Ben-Gurion University of the Negev
The Faculty of Natural Sciences
The Department of Computer Science

Detecting Irrelevant Questions in VQA via Presupposition Augmentation

Thesis submitted in partial fulfillment of the requirements
for the Master of Sciences degree

Nitzan Cohen

Under the supervision of **Prof. Michael Elhadad**

March 2023

Ben-Gurion University of the Negev
The Faculty of Natural Sciences
The Department of Computer Science

Detecting Irrelevant Questions in VQA via Presupposition Augmentation

Thesis submitted in partial fulfillment of the requirements
for the Master of Sciences degree

Nitzan Cohen

Under the supervision of **Prof. Michael Elhadad**

Signature of student: _____ Date: _____
Signature of supervisor: _____ Date: _____
Signature of chairperson of the
committee for graduate studies: _____ Date: _____

March 2023

Detecting Irrelevant Questions in VQA via Presupposition Augmentation

Nitzan Cohen

Master of Sciences Thesis

Ben-Gurion University of the Negev

2023

Abstract

In this work, we study the relevance of a question in the context of an image. The setting is given an image, and a question in natural language, we want to assess whether the question is relevant to the image. For example, given an image that shows a dog running on a beach, a question asking about the location of a washing machine in the kitchen should be flagged as irrelevant. We define relevance from a presupposition perspective and apply relevance theory to model the task of Visual Question Answering (VQA) in conversational settings. We provide a manual analysis of the challenging aspects of questions in GQA (a standard dataset used to benchmark progress on the VQA task) and categorize them into eight different classes.

We introduce a method to validate whether a presupposition is holding in a given context by analyzing the scene graph associated with the image, while addressing arising pragmatic and semantic phenomena, scoring 99.29% F1 score on a test set we developed to model the task.

We introduce a new data set of irrelevant questions which specify for each question the reason for the intended lack of relevance. The data set contains five types of failures, each of which is implemented in different predefined templates and applied to existing questions from GQA.

We leverage the generated data set to demonstrate the capability of detecting irrelevant questions by training an available pre-trained vision and language model to detect irrelevant questions. This method detected 2% of the GQA Test-dev split as actually irrelevant by using our detector on the original GQA data. Manual verification confirmed that this classification was correct with 71% precision. The data set, code and pre-trained

weights are available online. [1]

[1] https://github.com/NitCoh/Detecting-Irrelevant-Questions-in-VQA-via-Presupposition-Augmentation

Acknowledgements

I would like to thank to my advisor, Prof. Michael Elhadad for the support. I had a great time and opportunity to work with knowledgeable, patient and determined advisor. Thanks to Yarin Kuper and the rest of the NLP group of CS BGU.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The use of multiple input modalities (text, image, audio, video) has gained attention in recent years. Specifically, tasks involving linguistic and vision information include image-captioning, visual entailment, visual question answering, image-text retrieval, and phrase grounding.

The recent breakthroughs in deep artificial neural network architectures and the publishing of large-scale data sets with detailed annotation allowed rapid progression in this emerging field. Massive research and development of large pre-trained models functions as a proxy to solve different vision and language tasks.

The multi-modal tasks also require multiple cognitive abilities (reasoning, grounding, and understanding of compositional terms) for a model to achieve success. For vision-language data, tasks involve describing relations between visual objects, retrieving images based on text, answering a complex question regarding an image and many more. From these tasks, many practical applications are derived, for example, question-based assistance for blind people, image-text retrieval for search engines, textually summarizing visual scenes and many more.

In this work, we focus on the task of visual question answering (VQA), where the input consists of an image and a text-based question about the image, and the expected output is a textual answer to the question that relies on the content of the image.

The task of VQA requires a deep understanding of what is asked and what is depicted in the given scene. VQA systems require a diverse set of skills, among them detecting objects depicted in an image, relating lin-

guistic elements to objects shown in the image, counting objects, relating attributes to objects, determining whether a relation holds among objects, and in many cases common-sense reasoning to infer unseen facts or relations based on visible objects.

Recent advances in deep learning improved both NLP and Vision domains, ranging from new architectures to model complex relations, such as encoder-decoder and transformer architectures, and large-scale challenging VQA data sets such as VQA (1.0/2.0) [7] and CLEVR [8] and more compositional question data sets such as GQA [2].

Existing large-scale data sets contain questions that are all relevant with respect to a given image, i.e., questions that can be answered given the image, which functions as the context. When used as training data, these data sets result in VQA systems that rely on the assumption that a question asked is relevant to the image. Such systems provide an answer even when the question cannot be answered given the context. For example, asking "What is the color of the cat on the mat?", when there is no cat in the context. Moreover, some of the large scale datasets are synthetically generated using various methods (e.g., GQA is built from questions generated automatically from scene graph information). Based on manual inspection, we suspect mistakes are present when associating a question with an image.

From a wider perspective, real-world VQA systems must have the functionality to distinguish between answerable (**relevant**) and non-answerable (**irrelevant**) questions given the image. Otherwise, challenging users can ask many different questions, such as commonsense or general knowledge questions - "Who is the prime minister of Israel?" and receive unreliable answers instead of a more trustworthy "I don't know" answer.

In this work, we focus on detecting irrelevant questions by generating infelicitous questions, where their failure comes from a failed presupposition with respect to the context. Specifically, we augment true presuppositions in the context and falsify them with respect to the context. For example, the question "What is the color of the dog on the mat?" presupposes that the image depicts a dog, a mat and that the dog is sitting on the mat. If the context contains all the specified entities and relations, we say the question is answerable. However, if the context does not contain any mat, the existential assumption about a mat is wrong (given the context), hence the question is unanswerable.

During our work, via a manual annotation process, we discovered that there are challenging questions in GQA and that human annotators fail

to agree if the questions are even answerable. We classified the classes of disagreement between annotators derived from challenging questions where annotators disagree and used this observation data to define failure types. Given our list of failure types, we design a synthetic method to generate infelicitous questions.

With five different types of failed presupposition, we implement a pipeline that automatically generates irrelevant questions of different levels of difficulty given an image and a scene graph describing the content of the image. We use this pipeline to build a data set of irrelevant questions that use to train a classifier of questions as either relevant or irrelevant, and for irrelevant questions, the classifier provides a reason for the failure.

Our key contributions are:

1. We provide a manual analysis of the challenging aspects of different questions in GQA and categorize the types into eight different classes of difficulty.

2. We developed a novel distance metric to compare presupposition to the given context via Scene Graphs relying on the notion of accommodation process from Relevance Theory.

3. We provide a new data set of irrelevant questions that contain 100k generated questions of different types.

4. We train an irrelevant question detector and demonstrate its performance on the GQA Testdev split. We identify that about 2% of the questions in this standard data set are actually irrelevant.

Chapter 2

Related Work

In this chapter, we present the standard datasets that are used to research the task of Visual Question Answering (VQA). We then survey the techniques that have been developed to model this task and conclude with a discussion of the various metrics and evaluation methods developed to assess the success of models in VQA.

2.1 Data

Data sets in the domain of image-language analysis include large collections of images associated with different forms of language. For example, in image captioning datasets, each image is associated to a short textual caption (usually a single line of 5 to 10 words). In visual question answering data sets, for each image, there may be a set of question-answer pairs that relate to the content of the image. In addition, some of the data sets contain an additional element called a "scene graph" which includes a structural representation of the set of objects depicted in the image, with a list of attributes for each object and relation that hold among the objects. GPA utilizes the Scene Graph information to build questions about the images and finally generate data set composed of images, questions and answers.

2.1.1 Visual Genome

Visual Genome (VG) [1] is a dataset which collects images with dense annotation of objects, attributes and relationships. It encodes the annotation

in an object called Scene Graph [9]. Objects are associated to regions in the image (characterized by a bounding box) and represented by a sub-graph. The union of all sub-graphs is a scene graph describing the image. For example, a region containing a dog running on grass will be represented by the functional expression *running_on(dog, grass)* and the referenced scene graph will contain two nodes, *dog* and *grass* and a labeled edge labeled *running_on* connecting both.

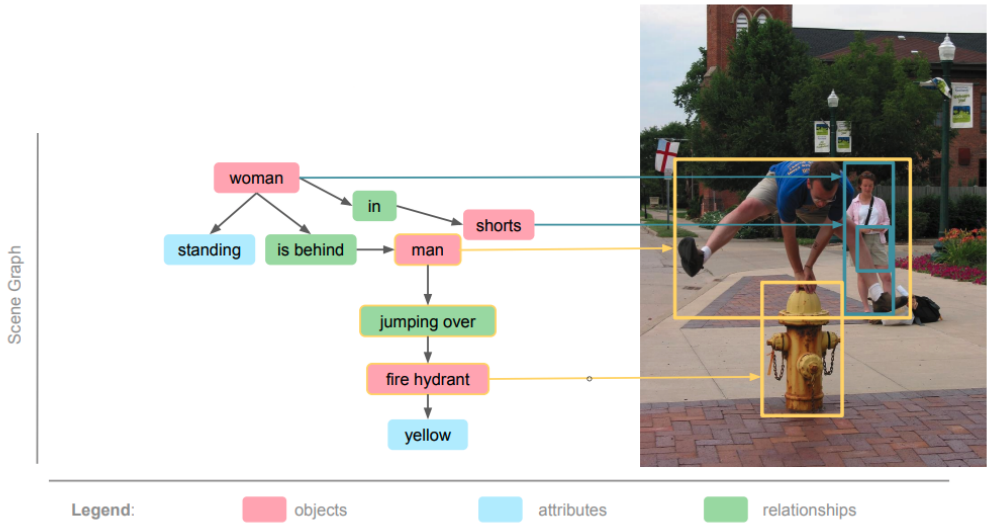


Figure 2.1: Scene Graph from [1]. Each object (in pink) refers to a region in the image, relationships connecting objects appear in green, and attributes appear in blue.

VG contains about 100K images, where each image annotation includes an average of 21 objects, 18 attributes and 18 pairwise relationships between objects. The set of annotated relationships contains 13,894 unique relationships grouped into categories: actions, verbs, prepositions, comparative and prepositional phrases.

This dataset is a core resource for solving cognitive tasks on images that require reasoning abilities. Additionally, it is a building block for other datasets, such as GQA, in order to generate compositional questions.

2.1.2 GQA

GQA [2] is a large scale, real-world dataset which uses Visual Genome scene graphs in order to generate visual reasoning and compositional questions. To build this dataset, the authors developed a question engine which traverses scene graphs and composes, through templates, pairs of question and answer.

The question engine utilizes 250 manually constructed templates and an additional 274 templates extracted from VQA 1.0 [10] for a total of 524 patterns. GQA is an important milestone as it is among the first dataset to address compositional generalization from real word instances at such scale. The dataset contains 22M diverse reasoning questions generated from 113K real-word images.

Each question is accompanied with an image, a functional program to represent its semantics and guide the reasoning steps required to answer, an answer. Questions are grouped in semantic types - for example, *e.g., "Do you see any boys?"* is an existence question. Additionally, each question has a list of entailed questions, i.e., if one can answer the question it should consistently be able to answer all the entailed questions.

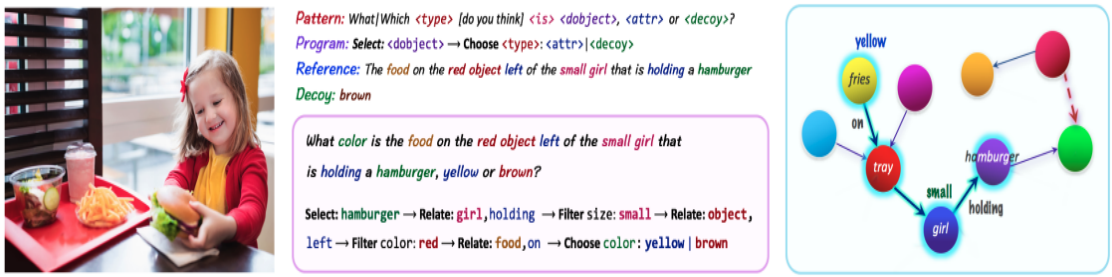


Figure 2.2: Example from GQA [2]. **Pattern** is the template which is used to create the question; **Program** is the functional program pattern which represents its semantics; **Reference** is a reference phrase populated to generate the question.

Objects and relations are controlled in the data set so that they are distributed uniformly across different semantic types. This prevents unwanted effects where a model could predict the answer to questions by relying on the frequency of objects in the answer instead of reasoning about the specific aspects of a given scene. The data set focuses on four types of relations: spatial, descriptive verbs, comparative and prepositional phrases.

In addition to objects and relations, each node in the scene graph has additional attributes such as color (red), states (standing, walking), size (small), materials (wood). There are 13,041 unique attributes in VG which allow more detailed descriptions about objects and can help for disambiguation when an image depicts multiple instances of the same object type (e.g., several persons in one image). The diversity of the information in a scene-graph is a fertile soil to compose multiple types of question-answer pairs. The process of traversing scene graphs allows the engine to generate nested compositional questions, *e.g., "Is the tree in front of the building that is made of stone?"* . However, this can lead to very unbalanced version (in terms of question groups and different answers) data set. Thus, a balanced version of GQA exist with 1.7M questions and 1,878 possible answers.

As of 2022, Coarse-to-Fine Reasoning [11] achieves the best results on the test-dev subset of GQA, with 72.1% accuracy, followed by *Neural State Machine* [12] with 62.95% and LXMERT [13] with 60%.

2.1.3 VQA

Visual Question Answering (VQA) [7] is a visual question answering data set and is among the most widely used data sets in the field. Its latest version *VQA 2.0* contains 204,721 images taken from the MS-COCO dataset [14], 50,000 abstract scenes, 1,105,904 questions and 11,059,040 ground truth answers (each question includes multiple valid answers). The data set was collected using human workers who generated both questions and answers given an image. Three questions from distinct workers are gathered for each image or abstract scene. The interface which guides the worker instructed them to ask questions that require consulting the image in order to be answered correctly.

The first version *VQA 1.0* suffered from unimodal text bias, and it was found that a question could be answered without looking at the image, but only by relying on correlations found within the data set. For example, because most bananas are *yellow*, a model does not need to utilize information from the image in order to achieve high accuracy when answering the question *"What is the color of the banana?"*. Simply by answering always *yellow* it will reach high accuracy.

In [3], the authors took steps to reduce such bias by providing for each question a complementary image which changes the answer. One way to quantify such unimodal bias is to train a unimodal model (a model that does not inspect the images) and check its accuracy, as stated by [15]. Even

in VQA2, a text-only model was able to predict the correct answer with 44% accuracy over the test set.

Figure 2.3: Examples from VQA 2.0 [3], where each question has the original image, a complementary image and the answer is shifted when referring for each of them.

2.2 The Visual Question Answering Task

Visual Question Answering as a task was first introduced in [10]. The goal is to answer a question utilizing an image as the source of information and to produce a natural-language answer. The input is a question as a free-form text and a plain image. The two dominant formulations of VQA are either as a classification task (choose from a fixed number of possible answers) or as a free-form answer generation task, where a system must generate an answer in natural language [16]. The former is frequently used due to its simplicity.

VQA requires from a system to develop diverse set of skills in order to correctly answer different questions and generalize to other domains: *e.g.*, common-sense reasoning, visio-linguistic compositional reasoning, counting objects, relating objects to attributes, determining whether a relation holds among objects and more.

2.3 VQA Models

Rapid progress in deep learning pushed single modalities models in computer vision and natural language processing to achieve human performance and even surpass it on different tasks. Multi-modal models are challenging because they need a mechanism to fuse information from each modality. For example, a VQA system needs to be able to ground objects and relationships from an image, understand a given question and relate the objects and relationships found within the question to the image.

Such models are referred to as *joint embedding models*. The fusion mechanism between modalities is often addressed as the alignment between representations of the given modalities. Existing multi-modal encoders address the alignment in their proposed architectures via different approaches [17], [18], [19] and [12].

In addition to joint embeddings models, in GQA, questions are parsed into functional programs that describe the steps required to get to the answer given a description of the scene graph depicted in the image. The transformation from text to programs allows us to model the reasoning steps required to get an answer. The functional program is composed from basic functional blocks such as *select*, *relate*, *query*, etc. Those methods can generalize to compositional questions requiring many reasoning steps to answer and also provide interpretability.

In this section we describe few of the approaches used in the vision and language domains to fuse the information. We then describe approaches to encode Vision and Language modalities and solve the VQA task.

2.3.1 Attention

Attention was introduced to assist the decoder of Neural Machine Translation models to pay attention to the most relevant parts in the encoded sequence at each time step. Generally, attention mechanisms allow us to model dependencies between a token and a given sequence. Specifically, an attention function maps a query and a set of key-value pairs to an output by computing a weighted sum over a compatibility function of the vector representing the token to all other vectors in the sequence. Self-attention allows modeling dependencies between each token in the sequence to all tokens in the sequence and it was first used as a core sequence modeling mechanism as part of the Transformer [20] architecture.

The attention formula is given by:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

Where queries are packed into the Q matrix and key-value pairs are packed into the K and V matrices and d_k is a scaling factor determined by the dimension of queries and keys.

Transformer is one of the most widely used architecture in the Vision and Language multi modal settings. A transformer is used to encode the textual information from the questions. It is made of a stack of multiple self-attention layers (usually 6 to 12 layers) and a multi-head attention mechanism (that is, multiple attentions are computed for each layer in parallel).

In order to fuse the information from the vision modality and the textual modality, Cross Attention was introduced as a modeling mechanism in the Vision+Language encoders. While Self-Attention is performed on queries, keys and values generated from the same modality, Cross Attention is performed on queries generated from one modality and key-value pairs generated from the other modality.

2.3.2 Vision and Language Encoders and Pre-Trained Models

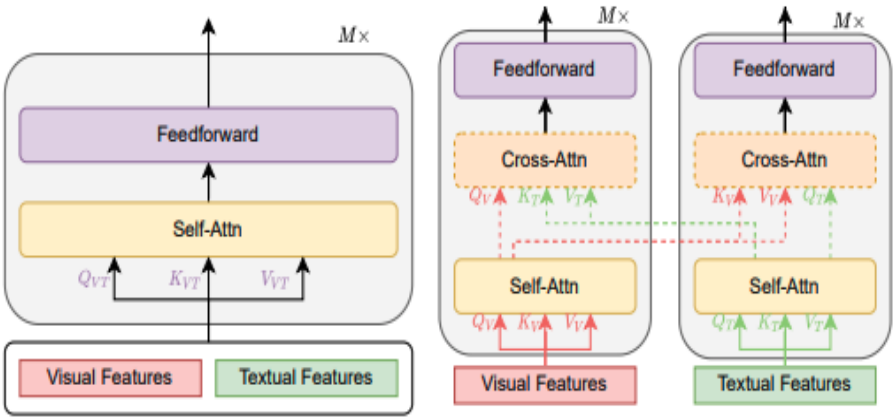
Vision and Language (V+L) encoders form the backbone of most VQA systems based on the joint embedding approach. In order to complete the task of answering a question based on an image, the joint embedding approach aims to encode the visual information and the textual information and align the representations together, then answer the question based on the fused information in an encoder-decoder manner.

V+L encoders provide the basis of this requirement. There are two dominant architectural approaches in V+L encoders, Uni-Stream and Dual-Stream. In Uni-stream, the visual and textual features are concatenated together and fed to the Attention layer, thus allowing each token to attend to both modalities. In contrast, the Dual-Stream approach utilizes Cross Attention, thus allowing each modality to attend to the other modality independently.

Additionally, V+L encoders align the modalities with shallow or deep ap-

proaches. In the shallow approach (e.g., CLIP [21]), representations are aligned by computing a compatibility function between the modalities only in the final layer. In the deep approach (e.g., LXMERT [13]), representations are aligned in all cross-modality layers.

Different encoders integrate recent breakthroughs in uni-modal fields to primarily encode the image using uni-modal feature extractors and text and then pass it to the cross-modality layers. For example, [10] used CNNs to encode the image and LSTMs to encode the text. In contrast, LXMERT uses a region-based approach utilizing an object-relationship encoder to model relationships between objects in the image and a Transformer to encode the text.



(a) Single-Stream Architecture (b) Dual-Stream Architecture

Figure 2.4: Illustration of the two architectural types for Vision and Language Encoders from [4].

2.3.2.1 Uni-Stream

The Uni-Stream architecture is more parameter efficient, since both modalities are concatenated together, then fed into a single Transformer block which causes the same set of parameters to be used for both modalities. Examples of this approach include *UNITER* [22], *VinVL* [23] and *ViLT* [24].

2.3.2.2 Dual-Stream

In Dual-Stream architecture each modality is sent to a different Transformer block and Cross Attention is utilized to allow each modality to attend to the other. Since the two blocks are not sharing parameters, the approach is less parameter efficient. Examples include *LXMERT* [13] and *CLIP* [21].

2.3.2.3 Pre Trained Models

The effectiveness of a large pre-training step has been demonstrated in BERT [25] and GPT [26], where both models were able to achieve SOTA results on many NLP tasks, such as *GLUE* [27] and SQuAD [28]. In order to learn contextual syntactic and semantic properties of a word, these models were first pre-trained on very large corpus in a self-supervised manner, e.g., by masking words from a sentence and training the model to predict them back (Masked Language Model). These models can then be fine-tuned in a supervised manner on a specific task such as part of speech tagging, sentiment analysis or question answering.

Inspired by this approach, large scale pre-train steps were introduced in many different vision and language models. By utilizing a large amount of image-text pairs models included self-supervised cross modal tasks such as *Masked Cross-Modality Language Model*, where words were masked from the textual description and the model utilizes the information from the vision side to predict them. Another pre-training task is *Image-Text Matching*, where a model is given a set of image regions and a sentence and needs to classify if they match together. When designing a cross modal task it is important to force the model to use both modalities. For example, in the Masked Cross-Modality LM task, the model has to use the information from the visual side in order to correctly predict the masked word.

2.3.3 VQA Models

Simple bag-of-words and LSTM+CNN models [10] have been introduced as the original architectures towards solving VQA task. Subsequent models integrated attention mechanisms, [29] introducing co-attention to jointly reason about visual and question information.

LXMERT [13] uses a cross-attention sub-layer followed by self-attention to exchange information and align entities between the two modalities.

ViLBERT [30] uses a similar idea, co-attention Transformer [20] layers, exchanging key-value pairs in multi-headed attention.

Another research direction attempts to deal with compositionality in questions. Neural State Machine [12] is a graph network that simulates the computation of an automaton to model the required reasoning steps to answer a question.

Coarse-To-Fine Reasoning (CFR) [11] is a framework built on the observation that many questions and images are composed of complex semantic information, which can have noise or ambiguous attributes. Thus, they add a filtering mechanism to extract for both modalities relevant features at different fine-grained levels. CFR achieved SOTA of 72.1% on the GQA test-dev and 72.14% on test split.

2.4 Evaluation Metrics and Benchmarks

Evaluation of VQA systems aims at understanding how good they are at answering questions. Accuracy is not enough, because systems can exploit statistical biases in the data. There are set of cognitive skills we expect a system to develop while training on VQA, e.g., compositional reasoning, common-sense reasoning, counting objects, classifying object attributes, determining if a relation exists. Utilizing external benchmarks, we can assess whether these cognitive skills are addressed by the system.

2.4.1 GQA

GQA Test-dev is a balanced subset composed of 12,578 questions and 398 images for evaluation. Testing accuracy on it is a direct metric of how well a model perform on GQA. Additionally, more metrics were introduced taking advantage of the rich semantics associated to each sample in the dataset (scene graph and functional program).

- *Consistency* measures the level of consistency between all entailed questions given a question: if a model can answer a question, it should be able to answer all entailed questions.
- *Validity* measures if a model gives valid answer, that is an answer which is in the set of all possible correct answers given the type of the question.

- *Plausibility* measures if a model answers are reasonable in the real world (based on the distribution of observed answers to the type of question).
- *Grounding* measures the model attendance to the relevant regions in the image in order to answer the question.
- *Distribution* is a comparison between true answer distribution to the model predicted distribution to measure the overall match between them.

2.4.2 SNLI-VE

Visual Entailment (VE) first introduced in [5] is an inference task similar to Textual Entailment (TE), where the goal of a model is to predict whether a premise entails a given hypothesis. In contrast to TE, the input in VE is an image-sentence pair, where a premise is defined by the image and hypothesis defined by the sentence.

In order to research this task, the SNLI-VE dataset was introduced based on the *Stanford Natural Language Inference* corpus (SNLI) [31] and the Flickr30k [32] data set. The data set was developed following four criteria.

- *Structured set of real-world images*, the data set should be based on real-world images.
- *Fine-grained*, any subtle change in the hypotheses that potentially could lead to distinct label should be enforced by the data set with fine-grained reasoning.
- *Sanitization*, no images overlapping between different partitions of the data set.
- *Account for any bias*, measures the data set bias.

SNLI is a large annotated data set built on Flickr30k captions. SNLI authors utilized the image caption from Flickr30k as a premise and collected via crowd-sourcing multiple hypotheses in three classes - entailment, neutral and contradiction. Then, data validation was conducted to measure the label agreement through the data set using the majority vote of five crowd-sourcing workers. By replacing the image caption for the corresponding image, SNLI-VE was created, containing 570K image-sentence pairs.

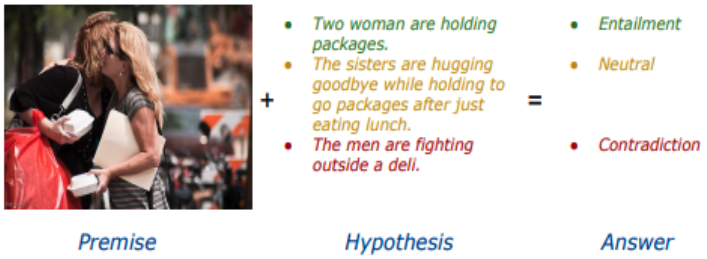


Figure 2.5: Example from SNLI-VE [5], where each premise is an image and the hypothesis is a sentence which have possibly three types of labels.

2.4.3 Winoground

Winoground [6] is a novel task and data set which probes the ability of a model to conduct visio-linguistic compositional reasoning. The data set was collected utilizing images from the Getty Images API [33], where expert annotators carefully designed the corresponding text for each example. Each example contains two pairs of an image and a caption and the goal is ultimately to match them correctly. However, both captions have the same set of words only in different order. This extra criterion, based on the authors experiments, turns Winoground into a very hard task for all top performing Vision+Language models. Formally, Let (C_0, I_0) and (C_1, I_1) be the two image-caption pairs. An example satisfies the Winoground schema is when (C_0, I_0) is preferred over (C_1, I_0) or (C_0, I_1) and (C_1, I_1) over (C_0, I_1) or (C_1, I_0), the comparison depending on the score we want to compute - *Text Score*, *Image Score* or *Group Score*.

Winoground defines three metrics to evaluate different aspects of a model visio-linguistic reasoning skills.

- *Text Score* measures whether a model can select the correct caption, given an image.
- *Image Score*, measures whether a model can select the correct image, given a caption.
- *Group Score*, where all four combinations of the two pairs should be correctly predicted in order the example to be considered as correct.

The data set contains 1,600 image-text pairs in total, where each example has four combinations of pairings while only two are correct. These com-

prise 400 examples with 800 unique captions and images. Winoground is not just hypothesized to be hard task, but experiments shows approximately 50% absolute difference on **Text Score** between humans and the best performing models - UNITER, VILLA, VinVL, ViLT, FLAVA and CLIP. On **Image Score** the situation is even worse, reaching approximately 70% gap between human performance and highest performing model.

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Figure 2.6: Example from Winoground [6]. Two captions contain the same words but in a different order. The example is tagged with the linguistic tag of Object, indicates the type of the swap.

For further analysis, each example is annotated with a total of 70 linguistic tags indicating the type of swap that happened in the text. The set of tags can be categorized into three main groups: objects, relations and swaps involving both. Object swaps reorder elements in the caption such as noun phrases that tend to refer to real-world objects. Relation swaps reorder elements such as verbs, adjectives and prepositions. Swaps of both can be conducted with two swap steps, e.g., object and then relation, or a single swap bridging both. Moreover, examples were additionally tagged from a set of three non-mutually exclusive visual reasoning tags, where the tags are related to the image not the annotated caption.

Chapter 3

Motivation and Starting Points

Relevance is the concept of one topic being related to another topic in a way that makes it useful to consider the second topic when considering the first. As defined by [34], "Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T."

Sperber and Wilson's Relevance Theory [35] attempts to capture the notion of relevance in communicative situations through contextual effects. One of the key ideas of the theory is that the interpretation of all utterances is derived in context. Relevance is conceived as relative and subjective, as it depends upon the state of knowledge of a hearer when they encounter an utterance. Relevance from a pragmatics point of view is a property of utterances processed as inputs to cognitive processes [36]: "An input is relevant to an individual when it connects with available contextual assumptions to yield positive cognitive effects: for example, true contextual implications". After an addressee is decoding a sentence uttered and getting the proposition expressed, the hearer will build a 'context' of 'implicated premises' or assumptions for getting the cognitive positive effects that make the utterance relevant.

We apply these definitions specifically in the task of Visual Question Answering (VQA), where a system must answer free-form questions by reasoning about presented images. The utterance described as the question, the system as the hearer and the context is perceived as the image. In an expected grounding process, a system will get a question, build 'implicated premises' based on what is asked and utilize the image as its source of knowledge to ground the relevance of the question to the image by validating those premises and finally output an answer. In order to answer a

question, the implicated premises must be true in the context of the image (i.e., yield positive cognitive effects).

Premises are propositions described in terms of pragmatic felicity conditions and there are different aspects to determine that a question is relevant given an image. We describe four such conditions followed up by boolean questions to demonstrate their meaning:

- **Reference**, i.e., are the entities to which the questions refer relevant to the image? Are the conditions for using definite or indefinite and singular or plural quantifiers met?
- **Perception**, i.e., is the question related to information that can be derived based on visual info?
- **Presupposition**, i.e., a semantic and pragmatic phenomenon, where an implicit assumption about the context is related to an utterance whose truth is taken for granted. Are all presuppositions arising in the question supported?
- **Implicatures** are inferences suggested by the speaker's utterance, but not included in its literal meaning [37]. Do all implicatures derived from the question hold?

The task of visual question answering requires diverse set of skills to correctly answer different questions. One of the basic cognitive skills required is compositional reasoning. In order to deal with new unseen object relations, object attributes, subject relation object interactions a system needs to be able to understand the phrase by composing the meaning from its parts. For example, without ever encountering "girl throwing ball", its meaning could be understood by composing the meaning of its elementary parts - "girl", "throwing" and "ball". In the realm of VQA, compositional reasoning extends to visio-linguistic compositional reasoning and requires additionally the capability to align between the two modalities.

In this work, we focus on constructing a novel data set through text-based augmentations which trigger controlled presuppositions. For example, *"What color does the dog running on the grass have?"* triggers presuppositions about the existence of a dog, and the fact that the dog runs on the grass. We hypothesize that a VQA system which can properly classify a question as failing to satisfy presuppositions in the context of an image will generalize better for compositionality.

We focus on four types of relations - spatial, descriptive verbs, comparative and prepositional phrases.

Our main research questions are:

- Are there challenging questions to answer in GQA and why are they challenging?
- Can we integrate pragmatic and semantic phenomena to learn a distance metric in order to derive the truth value of a presupposition in a context of an image?
- Does training a VQA system on robustness to infelicitous questions improve its overall performance in the original benchmark?
- Are there questions in GQA in which the annotated answer is not the only one available?

3.1 Semantics & Pragmatics Point of View

3.1.1 Presuppositions

We take a lot of information for granted while speaking, i.e., we presuppose information. As we wrote this thesis, we presupposed that readers would understand English and most of the time speakers tend to do so unconsciously. Formally, the phenomenon of presupposition emerges when a speaker assumes linguistically information as being taken for granted. The presuppositions of an utterance are the pieces of information that the speaker assumes (or acts as if she assumes) in order for her utterance to be meaningful in the current context (cf Pots in [38]).

There are two kinds of presuppositions, pragmatic and semantic. Semantic presuppositions are properties of natural language expressions, while pragmatic presuppositions are properties of information emerging from the speech. The two are related to one another in the sense that when a semantic presupposition is used in an utterance, it typically triggers a pragmatic presupposition. For example, if a tourist says something to you in Spanish, they pragmatically presuppose that you understand Spanish.

Presupposition is distinct from other inference types, such as implicature and entailment. Presuppositions have the characteristic property that they survive when embedded under operators like negation and questions [39].

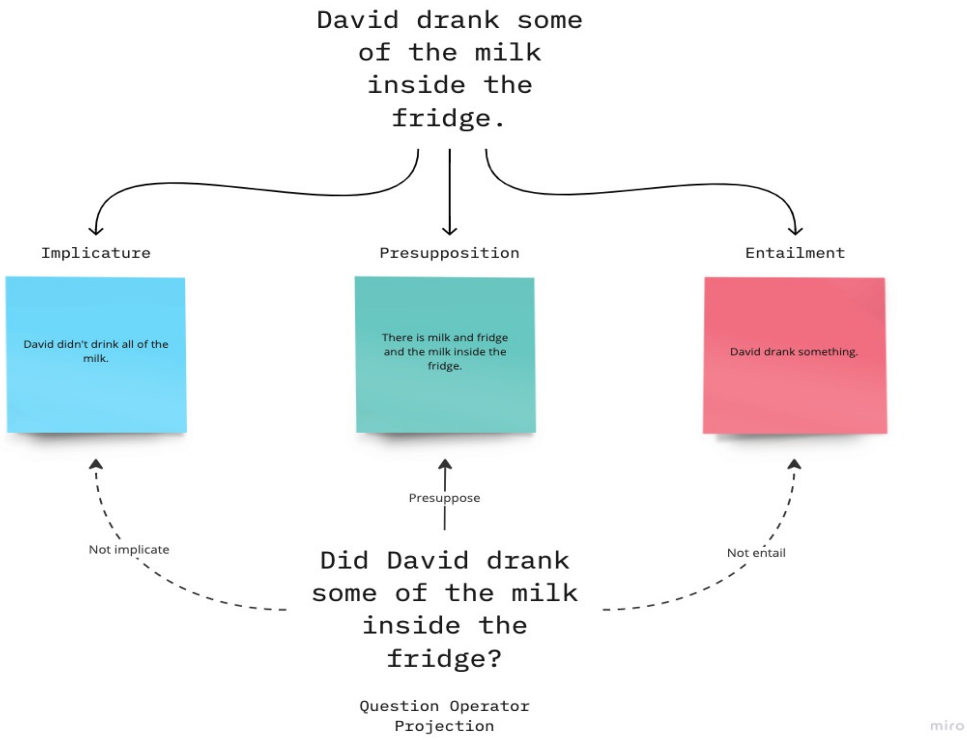


Figure 3.1: An example of presuppositions that survive when embedded under a question operator vs. other inference types.

Semantic presupposition is detected through the presence of lexical items and syntactic constructs called *presupposition triggers*. These items form a large class including definite noun phrases and factive verbs. For example, when a speaker asks *"What color is the dog?"*, an existential presupposition arises, indicating the presence of a dog and triggered by a definite description "the dog". A presupposition that does not hold with respect to the current context yields a pragmatic situation called *presupposition failure*. For example, asking *"What color is the dog?"* in the context of an image full of cats without any dog yields an existential presupposition failure.

Presupposition triggers are frequent and there are many lexical classes widely recognized. The following list illustrates the relevant classes:

- **Factive Verbs** - "David **knows** that he is signing a two-way contract." → David is signing two-way contract.
- **Temporal Clauses** - "The guy drank water **after** he went to the gym." → The guy went to the gym.

- **Definite Descriptions** - "**The** cake inside **the** fridge is full of colors."
  → There is a cake inside the fridge and it is unique.
- **Questions** - There are many types of question presupposition triggers, we focus on two cases:
  - **Alternative Questions** - When presenting alternatives, exclusive-or questions tend to trigger a presupposition of the truth of one the alternatives.
    "Is the ball under the table green or blue?"
    → The ball under the table is either green or blue.
  - **WH Questions** - Questions containing interrogative pro-forms (denoting the unknown item in the question) tend to trigger a corresponding presupposition containing an in-definite pro-form.
    "Who is hiding under the table?"
    → Someone is hiding under the table.

Presuppositions triggers are associated to specific types of semantic presuppositions. The relevant types of semantic presuppositions we consider in this work are:

- **Existential Presupposition** - Is the speaker committed to the existence of the entities assumed to be present by the usage of noun phrases:
  "The cat is on the mat."
  → The cat and the mat exist and the cat is on the mat.
- **Factive Presupposition** - It is assumption assumed to be committed to the existence of factive verbs such as 'know', 'regret', 'realize', 'glad', 'aware'. "Samuel regrets revealing his financial secrets to us."
  → Samuel revealed his financial secrets to us.
- **Structural Presupposition** - It is the assumption associated with use of certain words, interrogative pro-forms for example. WH questions in English are conventionally interpreted with the presupposition that the information after the WH-form is already known.
  - "Where is the dog?"
    → There is a dog in the context.
  - "Who is hiding under the table?"
    → Someone is hiding under the table.

Presuppositions must hold in some context. In conversational settings the context is the shareable knowledge that all participants have or public available knowledge. For example, if there are a group of friends feasting in a restaurant together and one of them declares *"I called the waitress to order a dessert"*, then all of them understand that he called their specific waitress. In VQA on the other hand, the context is the given image and the aspect of aligning between textual entities and visual entities is emerging while trying to validate each of the presuppositions truth value. This aspect is known as visual grounding. If a VQA system fails to visually ground the presupposed entities in the question, then there will be a case of presupposition failure. In conversational settings, people are able to validate the truth value of a presupposition even if the meaning is vague. For example, if someone says *"The glass is sitting on the counter"*, it is obvious from commonsense that a glass doesn't actually sit, it is just placed on the counter, however a hearer will most likely extend his common ground of assumptions and still understand the meaning of this presupposition. We will delve deeper into understanding this phenomenon in the next section.

In this work, we focus on semantic presuppositions with an emphasis on existential and structural presuppositions. In the questions we are dealing with, we will mainly encounter existential presupposition arising from the existence of subject, object and relation in-between and we will represent this presupposition as a triplet - $(subject, relation, object)$.

3.1.2 Accommodation

Accommodation was first discussed by Karttunen [40]. Karttunen introduces the concept in an ordinary conversation, where it does not always proceed in an ideal fashion. People tend to make shortcuts and leaps by using sentences whose presuppositions are not satisfied in the conversational context. However, if the current conversational context does not suffice, the listener is expected to extend it as required. Similarly, Heim [41] defines that accommodation is the process whereby a context is adjusted to make an update possible when presuppositions do not hold. Consider the following example: *"Rachel observed a leakage in her apartment and informed the owner"*. In conversational settings, in order to determine the intended meaning of *"the owner"*, the hearer has to infer (i) that there is an owner and (ii) that the said owner is the owner of the apartment.

Figure 3.2: Accommodation in process demonstrated on a question from GQA - "Is the pipe on the wall green or brown?"

In VQA settings, the accommodation process takes place while visually grounding the textual entities and match them to the presented visual entities. For example in the question at 3.2, an existential presupposition arises as $(pipe, on, wall)$. If we analyze the compositional terms in the question, then a correct way to rephrase the question might be "Is the pipe attached to the wall green or brown?" (there are multiple pipes in the scene). However, through the accommodation process humans can adjust and update the available assumptions, as if $(pipe, on, wall)$ holds even though the pipe was not introduced in the discourse earlier to warrant the usage of the definite expression "the pipe".

3.1.3 Compositionality

Compositionality is a concept in the philosophy of language, which states that the meaning of complex expressions is built up from the meaning of their sub-expressions. We can understand a large collection of complex expressions the first time we encounter them, simply by combining the meaning from their constituents.

Compositionality is a desirable property when developing AI systems in general, in particular VQA systems. Since when a VQA system is trained,

we cannot represent every possible expression in our training data, i.e., our sample cannot fully represent all possible compositional terms in the underlying distribution. Thus, in order to generalize to new unseen questions including new unseen expressions, compositionality is a promising strategy as opposed to pure end to end solutions. For example, when a VQA system encounters the question *"What is the color of the shirt of the girl that is throwing the ball?"* in the train set, we expect the system to be able to answer similar questions such as *"What is the color of the shirt of the **boy** that is throwing the ball?"*, even though it never encountered the expression *the boy that is throwing the ball*.

As discussed above, VQA systems in the past tended to exploit statistical biases in the training data in order to answer questions. For example for the question *"What color is the banana on the counter?"*, the answer will be "yellow" with high probability without even considering a given image. Quantifying and measuring compositionality can provide us with in-depth understanding of the model and the ability to test it is valuable.

3.2 Infelicitous Questions

We define that a question is *infelicitous* when some of the presuppositions triggered by the question in the context of a given image do not hold. Formally, given (Q, I) a question and an image pair and all presuppositions $(P_1, ..., P_n)$ arising in Q. We test whether $(P_1, ..., P_n)$ hold in the context of I. If at least one presupposition does not hold, i.e, has false value, we say the question is infelicitous or irrelevant.

Recent work has showed good results on various VQA data sets, such as GQA. These data sets only contain relevant questions by construction, however a VQA system in the wild is expected to provide an answer even if the submitted question is not relevant with respect to the image. Current models which cast VQA as a multi-class classification task will still predict the answer even if it doesn't exist. For example, when asking *"Where is the dog?"* and there is no dog in the given image, existing approaches will output a baseless answer.

The task of detecting infelicitous questions in the context of VQA was studied by [42]. In order to evaluate it, they constructed the Visual True and False Question (VTFQ) dataset. In attempt to improve or probe compositionality through data augmentations, [43] construct an irrelevant questions dataset by substitution of the context (the image).

In contrast, in our approach, we consider text-based augmentations, specifically presupposition augmentation (discussed in later sections) to generate infelicitous questions. We rely on previous work [44] from our group (BGUNLP), which also attempted to synthesize infelicitous questions, but with the focus on creating compositional split from GQA and sampling objects that are likely to appear, but do not appear in the scene. They generated different type of failures utilizing this sampling mechanism to create a variety of types of infelicitous questions. During their work they re-trained LXMERT on GQA train balanced subset augmented with 80k additional synthetic questions marked as irrelevant as their answer.

As a starting point to this work, we designed a manual annotation process and annotated 500 questions, 250 predicted with irrelevant answer and 250 predicted with a relevant answer from GQA. We generated overlap splits of the subset such that every question was seen by two distinct annotators.

The manual annotation process must be carefully designed, since a question can be irrelevant for many reasons, as mentioned above. On the one hand, directly instructing annotators to determine whether a question is irrelevant according to our definition might lead to biased results. On the other hand, if annotators are not given any guidance, we may end up with low levels of agreement among annotators. In order to find a balance between the two, we empirically analyzed how people deal with ambiguity or complexity when answering questions about what they see in images. Our aim was to gather information on what causes questions to be perceived as difficult or easy given visual scenes.

Question: *What does the person wear?*

Answer: _____

Is the question challenging? ____

If yes, please explain how you reached your answer:

Figure 3.3: An example from our annotation framework

For each question, we ask annotators if they find it a challenging question, and if yes, they must elaborate on the reasons. From the frequently arising reasons, we gather information about the types of failures that make a question challenging. When asking if a question is challenging, we are forcing annotators to not accommodate questions, which will give rise to uncertainty about challenging questions. Since challenging is a subjective term, one can perceive a question as challenging and the other as not. Thus, through manual analysis of the questions where the level of agreement is low, we can derive the types of failures, i.e, categorize them according to classes of disagreement.

Both annotators marked questions predicted as irrelevant as challenging for 37.8% of the cases. At least one annotator marked questions predicted as irrelevant as challenging in 70.1% of the cases.

In order to measure the total level of agreement between annotators, we computed Cohen Kappa score for each question between the answers of both annotators, which was 0.29, which indicates poor agreement (Kappa score below 0.40 is considered poor). However, the comparison included all answers for all questions and relevant questions where an answer can be given. In the relevant case, two annotators can give almost the same answer, but not exactly, a synonym for example, thus can drop the level of agreement dramatically.

After manual analysis of 177 questions where both annotators do not agree, we came up with 8 classes of disagreement, which we later utilized as our core failure types when constructing the new data set. The full list along with examples is as follows:

1. **Alternative questions where answer is no.**
   $\rightarrow$ "Are there bags or cars?"

2. **Relation to object that is not visible but could be inferred.**
   $\rightarrow$ "What type of fruit is on the bed?", when there is a sheet but not a bed in the image.

3. **Uncertainty mapping a lexical term in the question to objects or properties in the image.**
   $\rightarrow$ hat vs helmet, trouser vs shorts, street vs lane, small vs large clock.

4. **Multiple objects in scene with same answer to a singular question.**
   $\rightarrow$"How big is the poster?", when there are two posters of the same size in the scene.

5. **Preposition or relation uncertainty.**
   $\rightarrow$ "What is around the man on the left?", (i) the term around vs close by, for example, can be found as challenging in a strict examination. (ii) the term "on the left" as well, from which perspective - viewer or object in scene.

6. **Multiple objects match the type but the property can be verified only on one.**
   $\rightarrow$ "Are the eyes of the elephant open?", when two elephants are visible in the scene but the eyes of only one are visible.

7. **Uncertainty in describing an object in the scene with an appropriate term.**
   $\rightarrow$ "What does this bird walk on?", where the referred object looks like a tree, rock or crocodile.

8. **Uncertainty in answering where many answers are possible.**
   $\rightarrow$ "What does the biker wear?", where the biker wears jeans, t-shirt and shoes.

Based on these preliminary annotation results and the classes of disagreement we identified, our aim is to (i) study presupposition augmentations from the perspective of relations between object and subject and (ii) generalize the approach to obtain a large-scale dataset.

Chapter 4

Presupposition Augmentation via Scene Graphs

We present in this chapter the method we have designed to transform an existing $(question, image)$ pair into a set of infelicitous questions with respect to the same image. The process consists of synthesizing failed presuppositions from the original question and then regenerate a modified question which would trigger the unwarranted synthetic presupposition.

Using this augmentation method, we create a dataset of questions that cannot be answered in the context of specific images mixed with original relevant questions. In the next chapter, we show how this new dataset can be used to train a new model that can detect irrelevant questions and behave in a more robust manner than existing VQA models.

4.1 Using Scene Graphs for Candidate Presupposition Generation

While presuppositions can be triggered by a variety of mechanisms, we focus in the following method on existential presuppositions triggered by definite descriptions. These descriptions can appear in different positions within sentences: as subject, object or as part of the expression of the relation between subject and object. Not all questions containing a subject or object presuppose something, for example, the utterance *"Is there a dog?"* does not trigger an existential presupposition. In contrast, *"Is the dog running on the grass?"* triggers two such presuppositions.

We represent presuppositions in three forms:

- A singleton describing the object or subject the question presupposes. For example, the presupposition associated to *"Where is the dog?"* is represented as (dog).
- A pair $(subject, relation)$. For example, *"What is the dog running on?"* is associated to $(dog, running\ on)$.
- A triplet, $(subject, relation, object)$. For example, *"What is the color of the dog running on the grass?"* is associated to $(dog, running\ on, grass)$.

Given such a representation of presuppositions, a general augmentation method consists of choosing a candidate element and substituting the term with a new one, and finally generate a new question triggering the modified presupposition.

However, we must ensure that the truth value of the new presupposition is false to cause failure. Technically, the step which consists of selecting a component with a new one is tricky, since not all datasets contain additional data structures to describe what is inside a given image, and even when these data structures are available, they may be partial. For example, assume an image showing a dog running on the grass and a relevant question *"Is the dog running on the grass?"*. We want to replace the presupposition $(dog, running\ on, grass)$ with $(cat, running\ on, grass)$ and generate the candidate question *"Is the cat running on the grass?"* as a candidate irrelevant question. But, if the image actually contains a depiction of a cat, this new synthetic question may end up actually being relevant.

In such cases, silver labels generated by automatic object detectors and object-relationship detectors can be leveraged. However, current Vision and Language encoders barely distinguish between triplets, thus, automatically validating whether a presupposition holds in an image remains technically challenging.

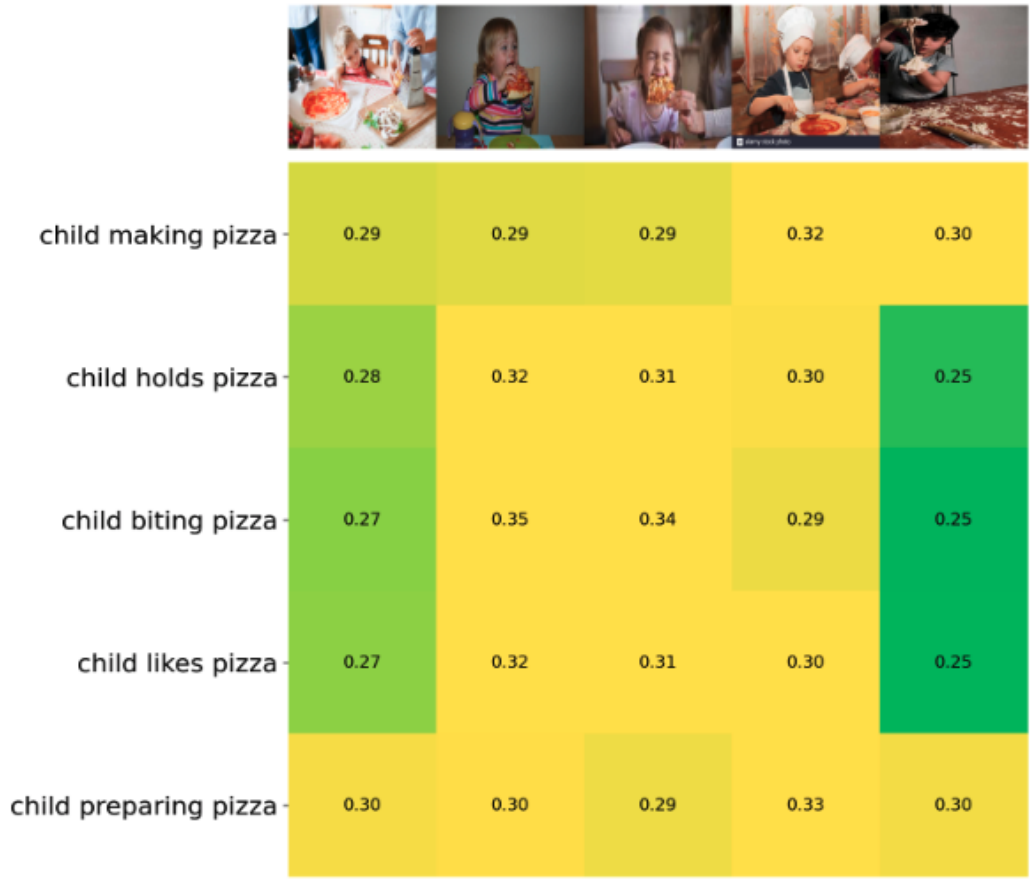


Figure 4.1: CLIP score on different relational presuppositions.

To illustrate this difficulty, consider the example in Figure 4.1. We measure the distance between an image and triplet as measured by the state of the art system called CLIP. The left most image when compared with the triplet $(child, making, pizza)$ scores 0.29. When comparing the same image with the triplet $(child, holds, pizza)$, where in the image the child isn't actually holding the pizza but instead touching it, we still get a score of 0.28, nearly as close. Such small differences in scores are not robust enough or calibrated to enable the factual validation of irrelevant presuppositions.

To overcome this limitation, we rely on the presence of Scene Graphs in the GQA dataset associated to each image. We use the scene graph instead of the image as a representation of the context when assessing whether a candidate synthetic presupposition holds.

After substituting the selected part in the presupposition and generating a

new candidate, we traverse the scene graph and generate singletons, pairs and triplets as candidate false presuppositions. We then compare each of the generated candidates with the scene graph to derive its truth value.

Figure 4.2: "What is the color of the pipe attached to the wall?"

4.2 Assessing the Truth of Presuppositions in Context

When semantically comparing the augmented candidate pairs and triplets to those found in the scene-graph, the process of accommodation comes into play. For example, consider the question and image pair in Figure 4.2. The presupposition arising from the question can be found in the scene graph as $(pipe, attached\ to, wall)$.

We consider whether the alternative synthetic triplet candidate $(pipe, on, wall)$ holds in the context of the image (this candidate triplet is obtained by replacing the relation $attached\ to$ with on). It seems, however, that this triplet can be accommodated - and we empirically confirm that human annotators tend to classify this synthetic triplet as true in the context of the image.

A direct comparison of triplets, which would only be based on their syntax, would lead to unreliable conclusions in terms of their truth value. Even though $attached\ to$ and on are not similar relations, when they are

used in the context of $pipe$ and $wall$, the resulting combination yields similar contextual meanings. This effect is probably connected to commonsense and is not categorical: different people provide different judgments.

In conclusion, while comparing different relational presuppositions, we must address this phenomenon. We describe next how we learn how to compare such triplets using natural language processing techniques.

4.3 Scene Graph Validation

When comparing between a presupposition to singletons, pairs and triplets from the scene graph, we seek to determine if they have close semantic meaning. The three forms described earlier must be approached with different measures.

4.3.1 Embedding-based Comparison of Triplets

When comparing the meaning of a single object to another object without any context, word embedding methods such as Word2Vec [45] and GloVe [46] can be utilized. However, these methods are heavily based on the frequencies of occurrence of the words in different contexts, making the comparison of antonyms and hypernym - hyponyms difficult. Each object in the scene graph in Visual Genome is labeled with its corresponding WordNet [47] synonym label. We utilize this information as our core comparison mechanism between objects: if the synonym labels don't match, we select a representative term from the lemmas describing the synonym set in WordNet and use it to compare through cosine similarity with respect to GloVe word representations.

We compare the second and third forms of relational presuppositions using Sentence Transformers [48]. Specifically, we aim to use paraphrasing model to understand if two relational presuppositions are semantically equivalent. In order to validate if those models are sufficient for comparison, we manually generated 50 subject and object pairs with two sets of relations. The first set is the positive set, where embedded under the subject and object they will most likely be found semantically equivalent under the context of most images. In contrast, the negative set contains relations that are semantically different when embedded. For example, $(man, shoes)$ with positive set $\{wear, has, with, in\}$ and negative set $\{give, throw, claim\}$.

Set Type	Cosine Similarity Average	Cosine Similarity Std Deviation
Positive	0.915	0.049
Negative	0.836	0.085

Table 4.1: Sentence Transformer *paraphrase-mpnet-base-v2* evaluated on positive and negative sets.

At first glimpse, it seems that paraphrasing models can help here to determine if two relational presuppositions are semantically equivalent, therefore it is holding as our hypothesis.

In order to test that claim, we selected *paraphrase-mpnet-base-v2* as our pre-trained model to evaluate over the 50 pairs we use as a development set. For each subject and object pair, we computed the average cosine similarity between each member in the positive set and averaged over all 50 pairs. For the negative set, we averaged the positive set to generate a representative and computed the average cosine similarity for each member in the negative set to the representative, averaged over all 50 pairs.

We expect the model to distinguish between positive and negative sets with respect to their cosine similarities. However, we observe in Table 4.1 that the ranges of positive and negative sets are too close: i.e., the positive $[0.866, 0.964]$ and negatives $[0.751, 0.921]$ cosine similarity ranges overlap. Thus, we conclude that the metric the model represent isn't sufficient for our needs. We therefore try a different method.

4.3.2 Learning an Accommodation-based Distance Metric

Sentence Transformers use BERT-based Siamese Networks [49] to learn a distance metric by leveraging different Contrastive Learning [50] objectives. We use the same objective to fine-tune pre-trained models to fit our metric needs.

The dataset of 50 pairs that we have collected isn't sufficient in size for the training to converge. In order to enrich it, we utilize WordNet synonyms for each pair's subject and object. Additionally, we enriched our negatives by crossing pairs.

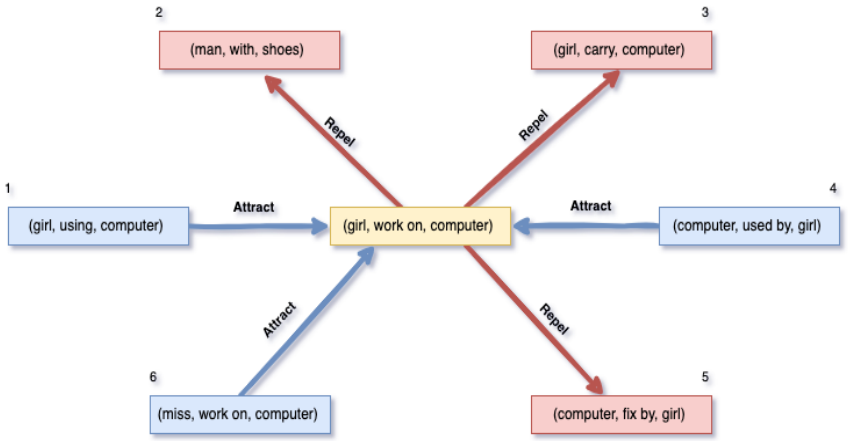


Figure 4.3: Data Enrichment Visualization

Figure 4.3 provides a visualization of the six different types of data enrichment. Edges annotated by *"Attract"* connect positive pairs and those with *"Repel"* denote negative pairs. Type (1) is a pair from the manually annotated positive set. Type (2) is a cross dataset negative pairing. Type (3) is a pair from the manually annotated negative set.

Sometimes, the presupposition can be accommodated even if we substitute the subject and object, for example $(man, with, shoes)$, $(shoes, of, man)$, $(shoes, wear\ by, man)$. Moreover, we don't know in which way an annotator of a scene graph chose to describe the relationship.

In order to address that issue, for each of the 50 pairs, we manually annotate additional positive and negative sets when substituting the subject and object.

Type (4) is a substituted subject and object pair from the manually annotated positive set. Type (5) is a substituted subject and object pair from the manually annotated negative set. Type (6) is a pair generated by substituting the subject with synonyms from WordNet. Implementing those types of enrichment, we were able to generate $43,091$ positive pairs and $207,885$ negative pairs.

Model	Positive CosSim Avg/Std	Negative CosSim Avg/Std
pre-trained	(0.647, 0.179)	(0.292, 0.232)
pre-trained-FT	(0.968, 0.044)	(0.215, 0.144)

Table 4.2: Pre-trained Sentence Transformer *paraphrase-mpnet-base-v2* and its fine-tuned version on the enriched data set evaluated with cosine similarity metric on positive and negative pairs from the test set.

We fine-tuned the model for 3 epochs with low learning rate $(1e-06)$ which improved F1 score from 55.91% to 99.29% on the test set. Additionally, observing Table 4.2 we conclude that the overlap between negative and positive ranges after the fine-tuning procedure is empty. The fine-tuned model can distinguish between accommodated to non-accommodated relational presuppositions.

With this method, we can generate a set of false presuppositions which do not hold with high probability in the context of the scene graph depicting an image. We now turn to the next step which consists of turning these false presuppositions into irrelevant questions.

Chapter 5

Irrelevant Question Generation

In this chapter we describe how we constructed a dataset of irrelevant questions with respect to compositionality and different types of failures introduced in each question generated. At first, we select the types of question failures we wish to generate. Then, we partition the generation pipeline into two sub-pipelines. The first is based on predefined question templates and the second is based on GQA questions and their functional programs.

Finally, we demonstrate how to leverage the augmented dataset we have created in order to train an irrelevant question detector and filter irrelevant questions from existing VQA datasets.

5.1 Synthesizing Irrelevant Questions

Our goal is to generate new questions which rely on failed presuppositions given the scene graph associated to a given image.

We use two general methods to synthesize such questions: (1) using predefined templates to generate new questions on the basis of carefully selected triplets (sub, rel, obj) on the basis of the scene graph only; (2) starting from an existing question in the GQA dataset and corrupting it to according to specific transformation rules.

5.1.1 Pre-Defined Question Templates

For the first method, we rely on the a-priori class of presupposition failures we have identified in Section 3.2:

- **Type 1** - Definite reference to a non-existent object.
- **Type 2** - Definite singular reference to an object when the scene contains multiple instances of this object.
- **Type 3** - Definite plural reference to an object when the scene contains only one instance.
- **Type 4** - Definite reference to an object that exists in the scene, but the instance does not hold in the relation expressed in the description.
- **Type 5** - Interrogative pro form coupled with a definite reference to an object or subject that exists in the scene, and the instance does hold in the relation expressed, but the relation refers to multiple options.

Accordingly, we build a pipeline to generate these five types of presupposition failures, focusing on definite description and interrogative pro forms as presupposition triggers.

We implement question synthesis by iterating over scene graphs and generating irrelevant questions from pre-defined templates. We constructed for each type two to four templates, each of which associated to a different presupposition failure depending on the presence of interrogative pro-forms or definite description coupled with the different positions in the question (subject, object, relation). The results of this method are shown in Figure 5.1.

Question	Type	Sub-Type	Failure Reason
Where is the white kite?	1	3	No white kite in image
What is the man toward?	2	2	There are multiple man in image
What are the roads to the right of?	3	1	There is only one road in image
What does the man away from the group on?	4	2	No man away from group in image
What does the woman wear?	5	1	The question refer to at least two objects: ['boot', 'hat'] in the image

Figure 5.1: Examples of irrelevant questions for each type. Sub-types correspond to the template used for the synthesis and different failure reasons.

For *Type 1* questions, we have three templates. The first is focusing on the object of a relation, e.g, $(wearing, jacket)$. The main idea is to create a new pair $(relation, object)$ which does not hold in the scene graph. To implement, we sample a pair $(relation, object)$ and we focus on two conditions: *plausibility* and *failure*. Plausibility means that the selected object can appear with respect to the relation in the real world. We take the list of such pairs in Visual Genome as our source and sample from this distribution. The second condition ensures that the sampled object is not holding in the

given scene graph.

The second template is much like the first but with focus on subject and relation instead. The third is focusing on the subject or object, we implement it by randomly sampling a noun which doesn't hold in the given scene graph.

For *Type 2* the synthetic failure arises from using a definite singular reference to an object when the scene contains multiple instances of this type. To implement it, we count nouns in the scene graphs. However, sometimes annotators of VG scene graphs annotated the same object multiple times. In order to avoid relating to the same object as multiple instances, we merge objects when their center is δ close to each other and consider these references as a single instance. We select $delta$ heuristically. Moreover, we filter mass nouns, i.e., non-countable objects such as *sky*. If we find a noun which appears more than once, we embed it in three different templates. The first has a singular reference to the object which causes the failure. For the second and third we sample a relation to the described noun applying the same plausibility and failure conditions mentioned above for *Type 1*.

Type 3 is similar to *Type 2* in terms of implementation, except that the failure arises from using a definite plural reference when the scene graph contains only a single object of that kind.

For *Type 4* the failure arises from definite reference to an object or subject in a relation, but the instance does not hold in the relation expressed. Thus, our focus is on substituting relations expressed in triplets mined from the scene graphs in order to trigger a failure. We approach it in four different templates. Let (s, r, o) be a triplet taken from a scene graph. For the first template, we validate if after substitution of the subject to be the object and object to be subject, (o, s) is a plausible pair of subject and object. If it is, we sample a new plausible relation randomly r' and validate if (o, r', s) is not holding in the scene graph.

Let (s_1, r_1, o_1) and (s_1, r_2, o_2) be triplets taken from a scene graph. For the second template, we generate a compositional question by asking about r_2 and generate a failure by substituting r_1 with a new relation r' according to the approach described in Section 5.1.3, where (s_1, r', o_1) is not holding in the scene graph. For example, $(woman, talking\,to, man)$, $(woman, wearing, shirt)$ $\rightarrow$ *"What does the woman buying from the man wear?"*.

For the third and fourth template, we split (s, r, o) to two pairs: (s, r) and (r, o), substitute r with a new relation r' sampled randomly with respect to

plausibility and validate if (s, r') and (r', o) do not hold in the scene graph.

Type 5 is different from the rest. The interrogative pro forms e.g, "What" and "Who" presented in two different templates, generate a true assumption about the existence of an object or subject and the relation expressed. But the assumption contains additional uniqueness regarding the subject or object. Each question is asking about only one subject or object. Thus, the failure is emerging from the ambiguity arising when the context has multiple options for subject or object.

For example, in Figure 5.1 the question *"What does the woman wear?"* is referring to one object, where the woman is wearing multiple items, clearly one can elaborate on all possible answers. But the questions in VQA refer to only one. This can cause a drop in accuracy for a VQA system when it faces a question with such ambiguity and not answering the annotated answer.

The full template list is given in Appendix.

5.1.2 GQA Questions

This second pipeline differs from the previous one in two aspects. The first is that we iterate on real questions from GQA to generate new questions and second, only *Type 1* and *Type 4* are implemented due to insufficient reliability when trying to implement other failure types from existing GQA questions.

There are two main reasons why we chose to utilize questions from GQA to generate new ones. The first is variance: we don't want a model trained with irrelevant questions to be able to predict an irrelevant question just from the structure, or to exploit statistical bias such as different interrogative pro-forms that appear in our templates while GQA contains many different templates for questions. The second is compositionality: GQA contains 94.34% questions with at least two steps required to infer the answer. We aim to synthesize irrelevant questions that are similar to the original GQA questions in this respect.

GQA contains two types of questions: **Structural type** is derived from the final operation in the provided functional program. **Semantic type** refers to the main subject of the question: **object**, **attribute**, **category**, **relation** and **global**.

In our pipeline, we focus only on *query-rel* structural questions, i.e., open questions (as opposed to Y/N questions) asking about the subject or object

of a described relation. For example: *"What is the girl wearing?"*. The *query-rel* subset composes 34.1% of the GQA dataset, the most frequent type. We selected this type because by augmenting the relation, subject or object, we are able to change the meaning of the compositional terms found within the questions and still cause a presupposition failure.

However, not all questions directly presuppose relational presuppositions. For example, the question *"Who is wearing a shirt?"* does not trigger any existential presupposition, but only a structural one (someone is wearing a shirt). Thus, when we augment an existing question by replacing the subject or object, we replace determiners such as "a" or "an" with a definite description or inject additional definite description to trigger an existential presupposition. For example, we generate *"Who is wearing the sunglasses?"*. Finally, we use T5 [51] as a Grammar Error Correction tool to produce grammatically correct questions.

Question	Type	Failure Reason
Who is sitting around the table the bottles are on?	Original	
Who is sitting around **the cake stand** the bottles are on?	1	No cake stand in image
Who is sitting around **the kids** the bottles are on?	1	No kids in image
Who is sitting around the table the bottles are **under**?	4	No bottles under table in image

Figure 5.2: Examples of irrelevant questions generated from GQA original question, their type and failure reason.

The implementation is very similar to Section 5.1.1, except that we don't have access to the triplets in the question describing subjects, objects and relations. Instead, we use the functional program to extract triplets as potential presuppositions.

For *Type 1* we randomly sample plausible subject or object and validate the augmented triplet doesn't hold in the scene graph provided with the question.

For *Type 4*, given a triplet, we substitute the relation according to the method described in Section 5.1.3 and validate it doesn't hold in the scene graph provided with the question.

5.1.3 Filtering Entailed Presuppositions

Let (s, r, o) be a triplet taken from a scene graph. We select a new relation r' to generate the candidate triplet (s, r', o). Although, (s, r, o) and (s, r', o) can have completely different meaning, (s, r, o) could entail (s, r', o).
For example, $(horse, eating\ from, pasture) \rightarrow (horse, in, pasture)$.

Our method of testing whether whether the candidate triplet holds in the scene graph would fail to detect this situation because inferred information is often not mentioned explicitly in scene graphs.

To address this phenomenon, when selecting candidate relations, we utilize an NLU model [52] to classify whether an existing presupposition entails the augmented presupposition. Specifically, when considering candidates for the augmented triplet from all possible relations in VG between s and o, we filter out any candidate which the model predicts as entailed.

5.2 Detecting Irrelevant Questions

Utilizing the new irrelevant questions dataset generated from the two above pipelines, we fine-tune LXMERT for the VQA task with additional *"irrelevant"* questions blended into the GQA Train Balanced split, GQA Validation Balanced split and our dataset.

Data	Model	Train Size	Valid. Size	# Answers
Original GQA	LXMERT$_{base}$	943K	132K	1842
GQA + Irrelevant Qs	LXMERT$_{base}$	1023K	152K	1843

Table 5.1: LXMERT fine-tuning experiments data details.

Table 5.1 shows the size of the train and validation set in each experiment. In the first experiment, in order to create a baseline for comparison using we simply fine-tune LXMERT with GQA Train Balanced split as our training set and GQA Validation Balanced split as our validation set. In the second run, we use an additional 100K irrelevant questions, split $80\% - 20\%$ into train and validation. Evaluation is done both on the new validation set after adding $20k$ irrelevant questions and GQA Test-Dev Balanced split separately which we later use for further analysis.

Irrelevant Size	Type 1	Type 2	Type 3	Type 4	Type 5
Original					
100k	0.916	0.816	0.859	0.931	0.795

Table 5.2: LXMERT fine-tuned with additional irrelevant questions F1 score by type.

All models are trained on NVIDIA Tesla T4 for 3 epochs, learning rate of $5e-05$, batch size of 64 and linear decay of learning rate.

5.2.1 Analysis: Detection of Irrelevant Question Types

Observing the results from Table 5.2, the model trained with additional 100K irrelevant questions is capable of detecting all types of irrelevant questions with high F1 score (0.795 to 0.931). *Type 2*, *Type 3* and *Type 5* are relatively low compared to the other types. This is expected, since *Type 2* and *Type 3* requires the model to develop counting skills in order to detect the failure which is known to be difficult for existing model architectures. Additionally, *Type 5* questions are the only questions in which the failure is emerging from the ambiguity of many presuppositions and not a presupposition failure, thus this kind of questions are very similar to relevant questions - making their detection more difficult.

Moreover, we observe that *Type 4* questions have relatively high F1 score (0.931). Recall, *Type 4* are questions generated and validated using Scene Graph Validation (Section 4.3.2), thus indicating a strong and consistent validation process with low noise.

5.2.2 Impact of Irrelevant Questions on Relevant Questions Answering Performance

Our baseline model reaches an accuracy of 58.55% on GQA Testdev, while after training with additional 100K irrelevant questions, we observe a drop of 0.71% in performance to 57.84%.

Our original assumption was that the capability to detect irrelevant questions would extend the capabilities of the model and accordingly lead to improved performance on relevant questions as well.

We explore the hypothesis that this drop in performance may be caused by the classification of original GQA questions as irrelevant. In other words,

we suspect there may be irrelevant questions in the original GQA Testdev split.

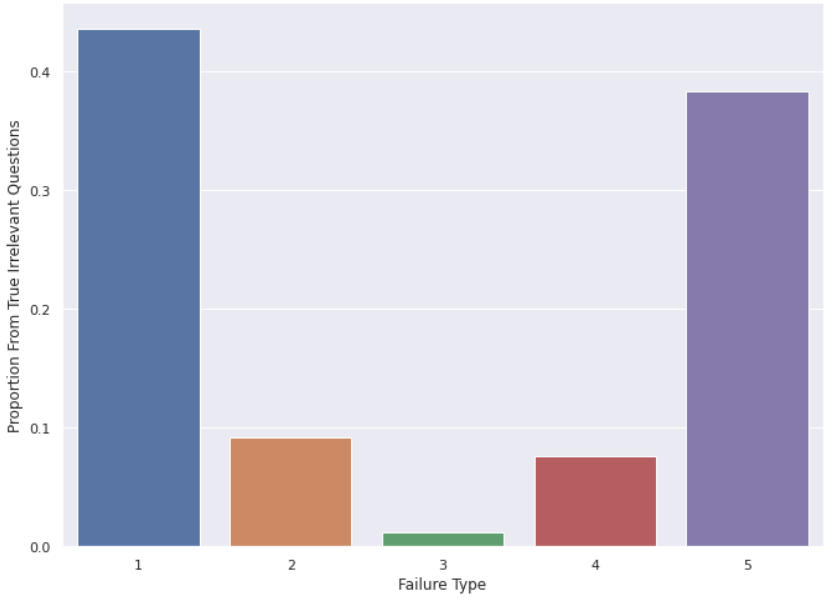


Figure 5.3: Failure types normalized frequency distribution of true irrelevant questions annotated from the subset the model predicted as "irrelevant" from GQA Testdev balanced split.

To test this hypothesis, we performed a manual annotation campaign on the subset of questions from GQA Testdev that our model predicted as "irrelevant", which corresponds to 2.82% of the original GQA Testdev split. Our annotation identifies that 70.42% of these original GQA questions are actually irrelevant questions, confirming our hypothesis. After cleaning the "irrelevant" questions, the model accuracy on GQA Testdev indeed improved by about 2% - thus confirming our original assumption that detecting irrelevant questions is also beneficial in making the baseline model more robust to answer relevant questions as well.

Moreover, in our manual annotation process, we didn't include only straightforward annotations of "irrelevant"/"relevant", we also annotated the types of failures. Figure 5.3 shows that within original GQA test questions, *Type 5* has a frequency of 38.4%, which is relatively high frequency. This leads to an additional suspicion that many questions in GQA in general cannot be answered because of ambiguity arising when there are multiple answers available and the model needs to provide only one. In such cases, multiple answers may be valid, but only one would be recognized by the

GQA validation procedure.

Chapter 6

Conclusion

In this work, we studied the effect of irrelevant questions in the context of VQA. We found or analysis on relevance definitions from semantics, pragmatics and human-to-human conversational theory. Our study focused on questions made irrelevant by presupposition failure. We restricted our attention to existential presuppositions.

We empirically discovered that classifying questions as irrelevant is a vague task and without additional guidance, humans annotating a question as irrelevant is a process reach low levels of agreement. We identified common classes of disagreement among annotators through a manual annotation process and established that existing questions from the standard GQA dataset are indeed challenging to answer.

We then designed different methods to generate presupposition failures based on the classes of disagreement identified by our empirical analysis. The synthesis of false presuppositions requires a complex validation process, we had to address multiple semantic and pragmatic phenomena.

We exploited semantic information associated with images in the form of scene graph and QA pairs from GQA to automatically synthesize presupposition failures. Using the generated irrelevant dataset, we fine-tuned a multimodal pretrained model to detect irrelevant questions.

The model trained on irrelevant questions in addition to the original GQA dataset performs well at detecting irrelevant questions from relevant ones. It also improves the baseline on relevant questions by about 2%.

Our model also helped us detect problematic samples within the original GQA dataset: about 2% of questions are actually irrelevant, and we provide an analysis that some questions in GQA may have more than one

correct answer.

6.1 Future Work

In our presupposition validating method, we are using text-based models to derive semantically whether a presupposition is holding within the context of an image. We used associated scene graph for each image to do that and learn accommodation-based metric to derive a presupposition truth value from a context. However, scene graphs describe the image only partially. A more complete way to perform this validation will require a model to learn such metric in vision and language joint embedding space. This is a task currently pursued under the name of Visual Commonsense Reasoning (VCR), and we expect recent results from VCR will improve that part of our approach.

The dataset we manually collected to assess the truth of a triplet in a scene graph is very small, even though we enriched via different augmentation methods. Since accommodation is a phenomenon related to a given context, the effectiveness of presupposition validation should be studied in a more extensive and large-scale procedure with more diverse contexts.

We have seen evidence that training a VQA system with the additional capability to answer a question as irrelevant can lead to overall increase in performance. However, the effect of robustness to infelicitous questions on the generalization of visio-linguistic compositional reasoning of a VQA system is still not yet to be measured and requires more future work. The recently designed WinoGround dataset will likely help in this aspect.

Finally, we discovered that there are questions in GQA where the annotated answer is not the only possible correct answer. Systems, however, are provided credit for only one answer. More work is required to discover which questions in GQA contain such ambiguity, and what is the set of available answers that a VQA system can provide to answer the question correctly.

Appendix A

Pre Defined Templates

The following lists are the full list of templates used in our pre-defined irrelevant question generation pipeline.

A.0.1 Type 1 + 2 + 3

- Who is [REL] the [OBJ]?
- What is the [SUBJ] [REL]?
- Where is the [SUBJ/OBJ]?

A.0.2 Type 4

- Where is the [SUBJ] [REL] the [OBJ]?
- What does the [S1] [R1] the [O1] [R2]?
- What is the [SUBJ] [REL]?
- Who is [REL] the [OBJ]?

A.0.3 Type 5

- What does the [SUBJ] [REL]?
- Who is [REL] the [OBJ]?

Bibliography

[1] Krishna, Ranjay, Zhu, Yuke, Groth, Oliver, Johnson, Justin, Hata, Kenji, Kravitz, Joshua, Chen, Stephanie, Kalantidis, Yannis, Li, Li-Jia, Shamma, David A, Bernstein, Michael, and Fei-Fei, Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL `https://arxiv.org/abs/1602.07332`.

[2] Hudson, Drew A. and Manning, Christopher D. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019.

[3] Goyal, Yash, Khot, Tejas, Summers-Stay, Douglas, Batra, Dhruv, and Parikh, Devi. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Chen, Feilong, Zhang, Duzhen, Han, Minglun, Chen, Xiuyi, Shi, Jing, Xu, Shuang, and Xu, Bo. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*, 2022.

[5] Xie, Ning, Lai, Farley, Doran, Derek, and Kadav, Asim. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[6] Thrush, Tristan, Jiang, Ryan, Bartolo, Max, Singh, Amanpreet, Williams, Adina, Kiela, Douwe, and Ross, Candace. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[7] Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. VQA:

Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[8] Johnson, Justin, Hariharan, Bharath, Van Der Maaten, Laurens, Fei-Fei, Li, Lawrence Zitnick, C, and Girshick, Ross. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[9] Johnson, Justin, Krishna, Ranjay, Stark, Michael, Li, Li-Jia, Shamma, David, Bernstein, Michael, and Fei-Fei, Li. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[10] Agrawal, Aishwarya, Lu, Jiasen, Antol, Stanislaw, Mitchell, Margaret, Zitnick, C. Lawrence, Batra, Dhruv, and Parikh, Devi. Vqa: Visual question answering, 2016.

[11] Nguyen, Binh X, Do, Tuong, Tran, Huy, Tjiputra, Erman, Tran, Quang D, and Nguyen, Anh. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.

[12] Hudson, Drew and Manning, Christopher D. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Tan, Hao and Bansal, Mohit. Lxmert: Learning cross-modality encoder representations from transformers, 2019.

[14] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Cadene, Remi, Dancette, Corentin, Cord, Matthieu, Parikh, Devi, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

[16] Cho, Jaemin, Lei, Jie, Tan, Hao, and Bansal, Mohit. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[17] Andreas, Jacob, Rohrbach, Marcus, Darrell, Trevor, and Klein, Dan. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

[18] Andreas, Jacob, Rohrbach, Marcus, Darrell, Trevor, and Klein, Dan. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

[19] Hudson, Drew A and Manning, Christopher D. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[20] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Polosukhin, Illia. Attention is all you need, 2017.

[21] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[22] Chen, Yen-Chun, Li, Linjie, Yu, Licheng, El Kholy, Ahmed, Ahmed, Faisal, Gan, Zhe, Cheng, Yu, and Liu, Jingjing. Uniter: Learning universal image-text representations. 2019.

[23] Zhang, Pengchuan, Li, Xiujun, Hu, Xiaowei, Yang, Jianwei, Zhang, Lei, Wang, Lijuan, Choi, Yejin, and Gao, Jianfeng. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[24] Kim, Wonjae, Son, Bokyung, and Kim, Ildoo. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[25] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Radford, Alec, Narasimhan, Karthik, Salimans, Tim, Sutskever, Ilya, et al. Improving language understanding by generative pre-training. 2018.

[27] Wang, Alex, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, and Bowman, Samuel R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[28] Rajpurkar, Pranav, Zhang, Jian, Lopyrev, Konstantin, and Liang, Percy. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[29] Lu, Jiasen, Yang, Jianwei, Batra, Dhruv, and Parikh, Devi. Hierarchical question-image co-attention for visual question answering, 2017.

[30] Lu, Jiasen, Batra, Dhruv, Parikh, Devi, and Lee, Stefan. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. URL `http://arxiv.org/abs/1908.02265`.

[31] Bowman, Samuel R, Angeli, Gabor, Potts, Christopher, and Manning, Christopher D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[32] Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[33] Getty image api. `https://www.gettyimages.com/`.

[34] Hjørland, Birger and Christensen, Frank Sejer. Work tasks and socio-cognitive relevance: A specific example. *Journal of the American Society for Information Science and Technology*, 53(11):960–965, 2002.

[35] Sperber, Dan and Wilson, Deirdre. *Relevance: Communication and Cognition*. Harvard University Press, USA, 1986. ISBN 0674754761.

[36] Sperber, Dan and Wilson, Deirdre. The oxford handbook of contemporary philosophy. In Jackson, Frank and Smith, Michael, editors, *The Oxford handbook of contemporary philosophy*, chapter Pragmatics, pages 468–505. Oxford University Press, 2005.

[37] Grice, Herbert P. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[38] Potts, Christopher. Presupposition and implicature. *The handbook of contemporary semantic theory*, 2:168–202, 2015.

[39] Karttunen, Lauri. Presuppositions of compound sentences. *Linguistic inquiry*, 4(2):169–193, 1973.

[40] Karttunen, Lauri. Presupposition and linguistic context. 1974.

[41] Heim, Irene Roswitha. *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst, 1982.

[42] Ray, Arijit, Christie, Gordon, Bansal, Mohit, Batra, Dhruv, and Parikh, Devi. Question relevance in vqa: Identifying non-visual and false-premise questions, 2016.

[43] Mahendru, Aroma, Prabhu, Viraj, Mohapatra, Akrit, Batra, Dhruv, and Lee, Stefan. The promise of premise: Harnessing question premises in visual question answering, 2017.

[44] Kuper, Yarin. Detecting infelicitous questions in vqa. Master's thesis, Department of Computer Science, Ben Gurion University, 2022.

[45] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[46] Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[47] Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[48] Reimers, Nils and Gurevych, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[49] Koch, Gregory, Zemel, Richard, Salakhutdinov, Ruslan, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

[50] Hadsell, Raia, Chopra, Sumit, and LeCun, Yann. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[51] Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, Liu, Peter J, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[52] He, Pengcheng, Liu, Xiaodong, Gao, Jianfeng, and Chen, Weizhu. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

זיהוי שאלות לא רלוונטיות במענה על שאלות חזותיות באמצעות שינוי הנחות היסוד

ניצן כהן

עבודת גמר לתואר מוסמך למדעי הטבע

אוניברסיטת בן־גוריון בנגב

2023

תקציר

בעבודה זו אנו לומדים את הרלוונטיות של שאלות בהקשר של תמונה. אנו מגדירים זאת מנקודת מבט של הנחות יסוד ומיישמים את תיאוריית הרלוונטיות כדי למדל את המשימה של מענה לשאלות חזותיות במסגרת שיחה. אנו מספקים ניתוח ידני של ההיבטים המאתגרים של ההיבטים המאתגרים של שאלות במערך הנתונים קיים ומסווגים אותן ל־8 מחלקות שונות.

אנו מציגים שיטה חדשה לאמת אם הנחת יסוד מתקיימת בהקשר באמצעות גרפי סצנה, תוך התייחסות לתופעות פרגמטיות וסמנטיות המתעוררות, תוך ציון F של 99 אחוזים במערך המבחן.

אנו מציגים מערך נתונים חדש של שאלות לא רלוונטיות ביחס לתמונה ולכל אחת אנו מציינים את הסיבה לכשל המיועד. מערך הנתונים מכיל 5 סוגי כשלים, אשר כל אחד מהם מיושם בתבניות שונות המוגדרות מראש ובשאלות קיימות ממערך נתונים קיים.

אנו משתמשים במערך הנתונים החדש כדי להדגים את היכולת לזהות שאלות לא רלוונטיות על ידי אימון מודל של ראיה ושפה כדי לזהות שאלות לא רלוונטיות. המודל מצליח לזהות כמעט 2 אחוזים של שאלות לא רלוונטיות ממערך נתונים קיים, אשר לאחר וידוא ידני של התוצאות, הגיע לדיוק של 71 אחוזים.

אוניברסיטת בן־גוריון בנגב
הפקולטה למדעי הטבע
המחלקה למדעי המחשב

זיהוי שאלות לא רלוונטיות במענה על שאלות חזותיות באמצעות שינוי הנחות היסוד

חיבור זה מהווה חלק מהדרישות לקבלת התואר מוסמך למדעי הטבע (M.Sc)

ניצן כהן

בהנחיית פרופסור מיכאל אלחדד

מרץ 2023