对抗样本攻击与防御报告

胡若凡 3200102312

1. 序言

随着深度学习在各种应用领域的快速发展和显著成功，针对其的攻击也愈发严重。研究发现，由于深度神经网络（下称DNN）的黑盒性等原因，攻击者可以设计出针对样本的**微小扰动**，尽管它们对于人类可能是细微或者直观无影响的，但是却能导致输出结果的迥异性。

在这篇综述中，分析了有关 DNN在计算机视觉等领域上有关对抗样本攻击的最新研究成果，总结了生成对抗性样本的方法，并提出了这些方法的分类法和对应的防御方法。

并且，对抗样本攻击基于了以下假设：

- 攻击者只在测试和部署阶段攻击，并不在训练阶段攻击
- 攻击者攻击的模型是DNN模型
- 攻击者的目标是降低训练结果的精确度

2. 机器学习VS深度学习

2.1 深度学习概念

- 概念上，深度学习是一种机器学习的方法，其利用神经元函数构建深度神经网络，能够充分利用现代计算机的**并行性**算力，享受硬件带来的优化结果。在我们对计算机视觉的讨论之中，**CNN**与**RNN**是两种十分著名的方法
- 模型上，有一些出名的例如LeNet , VGG , AlexNet , GoogLeNet , and ResNet
- 数据集上，在图像领域，有一些出名的数据集**MNIST, CIFAR-10,ImageNet。**MNIST是手写图像识别，后两者是图像识别任务

2.2 机器学习概念

- 任务目标：关于机器学习对抗样本的攻击与防御，主要关注二元分类问题，例如病毒检测系统、入侵检测系统和入侵预防系统。而手工制作特征的机器学习系统是主要的目标，例如垃圾邮件过滤器、入侵检测、生物特征认证和欺诈检测。例如课堂上就提到过的，为了避免检测，垃圾邮件通常会通过添加字符进行修改。
- 任务分类：Barreno等人对机器学习的安全问题进行了初步调查，得到了如下的分类：（1）影响：攻击是否会污染训练数据；（2）安全违规：对抗性样本属于假positive还是假negative；（3）特异性：攻击是针对特定实例还是广泛类别的。
- 模型区别：ML与DL相比，ML需要 **特征提取** 方式，DL 只需要 **数据输入**

3. 对抗样本概念

3.1 Threat Model

威胁模型可以进一步分解为四个方面：对抗性伪造、攻击者的知识、对抗特异性和攻击频率。下面对每个概念做一个记录

- **Adversarial Falsification**

  分为False positive和False negative两个种类，即伪造的结果

- **Adversary's Knowledge**

  分为black box和whte box两个种类，前者只知道输入与得到的结果，后者可以知道模型的具体设计细节

- **Adversarial Specificity**

  分为Targeted attacks和Nontargeted attacks，前者有明确的misguide导向，后者只要使输出的结果部署正确结果即可

- **Attack Frequency**

  分为One-time attacks和Iterative attacks，两者相比之下，相较于一次性攻击，迭代攻击通常可以产生更好的对抗性示例，但它需要更多的交互（更多的查询）与受害者分类器，并花费更多的计算时间来生成它们。对于一些计算密集型任务（例如，强化学习），一次性攻击可能是唯一可行的选择。

3.2 Perturbation

微小扰动是对抗性样本的一个基本前提。对抗性样本被设计为接近原始样本并且不可察觉的，这样才能达到对抗样本设计者的设计目的。下面对几个具体概念进行记录

- **Perturbation Scope**

  这里分为Individual attacks和Universe attacks。个体攻击为每个干净输入产生不同的扰动，通用攻击只为整个数据集创建一个通用扰动

- **Perturbation Limitation**

  这里分为Optimized perturbation和Constraint perturbation。优化扰动将扰动作为优化问题的目标。这些方法旨在最小化扰动，使得人类无法辨认扰动。 约束扰动将扰动设置为优化问题的约束条件

- **Perturbation Measurement**

  这里分为lp范数和PASS。举例而言，lp通过p-范数距离测量扰动的大小，计算在对抗性样例中改变的像素数，心理学感知对抗相似性分数（PASS）是最新的一种评判标准，其与人类感知一致，可以测量对抗性样例与原始样本之间的相似性。该度量标准是基于对称加权KL散度提出的。

3.3 Benchmark

数据集和受害模型的多样性，使得研究人员难以判断对抗样本的存在是由于数据集还是模型所致。下面记录常用的数据集与受害模型

- **Data Sets:**

  MNIST、CIFAR-10和ImageNet是最广泛使用的图像分类数据集，用于评估对抗性攻击。因为MNIST和CIFAR-10由于其简单性和小尺寸而易于攻击和防御，所以ImageNet是迄今为止最好的数据集来评估对抗性攻击。需要设计良好的数据集来评估对抗性攻击。

- **Victim Models**

  对手通常会攻击几个众所周知的深度学习模型，例如LeNet、VGG、AlexNet、GoogLeNet、CaffeNet和ResNet。在第IV和V节中，我们将根据这种分类法调查最近的对抗性样本研究。

4. 对抗样本攻击模型

在本篇论文之中，作者一共列举了几种生成对抗性样本的代表性方法，下面我将对这些方法做一个自己的学习记录

4.1 L-BFGS Attack

在L-BFGS Attack中，目标函数的输入是原始输入数据和错误分类的标签。错误分类的标签的产生是通过查询模型进行预测得到的，然后通过将其转化为one-hot编码来创建误分类的标签。

接下来，L-BFGS算法被用于最小化目标函数，并对c进行**linear-searching**生成对抗样本。在生成样本时，使用一些约束条件，例如添加噪声以确保对抗样本与原始输入类似。

$$\min_{x'} c\|\eta\| + J_\theta(x', l')$$
$$\text{s.t. } x' \in [0, 1].$$

4.2 FSGM

FGSM是梯度上升的单步（Fast）优化方法，优化方向沿着训练模型梯度下降的反方向。该算法中，假定深度网络容易收到对抗性扰动的主要原因是它们的线性性质，高维空间中的线性行为。如果我们在输入图像中加上计算得到的梯度方向，修改后的图像经过分类网络时的损失值就比修改前的图像经过分类网络时的损失值要大，模型预测对的概率就会变小。

4.3 BIM & ILLCM

在原始的 FGSM 中，攻击者只使用一个梯度来生成对抗性样本，这种攻击方式比较局限。BIM 和 ILLCM 在每个迭代步骤中都计算梯度，根据梯度进行多次更新，通过引入更多的扰动，从而产生更具有挑战性的对抗性样本。

具体来说，BIM 方法中，攻击者通过多次应用 FGSM 来生成对抗性样本。在每一轮迭代中，攻击者利用当前的对抗性样本计算新的梯度，并将其应用于原始数据以生成下一个对抗性样本。这样可以产生更具挑战性的对抗性样本，因为攻击者可以利用初始样本的微小差异来引入更多的扰动。

在 ILLCM 方法中，攻击者在每次迭代中选择网络最不可能产生的目标类别作为目标类别，以此来欺骗网络输出。这种方法更加有效，因为攻击者试图在每次迭代中最大程度地增加目标类别的概率，这样结果会产生更具有挑战性的对抗性样本。

4.4 JSMA

JSMA是一种基于雅可比矩阵的对抗样本攻击方法。它利用了图像中每个像素对模型输出的影响，从而创建对抗样本。

在这个方法中，雅可比矩阵是由神经网络的输出与输入相对应的梯度组成的。使用这个雅可比矩阵可以去导致每个像素对模型输出的影响。作者发现，通过每个样本仅修改4.02%的输入特征，实现了97% 的对抗性成功率。但是，由于其显着的计算成本，此方法运行非常缓慢。

4.5 DeepFool

DeepFool 的基本思想是将原始样本移动到最近的决策边界上，从而产生对抗性样本。它是一种迭代攻击方法，因此可以生成非常强大和具有挑战性的对抗性样本。

具体而言，DeepFool 方法会在决策边界上寻找一个最小扰动向量，该扰动向量将原始样本移到了另一类别的决策区域。这个扰动向量是通过计算网络梯度并沿着梯度方向进行更新生成的。然后，攻击者重复执行这个过程，直到达到预设的最大迭代次数或者满足某些收敛条件。

4.6 CPPN EA Fool

CPPN EA FOOL 的主要流程如下：

准备一个原始的输入图像 x 和一个深度学习模型F(x)。定义 CPPN 神经网络结构，包括输入和输出层，中间的激活函数层以及权重参数和偏置项；通过进化算法来优化 CPPN 网络的权重和偏置参数，以最大程度地欺骗深度学习模型。在每代优化中，对 CPPN 网络生成的对抗样本进行评估，并选择最具欺骗性能力的样本进行下一代优化；重复，直到达到预设的最大迭代次数或者某些收敛条件。返回生成的对抗性图像作为攻击结果。

4.7 C&W's Attack

C&W's Attack 方法使用 L-BFGS（Limited-memory Broyden–Fletcher–Goldfarb–Shanno）算法，通过反复迭代来使得目标函数最小化，并满足上述约束。在每次迭代中，算法会根据当前的扰动向量对深度学习模型进行前向传播和反向传播，并利用梯度信息来更新扰动向量。

C&W's Attack 方法的优点在于通过优化来最小化扰动大小，从而在保证攻击成功率的同时尽可能减小扰动的幅度和可见性，使得攻击成功率和图像质量得到了有效的平衡

4.8 Zeroth-Order Optimization

Zeroth-Order Optimization，是一种零阶优化方法，用于生成对抗性样本。与其他对抗性攻击方法不同，ZOO 不需要直接访问或计算深度学习模型的梯度信息，而是基于模型输出的黑盒反馈信息来生成对抗样本。

Zeroth-Order Optimization 的基本思想是通过构造一个代理模型来对原始模型进行模拟，并根据模型输出的分类结果进行反馈，从而优化损失函数来生成对抗性样本。

4.9 Universal Perturbation

Universal Perturbation 是一种基于迭代算法的对抗性攻击方法，其主要目标是生成一组共性的扰动向量，使得这些扰动向量可以成功地攻击深度学习模型中的任意输入样本。下面我记录下对该方法流程的学习：

a. 随机选取一个未攻击成功的样本，并将其添加到所有已攻击成功的样本列表中。

b. 通过梯度下降或其他优化算法求解一个最小化扰动向量的问题，使得通用扰动向量可以成功地攻击所有已攻击成功的样本，并尽可能地减小扰动向量的幅度。

c. 如果通用扰动向量符合预设的停止条件，则跳出循环；

4.10 One-Pixel Attack

One-Pixel Attack 是一种基于遗传算法的对抗性攻击方法，其主要目标是在保证图像质量不变的情况下，通过修改图像中某个像素的 RGB 值来生成对抗性样本。与其他对抗性攻击方法不同，One-Pixel Attack 只修改了图像中的一个像素，因此攻击效果非常难以察觉。

4.11 Feature Adversary

Feature Adversary 是一种基于特征空间的对抗性攻击方法，其主要目标是通过最小化特征空间中输入样本与目标类别之间的距离来生成对抗性样本。与其他对抗性攻击方法不同，Feature Adversary 在特征空间上操作，而非像素空间上的像素值。

Feature Adversary 的基本思想是通过反向传播算法在特征空间中求解一个最小化目标类别特征与输入样本特征距离的问题，并以此来生成对抗性样本。

4.12 Hot/Cold

本方法我不是很能理解，其大致思路应该为，通过模拟退火算法来控制扰动大小，以便生成高质量的对抗性样本。随着温度的降低，扰动大小也逐渐减小，以使攻击效果更加准确和鲁棒性

$$\arg\min_{\mathcal{H}} \left\| \frac{\overline{x}}{\|\overline{x}\|} - \frac{\overline{\phi(x', x)}}{\|\overline{\phi(x', x)}\|} \right\|$$

4.13 Model-Based Ensembling Attack

Model-Based Ensembling Attack 是一种基于集成模型的对抗性攻击方法，其主要目标是通过结合多个深度学习模型的输出结果来生成具有强大攻击性能的对抗性样本。与非目标对抗性样本相比，在深度模型之间传递目标对抗性样本要困难得多。通过使用基于模型集成的攻击方法，他们可以生成可转移的对抗性样本，以攻击黑盒模型。

4.14 Ground-Truth Attack

Ground-Truth Attack 是一种基于真实标签信息的对抗性攻击方法，其主要目标是生成具有强大攻击性能的对抗性样本，以使深度学习模型的分类结果与真实标签不同。与其他对抗性攻击方法不同，Ground-Truth Attack 利用真实标签信息来制定攻击策略，并生成更高质量的对抗性样本。

5. 对抗样本应用

之前的文章一直是基于一个假设，我们的对抗样本攻击大多部署在计算机视觉和图像分类的相关问题上，这个部分，文章则是关注对抗样本还能用在哪些其他的部分。本处主要关注三个问题。对抗样本在什么情况下应用于新任务？如何在新任务中生成对抗样本？是提出新方法还是将问题转化为图像分类任务并通过上述方法解决。

5.1 Reinforcement Learning

强化学习 DNN已经被用于通过训练原始输入（例如图像）的策略来进行强化学习。如果应用FGSM来攻击深度强化学习网络和算法，被证明可以取得一个较好的结果，Huang's Attack with l1 norm在白盒和黑盒攻击上都取得了十分良好的效能。

5.2 Generative Modeling

在生成模型上，可以对自编码器（AE）的编码器输入中注入扰动来进行一定的攻击。实验者测试了VAE and VAE-GAN在MNIST, SVHN上的性能，实验发现，Latent Attack达到了一个十分优秀的攻击效果

5.3 Face Recognition

基于DNN的人脸识别系统和人脸检测系统由于其高性能而广泛应用于商业产品中，Sharif等人测试了使用eyeglass frames来进行攻击，它们在攻击中使用了上述提到的L-BFGS 攻击方法。实验的结果证明，他们成功地对抗FRS（over 80%）并以高成功率（取决于目标）误导了FRS作为特定的面部。

5.4 Object Detection

Xie等人提出了一种DAG的通用算法，在每次迭代中对前一次迭代进行优化，并且为了应付大数据，作者使用了regional proposal network，同样达到了十分优秀的效果。

5.5 Semantic Segmentation

分割也可被视作是图像领域多维的一个重要任务，实验发现，对抗样本攻击可以在这个方面渠道良好的攻击效果

5.6 NLP

自然语言处理中的许多任务都可以通过对抗样本来攻击。人们通常通过在句子中添加/删除单词来生成对抗性样本，Jia和Liang曾在段落末尾添加了分散注意力的（对抗性）句子，结果发现深度学习模型无法区分段落中微妙但关键的差异，而后，他们添加与问题类似但不与正确答案矛盾的语法句子和加带有任意英语单词的句子可以达到一个较好的干扰效果。

6. 对抗性样本防御

对抗性样本防御策略有以下：（1）反应性：在构建深度神经网络之后检测对抗性样本；（2）预防性：在对抗者生成对抗性样本之前增强深度神经网络的鲁棒性。在本节中，文章讨论了三种反应性对抗措施（对抗检测、输入重构和网络验证）和三种预防性对抗措施（网络蒸馏、对抗训练和分类器增强）。此外，还记录了一个合成防御方法。

6.1 Network Distillation

Ensembling Defenses 是一种基于模型集成的对抗性样本防御方法，其主要目标是通过结合多个深度学习模型的输出结果来提高防御能力，从而有效地缓解对抗性攻击的影响。与其他对抗性样本防御方法不同，Ensembling Defenses 利用多个深度学习模型之间的差异来提高防御能力，并具有较强的可扩展性和实用性。

Ensembling Defenses 的基本思想是使用多个深度学习模型来对输入图像进行分类，并利用它们之间的差异来缓解对抗性攻击的影响。具体来说，该算法包括以下步骤：

- 训练多个深度学习模型，每个模型都有不同的参数和结构。
- 对每个模型，在训练集上进行交叉验证，计算每个图像属于每个类别的概率分布。
- 利用这些概率分布建立一个多分类回归模型，以最小化对抗性损失函数。
- 在测试过程中，对输入图像进行分类，并对多个模型的输出结果进行加权平均或投票决策，以得出最终的分类结果。
- 如果模型集成的分类结果为正常类别，则输出该结果；否则进行其他的对抗性防御方法

6.2 Adversarial (Re)training

Ensembling Defenses其主要目标是通过对深度学习模型进行重新训练来提高其鲁棒性，从而有效地缓解对抗性攻击的影响。与其他对抗性样本防御方法不同，Adversarial (Re)training 利用对抗性样本来重新训练模型，从而提高其鲁棒性和防御能力。

Adversarial (Re)training 的基本思想是使用对抗性样本来重新训练深度学习模型，并利用这些样本来识别和修复原模型中的漏洞。具体来说，该算法包括以下步骤：

- 使用原始数据集来训练深度学习模型。
- 使用对抗性攻击算法生成一组对抗性样本，并将其添加到原始数据集中。
- 使用新的带标签的数据集来重新训练模型，以最小化对抗性损失函数。
- 在测试过程中，对输入图像进行分类，并采用一定策略来检测对抗性样本。
- 如果输入图像为正常类别，则输出该结果；否则进行其他的对抗性防御方法。、

文章的例子说明，对抗性训练增加了MNIST案例下，神经网络在单步攻击（例如FGSM）下的鲁棒性，但在迭代攻击（例如BIM和ILLC方法）下不起作用，在 MNIST 和 ImageNet 数据集上经过对抗性训练的模型对白盒对抗性示例比迁移示例（黑盒）更加稳健

6.3 Adversarial Detecting

Adversarial Detecting 主要目标是识别和过滤出输入图像中的对抗性样本，从而有效地缓解对抗性攻击的影响。与其他对抗性样本防御方法不同，Adversarial Detecting 不需要对原始模型进行修改或重新训练，在原有模型的基础上引入额外的检测步骤，能够有效地提高防御能力。

Adversarial Detecting 的基本思想是使用一种合适的检测方法，对输入图像进行检测，并判断其是否为对抗性样本。具体来说，该算法包括以下步骤：

- 训练一个深度学习模型，用于对输入图像进行分类。
- 使用多种对抗性攻击算法生成一组对抗性样本，并将其添加到原始数据集中。
- 使用训练好的深度学习模型对带有对抗性样本的新数据集进行分类，记录分类的结果。
- 选择一种检测方法并利用对抗性样本和正常样本中的差异，建立一个分类器来识别对抗性样本。
- 在测试过程中，对输入图像进行分类，并使用检测方法来判断其是否为对抗性样本。
- 如果输入图像为正常类别，则输出该结果；否则进行其他的对抗性防御方法。

6.4 Input Reconstruction

Input Reconstruction 是一种对抗性样本防御方法，其主要目标是通过对输入图像进行重建或修复来消除对抗性攻击的影响，从而提高深度学习模型的鲁棒性和准确性。与其他对抗性样本防御方法不同，Input Reconstruction 利用对抗性样本中的扰动信息来恢复原始图像，并将其作为新的输入进行分类。

Input Reconstruction 的基本思想是使用图像重建算法来恢复对抗性样本中的扰动信息，并将其添加到原始图像中，得到新的输入图像。具体来说，该算法包括以下步骤：

- 使用对抗性攻击算法生成一组对抗性样本，并记录其扰动向量。
- 使用图像重建算法对扰动向量进行反演，以还原出原始图像。
- 将重建后的图像添加到原始图像中，得到新的输入图像。
- 对新的输入图像进行分类，并输出分类结果。
- 如果分类结果为正常类别，则输出该结果；否则返回第 2 步进行重建或修复。

6.5 Classifier Robustifying

Classifier Robustifying 通过增强深度学习模型的鲁棒性来有效地缓解对抗性攻击的影响。与其他对抗性样本防御方法不同，Classifier Robustifying 通过改进分类器的架构或调整其参数来提高其鲁棒性，从而提高其抵御对抗性攻击的能力。

Classifier Robustifying 的基本思想是使用一种合适的方法，对深度学习模型进行改进，以加强其鲁棒性和防御能力。具体来说，该算法包括以下步骤：

- 对原始模型的架构或参数进行改变，以提高其鲁棒性和防御能力。
- 使用训练数据集对改进后的分类器进行重新训练，以最小化对抗性损失函数。
- 在测试过程中，对输入图像进行分类，并检测其是否为对抗性样本。
- 如果输入图像为正常类别，则输出该结果；否则进行其他的对抗性防御方法。

6.6 Network Verification

Network Verification 用形式化验证技术来证明深度学习模型的正确性和鲁棒性，从而有效地缓解对抗性攻击的影响。与其他对抗性样本防御方法不同，Network Verification 利用数学证明来保证深度学习模型的准确性和安全性。

Network Verification 的基本思想是使用形式化验证技术来建立一个模型的语义模型，并使用该模型来证明模型的正确性和鲁棒性。具体来说，该算法包括以下步骤：

- 对深度学习模型进行语义建模，将模型转化为一组约束条件。
- 使用定理证明器或 SMT 求解器来证明模型是否存在任何漏洞或不安全的区域。
- 如果模型存在漏洞或不安全区域，则利用这些信息进行改进，并重新进行网络验证。
- 在测试过程中，使用经过验证的深度学习模型进行分类，并检测输入图像是否为对抗性样本。
- 如果输入图像为正常类别，则输出该结果；否则进行其他的对抗性防御方法。

6.7 Ensembling Defenses

Ensembling Defenses 将多个对抗性样本防御方法组合起来，以提高深度学习模型的鲁棒性和准确性。与其他单一的对抗性样本防御方法不同，Ensembling Defenses 组合多种防御方法来抵抗各种类型的攻击。

Ensembling Defenses 的基本思想是通过融合多个不同的对抗性样本防御方法来提高深度学习模型的鲁棒性和准确性。具体来说，该算法包括以下步骤：

- 选择多种不同的对抗性样本防御方法，并对每种方法进行优化和调整。
- 对训练数据集分别使用这些方法进行训练，得到多个不同的深度学习模型。
- 在测试过程中，使用所有的深度学习模型对输入图像进行分类，并对结果进行融合。
- 如果融合后的结果为正常类别，则输出该结果；否则进行其他的对抗性防御方法。

7. 展望

7.1 Transferability

传递性是指对抗样本可在不同的机器学习模型和数据集之间转移。攻击者可以利用对一个替代模型生成的对抗样本来攻击原真实模型，尤其是当原真实模型和训练数据集不可访问时。而防御者可以通过限制对抗样本的传递性来防御这种黑盒攻击。

而传递性三个级别，从易到难分别是：

（1）在不同训练集上的同一模型之间的传递；

（2）在训练于相同任务的不同模型之间的传递；

（3）在不同任务的深度神经网络之间的传递。各种研究表明，对抗样本的传递性是存在的，攻击者可以通过多种方式利用它来攻击目标模型

7.2 Existence of Adversarial Examples

对抗样本是否是深度神经网络的固有属性，目前尚无定论。数据不完备、模型能力、模型的鲁棒性可能是其中的原因。许多应用程序中也出现了对抗样本，但并非所有应用程序都可以使用相同的方法来处理。目前的研究主要集中在图像分类任务上，还没有探讨不同应用之间的关系和是否存在一种通用的攻击/防御方法适用于所有应用。

7.3 Robustness Evaluation

为了满足安全关键领域中对DNN鲁棒性的要求，需要评估DNN模型的鲁棒性，也就是要建立一种方法来评估DNN模型的鲁棒性，并提供攻击和防御的基准平台。目前存在的问题是评测工具缺少防御策略。未来研究方向包括解决这些问题，并建立一个通用的鲁棒性评估方法