上机实验四：深度神经网络后门攻击

任务要求：对实验二构建的CIFAR-10数据集分类神经网络进行训练数据集污染，实现后门攻击。

后门攻击：攻击者希望模型只对某种特定特征数据作出错误的判断。通过污染（poison）一小部分训练数据，诱导网络模型学习后门触发器和目标标签之间的相关性。

实现步骤：

1. 对训练数据集除airplane这一类别之外的其他九类数据进行污染，即添加后门攻击的触发开关。例如，如下图所示，右下角的白色方块即为后门触发条件。将所有污染后的训练样本标记为攻击目标，即airplane。联合原始数据集以及受到污染的九类数据样本，对CIFAR-10分类神经网络进行训练。



1. 在受到后门攻击的神经网络上，测试十类干净数据（即使用原始测试数据）的分类准确率，及九类植入后门触发开关的测试数据的攻击成功率（被分类为airplane即视为攻击成功）。
2. 提示：后门触发开关的大小、形状、位置、颜色、隐形程度均可自定义。Be Creative!后门触发开关越隐蔽越好。

报告要求：

1. 污染训练数据时，要对除airplane的剩余9类选取固定比例R的样本进行污染，即在每一类中，受污染的样本个数/该类的总样本数=R。测试不同的比例R，对攻击成功率的影响。
2. 画出横坐标为R，纵坐标为后门攻击成功率的折线图，R的范围自定。
3. 实验报告包含网络设计、后门触发条件示意图、实验结果图，以及必要的分析等。
4. 代码和实验报告一起打包成压缩包提交到NAS。
5. 截止时间为4月22号下午5:00，逾期视为0分。