|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗读  **文章标题：**  Fast is better than free: Revisiting adversarial training  **中文标题：**  快比免费更好：重新审视对抗训练  **发表于：**  ICLR2020  **作者：**  Eric Wong, Leslie Rice  **单位：**  Machine Learning Department  Carnegie Mellon University  Pittsburgh, PA 15213, USA | 提出所谓的灾难性过拟合现象。  提出了通过随机重启，及Cyclic learning rate来改善对抗训练  （这tm是ICLR2020的工作？） |
| **摘要及介绍** | **实验** |
| 对抗性训练是一种学习鲁棒深度网络的方法，由于需要通过投影梯度体面（PGD）等一阶方法构造对抗性例子，因此通常被认为比传统训练更昂贵。  在本文中，我们令人惊讶的发现，使用一个更弱和更便宜的对手来训练经验稳健模型，这种方法以前被认为是无效的，使得该方法在实践中并不比标准的标准训练更昂贵。  具体地说，我们证明了使用快速梯度符号法（FGSM）的对抗性训练，当与随机初始化相结合时，与基于PGD的训练一样有效，但成本明显更低。  此外，我们还证明，通过使用标准技术有效训练深度网络，可以进一步加速FGSM对抗训练，允许我们在6分钟内获得一个鲁棒性cifar10分类器，在面对PGD攻击（epsilon=8/255）时可以取得45%的鲁棒正确率，在12小时内取得一个鲁棒性Imagenet分类器，在面对epsilon=2/255的PGD攻击时获得43%的鲁棒性精度。与过去基于“免费”对抗性训练的工作相比，后者需要10个小时和50个小时才能达到相同的阈值。  最后，我们确定了一种被称为“灾难性过拟合”的失败模式，这可能导致之前使用FGSM对抗性训练的尝试失败。 |  |