Report on Generating Adversarial Examples

千英卓 2017013622

Introduction

In this experiment, we're expected to generate adversarial examples for VGG-16 model in CIFAR-10 image classification task. We're required to perform both targeted and untargeted attck with cross entropy loss and C&W attack loss, as well as regularizations.

Loss Function

The cross entropy loss and C&W attack loss are described below:
cross entropy loss:

$$\mathcal{J} = -crossentropy\_loss(y_{pred}, y_{true})$$

(untargeted)

$$\mathcal{J} = crossentropy\_loss(y_{pred}, y_{target})$$

(targeted)

C&W attack loss:

$$\mathcal{J} = \max\{[logit(\hat{x})]_{y_{true}} - \max_{y \neq y_{true}} [logit(\hat{x})]_y, -K\}$$

(untargeted)

$$\mathcal{J} = \max\{[logit(\hat{x})]_{y_{true}} - [logit(\hat{x})]_{y_{target}}, -K\}$$

(targeted)

Experiment Result

The results of experiment is shown in the table below.

Optimization method	α	β	γ	κ	Optimization epochs	Attack Success Rate	L_1	L_2	L_inf
cross entropy loss	0.001	0	0	\	500	0.989	9.85	0.233	0.00720
	0.001	0.01	0	\	1000	0.875	5.62	0.145	0.00388
	0.001	0	1	\	1000	0.849	5.48	0.141	0.00308
	0.001	0.01	1	\	1000	0.853	4.60	0.124	0.00293
C&W attack loss	0.001	0	0	0	500	1.0	10.43	0.247	0.00739
	0.001	0.01	0	0	500	1.0	10.05	0.239	0.00742
	0.001	0	1	0	500	1.0	10.17	0.242	0.00754
	0.001	0.01	1	0	500	1.0	9.95	0.237	0.00708

(untargeted)

Optimization method	α	β	γ	κ	Optimization epochs	Attack Success Rate	L_1	L_2	L_inf
cross entropy loss	0.001	0	0	\	4000	1.0	22.87	0.508	0.0200
	0.001	0.01	0	\	1000	0.990	22.15	0.495	0.0193
	0.001	0	1	\	1000	1.0	23.09	0.515	0.0203
	0.001	0.01	1	\	1000	1.0	21.67	0.486	0.0194
C&W attack loss	0.001	0	0	0	1000	0.989	26.37	0.577	0.0211
	0.001	0.01	0	0	1000	0.989	25.68	0.564	0.0209
	0.001	0	1	0	1000	0.989	25.41	0.559	0.0212
	0.001	0.01	1	0	1000	0.979	25.05	0.553	0.0208

(targeted)

Choose the most successful experiment setup you think for untargeted attack and targeted attack respectively. Explain why.

I choose the experiment setup with C&W attack loss, α=0.001, β=0.01, γ=1, κ=0 for untargeted attack, as it achieves the success rate of 1.0 with relatively small pertubation

scale. The setup chosen for targeted attack is cross entropy loss with $\alpha=0.001$, $\beta=0.01$ and $\gamma=1$, as it reaches the success rate of 1.0 with a clear margin over other setups in scale of distortion.

Discuss how regularization influences attack. Does it make attack harder? Why?

As can be seen from the loss function, regularization adds a deviation to the gradient of original cross entropy or C&W attack loss. Therefore, it does make attack harder, as is shown in the result above(untargeted cross entropy loss without regularization reaches much higher success rate in less epochs). Seeing from a practical perspective, regularization restricts the sclae of pertubation added, therefore making it harder to trick the discriminator.

Which kind of attack do you consider more difficult, untargeted attack or targeted attack? Why?

I think the targeted attack is more difficult, because the aim of it is one step further than the untargeted one: it should trick the discriminator to classify the image into a specific false label instead of any random false one. Any successful targeted attack is a successful untargeted one, but As is shown in the result, targeted attack requires more training epochs and larger scale of distortion.