Question Modifiers in Visual Question Answering

William Britton, Somdeb Sarkhel, Deepak Venugopal
University of Memphis, Adobe Research, University of Memphis
wjbrittonv@gmail.com, sarkhel@adobe.com, dvngopal@memphis.edu

Abstract

Visual Question Answering (VQA) is a challenge problem that can advance AI by integrating several important sub-disciplines including natural language understanding and computer vision. Large VQA datasets that are publicly available for training and evaluation have driven the growth of VQA models that have obtained increasingly larger accuracy scores. However, it is also important to understand how much a model understands the details that are provided in a question. For example, studies in psychology have shown that syntactic complexity places a larger cognitive load on humans. Analogously, we want to understand if models have the perceptual capability to handle modifications to questions. Therefore, we develop a new dataset using Amazon Mechanical Turk where we asked workers to add modifiers to questions based on object properties and spatial relationships. We evaluate this data on LXMERT which is a state-of-the-art model in VQA that focuses more extensively on language processing. Our conclusions indicate that there is a significant negative impact on the performance of the model when the questions are modified to include more detailed information.

Keywords: visual question answering, modifiers, deep models, perception

1. Introduction

Visual Question Answering (VQA) is a highly challenging multi-disciplinary task that requires integration of several key disciplines including natural language understanding, computer vision and knowledge representation. Specifically, in VQA, a system is required to answer questions (that may be open-ended) that are based on a specific image. Due to its multi-disciplinary nature, it can be argued that advances made in this task can potentially be a significant step forward for AI in general.

Based on the VQA benchmark dataset (Antol et al., 2015), several new models for VQA have been developed over the past few years (Goyal et al., 2017; Selvaraju et al., 2020; Tan and Bansal, 2019). While the goal is to enhance both visual and language understanding, it was observed in some cases that due to the type of questions, some priors based on language can significantly improve accuracy. For example, as pointed out in (Goyal et al., 2017), a simple prior such as assigning the answer "yes" a high probability for a question that begins with "do you see" can often achieve very high scores. However, the use of such priors do not reflect true understanding and are unlikely to generalize to real-world settings. To address this, in (Goyal et al., 2017), a *balanced* VQA dataset was developed to evaluate systems. Here, a large dataset was collected using Amazon Mechanical Turk where, the same question has different answers for different images. Further, the images with complementary answers are selected such that they are similar to each other. Thus, a model needs to recognize subtle visual characteristics in the image to answer get both the complementary answers correct. This shared benchmark dataset commonly known as VQA2.0 forces models to pay more attention to visual understanding in question answering. Using this dataset, a more explainable VQA model was developed in (Goyal et al., 2017) through the use of counterexamples. That is, the model answers the question and at the same time chooses counterexample images. Thus, a system that truly understands the image must also be able to come up with examples that are meaningful but ones that do not answer the question. For example, for a question about a red flower, a white flower image may be a suitable counterexample.

Figure 1: Example Question: Where is the child sitting? More detailed questions: Where is the child who is holding a bottle sitting? Where is the child drinking from a bottle? All three questions have the same answer but the model needs to understand the details that are provided in the question.

In this paper, our goal is to evaluate the sensitivity of a VQA model to details specified within a question. Typically, when the question adds modifiers, the model needs to reason about more information to arrive at the correct answer. In general, from studies in psychology, it is known that executive function (ie, working memory, inhibition, planning) contributes to decoding

and reading comprehension (Nouwens et al., 2021), and that syntactic complexity increases neural computational demand (Just et al., 1996). At the same time, additional modifiers in a question could also resolve ambiguity which in turn can help the VQA model. For instance, in the example shown in Fig. 1, the child holding the bottle and the child sitting are likely to be the key visual details that can be extracted (i.e., the parts of the image with more attention if we consider attention-mechanisms (Lu et al., 2016)). For a question such as *where is the child sitting?* that does not specify that the child is holding a bottle, a model can easily ignore the visual attention on this. However, when the question specifies this detail, as in *where is the child who is holding the bottle sitting* or *where is the child drinking from a bottle*, the model should be more selective and choose the correct visual details to answer the question. Further, a question such as *Is the child drinking milk?* may be ambiguous, however a more detailed question such as *Is the child drinking milk from a bottle?* can help the model since the visual representation can focus on the bottle. Further, modifiers could also help identify cases where a model is using incorrect reasoning even if it produces the right answers. For instance, suppose a model answers "yes" for *is there a child wearing red sitting?* correctly but also answers "yes" for *is there a child wearing blue sitting?* incorrectly, this means that it may be using incorrect reasoning to answer the question. That is, once a child and the activity of sitting is detected, the model defaults to answering "yes" without evaluating additional details in the question. While pragmatically, *Gricean maxims* (Grice, 1975) would dictate that typically human questions would not contain superfluous information such as "holding the bottle" when there is only one child, the purpose of this work is to test the sensitivity of a model to specific types of modifiers. Therefore, although many of these modified questions may contain more information than simply necessary to answer them, this additional information allows for the testing of the model's ability to handle new concepts added to a question (which would be necessary in discerning between two similar objects in a scene with different properties or relations).

In this work, we develop a dataset where a question is rephrased with modifications added to the question. The modifiers are added to generate question pairs, where one question in the pair is answer-preserving and the other changes the original answer to the question. To generate these modified questions, we use Amazon Mechanical Turk (AMT) workers. We ask them to add different types of modifiers, namely, modifiers w.r.t properties of objects or modifiers w.r.t object relationships in the image. Using this dataset, we evaluate the well-known LXMERT (Tan and Bansal, 2019) model for VQA. Our evaluation indicates that there is a significant difference between the model performance with and without modifiers added to the questions. Further, for questions that require a yes/no answer, adding modifiers shows a smaller degradation of performance as compared to non-binary answers.

2. Related Work

The original VQA dataset (Antol et al., 2015) was the first large dataset for open-ended answers with over 200K images. To overcome some of the biases, where simple language priors could artificially yield very high accuracy, the balanced VQA dataset or the VQA2.0 dataset (Goyal et al., 2017) was developed subsequently. Both these datasets were collected from AMT and consist of both natural and abstract scenes. The model developed for the original VQA dataset combined LSTMs to process the text with CNNs to process the image specified in the question. On VQA2.0 this model along with other models such as those using hierarchical co-attention (Lu et al., 2016) for the questions and images, and bilinear pooling (Fukui et al., 2016) (which was the winner for the 2016 VQA dataset challenge) gave significantly worse results. LXMERT (Tan and Bansal, 2019) that is based on the BERT architecture that popularized attention-mechanisms for language understanding is one of the state-of-the-art systems for VQA. It uses cross-modality training to capture interaction between the language and visual elements. The evaluation proposed in (Selvaraju et al., 2020) is related to our approach in principle, i.e., sub-questions are generated to augment questions. Specifically, in this work, reasoning and perception sub-questions are generated through AMT. The model is therefor evaluated on its ability to perform higher-level reasoning consistently. Another type of approach that has been used is to annotate attentions in the data (Das et al., 2016). Specifically, using human workers, the important aspects required to answer questions is annotated in the data. The model is therefore evaluated on whether it is using the right reasons to arrive at its answer. Yet another approach in (Park et al., 2018) annotated the natural language explanations for an answer. However, this is harder to use since for most neural network based models (which are the ones that give state-of-the-art performance),it is hard to convert the reasoning used by the model to natural language. Our work is also related to studies in image specificity (Jas and Parikh, 2015) where the goal is to know and predict if descriptions of an image are likely to be specific or not. In our case, we use a related idea, however, our goal is to understand if models are indeed capable of processing specific information which is one way of assessing if they are understanding the information that is provided to them. Finally, our work is also related the study of perturbations in questions (Khashabi et al., 2020) that has been used to analyze robustness in question answering. Here, the perturbations for questions are related to modifiers that are related to perception in VQA (Selvaraju et al., 2020).

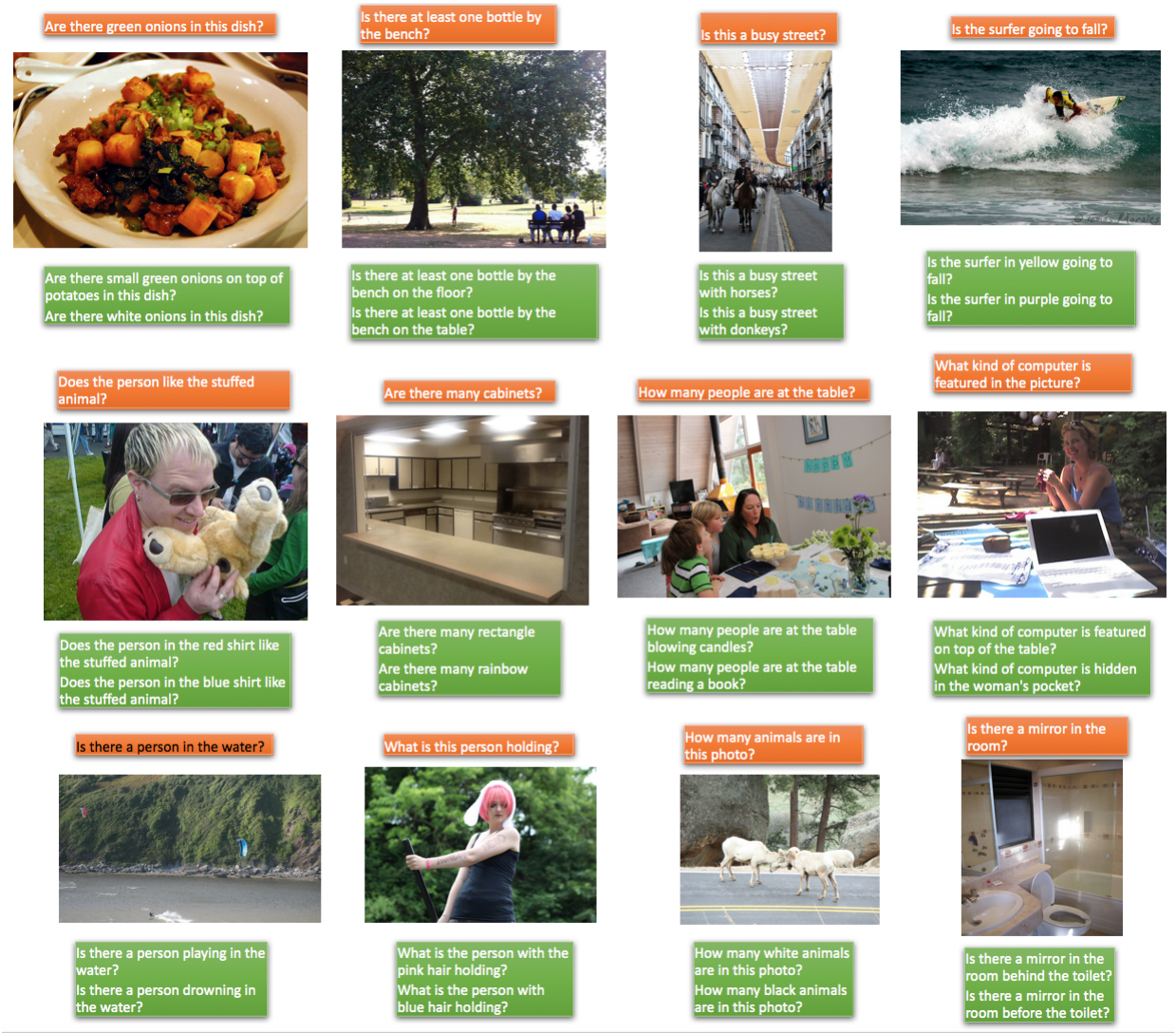
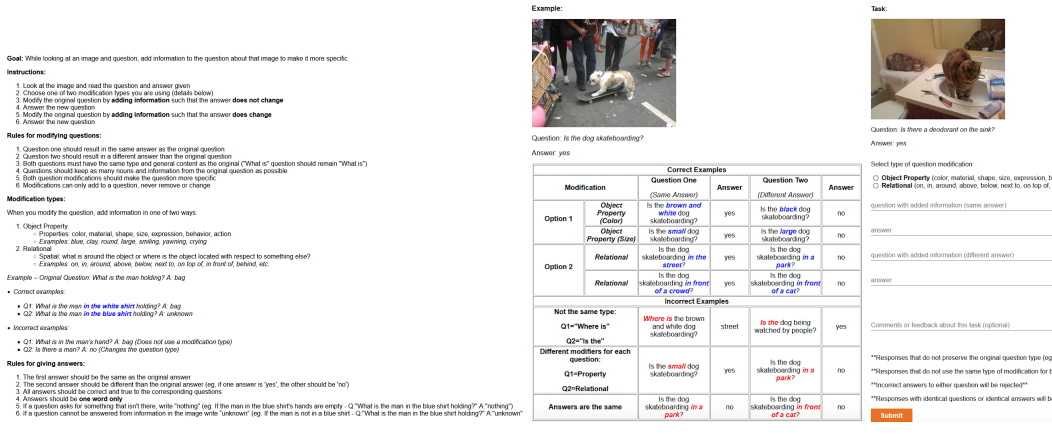


Figure 2: Samples from the dataset. The question in orange (placed above an image) is the original question and the questions in green (below an image) are the modified questions. The first modified question has the same answer as the original while the second one has a different answer.



(a) (b) (c)

Figure 3: The MTurk Interface.

$\lvert\mathbf{Q}\rvert$	$\lvert\mathbf{Q}_b\rvert$	$\lvert\mathbf{Q}_w\rvert$	$\lvert\mathbf{Q}_r\rvert$	$\lvert\mathbf{Q}_o\rvert$
3065	1968	1564	1105	2427

Table 1: Summary of data collected from MTurk. $\mathbf{Q}$ denotes the set of all questions that remain after quality validation, where each element of $\mathbf{Q}$ is a triple (q_0, q_1, q_2), q_0 is the original question, q_1 the modified question that preserves the answer of q_0 and q_2 is the modified question that changes the answer. $\mathbf{Q}_b \subset \mathbf{Q}$ is the set of questions with binary answers (yes/no), $\mathbf{Q}_w \subset \mathbf{Q}$ is the set of questions with non-binary answers, $\mathbf{Q}_r \subset \mathbf{Q}$ is the set of questions with relational modifiers and $\mathbf{Q}_o \subset \mathbf{Q}$ is the set of questions with object modifiers.

3. Question-Modification Evaluation

3.1. Dataset Creation

For every question q, we generated a pair of questions with modifiers added to q, one which preserved the same answer as q and the other which changed the answer. We consider two types of modification. The first one in which we specify a property for an object in the image and the other where we specify a relationship in the image. Object property modifiers identify additional details about an object in the image/question such as color, material, shape, size, expression, or behavior. Relational modifiers identify relative spatial relationships between an object and another object or its surroundings. These typically result in prepositional phrases such as *on, in, around, above, below, next to, in front of, behind, etc.* plus their object. Object properties and spatial relationships are well known to be highly significant for perception and thus are fundamental for VQA (Selvaraju et al., 2020; Bansal et al., 2020). Furthermore, questions can be modified based on object properties and spatial relationships by human workers without requiring significant expertise and therefore, we are likely to have a high quality dataset using AMT. Sample images and questions from the dataset are shown in Fig. 2. We used AMT to generate 3065 question pairs (6130 total questions) [Tab. 1] where the ratios of questions of each of the 21 types specified in (Goyal et al., 2017) (e.g., "Is there", "How many", "What is", "What color") is approximately equal to their ratios in the original data. We selected questions from the training partition of the VQA data due to the fact that ground truth answers are not available for the test set. As with the VQA dataset, all generated questions are open-ended in that answer choices are not provided. For simplicity and to reduce confounding factors, only questions that had one word answers (83.5% of questions in the VQA data) were selected as sample candidates and AMT workers were instructed to provide one word answers. As the new rephrased questions are not the same as the training set questions, the model did not have answers to the rephrased questions available. The interface we used for AMT is shown in Fig. 3. The dataset with the questions is available here [1] If we let q_0 be the original question from the training set, q_1 be the modified question yielding the same answer as q_0, q_2 be the modified question yielding a different answer than q_0, (a_0,a_1,a_2) be the answers to (q_0,q_1,q_2), and T_q be the question type for some q, the quality validation performed on the modified questions from AMT discarded all samples that met the following criteria:

1. $(T_{q_0} \neq T_{q_1}) \vee (T_{q_0} \neq T_{q_2}) \vee (T_{q_1} \neq T_{q_2})$
2. $a_0 \neq a_1$
3. $a_0 = a_2$
4. $q_0 = q_1 \vee q_1 = q_2$.

Due to the complexity of the images and the specific modification task asked of workers, many q_2 asked for information not available in the image resulting in an answer of "unknown". Since these questions differ fundamentally from questions that have answers determinable from the image, each "unknown" q_2 and its corresponding q_1 were removed from the main analysis for a separate analysis resulting in set Q_u. There are, however, implications for these "unknown" questions which will be discussed. In total, 1299 q_2 had "unknown" answers resulting in 1766 question pairs (3532 total questions) for the main analysis. Unless otherwise stated, all analysis will be for the dataset with "unknown" q_2 and their corresponding q_1 removed.

Original (q_0)	Modified (q_1)	Modified (q_2)	$\delta(q_0, q_2)$
0.647	0.61	0.442	0.453

Table 2: Evaluating influence of question modifiers. Accuracy for LXMERT is computed for the original questions (q_0) and both the answer preserving (q_1) and non answer preserving (q_2) rephrased questions. $\delta(q_0, q_2)$ is a measure of the percentage of instances for which the answer predicted for q_0 is different from that predicted for q_2. Larger values of $\delta(q_0, q_2)$ are better since they indicate that the model understands that it needs to change its prediction for the modified question q_2.

$\mathbf{Q}_r$	$\mathbf{Q}_o$	$\mathbf{Q}_b$	$\mathbf{Q}_w$
0.497	0.540	0.628	0.398

Table 3: Evaluating accuracy for modifiers and answer types.

3.2. Evaluation

For our evaluation, we used LXMERT which yields state-of-the-art results for the VQA task. We used the pre-trained model for LXMERT. In order to set a baseline and assure the sample retained accuracy from

[1] https://drive.google.com/file/d/11plw7P82ew9AkuJ2q9zDorVjlzhyg_iP/view?usp=sharing

the original dataset, the LXMERT pre-trained model was run for the full one word answer dataset yielding 65.0% accuracy. The pre-trained model was then run for the 3065 original q_0 from our sample and obtained 64.7% accuracy. Finally, the pre-trained model was run for all q_1 and q_2. Model performance was measured for predictions of all q_1 against the original q_0, and accuracy is reported for all q_2. Additionally, performance was measured in 2x2 dimensions: modification type (relational vs object property) and answer type (yes/no vs other). It should be noted that while "answer type" is used to refer to these two categories, each refer to a subset of question types (eg. "Is there" questions yield "yes/no" answers), therefore, this is a meaningful dimension on which to measure modified questions. McNemar's exact test (Fagerland et al., 2013) is commonly used in evaluating paired binomial data in medicine, but also sees use in machine learning when evaluating if classification algorithms perform differently from each other (Dietterich, 1998). McNemar's exact test measures marginal homogeneity between paired samples. That is, whether success or failure is more likely in one condition or another. Therefore, for all q_1, significance of accuracy change from the original q_0 was measured using McNemar's exact test. Accuracy differences, that is, overall performance on between categories in each dimension, such as between the object property and relational modification types and between "yes/no" answer types (binary questions) and "other" answer types (non-binary questions) were measured using independent samples t-test.

3.2.1. Overview

For all $q \in \mathbf{Q}$, the model performed with an accuracy of 52.6%. For q_1, that is, rephrased questions resulting in the same answer as q_0, model accuracy dropped to 61.0% from the original 64.7% (McNemar p=0.0003), while for q_2 accuracy was 44.2% [Tab. 2]. That is, adding modifiers to the questions resulted in a significant accuracy drop. Additionally, question modification caused 22.7% of all predicted answers to q_1 to change. Looking at whether adding modifiers to the questions helped (changed incorrect q_0 prediction to correct q_1) or hurt (changed correct q_0 prediction to incorrect q_1 prediction), 11.4% of correct answers (7.0% of all q_1 predictions) were changed from an incorrect prediction for q_0 to a correct prediction for q_1, while 27.3% of incorrect answers (10.6% of all q_1 predictions) were changed from an originally correct prediction for q_0 to an incorrect prediction for q_1. Next, modifying questions by adding details to yield a different answer than the original (q_2) caused 45.4% of answers to change from the original prediction for q_0. That is, over half the predicted answers to q_2 stayed the same as the original prediction for q_0.

3.2.2. Modification Type

In this section, we discuss how modification type affects model performance. In general, 553 question pairs modified object property, while 1213 question pairs used relational modification. Overall accuracy for questions modifying object property was 54.0%. More specifically, q_1 accuracy fell to 62.6% from an original q_0 accuracy of 65.2% (McNemar p=0.037), while q_2 accuracy was 45.3%. On the other hand, overall accuracy for questions using relational modification was 49.7%. Model accuracy for q_1 was 57.6% compared to an original q_0 accuracy of 63.6% (McNemar p=0.006), while accuracy for q_2 was 41.8%. That is, when either modification type was added to a question, model performance fell. We also compared model performance and behavior (such as changing answers from the q_0 answer) between the two modification types. In general, the model seems to handle object property modification (54.0%) significantly better (t-test p=0.046) than relational modification (49.7%). Lastly, when the model had the answer correct for q_1 the answer changed for q_2 45.5% of the time for object property modification and 42.1% of the time for relational modification although this difference was not significant (t-test p=0.307), suggesting that the model does not tend to alter its correct q_1 answers when answering q_2 more frequently for either of the modification types. Tab. 3 summarizes the accuracy obtained for the two modification types and Fig. 4 shows the corresponding McNemar statistic and significance scores.

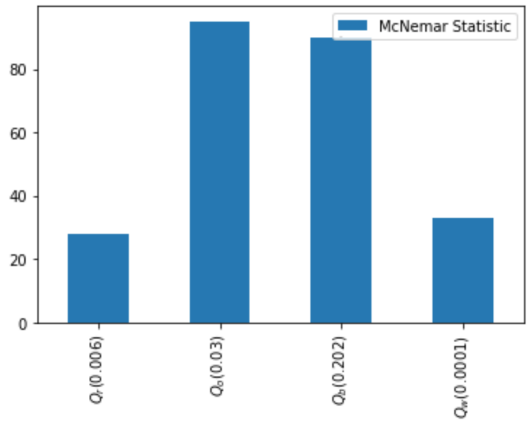


Figure 4: Evaluating the significance of question modifiers. The McNemar statistic and p-value is computed to compare the original predictions with the predictions made when modifiers are added. $\mathbf{Q}_r$ indicates that the significance of adding relational modifiers is measured and $\mathbf{Q}_o$ indicates that significance of adding object modifiers are measured. Similarly, for answer types, $\mathbf{Q}_b$ indicates that the significance adding modifiers to yes/no answer questions is measured and $\mathbf{Q}_w$ indicates the same measure for non-binary answers. The p-values are shown in brackets.

3.2.3. Answer Type

Next, we will evaluate how well the model handles modification in the context of answer type. For binary

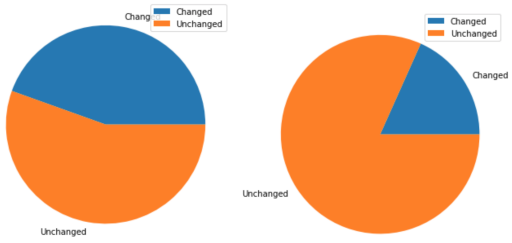
	Relational Modifiers	Object Modifiers
Yes/No Answers	0.655	0.697
Open-ended Answers	0.475	0.537

Table 4: Evaluating the joint influence of modifiers and answer types. The accuracy values for LXMERT are shown for each case.

(yes/no) questions, the model obtained an overall accuracy of 62.8%. While there was an observed drop in accuracy of binary q_1 answers (68.4%) from the original q_0 answers (70.3%), this was not significant (McNemar p=0.202), suggesting the model is more resilient at answering yes/no questions when more detailed information is added to the original question. As with modification type and other q_2 observations in general, accuracy for q_2 did fall for binary questions (57.2%), albeit to a lesser extent. Conversely, the model does not do as well on non-binary questions (39.8% total) and performance for q_1 (51.8%) decreased (McNemar p<0.0001) from q_0 performance (57.7%). Additionally, non-binary q_2 accuracy (27.9%) was the lowest amongst the four dimensions. Between answer types, the performance difference was significant (t-test p<0.0001) which does suggest that, as expected, the model is better at answering questions with binary possible answers. Tab. 3 summarizes the accuracy obtained for the two answer types and Fig. 4 shows the corresponding McNemar statistic and significance scores.

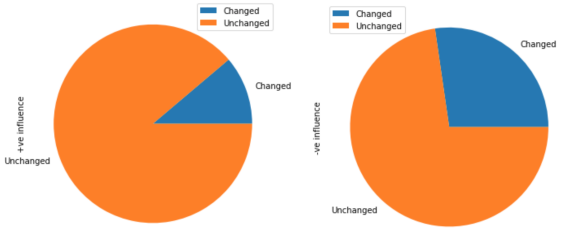
3.2.4. Joint Influence of Modification Type and Answer Type

An additional evaluation of the model at the intersection of modification type and answer type was also conducted. First, the intersection of object property and binary questions was evaluated yielding 64.3% general accuracy, 69.7% for q_1 that was not a significant drop (McNemar p=0.805) from q_0 accuracy (70.3%), and 58.8% for q_2 accuracy. Not surprisingly, given the earlier analysis, the model tended to perform best on both questions modified with object property as well as binary questions with no significant loss in accuracy. Evaluating the intersection of relational modification and binary questions, general accuracy was 59.6%, q_1 accuracy (65.5%) fell significantly (McNemar p=0.048) from q_0 accuracy (70.3%), and q_2 accuracy was 53.7%. That is, even for binary questions, which the model tends to perform much better on, questions modified according to object relationships still had a negative effect on model accuracy. For the intersection of object property and non-binary questions, general accuracy was 41.1%, a significant decrease (McNemar p=0.002) in q_1 accuracy (53.7%) from q_0 accuracy (58.9%) was observed, and q_2 accuracy was 28.5%. A decrease in accuracy for non-binary questions, but not for binary questions when using object property modifiers suggests that the model has a harder time comprehending modifications to object property in the context of the less concrete non-binary



(a) (b)

Figure 5: Sensitivity of model to specific details that are absent in the image (questions written by AMT workers for which they selected the answer as "unknown"). Given that the model answers q_1 correctly, we compute the percentage of instances where the answer to q_2 is different from the answer to q_1. (a) shows these results for when q_2 is not "unknown" and (b) when the answer to q_2 is "unknown". Larger values are better since it indicates the ability of the model to recognize that the changed modifiers requires a change in its prediction.



(a) (b)

Figure 6: (a) shows the results where given that the model answers q_0 correctly, we compute the percentage of instances where the model answers q_1 incorrectly. (b) shows the results where given that the model answers q_0 incorrect, we compute the percentage of instances where the model answers q_1 correctly. Smaller values in (a) are better since they indicate that modifier did not force the model to answer incorrectly. Larger values in (b) are better since they indicate that modifier helped the model correct previous errors.

questions. Lastly, examining the intersection of relational modification and non-binary questions yielded 37.0% in general. There was a significant drop (McNemar p=0.005) for q_1 (47.5%) from q_2 (55.0%), and q_2 accuracy was 26.4%. Inversely similar to the first intersection, the intersection of the two categories that the model had the hardest time with separately yielded the lowest accuracy. Accuracy for q_1 for the intersections are presented in Tab. 4.

3.2.5. "Unknown" questions

Finally, we analyzed question pairs that were taken out of the main dataset due to q_2 asking about details not present or determinable from the image resulting in a q_2 of "unknown". Overall accuracy for these questions

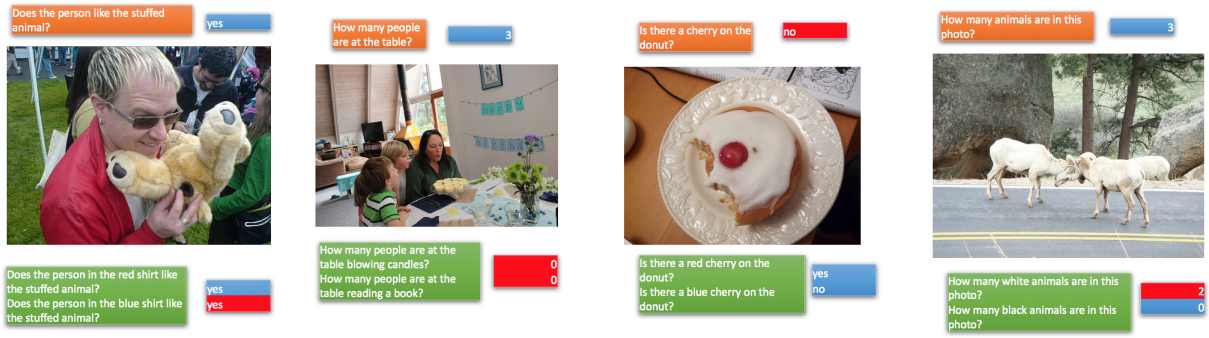


Figure 7: Example predictions made by the model for (q_0, q_1, q_2). Correct predictions are marked by blue and wrong ones by red.

that modified q_2 in such a way that the added details were absent from the image was comparatively very low (30.3%). Therefore, it appears that the model is much less able to determine what it cannot know from the image. For the main dataset $\mathbf{Q}$, 44.5% of questions where q_1 was answered correctly resulted in a q_2 answer from the q_1 prediction. On the other hand, for the "unknown" questions in $\mathbf{Q}_u$ only 18.3% of correctly answered q_1 answers resulted in a changed q_2 answer. As can be easily seen, this difference in changed answers was significant (t-test p<0.0001), and lends support to the proposition that if the model is asked a question not determinable from a image, it uses the context of the details in the question that are present in the image to answer, rather than reasoning that it cannot answer affirmatively. The results are summarized in Fig. 5.

3.2.6. Example Predictions

Some example predictions are shown in Fig. 7. As seen here, in some cases, adding modifiers helped the model correct its errors (the third image in the figure) while in others, it confused the model (the second image in the figure). Further, in other cases, while it seemed like the model was correct, the fact that it got one of q_1 or q_2 wrong seems to indicate that it does not deeply perceive the details in the question. Fig. 6 summarizes the cases where adding modifiers helped or hurt the performance of the model.

4. Conclusion

In summary, adding detail with respect to two categories of syntax modification to original VQA dataset questions caused statistically significant decreases in accuracy for the LXMERT VQA model on answering these modified questions compared to accuracy on the original questions. More specifically, for questions of each type of modification, that is, object property and spatial relational, model performance on questions with additional syntactic modifiers corresponding to more detailed features in the associated image fell. An implication of this observation, namely, that LXMERT is not as sensitive to higher syntactic detail when processing a question about a corresponding image. Interestingly, model accuracy was observed to be different between each modification type as well. That is, the model performed better for questions that modified object property as opposed to spatial relationships. This may be due to the fact that the majority of object property modifiers were syntactically simpler than relational modifiers since object property modifiers typically added details such as color, shape, size, etc. to an object which tended to result in only one modifier word being added more often than relational modifiers, which typically require a modifier phrase such as prepositional phrases or relative clauses. Comparing the effect of adding syntactic detail for binary (yes/no) and non-binary questions, a significant change in accuracy was not observed for binary questions, while a significant change was observed for non-binary questions which suggests that the model is able to process additional syntactic complexity for binary questions better than for the non-binary questions. This difference is important to note, however, should not be surprising due to the fact that VQA models in general tend to perform particularly well on binary questions when compared to less restrictive potential answer sets (Goyal et al., 2017). Lastly, for "unknown" questions, when the model predicted a correct answer to the first modified question q_1 and was presented with q_2 questions which added details that were not present in the corresponding image, it would significantly more frequently choose to keep its answer the same as its q_1 answer rather than change its answer. Therefore, suggesting that when the model is not able to ground its interpretation of a question about an image in that image, it will more frequently default to retaining its answer to q_1. This could mean that the model ignores details that it is not able to ground in an image rather than and that in effect, the model is not sensitive to when it cannot determine the answer to a question.

In future, we plan to use the dataset that we generated to develop novel models that can leverage modifiers in questions to obtain more accurate results. Further, we also plan to develop explanations based on questions with contrasting answers to help improve transparency

of the model.

5. Acknowledgements

Deepak Venugopal acknowledges funding by NSF grants IIS award #2008812, NSF award #1934745 and Adobe Research. The opinions, findings, and results are solely the authors' and do not reflect those of the funding agencies.

6. Bibliographical References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Bansal, A., Zhang, Y., and Chellappa, R. (2020). Visual question answering on image sets. In *ECCV (21)*, volume 12366 of *Lecture Notes in Computer Science*, pages 51–67. Springer.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. (2016). Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, pages 932–937. The Association for Computational Linguistics.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Fagerland, M. W., Lydersen, S., and Laake, P. (2013). The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(1):91.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468. The Association for Computational Linguistics.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

Jas, M. and Parikh, D. (2015). Image specificity. In *CVPR*, pages 2727–2736. IEEE Computer Society.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, 274(5284):114–116.

Khashabi, D., Khot, T., and Sabharwal, A. (2020). More bang for your buck: Natural perturbation for robust question answering. In *EMNLP (1)*, pages 163–170. Association for Computational Linguistics.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297.

Nouwens, S., Groen, M. A., Kleemans, T., and Verhoeven, L. (2021). How executive functions contribute to reading comprehension. *British Journal of Educational Psychology*, 91(1):169–192.

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society.

Selvaraju, R. R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M. T., Nushi, B., and Kamar, E. (2020). Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011.

Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.