**Improving the Robustness of Deep Neural Networks**

**via Adversarial Training with Triplet Loss**

**一、Introduction**

本文引入Triplet loss到对抗训练中。

**1.Triplet loss是什么**

Triplet loss是神经网络的损失函数，其中一组输入有三类，baseline(anchor)、positive (truthy)和negative (falsy)。最小化anchor到positive的距离，最大化anchor到negative的距离。从而使得具有相同标签的数据点之间的距离比具有不同标签的数据点之间的距离更近。

以人脸识别为例**，**Triplet loss的目的是学习一个好的embedding，使得同一个人的人脸在嵌入空间中尽量接近，不同人的人脸在嵌入空间中尽量远离。

Triplet loss的主要挑战在于如何选取有代表性的三元组。

**2.引入Triplet loss有什么作用**

Triplet loss作为最流行的距离度量学习方式之一，可以**平滑分类边界**，达到增强鲁棒性的目的。

**Contributions**

1. ·引入Triplet loss到对抗训练中（）。并设计了集成版本。
2. ·用改进的Triplet loss作为正则项，应用到现有防御方法中，提高鲁棒性。
3. ·大量实验表明，本文方式在不损失精度的前提下，提高了鲁棒性；且改进的triplet loss也能提高其他防御方法的鲁棒性。

**二、Related work**

**1. Attack methods**

基于梯度的攻击：要求已知目标模型的结构，且目标模型是可微的。

基于优化的攻击：最小化扰动的范数，使DNN对对抗样本分类错误。

**2. Defense methods**

模糊梯度：是gradient masking的一种，试图阻止敌手访问有效的梯度。但18年的一篇文章证明了这种防御是失败的。

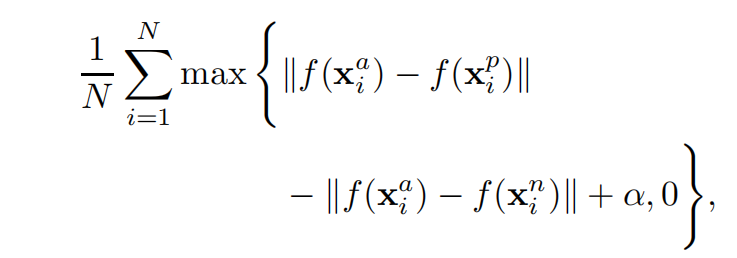
对抗训练：用对抗样本训练模型，直至能正确分类。但存在分类边界失真的问题，因此本文引入了triplet loss缓解失真。

**三、Methodology**

**1. Triplet loss**

一个triplet三元组包含了三个样本，分别来自两个不同类，triplet用clipboard.png表示。其中是anchor，clipboard.png有相同的label，而clipboard.png的label不同。

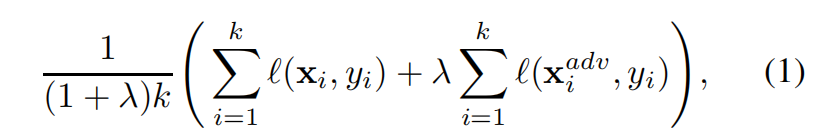
有相同label的那对之间的距离应当缩小，不同label的那对之间距离应当增大，前者距离要小于后者。因此triplet loss可以表示为：



[其中，N是triplets的数目，f(·)是神经网络的输出，||a-b||表示a、b之间的距离。]

**2. Adversarial training with triplet loss (****)**

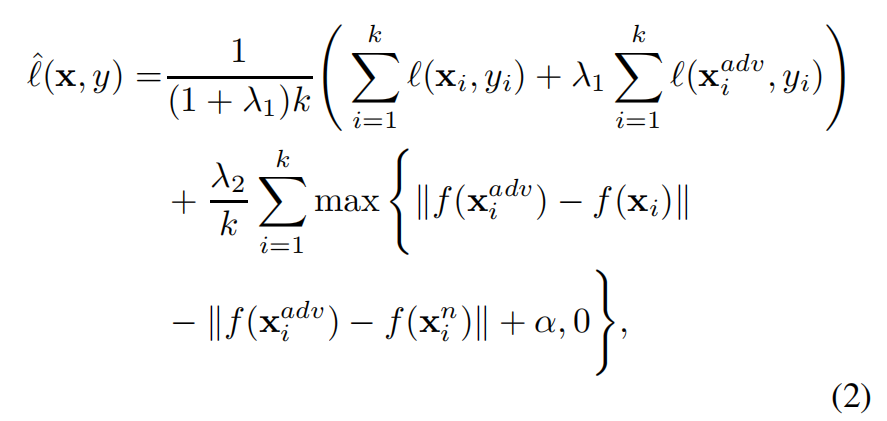
原始对抗训练的损失函数可表示为：



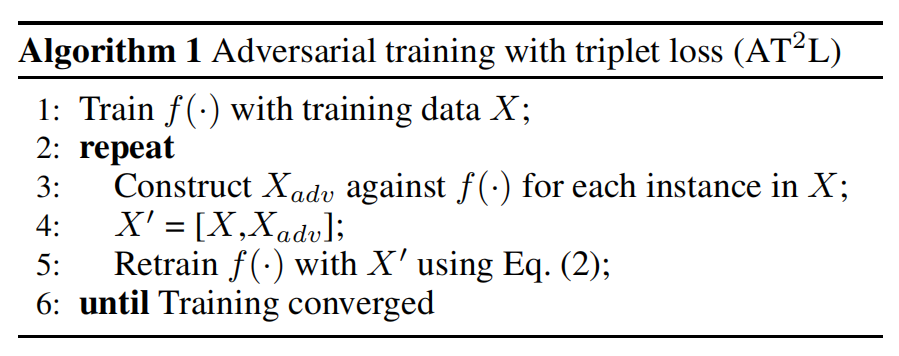
[其中，λ是超参数，k是mini-batch的步长，y是xi的标签。]

对1中的triplet loss进行改进，将生成的对抗样本xadv作为anchor：clipboard.png

将这个改进后的triplet loss加入到原始对抗训练的损失函数(1)中即可：



算法：

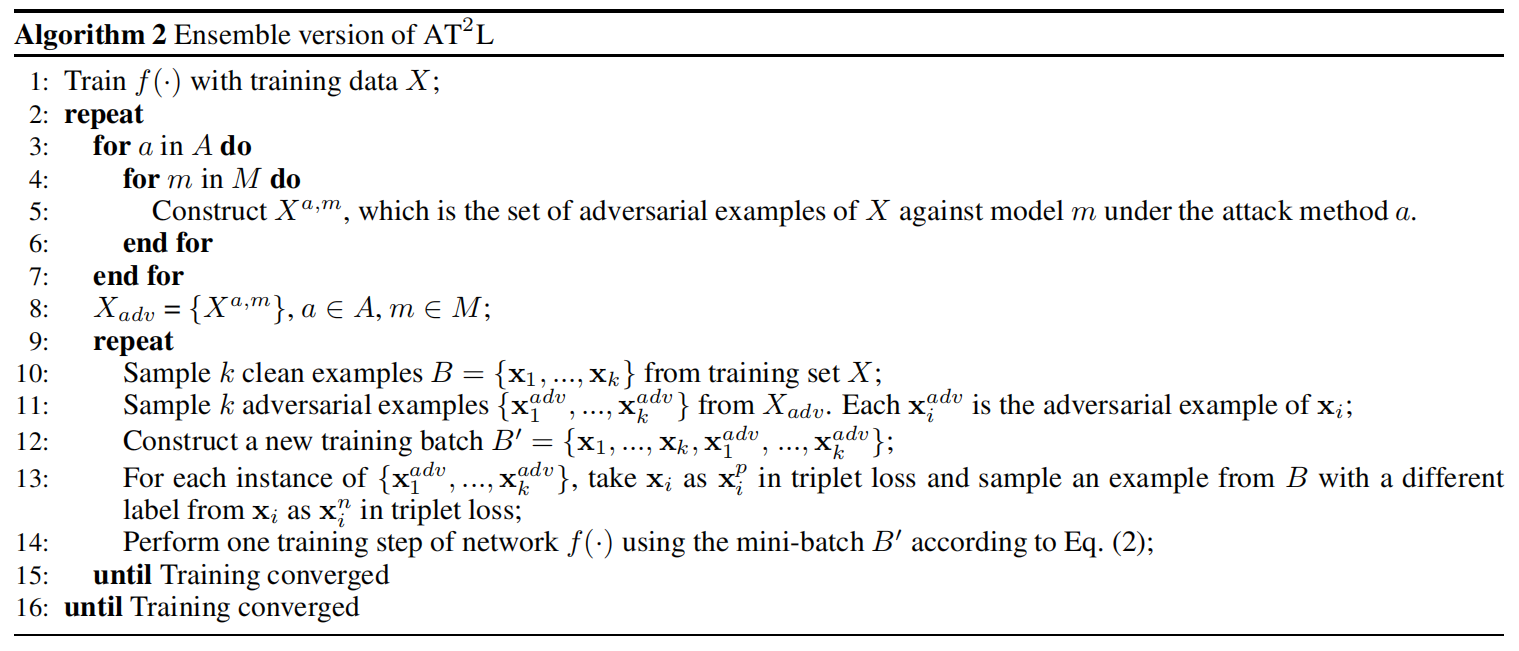


**3. Ensemble** 

为了将扩展到unknown attacks上，提出了集成版。

用A表示各种攻击方式的集合，在所有这些攻击上生成对抗样本。集成版不仅考虑了各种攻击，还考虑了不同模型结构（用M表示）生成的对抗样本，进一步提高了鲁棒性。

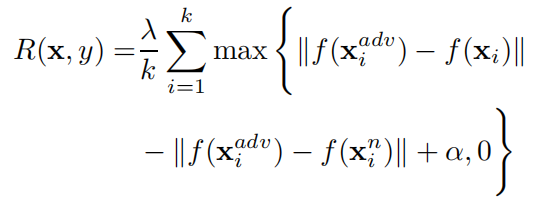
算法：



**4. Triplet regularization**

本文改进的triplet loss可以作为正则项，加到现有防御方式的loss上，如模糊梯度。

Triplet regularization表示为：

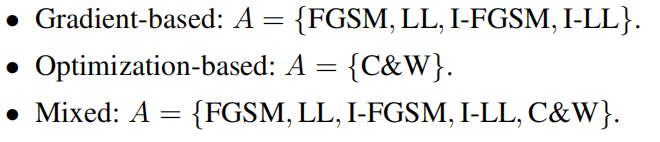


**四、Experiments**

**1. 实验设置**

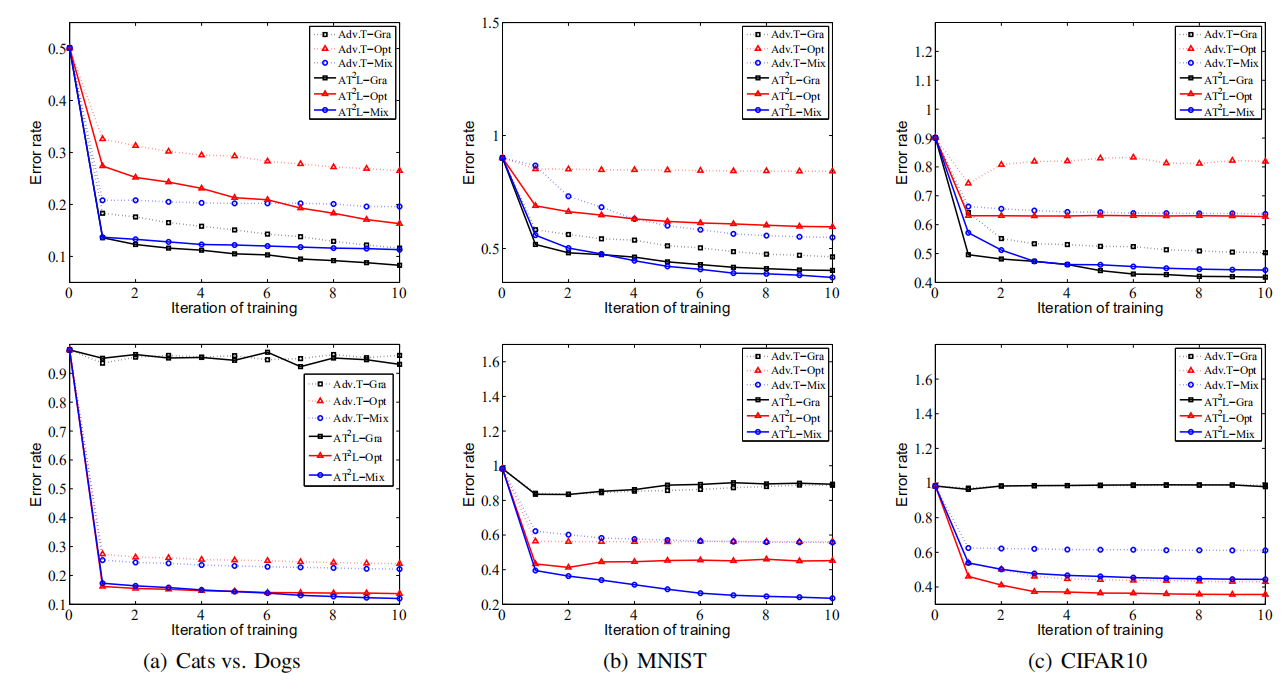
数据集：选择了适用于二分类的Cats vs.Dogs和多分类的MNIST、CIFAR10。

攻击方式：（下图+LS-PGA+Deepfool）



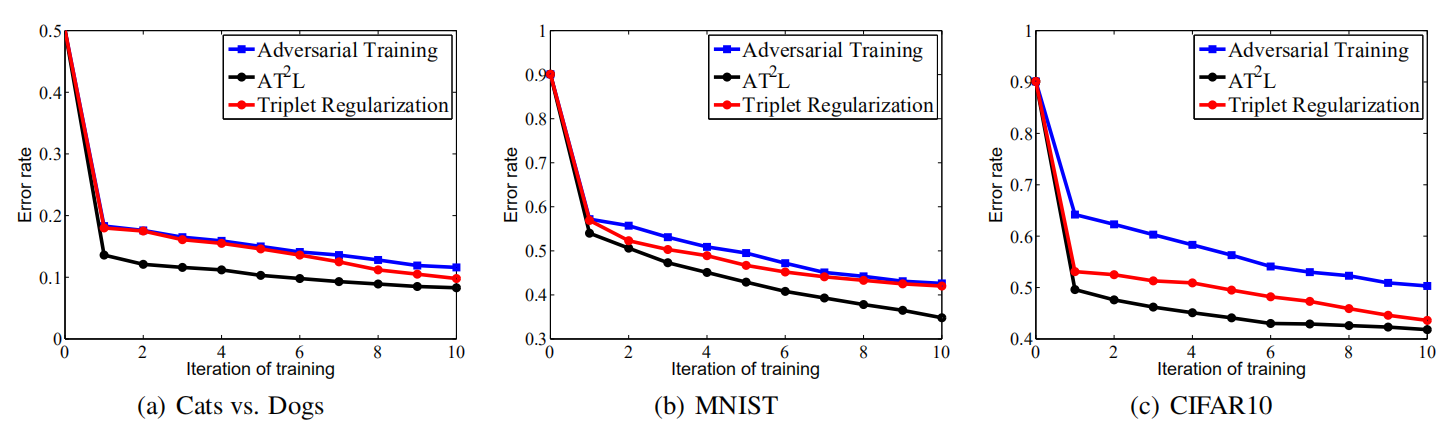
**2. 实验结果：**

第一、二行的攻击方式分别采用了FGSM，C&W（FGSM是gradient-based attack，C&W是optimization-based attack。）。仅展示白盒攻击：



结论：当受到这两种攻击时，使用相同攻击生成的对抗样本训练的鲁棒性最好。

**3. 实验结果：Triplet regularization**



观察：（1）随着轮数的增加，triplet regularization导致的错误率越来越低。

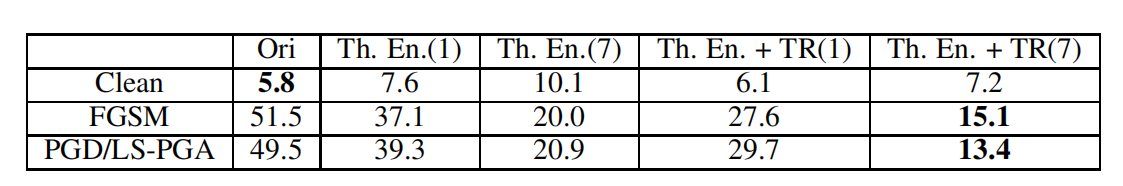
（2）AT2L错误率最低。

结论：（1）Triplet regularization可以提高鲁棒性，其效果不会比原始对抗训练更差。

（2）AT2L=Adversarial training+ Triplet regularization，效果最好。

**4. 实验结果：Triplet regularization+现有防御方式**

以thermometer encoding防御为例，五列分别代表不加防御、使用thermometer encoding训练一轮、使用thermometer encoding训练七轮、thermometer encoding+triplet regularization训练一轮、thermometer encoding+triplet regularization训练七轮：



结论：thermometer encoding加上triplet regularization后，显著降低了错误率。