Investigating Adversarial Attacks on Neural Style Transfer

Irakli Grigolia
Computer Science
Worcester Polytechnic Institute
Worcester, MA
igrigolia@wpi.edu

Alexander Galvan
Mechanical Engineering
Worcester Polytechnic Institute
Worcester, MA
adgalvan@wpi.edu

Yunhao Zhang
Mechanical Engineering
Worcester Polytechnic Institute
Worcester, MA
yzhang44@wpi.edu

Chaitanya Gaddipati
Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA
cgaddipati@wpi.edu

***Abstract*—In this paper, we investigate the behavior of neural style transfer algorithms under adversarial attacks, focusing on three state-of-the-art networks: Neural Neighbor Style Transfer (NNST), AdaIN Style Transfer, and Fast Style Transfer. Our objective is to understand the vulnerabilities of these networks to various adversarial attacks, including Projected Gradient Descent (PGD) Attack, Carlini-Wagner (CW) Attack, and DeepFool Attack. We aim to compare the results of these attacks, identify potential weaknesses in the neural transfer networks, and explore methods for improving the reliability and robustness of the neural style transfer algorithms. The performance metrics we employ include robustness, execution time, visual quality of the stylized images, and perceptual loss to ensure content and style preservation.**

I. Introduction

A. Background and Motivation

Neural style transfer has emerged as a popular and intriguing application of deep learning, enabling the creation of artistically stylized images by combining the content of one image with the style of another. The development of neural style transfer algorithms has led to a wide range of applications in the fields of art, design, advertising, and entertainment. However, the growing prevalence of these algorithms has also raised concerns regarding their vulnerability to adversarial attacks.

Adversarial attacks are malicious manipulations of input data designed to mislead or degrade the performance of machine learning models. While these attacks have been extensively studied in the context of image classifiers, there is limited research on their impact on neural style transfer algorithms. Understanding the vulnerabilities of neural style transfer algorithms to adversarial attacks is essential for ensuring their security, robustness, and reliability in real-world applications.

Moreover, the development of robust neural style transfer algorithms is critical for maintaining the integrity of the generated stylized images and preventing the unauthorized manipulation of their content and style. Identifying potential weaknesses and exploring techniques for improving the resilience of these algorithms is an important step toward securing and enhancing their practical use.

Motivated by these concerns, our research investigates the behavior of state-of-the-art neural style transfer algorithms under adversarial attacks, focusing on their vulnerabilities and the development of methods for improving their robustness.

B. Problem Statement and Objectives

The problem statement for our research can be formulated as follows: How do adversarial attacks affect the performance of state-of-the-art neural style transfer algorithms, and what techniques can be considered to potentially improve their robustness?

To address this problem, our research objectives are as follows:

1. Investigate the vulnerabilities of three state-of-the-art neural style transfer networks, Neural Style Transfer, Neural Neighbor, and AdaIN, to the four most popular adversarial attacks, including Projected Gradient Descent (PGD) Attack, Fast Gradient Sign Method (FGSM) Attack Carlini-Wagner (CW) Attack, and DeepFool Attack.
2. Evaluate the impact of these attacks on the performance of the neural style transfer algorithms in terms of robustness, execution time, visual quality, and perceptual loss.
3. Identify potential weaknesses in the neural style transfer algorithms and explore potential techniques that may contribute to improving their resilience to adversarial attacks.

II. Related Work

A. Neural Style Transfer Algorithms

Neural style transfer is a deep learning-based technique that combines the content of one image with the style of another to create an artistically stylized output. Pioneered by Gatys et al. (2016), the original Neural Style Transfer (NST) algorithm utilized a pre-trained VGG-19 network to optimize a loss function that combined content and style losses. Since the introduction of NST, numerous other algorithms have been proposed to address its limitations, such as computational complexity and slow processing times. AdaIN Style Transfer introduces an adaptive instance normalization layer to align the mean and variance of the content and style features, enabling faster style transfer with fewer artifacts. Fast Style Transfer employs a feed-forward network trained to apply a specific style, resulting in significantly reduced computation time compared to iterative optimization-based methods.

B. Adversarial Attacks on Neural Networks

Adversarial attacks are designed to exploit the vulnerabilities of machine learning models by introducing small perturbations to the input data that are imperceptible to humans but can cause significant degradation in model performance. Szegedy et al. (2014) first demonstrated the existence of adversarial examples in the context of image classification tasks, and subsequent research has produced a variety of adversarial attack techniques. Some notable examples include the Projected Gradient Descent (PGD) Attack (Madry et al., 2018), Carlini-Wagner (CW) Attack (Carlini and Wagner, 2017), and DeepFool Attack (Moosavi-Dezfooli et al., 2016).

While adversarial attacks have been extensively studied in the context of image classifiers, relatively little research has been conducted on their impact on neural style transfer algorithms. Our work aims to bridge this gap by investigating the vulnerabilities of neural style transfer algorithms to adversarial attacks and exploring potential techniques for enhancing their robustness.

III. Methodology and Results

A. Dataset

In this study, we aimed to explore the vulnerabilities of neural style transfer algorithms to adversarial attacks. To achieve this, we utilized a dataset comprising of eleven images that were carefully selected to represent a diverse range of styles and subjects. The dataset included photographs of famous landmarks such as the Colosseum, Golden Gate Bridge, and Eiffel Tower, and paintings by renowned artists such as Monet, Picasso, Van Gogh, Rene Magritte, Dali, and Giorgio de Chirico. The images of famous landmarks were selected to test the algorithms' ability to preserve content while applying a particular style, while the paintings were chosen to examine the algorithms' capability to capture the unique styles and characteristics of different artists and art movements. Our goal was to evaluate the performance and robustness of neural style transfer algorithms under adversarial attacks using a diverse set of styles and subjects.

Overall, the selected images provide a diverse set of styles and subjects, making them suitable for testing the performance and robustness of neural style transfer algorithms under adversarial attacks.

B. Neural Style Transfer (NST)

In this section, we provide an overview of the three state-of-the-art neural style transfer algorithms that serve as the focus of our investigation: Neural Style Transfer (NST), AdaIN Style Transfer, and Fast Style Transfer.

The original Neural Style Transfer (Gatys et al., 2016) algorithm is an optimization-based approach that leverages a pre-trained VGG-19 network to extract content and style features from input images. NST defines a loss function composed of a content loss and a style loss, which are used to measure the differences between the content of the generated image and the content image, and between the style of the generated image and the style image, respectively. The optimization process iteratively updates the generated image to minimize the combined loss, ultimately producing a stylized output that combines the content of the input image with the style of the reference image.

C. Neural Neighbor Style Transfer

Neural Neighbor Style Transfer (NNST) is a novel neural style transfer algorithm that utilizes a neighbor embedding technique to transfer the style of a reference image to a content image. The NNST algorithm works by first mapping the content and style images into a high-dimensional embedding space and then identifying the nearest neighbor of the style image for each content patch. This mapping is used to transfer the style of the reference image to the content image by iteratively updating the content image to minimize the distance between the content and its nearest style neighbor. Unlike other state-of-the-art neural style transfer methods, NNST does not rely on a loss function to optimize the stylized image. Instead, it leverages the properties of the embedding space to transfer the style information, which makes the algorithm computationally efficient and fast. The NNST algorithm produces stylized images that are visually appealing and preserves the content while incorporating the style of the reference image.

D. AdaIN Style Transfer

AdaIN Style Transfer is an alternative approach that introduces an adaptive instance normalization (AdaIN) layer into the neural network architecture. The AdaIN layer aligns the mean and variance of the content and style features, enabling the transfer of the style reference image's style onto the content image. Unlike NST, AdaIN Style Transfer employs a single feed-forward pass through the network, which reduces computational complexity and processing time compared to optimization-based approaches. The AdaIN method produces visually pleasing results with fewer artifacts and is more efficient than the original NST algorithm.

E. Fast Style Transfer

Fast Style Transfer is another feed-forward approach that uses a pre-trained style-specific network to apply a given style to input content images. Instead of iteratively optimizing the loss function like NST, Fast Style Transfer trains a separate network for each desired style, which can then be applied to content images with a single forward pass through the network. This approach significantly reduces the computation time required for style transfer, making it suitable for real-time applications. However, one drawback of Fast Style Transfer is that it requires training a new network for each style, which may be computationally expensive when dealing with a large number of styles.

F. Adversarial Attack Techniques

In this section, we discuss the adversarial attack techniques that will be used to assess the vulnerabilities of the neural style transfer algorithms under investigation: Projected Gradient Descent (PGD) Attack, Fast Gradient Sign Method (FGSM) Attack, Carlini-Wagner (CW) Attack, DeepFool Attack.

G. Projected Gradient Descent (PGD) Attack

In the context of classification networks, the Projected Gradient Descent (PGD) Attack attack is designed to generate adversarial examples that can deceive a model by slightly modifying the input image. This is done by iteratively updating the adversarial image while making sure the perturbations are limited to a certain extent, defined by the epsilon parameter. The attack's objective is to maximize the loss between the true label and the predicted label, which results in the model making incorrect predictions on the adversarial examples.

For style transfer, we adapted the PGD attack to work with the content and style losses instead of classification losses. The attack's objective was modified to maximize the content and style losses between the input content and style images and their corresponding stylized outputs. By iteratively updating the adversarial image with the gradients calculated from these losses, we generated adversarial examples that affect the style transfer process.

Our findings reveal that the attack had a partial impact on the generated images. While the overall style of the original image was mostly preserved, the visual quality of the stylized output was noticeably degraded. This degradation in quality can be attributed to the perturbations introduced by the PGD attack, which interferes with the content and style features extracted by the NST network. As a result, the generated images exhibit artifacts, increased noise, and a loss of detail in both content and style aspects. Furthermore, the perceptual loss metric indicates a deviation from the ideal balance between content and style preservation, suggesting that the adversarial perturbations have managed to exploit weaknesses in the NST network's optimization process. Our experiments also revealed that the effectiveness of the PGD attack varied across different images, with some images being affected more significantly than others. This variation can be attributed to several factors, including the complexity and characteristics of the content and style images, as well as the inherent robustness of the NST network's learned features.

The network's ability to extract and preserve essential content and style features depends on the quality of its training data and optimization process. If the network has learned more robust features during training, it may be more resilient to adversarial attacks, leading to a lower impact on the generated images. The varying impact of the PGD attack highlights the complex interplay between content and style characteristics and the robustness of the NST network's learned features.

In conclusion, these results demonstrate the susceptibility of the NST network to adversarial attacks and emphasize the need for developing more resilient neural style transfer algorithms that can effectively adapt and respond to adversarial perturbations.

H. FGSM Attack

The Fast Gradient Sign Method (FGSM) Attack is another popular adversarial attack in the domain of classification networks. It generates adversarial examples by adding a scaled perturbation to the input image, where the perturbation is computed from the gradient of the loss function with respect to the input image. The scaling factor for the perturbation is determined by the epsilon parameter, which controls the magnitude of the adversarial noise.

In our adaptation of the FGSM attack for style transfer, we computed the gradients based on the content and style losses between the input content and style images and their stylized outputs. Instead of maximizing the classification loss, we maximized these combined losses to create adversarial examples that can influence the style transfer process. The resulting adversarial images have perturbations that affect the output of the style transfer network when used as input style images.

I. Carlini-Wagner (CW) Attack

The Carlini-Wagner (CW) attack is an optimization-based adversarial technique that finds minimal perturbations to cause targeted misclassification. It is generally more effective than single-step attacks. However, when applied to style transfer, the CW attack failed to achieve the desired outcome. This is because the attack's objective function is primarily designed for classification tasks and may not be suitable for style transfer, which involves minimizing content and style losses as continuous quantities.

Furthermore, the CW attack struggles to find the delicate balance between content and style in neural style transfer algorithms. The attack may introduce overly strong or weak perturbations, making it difficult to achieve the desired output. Additionally, the complex nature of the style transfer process, involving multiple layers and interactions in the neural network, can challenge optimization-based attacks like the CW attack. In conclusion, the CW attack may not be well-suited for style transfer tasks due to problem structure differences and the process's complexity.

J. DeepFool Attack

DeepFool is an adversarial attack method designed for classification tasks, aiming to find the minimal perturbation required to fool a neural network. In the context of style transfer, the goal would be to create an adversarial image that results in a significantly different style-transferred output. However, implementing DeepFool for style transfer was unsuccessful due to several reasons.

Firstly, style transfer is a regression-like problem involving continuous content and style losses, making it challenging to apply DeepFool directly. Additionally, defining an appropriate objective function is difficult since there is no clear decision

boundary between content and style losses. The complexity of style transfer and the robustness of style transfer algorithms also contribute to the challenges faced in applying DeepFool.

In conclusion, applying DeepFool to style transfer tasks is not straightforward due to differences in problem structure and the inherent complexity of style transfer. Further research and adaptation of the algorithm may be necessary to make it effective for style transfer tasks.

IV. Metrics

A. Execution Time and Visual Loss

We investigated the behavior of three state-of-the-art neural style transfer networks under adversarial attacks: Neural Neighbor Style Transfer (NNST), AdaIN Style Transfer, and Fast Style Transfer. Each network has distinct characteristics, affecting their performance under adversarial attacks differently.

NNST relies on finding matching patches between content and style images for a faithful style transfer, whereas AdaIN Style Transfer employs instance normalization layers for faster results. Fast Style Transfer uses a feedforward approach for even greater computational efficiency. Our results showed that the Projected Gradient Descent (PGD) attack was most effective on the AdaIN Style Transfer network, exploiting its vulnerability to perturbations in input data.

In terms of execution time, Fast Style Transfer performed the best, both before and after adversarial attacks. Adversarial attacks increased execution times for all networks, but the relative differences remained consistent. To assess visual quality and content and style preservation, we used perceptual loss as a performance metric.

Our findings revealed that the PGD attack caused the most significant increase in perceptual loss for the AdaIN network, indicating the strongest adversarial effect. NNST and Fast Style Transfer networks exhibited smaller increases in perceptual loss, suggesting greater robustness to the PGD attack.

B. Attacks Success Rates

The PGD and FGSM attacks achieved partial success in perturbing the style transfer process. PGD, an iterative method, introduced noticeable perturbations, while FGSM, a single-step method, produced results similar to original content images, indicating NST algorithms' relative robustness. However, increasing epsilon could result in more noticeable perturbations at the cost of visible artifacts.

The CW attack, targeting the smallest perturbation for misclassification, was unsuccessful, and the DeepFool attack could not be implemented for neural-style transfer networks. Adapting these attacks to the unique characteristics of neural style transfer networks is challenging, highlighting the complexity of applying adversarial attacks to neural style transfer and providing insights into their robustness and vulnerabilities.

C. Neural Net Architecture Comparison and Possible Improvements

Neural Neighbor Style Transfer (NNST) is a patch-based approach focusing on matching patches between content and style images, unlike traditional NST or other deep learning techniques. AdaIN Style Transfer uses a single-pass approach with a pre-trained encoder and decoder, making it faster but more susceptible to adversarial attacks. Fast Style Transfer employs a feed-forward network trained on a specific style, optimized for speed and real-time applications, but limited in generalization to other styles. Some networks, like AdaIN, are more sensitive to certain attacks due to architectural differences.

To improve neural style transfer networks' robustness against adversarial attacks, consider adversarial training, defensive distillation, and regularization techniques. Adversarial training incorporates adversarial examples during training, defensive distillation trains a distilled version of the network to reduce sensitivity to adversarial input perturbations, and regularization techniques prevent overfitting and improve generalization capabilities.

These strategies have the potential to enhance neural style transfer networks' robustness, making them more resilient against adversarial attacks. However, due to time limitations, the suggested improvements have not been tested. Further research and experimentation are required to validate these approaches and identify the most promising methods for strengthening neural style transfer algorithms under adversarial attacks.

V. Conclusion and Future Work

In this paper, we analyzed three state-of-the-art neural style transfer networks (NNST, AdaIN Style Transfer, and Fast Style Transfer) under four adversarial attacks (PGD, FGSM, CW, and DeepFool) to understand their vulnerabilities and identify methods for enhancing reliability and robustness.

Our findings revealed that PGD and FGSM attacks partially perturbed the style transfer process, while CW attack was unsuccessful and DeepFool attack couldn't be implemented. AdaIN Style Transfer was most sensitive to the PGD attack, and Fast Style Transfer showed the best execution times. We proposed strategies like adversarial training, defensive distillation, and regularization techniques to improve network robustness.

Future work should focus on validating these approaches and exploring novel adversarial attacks specifically designed for style transfer networks. Investigating the transferability of adversarial attacks across different networks could help identify common weaknesses and inform the development of more general defenses. This study lays the groundwork for research focused on improving the robustness of neural style transfer networks, ensuring their reliable and secure application in various contexts.

Comparison of execution time

Network	Execution Time (Before Attack)	Execution Time (After Attack)	Increase in Time
NNST	28	30	2
AdaIN	1.7	2	0.3
Fast Style Transfer	0.2	0.3	0.1

Fig. 1. Time Comparison

Fig. 2. Dataset

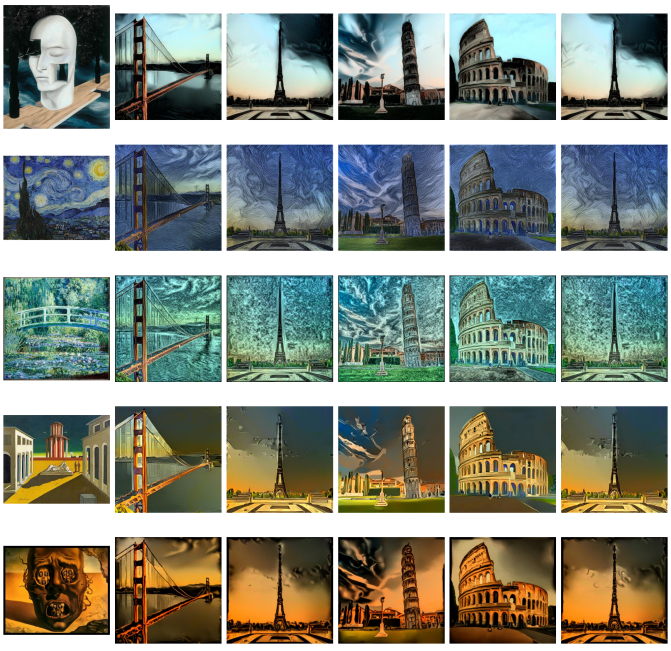

Fig. 3. Neural Neighbor Style Transfer Results

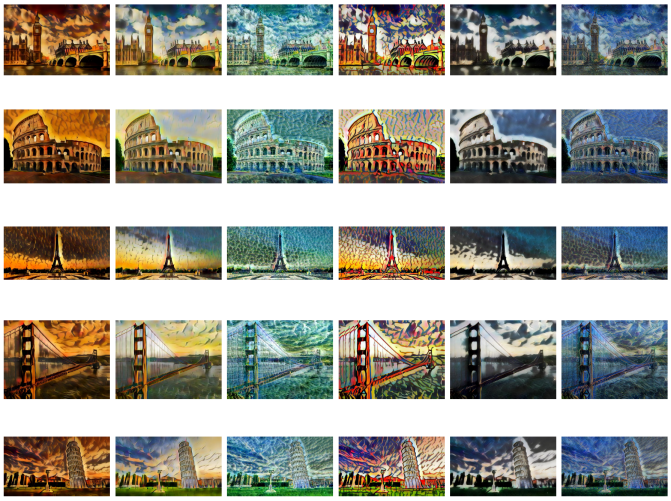

Fig. 4. AdaIN Style Transfer Results

Fig. 5. Fast Style Transfer Results

Fig. 6. PGD on NNST

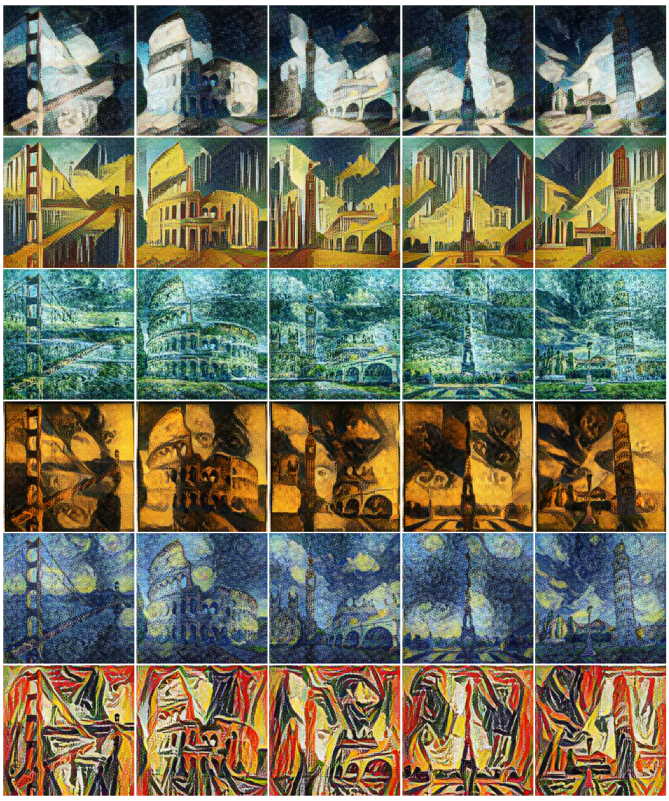

Fig. 8. PGD on Fast Style Transfer

Fig. 7. PGD on AdaIN

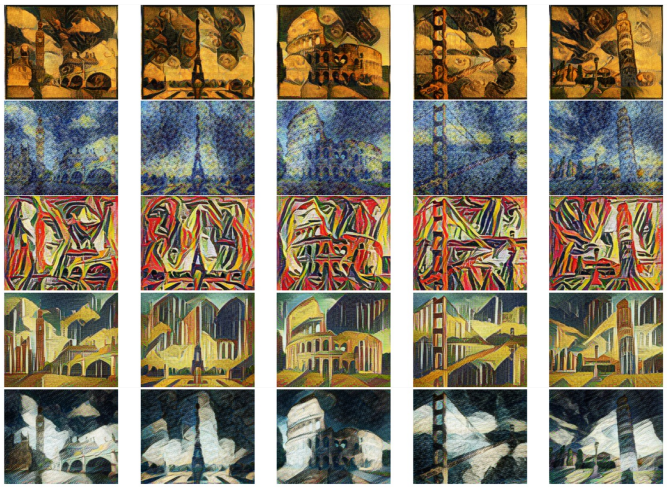

Fig. 9. FGSM attack on Fast Style Transfer

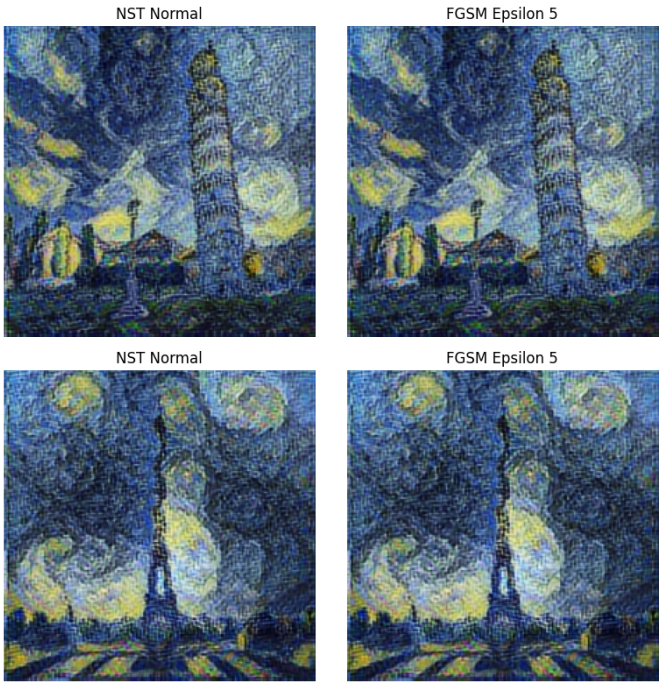

Fig. 10. Original VS Epsilon 5

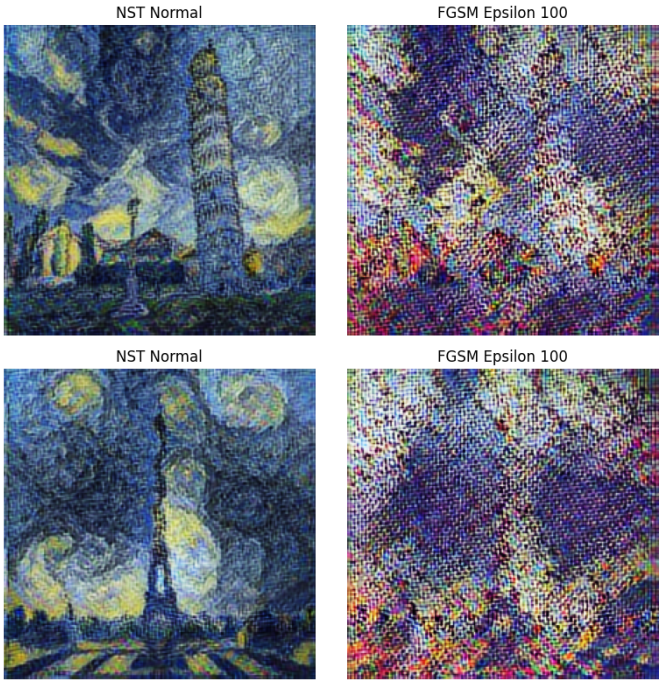

Fig. 11. Original VS Epsilon 100

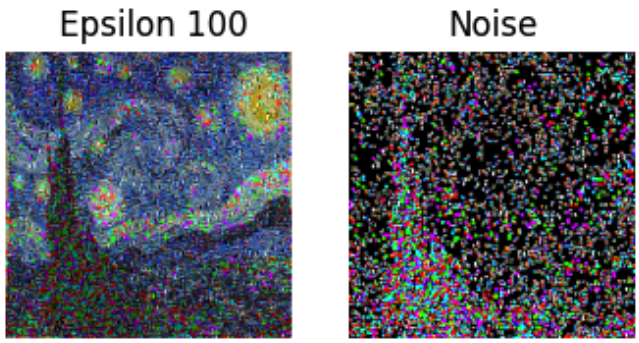

Fig. 12. Attack and Noise

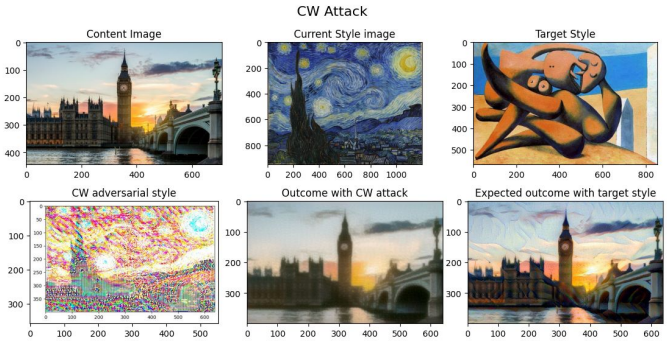

Fig. 13. CW attack Images

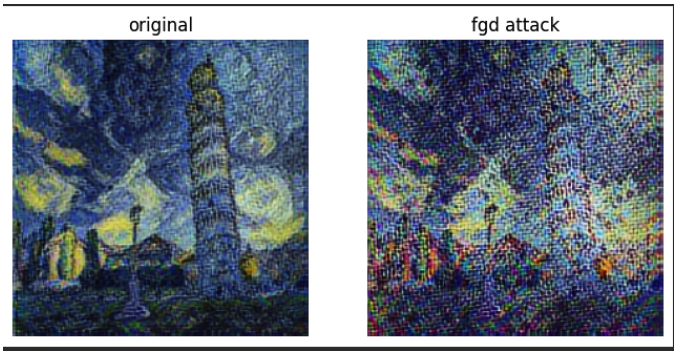

Fig. 14. Noise

REFERENCES

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. UC Berkeley, Google Research, UC San Diego.

[2] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge (2016). Image Style Transfer Using Convolutional Neural Networks. The University of Tübingen, Tübingen, Germany.

[3] Xun Huang, Serge Belongie (2017). Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. Cornell University, Ithaca, NY, USA.

[4] Justin Johnson, Alexandre Alahi, Li Fei-Fei (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Stanford University, Stanford, CA, USA.

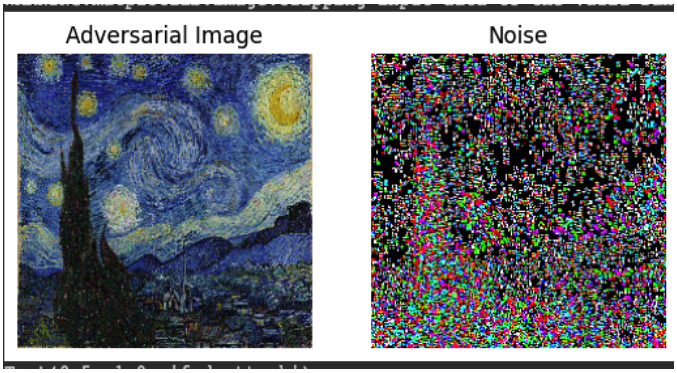

Fig. 15. Adversarial Image and Noise

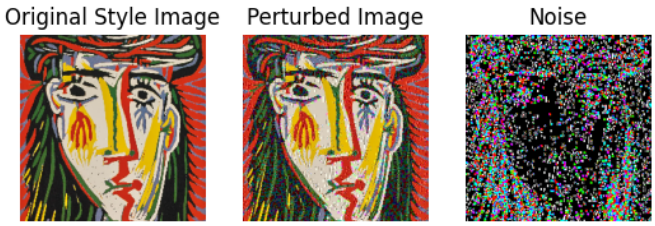

Fig. 16. Original VS Adversarial Image

[5] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, Ming-Hsuan Yang (2020). Less is More: Faithful Style Transfer without Content Loss. Adobe Research, San Jose, CA, USA.

[6] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy (2014). Explaining and Harnessing Adversarial Examples. Google Inc., Mountain View, CA, USA.

[7] Nicolas Carlini, David Wagner (2017). Towards Evaluating the Robustness of Neural Networks. University of California, Berkeley, CA, USA.

[8] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. Massachusetts Institute of Technology, Cambridge, MA, USA.

[10] Xu, H., Ma, Y., Liu, H. C., Deb, D., Liu, H., Tang, J. L., and Jain, A. K. (2020). Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17, 151-178.