IJCAI-19 阿里巴巴人工智能对抗算法竞赛总结

wanghao

2019 年 5 月 31 日

1 比赛过程

虽然比赛结果不好，但是还是作下记录。

本次比赛分为三个赛道，防御、无目标攻击、目标攻击。刚看到题目的时候，跟队友讨论了一波，发现防御的跑个分类模型就能得到结果，看起来似乎很简单。于是就决定主要做防御赛道，为比赛 gg 埋下了伏笔。

1.1 初赛

首先在网上找到了防御的三篇论文《Defense Against Adversarial Images using Web-Scale Nearest-Neighbor Search》、《PIXEL DEFEND: LEVERAGING GENERATIVE MODELS TO UNDERSTAND AND DEFEND AGAINST ADVERSARIAL EXAMPLES》、《Retrieval-Augmented Convolutionsl Neural Networks for Improved Robustness against Adversarial Examoles》。觉得第三篇用了图片压缩技术可能在比赛中用不到就只看了前两篇。第一篇论文通过搜索与对抗样本相似的干净样本来防御，pass。第二篇论文通过概率的方法逐像素恢复原图，觉得不靠谱，pass。这几篇文章中提到了 FGSM,i-FGSM,CW,DeepFool 等攻击方法，初步了解了 fgsm。

同时，我们训练了 resnet101,inception3,vgg,densenet 等基础分类网络，分辨率均为 224。除了 resnet, 其他的效果都不好，resnet101 分数达到了 14.3263。可能是攻击模型中没有 resnet?

看完论坛中大佬的 baseline 开始补论文,fgsm, 对抗训练，集成对抗训练，hgd，随机 padding。很多论文中都说 fgsm 的黑盒迁移性最好，后续就只用了 fgsm, 连 i-fgsm,pgd 都没试。。坑。。然后对 resnet101 只用 fgsm 进行了集成对抗训练，效果只提到了 14.8641. 侥幸进了复赛。赛后向前排大佬请教，发现 fgsm 的扰动上限设置小了，怪不得毫无效果。

github 找到 hgd 的模型代码，刚好是 pytorch 的，拿来就直接用了。

1.2 复赛

没仔细看 hgd 的训练过程，知道比赛才发现 hgd 用的是多个分类模型算的损失。。。

提交之前训好的 224 分辨率，resnet101 模型，效果不好,gg。直接怀疑 resize 的有效性。重新训练 299 的分类模型。

提交休战期间训练好的 hgd, 效果不好，gg。

提交有随机 padding 但是没有对抗训练的模型，gg。

然后开始找其他去噪模型，看 hgd 中提到的 DAE，论文中说把自编码器和分类网络压在一起形成新的网络，然后同样能产生对抗样本，心里蒙上了一层阴影。然后开始搜一般的盲去噪网络，大半没看懂，

放弃。后来找到 comdefend，文中说抗干扰能力很强，还不需要对抗样本。发现 comdefend 中公式含义不清晰，然后误以为网络中有二值化的操作，思考之后发现这个模型的想法真的不错。复现完 comdefend 模型。训练完后，经过去噪网络后的图片达不到论文中描述的 psnr，训练 3 小时一轮。

提交训好的 comdefend 模型，效果一般，gg。

无法忍受 comdefend 模型训练之慢，仿照 hgd 网络，去掉所有的跨越编解码器的 shortcut，用残差块作为基础网络，加入噪声和 sigmoid 和二值化 (不应该加二值化，对论文没理解对)，自己设计了个新的去噪网络。新网络的训练效果让我一度对他产生了很大的希望，速度快，准确率高。

提交新的模型，效果一般，gg。

怀疑人生之后，忍无可忍，提交了模型融合。多种去噪网络和多种分类网络，包括集成对抗训练的 resnet101 和随机 padding 层。终于达到了最高分，从一堆 5、6 分的渣渣模型变成了 9 分。虽然 9 分也渣。这一天刚好是 5.20 号，达到了这次比赛的最好成绩。

还剩下 10 天，天天脑子里就想的就是：

去噪，不用对抗样本！

去噪，不用对抗样本！

去噪，不用对抗样本！

期间尝试在 resnet 进入全连接层之前做噪声攻击，然后二值化，遇到了后述加噪声模拟攻击的问题。

终于有一天跟师兄讨论，重新理了一遍去噪网络的作用，用数学形式写了一下，推出了一个包含对抗样本的损失函数，联想到 comdefend 加噪声的方式，忽然间想到可以在编码器后的输出加入噪声模拟攻击即可。兴奋地和队友讨论，增强信心之后，实现了这个思路。期间发现了梯度截断的问题，借鉴了二值化网络的方法，用了 hardtanh。

提交最新的模型,gg。

发现在编码器后面加噪声并不能解决问题，前面的编码器网络的参数极有可能在原来的基础上集体变大 10 倍、100 倍，使网络的输出整体变大以达到抗干扰的目的。再度陷入迷茫，决定采用 fgsm 生成的对抗样本，不使用噪声模拟攻击。

提交新模型，gg。

返回的分数跟训练效果完全不一致，缩小为测试的四倍，开始怀疑对抗样本生成有问题。在 l1 限制的 fgsm 上进行修改，只取梯度绝对值前 20% 的像素点改变梯度。用此方法攻击提交的模型，模型效果果然很差，修改后继续训练。此时 5.28 号。

一直在想怎么根据各点的梯度大小动态调整要攻击的像素点，懵逼地发现我应该用 l2 限制的 fgsm。。修改后继续训练。此时 5.29 号，以为 5.31 号才结束，心里稳定得一批 (fyzz)，看着群里大佬熬夜苦战。

5.30 号，惊悚发现早上 10 点结束。

提交防御新模型，返回 nvidia-docker error，心态炸裂，发现忘记把压缩位数改为 8 了，超过 10 点，gg。9.0309 分，排名 70。

提交 l2 的 fgsm 和降 80% 像素点梯度置零的方法，无目标攻击通道提升了几名，gg。44.3563 分，排名 96。

2 比赛总结

1. 同时做攻击赛道和防御赛道能较早发现从头持续到尾的错误，单人作战极易在错误的道路上越走越远，申清题意。

2. 要尝试突破原有认知, 不能看多了 fgsm 迁移性好就不尝试其他的攻击方法，不能看到论文中大部分扰动很小就不尝试大扰动。

3. 多模型真的很有用。论文的总结性的话只能信一半, 不同的数据集，效果不一定一样。

4. 不确定的或可疑的信息，一定要确认，否则当作不存在。

5. 比赛时，不同的方法一定要多尝试。继续战斗。