Natural Adversarial Examples And Boosting It's Transferability

Ajay R Nair
CET-IS
IIST INTERN

Guided By :-
Dr. Deepak Mishra
Dr. Mahesh Sreekumar Rajasree

AVIONICS

IIST

Outline

Introduction

Key Concepts:

1. **Adversarial Examples:**
   - These are inputs deliberately modified to mislead machine learning models into making incorrect predictions.
   - Typically, these modifications are subtle and often imperceptible to humans but can cause significant errors in model outputs.
2. **Transferability:**
   - This refers to the ability of adversarial examples to deceive different models beyond the one they were originally crafted to attack.

Introduction

Objective:

- The primary aim is to improve the robustness of machine learning models by understanding and enhancing the transferability of adversarial examples.
- This involves creating adversarial examples that not only deceive the model they were designed for but also fool other models.

Examples Of Adversarial Images

Figure 1: Image from IMAGENET- A, where the black text is the original image and red colour text is the ResNet-50 prediction

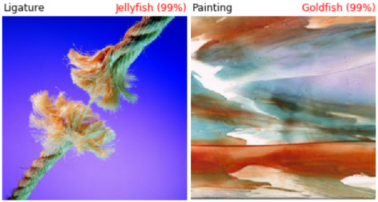


Figure 2: Image from IMAGENET- O, where the black text is the original image and red colour text is the ResNet-50 prediction

Literature Survey

Findings from ’Boosting the Transferability of Adversarial Examples’ (Liang, CVPR 2023):

Approach:

- The paper introduces novel techniques to create more transferable adversarial examples.
- These methods involve optimizing the perturbations applied to images to ensure they remain effective across different models.

Results:

- The new techniques significantly improve the success rate of adversarial attacks across multiple models compared to traditional methods.

Literature Survey

Findings from 'Natural Adversarial Examples' (Hendrycks et al., CVPR 2021):
Approach:

- IMAGENET-A includes difficult images that models trained on ImageNet often struggle to classify correctly.
- IMAGENET-O contains images of categories not found in the ImageNet-1K dataset, testing models' ability to handle new, unseen categories.

Results:

- Models perform poorly on IMAGENET-A, showing they are vulnerable to naturally difficult examples.
- On IMAGENET-O, models often confidently misclassify these new, out-of-distribution images, showing weaknesses in detecting anomalies.

What I Have Learned

- Adaptive Instance Normalization is an instance normalization approach primarily used in style transfer, whose main objective is manipulating the style of an image while maintaining its content
- This means that we can manipulate the style features of DNN by controlling the IN layer (Instance Normalization (IN) is a normalization technique used to stabilize and speed up the training of neural networks.)
- As we know that current transferable attack don't distinguish between style and content features which limits their transferability

What I Have Learned

- So that, stylized model is created by inserting an IN layer into the original surrogate network.
- Using different parameters for the IN layer, we can inject different styles.
- Therefore,to improve transferability, we use stylized networks (neural networks that incorporate style information into their architecture) as surrogate models.
- This method can be easily combined with existing methods.

What I Have Learned

- Adversarial attacks can mislead DNNs by adding small perturbations to benign images.
- The attackers use a white box DNN (a model where the internal workings are fully accessible and understandable) as a surrogate model
- The attack transferability means the generated adversarial examples can attack other black-box DNNs (neural network model whose internal workings are not easily interpret able or understandable by humans)

What I Have Learned

Creation Of Styless Model

- We split the original surrogate model into two parts: F1 and F2, and insert an IN layer to create a stylized network.
- The IN layer has two parameters: μ and σ .Initially, set these parameters to the mean and variance of the input.
- By perturbing the parameters of the IN, we can generate different stylized networks.

What I Have Learned

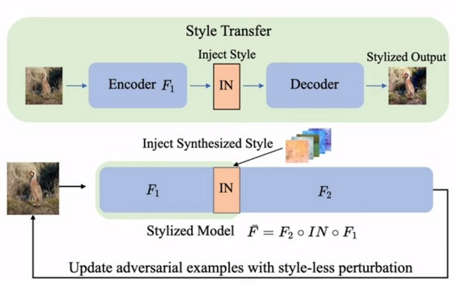


Figure 3: An overview of our StyLess attack. We create stylized model F by injecting synthesized style features into the surrogate model (F = F 2 ◦ F 1) using an adaptive IN layer.

Implementation And Output

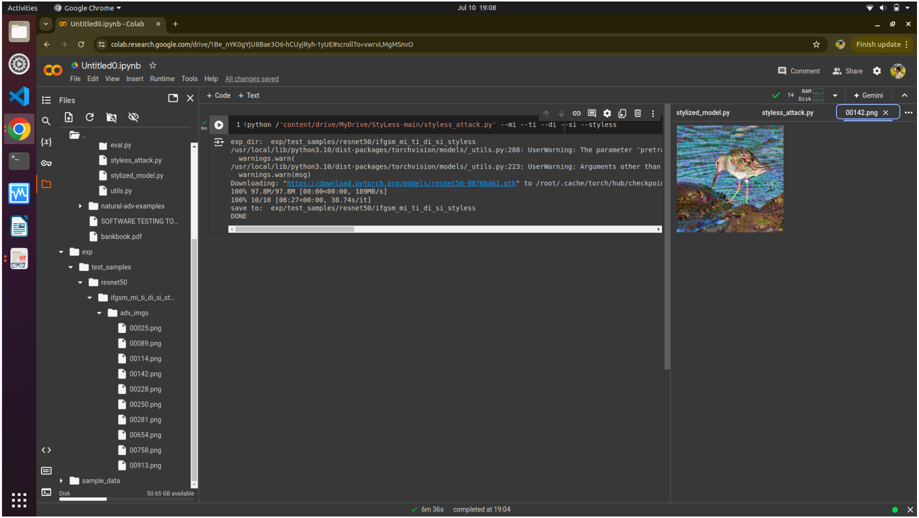


Figure 4: Input

Implementation And Output

Initialization and Loading Models

- The `StylizedNet` class is initialized with a specific model (e.g., ResNet50) and loads a pre-trained version of it.
- An instance normalization layer is generated and inserted into the model to introduce style variations.

FGSM Attack Implementation

- The FGSM attack is implemented to create adversarial examples by adding a small perturbation to the input image in the direction of the gradient of the loss with respect to the input image.

Implementation And Output

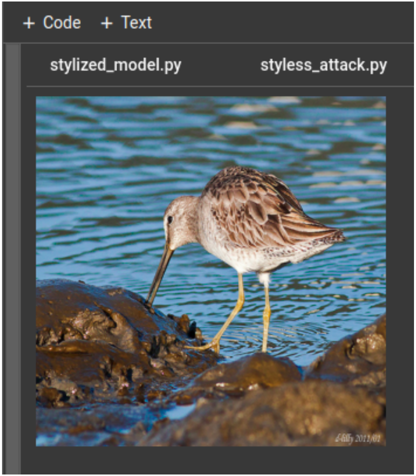
Applying the Stylized Layer

- The stylized layer is set and applied during the forward pass of the model, allowing the network to handle style variations in the input images.

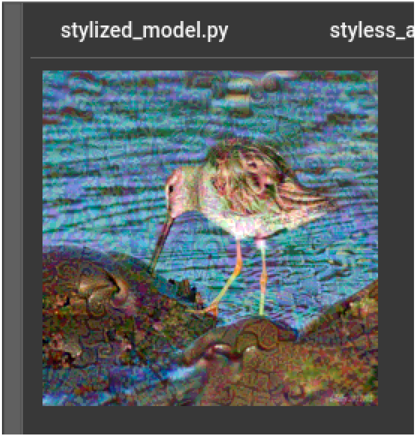
Saving and Loading Parameters

- The parameters for the stylized layers can be saved and loaded, allowing for consistent application of specific styles across different runs.

Implementation And Output



(a) Original Image

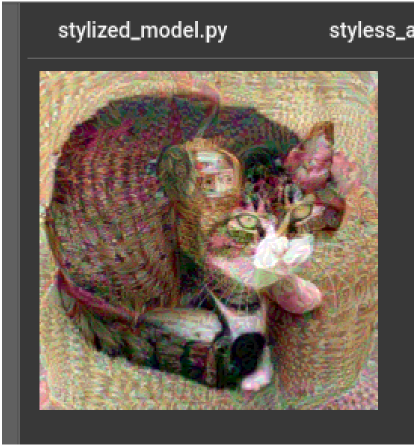


(b) FGSM Attacked Image

Figure 5: Output

Implementation And Output

(a) Original Image



(b) FGSM Attacked Image

Figure 6: Output

Implementation And Output

(a) Original Image

(b) FGSM Attacked Image

Figure 7: Output

Implementation And Output

- The Fast Gradient Sign Method (FGSM) is a technique used in adversarial attacks on neural networks.
- It generates adversarial examples by perturbing the input image in the direction of the gradient of the loss with respect to the input.
- This small perturbation is intended to cause the model to misclassify the input.
- This approach allows the model to be more robust to style variations and adversarial attacks by incorporating style-specific features into the training process.

Conclusion

- Stylized networks using adaptive instance normalization proved effective in boosting transferability.
- The Fast Gradient Sign Method (FGSM) was used to generate adversarial examples by adding small perturbations to input images.
- The incorporation of style variations into the neural network's training process improved the network's robustness to adversarial attacks.
- The techniques developed demonstrated higher success rates in fooling multiple models, making the models more resilient against adversarial attacks.

Future Enhancement

- **Integration of advanced adversarial training methods:-** This would be the case wherein the current method will be combined with sophisticated adversarial training techniques to build more robust models.
- **Investigation of Other Stylization Approaches:-** Test other stylization approaches rather than Adaptive Instance Normalization for the generation of further transferable adversarial examples.
- **Research in Cross-Domain Adversarial Attacks:-** Extend the current methods to other domains and datasets and test their generalizability.

Internship Summary

- The project successfully demonstrated that stylized networks improve the transferability of adversarial examples.
- FGSM attacks, combined with stylized networks, show higher success rates across multiple models, enhancing robustness against adversarial attacks.
- The techniques developed in this project provide a robust method for creating transferable adversarial examples, contributing to the ongoing efforts in improving the security and reliability of machine learning models.

Thank you!