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. With the emergence of large scale datasets (Antol et al., 2015; Goyal et al., 2017; Krishna et al., 2016; Malinowski and Fritz, 2014; Zhu et al., 2016), There has been outstanding progress in VQA systems in terms of accuracy obtained on the associated test sets. However these systems are seen to somewhat fail when applied in real-world situations (Gurari et al., 2018; Agrawal et al., 2016) majorly due to a significant domain shift and an inherent language/image bias. A direct application of VQA is to answer the questions for images captured by blind people. The VizWiz (Gurari et al., 2018) 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. Due to unsuitable images or irrelevant questions most of these questions are unanswerable. These questions differ from those in other datasets mainly in the type of answer they are expecting. The questions are often subjective and require the algorithm to actually read (OCR)/ detect/ count, moreover understand the image before answering. We believe models trained on such a challenging dataset must be interpretable and should be analyzed for robustness to ensure they are accurate for the right reasons.

2 Model Interpretability

Deep Neural Networks often lack interpretability but are widely used owing to their high accuracy on the representative test sets. In most applications a high test-set accuracy is sufficient, but in certain sensitive areas, understanding causality is crucial. When deploying such VQA models to aid the blind, utmost care needs to be taken to prevent the model from answering wrongly to avoid possible accidents. In the past, various saliency methods have been used to interpret models which have textual inputs. Vanilla Gradient Method(Simonyan et al., 2013) visualizes the gradients of the loss with respect to each input token(word in this case). SmoothGrad (Smilkov et al., 2017) averages the gradient by adding Gaussian noise to the input. Layerwise Relevance Propagation (LRP) (Binder et al., 2016), DeepLift (Shrikumar et al., 2017) are similar methods used for this purpose.

3 Integrated Gradients (IG)

Vanilla, LRP and DeepLift violate the axioms of Sensitivity and Implementational Invariance as discussed by Sundararajan et al. 2017. As Integrated

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 64–68
Seattle, USA, July5 - 10, 2020. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Gradients (IG)(Sundararajan et al., 2017) satisfies the necessary axioms, we use it for the purpose of interpretability. IG 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. These attributions are computed with respect to a baseline input. In this paper, we use an empty question as the baseline. 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. By content-free, we refer to words that are context independent like prepositions (e.g., "on", "in"), determiners (e.g., "this", "that") and certain qualifiers (e.g., "much", "many") among others.

4 Related Work

The main idea of adversarial attacks is to carefully perturb the input without making perceivable changes, in order to affect the prediction of the model. There has been significant research on adversarial attacks concerning images(Goodfellow et al., 2014; Madry et al., 2017). These attacks exploit the oversensitivity of models towards changes in the input image. Sharma et al. 2018 study attention guided implementations of popular image-based attacks on VQA models. Xu et al. 2018 discuss methods to generate targeted attacks to perturb input images in a multimodal setting. Ramakrishnan et al. 2018 observe that VQA models heavily rely on certain language priors to directly arrive at the answer irrespective of the image. They further develop a bias-reducing approach to improve performance. Kafle and Kanan 2017 study the response of VQA models towards various question categories to indicate the deficiencies in the datasets. Huang et al. 2019 analyze the robustness of VQA models on basic questions ranked on the basis of similarity by LASSO based optimization method. Finally, Mudrakarta et al. 2018 use attributions to determine word importance and leverage them to craft adversarial questions. We adapt their ideas to the conversational aspect of questions in VizWiz to better suit our task. In this paper we restrict ourselves to attacks in the language domain, i.e. we only perturb the input questions and analyze the network's response.

5 Robustness Analysis

5.1 Model and Data Specifications

The VizWiz dataset (Gurari et al., 2018) consists of 20,523 training set image-question pairs and 4,319 validation pairs (Bhattacharya and Gurari, 2019). Whereas the VQA v2 dataset (Goyal et al., 2017) consists of 443,757 training questions and 214,354 validation questions. 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 (Zhang et al., 2018) for the purpose of computing attributions. This model is structurally similar to the Q+I+A (Kazemi and Elqursh, 2017) (which was used to benchmark on VizWiz). We select this model for ease in reproducibility and for consistency with the original paper (Gurari et al., 2018). 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 (Singh et al., 2019) (the VizWiz 2018 challenge winner) pretrained on VQA v2 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).

5.2 Observations

We compute the total attribution that every word receives as well as average attribution for every word based on it's frequency of occurrence. We only take into account content free words, with the intention of preserving the meaning of the original question when these words are added to it. We observe that among the content-free words, 'what', 'many', 'is' 'this', 'how' consistently receive high attribution in a question. We use these words along with some other context independent words to de-

Question :
what is the color of this fruit ?

Predicted Label:
Banana

Question :
in not many words what is the color of this fruit ?

Predicted Label:
1

Figure 1: Attributions overlaid on the corresponding input words. The output of the model changes from 'yellow' to 1 which is driven by the word 'many'.

Question :
what is this ?

Predicted Label:
Train

Question :
answer this for me what is this ?

Predicted Label:
No

Figure 2: The output of the model is driven by the word 'answer' acting as an adversary.

sign the attacks. We use these words to create seemingly natural phrases to be prepended or appended to the question. We observe that the model alters it's prediction under the influence of these added words.

5.3 Suffix Attacks

We present Suffix Attacks, wherein 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 (U).

5.4 Prefix Attacks

We expand the Prefix attacks of Mudrakarta et al. 2018 in a conversational vein to suit our task. These are seen to be more effective as prefix allows us to add important words like 'What' and 'How' to the start of a question which confuses the model to a greater extent than suffix attacks.

5.5 Evaluation and Analysis

The Pythia v3 (Singh et al., 2019) model achieves an accuracy of 53% while the Q+I+A model achieves 48.8% when evaluated on clean samples from the val-set. We tabulate the results obtained by using these phrases as prefixes and suffixes. It is worth noting that when tested on empty questions (which is the baseline for our task) Pythia retains an accuracy of 35.43% while Q+I+A retains 38.35%. Thus our strongest attacks which are meaningful combinations of the basic 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. Our strongest attack (see Table 1) renders 97% of the questions unanswerable, which is a significant increase from 58% when evaluated on clean questions.

6 Performance on other attacks

6.1 Word Substitution

We observe that when we evaluate the model by substituting certain words of the input question by low-attributed words, which change the meaning of the question, the answer predicted in most cases

Pythia v0.3 (Singh et al., 2019)		
Prefix Phrase	Accuracy	% U
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 (Singh et al., 2019)		
Suffix Phrase	Accuracy	% U
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 (Kazemi and Elqursh, 2017)		
Suffix Phrase	Accuracy	% U
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

Q+I+A (Kazemi and Elqursh, 2017)		
Prefix Phrase	Accuracy	% U
describe this for me	46.72	76.8
answer this for me	45.90	79.8
what is the answer to	44.72	80.6
in not many words	44.50	81.4
answer this for me-in not many words	**42.1**	**81.13**

Table 4: Prefix attacks on Q+I+A

is 'unanswerable'. This means that the model does not over-rely on images and is robust in this aspect.

6.2 Input Reduction

We follow the approach of Feng et al. 2018 to iteratively remove less important words from the input question. With the removal of around 50% words from a question, the accuracy drops close to 46% and renders 72% of the questions unanswerable. The Pythia model is fairly robust in this sense too, as it's output becomes 'unanswerable' after considerable input reduction.

6.3 Absurd Questions

To evaluate the effect of absurd attacks on these models, we make a short, non-exhaustive list of objects that do not appear in the validation set of VizWiz(questions, answers and captions) but are present in the training set. We use these objects to form questions similar to the training set questions which contained these objects. A good model should be able to detect absurd questions. For absurd questions like "which country's flag is this ?" (where "flag" does not occur in the validation set of VizWiz) Pythia predicts over 90% of these (clean image)-(absurd question) pairs as 'unanswerable' which is the desired outcome.

7 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

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Nilavra Bhattacharya and Danna Gurari. 2019. Vizwiz dataset browser: A tool for visualizing machine learning datasets. *arXiv preprint arXiv:1912.09336*.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer.

Shi Feng, Eric Wallace, II Grissom, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber, et al. 2018. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. 2019. A novel framework for robustness analysis of visual qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8449–8456.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. 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. Association for Computational Linguistics.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551.

Vasu Sharma, Ankita Kalra, Vaibhav, Sumedha Chaudhary, Labhesh Patel, and Louis-Phillippe Morency. 2018. Attend and attack : Attention guided adversarial attacks on visual question answering models.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

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

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

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

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4961.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.