REPORT FOR ARTIFICIAL NEURAL NETWORK ASSIGNMENT 4

Xiang Zhang
Department of Computer Science and Technology
Tsinghua University
xiang-zh17@mails.tsinghua.edu.cn

November 17, 2019

ABSTRACT

In this assignment, two types of loss functions, i.e. cross entropy and Carlini-Wagner loss are employed to generate adversarial examples against a custom VGG-like model on CIFAR-10 dataset, in both untargeted and targeted scenarios. The performance of these two loss functions along with different hyperparameters are evaluated, and hence a relatively optimal setting is selected respectively for targeted and untargeted attack.

1 Loss functions in targeted attack

1.1 Cross entropy

In untargeted attack, we intend to add noises that may alter the probability distribution of the classifier output, in a way that the most probable output deviates from the original value, which is achieved by maximize

$$\mathcal{J} = CrossEntropy(y_{pred}, y_{true}) \tag{1}$$

For targeted attack, we would like to modify the output distribution such that the target label has the highest probability. This can simply be achieved by regarding target label as the ground truth in regular training procedure, i.e. to minimize

$$\mathcal{J} = CrossEntropy(y_{pred}, y_{target}) \tag{2}$$

1.2 Carlini-Wagner loss

Carlini-Wagner loss comes with an adjustable confidence parameter κ. For untargeted attack, the loss is defined as

$$\mathcal{J} = \max\{[logit(\hat{x})]_{y_{true}} - \max_{y \neq y_{true}} [logit(\hat{x})]_y, -\kappa\} \tag{3}$$

This means we want the probability of original label lower than other labels, i.e. the network can guess anything other than the previous one. In terms of targeted attack, it is desired that the probability of other labels is below that of target label, hence the loss is adapted as

$$\mathcal{J} = \max\{\max_{y \neq y_{target}} [logit(\hat{x})]_y - [logit(\hat{x})]_{y_{target}}, -\kappa\} \tag{4}$$

From the equation we can conclude that the larger the parameter κ, the higher confidence is expected.

2 Experiments

All the experiments are only tuned with hyperparameter α, β, γ, whereas number of epochs remain default. The asterisk mark(*) indicates that the model is selected as the best one.

2.1 Untargeted attack

Table 1: Numerical results of untargeted attack

Optimization method	α	β	γ	κ	#Epoch	Success rate	L_1	L_2	L_∞
Cross Entropy	0.001	0	0	/	500	**1.0**	10.149667	0.240225	0.007369
	0.001	0.01	0	/	500	0.760417	0.014175	0.005254	0.000940
	0.001	0	1	/	500	0.694737	0.010898	0.004366	0.000826
	0.001	0.01	1	/	500	0.720430	**0.005102**	**0.002581**	**0.000590**
	0.01	0.001	0	/	500	0.755319	1.345098	0.052772	0.001544
	0.01	0.0005	0	/	500	0.778947	2.101259	0.069063	0.002023
	0.01	0.0003	0	/	500	0.780220	2.668563	0.092550	0.002241
	0.01	0.0001	0	/	500	0.793478*	3.794203	0.106492	0.002728
	0.01	0.00005	0	/	500	0.851064	5.238882	0.137117	0.004130
Carlini-Wagner	0.001	0	0	0	500	**1.0**	10.484968	0.247539	0.007669
	0.001	0.01	0	0	500	0.755319	0.102795	0.016459	0.001877
	0.001	0	1	0	500	0.741954	0.124436	0.017514	0.001687
	0.001	0.01	1	0	500	0.711340	**0.055711**	**0.011314**	**0.001577**
	0.001	0.01	0	0.1	500	0.774194	0.174025	0.018895	0.001687
	0.001	0.01	0	0.2	500	0.755319	0.095661	0.015669	0.001710
	0.001	0.01	0	0.3	500	0.706522	0.098252	0.015105	0.001577

2.2 Targeted attack

Table 2: Numerical results of targeted attack

Optimization method	α	β	γ	κ	#Epoch	Success rate	L_1	L_2	L_∞
Cross Entropy	0.001	0	0	/	500	**1.0**	23.784805	0.528013	0.020455
	0.001	0.01	0	/	500	0.211111	0.046100	0.006052	0.000697
	0.001	0	1	/	500	0.225806	0.065528	0.007462	0.000675
	0.001	0.01	1	/	500	0.215054	**0.024415**	**0.004225**	**0.006747**
Carlini-Wagner	0.001	0	0	0	500	**1.0**	23.829072	0.528949	0.020588
	0.001	0.01	0	0	500	0.197802	0.105969	0.007057	0.000690
	0.001	0	1	0	500	0.230769	0.055555	0.006809	0.000690
	0.001	0.01	1	0	500	0.247423	**0.042450**	**0.005819**	**0.000849**
	0.001	0.01	1	0.05	500	0.25	0.338960	0.017830	0.001023
	0.001	0.01	1	0.01	500	0.255319	0.359741	0.019104	0.001376
	0.005	0.01	1	0.01	500	0.510638*	3.781185	0.108126	0.007301
	0.01	0.01	1	0.01	500	0.651685	7.435118	0.194737	0.011633

The evaluation of adversarial attack examples is not only based on success rate, but the amount of distortion, i.e. noises added to original input should also be included. Ideally we would like higher success rate along with lower distortion, yet these two metrics are contradictory, in which case trade-offs are inevitable and a proper equilibrium point between them should be determined.

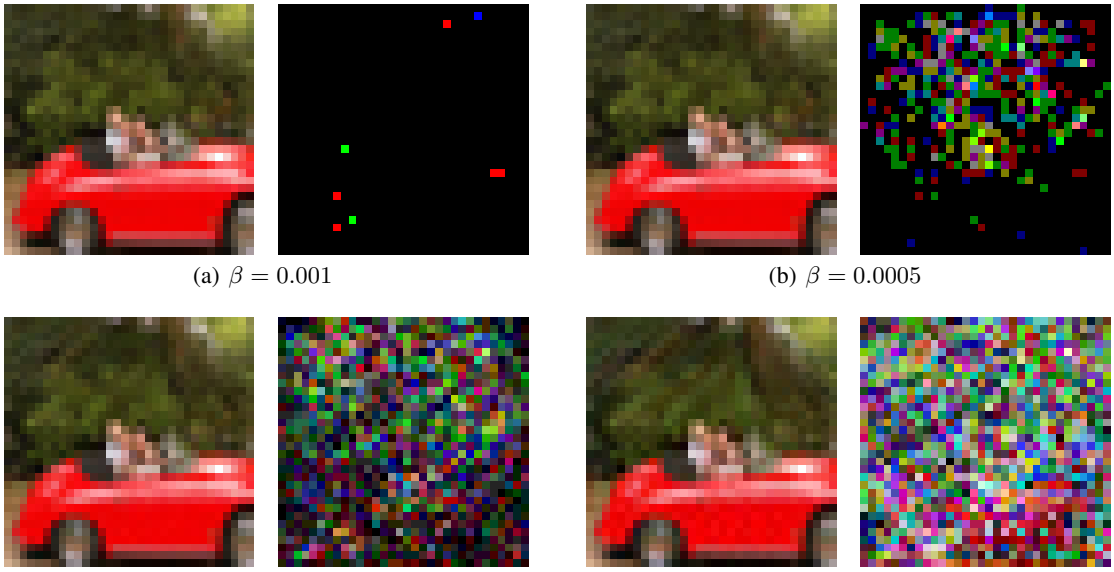
From **Table 1, 2**, conspicuously without regularization, the attack model can obtain a success rate of 100%, yet huge amounts of noises are introduced (with L_1 norm higher than 10). With the default regularization settings (the first four rows in each category), the noises considerably diminish, at the cost of a lower success rate. It is common for both kinds of attacks that elastic net approach ($\beta = 0.01$, $\gamma = 1$) has the best capability to suppress the noise.

Since the original settings are relatively conservative, more experiments are conducted to explore the effects of different regularization level, which derive from the best case present in targeted or untargeted attack. From untargeted attack, L_1 regularization with cross entropy loss performs best in default settings, thus another 5 experiments with distinctive weight of L_1 term are carried out. For targeted attack, elastic net approach with Carlini-Wagner loss (with pre-tuned κ) obtains the best result, thus another three experiments are designed.

2.3 The effects of κ

κ is a parameter in Carlini-Wagner loss denoting the desired confidence. From **Table 1, 2**, properly increasing κ can bolster the performance, yet excessively large κ can exert a contrary effect. From (3)(4), κ controls the probability distance between other labels and target label, which can be represented visually by the height of peak in the distribution of $logit(\hat{x})$. Higher κ can help to opt out samples with better confidence, whereas superfluous value can impair the training process, especially for targeted attack, where generating a distribution with higher peak of target label is much more difficult. In this case, $\kappa = 0.1$ and 0.01 performs well for untargeted and targeted attack respectively.

2.4 Regularization



(a) $\beta = 0.001$ (b) $\beta = 0.0005$

(c) $\beta = 0.0001$ (d) $\beta = 0.00005$

Figure 1: Visualization of image distortion with different regularization settings ($\alpha = 0.01$, $\gamma = 0$)

In this experiment, the regularization term in loss function penalizes the noises to be added to input image, making it harder to conjure up an adversarial example. From the results, when regularization is disabled by setting β and γ to zero, both targeted and untargeted attack can achieve a 100% success rate regardless of optimization method (cross entropy or Carlini-Wagner) utilized, whereas the noise suffers from a high L_k norm, which renders the attack images visually discernable from normal ones.

Several untargeted attack examples are selected in **Fig. 1**. Note that the distortion is a rather subjective term, and is mainly connected with the basic properties of original images, such as the complexity of colors/textures, and the observer as well. Images that has monotonic colors may accentuate the noise better. In this case, outputs with L_1 norm lower than 5 ($\beta \leq 0.0001$ in the figure) appear to be visually satisfactory, i.e. the noise is not readily perceivable. We mark $L_1 \leq 5$ as acceptable and this criterion will be adopted to select best models for the task.

2.5 Untargeted versus targeted attack

In comparison, targeted attack is more difficult than untargeted attack. From a probabilistic perspective, untargeted attack offers various options, i.e. any distribution of output yielding a prediction label other than ground truth is considered proper, while targeted attack stipulates that the final output must be the specified target, limiting the number of valid distributions, thus rendering it more intractable.

2.6 Best models

For untargeted attack, $\alpha = 0.01$, $\beta = 0.0001$, $\gamma = 0$ with cross entropy loss is selected as the best model, which yields a success rate of 79.3478%. Note $\beta = 0.00005$ fails here because its L_1 norm exceeds 5, and in **Fig. 1(d)**

observable artifacts are present in the attack image. For targeted attack, $\alpha = 0.005$, $\beta = 0.01$, $\gamma = 1$, $\kappa = 0.01$ with Carlini-Wagner loss is viewed as the best, which has 51.0638% success rate. These models reach an equilibrium point where success rate is as high as possible with controlled distortion rate.

3 Conclusion

In this assignment, the performance of untargeted and targeted adversarial attacks is evaluated with respect to different loss functions (cross entropy and Carlini-Wagner) and levels of regularization. It is found that the success rate of attack and amount of distortion are two contradictory metrics, rendering it difficult to optimize them simultaneously, which necessitates a trade-off between numerical performance and visual effects. Controlling the weight of regularization in the loss function is a direct approach to find this balance. Plus, targeted attack is more difficult than untargeted one, in which case, it requires more distortion to obtain an equivalent success rate to its counterpart. Finally, two models with relatively high success rate and controlled distortion are selected as the best models respectively for targeted and untargeted task.