白盒攻击与黑盒攻击实验报告

马江岩

2021 年 4 月 30 日

摘要

本次实验，我对在 CIFAR-10 数据集上训练的 PreActResNet-18 神经网络分别进行了 FGSM 攻击、PGD 攻击、CW 攻击和 ZOO 攻击，并探索了部分参数对攻击效果的影响。在规定的参数下，我使用不同的攻击方法取得的成功率如表 1 所示。

表 1: 不同攻击方法的成功率

攻击方法	成功率
FGSM	90.0%
PGD	100.0%
CW	94.5%
ZOO	100.0%

1 FGSM 攻击

1.1 原理

FGSM 攻击 [1] 的原理的表达式如下：

$$\mathbf{x}' = \mathbf{x} + \varepsilon \cdot \operatorname{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}, y)\right)$$

其中，$\mathbf{x}$ 表示原图像，y 表示原图像的标签，$J(\mathbf{x}, y)$ 表示损失函数，ε 为参数，表示扰动的最大范围，$\mathbf{x}'$ 表示生成的对抗样本。可以看出，FGSM 攻击的原理，就是沿着损失函数梯度上升的方向，将原图像的数据变化 ε。

1.2 实验配置及结果

在实验中，我取 $\varepsilon = 0.031$，使用 SoftMarginLoss 损失函数，获得了 90.0% 的攻击成功率。实验在 Google Colab[1]上完成。在不同的 ε 下，FGSM 攻击中生成的部分对抗样本如图 1 所示。

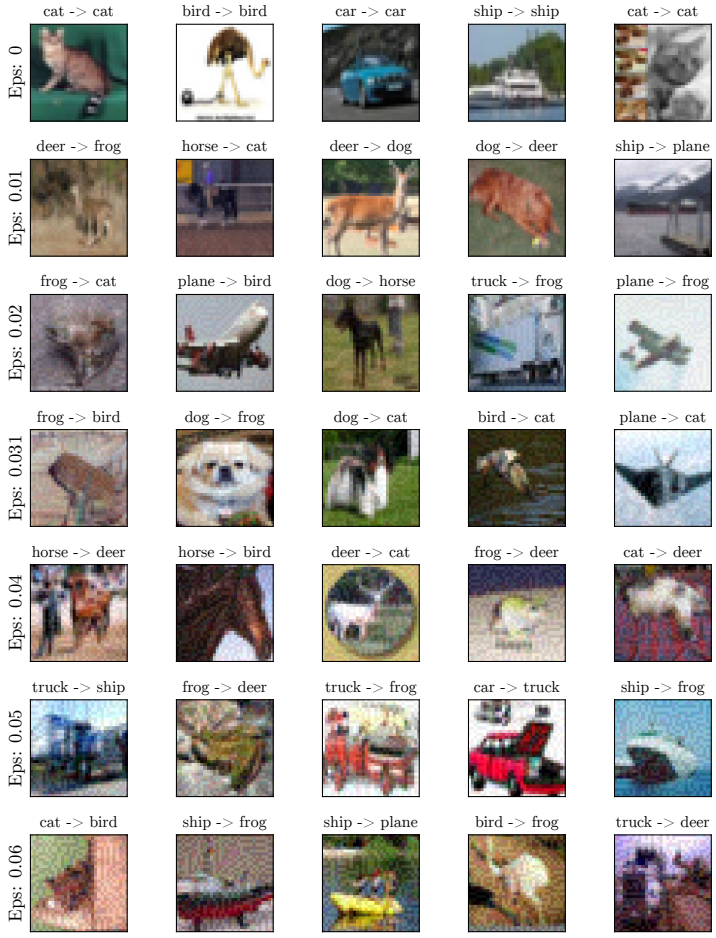

图 1: 使用 FGSM 攻击生成的部分对抗样本。

[1] https://colab.research.google.com/

1.3 扰动范围和损失函数对成功率的影响

在实验中，我在不同的 ε 下，尝试换用不同的损失函数，评估 FGSM 攻击的成功率。所得到的结果如图 2 所示。

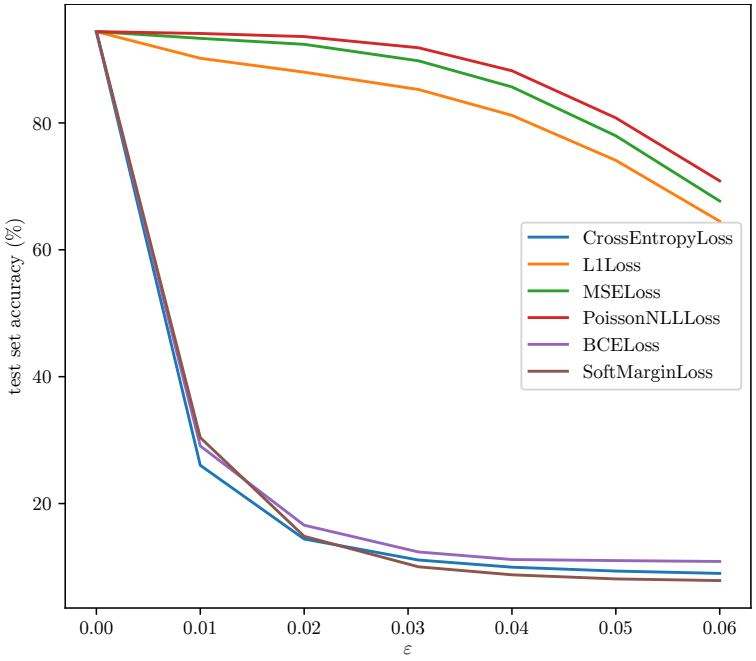

图 2: 不同的 ε 和损失函数对测试集准确率的影响。

首先可以看出，随着 ε 的增大，神经网络在生成对抗样本后的测试集上的分类准确率是逐渐降低的。这是因为，对原图像的扰动越大，生成的对抗样本就更可能越过决策边界，攻击的成功率也就越高。但随着 ε 的增大，对抗样本与原图像的差距也会越来越大，越容易被人眼所察觉到，这会降低生成的对抗样本的质量（这从图 1 中可以看出）。因此，我们应当选择合适的 ε，在攻击成功的基础上，尽可能减小对原图像的扰动。

另外，使用不同的损失函数，得到的测试集准确率也有明显的差异。一般来说，我们在攻击时，选择与模型训练时所用的相同的损失函数，更容易得到较高的攻击成功率。这是因为，我们采用与神经网络训练时相似的参数条件，能够更准确地计算出模型的梯度和决策边界，以便我们进行攻击。我所攻击的神经网络采用的是 CrossEntropyLoss 损失函数，如图 2 所示，在 CrossEntropyLoss 损失函数下，模型的测试集准确率更低；而在诸如 MSELoss、L1Loss 损失函数下，模型的测试集准确率明显更高。但同时我们也注意到，在其他的一些损失函数上，FGSM 攻击也能取得很好的效果；甚至在 SoftMarginLoss 损失函数下，模型的测试集准确率

比在 CrossEntropyLoss 损失函数下还要低。这或许与不同损失函数的计算方式有关，某些损失函数之间具有一定的相似性，在不同于神经网络所用的原损失函数的损失函数下，有时我们也能取得较好的攻击成功率。所以，我们在进行 FGSM 攻击时，应当优先选择与训练神经网络时相同的损失函数，也可以去尝试其他的损失函数。

2 PGD 攻击

2.1 原理

PGD 攻击 [2] 是一种迭代攻击，其原理的表达式如下:

$$\mathbf{x}_{t+1} = \prod_{\mathbf{x}+S} \left(\mathbf{x}_t + \alpha \cdot \operatorname{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}, y)\right)\right),$$

其中，$\mathbf{x}_t$ 表示经过 t 次扰动后所得到的对抗样本。每次迭代后，我们都将扰动裁剪到规定的范围之内。可以看出，PGD 攻击其实是 FGSM 攻击的迭代版，每次迭代都可以看作一次 FGSM 攻击。

2.2 实验配置及结果

在实验中，我取 $\varepsilon = 0.031$，$\alpha = 2/255$，迭代次数为 10 次，使用 CrossEntropyLoss 损失函数，获得了 100.0% 的攻击成功率。同时，我还探究了不同的 ε 和 α 对攻击成功率的影响，见下一节。

2.3 扰动范围和步长对成功率的影响

在不同的 ε 和 α 下，模型在测试集上的分类准确率如图 3 所示。

首先，ε 对测试集准确率的影响是与 FGSM 攻击的情况相似的。ε 越大，对原图像的扰动越大，攻击成功率也就越高；但这也会导致对抗样本与原图像的差别越大，越容易被人眼所察觉。

对于 α，由于 ε 较大时 PGD 攻击准确率大多都能达到接近 100% 的水平，因此我们不妨来研究 $\varepsilon = 0.01$ 时的情况。随着 α 的增大，我们注意到，测试集准确率呈先减小后增加的趋势。这是因为，当 α 较小时，对图像的扰动幅度较小，攻击的成功率也就较低，此时增大 α 能够有效地提高攻击的成功率。但当 α 足够大时，我们再增大 α，成功率反而会下降。这是因为我们对图像的扰动限定在 ε 的范围内，当 α 足够大时，对图像的扰动幅度已经不会再增加。但是步长过大，反而会使损失函数难以上升到最大值，收敛速度慢。因此步长 α 不是越大越好，我们应根据情况和 ε 的大小，选择合适的 α。

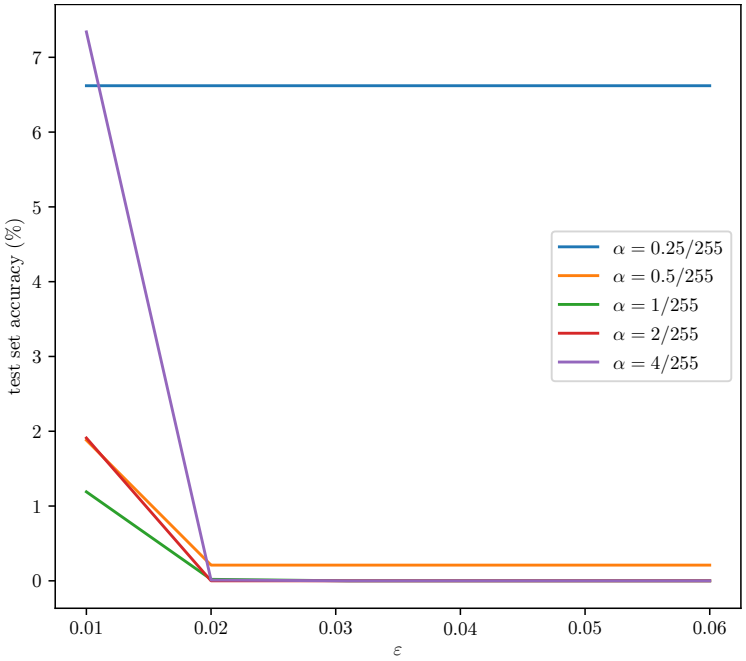

图 3: 不同的 ε 和 α 对测试集准确率的影响。

此外，我们还可以注意到一个现象：当 ε 过大或 α 过小时，随着 ε 的增大，模型在测试集上的准确率不再发生变化。这是因为，在固定的迭代次数下，若 ε 过大或 α 过小，则图像的扰动幅度不会达到 ε 的限制，因此 ε 的增大也就没有意义了。由此可见，我们不仅要选择合适的 ε 和 α，还要注意它们之间的大小关系，使两者比例适中。

3 CW 攻击

3.1 原理

CW 攻击 [3] 是一种基于优化的攻击，它的目标为

$$\text{minimize} \left\| \frac{1}{2}(\tanh(\mathbf{w})+1) - \mathbf{x} \right\|_2^2 + c \cdot f\left(\frac{1}{2}(\tanh(\mathbf{w})+1)\right),$$

其中

$$f(\mathbf{x}') = \max\left(\max\{Z(\mathbf{x}')_i \colon i \neq t\} - Z(\mathbf{x}')_t, -\kappa\right).$$

式中，$\frac{1}{2}(\tanh(\mathbf{w})+1) - \mathbf{x}$ 代表干净样本和对抗样本的差，通过双曲正切变换，使得 $\mathbf{x}$ 可以在 $(-\infty, \infty)$ 上变化，有利于优化；$Z(\mathbf{x})$ 表示样本 $\mathbf{x}$ 输入模型后未经 softmax 处理的输出

向量，在分类正确的情况下，其最大的分量对应着正确的类别；将类别 t 所对应的逻辑值记为 $Z(\mathbf{x}')$，将不同于 t 的类别的最大逻辑值记为 $\max\{Z(\mathbf{x}')_i : i \neq t\}$；$\kappa$ 为置信度，κ 越大，模型在错误类别上的置信度越大，但这样的对抗样本也更难寻找；c 是一个超参数，用来衡量两个损失函数的关系。

3.2 实验配置及结果

由于 CW 攻击十分耗时，速度较慢，我并没有反复调节超参数以达到最优的效果，而是仅仅选择了一组超参数进行实验，以了解 CW 攻击的基本原理。我取 $c = 0.1$，$\kappa = 0$，最大迭代次数为 1000，学习率为 0.01，对模型进行无目标攻击，获得了 94.5% 的成功率。

4 WOO 攻击

4.1 原理

WOO 攻击 [4] 属于黑盒攻击。在黑盒攻击中，我们无法直接获得模型的参数信息，也就无法直接计算梯度。因此，为了进行黑盒攻击，我们的思路有两种：一是自己训练模型，并将自己训练的模型的对抗样本迁移到目标模型上，这样攻击速度较快，但成功率较低；另一种是通过不断查询获得模型在某些样本上的输出，从而估计梯度，这样速度较慢，但成功率较高，可以达到 100%。

WOO 攻击是基于第二种思路的，其梯度估计公式如下：

$$\hat{g}_i := \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h}, \tag{1}$$

其中，h 是一个很小的常数，$\mathbf{e}_i$ 是标准基向量，其第 i 个分量为 1，其余分量为 0。对任意 $\mathbf{x} \in \mathbb{R}^p$，我们需要 $2p$ 次查询以得到 $\mathbf{x}$ 每个分量的梯度。事实上，仅需要多查询 1 次，我们还可以估计出 $f(\mathbf{x})$ 的 Hessian 矩阵：

$$\hat{h}_i := \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_{ii}^2} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - 2f(\mathbf{x}) + f(\mathbf{x} - h\mathbf{e}_i)}{h^2}. \tag{2}$$

ZOO 攻击的伪代码如 Algorithm 1 所示。

Algorithm 1 ZOO-Newton: Zeroth Order Stochastic Coordinate Descent with Coordinate-wise Newton's Method

Require: Step size η

while Not converged **do**

Randomly pick a coordinate $i \in \{1, \ldots, p\}$

Estimate $\hat{g}_i$ and $\hat{h}_i$ using (1) and (2)
if $\hat{h}_i \leqslant 0$ **then**
$\delta^* \leftarrow -\eta \hat{g}_i$
else
$\delta^* \leftarrow -\eta \frac{\hat{g}_i}{\hat{h}_i}$
end if
Update $\mathbf{x}_i \leftarrow \mathbf{x}_i + \delta^*$
end while

4.2 实验配置及结果

使用 ZOO 攻击，我在 CIFAR-10 测试集上获得了 100% 的攻击成功率。

ZOO 攻击生成的部分对抗样本如图 4 所示。可以看出，ZOO 攻击生成的样本，质量还是比较高的。

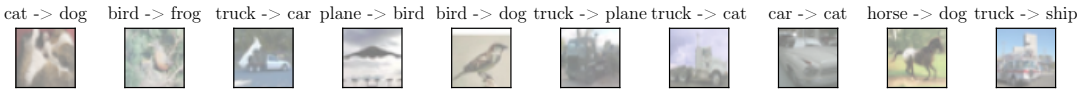

图 4: ZOO 攻击生成的部分对抗样本。

5 总结与收获

在本次实验中，我通过不同的方式实现了对图像识别神经网络的白盒攻击与黑盒攻击，并探究了超参数对攻击效果的影响，收获颇丰。

白盒攻击，即攻击者在可以了解到模型的内部结构和训练参数的情况下进行攻击。FGSM 攻击，是每次通过损失函数的回传，计算出原图像数据的梯度，并使图像数据沿梯度上升的方向变化 ε。ε 是扰动的大小，ε 越大，对原图像的扰动越大，攻击成功率就越高，但对抗样本与原图像的差距也就越大，更容易被人眼所察觉。PGD 攻击是 FGSM 攻击的迭代版本，每次迭代都相当于一次 FGSM 攻击。它通过多次迭代，能够更准确地找到更高质量的对抗样本。它主要有两个超参数：ε 和 κ。ε 是扰动的最大范围，PGD 攻击每次迭代的扰动都不超过这个范围，否则会被裁剪。α 是步长，即每次迭代图像数据的变化大小。为达到良好的攻击效果，我们应当选择合适的 ε，在保证攻击成功的情况下，使对抗样本的扰动尽可能的小；同时根据 ε 和迭代次数的大小，选择合适的 α，使得损失函数更快地趋向最大值。CW 攻击是基于优化的攻击，通过减小损失函数，我们希望达成两个目的：一是希望对抗样本和干净样本的差距越小

越好；二是希望模型将对抗样本分错，且在错误类别上的概率越大越好。它的优点是十分有效，缺点是比较耗时。

黑盒攻击，即攻击者在不知道模型内部结构和训练参数的情况下，通过模型的输入输出与模型交互，从而实现对模型的攻击。它通常有两种思路：一是利用对抗样本的可迁移性，在与模型类似的网络结构下，自己训练模型，生成对抗样本，然后用来攻击模型；二是利用模型的输入输出，推测出模型的一些参数和数据梯度等信息，来达到攻击的效果。我所使用的 ZOO 攻击属于第二种思路，它通过在图像数据上加一个微小的扰动 $h\mathbf{e}_i$，计算出输出 $f(\mathbf{x}+h\mathbf{e}_i)$ 和 $f(\mathbf{x}-h\mathbf{e}_i)$，从而估计出数据的梯度；并结合 $f(\mathbf{x})$，可以进一步估计出数据的 Hessian 矩阵。随后，根据计算出的梯度，实施攻击。ZOO 攻击是十分强大的，可以在不知道模型内部结构和参数的情况下，达到 100% 的准确率；但计算量也很大，比较耗时。

神经网络最大的弱点，便在于它易被攻击，鲁棒性弱。因此，研究各种各样的攻击和防御手段，对提高神经网络的鲁棒性，意义重大。相信随着相关研究的进行，神经网络的可靠性和实用性也将取得进一步的突破。

参考文献

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)*, pages 39–57. IEEE, 2017.

[4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.