BLACK-BOX ADVERSARIAL ATTACKS WITH STYLE INFORMATION

Christodoulos Constantinides
Data Science Institute
Columbia University
New york, NY 10025, USA
cc4718@columbia.edu

Lingyu Zhang
Department of Electrical Engineering
Columbia University
New york, NY 10025, USA
lz2814@columbia.edu

ABSTRACT

Convolutional Neural Networks have advanced the field of Computer Vision. Machines can classify objects of an image with high confidence. However, they are vulnerable to noise injected in the image in an intelligent way, known as adversarial examples. More robust models have been introduced to mitigate this problem. Currently the most effective attacks are white-box attacks where the attacker has access to the model. However, in many scenarios, the gradients of the target model are not available. Meanwhile, recent work in Machine Learning has enabled style transfer from one image to another. In this work we propose two types of black-box attacks based on style transfer and investigate how robust classifiers behave on them.

1 INTRODUCTION

Artists have been painting scenes with their own unique style. Recent work has enabled the style transfer of an image to the content of another image Gatys et al. (2015) Karras et al. (2019). The way these images are created are very similar to how adversarial attacks try to inject noise in an image to confuse the classifier. The only difference in our case is the objective. In adversarial attacks the objective is to lower the confidence of the classifier, while in style transfer is to increase the similarity of specific features between the images that characterize the style. Recent works towards more robust classifiers Ilyas et al. (2019) suggest that classifiers rely on non-robust features that are not perceptible by humans but strongly correlate with different classes. They remove these features from the dataset and were able to obtain classifiers robust to adversarial noise.
Our focus in this work is to see whether these classifiers are robust against images with injected style. We will focus on black box attacks, where we generate adversarial examples without using the classifier we want to fool.

2 METHODS

2.1 STYLE TRANSFER

In this work, we follow the style transfer method indtroduced in Gatys et al. (2015). The convolutional layers extract feature maps of the image. These feature maps represent various level of features of the image which provide information about the content and style of an image. The deeper we go the more abstract the features.

Our objective is to minimize the style difference between the style image and the content image while preserving the content of the content image.

Content objective
We want to produce a new image that has the same content as in the content image. The feature maps capture spatial information (content)

$$L_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{ij} (F_{ij}^l - P_{ij}^l)^2 \tag{1}$$

Where p: content image, x: style image, F^l: Feature map of $\vec{p}$ at layer l, P^l: Feature map of $\vec{x}$ at layer l

Style objective
The style objective ensures similarity between various level of features between images. We first need to compute the Gram Matrix which captures the correlation/similarity between two vectors. The bigger the value, the more correlated they are. We'll take all the feature maps from a specific layer, flatten them and compute their dot product between all the pairs.

$$G_{ij} = \sum_k F_{ik}^l F_{jk}^l \tag{2}$$

The style loss is defined as

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{i,j}^l - A_{i,j}^l)^2 \tag{3}$$

Where G^l is the Gram matrix of generated image G at layer l and A^l is the Gram matrix of the style image A at layer l.
The total loss is defined as

$$L_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha L_{content}(\vec{p}, \vec{x}) + \beta L_{style}(\vec{a}, \vec{x}) \tag{4}$$

Image generation procedure
In the original neural transfer work, the images were initialized with random noise. We initialize with the content image for faster convergence. At every time set, the image is passed through a pretrained CNN and we use the computed feature maps and activations to compute the loss defined above. We then update the loss w.r.t the input image only by performing gradient descent. We don't modify anything from the models themselves, only from the image.

2.2 Adversarial Attacks

Despite the widely successful applications, neural networks have been shown to be susceptible to imperceptible perturbations in the pixel space Szegedy et al. (2014), known as adversarial examples. Most common and effective adversarial attacks Goodfellow et al. (2015)Madry et al. (2018)are white-box attacks, they injecting noise to the image through gradient ascent to increase vision task loss. However, they all rely on the the gradients of the target model to generate adversarial examples. In many scenarios, we do not have the weights of the target model.
We hypothesize that style information computed from a target image can contain features that can potentially fool the vision model. In this section we describe two algorithms for black-box adversarial attack with style information.

2.2.1 Projected Gradient Descent Style attack

A good adversarial example should balance the trade-off between effecting the vision model's outputs and distorting the original image. We aim at fooling a classifier by minimizing a style loss computed from another image. One way to control the distortion of the attacked images is to bound the norm of the style noise. A common norm for bounding perturbations is the L_∞ norm. We formulate the objective in the following:

$$\delta* = \arg\min_{\delta} L_{\text{style}} \tag{5}$$

$$s.t. ||\delta||_\infty < \epsilon \tag{6}$$

However, the L_∞ norm is not smooth so we adopt the projected gradient descent for optimization. At every time step, δ is moved along the direction of the computed gradients, with a certain step size. When it exceeds he ϵ bound, it is projected back to the surface of the bound.

2.2.2 STYLE-CONTENT LAGRANGIAN ATTACK

The L_∞ norm does correlate with the human visual system perfectly, it is possible to achieve higher attack rates with fewer modified pixels if the noise is unbounded. We also consider directly using the style-content Lagrangian as the objective function, and use the Lagrange multiplier to control the distortion-attack tradeoff.

2.3 ADVERSARIAL ROBUST MODELS

It has been recently shown that computer vision models when trained, look into two kinds of features: features that define the class and features derived by the pattern of data distribution that are imperceptible to humans but are highly predictive but at the same time brittle to small changes Ilyas et al. (2019). These small brittle features are even transferable between models because they are patterns that naturally occur in the datasets. The robust models used come from the python package robustness Engstrom et al. (2019). The datasets trained on are ImageNet Deng et al. (2009) and CIFAR-10 Krizhevsky et al. which are widely used for Computer Vision tasks. We only used ResNet as a backbone He et al. (2016)

The authors have identified these features by constructing an adversarial dataset where they injected human imperceptible noise to the dataset and labeled it with the adversarial target class. The trained model on the adversarial dataset oddly performs very well on unseen data but is not robust on adversarial examples. By using this model the authors removed the adversarial features from the dataset and trained a robust classifier on this new clean dataset. Since the imperceptible non-robust features will be ignored by the robust model, we hypothesize that the robust models should be more susceptible to style to targeted style attacks, because the visible style features will have a larger influence on the model's outputs.

In the next section we provide our experiments on evaluating the robustness of these models on style transfered images.

3 EXPERIMENTAL SETUP

We investigate the effect of style information on robust and non-robust classifier models. For style loss computation, we use a pretrained VGG-19 net Simonyan & Zisserman (2015) and extract 5 different layers. For the content loss in the unbounded style-content Lagrangian attack, we directly use the l_2 distance in the pixel space. In the following experiments we combine all the layers together to get the style but we also try only the last layer.

4 RESULTS

4.1 STYLE TRANSFER ATTACK VISUALIZATION

There are apparent visible perturbations on the attacked image. Nonetheless, it is obvious to the human eye that they are all images of a cat.

4.2 BOUNDED ATTACK EFFECT ON NON ROBUST MODELS

We find that for non-robust models, bounded attack generally cannot fool the model into predicting the desired class. However, giving the style information of a specific target class will lower the model's confidence in its prediction.

4.3 BOUNDED ATTACK EFFECT ON ROBUST MODELS

We find that for robust models, bounded attack generally has little effect.

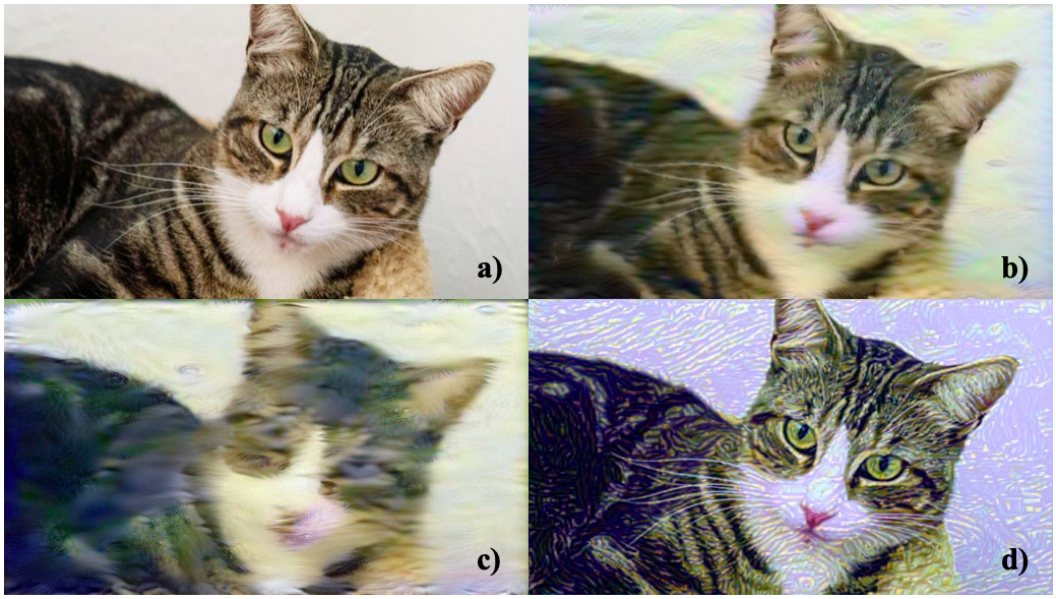


Figure 1: Style transfer attack using all layers to compute style information. a) The original image. b) Bounded PGD attack with target dog image. c) Style-content Lagrangian attack with target dog image. d) Bounded PGD attack with untargeted art image.

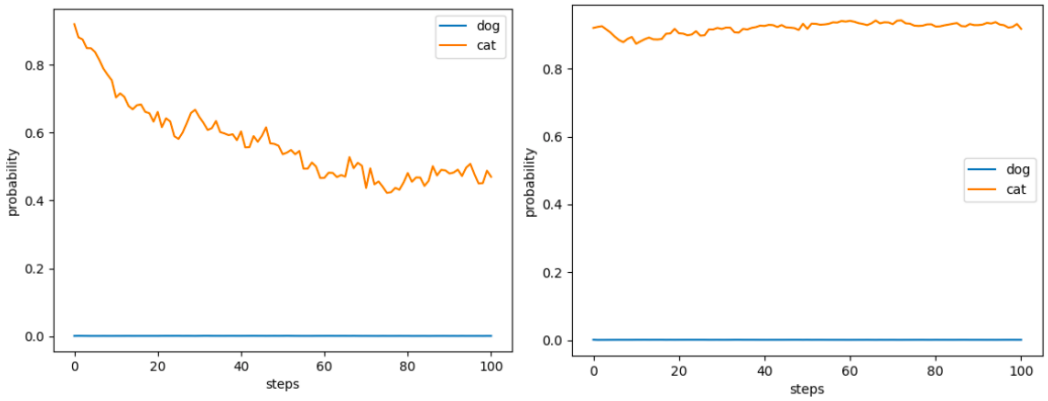


Figure 2: Left: Probability curves of bounded targeted attack on non-robust models. Right: Probability curves of bounded untargeted attack on non-robust models.

4.4 Unbounded attack effect on non robust models

We find that unbounded style attacks on non-robust models are very effective. We were able to decrease the probability of the cat class from over 80% to under 2% in 10 steps. We also observe a gradual increase in the probability of the dog class, due to the increasing amount of dos style information injected in to the image.

4.5 Unbounded attack effect on robust models

We find that the unbounded style attacks on robust models are also effective. The confidence of the class of the style image increases to even higher than non-robust models. This aligns with our hypothesis that because non-robust features are not picked up by the robust models, it is more sensitive to visible style features. We also observe a rise in the content class, suggesting that with the content loss constraint, the robust model is able to recognize some recovered content features.

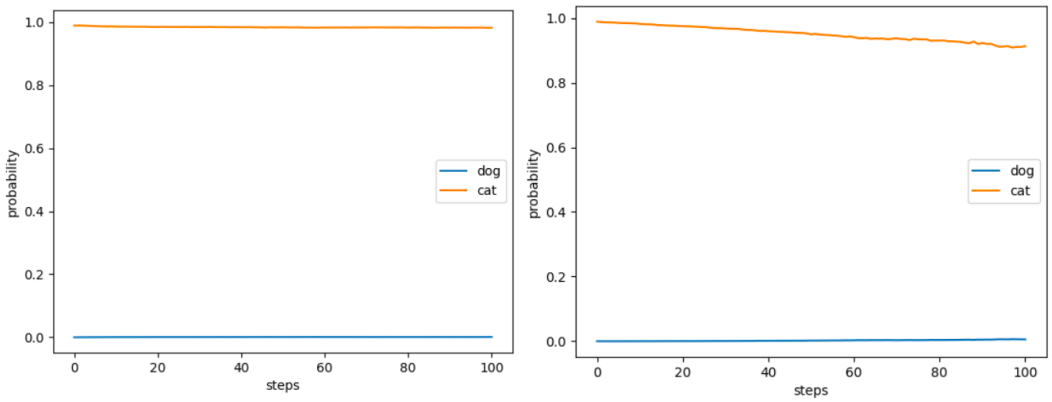


Figure 3: Left: Probability curves of bounded targeted attack on robust models. Right: Probability curves of bounded untargeted attack on robust models.

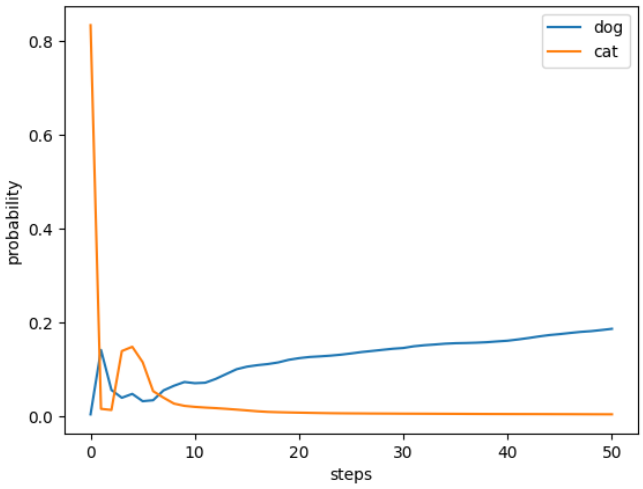


Figure 4: Probability curves of unbounded targeted attack on non-robust models. The classifier is pretrained on imagenet, and we used all 5 layers from the VGG net to compute style loss.

The robust models with different ϵ and l_2 and $l^{\inf}$ loss were assessed on style transfer generated images with different style loss coefficients β.

We also experimented with different style loss coefficients β to emphasize more on the style and less on the content. From the results we can see that increasing the coefficient decreases the confidence of the correct class in some of the robust models. On the style transfered images with $\beta = 1$ the only classifier that is not performing well is l_{infty} loss trained on ImageNet and $\epsilon = 8/255$. For $\beta = 2$ we see some confidence drop on all the classifiers and on $\beta = 4$ major confidence drop.

We also visualized the Gradient Maps of the predictions Selvaraju et al. (2017) using Grad-CAM++ Selvaraju et al. (2019). Gradient Maps is an interpretability technique used on CNNs to see where the network mostly focuses to make a prediction. We plotted these maps for different iterations. We can see that the focus of the network is initially concentrated in key points that characterize the cat. When the style is injected, the network's focus is spread all over the image, giving bad results.

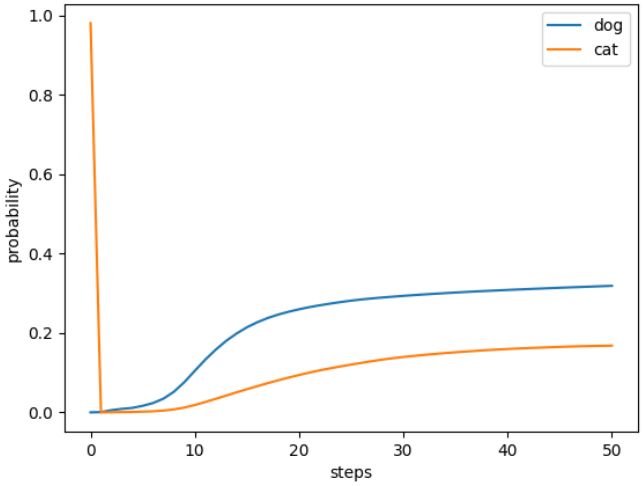
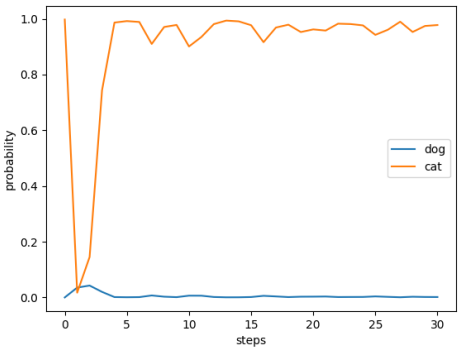
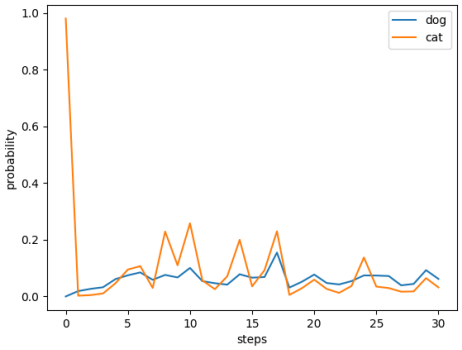


Figure 5: Probability curves of unbounded targeted attack on robust models. The classifier is pretrained on imagenet with nonrobust features removed, and we used all 5 layers from the VGG net to compute style loss.

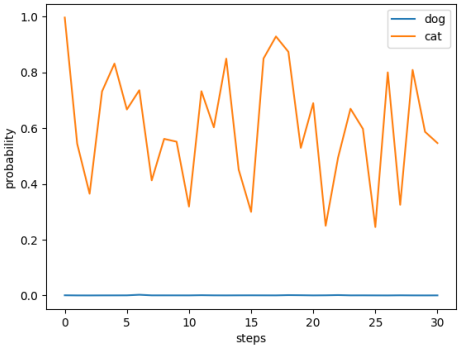
Class probabilities for style transfer with style loss coefficient $\beta = 1$



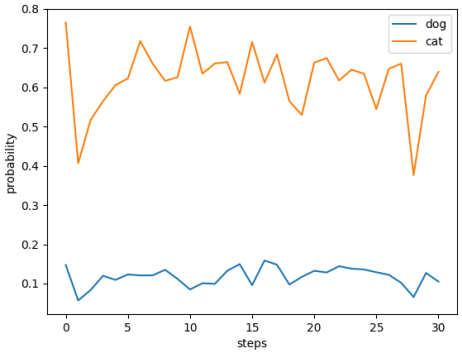
Robust trained with l_2 loss on imagenet



Robust trained with l_∞ loss on imagenet, ϵ =8/255

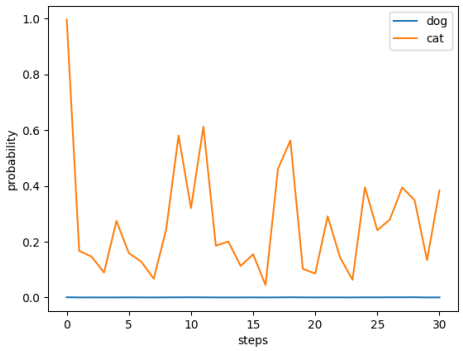


Robust trained with l_2 loss on CIFAR

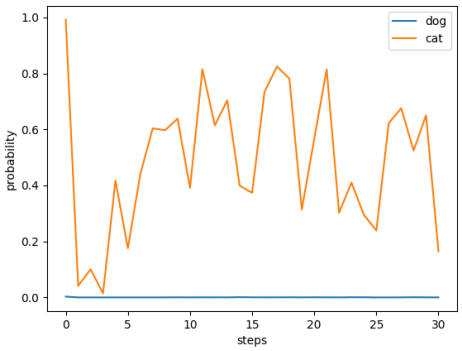


Robust trained with l_∞ loss on CIFAR, ϵ =8/255

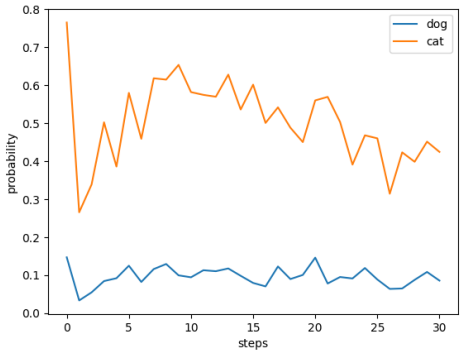
Class probabilities for style transfer with style loss coefficient $\beta = 2$



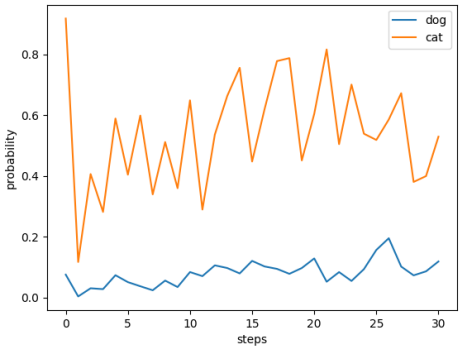
Robust trained with l_2 loss on CIFAR, $\epsilon = 0.25$



Robust trained with l_2 loss on CIFAR, ϵ =0.5

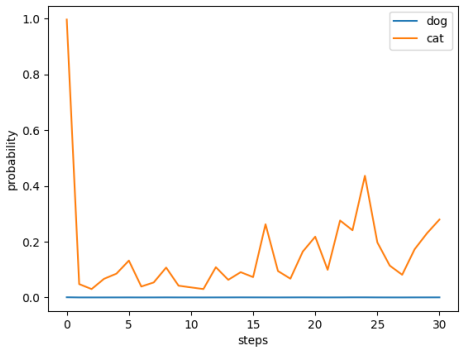


Robust trained with l_2 loss on CIFAR $\epsilon = 1$

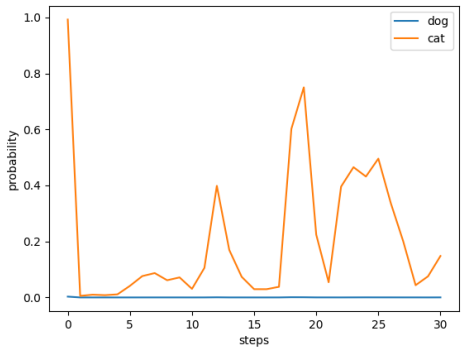


Robust trained with l_∞ loss on CIFAR, ϵ =8/255

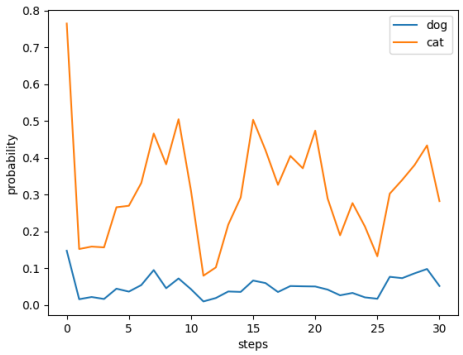
Class probabilities for style transfer with style loss coefficient $\beta = 4$



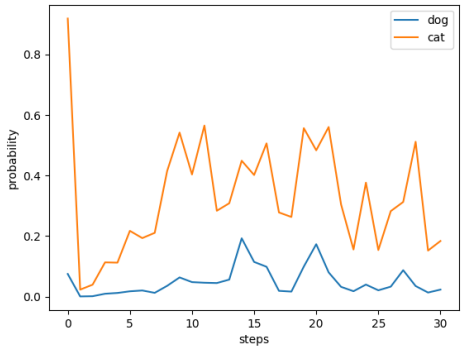
Robust trained with l_2 loss on CIFAR, $\epsilon = 0.25$



Robust trained with l_2 loss on CIFAR, ϵ =0.5



Robust trained with l_2 loss on CIFAR $\epsilon = 1$

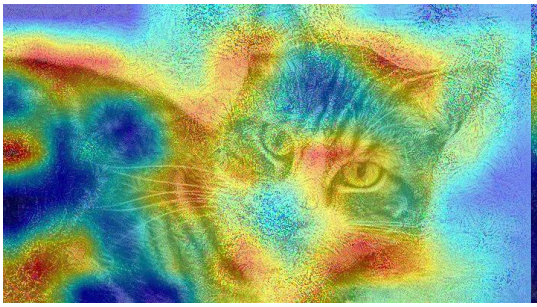


Robust trained with l_∞ loss on CIFAR, ϵ =8/255

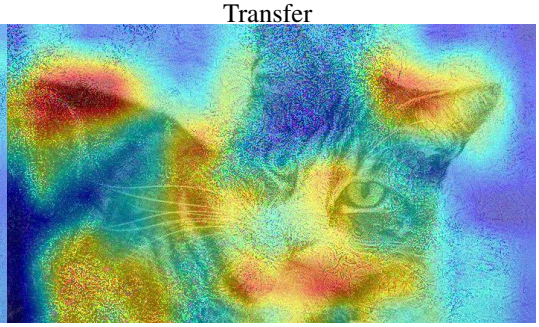
Gradient Maps for different style transfer iterations

Gradient Map of Original Image

Gradient Map after 100 iterations of Style Transfer



Gradient Map after 200 iterations of Style Transfer



Gradient Map after 300 iterations of Style Transfer

5 CONCLUSION

In this work, we proposed two types of black-box adversarial attacks based on style information. We find that bounded style attacks are weaker than unbounded ones. Moreover, robust models are more susceptible to style attacks because their non-robust features have little effect on the output, and visible style features have larger effect.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *ArXiv*, abs/1508.06576, 2015.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *ArXiv*, abs/1905.02175, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

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