Freelb

本文从对抗性训练的安全效益出发，研究了对抗性训练对泛化的影响。当对抗性训练增强健壮性时，它是广泛的。Ly被计算机视觉研究人员接受，它与泛化不一致，对未损坏的图像的分类精度在cifar-10上下降了10%，在Imagenet上下降了15%(Madry)。etal.，2018；Xie等，2019）。令人惊讶的是，人们观察到语言模型的相反结果(Miyaho等，2017；Cheng等，2019)，显示对抗训练可以改善两个基因化和鲁棒性。

Fsgd，Fgd，pgd，假设第二层分别e1，e2是纯度最高的猫，狗，e1如果想接近e2只用在线性意义上往一个方向偏。（这个方向不是狗本身）

靠谱的pgd防御：

做宝搞：最大攻击必须在知道当前模型参数才能求，而防御又必须用现有参数去

先建立在没变时，再变一下正常训练，再变一下接着训练，损失是n者和的平均。每一次都要改。

（也有个人提的是损失是最后一个。最后再改）

Freelb是先gt（样本在变）再gδ，再改δ，做完K次后再更昕seta

先seta梯度（初期有个随机扰动）。再基于现在未改的seta但已经改了嵌入算δ梯度，再改δ，也就是说小操作中seta未动，但会基于seta积累梯度，扰动一直在变，并且δ在不断的改，总共做K次。最后更新一次。并且seta虽然在pgd时不改但一直在积累梯度。

注意embedding后再做扰动。