

文章链接：<https://arxiv.org/pdf/1911.07682.pdf(accepted> by AAAI2020)

代码：https://github.com/CZHQuality/AAA-Pix2pix.

Abstract

解决问题：目前的集成攻击（ensemble attack）只考虑利用有限数量的源模型来设计敌手，可移植性较差。因而本文提出黑盒攻击，Serial-Mini-Batch-Ensemble-Attack(SMBEA)，将大量的预训练source models分为几个mini-batches，对每个单独的batch，设计3个ensemble策略以提高intra-batch的可转移性，如图1。同时，提出了一个新的递归算法，可以积累过去batch的长期梯度记忆到下一个batch。

Introduction

在设计将预训练模型的攻击转移到目标模型上时，source model和target model的结构和参数不同，但会共享相似的decision boundaries。如果敌手可以fool几个source models，那么就可以捕捉固有的转移对抗信息以攻击其他黑盒target models。

目前设计转移对抗样本的方法有两种：

1. ensemble attacks（2017年提出）:通过同时fool多种白盒源模型来设计可转移的对抗样本。
2. Generative methods（2018年提出）：依赖于GAN，以生成可以fool目标模型的敌手，同时训练了一个鉴别器（discriminator），将合成adversaries与orginal clean images区分开来，最大限度地降低了感知能力。

缺点：

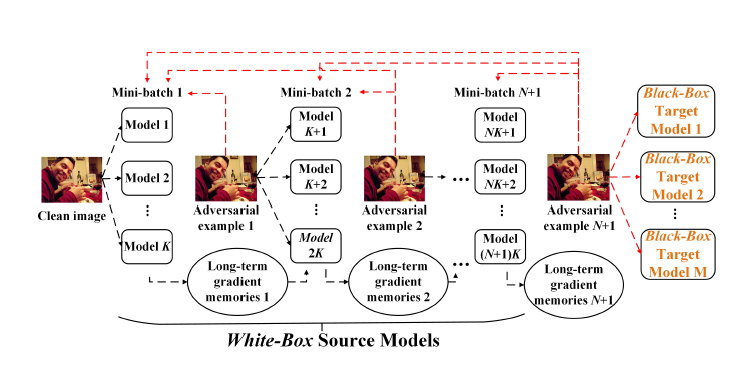
（1）对于ensemble attacks，只考虑了有限数量的source models，state-of-the-art攻击采用少于8种source models来设计adversaires，且是并行结合的，导致source models的数量受GPU存储限制。

（2）虽然并行计算技术支持对大量模型的并发攻击，但计算和反向传播大量模型的损失函数梯度很慢且很难。

（3）而使用GAN时，训练比较难，需要大量expensive标签的训练样本。

设计本文方法灵感：

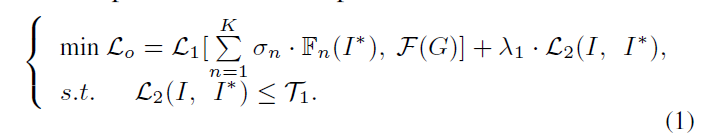
1. 构建一个对抗性的例子类似于训练一个模型，而对抗性例子的可转移性类似于模型的可泛化性，因此fool尽可能多的source models来增加敌手可转移性是可能的
2. 相比于扰动大小，扰动的空间结构对final fooling能力的影响更强。因此本文保留学习的对抗结构信息，而是优化fooling能力和可转移性。



Method

Inra-batch ensemble strategies

本文仅考虑targeted attacks，将对pixel-to-pixel任务的攻击视为一个约束性的优化问题。为了简化，首先介绍一个基本的ensemble策略（在输出空间融合了多种模型，优化目标由多个source model最终预测的元素加权和计算得到）：

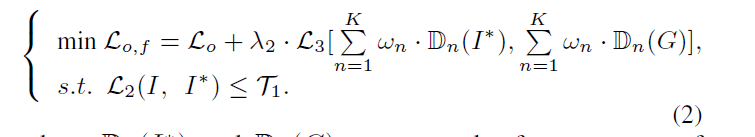


优化目标：改变模型对I*\**的ensemble预测，使其被预测为guide image。

其中，*I*和*I\**分别表示clean image和adversarial example。G表示引导图片（guide image），F(G)表示G的真实输出，

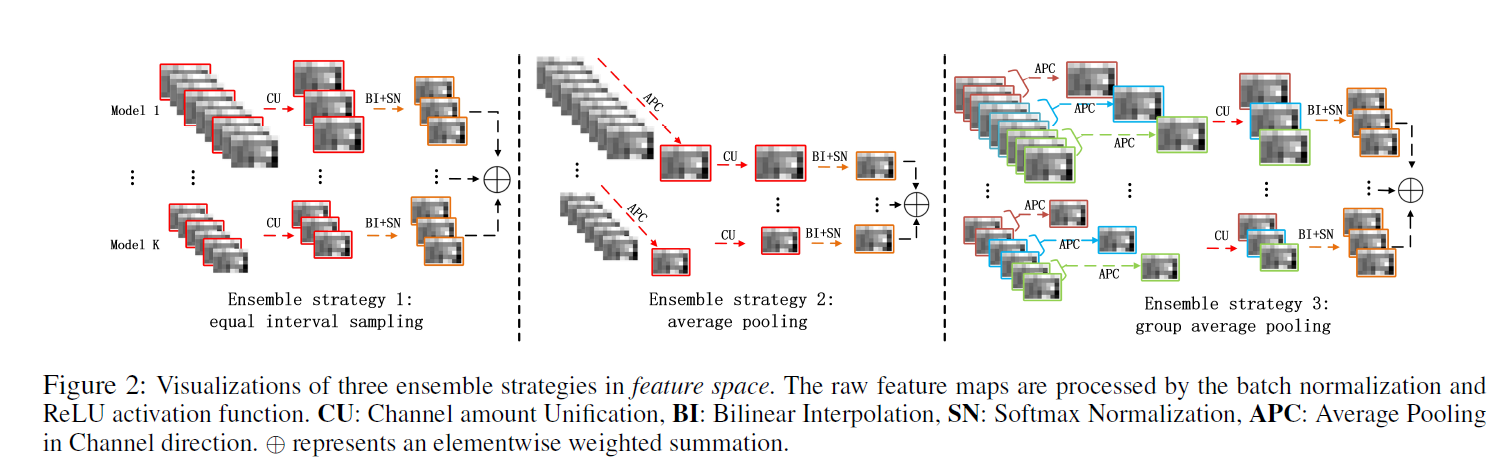
是在mini-batch的第n个source model，是在所设计的敌手I\*上的第n个模型的最终预测。是权重，。

改进：上述优化只fool了输出空间的最终预测值，本文将其扩展到特征空间，使其不仅能fool最终预测，还能fool特征maps：



其中和分别表示在I\*上的第n个source model的特征映射和引导图片G的特征映射。

由于不同的模型使用不同的网络结构，因而feature map有不同的分辨率和通道。为解决这个问题，本文使用了3种不同的特征空间ensemble策略，如图2.



第一个ensemble strategy从每个模型取p个feature maps，第n个模型的取样区间（为从第n个模型选择的一个特征层的特征通道总数），通过这种方法，可以从不同的模型获取相同数目的feature maps(channel)。然后使用（bilinear interpolation）调整不同模型的feature map为相同分辨率（如，高\*权重）。然后使用softmax函数进行归一化。最后，对不同feature map的元素加权求和后，得到特征空间ensemble result。

另外两种策略类似，除了第一步不同。第2种，计算的是每个通道方向上的Pn个特征映射的平均pooling，然后获取每个模型的2D one-channel特征映射。

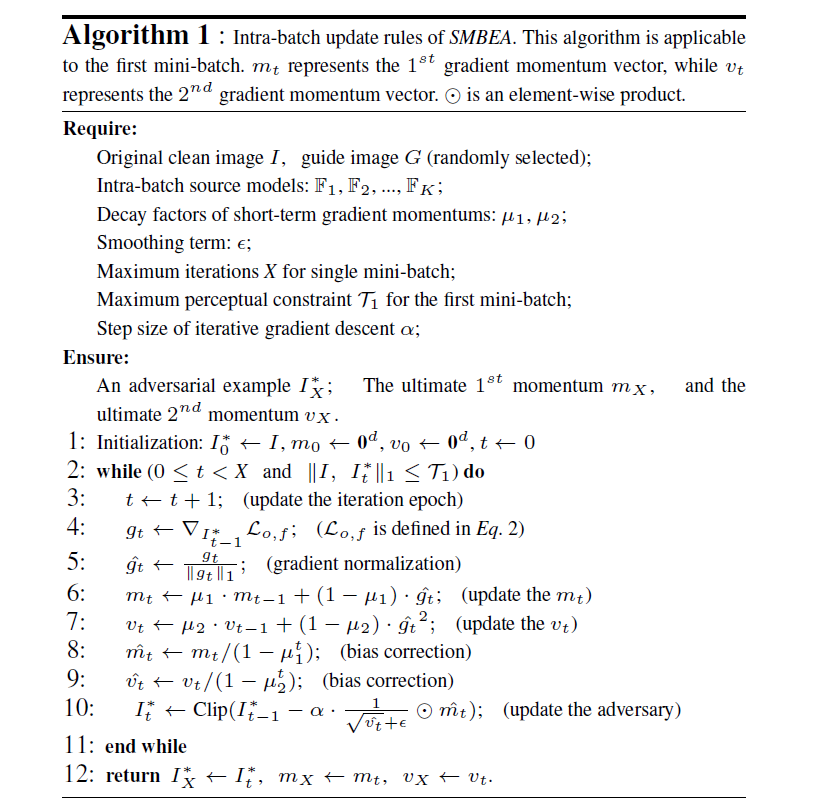
第3种，将Pn个特征映射分为p个group，计算每个group的通道方向平均值，然后从每个模型获取p个候选特征映射。

Intra-batch update rules

第一种动力：积累前几次迭代的梯度，稳定梯度下降方向，降低局部极值。

第二种动力：使learning rate适应不同的参数，为频繁更新的像素值设计小的更新步长，非频繁更新的像素值设计较大的更新步长。。因为在pixel-to-pixel攻击中，目标是更新adversarial example的图片像素，但在攻击过程中，小部分的像素值会频繁更新，其余像素值只偶尔更新，因而频繁更新的像素值达到扰动约束边界时，而非频繁更新像素值则还还远低于边界，限制了fool能力和扰动能力的trade-off。

本文采取结合了第一种和第二种动力，基于Adam优化（2015年提出的随机梯度算法），增加了梯度的归一化，使得step-size不受梯度尺寸转换的影响，既获得了好的可转移性能，又很好地平衡了fooling ability和perceptibility。



Inter-batch update rules

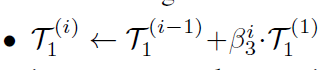
Intra算法只能保证adversarial可以fool有限数目的source models，因而本文设计inter算法打破这个限制，递归积累前面的batches的长期梯度记忆到后面的batches。长期梯度记忆可以保留已经学习的对抗信息，避免过拟合，如算法2。

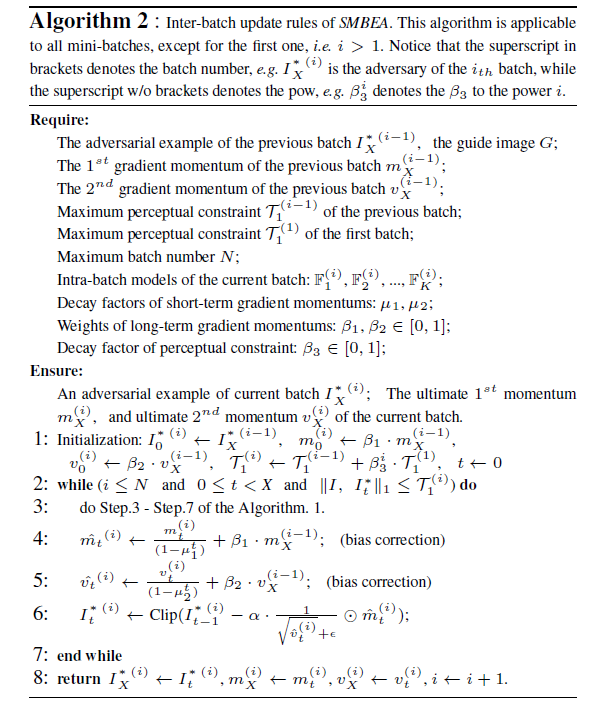
算法1和算法2的却别在于initialization和bias correction。

1. 初始化

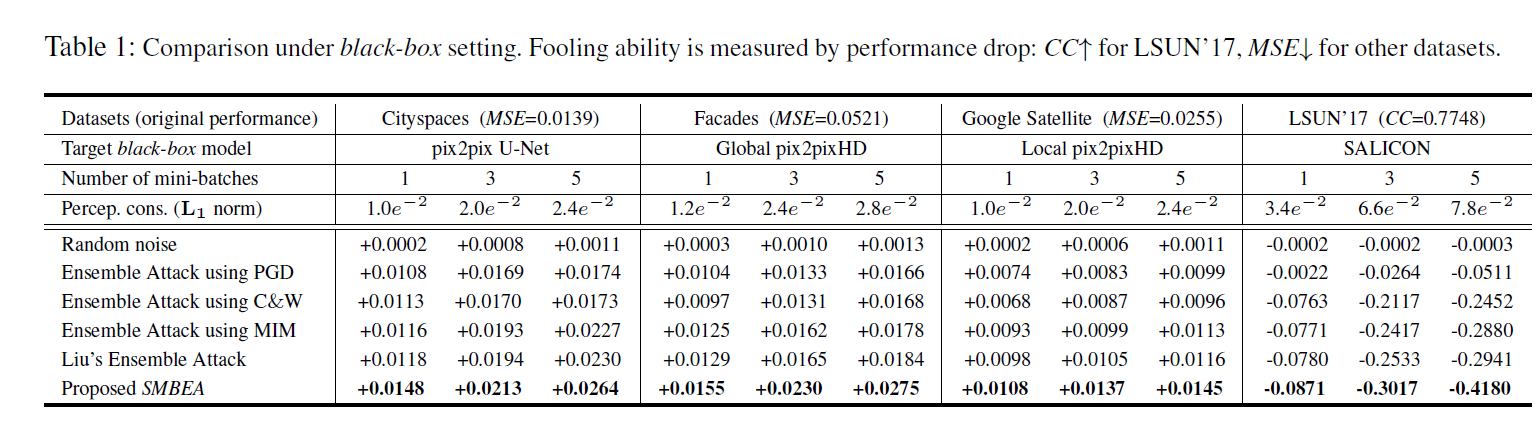
C:\Users\LISHUANG\AppData\Local\Temp\1588329171(1).png：将前面的batch作为当前batch的初始状态。

C:\Users\LISHUANG\AppData\Local\Temp\1588329430(1).png使用前面的batch的两种动力分别作为当前batch的初始动力。防止过拟合，保留学习的对抗信息。

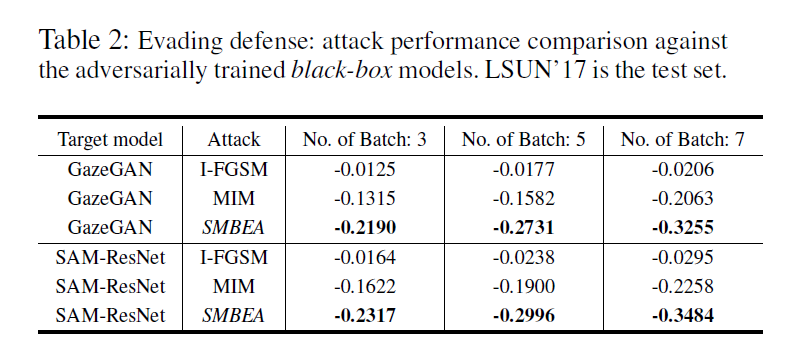
：通过增加因子C:\Users\LISHUANG\AppData\Local\Temp\1588335366(1).png到前面分支T1的扰动约束，递归更新当前batch的最大扰动约束。防止过早收敛而拟合不足。



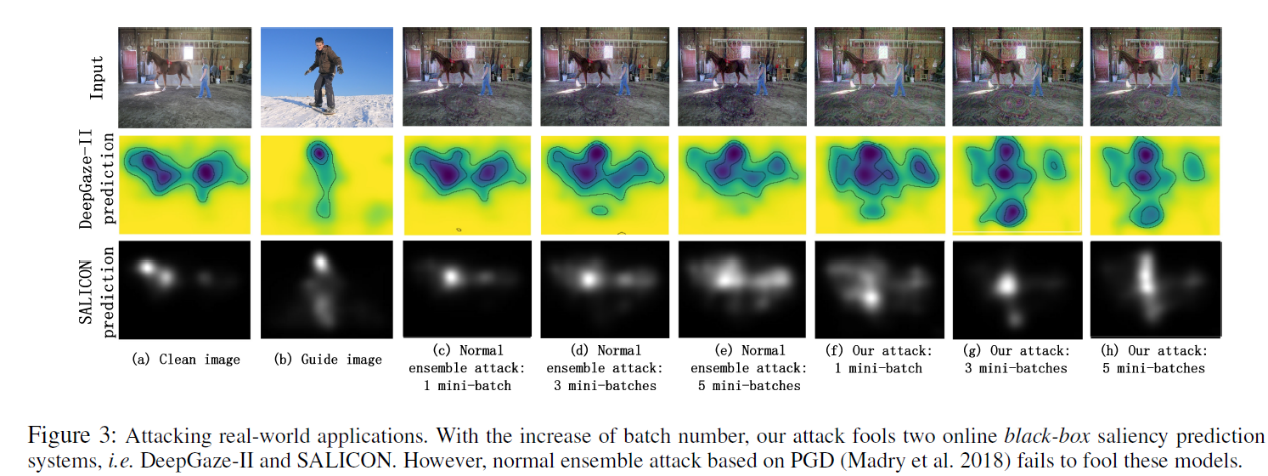
Experiments



在黑盒设定下，对比方法都采用了state-of-the-art的梯度反向传播攻击算法，且不同的对比方法，使用相同的扰动约束，可以看到本文在不同的dataset的性能表现都更优。



可以看到，本文方法所训练出的对抗样本，攻击性更强。



和PGD方法比较，对在线黑盒预测模型攻击，随着batch的增加，本文方法误导模型预测图片为guide图片更强，而普通方法failed。

Conclusion

本文实际上是一种基于随机梯度的攻击方法，只针对像素级别的图像学习模型，提取不同模型的feature map共性，以训练出adversarial，有种分而治之的思想，以降低花销。