《计算机科学导论》大作业

Project 01: 图像的分类模型攻击与消息隐写

大作业背景

计算机视觉与计算机安全是计算机科学领域两大重要的子分支。随着计算机科学的不断发展，这两大领域出现了诸多交叉方向。在这其中，对于计算机视觉深度学习模型的对抗攻防，以及对于数字图像的消息隐写是两大受到了广泛关注的问题。本次大作业要求结合在本课程中学到的关于计算机视觉与计算机安全知识，对于上述两大方向进行基本的实践。

任务一：图像分类模型攻击

- **背景：**在计算机视觉相关内容的学习中，我们了解到，在深度学习的范式下通过大量图像数据的训练，可以得到一个准确率很高的基于卷积神经网络（CNN）的分类模型。然而，由于模型的识别能力过度依赖于训练数据分布，输入图像的微小扰动都可能会导致分类结果的巨大变化，我们将这种使得模型分类出错的方法叫做“模型攻击”。
  - **黑盒攻击：**其中，不考虑模型结构， 只通过扰动输入观察模型输出的方法我们称为黑盒攻击。
  - **攻击评估：**对于每一次攻击前后，模型分类错误则意味着攻击成功，对应真实类别分类概率分数下降程度越高，则本次攻击越有效。
- **参考内容：**
  - 一个在Cifar-10上预训练好的CNN分类模型（./attack/cifar_net.pth）
  - 图片的测试和部分攻击、可视化代码 (./attack/attack.py)。成功的攻击会以图片形式保存在./attack/good_attack文件夹中，图片命名格式为A_to_B_score.jpg，表示该次攻击使得模型将该图片从A类识别错为B类，且A类的分类概率下降程度为score。
  - Cifar-10 image set (Project01-CifarData.zip)，其中包含41份划分好的测试数据
- **要求：**
  - 根据你所在的小组对应的ID，解压Project01-CifarData.zip，在解压出的文件夹中选取对应的测试数据文件夹（文件夹名称和ID一致），重命名为“test”并置于data/文件夹下（./attack /data/test）。
  - test中一共有10类，每类有100张图片，对这1000张图片进行攻击。
- **指标：**模型分类错误表示攻击成功，对应真实类别分类概率分数下降程度越高，则本次攻击越有效。
- **任务详情：**编写相关代码针对输入图像以下三种不同的攻击，观察模型分类错误情况：
  - **噪声攻击：**在输入图像上添加满足不同分布类型的噪声（高斯、随机噪声）进行攻击。对于高斯噪声，设定均值为零，选取十个在$[0.001,0.01]$之间的方差，用表格列出

攻击后的分类准确度。对于随机噪声（在图像上随机选取N个像素点，将像素点的值置1，可参考[2]的实现代码）选取不同的$N \in [100,400]$，用表格列出攻击后的分类准确度。观察并量化比较这两类噪声的攻击效果，并分析原因。（20分）

- **Mask攻击：**在输入图像上选择一个大小为$8 \times 8 \sim 4 \times 4$范围内的方形区域 (在该范围内自选一种面积，如$8 \times 8$或者$5 \times 5$)，将该区域的像素值均置为0来进行攻击，你选用的Mask区域大小是多少，选择的理由是什么？观察对于不同类别，在哪些区域进行mask攻击是最有效的，通过观察和可视化，你能够得出什么结论？（10分）
- **单像素攻击：**在输入图像上对一个像素点在RGB通道上的值进行"轻微扰动"进行模型攻击。这种攻击方式在你挑选的测试图片中能否成功，如果能，对干扰前后的图片进行比较和展示，并选取5个不同的扰动强度p (0<p<0.5)进行测试和分析。如果不能，则寻找预测类别置信度下降最大的10张图片进行分析。（10分）

  （补注：“轻微扰动”定义：Perturbed(I, x, y, p)= p * sign(I (:, x, y)-0.5) + 0.5 ,其中I表示图像，图像像素点的值∈ [0,1]，x, y表示单像素坐标，sign为符号函数，p为扰动强度, (0<p<0.5), 相关参考代码可见 attack.py）
- **【Bonus】模型防御：**根据你所了解的图像分类模型训练和测试过程，针对以上攻击手段，是否有策略使得模型的防御能力更强呢？（10分）

任务二：消息隐写

- **背景：**LSB全称为 least significant bit，即最低有效位。LSB图片隐写术是在信息安全领域一种常见的信息隐藏方法，其核心原理是，对于8比特图像，最低阶比特位的变化人眼是难以捕捉的，所以可以把秘密消息转换为二进制，再把二进制码嵌入到一张图像的最低阶比特位，这就完成了信息的隐藏。
- **要求：**本次作业需要同学们按照以下要求创建一段密文，并将密文用LSB图片隐写术嵌入到一张8比特RGB图像中。
- **参考资料：**本任务给定以下内容供参考：
  - 对于灰度图的隐写代码示例 (./conceal/LSB_grey.py)
  - 灰度图像示例 (./conceal/*.png)，hex形式表示的哈希示例（./conceal/hash_example.txt），RGB图像示例（./conceal/image_RGB.jpg）
- **任务详情：**编写相关代码实践并实现以上功能：
  - **创建密文：**Hash（散列函数）可以用于生成文本的摘要。一般而言，不同的信息通过Hash函数的处理会生成不同的消息摘要。Hash函数在区块链中也有着非常重要的应用。请同学们将**自己的队伍名称**利用一次SHA-256运算生成消息摘要，并以16进制表示的字符文本（每四个比特用0到f表示）保存为txt格式。生成过程可以用Python语言提供的相

关密码库，也可以使用在线加密工具如 https://www.sojson.com/hash.html 等。（10分）

- **LSB隐写：**在Cifar-10数据集中任意选择一张图片，使用Python语言完成信息的隐写。源文件LSB_grey.py实现的是对灰度图像进行LSB隐写，它包含的主要函数有：
  - 隐写函数LSB_encode，输入原图与消息文本，输出为增加了LSB消息的图片。
  - 隐写信息提取函数LSB_decode，输入隐写图片，输出LSB中隐藏的消息。

  Cifar-10数据集的图像是RGB图像而非灰度图像，请你对LSB_grey.py代码进行修改，让它实现对RGB图像的R（红色）通道进行隐写的功能（15分），并且能将隐写的信息提取出来。修改后的代码请以LSB_RGB.py为名保存（15分）

作业提交

- **报告要求**（10分）
  - 报告需要包含选题，各组员学号、姓名、邮箱地址，页眉表明课程信息、选题题号、小组组号和组名，页脚表明页数。
  - 报告以中文分章节撰写，每小题对应一节内容，包含任何必须的文字、公式、图片、表格等，表格和图片要有对应的标题，在文中要有相应的引用。
  - 报告结尾要包含致谢和参考文献两章，可在致谢章节中写上你的感受、建议或评论。
  - 如有变量，请明确定义，可以单列一章（或一个表格）来总结报告中涉及的所有变量和对应的含义。
- **提交材料**
  - 任务一：请提交你修改、补充后的attack.py，示例图像（成功攻击的图像），所有以上内容放在一个名为attack的文件夹内, 我们提供的预训练模型无需再提交。
  - 任务二：请提交源文件LSB_RGB.py，示例图像（包括原图、被隐写的图像、隐写的消息内容），所有以上内容放在一个名为conceal的文件夹内。
  - 对于整个实验过程按照报告要求撰写一份报告；将实验报告和attack/文件夹conceal/文件夹，联合进行打包，以“Project-队伍编号.zip”进行提交。如第01组同学请以“Project01-队号.zip”进行提交，比如第24组同学的命名应该为Project01-24.zip。

参考资料

[1]. Numpy和tensor的转化：https://www.cnblogs.com/wzyuan/p/9733433.html

[2]. 图像加噪：https://blog.csdn.net/qq_45769063/article/details/107137025

[3]. 对图像进行掩膜处理：https://blog.csdn.net/lucky__ing/article/details/78559916

[4]. PIL数据处理图片加遮挡、噪声、模糊: https://cloud.tencent.com/developer/article/1370938

[5]. 单像素攻击: Simple-Black-Box-Adversarial-Perturbations-for-Deep-Networks, CVPRW, 2017