Adversarial Attacks on Neural Networks

Milan Chaudhari (2015CSB1010)
Naman Goyal (2015CSB1021)

September 29, 2017

1 Problem Statement

Most existing machine learning classifiers are highly vulnerable to adversarial examples. An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to incorrectly classify it. In many cases, these modifications can be so subtle that a human observer does not even notice the modification at all, yet the classifier still makes a mistake.

Adversarial examples pose security concerns because they could be used to perform an attack on machine learning systems, even if the adversary has no access to the underlying model.

So we would like to implement non-targeted attack for this project, where we would design an adversarial attack that would modify source image in a way that image will be classified incorrectly by generally unknown machine learning classifier, which is known as non-targeted adversarial attack.

2 Data Set

The dataset is available from *Kaggle NIPS 2017 Competition*.

The labelled DEV dataset will be used for training and evaluation. The dataset is ImageNet-compatible, but contains different images which are not used in the original ImageNet dataset.

It contains 1000 images and is available as a part of development package and will be used to evaluate attack.

It is compatible with ImageNet classifiers in TF-Slim library. In practice it means that:

1. All images are 299x299 pixels.
2. Dataset uses 1001 labels. Label 0 is background. Labels 1-1001 correspond to ImageNet classes.

3 References

- *https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack*
- *http://rll.berkeley.edu/adversarial/*
- Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284.