Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models

Vasu Sharma[2], **Ankita Kalra**[1], **Vaibhav**[2] , **Simral Chaudhary**[2], **Labhesh Patel**[3], **LP Morency**[2]
Robotics Institute[1], Language Technologies Institute[2], Jumio Inc.[3]
[vasus, akalra1, vvaibhav, simralc, labhesh, morency] @andrew.cmu.edu

Abstract

In the present day world, there is large scale deployment of Deep Learning based models in a variety of AI critical applications but very little work has been done to test the robustness of such models to adversarial attacks. In this work we propose a way to generate adversarial samples for the task of Visual Question Answering(VQA) by guiding our adversarial sample generation using attention maps from the underlying VQA model. We examine attacks on the state of the art VQA model proposed by Kazemi and Elqursh [1] and demonstrate the effectiveness of our approach on the VQA dataset [2]. Our attention guided adversarial attack model beats the prior state of the art attack model by a substantial margin and establishes a new state of the art for this task.

1 Introduction and Problem Statement

In this era of growing attempts at breaking the sanctity of deep net models by artificially constructing adversarial samples which closely resemble realistic data samples but break the model's ability to perform correctly, it has become increasingly important to study the possible sources for such attacks and then incorporate a more robust training paradigm while training such models to explicitly prevent such attacks from happening.
In light of the above, we attempt to study the effect of adversarial attacks on the Visual Question Answering (VQA) models [[2], [3], [4], [5]] and test their robustness to the same. In this work, we propose an attention guided adversarial sample generation technique which generates adversarial image samples which are able to fool the VQA model despite being imperceptibly different from the true image. We focus our attacks on the state of the art VQA model by Kazemi and Elqursh [1].
Usually attacks can be classified into two categories, White-Box Attacks, where the weights of the subject network are already known; and Black-Box Attacks, where the network parameters and weights are unknown and is made for generic attacks [6]. In this work we consider white-box attacks, where problem becomes, given a question, an image and a model, we generate an adversarial image such that the victim VQA model produces an answer, which is different from the answer generated for the benign image. In this work we exploit the intermediate attention maps generated by these VQA models to make these attacks more targeted and potent. We launch untargetted attacks wherein the attack is considered successful as long as the answer generated for the benign image is different from that generated for the adversarial image.

With deep learning systems becoming increasingly ubiquitous in the real-world, an adversarial attack could cause tremendous damage. Work presented by Szegedy et. al[7], Goodfellow et. al [8], Liu et. al [9], Papernot et. al [[6], [10]], Tramer et al. [11], Athalye et al. [12] shows that neural networks can be fooled easily. Majority of this work mainly focuses on generating adversarial examples for computer vision classification tasks where the end goal is to estimate the probability distribution among the fixed set of classes. There has also been some work reported in generating adversarial examples for language modality by Jia et. al [13]. Such studies have helped in understanding the representation learning power of networks but mainly in image classification tasks and open-ended multimodal machine learning tasks still remain unexplored. In this work, we extend this line of work and test the popular VQA models and prove that these models are not resilient enough in their present form to be deployed in any safety critical tasks and more work needs to be done on trying to make them robust to such adversarial samples before deploying them in AI critical tasks.

The primary contributions of this work are the following:

1. Proposing a novel architecture to attack Visual Question Answering Models which attains state of the art results on attacking Show, Ask, Attend and Answer model [1] trained on the VQA dataset [2]
2. Proposing a mechanism to reuse attention maps from the VQA models to generate more targetted and effective attacks despite causing very minor perturbations to the input image
3. Proposing a new evaluation metric for evaluating such adversarial attack models which also accounts for the amount of noise added to create the adversarial samples.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

2 Baseline Models

In this section we present some of the baseline approaches to attack VQA models. Most works on adversarial attacks have focussed on the Image Classification problem. We adapt their techniques to the VQA task and use the same to attack the Show, Ask, Attend and Answer [1] model.

2.1 Fast Gradient Sign Method

Fast Gradient Sign Method (FGSM) [8] computes an adversarial image by adding a pixel-wide perturbation of magnitude in the direction of the gradient. This perturbation is computed with a single step, thus is very efficient in terms of computation time. The adversarial image is generated using the equation:

$$I_{adv} = I + \epsilon.sign(\nabla_I J(I, Q, A^*))$$

where I_{adv} is the adversarial image, I is the benign image, Q is the actual question, A^* is the ground truth answer, ϵ is the parameter which controls the update size and J is the loss function for the VQA model. Also note that the adversarial perturbation for every pixel is bound using the box constraint $0 \leq I_{adv}(x, y, c) \leq 1$.

2.2 Iterative Fast Gradient Sign Method

Iterative Fast Gradient Sign Method (IFGSM) [14] is a simple modification over the FGSM and performs the attack in an iterative fashion taking T steps instead of a single step using $\alpha = \epsilon/T$. This can be represented by the equations:

$$I^0_{adv} = 0$$
$$I^{t+1}_{adv} = I^t_{adv} + \alpha.sign(\nabla_I J(I^t_{adv}, Q, A^*))$$

2.3 Momentum Iterative Fast Gradient Sign Method

Momentum Iterative Fast Gradient Sign Method (MIFGSM) [15] was the winning attack in both non-targeted and targeted adversarial attacks competition in NIPS 2017 challenge. This method makes use of momentum to improve the performance of the iterative gradient methods, as described in the following algorithm:

$$I^0_{adv} = 0$$
$$g_0 = 0$$
$$g_{t+1} = \mu.g_t + \frac{\nabla_I J(I^t_{adv}, Q, A^*)}{||\nabla_I J(I^t_{adv}, Q, A^*)||_1}$$
$$I^{t+1}_{adv} = I^t_{adv} + \alpha.sign(g_{t+1})$$

Here a momentum term g_t controlled by the decay rate μ is used to guide the attack.

2.4 Carlini's Attack

Carlini and Wagner [16] propose a method to generate adversarial examples based on optimization methods using gradient descent. They optimize the following optimization function:

$$min_\delta ||\delta|| - \lambda_L J(I + \delta, Q, A^*)$$

under the constraint that $0 \leq I + \delta(x, y, c) \leq 1$. Here λ_L controls the tradeoff between amount of noise added and attack effectiveness. They use a transformation of variables to impose the box constraint as

$$\delta_i = \frac{1}{2}(tanh(w_i) + 1) - I_i$$

Under this transformation, the optimization problem now becomes

$$min_w ||\frac{1}{2}(tanh(w) + 1) - I|| - \lambda_L J(\frac{1}{2}(tanh(w_i) + 1), Q, A^*)$$

This is now solved using iterative gradient descent and the resulting perturbation is used to generate the adversarial sample.

3 Problem Formulation

Let the VQA model be represented by the function h which takes an Image (I) and a Question (Q) as input and generates an answer A. Let A^* be the correct answer to the question Q given the Image I. Now consider an adversarial sample Image I', such an adversarial sample is said to fool the VQA model h if the following conditions hold:

$$h(I, Q) = A = A^*$$
$$h(I', Q) = A' \neq A^*$$
$$||I' - I|| \leq T$$

Where T is the threshold of the Noise we are willing to tolerate in the adversarial image. This can be considered as an optimization problem which needs to be solved by the adversarial attack model to generate effective adversarial samples.

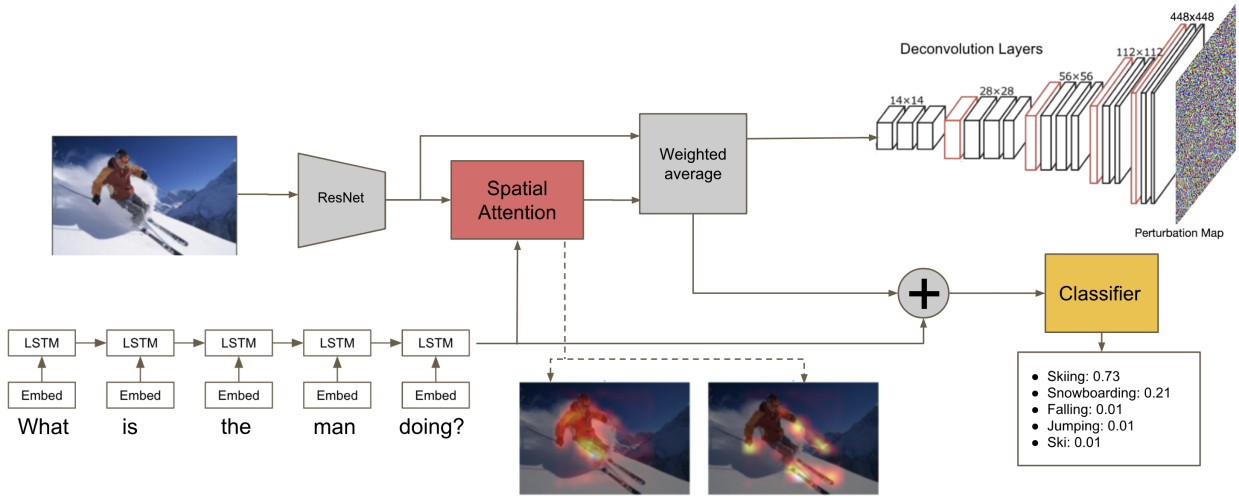

Figure 1: An overview of our Attend and Attack model: The lower part of the network is our *Attend* part which uses the VQA model to generate attention maps which becomes the input to the *Attack* part at the top right which now generates the perturbation maps. The novelty of this model lies in the use of the attention maps to generate the perturbation map

4 Attend & Attack: Attention targeted visual attacks

In this section we present the model we used to generate adversarial images to attack the VQA model. Attention based methods have become a common part of most state of the art VQA architectures [1, 17, 18, 19, 20]. Ablation studies have shown the attention mechanism to be a very crucial part of these models [19]. This motivates us to try and attack this attention mechanism by generating adversarial samples for which the models might not be able to attend to the correct objects in the image. We call our attention targeted visual attack model as **Attend & Attack**.

We base our adversarial attack model on the optimization problem presented in the problem formulation. The loss consists of two terms, the first of which penalizes large amounts of noise addition while the second term encourages the adversarial sample to fool the VQA model.

Let our adversarial sample generation mechanism be denoted by adv_θ where θ denotes the parameters of our adversarial attack model. Now, our model generates the adversarial sample image I' as:

$$N = adv_\theta(Q, I)$$

$$I' = I + N$$

Where our model must optimize the following loss function to effectively generate adversarial samples:

$$min_N(||N|| - \lambda L_{CE}(h(I', Q), A^*))$$

Here λ controls the relative importance of the noise and the cross entropy loss which we denote by L_{CE}. An overview of the architecture of our adversarial attack model (adv_θ) is presented in Figure 1. As can be seen from the Figure 1, our model consists of two main parts which control the *Attend* and *Attack* part of our model respectively. The *Attend* part reuses the trained VQA model to generate the attention maps given a question and the corresponding image. Let I be the original image and Q be the question, then the VQA model first generates image embeddings using the last convolution layer features from a resnet model [21]. The text embeddings are generated by passing the question word by word through a LSTM. The final state of the LSTM represents the question. The visual and text embeddings then go through 2 convolution layers and then a softmax is taken over the spatial locations to generate the attention weights. The content vector is then generated by taking a weighted sum of the image regions weighted by the attention weights. This context vector is then fed to the *Attack* part of the model which generates a full resolution perturbation map that is then added to the original image I to create the adversarial sample I'. The second part uses 4 stacked deconvolution layers with a stride of 2 each for the first 3 layers and a stride of 4 for the last layer, which generate a 448×448 full resolution perturbation map.

5 Proposed Evaluation Metric

Success of an adversarial attack model is studied by how effective it was in attacking the model it is targeting. Traditionally, decrease in VQA accuracy **(ACC)** and relative decrease in VQA accuracy, called as the attack success rate **(ASUCR)** have been used to evaluate success of adversarial attack models.

To the best of our knowledge, none of the standard evaluation metrics account for the amount of noise that has been added to create the adversarial samples. Most works in the existing literature, such as [8] and [22] choose an arbitrary threshold on the noise and use that to cap the amount of noise added but do not explicitly account for how much noise has been added to create the adversarial

sample. We believe that there should be a metric which rewards these low noise adversarial samples more than the ones with higher noise. In light of the above, we propose a novel evaluation metric which penalizes large amount of noise addition to create adversarial samples. We call this metric as the **Attack Effectiveness to Noise Ratio (ENR)**. ENR is defined as the ratio of the ASUCR and the average per pixel noise added to the image to distort it i.e.

$$ENR = \frac{ASUCR}{N}$$

where $N = E[(I' - I)^2]$ and the expectation is taken over all pixels across all channels. I' is the noise distorted and I is the original image.

6 Experimental Details

We present our results on the VQA [2] dataset. The VQA dataset consists of 204,721 images from the MS COCO dataset [23]. We evaluate our models on the real open ended challenge which consists of 614,163 questions and 6,141,630 answers. The dataset comes with predefined train, validation, and test splits. We use the train split to train our models and report our results on test-dev set of the VQA dataset.

As mentioned earlier we attack the Show, Ask, Attend and Tell model [1]. We resuse the attention computations from this VQA model and feed the attention weighted image features to the second part of our network to generate the perturbation maps. Note that the weights of the VQA model itself are frozen while training the adversarial attack model.
We use the central crop of size 448×448 for each of the images. The attention weighted image features have a resolution of $14 \times 14 \times 4096$. The 4 deconvolution layers have 512, 128, 32 and 3 channels respectively. The *tan-hyperbolic* nonlinearity is used in each layer. Adam optimizer is used with a smooth exponential decay of the learning rate with a half life of 50000 iterations. The maximum permissible noise threshold is chosen as 0.2 per pixel per channel. This threshold is chosen based on the approximate level of noise which is added by the approach of Carlini and Wagner [16] on this dataset and model to be comparable to the same.

7 Results

We present our visual attack results in Table 1. As can be clearly seen from the results, our model substantially outperforms the prior state of the art attack on such models. Another thing to note is that our model adds noise of only 0.0048 per pixel per channel which is much lower than that added by the attack by [16] where their model adds noise of almost 0.2 per pixel per channel. The other baseline models [8, 14, 15] add noise of around 0.68 per pixel per channel. This is also reflected in our ENR metric which clearly reflects the superiority of our model over the prior state of the art model for this task.
We further inspect which question types are the most susceptible to such adversarial attacks. Table 2 presents the breakup of the model performance over the various question types. We notice that Color category of the question seems easiest to attack while it is much harder to fool the model for counting type of questions.

Figure 2 presents an example of the working of our model and the Carlini attack model on a given Image. It can be clearly seen from the example that our attack model makes effective use of the attention maps and the noise addition is focussed heavily on the regions which the original VQA model was attending to. Carlini's attacks perturbation model however is highly distributed over the image and also the total noise added by Carlini's attack in much larger than that added by our attack.

We further analyze how the attention maps vary with the different types of attacks. We noticed that attacks which are successful in distorting the attention maps of the VQA model typically succeed in fooling the VQA model into answering the question incorrectly. An example of the same can be seen from Figure 3 where we clearly see that the models which manage to fool the attention mechanism of the VQA model succeed in attacking the same. This exposes a vulnerability in the VQA models, allowing attack models to craft adversarial samples which can fool the VQA models by misleading their attention mechanism.

Model	Dec in Acc(%)	ASUCR	ENR
FGSM [8]	13.2	0.216	0.315
IFGSM [14]	20.3	0.332	0.482
MIFGSM [15]	19.5	0.319	0.469
Carlini Attack [16]	12.1	0.199	0.996
Attend & Attack	**27.7**	**0.454**	**94.58**

Table 1: Comparison of results of our model (Actual Accuracy: 61%)

8 Conclusion

In this paper we present our attention guided adversarial attack generation model and demonstrated its effectiveness on attacking the Show, Ask, Attend and Answer model[1] trained on the VQA dataset [2]. We also see how our model is able to cause much better attacks despite perturbing the input image by a much lesser amount. The model achieves this by targetting the noise to the regions attended by the VQA model thereby misguiding the crucial attention mechanism used by the VQA model itself. We also proposed the new ENR metric which we showed accounts for the level of noise in addition to the success of the adversarial attack itself. We hope that this work will guide building models more robust to such attacks in the future.

Type of attack	Yes/No	Number	Color	Others	Overall	ASUCR	Dec. in Total Acc(%)
no attack	0.668	0.317	0.623	0.412	0.611	-	-
FGSM [8]	0.650	0.300	0.454	0.350	0.479	0.216	13.2
IFGSM [14]	0.600	**0.243**	0.344	**0.264**	0.408	0.332	20.3
MIFGSM [15]	0.607	0.250	0.355	0.271	0.416	0.319	19.5
Carlini's attack [16]	0.651	0.315	0.382	0.396	0.489	0.199	12.1
Attend & Attack	**0.370**	0.309	**0.128**	0.366	**0.334**	**0.454**	**27.7**

Table 2: Accuracy of Show, Ask, Attend and Answer model on various types of attacks.

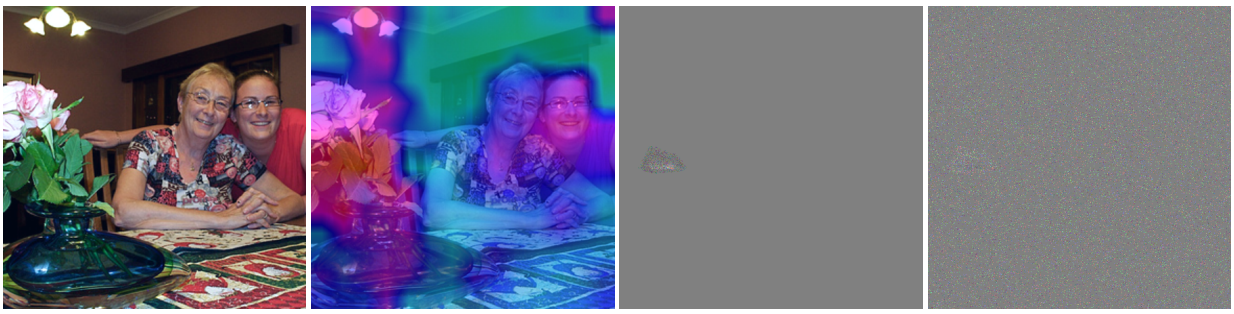

Figure 2: Images (from left to right) : Actual Image, Attention Maps, Perturbation (Our model), Perturbation (Carlini)
Question: what kind of flowers are in the vase?
Actual Answer: Roses, Answer after our attack: Sunflower, Answer after Carlini attack: Roses

Figure 3: Variation in attention maps with different kinds of attacks. **Question:** what is in the top right corner?
Images (from left to right)(Answer in brackets) : (Row 1:) Actual Image (Actual Answer: tree), Original attention (tree), FGSM(tree), IFGSM (coca-cola) (Row 2:) MIFGSM (chips), Carlini (tree), Attend & Attack (pole)

9 Acknowledgements

We would like to thank Sudeep Fadadu with his help with initial idea formulation and discussions. We would also like to thank Harsh Jamthani and Amir Zadeh for their useful insights on the project.
This material is based upon work partially supported by the National Science Foundation (Award 1734868). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, and no official endorsement should be inferred.

References

[1] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR*, abs/1704.03162, 2017. 1, 2, 3, 4

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 4

[3] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[4] 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. 1

[5] Vasu Sharma, Ankita Bishnu, and Labhesh Patel. Segmentation guided attention networks for visual question answering. In *Proceedings of ACL 2017, Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2017. 1

[6] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016. 1

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, abs/1312.6199, 2014. 1

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 2, 3, 4, 5

[9] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017. 1

[10] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016. 1

[11] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR) 2018*, 2018. 1

[12] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *CoRR*, abs/1707.07397, 2017. 1

[13] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. pages 2021–2031, 2017. 1

[14] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations (ICLR) 2017*, 2017. 2, 4, 5

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 4, 5

[16] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016. 2, 4, 5

[17] Yu Jiang*, Vivek Natarajan*, Xinlei Chen*, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 3

[18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018. 3

[19] D. Batra J. Lu, J. Yang and D. Parikh. Hierarchical question-image co-attention for visual question answering. 2016. 3

[20] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[22] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *CoRR*, abs/1712.07107, 2017. 3

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312. 4