Team Teaflow Non-targeted / Targeted Attack Summary

Caspar Zhang
casparz@gmail.com

October 6, 2018

In CAAD 2018 competition，our submissions on Non-targeted & Targeted Attack task are both based on Basic Iterative Method(BIM).

1. Gradient is smoothed spatially using Gaussian filter:

$$\boxed{\begin{aligned}&X_0^{adv} = X\\&X_{N+1}^{adv} = Clip_{X,\epsilon}(X_N^{adv} + Y_N)\\&Y_N = \alpha sign(\nabla_X J(X_N^{adv}, Y_{true}))\end{aligned}}$$

$$\Downarrow$$

$$\boxed{\begin{aligned}&X_0^{adv} = X\\&X_{N+1}^{adv} = Clip_{X,\epsilon}(X_N^{adv} + Y_N)\\&Y_N = \alpha sign(h_{smooth} * \nabla_X J(X_N^{adv}, Y_{true}))\end{aligned}}$$

The parameters of Gaussian filter are set separately for the two tasks：

1) non-targeted attack
$size(kernel) = 7$
$\sigma = 2$

2) targeted attack
$size(kernel) = 6$
$\sigma = 4$

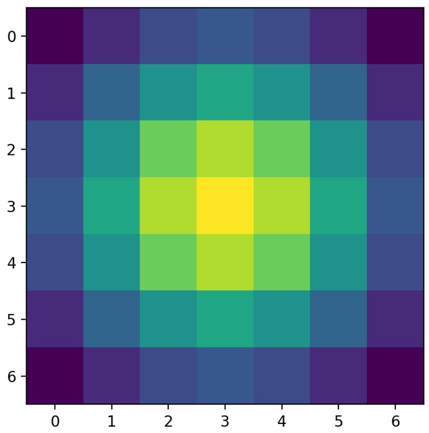

Figure 1 non-targeted attack

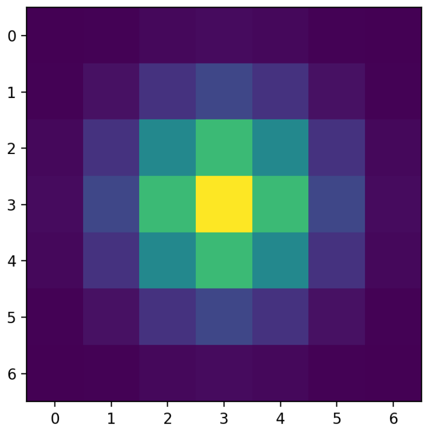

Figure 1 targeted attack

2. After some comparison tests, we decided to use the following models for ensembling：
    1) non-targeted attack
        i. adv_inception_v3_2017_08_18
        ii. ens4_adv_inception_v3_2017_08_18
        iii. adv_inception_resnet_v2_2017_12_18
    2) targeted attack
        i. ens_adv_inception_resnet_v2_2017_08_18
        ii. adv_inception_v3_2017_08_18
        iii. inception_v3_2016_08_28

3. Engineering direction improvement：
    1) Using TensorRT to speedup TensorFlow inference, try to perform more iterations under the time constraint
    2) Compiling Tensorflow with SSE, AVX & AVX2 instructions support

4. Benchmark
   For benchmarking, we use all images in the DEV dataset. For evaluation, we used the top-2 NIPS 2017 defense solutions and also some baselines (fgsm & adv_inception_resnet).

	TsAIL	iyswim	adv_inception_resnet
fgsm	16.8%	39.0%	39.5%
teaflow	**93.7%**	**64.3%**	**71.1%**

Table 1 Non-targeted Attack Success Rate

	TsAIL	iyswim	adv_inception_resnet
BIM	1.2%	2.4%	2.8%
teaflow	**46.7%**	**16.4%**	**19.6%**

Table 2 Targeted Attack Success Rate