信息网络专题研究课文献阅读报告（应用层）

一、 文献信息

1、 论文题目：DELVING INTO TRANSFERABLE ADVERSARIAL EXAMPLES AND BLACK-BOX ATTACKS

2、 作者：Yanpei Liu，Xinyun Chen，Chang Liu，Dawn Song

3、 发表途径：Published as a conference paper at ICLR

4、 发表时间：2017

二、 问题意义

1. 研究背景

无论是在网络还是操作系统领域，现实应用中我们或多或少对网络攻击和病毒软件都有所耳闻，在神经网络这一领域同样也存在类似的威胁。对抗样本的存在证实了这一点，它指的是在数据集中通过故意添加细微的干扰所形成的输入样本，可以导致机器学习模型接受并做出错误的分类决定。黑盒攻击是指攻击者对攻击的模型的内部结构、训练参数、防御方法等等一无所知，只能通过输入输出与模型进行交互的攻击方法。这二者对神经网络的安全性造成了威胁。

2. 研究问题

本文主要研究了对抗样本的可转移性以及黑盒攻击下的特性，并且关注点不只是停留在小样本即，而是针对大规模样本集。并在 Clarifai.com（最先进的在线分类模型）上实现黑盒攻击。

三、 思路方法

1. 对抗性深度学习与可转移性的简单介绍

此部分介绍了所使用的生成对抗样本的算法，分别为基于优化算法、快速梯度法、FGSM 三种算法，每种算法分别进行目标攻击对抗样本的生成和非目标攻击对抗样本的生成，目标攻击就是把公式中的 y 换成目标类 y’，并将梯度符号及损失符号变号。然后在五种攻击模型中测量了可转移性，给定两个模型，通过计算一个模型生成的对抗样本在另一个模型上分类正确的百分比，称作 accuracy，来测量非目标攻击的可转移性。accuracy 越低，说明攻击效果越好。而目标攻击的可转移性计算一个模型生成的对抗样本在另一个模型中分类为指定类别的百分比，称作 matching rate。matching rate 越高表明目标攻击效果越好。

2. 非目标攻击和目标攻击

A. 非目标攻击

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

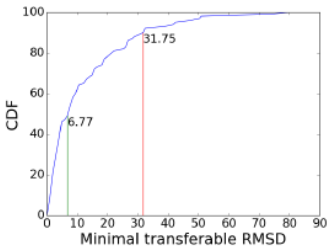
Panel A: Optimization-based approach

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.45	4%	13%	13%	20%	12%
ResNet-101	23.49	19%	4%	11%	23%	13%
ResNet-50	23.49	25%	19%	5%	25%	14%
VGG-16	23.73	20%	16%	15%	1%	7%
GoogLeNet	23.45	25%	25%	17%	19%	1%

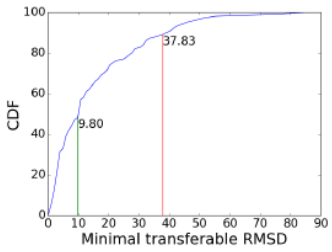
Panel B: Fast gradient approach

上表为基于优化算法和基于快速梯度法的结果，RMED 表示最左侧模型得到结果的平均失真率，而表中间每个格则是模型（行）生成的样本在模型（列）上迁移的结果。

超参数 B 和 RMSD 有着密切的关系，呈正相关。也可以看出，RMSD 同可转移性是正相关，即 RMSD 越大，扰动越强烈，转移性越好。所以要通过控制 BB 来控制最小转移性的 RMSD，实现干扰和转移性的折中。



(a) Fast Gradient



(b) Fast Gradient Sign

作者给出的实验结果如图所示，VGG-16，ResNet-15；左图为 FG 结果，右图为 FGSM 算法结果；CDF 为对应的累积分布函数，按原文可理解为转移性；FG 算法的结果要好于 FGSM。

B. 目标攻击

作者发现目标攻击对于用于训练的模型能够得到好的结果，但是对于其他模型的转移能力几乎没有，下图为基于优化算法的结果。

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.13	100%	2%	1%	1%	1%
ResNet-101	23.16	3%	100%	3%	2%	1%
ResNet-50	23.06	4%	2%	100%	1%	1%
VGG-16	23.59	2%	1%	2%	100%	1%
GoogLeNet	22.87	1%	1%	0%	1%	100%

作者发现即使增加扰动也不能提高转移性，包括基于梯度的算法。

3. 基于集成的方法

作者分别针对几个模型进行了独立的实验，实验结果表明，无论是目标攻击还是非目标攻击效果都有很大的提升，但是作者也发现对角线上的 accuracy 不是 0。作者假设这一现象存在的原因是不同模型之前的梯度相互正交，因此沿着这个方向可能需要很大的失真才能得到对抗样本。

4. 不同模型的几何特性

为了更好的理解对抗样本转移性，作者从几何的角度来观察，参考了 1000 个类标的大型数据集。据作者实验观察，不同模型之间的梯度方向几乎是正交的，不同模型梯度方向间夹角的余弦值，非对角值都接近于 0。

A. 单模型算法非目标攻击的决策边界

对于 VGG-16 模型，在与真实相对应的区域内存在一个小洞。这也许可以部分解释为什么非靶向畸变小的对抗性图像存在，但不能很好地转移。这个洞在其他模型决策平面中不存在。在这种情况下，非目标对抗性图像在这个洞不转移。

B. 基于目标集合的方法的决策边界

作者选择除了 ResNet-101 和随机正交方向之外的所有模型集合的目标对抗性方向，并在由这两个方向向量张成的平面上绘制决策边界，可以观察到，被预测为目标标签的图像区域，与集成中的四个模型很好地对齐。然而，对于不用于生成对抗性图像的模型（ResNet-101），它也有一个非空区域被成功误导到目标标签，虽然面积小得多。同时，各模型闭合曲线内的区域几乎具有相同的中心。

四、 实验结论

结果证实，即使对于大模型和大规模数据集，非目标对抗样本的可转移性也是显著的。另一方面，作者发现很难使用现有的方法来生成具有可转移性目标标签的目标对抗样本。所以作者开发了新的基于集成的方法，并证明它们可以生成具有可转移特性的目标对抗样本且成功率很高。这些新方法在生成非目标可转移的对抗样本方面的性能有大幅度的提升。此外，使用新方法生成的非目标和目标对抗样本都可以成功地攻击黑盒图像分类系统 Clarifai.com。最后，不同模型的几何特性的研究对更好地理解可转移对抗样本做出了有利贡献。

五、 启发思考

从计算机和互联网诞生开始，攻击就是无处不在的，防御方法也在不断升级。这本来就是矛与盾的关系，其实并没有一些所谓真正的安全，攻击有时候只是成本问题，事实表明这并不会阻碍计算机和互联网行业的蓬勃发展。目前为止，我们能做的有两点：一是关注该领域的发展动向，二是做好针对此类攻击的风险控制。