一、 提交的样本对比

（1） 部分原始图片

（2） 部分第一轮普通模型攻击样本

（3） 部分第二轮防御模型攻击样本

二、 赛题分析

本次的百度 AI 安全对抗赛比赛分为初赛和复赛两个阶段，比赛中，我们需要通过对官方指定的 120 张图片添加扰动，使目标模型（Target Model）分类错误。

1. 初赛

（1）模型：需要攻击的模型一共有三个，包括官方给出了两个白盒模型，ResNeXt50 和 MobileNetV2，还有一个我们不知道模型结构与参数的黑盒模型。

（2）数据集：需要攻击的图片（120 张），来自于 Stanford Dogs 数据集。

（3）评价指标：

$$D(I, I^a) = \begin{cases} 128 & M(I^a) = y \\ mean(\|I - I^a\|_2) & M(I^a) \neq y \end{cases} \quad (1)$$

$$Dist\_Score(A,m) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} D_i\left(I_j, I_j^a\right) \quad (2)$$

$$Score(A,m) = \frac{128 - Dist\_Score(A,m)}{128} \quad (3)$$

$$Total\_Score = Score(A,3) \quad (4)$$

公式（1）是计算攻击后的图片I^a与原图I之间的距离度量公式，公式（4）是得分的计算公式。我们可以简单的理解为，目标模型无法正确识别攻击后的图片，

则可以得分，攻击后的图片与原图的差异较小时所得的分就高，差异大时所得的分就低。

2. 复赛

（1）模型：需要攻击的模型一共有五个，包括三个基础的模型，两个 AutoDL 模型，一个防御增强模型。基础模型中包含了一个初赛时的白盒模型 ResNeXt50。

（2）数据集：需要攻击的图片（120 张），来自于 Stanford Dogs 数据集。

（3）评价指标：

$$Totat\_Score = Score(A,3) * 0.2 + Score(B,1) * 0.4 + Score(C,1) * 0.4 \quad （5）$$

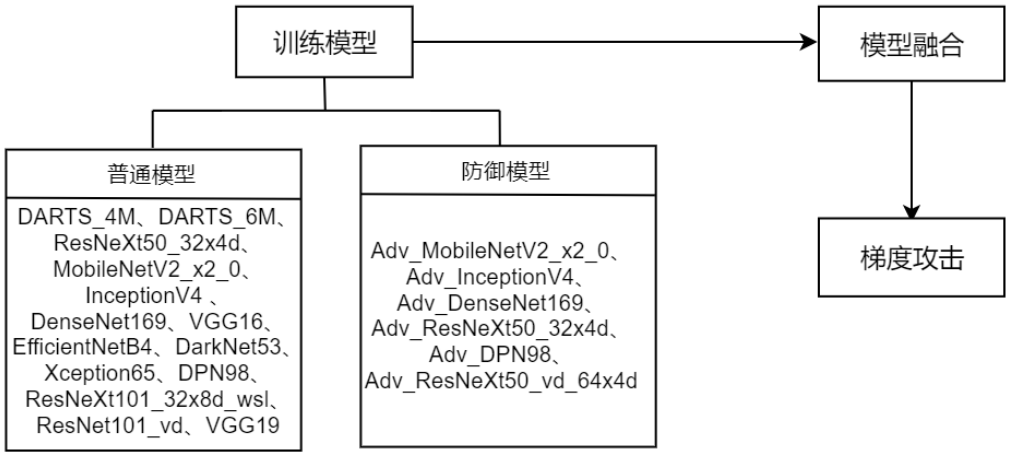
公式（5）是复赛的攻击得分计算，每类模型都有一个权重，基础(A)=0.2、AutoDL(B)=0.4、防御增强(C)=0.4。

初赛与复赛的赛题分析后，可以得出，我们目标有两点：

a. 使目标模型无法正确识别攻击后的图片。

b. 使攻击后的图片与原图的差异尽可能的小。

比赛的流程如下图



三、 模型训练

要攻击未知的模型，使用 FGSM 等依靠梯度的攻击方法，需要有训练好的模型来提供梯度。

在初赛中，仅仅使用官方提供的白盒模型 ResNeXt50，只能获得 49.2 分。

61	彭珏子	49.22291	64.99468	完成	2019-11-21 12:40

因此，想要获得高分，训练模型是必不可少的。好在 paddlepaddle 在 github 中给出了大量的预训练模型，网址如下。

https://github.com/PaddlePaddle/models/tree/develop/PaddleCV/image_classification

并且 paddlepaddle 在 github 中也给出了详细的训练过程，很好上手，其中，复赛涉及到的 AutoDL，官方也给出了其预训练模型。我们使用 Stanford Dogs 数据集（20580 张）进行模型训练。

初赛中，我们仅仅使用了官方给出的两个白盒模型和自己训练的 InceptionV4 获得了初赛第 9 名的成绩，得到了 89 分。

复赛中，我们训练了普通模型： DARTS_4M、DARTS_6M、ResNeXt50_32x4d、MobileNetV2_x2_0、InceptionV4 、DenseNet169、VGG16、EfficientNetB4、DarkNet53、Xception65、DPN98、ResNeXt101_32x8d_wsl、ResNet101_vd、VGG19 防御模型：Adv_MobileNetV2_x2_0、Adv_InceptionV4、Adv_DenseNet169、Adv_ResNeXt50_32x4d、Adv_DPN98、Adv_ResNeXt50_vd_64x4d，攻击整个过程中一共用到了 20 个模型。其中防御模型的训练，是将训练好的普通模型作为与训练模型，将普通模型攻击后的图片与原 120 张图片作为数据进行训练得到的。

四、 模型融合

在 Mi-FGSM 的论文**错误!未找到引用源。**中,对与模型在 Logits、Predictions、Loss 融合分别进行实验得出，在 Logits 融合的攻击成功率最高，效果最好。因此，我们也将多个模型在 Logits 进行融合。

复赛中，我们对图片进行了两轮攻击。第一轮是普通模型攻击，第二轮是防御模型攻击。在进行第二轮攻击时，很多防御模型并不能一次攻击成功，因此，在多次攻击中，我们调整不同模型 logits 的比重，从而减少攻击次数，也能降低 avg_mse。达到我们减少攻击后图片与原图的差异的目的。

根据复赛的评价指标，我们可以看出 AutoDL 的模型占据了 40%的比重，可以说也是很重要的。因此，我们在模型融合时，将 AutoDL 模型的初始比重设置的也比一般的模型要大，提交结果后也可以发现，AutoDL 模型比重较大时比与其他模型一样比重得分更高。

五、 攻击方法

在本次比赛中，感谢百度官方给出的 baseline,给我们了研究的方向与思路。官方的 baseline 中给了两个攻击方法 FGSM 与 PGD，我们后续的攻击算法也是基于这个 baseline 修改的。

我们所使用到的攻击方法：

- Mi-FGSM
- Si-Ni-FGSM

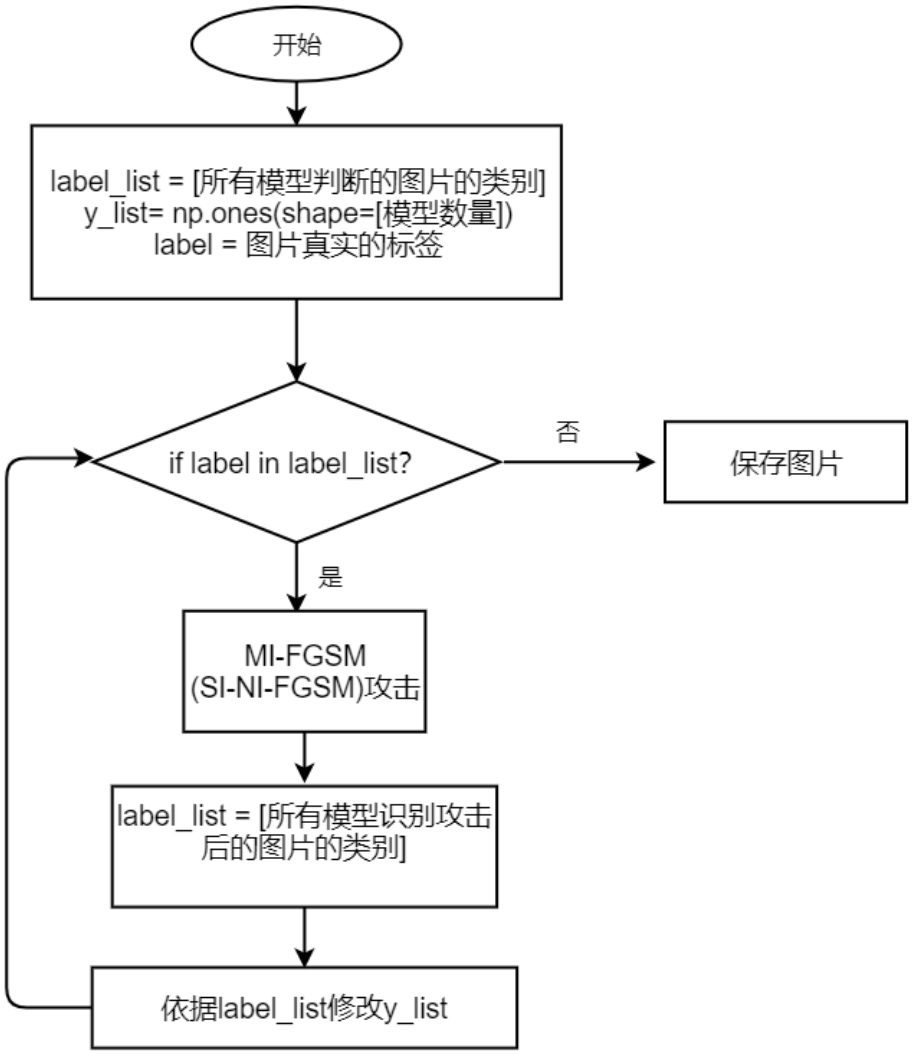
我们在初赛中选择将 FGSM 修改为 Mi-FGSM,Mi-FGSM 与 FGSM 攻击方法中的区别，可以简单的理解为在 FGSM 的基础上，增加了迭代的过程和动量。

复赛中，我们进行了两轮攻击，首先使用 DARTS_6M、ResNeXt50_32x4d、MobileNetV2_x2_0、InceptionV4 、DenseNet169、VGG16、EfficientNetB4、DarkNet53、Xception65、DPN98、ResNeXt101_32x8d_wsl、ResNet101_vd 等 12 个模型进行 Mi-FGSM 攻击。然后使用 DARTS_4M、Adv_MobileNetV2_x2_0、Adv_InceptionV4、Adv_DenseNet169、Adv_ResNeXt50_32x4d、Adv_DPN98、Adv_ResNeXt50_vd_64x4d、VGG19 等 8 个模型对第一步的结果进行 Si-Ni-FGSM 攻击。

复赛中，初始使用 Mi-FGSM 算法攻击时发现，有些模型加入之后，会增大扰动，得分时 avg_mse 变大，而在加入对抗训练模型之后，avg_mse 变大的更多。

因此，在使用对抗训练模型攻击时，改用 Si-Ni-FGSM 算法攻击，可以明显降低 avg_mse。

攻击流程图如下：



其中 y_list 就是我们 logits 融合的模型的权重。

六、 总结

在复赛中，我们一共用了 20 个模型，可以说是用的模型堆叠大法，感觉我们团队实力还有待加强。希望能多看看其他大佬们的方法，多多学习，早日进步。

我们队的代码地址：

https://aistudio.baidu.com/aistudio/projectdetail/247664

七、 致谢

首先感谢比赛的主办方百度公司给大家提供的这次机会和计算资源；同时感谢我的导师--云南大学软件学院周维教授对我们团队的大力支持和鼓励、云南大学软

件学院 1421 实验室为我们团队提供的 GPU 计算资源；最后感谢我们的成员@彭文钰师妹的全力付出、@程泰宁师弟的精神支持，百度官方@我觉得没有问题、@抄起键盘上战场对我们提的 issue、问题的耐心解答、指导和帮助。