Knowledge-Based Counterfactual Queries for Visual Question Answering

Theodoti Stoikou[1], Maria Lymperaiou[1] and Giorgos Stamou[1]

[1]AILS Lab, School of Electrical and Computer Engineering, National Technical University of Athens

Abstract

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

Keywords

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

1. Introduction

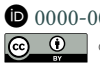
The indisputable rise in popularity of visiolinguistic (VL) learning [1, 2, 3] has offered a variety of impressive model implementations to the community in a short time [4, 5, 6, 7, 8, 9]. Visual Question Answering (VQA) is a VL task that has obtained a fundamental role in the evolution of various interactive VL AI systems, such as Visual Dialogue [10], Text-Image Retrieval [11] and Visual Commonsense Reasoning [12]. To this end, there is an extensive range of real-world applications that benefit significantly from the new advances around the VQA task, such as aiding systems for visually impaired individuals [13, 14] and self-driving cars [15].

VQA involves a textual question q from a pre-defined question set Q accompanied by an image I, the interaction of which yields a textual answer a. The race for continuously advancing VQA model performance unavoidably results in leaving open issues, especially attributed to the black-box nature of state-of-the-art implementations [16, 17, 18, 19, 20]. This limited access to

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023.

dido.stoikou@gmail.com (T. Stoikou); marialymp@islab.ntua.gr (M. Lymperaiou); gstam@cs.ntua.gr (G. Stamou)

0000-0002-3653-0041 (T. Stoikou); 0000-0001-9442-4186 (M. Lymperaiou); 0000-0003-1210-9874 (G. Stamou)



the reasoning that such models follow to make decisions emphasizes the risk of an arbitrary behavior on their behalf. This peril lies mainly in the possibility of bias integration, decisions that lack the proper focus, as well as the absence of explainability and fairness of results. Especially when pivotal decisions are made based on this type of systems, their opacity renders them impractical, and at times hazardous, for most applications. This uncertainty indicates the need for new robustness evaluation methods, that prioritize the transparency of VQA models.

Different approaches to debiasing and explainability of VQA models focus on diverse aspects of the issue. For example, [21] examines VQA robustness and explainability by addressing transformations on the visual modality, as they attribute the problem mostly to the visual bias as occurring from unwanted correlations between image concepts. In general, existing works primarily focus on the effect of *visual bias* rather than the impact of *linguistic bias*, as a reason behind the lack of robustness in VQA models. Other works follow attention-based strategies that require extensive knowledge on the model architecture, thus they cannot handle efficiently the *black-box* nature of these systems. To this end, various model-specific approaches are proven to be fruitful in their strict framework, but lack the capability to be generalized to the evaluation of any other model, thus limiting their efficiency scope to just one specific case.

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

Starting with the observation of *local model responses* in different linguistic perturbations for a single data sample, we further identify *global patterns* that refer to the overall behavior of the model when faced with a specific set of perturbed concepts. Following this, we propose global rules that characterize the response of a model and can underline its weaknesses, by indicating what concepts could harm its robustness and certainty. This process reveals possible biases that a model has integrated, and hence attributes explanations as to why a particular answer is generated in place of another one; thus, we are able to obtain insights to the reasoning process of a model, without the need of access to the model's inner architecture. Our method is generalizable to any VQA model and corresponding suitable dataset, as it approaches the issue in a totally model-agnostic strategy. To sum up, we contribute to the following:

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

q inputs.

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

2. Visual Question Answering (VQA)

Visual Question Answering (VQA), first introduced in [22], lies in the category of multimodal learning tasks, as it receives both visual and linguistic modalities as input. Specifically, a VQA model M receives images i from a set I and relevant questions q belonging to a predefined question set Q, and is expected to accurately answer those q by providing a natural language answer a. The aforementioned answers can be either open-ended (generated by M) or belong to a set of pre-defined candidates A. In general, questions q have an arbitrary nature and they enclose different computer vision sub-problems, such as object recognition and detection, attribute and scene classification, as well as counting [23]. Furthermore, more intricate questions concern more complex processes, such as spatial relationships among objects and commonsense reasoning. According to each specific case, visual questions q selectively target different areas of an image I, including background details and underlying context. In accordance, the focus regarding the linguistic input lies in different word concepts depending on each image-question pair. An example of the VQA task, including an image I and related questions q, as well as the answers a that a VQA model M returns to these questions, is demonstrated in Fig. 1.

Figure 1: Example of image and free-form questions retrieved from the Visual Genome dataset [24], targeted to the VQA task. The displayed answers were given as response by the ViLT model [5].

Our counterfactual approach regarding linguistic substitutions revolves around the following fundamental question: *"What is the response of M if we substitute word X with word Y in question q?"*. The implemented counterfactual $X \rightarrow Y$ substitution should be semantically *minimal* and

linguistically *feasible*. *Minimality* refers to substitutions that maintain a meaning close to the meaning of the original word *X*. For example, synonym words preserve this minimality constraint. In order to ensure semantic minimality of substitutions, we leverage lexical knowledge sources (such as WordNet [25]), which can provide the minimum possible $X \rightarrow Y$ transitions by selecting the closest concept Y to concept X that respects certain constrains. Linguistic *feasibility* instructs meaningful substitutions which always involve the same part of speech (POS); for example, nouns can only be substituted by nouns but not by verbs. In total, such $X \rightarrow Y$ substitutions are applied on the whole Q set, targeting one POS at a time.

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

3. Related work

VQA models Since the introductory work on VQA [22], several endeavors have extended this paradigm, either by suggesting advanced model architectures or by proposing more challenging datasets. State-of-the-art models addressing the VQA task are mainly based on VL transformer backbones; thus, models such as ViLBERT [6], VisualBERT [4], FLAVA [8], ALBEF [9], ViLT [5] and others have dominated the recent VQA literature demonstrating rapid improvements on relevant benchmark datasets. Regarding datasets, improvements on the original VQA (VQA-v2) suggest adding similar image pairs corresponding to the same question q, but leading to diverging answers [20]. Visual Genome (VG) is another large scale dataset including numerous scene images, object, attribute and relationship annotations, as well as visual question-answer pairs [24]. Our approach is tested on both VQA-v2 and VG datasets. Other popular VQA datasets are Flickr30k-Entities [26], COCO-QA [27], Visual7W [28] and others. For a detailed analysis on VQA and relevant topics we refer readers to recent specialized survey papers [1, 2, 3, 29, 30, 31].

Explainability in VQA Regarding the research topic of explainability and robustness in Visual Question Answering [32, 33, 34], multiple efforts have been proposed, including attention maps [35, 36], and other model-specific approaches [37]. The strategy through counterfactuals is a rather new one, while already existing attempts focus on visual perturbations [21], masking [18, 38], introducing counterfactuals in the training stage [39, 38] and relationship-driven approaches between original and counterfactual samples [40].

Linguistic perturbations There is a variety of prior works that perform word-level linguistic perturbations, even though they target purely linguistic tasks, mostly text classification [41, 42, 43, 44, 45], but also semantic similarity [46] and machine translation [47]. Our perturbations regarding synonym replacement and random noun deletion are inspired by [42], guiding substitutions with the usage of WordNet [25]. Color perturbations are adapted from [46],

upon which we construct an appropriate hierarchy based on color distance. The rest of our implemented replacements involving noun and verb substitutions are completely novel ideas.

4. Method

The input of our framework consists of a dataset D that contains aligned images I, textual questions forming a set Q and candidate textual answers A. We will later present results on Visual Genome (VG) [24] and VQA-v2 [20], which satisfy these requirements.

We select ViLT [5] as a proof-of-concept pre-trained VQA model M. Nevertheless, our proposed method is not restricted to ViLT, as it only considers inputs (questions) and outputs (answers). ViLT receives a question $q \in Q$ and an image $i \in I$ from D and then generates an answer a, rather than selecting one of the candidates $a \in A$; since this behavior is inherent to several VQA models, it is important to allow looser definitions of the accuracy metric. To be more precise, ViLT produces outputs a in the form of natural language text, and there are many different ways of expressing the same answer in English. In this case, heuristically comparing the generated answer with the ground truth answer from A defines if the prediction of M is accurate or not. By repeating the same prediction process for all Q, I pairs, and by obtaining successful or unsuccessful answers for them, we finally extract an accuracy score acc_Q, reflecting the ratio of correct answers over all generated answers.

Our word substitutions are guided from external knowledge sources, targeting different parts of speech (nouns, verbs and adjectives) at a time. Specifically, WordNet knowledge graph [25] provides hierarchical relationships between an abundance of common words widely present in VG and VQA-v2 vocabularies. Therefore, substitution pairs are created by connecting specific words with their WordNet matches, respecting hierarchical relationships as described in Section 4.1. Furthermore, we extend the Matplotlib color[2] relationships presented in [46], forming a hierarchy of *color relatedness*. This color hierarchy is based on color distances according to the RGB value of each Matplotlib color, with more details provided in Section 4.1.7

We then proceed with applying our designed perturbations on dataset questions $q \in Q$, resulting in *counterfactual questions* $q^* \in Q^*$. For each substitution, we obtain the *counterfactual question accuracy* acc^*_Q, as the response of M to the counterfactual questions $q^* \in Q^*$, and we compare it to the ground truth accuracy scores acc_Q. Throughout this process, we evaluate whether and how the response of M changes by measuring the difference between acc_Q and acc^*_Q, as an indicator of its robustness against replacements with semantically related concepts.

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

We present a visual outline of our approach in Figure 2. Words from $q \in Q$ colored in red denote the concepts to be substituted, while words in blue indicate the knowledge-driven substitutions that lead to counterfactual questions $q^* \in Q^*$.

[2] Matplotlib colors

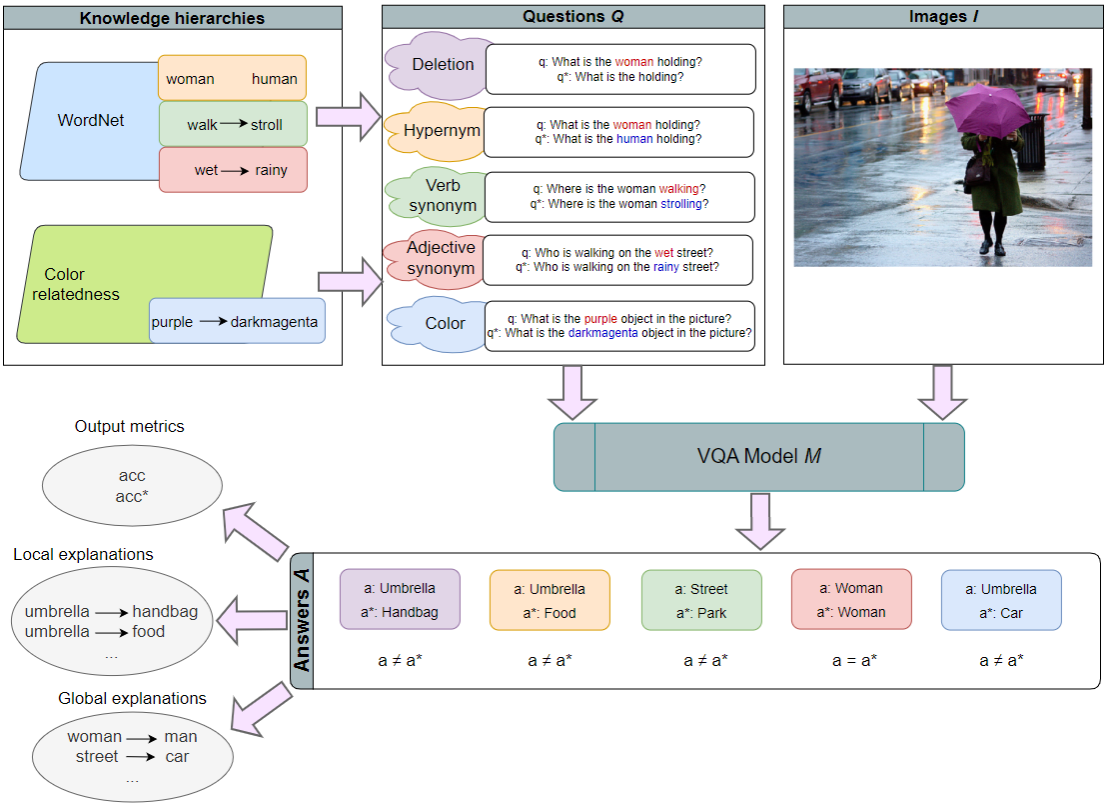


Figure 2: Overview of our proposed knowledge-based counterfactual VQA framework.

4.1. Perturbations

In our work, we perform a variety of substitutions or deletions in the linguistic representation of Q. This counterfactual strategy exploits multiple and diverse morphological attributes of Q and attempts to demonstrate the semantics that most affect the model's response a. Our target is to stimulate an altered response of M through these counterfactual perturbations, in order to identify how the behavior of M changes when faced with different concepts. Thus, we can infer potential biases or points of weak robustness in M. The aforementioned substitutions can be divided into the following categories, based on the knowledge source used and the targeted part of speech. A summary and representative examples of substitutions are provided in Table 1.

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

- **Synonyms:** We employ synonym transformations on adjectives and verbs of the original questions $q \in Q$. For example, "talk" and "speak" are *synonym verbs* according to WordNet, while "small" and "minuscule" are *adjective synonyms*.
- **Hypernyms - Hyponyms:** More *general*, as well as more *specific* noun concepts are provided via WordNet in the form of *hypernyms* and *hyponyms* respectively. For example,

a given noun word (e.g. "dog") we can extract its immediate noun hypernyms (e.g. "canine"), or its immediate hyponyms (e.g. "labrador").
- **Siblings:** We construct noun *sibling* substitutions by traversing the Wordnet knowledge tree one step upwards and then one step downwards. Siblings are defined as noun entities that share the same immediate parent. For example, "carrot" and "radish" are siblings, because they both have "plant root" as their parent concept according to WordNet.

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

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

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

Table 1
Question perturbations examples towards counterfactual queries.

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

Substitutions and deletions are an excellent way to quantify whether a VQA model M understands a specific question-image pair, or if its output is greatly dependent on biased

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

In relation to the **Synonym** substitutions, we aim to investigate the model's ability to efficiently handle mild morphological language alterations that maintain the same meaning. In this case, failing to properly respond (providing an alternative $a^* = a$) would disclose overfitting to specific semantics, which renders the model lexically inflexible and thus non-robust to semantically negligible perturbations. The **Hypernyms-Hyponyms** perturbations are dedicated to depicting the model's ability to generalize and specify correspondingly, while retaining a reliable level of robustness. Hypernym and hyponym relationships are notions profoundly understood in the real world and consequently embedded in large scale datasets, which are widely used for VQA models pre-training. Thus, M should also be able to properly comprehend and reason over them. Ideally, we would expect M to maintain the same response for hypernyms substitutions, whereas justifiably respond in specific ways for the hyponyms replacements, taking into account the specification of meaning. The commensurate amount of specifying skill is sought to be established through **Sibling** substitution experiment. Depending on each particular case, we expect M to modify or maintain its response appropriately, to confirm the level of understanding and distinguishment of different, but still related, meanings.

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

5. Experiments

We present results using the accuracy metric, which illustrates the extent of similarity of the model's predicted answer to the ground truth answer, both for the original Q of each dataset, as well as for the counterfactual question set Q^*. In our analysis, accuracy is not profound enough to provide specific situational explanations and insights on the model's behavior, when faced with particular concepts. However, accuracy still showcases a high-level approach on the model's efficiency fluctuations under the implemented counterfactual perturbations. Since

ViLT model is trained and optimized on the VQA-v2 dataset, it is somehow expected to perform better on it compared to VG (both datasets contain similar vocabularies). This observation is indeed validated by our results presented in Tables 2 & 3, which demonstrate a consistently higher acc_Q on the former versus the latter dataset, concerning all implemented experiments. We denote that acc_Q scores for each experiment contain the corresponding questions only, e.g. color experiments only contain questions that mention colors. This contributes to the differences in original acc_Q scores for each experiment. Nevertheless, in both datasets, we notice an analogous difference between acc_Q and acc^*_Q per experiment when M is presented with counterfactual questions $q^* \in Q^*$. This could generally indicate the existence of underlying biases: M presents a type of overfitting to the original $q \in Q$, which renders it less efficient when asked to handle minimally perturbed counterfactual questions. In all experiments, the accuracy reduction from the original acc_Q to acc^*_Q is approximately 15-20% or more.

An extended depiction of the retrieved accuracies for both datasets is presented in Table 2 (color-based substitutions) and Table 3 (WordNet-based substitutions and noun deletions). Specifically, for **Color Maximal** in VQA-v2 we observe a decline of 34.2% for *common* colors and a decline of 37.4% for *uncommon* ones. Even for semantically minimal substitutions, (**Color Minimal** experiment), the decline is 31% for both *common* and *uncommon* colors. As for VG, we observe a decline of 38.2% for *common* colors and a decline of 52.2% for *uncommon* colors when **Color Maximal** substitutions are performed. Correspondingly, a decline of 35% for both *common* and *uncommon* colors is reported for **Color Minimal** substitutions.

Table 2
Accuracies for color perturbations on VQA-v2 and Visual Genome (VG). Common refers to substitutions with in-dataset colors, while uncommon refers to substitutions involving any Matplotlib color.

	acc_Q%		acc^*_Q% (common)		acc^*_Q% (uncommon)	
Perturbation	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
Color Maximal	69.6	46.9	45.8	29.0	43.6	22.4
Color Minimal	70.0	47.5	48.3	30.9	48.3	30.9

Table 3
Accuracies for WordNet-based perturbations on VQA-v2 and Visual Genome (VG).

	acc_Q%		acc^*_Q%		acc_Q reduction %	
Perturbation	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
Synonym Adjectives	75.1	47.0	56.9	37.4	20.6	20.4
Synonym Verbs	76.8	52.3	64.1	44.4	16.5	15.1
Hypernym Noun	75.2	54.6	60.8	41.5	19.1	24.0
Hyponym Noun	75.1	53.5	56.3	36.1	25.0	32.5
Sibling Noun	76.9	54.0	54.0	33.4	29.8	38.1
Deletion Noun	76.9	53.9	59.1	36.5	23.1	32.3

The discovery of global patterns provides a more profound and targeted view on robustness

of the model M. To this end, we note that in all the cases we studied, we are not interested in the ground truth answer of a question q, but rather in the differentiation between the answer a^* to the counterfactual question in relation to the original answer a that M predicts, either if a is correct or not. We select this approach since we are interested in discovering the model's change in decision-making under the presence of counterfactual inputs, which is more informative than measuring how much a^* semantically deviates from the ground truth response.

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

5.1. Color Maximal explanations

In **Color Maximal** substitutions, we notice that M erroneously maintains the same answer $a^* = a$ when we replace the colors *gray* and *silver* with any other semantically maximal color, either common or uncommon (underlined). Therefore, we detect a bias in the model related to these two colors, as it does not make logical decisions after replacing them with others and does not properly reason over this substitution. A relevant example is presented in Figure 3a.

However, contrary to the above, M logically revises its answers when we replace the colors *green* and *red* with any distant colors, either common or uncommon (underlined) as presented in Figure 3b. Therefore, the model recognizes and qualitatively understands these substitutions and has not incorporated any problematic attachment regarding these two colors.

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

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

Figure 3: Local explanations for **Color Maximal** counterfactual perturbations.

This observation denotes that M is more sensitive towards intense and visually distinct colors and rather bypasses changes involving more neutral ones, focusing on object identities (e.g. "fire hydrant" and "posts" of Figure 3a). A more uncertain a^* (e.g. the model's answer a^* could be "nothing") would be more suitable, if all question semantics were equally taken into account.

5.2. Color Minimal explanations

Based on our experiments on semantically minimal color substitutions, we derive the following global rule: When we replace the colors *gray* and *purple* with any other closely related color, M tends to give the same answer $a^* = a$. Therefore, M does not give due importance to this change of colors, a fact that highlights a robust behavior related to the two aforementioned colors. The model maintains this invariant behavior equally when we perform replacements with *common* colors or with *uncommon* ones, as presented in Figure 4a.

In contrast, M redefines its answers when we replace the *green* color with any other, *common* or *uncommon*, semantically similar color. Consequently, it is being confused by such minimal changes, failing to provide a meaningful answer, as shown in Figure 4b. Even a more uncertain answer (e.g. "nothing") to counterfactual questions would be more suitable compared to the semantically divergent ones returned ("bus" and "bag" instead of "light"). Likewise, M presents a similar change in behavior when we replace the *pink* color with common minimal colors and the *silver* color with uncommon ones.

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

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

Figure 4: Local explanations for **Color Minimal** counterfactual perturbations.

5.3. Synonym Adjectives explanations

In general, adjective-noun pairs present in questions contain some joint special conceptual meaning which differs from the independent meaning of adjectives when they exist autonomously and separately in a sentence. In this case, we notice that the model M varies its answer a^* when we implement a synonym substitution of its question adjectives. This finding suggests that M can qualitatively perceive the meaning of such adjective-noun pairs and differentiate its response accordingly, as presented in Figure 5a.

Another finding is related to the ability of M to correctly adjust its answer, when it is presented with a lexically correct synonym to an adjective, which, however, is not quite appropriate for the given linguistic environment of the question. Accordingly, we conclude that M is capable of understanding the meaning of an adjective in relation to the context of the sentence ("typical food" is meaningful, but "distinctive food" is not), as presented in Figure 5b.

In addition, we derive a global rule that concerns the behavior of M when we replace an adjective having multiple meanings with one of its synonyms, which, although it is optimal with respect to the aforementioned meanings, is however not suitable to the semantic context of the substituted adjectives (such as "delicious" vs "delightful"). We note that in this case, M demonstrates a stable behavior against such substitutions, which proves that it is able to reason over adjectives in a contextualized manner, without being fooled by synonyms not suitable to the exact context of the question q. An example of this observation is provided in Figure 5c.

Size-related adjective substitutions are demonstrated in Figure 5d. We observe that M is particularly robust to such substitutions, therefore correctly capturing the underlying meaning without being biased towards specific words. This global rule demonstrates the flexibility of M towards appropriately handling semantically and contextually equivalent adjective substitutions.

Finally, M is proven to be unstable when it has to handle rare or difficult synonyms of adjectives, as the ones shown in Figure 5e. Consequently, it presents a lexical weakness in handling such rare adjectives and possibly a bias in specific words that are more familiar to it.

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

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

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

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

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

Figure 5: Local explanations for **Synonym Adjectives** counterfactual perturbations.

5.4. Synonym Verbs explanations

Regarding substitutions involving verb synonyms, M is not particularly stable when dealing with substitutions of verbs that present multiple meanings, as presented in Figure 6a. This indicates a difficulty in distinguishing the correct and desired meaning among multiple ones.

An even more specific rule we extract is that M falsely changes its original answer a when we replace the verb "see" with a qualitative synonym of it. A relevant example is provided in Figure 6b. This finding indicates an unwanted attachment of the model to the word "see", which is interpreted as bias. This is an example of a more general situation, where optimal synonyms may not be the best choice for a synonym in a specific contextual setting. In these kinds of instances, the model tends to change its response, as observed in this particular case.

The model M presents satisfactory robustness when it has to deal with easy or common verbs of the English vocabulary, which means that it has acquired a certain degree of versatility in simple vocabulary challenges, as shown in Figure 6c.

Finally, M is rather stable when the replaced verb corresponds to a noun counterpart (e.g.

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

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

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

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

Figure 6: Local explanations for **Synonym Verbs** counterfactual perturbations.

picture -verb-, picture -noun-) or even adjective counterpart (e.g. pictured), such as the ones of Figure 6d. Consequently, M is capable of capturing the general sense of such verbs in the context of the question; equivalent substitution of corresponding nouns (picture→visualization) or adjectives (pictured→visualized) would mostly yield the same counterfactual response a^*.

5.5. Hypernym Noun explanations

Throughout our Hypernym Noun substitutions, we conclude that M is particularly robust against substitutions involving living creatures, such as animals or humans, which shows that it can properly reason over hierarchical relationships governing such concepts, as in Figure 7a.

On the contrary, M does not clearly distinguish between concepts related to types of clothing, i.e. it tends to erroneously change its answer when replaced with a broader concept. Consequently, M does not generalize well on such entities and a bias towards more specific and clear types of clothing emerges. A relevant example is presented in Figure 7b.

As an extension of the above, M exhibits instability in hypernym substitutions that are very broad, inclusive, and polysemous. Therefore, when we replace a noun with an optimal hypernym that presents much greater conceptual generality, M is unable to qualitatively perceive the hierarchical relation that governs them, outputting a wrong answer, as in Figure 7c.

5.6. Hyponyms Noun explanations

Similar to the hypernyms substitution experiment, M is able to appropriately respond in cases where the substitutions of hyponyms refer to living entities. Therefore, the specialization in more specific living entities concepts is properly perceived, as presented in Figure 8a.

Correspondingly, M also shows stability in the substitutions of hyponyms that represent articles of clothing. Therefore, it specializes skillfully in more specific cloth-related entities and according to the case, it appropriately changes its response by adapting to the change. A relevant example is presented in Figure 8b.

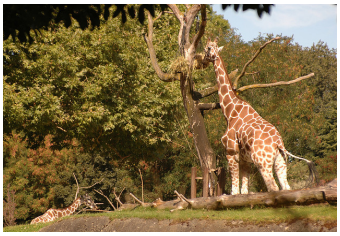
On the contrary, the model does not demonstrate robustness to hyponym substitutions referring to means of transport. Consequently, the model is biased toward such broader concepts and fails to adequately understand their specialization, as shown in Figure 8c.

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

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

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

Figure 7: Local explanations for **Hypernym Noun** counterfactual perturbations.



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

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

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

Figure 8: Local explanations for **Hyponyms Noun** counterfactual perturbations.

5.7. Sibling Noun explanations

With reference to Sibling Noun substitutions, we notice as a global pattern that M has insufficient separation ability when the sibling nouns refer to rooms of buildings or houses. As an example, in Figure 9a, M cannot properly differentiate between the described interior spaces and can be easily fooled by substitutions involving places of different functionality. M is also confused in the case of Figure 9d, when sibling means of transport are substituted. Specifically, M insists on its answer even though a concept not existing in the image appears (bike→truck). This indicates that M rather trespasses the linguistic modality context, providing an 'easy' answer based on the visual modality, since the only *bird* appearing in the image is a *parrot*. In this case, the relevant position of the bird *on the man's bike/truck* is ignored.

An interesting behavior is observed when sibling concepts involving animals are tested, as in Figure 9e. In this case, M seems to circumvent reasoning over the image, providing an answer based on knowledge it has most possibly acquired during its pre-training phase (zebras are black and white in color). Nevertheless, M is not fooled by the horse→zebra substitution, in which case it would conclude that *the zebra is brown*, which is a wrong factual statement. Another case that M is not being fooled is depicted in Figure 9c. In this case, M is very consistent in sibling entities that declare human body parts, which means that it correctly perceives their

differences and does not group them in an arbitrary way. Furthermore, M presents a correct reasoning process by differentiating its answer when the sibling concepts present very different meanings between them, as the concepts air→water in Figure 9b.

Overall, Siblings Noun substitution provided a rich set of insights, unequally relying on either q or I to derive an answer in many cases, rather than providing an uncertain outcome (such as answering "nothing" in the examples of Figures 9d, 9e, similarly to the correct reasoning of Figure 9c). In total, this indicates an unstable behavior of M towards different sibling pairs, yielding unpredictable outcomes under different substitutions of the same conceptual distance.

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

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

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

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

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

Figure 9: Local explanations for **Sibling Noun** counterfactual perturbations.

5.8. Deletion Noun explanations

Regarding the counterfactual questions concerning deletions of nouns, we firstly observe the following pattern: When the deleted noun has a determining role in another noun already present in the question, M maintains its original answer even after the deletion. Therefore, we detect a tendency towards attaching to the determined noun, while at the same time not paying due attention to the determiner noun. Hence, M answers such questions arbitrarily, even though its answer cannot be perceived as wrong; a human could have also answered the same, especially in yes/no questions. A related example is provided in Figure 10a.

Another pattern that we detect concerns questions that refer to the color of a noun, which has been deleted. In these cases, M tends to respond with the most dominant color in the image, without taking into account the absence of the noun that this color should define, as in Figure 10b. Of course, we regard this behavior as justified, since a human would most probably answer such questions in the same way. Similarly, in questions concerning the location of a noun, which has been deleted, M answers with the most dominant entity present in the given image. A relevant example is demonstrated in Figure 10c.

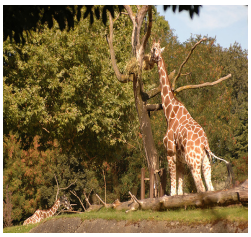
Finally, we list some nouns to which we notice that M does not pay due attention when asked to give an answer, as in Figure 10d. Specifically, even after deleting them, M tends to return the initial answer with great frequency. These words are: image, photographs, human, man, animal, room. In general, these are words that are encountered very often in questions and usually act in addition to other, more specific, entities. Once again, this behavior is justified.

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

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

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

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



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

Figure 10: Local explanations for **Deletion Noun** counterfactual perturbations.

6. Conclusion and Future Work

Counterfactual perturbations in VQA models can provide novel and useful insights regarding model robustness and explainability of results. In our work, we propose a knowledge-based counterfactual framework targeting substitutions on questions. Specifically, our framework suggests multiple types of word-level linguistic transformations in order to probe selected VQA models in a black-box fashion, and investigate whether the presence of counterfactual questions will lead to unexpected model responses. Through this process, underlying linguistic biases are revealed, while informative explanations regarding the model's behavior are provided, by extracting global rules in a qualitative manner, ultimately depicting those existing biases. Our results on Visual Genome and VQA-v2 datasets, using ViLT model as proof of concept, illustrate the merits for our approach, highlighting concepts that incite model biases, in a model-agnostic manner. As an immediate extension of our method, we aim to apply the same linguistic perturbations on dataset answers, addressing VQA models that reason over multiple choice answers. As future work, we plan to expand our approach to other related visiolinguistic tasks, such as Text - Image Retrieval, Visual Entailment, and Visual Commonsense Reasoning, while another direction involves crafting counterfactual perturbations targeting the visual modality.

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number 5537).

References

[1] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, Journal of Artificial Intelligence Research 71 (2021) 1183–1317. URL: http://dx.doi.org/10.1613/jair.1.11688. doi:10.1613/jair.1.11688.

[2] Y. Du, Z. Liu, J. Li, W. Zhao, A survey of vision-language pre-trained models (2022).

[3] M. Lymperaiou, G. Stamou, A survey on knowledge-enhanced multimodal learning, ArXiv abs/2211.12328 (2022).

[4] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019. URL: https://arxiv.org/abs/1908.03557. doi:10.48550/ARXIV.1908.03557.

[5] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, 2021. URL: https://arxiv.org/abs/2102.03334. doi:10.48550/ARXIV.2102.03334.

[6] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL: https://arxiv.org/abs/1908.02265. doi:10.48550/ARXIV.1908.02265.

[7] V. Kazemi, A. Elqursh, Show, ask, attend, and answer: A strong baseline for visual question answering, 2017. URL: https://arxiv.org/abs/1704.03162. doi:10.48550/ARXIV.1704.03162.

[8] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, Flava: A foundational language and vision alignment model, 2021. URL: https://arxiv.org/abs/2112.04482. doi:10.48550/ARXIV.2112.04482.

[9] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, S. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, 2021. URL: https://arxiv.org/abs/2107.07651. doi:10.48550/ARXIV.2107.07651.

[10] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, G. W. Taylor, Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction, 2019. arXiv:1811.09845.

[11] S. R. Dubey, A decade survey of content based image retrieval using deep learning, IEEE Transactions on Circuits and Systems for Video Technology (2021) 1–1. URL: http://dx.doi.org/10.1109/TCSVT.2021.3080920. doi:10.1109/tcsvt.2021.3080920.

[12] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, 2019. arXiv:1811.10830.

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

[14] C. Chen, S. Anjum, D. Gurari, Grounding answers for visual questions asked by visually impaired people, 2022. URL: https://arxiv.org/abs/2202.01993. doi:10.48550/ARXIV.2202.01993.

[15] H. Ben-Younes, Ã‰loi Zablocki, P. PÃ©rez, M. Cord, Driving behavior explanation with multi-level fusion, Pattern Recognition 123 (2022) 108421. URL: https://

www.sciencedirect.com/science/article/pii/S0031320321005975. doi:https://doi.org/10.1016/j.patcog.2021.108421.

[16] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, D. Parikh, Rubi: Reducing unimodal biases in visual question answering (2019). URL: https://arxiv.org/abs/1906.10169. doi:10.48550/ARXIV.1906.10169.

[17] V. Agarwal, R. Shetty, M. Fritz, Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing, 2019. URL: https://arxiv.org/abs/1912.07538. doi:10.48550/ARXIV.1912.07538.

[18] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, Y. Zhuang, Counterfactual samples synthesizing for robust visual question answering, 2020. URL: https://arxiv.org/abs/2003.06576. doi:10.48550/ARXIV.2003.06576.

[19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, 2017. URL: https://arxiv.org/abs/1707.07998. doi:10.48550/ARXIV.1707.07998.

[20] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2016. URL: https://arxiv.org/abs/1612.00837. doi:10.48550/ARXIV.1612.00837.

[21] Z. Boukhers, T. Hartmann, J. Jürjens, Coin: Counterfactual image generation for vqa interpretation, 2022. URL: https://arxiv.org/abs/2201.03342. doi:10.48550/ARXIV.2201.03342.

[22] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, Vqa: Visual question answering, 2015. URL: https://arxiv.org/abs/1505.00468. doi:10.48550/ARXIV.1505.00468.

[23] K. Kafle, C. Kanan, Visual question answering: Datasets, algorithms, and future challenges, Computer Vision and Image Understanding 163 (2017) 3–20. URL: https://doi.org/10.1016%2Fj.cviu.2017.06.005. doi:10.1016/j.cviu.2017.06.005.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, F.-F. Li, Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL: https://arxiv.org/abs/1602.07332. doi:10.48550/ARXIV.1602.07332.

[25] C. Fellbaum, Wordnet: An electronic lexical database (1998).

[26] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2641–2649. doi:10.1109/ICCV.2015.303.

[27] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, 2015. arXiv:1505.02074.

[28] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, 2016, pp. 4995–5004. doi:10.1109/CVPR.2016.540.

[29] Y. Zou, Q. Xie, A survey on VQA: Datasets and approaches, in: 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE, 2020. URL: https://doi.org/10.1109%2Fitca52113.2020.00069. doi:10.1109/itca52113.2020.00069.

[30] M. Banchhor, P. Singh, A survey on visual question answering, in: 2021 2nd Global Confer-

ence for Advancement in Technology (GCAT), 2021, pp. 1–5. doi:10.1109/GCAT52182.2021.9587797.

[31] H. Sharma, A. S. Jalal, A survey of methods, datasets and evaluation metrics for visual question answering, Image and Vision Computing 116 (2021) 104327. URL: https://www.sciencedirect.com/science/article/pii/S0262885621002328. doi:https://doi.org/10.1016/j.imavis.2021.104327.

[32] A. Panesar, F. I. Doğan, I. Leite, Improving visual question answering by leveraging depth and adapting explainability, in: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 252–259. doi:10.1109/RO-MAN53752.2022.9900586.

[33] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, G. Burachas, A study on multimodal and interactive explanations for visual question answering (2020). URL: https://arxiv.org/abs/2003.00431. doi:10.48550/ARXIV.2003.00431.

[34] J.-H. Huang, M. Alfadly, B. Ghanem, M. Worring, Assessing the robustness of visual question answering models, 2019. URL: https://arxiv.org/abs/1912.01452. doi:10.48550/ARXIV.1912.01452.

[35] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, 2016. URL: https://arxiv.org/abs/1606.00061. doi:10.48550/ARXIV.1606.00061.

[36] M. Jiang, S. Chen, J. Yang, Q. Zhao, Fantastic answers and where to find them: Immersive question-directed visual attention, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2977–2986. doi:10.1109/CVPR42600.2020.00305.

[37] F. Sammani, T. Mukherjee, N. Deligiannis, Nlx-gpt: A model for natural language explanations in vision and vision-language tasks, 2022. URL: https://arxiv.org/abs/2203.05081. doi:10.48550/ARXIV.2203.05081.

[38] L. Chen, Y. Zheng, Y. Niu, H. Zhang, J. Xiao, Counterfactual samples synthesizing and training for robust visual question answering, 2021. URL: https://arxiv.org/abs/2110.01013. doi:10.48550/ARXIV.2110.01013.

[39] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, A. van den Hengel, Counterfactual vision and language learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10041–10051. doi:10.1109/CVPR42600.2020.01006.

[40] Z. Liang, W. Jiang, H. Hu, J. Zhu, Learning to contrast the counterfactual samples for robust visual question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3285–3292. URL: https://aclanthology.org/2020.emnlp-main.265. doi:10.18653/v1/2020.emnlp-main.265.

[41] S. Ren, Y. Deng, K. He, W. Che, Generating natural language adversarial examples through probability weighted word saliency, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1085–1097. URL: https://aclanthology.org/P19-1103. doi:10.18653/v1/P19-1103.

[42] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019. URL: https://arxiv.org/abs/1901.11196. doi:10.48550/ARXIV.

1901.11196.

[43] S. Garg, G. Ramakrishnan, BAE: BERT-based adversarial examples for text classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6174–6181. URL: https://aclanthology.org/2020.emnlp-main.498. doi:10.18653/v1/2020.emnlp-main.498.

[44] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, Y. Qi, Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, 2020. URL: https://arxiv.org/abs/2005.05909. doi:10.48550/ARXIV.2005.05909.

[45] A. Karimi, L. Rossi, A. Prati, AEDA: An easier data augmentation technique for text classification, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2748–2754. URL: https://aclanthology.org/2021.findings-emnlp.234. doi:10.18653/v1/2021.findings-emnlp.234.

[46] M. Lymperaiou, G. Manoliadis, O. M. Mastromichalakis, E. G. Dervakos, G. Stamou, Towards explainable evaluation of language models on the semantic similarity of visual concepts, 2022. URL: https://arxiv.org/abs/2209.03723. doi:10.48550/ARXIV.2209.03723.

[47] J. Wan, J. Yang, S. Ma, D. Zhang, W. Zhang, Y. Yu, Z. Li, Paeg: Phrase-level adversarial example generation for neural machine translation, 2022. URL: https://arxiv.org/abs/2201.02009. doi:10.48550/ARXIV.2201.02009.