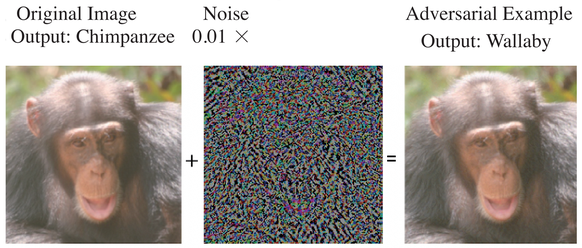
REINFORCING THE ROBUSTNESS OF A DEEP NEURAL NETWORK TO ADVERSARIAL EXAMPLES BY USING COLOR QUANTIZATION OF TRAINING IMAGE DATA

Shuntaro Miyazato, Xueting Wang, Toshihiko Yamasaki and Kiyoharu Aizawa
Dept. of Information and Communication Eng., The University of Tokyo

Introduction

Adversarial Example:
• Image classification using CNN is vulnerable to Adversarial Examples with small perturbation
• Adversarial Example is a possible threat to CNN used in the real-world (such as self-driving car and face-recognition).

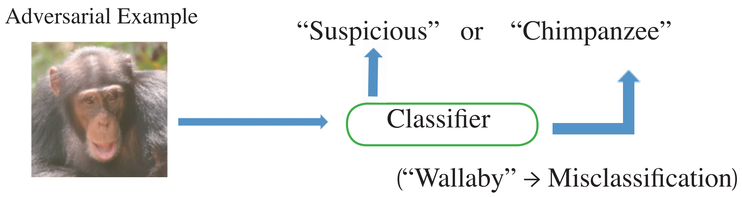


Black-Box Attack: (⇄ White-Box Attack)
• Attackers can generate Adversarial Example without knowing the parameters of the target models because of Transferability
• Transferability: Adversarial Examples for one model tend to induce misclassification through other models as well

Problem:
Detection of Adversarial Example (especially Black-Box Attack) or robustness to Adversarial Example is needed

Contribution

• Loss Mazimization (through color quantization of training data)
• We generated an ensemble of models which can classify Adversarial Examples correctly or reject them as undecidable

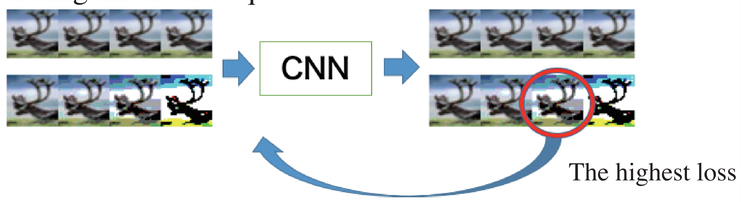


Proposed Method

Color Quantization:
Adversarial noise contains small features
→ Training data are quantized so that models learn only conspicuous features and become insensitive to small features

(TQ means test data are quantized as well for comparison)

Loss Maximization (LM):
Recent works show training by high-loss adversarial examples can generate more robust CNN
→ The loss is maximized immediately before backpropagation by selecting the level of quantization



Ensemble (En):
ResNet-20, VGG-16 and DenseNet trained using LM
• Majority rule (MR): Only if two or three models output the same inference, the image is accepted and the ensenble outputs it
• Unanimous rule (UR): Only if three models output the same inference, the image is accepted

Settings

Dataset: Cifar10 (10, 000 images as test data)
Attack Method: FGSM + Black Box (Attacker know VGG16 trained by the original images)
Comparative Method: Adversarial Training (ResNet-20 trainied using Adversarial Examples with noise which size is 8/255 or 2/255)

Results

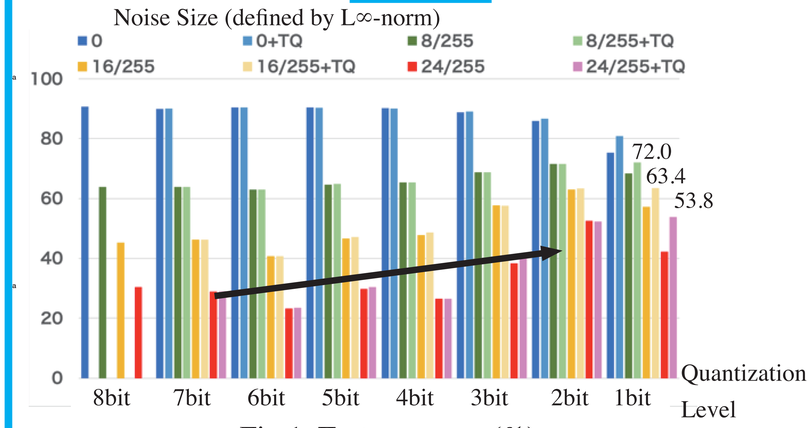


Fig 1. Test accuracy (%)

Table 1. Test accuracy (%) . The values in parentheses show the numbers of accepted data. F-AT means Adversarial Trainig.

Method	Noise Size			
	0	8/255	16/255	24/255
ResNet 8bit	90.7	63.9	45.2	30.4
VGG 8bit	91.2	64.6	57.0	45.7
DenseNet 8bit	92.3	64.8	49.1	29.5
F-AT 8/255	81.3	77.7	70.5	61.4
F-AT 2/255	88.3	80.6	71.1	56.6
LM-ResNet	86.0	74.7	65.1	54.3
LM-VGG	86.7	75.3	68.0	63.0
LM-Dense	87.1	75.8	65.3	50.7
MR-En	94.5(9850)	66.9(9273)	57.6(8976)	38.7(8819)
UR-En	**98.0**(8654)	79.5(6898)	72.7(5253)	41.7(4174)
LM-MR-En	90.0(9736)	78.5(9556)	70.4(9414)	62.5(9129)
LM-UR-En	95.9(8090)	**88.0**(7561)	**82.4**(6966)	**74.5**(5750)

Conclusion

• LM improves the robustness compared with simple quantization.
• The ensemble with LM increases the accuracy of the accepted examples to a better level than F-AT.