Visual Question Answering with Textual Representations for Images

Yusuke Hirota[1,a] Noa Garcia[1,b] Mayu Otani[2,c] Chenhui Chu[3,d] Yuta Nakashima[1,b]
Ittetsu Taniguchi[1,a] Takao Onoye[1,a]
[1]Osaka University [2]CyberAgent, Inc. [3]Kyoto University
[a]{y-hirota,i-tanigu,onoye}@ist.osaka-u.ac.jp [b]{noagarcia,n-yuta}@ids.osaka-u.ac.jp
[c]otani_mayu@cyberagent.co.jp [d]chu@i.kyoto-u.ac.jp

Abstract

How far can we go with textual representations for understanding pictures? Deep visual features extracted by object recognition models are prevailing used in multiple tasks, and especially in visual question answering (VQA). However, conventional deep visual features may struggle to convey all the details in an image as we humans do. Meanwhile, with recent language models' progress, descriptive text may be an alternative to this problem. This paper delves into the effectiveness of textual representations for image understanding in the specific context of VQA.

1. Introduction

A practical task to evaluate image understanding is visual question answering (VQA). VQA aims to answer questions about an image's visual content, requiring a machine to understand both the question and the image. To reason about the visual content, the way in which images are represented is essential. Due to the bottom-up attention's success [2], deep visual features extracted by object recognition models have been used as the de-facto standard for representing visual content.

With the recent progress of Transformer language models [14], research on vision-and-language has shifted to explore pre-training methods [12] to learn cross-modal representations on image-text pairs. However, these methods are still based on deep visual features, which may present some limitations to capture the rich semantic content from a picture [1]. In this paper, we study descriptive representations based on text as an alternative.

Specifically, we explore the effectiveness of textual representations of images and their competitiveness with current deep visual features. We conduct VQA experiments by representing images using text, instead of deep visual features and we investigate the use of synthetic samples on language-only representations. We rely on already annotated descriptions from two datasets [6, 17]. Automatically generating the image descriptions, although a necessary future step, is out of the scope of this paper.

2. Approach

Our input consists of a question and a detailed description of an image, which we encode through a Transformer language-only model. The output of the Transformer is fed into a classifier to predict an answer. We additionally propose the use of data augmentation techniques to increase the size and diversity of the training set.

2.1. Language-Only Data

The data for our language-only VQA framework consists of: (1) questions and answers from standard VQA datasets, (2) image descriptions representing image content, and (3) synthetic data obtained from data augmentation techniques.

Questions and Answers. For the questions and answers, we use VQA-CP [1] and VQA 2.0 [3] datasets. All images of VQA 2.0 and VQA-CP are from MSCOCO dataset [13]. Note that the VQA questions are generated from the images themselves rather than from the image descriptions.

Image Descriptions. We obtain image descriptions from two different corpus: COCO captions [6] and Localized Narratives [17]. COCO captions contains five captions about salient parts for each image in MSCOCO dataset [13]. Specifically, captions are obtained by asking annotators to describe the important parts of the scene, without mentioning unimportant details. Localized Narratives contains the narratives representing the entire image, including minor objects as opposed to COCO captions.

Synthetic Data. Additionally, we generate synthetic samples using data augmentation techniques. We explore multiple techniques grouped into two main categories: Data Augmentation for VQA and Data Augmentation for Language.

2.2. Data Augmentation for VQA

We adapt data augmentation techniques for VQA [8, 5] to our language-only setting. The aim is to generate samples that force the model to react to essential parts in the

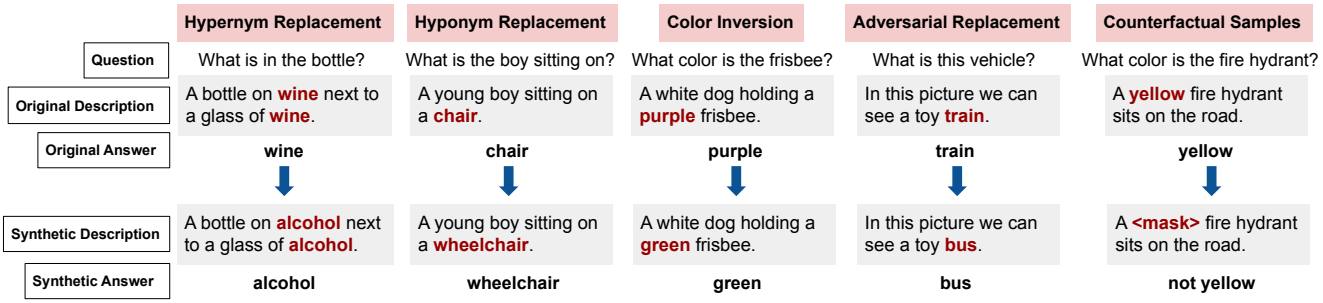

Figure 1. Generated synthetic examples using our proposed data augmentation for VQA techniques.

Table 1. Image Description Evaluation. Length denotes the mean number of tokens in the image descriptions.

Image Description	Length	Accuracy
None (Question-Only)	-	21.39
1 Caption	10.5	35.31
2 Captions	21.0	38.49
3 Captions	31.5	40.09
4 Captions	42.0	41.93
5 Captions	52.5	42.34
Narrative	42.9	36.45
Whole (Narrative + 5 Captions)	95.3	**43.64**

input. We propose four techniques: (1) hypernym and hyponym replacement, (2) color inversion, (3) adversarial replacement, and (4) counterfactual samples. An example of each technique is shown in Figure 1.

Hypernym and Hyponym Replacement replaces words corresponding to answers with their hypernym (or hyponym) and changes the answers accordingly to introduce similar yet semantically distinct mutations into the image descriptions. For a given word, a hypernym covers a wider range of concepts of the original word, e.g. *food* is a hypernym of *fruit*. Whereas a hyponym covers a narrower range of meanings, e.g. *apple* is a hyponym of *fruit*. **Color Inversion** substitutes a color word in a description with another color word and changes the answers accordingly. For Yes/No samples, **Adversarial Replacement** replaces object words $o \in O$ in description with *adversarial words*, where O is the set of 80 object classes in MSCOCO dataset [13]. If o (or its synonyms) are in the question, we change the answer from *yes* to *no*; otherwise the answer is not changed. We define adversarial word as the word that is the most similar yet with a different meaning to o. Adversarial word is selected as the closest word to $o \in O$ according to the Euclidean distance between their Glove embeddings [16]. **Counterfactual Samples** are modifications of questions or images that make the original question-answer pairs irrelevant. We generate counterfactual samples by adapting [5] to language-only description-question pairs. Specifically, we identify the critical words in a question or a description by leveraging Grad-CAM [18] and then remove the answers answered by only looking at the critical words. As a result, we obtain the question or description whose critical words are masked and the remaining answers.

2.3. Data Augmentation for Language

Given that the input to our VQA model is solely based on the language modality, we also explore NLP data augmentation techniques. Among all existing techniques, we adopt three of the most popular and successful ones: (1) EDA, (2) back translation, and (3) contextual word replacement/insertion. Each technique is applied to either the description or the question of the input.

As for each technique, **EDA** [19] is a text editing method that is composed of 4 operations; Synonym Replacement, Random Insertion, Random Swap, and Random Deletion. EDA has been shown to improve text classification performance in low-resource tasks. **Back Translation** [20] translates a sentence into another language and then translates it back into the original language. It can generate diverse paraphrases while preserving the original sentences' semantics. **Contextual Word Replacement/Insertion** replaces or inserts context-sensitive words that the deep bidirectional language model computes [14].

3. Experiments

Setup. We use a large RoBERTa [14] as our Transformer language model. As a classifier, we use a multi-layer perceptron with two fully-connected layers, and Swish activation function between them. We use softmax cross entropy over the answer vocabulary for the loss function. A detailed comparison of language models is provided in the appendix. Unless otherwise stated, the input of our model consists of the whole sequence of question, narrative, and five captions. Results are presented in terms of accuracy.

Comparison of Image Descriptions. We evaluate the performance of different language-only inputs. Specifically, we consider the following inputs: only the question, question and 1 to 5 randomly selected captions, question and narrative, and the whole input (question, narrative, and 5 captions). Results on the VQA-CP v2 test set are reported in Table 1, along with the average sequence length.

Table 2. Comparison of language-only representations with standard deep visual features. * indicates our re-implementations.

Model	VQA-CP v2 test				VQA 2.0 val			
	Yes/No	Number	Other	Overall	Yes/No	Number	Other	Overall
HAN [15]	52.25	13.79	20.33	28.65	-	-	-	-
MuRel [4]	42.85	13.17	45.04	39.54	-	-	-	65.14
UpDn [2]	42.27	11.93	46.05	39.74	81.18	42.14	55.66	63.48
ReGAT [11]	-	-	-	40.42	-	-	-	67.18
BAN* [10]	43.14	13.63	46.92	40.74	83.19	48.13	57.52	65.93
VisualBERT* [12]	43.30	15.07	47.83	41.51	84.55	48.19	57.29	66.33
NSM [9]	-	-	-	**45.80**	-	-	-	-
Ours (Narrative + 5 Captions)	45.13	20.06	49.33	43.64	87.91	56.47	59.43	**69.74**

Table 3. Data augmentation results on the VQA-CP v2 test set. DAV denotes Data Augmentation for VQA, and DAL denotes Data Augmentation for Language. In DAL, D and Q denote when applied to descriptions or questions, respectively. Gap is the overall accuracy difference compared to the accuracy when not using synthetic samples.

	Input Data	Num. Synthetic	Num. Total	Yes/No	Number	Other	Overall	Gap
	Narrative + 5 Captions	-	438,183	45.13	20.06	49.33	43.64	-
DAV	w/ Hyponym Replacement	132,570	570,753	45.65	25.36	50.52	45.26	+1.62
DAV	w/ Hypernym Replacement	23,869	462,052	47.28	17.69	49.10	43.70	+0.06
DAV	w/ Hyponym and Hypernym Replacement	183,944	622,177	45.80	21.46	51.15	45.06	+1.42
DAV	w/ Color Inversion	19,308	457,491	45.61	19.93	50.60	44.47	+1.06
DAV	w/ Adversarial Word Replacement	169,929	608,112	44.71	19.84	50.03	43.93	+0.29
DAV	w/ Counterfactual Samples	438,183	876,366	44.20	19.84	52.07	44.86	+1.22
DAL	w/ EDA (D)	438,183	876,366	44.68	20.64	50.08	44.02	+0.38
DAL	w/ EDA (Q)	438,183	876,366	46.86	23.50	50.62	45.39	+1.75
DAL	w/ Contextual Word Replacement (D)	438,183	876,366	44.69	19.40	48.91	43.18	−0.46
DAL	w/ Contextual Word Replacement (Q)	438,183	876,366	46.09	22.49	49.10	44.16	+0.52
DAL	w/ Contextual Word Insertion (D)	438,183	876,366	45.15	19.31	48.86	43.27	−0.37
DAL	w/ Contextual Word Insertion (Q)	438,183	876,366	45.86	21.44	51.10	45.05	+1.41
DAL	w/ Back Translation (D)	438,183	876,366	45.28	21.01	50.89	44.70	+1.06
DAL	w/ Back Translation (Q)	293,811	731,994	**62.43**	**27.15**	**51.84**	**51.16**	**+7.52**

Table 4. Results of applying back translation to different VQA models in the VQA-CP v2 test set. Gap represents the improvement by training with the synthetic samples.

Model	Yes/No	Number	Other	Overall	Gap
BAN [10]	43.14	13.63	46.92	40.74	-
w/ BT	47.87	16.27	48.76	43.57	+2.83
VisualBERT [12]	43.30	15.07	47.83	41.51	-
w/ BT	55.95	17.11	49.74	46.57	+5.06

The whole input, consisting on merging the narrative with the 5 captions performs the best, which indicates that the two datasets contain complimentary useful information for VQA. When comparing captions and narratives, we find that the former produces better results with fewer words. This confirms that the VQA dataset contains a substantial amount of questions about the general content of the image, rather than its specific details, as the COCO captions, in contrast to narratives, focus mostly on the prominent areas in the scene. In other words, the majority of questions that people make when they look at an image are about the prominent parts of the image.

Comparison against Deep Visual Features. We compare our language-only model with state-of-the-art VQA models based on deep visual features on both the VQA-CP v2 and the VQA 2.0 datasets. For a fair comparison, we do not include the models developed to mitigate language bias [7], as these modules can be added as a plug-in extension to any other method, including ours. For the same reason, we do not use data augmentation.

Results are reported in Table 2. Our model outperforms most of the baselines that use deep visual features on both VQA-CP v2 and VQA 2.0. For the accuracy on VQA-CP v2, NSM [9] performs slightly better than our language-only model. This result verifies that the textual representations of the images are effective and competitive with deep visual features. We provide the visualization of the output of both our language-only model and the deep visual feature-based models in the appendix.

Use of Synthetic Samples. We evaluate the performance when augmenting the VQA-CP v2 training set with synthetic samples. Results for each of the proposed data augmentation techniques are shown in Table 3. Whereas most

techniques, except contextual word replacement/insertion for descriptions, are able to increase the accuracy with respect to the baseline (when no data augmentation is used), back translation for questions has by far the best results, with a gain of 7.52 points on overall. As back translation is the only data augmentation for language that generates new samples while maintaining the original semantics, its impressive performance points us to the importance of 1) having diversity within the training questions set, whereas at the same time, 2) a correct relationship between the triplet question-description-answer semantics.

Since back translation for questions can be easily applied to deep visual feature-based models, we explore if its benefits are transferable to deep visual feature-based VQA models (BAN [10] and VisualBERT [12]). The results in Table 4 shows that training the models with synthesized back translated samples improves the performance by a large margin, with gaps of 2.83 and 5.06 points, respectively.

4. Limitations and Conclusion

Note that the direct comparison between our language-only model and the deep visual feature-based models may not be fair, as our method uses annotated sentences by humans. Meanwhile, deep visual features are trained in an end-to-end manner, which can provide strong supervision on what to see. This, on the other hand, benefits deep visual feature-based models. Yet, our results give interesting insights about the differences and analogies between deep visual features and textual representations, providing a baseline for the VQA tasks with the interpretable representation. Moreover, this study brings us the opportunity to introduce a new research direction for VQA in particular, and image understanding in general: to automatically generate image descriptions as image representations instead of (or combined with) deep visual features. Finally, one of the most surprising findings in this paper, is that the use of back translation boosts VQA models' performance, both when using text representations and deep visual features. This may benefit future research on VQA and implement new training protocols based on back translation augmentation.

Acknowledgements This work was supported by JSPS KAKENHI Number JP18H03264 and JP20K19822, and JST CREST Grant Number JPMJCR20D3, Japan.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015.

[4] Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. MUREL: multimodal relational reasoning for visual question answering. In *CVPR*, 2019.

[5] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, 2015.

[7] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP/IJCNLP*, 2019.

[8] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *EMNLP*, 2020.

[9] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, 2019.

[10] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018.

[11] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, 2019.

[12] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, 2014.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[15] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter W. Battaglia. Learning visual question answering by bootstrapping hard attention. In *ECCV (6)*, 2018.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*. ACL, 2014.

[17] Jordi Pont-Tuset, Jasper R. R. Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV (5)*, 2020.

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[19] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP/IJCNLP (1)*, 2019.

[20] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*, 2018.