Artificial Neural Network
Homework 4 Report

Jiahao Li
Department of Computer Science and Technology
Tsinghua University
lijiahao17@mails.tsinghua.edu.cn

1 Targeted Losses

The targeted adaptation for cross entropy loss is formulated as Equation (1). To understand this formula, recall that the model is trained to minimize the loss between the prediction and the ground truth categories. Similarly, in the targeted attack, the given target class can be regarded as the ground truth class, so the noise is adjusted to minimize the loss between the prediction and the targeted class.

$$\mathcal{J} = \text{crossentropy\_loss}\left(y_{\text{pred}}, y_{\text{target}}\right) \tag{1}$$

The targeted adaptation for C&W attack loss is shown in Equation (2). Similarly, targeting at a specific class, it is necessary to directly maximize the predicted probability on the targeted class, and minimize those on other classes.

$$\mathcal{J} = \max\left\{\max_{y \neq y_{\text{target}}} [\text{logit}(\widehat{x})]_y - [\text{logit}(\widehat{x})]_{\text{target}}, -\kappa\right\} \tag{2}$$

2 Experimental Results

Unless otherwise specified, all experiments are conducted following the same settings stated as follows. The parameter κ is set to 1, and α is set to 1. The optimization step is fixed to 500.

The experimental results for untargeted and targeted attack are presented in Table 1 and Table 2, respectively. The final attack success rate, L_1, L_2, and L_∞ norm are reported. The ↑ (↓) symbol indicates that the better result is achieved when the metric is higher (lower). The best result for each metric is highlighted in bold.

Method	α	β	γ	# Optim Steps	Attack Success Rate ↑	L_1 ↓	L_2 ↓	L_∞ ↓
cross entropy loss	1	0	0	500	**1.00**	19.31	0.4035	0.0083
	1	0.00001	0	500	0.98	17.84	0.3762	0.0089
	1	0	0.001	500	0.95	17.61	0.3702	**0.0080**
	1	0.00001	0.001	500	0.93	**16.86**	**0.3555**	0.0084
C&W attack loss	1	0	0	500	**1.00**	21.83	0.4458	**0.0086**
	1	0.01	0	500	**1.00**	10.89	0.3330	**0.0086**
	1	0	1	500	**1.00**	10.79	0.2975	**0.0086**
	1	0.01	1	500	**1.00**	**9.15**	**0.2935**	0.0087

Table 1: Experimental results for untargeted attack.

Method	α	β	γ	# Optim Steps	Attack Success Rate ↑	$L_1 \downarrow$	$L_2 \downarrow$	$L_\infty \downarrow$
cross entropy loss	1	0	0	500	**1.00**	32.04	0.6563	**0.0162**
	1	0.0001	0	500	**1.00**	32.03	0.6561	0.0163
	1	0	0.01	500	**1.00**	**31.99**	**0.6554**	**0.0162**
	1	0.0001	0.01	500	**1.00**	32.01	0.6558	0.0163
C&W attack loss	1	0	0	500	0.65	30.89	0.6284	**0.0145**
	1	0.01	0	500	0.68	16.82	0.4805	0.0146
	1	0	1	500	**0.72**	15.76	0.4157	0.0151
	1	0.01	1	500	**0.72**	**13.60**	**0.4080**	0.0152

Table 2: Experimental results for targeted attack.

In general, the C&W attack loss obtain better performance than the cross entropy loss, achieving a higher attack success rate as well as a lower perturbation. However, there is a serious problem with the given hyperparameters for cross entropy loss, where α is much larger than β and γ, resulting in the ineffectiveness of the regularization term. For example, in the untargeted attack settings with cross entropy loss, α is set to 1, and β is set to 0.00001, where α is 100,000 times as large as β. Eventually, in the first batch, the final cross entropy loss is -16.38, while the weighted regularization loss is only 0.0007, leading to a conspicuous perturbation in the final result.

3 Best Experimental Settings

From my perspective, the best adversarial example misleads the neural networks at the lowest cost, *i.e.*, the smallest perturbation from the original image.

The most successful experimental settings for untargeted and targeted attack remain the same. They both apply C&W attack loss with α of 1, β of 0.01, and γ of 1. In these successful settings, the final result achieved a superior attack success rate as well as a small perturbation, which makes a qualified adversarial example.

To intuitively validate the best experimental settings, I visualize the best qualitative results in untargeted attack, as shown in Figure 1. With a small perturbation, the noise centers on the main object and significantly misleads the classifier.

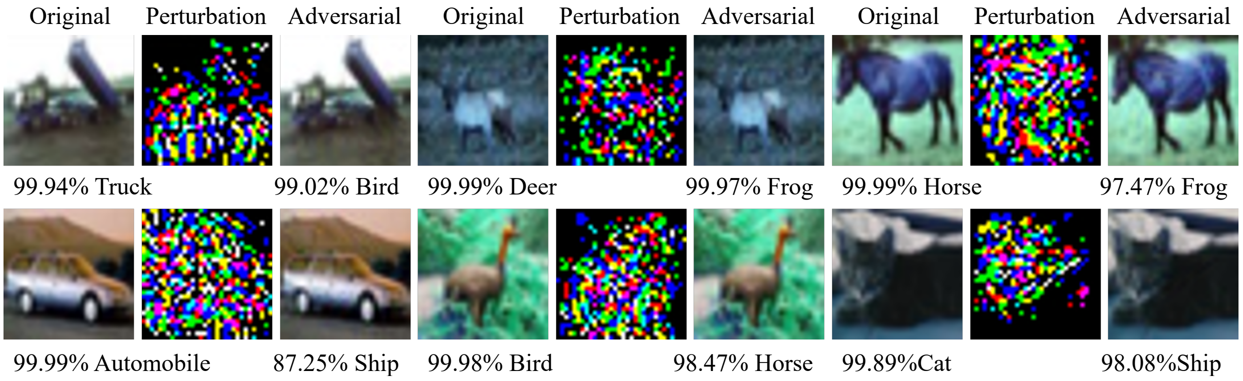

Figure 1: Qualitative results of untargeted attack.

4 Effect of Regularization

Targeting to confuse the classifier with the lowest cost, the regularization term is applied to prevent the perturbation from growing too large. It makes the attack more difficult, because it limits the scale of perturbation. The smaller the perturbation is, the easier the neural network can distinguish the image correctly, the harder for adversarial examples to attack the model.

5 Comparison of Untargeted and Targeted Attack

Targeted attack is more difficult, because it is required to mislead the neural networks to a specific class, rather than simply force a incorrect classification as required in untargeted attack. Besides, from the experimental results in Table 1 and Table 2, it can be observed that the L_1, L_2 and L_∞ norm in untargeted attack are lower than those in targeted attack, indicating that the targeted attack requires larger perturbations and is much harder.