Adversarial Examples: Attacks and Defenses for Deep Learning 论文阅读报告

- 姓名：蒋奕
- 学号：JY

这篇文章主要概述了对抗样本攻防的一些概念。偏向综述性质。

introduction

深度神经网络对于**精心设计**好的输入样本是脆弱的，这种样本就被称为**对抗样本**

生成对抗样本的方法

Threat model

分为：

Adversarial Falsification

- False positive attacks

生成一个反例，让模型误认为正例

- False negative attacks

生成一个正例，让模型误认为反例

Adversary's Knowledge

- White-box attacks

假定攻击者可以访问他们正在攻击的神经网络模型的结构和参数

- Black-box attacks

假定攻击者不能访问他们正在攻击的神经网络模型的结构和参数，只知道模型的输出

Adversarial Specificity

- Targeted attacks

是神经网络将生成的对抗样本分类成一个特定的类，通常最大化目标类的可能性。

- Non-targeted attacks

只要干预神经网络判定的label与原label不同即认为攻击成功。

Attack Frequency

- One-time attacks

只需优化一次即可生成对抗样本

- Iterative attacks

迭代的进行优化来生成对抗样本，效果更好，计算时间多。

Perturbation

分成

Perturbation scope

- Individual attacks

对于每一个不同的原始输入，加入不同的扰动，现有方法基本这样

- Universal attacks

对整个数据集应用同一个扰动，容易实现

Perturbation limitation

- Optimized Perturbation

把扰动作为优化问题的目标。为了最小化扰动，要求扰动小到极致

- Constraint Perturbation

将扰动作为优化问题的约束条件，只要求扰动小于设定的阈值就行了。

Perturbation measurement

- lp measure

1. l_0 ： 计算对抗样本中改变的像素的数目
2. l_2 ： 计算对抗样本与原始样本的欧氏距离
3. l_∞ ： 计算对抗样本中所有像素的最大改变值

- Psychometric perceptual adversarial similarity score (PASS)

符合人类的感知

Perturbation

Datasets

- MNIST, CIFAR-10 简单且数目少，所以比较容易去攻击和防御
- ImageNet是最好的数据集

Victim Models

LeNet, VGG, AlexNet, GoogLeNet, CaffeNet, and ResNet.

METHODS FOR GENERATING ADVERSARIAL EXAMPLES

L-BFGS

方法如下：

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

为了找到适合的常量C，L-BFGS算法通过线性搜索C > 0的所有情况，找到C的近似值。实验表明，生成的对抗样本也可以推广到不同的模型和不同的训练数据集中。但是使用的线性搜索方法代价很高并且是不切实际

Fast Gradient Sign Method (FGSM)

在每一个像素上仅仅执行了一步沿着梯度符号方向上的梯度更新。扰动形式为

$$\eta = \epsilon sign(\nabla_x J_\theta(x, l)),$$

扰动通过反向传播过程计算。高维神经网络的线性部分无法抵抗对抗样本。因此，一些正则化可以被用于深度神经网络。

FGSM是一种无目标攻击。

FGSM的改进方法：

Fast Gradient Value method

用

$$\eta = \nabla_x J(\theta, x, l)$$

替换原来的扰动。由于该方法没有常量，所以会生成具有较大差异的图像。

One-step Target Class Method (OTCM)

单步攻击是很容易迁移但是也很容易防御。将动量的思想放入FGSM中，来迭代的生成对抗样本。每次迭代的梯度计算公式为：

$$\mathbf{g}_{t+1} = \mu \mathbf{g}_t + \frac{\nabla_x J_\theta(x'_t, l)}{\|\nabla_x J_\theta(x'_t, l)\|},$$

该方法通过引入动量提高了攻击的有效性，通过使用单步攻击和压缩方法提高了其迁移性。

将FGSM拓展到了目标攻击，其公式如下：

$$x' = x - \epsilon sign(\nabla_x J(\theta, x, l')).$$

RAND-FGSM

因为gradient masking，FGSM对白盒攻击的鲁棒性更好。

提出了随机FGSM，在更新对抗样本时，对样本增加随机值来进行对抗样本的防御。公式如下:

$$\begin{aligned} x_{tmp} &= x + \alpha \cdot sign(\mathcal{N}(\mathbf{0}^d, \mathbf{I}^d)), \\ x' &= x_{tmp} + (\epsilon - \alpha) \cdot sign(\nabla_{x_{tmp}} J(x_{tmp}, l)) \end{aligned}$$

Basic Iterative Method (BIM)

之前的方法假定数据能够直接被送入神经网络中。然而，在很多情况下只能依靠一些设备来传送数据。对FGSM做了一点点小的改变，使用了多次迭代。在每次迭代过程中限制像素值来避免过大。

$$Clip_{x,\xi}\{x'\} = \min\{255, x + \xi, \max\{0, x - \epsilon, x'\}\}.$$

限制了在每次迭代过程中的对抗样本的改变的大小。对抗样本通过多次迭代产生:

$$\begin{aligned} x_0 &= x, \\ x_{n+1} &= Clip_{x,\xi}\{x_n + \epsilon sign(\nabla_x J(x_n, y))\} \end{aligned}$$

Iterative Least-Likely Class Method (ILLC)

将BIM拓展到了目标攻击，使用与原始标签最不像的类作为目标，通过最大化交叉熵损失函数的方法来实现。

$$
\begin{aligned}
x_0 &= x, \\
y_{LL} &= arg\min_y\{p(y|x)\}, \\
x_{n+1} &= Clip_{x,\epsilon}\{x_n - \epsilon sign(\nabla_x J(x_n, y_{LL}))\}.
\end{aligned}
$$

用一张手机拍摄的精心制作的图片欺骗了神经网络。

FGSM对光转化的鲁棒性更好，而迭代性方法不能够抵挡光转化。

Jacobian-based Saliency Map Attack (JSMA)

计算了原始样本x的Jacobian矩阵，计算方法如下：

$$
J_F(x) = \frac{\partial F(x)}{\partial x} = \left[\frac{\partial F_j(x)}{\partial x_i}\right]_{i \times j}.
$$

发现样本x的输入特征会对输出有着最显著的影响。一个被设计好的小的扰动就能够引起输出的较大改变，一个小的特征的改变就能够欺骗神经网络。但是由于计算Jacobian矩阵很慢，因此这个方法运行的太慢。

DeepFool

寻找从原始输入到对抗样本决策边界最近的距离。为了克服高维中的非线性，用线性估计执行迭代攻击。从仿射分类器开始，一个仿射分类器的最小扰动是到分类超平面的距离。一个仿射分类器f的扰动可能是

$$
\eta^*(x) = -\frac{f(x)}{\|w\|^2}w.
$$

如果分类器f是二分类器，使用一种迭代的方法来估计扰动，并考虑在每次迭代过程中f关于xi是线性的。最小的扰动的计算方法如下：

$$
\begin{aligned}
\arg\min_{\eta_i} \quad & \|\eta_i\|_2 \\
s.t. \quad & f(x_i) + \nabla f(x_i)^T \eta_i = 0.
\end{aligned}
$$

- 这个结果可以通过找到最近的超平面的方法拓展到多分类任务。这也可以拓展到更普遍的lp正则化。
- 相比于FGSM和JSMA，DeepFool提供了更少的扰动。
- 相比于JSMA，DeepFool降低了扰动的强度而不是选中的特征的数目。

CPPN EA Fool

在该方法中，一个不能够被人类识别是什么种类的对抗样本可以被深度神经网络以很高的置信度分成一个类。（假正例攻击）

- 使用进化算法（EA）来生成对抗样本。
- 为了解决使用EA算法的多分类任务，使用了multi– dimensional archive of phenotypic elites MAP–Elites
- 用两种不同的方法编码图片，直接编码和间接编码。
- 在每次迭代过程中MAP–Elites就像普通的进化算法一样。

C&W's Attack

对大多数现存的防御方法都是有效的。

- 首先定义了一个新的目标函数g，同时满足

$$\begin{aligned} \min_{\eta} \quad & \|\eta\|_p + c \cdot g(x+\eta) \\ s.t. \quad & x+\eta \in [0,1]^n, \end{aligned}$$

距离和惩罚条件可以得到更好的优化。通过实验，7个目标函数中有效的函数之一是：

$$g(x') = \max(\max_{i \neq l'}(Z(x')_i) - Z(x')_t, -\kappa)$$

- 相比于使用box约束来寻找最小扰动的L–BFGS攻击方法，引入了一个新的变量w避免box约束，其中

$$\eta = \frac{1}{2}(\tanh(w)+1) - x$$

  在深度学习中通用的优化器，被用来生成对抗样本。但是，如果‖η‖和g(x+η)的梯度不在同一范围，那么在梯度迭代搜索过程中找到一个合适的常量c很困难

- 基于距离测量方法提出了三种攻击方式

1. l_0 攻击

迭代进行

在每次迭代过程中不重要的像素被移除。像素的重要性取决于l2距离的梯度。如果剩下的像素不能够生成一个对抗样本的时候，迭代停止

2. l_2 攻击

蒸馏网络不能抵御，描述为：

$$\min_w \|\frac{1}{2}(\tanh(w)+1)\|_2 + c \cdot g(\frac{1}{2}\tanh(w)+1)$$

3. l_∞ 攻击

迭代攻击，在每次迭代的过程中用一个新的惩罚方式取代l2条件：

$$\min \quad c \cdot g(x+\eta) + \sum_i [(\eta_i - \tau)^+].$$

Zeroth Order Optimization (ZOO)

不需要梯度，可直接的部署到黑盒攻击中，而不需要模型迁移。

修改g函数为：

$$g(x') = \max(\max_{i \neq l'}(\log[f(x)]_i) - \log[f(x)]_{l'}, -\kappa),$$

并且使用对称差商估计梯度和Hessian：

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x+he_i) - f(x-he_i)}{2h},$$
$$\frac{\partial^2 f(x)}{\partial x_i^2} \approx \frac{f(x+he_i) - 2f(x) + f(x-he_i)}{h^2};$$

不需要接触要被攻击的深度学习模型但需要昂贵的代价来查询和估计梯度。

Universal Perturbation

普适性攻击方法：找到一个普适性的扰动向量满足：

$$\|\eta\|_p \leq \epsilon,$$
$$\mathcal{P}\big(x' \neq f(x)\big) \geq 1 - \delta.$$

在每次迭代中，使用DeepFool方法为每一个输入数据获得一个最小的样本扰动，并且更新该扰动到总扰动中去。直到大多数的样本攻击成功，迭代才停止。

One Pixel Attack

为了避免感知测量的问题，通过仅仅修改了一个像素，生成了对抗样本。优化问题变成了：

$$\begin{aligned} \min_{x'} \quad & J(f(x'), l') \\ s.t. \quad & \|\eta\|_0 \leq \epsilon_0, \end{aligned}$$

- 应用了微分进化来找到优化解。
- 可以用在不可微的目标函数上。

Feature Adversary

通过最小化内部神经网络层而不是输出层的表示距离来执行了一次目标攻击。该问题可以描述为：

$$\begin{aligned} \min_{x'} \quad & \|\phi_k(x) - \phi_k(x')\| \\ s.t. \quad & \|x - x'\|_\infty < \delta, \end{aligned}$$

- 一个固定的δ值对人类的感知来说已经足够了。
- 使用了L-BFGS-B来解决优化问题。对抗性图像更自然，更接近内部层的目标图像。

Hot/Cold

定义了一个新的标准，Psychometric Perceptual Adversarial Similarity Score (PASS)，衡量与人类的明显相似性。

Hot/Cold，忽略了像素上的不明显的差别，并用带有PASS的lp距离取代。PASS包括两个步骤，首先，将修改后的图片与原图片对齐，之后，测量两个图片的的相似度。

为了生成各种各样的对抗样本，作者定义了目标标签l'为hot 类，原始标签l为cold类。在每次迭代中，它们都会移向目标（热）类，同时远离原始（冷）类。他们的结果表明，生成的对抗性例子与FGSM相似，并且具有更多的多样性。

Natural GAN

用GAN作为方法的一部分来生成图片和文本的对抗样本

首先训练了一个WGAN模型，生成器G将随机噪声映射到输入域

还训练了一个“反相器”L将输入数据映射到密集的内部表示。

Model-based Ensembling Attack

提出了Model-based Ensembling Attack用来目标攻击。

相比于无目标攻击，目标攻击在深度模型上迁移是更加困难的。

使用Model-based Ensembling Attack，可以生成可转移的对抗样本来攻击一个黑盒模型。

Ground-Truth Attack

网络验证(Network Verificatio)总是检查对抗样本是否违反深度神经网络的属性，以及是否存在示例在一定距离内改变标签。

Ground-Truth Attack执行了一个二值搜索，并且迭代的调用Reluplex来找到一个最小扰动的对抗样本。

对抗防御方法

proactive

Network Distillation

蒸馏网络最初是被用于减小深度网络的尺寸，通过将大网络的知识转移到小网络中。在论文中，作者提到，网络蒸馏可以提取深度网络的知识，并能够提高鲁棒性。

Adversarial (Re)training

让神经网络更加鲁棒。在训练阶段引入了对抗样本。他们在训练的每一步生成对抗样本，然后把他们加入到训练集中。对抗训练可以为神经网络提供正则化，并且提高了其准确率。

对抗训练可以抵御单步攻击，但不能抵御迭代攻击。

Classifier Robustifying

由于对抗样本的不确定性，Bradshaw等人利用了贝叶斯分类来建立了一个更加鲁棒的神经网络。高斯变换（RBF核）也被用来提供不确定检测。提出来的神经网络叫做Gaussian Process Hybrid Deep Neural Networks (GPDNNs)。

reactive

Adversarial Detecting

在测试阶段来检测对抗样本。

Input Reconstruction

对抗样本可以通过重建的方式变换成干净数据。在转变后，对抗样本并不会影响深度模型的预测。

一个去噪的自动编码网络被训练用来压缩对抗样本成原始样本来消除对抗性扰动。通过两种方法来重建对抗样本，添加高斯噪声和用自动编码器压缩他们即MagNet的plan B。

Network Verificatio

验证深度网络的属性是一个对抗防御的方式，因为它可以检测出新的不可见的攻击。网络验证会检查一个神经网络的属性，输入是否违反或是否满足属性。

提出了一个使用ReLU激活函数的验证方法，叫做Reluplex，使用Satisfiability Modulo Theory (SMT 可满足性模理论)来验证神经网络，即，在一个小扰动下，没有现有的对抗样本可以让神经网络进行错误分类。

Ensembling Defenses

由于对抗样本的多样性，因此可以使用多种防御策略一起运行来进行防御。

MagNet包含一个或多个检测器以及一个重建器，分别为Plan A和Plan B。检测器用来找到离分类边界比较远的对抗样本。在一篇论文中，测量了在输入和编码输入的距离，以及（输入和编码输入的）softmax输出概率散度。对抗样本具有大的距离和概率散度。为了解决距离边界比较近的对抗样本，MagNet使用了一个基于自动编码器的重建器，重建器会将对抗样本映射到合理的样本。

挑战与讨论

迁移性

- 在机器学习与深度学习网络之间的迁移性。对抗样本可以在不同的参数、机器学习模型的训练数据集之间进行，甚至可以跨不同的机器学习技术进行。
- 在复杂模型和大数据集上的目标攻击和无目标攻击的迁移性。无目标攻击的对抗样本更具有迁移性。不同模型的决策边界彼此吻合得很好。Model-Based Ensembling Attack，以创建可迁移的目

标攻击的对抗样本。

- 到模型决策边界的距离平均大于同一方向上两个模型边界之间的距离。这可以解释对抗样本的迁移性存在的原因。

对抗样本的存在性

- 数据不完整。一个假设是，对抗样本是测试数据集中的角落案例，具有低概率性和低覆盖性。通过训练一个PixelCNN，发现对抗性例子的分布不同于干净的数据。即使对于简单的高斯模型，鲁棒的模型也可能比“标准”模型更复杂，需要更多的训练数据。
- 模型容量。对抗样本不仅仅是深度网络的一种现象，而且是所有分类器的一种现象。论文结果表明，在线性情况下，当决策边界接近训练数据的流形时，存在对抗实例。当然，也有不同的意见，有篇论文认为，对抗样本是因为分类器对特定任务的低灵活性，线性并不是一个合理的解释。还有论文表示，是因为稀疏不连续性导致分类器不稳定而产生的对抗样本。
- 没有鲁棒的模型。有论文表示深层神经网络的决策边界本质上是不正确的，不能检测出语义对象。

鲁棒性

- 一种评价深层神经网络鲁棒性的方法。许多深层神经网络计划部署在安全关键环境中。仅防御现有攻击是不够的。Zero-day攻击对深层神经网络的危害更大。需要一种评估深层神经网络鲁棒性的方法。
- 一个攻击和防御的基准平台。 大多数攻击和防御在没有公开代码的情况下描述了它们的方法，也没有说方法中使用的参数。这给其他研究人员复制他们的解决方案并提供相应的攻击/防御带来了困难。
- 稳健性评估的各种应用。类似于各种应用的对抗实例的存在，广泛的应用使得很难评估深度神经网络体系结构的鲁棒性。