CSci 5980 Project Proposal

Minimal Noise Generator under an Adversarial Framework

Xinpeng Shen shenx582@umn.edu

Yifan Hu huxxx988@umn.edu

Prabhjot Singh Rai rai00027@umn.edu

1 Introduction

Many state-of-art works on computer vision areas use convolutional neural network (CNN) as the main structure. However, CNNs are designed to deal with the translation of objects and are not capable of understanding images in terms of objects and their parts. Therefore, most advanced CNNs are just learning richer and richer description of the object's 2D projection. As a result of this nature, CNNs are known to be fooled easily. One can change the output of a CNN network by simply adding tiny random noise to the original images. Some works have been done to generate the noise via machine learning approaches [1,2]. This project intends to extend the previous work by generating minimal noises that can effectively fool the classifier and explores the information carried by the noise itself.

2 Methodology

The basic framework of this project will be similar to the framework in discussed in [2]. Where a generative adversarial network (GAN) is designed to train the image noise generator. Some CNN networks are designed as generators and one pre-trained classifier (e.g. ResNet) is designed as a discriminator. The generator produces noise that will be added to the original image. Then the original image and the salted image will then send to the discriminator. The discriminator will determine the label of the image, and the results are backpropagated to the generator network, thus, form a complete GAN training cycle.
Besides reproducing the referenced result, we plan to make the following improvements based on the work of [2]:

1. Add penalty to the scale of noise.
   In this project, we are looking for minimal noise that can efficiently fool the classifier. Therefore, the noise should be as close to 0 as possible. This should eliminate redundant noise.
2. Separate structure and texture noise.
   In this project, we are trying to find out how much information is provided by the structure of the image and how much is by the texture. This would help us better understand the mechanism of CNN networks.
3. Explain the noise.
   By applying the two steps above, the generator should only generate noises that perform a surgical strike on the classifier. Thus, the noise itself will provide information about which part of the image is contributing the most to the classification.

References

[1] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. arXiv preprint arXiv:1610.08401, 2016.

[2] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie. Generative Adversarial Perturbations. arXiv preprint arXiv:1712.02328, 2018.