Induced Attention Invariance: Defending VQA Models against Adversarial Attacks

Vasu Sharma[*2], **Ankita Kalra**[*1] , **LP Morency**[2]
Robotics Institute[1], Language Technologies Institute[2]
[vasus, akalra1, morency] @andrew.cmu.edu

Abstract

Deep learning models are increasingly being deployed in a wide number of safety critical applications, which makes protecting them from adversarial attacks a topic of paramount importance. In this paper we study state of the art adversarial attacks and their effect on top performing Visual Question Answering (VQA) models. Since most modern VQA architectures rely heavily on the use of attention, we hypothesize that defending VQA models against adversarial attacks involves protecting the attention maps from distortion due to these attacks. Based on this hypothesis, we propose a new loss term which we name as *'Induced attention Invariance loss'* (IAI) which is designed to reduce variations in the attention maps due to such distortions. We evaluate the advantages of this approach on a number of top performing VQA models and show that models trained with our proposed IAI loss are almost twice as robust as standard models against state of the art adversarial attacks. We further analyze the effect of adversarial attacks on attention maps and confirm that the adversarial attacks typically distort the attention maps, leading them to focus on the incorrect image regions while also reducing their sharpness. In comparison, models trained with our proposed IAI loss show increased robustness to these distortions.

1 Introduction

The past few years have seen a tremendous increase in the attention given to both designing adversarial attacks and building robust defense mechanisms to protect against them Szegedy et al. (2014); Goodfellow et al. (2015); Liu et al. (2017); Tramèr et al. (2018); Athalye et al. (2017); Papernot et al. (2016). This rise can be attributed to the increasing ubiquity of deep learning systems and their use in a large number of safety critical applications Cheng et al. (2018). This recent work on designing adversarial attacks has exposed the vulnerability of the state of the art deep learning models to adversarial attacks i.e these models can be fooled into misclassifying input samples that are only slightly different from correctly classified samples. These minor modifications are often invisible to the human eye making them extremely hard to detect but can still fool deep learning models. The potential for damage caused by such attacks is immense and it is of the utmost importance that deep learning models be made robust to such attacks before deploying them in real life applications Akhtar and Mian (2018).

In this paper, we focus on the problem of Visual Question Answering (VQA) which can be defined as the problem of answering a specific question based on the visual content of a given image Antol et al. (2015). VQA finds immense uses in systems to aid the visually impaired, in interactive education systems, remote navigation systems, query based image retrieval among others Kafle and Kanan (2017). The large number of potential applications of VQA necessitates the need for training these models in a way which makes them robust before they could be used in day to day life. A study of most modern day VQA models reveals an interesting trend: the ubiquity of the use of attention

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

modules across almost all of these models Kim et al. (2018), Kazemi and Elqursh (2017), Singh et al. (2018), J. Lu and Parikh (2016), Anderson et al. (2017), Agrawal (2017). Attention modules allows the VQA model to dynamically identify the key parts of the input image which should be focussed on to complete the assigned task like answering the question. This ubiquitous presence of attention mechanisms across VQA models leads us to hypothesize that attention maps are not only a key component of the VQA models but also a potential target for adversarial attacks.

Based on this hypothesis we introduce a new loss function named as the ***Induced Attention Invariance*** (IAI) loss which is designed to make the attention maps more robust to adversarial attacks. We evaluate our approach on three of the top performing VQA models when attacked by state of the art adversarial attack techniques. We further evaluate our hypothesis by analyzing the effect of these adversarial attacks on attention maps and how the proposed IAI loss helps protect them against these attacks.

2 Proposed Approach

In this section, we first present a formalization of the VQA problem, followed by the introduction of our proposed *Induced Attention Invariance* (IAI) Loss.

2.1 VQA problem formalization

The problem of Visual Question Answering involves answering a given question Q, based on the visual content of an image I. Most modern VQA models involve the use of an attention mechanism to identify the regions in a given image which are most relevant to answering a specific question. Information from the attention maps, the question and the image itself is fused using a variety of multimodal fusion techniques and then the answer is generated. The full VQA model is typically trained using cross entropy loss computed on the predicted probability distribution over the possible answer space.

2.2 *Induced Attention Invariance* Loss

In this paper, we hypothesize that attention maps play a central role in the vulnerability of present day VQA models to adversarial attacks. Our goal in proposing a new training loss is to make the attention maps less susceptible to perturbations in the input image. With this goal in mind, we propose a new loss function which is designed to make these attention maps more robust to such adversarial attacks. We call this new loss term as the *Induced Attention Invariance* (IAI) loss and define it as:

$$L_{IAI} = \mathbb{E}_I \left[\mathbb{E}_A \left[\left\| \left(\frac{\partial A}{\partial I} \right) \right\|^2 \right] \right]$$

Where $\mathbb{E}_I$ denotes an expectation computed over the input image I and $\mathbb{E}_A$ denotes an expectation over the attention map A. Note that it's important to use the pre-softmax values of the attention maps and not use the probability scores generated by the softmax since that will make the scores dependent on each other. Our IAI loss essentially penalizes the averaged squared norm of the partial derivatives of the attention map with respect to the input image thereby preventing the attention maps to have a large variance with respect to perturbations in the input image. The goal of penalizing this variance is to reduce the susceptibility of the attention maps to adversarial attacks. Now the new robust loss L_{robust} for training these VQA models can be defined as:

$$L_{robust} = L + \lambda L_{IAI}$$

where L is the traditional loss used to train VQA models and λ is the weighting factor for the Induced attention invariance loss which is denoted by L_{IAI}.

3 Experimental Setup

This section explains our experimental setup and design parameters. It also provides the details of our baseline VQA models and the adversarial attack techniques used in our experiments.

3.1 Baseline VQA models

We test our proposed approach on three different VQA architectures: Bilinear Attention Networks Kim et al. (2018), Show, Ask Attend and Answer Kazemi and Elqursh (2017) and Attention on Attention networks Singh et al. (2018). These architectures were chosen to evaluate our approach on both the state of the art and relatively simpler VQA models. We retrain these models using the new Robust Loss (L_{robust}) as defined in the previous section. For Bilinear Attention Networks and Attention on Attention networks, we noticed better results when the pre-trained network was simply finetuned with the new loss, however Show, Ask, Attend and Answer performs better when completely retrained from scratch.

3.2 Proposed Adversarial Attacks

In this work, we attack our models using 3 of the most common adversarial attack techniques namely Fast Gradient Sign Method **(FGSM)** Goodfellow et al. (2015), Iterative Fast Gradient Sign Method **(IFGSM)** Kurakin et al. (2017) and Momentum Iterative Fast Gradient Sign Method **(MIFGSM)** Dong et al. (2018). These techniques were some of the best performing attack models in the NIPS Adversarial Attack challenge with MIFGSM being the winner of the challenge. Each of these techniques try to use the gradient of the model to predict the perturbation needed in the input image to cause the model to mispredict the answer. The performance of each technique is measured using Attack success rate which represents the decrease in accuracy due to adversarial attack relative to the accuracy of the actual model.

3.3 Dataset and Methodology

We test our models on the benchmark VQA 2.0 dataset Goyal et al. (2017) and report the results on test-dev split of the dataset. The loss weighing parameter, λ was chosen based on performance on the validation set. Open source implementations of the VQA models were used to replicate their results and modified to allow training with the new robust loss.

4 Results and Discussion

4.1 Robustness study for VQA models

This section presents the experimental results for the evaluation of the models trained with IAI loss under adversarial attacks. Table 1 compares the robustness of VQA models when trained with and without our proposed IAI loss. The attack success rates clearly show a reduction when training with our proposed IAI loss, making them substantially more robust to adversarial attacks. We further note that training with the new loss has only a marginal impact on the accuracy of the models without attacks, while showing substantial increase in robustness when subjected to adversarial attacks.

VQA Model	IAI Loss	Accuracy without attack	Accuracy after FGSM	Accuracy after IFGSM	Accuracy after MIFGSM	Attack Success Rate
Show, Ask, Attend and Ans	No	60.1	48.0	40.3	41.5	28.0
	Yes	59.3	54.5	50.8	50.9	**12.1**
Attention on Attention	No	64.5	50.4	47.5	47.3	24.9
	Yes	63.8	58.0	55.4	55.3	**11.8**
Bilinear Attention Network	No	70.1	58.7	56.8	56.5	18.2
	Yes	69.5	64.8	63.3	62.8	**8.4**

Table 1: VQA models when trained with and without the proposed IAI loss. The results show that the models trained with IAI loss are twice as robust to adversarial attacks as models trained without it.

4.2 Analysis of attention maps under adversarial attacks

In this paper, we originally hypothesized that attention maps are a crucial part of the VQA models and could potentially make them vulnerable to adversarial attacks. Figure 1 provides an example of how the attention maps vary when subjected to adversarial attacks. As can be seen from the Figure, the attention maps for the the VQA model trained without the IAI loss becomes incorrectly localized

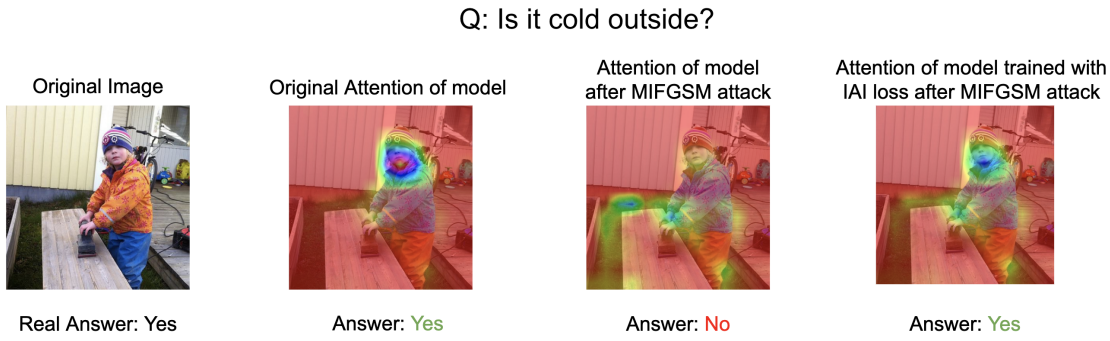

Figure 1: Analyzing attention maps before and after adversarial attacks for models trained with and without IAI loss. As can clearly be seen, the attention becomes incorrectly localized and more diffused after adversarial attack on model trained without IAI, but models trained with IAI show no remarkable difference

and more diffused when subjected to an adversarial attack but the same is not seen when the model was trained with IAI loss.

In these follow up experiments, we aim to study this trend quantitatively and analyze how these attention maps vary before and after an adversarial attack. We also study the impact of adversarial attacks on the attention maps of the VQA models trained with and without our IAI loss. We use the following metrics for this analysis:

- **Correlation** between the attention maps without the attack and the attention maps after the adversarial attack. Large correlation implies that the attacked attention maps localized to similar regions as the attention maps without the attack
- **Sharpness** of attention maps represents how focussed or peaky the attended region is in the attention map. We measure this using variance across the attention maps for the pre and post attack cases. A sharper and more focused attention map will have a large magnitude of attention at a few locations and much lower values elsewhere (since attention maps are normalized to sum to 1) leading to a larger spread in the range of values which in turn leads to increased variance for sharper attention maps

Figure 2 shows the results of the quantitative analysis using these 2 metrics. These results compare the VQA model without attack to the attacked models trained with and without our IAI loss. Similar results for IFGSM and FGSM attacks are shown in the Supplementary material. It can be seen from Figure 2 that attention maps are highly susceptible to adversarial attacks and become incorrectly localized (lower correlation) and highly diffused (lower sharpness) when subjected to adversarial attacks. This analysis supports our original hypothesis that attention maps are potentially susceptible to adversarial attacks. We also note that the VQA models trained with our proposed IAI loss seem to be more robust to distortions in the attention maps.

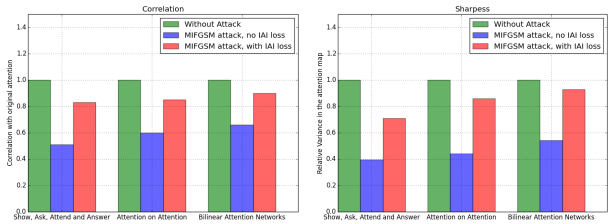

Figure 2: Quantitative analysis of the attention maps after MIFGSM attack: Variation in correlation (top) and sharpness (bottom). Our proposed IAI loss increases robustness of all 3 models bringing their performance closer to the scenario without attacks

References

Harsh Agrawal. 2017. Role of attention for visual question answering. https://computing.ece.vt.edu/~harsh/visualAttention/ProjectWebpage/.

Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

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)*.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. Synthesizing robust adversarial examples. *CoRR*, abs/1707.07397.

Chih-Hong Cheng, Frederik Diehl, Yassine Hamza, Gereon Hinz, Georg Nührenberg, Markus Rickert, Harald Ruess, and Michael Truong-Le. 2018. Neural networks for safety-critical applications — challenges, experiments and perspectives. *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1005–1006.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

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)*.

D. Batra J. Lu, J. Yang and D. Parikh. 2016. Hierarchical question-image co-attention for visual question answering.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *CoRR*, abs/1610.01465.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR*, abs/1704.03162.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial machine learning at scale. *International Conference on Learning Representations (ICLR) 2017*.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017. Delving into transferable adversarial examples and black-box attacks. *ICLR*.

Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277.

Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. 2018. Attention on attention: Architectures for visual question answering (VQA). *CoRR*, abs/1803.07724.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *ICLR*, abs/1312.6199.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR) 2018*.