Stochastic Optimization for Adversarial Training 研究方向探索

1.随机优化算法

内容：设计新的随机优化算法，在面对对抗攻击时抵抗能力更强；或者是对现有的随机优化算法进行改动，比如替换优化器等，进行比较分析。

设计新的优化算法：

比较分析：

对现有的算法的优化算法部分进行替换，比如SGD→Adam等，进行实验和比较分析

（1）优化器

Adam、AdaGrad、RMSProp、AdamW、Lookahead、RAdam等

（2）实验设置

数据集：MNIST、CIFAR10、ImageNet等

攻击方法：FGSM、PGD、FREE等

（3）评估指标（实验并采集数据）

鲁棒性评估： 评估模型在对抗样本上的准确率，以衡量优化器在对抗训练中的效果。

收敛速度： 记录训练过程中的损失函数值和准确率变化，以比较不同优化器的收敛速度。

稳定性： 观察不同优化器在训练过程中参数更新的稳定性和模型性能的波动情况。

（4）结果分析

性能比较： 对比不同优化器在对抗训练中的表现，找出最适合的优化器或组合策略。

参数调整： 探索不同优化器的参数设置，如学习率、动量、权重衰减等，优化其在对抗训练中的效果。

(5) 改进

混合优化器： 基于实验结果，可以尝试将多个优化器结合使用，形成混合优化策略。例如，在早期训练阶段使用RMSProp，后期切换到Adam。

自适应调整策略： 开发自适应调整策略，根据对抗训练过程中模型性能的变化，动态调整优化器的参数或切换优化器。

参考文献

Adam: A Method for Stochastic Optimization

Adversarial Training Methods for Semi-Supervised Text Classification

Improving the Adversarial Robustness of Transfer Learning by Implicit Regularization

On the Convergence and Robustness of Adversarial Training

Robust Optimization for Machine Learning

2.收敛性和鲁棒性理论分析

收敛性：收敛速度，收敛条件

鲁棒性：度量，正则化方法的影响

其他：凸优化，非凸优化，随机优化

实验：

收敛性实验：

实验设置: 选择几个常见的优化算法（如SGD、Adam），在标准数据集上进行对抗训练，记录每次迭代的损失值和准确率。

分析方法: 使用梯度范数、损失值变化速率等指标评估收敛性。

鲁棒性实验：

实验设置: 采用常见的对抗攻击方法（如FGSM、PGD），测试模型在不同扰动强度下的准确率。

分析方法: 计算对抗损失、对抗准确率，并绘制鲁棒性曲线。

参考文献

3.对抗训练中的正则化技术分析

4.不同模态的对抗训练分析

内容：比较同一对抗训练算法在不同模态下的表现，并根据表现调整和优化技术，应对每种模态独特的挑战。

（1）图像数据的对抗训练：

背景: 图像数据是对抗训练中最常见的模态之一。对抗样本在图像领域表现出明显的效果，例如添加细微的像素级扰动。

可以做的方向：

多尺度攻击: 研究如何在不同尺度上生成对抗扰动，从而使模型在多尺度特征上都具有鲁棒性。

GANs与对抗训练结合: 利用生成对抗网络（GANs）生成逼真的对抗样本，以提升对抗训练的效果。

（2）文本数据的对抗训练：

背景: 文本数据的对抗训练面临的挑战在于扰动需要保持语义不变且符合语言结构。

可以做的方向：

词级扰动: 对单词进行替换或修改，使其在语义上保持一致但对模型具有迷惑性。

句级扰动: 使用对抗生成器对整个句子进行修改，确保生成的句子仍然具有可读性和逻辑性。

（3）音频数据的对抗训练：

背景: 音频数据的对抗训练涉及对波形的微小扰动，要求在听觉上不明显但对模型造成影响。

可以做的方向:

频域扰动: 在频域上进行扰动，如修改特定频段的幅度。

时域扰动: 在时域上进行微小的时间偏移或添加噪声。

（4）多模态数据的对抗训练：

背景: 多模态数据结合了不同类型的数据（如图像和文本），需要综合考虑各模态的对抗性。

可以做的方向:

跨模态对抗训练: 研究如何在一个模态上生成对抗样本，并在其他模态上验证其影响。

联合对抗生成: 开发联合对抗生成方法，同时对多个模态施加扰动，以提高综合鲁棒性。

参考文献

Intriguing properties of neural networks.

Explaining and harnessing adversarial examples.

Adversarial examples for evaluating reading comprehension systems.

Hotflip: White-box adversarial examples for text classification.

CRAFT: Cross-modal Adversarial Filter Training for Robust Video Sentiment Classification..