What it Thinks is Important is Important: Robustness Transfers through Input Gradients

文章链接：<https://arxiv.org/pdf/1912.05699.pdf>

Accepted by CVPR 2020

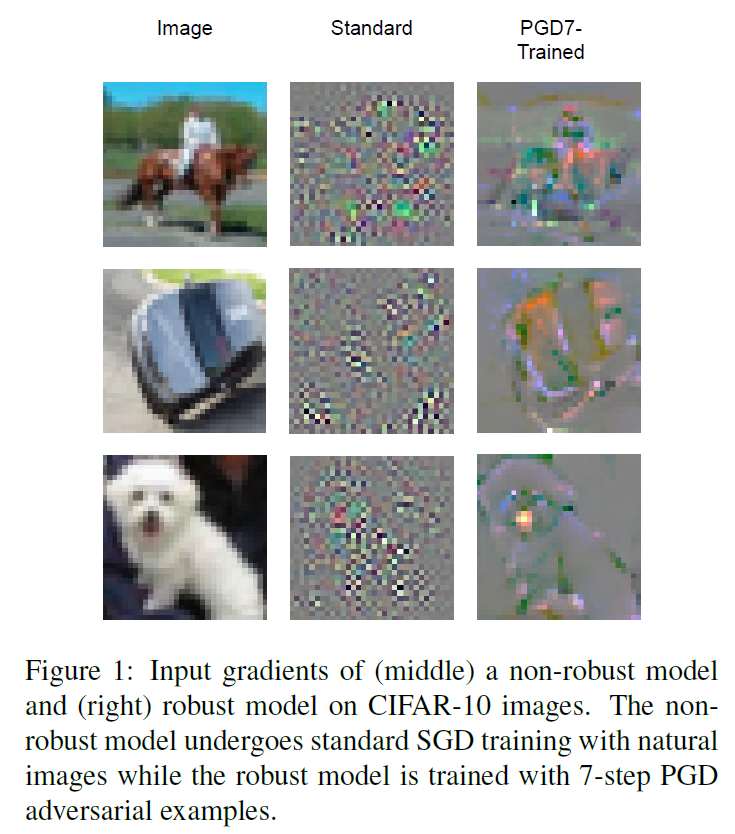
**1 Abstract**

本文在ICLR 2020 文章robust transfer learning（简称RTL）基础上进行改进。RTL中要求：teacher model 和student model架构相同。本文提出robustness可以在不同架构之间迁移。本文利用clean sample训练student model，计算input gradients（即loss对输入样本的梯度，简称IG），以匹配robust teacher model（已经对抗训练好的model）的IG，使得student model继承robustness。

**2 Introduction**

adversarial training（简称AT）每次训练都需要产生adversarial sample，因此需要耗费大量的计算开销。文本propose a robustness transfer method（简称IGAM） that is both task- and architecture-agnostic with input gradient as the medium of transfer。

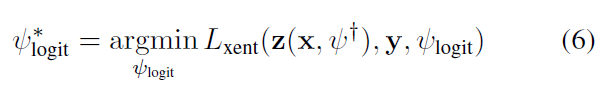
Input gradient：loss对输入样本的梯度。AT模型所计算的IG是具有样本结构特征的。然而，正常训练的却没有。IG的大小表示每个pixel对模型预测的影响程度或称为重要性。因此，IG的结构特征是衡量robustness的一个基准，本文matching IG，从而继承robustness。



**2 Approach**

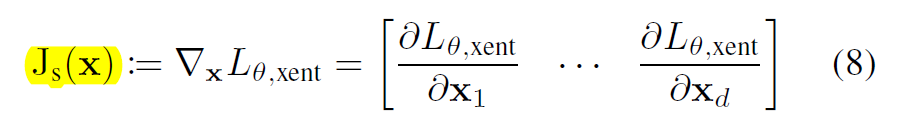
本文包含两个阶段：1）fine-tuning robust teacher model on target task；2）train student model by matching IG。

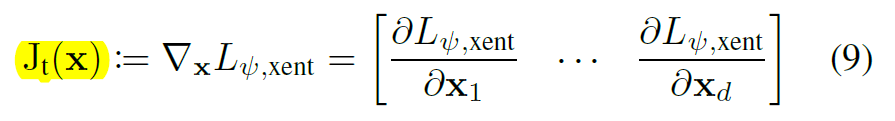
Phase 1:

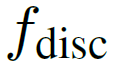


冻结robust teacher model的feature extractor层  ，fine-tuning最后的logit层 ，以适应target task。借鉴于ICLR20，因为 feature extractor层具有robustness，因此fine-tuning后的模型具有robustness。

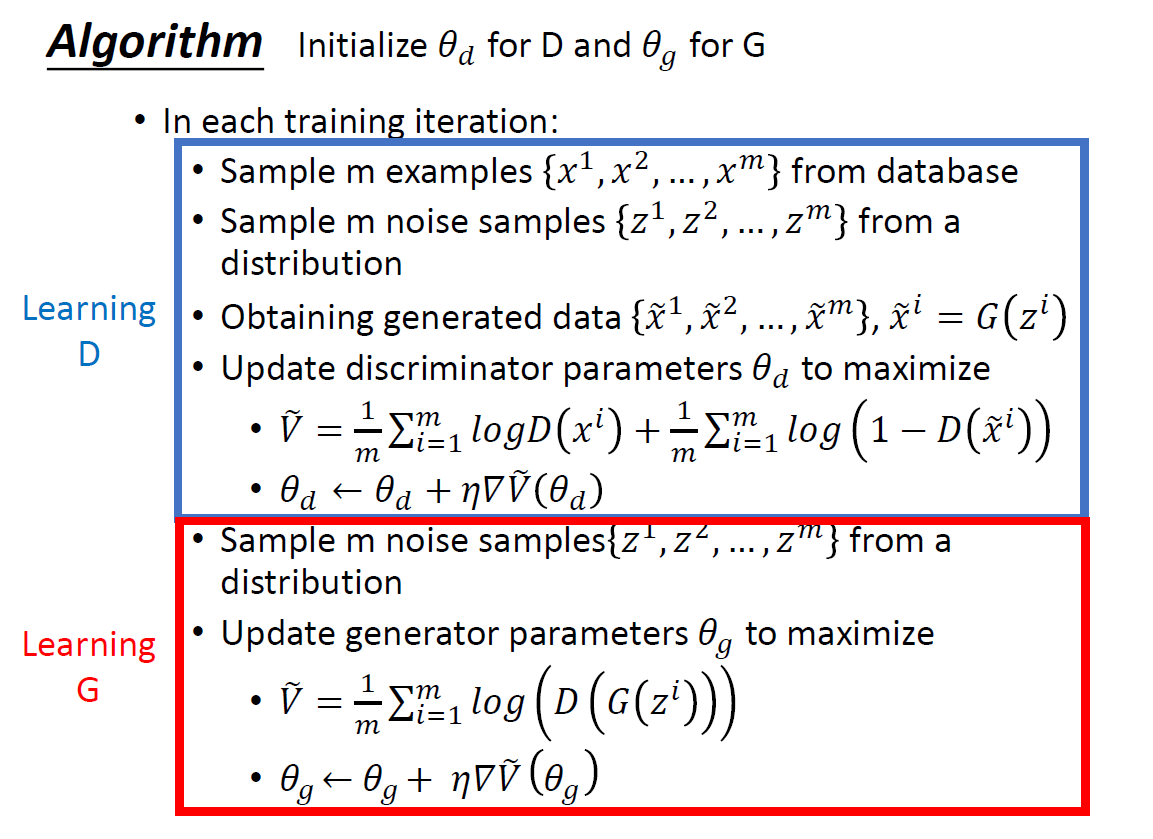
Phase 2：



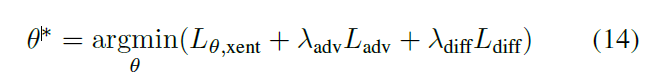


J\_s和J\_t分别表示student 和teacher模型计算的IGs。核心观点是，让J\_s来模仿J\_t。本文借鉴GAN（包括generator G和discriminator D）的思想，训练一个D（文中表示）来区分J\_s和J\_t，student model充当G（文中表示，teacher model为），来生成J\_s。

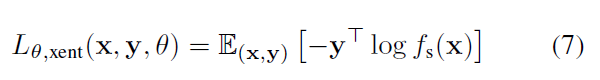
GAN：包括generator G和discriminator D，用来生成训练数据，例如图片。G生成图片，D区分生成的图片和真实图片。训练D使得区分能力越来越强；训练G使得生成的图片让D区分不来，即生成的图片和真实图片非常接近。下图是标准GAN的训练过程：



本文student模型训练的目标函数为：



其中，第一项为常规的cross entropy目标函数：



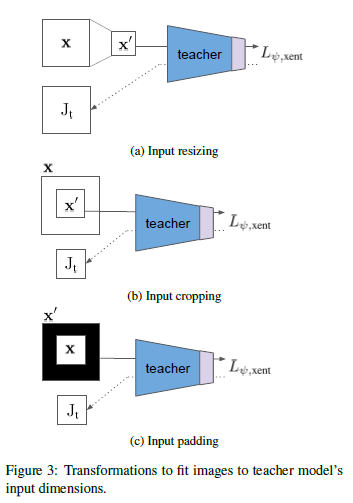
第二项为GAN中的目标函数，即D的区分能力。G需要最小化这个值，D需要最大化这个值；本文中， student模型最小化这个值，discriminator D最大化这个值：



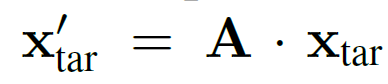
第三项为判定J\_s和J\_t的L\_2范数差异。



补充：上述模型都假设teacher and student models的输入维度是相同的。下面利用维度变化，来relax这个假设，如下图：

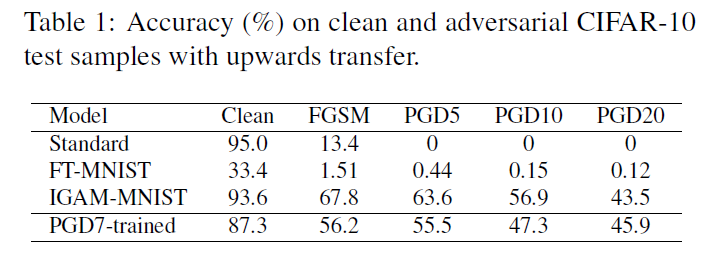


1. Input Resizing：利用线性变换，A为高维tensor

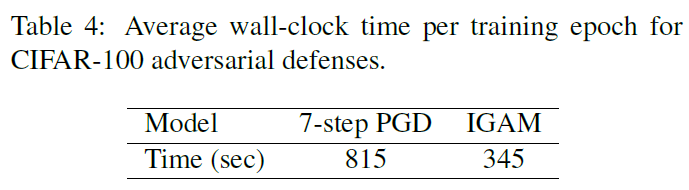


b）、c）分别为cropping，剪裁边上像素值；padding，在边上填充0。

**3 Experiment**



上图为从MNIST数据（2层CNN） 迁移至 CIFAR数据（Wide-Resnet 32）。Standard表示正常训练结果；FT-MNIST表示在2层CNN上用CIFAR-10数据finetuning的结果； IGAM-MNIST表示本文matching IG方法；PGD7-trained：表示利用CIFAR-10数据对抗训练的结果。本文方法具有更好的natural accuracy和robustness。（其他实验结果类似不一一列举）



与AT相比，本文提出方法具有更少的计算开销，减少了57%的开销。

**4 Conclusion**

个人观点：本文虽然没有利用AT，但是相比于ICLR20文章，本文需要更多的训练数据和计算资源，用来matching IG。优势在于不同架构下可以进行robustness迁移。

据了解，2020年已经有三篇AI顶会研究了robustness的迁移性(已经分享两篇，下一次分享另外一篇)，无需进行AT，进而提高训练效率。应该是一个趋势，因为AT对sample complexity和computation complexity要求过高，值得深入研究。