Supervised Adversarial Attack Classification

Kumar Kallurupalli, Nicholas Tan, Shalini Keshavamurthy
{kksreddy,ntan2012,skmurthy}@stanford.edu

Introduction

- Deep neural networks are very sensitive to manipulated inputs.
- One way to manipulate the inputs is through non-targeted attacks.
- Non-targeted attacks slightly modify source image in a way that image will be classified incorrectly by generally unknown machine learning classifier.
- The goal of our study is to implement a robust and lightweight system to detect and classify such attacks.

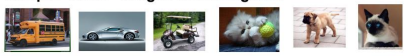
Data

Source - ImageNet dataset and generated adversarial attacks

Dataset split - 92%-5%-3%

Preprocess step -
- Reshape images to 256x256 RGB.
- Subtract mean from RGB layers

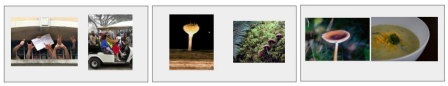
Sample "clean" images from ImageNet -



Adversarial attack - Gradient-based attacks
- Fast Gradient Sign Method (FGSM),
- Contrast,
- DeepFool,
- Projected Gradient Descent (PGD),
- LBFGS

Sample "adversarial" images from generated dataset

Associated "clean" images from original dataset

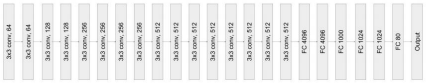


Method

- Train supervised binary classifier on two labelled classes: Adversarial images, Regular (clean) Images
- Use transfer learning on VGG19 to build features of image

Architecture

VGG-19-Extended NN
#Layers: 22 (19+3)
Activation: Relu
Final layer: Softmax



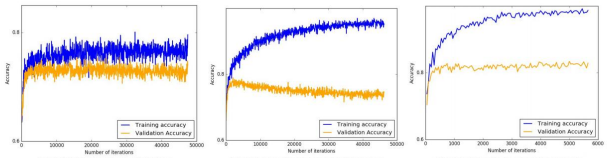
Custom NN
#Layers: 7
Activation: Relu
Final layer: Softmax

Sample Hyperparameters	
Learning rate	0.1, 0.01, 0.001
Optimizer	SGD, RMSProp
DropOut %	30%, 40%, 50%

Design / Tuning

Parameters / Hyper-Parameters tuned:
- Network Depth (Added layers of different types)
- Network Width (Added more nodes in layers)
- Convolutional Filter Sizes
- Learning Rate (Traversed Grid Exponentially)
- Optimization Algorithms (SGD, RMSProp, etc.)

In addition, we also performed tuning on the number of layers that were frozen while performing transfer learning.



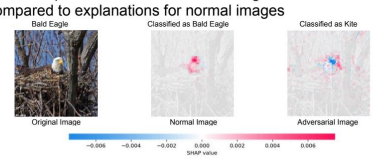
Training result shows High Bias | Training result shows High Variance | Training result shows better result

Results

Attack	Classifier trained on data from	Training Accuracy	Validation Accuracy	Test Accuracy
FGSM	FGSM + Normal Data	93.70%	82.2%	67%
DeepFool				68.12%
Contrast				56.73%

Explainability

Claim: Explanations for adversarial images will not be coherent compared to explanations for normal images



Tool: SHAP (SHapley Additive exPlanations) - open source. SHAP assigns each feature an importance value for a particular prediction. The values are integers where positive values and negative values indicates how much the feature was crucial or not for the prediction of the class.

Explanation for Adversarial Image looks a lot more scattered with a lot of negative correlation

Approach - Feed in images with explanation for most likely class overlayed on top

Architecture And Tuning - Same architecture used previously

Classifier trained on data from	FGSM + normal data with explanations
Training accuracy	93.7%
Validation accuracy	79.8%
Test accuracy	Same as in result section for the 3 attacks

Conclusion

- We trained the dataset on smaller architectures - an extended version of VGG19 and a custom 7-layer network.
- We tried different hyperparameters - optimizers, learning rate, dropouts, etc.
- It is very difficult to classify adversarial images.
- Model trained for one adversarial attack may not be extended to other networks (test accuracy less than 70%)

Future Work

- Try adversarial attack classification using
  - Deeper networks - Example: ResNet50 or Inception
  - Various hyperparameters
- Increase dataset size to 100 classes of ImageNet