Project Proposal
CSE 569 - Fundamentals of Statistical Learning

Graceful Misclassification of Adversarial Attacks on Neural Networks

Problem description

Neural networks have proven to be powerful models for classification tasks. However, they're also highly susceptible to adversarial attacks via input perturbation. This fact reduces their applicability in security applications, and poses significant risks for NN-based classifiers being implemented into modern technology such as autonomous vehicles.

Recent attempts to reduce the effectiveness of these attacks appeared to be successful against existing methods of generating adversarial samples, only to be quickly defeated by new attack algorithms.

Project summary

For this project, we propose a new method to increase the robustness of neural networks against attacks – not by attempting to prevent adversarial misclassification, but by limiting how far the misclassification is from the ground truth. For example, instead of a perturbed image of a bus being classified as an ostrich, it may instead be classified as a car.

The idea behind how this may be accomplished is simple. For each class, a series of additional classes can be provided for which the training data can be *gracefully misclassified* as. The addition of these classes will result in a greater fragmentation of the input space given the same set of training data. It is believed that by providing a loss function from each class to the provided set of graceful misclassification labels, the cost incurred from misclassification can be controlled and minimized.

OCR Banking domain

When performing character recognition to read the value of a bank check, there is a greater cost incurred by the bank if the '1' character is misclassified as a '9' than if a '2' is misclassified as a '3'.

This domain-specific cost information can be encoded into a loss function that in turn provides a weighted set of misclassification labels for each class. Using the MNIST data set, these new class labels can then be used to provide additional
TODO

Semantic misclassification

If the graceful misclassification technique succeeds in the OCR domain, further experimentation can be done on a more complex domain.