1-20

Improvement of Image Misclassification by Adversarial Perturbations Using Convolutional Neural Network

Futo SHINOHARA Shu SUMIMOTO Yuichi MIYATA Yoko UWATE Yoshifumi NISHIO
(Tokushima University)

1. Introduction

Convolutional neural networks (CNN) are the most popular techniques employed for computer vision tasks, including but not limited to image recognition, localization, video tracking, and image and videosegmentation. Among them, they are often used for image recognition, and computers have already proven to be superior to human image recognition. However, these deep networks have problem that they are particularly susceptible to adversarial perturbations, such as human malicious intent on input images. For example, it is essential to deal with adversarial perturbations, as misrecognizing a traffic sign during automa driving can lead to a major accident.

In this study, we recognize images by using convolutional neural network (CNN). The purpose is to improve the accuracy of image recognition using images which are intentionally to adversarial perturbations.

2. Proposed system

Image recognition of traffic signs is performed using a model called VGG16. VGG16 is a model of CNN consisting of 13 convolutional layers and 3 fully connected layers.

We train the network using the German Traffic Sign Recognition Benchmark (GTSRB) data set. We prepare two types of data sets, images without processing and images with random black noise of 1 to 10 pixels, and compared the test accuracy. Examples of test images are shown in Figs. 1 and 2. Figure 1 is original image. Figure 2 is noise-processed image.

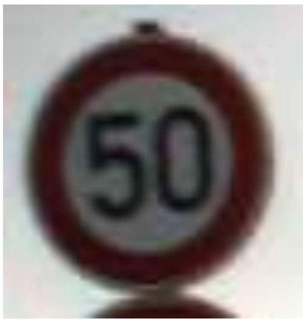

Figure 1: Original image.

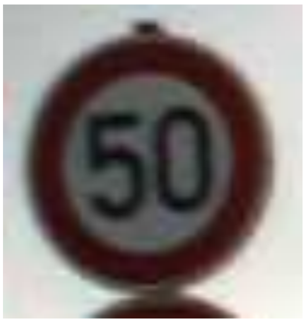

Figure 2: Processed image.

3. Simulation results

We define as the learning steps = 30, the number of learning images = 24,000. As a result of learning this network, Fig. 1 is classified as speed limit 50, however Fig. 2 is misclassified as speed limit 80.

When comparing the test accuracy of the original image and the processed images, the accuracy of the noise-processed images is reduced. In addition, test accuracy is restored by adding the noise-processed images to the images and learning the model again (Table 1). Figure 3 shows the comparison of the training accuracy of the first model and the relearning model.

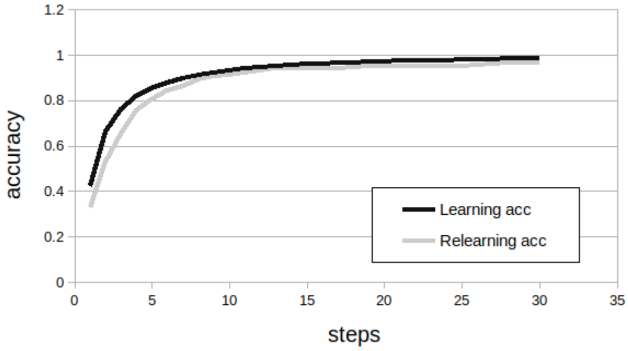

Figure 3: Training accuracy and validation accuracy.

Table 1: Training and test accuracies.

Learning images	Original	Original	Processed
Test images	Original	Processed	Processed
Training accuracy	0.98	0.98	0.97
Test accuracy	0.81	0.78	0.79

4. Conclusion

In this study, we compared with digitally processed images which are similar noise-processed images and original images for deep learning. We investigated the learning accuracy of noise-processed images and original images. It turned out that it is possible to reduce learning accuracy only by giving simple noise as a adversarial perturbations.

As our future works, it is expected that resistance will be further enhanced by enhancing the learning data set by adding other contrast adjustment, blurring, and other noise. We also would like to find out the noise patterns that are misclassified.

Furthermore, we would like to work on improving the accuracy by removing noise using a generative adversary network (GAN) for image preprocessing.