Exploring Weaknesses of VQA Models through Attribution Driven Insights

Shaunak Halbe
College of Engineering Pune
shaunak9@ieee.org

Abstract

Deep Neural Networks have been successfully used for the task of Visual Question Answering for the past few years owing to the availability of relevant large scale datasets. However these datasets are created in artificial settings and rarely reflect the real world scenario. Recent research effectively applies these VQA models for answering visual questions for the blind. Despite achieving high accuracy these models appear to be susceptible to variation in input questions.We analyze popular VQA models through the lens of attribution(input's influence on predictions) to gain valuable insights. Further, We use these insights to craft adversarial attacks which inflict significant damage to these systems with negligible change in meaning of the input questions. We believe this will enhance development of systems more robust to the possible variations in inputs when deployed to assist the visually impaired.

1. Introduction

Visual Question Answering (VQA) is a semantic task, where a model attempts to answer a natural language question based on the visual context. A direct application of VQA is to answer the questions for images captured by blind people. The VizWiz [2] is a first of its kind goal oriented dataset which reflects the challenges conventional VQA models might face when applied to assist the blind. The questions in this dataset are not straightforward and are often conversational which is natural knowing that they have been asked by visually impaired people for assistance.

Deep Neural Networks often lack interpretability but are widely used owing to their high accuracy on the representative test sets. When deploying these models to aid the blind, utmost care needs to be taken to prevent the model from answering wrongly to avoid possible accidents. To interpret VQA models we use the method of Integrated Gradients [5] which computes attributions for the input features based on the network's predictions. These attributions assign credit/blame to the input features (pixels in case of an image and words in case of a question) which are responsible for the output of the model. These attributions can help identify when a model is accurate for the wrong reasons like over-reliance on images or possible language priors.

We use these attributions which specify word importance in the input question to design adversarial questions, which the model fails to answer correctly. While doing so, we try to preserve the original meaning of the question and ensure the simplicity of the same. We design these questions manually by incorporating highly attributed content-free words in the original question,taking into consideration the free-formed conversational nature of the questions that any user of such a system might ask.

2. Robustness Analysis

2.1. Model and Data Specifications

The VizWiz dataset is significantly smaller than other VQA datasets and hence is not ideal to determine word importance for the content free words. In order to do justice to these words and to keep the analysis generalizable we use the VQA v2 dataset for computing text attributions.

We use the Counter model [6] for the purpose of computing attributions. This model is structurally similar to the Q+I+A [2] (which was used to benchmark on VizWiz). We select this model for ease in reproducibility and for consistency with the original paper [2]. We compute attributions over the validation set, of which the highly attributed words are selected to design prefix and suffix phrases which can be incorporated in original questions for adversarial effect. Further we verify and test these attacks on the following models : (1) Pythia [4] (the VizWiz 2018 challenge winner) pretrained on VQA v2 [1] and transferred to VizWiz (train split) and (2) Q+I+A model (which was used to benchmark on VizWiz) trained from scratch on VizWiz (train split).

2.2. Observations

We observe that among the content-free words, 'what', 'many', 'is' 'this', 'how' consistently receive high attribution in a question. We incorporate these words in the original question. Figure 1 shows the effect of a prefix attack containing heavily attributed words like 'many','in' which

Pythia v0.3 [4]		
Prefix Phrase	Accuracy	% Unanswerable
guide me on this	47.8	74.28
answer this for me	46.27	82.66
in not a lot of words	44.66	85.15
what is the answer to	43.46	86.10
in not many words	42.29	91.3
in not many words- what is the answer to	**38.16**	**97.06**

Table 1. Prefix attacks on Pythia v0.3

Pythia v0.3 [4]		
Suffix Phrase	Accuracy	% Unanswerable
guide me on this	49.8	69.2
answer this for me	48.82	75.19
answer this for me- in not a lot of words	45.3	82.47
answer this for me- in not many words	**42.5**	**88.46**

Table 2. Suffix attacks on Pythia v0.3

Q+I+A [2]		
Suffix Phrase	Accuracy	% Unanswerable
describe this for me	43.52	82.8
answer this for me	43.90	89.7
guide me on this	41.31	87.0
answer this for me- in not a lot of words	40.1	91.13
answer this for me- in not many words	**38.44**	**94.1**

Table 3. Suffix attacks on Q+I+A

steer the prediction of the model from yellow (expected answer) to 1 , a numerical value.

2.3. Attacks

In Suffix Attacks we append content free phrases to the end of each question and evaluate the strength of these attacks through the accuracy obtained by the model on validation set and the percentage of answers it predicts as unanswerable/unsuitable. We expand the Prefix attacks of Mudrakarta et.al. [3] in a conversational vein to suit our task.

2.4. Evaluation and Analysis

The Pythia v3 [4] model achieves an accuracy of 53% while the Q+I+A model achieves 48.8% when evaluated on clean samples from the validation set.It is worth noting that when tested on empty questions Pythia retains an accuracy of 35.43% while Q+I+A retains 38.35%. Thus our strongest attacks(in bold; see Table 1 for Pythia) and (in bold; see Table 3 for Q+I+A) drop the model's accuracy close to the empty question lower bound.

Figure 1. Demonstration of model's response to perturbations in input question with attributions overlaid on the corresponding words.

3. Conclusion

We analyzed two popular VQA models trained under different circumstances for robustness. Our analysis was driven by textual attributions, which helped identify shortcomings of the current approaches to solve a real world problem. The attacks discussed in this paper, illuminate the need for achieving robustness to scale up better to the task of visual assistance. To improve accessibility for the visually impaired, these VQA systems must be interpretable and safe for operation even under adverse conditions arising out of conversational variations. We believe these insights can be useful to surmount this challenging task.

References

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

[2] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *CVPR*, 2018.

[3] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018, 2019.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[6] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018.