ANN Homework 4 GAN

冯卓尔

2017011998

Give the cross entropy loss and C&W attack loss adapted to targeted attack. Explain them briefly.

cross entropy loss

$$J = cross\_entropy\_loss(y_{pred}, y_{target})$$

在untargeted attack中，任务是将样本攻击至分类为非y_true, 而在targeted attack中，任务变为将样本共计为y_target，那么在复用这个公式的时候，两者都要减小J，而对于untargeted attack的J减小，对应于targeted attack中的J增大，因此需要将负号去掉，这样两者的回归方向都是J减小方向。

C&W attack loss

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

同理，按照上述的原理，将所有的y_true换成y_target的同时，需要给这两项存在的位置添加负号，结果如上。

Experiment results

For untargeted attack

optimization method	α	β	γ	# optimization steps	Attack success steps	L_1	L_2	L_∞
cross entropy loss	1	0	0	500	1.00	19.3095	0.4035	0.0083
	1	0.00001	0	500	0.98	17.8523	0.3764	0.0088
	1	0	0.001	500	0.95	17.6136	0.3703	0.0080
	1	0.00001	0.001	500	0.93	16.8573	0.3554	0.0084
C&W attack loss	1	0	0	500	1.00	21.8275	0.4458	0.0086
	1	0.01	0	500	1.00	10.1301	0.3163	0.0086
	1	0	1	500	1.00	10.1450	0.2836	0.0086
	1	0.01	1	500	1.00	8.6111	0.2807	0.0087

For targeted attack

optimization method	α	β	γ	# optimization steps	Attack success steps	L_1	L_2	L_∞
cross entropy loss	1	0	0	4000	1.00	32.0635	0.6567	0.0162
	1	0.00001	0	4000	1.00	28.4081	0.6462	0.0162
	1	0	0.001	4000	1.00	24.5182	0.5602	0.0162
	1	0.00001	0.001	4000	1.00	21.1178	0.5397	0.0163
C&W attack loss	1	0	0	4000	1.00	31.0623	0.6598	0.0162
	1	0.01	0	4000	1.00	10.7392	0.4467	0.0161
	1	0	1	4000	1.00	12.9292	0.3987	0.0161
	1	0.01	1	4000	1.00	10.8013	0.4039	0.0163

Discuss what make a good adversarial example. Choose the most successful experiment setup you think for untargeted attack and targeted attack respectively. Explain why.

通过实验可以看出，untargeted attack与targeted attack的攻击成功率都几乎为1.00，然而我们可以发现对于targeted attack，在所有的参数设定条件下其attack success几乎都是1，相比于untargeted attack期望成功率更高。但是由于targeted attack的epoch达到了4000，因此这个结论或许并不绝对。因而我观察了一下targeted attack的第一个实验，发现在nr_epoch=800左右，模型的loss已经前四位小数为0，因此实际上可以认为targeted attack的效率较高。

从L_1 L_2 L_∞的意义中可以得知，我们知道他们噪声的强度，它们反映了模型的一致程度。因此观察untargeted attack与targeted attack，我们发现使用C&W attack loss能够获得最小的L参数，也就是较小的噪声，因此我们就获得了一下的几个结论：

对于untargeted attack， use C&W attack loss, alpha = 1, beta = 0.01, gamma = 1.00，此时对抗生成样本最优；对于targeted attack，use C&W attack loss, alpha = 1, beta = 0.01, gamma = 1.00，此时对抗生成样本最优。

其性能较好的原因可能如下：

1. 使用了C&W attack loss方法，这个方法能够提升攻击成功率
2. 使用了正则化的方法，对原图的攻击修改做出了限制，因而产生极端case的概率较小。

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

正则化影响攻击。具体来说，随着beta和gamma的增大，从untargeted attack的数据中可以看出，随着beta与gamma的正则化参数的加入（非零化），攻击的成功率变小了。

其原因为正则化是为了防止过拟合等条件对攻击的幅度进行约束的过程，当beta和gamma参数越大的时候，攻击的幅度变小了，因而生成的样本有可能还是被分类至原本的类别（针对untargeted attack）或者是被分类到其他的类别中去（targeted attack)。从而，表现为攻击成功率的下降。这一部分的佐证在untargeted attack实验1~4.

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

我觉得targeted attack是最难的，原因分为两个部分。

第一，从实验结果上来看，targeted attack需要一倍甚至五倍与untargeted attack的迭代步数来完成迭代，从计算量上来说，targeted attack比untargeted attack复杂。

第二，从理论构造上来看，targeted attack与untargeted attack的任务不同：untargeted attack只需要讲一个样本攻击至被分类到其他的类别中去即可，其成功地类型有N-1中（N为总共的分类数量）；而targeted attack需要将被攻击样本修改至其被分类至指定的类别中去，其成功率为untargeted的$\frac{1}{N-1}$。

附件中，我保存了targeted attack使用cross entropy loss的第一个实验，实验中，$\alpha = 1, \beta = 0.01, \gamma = 1$，我们可以看到在训练的xx代时，loss已经非常稳定了，我们可以认为在这时候几乎可以停止对抗生成了。

但是，缺乏数学工具与理论推导，这个结论仅仅只能从结果中窥测，所以我能够给出看法，但是无法下定论。