Top-Down Attention Model

Our baseline, the Bottom-Up and Top-Down Attention (UpDn) model [23], integrates two distinct yet complementary mechanisms to effectively process visual and language data. This model is adept at tasks involving image and language understanding, such as image captioning and visible question answering (VQA). The three critical components of the UpDn model are:

3.2.1 Language Encoder

The language encoder utilizes GloVe embeddings and Recurrent Neural Networks (RNNs) to construct question embeddings. The GloVe (Global Vectors for Word Representation) [28] embeddings provide a pre-trained word representation, capturing global word-word co-occurrence statistics from a corpus. The RNNs then process these embeddings sequentially, creating a question embedding for fusion with the image representations as the output of the last LSTM layer for the Nth word passed through a non-linear layer. We can observe the architecture of the language encoder in Figure 3.3.

3.2.2 Bottom-Up Mechanism

The bottom-up mechanism segments the image through Faster R-CNN to identify important objects as seen in Figure 3.4. Each identified object is associated with a 1024-dimension feature vector, its object class and its bounding box.

3.2.3 Top-Down Mechanism

The top-down mechanism applies attention guided by the linguistic context (i.e., the question embedding) to the image regions identified by the bottom-up mechanism. This process involves weighing the importance of different areas concerning the question content and the answer classification. It selectively focuses on parts of the image that are

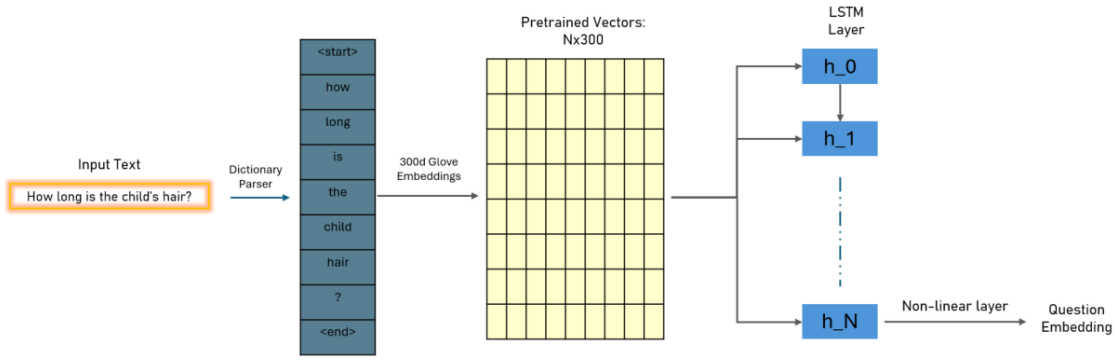


Figure 3.3: *Language encoder of the Bottom Up- Top Down attention architecture.*

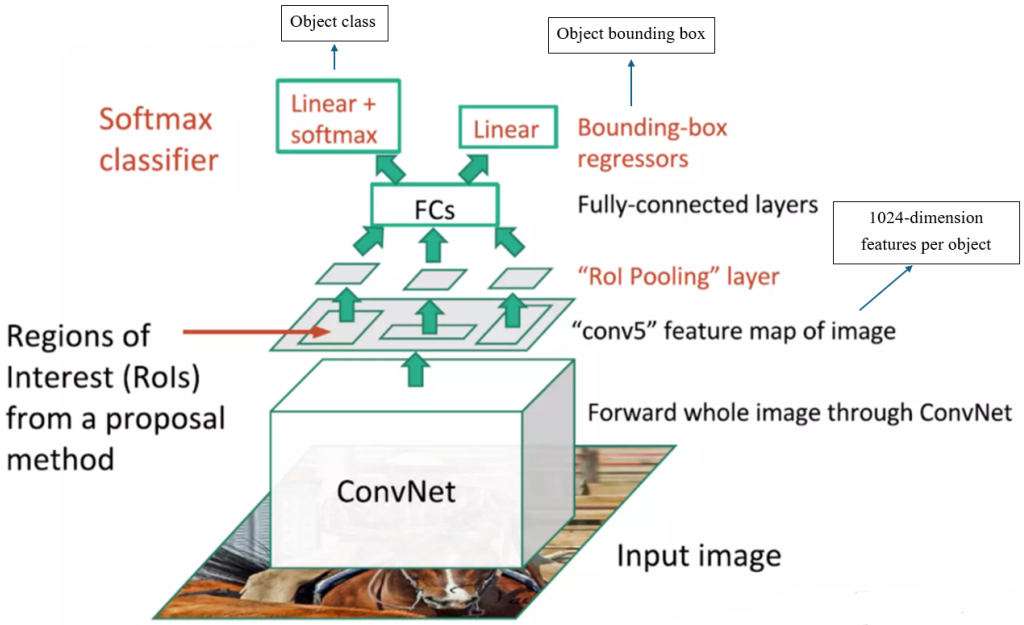


Figure 3.4: *Feature extraction using the Faster-RCNN framework. Adpated from* `https://www.slideshare.net/JinwonLee9/pr12-faster-rcnn170528`

more relevant to the question, thereby integrating language-driven attention with visual features. Attention is bottom-up, top-down attention is formulated mathematically as follows:

3.2.4 Integration and Output

As seen in Figure 3.5, the model combines the outputs of the bottom-up and top-down mechanisms to generate a final prediction using the element-wise product of the final image and question representations and passing it through a non-linear layer. This integration allows the model to correlate important image regions concerning the language

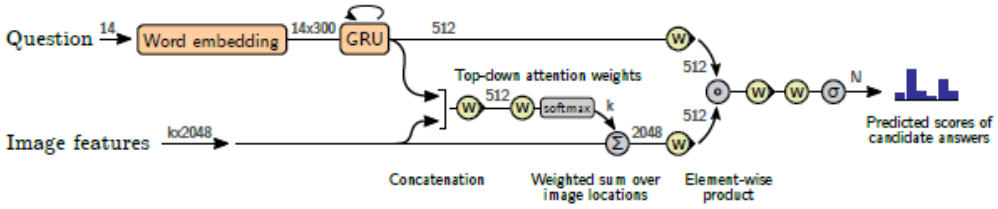


Figure 3.5: *Bottom Up- Top Down attention architecture. Adapted from: [23]*

content and create adaptable visolinguistic embeddings depending on the task to solve.

3.2.5 Summary

The Bottom-Up and Top-Down Attention model [23] effectively merges visual perception with language understanding, enabling detailed analysis of images guided by linguistic context. It represents a significant advancement in AI, particularly in tasks that require a joint understanding of visual and textual information.

3.3 Generalization in Visual Question Answering

Building on the robust foundation established by the bottom-up and top-down (UpDn) attention model, this thesis positions the UpDn framework as a baseline to examine various generalization methods in Visual Question Answering (VQA). It is imperative to assess the adaptability and scalability of VQA systems in diverse scenarios, a pursuit that this work undertakes with vigor.

As we transition from a theoretical exploration to an empirical investigation, the subsequent chapters will detail these out-of-distribution generalization methods and the datasets they were implemented on. The findings promise to contribute valuable insights to the field of VQA, paving the way for more intelligent, flexible, and capable AI systems in the future.

3.3.1 Out of distribution datasets for Visual Question Answering

3.3.2 VQA-CPv2 dataset

This dataset originates from a different split of the VQA v2 dataset[24]. The original VQA v1 dataset [44] was deemed to be extremely reliant on language. As a result, in [24] they created a second balanced version of the VQA dataset by collecting complementary images such that every question is associated with not just a single image but a pair of similar images resulting in two different answers. The dataset's metric is the accuracy in the 3 primary answer categories “Yes/No, "Number", “Other", and “Other" refers to all the questions that are not number based or comprise 48% of the dataset.

However, several studies showed that models trained on VQA v2 were still heavily driven by superficial correlations in the training data and lacked sufficient visual grounding. In [25], they found the cause of the above behavior to reside in the fundamentally

problematic nature of IID train-test splits in the presence of solid priors. Hence, they proposed a new setting for VQA, where train and test sets have different prior distributions of answers for every question type, as shown in Figure 1.

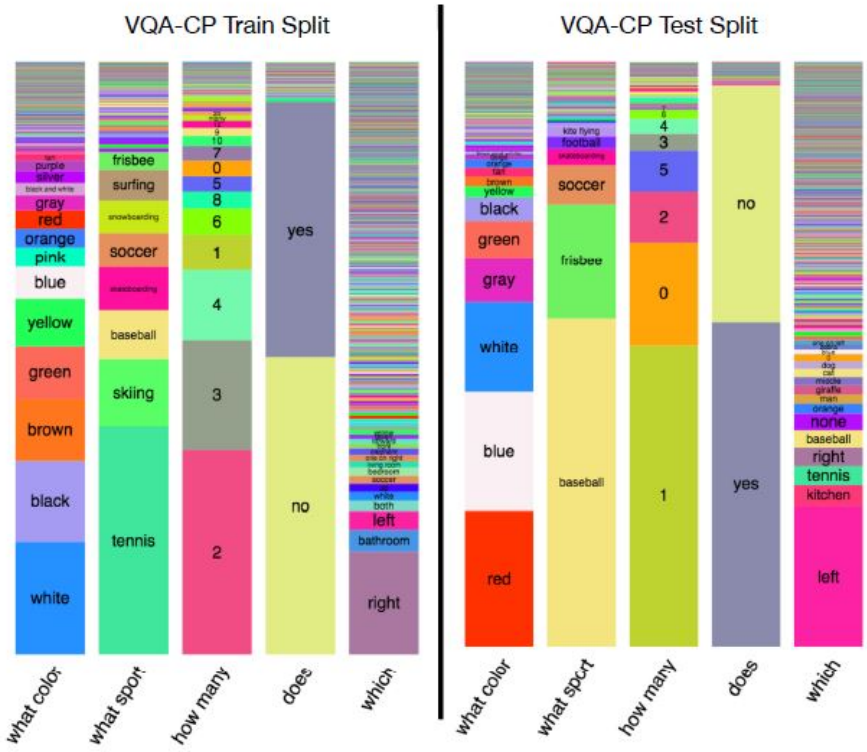


Figure 3.6: *Distribution of answers per question type varies significantly between VQA-CP v2 train (left) and test (right) splits*

Even though the VQA-CP v2 has been the cornerstone for various benchmarks, it has received extensive criticism in [1, 4, 2]. The main points of criticism can be summarized as follows:

- The VQA dataset comprises five different question categories that can be summed up in 3 (Yes/No, Number, Other) in evaluation. In [4], they assume that the test distribution shift is not only different but the inverse of the train distribution. By predicting the labels of the inverse train distribution with a weighted random selector, they achieve SOTA results in "Number" and "Yes/No" questions without even having a model; at the same time, they perform horribly in "Other" (0.02 accuracy). Hence, they propose that the only viable metric for evaluating the dataset is accuracy in "Other." A similar result is concluded in [1]. Consequently, in contrast to older methods, most of the newest models in VQA-CP focus on that proposed metric.

- Knowing the distribution shift beforehand leads to model architectures designed to exploit it and achieve superficial performance, defeating the purpose of OOD generalization where the distribution of the test data should be unknown.

- VQA-CPv2 lacks an ID validation split. Retraining in VQA v2 is not a valid method for measuring ID performance. However, this can be easily fixed by holding a part of the training data as an ID validation spit.

- It lacks a proper performance metric besides accuracy that can better capture visual grounding and generalization improvements.

3.3.3 GQA-OOD dataset

The GQA dataset[19] was developed as a continuation to CLEVR[45] to evaluate visual grounding and reasoning and compositional question answering in real-world scenarios. It was constructed from scene graphs and images from the Visual Genome dataset for visual explanations. It contains extra annotation information for images (scene graphs) and semantic programs for the questions. Its question answering is primarily grounded in the image contents is based on composite questions relevant to certain attributes and relations between objects as seen in Figure 3.7. In [2], a new dataset split is proposed,

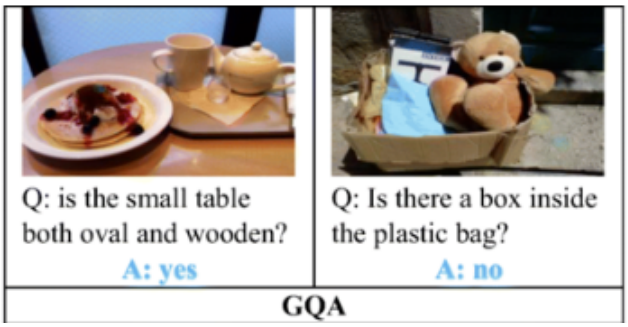


Figure 3.7: *GQA dataset examples*

which addresses the issues residing in VQA-CPv2 and tests the performance of VQA models in both ID and OOD conditions. GQA-OOD contains both ID and OOD test/val splits.

The distribution shift of the dataset is implemented as follows:

- **Head of the Distribution**: This portion comprises the most common or frequent answers within a specific context or local group. These are the answers that occur more frequently than a defined threshold. The head represents the ID scenario where the answers are within the expected distribution of the dataset.

- **Tail of the Distribution**: In contrast, the tail consists of the rarest answers in the same local group. These answers occur less frequently, falling below the threshold defined by the average sample count for the group. Specifically, tail classes are defined as classes i with $|a_i| \leq \kappa \times \overline{a}$, where $|a_i|$ is the number of samples belonging to class i, $\overline{a}$ is the average sample count for the group, and κ is a factor set empirically (e.g., $\kappa = 1.2$). The tail represents the OOD scenario, capturing the answers that are not commonly expected or are outliers in the dataset.

Additionally, new performance metrics are introduced:

- Acc-tail: The accuracy on OOD samples, which are the samples of the tail of the answer class distribution.

- Acc-head: The accuracy of the distribution head.

- Acc-all: The ID accuracy.

- $\Delta = \frac{Acc_{Head} - Acc_{Tail}}{Acc_{head}}$: Δ shows the discrepancy of performance between the ID and OOD settings.

The dataset's effectiveness was tested on several general SOTA VQA models [21, 20, 23], which seemed to suffer a 10% drop in accuracy-tail. Additionally, most successful debiasing models in VQA-CPv2, such as [14, 13, 30], seemed to underperform under those settings.

3.3.4 Generalization Methods in the literature.

The methods for generalization in VQA tested in the datasets above can be split into the following categories : Language debiasing ensemble-based methods, Data augmentation methods, Enhancing visual information methods, and Answer Reranking.

Ensemble based methods

The ensemble-based methods' LMH[14], AdvReg[30], Rubi [13] main goal is to utilize a language-only model that theoretically incorporates the entire language bias of the dataset and tries to force the fusion model to predict different answer distributions while simultaneously achieving the primary task (indicating the correct answers). At test time, they keep only the multimodal part of the model to perform the inference. We can see the architecture of such models in 3.8. The most notable of those methods is LMH, which is

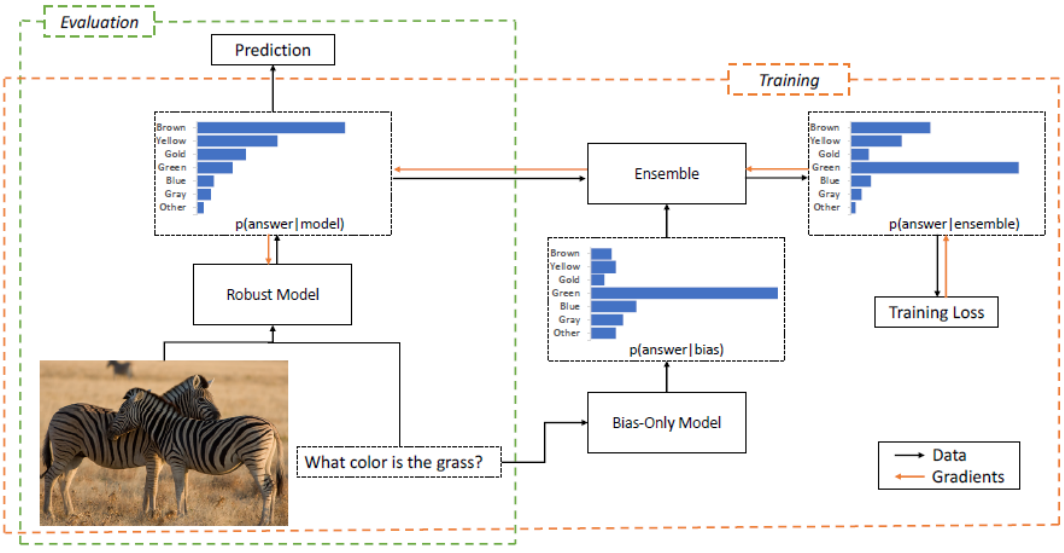


Figure 3.8: *An Overview of the ensemble-based models Adapted from [14].*

based on the following mathematical formulation: Let $x_{\sim b}$ be a view of the example that captures all information about that example except the bias. Assume that $x_{\sim b}$ and x_b are

conditionally independent given the label, c. Then to compute $p(c|x)$ we have:

$$\begin{aligned} p(c|x) &= p(c|x^b, x^{-b}) \\ &\propto p(c|x^{-b})p(x^b|c, x^{-b}) \\ &= p(c|x^{-b})p(x^b|c) \\ &= \frac{p(c|x^{-b})p(c|x^b)p(x^b)}{p(c)} \\ &\propto \frac{p(c|x^{-b})p(c|x^b)}{p(c)} \end{aligned}$$

The $p(c, x)$ is the output of the ensemble of the language and the multimodal model that will be used as the primary loss function. The $p(c|x^{-b})$ is the unbiased multimodal model they use for inference. The $\frac{p(c|x^b)}{p(c)}$ is the language only biased model. By maximizing the left term and minimizing the prediction of the language model, they theoretically produce an unbiased multimodal model that is not constrained by spurious language-based correlations.

Even though those methods have achieved extraordinary overall accuracy in VQA-CPv2, they have received severe criticism[2, 1, 4, 3]. Through multiple experiments, it has been observed that they heavily rely on improving performance in VQA-CPv2 through "Yes/No" or "Number" questions by knowing the distribution shift beforehand and predicting the inverse answer distribution while simultaneously underperforming in ID settings.

Data augmentation methods

Data augmentation methods rely on two different methodologies.

1. Expand the dataset by automatically generating new samples in the image or question space with the same or different ground-truth answers. Then, use those new samples to train the model or perform a second regularization task(by an added loss function).

2. Use adversarial perturbations to create more robust multimodal representations

3. Generate new visually generated questions explicitly or implicitly in joint training.

Distinct sample generation

The most notable data augmentation method, CSS[9], relies on creating counterfactual samples. First, they find the most important objects of the image(VSS) or sentence(QSS). They mask all the unimportant objects or words and pass the new positive sample through the model, obtaining the K-best answers. The complement of the K-best answers is the "ground truth" of the negative samples(Dynamic Answer Assigning). A negative sample is an image-sentence pair with masks on the critical objects. Finally, they train our model with the counterfactual/negative sample. We can see the negative examples for Q-CSS or V-CSS in Figure 3.9

Lastly, in [9], for counterfactual samples, the supervised loss is computed by defining a new VQ pair $(I^-; Q)$ from counterfactual input I^- and question Q. Ground truth answers are assigned using dynamic answer assigning (DA ASS), which involves: feeding $(I^+; Q)$ to the VQA model, obtaining a predicted answer distribution $P^+_{\text{vqa}}(a)$, selecting top-N answers a^+, and defining $a^- = \{a_i \mid a_i \in a, a_i \notin a^+\}$. In cases where $a \subseteq a^+$, a^- is set to empty.

The most notable aspect of their research lies in their method of identifying critical question tokens or visual objects. For question tokens, they determine their relevance by calculating the cosine similarity with the answer using GloVe embeddings. For visual objects, they employ Grad-CAM [46] to focus on the object that yields the highest gradient concerning the ground truth answer. CSS[9] seems to improve uniformly across all metrics and baseline models in VQA-CPv2.

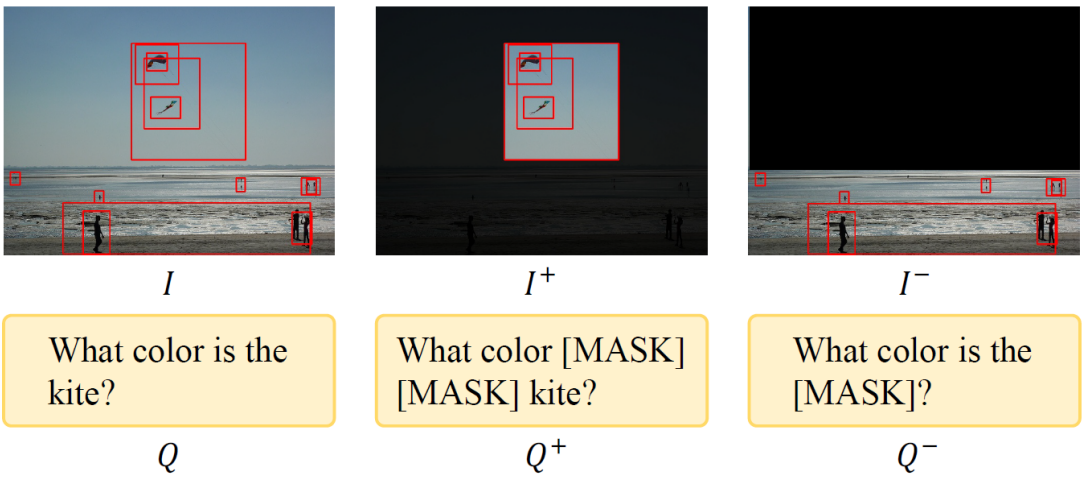


Figure 3.9: *The original image, sample with a negative question, and sample with a negative image.*

Another method utilizing counterfactual samples is gradient supervision[15]. It has not yet been utilized in our datasets, but it has been used in many different datasets, performing very well; it assumes we have a method of obtaining counterfactual samples for our dataset and is formulated as follows. Let the gradient of the network f with respect to its input at a point x_i be denoted as $g_i = \nabla_x f(x_i)$. The Gradient Supervision (GS) loss is a regularization loss L_{GS} that aims to align g_i with a "ground truth" gradient vector $\hat{g}_i$, and is defined by the equation:

$$L_{GS}(g_i, \hat{g}_i) = 1 - \frac{g_i \cdot \hat{g}_i}{\|g_i\|\|\hat{g}_i\|} \tag{3.1}$$

This equation represents a cosine distance between g_i and $\hat{g}_i$. For a pair of counterfactual examples (x_i, y_i) and (x_j, y_j), the "ground truth" gradient at x_i is given as $\hat{g}_i = x_j - x_i$. This vector indicates the transformation in the input space that should change the network's output from y_i to y_j. Minimizing this equation encourages the network's gradient at the training points to align with this vector.

SSL[8] creates new sample pairs by random matching of images and sentences in the dataset as seen in Figure 3.11. Since the probability of fitting random image/sentence

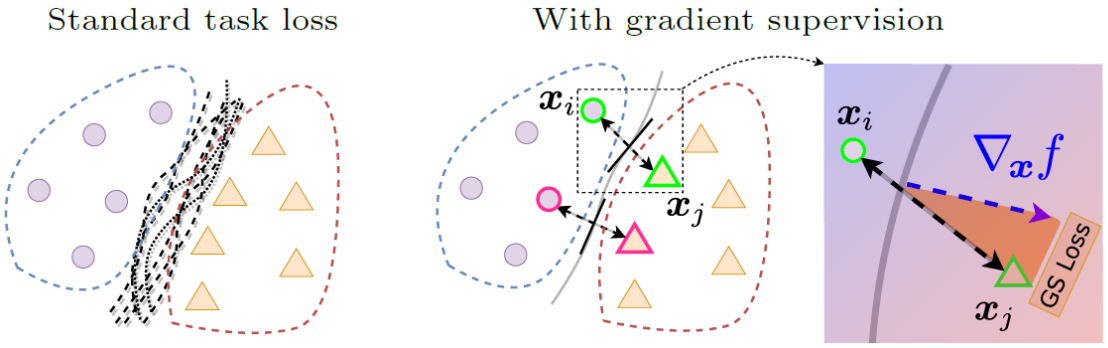


Figure 3.10: *The original image, sample with a negative question, and sample with a negative image.*

pairs is close to 0, a loss function is introduced, forcing the model to predict an empty answer distribution for "irrelevant" samples. Their method minimizes negative sample sensitivity to the primary ground truth answer.

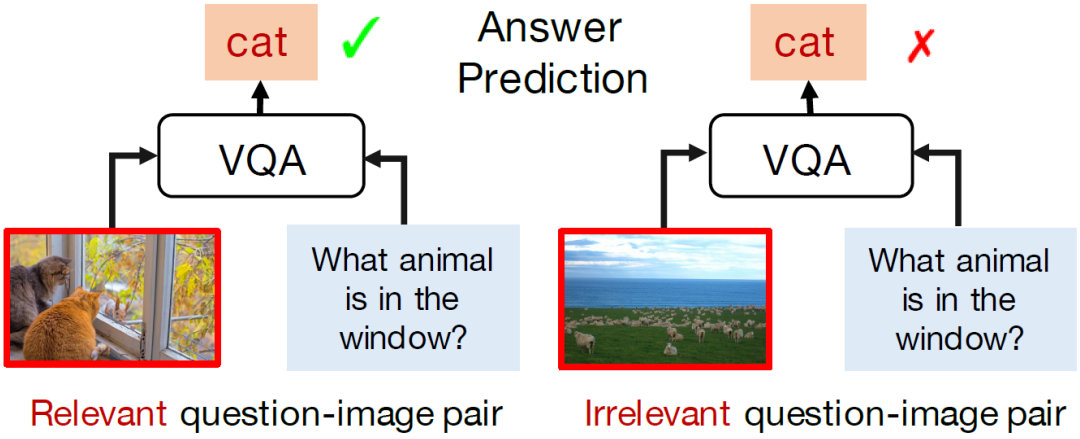


Figure 3.11: *Schematic of the SSL framework.*

MUTANT[10] is a recent model with exceptional results that relies on four methods. The first and most important is a data augmentation method that creates "mutations" of original Image-Question pairs. They use Object Removal (for number question types), Color Change (for color question types), Negation Adversarial Substitution, and Word masking as seen in Figure 3.12 and automatically change the answers to fit the corresponding mutation. It achieves significant performance increase across all metrics but is question-type specific.

SwapMix[7] is an augmentation-based model used on the GQA dataset[19]. At first, they find the most important objects of the image straight from the original program and scene graph annotations and match them with the fast-rcnn extracted objects by choosing the most similar objects based on the IoU overlap. They create new augmented samples by exchanging "non-important" objects with similar objects (same class or attribute) from the corpus, as shown in Figure 3.13. They observe that VQA models heavily rely on irrelevant image subcontext and change their answers by perturbating unrelated regions

Mutation	Q	A	Q_{mutant}	A_{mutant}
Negation	Is this bread?	yes	Is this not bread	no
	What is the color of the woman's shirt?	black	What is not the color of the woman's shirt?	white
	Are there deciduous trees?	no	Are there no deciduous trees?	yes
	Is there a boy?	no	Is the no boy?	yes
Adversarial	Who is riding the boat?	man	Who is riding the desk	*"can't say"*
	How big is the plane?	large	How big is the book?	*"size"*
	How many pillows are on the bed?	four	How many pillows are on the table?	*"number"*
Masking	What type of drink is being displayed?	wine	What type of [MASK] is being displayed?	*"beverage"*
	How many bins?	two	How many [MASK] ?	*"number"*
	What is the green stuff on the sandwich?	lettuce	What is the green stuff on the [MASK]?	*"food"*

Figure 3.12: *Mutant question type specific augmentation methods.*

of the image. They utilize the above augmentations to mitigate the sensitivity of VQA models in similar object swappings. Swapmix **does not** seem to improve results in our datasets and performs better in cases where actual perturbations of relevant objects are implemented.

3.3.5 Adversarial perturbations

VILLA [6] uses adversarial perturbations to create more robust visiolinguistic representations. It is trained with a more computationally efficient version of PGD called FreeLB [47]. However, it does not seem to improve the accuracy on any metric in VQA-CPv2 with LXMERT [21] as its baseline.

Mango [5] is an improvement on VILLA. As in VILLA, it is based on adding learnable noise to the embeddings, as seen in Figure 3.14. However, it uses a method developed in [48], where the adversarial noise is learned as a neural module applied to Gaussian noise instead of being known by PGD or its variants.

An overview of the method indicates that while there have been reported enhancements in overall accuracy across our datasets, they have not provided specific accuracy details for the 'Other' category in VQA-CPv2 and the Acc-tail in GQA-OOD. Additionally, the absence of accessible code for both models means we cannot replicate the studies and present corresponding results.

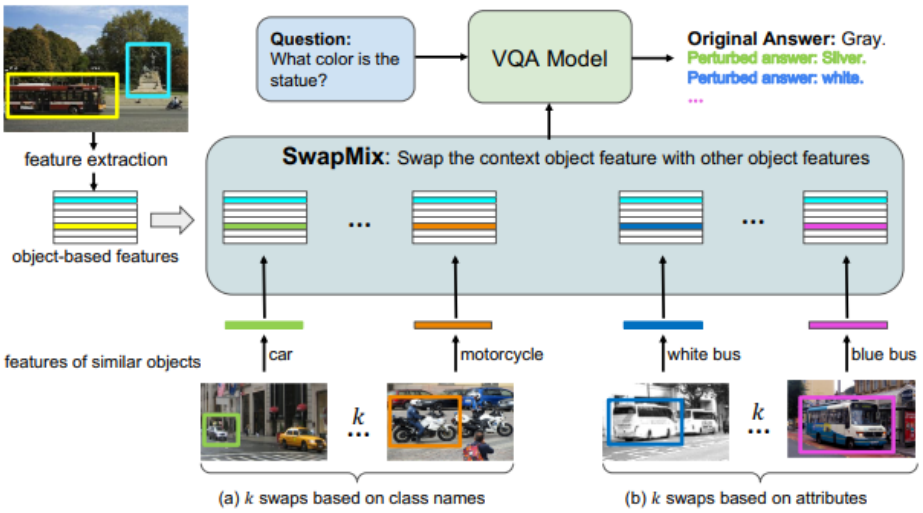


Figure 3.13: *A schematic of the object swapping of based on similar objects or similar attributes*

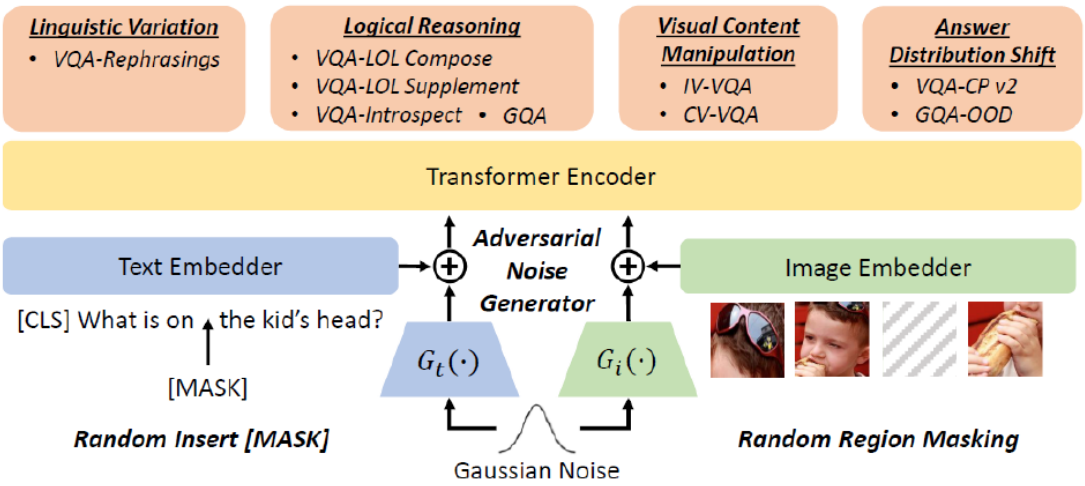


Figure 3.14: *Schematic of Mango utilizing adversarial perturbations in images and text for VQA.*

Enhancing visual information methods

In SCR [49], the base UpDn [23] VQA system first detects a set of objects and predicts an answer. They then analyze the correct answer's sensitivity to the detected objects via visual explanation (either from captions or human attention maps) and extract the most influential object, further strengthened via an influence enhancement loss. They also analyze the competitive incorrect answers' sensitivities to the most influential object and criticize the sensitivity until the VQA system answers the question correctly.

In the "X-GGM" research paper [31], the authors propose a two-step process for enhancing model performance using image data. Initially, the model identifies all distinct

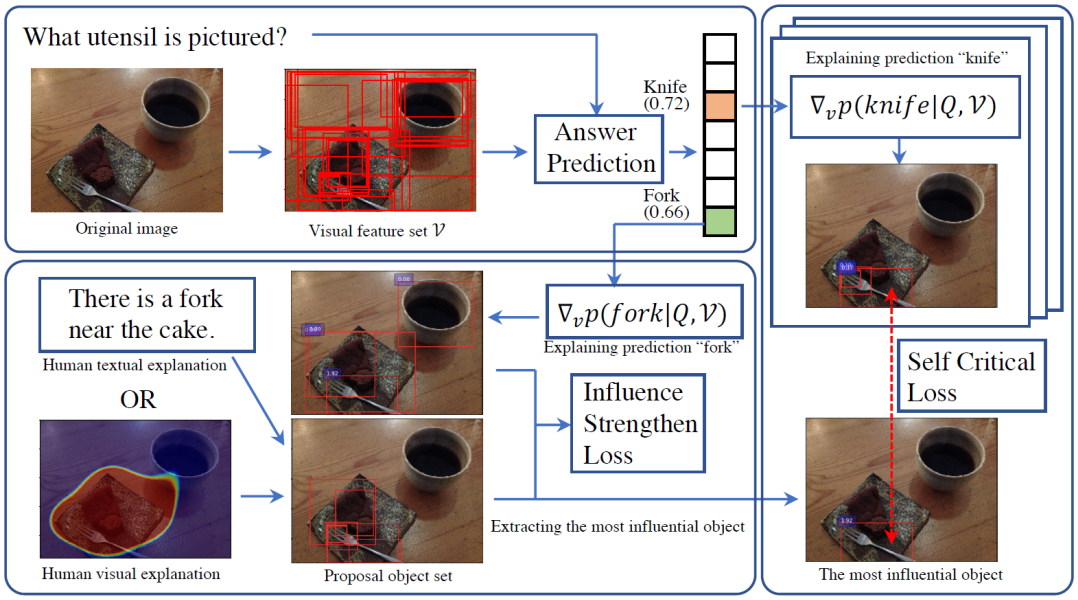


Figure 3.15: *In the UpDn VQA system, objects are identified and analyzed for their influence on the correct answer ('Fork'). The most influential object is strengthened using the 'influence strengthen loss' method. Finally, potential incorrect answers ('Knife') are assessed based on their response to the most influential object until the VQA system accurately answers the question. The numerical values on each bounding box indicate the sensitivity of the answer to the associated object*

objects in an image, each tagged with a unique attribute, such as "green bowl." This results in a one-to-one mapping between each object and its attribute. The next step involves creating a correlation matrix. This matrix is built by comparing the objects' BERT embeddings [**?**] and attributes to establish their similarities, forming an NxN "ground truth" matrix. The model then employs one of two fusion methods in each iteration, building upon the foundation set by the LXMERT model. The first method involves predicting this "ground truth" matrix using multimodal features and object embeddings. The second method aims to predict the object embeddings based on the ground truth matrix and multimodal features. Additionally, the model introduces Gaussian Noise to the predicted relational matrix or the embeddings, enhancing its robustness.

This approach has demonstrated notable improvements in key metrics across both datasets used in the study. However, a challenge arises in accurately evaluating the model, primarily because the researchers have yet to release their code. Furthermore, the published results of their baseline model, LXMERT, show inconsistencies compared to other field studies.

Answer Reranking

The first notable model is called RankVQA[11]. After passing the image-text pair through a VQA model, they selected the K-candidate answers. Additionally, the image caption is produced by combining question information offline. Finally, an Answer Reranking Module computes the re-rank score by measuring the answer-image matching

degree and the answer-caption matching degree and back-propagates the score to guide the VQA module.

A similar model[12] uses answer re-ranking [11] and combines it with a method called Visual Entailment [50]. In SAR[12], they use a pre-trained VQA model to select the best K-answers, and they build a new dataset MxK of tuples (Im, ques, candidate answer) with ground truth as the probability given by the first VQA model to the candidate answer. Then, by combining the question and answer into a single caption, they use Visual Entailment to produce a new re-ranked answer distribution. This model, along with Mutant[10], is by far the most successful in the literature but computationally expensive.

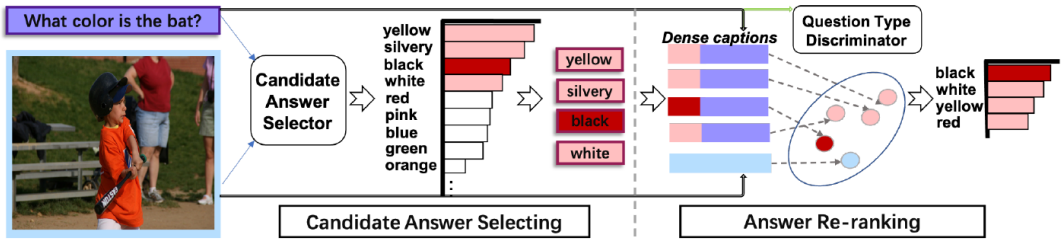


Figure 3.16: *Answer reranking showcasing the candidate selected answers and the visual entailment of image and question-answer captions to predict the correct answer.*

Essentially, both methods combine the Question and answers into captions and transform the VQA task in a 1-class alignment task as shown in Figure 3.16. Each data point is essentially transformed into K data points for each possible answer of candidate answers. In SAR specifically, they reduce the computational complexity by using a pre-trained VQA model to choose only the most probable answers. However, their complexity is still proportional to K times M ($\Theta(K * M)$), where M represents the complexity of the baseline model and K is the chosen “best" answer.

Chapter 4

Literature Reimplementations and Visual Question Generation.

This chapter includes our first experiments regarding reimplementing various research papers to improve generalization in the VQACPv2 dataset [25]. Since intensive criticism has been towards exploiting biases in the CPv2 dataset, these methods will be reimplemented to the GQA-OOD dataset [2] to evaluate their performance in a different out-of-distribution setting and will be discussed thoroughly.

Additionally, we present our initial experiments using visual question generation as an augmentation method for VQA tasks and discuss our findings.

4.1 Paper Reproductions for the GQA-OOD and VQA-CPv2 datasets

4.1.1 Reimplementation Choices

This section will enumerate the research papers selected for reimplementation and explain these choices. We will also highlight those studies that were not chosen for reimplementation, along with the reasoning behind their exclusion based on the analysis above:

- MANGO [5] , X-GGM [31], RankVQA [11] did not provide source code, and the results of their baselines are inconsistent with the rest of the literature.
- The SAR model [12], initially deployed on the VQA-CPv2 dataset, proved to be excessively computationally demanding to complete its training. Consequently, we determined that the marginal performance improvement did not justify further exploration or investment into this method.
- While SwapMix[7] did not enhance the standard accuracy in the VQA dataset on which it was trained, it showed improvement only in context reliance and attribute metrics. Therefore, we decided to omit it from our study.
- The SCR paper [49] requires human textual or visual explanations that were not available in the GQA-OOD dataset.
- The MUTANT [10] augmentation technique is noteworthy; however, its applicability is limited due to its focus on the unique question types found in VQA-CPv2, such as color-based questions, which are infrequent in the GQA-OOD dataset.

- Visual Question generation methods were utilized in our own visual question generation implementation.
- The ensemble and various unique sample generation methods were not specific to any dataset and either had accessible source code or were simpler to implement. Therefore, we re-implemented some of these methods.

We reimplemented the LMH and Rubi methods for ensemble-based methods in our experiments. We also tested the SSL and CSS as augmentation-based methods in both datasets. Finally, in terms of visual enhancement, we created our visual question-generation methods in the VQA-CPv2 dataset.

4.1.2 Results

Model	Baseline	Reported VQA-CPv2 results				Reimplemented VQA-CPv2 results			
		Overall	Yes/No	Num	Other	Overall	Yes/No	Num	Other
Baselines									
UpDn[23]	UpDn	40.14	42.27	11.93	46.05	39.50	43.46	12.17	44.92
Ensemble based methods									
Rubi[13]	UpDn	44.23	67.05	17.48	39.61	44.11	64.85	11.83	42.11
LMH[14]	UpDn	52.01	72.58	31.12	46.97	51.93	72.58	31.12	45.03
Data augmentation methods									
CSS[9]	UpDn	41.16	43.96	12.78	47.48	39.36	42.12	12.32	45.34
LMH+CSS[9]	UpDn	58.95	84.37	49.42	48.21	57.92	86.12	51.09	45.01
SSL*[8]	UpDn	58.11	86.53	29.87	50.03	57.16	85.67	30.12	49.31

Model	Baseline	GQA-OOD test results			
		Acc tail	Acc head	Acc all	Delta
Baselines					
UpDn [23]	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 4.1}$
Ensemble based methods					
Rubi [13]	UpDn	$30.78_{\pm 2.3}$	$39.52_{\pm 2.4}$	$36.2_{\pm 3.1}$	$22.10_{\pm 4.7}$
LMH [14]	UpDn	$27.621_{\pm 2.5}$	$38.41_{\pm 1.9}$	$34.31_{\pm 2.7}$	$28.09_{\pm 3.9}$
Data augmentation methods					
CSS [9]	UpDn	$41.75_{\pm 1.6}$	$49.11_{\pm 1.9}$	$46.31_{\pm 1.3}$	$14.95_{\pm 2.8}$
LMH+CSS [9]	UpDn	$29.19_{\pm 3.1}$	$37.67_{\pm 2.2}$	$34.45_{\pm 2.3}$	$22.49_{\pm 3.7}$
SSL*[8]	UpDn	↓	↓	↓	↓

We reimplemented the papers in fixed seeds accodring to their repositories for a fair comparison, and in GQA-OOD, they were tested on multiple seeds.

4.1.3 Notable Differences in VQACPv2

Most of the results were similar to those reported in the VQACPv2 dataset.

In SSL, the regularization loss becomes active after pretraining the model only with the original VQA Loss. However, the training was highly unstable, significantly deteriorating the results after the first epoch. The chosen weight parameter a = 3 determined for optimal model performance in the paper resulted in deteriorating results. Consequently, to obtain the desired result, we had to use a=2.

The CSS framework did not perform up to par with the reported results and showed no improvement compared to the baseline.

4.1.4 Results in GQA OOD

The ensemble-based models significantly underperformed in the GQA-OOD datasets, showcasing that language-biased ensemble models primarily took advantage of the inverse answer distribution in the test set to obtain their results. The CSS framework for counterfactual samples did not improve the results in either of the two datasets. The SSL framework is highly unstable for high values of a (the weight of our loss function). Testing for different values of a, we concluded that only for small values (lower than 1) the model does not severely underperform compared to the baseline.

4.1.5 Discussion

After carefully reading the literature reimplementing certain methods, we concluded the following:

- Addressing the language biases needs careful consideration because they could highly deteriorate results if based on knowing the answer distribution shift in the test set, and hence show superficial improvements in results instead of enhancing results in out-of-distribution conditions [2, 4]. Ensemble-based models based on a biased language model, in particular, consistently deteriorate results in the GQA-OOD condition and are only fit for situations where language biases are too prevalent in the data.

- Models trained with adversarial or object-swapping perturbations primarily enhance their generalization abilities under conditions similar to their training.

- Augmentation-based methods could potentially improve results. So far in the literature, only [10].has shown significant performance increases. However, its augmentation strategies are question/answer type-specific and unsuitable for general VQA tasks.

- Methods for answer reranking/visual entailment can enhance the accuracy of answer classification, especially when dealing with a large pool of possible answers (denoted as K). However, the complexity of these methods scales with the number of answer candidates. Consequently, while they are potentially more effective, they suffer from significant computational inefficiency due to their dependency on the number of answer candidates.

Our results aligned with the criticism towards the VQA-CPv2 dataset in [1, 4, 2]; hence, in the following Chapter, we will focus strictly on the GQA-OOD dataset for our proposed method.

4.2 Paraphrasing using Visual Question Generation

4.2.1 Brief Overview

The critical aspect of Visual Question Answering (VQA) is developing robust models that accurately interpret and respond to various questions about different images. Augmentation methods, such as Visual Question Generation (VQG), play a vital role in this endeavor.

VQG involves automatically generating relevant and diverse questions about a given image. This process focuses on creating contextually appropriate and meaningful questions, enhancing the capabilities of VQA models. VQG combines methods to improve visual information and data augmentation. It generates questions from images and additional metadata, serving as a dual task and a data augmentation method in VQA. VQG aims to create a more diverse and comprehensive dataset by generating a broad spectrum of questions for each image.

This section will discuss our experimental process, results, and conclusions on VQG's effectiveness in generalizing better in out-of-distribution settings, especially when dealing with limited training data and fixed model architecture.

4.2.2 Related Work

In [51], a purely supervised method is employed using CNN/RNN encoders, decoders, and attention mechanisms, trained with crossEntropyLoss using ground truth questions.

[52] utilizes a Variational Autoencoder (VAE) to generate questions from images, answers, and answer categories, trained with supervised methods (crossEntropyLoss and Scheduled Sampling [53]) and unsupervised loss functions (Reconstruction loss and Information Maximizing loss).

In contrast, some models use VQG as a supplementary method to VQA:

[54] introduces a dual architecture using MUTAN for simultaneous task-solving. Both streams share the same encoder/decoder pairs, using VQG as a regularization task for robust representation creation.

[55] employs a cyclic consistency framework, where the VQG model generates questions Q' based on the VQA model's predictions A', and vice versa. Models are jointly trained through various loss functions, including cyclic loss.

VQG models have been tested on the original VQA dataset but not extensively in OOD conditions [25, 2].

4.2.3 Baseline and Dataset.

We used the VQA-CPv2 [25] dataset for our experiments. This dataset is focused on having specific question type categories as shown in Figure 3.6. Q

We used the Bottom-Up-Top-Down attention model as our baseline for VQA [23] and a Visual Question Generator model with the following architecture for the augmented sample extraction. The UpDn model used had slightly different and more optimized hyperparameters than the one we reimplemented in the previous section and hence produced overall slightly better results than the ones reported in [14, 9].

Our architecture is shown in Figure 4.1 is based on the cycle consistency paper architecture [55] and includes the following components:

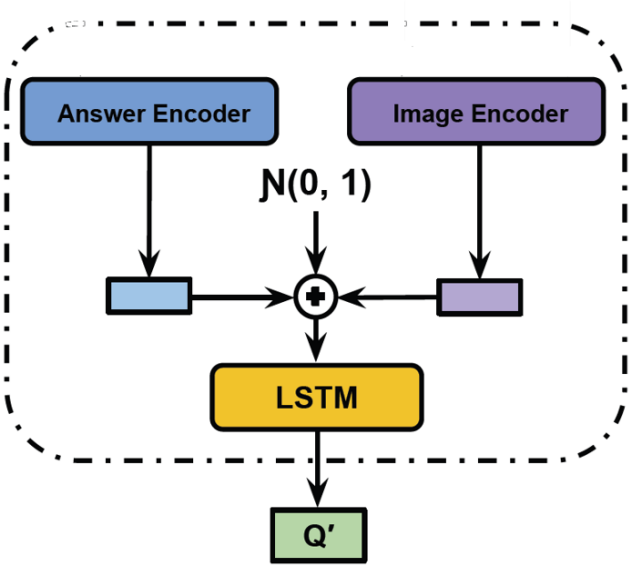


Figure 4.1: *Our baseline generator architecture.*

- **Language Encoder:** Converts the answers into series Glove embeddings and passes them through an LSTM to create an answer embedding in the joint embedding space.
- **Image Encoder:** Extract Faster-RCNN features from the image pass through a non-linear layer and take the average feature as the image embedding in the joint embedding space.
- **Fusion Layer:** A non-linear fusion layer combines the information of the image and the answer to a joint embedding.
- **Generator LSTM Network:** A generator LSTM that uses the joint embedding to predict the augmented question.

4.2.4 Experimental Design

During our experimental process, we implemented the following steps:

- **Teacher Forcing vs No-Teacher Forcing vs and Curriculum Learning:** We experimented with all three methods to see which fit our data best.
- **Choosing specific question types for augmentation:** An ablation of the difference performance results per question type in the dataset.
- **Beam Search:** We used beam search for a more diverse sample generation.

Training method

In the initial experiments, we used teacher forcing, no teacher forcing, and curriculum learning and compared their results. Without teacher forcing, the model always fails to generate syntactically, logically, and semantically correct questions the majority of the time. With the curriculum learning strategy, a lot of the results are syntactically correct, but we still observed low-quality results.

Some **grammatically incorrect generated** examples with curriculum learning and no teacher forcing can be examined below:

No Teacher Forcing:

1. Does the man wear a polo shirt? -> Is there a polo?
2. Is the man wearing sandals? -> Is there a wearing sandal?
3. Who is holding an umbrella -> What is the umbrella holding an umbrella?

Curriculum Learning:

1. What animals are they riding in the picture -> what animals are riding the picture?
2. Is the train moving -> Are there trains moving?

Consequently, the rest of the experiments were implemented only with teacher forcing, as did all the three referenced papers below. H We generated 1 example per data point and trained our VQA model with an augmented dataset.

The results for VQA training using teacher forcing are shown in Table 1. The question augmentation seems to degrade the performance of the baseline highly, especially in the "Other" question type.

Some teacher forcing results can be examined in Figure 4.2. The results vary and primarily belong to 3 different categories:

1. Paraphrased questions.
2. Questions that provide new information.
3. Syntactically and grammatically correct but semantically incorrect questions,

Figure 4.2: *In this figure, we can see that the first example is a paraphrased version of the original. In the second example for the answer 'ball,' the generated question refers to the boy instead of the man, and in the last question, the model falsely refers to the dog's necklace instead of the drawing in the sand.*

Analysis on the question types

After the above experiments, I examined what degraded the performance by comparing baseline and augmented model results in every question type. There was a diversity in 'Yes/No' questions since it is hard for the model to understand negation and generate counterfactual images(in case of a No answer).

Also, there was a consistent 15% average drop in color-type questions. Each answer has multiple colors; consequently, it is unlikely that the same answer distribution refers to more than one object in the image. As a result, at best, our model generates the question 'What color is the correct obj?', which is our ground truth, and at worst, it refers to a wrong object.

We generated augmented samples for only 'Other' type questions without augmenting color-type questions. Those questions contain approximately 37% of the total dataset. The results are shown in Figure 4.3. The performance seems to improve compared to the previous approach and is comparable to the baseline. However, we are generating only 37% extra samples.

Beam search generation

Our initial question generation model produced semantically correct results but it does not manage to produce diverse samples. We used beam search to extract the top k questions for each data point to create more diverse samples. Beam search improves the sample generation using a less greedy approach and forces the VQG

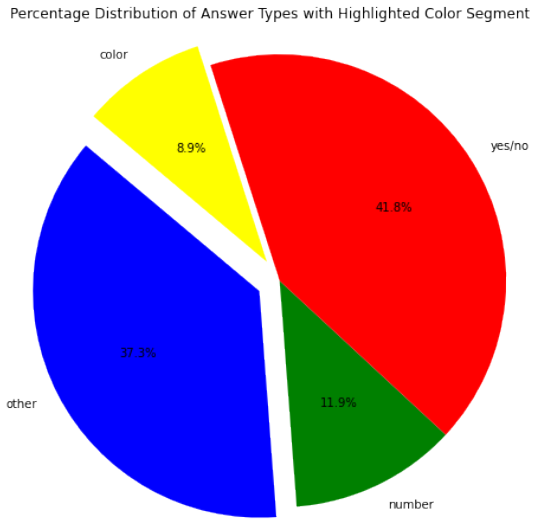


Figure 4.3: *Percentages of various answer types.*

model to create multiple discrete questions. Additionally, as shown in Figure 4.1, we tried adding random noise to the initial fusion embeddings. However, it did perform as expected and did not result in a diverse sample generation.

Using a 10-stream beam search, we chose the top 3 questions for each data point and trained the model with the new augmented data. Even though Beam-search generated more diverse samples, as showcased in Figure 4.4, it produced lower results than the 1-sample generation, indicating that even though we managed to create more varied samples, they either resulted in overfitting the dataset or created semantically incorrect questions.

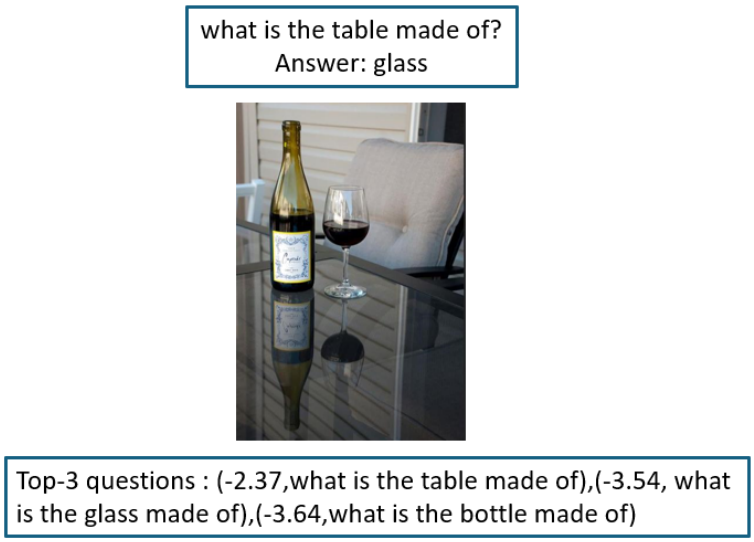


Figure 4.4: *Top 3 results for the image.*

4.2.5 Results

Here we present the results based on our augmented samples compared to our baseline model.

Model	All	Yes/No	Num	Other
UpDn(Baseline)	41.53	43.45	13.64	48.18
Updn + base_VQG (all questions)	38.7	42.34	9.4	43.21
UpDn + base_VQG	41.21	42.73	14.08	48.01
UpDn + 3-beam	40.64	42.60	13.32	47.12

4.2.6 Discussion

Question paraphrasing using visual question generation for specific answer-question pairs does not improve or even severely deteriorate results in OOD conditions, provided that we are constrained by our trainset for diverse sample generation. To achieve better OOD results, a strategy that creates new question-answer pairs like in [10] should be implemented since creating questions with the same expected answer is severely constraining the generation process.

Chapter 5

Regularization with visual object masking

5.1 Proposal

Building on the insights from the previous chapter, it becomes evident that a deeper understanding of visual context is crucial for enhancing the performance of models under Out-of-Distribution (OOD) conditions. This aligns well with findings from the GQA-OOD study, where the authors developed a VIS-Oracle model using a compact LXMERT. This model, leveraging the complete human-annotated scene graph information (encompassing ground truth objects, attributes, and relations), attained a remarkable 90% accuracy in OOD scenarios [2]. A model endowed with perfect visual perception can effectively employ semantic reasoning without depending on spurious correlations. With this motivation, we propose a new method that focuses on teaching the model to attend to the most important objects and ignore irrelevant ones while simultaneously filtering the noisy information of the objects with extracted textual information from the image.

Hence, to achieve better performance in both ID and OOD settings we should mostly focus on improving the image processing instead focus on the language part of the model. To generalize perfectly, a model should ideally capture the real-world causal mechanisms behind the data. It is important for the VQA model to learn to attend to the proper image regions related to the question to reason properly, instead of relying on spurious correlations in the training data.

With this motivation, we propose a new method that teaches the model to attend to the most image important objects and ignore irrelevant ones.

5.2 Methodology

5.2.1 Dataset

For our dataset, we use GQA-OOD, which is formulated as a VQA single-label classification task. Additional annotations are provided that include decomposed questions as semantic programs that refer to objects included in the scene graph annotations.

5.2.2 Baseline

Our baseline model is the Bottom-Up Top-Down (UpDn) Attention Model. As explained in the previous section, the UpDn model leverages bottom-up and top-down mechanisms to effectively integrate visual and textual features. The bottom-up mechanism focuses on extracting salient image features, while the top-down attention guides this process based on the textual query.

The network's final output shown in Figure 3.5, which represents the probability distribution over potential answers, is assigned to a binary cross-entropy loss function. The loss function is formulated as follows:

$$L_{\text{BCE}} = -\sum_{i=1}^{C} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{5.1}$$

where C is the number of answer classes, y_i is the ground truth label for class i, and p_i is the predicted probability for class i by the model. This loss function calculates the cross-entropy between the predicted probabilities and the ground truth labels, effectively penalizing the predictions that diverge from the actual labels.

5.2.3 Positive and Negative sample construction.

Inspired by the work in [9, 15], we construct similar and counterfactual samples for each image based on important image regions.

If we have stronger annotations in some form of visual explanation, we can directly locate the important regions. For example, in [15], they utilize human attention maps relative to the question. The GQA dataset provides the ground-truth reasoning steps(programs) for each question and the selected objects after each step. We use those reasoning steps to filter out all the relevant and irrelevant objects for the question. Then, we can use a bounding box overlap mechanism to match the Faster-RCNN objects with the ground-truth ones.

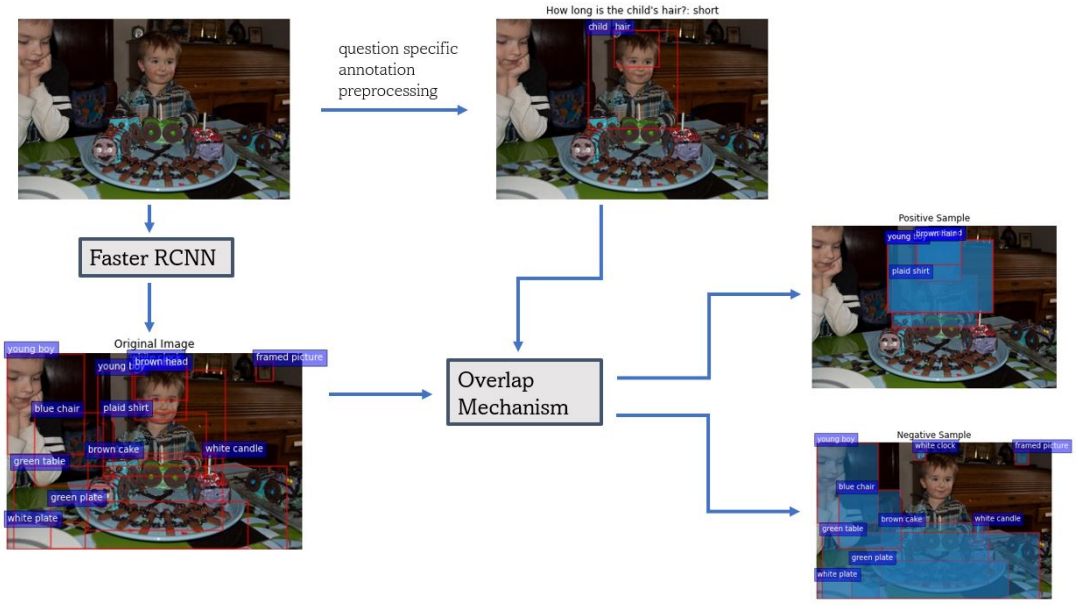
After selecting the most important regions, we want to find all relevant objects. Several methods in the GQA dataset use the IoU(Intersection over Union) overlap similarly as additional information for their respective methodologies[16, 7]. However, the IoU ratio might not necessarily be the proper overlap mechanism for our method. In contrast to [16], we do not want to perform a 1-1 matching between ground truth and extracted objects. Instead, we mainly wish to filter out the most relevant information from our negative samples. Faster-RCNN bounding box regression could be better, and there is

severe overlap between different object regions. Moreover, more oversized objects contain smaller important objects, and negative samples will retain the smaller objects' information even if we mask them. We can use the overlap ratio as an additional metric to IoU since it can better filter out nearby objects:

$$Overlap = \frac{ObjectArea \cap ExtractedObjectArea}{ObjectArea}$$

After extracting the "relevant objects," we can mask the non-important ones to create our positive sample and, conversely, mask the important ones to create our negative sample. An example of the above process with $IoU_{thresh} = 0.1, Overl_{thresh} = 0.2$ can be seen in Figure 5.1.

Figure 5.1: *Is the hair brown and thin?: yes*



5.2.4 Regularization Tasks and Loss Functions

Using various loss functions, we explore four critical aspects of visual question answering (VQA). These aspects include evaluating the counterfactuality of negative examples, the potential of using positive samples for augmentation, the validity of the assumptions underlying the triplet loss, and the regularization of the potential of random masking mechanisms. Let's delve into each aspect:

Counterfactual Losses

We use our extracted negative samples as the counterfactual samples for the counterfactual losses.

The first counterfactual loss we experimented with is designed to match negative images with questions [8], focusing on counterfactual examples where the question-image

pairs are irrelevant. The **Self-Supervised Learning loss (SSL)** is formulated as:

$$L_{\text{SSL}} = -\frac{1}{N}\sum_{i=1}^{N}\log(1 - P(A_i|Q_i, I_i'))$$

where P represents the prediction function, A_i is the answer, Q_i is the question, and I_i' is the counterfactual image.

In the context of L_{qd}, minimizing $-\log(1 - P(A|Q, I_0))$ is mathematically equivalent to minimizing $P(A|Q, I_0)$ which is claimed to be more stable during training in [8]. This formulation inherently implies that, for counterfactual samples, the model should assign a probability of zero to all potential answers:

$$L_{qd} = \frac{1}{N}\sum_{i=1}^{N}P(A_i|Q_i, I_{i0})$$

The **Counterfactual Gradients Supervision method** [15] involves using counterfactual examples as negative samples. The loss function for CF-GS, which aims to align network gradients with a ground truth gradient vector, is defined as:

$$L_{\text{GS}}(g_i, \hat{g}_i) = 1 - \frac{g_i \cdot \hat{g}_i}{||g_i||||\hat{g}_i||} \tag{5.2}$$

where $g_i = \nabla_x f(x_i)$ is the gradient of the network concerning its input at a point x_i, and $\hat{g}_i = x_j - x_i$ is the "ground truth" gradient vector.

Lastly, in [9], they use a **supervised loss for the counterfactual samples**. The counterfactual labels are the inverse of the topK predictions of the model by passing the positive sample. In the context of GQA, where multi-label classification is not applicable, the loss of the dynamic answer-assigning mechanism mentioned in the previous sections simplifies. For a VQA pair, we assign $a = 1$ if the correct answer is not predicted correctly. Conversely, if the answer is predicted correctly, we assign $a = 0$. This represents an **inverted labeling** approach, where $a = 0$ indicates correct predictions and $a = 1$ indicates incorrect predictions, referring to the dynamic answer assigning method mentioned above.

Regular Supervised Loss for Positive Samples

This loss is used for positive samples, turning the supervised method into an augmentation technique. The overall loss is a weighted sum of the VQA classification losses L_{vqa} and the BCE loss for positive samples L_{pos}.

Triplet Loss

Theoretically, positive samples are more informative for correctly answering questions and highlighting relevant image regions, unlike negative samples where crucial regions are obscured. In the embedding space, the goal is to position positive samples closer to the original samples while distancing them from negative samples, as seen in Figure 5.2. This concept is mathematically captured using a self-supervised triplet loss. This

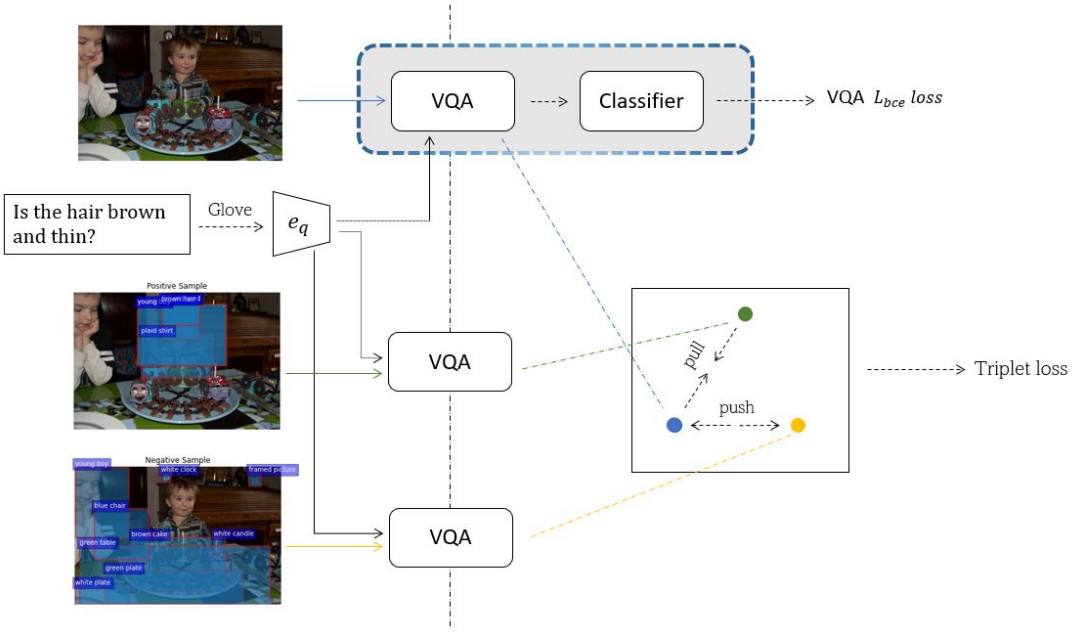
approach enhances the final visuolinguistic representation by pulling the original samples closer to positive ones and pushing them away from negative ones.

The triplet loss, which leverages cosine similarity in the multimodal joint embedding space, is formulated as follows:

$$L_c = \mathbb{E}_{p,n,a}\left[-\log\left(\frac{e^{s(a,p)}}{e^{s(a,p)} + e^{s(a,n)}}\right)\right]$$

where p and n represent positive and negative samples respectively, a is the anchor, and s denotes the cosine similarity function. The triplet loss above is similar to the common triple loss used in [17] but is more robust since it normalizes the embedding space [56]. Figure 5.2 shows the triplet loss application.

Figure 5.2: *Is the hair brown and thin?: yes*



Training objective and Inference

For all our methods, the loss function is formulated as the sum of the BCE loss with our weighted regularization loss by a parameter λreg:

$$L = L_{\text{vqa}} + \lambda_{\text{reg}} \times L_{\text{reg}}$$

The test set has no programs, ground truth objects, or annotations. On inference, we can directly feed our original image sample (without any masking) to our VQA model and get the expected answer.

5.3 Experiments

For our baseline, we used the default hyperparameters found in Bottom-Up-Top Down Attention in the literature [14].

To construct our positive and negative samples, we filtered the important image objects relative to the question as shown in Figure 5.1. We kept the objects with $IoU_{overl} \geq 0.1$ and $Overlap \geq 0.2$ with the extracted annotated ground truth relevant objects.

5.3.1 Counterfactual Learning

In this section, we used the counterfactual samples constructed from the negative image with the important objects masked and the question.

Gradient Supervision We used the gradient supervision method proposed in [15]. We experimented with the weight hyperparameter for values $\lambda = 1, 0.5, 0.1$. However, we witnessed lower performance for higher values of λ and similar performance to the baseline for lower values.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Gradient Sup($\lambda = 1$)	UpDn	41.76	50.60	47.24	21.15
BCE+Gradient Sup($\lambda = 0.5$)	UpDn	40.84	49.5	46.21	17.49
BCE+Gradient Sup($\lambda = 0.1$)	UpDn	42.03	49.41	46.59	18.28

Self-Supervised Loss The self-supervised loss is used after pre-training the UpDn model and was shown to deteriorate the results after the first epoch in the validation set for multiple values of a.

Supervised Counterfactual Loss For our supervised counterfactual loss we followed the methodology in [9]. We passed our positive samples through the model and obtained the positive answers $a^+ \leftarrow$ top-N(argsort$_{a_i \in A}(P_{vqa}(a_i))$). Then we selected the topK answers for K = 1, and $a^- := \{a_i | a_i \in a, a_i \neq a^+\}$ a^+ is gt answer set. For 1-label classification, that is essentially 0 if the positive pas correctly answers the question and 1 if it fails. The supervised loss for counterfactual samples did not improve the overall results of the baseline model.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Counterfactual Loss($\lambda = 1$)	UpDn	41.70	49.52	46.55	15.78

Based on the assumption of incompatibility between the image and question, neither of the above methods improved the results. The above potentially indicates that our negative samples **should not be counted as counterfactual**. A possible reason for that is that there is some over-leak of information from the masked objects in the image because of the CNN feature extraction.

5.3.2 Triplet Loss

Lambda value.

The table shows a clear trend related to tuning the hyperparameter λ_c, which controls the contribution of the triplet loss L_c in the overall loss function. The baseline model, UpDn, without any triplet loss, shows certain performance levels across different metrics. Even though for higher values of λ we witness a performance drop, there is a noticeable performance improvement as for smaller values of λ_c.

Starting with $\lambda_c = 10$, we observe a major decline in performance compared to the baseline. This suggests that a high value of λ_c may overly penalize the model with contrastive loss, leading to poorer results. As λ_c is decreased to 0.5 and further to 0.2, there is a gradual improvement in all metrics (Acc tail, Acc head, Acc all, Delta). This indicates that reducing the influence of the triplet loss relative to the base VQA classification loss (L_{vqa}) helps in achieving better performance.

The most notable improvement is observed at $\lambda_c = 0.1$, where the performance across all metrics is the best. This value of λ_c seems to strike an optimal balance, effectively incorporating the triplet loss's regularization benefits without overwhelming the VQA model's primary task.

Furthermore, the reduced variance in results across different seeds with the introduction of the triplet loss, especially at lower values of λ_c, implies enhanced stability and reliability of the model. It suggests that the model performs better on average and is more consistent across different runs, a desirable trait in machine learning models.

We experimented with different λ values and reported the following results:

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
Triplet($\lambda = 0.1$)	UpDn	$\mathbf{44.097}_{\pm 0.7}$	$\mathbf{50.851}_{\pm 0.7}$	$\mathbf{48.283}_{\pm 0.6}$	$\mathbf{15.327}_{\pm 1.1}$

5.3.3 Augmentation method

Weeusede only our positive samples for training, reducing the "supervised" method to an **augmentation** technique.

We used $\lambda_{pos} = 1$ for our experiment as in [9].

We observe improvement across all metrics and significantly less variance across seeds. In contrast to the triple method, our secondary augmentation loss decreases relatively linearly compared to our VQA loss. Moreover, it is clear that strictly using only augmented samples is not sufficient for training our VQA model.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
Augmented Only	UpDn	36.50	48.346	43.85	15.944
BCE + Augmented	UpDn	$\mathbf{44.003}_{\pm 0.6}$	$\mathbf{51.01}_{\pm 0.7}$	$\mathbf{48.346}_{\pm 0.6}$	$\mathbf{15.934}_{\pm 1.5}$

5.3.4 Random Masking

Contrastive Learning using Random Masking.

To test the effectiveness of our framework, instead of using Customally selected masks for important objects, we will use random masks for comparison.

Important objects consist of 18% of the total objects. Hence, our random masks are derived from a **Bernoulli** distribution with $P(F) = 0.82$, where F is the masking of an object.

A modest performance improvement was observed with the addition of the triplet loss, which functions as a regularization term to our primary supervised loss; given the nature of our experiment, where randomly masked images retain a substantial 82% of the objects, it's reasonable to assume that these images (our negative samples) hold considerably more informational content than the positive samples, which only include 18% of the objects. In a typical setting, this imbalance might lead to a degradation in performance, as the loss function could struggle to effectively discern between positive and negative samples due to the overwhelming presence of information in the negative samples.

However, our results do not align with this theoretical expectation, as they indicate a slight enhancement in model performance. This suggests that the interaction between our triplet loss (as a regularization term) and the primary supervised loss might be more complex than initially assumed. It appears that the triplet loss is contributing positively, yet there's an indication that its full potential is not being harnessed. Introducing a level of randomness in the masking process could potentially optimize the effectiveness of this regularization approach. A randomized approach could promote a more balanced learning process by preventing the model from becoming too reliant on specific features of the unmasked image portions. Consequently, further exploration of various degrees and methodologies of randomization in masking could be pivotal in fine-tuning the triplet loss's role in enhancing the overall learning strategy.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Triplet with Heur Masking	UpDn	$\mathbf{44.097}_{\pm 0.7}$	$50.851_{\pm 0.7}$	$48.283_{\pm 0.6}$	$\mathbf{15.327}_{\pm 1.1}$
BCE+Triplet with Random Masks	UpDn	$43.52_{\pm 0.8}$	$\mathbf{51.029}_{\pm 0.5}$	$\mathbf{48.402}_{\pm 0.6}$	$15.683_{\pm 1.6}$

Augmentation with random maskings

To test the effectiveness of our masking method and augmentation loss, we will use the same random masking mechanism with $P(F) = 0.82$. Interestingly, we witnessed a significant performance improvement. The results are not discouraging since the overall bce loss of our Customally selected augmented samples during training is relatively close to the original bce loss. In contrast, the bce loss of the random samples is significantly higher. Therefore, we can assume that the Custom masking mechanism is qualitatively correct, but we should add additional randomness to the masking method.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+ Custom Masking	UpDn	$44.003_{\pm 0.6}$	$51.01_{\pm 0.7}$	$48.346_{\pm 0.6}$	$15.934_{\pm 1.5}$
BCE+ Random Masking	UpDn	$\mathbf{44.803}_{\pm 0.7}$	$\mathbf{52.352}_{\pm 0.8}$	$\mathbf{49.482}_{\pm 0.4}$	$\mathbf{16.898}_{\pm 3.3}$

Experiments with different percentages.

The random masking method seems intuitively similar to the classic Dropout[57] technique. We observe improvements across different P(F) values, none of which surpasses performance for $P(F) = 0.82$.

GQA OOD						
Loss	Baseline	Rand.Mask%	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	-	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Custom	UpDn	-	$44.003_{\pm 0.6}$	$51.01_{\pm 0.7}$	$48.346_{\pm 0.6}$	$15.934_{\pm 1.5}$
BCE+ Random	UpDn	0.18	$43.038_{\pm 0.9}$	$51.615_{\pm 1.2}$	$48.353_{\pm 1.0}$	$19.946_{\pm 2.4}$
BCE+ Random	UpDn	0.5	$43.932_{\pm 1.0}$	$51.529_{\pm 0.9}$	$48.641_{\pm 0.9}$	$17.31_{\pm 1.3}$
BCE+ Random	UpDn	0.82	$\mathbf{44.803}_{\pm 0.7}$	$\mathbf{52.352}_{\pm 0.8}$	$\mathbf{49.482}_{\pm 0.4}$	$\mathbf{16.898}_{\pm 3.3}$

Adding randomness to the masking method.

To include randomness in our Custom masking method, we randomly choose between a random mask with $P(F) = 0.82$ and our original mask for each augmented sample. The following results suggest that combining those two methods properly could be beneficial for better regularization.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Augmented+Random	UpDn	$\mathbf{44.803}_{\pm 0.7}$	$52.352_{\pm 0.8}$	$49.482_{\pm 0.4}$	$\mathbf{16.898}_{\pm 3.3}$
BCE+Augmented+Mixed	UpDn	$44.52_{\pm 0.7}$	$\mathbf{52.698}_{\pm 0.2}$	$\mathbf{49.589}_{\pm 0.2}$	$18.408_{\pm 2.4}$

Reducing augmented loss.

In our following experiment, we reduce the λ_{pos} hyperparameter to 0.5 to evaluate the augmented loss's importance in our framework.

GQA OOD					
Loss	Baseline	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Augmented+Random	UpDn	$\mathbf{44.803}_{\pm 0.7}$	$\mathbf{52.352}_{\pm 0.8}$	$\mathbf{49.482}_{\pm 0.4}$	$16.898_{\pm 3.3}$
BCE+Augmented+Random+0.5	UpDn	$44.496_{\pm 1.3}$	$51.284_{\pm 0.5}$	$48.704_{\pm 0.5}$	$\mathbf{15.369}_{\pm 4.0}$

5.3.5 Final results.

The last table includes our best-performing models relative to the baseline. The first model consists of an augmented bce loss for randomly masked images, essentially a more fine-grained dropout. The second model contains the augmented bce loss with random maskings and our triplet loss. The last model contains the augmented bce loss for our positive image objects and our triplet loss. Most of the models behave similarly, indicating that perhaps the masking process is the most important factor for improvement in initiating specific regularization. The proposed models improve attention in our test set, as shown in Figure 5.3.

GQA OOD					
Loss	Model	Acc tail	Acc head	Acc all	Delta
BCE	UpDn	$42.545_{\pm 1.6}$	$49.668_{\pm 1.2}$	$46.96_{\pm 1.5}$	$16.928_{\pm 3.5}$
BCE+Augm Rand	UpDn	$44.803_{\pm 0.7}$	$\mathbf{52.352}_{\pm 0.8}$	$49.482_{\pm 0.4}$	$16.898_{\pm 3.3}$
BCE+ Augm Rand +Triplet with Heur.	UpDn	$\mathbf{45.2}_{\pm 1.3}$	$52.284_{\pm 0.5}$	$\mathbf{49.704}_{\pm 0.5}$	$15.369_{\pm 4.0}$
BCE+ Augm + Triplet with Heur.	UpDn	$44.73_{\pm 0.47}$	$52.26_{\pm 0.65}$	$49.41_{\pm 0.42}$	$\mathbf{14.4}_{\pm 2.06}$

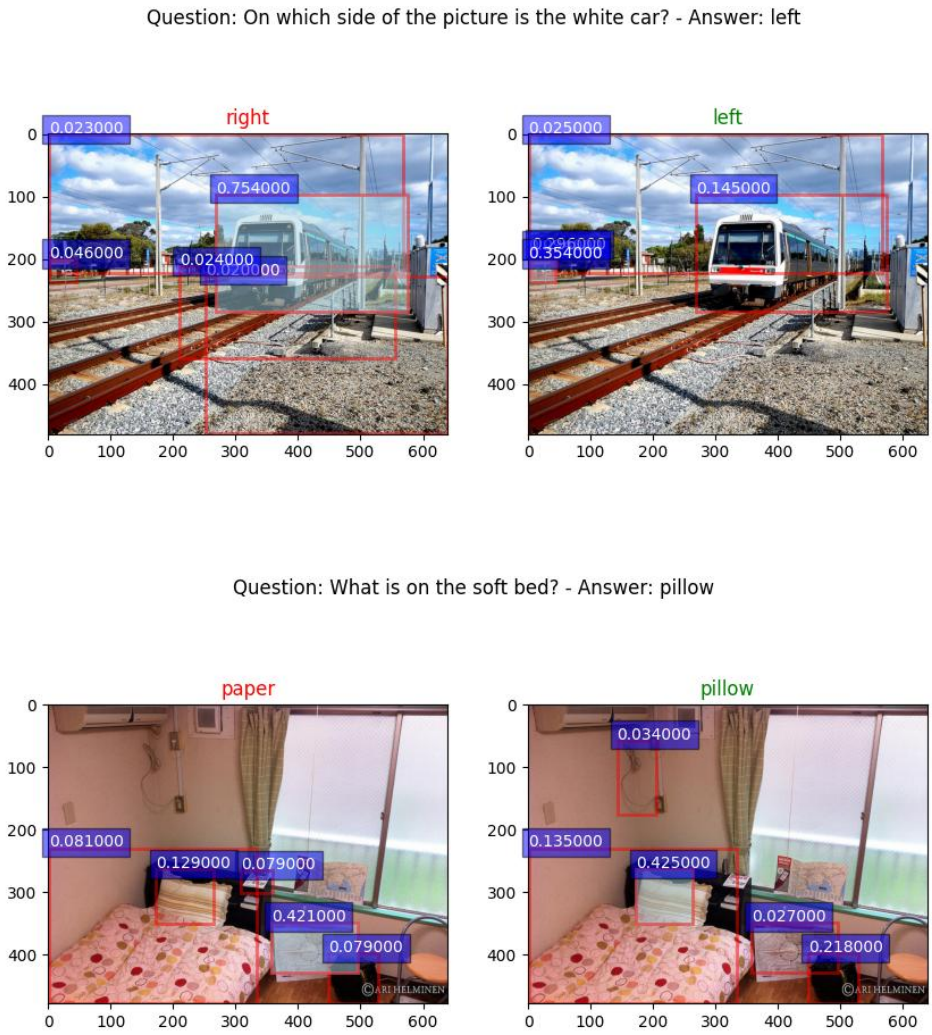


Figure 5.3: *Comparison of the last model's attention maps for the image objects.*

Chapter 6

Discussion and Future Work

6.1 Discussion

In our thesis, we delve into the field of visual question-answering (VQA) models, focusing on the impacts of various methodologies and their effectiveness in out-of-distribution scenarios. We reimplemented multiple methods in the literature, conducted initial experiments for question paraphrasing, and developed our own methodology for generalization methods in VQA based on visual object masking. Through a comprehensive literature review, multiple experiments, and critical analysis, we have drawn significant insights and findings that contribute to the field of generalization in Visual Question Answering.

Literature Review and Reimplementations

Our research highlights the importance of carefully addressing language biases in visual question-answering (VQA) models. This issue is especially pronounced in ensemble models based on biased language, which show reduced effectiveness in GQA-OOD scenarios and are limited in broader applications. We should note that constructing methods dependent on known answer distribution shifts in the test set tends to produce superficial improvements that could severely deteriorate results in different out-of-distribution (OOD) conditions.

Our review acknowledges the potential of augmentation-based methods in VQA. However, as currently documented in the literature, significant performance gains are restricted to specific augmentation strategies [10] and are not universally applicable to all VQA tasks.

Regarding answer reranking and visual entailment [50, 11], binding the information of question-answer pairs, enhances answer classification accuracy. However, those methods add significant computational costs that could potentially be avoided by a more efficient approach for combining the two linguistic representations.

Visual Question Generation

Our experiments in visual question generation showcase that trying to perturbate the question without changing the answers resulted in suboptimal performance. Hence, for methodologies that focus on question generation to be effective, methodologies should be implemented that construct semantically similar question-answer pairs with **different** answers in order to fully utilize the linguistic information in our data.

Masking methodology

Our extensive experimentation in visual question answering has led to several pivotal findings. The experiments indicate that the negative samples used in our study should

not be considered entirely counterfactual since they failed to improve results and even deteriorated them when utilizing them in counterfactual losses. A potential reason for that could be they potentially retain significant image information extracted through CNN models. Therefore, the model could theoretically infer the answer even when masking essential objects. This observation suggests a potential limitation in our approach to constructing negative samples, which might be causing an overleak of information.

The heuristic and random masking processes have proved effective in improveming performance in both ID and OOD settings. The augmented BCE loss, especially when combined with random masking, functions similarly to a refined dropout method and significantly boosts the model's generalization capabilities. This finding underscores the value of random masking in model training. Conversely, the heuristic masking approach appears to aid the model in focusing more on relevant image regions, thereby improving its attention capabilities.

Furthermore, our triplet loss methodology improves the overall results and can be combined effectively with the augmentation approach. Although as a sole regularization method, it does not perform as optimally as the augmentation-based methods, it showcases the big improvement in the Delta value, which is crucial for achieving similar performance in out-of-distribution (OOD) and in-distribution (ID) settings. However, our experiments with counterfactual losses suggest that we might not fully utilize the triplet loss because of potentially sub-optimal negative examples.

6.2 Future work

In the future, extensions to our work can include:

- **Negative Sample Construction Improvements:** Training with the counterfactual methods for our negative did not improve overall results, indicating that the triplet loss was not fully utilized. Our experiments could involve more fine-grained negative images or questions for our triplet pairs. We could also experiment with masking important question tokens besides masking the image content or/and masking objects that have the same class as our important objects. For example, regarding the VQA sample in Figure 5.2, in the first case both the boys object regions would be masked and in the second case masking the word "hair" would create a semantically different Q-I sample. Constructing "harder" negative samples could lead to more optimal results concering both the triplet loss methodology and the counterfactual learning based methods.

- **Replacing UpDn with other task-agnostic attention-based viso-linguistic models:** We could experiment with the bilinear attention models [18], which achieve higher perforce across most vqa tasks [18, 19, 2] or similar transformers-based models [20, 21] provided that we overcome the issue of test data leak in their pretraining. Bilinear attention models have showcased significant improvements compared to the original UpDn model as seen in [21, 18, 20] since their *NXM* attention mechanism allows for a more nuanced token-object interaction. Provided that we

have enough computational resources, testing our methodologies to other agnostic visio-linguistic models would be great form of validating our architecture agnostic proposal for generalization.

- **Creating new augmented question-answer pairs through semantic recomposition:** Inspired by the reranking methods in the literature review [11, 12] and our dissapointing results in the VQG method, a question-answer pair could contain additional semantic information if perturbated properly. More specifically, most question-answer pairs in VQA are included in a specific question type (f.e. what type of) and can be deconstructed using spacy position tagger [22] and reformulated in a way that we can create semantically similar but different Q-A pairs. Additionally, we could also replace certain words with similar words in the glove embedding space that are of the type, as seen in Figure 6.1, to further enhance the perturbation process. This methodology could be achieved both by using LLMs like GPT-3 and with a rule based system similar to the construction of several VQA datasets such as [19] as showcased in Figure 6.1. By augmenting our dataset with similar samples that showcase semantic differences and different answers, we could enhance the reasoning and semantic understanding of our VQA models and prevent them from relying on spurious language-biased correlations or skewed answer distributions.

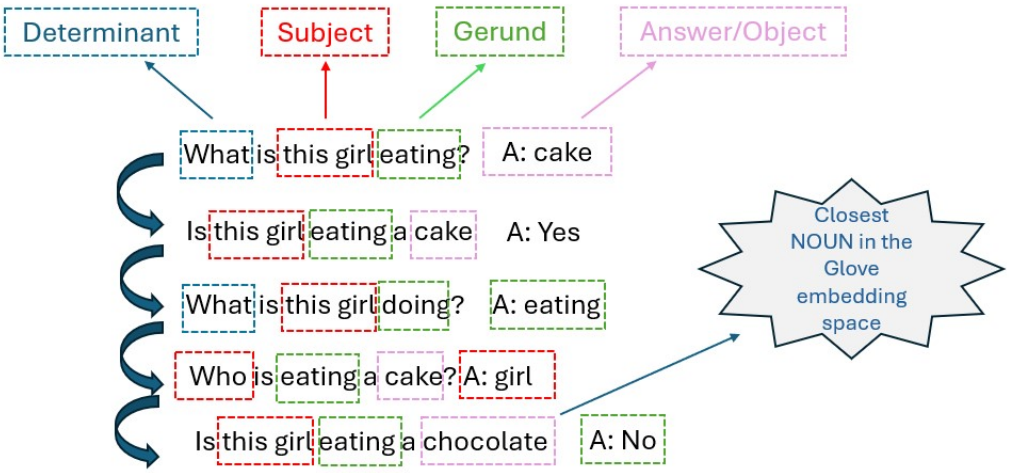


Figure 6.1: *Example for rule-based perturbations based on syntactical and grammatical recomposition.*

Bibliography

[1] Damien Teney, Ehsan Abbasnejad και Anton Hengel. *Unshuffling Data for Improved Generalization in Visual Question Answering*. *Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV)*, σελίδες 1397–1407, 2021.

[2] Corentin Kervadec, Grigory Antipov, Moez Baccouche και Christian Wolf. *Roses are Red, Violets are Blue... But Should VQA expect Them To?* *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 2775–2784, 2021.

[3] Han Xinzhe, Shuhui Wang, Chi Su και Qi Tian. *Greedy Gradient Ensemble for Robust Visual Question Answering*. *Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV)*, σελίδες 1564–1573, 2021.

[4] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan και Antonvan den Hengel. *On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law*. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).*, 2020.

[5] Linjie Li, Zhe Gan και Jingjing Liu. *A Closer Look at the Robustness of Vision-and-Language Pre-trained Models*. 2020.

[6] Zhe Gan, Yen Chun Chen, Linjie Li, Chen Zhu, Yu Cheng και Jingjing Liu. *Large-Scale Adversarial Training for Vision-and-Language Representation Learning*. 2020.

[7] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li και Alan Yuille. *SwapMix: Diagnosing and Regularizing the Over-Reliance on Visual Context in Visual Question Answering*. *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 5068–5078, 2022.

[8] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Wang Xian Bin και Yongdong Zhang. *Overcoming Language Priors with Self-supervised Learning for Visual Question Answering*. *Proceedings of Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*, σελίδες 1083–1089, 2020.

[9] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu και Yueting Zhuang. *Counterfactual Samples Synthesizing for Robust Visual Question Answering*. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 10797–10806, 2020.

[10] Tejas Gokhale, Pratyay Banerjee, Chitta Baral και Yezhou Yang. *MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering*. *Pro-*

ceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), σελίδες 878–892, 2020.

[11] Yanyuan Qiao, Zheng Yu και Jing Liu. *Rankvqa: Answer Re-Ranking For Visual Question Answering*. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) 2020*, 2020.

[12] Qingyi Si, Zheng Lin, Mingyu Zheng, Peng Fu και Weiping Wang. *Check It Again: Progressive Visual Question Answering via Visual Entailment*. *arXiv:2106.04605*, 2021.

[13] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord και Devi Parikh. *RUBi: Reducing Unimodal Biases in Visual Question Answering*. *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).*, 2019.

[14] Christopher Clark, Mark Yatskar και Luke Zettlemoyer. *Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases*. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, σελίδες 4060–4073, 2019.

[15] Damien Teney, Ehsan Abbasnejad και Anton Hengel. *Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision*, σελίδες 580–599. 2020.

[16] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot και Christian Wolf. *How Transferable are Reasoning Patterns in VQA?* *Proceedings of the 2021 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger και Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. *Proceedings of Machine Learning Research*, 2021.

[18] Jin Hwa Kim, Jaehyun Jun και Byoung Tak Zhang. *Bilinear Attention Networks*. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[19] Drew Hudson και Christopher Manning. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 6693–6702, 2019.

[20] Jiasen Lu, Dhruv Batra, Devi Parikh και Stefan Lee. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[21] Hao Tan και Mohit Bansal. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, σελίδες 5103–5114, 2019.

[22] Explosion AI. *Tagger - spaCy API Documentation*. https://spacy.io/api/tagger, 2023.

[23] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould και Lei Zhang. *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 6077–6086, 2018.

[24] Goyal, Yash, Khot, Tushar, Summers-Stay, Douglas, Batra, Dhruv, Parikh και Devi. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Aishwarya Agrawal, Dhruv Batra, Devi Parikh και Aniruddha Kembhavi. *Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering*. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[26] Stuart Russell και Peter Norvig. *Artificial intelligence: a modern approach*, 2002.

[27] Geoffrey Hinton και Terrence J Sejnowski. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[28] Robin Brochier, Adrien Guille και Julien Velcin. *Global Vectors for Node Representations*. *The World Wide Web Conference*, WWW '19. ACM, 2019.

[29] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative Adversarial Networks*. 2014.

[30] Sainandan Ramakrishnan, Aishwarya Agrawal και Stefan Lee. *Overcoming Language Priors in Visual Question Answering with Adversarial Regularization*. 2018.

[31] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan και Nanning Zheng. *X-GGM: Graph Generative Modeling for Out-of-distribution Generalization in Visual Question Answering*. *Proceedings of MM 2021: ACM Multimedia Conference*, σελίδες 199–208, 2021.

[32] Carl Doersch. *Tutorial on Variational Autoencoders*, 2016.

[33] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018.

[34] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao και Junwei Han. *Oriented R-CNN for Object Detection*. *Proceedings of International Conference on Computer Vision*, 2021.

[35] Sebastian Ruder. *An overview of gradient descent optimization algorithms*, 2017.

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever και Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[37] *Understanding LSTM Networks*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[38] Ross Girshick. *Fast R-CNN*. *Proceedings of International Conference on Computer Vision*, 2015.

[39] Shaoqing Ren, Kaiming He, Ross Girshick και Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[40] Tim Rocktäschel, Edward Grefenstette, Karl Hermann, Tomáš Kočiský και Phil Blunsom. *Reasoning about Entailment with Neural Attention*. 2015.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser και Illia Polosukhin. *Attention Is All You Need*. *Proceedings of th31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[42] Liunian Li, Mark Yatskar, Da Yin, Cho Jui Hsieh και Kai Wei Chang. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. *arXiv e-prints*, 2019.

[43] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár και C. Zitnick. *Microsoft COCO: Common Objects in Context*. *European Conference on Computer Vision*, τόμος 8693, 2014.

[44] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick και Devi Parikh. *VQA: Visual Question Answering*. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[45] Justin Johnson, Bharath Hariharan, Laurensvan der Maaten, Li Fei-Fei, C. Lawrence Zitnick και Ross Girshick. *CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning*. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[46] Ramprasaath Rs, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh και Dhruv Batra. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. *International Journal of Computer Vision*, 128, 2020.

[47] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein και Jingjing Liu. *FreeLB: Enhanced Adversarial Training for Natural Language Understanding*. *arXiv:1909.11764*, 2019.

[48] Evgenia Rusak, Lukas Schott, Roland Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge και Wieland Brendel. *A Simple Way to Make Neural Networks Robust Against Diverse Image Corruptions*, σελίδες 53–69. 2020.

[49] Jialin Wu και Raymond J. Mooney. *Self-Critical Reasoning for Robust Visual Question Answering*. *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[50] Ning Xie, Farley Lai, Derek Doran και Asim Kadav. *Visual Entailment: A Novel Task for Fine-Grained Image Understanding*. 2019.

[51] Feng Liu, Tao Xiang, Timothy Hospedales, Wankou Yang και Changyin Sun. *Inverse Visual Question Answering: A New Benchmark and VQA Diagnosis Tool*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 2018.

[52] Ranjay Krishna, Michael Bernstein και Li Fei-Fei. *Information Maximizing Visual Question Generation*. *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 2008–2018, 2019.

[53] Samy Bengio, Oriol Vinyals, Navdeep Jaitly και Noam Shazeer. *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks*. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

[54] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang και Xiaogang Wang. *Visual Question Generation as Dual Task of Visual Question Answering*. *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[55] Meet Shah, Xinlei Chen, Marcus Rohrbach και Devi Parikh. *Cycle-Consistency for Robust Visual Question Answering*. *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 6642–6651, 2019.

[56] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu και Yueting Zhuang. *Counterfactual Samples Synthesizing for Robust Visual Question Answering*. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever και Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.