Exploring the Limits of Classifiers in a Hostile Environment

Ishani J Vyas

University of California, Berkeley
Berkeley, USA
ishanivyas@berkeley.edu

***Abstract*—The purpose of this project is to evaluate a representative set of classifiers in as pessimistic a way as possible. I attempt to quantify both the level of effort necessary to deceive each classifier as well as how effective the deception is at that level of effort. The attacks are broken down along two dimensions: how substantial must a change be to change each classifier's predicted label, and how vulnerable is the classifier to an attacker that seeks to force it to predict a specific label.**

***Index Terms*—machine-learning, classification, adversarial, survey**

I. Introduction

The purpose of this project is to evaluate a representative set of classifiers in as pessimistic a way as possible. Knowledge of the limitations of classifiers can help minimize inadvertent negative impacts from applications of machine learning to solving real world problems. Without this useful knowledge, increased application of machine learning also entails increased exposure to risks similar to the internet security problems experienced when ARPAnet was still developing into the World Wide Web.

The project selects a machine-learning problem, constructs classifiers to apply to the problem, briefly evaluates their effectiveness, and then constructs a general attack on the classifiers that attempts to quantify how gullible and vulnerable to manipulation they can be.

II. Setting

A. Data Set

A representative problem, "Fashion" MNIST, was chosen for this project. This dataset seeks to provide a higher bar for classifiers while staying within the parameters of the original MNIST problem: 28x28 images as input with 10 labels for output. Shapes and luminosity play a much more prominent role in this data set, in contrast to the high-contrast lines of the original MNIST problem. "Fashion" MNIST also comes with a pre-defined test set of 10,000 inputs in addition to the training set of 60,000 images. In the course of evaluating the classifiers, it was discovered that some of the classifiers did not operate fast enough when trained with the full training set and tested with the full test set. To alleviate this problem, the training set was reduced to 10,000 images and the test set to 3,000.

B. Classifiers

To survey the domain of classifiers, the following classifiers were chosen for evaluation: Ridge-regression based, Naive Bayes, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Support Vector Machine, DecisionTree, RandomForest, AdaBoost Decision Tree, and Neural Network (Multilayer Perceptron). The classifiers were not extensively optimized for top performance. This seemed reasonable since only the correctly-classified inputs were selected for attack.

III. Strategy

There are many ways to approach deceiving classifiers. Exhaustively exploring all of them is infeasible. To simplify the problem for this project, the set of successfully-classified inputs of each classifier is used to build deceptive inputs which are fed back to the classifier. Specifically, a correctly- classified input $\vec{X_1}$ with label L_1 is combined with another correctly-classified input $\vec{X_2}$ which was classified as L_2 (where $L_2 \neq L_1$) with the aim of changing the classifier's predicted label. The adversary starts by introducing small changes to $\vec{X_i}$ to produce new deceptive inputs $\vec{D}$. The attacker then proceeds to increase the size of the change until the classifier picks the wrong label. A "scale" variable controls the amount of the deceptive input that comes from $\vec{X_1}$ and how much comes from $\vec{X_2}$ according to the equation:

$$\vec{D} = (1-s) * \vec{X_1} + s * \vec{X_2} * \mathcal{N}, \quad s \in [0,1] \qquad (1)$$

$\mathcal{N}$ is a noise term used to explore the effect of noise on the interaction between the adversary and the classifier.

The project takes a white-box adversarial approach to attacking the classifiers. The adversary is free to choose as simple a goal as possible in manipulating each classifier. White-box also means the adversary can see detailed statistics for the classifier under evaluation and use that information to construct deceptive inputs. Specifically, this project uses the label-confusion matrix to select attack vectors. The attacker's knowledge of what kinds of mistakes the classifier is likely to make helps to select labels where a small amount of change may be quite effective.

An adversary attempting to deceive the classifier can have three outcomes. The first outcome is that the classifier does not change its prediction. The second is that the classifier changes

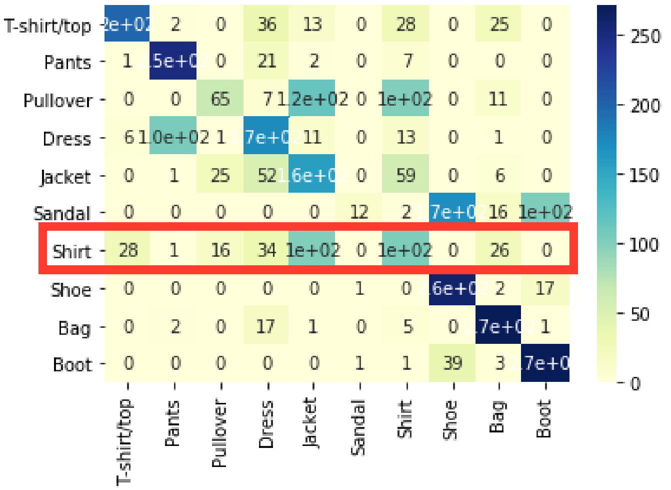

Fig. 1. A confusion matrix for the Quadratic Discriminant Analysis classifier. This shows that “Shirt” images frequently get confused with “Jacket”, “Dress”, “T-shirt/top”, and “Pullover”.

its prediction, but it predicts a label other than the one the attacker seeks to apply (L_2). Depending upon the goals of the attacker, this could be considered partial success. The final possibility is that the attacker succeeds in manipulating the classifier into appling L_2 to the input.

IV. Evaluation Method

For the purposes of this project, the ideal attack on a classifier would allow the attacker to control the output of the classifier by making only an imperceptible change in an image that it was previously capable of classifying. Attacks which change the label of incorrectly-classified results are less useful.

To quantify the amount of change applied to an image, the L2-norm is used. If imperceptible changes are needed to change the decision of the classifier, then the classifier can be considered gullible. The second dimension simply establishes how well the attacker can control the decision of the classifier. If the attacker can force the classifier to choose a particular label, then the classifier is more vulnerable to manipulation.