Summer Semester 2020

University of Potsdam

Institute of

Informatik and Computational Science

MSc Data Science

Research Module

Adversarial Examples for Eye-State Classification

Expose of the Project

Name : Sefika Efeoglu

Martikel-Nr : 799932

efeoglu@uni-potsdam.de

Table of Content

1. Introduction

Deep learning is a class of machine learning which uses neural network algorithms for unsupervised and supervised learning. Deep neural networks have many layers, and one of the most used variants of it is convolutional neural networks. It is used in speech recognition, object detection, computer vision, natural language processing, self-driving area with some variants of neural networks such as RNN, CNN, GAN. [1] One important and appealing application domain is self-driving cars, for example, deep learning techniques can help self-driving cars explore the environment, such as traffic signs and surrounding objects, using the images taken from cameras on the car.

Image classification which is the task of assigning an input image one label from a fixed set of categories is one of the most common problems in the computer vision area. However, [2] it has some challenges such as viewpoint variation, scale variation, deformation, intra-class variation and background clatter. Therefore, image classifiers (machine learning models) might have classification problems which are defined as misclassification due to noise and natural transformation of real world examples. That is to say, when machine learning classifiers are worked in the real world tasks, they are prone to fail when the training and test distribution differ.

As a result, to consolidate its preference in the real world applications, the lack of robustness, confidence or uncertainty estimators are problems that deep learning models need to overcome. These requirements for artificial intelligence systems are becoming harder.

2. Motivation and Problem Setting

The safe control between vehicle and driver is a significant prerequisite for automated driving whether the driver is able to take control can be evaluated using eye state detection. It is important that such systems are robust against changing real world conditions like lighting and natural transformation. These real world conditions, which might not be aware of humans, lead to fooling a machine learning or deep learning model. [3] Deep neural networks which are the kind of machine learning models that have recently resulted in dramatic performance improvements in a wide range of applications are vulnerable to tiny perturbations of their inputs (images). This leads to misclassification problems, namely error in the accuracy of the model. Adversarial examples are specialised inputs created with the

purpose of confusing a neural network, resulting in the misclassification of a given input. These notorious inputs are indistinguishable to the human eye, but cause the network to fail to identify the contents of the image. Additionally, the aim of adversal examples is to disturbed the well trained machine learning model. [4] However, small adversarial perturbation should not result in a significant impact on the out of the model for a trained and robust machine learning model. Consequently, generating adversarial perturbation as negative training examples can improve the robustness of the model.

With respect to our dataset, images of eyes consist of various eye-states which are labelled as open, partially open, closed, and not visible. Natural transformations like angle of input images, viewpoints might lead to misclassification problems in machine learning models. The machine learning model has sensitive measurement to decide the eye-states, so small perturbation of an image fools the deep learning model.

3. Objectives

Using adversarial examples, the project aims to improve the robustness and accuracy of a machine learning model which detects the eye-states against small perturbation of an image, and to solve the misclassification problem caused by natural transformation.

4. Methodology (Approach)

To address the misclassification issue and improve the robust of the eye-states detection, the project will benefit from the advantages of adversarial examples.

Fast Gradient Sign Method (FGSM) will be applied to compute adversarial perturbation as well as the experiment in [5] will be repeated for Eye-States Dataset in this project. Therefore, in this section, Parseval Networks and the experiments in [5] will be focused on. The steps in Figure 1 will be followed to develop the project;

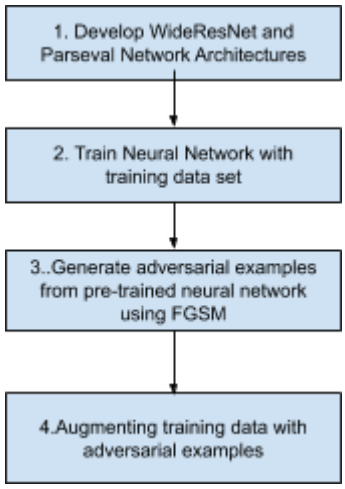

Figure 1: The steps of the methodology in the project.

The rest of this part has given an overview about dataset, image classifiers (neural networks), FGSM and Adversarial Training.

Eye-State Dataset

Eye-State Dataset consists of 4 labels; closed, open, partially open and not visible and 5722 images. Table 1 below gives information about the number of labels.

Label	Count
closed	1500
open	1500
partially Open	1376
not visible	1346

Table 1: İnformation about data

Image Classifiers

Parseval Networks and Wide Residual Network will have been trained through the project. Wide Residual Network has 16 depth and 2 width. Parseval Network has orthogonality constraint and orthogonal weight initializer as difference from Wide Residual Network. The architecture of Wide ResNet is in the appendix section. The orthogonality constraint is explained in further sections after FGSM.

Fast Gradient Sign Method

Fast Gradient Sign Method (FGSM) is to add the noise (not random noise) whose direction (sign) is the same as the gradient of the cost function with respect to the data. The noise is scaled by epsilon, which is usually constrained to be a small number via max norm. The value of gradient does not matter in Formula 4, but the direction (+/-) of it affects.
To find adversarial examples, the formula below is used in this method:

$$X_{adversarial} = X + \varepsilon.sign(\nabla_x J(\Theta, X, Y))$$

Formula 1: shows how adversarial examples are found using FGSM. To calculate perturbation, the sign of gradient of loss is multiplied by epsilon.

To generate adversarial examples, adversarial perturbation, which is calculated using $\varepsilon.sign(\nabla_x J(X, Y))$ in Formula 1, is added to the image (X). J(X,Y) is calculated by a machine learning model, and then the gradient of it is computed to find the direction.

Orthogonality Constraint for Parseval Network

Parseval Network has an orthogonality constraint on the weight matrices. This is called an optimization algorithm. This constraint is called in convolution layers of the network and provides faster converges on layer-wise regularization. This weight update is runned in hidden layers on the training after each gradient update. The results of this approach can be seen in preliminary results.

$$W_k = (1 + \beta)W_k - \beta W_k W_k^T W_k$$

Formula 2 shows the orthogonality constraint.

Data Augmentation using Adversarial Examples

The networks which are defined above will be trained to build more robust classifiers for Eye-State Dataset using adversarial training approach. Adversarial Training approach is defined as mini-batch training, which means the half of true examples changing with their adversarial examples per batch on the training.

The Evaluation of Non-Adversarial and Adversarial Training

To evaluate both training approaches, the test dataset which has different SNRs (Signal to Noise Ratios) will be evaluated on all models that are obtained using two different training approaches. Signal to Noise ratio is used in imaging, and provides to interpret the accuracy values of the models easily. In addition to this, it is used in engineering and science. The SNR is calculated by $20log_{10}\frac{||X||_2}{||\delta_x||_2}$ [5]. $||X||_2$ is defined as a signal which refers to an image in our problem, whereas $||\delta_x||_2$ is called a noise of signal which is calculated by FGSM for each image and is defined as perturbation.

4 different SNRs are used for the evaluation of the models and adversarial training approach. To obtain more accurate results, 10 folds cross-validation will be applied on adversarial training.

5. Expected Scope of the Work

Neural network models will have been implemented through the project. The experiments are developed as follows,

- Training of neural network with clean Eye-State dataset
- Calculation of perturbation using Fast Gradient Sign Method as well as generation of the adversarial examples using perturbation and an input image.
- Augmenting training data with adversarial examples.
- To obtain the perturbation of an image, cleverhans library will be used through the project.

An improvement between the evaluation results of adversarial training and non-adversarial training on different SNR levels is expected at the end of this project.

Preliminary Results of the Project

Figure 2 illustrates the learning curves of Parseval with orthogonality constraint and Wide Residual Network on non-adversarial training.

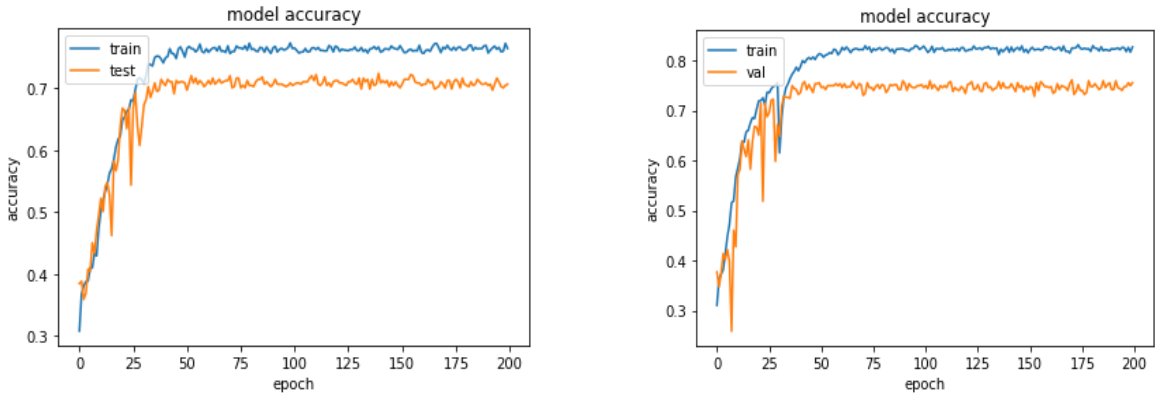

Figure 2: learning curves of models (Wide ResNet (left) and Parseval Network with OC (right)).

The table below shows the preliminary results of the project on 4 different SNR levels and clean test datasets.

Models\SNR	Clean	~50	~45	~40	~33
WideResNet	0.731239	0.694415	0.665794	0.600873	0.474346
Parseval_OC	0.775218	0.731763	0.700873	0.62164	0.4726

Table 2 : shows the evaluation results of non-adversarial training against different noise levels. 10 folds cross validation is applied to obtain these results. It is clear that orthogonality constraint provides faster and more accurate results when the results of Parseval(OC) are compared with Wide ResNet.

References

1. “Adversarial Deep Learning for Autonomous Driving.” Berkeley DeepDrive | We Seek to Merge Deep Learning with Automotive Perception and Bring Computer Vision Technology to the Forefront., deepdrive.berkeley.edu/project/adversarial-deep-learning-autonomous-driving.
2. CS231n Convolutional Neural Networks for Visual Recognition, cs231n.github.io/classification/.
3. “Adversarial Examples, Explained.” KDnuggets, www.kdnuggets.com/2018/10/adversarial-examples-explained.html.
4. Sun, Sining, et al. “Training Augmentation with Adversarial Examples for Robust Speech Recognition.” Interspeech 2018, 2018, doi:10.21437/interspeech.2018-1247.
5. Agarwal, Chirag, et al. “Parseval Networks: Improving Robustness to Adversarial Examples by Encouraging Discriminative Features.” 2019 IEEE International Conference on Image Processing (ICIP), 2019, doi:10.1109/icip.2019.8803601.