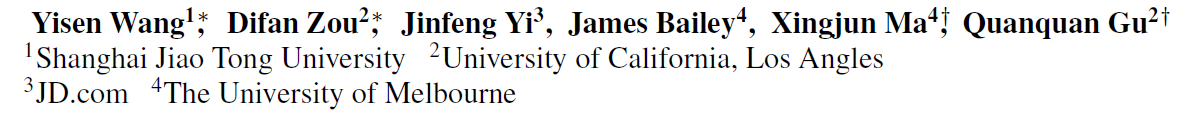
IMPROVING ADVERSARIAL ROBUSTNESS REQUIRES

REVISITING MISCLASSIFIED EXAMPLES



本文作者Quanquan Gu (UCLA) 每年发表多篇AI顶会，在adversarial learning和secure multi-party ML有很多厉害成果，个人主页<http://web.cs.ucla.edu/~qgu/index.html>。

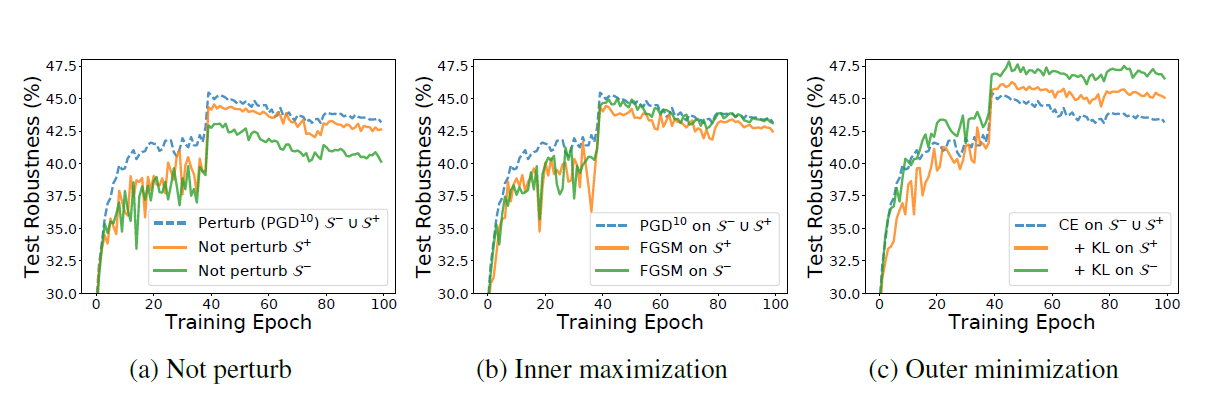
**0 Abstract**

本文认为对抗训练（即min-max优化）忽略了一个重要问题：对抗训练首先要计算寻找对抗样本（即max过程），然而对抗样本是通过扰动正确分类的样本所获得的；在训练过程中，很多样本不能被正确分类，然而现存的对抗训练方法，利用同样的方法寻找误分类样本的对抗样本。本文通过实验发现，正确分类和误分类样本对最终鲁棒性具有不同的影响。详细来说，误分类样本对鲁棒性具有更强的影响；对于误分类样本，不同的max方法，对最终鲁棒性影响很小，但是min方法有很强的影响。因此，作者提出新的对抗训练方法，MART，该方法区别对待正确分类样本和误分类样本，极大提高了现存方法的鲁棒性。

**1 Intuition**

虽然对抗训练是抵抗对抗样本最有效的防御方法，但是现存的方法仍然存在一个natural accuracy和robust accuracy 之间的Gap。举例来说，在CIFAR-10上通过对抗训练获得的模型，natural accuracy可以达到90%，但是robust accuracy只能达到50%左右。因此很多文章在对抗训练的基础模型上，研究提高robust accuracy方法，本文也属于此范畴。还有很多工作包括：增加model capacity来提高robustness；增加sample complexity来提高robustness，比如利用unlabelled data；利用pretraining来提高robustness；等等。

本文为了说明，误分类样本和正确分类样本具有不同的影响，首先利用标准对抗训练得到了natural accuracy为87%的CIFAR-10模型。然后选择全部13%的误分类样本和随机选择13%的正确分类样本，重新进行训练。发现三个现象：



现象a：同时利用正确和误分类样本的对抗样本进行训练，模型robust精确度最高。但如果误分类样本的训练不用对抗样本，robust精确度大大降低。

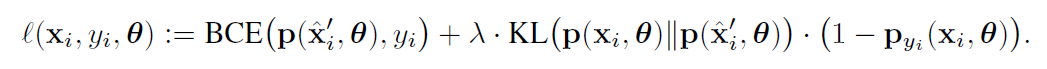
现象b：不同的max方法，如FGSM和PGD，对模型最终鲁棒性影响不大。但是，相比于正确分类样本，误分类样本在利用较弱的FGSM攻击方法时鲁棒性下降较高。

现象c：不同的min方法，即在CE（cross entropy）损失下再增加KL散度的正则项，对模型最终鲁棒性影响很大。尤其是对误分类样本增加KL正则项，鲁棒性大幅度提升。

因此，本文区别对待正确和误分类样本，构造特定的损失函数进行优化，来提高鲁棒性。

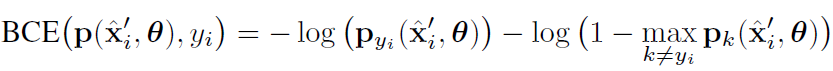
**3 Method**

根据上面的发现，作者在现有的对抗训练方法上，在损失函数中增加了正则项，以区别对待正确和误分类的样本。对于每个样本，新的损失函数如下所示：

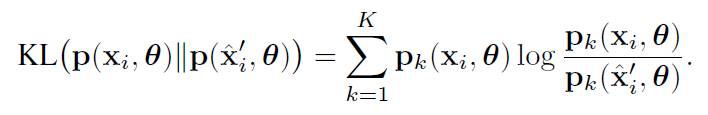


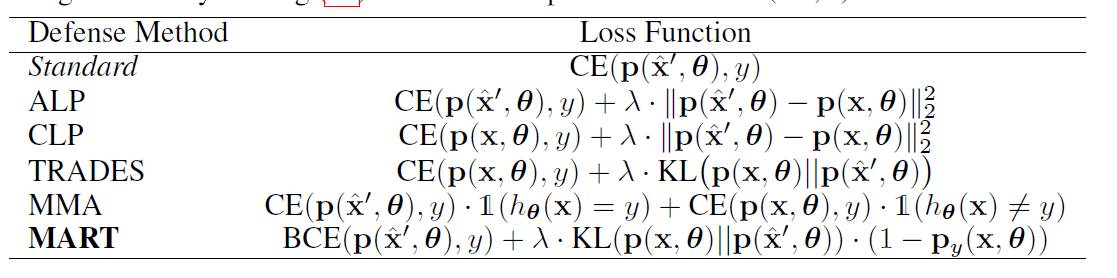
其中， 表示利用PGD生成的对抗样本， 表示真实的标签。该损失函数包括：

1）第一项为boosted cross entropy，即使得对抗样本预测值接近真实标签，同时使得最大的错误预测概率越小越好：



2）第二项为KL散度和对误分类样本的soft decision的乘积。soft decision即对于误分类样本，1*-p*接近为1，使其考虑KL散度；对于正确分类样本，1*-p*接近为0，使其忽略KL散度。KL散度公式如下：

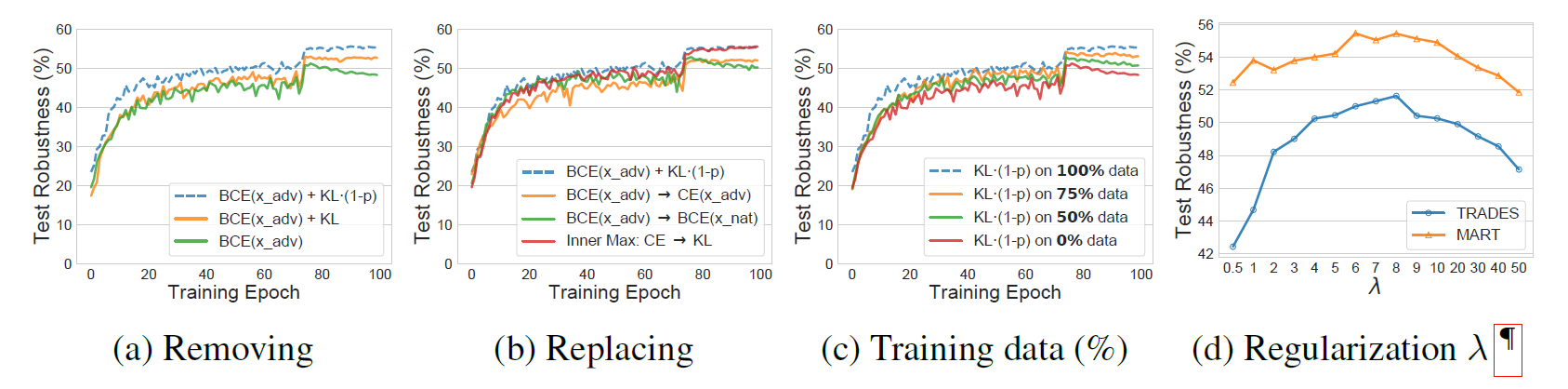


**与现有方法比较：**  


如图可以发现，现有基于对抗训练的不同防御方法，在损失函数上仅有细微的差别。Standard（ICLR18）和TRADES（NIPS19）是最经典的两篇对抗训练文章，与现有方法相比，本文第一个研究误分类样本对鲁棒性的影响，实验结果（如下）有很大提升。

**4 Experimental**

**MART不同参数设置对实验结果的影响：**



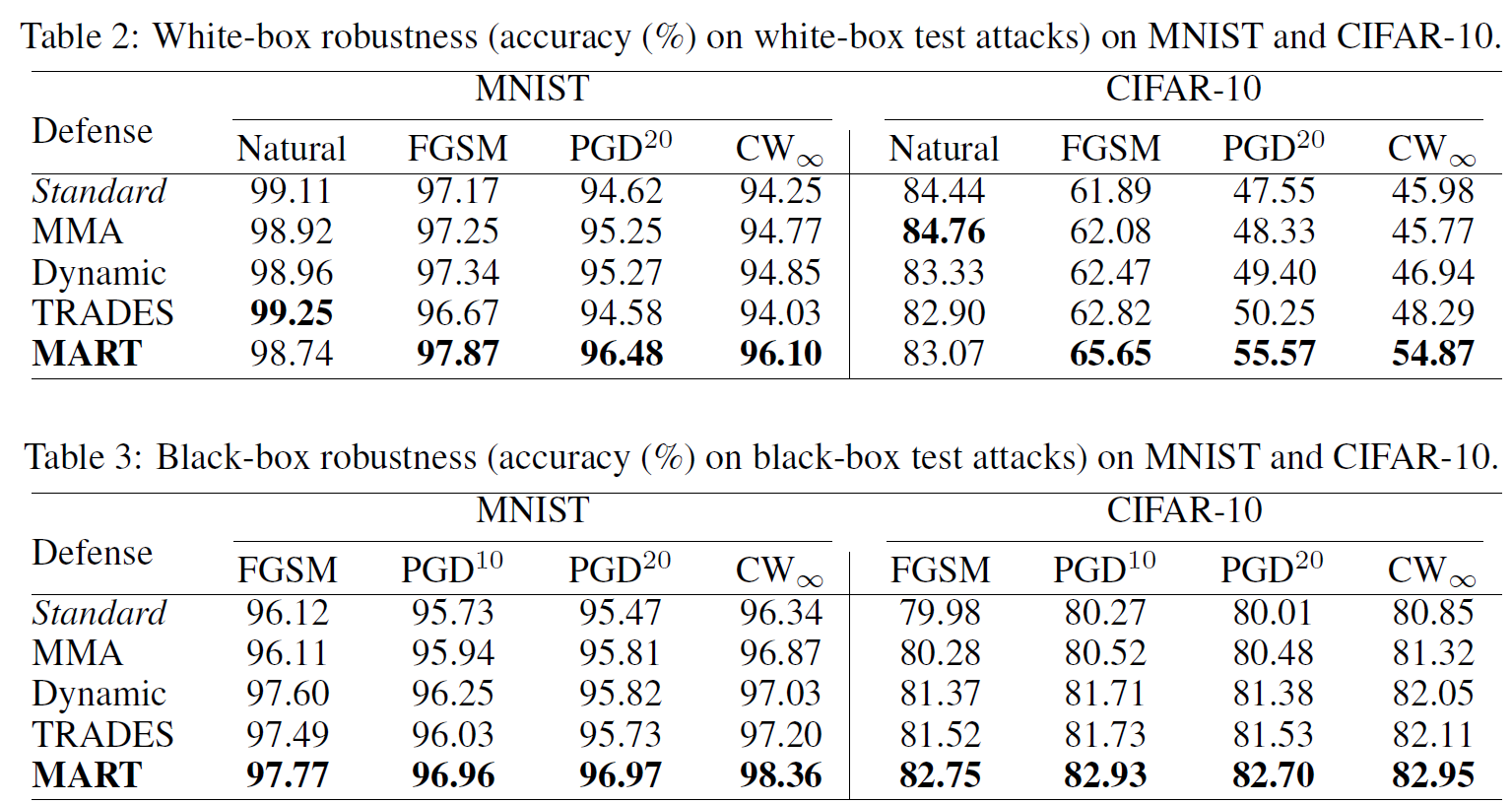
实验结果a：考虑损失函数中正则项KL（1-p）的作用。KL正则项能够让模型避免过拟合，提高鲁棒精确度；1-p能够从训练开始就提高鲁棒精确度。

实验结果b：考虑损失函数中BCE（x\_adv）的作用。将BCE换位传统的CE或是将对抗样本换为natural 样本都会极大地降低鲁棒精确度。Max方法中利用CE或为KL，对鲁棒精确度几乎无影响。

实验结果c：考虑只在部分训练数据上增加KL（1-p）正则项，所应用的数据量越小，鲁棒精确度越低。

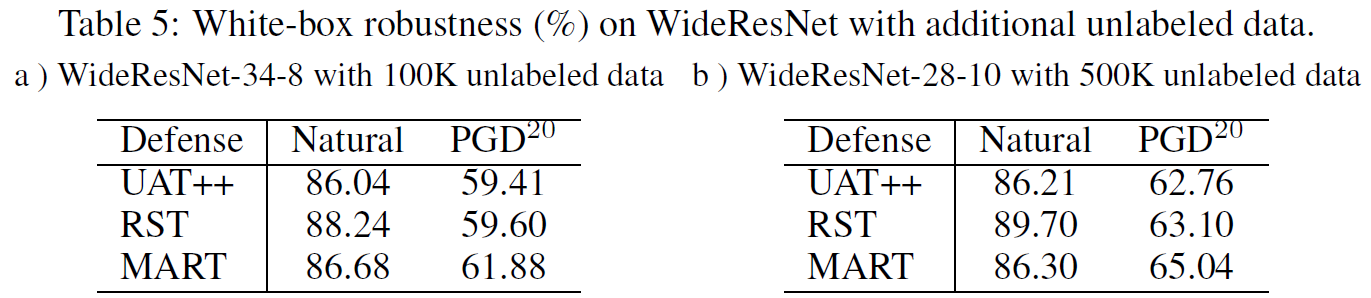
实验结果d：正则项参数对鲁棒精确度的影响，MART更稳定。

**与现有对抗训练鲁棒精确度比较：**



实验显示，本文方法在MNIST和CIFAR-10数据集上，鲁棒精确度高于现有方案。CIFAR-10数据有更加明显的提高，原因是CIFAR-10数据包含更多误分类的训练样本。

**与unlabelled data提高鲁棒性方法相结合：**



NIPS19年有四篇文章分别发现，unlabelled data对鲁棒精确度的作用。基本方法为，首先利用labeled data训练natural模型，再对unlabelled data进行标注，再用所有数据进行对抗训练。本文与现有两种unlabelled data增强鲁棒性方法结合，显著提高了对抗鲁棒性。

**5 Conclusion**

本文提出分别考虑正确和误分类样本对最终模型鲁棒性的影响，进而设计了新的对抗训练方法，实现了更高的模型鲁棒性。

个人观点：提高对抗训练的鲁棒性是当下非常火热的研究方向，然后这些分别提高了对抗训练的鲁棒性的parallel研究，这些方法叠加能否实现最佳的效果，并没有依据。另一方面，提高对抗训练的效率又是另一个研究热点，能否有方案既能提高效率又能提高鲁棒性。