Adversarial examples for neural nets

Qimin Chen
A53284263
qic003@ucsd.edu

1 Fast Gradient Sign Method

1.1 Error rate

The error rate of the neural net on the original data set is 0.024 and the error rate on the perturbed data set is 0.986.

1.2 Examples

Figure 1 shows the an example of an error due to perturbation, left figure is the original image with label 4 and right figure is the perturbed version with predicted label 9.

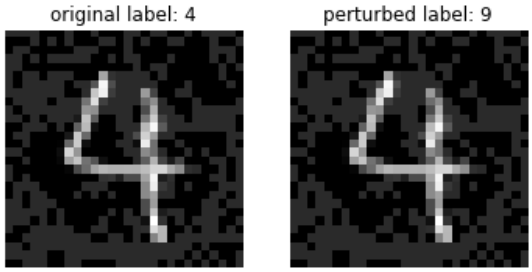

Figure 1: **Left:** original image with label 4. **Right:** perturbed version with label 9.

2 New method and pseudocode

2.1 Description

The gradient calculation remains the same while I make some changes to the adversarial examples generation. The initial perturbed $x^{(0)}$ is assigned as the same value of $x^{(0)}$ and the current perturbed $x^{(i)}$ is now depend on previous perturbed $x^{(i-1)}$ and the perturbation amount ϵ instead of depending on corresponding original image $x^{(i)}$.

2.2 Pseudocode

Below shows the pseudocode of perturbed x generation.

Algorithm 1 Modified Perturbed x Generation

Input: first test example $x^{(0)}$.
Output: perturbed x.

1: $x^{(0)}_{perturbed} = x^{(0)}$;
2: **for** i in range $(1, 1000)$ **do**
3: $x, y = x^{(i-1)}_{perturbed}, y^i_{test}$
4: forward x
5: $x^{(i)}_{perturbed} = x^{(i-1)}_{perturbed} + \epsilon * sign(\nabla L(f, x^{(i-1)}_{perturbed}, y))$
6: **end for**

3 Experimental results

The confidence intervals is computed by following:

$$(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}})$$

where $\bar{x}$ is accuracy mean, z^* is critical value which is 1.96 in 95% confidence intervals, σ is the standard deviation of accuracy and n is the number of experiments.

Table 1 shows the accuracy of the neural networks on adversarial examples generated by my algorithm vs. the fast gradient sign method over 10 times experiments. As it can be seen from table that FGSM can totally misclassify the data set with adversarial amount $\epsilon = 0.15$.

ϵ	my algorithm	FGSM
0.05	0.109 ± 0.009	0.002 ± 0.001
0.10	0.094 ± 0.012	0.014 ± 0.002
0.15	0.086 ± 0.017	0.000 ± 0.000
0.20	0.082 ± 0.014	0.000 ± 0.000

Table 1: Averaged accuracy on adversarial examples between my algorithm and FGSM over 10 times experiments.

4 Critical evaluation

My method does not perform better than the Fast Gradient Sign Method.

My method can be improved by carefully initiating the value of $x^{(0)}_{perturbed}$ instead of just assigning the original $x^{(0)}$ to it and I would like to try a more complicated network for example AlexNet.

model	training time / epoch	# images	quality
StarGAN	$\sim$ 37 mins	$202,599$	realistic / clear
AEcDCGAN	$\sim$ 26 mins	$275,190$	oil-painting / less clear

Table 2: lol.