作业二：神经网络测试工具的设计和实现

MF1833078 王月欢

一、 背景描述

随着大数据时代的到来和计算机硬件计算性能的提高，机器学习技术得到了空前的发展，越来越多的机器学习程序也被应用到日常生产生活的各个领域中，其中包括了如无人驾驶、恶意软件检测等很多安全性要求比较高的领域。机器学习是人工智能研究中的一个重要分支，致力于研究如何利用数据来改善自身性能的算法，使计算机程序在某一方面能够达到和人类表现相近的性能。随着机器学习程序的广泛应用，其在实际运行中的质量保障也越来越引起了人们的重视。当我们放心地把某一项任务交给计算机程序时，相信它能够做出和人类相同的判断和处理，但是这样的情况不是能够得到完全保障的，其中，机器学习程序就很容易受到对抗样本的攻击。以计算机视觉中的图片多分类的应用为例，在分类正确的图片中加入人眼接受范围内的扰动就可以使机器学习程序得到错误的分类结果，从而导致程序错误行为的产生。

二、 功能描述

在当前流行数据集上 mnist 上，对经典的神经网络模型 Lenet 进行测试，生成模型的对抗样本。

三、 需求分析

本工具的主要功能为产生神经网络的对抗样本，使神经网络产生错误的行为。

输入：原始样本

输出：对抗样本

方法：FGSM、JSMA、C&W 神经网络攻击方法。

使用开源框架：cleverhans

四、 技术路线

利用神经网络内部的梯度信息，生成扰动，使得神经网络产生错误行为，同时控制原始样本和生成样本的距离，使得生成的对抗样本对人类来说仍可分类正确。

五、 详细设计

FGSM 算法：快速梯度符号法，通过生成扰动来使神经网络分类错误，寻找扰动的方法是通过神经网络的训练误差函数关于目标错误分类对原始样本求梯度符号化之后加到原样本上。

生成样本为原始样本加上扰动：

$$\vec{x^*} = \vec{x} + \delta_{\vec{x}}$$

生成扰动的公式为：

$$\delta_{\vec{x}} = \varepsilon \operatorname{sgn}(\nabla_{\vec{x}} c(F, \vec{x}, y))$$

JSMA 算法：基于雅可比显著图的攻击方法，首先根据雅可比矩阵，计算样本的显著图，代表了图片中每一个元素对分类错误标签产生的影响，然后选择影响较大的元素，通过改变像素值来添加扰动。

生成样本为原始样本加上扰动：

$$\vec{x^*} = \vec{x} + \delta_{\vec{x}}$$

根据样本的显著图来选择改变的样本点

$$S(\vec{x},t)[i] = \begin{cases} 0 \text{ if } \frac{\partial F_t}{\partial \vec{x}_i}(\vec{x}) < 0 \text{ or } \sum_{j \neq t} \frac{\partial F_j}{\partial \vec{x}_i}(\vec{x}) > 0 \\ \frac{\partial F_t}{\partial \vec{x}_i}(\vec{x}) \left| \sum_{j \neq t} \frac{\partial F_j}{\partial \vec{x}_i}(\vec{x}) \right| \text{ otherwise} \end{cases}$$

C&W 方法：将攻击方式转换为一个更加高效的优化问题，能够以添加更小扰动的代价得到更高效的对抗样本。

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x + \delta \in [0,1]^n \end{aligned}$$

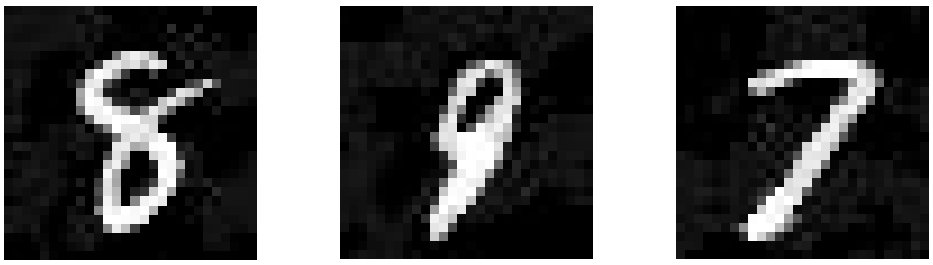
六、 实现结果

运行方式：

```
python3 main.py [option]
option:
    fgsm：FGSM 攻击方法
    jsma:  JSMA 攻击方法
    cw: C&W 攻击方法
```

生成的对抗样本示例：

FGSM:

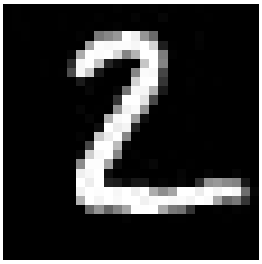


5　　3　　2

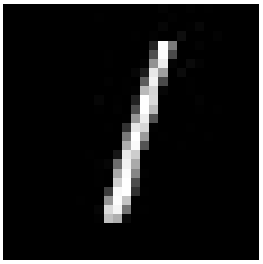
JSMA:

6　　1　　0

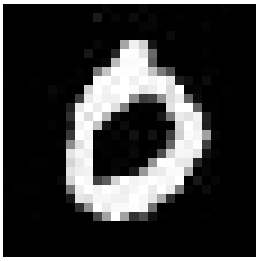
C&W:



0



3



2

七、 结论

三种产生对抗样本的方法中，FGSM 方法速度最快，JSMA 方法只改变一些像素的值，速度较慢，C&W 方法产生的扰动最小，速度慢。另外生成的对抗样本可以用于神经网络模型的再训练，从而提高模型的性能。

项目代码，github 链接：https://github.com/wazxser/DNN_testing