# 深度学习对抗攻击算法之FGSM

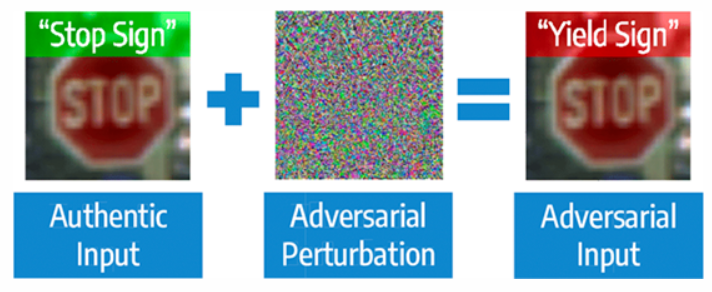
班级：30141806 学号：1120182525 姓名：梁瑛平

### 对抗攻击简介

当前网络安全领域攻击与防御的协同进化如火如荼。像人工智能（AI）和机器学习（ML）这种先进的技术同时为恶意的攻击者也带来了攻击技术演进的机会。简单来看，对网络安全的需求比以往任何时候都更加重要。AI/ML工具在帮助抗击网络犯罪方面可能走了很长一段路，但是这些技术并非无所不能，也会被恶意黑客利用。人工智能将致力于极大地提高网络安全性，但黑客也可将其用于网络犯罪活动，这是对网络安全的真正威胁。AI可以有效地分析用户行为，推导模式并识别网络中的各种异常或不正常情况。有了这些数据，可以快速轻松地识别网络漏洞。反之，现在依赖于人类智能的职责将易于受到模仿合法的基于AI算法的恶意程序的攻击。天地和兴认为，一些企业正在热衷于将其基于AI/ML的概念或产品推向市场。但AI/ML的局限性，导致他们可能会忽略算法正在产生错误或虚假的安全感。

对于人类而言，仅仅通过所接受到的视觉信息并不能完全帮助我们做出正确、迅速的判定，还需要结合我们的生活经验做出相应的反应，以确定哪些信息是真实可靠的，而哪些信息是虚假伪造的，从而选取最适合的信息并作出最终决策。

尽管深度学习在很多计算机视觉领域的任务上表现出色，Szegedy等人第一次发现了深度神经网络在图像分类领域存在有意思的弱点。他们证明尽管有很高的正确率，现代深度网络是非常容易受到对抗样本的攻击的。这些对抗样本仅有很轻微的扰动，以至于人类视觉系统无法察觉这种扰动（图片看起来几乎一样）。这样的攻击会导致神经网络完全改变它对图片的分类。此外，同样的图片扰动可以欺骗好多网络分类器。这类现象的深远意义吸引了好多研究员在对抗攻击和深度学习安全性领域的研究。



### 对抗攻击方式

按照攻击对象，对抗攻击可以分为两类：白盒攻击和黑盒攻击。

白盒攻击（White-box Attacks）指的是攻击者已知模型内部的所有信息和参数，基于给定模型的梯度生成对抗样本，对网络进行攻击。

而黑盒攻击（Black-box Attacks）指的是当攻击者无法访问模型详细信息时，白盒攻击显然不适用，黑盒攻击即不了解模型的参数和结构信息，仅通过模型的输入和输出，生成对抗样本，再对网络进行攻击。

现实生活中相应系统的保密程度还是很可靠的，模型的信息完全泄露的情况也很少，因此白盒攻击的情况要远远少于黑盒攻击。但二者的思想均是一致的，通过梯度信息以生成对抗样本，从而达到欺骗网络模型的目的。

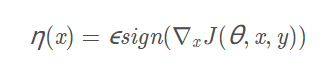
### 对抗攻击经典算法——FGSM

Goodfellow等人提出一种经典的对抗样本生成方法——FGSM 算法。该算法是一种无目标攻击算法，通过在原有样本的基础上计算得到一个扰动，使得加上扰动后的样本被分类到错误的类别。为了方便地表示FGSM算法，定义扰动后的样本为 x’=x+r，其中 x 表示原样本， r 为扰动。显然，对于可区分的类别来讲，当满足||r||\_∞ < e（e表示足够小的值），我们希望分类器将 x 和 x’ 识别为相同的类别。

算法假设神经网络由于过于线性的缘故无法抵抗线性对抗扰动，譬如LSTMs（Hochreiter & Schmidhuber, 1997），ReLUs（Glorot et al., 2011）和maxout 网络（Goodfellow et al., 2013）等，为了便于优化，都采用了非常线性的方式，线性的本质使得通过分析线性模型的扰动可以实现对神经网络的攻击。设 p 表示模型的参数，x 表示模型的输入，y 表示输入 x 时模型对应的输出，J(p, x, y) 表示模型训练时的损失。此时可以将当前 p 下的损失函数线性化，从而得到最优的最大模扰动。

**非线性模型：**

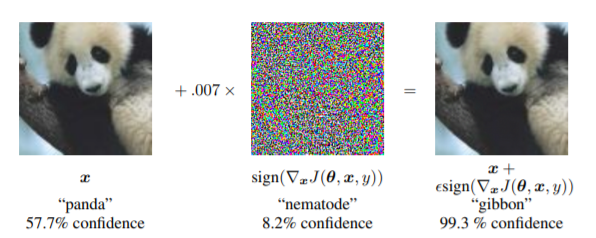
LSTM、ReLU、maxout网络都是线性的，sigmoid也大多工作在线性区域，所有这些具有线性性质的模型都能以下FGSM方法简单获得对抗样本：



这很容易理解：在输入空间中把样本朝着增加成本函数的方向移动。这一方案在实验中也取得了很好的效果。

**算法假设：**

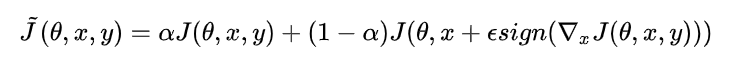
而非线性模型的线性扰动就是一个非线性微分线性化的过程，利用梯度下降的方式来实现。基于此，作者提出了一个快速生成对抗样本的方法，即fast gradient sign method(FGSM).方法的实现:假设一个模型参数θ，x是输入，y是标签（目标输出），损失函数J(θ, x, y)，可以在θ的附近线性化这个损失函数获得一个最佳正则限制扰动，η = esign (∇xJ(θ, x, y)) .梯度可以通过反馈的方式得到。一组实现如下图所示：



而实验表明，FGSM这个简单廉价的算法确实可以生成对抗样本，这也证明了对抗样本是由线性特性引起的。

**深度网络的对抗训练：**

作者在文中表明，相比于线性模型，深度网络至少可以在训练网络过程中来抵御对抗扰动攻击。文章给一种利用FGSM进行对抗训练的方法：



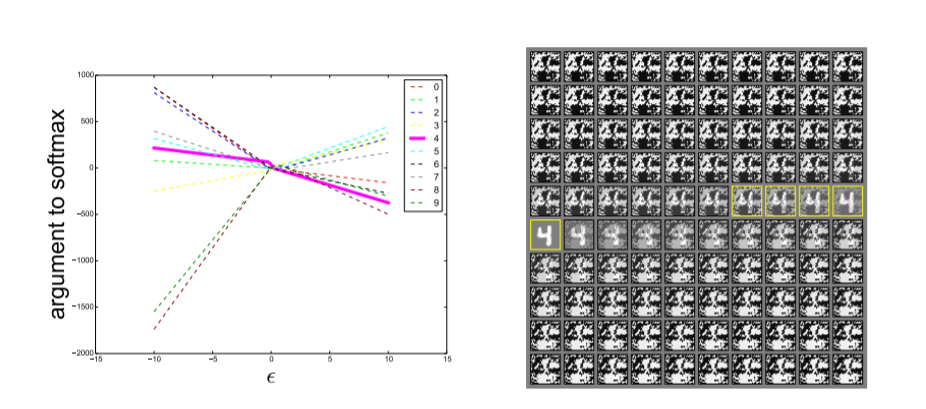
这种对抗训练的方法意味着在训练过程中不断更新对抗样本，从而使得当前模型可以抵御对抗样本。但是作者表示在训练集上对抗训练的错误率error rate没有达到过0%，作者主要从以下两个方面解决：

* 增大模型，即使用1600个unit代替240个unit
* 在validation set上也使用early stopping算法

文章表明，在不进行对抗训练的情况下，模型识别FGSM攻击方法生成样本的错误率是89.4%，但是通过对抗训练，同样的模型识别对抗样本的错误率下降到17.9%作者还探讨了在模型中加零均值零方差的实验，实验表明，这种方式对于抵抗对抗样本效果不好。

**对抗样本泛化的原因：**

很多论文都表明，对抗样本具有可迁移性。具体来说，在一个特定模型上产生的对抗样本通常也容易被其他模型误分类，即使这些模型的结构不同或者模型在不同的训练集上训练。甚至，不同的模型对对抗样本误分类的结果相同！作者表明，非线性或者过拟合的假设不能解释上述的现象，即，为什么拥有无限能力的极度非线性模型会以相同的方式标注数据分布点？在本文提出的线性解释下，作者认为对抗样本在广泛的子空间存在。



上图表明，在不同的参数下，可以看到FGSM可以在一维的连续子空间内产生对抗样本，而不是特定的区域。这就解释了为什么对抗样本特别多，以及对抗样本可迁移性存在的原因。

另外，为了解释为什么不同的分类器将对抗样本误分类到同一个类，作者假设目前的方法训练神经网络都类似于在同一个训练集上学习的线性分类器。由于机器学习算法的泛化能力，所以线性分类器可以在训练集的不同子集上训练出大致相同的分类权重。底层分类权重的稳定性反过来又会导致对抗样本中的稳定性。

### 对抗攻击实验

**基准数据集：**

Stanford Dogs数据集包含来自世界各地的120种犬种的图像。该数据集是使用ImageNet中的图像和注释方法构建的，通常被用于细粒度图像分类的任务。细粒图像分类这是一个具有挑战性的任务，因为某些犬种在颜色和年龄不同的情况下具有接近相同的特征。

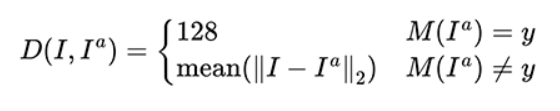
* 分类数量：120
* 图片数量：20,580

**对抗对象：**

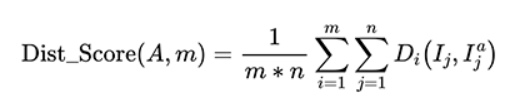
对抗的模型对象为5个使用Stanford Dogs数据集训练的Target Model 。一个是与初赛相同的ResNeXt50白盒模型（白盒），一个是人工加固的模型（灰盒），另外三个均为黑盒模型，其中包括由AutoDL技术训练的模型。

**评估标准：**

对每个生成的对抗样本，会采用m个（m初赛为3复赛为5）基础防御模型对该样本进行预测，并根据识别结果计算相应的扰动量，具体距离度量公式如下：

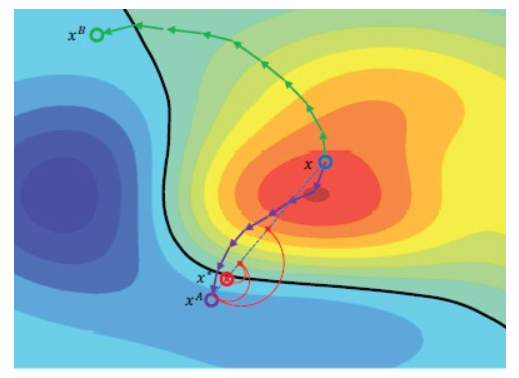


其中M表示防御模型，y表示样本I的真实标签。如果防御算法对样本识别正确，此次攻击不成功，扰动量直接置为上限128。如果攻击成功，计算对抗样本和原始样本的扰动量，采用平均L2距离。每个对抗样本都会在m个防御模型上计算扰动量，n代表样本个数，最后对所有的扰动量进行平均，做为本次攻击的整体距离得分，得分越小越好。具体计算公式如下：



**算法实践：**

为了攻击黑盒模型，我在本地训练了一个加固模型，作为灰盒模型的逼近。训练加固模型涉及到训练集的选取和训练方法的选取。 训练集的构成主要有两部分，第一部分我采用不同的方法攻击初赛的白盒模型(此白盒模型与灰盒模型具有相同的网络结构)，将生成的n个样本集作为训练集的一部分，思路如图3.2所示。这样基于的假设是，不同的攻击方法生成的对抗样本在真实的灰盒模型表现上会有差异，有些图片依然能被灰盒模型正确识别。将n个样本集集合，就可以构建出完全将灰盒攻击成功的图片集。大粒度 此方法受启发于[3]，论文中采取双路寻优，一路采用常规方法梯度上升，如图中绿线所示。另一路先采取梯度下降到达这一分类局部最优再进行梯度上升，以期找到更快的上升路径，如图中蓝线所示。



**添加高斯噪声：**

此方法受启发于[4]，论文作者认为攻击模型的梯度具有噪声，损害了迁移能力。论文作者用一组原始图片加噪声后的梯度的平均代替原来的梯度，效果得到提升。

**实验结果：**

使用上述方法后，我的结果在95-96分之间波动，为进一步提升成绩，我选用最高分96.53分图片进行后处理。后处理方法为：将攻击后的图片与原图片进行对比，对一定阈值以下的扰动进行截断。



### 参考文献

1. Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
2. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
3. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. Ieee Access, 2018, 6: 14410-14430.
4. Shi Y, Wang S, Han Y. Curls & whey: Boosting black-box adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6519-6527.