Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Linguistic Counterfactuals for Visual Question Answering

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΟΔΟΤΗ ΣΤΟΪΚΟΥ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Linguistic Counterfactuals for Visual Question Answering

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΟΔΟΤΗ ΣΤΟΪΚΟΥ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Ιουλίου 2023.

Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.	Αθανάσιος Βουλόδημος Επ. Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

Θεοδότη Στόϊκου

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

Περίληψη

Η απάντηση οπτικών ερωτήσεων (Visual Question Answering - VQA) είναι μια δημοφιλής εργασία που συνδυάζει την όραση και τη γλώσσα, με πολλές σχετικές υλοποιήσεις στη βιβλιογραφία. Παρόλο που υπάρχουν κάποιες προσπάθειες που προσεγγίζουν ζητήματα εξηγησιμότητας και ευρωστίας σε μοντέλα VQA, πολύ λίγες από αυτές χρησιμοποιούν αντιπαραδείγματα ως μέσο διερεύνησης τέτοιων προκλήσεων με τρόπο γενικεύσιμο ως προς τα μοντέλα. Στην παρούσα διπλωματική εργασία, προτείνουμε μια συστηματική μέθοδο για την εξήγηση της συμπεριφοράς και τη διερεύνηση της ευρωστίας των μοντέλων VQA μέσω αντιπαραδειγματικών διαταραχών. Για το λόγο αυτό, αξιοποιούμε δομημένες βάσεις γνώσης για να εκτελέσουμε ντετερμινιστικές, βέλτιστες και ελεγχόμενες αντικαταστάσεις σε επίπεδο λέξεων που στοχεύουν στη γλωσσική μορφολογία εισόδου, και στη συνέχεια αξιολογούμε την απόκριση του μοντέλου έναντι τέτοιων αντιφατικών εισόδων. Τέλος, εξάγουμε ποιοτικές τοπικές και συνολικές εξηγήσεις με βάση τις αντιπαραδειγματικές αποκρίσεις, οι οποίες τελικά αποδεικνύονται κατατοπιστικές για την ερμηνεία της συμπεριφοράς του μοντέλου VQA. Πραγματοποιώντας μια ποικιλία τύπων διαταραχών, που στοχεύουν σε διαφορετικά μέρη του λόγου της ερώτησης εισόδου, αποκτούμε γνώσεις σχετικά με τη συλλογιστική του μοντέλου, μέσω της σύγκρισης των απαντήσεών του σε διαφορετικές αντιπαραθετικές συνθήκες. Συνολικά, αποκαλύπτουμε πιθανές προκαταλήψεις στη διαδικασία λήψης αποφάσεων του μοντέλου, καθώς και αναμενόμενα και απροσδόκητα μοτίβα, τα οποία επηρεάζουν ποσοτικά και ποιοτικά την απόδοσή του, όπως υποδεικνύεται από την ανάλυσή μας.

Λέξεις κλειδιά

Οπτική Απάντηση Ερωτήσεων, Γράφοι Γνώσης, Εξηγήσιμη Τεχνητή Νοημοσύνη, Αντιπαραδειγματικές Εξηγήσεις, Ευστάθεια.

Abstract

Visual Question Answering (VQA) has been a popular task that combines vision and language, with numerous relevant implementations in literature. Even though there are some attempts that approach explainability and robustness issues in VQA models, very few of them employ counterfactuals as a means of probing such challenges in a model-agnostic way. In this diploma thesis, we propose a systematic method for explaining the behavior and investigating the robustness of VQA models through counterfactual perturbations. For this reason, we exploit structured knowledge bases to perform deterministic, optimal and controllable word-level replacements targeting the linguistic modality, and we then evaluate the model's response against such counterfactual inputs. Finally, we qualitatively extract local and global explanations based on counterfactual responses, which are ultimately proven insightful in interpreting VQA model behaviors. By performing a variety of perturbation types, targeting different parts of speech of the input question, we gain insights into the reasoning of the model, through the comparison of its responses in different adversarial circumstances. Overall, we reveal possible biases in the decision-making process of the model, as well as expected and unexpected patterns, which impact its performance quantitatively and qualitatively, as indicated by our analysis.

Key words

Visual Question Answering, Knowledge Graphs, XAI, Counterfactual Explanations, Robustness.

Ευχαριστίες

Μέσα από την ολοκλήρωση αυτής της διπλωματικής εργασίας, ολοκληρώνεται και η φοίτησή μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Είμαι εξαιρετικά ευγνώμων για τα όσα μου έμαθε και μου πρόσφερε αυτό το ταξίδι. Και ακόμα περισσότερο, είμαι ευγνώμων που μου δόθηκε η δυνατότητα να υλοποιήσω την παρούσα δουλειά με τους ανθρώπους που συνεργάστηκα. Η εργασία αυτή μου ενέπνευσε αγάπη για την επιστήμη της τεχνητής νοημοσύνης και πίστη στην θετική έκβαση που μπορεί να έχει μια προσπάθεια. Πίστη ότι αληθινά είναι εφικτή η συμβολή στην ερευνητική κοινότητα, μέσα από την υλοποίηση μιας νέας ιδέας που καθοδηγείται από σημαντικά επιστημονικά κίνητρα.

Ευχαριστώ θερμά για όλα τον Καθηγητή μου, κύριο Γιώργο Στάμου, ο οποίος στάθηκε δίπλα μου από την αρχή αυτού του ταξιδιού. Τον ευχαριστώ για την πολύτιμη καθοδήγησή του, για τις σπουδαίες ακαδημαϊκές ευκαιρίες στις οποίες μου έδωσε πρόσβαση μέσω από το εργαστήριό του και για το ανθρώπινο ενδιαφέρον και την αστείρευτη υποστήριξή του για την εξέλιξή μου. Ευχαριστώ επίσης βαθιά τη Μαρία Λυμπεραίου, για την πολύωρη και πολυεπίπεδη υποστήριξή της στην διπλωματική μου εργασία, για τις πολλές φορές που με βοήθησε καθοριστικά τόσο στο πλαίσιο αυτής της δουλειάς, όσο και σε άλλες στιγμές της σταδιοδρομίας μου, και για την ζεστή και ειλικρινή σχέση φιλίας που αποκτήσαμε μέσα από αυτό το ταξίδι.

Κλείνοντας, θέλω να ευχαριστήσω τους προσωπικούς μου ανθρώπους για την αγάπη τους. Ευχαριστώ για πάντα την οικογένειά μου, τη μητέρα μου Ελένη, τον πατέρα μου Χάρη και την αδερφή μου Νίκη Αλεξάνδρα, που δε με άφησαν ποτέ μόνη και που στηρίζουν και αγαπούν φανατικά εμένα και όλα μου τα όνειρα. Ευχαριστώ βαθιά τον Θοδωρή μου, που κάνουμε όλα μας τα όνειρα μαζί από την πρώτη μέρα και που πιστεύει πάντα σε εμένα και σε εμάς. Ευχαριστώ τους φίλους μου για όλες τις στιγμές γέλιου και ξεγνοιασιάς που μοιραστήκαμε.

Θεοδότη Στόϊκου,

Αθήνα, 7η Ιουλίου 2023

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εκτεταμένη Ελληνική Περίληψη

1.1 Εισαγωγή

Η ραγδαία ανάπτυξη της τεχνολογίας τις τελευταίες δεκαετίες έχει διαδραματίσει σημαντικό ρόλο στη μετασχηματισμό της ανθρώπινης δραστηριότητας σε διάφορους τομείς, όπως η κοινωνία, η εργασία και η λήψη αποφάσεων γενικότερα. Στον τομέα της επιστήμης των υπολογιστών, μια βασική αντανάκλαση αυτής της σημαντικής εξέλιξης της γνώσης και της τεχνογνωσίας μπορεί να συνοψιστεί στην πρόοδο της τεχνητής νοημοσύνης. Πιθανώς το πιο αξιοσημείωτο όφελος που μας προσφέρει η τεχνητή νοημοσύνη και τα αποτελέσματα των μοντέλων μηχανικής μάθησης είναι η ποιότητα, η ευκολία και η ταχύτητα παραγωγής χρήσιμων αποτελεσμάτων που αναδιαμορφώνουν τους τρόπους λήψης αποφάσεων και καθορισμού της δράσης. Παρατηρούμε ότι αυτά τα αποτελέσματα και τα οφέλη είναι εμφανή σε πολλαπλούς τομείς της δράσης και της προόδου, όπως η υγεία, η δικαιοσύνη, η οικονομία, η πολιτική, η επιστήμη, η εκπαίδευση, ακόμη και η τέχνη και η δημιουργικότητα.

Η ευρεία χρήση των Συστημάτων Μηχανικής Μάθησης στη λήψη αποφάσεων καθώς και η αποφασιστική τους επιρροή στη λήψη απόφασης και στη συνέχεια στη δράση καθιστούν αναμφίβολα ύψιστη προτεραιότητα την διασφάλιση μεθόδων που προάγουν τη διαφάνεια, τη δικαιοσύνη και την αξιοπιστία. Στο πλαίσιο αυτό, ο ταχέως αναπτυσσόμενος τομέας της επεξηγήσιμης τεχνητής νοημοσύνης προσφέρει τα μέσα για την αύξηση της εμπιστοσύνης στα συστήματα μηχανικής μάθησης. Η ποιοτική ερμηνεία των παραγόμενων αποτελεσμάτων αυτών των μοντέλων παρέχει μια διεισδυτική οπτική στη διαδικασία συλλογισμού που ακολουθήθηκε για την παραγωγή τους.

Παράλληλα με την ανάγκη διεύρυνσης της επεξηγηματικότητας των μοντέλων μηχανικής μάθησης, υπάρχει επίσης η ανάγκη να διερευνηθεί η ευρωστία τους, δηλαδή ο βαθμός στον οποίο μπορούν να είναι ανθεκτικά, ευέλικτα και να προσαρμόζονται με ποιοτικό τρόπο σε σημαντικές ή μικρότερες τροποποιήσεις της εισόδου τους. Η παρούσα μελέτη αναδεικνύει πιθανές προκαταλήψεις που μπορεί να είναι ενσωματωμένες στα μοντέλα, οι οποίες σίγουρα επηρεάζουν τόσο τη δικαιοσύνη όσο και την αξιοπιστία τους. Έτσι, η συμπεριφορά και η απόκριση των μοντέλων καθίσταται μεροληπτική, γεγονός που υποδηλώνει την επείγουσα ανάγκη για την ανάπτυξη συστηματικών μεθόδων που μπορούν να εντοπίζουν και να αναδεικνύουν ποιοτικά τέτοια προβλήματα.

Η πολυτροπική μάθηση είναι ένας τομέας της τεχνητής νοημοσύνης που συνδυάζει πολλαπλές μορφές εισόδου. Τα μοντέλα που εκτελούν εργασίες πολυτροπικής μάθησης έχουν κερδίσει δημοτικότητα τα τελευταία χρόνια χάρη στην ευελιξία τους να χειρίζονται πολλαπλές μορφές εισόδου και κατά συνέπεια πιο σύνθετα προβλήματα που απαιτούν πολυπαραγοντικές προσεγγίσεις επίλυσης. Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία ασχολείται με την εργασία Απάντηση Οπτικών Ερωτήσεων - Visual Question Answering (VQA), η οποία συνδυάζει εικόνα και φυσική γλώσσα ως είσοδο. Συγκεκριμένα, τα μοντέλα απάντησης οπτικών ερωτήσεων δέχονται ερωτήσεις ανοικτού τύπου που αφορούν εικόνες και καλούνται να επιστρέψουν μια ποιοτική απάντηση σχετικά με το συνδυασμό αυτών των ερωτήσεων και εικόνων.

Στην παρούσα διπλωματική εργασία αναπτύσσουμε μια συστηματική μέθοδο για την αξιολόγηση της ευστάθειας τέτοιων μοντέλων και διερευνούμε την ύπαρξη πιθανών προκαταλήψεων στη

διαδικασία συλλογισμού που ακολουθούν, καθώς επίσης παρέχουμε και μια επεξηγηματική προσέγγιση για τη μελέτη των αποτελεσμάτων τους. Η μέθοδός μας βασίζεται στη χρήση βέλτιστων αντιπαραθετικών ερωτήσεων με γλωσσική υπόσταση και αναδεικνύει μια ποικιλία μεθόδων με τις οποίες μπορεί να αξιολογηθεί η απόκριση των μοντέλων VQA με τρόπο που να είναι πλήρως γενικεύσιμος και επεκτάσιμος σε οποιοδήποτε μοντέλο εκτελεί την εργασία VQA, ενώ παράλληλα αναγνωρίζει τη μη διαυγή φύση τέτοιων μοντέλων και τα χειρίζεται ως μαύρα κουτιά. Ο έλεγχος, η ποιότητα και ο σχεδιασμός των προτεινόμενων αντιπαραδειγματικών διαταραχών διασφαλίζεται από τη χρήση ιεραρχικών πηγών γνώσης, οι οποίες καθοδηγούν πλήρως τα πειράματά μας.

1.1.1 Κίνητρο και Συνεισφορά

Η αδιαμφισβήτητη άνοδος της δημοτικότητας της οπτικογλωσσικής (VL) μάθησης [48, 19, 44] έχει προσφέρει στην κοινότητα μια ποικιλία εντυπωσιακών υλοποιήσεων μοντέλων σε σύντομο χρονικό διάστημα [39, 36, 41, 35, 59, 38]. Η απάντηση οπτικών ερωτήσεων (VQA) είναι μια εργασία οπτικογλωσσικής μάθησης που έχει αποκτήσει θεμελιώδη ρόλο στην εξέλιξη διαφόρων διαδραστικών συστημάτων τεχνητής νοημοσύνης VL, όπως ο οπτικός διάλογος [22], η ανάκτηση κειμένου-εικόνων [21] και η οπτική κοινή λογική [67]. Σε αυτό το πλάισιο, υπάρχει ένα εκτεταμένο φάσμα εφαρμογών του πραγματικού κόσμου που επωφελούνται σημαντικά από τις νέες εξελίξεις γύρω από την εργασία VQA, όπως συστήματα υποστήριξης ατόμων με προβλήματα όρασης [8, 13] και αυτοκινούμενα αυτοκίνητα [10].

Η διαδικασία VQA περιλαμβάνει μια ερώτηση κειμένου q από ένα προκαθορισμένο σύνολο ερωτήσεων Q συνοδευόμενη από μια εικόνα I, η αλληλεπίδραση των οποίων παράγει μια απάντηση κειμένου a. Ο αγώνας για τη συνεχή βελτίωση της απόδοσης του μοντέλου VQA οδηγεί αναπόφευκτα στο να παραμένουν ανοιχτά ζητήματα, ιδίως λόγω της φύσης του μαύρου κουτιού των σύγχρονων υλοποιήσεων [12, 2, 15, 7, 26]. Αυτή η περιορισμένη πρόσβαση στη συλλογιστική που ακολουθούν αυτά τα μοντέλα για τη λήψη αποφάσεων υπογραμμίζει τον κίνδυνο αυθαίρετης συμπεριφοράς εκ μέρους τους. Αυτός ο κίνδυνος έγκειται κυρίως στην πιθανότητα ενσωμάτωσης μεροληψιών, αποφάσεων που δεν έχουν την κατάλληλη εστίαση, καθώς και στην απουσία εξηγήσιμης και δίκαιης απόδοσης των αποτελεσμάτων. Ειδικά όταν λαμβάνονται καίριες αποφάσεις με βάση αυτού του είδους τα συστήματα, η αδιαφάνεια τους τα καθιστά μη πρακτικά, και ενίοτε επικίνδυνα, για τις περισσότερες εφαρμογές. Αυτή η αβεβαιότητα υποδεικνύει την ανάγκη για νέες μεθόδους αξιολόγησης της ευρωστίας, οι οποίες δίνουν προτεραιότητα στη διαφάνεια των μοντέλων VQA.

Διαφορετικές προσεγγίσεις σχετικά με την αποκατάσταση των μεροληψιών και την εξηγησιμότητα των μοντέλων VQA εστιάζουν σε διαφορετικές πτυχές του ζητήματος. Για παράδειγμα, η [11] εξετάζει την ευρωστία και την επεξηγησιμότητα της VQA αντιμετωπίζοντας τους μετασχηματισμούς στην οπτική τροπικότητα, καθώς αποδίδει το πρόβλημα κυρίως στην οπτική μεροληψία ως προκύπτον από ανεπιθύμητες συσχετίσεις μεταξύ των εννοιών της εικόνας. Σε γενικές γραμμές, οι υπάρχουσες εργασίες επικεντρώνονται κυρίως στην επίδραση των οπτικών μεροληψιών και όχι στην επίδραση της γλωσσικών μεροληψιών, ως αιτία πίσω από την έλλειψη ευρωστίας στα μοντέλα VQA. Άλλες εργασίες ακολουθούν στρατηγικές βασισμένες στην προσοχή που απαιτούν εκτεταμένη γνώση της αρχιτεκτονικής του μοντέλου, επομένως δεν μπορούν να χειριστούν αποτελεσματικά τη φύση μαύρου κουτιού αυτών των συστημάτων. Για το σκοπό αυτό, διάφορες προσεγγίσεις που αφορούν συγκεκριμένα μοντέλα αποδεικνύονται αποδοτικές στο αυστηρό τους πλαίσιο, αλλά δεν έχουν τη δυνατότητα να γενικευτούν για την αξιολόγηση οποιουδήποτε άλλου μοντέλου, περιορίζοντας έτσι το πεδίο της αποτελεσματικότητάς τους σε μια μόνο συγκεκριμένη περίπτωση.

Υποστηρίζουμε ότι η επίλυση των προκλήσεων επεξηγησιμότητας στα μοντέλα VQA απαιτεί μια αντιπαραδειγματική προσέγγιση, η οποία υλοποιείται με διαταραχές σε επίπεδο λέξεων στις ερωτήσεις q- έτσι, αποκλίνουμε από την καθιερωμένη εξερεύνηση της οπτικής ιδιομορφίας, εξετάζοντας το ρόλο της γλώσσας σε πιθανές προκαταλήψεις και ψευδείς συσχετίσεις που κρύ-

βονται στα μοντέλα VQA, ενώ παράλληλα ανιχνεύουμε και ερμηνεύουμε τη διαδικασία λήψης αδιαφανών αποφάσεων. Οι προτεινόμενες από εμάς αντιπαραδειγματικές διαταραχές διαμορφώνονται ως εξής: *"Ποια είναι η απόκριση του μοντέλου VQA αν αντικαταστήσουμε τη λέξη X με τη λέξη Y στην ερώτηση q;"* Συγκεκριμένα, θεωρώντας τις λέξεις ως έννοιες, εκτελούμε τον ελάχιστο δυνατό εφικτό μετασχηματισμό για να διεγείρουμε μια αλλαγή στην απόκριση του μοντέλου- στη συνέχεια, γίνονται εύστοχες συγκρίσεις καταγράφοντας τη συμπεριφορά του μοντέλου σε διάφορους τέτοιους μετασχηματισμούς. Οι αντιπαραδειγματικές διαταραχές που εκτελούμε καθοδηγούνται πλήρως από την ντετερμινιστική διασφάλιση των δομών *ιεραρχικής γνώσης*. Με την αξιοποίηση αυτών των πηγών γνώσης, παρέχουμε μετασχηματισμούς που όχι μόνο είναι βέλτιστα στοχευμένοι σε κάθε συγκεκριμένη γλωσσική έννοια, αλλά είναι επίσης πλήρως *επεξηγήσιμοι* όσον αφορά τη στρατηγική που ακολουθείται για την εφαρμογή τους.

Ξεκινώντας από την παρατήρηση των *τοπικών αποκρίσεων του μοντέλου* σε διαφορετικές γλωσσικές διαταραχές για ένα ενιαίο δείγμα δεδομένων, προσδιορίζουμε περαιτέρω τα *καθολικά μοτίβα* που αναφέρονται στη συνολική συμπεριφορά του μοντέλου όταν αντιμετωπίζει ένα συγκεκριμένο σύνολο διαταραγμένων εννοιών. Ακολούθως, προτείνουμε καθολικούς κανόνες που χαρακτηρίζουν την απόκριση ενός μοντέλου και μπορούν να υπογραμμίσουν τις αδυναμίες του, υποδεικνύοντας ποιες έννοιες θα μπορούσαν να βλάψουν την ευρωστία και την αξιοπιστία του. Η διαδικασία αυτή αποκαλύπτει πιθανές μεροληψίες που έχει ενσωματώσει ένα μοντέλο και, ως εκ τούτου, αποδίδει εξηγήσεις ως προς το γιατί παράγεται μια συγκεκριμένη απάντηση στη θέση μιας άλλης- έτσι, είμαστε σε θέση να αποκτήσουμε πληροφορίες για τη διαδικασία συλλογισμού ενός μοντέλου, χωρίς να χρειάζεται πρόσβαση στην εσωτερική αρχιτεκτονική του.

Η μέθοδός μας μπορεί να γενικευτεί σε οποιοδήποτε μοντέλο VQA και αντίστοιχο κατάλληλο σύνολο δεδομένων, καθώς προσεγγίζει το θέμα με μια στρατηγική που δεν διαφοροποιείται καθόλου από το μοντέλο. Συνοψίζοντας, συνεισφέρουμε στα εξής:

1. Σχεδιάζουμε αντιπαραδειγματικές εισόδους εφαρμόζοντας μια ποικιλία δομημένων *αντικαταστάσεων σε επίπεδο λέξεων* στις ερωτήσεις $q \in Q$, οι οποίες καθοδηγούνται από ιεραρχικές πηγές γνώσης. Η προσέγγισή μας είναι ανεξάρτητη από το μοντέλο, καθώς αντιμετωπίζουμε οποιοδήποτε μοντέλο VQA ως μαύρο κουτί.

2. Λαμβάνουμε *τοπικές επεξηγήσεις* που προέρχονται από απροσδόκητες αποκρίσεις του μοντέλου σε αντιπαραδειγματικές εισόδους q.

3. Συνοψίζοντας τις τοπικές συμπεριφορές του μοντέλου για όλα τα $q \in Q$, εξάγουμε κάποιες *καθολικές επεξηγήσεις* που αποκαλύπτουν τη συνολική απόκριση του μοντέλου σε κάθε μία από τις σχεδιασμένες αντιπαραδειγματικές εισόδους.

1.1.2 Όργάνωση Εκτεταμένης Ελληνικής Περίληψης

Η ελληνική περίληψη θα χωριστεί στα παρακάτω υποκεφάλαια σε πλήρη συμφωνία και με το αγγλικό κείμενο.

Στο πρώτο υποκεφάλαιο 1.2 θα περιγράψουμε σε βάθος την εργασία Απάντησης Οπτικών Ερωτήσεων στον αναγνώστη και θα εξηγήσουμε εν συντομία τη μεθοδολογία που ακολουθείται από τη μέθοδο των αντιπαραδειγματικών επεξηγήσεων που προτείνουμε στην παρούσα διπλωματική εργασία, καθώς και την ιδέα που την δημιουργεί και τα κίνητρα που την καθιστούν χρήσιμη για το πρόβλημα που θέλουμε να αντιμετωπίσουμε. Επιπλέον, θα παρουσιάσουμε διάφορα μοντέλα VQA, με βάση τις διαφορετικές αρχιτεκτονικές, πεδίο εφαρμογής και σχεδιαστικές επιλογές, καθώς και διάφορα ευρέως χρησιμοποιούμενα και δημοφιλή σύνολα δεδομένων που εξυπηρετούν την εργασία που μας ενδιαφέρει και έχουν συμβάλει σημαντικά στην πρόοδο αυτού του πεδίου χάρη στη συνεισφορά τους στην εκπαίδευση και την αξιολόγηση.

Μια εποικοδομητική βιβλιογραφική ανασκόπηση θα βοηθήσει τον αναγνώστη να κατανοήσει σε βάθος διάφορες έννοιες που σχετίζονται στενά με το θέμα που πραγματεύεται η παρούσα δι-

πλωματική εργασία. Ως εκ τούτου, στο υποκεφάλαιο 1.3, θα παρουσιάσουμε μια εμπεριστατωμένη αλλά σύντομη βιβλιογραφική αναφορά στα διάφορα πεδία που σχετίζονται με την οπτικογλωσσική μάθηση και την επεξηγηματικότητα, που είναι σημαντικά όσον αφορά την παρούσα εργασία. Θα αναφερθούμε εκτενώς στα προ-εκπαιδευμένα μοντέλα που συνδυάζουν οπτικές και γλωσσικές μορφές εισόδου, αναλύοντάς τα στο πλαίσιο διαφόρων κατηγοριών. Επιπλέον, θα παραθέσουμε ένα περιγραφικό υπόβαθρο για το πεδίο της γνώσης στην οπτικογλωσσική μάθηση.

Στο επόμενο υποκεφάλαιο 1.4, θα εμβαθύνουμε στη μέθοδο Counterfactual Visual Question Answering που εισάγουμε. Θα παρουσιάσουμε εκτενώς τη δομή της προσέγγισής μας, θα αναφερθούμε στο proof-of-concept μοντέλο και στα σύνολα δεδομένων στα οποία εφαρμόζουμε τη γενικεύσιμη μέθοδό μας, και θα συζητήσουμε χρήσιμους συμβολισμούς, βήματα επεξεργασίας και τη διαδικασία που ακολουθούμε για να παράγουμε τις τελικές αντιπαραδειγματικές επεξηγήσεις μας. Αρχικά, θα παρουσιάσουμε μια διεισδυτική διερευνητική ανάλυση δεδομένων την οποία εφαρμόσαμε στα σύνολα δεδομένων μας ως προκαταρκτικό βήμα για να καθοδηγήσουμε τη σχεδιαστική μας προσέγγιση. Θα προβούμε σε ενδελεχή παρουσίαση των διαφόρων τύπων γλωσσικών αντιπαραδειγματικών διαταραχών που έχουμε σχεδιάσει και εφαρμόσει, ενώ επιπλέον θα αναλύσουμε τη χρήση των πηγών γνώσης στο αντιπαραδειγματικό μας σύστημα. Επιπλέον, θα εξηγήσουμε και θα σχολιάσουμε τα διάφορα επίπεδα και τους τομείς στους οποίους οι αντικαταστάσεις που υλοποιούμε παρέχουν χρήσιμες πληροφορίες τις οποίες θα αξιοποιήσουμε στη συνέχεια για να αποκτήσουμε μια ποιοτική ερμηνεία και αξιολόγηση των αποτελεσμάτων που παράγονται από τα μοντέλα που διερευνούμε, προσεγγίζοντας έτσι βέλτιστα το ζήτημα της επεξηγηματικότητας, της ανίχνευσης μεροληψιών και της αξιολόγησης της ευρωστίας.

Το υποκεφάλαιο 1.5 θα συνοψίσει τη συμβολή μας στο πρόβλημα. Θα παρουσιάσουμε ενδεικτικά τα αποτελέσματά μας με βάση τα πειράματα που υλοποιήθηκαν. Παρουσιάζοντας τις μετρικές αξιολόγησης και εξηγώντας πώς μπορούν να καταστούν εποικοδομητικές για το έργο μας, θα εμβαθύνουμε περαιτέρω σε μια λεπτομερέστερη ανάλυση των αποτελεσμάτων μας. Αυτό θα προσεγγιστεί μέσω της παρουσίασης των παρατηρούμενων μεροληψιών που ανιχνεύονται στην απόκριση του μοντέλου στις αντιπαραδειγματικές ερωτήσεις μας, οι οποίες κατηγοριοποιούνται από τα διάφορα πειράματα που υλοποιήθηκαν. Θα αναλύσουμε πώς αυτές οι μεροληψίες υποδεικνύουν μια συνολική συγκεκριμένη συμπεριφορά του μοντέλου, η οποία μπορεί να χαρακτηριστεί ως μη αξιόπιστη και μεροληπτική σε ορισμένα επίπεδα. Το ουσιαστικό όφελος της μεθόδου μας θα αναδειχθεί σε αυτό το υποκεφάλαιο, καθώς θα αναδείξουμε τις εντοπισθείσες μεροληψίες που διαμορφώνονται ως καθολικοί κανόνες που αποτυπώνουν τη συμπεριφορά του μοντέλου και τελικά καθοδηγούν και προδιαγράφουν τις διορθώσεις και τις βελτιώσεις που πρέπει να εφαρμοστούν προκειμένου να ενισχυθεί η αξιοπιστία και η αμεροληψία του υπό εξέταση μοντέλου.

Τέλος, στο τελευταίο υποκεφάλαιο 1.6 θα παρουσιάσουμε ένα τελικό συμπέρασμα της εργασίας μας, συνοψίζοντας τη μέθοδο και τη συμβολή μας, ενώ παράλληλα θα προτείνουμε κατευθυντήριες γραμμές και ιδέες για μελλοντικές προοπτικές που σχετίζονται με υποσχόμενες δυνατότητες επέκτασης που αφορούν τον τομέα που μελετάμε.

1.2 Απάντηση Οπτικών Ερωτήσεων

Η απάντηση οπτικών ερωτήσεων (VQA), που παρουσιάστηκε για πρώτη φορά στο [5], ανήκει στην κατηγορία των πολυτροπικών εργασιών μάθησης, καθώς δέχεται ως είσοδο τόσο οπτικές όσο και γλωσσικές παραμέτρους. Συγκεκριμένα, ένα μοντέλο VQA M λαμβάνει εικόνες i από ένα σύνολο I και σχετικές ερωτήσεις q που ανήκουν σε ένα προκαθορισμένο σύνολο ερωτήσεων Q, και αναμένεται να ανταποκριθεί με ακρίβεια σε αυτές τις ερωτήσεις q παρέχοντας μια απάντηση σε φυσική γλώσσα a. Οι προαναφερθείσες απαντήσεις μπορεί να είναι είτε ανοιχτές (παράγονται από το M) είτε να ανήκουν σε ένα σύνολο προκαθορισμένων υποψηφίων A. Η εργασία VQA είναι καθοδηγούμενη από το στόχο, δηλαδή η ανάπτυξή της υποκινείται από τη στόχευση επίλυσης συγκεκριμένων προβλημάτων και βελτίωσης ευρείας κλίμακας ζητημάτων που σχετίζονται με τη

συνεργασία και τη συνύπαρξη εισόδων εικόνας και φυσικής γλώσσας.

Σύμφωνα με κάθε συγκεκριμένη περίπτωση, οι οπτικές ερωτήσεις q στοχεύουν επιλεκτικά σε διαφορετικές περιοχές μιας εικόνας I, συμπεριλαμβανομένων των λεπτομερειών του φόντου και του υποκείμενου πλαισίου. Αντίστοιχα, η εστίαση όσον αφορά τη γλωσσική είσοδο έγκειται σε διαφορετικές λεκτικές έννοιες ανάλογα με κάθε ζεύγος εικόνας-ερώτησης.

Γενικά, οι ερωτήσεις q έχουν αυθαίρετη φύση και περικλείουν διαφορετικά υποπροβλήματα όρασης υπολογιστών, όπως η αναγνώριση λεπτομερούς ανάλυσης (Εικ. 3.2), η αναγνώριση αντικειμένων (Εικ. 3.3) και η ανίχνευση αντικειμένων (Εικ. 3.4), η αναγνώριση δραστηριοτήτων (Εικ. 3.5), η συλλογιστική βάσει γνώσης (Εικ. 3.6), η ταξινόμηση χαρακτηριστικών (Εικ. 3.7), η ταξινόμηση σκηνών (Εικ. 3.8), καθώς και καταμέτρηση (Εικ. 3.9) [32]. Επιπλέον, οι πιο περίπλοκες ερωτήσεις αφορούν πιο σύνθετες διαδικασίες, όπως οι χωρικές σχέσεις μεταξύ αντικειμένων (Εικ. 3.10) και η κοινή λογική (Εικ. 3.11) [23, vqa]. Σε γενικές γραμμές, ένα μοντέλο VQA που επιδεικνύει υψηλά επίπεδα ευρωστίας, αποδοτικότητας και ευελιξίας είναι αρκετά ικανό ώστε να ανταποκρίνεται στις απαιτήσεις για την επίλυση ενός μεγάλου εύρους κλασικών εργασιών όρασης υπολογιστών παράλληλα με την ορθή συλλογιστική επί του συνδυασμού εικόνων και σχετικών, συχνά περίπλοκων ή δύσκολων, ερωτήσεων [5], [33].

Η απάντηση ερωτήσεων, ακόμη και εκείνων με δυαδικές απαντήσεις, είναι ένα πολύπλοκο έργο που απαιτεί μια αυστηρή διαδικασία. Όταν πρόκειται για ερωτήματα που σχετίζονται με εικόνες, οι απλές απαντήσεις που αποτελούνται από λίγες μόνο λέξεις συχνά αρκούν. Η αποτελεσματικότητα ενός αλγορίθμου σε τέτοιες περιπτώσεις μπορεί να μετρηθεί από τον αριθμό των σωστών απαντήσεων που παράγει. Ενώ οι ερωτήσεις ανοικτού τύπου απαιτούν μια απάντηση ελεύθερης μορφής, οι ερωτήσεις πολλαπλής επιλογής απαιτούν από τους αλγορίθμους να επιλέγουν από ένα προκαθορισμένο σύνολο πιθανών απαντήσεων.

1.2.1 Εφαρμογές

Το VQA έχει ποικίλες πιθανές εφαρμογές, ιδίως ως βοήθημα για άτομα με προβλήματα όρασης για την πρόσβαση σε πληροφορίες σχετικά με εικόνες τόσο στο διαδίκτυο όσο και στον πραγματικό κόσμο. Μπορεί επίσης να βελτιώσει την αλληλεπίδραση ανθρώπου-υπολογιστή ως ένα φυσικό μέσο για την αναζήτηση οπτικού περιεχομένου και να επιτρέψει την ανάκτηση εικόνων χωρίς μεταδεδομένα ή ετικέτες. Επιπλέον, το VQA παρουσιάζει μια σημαντική ερευνητική πρόκληση, καθώς ένα επιτυχημένο σύστημα πρέπει να είναι ικανό να επιλύει διάφορα προβλήματα όρασης υπολογιστών, καθιστώντας το αναπόσπαστο μέρος ενός Turing Test για την κατανόηση εικόνων [33].

1.2.2 Σχετικές Εργασίες

Ο πρωταρχικός στόχος του VQA είναι η εξαγωγή σημασιολογίας εικόνας που σχετίζεται με ένα δοθέν ερώτημα, η οποία περιλαμβάνει τόσο μικρές λεπτομέρειες όσο και αφηρημένα χαρακτηριστικά της σκηνής. Ενώ οι εργασίες όρασης υπολογιστών, όπως η αναγνώριση αντικειμένων, η αναγνώριση δραστηριοτήτων και η ταξινόμηση σκηνών, εστιάζουν στην εξαγωγή πληροφοριών από εικόνες, το πεδίο εφαρμογής τους είναι σχετικά περιορισμένο σε σύγκριση με το VQA. Εκτός από το VQA, υπάρχουν και άλλες προσεγγίσεις που συνδυάζουν την όραση και τη γλώσσα. Μια από τις πιο διερευνημένες μεθόδους είναι η δημιουργία λεζάντας εικόνας, όπου ένας αλγόριθμος στοχεύει στη δημιουργία της περιγραφής μιας εικόνας σε φυσική γλώσσα.

1.2.3 Μοντέλα και Σύνολα Δεδομένων για την Απάντηση Οπτικών Ερωτήσεων

Από την αρχική εισαγωγή του VQA [5], πολλές προσπάθειες έχουν επεκτείνει αυτό το πρότυπο, είτε προτείνοντας προηγμένες αρχιτεκτονικές μοντέλων είτε υποδεικνύοντας πιο απαιτητικά σύνολα δεδομένων. Τα σύγχρονα μοντέλα που ασχολούνται με την εργασία VQA βασίζονται κυρίως

σε οπτικογλωσσικούς μετασχηματιστές- έτσι, μοντέλα όπως τα ViLBERT [41], VisualBERT [39], FLAVA [59], ALBEF [38], ViLT [36] και άλλα έχουν κυριαρχήσει στην πρόσφατη βιβλιογραφία VQA επιδεικνύοντας ταχείες βελτιώσεις σε σχετικά σύνολα δεδομένων αναφοράς.

Έχουν προταθεί διάφορα κυρίαρχα σύνολα δεδομένων VQA, όπως το Visual Genome, το σύνολο δεδομένων VQA [5], το DAQUAR [47], το COCO-QA [54], το FM-IQA [24] και άλλα. Το Visual Genome (VG) είναι ένα σύνολο δεδομένων μεγάλης κλίμακας που περιλαμβάνει πολυάριθμες εικόνες σκηνών, σημειώσεις αντικειμένων, χαρακτηριστικών και σχέσεων, καθώς και οπτικά ζεύγη ερωτήσεων-απαντήσεων [37]. Επιπλέον, προκειμένου να αντιμετωπιστεί η στατιστική μεροληψία που υπάρχει στα τρέχοντα σύνολα δεδομένων VQA, έχουν καταβληθεί προσπάθειες για τη δημιουργία συνόλων δεδομένων όπως το VQA v2 [26] και το VQA-CP [4]. Οι βελτιώσεις του αρχικού VQA (VQA-v2) προτείνουν την προσθήκη παρόμοιων ζευγών εικόνων που αντιστοιχούν στην ίδια ερώτηση q, αλλά οδηγούν σε αποκλίνουσες απαντήσεις [26]. Αυτά τα σύνολα δεδομένων αποσκοπούν στο να εμποδίσουν τη γλωσσική μεροληψία και να βελτιώσουν τη δυνατότητα οπτικής κατανόησης του μοντέλου [70]. Η προσέγγισή μας σε αυτή τη διπλωματική εργασία δοκιμάζεται τόσο στα σύνολα δεδομένων VQA-v2 όσο και στα σύνολα δεδομένων Visual Genome.

1.2.4 Αξιολόγηση

Οι μετρικές αξιολόγησης για το ανοικτού τύπου VQA αποτελούν θέμα συνεχιζόμενης έρευνας. Η VQA μπορεί να αξιολογηθεί είτε ως εργασία ανοικτού τύπου, όπου οι αλγόριθμοι δημιουργούν μια συμβολοσειρά για να απαντήσουν σε μια ερώτηση, είτε ως εργασία πολλαπλών επιλογών. Για την αξιολόγηση της απόδοσης του μοντέλου σε ερωτήσεις πολλαπλής επιλογής, μια απλή και αποτελεσματική προσέγγιση είναι η χρήση της απλής ακρίβειας, η οποία περιλαμβάνει τον υπολογισμό του λόγου των σωστών απαντήσεων προς τον συνολικό αριθμό των απαντήσεων. Ωστόσο, η αξιολόγηση των αλγορίθμων VQA ανοικτού τύπου με τη χρήση της απλής ακρίβειας μπορεί να είναι πολύ αυστηρή, επειδή ορισμένα σφάλματα είναι πιο σοβαρά από άλλα. Για παράδειγμα, ένα σύστημα που εξάγει μια πολύ παρόμοια απάντηση με τη θεμελιώδη αλήθεια μπορεί να εξακολουθεί να θεωρείται εσφαλμένο, παρόλο που είναι αρκετά κοντά. Επιπλέον, ορισμένες ερωτήσεις μπορεί να έχουν πολλαπλές σωστές απαντήσεις, γεγονός που οδηγεί σε περαιτέρω προβλήματα με τη χρήση της απόλυτης ακρίβειας [33].

1.2.5 Ο Ρόλος των Γλωσσικών Μεροληψιών

Η επιτυχία των συστημάτων απάντησης οπτικών ερωτήσεων (VQA) εξαρτάται από την αποτελεσματική αξιοποίηση τόσο της εικόνας όσο και των γλωσσικών δεδομένων για την επίτευξη ισχυρών επιδόσεων. Ωστόσο, μελέτες έχουν δείξει ότι τα τρέχοντα συστήματα VQA μπορεί να μην αξιοποιούν αποτελεσματικά τόσο την όραση όσο και τη γλώσσα, αλλά βασίζονται σε μεγάλο βαθμό στη γλώσσα. Μελέτες έχουν δείξει ότι τα μοντέλα που χρησιμοποιούν μόνο ερωτήσεις αποδίδουν σημαντικά καλύτερα από εκείνα που χρησιμοποιούν μόνο εικόνες, ειδικά σε ανοικτού τύπου δεδομένα COCO-VQA [30], [5]. Τα αποτελέσματα αυτού του πειράματος σε συνδυασμό με τα αποτελέσματα που αφορούν άλλα σύνολα δεδομένων καθώς και άλλες σχετικές μελέτες [3], [68] υποδεικνύουν ότι η γλωσσική προκατάληψη βλάπτει σημαντικά την απόδοση και τις δυνατότητες ευρωστίας των μοντέλων VQA. Αυτό οφείλεται στη φύση των ερωτήσεων στα σύνολα δεδομένων VQA, η οποία συχνά περιορίζει τις αναμενόμενες απαντήσεις μετατρέποντας ουσιαστικά τις ερωτήσεις ανοικτού τύπου σε ερωτήσεις πολλαπλής επιλογής, καθώς και στην ισχυρή προκατάληψη στα σύνολα δεδομένων. Ως εκ τούτου, τα τρέχοντα συστήματα VQA βασίζονται περισσότερο στην ερώτηση παρά στο περιεχόμενο της εικόνας και η γλωσσική προκατάληψη στα σύνολα δεδομένων επηρεάζει σημαντικά την απόδοσή τους, περιορίζοντας την ανάπτυξή τους. Για να αντιμετωπιστεί αυτό το ζήτημα, τα νέα σύνολα δεδομένων VQA θα πρέπει να προσπαθήσουν να αντισταθμίσουν την προκατάληψη είτε με ερωτήσεις που επιβάλλουν την ανάλυση του περιεχομένου της εικόνας είτε καθιστώντας τα σύνολα δεδομένων λιγότερο μεροληπτικά [33].

1.2.6 Αντιπαραδειγματικές Επεξηγήσεις για την Απάντηση Οπτικών Ερωτήσεων

Η αντιπαραδειγματική προσέγγισή μας όσον αφορά τις γλωσσικές αντικαταστάσεις περιστρέφεται γύρω από το ακόλουθο θεμελιώδες ερώτημα: *"Ποια είναι η απάντηση του M αν αντικαταστήσουμε τη λέξη X με τη λέξη Y στην ερώτηση q;"*. Η υλοποιούμενη αντιπαραδειγματική αντικατάσταση $X \rightarrow Y$ θα πρέπει να είναι σημασιολογικά *ελάχιστη* και γλωσσικά *εφικτή*. Η *ελαχιστότητα* αναφέρεται σε αντικαταστάσεις που διατηρούν ένα νόημα κοντά στο νόημα της αρχικής λέξης X. Για παράδειγμα, οι συνώνυμες λέξεις διατηρούν αυτόν τον περιορισμό της ελαχιστότητας. Προκειμένου να διασφαλίσουμε τη σημασιολογική ελαχιστότητα των αντικαταστάσεων, αξιοποιούμε πηγές λεξιλογικής γνώσης (όπως το WordNet [23]), οι οποίες μπορούν να παρέχουν τις ελάχιστες δυνατές μεταβάσεις $X \rightarrow Y$ επιλέγοντας την πλησιέστερη έννοια Y στην έννοια X που σέβεται ορισμένους περιορισμούς. Η γλωσσική *εφικτότητα* υποδεικνύει ουσιαστικές αντικαταστάσεις που αφορούν πάντα το ίδιο μέρος του λόγου - για παράδειγμα, τα ουσιαστικά μπορούν να αντικατασταθούν μόνο από ουσιαστικά αλλά όχι από ρήματα. Συνολικά, τέτοιες αντικαταστάσεις $X \rightarrow Y$ εφαρμόζονται σε ολόκληρο το σύνολο Q, στοχεύοντας σε ένα μερός του λόγου κάθε φορά.

Τέτοιου είδους αντιπαραδειγματικές ερωτήσεις είναι σε θέση να ενεργοποιήσουν εναλλακτικές απαντήσεις του μοντέλου. Επομένως, μια αντικατάσταση έννοιας $X \rightarrow Y$ στην είσοδο μπορεί να οδηγήσει σε μια εναλλακτική απάντηση $X' \rightarrow Y'$ στην έξοδο, ή όχι. Η διερεύνηση πιθανών αλλαγών στην έξοδο είναι ιδιαίτερα κατατοπιστική όσον αφορά τη διαδικασία συλλογισμού που ακολουθεί το μοντέλο M, αναδεικνύοντας έννοιες ή οικογένειες εννοιών που έχουν μεγαλύτερη ή μικρότερη επιρροή στη διαδικασία λήψης αποφάσεων του M. Ως εκ τούτου, οι αντιπαραδειγματικές αντικαταστάσεις που εφαρμόζονται στο q παρέχουν χρήσιμες εξηγήσεις για την παρατηρούμενη συμπεριφορά του μοντέλου και ενισχύουν το βαθμό ερμηνευσιμότητάς του, ενώ παράλληλα το χειρίζονται ως δομή μαύρου κουτιού.

1.3 Οπτικογλωσσική Μάθηση & Προσεγγίσεις Επεξηγησιμότητας

Τα τελευταία χρόνια, τα μοντέλα προ-εκπαίδευσης (PTM) έχουν φέρει επανάσταση σε τομείς όπως η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας. Έχει αποδειχθεί ότι είναι εξαιρετικά αποτελεσματικά στη βελτίωση της απόδοσης σε καθιερωμένες εργασίες, αποφεύγοντας παράλληλα την ανάγκη εκπαίδευσης νέων μοντέλων από το μηδέν. Η προσαρμογή των μοντέλων προ-εκπαίδευσης στον τομέα της μάθησης όρασης και γλώσσας (V-L) έχει γίνει κεντρικό σημείο της έρευνας για την πολυτροπική μάθηση, καθώς οι ερευνητές επιδιώκουν να βελτιώσουν τις επιδόσεις σε καθιερωμένες εργασίες. Σημαντική πρόοδος έχει σημειωθεί στη διερεύνηση της εφαρμογής προ-εκπαιδευμένων μοντέλων σε πολυτροπικές εργασίες.

1.3.1 Προεκπαιδευμένα Οπτικογλωσσικά Μοντέλα

Οι ερευνητές της τεχνητής νοημοσύνης επιδιώκουν εδώ και καιρό να δημιουργήσουν μηχανές που να μπορούν να σκέφτονται και να ανταποκρίνονται όπως οι άνθρωποι. Για να το επιτύχουν αυτό, οι ερευνητές έχουν προτείνει διάφορες εργασίες για την εκπαίδευση και την αξιολόγηση των μηχανών, όπως η αναγνώριση προσώπου, η κατανόηση της ανάγνωσης και ο διάλογος ανθρώπου-μηχανής. Ωστόσο, λόγω τεχνολογικών περιορισμών, η εκπαίδευση σε μεγάλο όγκο επισημασμένων δεδομένων είναι συχνά απαραίτητη για τη δημιουργία ικανών μοντέλων. Επιπλέον, μοντέλα βαθιάς μάθησης, όπως τα RNN, CNN και Transformer, έχουν εφαρμοστεί για την επίλυση εργασιών V-L, αλλά είναι συνήθως σχεδιασμένα για συγκεκριμένες εργασίες, γεγονός που οδηγεί σε ελλιπή μεταφερσιμότητα [14].

Για την αντιμετώπιση αυτού του προβλήματος, οι ερευνητές προ-εκπαιδεύουν ένα τεράστιο μοντέλο σε σύνολα γενικών δεδομένων μεγάλης κλίμακας και το τελειοποιούν σε συγκεκριμένες μεταγενέστερες εργασίες για να αυξήσουν τη μεταφερσιμότητα. Τα μοντέλα προ-εκπαίδευσης που χρησιμοποιούν τη δομή Transformer [63] έχουν απαλύνει αυτό το πρόβλημα αρχικά μέσω της

προ-εκπαίδευσης με αυτοεπιβλεπόμενη μάθηση σε μη επισημειωμένα δεδομένα και στη συνέχεια με τη λεπτομερή ρύθμιση με μια μικρή ποσότητα επισημειωμένων δεδομένων σε μεταγενέστερες εργασίες. Μοντέλα προ-εκπαίδευσης όπως τα BERT [17], ViT [18] και Wave2Vec [57] έχουν επιτύχει σε μονοτροπικά πεδία όπως η Επεξεργασία Φυσικής Γλώσσας, η Όραση Υπολογιστών και η Ομιλία, αλλά το ερώτημα παραμένει αν μπορούν να εφαρμοστούν σε πολυτροπικά καθήκοντα. Οι ερευνητές έχουν διερευνήσει αυτό το πρόβλημα στον τομέα της Όρασης-και-Γλώσσας (VLP) και έχουν σημειώσει σημαντική πρόοδο στην εκμάθηση της σημασιολογικής αντιστοιχίας μεταξύ διαφορετικών μορφών μέσω έξυπνα σχεδιασμένων αρχιτεκτονικών μοντέλων [14]. Ο μετασχηματιστής (Transformer) έχει γίνει η ραχοκοκαλιά των περισσότερων προ-εκπαιδευμένων γλωσσικών μοντέλων (PLM), όπως το BERT και το GPT-3, τα οποία έχουν επιτύχει νέα κορυφαία αποτελέσματα σε διάφορες μεταγενέστερες εργασίες, συμπεριλαμβανομένης της απάντησης οπτικών ερωτήσεων και της δημιουργίας λεζάντας εικόνας [20].

Η διαδικασία προ-εκπαίδευσης ενός VL-PLM αποτελείται από τις ακόλουθες κύριες υποενότητες: Εξαγωγή χαρακτηριστικών - κωδικοποίηση, αρχιτεκτονική μοντέλου, αποτελεσματικοί στόχοι προ-εκπαίδευσης, σύνολα δεδομένων προ-εκπαίδευσης και μεταγενέστερες οπτικογλωσσικές εργασίες. Ορισμένες από αυτές τις εργασίες είναι οι ακόλουθες: Οπτική Απάντηση Ερωτήσεων (VQA), Οπτική Συλλογιστική και Συνθετική Απάντηση Ερωτήσεων (GQA), Οπτική Συμπλήρωση (VE), Οπτική Συλλογιστική Κοινής Λογικής (VCR), Οπτική-Γλωσσική Ανάκτηση (VLR), Οπτική Υποτιτλισμός (VC), Οπτικός Διάλογος (VD) και Πολυτροπική Μετάφραση (MMT). Η ανάλυση που παρουσιάζεται παραπάνω, η οποία σκιαγραφεί τη διαδικασία προ-εκπαίδευσης, βασίζεται στις σχετικές έρευνες για την οπτικογλωσσική προ-εκπαίδευση [14], [20].

1.3.2 Γνώση στην Οπτικογλωσσική Μάθηση

Η ραγδαία πρόοδος στην οπτικογλωσσική μάθηση (VL) έχει οδηγήσει στην εμφάνιση διαφόρων μοντέλων και τεχνικών που επιδεικνύουν αξιοσημείωτες ικανότητες στην επίλυση εργασιών που απαιτούν τη συνεργασία της όρασης και της γλώσσας. Ωστόσο, τα υπάρχοντα σύνολα δεδομένων προ-εκπαίδευσης VL έχουν περιορισμούς όσον αφορά την ποσότητα της οπτικής και γλωσσικής γνώσης που περιλαμβάνουν. Κατά συνέπεια, οι δυνατότητες γενίκευσης πολλών μοντέλων VL είναι περιορισμένες. Για να ξεπεραστεί αυτή η πρόκληση, έχουν εισαχθεί υβριδικές αρχιτεκτονικές, οι οποίες ενσωματώνουν εξωτερικές πηγές γνώσης, όπως οι γράφοι γνώσης (KG) και τα μεγάλα γλωσσικά μοντέλα (LLM). Αυτές οι πηγές γνώσης βοηθούν στη γεφύρωση των κενών σε ελλιπείς πληροφορίες και ενισχύουν τη συνολική απόδοση των μοντέλων VL [45].

Ενώ τα μοντέλα VL έχουν κάνει σημαντικά βήματα στην επίτευξη κατανόησης του πραγματικού κόσμου μέσω εκτεταμένων διαδικασιών προ-εκπαίδευσης, εξακολουθούν να παρουσιάζουν ορισμένους περιορισμούς. Συγκεκριμένα, η κατανόηση της κοινής λογικής, των πραγματικών πληροφοριών, των χρονικών παραμέτρων και της καθημερινής γνώσης παραμένει σχετικά περιορισμένη, εγείροντας ερωτήματα σχετικά με την επεκτασιμότητα των εργασιών VL. Με την αξιοποίηση των γραφημάτων γνώσης και άλλων εξωτερικών πηγών γνώσης, αυτά τα κενά μπορούν να αντιμετωπιστούν αποτελεσματικά παρέχοντας με σαφήνεια τις πληροφορίες που λείπουν και ενδυναμώνοντας τα μοντέλα VL με νέες δυνατότητες. Επιπλέον, η ενσωμάτωση των γραφημάτων γνώσης όχι μόνο καλύπτει αυτά τα κενά γνώσης, αλλά και ενισχύει την επεξηγηματικότητα, τη δικαιοσύνη και την αξιοπιστία των διαδικασιών λήψης αποφάσεων, που αποτελούν κρίσιμα ζητήματα κατά την ανάπτυξη σύνθετων υλοποιήσεων VL [46].

Τα τελευταία χρόνια, υπάρχει αυξανόμενο ενδιαφέρον για την ενσωμάτωση εξωτερικής γνώσης, η οποία συμβολίζεται ως K, σε οπτικογλωσσικά μοντέλα (VL). Η προσέγγιση αυτή έχει κερδίσει αναγνώριση για τις δυνατότητές της να βελτιώσει την απόδοση, την επεκτασιμότητα, ακόμη και την επεξηγηματικότητα των υφιστάμενων εργασιών VL. Η ακόλουθη περιγραφή των διαφόρων κατηγοριών πρόσθετης γνώσης, καθώς και των τύπων της μορφής της εξωτερικής γνώσης, βασίζεται στη σύνοψη και τα συμπεράσματα που παρέχονται στις σχετικές έρευνες σχετικά με τον αντίκτυπο της ενσωμάτωσης της γνώσης στην οπτικογλωσσική μάθηση [45, 46]. Η ένταξη πρό-

σθετης γνώσης μπορεί να προσφέρει τα προαναφερθέντα οφέλη, μέσω των διαφόρων τύπων που περιλαμβάνει, όπως αναφέρονται ακολούθως: Ιεραρχική γνώση, Λεξιλογική γνώση, Ονομαστικές οντότητες, Πραγματικές γνώσεις, Γνώση κοινής λογικής, Γνώση γεγονότων/χρόνου και Οπτική γνώση.

Η επιλογή της εξωτερικής πηγής γνώσης διαδραματίζει κρίσιμο ρόλο στον καθορισμό του τρόπου πρόσβασης και χρήσης της γνώσης στο πλαίσιο των μοντέλων VL. Η εξωτερική γνώση μπορεί να αναλυθεί περαιτέρω εάν ταξινομηθεί στις ακόλουθες ομάδες: Άμεση - Ρητή γνώση, Έμμεση γνώση και Γνώση που αντλείται από τον Παγκόσμιο Ιστό.

1.3.3 Επεξηγησιμότητα στην Απάντηση Οπτικών Ερωτήσεων

Όσον αφορά το ερευνητικό θέμα της επεξηγηματικότητας και της ευρωστίας στην απάντηση οπτικών ερωτήσεων [50, 6, 27], έχουν προταθεί πολλαπλές προσπάθειες, συμπεριλαμβανομένων των χαρτών προσοχής [42, 29], και άλλων προσεγγίσεων που αφορούν συγκεκριμένα μοντέλα [56]. Η στρατηγική μέσω των αντιπαραδειγμάτων είναι μάλλον καινούργια, ενώ οι ήδη υπάρχουσες προσπάθειες επικεντρώνονται στις οπτικές διαταραχές [11], στη συγκάλυψη (masking) [15, 16], στην εισαγωγή αντιπαραδειγμάτων στο στάδιο της εκπαίδευσης [1, 16] και στις προσεγγίσεις που βασίζονται στη συσχέτιση μεταξύ αρχικών και αντιπαραδειγμάτων [40].

1.3.4 Γλωσσικές Αντικαταστάσεις

Υπάρχει μια ποικιλία προηγούμενων εργασιών που εκτελούν γλωσσικές αντικαταστάσεις σε επίπεδο λέξης, παρόλο που στοχεύουν σε καθαρά γλωσσικά προβλήματα, κυρίως στην ταξινόμηση κειμένων [55, 65, 25, 49, 34], αλλά και στη σημασιολογική ομοιότητα [43] και στη μηχανική μετάφραση [64]. Οι διαταραχές μας όσον αφορά την αντικατάσταση συνωνύμων και την τυχαία διαγραφή ουσιαστικών είναι εμπνευσμένες από την [65], με καθοδήγηση των αντικαταστάσεων με τη χρήση του WordNet [23].

Οι χρωματικές διαταραχές είναι προσαρμοσμένες από το [43], βάσει του οποίου κατασκευάζουμε μια κατάλληλη ιεραρχία με βάση την απόσταση χρώματος. Οι υπόλοιπες αντικαταστάσεις που υλοποιήσαμε και αφορούν σε ουσιαστικά και ρήματα είναι εντελώς νέες ιδέες.

1.4 Μέθοδος

Το πλαίσιο που προτείνουμε λειτουργεί ως εξής: Γενικά, ένα μοντέλο VQA λαμβάνει μια ερώτηση (Q) και μια σχετική εικόνα (I), και με βάση αυτά, επιλέγει μια κατάλληλη απάντηση (A) μεταξύ προκαθορισμένων υποψηφίων απαντήσεων. Η αντιπαραδειγματική μέθοδος που χρησιμοποιούμε αναφέρεται στην ελάχιστη δυνατή εφικτή γλωσσική αντικατάσταση για την επίτευξη μιας αλλαγής στην απάντηση ενός μοντέλου. Στην εργασία μας εστιάζουμε σε κειμενικά αντιπαραδείγματα, που επηρεάζουν είτε το Q είτε το A κάθε φορά, καθώς το κείμενο είναι ιδιαίτερα κατάλληλο για εννοιολογικές αντικαταστάσεις (μια λέξη μπορεί εύκολα να αντικατασταθεί από μια άλλη λέξη) σε αντίθεση με τις οπτικές αντικαταστάσεις (ένα αντικείμενο δεν μπορεί εύκολα να αντικατασταθεί ποιοτικά από ένα άλλο αντικείμενο). Η είσοδος του συστήματός μας αποτελείται από το σύνολο δεδομένων D που περιέχει ευθυγραμμισμένες εικόνες (I), ερωτήσεις κειμένου (Q) και υποψήφιες απαντήσεις κειμένου (A). Οι είσοδοι του D εισάγονται σε ένα προεκπαιδευμένο VQA μοντέλο M. Επιλέξαμε το ViLT ως τεκμήριο και πλάισιο εφαρμογής της ιδέας μας, και πραγματοποιήσαμε την ανάλυση στα σύνολα δεδομένων Visual Genome και VQA-v2. Στην εργασία μας, εφαρμόζουμε λεκτικές αντικαταστάσεις που καθοδηγούνται από εξωτερικές πηγές γνώσης. Χρησιμοποιούμε τον γράφο γνώσης WordNet (ο οποίος παρέχει ιεραρχικές σχέσεις μεταξύ κοινών λέξεων) και μια ιεραρχία των χρωματικής συσχέτισης (η οποία έχει επεκταθεί από τις σχέσεις color2 της Matplotlib).

Μια καίρια πτυχή της εργασίας μας έγκειται στη βαθιά κατανόηση των αντίστοιχων συνόλων δεδομένων στα οποία εκτελείται ένα μοντέλο. Για τον λόγο αυτό, εφαρμόσαμε μια διεξοδική διερευνητική ανάλυση δεδομένων στα σύνολα δεδομένων Visual Genome και VQA-v2, προκειμένου να αποκτήσουμε πληροφορίες που καθοδήγησαν τον σχεδιασμό των αντικαταστάσεών μας. Αργότερα, θα παρουσιαστούν και θα αναλυθούν ορισμένες ενδεικτικές περιπτώσεις οπτικοποίησης δεδομένων που απεικονίζουν χρήσιμα χαρακτηριστικά των συνόλων δεδομένων Visual Genome και VQA-v2.

Η συνεισφορά και η καινοτομία της εργασίας μας μπορούν να συνοψιστούν ως εξής: Καθ' όλη τη διάρκεια αυτής της διαδικασίας, αξιολογούμε αν και πώς αλλάζει η απόκριση του , ως ένδειξη της ευρωστίας του έναντι αντικαταστάσεων με σημασιολογικά συναφείς έννοιες. Οι μεμονωμένες τιμές ακρίβειας δεν είναι αρκετά κατατοπιστικές για να εξηγήσουν γιατί παρατηρούμε τέτοιες συμπεριφορές. Για το σκοπό αυτό, εξετάζουμε ξεχωριστά τα δείγματα όπου το προβλεπόμενο A αλλάζει υπό την παρουσία μιας διαταραχής, λαμβάνοντας τοπικές εξηγήσεις που προέρχονται από απροσδόκητες αποκρίσεις του μοντέλου σε αντιπαραδειγματικές εισόδους Q ή A, της μορφής: "αν η έννοια αλλάζει στο Q (ή στο A), τότε το A -εσφαλμένα- αλλάζει". Η συνάθροιση τέτοιων τοπικών κανόνων οδηγεί σε καθολικές επεξηγήσεις, οι οποίες προκύπτουν από σχέσεις "αν-τότε" που ισχύουν για πολλαπλά δείγματα του D. Αυτές αποκαλύπτουν τη συνολική απόκριση του μοντέλου σε κάθε μία από τις σχεδιασμένες αντιπαραδειγματικές εισόδους. Σχεδιάζουμε τις αντιφατικές εισόδους Q και A χρησιμοποιώντας μια ποικιλία δομημένων αντικαταστάσεων σε επίπεδο λέξεων, όπως καθοδηγείται από ιεραρχικές πηγές γνώσης, εφαρμόζοντας έτσι ντετερμινιστικές, ελεγχόμενες, βέλτιστες αντιπαραδειγματικές διαταραχές. Η προσέγγισή μας είναι ανεξάρτητη από το μοντέλο, καθώς αντιμετωπίζουμε οποιοδήποτε μοντέλο VQA ως μαύρο κουτί.

Για να παραθέσουμε ένα διεισδυτικό παράδειγμα των διαταραχών μας, ας θεωρήσουμε την ερώτηση "Τι χρώμα έχει η γάτα;", που αναφέρεται σε μια σχετική εικόνα. Πραγματοποιώντας τον υπερνυμικό μετασχηματισμό από "γάτα" σε "ζώο", επιδιώκουμε να αξιολογήσουμε κατά πόσον το μοντέλο θα επηρεαστεί. Συγκεκριμένα, εάν η αρχική του απάντηση μεταβάλλεται λανθασμένα ή εάν παρατηρήσουμε ότι χειρίζεται τέτοιες αντικαταστάσεις με μειωμένη βεβαιότητα, μπορούμε να συμπεράνουμε ότι το μοντέλο δεν αντιλαμβάνεται αποτελεσματικά τέτοιες ιεραρχικές σχέσεις υπερώνυμων-υπωνύμων.

1.4.1 Περιγραφή του Πλαισίου

Η είσοδος του πλαισίου μας αποτελείται από ένα σύνολο δεδομένων D που περιέχει ευθυγραμμισμένες εικόνες I, ερωτήσεις κειμένου που αποτελούν ένα σύνολο Q και υποψήφιες απαντήσεις κειμένου A. Αργότερα θα παρουσιάσουμε αποτελέσματα για το Visual Genome (VG) [37] και VQA-v2 [26], τα οποία ικανοποιούν αυτές τις απαιτήσεις.

Επιλέγουμε το ViLT [36] ως ένα προ-εκπαιδευμένο μοντέλο VQA M. Παρόλα αυτά, η προτεινόμενη μέθοδος μας δεν περιορίζεται στο ViLT, καθώς εξετάζει μόνο τις εισόδους (ερωτήσεις) και τις εξόδους (απαντήσεις). Το ViLT λαμβάνει μια ερώτηση q Q και μια εικόνα i I από το D και στη συνέχεια παράγει μια απάντηση a, αντί να επιλέγει έναν από τους υποψηφίους $a \in A$- δεδομένου ότι αυτή η συμπεριφορά είναι εγγενής σε πολλά μοντέλα VQA, είναι σημαντικό να επιτρέπονται πιο ελαστικοί ορισμοί της μετρικής ακρίβειας. Για μεγαλύτερη σαφήνεια, το ViLT παράγει εξόδους a με τη μορφή κειμένου φυσικής γλώσσας και υπάρχουν πολλοί διαφορετικοί τρόποι να εκφραστεί η ίδια απάντηση στα αγγλικά. Σε αυτή την περίπτωση, η ευριστική σύγκριση της παραγόμενης απάντησης με την απάντηση της θεμελιώδους αλήθειας από το A καθορίζει αν η πρόβλεψη του M είναι ακριβής ή όχι. Επαναλαμβάνοντας την ίδια διαδικασία πρόβλεψης για όλα τα ζεύγη Q, I, και λαμβάνοντας επιτυχείς ή ανεπιτυχείς απαντήσεις για αυτά, εξάγουμε τελικά ένα σκορ ακρίβειας acc_Q, το οποίο αντικατοπτρίζει την αναλογία των σωστών απαντήσεων επί όλων των παραγόμενων απαντήσεων.

Οι αντικαταστάσεις λέξεων καθοδηγούνται από εξωτερικές πηγές γνώσης, στοχεύοντας σε διαφορετικά μέρη του λόγου (ουσιαστικά, ρήματα και επίθετα) κάθε φορά. Συγκεκριμένα, ο γρά-

φος γνώσης WordNet [23] παρέχει ιεραρχικές σχέσεις μεταξύ μιας πληθώρας κοινών λέξεων που υπάρχουν ευρέως στα λεξιλόγια των VG και VQA-v2. Ως εκ τούτου, τα ζεύγη αντικατάστασης δημιουργούνται συνδέοντας συγκεκριμένες λέξεις με τις αντιστοιχίες τους στο WordNet, τηρώντας τις ιεραρχικές σχέσεις όπως περιγράφεται στην ενότητα **??**. Στη συνέχεια, προχωράμε στην εφαρμογή των διαταραχών που σχεδιάσαμε στις ερωτήσεις του συνόλου δεδομένων $q \in Q$, με αποτέλεσμα να προκύπτουν οι *αντιπαραδειγματικές ερωτήσεις* $q^* \in Q^*$.

1.4.2 Αξιολόγηση

Αρχικά, εκτελούμε το ViLT στις αρχικές εισόδους (*I, Q, A*) για να λάβουμε σκορ ακρίβειας βασικής αλήθειας *acc* ως βασική γραμμή. Για κάθε αντικατάσταση, λαμβάνουμε την *ακρίβεια αντιπαραδειγματικών ερωτήσεων* acc_Q^*, ως την απάντηση του M στις αντιπαραδειγματικές ερωτήσεις q^* Q^*, και τη συγκρίνουμε με τις βαθμολογίες ακρίβειας θεμελιώδους αλήθειας acc_Q. Καθ' όλη τη διάρκεια αυτής της διαδικασίας, αξιολογούμε αν και πώς αλλάζει η απόκριση του M, μετρώντας τη διαφορά μεταξύ acc_Q και acc_Q^*, ως δείκτη της ευρωστίας του έναντι αντικαταστάσεων με σημασιολογικά συναφείς έννοιες. Μετράμε ευριστικά αν ένα ζεύγος απαντήσεων είναι σωστό, με αποτέλεσμα η αντιληπτή ακρίβεια να αυξάνεται σημαντικά. Αυτό ωστόσο δεν αποτελεί ζήτημα για την ανάλυση των αποτελεσμάτων μας. Μας ενδιαφέρει πώς μεταβάλλεται η ακρίβεια με τις διαταραχές, όχι τα απόλυτα μέτρα ακρίβειας κάθε πειράματος.

Παρόλο που είναι χρήσιμα για λόγους συγκριτικής αξιολόγησης, τα μεμονωμένα αποτελέσματα ακρίβειας δεν είναι αρκετά κατατοπιστικά για να εξηγήσουν *γιατί* παρατηρούμε τέτοιες διαφορές μεταξύ των αρχικών q και των αντιπαραδειγματικών εισόδων q^*. Για το σκοπό αυτό, εξετάζουμε ξεχωριστά τα δείγματα στα οποία το παραγόμενο a αλλάζει υπό την παρουσία μιας διαταραχής, λαμβάνοντας τοπική εξήγηση της μορφής *αν η έννοια αλλάζει στο q, τότε το a -λανθασμένα- αλλάζει.* Η συνάθροιση τέτοιων τοπικών κανόνων οδηγεί σε *καθολικές επεξηγήσεις*, αντλώντας σχέσεις *αν-τότε* που εφαρμόζονται σε πολλαπλά δείγματα του D.

1.4.3 Διερευνητική Ανάλυση Δεδομένων

Μια καίρια πτυχή της εργασίας μας έγκειται στη βαθιά κατανόηση του αντίστοιχου συνόλου δεδομένων στο οποίο λειτουργεί ένα μοντέλο. Για τον λόγο αυτό, εφαρμόσαμε μια διεξοδική διερευνητική ανάλυση δεδομένων στα σύνολα δεδομένων Visual Genome και VQA-v2, προκειμένου να αποκτήσουμε γνώσεις που θα καθοδηγούσαν τον σχεδιασμό των αντικαταστάσεών μας. Παρακάτω, θα παρουσιαστούν ορισμένες ενδεικτικές περιπτώσεις οπτικοποίησης δεδομένων που απεικονίζουν χρήσιμα χαρακτηριστικά των δύο προαναφερθέντων συνόλων δεδομένων.

Μια προκαταρκτική αρχική παρατήρηση είναι η ομοιότητα των ευρημάτων μας για τα δύο σύνολα δεδομένων. Παρά τις μικρές διαφορές που αναπόφευκτα υπάρχουν λόγω της ιδιαιτερότητας του καθενός, μπορούμε να συμπεράνουμε με ασφάλεια ότι παρουσιάζουν παρόμοια χαρακτηριστικά όσον αφορά την κατανομή των ερωτήσεων, συνεπώς είναι λογικό να σχεδιάσουμε ένα σύστημα αντιπαραδειγμάτων που να αξιοποιεί ομοιόμορφα και ισότιμα και τα δύο σύνολα δεδομένων. Δεδομένου ότι τα συγκεκριμένα σύνολα δεδομένων που χρησιμοποιούνται στην παρούσα διπλωματική εργασία αποτελούν δύο από τα πιο διαδεδομένα διαθέσιμα σύνολα δεδομένων για το έργο Visual Question Answering, η παρατήρηση αυτή θα μπορούσε να υποδηλώνει γενικότερα ότι τα διαθέσιμα δεδομένα για το πρόβλημα αυτό παρουσιάζουν μια ανάλογη ομοιομορφία στη γλωσσική και εννοιολογική κατανομή των ερωτήσεων.

Τα σημαντικότερα αποτελέσματα που μπορούμε να αντλήσουμε από αυτή τη μελέτη είναι αυτά που αναλύονται παρακάτω:

- **Συχνότητες μεγέθους ερωτήσεων**: Όσον αφορά το πιο συχνά παρατηρούμενο μέγεθος των ερωτήσεων, παρατηρούμε ότι και τα δύο σύνολα δεδομένων χαρακτηρίζονται συνήθως από μικρές ερωτήσεις εισόδου. Πιο συγκεκριμένα, δεν ξεπερνούν κατά μέσο όρο το μήκος των

επτά λέξεων. Αυτό είναι ένα θετικό εύρημα για την εργασία μας: τα γλωσσικά μοντέλα δεν περικόπτουν συνήθως μικρές εισόδους, πράγμα που σημαίνει ότι δεν θα υπήρχε η εμφάνιση ανεπιθύμητων περικοπών που θα μπορούσαν πιθανότατα να βλάψουν την απόδοση του μοντέλου και να προκαλέσουν μεγάλη δυσκολία στην ολκλήρωση του στόχου μας.

- **Ποσοστό των κατηγοριών ανά μέρη του λόγου στις ερωτήσεις**: Η διερεύνηση της κατανομής των διαφόρων μερών του λόγου στις ερωτήσεις εισόδου παρέχει μια εικόνα για τον αντίκτυπο που θα έχουν τελικά οι παρεμβάσεις μας. Είναι προφανές ότι τα ουσιαστικά παρουσιάζουν τις υψηλότερες εμφανίσεις τόσο στα σύνολα δεδομένων Visual Genome όσο και στα σύνολα δεδομένων VQA-v2. Αυτό αποτελεί κυρίαρχη ένδειξη ότι οι αντικαταστάσεις των ουσιαστικών θα έχουν τη μεγαλύτερη επιρροή στα πειράματά μας. Εκμεταλλευόμενοι αυτή την εδραιωμένη διαίσθηση, προχωρούμε στη δόμηση των διαταραχών που θα αναλύσουμε αργότερα. Επιπροσθέτως, παρατηρούμε ότι τα ρήματα και τα επίθετα παρουσιάζουν παρόμοιο επίπεδο συχνότητας εμφάνισης, γεγονός που υποδηλώνει ότι οι αντικαταστάσεις ρήματος και επιθέτου θα είναι εξίσου διεισδυτικές και επιδραστικές για τα πειράματά μας.

- **Τύποι ερωτήσεων**: Αναλύοντας τους τύπους των ερωτήσεων που υπάρχουν στα δύο σύνολα δεδομένων που χρησιμοποιούμε, αποκτούμε μια πολύτιμη εικόνα του περιβάλλοντος του προβλήματός μας, με βάση την ακόλουθη έννοια: διερευνώντας τις ερωτήσεις, αποκτούμε μια εικόνα για το ποιες μπορεί να είναι οι απαντήσεις ως προς τον τύπο τους. Έτσι, οι κατανομές τύπων ερωτήσεων παρέχουν μια προσδοκία για τις αντίστοιχες απαντήσεις. Πρωτίστως, οι ερωτήσεις του τύπου "Τι" αποτελούν τη συντριπτική πλειοψηφία του συνόλου των ερωτήσεων. Αυτό δείχνει ότι μια αντικατάσταση του είδους *"Τι"* σε *"Πώς"* θα άλλαζε δραματικά το νόημα της ερώτησης και πιθανώς θα οδηγούσε σε άσκοπες ερωτήσεις που θα μπορούσαν να προκαλέσουν σύγχυση στο μοντέλο και να βλάψουν αδίκως τη μελέτη μας.

- **Συνηθέστερες λέξεις**: Κατ' αρχάς, σημειώνεται ότι έχει ολοκληρωθεί μια υποχρεωτική προεπεξεργασία αυτού του είδους των διερευνητικών αποτελεσμάτων, προκειμένου να αποκλειστούν ανούσιες κοινές λέξεις που δεν παρέχουν αξιόλογες πληροφορίες για το περιεχόμενο των ερωτήσεων (π.χ. "τι", "εικόνα", "η" κ.λπ.). Λόγω της εξαιρετικά υψηλής συχνότητάς τους, δεν μπορούν να αποδειχθούν χρήσιμες για τα πλαίσια της παρούσας ανάλυσης δεδομένων. Με βάση τα εξαγόμενα αποτελέσματα, παρατηρούμε ότι η λέξη "χρώμα" είναι μία από τις πιο συχνές λέξεις και στα δύο σύνολα δεδομένων. Αυτή η διαπίστωση ενέπνευσε την πρωτότυπη αντικατάσταση χρώματος που σχεδιάσαμε και θα παρουσιάσουμε περαιτέρω στη συνέχεια. Επιπλέον, η κυρίαρχη συχνότητα της λέξης "πολλά" υποδηλώνει την παρουσία πολλών αριθμητικών ερωτήσεων, άρα υποδεικνύει το υψηλό επίπεδο επιρροής τους. Τέλος, η εκτεταμένη παρουσία λέξεων όπως " άνδρας" ή "άνθρωποι" υποδηλώνει μια έντονη ανθρωποκεντρική προσέγγιση στο περιεχόμενο των ερωτήσεων, η οποία εφιστά επίσης την προσοχή σε συναφείς τύπους πειραματικών αντικαταστάσεων.

1.4.4 Αντικαταστάσεις

Στην εργασία μας, πραγματοποιούμε διάφορες αντικαταστάσεις ή διαγραφές στη γλωσσική αναπαράσταση του Q. Αυτή η αντιπαραδειγματική στρατηγική εκμεταλλεύεται πολλαπλά και ποικίλα μορφολογικά χαρακτηριστικά της Q και προσπαθεί να καταδείξει τη σημασιολογία που επηρεάζει περισσότερο την απάντηση a του μοντέλου. Στόχος μας είναι να διεγείρουμε μια τροποποιημένη απόκριση του M μέσω αυτών των αντιπαραδειγματικών διαταραχών, προκειμένου να εντοπίσουμε πώς αλλάζει η συμπεριφορά του M όταν έρχεται αντιμέτωπο με διαφορετικές έννοιες. Έτσι, μπορούμε να συμπεράνουμε πιθανές μεροληψίες ή σημεία ασθενούς ευρωστίας του M. Οι προαναφερθείσες αντικαταστάσεις μπορούν να χωριστούν στις ακόλουθες κατηγορίες, με βάση τη χρησιμοποιούμενη πηγή γνώσης και το στοχευόμενο μέρος του λόγου.

Α. Ιεραρχία του Wordnet: Οι αντικαταστάσεις λέξεων με βάση τη χρήση ιεραρχικών πηγών γνώσεων περιλαμβάνουν την αντικατάσταση ενός ουσιαστικού από τις ερωτήσεις $q \in Q$ με ένα

ιεραρχικά συγγενές ουσιαστικό (*υποώνυμο, υπερώνυμο, αδελφικό*), ή ρήματα και επίθετα με τα *συνώνυμα* τους. Η ποιότητα και η καταλληλότητα των αντικαταστάσεων μας διαβεβαιώνεται από τη χρήση της ντετερμινιστικής δομής της ιεραρχίας του Wordnet, που εγγυάται ελεγχόμενες και βέλτιστες αντικαταστάσεις για κάθε λέξη.

- **Συνώνυμα**: Χρησιμοποιούμε μετασχηματισμούς συνωνύμων στα επίθετα και τα ρήματα των αρχικών ερωτήσεων $q \in Q$. Για παράδειγμα, τα "talk" και "speak" είναι *συνώνυμα ρήματα* σύμφωνα με το WordNet, ενώ τα "small" και "minuscule" είναι *συνώνυμα επίθετα*. Η ιεραρχία του Wordnet παρέχει μια οργανωμένη κατά συνάφεια προτεραιότητα των συνωνύμων και την αξιοποιούμε επιλέγοντας το πιο σχετικό. Με αυτόν τον τρόπο, διασφαλίζονται η βελτιστότητα και η ελεγξιμότητα των αντικαταστάσεων.

- **Υπερώνυμα & Υπώνυμα ουσιαστικά**: Περισσότερο *γενικές*, καθώς και περισσότερο *ειδικές* έννοιες ουσιαστικών παρέχονται μέσω του WordNet με τη μορφή *Υπερωνύμων* και *Υπωνύμων* αντίστοιχα. Για παράδειγμα, από μια δεδομένη ουσιαστική λέξη (π.χ. "dog") μπορούμε να εξάγουμε τα άμεσα υπερώνυμα του ουσιαστικού (π.χ. "canine"), ή τα άμεσα υποώνυμα του (π.χ. "labrador").

- **Αδελφικά ουσιαστικά**: Κατασκευάζουμε τις αντικαταστάσεις *αδελφικού* ουσιαστικού διατρέχοντας το δέντρο γνώσης του Wordnet ένα βήμα προς τα πάνω και στη συνέχεια ένα βήμα προς τα κάτω. Ως αδέρφια ορίζονται οι οντότητες ουσιαστικών που μοιράζονται τον ίδιο άμεσο γονέα. Για παράδειγμα, το "carrot" και το "radish" είναι αδελφικά ουσιαστικά, επειδή και τα δύο έχουν ως γονική τους έννοια τη "plant root", σύμφωνα με το WordNet.

Β. Ιεραρχία συγγένειας χρωμάτων: Τα χρώματα που είναι σημασιολογικά παρεμφερή, και επομένως παρουσιάζουν τιμές RGB κοντά το ένα στο άλλο, είναι κοντά και στην ιεραρχία συγγένειας χρωμάτων. Για παράδειγμα, τα χρώματα "violet" και "orchid" της Matplotlib βρίσκονται κοντά στην ιεραρχία χρωμάτων (η ενδιάμεση χρωματική τους απόσταση είναι 6,16), ενώ τα χρώματα "violet" και "deepskyblue" τοποθετούνται πολύ μακριά το ένα από το άλλο (η χρωματική τους απόσταση είναι 207,88). Τα χρώματα μπορούν να αντικατασταθούν είτε με απομακρυσμένα είτε με παρόμοια χρώματα από αυτή την ιεραρχία συγγένειας χρωμάτων, οδηγώντας στις αντικαταστάσεις χρωμάτων **Μέγιστη Αντικατάσταση Χρώματος** και **Ελάχιστη Αντικατάσταση Χρώματος**. Και οι δύο αντικαταστάσεις Maximal/Minimal μπορούν είτε να περιλαμβάνουν *συνηθισμένα* χρώματα, τα οποία εμφανίζοντα ήδη στο σύνολο δεδομένων είτε *ασυνήθιστα* χρώματα, τα οποία εμφανίζονται στον κατάλογο χρωμάτων του Matplotlib αλλά όχι απαραίτητα στα λεξιλόγια των VG και VQA-v2:

- **Μέγιστη αντικατάσταση χρώματος**: Στις ερωτήσεις που αναφέρουν κάποιο συγκεκριμένο χρώμα αντιπαραβάλλουμε την έξοδο a της M με βάση την είσοδο της αρχικής ερώτησης $q \in Q$ έναντι της εξόδου a^* της αντικατεστημένης ερώτησης $q^* \in Q^*$. Στο q^* το αρχικό χρώμα αντικαθίσταται από ένα χρώμα που απέχει πολύ από αυτό, παραδείγματος χάρην "violet" $\rightarrow$ "deepskyblue". Σε αυτή την κατηγορία, δοκιμάζουμε επίσης το μοντέλο σε λιγότερο συχνές περιπτώσεις χρωμάτων (π.χ. "azure", "turquoise", "salmon"). Αυτή η αντικατάσταση αποκλίνει από το αρχικό αντιπαραδειγματικό ερώτημα που ζητά *ελάχιστες* αλλαγές- ωστόσο, η σύγκριση με σχετικές *ελάχιστες* αλλαγές θα αναδείξει τις διαφορές που επιβάλλουν οι ποικίλες αποστάσεις χρωμάτων στο τελικό acc^*_Q.

- **Ελάχιστη αντικατάσταση χρώματος**: Σε συμφωνία με τα παραπάνω, εκτελούμε αντικαταστάσεις χρωμάτων με χρώματα που απέχουν λίγο μεταξύ τους, π.χ. "violet" $\rightarrow$ "orchid". Και πάλι δοκιμάζουμε το μοντέλο με λιγότερο συχνές αντικαταστάσεις χρωμάτων.

C. Διαγραφές: Επιλέγουμε τυχαία ένα ουσιαστικό σε κάθε ερώτηση $q \in Q$ και το αφαιρούμε.

1.4.5 Προστιθέμενη Αξία των Αντικαταστάσεών μας

Οι αντικαταστάσεις και οι διαγραφές είναι ένας εξαιρετικός τρόπος για να ποσοτικοποιήσουμε αν ένα μοντέλο VQA M κατανοεί ένα συγκεκριμένο ζεύγος ερωτήσεων-εικόνων ή αν η έξοδός του εξαρτάται σε μεγάλο βαθμό από μεροληπτικές εκτιμήσεις. Με αυτόν τον τρόπο, μπορούμε να αποκαλύψουμε παραπλανητικές συσχετίσεις που λανθασμένα ενσωματώνονται στο M.

- **Color Substitutions**: Οι αντικαταστάσεις χρώματος παρακινούνται από την αξιοσημείωτη ποσότητα των ερωτήσεων που σχετίζονται με το χρώμα και υπάρχουν στα σύνολα δεδομένων εισόδου μας (VG και VQA-v2). Με το να ρωτάμε το M ερωτήσεις που περιλαμβάνουν χρωματικές αντικαταστάσεις που απέχουν πολύ από το αρχικό χρώμα (πείραμα **Μέγιστη αντικατάσταση χρώματος**), στοχεύουμε να ανιχνεύσουμε αν το M θα αντιληφθεί σωστά και λογικά αυτή τη σημασιολογικά τεράστια αλλαγή. Θα περιμέναμε ότι το M θα άλλαζε την απάντησή του a^* στις περισσότερες περιπτώσεις του πειράματος αυτού- το αντίθετο θα υποδείκνυε ένα υποβόσκον μοτίβο αγνόησης των χρωματικών χαρακτηριστικών. Ομοίως, εκτελούμε το πείραμα αντικατάστασης **Ελάχιστη αντικατάσταση χρώματος** προκειμένου να διερευνήσουμε τη συμπεριφορά του μοντέλου όταν αντιμετωπίζει μικρές αλλαγές στην έννοια του χρώματος. Αναμένουμε ότι οι αντικαταστάσεις από αυτό το πείραμα θα έχουν μικρή έως μηδαμινή επίδραση στην απόκριση a^* του μοντέλου. Μια αντίθετη συμπεριφορά θα αποκάλυπτε μια υφιστάμενη προκατάληψη όσον αφορά συγκεκριμένα χρώματα, γεγονός που θα οδηγούσε στο συμπέρασμα ότι το M δεν μπορεί να προσαρμοστεί σωστά και ισχυρά σε μικρές αλλαγές χρώματος και να γενικεύσει ανάλογα. Φυσικά, οι *ασυνήθιστες* αντικαταστάσεις χρωμάτων και στις δύο περιπτώσεις Minimal/Maximal θέτουν ένα πιο δύσκολο πρόβλημα, καθώς το M πρέπει να ανταποκριθεί προσαρμοζόμενο σε χρωματικές έννοιες εκτός συνόλου δεδομένων.

- **Συνώνυμες αντικαταστάσεις**: Σε σχέση με τις αντικαταστάσεις **Συνωνύμων**, στοχεύουμε να διερευνήσουμε την ικανότητα του μοντέλου να χειρίζεται αποτελεσματικά ήπιες μορφολογικές γλωσσικές αλλαγές που διατηρούν το ίδιο νόημα. Σε αυτή την περίπτωση, η αποτυχία να ανταποκριθεί σωστά (παρέχοντας μια διαφοροποιημένη απάντηση $a^* = a$) θα αποκάλυπτε υπερβολική προσκόλληση σε συγκεκριμένη σημασιολογία, γεγονός που καθιστά το μοντέλο λεξιλογικά άκαμπτο και συνεπώς μη ανθεκτικό σε σημασιολογικά αμελητέες διαταραχές.

- **Αντικαταστάσεις υπερωνύμων-υπωνύμων**: Οι διαταραχές **Υπερνωνύμων-Υπωνύμων** είναι προορισμένες να απεικονίζουν την ικανότητα του μοντέλου να γενικεύει και να προσδιορίζει αντίστοιχα, διατηρώντας ένα αξιόπιστο επίπεδο ευρωστίας. Οι σχέσεις υπερωνύμων και υπωνύμων είναι έννοιες βαθιά κατανοητές στον πραγματικό κόσμο και κατά συνέπεια ενσωματωμένες σε σύνολα δεδομένων μεγάλης κλίμακας, τα οποία χρησιμοποιούνται ευρέως για την προ-εκπαίδευση των μοντέλων VQA. Συνεπώς, το M θα πρέπει επίσης να είναι σε θέση να τις κατανοεί και να τις αιτιολογεί σωστά. Ιδανικά, θα περιμέναμε από το M να διατηρήσει την ίδια απόκριση για τις αντικαταστάσεις υπερωνύμων, ενώ δικαιολογημένα θα ανταποκρινόταν με συγκεκριμένους τρόπους για τις αντικαταστάσεις υπωνύμων, λαμβάνοντας υπόψη τον προσδιορισμό της σημασίας. Η ανάλογη ποσότητα ικανότητας εξειδίκευσης επιδιώκεται να διαπιστωθεί μέσω του πειράματος **αντικατάστασης αδελφικών ουσιαστικών**. Ανάλογα με την κάθε συγκεκριμένη περίπτωση, αναμένουμε ότι το M θα τροποποιήσει ή θα διατηρήσει κατάλληλα την απόκρισή του, ώστε να επιβεβαιώσει το επίπεδο κατανόησης και διάκρισης διαφορετικών, αλλά παρόλα αυτά συναφών, σημασιών.

- **Διαγραφές**: Τέλος, υλοποιήσαμε το πείραμα **Διαγραφές** αναμένοντας ιδανικά την υποβάθμιση της απόδοσης του M. Το μέγεθος της πτώσης του acc_Q^* εξαρτάται από τη σημασία του διαγραμμένου ουσιαστικού για το νόημα της ερώτησης. Κατά συνέπεια, ένα αμερόληπτο M θα πρέπει να είναι σε θέση να προσδιορίσει αυτή τη σημασία και να ενεργήσει ανάλογα,

χωρίς να καταλήξει σε αδικαιολόγητα συμπεράσματα που εκφράζονται μέσω μιας αδικαιολόγητης απάντησης.

1.5 Πειράματα και Αποτελέσματα

Παρουσιάζουμε τα αποτελέσματα χρησιμοποιώντας τη μετρική ακρίβειας, η οποία απεικονίζει το βαθμό ομοιότητας της προβλεπόμενης απάντησης του μοντέλου με την απάντηση της θεμελιώδους αλήθειας, τόσο για το αρχικό Q κάθε συνόλου δεδομένων, όσο και για το αντιπαραδειγματικό σύνολο ερωτήσεων Q^*. Στην ανάλυσή μας, η ακρίβεια ως μετρική δεν διαθέτει το απαιτούμενο βάθος ώστε να παρέχει συγκεκριμένες εξηγήσεις σε ποικίλες καταστάσεις και πληροφορίες σχετικά με τη συμπεριφορά του μοντέλου, όταν έρχεται αντιμέτωπο με συγκεκριμένες έννοιες. Ωστόσο, η ακρίβεια εξακολουθεί να επιδεικνύει μια υψηλού επιπέδου προσέγγιση σχετικά με τις διακυμάνσεις της αποδοτικότητας του μοντέλου υπό τις υλοποιημένες αντιπαραδειγματικές διαταραχές. Δεδομένου ότι το μοντέλο ViLT έχει εκπαιδευτεί και βελτιστοποιηθεί στο σύνολο δεδομένων VQA-v2, αναμένεται κατά κάποιο τρόπο να έχει καλύτερες επιδόσεις σε αυτό σε σύγκριση με το VG (και τα δύο σύνολα δεδομένων περιέχουν παρόμοια λεξιλόγια). Αυτή η παρατήρηση επιβεβαιώνεται πράγματι από τα αποτελέσματά μας που παρουσιάζονται στους πίνακες 6.1 & 6.2, τα οποία καταδεικνύουν σταθερά υψηλότερο acc_Q στο πρώτο έναντι του δεύτερου συνόλου δεδομένων, όσον αφορά όλα τα πειράματα που υλοποιήθηκαν. Σημειώνουμε ότι οι βαθμολογίες acc_Q για κάθε πείραμα περιέχουν μόνο τις αντίστοιχες ερωτήσεις, π.χ. τα πειράματα για το χρώμα περιέχουν μόνο ερωτήσεις που αναφέρουν χρώματα. Αυτό συμβάλλει στις διαφορές στα αρχικά σκορ acc_Q για κάθε πείραμα.

Παρ' όλα αυτά, και στα δύο σύνολα δεδομένων, παρατηρούμε μια ανάλογη διαφορά μεταξύ acc_Q και acc^*_Q ανά πείραμα, όταν το M υποβάλλεται σε αντιπαραδειγματικές ερωτήσεις $q^* \in Q^*$. Αυτό θα μπορούσε γενικά να υποδηλώνει την ύπαρξη υποβόσκουσας μεροληψίας: η παρουσιάζει ένα είδος υπερπροσαρμογής στην αρχική $q \in Q$, γεγονός που την καθιστά λιγότερο αποτελεσματική όταν καλείται να χειριστεί ελάχιστα διαταραγμένες αντιπαραδειγματικές ερωτήσεις. Σε όλα τα πειράματα, η μείωση της ακρίβειας από το αρχικό acc_Q στο acc^*_Q είναι περίπου 10-35% ή περισσότερο.

Η ανεύρεση καθολικών μοτίβων παρέχει μια πιο βαθιά και στοχευμένη θεώρηση της ευρωστίας του μοντέλου M. Για το σκοπό αυτό, σημειώνουμε ότι σε όλες τις περιπτώσεις που μελετήσαμε, δεν μας ενδιαφέρει η απάντηση της θεμελιώδους αλήθειας μιας ερώτησης q, αλλά μάλλον η διαφοροποίηση της απάντησης a^* στην αντιπαραδειγματική ερώτηση σε σχέση με την αρχική απάντηση a που προβλέπει το M, είτε αν η a είναι σωστή είτε όχι. Επιλέγουμε αυτή την προσέγγιση, καθώς μας ενδιαφέρει να ανακαλύψουμε την μεταβολή του μοντέλου στη λήψη αποφάσεων υπό την παρουσία αντιπαραδειγματικών εισόδων, η οποία είναι πιο κατατοπιστική από τη μέτρηση του κατά πόσο η a^* αποκλίνει σημασιολογικά από την απάντηση της θεμελιώδους αλήθειας.

Με βάση την ενδελεχή διερεύνηση των αποτελεσμάτων των πειραμάτων μας και τη συγκέντρωση των *τοπικών εξηγήσεων*, όπως παρουσιάζονται στο παρόν κεφάλαιο αυτής της διπλωματικής εργασίας, έχουμε συμπεράνει κάποιους ουσιαστικούς *καθολικούς κανόνες* που αφενός αποτυπώνουν την ευρωστία του M στις αντιφατικές ερωτήσεις μας $q^* \in Q^*$, αφετέρου παρέχουν αξιόπιστες επεξηγήσεις που αποκαλύπτουν τη συλλογιστική πορεία του μοντέλου πίσω από τη λήψη των αποφάσεών του. Επιπλέον, αναλύουμε τις υποβόσκουσες υφιστάμενες προκαταλήψεις του M που προκύπτουν λογικά από αυτούς τους καθολικούς κανόνες.

Για μια λεπτομερέστερη επεξήγηση της μεθοδολογίας εξαγωγής των αποτελεσμάτων που παρουσιάζουμε παρακάτω, παραθέτουμε τα σχετικά Wordclouds, συνοδευόμενα από τις αντίστοιχες αριθμητικές τιμές που εκφράζουν τις κανονικοποιημένες συχνότητες εμφάνισης των μεταβάσεων που προκύπτουν από τις αντικαταστάσεις που εφαρμόζουμε.

1.6 Συμπεράσματα και Μελλοντικές Προεκτάσεις

Οι αντιπαραδειγματικές διαταραχές στα μοντέλα VQA μπορούν να παρέχουν νέες και χρήσιμες πληροφορίες σχετικά με την ευρωστία του μοντέλου και την ερμηνευσιμότητα των αποτελεσμάτων. Στην εργασία μας, προτείνουμε ένα πλαίσιο αντιπαραδειγματικής προσέγγισης βασισμένο στη γνώση, το οποίο στοχεύει σε αντικαταστάσεις σε ερωτήσεις. Συγκεκριμένα, το πλαίσιό μας προτείνει πολλαπλούς τύπους γλωσσικών μετασχηματισμών σε επίπεδο λέξεων, προκειμένου να εξετάσουμε επιλεγμένα μοντέλα VQA με τρόπο που αποδέχεται την φύση "μαύρου κουτιού" που τα χαρακτηρίζει, και να διερευνήσουμε κατά πόσον η παρουσία αντιπαραδειγματικών ερωτήσεων θα οδηγήσει σε απροσδόκητες αποκρίσεις του μοντέλου. Μέσω αυτής της διαδικασίας, αποκαλύπτονται οι υποβόσκουσες γλωσσικές προκαταλήψεις, ενώ παρέχονται κατατοπιστικές εξηγήσεις σχετικά με τη συμπεριφορά του μοντέλου, με την εξαγωγή καθολικών κανόνων με ποιοτικό τρόπο, που τελικά απεικονίζουν αυτές τις υπάρχουσες προκαταλήψεις. Τα αποτελέσματά μας στα σύνολα δεδομένων Visual Genome και VQA-v2, χρησιμοποιώντας το μοντέλο ViLT ως τεκμήριο εφαρμογής της ιδέας, καταδεικνύουν τα πλεονεκτήματα της προσέγγισής μας, αναδεικνύοντας τις έννοιες που υποκινούν τις προκαταλήψεις του μοντέλου, με τρόπο που δεν σχετίζεται με την επιλογή του εκάστοτε μοντέλου.

Σχεδιάζουμε το πλαίσιο εργασίας μας αναπτύσσοντας ένα πλήρως γενικεύσιμο σύνολο αντιπαραδειγματικών αντιθετικών μετασχηματισμών, πράγμα που σημαίνει ότι μπορεί να εφαρμοστεί σωστά σε οποιοδήποτε μοντέλο VQA. Αυτή η ιδιότητα της προτεινόμενης μεθόδου μας διαβεβαιώνεται μέσω της προσέγγισης με προσανατολισμό στο μαύρο κουτί που ακολουθούμε όσον αφορά τη διαταραχή του εν λόγω μοντέλου: δεν επιχειρούμε να διεισδύσουμε στην αρχιτεκτονική, τις σχεδιαστικές επιλογές ή τις παραμέτρους κάθε μοντέλου VQA, ούτε επιστρατεύουμε διαδικασίες που διερευνούν την εσωτερική διαδικασία συλλογισμού των μοντέλων μηχανικής μάθησης. Αντ' αυτού, εξετάζουμε κάθε μοντέλο εφαρμόζοντας διάφορους τύπους αντιπαραδειγματικών διαταραχών και παρατηρώντας εκτενώς τα παραγόμενα αποτελέσματα. Ως εκ τούτου, οι γνώσεις αντλούνται μέσω της ανάλυσης της απόκρισης του μοντέλου και όχι μέσω της εμβάθυνσης στην εσωτερική δομή του. Επιπλέον, οι προτεινόμενες από εμάς αντιπαραδειγματικές διαταραχές καθοδηγούνται πλήρως από τη χρήση πηγών βασισμένων στη γνώση, οι οποίες ελέγχουν και διαμορφώνουν τα πολλαπλά πειράματα που έχουμε εφαρμόσει. Εγγυώνται επίσης τους βέλτιστους μετασχηματισμούς που προτείνουμε, καθώς και διαβεβαιώνουν την ντετερμινιστική διάσταση των επιλογών αντικατάστασης που εισάγουμε.

Η κατάλληλη σύγκριση των αρχικών και αντιπαραδειγματικών αποτελεσμάτων, καθώς και η διορατικότητα που παρέχεται από τη διακύμανση της ακρίβειας των πειραμάτων, αναδεικνύουν την έκταση της ευρωστίας του μοντέλου. Ειδικότερα, η μελέτη αυτή έγκειται στην αξιολόγηση του κατά πόσον το μοντέλο παρουσιάζει την κατάλληλη ευελιξία ώστε να προσαρμόζεται ποιοτικά σε μικρές ή μεγάλες αλλαγές στην είσοδο και, κατά συνέπεια, να σημειώνει πανομοιότυπη ή παρόμοια ακρίβεια με αυτή που καταγράφηκε κατά την απόδοση στο αρχικό σύνολο δεδομένων.

Το προαναφερθέν ερευνητικό πεδίο που σχετίζεται με την ευρωστία και το οποίο προσεγγίζουμε, σχετίζεται στενά με τη διερεύνηση και την ανακάλυψη των κεκαλυμμενων υφιστάμενων προκαταλήψεων στο εν λόγω μοντέλο. Η προσεκτική παρατήρηση του παραγόμενου αποτελέσματος σε αντιπαραδειγματικά ερωτήματα σε αντιδιαστολή με τα αρχικά αποτελέσματα στο αρχικό σύνολο δεδομένων οδηγεί σε χρήσιμες τοπικές εξηγήσεις που απεικονίζουν συγκεκριμένες περιπτώσεις απροσδόκητης απόκρισης του μοντέλου που επισημαίνουμε. Στη συνέχεια, μια συνολική συνάθροιση αυτών των τοπικών παρατηρήσεων οδηγεί στην εξαγωγή καθολικών κανόνων που περιγράφουν και χαρακτηρίζουν πλήρως τη γενική συμπεριφορά του μοντέλου. Αυτές οι καθολικές επεξηγήσεις εκφράζουν τις ανιχνευμένες προκαταλήψεις που προέκυψαν από την εφαρμογή των πολυεπίπεδων διαταραχών μας.

Η πεμπτουσία της μεθόδου που παρουσιάζουμε έγκειται στη διαμόρφωση διεισδυτικών και στοχευμένων επεξηγήσεων που δικαιολογούν την απόκριση του μοντέλου και διαφωτίζουν τη συλλογιστική διαδικασία που ακολουθείται από το συνδυασμό των επιμέρους δομικών μονάδων του

μοντέλου. Μέσω αυτού του σημαντικού πλεονεκτήματος, η μέθοδός μας αποδεικνύεται ότι αυξάνει τη διαφάνεια και βελτιώνει το επίπεδο αξιοπιστίας του εν λόγω μοντέλου, καθώς παρέχει πληροφορίες που μπορούν να ληφθούν σοβαρά υπόψη κατά την αξιολόγηση της εμπιστοσύνης που πρέπει να αποδώσει κανείς στις αποφάσεις που εξάγονται από αυτό. Αυτή η νέα οπτική γωνία που προσφέρουμε αντιμετωπίζει με επιτυχία το αρχικά εισαχθέν προβληματικό ζήτημα που αποτέλεσε πηγαίως το κίνητρο για την έρευνά μας, δηλαδή την ανεπιθύμητη αδιαφάνεια που αναπόφευκτα συνοδεύει τη συγκεκαλυμμένη διαδικασία λήψης αποφάσεων που ακολουθείται από την εσωτερική συλλογιστική πολλών μοντέλων μηχανικής μάθησης, που εκτελούν οπτικογλωσσικές εργασίες, και συγκεκριμένα την απάντηση οπτικών ερωτήσεων, όπως διερευνάται στην εργασία μας.

Μετά την εκτενή παρουσίαση της παρούσας διπλωματικής εργασίας, θα θέλαμε να προτείνουμε κάποιες υποσχόμενες μελλοντικές ερευνητικές κατευθύνσεις σχετικά με τα εξεταζόμενα θέματα, που θεωρούμε ότι είναι γόνιμες για περαιτέρω επεξεργασία. Κατ' αρχάς, ως άμεση επέκταση της μεθόδου μας, προτείνουμε την εφαρμογή των ίδιων γλωσσικών διαταραχών στις απαντήσεις του συνόλου δεδομένων, απευθυνόμενοι σε μοντέλα VQA που συλλογίζονται πάνω σε απαντήσεις πολλαπλών επιλογών. Επιπλέον, προτείνουμε την επέκταση της προσέγγισής μας σε άλλες συναφείς οπτικογλωσσικές εργασίες, όπως η Ανάκτηση Κειμένου - Εικόνας, η Οπτική Επικαλυπτικότητα και η Οπτική Συλλογιστική Κοινής Λογικής, καθώς πιστεύουμε ότι έχει τη δυνατότητα να προσφέρει πολύτιμες γνώσεις και να αποδώσει μια τόσο αποτελεσματική όσο και καινοτόμο πτυχή επεξηγησιμότητας σε τέτοιες δημοφιλείς σύγχρονες εργασίες μηχανικής μάθησης με μεγάλο εύρος επιρροής. Τέλος, μια άλλη άξια αναφοράς προτεινόμενη κατεύθυνση περιλαμβάνει τη διαμόρφωση αντιπαραδειγματικών διαταραχών που στοχεύουν στην οπτική τροπικότητα, οι οποίες απευθύνονται τόσο σε επίπεδο εικονοστοιχείων όσο και σε εννοιολογικές αντικαταστάσεις. Ένα παράδειγμα τέτοιων πολλά υποσχόμενων επεκτάσεων της ανάλυσής μας θα μπορούσε να περιλαμβάνει την απόπειρα οπτικής διαταραχής μέσω της σύνθεσης εικόνων με τη χρήση μοντέλων διάχυσης και το prompting μέσω μεγάλων γλωσσικών μοντέλων. Τέτοιου είδους μετασχηματισμοί μπορούν να καταστούν χρήσιμοι στην αποκάλυψη οπτικών προκαταλήψεων και έτσι να διερευνηθεί μια παράλληλη πλευρά του πεδίου της επεξηγησιμότητας σε οπτικογλωσσικές εργασίες.

Chapter 2

Introduction

The rapid development of technology over the last decades has played a major role in the transformation of human activity in various areas such as society, work, and decision-making in general. In the field of computer science, a major reflection of this significant evolution of knowledge and know-how can be summarised in the progress of artificial intelligence. Probably the most remarkable benefit that artificial intelligence and the results of machine learning models offer us is the quality, ease, and speed of producing useful results that reshape our ways of making decisions and determining action. We see these results and benefits evident in very multiple areas of action and progress, such as health, justice, economics, politics, science, education, and even art and creativity.

The widespread use of Machine Learning Systems in decision making as well as their decisive influence on decision and subsequently on action undoubtedly render it of the highest priority to ensure methods that promote transparency, fairness, and trustworthiness. In this context, the rapidly growing field of Explainable Artificial Intelligence offers the means to increase trust in machine learning systems. The qualitative interpretation of the generated results of these models provides an insightful look at the reasoning process that has been followed to produce them.

In parallel with the need to broaden the explainability of machine learning models, there is also a need to investigate their robustness, i.e. the extent to which they can be resilient, flexible, and adaptable in a qualitative manner to significant or minor modifications of their input. This study highlights possible biases that may be embedded in the models which certainly affect both their fairness and their reliability. Thus, the behavior and response of the models are rendered biased, indicating the urgent need for the development of systematic methods that can qualitatively identify and highlight such problems.

Multimodal learning is an area of artificial intelligence that combines multiple input modalities. Models that perform multimodal learning tasks have gained popularity in recent years thanks to their versatility to handle multiple input formats and consequently more complex problems that require multifactor solution approaches. In this context, this diploma thesis addresses the Visual Question Answering (VQA) task, which combines image and natural language as input. In particular, Visual Question Answering models accept open-ended questions involving images and are asked to return a qualitative answer on the combination of these questions and images.

In this thesis, we develop a systematic method to evaluate the stability of such models and investigate the existence of possible biases in the reasoning process they follow as well as an explanatory approach to studying their results. Our method is based on the use of optimal counterfactual questions in linguistic modality. It highlights a variety of ways in which the response of VQA models can be evaluated in a way that is fully generalizable and extensible to any model performing the VQA task, while also recognizing the non-transparent nature of such models and handling them as black boxes. The control, quality, and design of our proposed counterfactual perturbations are guaranteed by the use of hierarchical knowledge sources, that fully guide our experiments.

2.1 Motivation

The indisputable rise in popularity of visiolinguistic (VL) learning [48, 19, 44] has offered a variety of impressive model implementations to the community in a short time [39, 36, 41, 35, 59, 38]. Visual Question Answering (VQA) is a VL task that has obtained a fundamental role in the evolution of various interactive VL AI systems, such as Visual Dialogue [22], Text-Image Retrieval [21] and Visual Commonsense Reasoning [67]. To this end, there is an extensive range of real-world applications that benefit significantly from the new advances around the VQA task, such as aiding systems for visually impaired individuals [8, 13] and self-driving cars [10].

VQA involves a textual question q from a pre-defined question set Q accompanied by an image I, the interaction of which yields a textual answer a. The race for continuously advancing VQA model performance unavoidably results in leaving open issues, especially attributed to the black-box nature of state-of-the-art implementations [12, 2, 15, 7, 26]. This limited access to the reasoning that such models follow to make decisions emphasizes the risk of an arbitrary behavior on their behalf. This peril lies mainly in the possibility of bias integration, decisions that lack the proper focus, as well as the absence of explainability and fairness of results. Especially when pivotal decisions are made based on systems of such type, their opacity renders them impractical, and at times hazardous, for most applications. This uncertainty indicates the need for new robustness evaluation methods, that prioritize the transparency of VQA models.

Different approaches to debiasing and explainability of VQA models focus on diverse aspects of the issue. For example, [11] examines VQA robustness and explainability by addressing transformations on the visual modality, as they attribute the problem mostly to the visual bias as occurring from unwanted correlations between image concepts. In general, existing works primarily focus on the effect of *visual bias* rather than the impact of *linguistic bias*, as a reason behind the lack of robustness in VQA models. Other works follow attention-based strategies that require extensive knowledge on the model architecture, thus they cannot handle efficiently the *black-box* nature of these systems. To this end, various model-specific approaches are proven to be fruitful in their strict framework but lack the capability to be generalized to the evaluation of any other model, thus limiting their efficiency scope to just one specific case.

In this work, we are motivated by the identification of the following alarming issue: The black-box nature of state-of-the-art visiolinguistic (VL) models blocks transparency and poses risks of spurious correlations, biases, and opaque decision-making. The importance, weight, and scope of the decisions made by these models highlight the need for extensive engagement with the issue, and this guides and motivates our work. We focus on the VQA models and emphasize the necessity of understanding the reasoning they may follow to produce the results they produce, even though we are aware of their non-transparent nature. We stress that before adopting their results, we should understand the behavior of VQA models. In this context, explainable models promote fairness, accuracy, and trust in AI-powered decision-making. Our method lies in producing linguistic perturbations based on hierarchical knowledge sources. We extract local explanations that we later generalize into global rules of biases and global explanations. Through this process, we develop a model-agnostic method that provides explanations for black-box VQA models through optimal and controllable counterfactual perturbations.

2.2 Organization

Before we start to study, analyze and discuss our work at length we should give a brief but illustrative and useful overview of VQA as a problem and task as well as refer to the approach of counterfactual explanations for this specific visiolinguistic task. In this context, in chapter 3 we will describe in depth the Visual Question Answering task to the reader and briefly explain the methodology followed by the counterfactual explanations method we propose in this thesis, as well as the idea that

gives rise to it and the motivation that makes it helpful for the problem we wish to address. For the most part, we will be presenting various diverse VQA models, based on the different architectures, scope, and design choices, as well as vastly used and popular datasets that are fine-tuned to the Visual Question Answering task and have greatly contributed to the progress of this field thanks to their assistance in training and evaluation. In addition, we will refer to the evaluation metrics and approaches that benefit the VQA task, discuss the essential role of language bias and its impact on performance and efficiency and illustratively explain the added value of the counterfactual explanations method for VQA, which we have deployed in this work.

A constructive literature review will assist the reader in gaining an in-depth understanding of various concepts closely related to the topic addressed in this thesis. As such, in chapter 4, we will be presenting a thorough yet short bibliographical report on the different fields relevant to visiolinguistic learning and explainability, that are of importance to us regarding this work. We will extensively refer to the pre-trained models that combine visual and linguistic input modalities, by analyzing them in the scope of various categories that characterize their pipeline. Moreover, we will provide a descriptive background on the field of knowledge in visiolinguistic learning. In addition, we will address the field of explainability in the Visual Question Answering task by listing a variety of directions that have been followed in relevant works, as well as their qualitative differences. We will highlight examples of the use of explainability grounded in the application of counterfactual transformations, thus highlighting the novelty of our proposed method. Furthermore, we will refer to the area of linguistic perturbations by briefly discussing previous works based on such methods and noting the various relevant sources of inspiration for the development and elaboration of our method.

In chapter 5 we will take a deeper look into the Counterfactual Visual Question Answering method that we introduce. We will extensively present the structure of our approach, refer to the proof-of-concept model and datasets we apply our generalizable method, and discuss useful notation, processing steps, and the procedure we follow to produce our ultimate counterfactual explanations. Initially, we will present an insightful exploratory data analysis which we have implemented on the datasets as a preliminary step to guide our design approach. We will make a thorough presentation of the various types of linguistic counterfactual perturbations we have designed and implemented, while also analyzing the use of knowledge sources in our counterfactual VQA framework. Moreover, we will explain and discuss the various levels and areas in which the substitutions we implement provide useful insights that we will subsequently exploit to obtain a qualitative interpretation and evaluation of the results produced by the models we investigate, thus optimally approaching the issue of explainability, of bias detection and robustness evaluation.

Chapter 6 will constitute our contributions to the problem. We will illustratively present our results based on the iChapterted experiments. By presenting the evaluation metrics and explaining how they can be rendered insightful for our task, we will furthermore delve into a deeper analysis of our outputs. This will be addressed through the presentation of observed biases that are discovered in the model's in-question response to our counterfactual questions, categorized by the various implemented experiments. We will elaborate on how these biases indicate a general specific model behavior, that can be characterized as non-trustworthy and partial at some levels. The essential benefit of our method will be showcased in this section, as we will highlight the detected biases formed as global rules that depict the model's behavior and ultimately guide and prescript the corrections and improvements that need to be applied to enhance the reliability and impartiality of the model under investigation.

Finally, in chapter 7 we will present a conclusion of our work, summarizing our method and contribution, while also proposing guidelines and ideas for future directions related to promising extension prospects concerning our field.

2.3 Contributions

We argue that resolving explainability challenges in VQA models calls for a counterfactual approach, implemented as word-level perturbations on the questions q; thus, we diverge from the well-sought exploration of the visual modality, examining the role of the language on possible biases and spurious correlations hidden in VQA models, while tracing and interpreting their opaque decision-making process. Our proposed counterfactual perturbations are framed as: *"What is the response of the VQA model if we substitute word X with word Y in question q?"* Specifically, by viewing words as concepts, we perform the minimum possible feasible transformation to stimulate a change in the model's response; then, insightful comparisons are made by recording the model's behavior to various such transformations. The counterfactual perturbations that we perform are fully guided by the deterministic assurance of *hierarchical knowledge* structures. By deploying these knowledge sources, we provide transformations that are not only optimally targeted to each specific linguistic concept but are also fully *explainable* in terms of the strategy followed for their implementation.

Starting with the observation of *local model responses* in different linguistic perturbations for a single data sample, we further identify *global patterns* that refer to the overall behavior of the model when faced with a specific set of perturbed concepts. Following this, we propose global rules that characterize the response of a model and can underline its weaknesses, by indicating what concepts could harm its robustness and certainty. This process reveals possible biases that a model has integrated, and hence attributes explanations as to why a particular answer is generated in place of another one; thus, we can obtain insights into the reasoning process of a model, without the need for access to the model's inner architecture.

Overall, spurious correlations, biases, and opaque decision-making are issues approached by our work: we aim to probe the response of VQA models when a *counterfactual textual input* is inserted into the model in place of the ground truth one. For this reason, we implement a vast variety of morphological linguistic transformations on the questions, to extract insightful observations on the model's response. Our method is generalizable to any VQA model and corresponding suitable dataset, as it approaches the issue in a totally model-agnostic strategy. Specifically, we utilize counterfactual explanations to frame robustness evaluation as the answer to the following question: What concepts need to change in a question, for it to cause a change in the model's answer? We, therefore, contribute to the following:

1. We design counterfactual inputs applying a variety of structured *word-level replacements* on the questions $q \in Q$, as instructed by hierarchical knowledge sources. Our approach is model-agnostic, as we treat any VQA model as a black box.

2. We obtain *local explanations* derived from unexpected model responses to counterfactual q inputs.

3. By summarizing local model behaviors for all $q \in Q$, we extract some *global explanations* that reveal the overall model response to each of the designed counterfactual inputs.

Part of this thesis's work has been accepted as a scientific paper in AAAI-MAKE 2023 (AAAI Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering) and has been uploaded to the public archive arxiv.org [60].

Chapter 3

Visual Question Answering

Visual Question Answering (VQA), first introduced in [5], lies in the category of multimodal learning tasks, as it receives both visual and linguistic modalities as input. Specifically, a VQA model M receives images i from a set I and relevant questions q belonging to a predefined question set Q, and is expected to accurately answer those q by providing a natural language answer a. The aforementioned answers can be either open-ended (generated by M) or belong to a set of pre-defined candidates A. The VQA task is goal-driven, meaning that its development is motivated by the targeting to solve specific problems and improve wide-ranging issues related to cooperation and coexistence of image and natural language inputs.

According to each specific case, visual questions q selectively target different areas of an image I, including background details and underlying context. In accordance, the focus regarding the linguistic input lies in different word concepts depending on each image-question pair. An example of the VQA task, including an image I and related questions q, as well as the answers a that a VQA model M returns to these questions, is demonstrated in Fig. 3.1.

Figure 3.1: Example of image and free-form questions retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

In general, questions q have an arbitrary nature and they enclose different computer vision subproblems, such as fine-grained recognition (Fig. 3.2), object recognition (Fig. 3.3) and detection (Fig. 3.4), activity recognition (Fig. 3.5), knowledge-based reasoning (Fig. 3.6), attribute classification (Fig. 3.7), scene classification (Fig. 3.8), as well as counting (Fig. 3.9) [32]. Furthermore, more intricate questions concern more complex processes, such as spatial relationships among objects (Fig. 3.10) and commonsense reasoning (Fig. 3.11) [23, vqa]. Some illustrative examples of the above categories

of questions are shown above. In general, a VQA model that demonstrates high levels of robustness, efficiency, and agility is flexible enough to meet the requirements to resolve a vast range of classical computer vision tasks alongside properly reasoning over the combination of images and relevant, often elaborate or difficult, questions [5], [33].

Answering questions, even those with binary answers, is a complex task that requires a rigorous process. When it comes to image-related queries, straightforward responses consisting of just a few words can often suffice. The effectiveness of an algorithm in such cases can be measured by the number of correct answers it produces. While open-ended questions require a free-form response, multiple-choice questions require algorithms to select from a predetermined set of potential answers.

3.1 Applications

VQA has diverse potential applications, particularly as an aid for blind and visually impaired individuals to access information about images both online and in the real world. It can also enhance human-computer interaction as a natural means to query visual content and enable image retrieval without metadata or tags. Additionally, VQA presents an important research challenge, as a successful system must be capable of solving various computer vision problems, making it an integral part of a Turing Test for image comprehension [33].

3.2 Relevant Tasks

The primary objective of VQA is to extract image semantics relevant to a given question, encompassing both minute details and abstract scene attributes. While computer vision tasks such as object recognition, activity recognition, and scene classification focus on extracting information from images, they are relatively narrow in scope compared to VQA. Object recognition algorithms are trained to classify images into specific semantic categories but only identify the dominant object in the image and not it's spatial positioning or role in the scene. Object detection is a more comprehensive version of object recognition, localizing specific semantic concepts by placing bounding boxes around each instance of an object in an image. Semantic segmentation goes one step further by classifying each pixel as belonging to a particular semantic class, while instance segmentation distinguishes between separate instances of the same class. However, these methods alone are insufficient for complete scene understanding, as they do not account for the role of an object in a larger context or handle label ambiguity. In contrast, VQA requires a system to answer arbitrary image-related questions that may necessitate object relationship reasoning, with the appropriate label specified by the query [33].

In addition to VQA, there are other approaches that combine vision and language. One of the most researched methods is image captioning, where an algorithm aims to generate a natural language description of an image. Unlike VQA, image captioning allows for arbitrary granularity in image analysis and can potentially describe complex attributes and object relationships in detail. However, VQA has the advantage of having specific and unambiguous answers for many question types, making it more easily evaluated compared to captioning. Although some questions in VQA may still be equivocal, the answer can often be evaluated by one-to-one matching with the ground truth answer [33].

3.3 VQA Models and Datasets

Since the introductory work on VQA [5], several endeavors have extended this paradigm, either by suggesting advanced model architectures or by proposing more challenging datasets. State-of-the-

art models addressing the VQA task are mainly based on VL transformer backbones; thus, models such as ViLBERT [41], VisualBERT [39], FLAVA [59], ALBEF [38], ViLT [36] and others have dominated the recent VQA literature demonstrating rapid improvements on relevant benchmark datasets.

A VQA dataset should be large and diverse enough to account for real-world variability. This is because real-world situations indeed are characterized by a vast variety of different concepts, thus showcasing the need for an abundance of questions and answers, both in terms of absolute number and conceptual diversification. An efficient VQA dataset should also possess an evaluation scheme that cannot be easily manipulated, ensuring that algorithms can accurately answer questions with definitive answers. Hence, both accuracy and fairness are promoted and guaranteed, as it would mean that a VQA model would properly reason over the dual input modality, instead of integrating unwanted correlations that eventually lead to biased model behavior. If a dataset contains biases in question or answer distribution, an algorithm may perform well without truly solving the VQA problem, leading to undesirable issues of incapability for generalization, due to overfitting [33].

Various dominant VQA datasets have been proposed, including Visual Genome, the VQA Dataset [5], DAQUAR [47], COCO-QA [54], FM-IQA [24] and more. The COCO dataset has been widely used in VQA research, as it provides a diverse range of images with multiple objects and a variety of scene contexts. The use of Microsoft Common Objects in Context (COCO) images in VQA datasets allows for the evaluation of algorithms on real-world images with complex visual content. Additionally, having multiple captions for each image in COCO allows for the generation of multiple questions for each image in VQA datasets, which helps to increase the diversity of questions that can be asked about an image. Thus, all of the aforementioned datasets, excluding only DAQUAR, include images retrieved from the COCO dataset. Specifically, Visual Genome additionally uses images retrieved from Flickr100M [33]. Visual Genome (VG) is a large-scale dataset including numerous scene images, object, attribute, and relationship annotations, as well as visual question-answer pairs [37]. Moreover, in order to address the statistical bias present in current VQA datasets, efforts have been made to establish datasets such as VQA v2 [26] and VQA-CP [4]. Improvements on the original VQA (VQA-v2) suggest adding similar image pairs corresponding to the same question q, but leading to diverging answers [26]. These datasets aim to block language bias and improve the model's visual comprehension potential [70].

Our approach is tested on both VQA-v2 and Visual Genome datasets. Other popular VQA datasets are Flickr30k-Entities [51], Visual7W [69], and others. For a detailed analysis of VQA and relevant topics, we refer readers to recent specialized survey papers [48, 19, 44, 71, 9, 58].

3.4 Evaluation

Evaluation metrics for open-ended VQA have been a topic of ongoing research. VQA can be evaluated either as an open-ended task, where algorithms generate a string to answer a question or as a multiple-choice task. To assess the performance of the model on multiple-choice questions, a straightforward and efficient approach is to use simple accuracy, which involves calculating the ratio of correct answers to the total number of answers. However, the evaluation of open-ended VQA algorithms using simple accuracy can be too strict because some errors are more severe than others. For instance, a system that outputs a very similar answer to the ground truth may still be considered incorrect, even though it is quite close. Additionally, some questions may have multiple correct answers, leading to further issues with the use of exact accuracy [33].

As a result, alternative evaluation metrics have been proposed for open-ended VQA algorithms. Additionally, some recent works have proposed task-specific evaluation metrics for VQA to address some of the limitations of using simple accuracy. More particularly, some metrics take into account the diversity of correct answers, the relevance of the answer to the question, or the reasoning and inference required to answer certain types of questions. Some relevant instances are the VQA accuracy [5], which evaluates the answers generated for the VQA open-ended task, an approach that

is based on human evaluation. However, there is still no universally accepted evaluation metric for VQA, and the choice of metric can have a significant impact on the perceived performance of different algorithms. There is ongoing research and discussion in the VQA community about the best ways to evaluate VQA models, while different metrics may be more appropriate for different types of questions or tasks.

3.5 The Role of Language Bias

The success of Visual Question Answering (VQA) systems depends on effectively utilizing both image and language data streams to achieve robust performance. However, studies have shown that current VQA systems may not be effectively utilizing both vision and language, but heavily rely on language. Studies have shown that models that only use questions perform significantly better than those that only use images, especially in open-ended COCO-VQA [30], [5]. The results of this experiment concatenated with the outputs regarding other datasets as well as other relevant studies [3], [68] indicate that language bias critically harms the performance and robustness potential of VQA models. This is due to the nature of questions in VQA datasets, which often constrains the expected answers essentially turning open-ended questions into multiple-choice questions, as well as strong bias in the datasets. Therefore, current VQA systems rely more on the question than on image content, and language bias in datasets significantly impacts their performance, limiting their deployment. To address this issue, new VQA datasets should strive to compensate for bias by either having questions that force analysis of image content or making datasets less biased [33].

3.6 Counterfactual Explanations for VQA

Our counterfactual approach regarding linguistic substitutions revolves around the following fundamental question: *"What is the response of M if we substitute word X with word Y in question q?"*. The implemented counterfactual $X \rightarrow Y$ substitution should be semantically *minimal* and linguistically *feasible*. *Minimality* refers to substitutions that maintain a meaning close to the meaning of the original word *X*. For example, synonym words preserve this minimality constraint. In order to ensure semantic minimality of substitutions, we leverage lexical knowledge sources (such as WordNet [23]), which can provide the minimum possible $X \rightarrow Y$ transitions by selecting the closest concept Y to concept X that respects certain constraints. Linguistic *feasibility* instructs meaningful substitutions which always involve the same part of speech (POS); for example, nouns can only be substituted by nouns but not by verbs. In total, such $X \rightarrow Y$ substitutions are applied on the whole Q set, targeting one POS at a time.

Such counterfactual questions are able to trigger alternative model responses. Therefore, a $X \rightarrow Y$ concept substitution in the input may result in an alternative $X' \rightarrow Y'$ response in the output, or not. Probing potential output changes is highly informative with respect to the reasoning process followed by the model M, highlighting concepts or concept families that are more or less influential to the decision-making process of M. Hence, the counterfactual substitutions implemented on q provide useful explanations for the model's observed behavior and enhance its interpretability extent, while also handling it as a black-box structure.

Question: What kind of material is the building closest made of?
Answer: Brick.

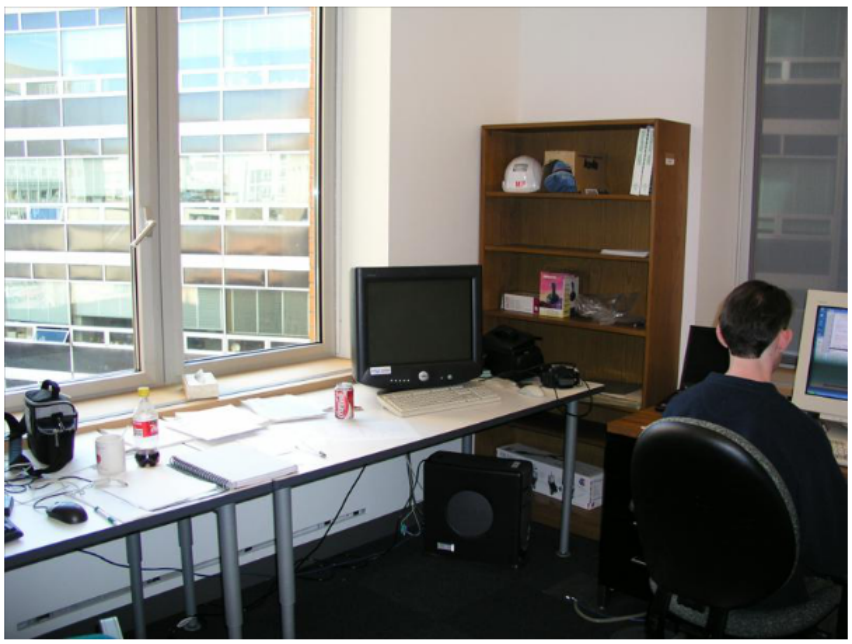
Figure 3.2: Example of image and fine-grained recognition related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: What is in front of the computer?
Answer: Keyboard.

Figure 3.3: Example of image and object recognition related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: Where is the carpet?
Answer: On the floor.

Figure 3.4: Example of image and object detection related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].



Question: What is the man doing?
Answer: Working.

Figure 3.5: Example of image and activity recognition related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: How many of these sandwiches would a vegetarian eat?
Answer: None.

Figure 3.6: Example of image and knowledge-base reasoning related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: What color is the canister set?
Answer: White.

Figure 3.7: Example of image and attribute classification related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: Where is the scene set?
Answer: Outside on a street.

Figure 3.8: Example of image and scene classification related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: How many forks are in the white plate?
Answer: One.

Figure 3.9: Example of image and counting related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: What is between the two women?
Answer: A blue and white boat.

Figure 3.10: Example of image and spatial relationship among objects related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Question: Why can the building be seen across the street?
Answer: The window blind is open.

Figure 3.11: Example of image and commonsense reasoning related question retrieved from the Visual Genome dataset [37], targeted to the VQA task. The displayed answers were given as responses by the ViLT model [36].

Chapter 4

Visiolinguistic Learning & Explainability Approaches

In recent years, pre-training models (PTMs) have revolutionized fields such as computer vision and natural language processing. They have been proven to be highly effective in improving performance in downstream tasks while avoiding the need to train new models from scratch. The adaptation of pre-training models to the field of Vision-and-Language (V-L) learning has become a focal point of multimodal learning research, as researchers seek to improve performance in downstream tasks. Significant progress has been made in exploring the application of pre-trained models to multi-modal tasks.

4.1 Vision-Language Pre-trained Models

AI researchers have long sought to create machines that can think and respond like humans. To achieve this, researchers have proposed various tasks to train and evaluate machines, such as face recognition, reading comprehension, and human-machine dialogue. However, due to technological limitations, training on a large amount of labeled data is often necessary to create capable models. Additionally, deep learning models, such as RNN, CNN, and Transformer, have been applied to solve V-L tasks, but they are typically designed for specific tasks, which leads to poor transferability [14].

To address this, researchers pre-train a huge model on large-scale general datasets and fine-tune it on specific downstream tasks to increase transferability. Pre-training models using the Transformer [63] structure have alleviated this issue by first pre-training through self-supervised learning on unlabelled data and then fine-tuning with a small amount of labeled data on downstream tasks. Pre-training models such as BERT [17], ViT [18], and Wave2Vec [57] have been successful in uni-modal fields such as Natural Language Processing, Computer Vision, and Speech, but the question remains whether they can be applied to multi-modal tasks. Researchers have explored this problem in the field of Vision-and-Language (VLP) and have made significant progress in learning the semantic correspondence between different modalities through cleverly designed model architectures [14]. Transformer has become the backbone of most pre-trained language models (PLMs), such as BERT and GPT-3, which have achieved new state-of-the-art results on various downstream tasks, including visual question answering and image captioning [20].

The pre-training procedure of a VL-PLM consists of the following major subsections. The analysis presented below, which outlines the pre-training process, is based on the relevant surveys on visiolinguistic pre-training [14], [20].

1. **Feature Extraction - Encoding**. First and foremost, the process involves encoding both images and text into latent representations (embeddings) that maintain their meaning. This step includes the preprocessing and representation approaches for images, text, or even video. To leverage uni-modal pre-trained models in VLP, visual or text features can be sent to a transformer encoder. This can be achieved by using the standard transformer encoder with random initialization or a pre-trained visual transformer like ViT [18] and DeiT [62] for ViT-based patch features, and a pre-trained textual transformer such as BERT for textual features. The methods of encoding images and texts differ because of the disparity between the two modali-

ties. While VL-PTMs typically use a transformer-based PTM as a text encoder, learning visual representations based on visual content remains an open challenge.

2. **Model Architecture**. High-performing architecture is then designed to model the interplay between the two modalities. We mention here two different viewpoints regarding the architectural design choices that propose a dilemma: One, a single-stream versus a dual-stream approach, focusing on the fusion of multiple modalities. Two, an encoder-only versus an encoder-decoder approach, examining the overall architectural design. Once images and texts have been encoded into separate embeddings, the subsequent step is to construct an encoder that can combine information from both modalities. For instance, if the model is tasked with answering a question about an image, it must merge linguistic details from both the question and answer, pinpoint the matching area in the corresponding image, and eventually synchronize linguistic meanings with visual evidence. The encoder can be classified as a fusion encoder, dual encoder, or a hybrid of both, depending on how information from various modalities is aggregated. Regarding the second aforementioned issue, the encoder-only architecture is popular among VLP models, as it directly uses the cross-modal representations to generate final outputs. However, some models prefer a transformer encoder-decoder architecture, which involves feeding the cross-modal representations to a decoder before passing them to an output layer.

3. **Efficient Pre-Training Objectives**. Third, effective pre-training tasks are developed to train the VL-PTMs. The fundamental aspect of VLP is pre-training objectives, which guide the model in learning information that is associated with vision and language. In this step, a sequence of pre-training objectives is deployed to teach VLP models universal vision-language representation. There are four main types of objectives: completion, matching, temporal, and particular. Completion involves reconstructing masked elements using unmasked parts to understand the modality. Matching aims to generate a shared hidden space for vision and language, while temporal learning involves reordering disrupted input sequences. Finally, particular types include objectives such as visual question answering and image captioning.

4. **Pre-Training Datasets**. The next step lies in the selection of the appropriate pre-training dataset. In order to achieve success in VLP, pre-training datasets are crucial. There are two main categories of pre-training datasets: image-language pre-training and video-language pre-training. The size and quality of these datasets can be more important than the actual training strategies and algorithms used. Pre-training datasets vary in size and sources across different research and are often constructed by combining public datasets across different cross-modal tasks or scenarios. However, some works use self-constructed datasets, which can be larger but may contain more noise. Such examples are VideoBERT [61], ImageBERT [52], ALIGN [28], and CLIP [53].

5. **Downstream Visiolinguistic Tasks**. Finally, once pre-trained, the VL-PTMs can be fine-tuned for downstream V-L tasks. There is a diverse range of tasks that necessitate a collaborative understanding of vision and language. Some of these tasks are briefly but illustratively described below:

    - **Visual Question Answering (VQA)**: VQA is a task that requires correctly answering a question based on visual input, such as an image or video. This task is often treated as a classification problem, where the model must select the most appropriate answer from a set of options. To achieve high accuracy, it is essential for the model to understand the logical relationships between the visual input and the question being asked. Various studies, including [5, 66, 33, 31], have explored this task.
    - **Visual Reasoning and Compositional Question Answering (GQA)**: GQA is a more advanced version of visual question answering (VQA). It aims to push the boundaries of

research on understanding natural scenes through visual reasoning. GQA's dataset contains images, questions, and answers that have corresponding semantic representations, allowing for a more structured analysis of the model's performance from multiple perspectives. Unlike traditional VQA's single evaluation metric, GQA uses multiple evaluation metrics such as consistency, validity, plausibility, distribution, and grounding, providing a more comprehensive understanding of the model's ability to reason and comprehend natural scenes.

- **Visual Entailment (VE):** The VE task involves an image as a premise and a text as a hypothesis, and the objective is to determine whether the image semantically entails the text. The task has three labels: Entailment, Neutral, and Contradiction.
- **Visual Commonsense Reasoning (VCR):** VCR involves a machine's ability to infer common sense information and cognitive understanding when presented with an image. It is in the form of multiple-choice questions with several possible answers, where the model must choose the most appropriate one and provide a reason for that choice from various options. VCR can be broken down into two tasks: selecting the best answer from a set of expected answers (question answering) and providing the reasoning behind the chosen answer (answer justification).
- **Vision-Language Retrieval (VLR):** VLR is a task that involves matching and understanding both visual (image or video) and linguistic domains using appropriate strategies. VLR encompasses two subtasks, namely vision-to-text and text-to-vision retrieval. The goal of vision-to-text retrieval is to retrieve the most relevant text description from a large pool of descriptions based on a given visual input, while text-to-vision retrieval is to find the most relevant visual input based on a given text description. VLR is widely applied in various areas, including domain-specific searches, multiple search engines, and context-based vision retrieval design systems.
- **Visual Captioning (VC):** VC refers to the task of producing linguistically and semantically suitable textual descriptions for a given visual input, be it an image or a video. The generation of informative and meaningful captions for a visual input demands an extensive understanding of the language, as well as a consistent comprehension of the objects, entities, and interactions that occur in the visual input.
- **Visual Dialogue (VD):** VD refers to a task where given an image and a dialogue history containing a sequence of question-answer pairs, the objective is to provide a free-form natural language response to a follow-up question. This task is similar to the Turing Test, where the model's performance is measured by its ability to provide human-like responses.
- **Multi-Modal Machine Translation (MMT):** MMT is a task that combines translation and text generation with information from multiple modalities, including images. The purpose of including visual features is to reduce the ambiguity that may arise in traditional text-only machine translation and to preserve the context of the text descriptions. By incorporating both visual and linguistic embeddings, MMT creates a robust representation space that enhances the semantic information in the translation process.

4.2 Knowledge in Visiolinguistic Learning

The rapid progress in visiolinguistic (VL) learning has resulted in the emergence of various models and techniques that demonstrate remarkable capabilities in solving tasks that require the collaboration of vision and language. However, the existing VL pre-training datasets have limitations in terms of the amount of visual and linguistic knowledge they encompass. Consequently, the generalization capabilities of many VL models are constrained. To overcome this challenge, hybrid architectures

have been introduced, incorporating external knowledge sources like knowledge graphs (KGs) and Large Language Models (LLMs). These knowledge sources help bridge the gaps in missing information and enhance the overall performance of VL models [45].

While VL models have made significant strides in acquiring real-world understanding through extensive pre-training procedures, they still exhibit certain limitations. Specifically, their comprehension of common sense, factual information, temporal aspects, and everyday knowledge remains somewhat restricted, raising questions about the scalability of VL tasks. By leveraging knowledge graphs and other external knowledge sources, these gaps can be effectively addressed by explicitly providing the missing information and empowering VL models with new capabilities. Furthermore, the integration of knowledge graphs not only fills these knowledge gaps but also enhances explainability, fairness, and the reliability of decision-making processes, which are critical considerations in the development of complex VL implementations [46].

Over the past few years, there has been a growing interest in integrating external knowledge, denoted as K, into visiolinguistic (VL) models. This approach has gained recognition for its potential to enhance the performance, extendability, and even explainability of existing VL tasks. The following description of the various categories of additional knowledge, as well as the types of the external knowledge form, are based on the summarization and conclusions provided in the relevant surveys on the impact of knowledge integration in visiolinguistic learning [45, 46]. Incorporating additional knowledge can provide these benefits, through the various types it includes, as outlined below:

- **Hierarchical knowledge** encompasses *is-a* relationships that form a structured tree-like arrangement. Within this hierarchy, the root represents the most general concept and acts as the parent node for all others. On the other hand, the leaves of the hierarchy represent the most specific concepts.

- **Lexical knowledge** functions as an organized dictionary that provides linguistic rules and aids in resolving challenges like word sense disambiguation. When combined with hierarchical knowledge, lexical knowledge also offers hypernym/hyponym relationships, further enhancing its utility.

- **Named entities** encompass a diverse range of proper names representing specific entities, such as individuals, places, companies, organizations, and more.

- **Factual knowledge** encompasses comprehensive information about the world, including general encyclopedic knowledge as well as specific scientific facts in domains such as medicine, biology, chemistry, and others. These facts can be combined with named entities to form more multifunctional and complete statements.

- **Commonsense knowledge** refers to the inherent and intuitive understanding of the world that humans possess. It encompasses the basic, self-evident perceptions and assumptions that guide human everyday interactions and decision-making processes. As a general term, commonsense knowledge is derived from our collective experiences, observations, and social interactions, enabling us to make sense of the world and navigate various situations. It includes a wide range of implicit knowledge about cause-and-effect relationships, social norms, contextual understanding, and practical reasoning. This type of knowledge forms the foundation for human communication, problem-solving, and decision-making, and is an essential component in the development of intelligent systems that aim to emulate human-like understanding and reasoning.

- **Event/temporal knowledge** encompasses information about the chronological order of events, combining both factual and commonsense knowledge. It provides an understanding of the temporal relationships and sequences that occur in the world. Event/temporal knowledge is

crucial for various tasks, such as understanding narratives, analyzing historical data, forecasting future events, and reasoning about time-dependent phenomena. By incorporating event/temporal knowledge into models and systems, we can enhance their ability to comprehend and reason about the temporal aspects of real-world scenarios, leading to more robust and context-aware applications.

- **Visual knowledge** encompasses a collection of images accompanied by relevant annotations that establish a connection between visual perception and commonsense understanding. It includes information about various attributes of objects, such as shape, color, texture, and more, linking them to their visual representations. By integrating visual knowledge into models and systems, we can bridge the gap between visual stimuli and conceptual understanding. Visual knowledge enables a model to recognize and interpret visual features, extract meaningful information from images, and associate visual attributes with their corresponding conceptual descriptions. This integration enhances the capability of models to perceive and reason about visual data, leading to more comprehensive and contextually-aware visual understanding and analysis and more complex reasoning efficacy.

The selection of the external knowledge source plays a crucial role in determining how knowledge is accessed and utilized within VL models. External knowledge can be further analyzed if categorized into the following groups:

- **Explicit knowledge** encompasses the structured information stored in knowledge graphs (KGs). This knowledge is represented symbolically as triplets *(h, r, t)*, where entities *(h, t)* are connected by relationships *(r)*. Extracting answers from a knowledge graph follows a transparent and deterministic process, allowing for the recovery of the exact path taken. By relying on clear, structured facts, explicit knowledge within a knowledge graph effectively addresses gaps that cannot be filled through transfer learning alone.

- **Implicit knowledge** refers to information that is stored in a non-symbolic form, such as neural network weights. It encompasses unstructured knowledge that is learned and encoded within the weights of neural networks during model pre-training. This approach allows for the integration of multiple and diverse data sources on a large scale, without relying on human supervision. While neural architectures have gained significant popularity in deep learning research, their primary focus may not necessarily be knowledge representation. However, through unsupervised or self-supervised pre-training of transformer models, implicit knowledge can be acquired and utilized for various linguistic and multimodal tasks. By incorporating extensive linguistic and visual data during the pre-training phase, it is possible to create unstructured knowledge bases.

- **Web-crawled knowledge** refers to unstructured knowledge gathered from the web, combining the advantages of implicit and explicit knowledge bases. It eliminates the need for labeled data and expensive pre-training typically associated with implicit knowledge. Accessing online sources is convenient, allowing for control and customization of the retrieved knowledge according to the task requirements. However, the quality of web-scraped knowledge poses a challenge as it is difficult to validate the reliability of web sources. Transparency is partial, allowing sentence tracking for predictions, but the reasoning process is not as explicit as in structured graphs.

4.3 Explainability in Visual Question Answering

Regarding the research topic of explainability and robustness in Visual Question Answering [50, 6, 27], multiple efforts have been proposed, including attention maps [42, 29], and other model-specific approaches [56]. The strategy through counterfactuals is a rather new one, while already

existing attempts focus on visual perturbations [11], masking [15, 16], introducing counterfactuals in the training stage [1, 16] and relationship-driven approaches between original and counterfactual samples [40].

4.4 Linguistic perturbations

There is a variety of prior works that perform word-level linguistic perturbations, even though they target purely linguistic tasks, mostly text classification [55, 65, 25, 49, 34], but also semantic similarity [43] and machine translation [64]. Our perturbations regarding synonym replacement and random noun deletion are inspired by [65], guiding substitutions with the usage of WordNet [23]. Color perturbations are adapted from [43], upon which we construct an appropriate hierarchy based on color distance. The rest of our implemented replacements involving noun and verb substitutions are completely novel ideas.

Chapter 5

Method

The model and framework work as follows: Generally, a VQA model receives a question (Q) and a corresponding image (I), and based on that, it selects an appropriate answer (A)among pre-defined candidates. The counterfactual method which we employ refers to the minimum possible feasible adversarial attack to achieve a change in the response of a model. In our work we focus on textual counterfactuals, affecting either Q or A at a time, since text is highly suitable for conceptual substitutions (a word can easily be replaced by another word) contrary to visual substitutions (an object cannot easily be plausibly replaced by another object). The input of our framework consists of dataset D which contains aligned images (I), textual questions (Q), and candidate textual answers (A). The inputs of D are inserted into a pre-trained VQA model M fine-tuned for VQA; We selected ViLT as a proof-of-concept, and we performed the analysis on the Visual Genome and VQA-v2 datasets. In our work, we employ word replacements that are guided by external knowledge sources. We use the WordNet knowledge graph (which provides hierarchical relationships between common words) and a hierarchy of color relatedness (that is extended from the Matplotlib color2 relationships).

Below is an overview of our linguistic perturbation techniques and ideas that will be applied independently to either Q's or A's. On Q's substitutions, we replace the nouns with other neighboring noun concepts using WordNet. These replacements include child concepts (more specific concepts in the hierarchy, as provided by WordNet hyponyms), parent concepts (more generic concepts, as provided by WordNet hypernyms), and sibling concepts (concepts that share the same immediate parent to the ground truth concept). As for verbs and adjectives, we once again exploit WordNet to replace extracted verb and adjective concepts with their synonyms, while also exploring meaningful antonyms perturbations. Colors present in Q's are substituted with similar ones based on our constructed color-relatedness hierarchy. Another approach we investigate is the deletion of nouns instead of substitutions. We repeat the same concept replacement/deletion process for candidate answers A's while considering the initial (non-counterfactual) Q's. We measure the performance of a model using accuracy scores. For each substitution, we obtain the counterfactual question accuracy *acc*(Q), as the response of the VQA model M to the counterfactual Q, and we compare it to the ground truth accuracy scores *acc*. Similarly, we obtain the counterfactual answer accuracy *acc*(A), which we compare with both *acc*(Q)and *acc*.

A pivotal aspect of our work lies in the deep understanding of the corresponding datasets on which a model performs. For this reason, we have implemented a thorough exploratory data analysis on the Visual Genome and VQA-v2 datasets, to gain insights that would guide the design of our perturbations. Later on, some illustrative instances of data visualization that depict useful attributes of the Visual Genome and VQA-v2 datasets will be presented and analyzed.

Our work contribution and novelty can be summarized as follows: Throughout this process, we evaluate whether and how the response of M changes, as an indicator of its robustness against replacements with semantically related concepts. Standalone accuracy scores are not informative enough to explain why we observe such behaviors. To this end, we separately examine samples where the predicted A changes under the presence of a perturbation, obtaining local explanations derived from unexpected model responses to counterfactual Q or A inputs, in the form: "if con-

cept changes in Q (or A), then A -erroneously- changes". The aggregation of such local rules leads to global explanations, deriving if-then relationships that apply to multiple samples of D. Those reveal the overall model response to each of the designed counterfactual inputs. We design counterfactual Q and A inputs using a variety of structured word-level replacements, as instructed by hierarchical knowledge sources, thus employing deterministic, controllable, optimal counterfactual perturbations. Our approach is model-agnostic, as we treat any VQA model as a black box.

To provide an insightful example of our perturbations, let us consider the question "What color is the cat?", referring to a relevant image. By performing the hypernym transformation from "cat" to "animal", we seek to evaluate whether the model will be affected. Specifically, if its initial response erroneously changes, or if we observe it handles such perturbations with reduced certainty, we can infer that the model does not efficiently perceive such hypernym-hyponym hierarchical relationships.

5.1 Framework Description

The input of our framework consists of a dataset D that contains aligned images *I*, textual questions forming a set Q, and candidate textual answers *A*. We will later present results on Visual Genome (VG) [37] and VQA-v2 [26], which satisfy these requirements.

We select ViLT [36] as a proof-of-concept pre-trained VQA model M. Nevertheless, our proposed method is not restricted to ViLT, as it only considers inputs (questions) and outputs (answers). ViLT receives a question $q \in Q$ and an image $i \in I$ from D and then generates an answer a, rather than selecting one of the candidates $a \in A$; since this behavior is inherent to several VQA models, it is important to allow looser definitions of the accuracy metric. To be more precise, ViLT produces outputs a in the form of natural language text, and there are many different ways of expressing the same answer in English. In this case, heuristically comparing the generated answer with the ground truth answer from *A* defines if the prediction of M is accurate or not. By repeating the same prediction process for all *Q, I* pairs, and by obtaining successful or unsuccessful answers for them, we finally extract an accuracy score acc_Q, reflecting the ratio of correct answers over all generated answers.

Our word substitutions are guided by external knowledge sources, targeting different parts of speech (nouns, verbs, and adjectives) at a time. Specifically, the WordNet knowledge graph [23] provides hierarchical relationships between an abundance of common words widely present in VG and VQA-v2 vocabularies. Therefore, substitution pairs are created by connecting specific words with their WordNet matches, respecting hierarchical relationships as described in Section **??**. Furthermore, we extend the Matplotlib color[2] relationships presented in [43], forming a hierarchy of *color relatedness*. This color hierarchy is based on color distances according to the RGB value of each Matplotlib color, with more details provided in Section **??**.7

We then proceed with applying our designed perturbations on dataset questions $q \in Q$, resulting in *counterfactual questions* $q^* \in Q^*$.

We present a visual outline of our approach in Figure 5.1. Words from $q \in Q$ colored in red denote the concepts to be substituted, while words in blue indicate the knowledge-driven substitutions that lead to counterfactual questions $q^* \in Q^*$.

5.2 Evaluation

Initially, we run ViLT on the original (*I, Q, A*) inputs to obtain ground truth accuracy scores *acc* as a baseline. For each substitution, we obtain the *counterfactual question accuracy* acc^*_Q, as the response of M to the counterfactual questions $q^* \in Q^*$, and we compare it to the ground truth accuracy

[2] Matplotlib colors

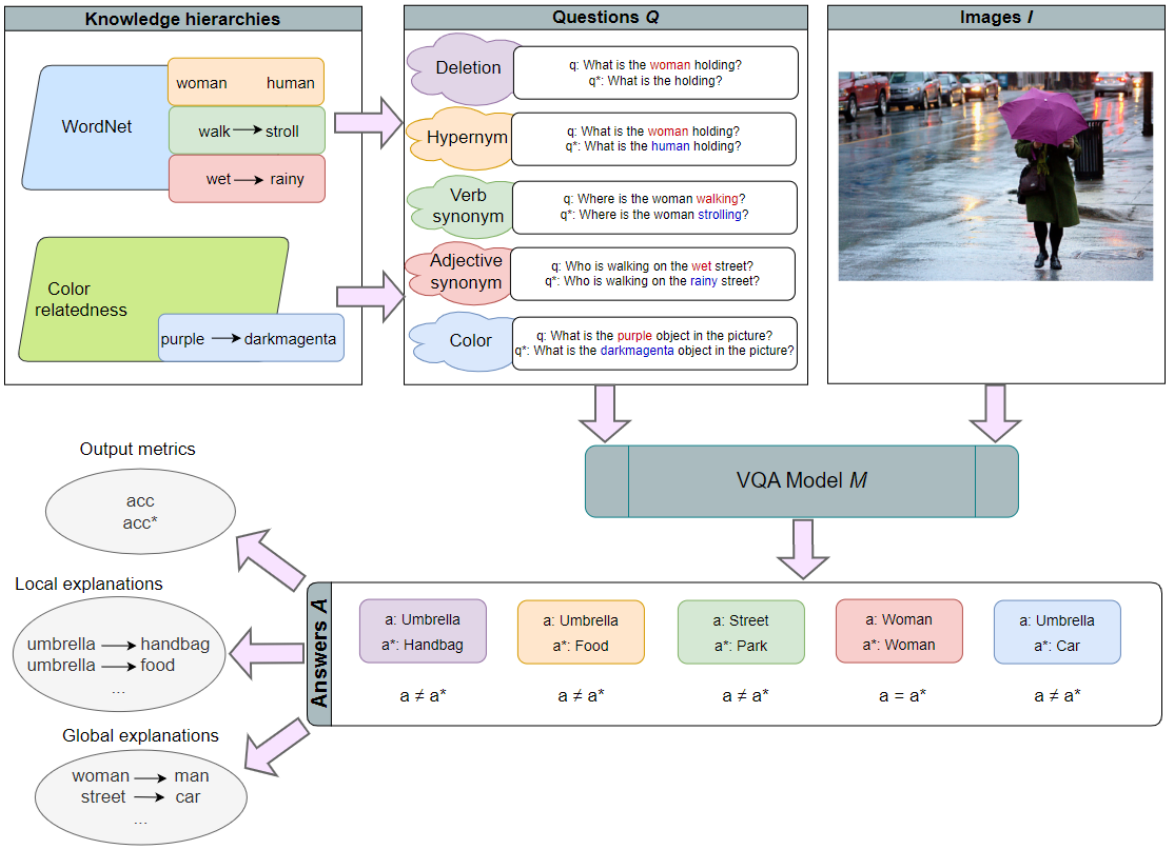


Figure 5.1: Overview of our proposed knowledge-based counterfactual VQA framework.

scores acc_Q. Throughout this process, we evaluate whether and how the response of M changes by measuring the difference between acc_Q and acc^*_Q, as an indicator of its robustness against replacements with semantically related concepts. We heuristically measure whether a pair of answers is correct, and as a result, the perceived accuracy is greatly increased. This however is not an issue in our analysis of the results. We are interested in how the accuracy changes with perturbations, not each experiment's absolute measures of accuracy.

Even though useful for benchmarking reasons, standalone accuracy scores are not informative enough to explain *why* we observe such differences between original q and counterfactual inputs q^*. To this end, we separately examine samples where the generated a changes under the presence of a perturbation, obtaining local explanation in the form *if concept changes in q, then a -erroneously- changes*. The aggregation of such local rules leads to *global explanations*, deriving *if-then* relationships that apply to multiple samples of D.

5.3 Exploratory Data Analysis

A pivotal aspect of our work lies in the deep understanding of the corresponding dataset on which a model performs. For this reason, we have implemented a thorough exploratory data analysis on the Visual Genome and the VQA-v2 datasets, to gain insights that would guide the design of our perturbations. Below, some illustrative instances of data visualization that depict useful attributes of the two aforementioned datasets will be presented.

A preliminary initial observation is the similarity of our findings for the two datasets. Despite the small differences that are inevitably present due to the particularity of each, we can safely conclude that they present similar characteristics in terms of the distribution of questions, hence it is reasonable to design a system of counterfactuals that utilizes both datasets uniformly and equally. Given that the particular datasets utilized in this thesis constitute two of the most prevalent available

datasets for the Visual Question Answering task, this observation could more generally suggest that the available data for this problem present a proportional uniformity in the linguistic and conceptual distribution of the questions.

The most significant results that we can draw from this study are those analyzed below:

5.3.1 Question Size Frequencies

Regarding the most seen size of the questions, we notice that both datasets are typically characterized by small input questions. More specifically, they do not exceed the length of seven words on average. This is a positive finding for our work: language models do not typically truncate small inputs, meaning that there would be no occurrence of unwanted truncations that could most possibly harm the model's performance and cause major difficulty in our debiasing mission.

Table 5.1: Question Sizes - Visual Genome and VQA-v2 Datasets

Question Size	Occurrences	
	Visual Genome	**VQA-v2**
1	6	21
2	40,005	11,875
3	204,556	59,480
4	573,664	115,836
5	276,955	94,804
6	143,450	68,642
7	114,020	45,088
8	47,468	22,277
9	20,711	11,387
10	12,860	6,704
11	6,237	3,508
12	2,620	1,752
13	1,404	1,044
14	704	615
15	316	326
16	149	198
17	102	114
18	53	59
19	21	30
20	11	17
21	3	1
22	3	-
23	3	-
25	1	-

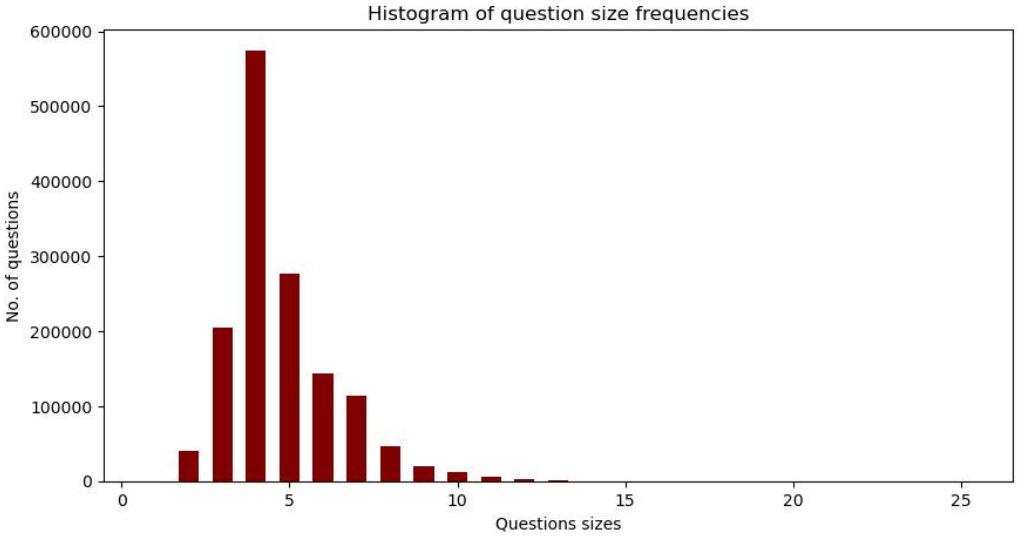


Figure 5.2: Question Size Frequencies - Visual Genome

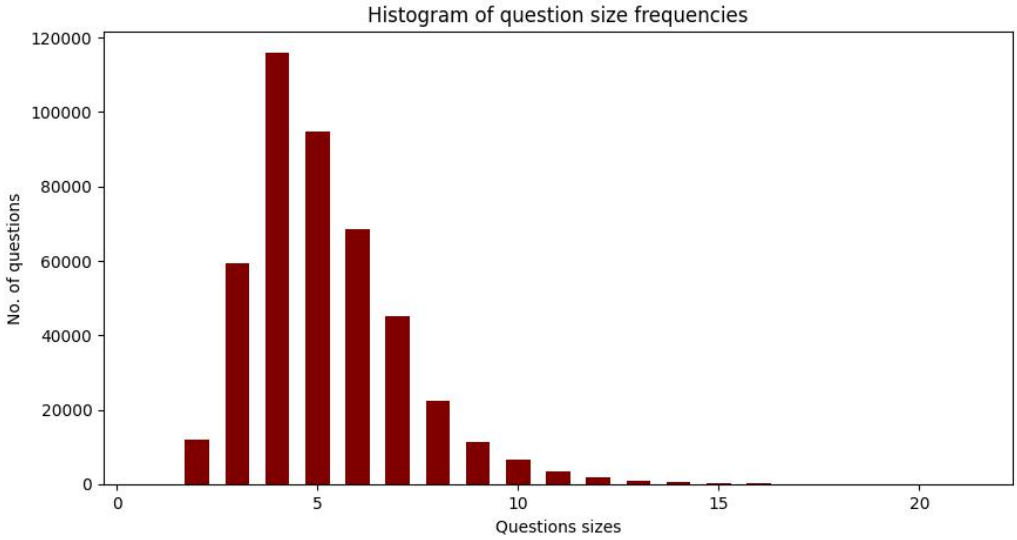


Figure 5.3: Question Size Frequencies - VQA-v2

Figure 5.4: Question Size Frequencies for Visual Genome and VQA-v2

5.3.2 Percentage of Part-Of-Speech Categories in Questions

The exploration of the distribution of the different parts of speech in the input questions provides an insight into the impact that our perturbations will eventually have. It is apparent that nouns exhibit the highest occurrences in both Visual Genome and VQA-v2 datasets. This is a dominant indicator that noun perturbations will be the most influential in our experiments. By exploiting this established intuition, we proceed toward structuring the perturbations, which we will analyze later. Additionally to that, we notice that verbs and adjectives present a similar level of frequency occurrences, which denotes that verb and adjective substitutions will be equally insightful and impactful for our experiments.

Table 5.2: Total Occurrences of Different Part-of-speech Categories - Visual Genome and VQA-v2 Datasets

Part-of-speech Category	**Occurrences**	
	Visual Genome	**VQA-v2**
NOUN	2,177,012	762,365
DET	1,671,227	553,789
PUNCT	1,452,446	448,091
AUX	1,416,747	450,666
ADP	744,451	245,259
PRON	735,193	230,480
VERB	577,006	202,657
SCONJ	494,135	74,636
ADJ	367,191	156,549
PART	78,286	26,959
ADV	64,459	35,077
PROPN	27,932	11,844
CCONJ	22,674	14,821
NUM	12,693	6,609
SYM	203	205
X	97	23
INTJ	59	134
SPACE	0	1,169

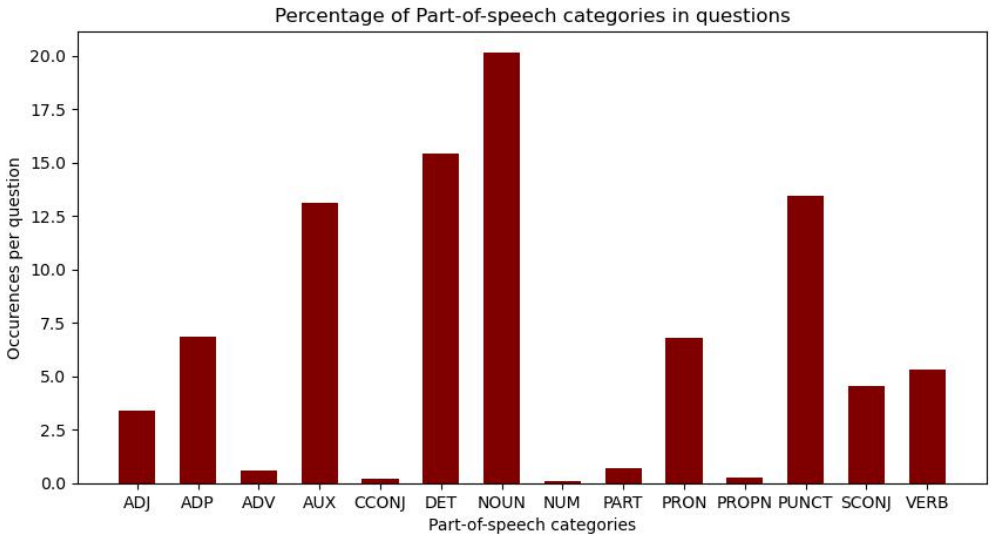


Figure 5.5: Percentage of Part-Of-Speech Categories in Questions - Visual Genome

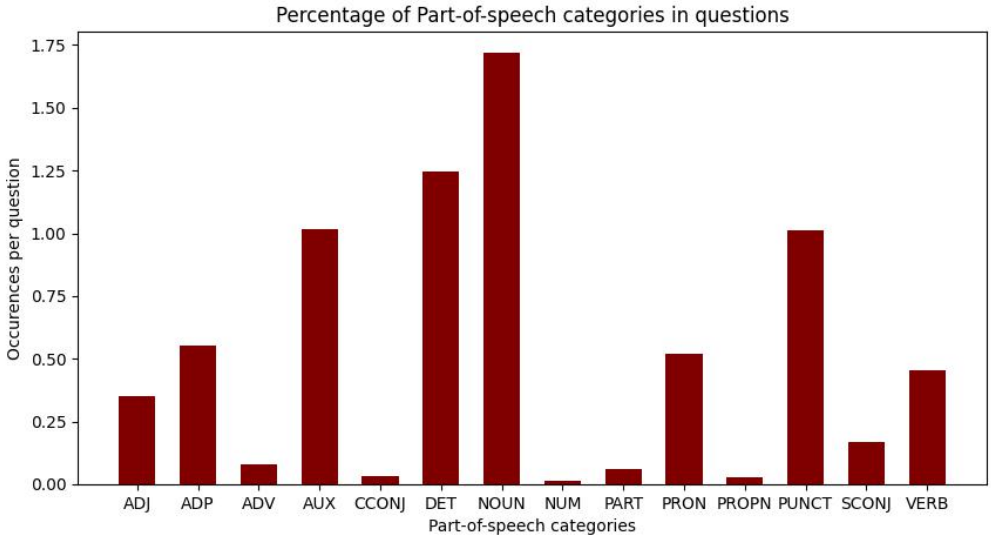


Figure 5.6: Percentage of Part-Of-Speech Categories in Questions - VQA-v2

Figure 5.7: Percentage of Part-Of-Speech Categories in Questions for Visual Genome and VQA-v2

5.3.3 Types of Questions

By analyzing the types of questions that are present in the two datasets we utilize, we gain a valuable image of our problem environment, based on the following notion: by exploring the questions, we gain an image of what the answers may be as a type. Thus, question types distributions provide an expectation of the corresponding answers. First and foremost, the "What" questions are the overwhelming majority of the total. This indicates that a *"What"* to *"How"* kind of perturbation would dramatically alternate the meaning of the question and probably lead to pointless questions that could cause confusion to the model and unjustifiably harm our study.

Table 5.3: Types of Questions - Visual Genome and VQA-v2 Datasets

Question Word	Number of Questions	
	Visual Genome	**VQA-v2**
What	868,265	182,826
Where	244,417	12,398
How	157,232	53,424
Who	78,534	3,322
When	50,869	-
Why	38,663	4,891
What's	5,575	-
Is	-	113,162
Are	-	33,136
Does	-	12,021
Which	-	5,382
Do	-	4,976

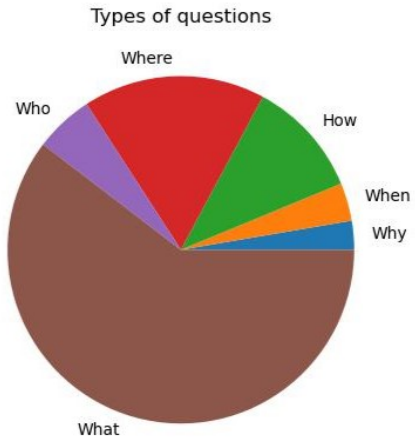


Figure 5.8: Types of Questions - Visual Genome

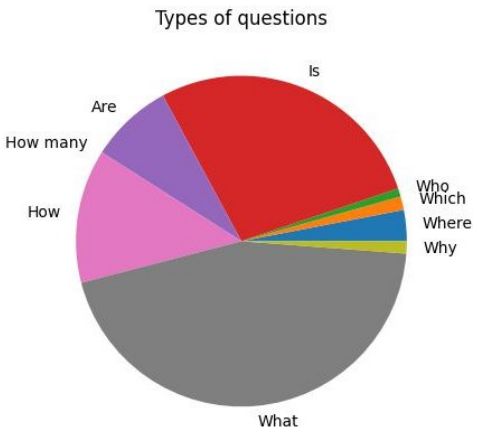


Figure 5.9: Types of Questions - VQA-v2

Figure 5.10: Types of Questions for Visual Genome and VQA-v2

5.3.4 Most Common Words

Firstly, it is noted that a mandatory preprocessing of this type of exploratory results has been completed, in order to exclude meaningless common words that do not provide valuable information on the questions' content (i.e. "what", "image", "the" etc.). Due to their extremely high frequency, they cannot prove to be insightful for the scope of this data analysis. Based on the extracted results, we notice that "color" is one of the most common words in both datasets. This finding has inspired the original color substitution that we have designed and will further demonstrate later. Moreover, the dominant frequency of the word "many" indicates the presence of a lot of numeric questions, hence it alludes to their high level of influence. Finally, the vast presence of words like "man" or "people" indicates a vastly anthropocentric approach in the questions' content, which also draws attention to related types of experimental perturbations.

Table 5.4: Word Frequencies - Visual Genome Dataset

Word	Occurrences 1	Word	Occurrences 2	Word	Occurrences 3
Color	195,855	Under	8,183	They	4,239
Many	109,752	Sign	7,886	Water	4,145
Man	50,169	Being	7,355	Vehicle	4,134
Picture	43,005	Main	7,332	These	4,121
People	41,679	Side	7,302	Street	4,113
Photo	35,283	Shape	7,293	Large	4,112
To	29,640	Have	7,222	Tennis	4,104
Made	29,579	Girl	6,769	Ground	4,053
Kind	28,158	Yellow	6,704	That	4,041
Type	24,299	Train	6,599	Item	3,944
White	22,389	Time	6,566	Day	3,914
Behind	22,035	Dog	6,504	Part	3,882
And	21,428	You	6,465	Pattern	3,824
Wearing	21,364	Cat	6,448	Back	3,778
Man's	20,989	Brown	6,441	Over	3,757
Woman	20,800	Shirt	6,241	Orange	3,700
Has	16,504	Men	6,228	Room	3,653
Person	16,237	Food	6,191	Bottom	3,611
Animal	14,005	Above	6,133	He	3,597
Front	13,302	Around	6,083	Floor	3,572
Sitting	13,261	Hanging	5,877	Giraffe	3,558
Black	12,961	Two	5,877	Material	3,510
Red	12,412	What's	5,575	Riding	3,501
Holding	12,348	Covering	5,362	Bus	3,498
Standing	11,826	Object	5,192	Person's	3,457
Green	11,740	From	5,033	Plane	3,364
Top	11,148	Looking	4,959	Clock	3,340
Blue	10,888	His	4,919	Taking	3,301
Next	10,484	Out	4,821	By	3,237
Left	10,300	Building	4,805	Would	3,190
Woman's	10,101	Wall	4,697	Fence	3,190
Number	10,027	Shade	4,682	Hair	3,142

Table 5.5: Word Frequencies - VQA-v2 Dataset

Word 1	**Occurrences 1**	**Word 2**	**Occurrences 2**
Many	49,266	Dog	4,534
Color	43,550	Made	4,525
Man	19,730	At	4,512
People	17,161	Man's	4,509
These	11,665	Room	4,476
Kind	11,444	Cat	4,474
Or	11,386	For	4,228
Wearing	10,537	Food	4,225
Person	10,374	Sitting	3,979
Does	10,365	Sport	3,740
Have	9,415	Look	3,662
You	9,331	Being	3,526
Type	8,509	Sign	3,480
They	7,081	With	3,474
Animal	6,907	Time	3,439
Woman	6,815	Number	3,426
His	6,034	Can	3,426
Animals	6,017	Her	3,382
He	5,812	Who	3,324
Any	5,620	All	3,305
Which	5,387	And	3,291
An	5,023	Photo	3,270
Do	4,978	See	3,238
That	4,929	Boy	3,217
Picture	4,916	Holding	3,205
Why	4,891	She	3,164

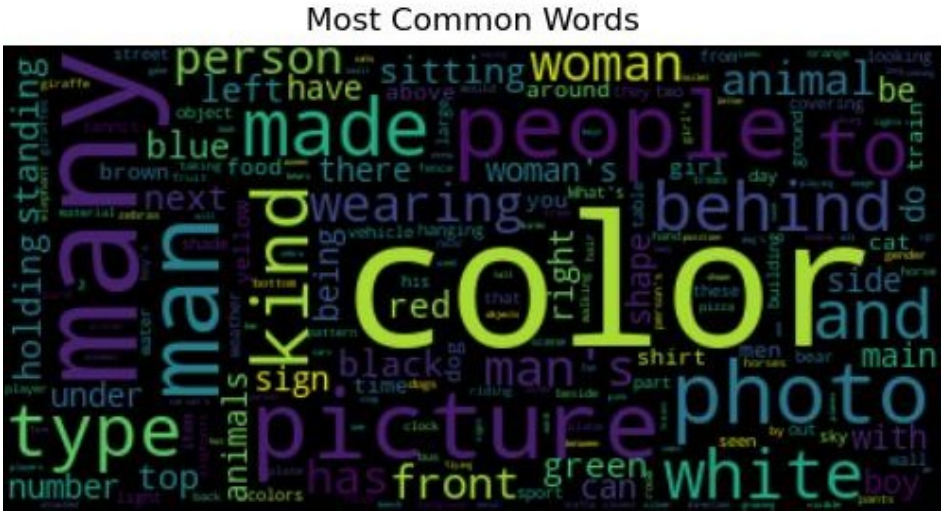


Figure 5.11: Most Common Words - Visual Genome

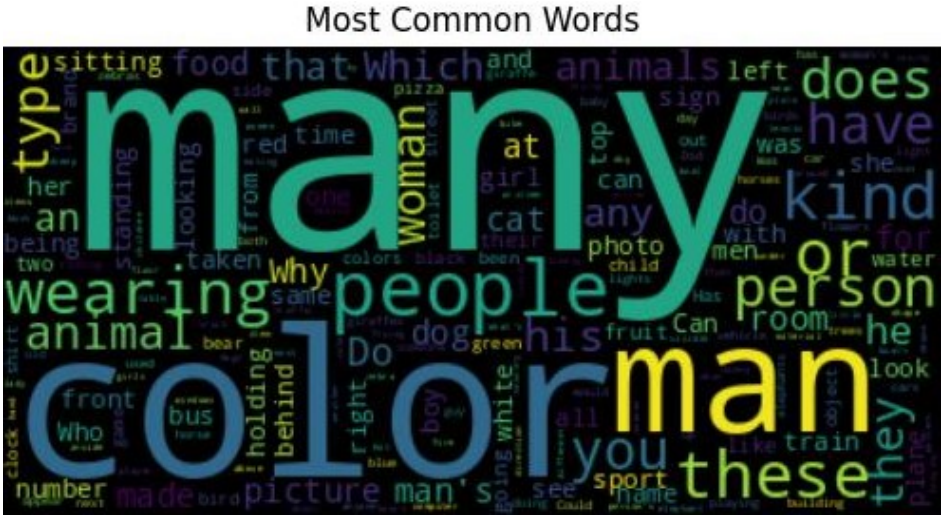


Figure 5.12: Most Common Words - VQA-v2

Figure 5.13: Most Common Words for Visual Genome and VQA-v2

5.4 Perturbations

In our work, we perform a variety of substitutions or deletions in the linguistic representation of Q. This counterfactual strategy exploits multiple and diverse morphological attributes of Q and attempts to demonstrate the semantics that most affect the model's response a. Our target is to stimulate an altered response of M through these counterfactual perturbations, to identify how the behavior of M changes when faced with different concepts. Thus, we can infer potential biases or points of weak robustness in M. The aforementioned substitutions can be divided into the following categories, based on the knowledge source used and the targeted part of speech. A summary and representative examples of substitutions are provided in Table 5.6.

A. Wordnet hierarchy: The knowledge-driven word substitutions involve replacing a noun word from questions $q \in Q$ with a hierarchically related word (*hyponym*, *hypernym*, *sibling*), or verbs and adjectives with their *synonyms*. The quality and relevance of our substitutions are reassured by the use of the deterministic structure of the Wordnet hierarchy, guaranteeing controllable and optimal word-level substitutions.

- **Synonyms**: We employ synonym transformations on adjectives and verbs of the original questions $q \in Q$. For example, "talk" and "speak" are *synonym verbs* according to WordNet, while "small" and "minuscule" are *adjective synonyms*. The Wordnet hierarchy provides a relevance-wise organized priority of synonyms and we select the most relevant one. This way, optimal and controllable substitutions are guaranteed.

- **Hypernyms - Hyponyms**: More *general*, as well as more *specific* noun concepts are provided via WordNet in the form of *hypernyms* and *hyponyms* respectively. For example, for a given noun word (e.g. "dog") we can extract its immediate noun hypernyms (e.g. "canine"), or its immediate hyponyms (e.g. "labrador").

- **Siblings**: We construct noun *sibling* substitutions by traversing the Wordnet knowledge tree one step upwards and then one step downwards. Siblings are defined as noun entities that share the same immediate parent. For example, "carrot" and "radish" are siblings, because they both have "plant root" as their parent concept according to WordNet.

B. Color relatedness hierarchy: Colors that are semantically similar, therefore presenting RGB values close to each other, will be also close within the color-relatedness hierarchy. For example, "violet" and "orchid" Matplotlib colors lie close within the color hierarchy (their in-between color distance is 6.16), while "violet" and "deepskyblue" are placed far away from each other (their color distance is 207.88). Colors can be replaced with either distant or else similar colors from this color-relatedness hierarchy, leading to the following **Color Maximal** and **Color Minimal** color substitutions. Both Maximal/Minimal substitutions may either involve *common* colors, which already exist in the dataset, or else *uncommon* colors, which belong to the Matplotlib color list but not in VG/VQA-v2 vocabularies:

- **Color Maximal**: On questions that mention some specific color we contradict the output a of M based on the input of the original question $q \in Q$ vs the output a^* of the perturbed question $q^* \in Q^*$. In q^* the original color is substituted with one that is greatly distant to it, such as "violet" $\rightarrow$ "deepskyblue". In this category, we also challenge the model using less frequent color instances (i.e. "azure", "turquoise", "salmon"). This substitution diverges from the initial counterfactual question requesting *minimal* changes; nevertheless, the comparison with related *minimal* changes will highlight the differences that varying color distances impose on the final acc^*_Q.

- **Color Minimal**: In accordance with the above, we perform color substitutions with the least distant colors, such as "violet" $\rightarrow$ "orchid". Again we also challenge the model with less frequent color substitutions.

C. **Deletions**: We randomly select a noun in each question $q \in Q$ and remove it.

Table 5.6: Question perturbations examples towards counterfactual queries.

Perturbation	Question
Original	Do you see the white small dog?
Color Maximal	Do you see the **black** small dog ?
Color Minimal	Do you see the **beige** small dog ?
Synonym Adjectives	Do you see the white **tiny** dog ?
Synonym Verbs	Do you **watch** the white small dog ?
Hypernym Noun	Do you see the white small **canine** ?
Hyponym Noun	Do you see the white small **labrador** ?
Sibling Noun	Do you see the white small **wolf** ?
Deletion Noun	Do you see the white small _ ?

5.5 Added Value of Designed Perturbations

Substitutions and deletions are an excellent way to quantify whether a VQA model M understands a specific question-image pair, or if its output is greatly dependent on biased estimations. This way, we can reveal spurious correlations that are mistakenly integrated into M.

- **Color Substitutions**: Color substitutions are motivated by the quantity of color-related questions that exist in our input datasets (VG and VQA-v2). By interrogating M with color perturbations that are greatly distant to the original color (**Color Maximal** substitutions experiment), we aim to detect whether M will correctly and reasonably perceive this semantically massive change. We would expect M to change its response a^* in most cases of **Color Maximal** experiment; the opposite would indicate an underlying pattern of ignoring color attributes. Similarly, we perform the **Color Minimal** substitution experiment to investigate the model's behavior when faced with minor alterations in the color concept. We expect the substitutions from this experiment to have little to no influence on the model's response a^*. An opposite behavior would reveal an existing bias regarding specific colors, which would lead to the conclusion that M cannot properly and robustly adapt to minor color changes and generalize accordingly. Of course, *uncommon* color substitutions in both Minimal/Maximal cases impose a more difficult problem, as M needs to adaptively respond to out-of-dataset color concepts.

- **Synonym Substitutions**: With the **Synonym** substitutions, we aim to investigate the model's ability to efficiently handle mild morphological language alterations that maintain the same meaning. In this case, failing to properly respond (providing an alternative $a^* = a$) would disclose overfitting to specific semantics, which renders the model lexically inflexible and thus non-robust to semantically negligible perturbations.

- **Hypernyms-Hyponyms Substitutions**: The **Hypernyms-Hyponyms** perturbations are dedicated to depicting the model's ability to generalize and specify correspondingly while retaining a reliable level of robustness. Hypernym and hyponym relationships are notions profoundly understood in the real world and consequently embedded in large-scale datasets, which are widely used for VQA models pre-training. Thus, M should also be able to properly comprehend and reason over them. Ideally, we would expect M to maintain the same response for hypernyms substitutions, whereas justifiably respond in specific ways for the hyponyms replacements, taking into account the specification of meaning. The commensurate amount of specifying skill is sought to be established through **Sibling** substitution experiment.

Depending on each particular case, we expect M to modify or maintain its response appropriately, to confirm the level of understanding and distinguishment of different, but still related, meanings.

- **Deletions Substitutions**: Finally, we implemented the **Deletion** experiment expecting ideally the performance of M to degrade. The amount of the acc^*_Q decline depends on the importance of the deleted noun for the meaning of the question. Consequently, an unbiased M should be able to determine this importance and act accordingly, without reaching unwarranted conclusions that are expressed through an indefensible response.

Chapter 6

Experiments and Results

We present results using the accuracy metric, which illustrates the extent of similarity of the model's predicted answer to the ground truth answer, both for the original Q of each dataset, as well as for the counterfactual question set Q^*. In our analysis, accuracy is not profound enough to provide specific situational explanations and insights on the model's behavior, when faced with particular concepts. However, accuracy still showcases a high-level approach to the model's efficiency fluctuations under the implemented counterfactual perturbations. Since ViLT model is trained and optimized on the VQA-v2 dataset, it is somehow expected to perform better on it compared to VG (both datasets contain similar vocabularies). This observation is indeed validated by our results presented in Tables 6.1 & 6.2, which demonstrate a consistently higher acc_Q on the former versus the latter dataset, concerning all implemented experiments. We denote that acc_Q scores for each experiment contain the corresponding questions only, e.g. color experiments only contain questions that mention colors. This contributes to the differences in original acc_Q scores for each experiment.

Nevertheless, in both datasets, we notice an analogous difference between acc_Q and acc_Q^* per experiment when M is presented with counterfactual questions $q^* \in Q^*$. This could generally indicate the existence of underlying biases: M presents a type of overfitting to the original $q \in Q$, which renders it less efficient when asked to handle minimally perturbed counterfactual questions. In all experiments, the accuracy reduction from the original acc_Q to acc_Q^* is approximately 10-35% or more.

An extended depiction of the retrieved accuracies for both datasets is presented in Table 6.1 (color-based substitutions) and Table 6.2 (WordNet-based substitutions and noun deletions). Specifically, for **Color Maximal** in VQA-v2 we observe a decline of 35.2% for *common* colors and a decline of 37.1% for *uncommon* ones. Even for semantically minimal substitutions, (**Color Minimal** experiment), the decline is around 32% for both *common* and *uncommon* colors. As for VG, we observe a decline of 35.5% for *common* colors and a decline of 41.8% for *uncommon* colors when **Color Maximal** substitutions are performed. Correspondingly, a decline of 33.3% for *common* and a larger 41.4% decline of accuracy for the *uncommon* colors is reported for **Color Minimal** substitutions.

Table 6.1: Accuracies for color perturbations on VQA-v2 and Visual Genome (VG). Common refers to substitutions with in-dataset colors, while uncommon refers to substitutions involving any Matplotlib color.

	acc_Q%		acc_Q^*% (common)		acc_Q^*% (uncommon)	
Perturbation	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
Color Maximal	69.4	47.1	45.0	30.4	43.7	27.4
Color Minimal	69.6	47.8	47.5	31.9	46.1	28.0

The discovery of global patterns provides a more profound and targeted view of the robustness of the model M. To this end, we note that in all the cases we studied, we are not interested in the ground truth answer of a question q, but rather in the differentiation between the answer a^* to the counterfactual question in relation to the original answer a that M predicts, either if a is correct or

Table 6.2: Accuracies for WordNet-based perturbations on VQA-v2 and Visual Genome (VG).

	acc_Q%		acc^*_Q%		acc_Q reduction %	
Perturbation	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
Synonym Adjectives	75.3	45.6	57.4	36.3	23.7	20.3
Synonym Verbs	76.5	51.5	63.9	45.6	16.5	11.5
Hypernym Noun	75.1	51.7	61.4	40.0	18.2	22.7
Hyponym Noun	75.1	52.1	56.9	35.4	24.2	31.9
Sibling Noun	76.7	51.7	54.2	32.8	29.3	36.6
Deletion Noun	76.8	51.6	59.2	36.9	22.9	28.6

not. We select this approach since we are interested in discovering the model's change in decision-making under the presence of counterfactual inputs, which is more informative than measuring how much a^* semantically deviates from the ground truth response.

Based on the thorough investigation of our experiments' results and the aggregation of the following *local explanations*, as presented in the upcoming Figures, we have deduced some meaningful *global rules* that both embody the robustness of M to our counterfactual questions $q^* \in Q^*$, while providing reliable explanations that reveal the model's reasoning behind its decision-making. Furthermore, we analyze underlying existing biases of M that logically derive from these global rules. In the following Figures, we highlight the original q, a with red and the counterfactual q^*, a^* with blue.

For a more detailed explanation of the methodology for extracting the results we present below, we provide the Wordclouds, accompanied by the corresponding numerical values expressing the normalized frequencies of occurrences of the transitions resulting from the substitutions we implement. Wordclouds and tables for the following explanations can be found in section Bias Extraction Data 6.9.

6.1 Color Maximal explanations

In **Color Maximal** substitutions, we notice that M erroneously maintains the same answer $a^* = a$ when we replace the colors *gray* and *silver* with any other semantically maximal color, either common or uncommon (underlined). Therefore, we detect a bias in the model related to these two colors, as it does not make logical decisions after replacing them with others and does not properly reason over this substitution. A relevant example is presented in Figure 6.1a.

However, contrary to the above, M logically revises its answers when we replace the colors *green* and *red* with any distant colors, either common or uncommon (underlined) as presented in Figure 6.1b. Therefore, the model recognizes and qualitatively understands these substitutions and has not incorporated any problematic attachment regarding these two colors.

This observation denotes that M is more sensitive towards intense and visually distinct colors and rather bypasses changes involving more neutral ones, focusing on object identities (e.g. "fire hydrant" and "posts" of Figure 6.1a). A more uncertain a^* (e.g. the model's answer a^* could be "nothing") would be more suitable if all question semantics were equally taken into account.

6.2 Color Minimal explanations

Based on our experiments on semantically minimal color substitutions, we derive the following global rule: When we replace the colors *gray* and *purple* with any other closely related color, M tends to give the same answer $a^* = a$. Therefore, M does not give due importance to this change of

(a) **q**: What is surrounding the silver/black/navy fire hydrant?
a: posts/posts/posts.

(b) **q**: How many glasses have red/gold/darkturquoise wine?
a: 6/0/0.

Figure 6.1: Local explanations for **Color Maximal** counterfactual perturbations.

colors, a fact that highlights a robust behavior related to the two aforementioned colors. The model maintains this invariant behavior equally when we perform replacements with *common* colors or with *uncommon* ones, as presented in Figure 6.2a.

In contrast, M redefines its answers when we replace the *green* color with any other, *common* or *uncommon*, semantically similar color. Consequently, it is being confused by such minimal changes, failing to provide a meaningful answer, as shown in Figure 6.2b. Even a more uncertain answer (e.g. "nothing") to counterfactual questions would be more suitable compared to the semantically divergent ones returned ("bus" and "bag" instead of "light"). Likewise, M presents a similar change in behavior when we replace the *pink* color with common minimal colors and the *silver* color with uncommon ones.

(a) **q**: What organization's logo is on the purple/blue/plum banner?
a: olympics/olympics/olympics.

(b) **q**: What is being held green/forestgreen/olive?
a: light/bus/bag.

Figure 6.2: Local explanations for **Color Minimal** counterfactual perturbations.

6.3 Synonym Adjectives explanations

In general, adjective-noun pairs present in questions contain some joint special conceptual meaning which differs from the independent meaning of adjectives when they exist autonomously and separately in a sentence. In this case, we notice that the model M varies its answer a^* when we implement a synonym substitution of its question adjectives. This finding suggests that M can qualitatively perceive the meaning of such adjective-noun pairs and differentiate its response accordingly, as presented in Figure 6.3a.

Another finding is related to the ability of M to correctly adjust its answer, when it is presented with a lexically correct synonym to an adjective, which, however, is not quite appropriate for the

given linguistic environment of the question. Accordingly, we conclude that M is capable of understanding the meaning of an adjective in relation to the context of the sentence (”typical food” is meaningful, but ”distinctive food” is not), as presented in Figure 6.3b.

In addition, we derive a global rule that concerns the behavior of M when we replace an adjective having multiple meanings with one of its synonyms, which, although it is optimal with respect to the aforementioned meanings, is however not suitable to the semantic context of the substituted adjectives (such as ”delicious” vs ”delightful”). We note that in this case, M demonstrates a stable behavior against such substitutions, which proves that it is able to reason over adjectives in a contextualized manner, without being fooled by synonyms not suitable to the exact context of the question q. An example of this observation is provided in Figure 6.3c.

Size-related adjective substitutions are demonstrated in Figure 6.3d. We observe that M is particularly robust to such substitutions, therefore correctly capturing the underlying meaning without being biased towards specific words. This global rule demonstrates the flexibility of M towards appropriately handling semantically and contextually equivalent adjective substitutions.

Finally, M is proven to be unstable when it has to handle rare or difficult synonyms of adjectives, as the ones shown in Figure 6.3e. Consequently, it presents a lexical weakness in handling such rare adjectives and possibly a bias in specific words that are more familiar to it.

(a) **q**: Is this a hot/raging dog? **a**: yes/no.

(b) **q**: Of what meal is this kind of food typical/distinctive? **a**: lunch/hot dog.

(c) **q**: How delicious/delightful does this look? **a**: very/not very.

(d) **q**: Is this a small/little town? **a**: yes/yes.

(e) **q**: Does the man look happy/felicitous? **a**: no/yes.

Figure 6.3: Local explanations for **Synonym Adjectives** counterfactual perturbations.

6.4 Synonym Verbs explanations

Regarding substitutions involving verb synonyms, M is not particularly stable when dealing with substitutions of verbs that present multiple meanings, as presented in Figure 6.4a. This indicates a difficulty in distinguishing the correct and desired meaning among multiple ones.

An even more specific rule we extract is that M falsely changes its original answer a when we replace the verb ”see” with a qualitative synonym of it. A relevant example is provided in Figure 6.4b. This finding indicates an unwanted attachment of the model to the word ”see”, which is interpreted as bias. This is an example of a more general situation, where optimal synonyms may not be the best choice for a synonym in a specific contextual setting. In these kinds of instances, the model tends to change its response, as observed in this particular case.

The model M presents satisfactory robustness when it has to deal with easy or common verbs of the English vocabulary, which means that it has acquired a certain degree of versatility in simple vocabulary challenges, as shown in Figure 6.4c.

Finally, M is rather stable when the replaced verb corresponds to a noun counterpart (e.g. picture -verb-, picture -noun-) or even an adjective counterpart (e.g. pictured), such as the ones of Figure 6.4d. Consequently, M is capable of capturing the general sense of such verbs in the context of the question; equivalent substitution of corresponding nouns (picture→visualization) or adjectives

(a) q: What says/state STAPLES?
a: nothing/New York.

(b) q: Do you see/understand any motorcycle helmets?
a: no/yes.

(c) q: Are the walls done/made in a summery color?
a: yes/yes.

(d) q: What kind of birds are pictured/visualized?
a: parrots/parrots.

Figure 6.4: Local explanations for **Synonym Verbs** counterfactual perturbations.

(pictured→visualized) would mostly yield the same counterfactual response a^*.

6.5 Hypernym Noun explanations

Throughout our Hypernym Noun substitutions, we conclude that M is particularly robust against substitutions involving living creatures, such as animals or humans, which shows that it can properly reason over hierarchical relationships governing such concepts, as in Figure 6.5a.

On the contrary, M does not clearly distinguish between concepts related to types of clothing, i.e. it tends to erroneously change its answer when replaced with a broader concept. Consequently, M does not generalize well on such entities and a bias towards more specific and clear types of clothing emerges. A relevant example is presented in Figure 6.5b.

As an extension of the above, M exhibits instability in hypernym substitutions that are very broad, inclusive, and polysemous. Therefore, when we replace a noun with an optimal hypernym that presents much greater conceptual generality, M is unable to qualitatively perceive the hierarchical relation that governs them, outputting a wrong answer, as in Figure 6.5c.

(a) q: Where is the cat/feline?
a: bed/bed.

(b) q: Are all the players wearing black shirts/garment?
a: no/yes.

(c) q: Are there multiple vegetables on the plate/base?
a: yes/no.

Figure 6.5: Local explanations for **Hypernym Noun** counterfactual perturbations.

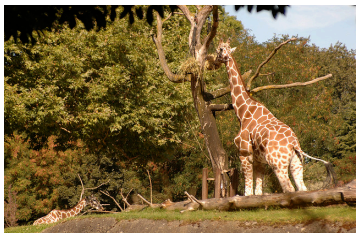
6.6 Hyponyms Noun explanations

Similar to the hypernyms substitution experiment, M is able to appropriately respond in cases where the substitutions of hyponyms refer to living entities. Therefore, the specialization in more specific living entities concepts is properly perceived, as presented in Figure 6.6a.

Correspondingly, M also shows stability in the substitutions of hyponyms that represent articles of clothing. Therefore, it specializes skillfully in more specific cloth-related entities, and according

to the case, it appropriately changes its response by adapting to the change. A relevant example is presented in Figure 6.6b.

On the contrary, the model does not demonstrate robustness to hyponym substitutions referring to means of transport. Consequently, the model is biased toward such broader concepts and fails to adequately understand their specialization, as shown in Figure 6.6c.



(a) **q**: Are the animals/acrodont eating?
a: yes/yes.

(b) **q**: Are all the players wearing black shirts/camise?
a: no/yes.

(c) **q**: What are objects behind the motorcycles/minibike?
a: sign/sign.

Figure 6.6: Local explanations for **Hyponyms Noun** counterfactual perturbations.

6.7 Sibling Noun explanations

With reference to Sibling Noun substitutions, we notice as a global pattern that M has insufficient separation ability when the sibling nouns refer to rooms of buildings or houses. As an example, in Figure 6.7a, M cannot properly differentiate between the described interior spaces and can be easily fooled by substitutions involving places of different functionality. M is also confused in the case of Figure 6.7d, when sibling means of transport are substituted. Specifically, M insists on its answer even though a concept not existing in the image appears (bike→truck). This indicates that M rather trespasses the linguistic modality context, providing an 'easy' answer based on the visual modality, since the only *bird* appearing in the image is a *parrot*. In this case, the relevant position of the bird *on the man's bike/truck* is ignored.

An interesting behavior is observed when sibling concepts involving animals are tested, as in Figure 6.7e. In this case, M seems to circumvent reasoning over the image, providing an answer based on knowledge it has most possibly acquired during its pre-training phase (zebras are black and white in color). Nevertheless, M is not fooled by the horse→zebra substitution, in which case it would conclude that *the zebra is brown*, which is a wrong factual statement. Another case that M is not being fooled is depicted in Figure 6.7c. In this case, M is very consistent in sibling entities that declare human body parts, which means that it correctly perceives their differences and does not group them in an arbitrary way. Furthermore, M presents a correct reasoning process by differentiating its answer when the sibling concepts present very different meanings between them, as the concepts air→water in Figure 6.7b.

Overall, Siblings' Noun substitution provided a rich set of insights, unequally relying on either q or I to derive an answer in many cases, rather than providing an uncertain outcome (such as answering "nothing" in the examples of Figures 6.7d, 6.7e, similarly to the correct reasoning of Figure 6.7c). In total, this indicates an unstable behavior of M towards different sibling pairs, yielding unpredictable outcomes under different substitutions of the same conceptual distance.

6.8 Deletion Noun explanations

Regarding the counterfactual questions concerning deletions of nouns, we firstly observe the following pattern: When the deleted noun has a determining role in another noun already present in the

(a) **q**: Is the bathroom/workroom organized?
a: yes/yes.

(b) **q**: Are those kites in the air/water?
a: yes/no.

(c) **q**: What is she wearing on her head/throat?
a: helmet/nothing.

(d) **q**: What bird is on the man's bike/truck?
a: parrot/parrot.

(e) **q**: What color is the horse/zebra?
a: brown/black and white.

Figure 6.7: Local explanations for **Sibling Noun** counterfactual perturbations.

question, M maintains its original answer even after the deletion. Therefore, we detect a tendency towards attaching to the determined noun, while at the same time not paying due attention to the determiner noun. Hence, M answers such questions arbitrarily, even though its answer cannot be perceived as wrong; a human could have also answered the same, especially in yes/no questions. A related example is provided in Figure 6.8a.

Another pattern that we detect concerns questions that refer to the color of a noun, which has been deleted. In these cases, M tends to respond with the most dominant color in the image, without taking into account the absence of the noun that this color should define, as in Figure 6.8b. Of course, we regard this behavior as justified, since a human would most probably answer such questions in the same way. Similarly, in questions concerning the location of a noun, which has been deleted, M answers with the most dominant entity present in the given image. A relevant example is demonstrated in Figure 6.8c.

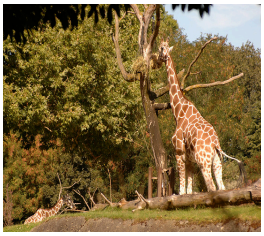
Finally, we list some nouns to which we notice that M does not pay due attention when asked to give an answer, as in Figure 6.8d. Specifically, even after deleting them, M tends to return the initial answer with great frequency. These words are: image, photographs, human, man, animal, and room. In general, these are words that are encountered very often in questions and usually act in addition to other, more specific, entities. Once again, this behavior is justified.

All in all, we observe that the random deletion of a noun results in a rather expected model behavior, driven by dominant visual concepts present in the given image.

(a) **q**: Is the womanś ~~hair~~ tied back?
a: no/no.

(b) **q**: What color is the ~~bathroom~~?
a: yellow/white.

(c) **q**: Where are the ~~cakes~~?
a: table/table.



(d) **q**: How many animals are in this ~~photo~~?
a: 2/2.

Figure 6.8: Local explanations for **Deletion Noun** counterfactual perturbations.

6.9 Bias Extraction Data

6.9.1 Tables

Table 6.3: Same Common Maximal

Initial color	Perturbed color	Same answer percentage
black	gray	0.699
gray	black	0.613
black	white	0.575
white	black	0.538
red	black	0.507
yellow	black	0.490
black	red	0.487
black	green	0.463
red	green	0.434
red	yellow	0.431

Table 6.4: Same Uncommon Maximal

Initial color	Perturbed color	Same answer percentage
black	gray	0.689
black	steelblue	0.673
yellow	mediumblue	0.648
yellow	slateblue	0.629
red	honeydew	0.627
yellow	sienna	0.624
red	deepskyblue	0.624
gray	black	0.613
black	beige	0.612
yellow	lightskyblue	0.611

Table 6.5: Different Common Maximal

Initial color	Perturbed color	Different answer percentage
green	red	0.677
green	black	0.592
black	yellow	0.572
yellow	red	0.570
red	yellow	0.569
red	green	0.566
black	green	0.537
black	red	0.513
yellow	black	0.510
red	black	0.493

Table 6.6: Different Uncommon Maximal

Initial color	Perturbed color	Different answer percentage
green	red	0.752
green	maroon	0.732
green	orange	0.725
green	salmon	0.708
green	midnightblue	0.695
yellow	chocolate	0.688
green	snow	0.678
green	magenta	0.669
green	orangered	0.664
red	yellowgreen	0.655

Table 6.7: Same Common Minimal

Initial color	Perturbed color	Same answer percentage
white	gray	0.658
gray	white	0.617
yellow	white	0.531
yellow	gray	0.520
red	gray	0.519
red	white	0.517
white	yellow	0.479
white	red	0.476
yellow	green	0.445
gray	red	0.433

Table 6.8: Same Uncommon Minimal

Initial color	Perturbed color	Same answer percentage
red	maroon	0.831
white	ivory	0.802
red	crimson	0.783
red	magenta	0.771
white	ghostwhite	0.762
green	lime	0.757
green	greenyellow	0.746
green	chartreuse	0.739
white	palegreen	0.731
white	floralwhite	0.730

Table 6.9: Different Common Minimal

Initial color	Perturbed color	Different answer percentage
green	yellow	0.647
green	white	0.641
gray	yellow	0.632
gray	green	0.594
green	gray	0.593
white	green	0.587
gray	red	0.567
yellow	green	0.555
white	red	0.524
white	yellow	0.521

Table 6.10: Different Uncommon Minimal

Initial color	Perturbed color	Different answer percentage
green	ghostwhite	0.729
green	wheat	0.697
green	yellowgreen	0.684
white	green	0.672
green	goldenrod	0.664
green	white	0.657
green	antiquewhite	0.652
green	silver	0.651
green	palegoldenrod	0.647
green	floralwhite	0.643

Table 6.11: Same Answer Adjectives

Initial word	Perturbed word	Same answer percentage
little	small	0.91
larger	bigger	0.90
large	big	0.88
middle	center	0.87
likely	probable	0.86
nearest	close	0.86
total	sum	0.852
prominent	outstanding	0.85
polar	diametric	0.83
small	little	0.80

Table 6.12: Different Answer Adjectives

Initial word	Perturbed word	Different answer percentage
blurry	bleary	0.77
cloudy	nebulose	0.72
warm	affectionate	0.66
empty	discharge	0.62
tan	sunburn	0.58
old	erstwhile	0.57
reflective	brooding	0.56
thin	dilute	0.53
touching	stir	0.53
bright	brilliant	0.51

Table 6.13: Same Verbs

Initial word	Perturbed word	Same answer percentage
played	acted	0.908
depicted	pictured	0.886
pictured	visualized	0.804
taken	occupied	0.802
hit	strie	0.801
appear	look	0.794
read	say	0.785
eat	feed	0.715
take	return	0.705
photographed	snap	0.704

Table 6.14: Different Verbs

Initial word	Perturbed word	Different answer percentage
positioned	put	0.948
shaped	determined	0.703
see	understand	0.694
say	state	0.637
arranged	set up	0.629
dressed	clothed	0.591
stop	halt	0.581
get	acquire	0.526
makes	brand	0.497
connected	link	0.497

Table 6.15: Same Hypernyms

Initial word	Perturbed word	Same answer percentage
guy	man	0.948
animal	organism	0.945
lady	woman	0.940
fridge	refrigerator	0.938
girl	woman	0.913
surfer	swimmer	0.911
jacket	coat	0.906
game	activity	0.900
cats	feline	0.899
tablecloth	table linen	0.893

Table 6.16: Different Hypernyms

Initial word	Perturbed word	Different answer percentage
weather	atmospheric phenomenon	0.875
lamp	source of illumination	0.850
tires	hoop	0.797
reflection	consideration	0.783
plant	building complex	0.783
windows	operating system	0.782
distance	spacing	0.776
bowls	bowling	0.776
plants	building complex	0.760
shot	propulsion	0.755

Table 6.17: Same Hyponyms

Initial word	Perturbed word	Same answer percentage
concrete	cement	0.875
sport	archery	0.874
appliance	gadgetry	0.834
motorcycle	minibike	0.831
tablecloth	tea cloth	0.830
station	terminal	0.825
sweater	cardigan	0.825
blanket	afghan	0.821
type	breed	0.821
slope	ascent	0.819

Table 6.18: Different Hyponyms

Initial word	Perturbed word	Different answer percentage
season	seedtime	0.981
expression	leer	0.973
shape	angularity	0.903
day	date	0.873
time	day	0.854
shot	discharge	0.844
spectators	browser	0.843
hydrants	fireplug	0.843
scene	darkness	0.839
carpet	broadloom	0.812

Table 6.19: Same Siblings

Initial word	Perturbed word	Same answer percentage
object	thing	0.909
kid	young person	0.898
dirt	soil	0.892
tennis	volleyball	0.877
child	young person	0.872
animal	zooid	0.843
ocean	waterway	0.841
carpet	rug	0.825
field	yard	0.821
runway	track	0.816

Table 6.20: Different Siblings

Initial word	Perturbed word	Different answer percentage
season	youth	0.986
weather	wilt	0.959
expression	poker face	0.917
shadows	venue	0.914
time	youth	0.911
colors	yellow jack	0.910
doing	run	0.906
clocks	texas storksbill	0.884
shadow	venue	0.878
brick	wattle and daub	0.873

Table 6.21: Same Deletions

Initial word	Same Answer Percentage
fire	0.908
picture	0.907
guys	0.891
pic	0.888
control	0.878
man's	0.878
photo	0.877
surfers	0.874
thing	0.873
animal	0.870

Table 6.22: Different Deletions

Initial word	Different Answer Percentage
season	0.985
flooring	0.904
time	0.899
expression	0.899
country	0.889
year	0.874
weather	0.864
gravel	0.860
shadows	0.854
cut	0.848

6.9.2 Wordclouds

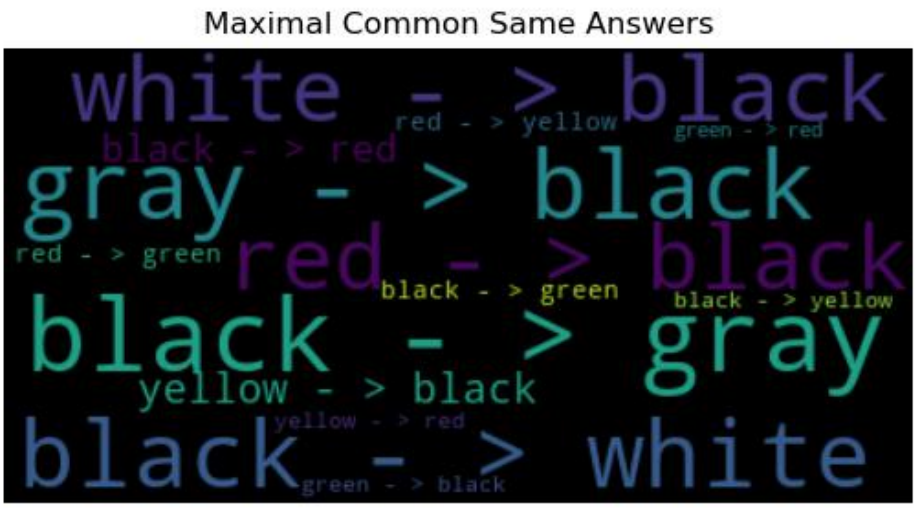


Figure 6.9: Maximal Common Same

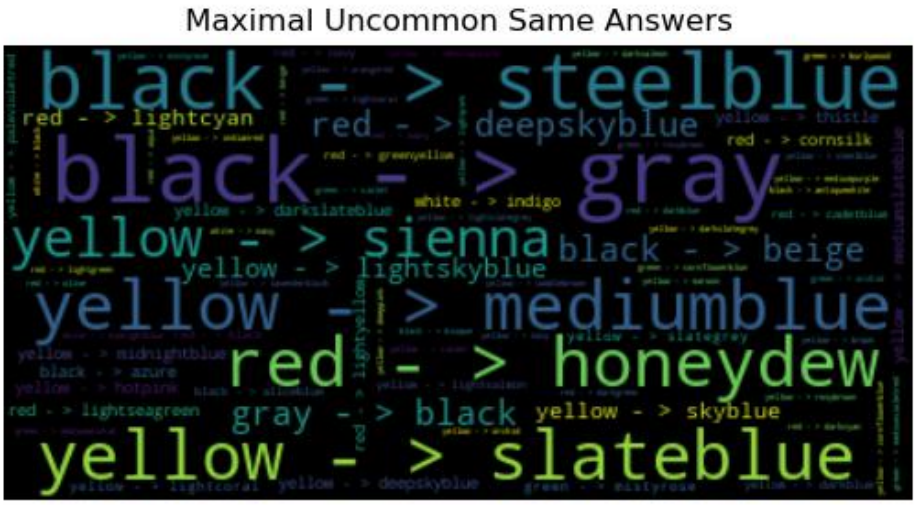


Figure 6.10: Maximal Uncommon Same

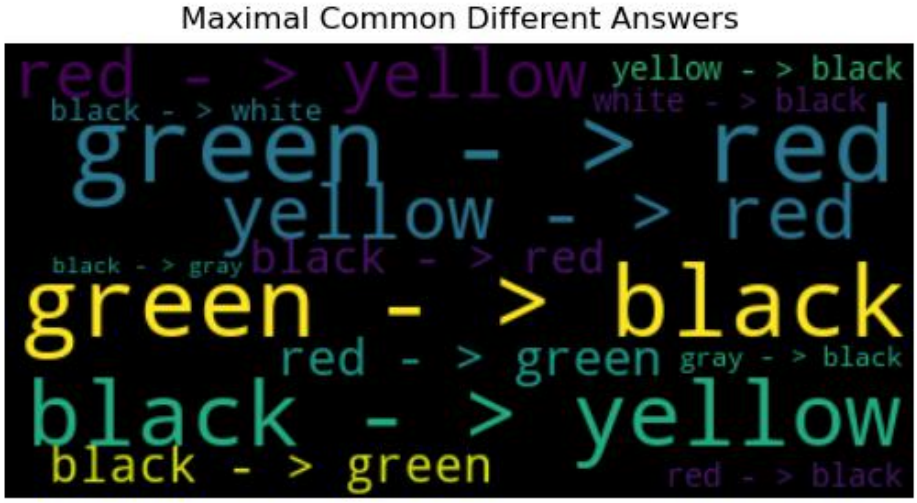


Figure 6.11: Maximal Common Different

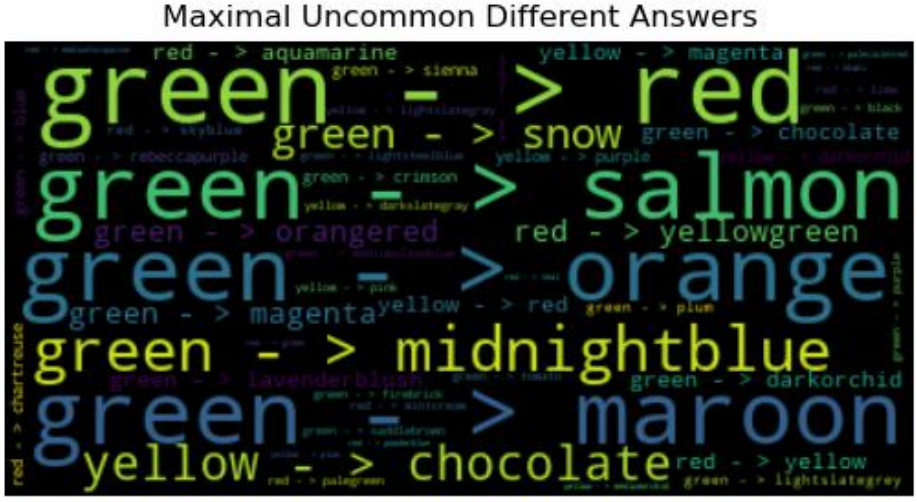


Figure 6.12: Maximal Uncommon Different

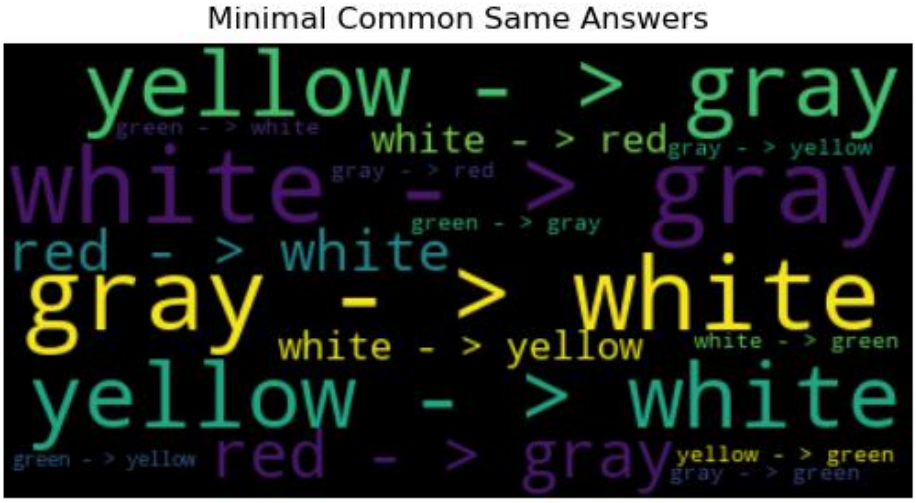


Figure 6.13: Minimal Common Same

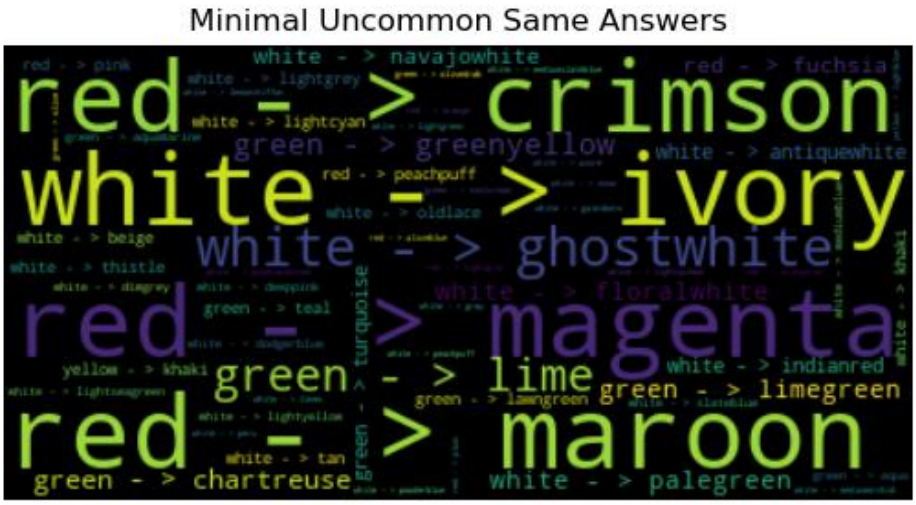


Figure 6.14: Minimal Uncommon Same

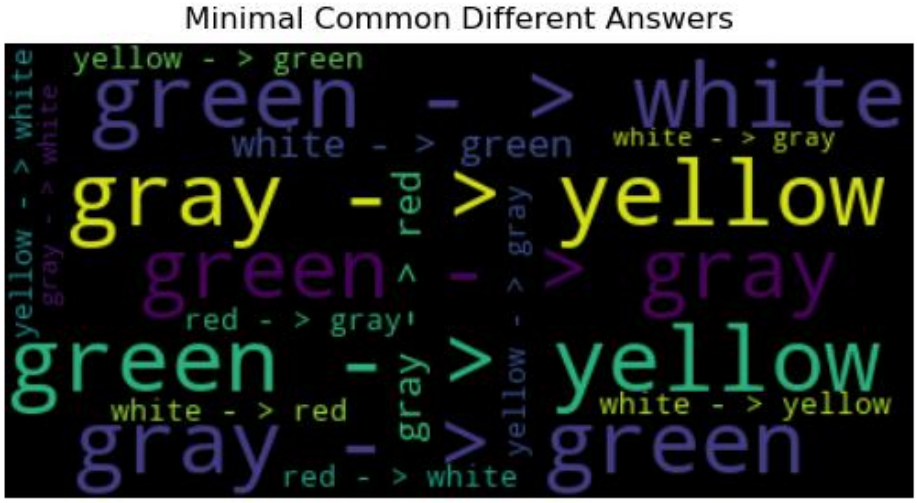


Figure 6.15: Minimal Common Different

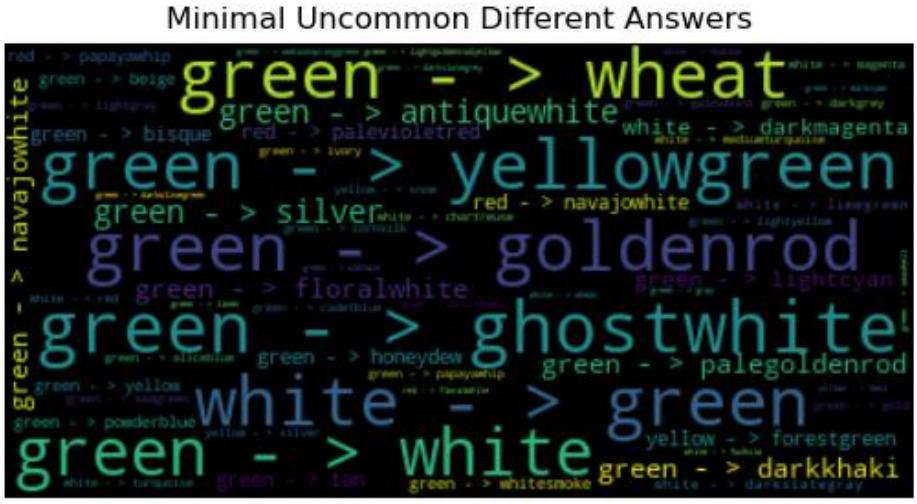


Figure 6.16: Minimal Uncommon Different

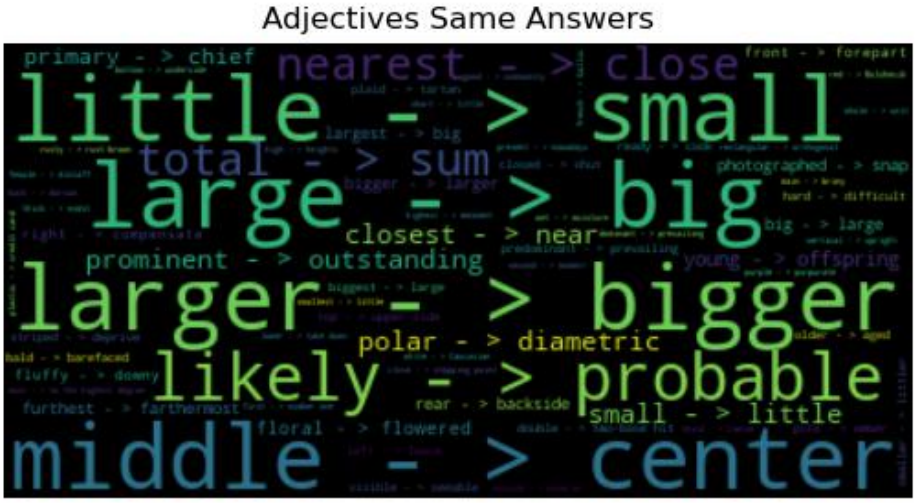


Figure 6.17: Adjectives Same

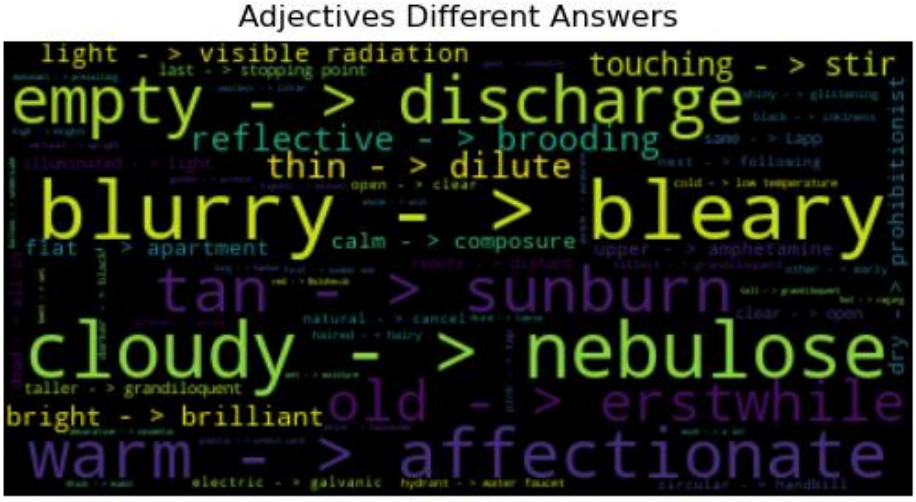


Figure 6.18: Adjectives Different

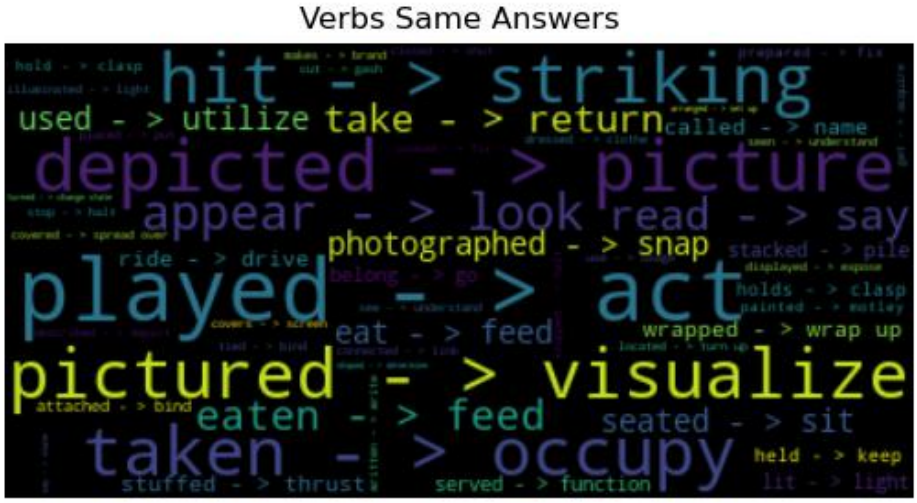


Figure 6.19: Verbs Same

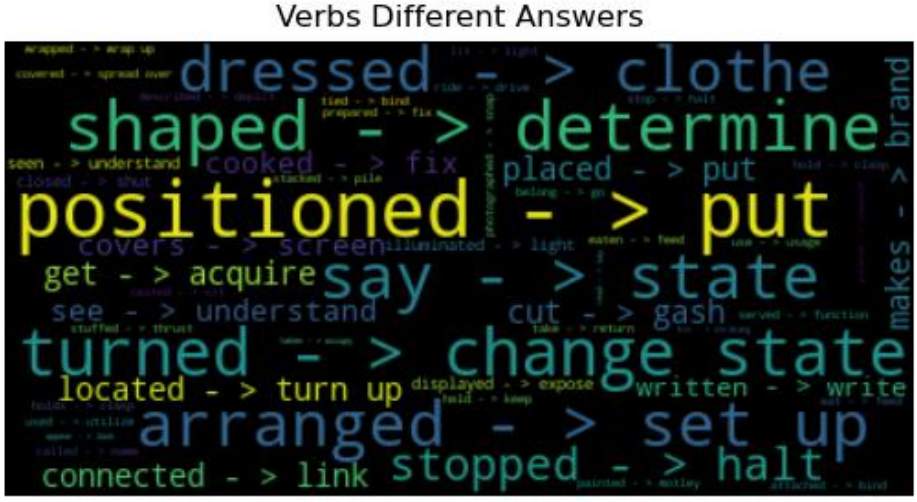


Figure 6.20: Verbs Different

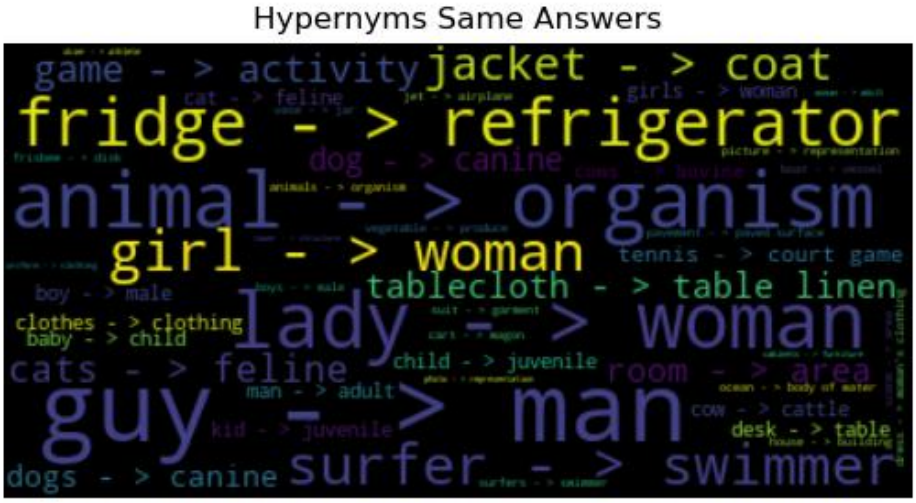


Figure 6.21: Hypernyms Same

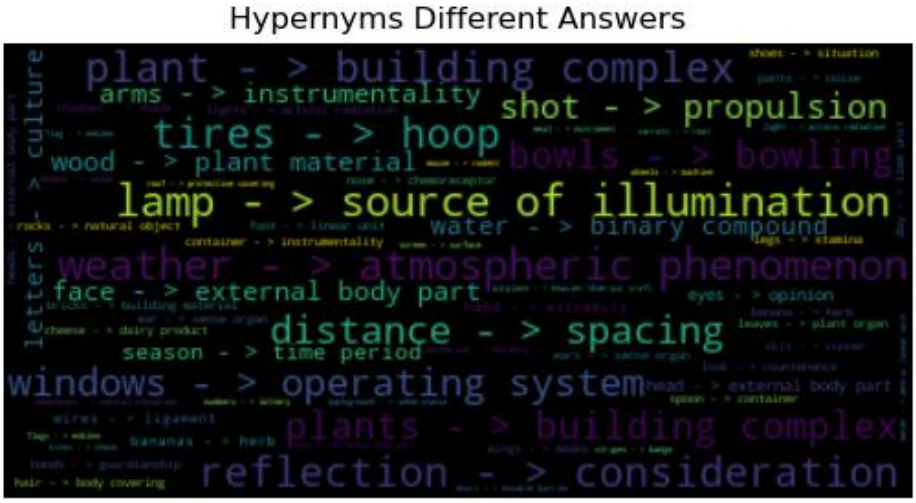


Figure 6.22: Hypernyms Different

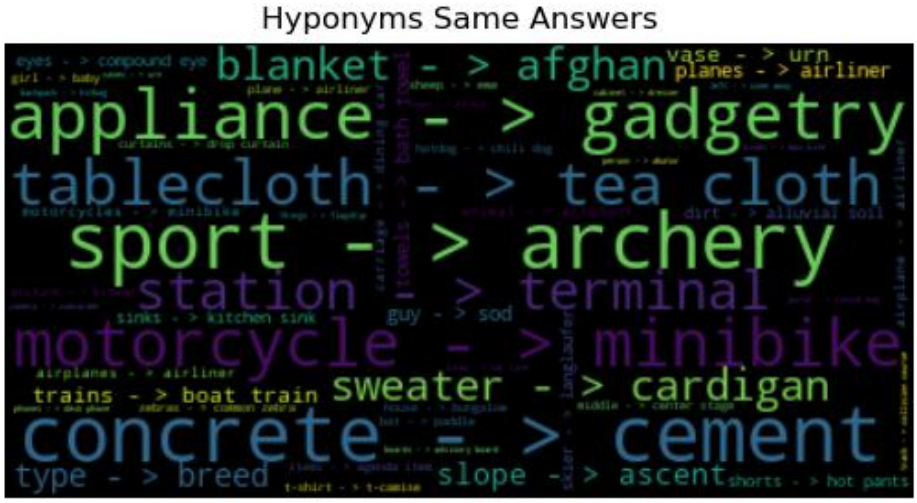


Figure 6.23: Hyponyms Same

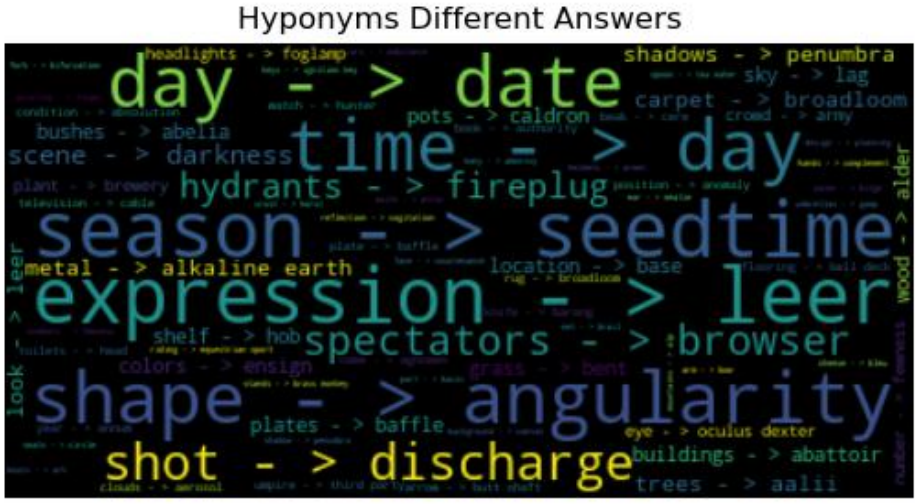


Figure 6.24: Hyponyms Different

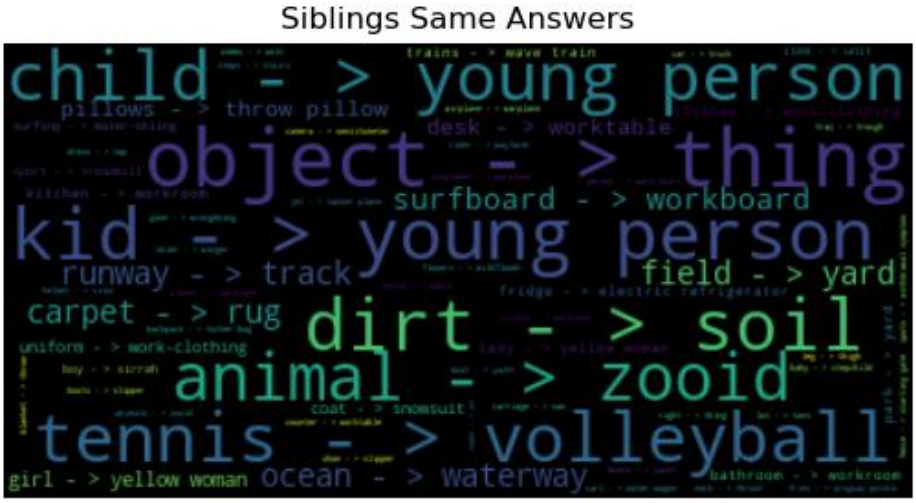


Figure 6.25: Siblings Same

Figure 6.26: Siblings Different

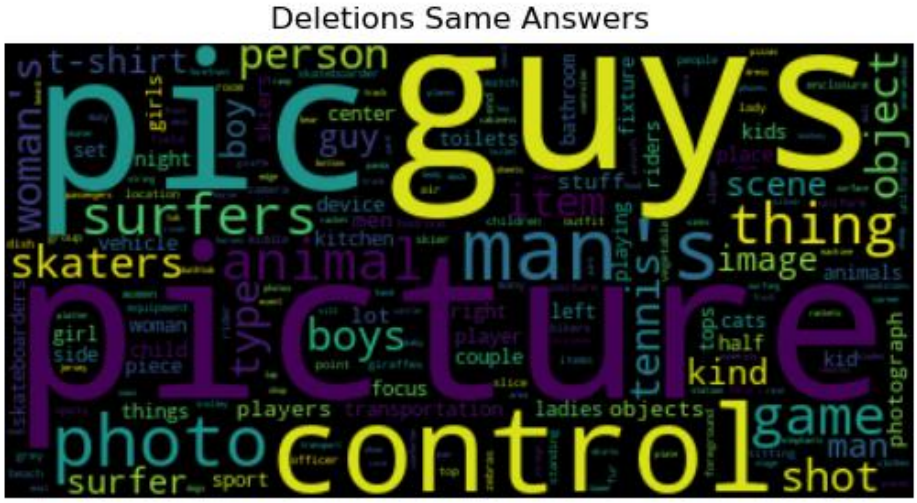


Figure 6.27: Deletions Same

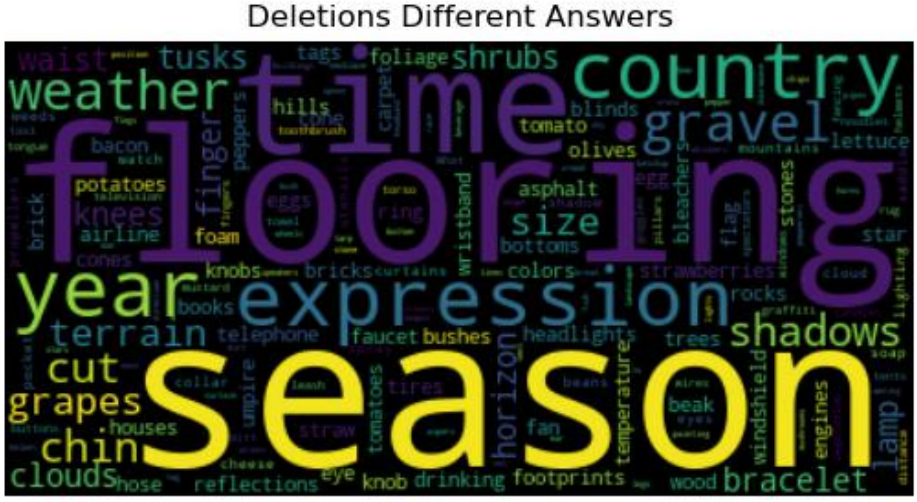


Figure 6.28: Deletions Different

Chapter 7

Conclusion and Future Work

Counterfactual perturbations in VQA models can provide novel and useful insights regarding model robustness and explainability of results. In our work, we propose a knowledge-based counterfactual framework targeting substitutions on questions. Specifically, our framework suggests multiple types of word-level linguistic transformations to probe selected VQA models in a black-box fashion and investigate whether the presence of counterfactual questions will lead to unexpected model responses. Through this process, underlying linguistic biases are revealed, while informative explanations regarding the model's behavior are provided, by qualitatively extracting global rules, ultimately depicting those existing biases. Our results on Visual Genome and VQA-v2 datasets, using ViLT model as proof of concept, illustrate the merits of our approach, highlighting concepts that incite model biases, in a model-agnostic manner.

We design our framework by deploying a generalizable set of counterfactual adversarial transformations, meaning that it can properly be applied to any VQA model. This property of our proposed method is reassured through the black-box-oriented approach we follow regarding perturbating the model in question: we do not attempt to penetrate each VQA model's architecture, design choices, or parameters, nor do we recruit procedures that investigate the inner reasoning process of machine learning models. Instead, we probe each model by applying various types of counterfactual perturbations and extensively observing the produced outputs. Hence, the insights are drawn by analyzing the model's response, rather than delving into its inner structure. Moreover, our proposed counterfactual perturbations are fully guided by the utilization of knowledge-based sources, that control and configure the multiple experiments we have implemented. They also guarantee the optimal transformations that we propose, as well as reassure the deterministic aspect of the substitution choices we introduce.

A proper comparison of initial and counterfactual outputs, as well as the insight provided by the accuracy variance of the experiments, showcases the robustness extent of the model. Particularly, this study lies in assessing whether the model presents the adequate agility to qualitatively adapt to minor or major changes in the input, and consequently note identical or closely similar accuracy as recorded in the performance on the initial dataset.

The aforementioned robustness-related research field that we approach, is closely correlated to the investigation and discovery of underlying existing biases in the model in question. Close observation of the produced result to counterfactual questions in opposition to the original results to the initial dataset lead to valuable local explanations that depict specific cases of unexpected model responses that we point out. Subsequently, a comprehensive aggregation of those local observations leads to the extraction of global rules that fully describe and characterize the model's general behavior. These global explanations express the detected biases that have arisen from the application of our multilevel perturbations.

The quintessence of our presented method lies in the configuration of insightful and targeted explanations that justify the model's response and shed light on the reasoning process followed by the model's combination of components. Through this major benefit, our method is proven to increase transparency and improve the reliability level of the model in question, as it provides insights that can be seriously considered when assessing the trust one should pay to the decisions extracted by

it. This new point of view that we offer successfully addresses the initially introduced problematic issue that originally motivated our research, i.e. the unwanted opacity that inevitably accompanies the obscure decision-making process followed by the internal reasoning of many machine learning models, performing visiolinguistic tasks, and specifically Visual Question Answering, as explored in our work.

Following the extensive presentation of this thesis work, we would like to suggest some promising future research directions regarding the examined topics, that we consider to be fruitful for further elaboration. Firstly, as an immediate extension of our method, we propose the application of the same linguistic perturbations on dataset answers, addressing VQA models that reason over multiple choice answers. Additionally, we recommend the extension of our approach to other related visiolinguistic tasks, such as Text - Image Retrieval, Visual Entailment, and Visual Commonsense Reasoning, as we believe it has the potential to provide valuable insights and attributes both efficient and innovative explainability aspects to such popular contemporary machine learning tasks with a great range of influence. Finally, another worth-mentioning proposed direction involves crafting counterfactual perturbations targeting the visual modality, both directed towards pixel-level and conceptual substitutions. An example of such promising extensions of our analysis could include attempting vision perturbation through image synthesis with the use of diffusion models and prompting through Large Language Models. Such kinds of transformations can be rendered useful in revealing visual biases and thus exploring a parallel side of the explainability in the visiolinguistic tasks field.

Bibliography

[1] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. van den Hengel, "Counterfactual vision and language learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 041–10 051.

[2] V. Agarwal, R. Shetty, and M. Fritz, "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing," 2019. [Online]. Available: https://arxiv.org/abs/1912.07538

[3] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," 2016.

[4] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," 2018.

[5] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," 2015. [Online]. Available: https://arxiv.org/abs/1505.00468

[6] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, and G. Burachas, "A study on multimodal and interactive explanations for visual question answering," 2020. [Online]. Available: https://arxiv.org/abs/2003.00431

[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2017. [Online]. Available: https://arxiv.org/abs/1707.07998

[8] K. Baker, A. Parekh, A. Fabre, A. Addlesee, R. Kruiper, and O. Lemon, "The spoon is in the sink: Assisting visually impaired people in the kitchen," in *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*. Gothenburg, Sweden: Association for Computational Linguistics, Oct. 2021, pp. 32–39. [Online]. Available: https://aclanthology.org/2021.reinact-1.5

[9] M. Banchhor and P. Singh, "A survey on visual question answering," in *2021 2nd Global Conference for Advancement in Technology (GCAT)*, 2021, pp. 1–5.

[10] H. Ben-Younes, Ã‰loi Zablocki, P. PÃ©rez, and M. Cord, "Driving behavior explanation with multi-level fusion," *Pattern Recognition*, vol. 123, p. 108421, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321005975

[11] Z. Boukhers, T. Hartmann, and J. Jürjens, "Coin: Counterfactual image generation for vqa interpretation," 2022. [Online]. Available: https://arxiv.org/abs/2201.03342

[12] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases in visual question answering," 2019. [Online]. Available: https://arxiv.org/abs/1906.10169

[13] C. Chen, S. Anjum, and D. Gurari, "Grounding answers for visual questions asked by visually impaired people," 2022. [Online]. Available: https://arxiv.org/abs/2202.01993

[14] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," 2022.

[15] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," 2020. [Online]. Available: https://arxiv.org/abs/2003.06576

[16] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples synthesizing and training for robust visual question answering," 2021. [Online]. Available: https://arxiv.org/abs/2110.01013

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[19] Y. Du, Z. Liu, J. Li, and W. Zhao, "A survey of vision-language pre-trained models," 02 2022.

[20] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," 2022.

[21] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1–1, 2021. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2021.3080920

[22] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction," 2019.

[23] C. Fellbaum, "Wordnet: An electronic lexical database," 1998.

[24] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," 2015.

[25] S. Garg and G. Ramakrishnan, "BAE: BERT-based adversarial examples for text classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6174–6181. [Online]. Available: https://aclanthology.org/2020.emnlp-main.498

[26] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," 2016. [Online]. Available: https://arxiv.org/abs/1612.00837

[27] J.-H. Huang, M. Alfadly, B. Ghanem, and M. Worring, "Assessing the robustness of visual question answering models," 2019. [Online]. Available: https://arxiv.org/abs/1912.01452

[28] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021.

[29] M. Jiang, S. Chen, J. Yang, and Q. Zhao, "Fantastic answers and where to find them: Immersive question-directed visual attention," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2977–2986.

[30] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4976–4984.

[31] ——, "An analysis of visual question answering algorithms," 2017.

[32] ——, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, oct 2017. [Online]. Available: https://doi.org/10.1016%2Fj.cviu.2017.06.005

[33] ——, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, oct 2017. [Online]. Available: https://doi.org/10.1016%2Fj.cviu.2017.06.005

[34] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2748–2754. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.234

[35] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," 2017. [Online]. Available: https://arxiv.org/abs/1704.03162

[36] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," 2021. [Online]. Available: https://arxiv.org/abs/2102.03334

[37] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016. [Online]. Available: https://arxiv.org/abs/1602.07332

[38] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," 2021. [Online]. Available: https://arxiv.org/abs/2107.07651

[39] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," 2019. [Online]. Available: https://arxiv.org/abs/1908.03557

[40] Z. Liang, W. Jiang, H. Hu, and J. Zhu, "Learning to contrast the counterfactual samples for robust visual question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3285–3292. [Online]. Available: https://aclanthology.org/2020.emnlp-main.265

[41] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019. [Online]. Available: https://arxiv.org/abs/1908.02265

[42] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," 2016. [Online]. Available: https://arxiv.org/abs/1606.00061

[43] M. Lymperaiou, G. Manoliadis, O. M. Mastromichalakis, E. G. Dervakos, and G. Stamou, "Towards explainable evaluation of language models on the semantic similarity of visual concepts," 2022. [Online]. Available: https://arxiv.org/abs/2209.03723

[44] M. Lymperaiou and G. Stamou, "A survey on knowledge-enhanced multimodal learning," *ArXiv*, vol. abs/2211.12328, 2022.

[45] ——, "The contribution of knowledge in visiolinguistic learning: A survey on tasks and challenges," 2023.

[46] ——, "A survey on knowledge-enhanced multimodal learning," 2023.

[47] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," 2015.

[48] A. Mogadala, M. Kalimuthu, and D. Klakow, "Trends in integration of vision and language research: A survey of tasks, datasets, and methods," *Journal of Artificial Intelligence Research*, vol. 71, p. 1183–1317, Aug 2021. [Online]. Available: http://dx.doi.org/10.1613/jair.1.11688

[49] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," 2020. [Online]. Available: https://arxiv.org/abs/2005.05909

[50] A. Panesar, F. I. Doğan, and I. Leite, "Improving visual question answering by leveraging depth and adapting explainability," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 252–259.

[51] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649.

[52] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," 2020.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[54] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," 2015.

[55] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1085–1097. [Online]. Available: https://aclanthology.org/P19-1103

[56] F. Sammani, T. Mukherjee, and N. Deligiannis, "Nlx-gpt: A model for natural language explanations in vision and vision-language tasks," 2022. [Online]. Available: https://arxiv.org/abs/2203.05081

[57] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.

[58] H. Sharma and A. S. Jalal, "A survey of methods, datasets and evaluation metrics for visual question answering," *Image and Vision Computing*, vol. 116, p. 104327, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885621002328

[59] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," 2021. [Online]. Available: https://arxiv.org/abs/2112.04482

[60] T. Stoikou, M. Lymperaiou, and G. Stamou, "Knowledge-based counterfactual queries for visual question answering," 2023.

[61] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," 2019.

[62] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2021.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[64] J. Wan, J. Yang, S. Ma, D. Zhang, W. Zhang, Y. Yu, and Z. Li, "Paeg: Phrase-level adversarial example generation for neural machine translation," 2022. [Online]. Available: https://arxiv.org/abs/2201.02009

[65] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," 2019. [Online]. Available: https://arxiv.org/abs/1901.11196

[66] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," 2016.

[67] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," 2019.

[68] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," 2016.

[69] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," 06 2016, pp. 4995–5004.

[70] Y. Zou and Q. Xie, "A survey on VQA: Datasets and approaches," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. IEEE, dec 2020. [Online]. Available: https://doi.org/10.1109%2Fitca52113.2020.00069

[71] ——, "A survey on VQA: Datasets and approaches," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. IEEE, dec 2020. [Online]. Available: https://doi.org/10.1109%2Fitca52113.2020.00069