Using Pre-Training Can Improve Model Robustness and Uncertainty

文章链接：[https://arxiv.org/pdf/1901.09960.pdf](https://arxiv.org/pdf/1912.05699.pdf)

code：[https://github.com/hendrycks/pre-training](https://github.com/hendrycks/pre-training )

**1 Abstract**

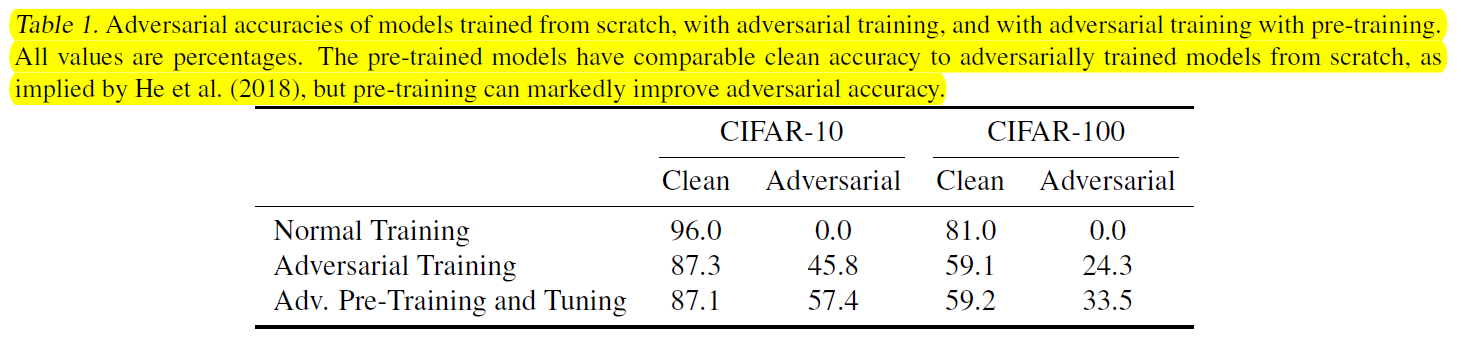
Kaiming He大佬提出pre-training不能提高传统模型的performance，但是本文认为即便如此，其可以提高模型的robustness（包括：adversarial examples、label corruption、class imbalance）和uncertainty（包括：out-of-distribution detection、confidence calibration）。下文分别阐述：

（部分内容没有理解，就没有写下来，因为每一部分都是单独的研究方向，标黄部分是最重要部分。）

**2 Introduction**

2.1 Robustness to Adversarial Perturbations

这部分利用ImageNet数据进行adversarial pre-training，再transfer到CIFAR-10和-100数据，利用少量数据进行adversarial fine-tuning整个模型。相比于在CIFAR-10和-100数据上进行adversarial training from scratch，实验表明，adversarial pre-training可以显著地提高robustness。如下表：

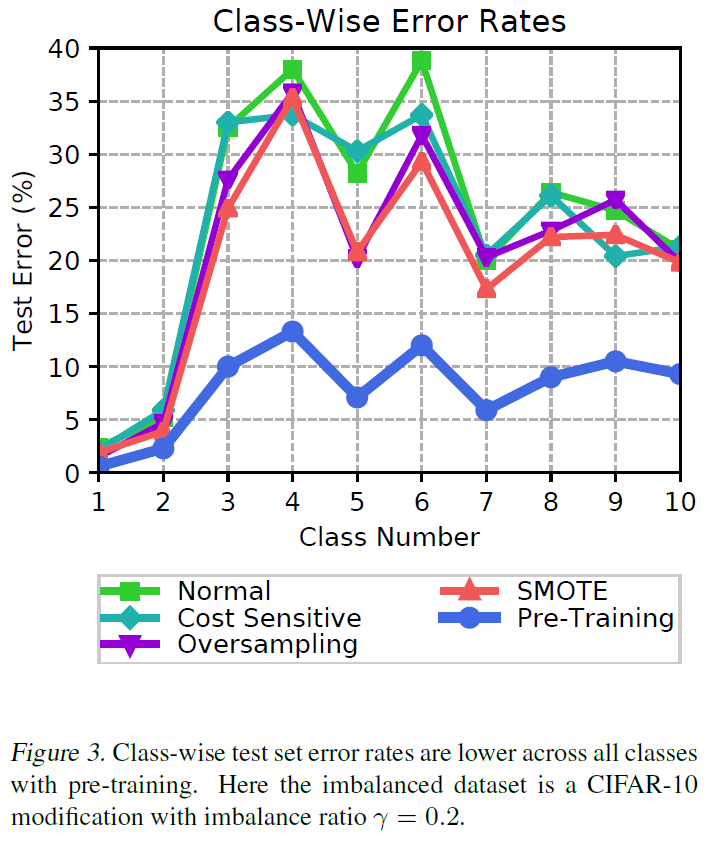


2.2 Robustness to Label Corruption

是指模型训练时，训练数据标签是错误的。

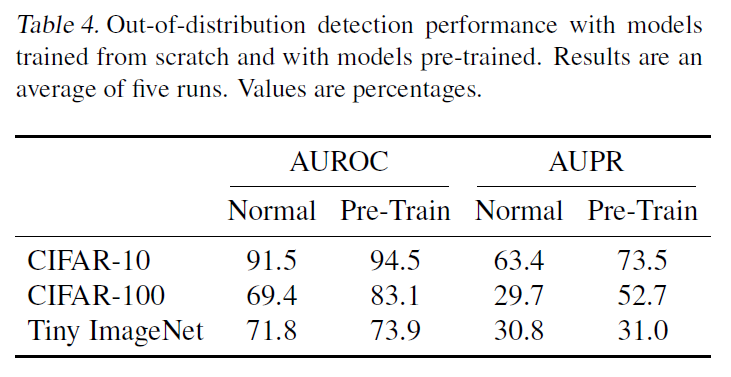
2.3 Robustness to Class Imbalance

Class Imbalance是指多分类任务时，每一类的数据量大小不同，导致Imbalance。模型会倾向于去学习占大多数数据的类别，因此数据量越少的类误差越大。如下图所示：只需了解绿色和蓝色线，分别为training from scratch和pre-training（这里不需要adversarial training）。



2.4 Uncertainty：Out-of-Distribution Detection

Out-of-distribution （OOD） detection是指：在test阶段，模型对输入样本产生一个异常分数，来指示其In- 或是 Out-of-distribution（训练数据分布）。本文用softmax输出的最大值表示，OOD样本具有较小的最大值。本文用test dataset做为in distribution样本，利用ROC和PR曲线AUC评估结果（网上非常多解释），值越大越好。下表可以看出，pre-training提高了检测成功率。



2.5 Uncertainty：Calibration

**4 Conclusion**

个人观点：本文作者不仅考虑针对adversarial sample的robustness，同时考虑其他常见robustness和uncertainty，更接近现实场景。本文作者已相继在顶会发表多篇文章，利用不同方法提高robustness和uncertainty。