Can Pre-training help VQA with Lexical Variations?

Shailza Jolly
TU Kaiserslautern, Germany
DFKI GmbH, Germany
shailza.jolly@dfki.de

Shubham Kapoor[1]
Amazon Research, Germany
kapooshu@amazon.com

Abstract

Rephrasings or paraphrases are sentences with similar meanings expressed in different ways. Visual Question Answering (VQA) models are closing the gap with the oracle performance for datasets like VQA2.0. However, these models fail to perform well on rephrasings of a question, which raises some important questions like *Are these models robust towards linguistic variations? Is it the architecture or the dataset that we need to optimize?* In this paper, we analyzed VQA models in the space of paraphrasing. We explored the role of language & cross-modal pre-training to investigate the robustness of VQA models towards lexical variations. Our experiments find that pre-trained language encoders generate efficient representations of question rephrasings, which help VQA models correctly infer these samples. We empirically determine why pre-training language encoders improve lexical robustness. Finally, we observe that although pre-training all VQA components obtain state-of-the-art results on the VQA-Rephrasings dataset, it still fails to completely close the performance gap between original and rephrasing validation splits.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) is an image conditioned question answering task which has gained immense popularity in vision & language community. Since the introduction of the VQA challenge[2], there has been significant progress in the field of VQA, where new model architectures and training techniques are closing the gap between the model and oracle accuracy on benchmarking datasets like VQA2.0 (Goyal et al., 2017). A majority of models obtained higher gains by introducing semantically rich visual features (Anderson et al., 2018), efficient attention schemes (Lu et al., 2016; Yang et al., 2016), and advance multimodal fusion techniques (Fukui et al., 2016; Yu et al., 2017).

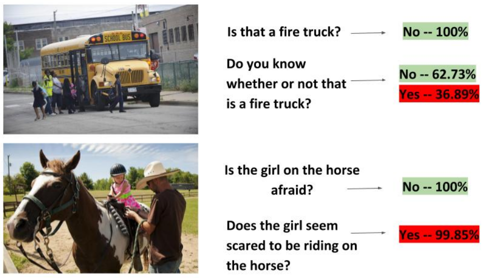
Is that a fire truck?
No -- 100%
Do you know whether or not that is a fire truck?
No -- 62.73%
Yes -- 36.89%
Is the girl on the horse afraid?
No -- 100%
Does the girl seem scared to be riding on the horse?
Yes -- 99.85%

Figure 1: Example from VQA-Rephrasings dataset (Shah et al., 2019). The answers are obtained using Pythia (Jiang et al., 2018) where green text refers to correct answer and red text refers to wrong answer.

However, to deploy these state-of-the-art VQA models into real-world settings, the models must be robust to linguistic variations that originate from interactions with real users. Recently, Shah et al. (2019) showed that state-of-the-art VQA models (Jiang et al., 2018; Kim et al., 2018) are extremely sensitive to the lexical variations which result in a significant performance drop on the VQA test datasets when the questions are replaced with their rephrases. Figure 1 shows the shift in confidence scores of answers for a rephrasing of the original question. To handle these scenarios, they provided a model-agnostic cyclic-consistency (CC) approach that generates question rephrases on the fly during training, which makes the underlying VQA model lexically robust. The best-reported model with their approach achieves 56.59% VQA accuracy on question rephrasings.

Nevertheless, all the models that Shah et al. (2019) experimented with their CC framework in-

[1] The work was done prior to joining Amazon.
[2] https://visualqa.org/challenge.html

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2863–2868
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

corporate an RNN based language encoder. Recently, transformer-based models (Vaswani et al., 2017) led to immense improvements in the whole NLP task spectrum (Wang et al., 2018a). Multi-headed self-attention, the core of transformer architecture, encodes the relationship of a word with its neighbors in several different representational subspaces, thus making these representations robust to linguistic variations.

Since existing datasets expose VQA models to a small subset of the language distribution, it leads to incorrect inference when the model receives rephrasings of the original question. Although training on large datasets may overcome the problem, however, building such extensive annotated datasets is time-consuming & cost-intensive. Pre-trained models like ULMFiT (Howard and Ruder, 2018), BERT (Devlin et al., 2018), and GPT (Radford et al., 2018) have improved performances on various NLP tasks (Rajpurkar et al., 2016; Wang et al., 2018a) trained with limited data. Recently, Tan and Bansal (2019); Lu et al. (2019); Chen et al. (2019) used cross-modal pre-training methods to alleviate this problem in VQA.

In this paper, we study the impact of using pre-training methods to make VQA models linguistically robust. Our contributions are summarized as follows:

- We show that pre-trained language encoders make VQA models lexically robust. We also analyze how pre-trained encoders efficiently extract the same semantic information from syntactically different sentences.
- We show that pre-training is the key to achieve lexical robustness even with complex transformer-based VQA architectures.

To the best of our knowledge, our work is the first one that explores the effect of pre-training to tackle lexical variations, especially for paraphrases, in VQA architectures.

2 Background

In this section, we explain the building blocks of our experiments in this study.

SBERT (Reimers and Gurevych, 2019)[3] is a BERT-based language encoder that generates semantically rich sentence embeddings. It uses siamese and triplet networks (Schroff et al., 2015) to finetune BERT (Devlin et al., 2018), which is a pre-trained transformer encoder trained on large amounts of monolingual data. It obtains state-of-the-art results on common semantic textual similarity and transfer learning tasks.

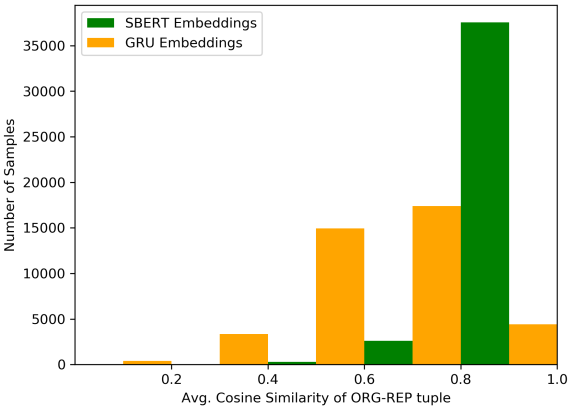

Figure 2: Distribution of cosine similarity of ORG-REP tuples, where each tuple comprises of 1 original sentence and its 3 rephrasings. We calculate the average cosine similarity of rephrasings with its original sentence.

BUTD (Anderson et al., 2018)[4] uses a GRU to encode input questions and uses them to attend image RoI features, enabling region-based attention to generate the answer. BUTD is the base architecture for many other VQA architectures like Pythia (Jiang et al., 2018) and BAN (Kim et al., 2018).

LXMERT (Tan and Bansal, 2019) is a vision-language cross-modality pre-training framework. In contrast to single modality pre-training like BERT, LXMERT focuses on vision-language interactions, which helps to understand better visual contents, language semantics, and the relationship between them. It contains three transformer encoders, namely an object relationship encoder, a language encoder, and a cross-modality encoder, pre-trained using five different vision-language tasks. It must be noted that LXMERT is just a placeholder for transformer-based VQA architectures to investigate if a model architecture plays any role in improving lexical robustness.

3 Experiments

3.1 Dataset

We used the training split of the VQA2.0 dataset (VQA2.0-train) for training the models in this work and evaluated them against the two splits of the VQA-Rephrasings (VQA-R) dataset. It contains

[3] https://github.com/UKPLab/sentence-transformers

[4] https://github.com/hengyuan-hu/bottom-up-attention-vqa

Model	VQA-Rephrasings									
	ORI					REP				
	OA	NUM	Y/N	O	RG	OA	NUM	Y/N	O	RG
BUTD	63.13	41.53	81.27	54.98	-	54.27	33.08	75.73	43.52	-
BUTD+SBERT	62.50	40.22	81.46	53.91	-0.99	57.21	35.91	77.46	47.40	+5.42
LXMERT (a)	63.86	43.38	81.86	55.54	-	54.79	33.86	75.73	44.36	-
LXMERT (b)	64.86	44.32	83.22	56.28	+1.56	58.21	39.25	78.8	47.55	+6.24
LXMERT (c)	73.61	55.88	88.56	66.9	+15.26	66.27	50.63	83.32	57.42	+20.95

Table 1: VQA Accuracy results on both splits of VQA-R. OA refers to overall accuracy. NUM, Y/N and O refers to accuracies for number, yes/no and other answer class. RG refers to relative gain. RG for BUTD+SBERT and LXMERT (c) (and LXMERT (b)) are computed w.r.t BUTD and LXMERT (a) respectively.

a randomly sampled 40,504 question-image pairs from VQA2.0-val. Shah et al. (2019) collected three rephrasings for each question using human annotators, which amount to 121,512 pairs. During data collection, the authors ensured that the rephrasings are syntactically correct and semantically aligned with original questions. We call the original split as ORI and rephrasings split as REP in our experiments.

3.2 Implementation Details

Unlike original BUTD architecture, we use only 36 RoI per image to obtain visual features and use ReLU activation units. We train the model using Adamax (Kingma and Ba, 2014) with an initial learning rate of 2 x 10^{-3} on the full training set, and the standard VQA accuracy (Antol et al., 2015) is reported for each split of VQA-Rephrasings dataset. In our experiments, we replace the GRU of BUTD with SBERT to obtain BUTD+SBERT. We pass the question embeddings from SBERT through a fully-connected (FC) layer, which is later combined with image embeddings to produce a multi-modal representation of the image-question pair. The size of SBERT embeddings is 768, and the FC layer size is 512.

We train three variants of LXMERT: (a) all parameters are randomly initialized (b) only language encoder is initialized with BERT weights (c) all parameters except VQA task head are initialized with the pre-trained LXMERT weights[5]. It is worth mentioning that we don't use any part of VQA2.0-val during training or finetuning to ensure the fairness of results on each split of VQA-R. In our experiments, we use the default hyperparameters set in the original implementation. LXMERT variant (a), (b), and (c) converged at 17 (30 hours), 10 (18 hours), and 4 epochs (8 hours) respectively on Nvidia V100 GPU.

4 Results and Analysis

4.1 Syntactic Variation causes Data Distribution Shift

Machine learning models perform generally well on test samples drawn from a distribution similar to their training data and fail to generalize when test data distribution differs. However, Wang et al. (2018b); Agrawal et al. (2016) showed that networks are misled by contextual heuristics in training data instead of learning underlying generalizations. McCoy et al. (2019) showed a similar trend in NLI and found that state-of-the-art language models like BERT indeed adopt underlying heuristics, thus failing to generalize for test samples. We observe that the VQA2.0-train and VQA2.0-val have similar distributions whereas the distribution of VQA-R is different[6]. Since we train the language encoder of BUTD using VQA2.0-train, it performs significantly better on ORI than REP (in Table 1). Therefore, a shift in the lexical distribution of REP is a contributing factor towards this artifact.

[5] https://github.com/airsplay/lxmert

[6] Distributions of question lengths are given in the supplementary material

4.2 Pre-trained Language Encoders generate Lexically Robust Representations

Although REP and ORI contain the same amount of semantic information, a significant performance drop for REP is due to the poor representation of input questions by the GRU. One can alleviate this problem by introducing a better language encoder. Therefore, we replace the GRU of the BUTD with SBERT, which is robust to lexical variations and efficiently extracts the overall semantics. As shown in Table 1, our approach (BUTD+SBERT) improves the accuracy of REP by 5.41% relative to BUTD and performs slightly better than BAN+CC which is the reported state-of-the-art model of Shah et al. (2019). One must note that the architecture of BUTD is relatively simpler than BAN, and our approach doesn't train any auxiliary component like the question generation module in CC.

However, BUTD+SBERT obtains a comparable performance on ORI, whose distribution is similar to VQA2.0-train. Since we train GRU on VQA2.0-train, it generates semantically rich question embeddings of ORI than the generalized embeddings from SBERT, which never interacts with VQA language data. Tan and Bansal (2019) observed a similar trend in VQA2.0-dev accuracies when they used BERT as the language encoder. Considering SBERT doesn't directly improve VQA models, it raises a question *What are the underlying factors that allow SBERT to improve the REP accuracy?*

We investigate it by generating the SBERT & GRU embeddings for the original question and its three rephrases, and calculate the average cosine similarity of the rephrases with their original counterpart. As shown in Fig. 2, we observe that SBERT moves the embeddings of rephrases significantly closer to the original question in its representational vector space; whereas, GRU fails to extract the underlying common semantics due to its lexical sensitivity. The average cosine similarity of ORG-REP tuple for SBERT and GRU is 91% and 60% respectively. Hence, we conclude that major accuracy gains for REP are derived from the pre-trained language encoder, thus making our approach model-agnostic.

4.3 Pre-trained Language Encoders latch on Keywords

A sentence and its rephrases share some common keywords which control their semantics. A lexically robust language encoder must latch on these keywords to generate semantically rich vector representations. In our experiment[7], we build an ordered sequence of keywords $S1$ extracted from a complete sentence $S2$. We encode $S1$ and $S2$ using a language encoder and measure the cosine similarity of the pair. We hypothesize that a lexically robust language encoder generates similar representations of $S1$ and $S2$ in its vector space. We found that the average cosine similarity over the whole VQA-R dataset for SBERT and GRU is 0.85 and 0.64 respectively[8]. The ability to stress on keywords makes SBERT circumvent syntactic deviations in paraphrases and embed them closer to each other in its vector space.

4.4 Transformers are Good but Pre-training makes them Great

As shown in Table 1, LXMERT (c) achieves state-of-the-art results on both ORI and REP. LXMERT's pre-training, in comparison to SBERT, is conditioned on both vision & language modality, which generates better multi-modal representations. Since a single image is associated with multiple questions, cross-modal attention helps obtain efficient language representations, making VQA models robust towards question rephrasings.

However, the high performance of LXMERT (c) raises an important question *Are the gains coming from pre-training or LXMERT architecture?* Since LXMERT (a) achieves similar performance to BUTD on REP split, it shows that even a complex cross-modality architecture is not enough to make VQA models lexically robust. However, when we train LXMERT initialized with BERT weights, we observe relative gains of 1.56% in ORI, and 6.24% in REP. Furthermore, when we finetune LXMERT with pre-trained language, vision, and cross-modality encoders, the gains in REP grows further to 20.95% relative to LXMERT (a).

Single modality pre-training, like BERT, only captures intra-modal relationships, while VL pre-training, like LXMERT (c), learns cross-modality relationships. Since cross-modal attention aligns entities across input modalities, it induces semantically rich and robust joint representations, thus outperforming BERT only initialization. These results validate that pre-training is a crucial component for obtaining lexical robustness even for highly complex architectures.

[7] We use *rake-nltk* to extract keywords.

[8] We show the distribution of average cosine similarity of $S1$ and $S2$ over whole VQA-R in supplementary material.

5 Discussion

Since pre-trained language models like BERT are trained on large and diverse data, it is generally hypothesized that such models are very robust to linguistic variations. Our results show that pre-trained language encoders like SBERT indeed improve the performance of REP split by 5.42% relative to a GRU encoder; however, it still underperforms by 9.37% relative to semantically similar ORI questions, modeled by a GRU encoder. We observed a similar trend with task-specific multimodal pretraining as well, where LXMERT (c) struggles to close the relative performance gap of about 10% between REP and ORI. In this work, we show that pre-training indeed improves the linguistic robustness of VQA models while simultaneously revealing the limitations of pre-trained language encoders for standard tasks.

6 Conclusion and Future Work

In this paper, we show that pre-trained language encoders, like SBERT, produce semantically similar embeddings for multiple rephrases of a sentence by latching on keywords, thus making VQA models robust to lexical variations. Combining cross-modal pre-training with transformer-based VQA architectures obtains state-of-the-art results on the VQA-Rephrasings dataset.

In the future, we plan to investigate the factors that prevent closing the accuracy gap between ORI & REP despite using extensive cross-modal pretraining. Further, we will study why some answer classes like *number* benefits the most from pretraining while others achieve significantly less relative performance gains.

Acknowledgments

Shailza Jolly was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab program.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning.

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

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13–23. Curran Associates, Inc.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. 2018b. Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, 3(1):151–188.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*.