

**《数据安全与隐私保护》课程报告**

**题目： 图像对抗攻击整合与图像隐写初探索**

**姓名： 林浩、蒋维、杨利伟**

**学号： 21312596、 21312151 、2131265**

**目 录**

[摘要 1](#_Toc156146194)

[第一章 背景和相关工作 2](#_Toc156146195)

[第二章 作品设计和意义 6](#_Toc156146196)

[第三章 作品架构分析 9](#_Toc156146197)

[第四章 创新性说明 15](#_Toc156146203)

[第五章 总结 16](#_Toc156146204)

[项目分工 17](#_Toc156146205)

[参考文献 18](#_Toc156146206)

# 摘要

项目实现了一个集成多种攻击方式以及图像隐写功能的图像对抗攻击程序，旨在研究数据安全与机器学习模型的鲁棒性。程序的核心功能是通过对抗攻击生成图像样本，并使用图像隐写进行信息隐藏。其特性包括多种对抗攻击策略的应用、用户友好的界面设计，以及灵活的参数调整能力。创新之处在于将对抗学习与图像隐写术有效结合，为安全性研究提供新的视角。此外，程序在教育和研究领域具有高度的实用价值，能够为学习者和研究人员提供一个实验和测试不同图像对抗攻击技术的平台。

# 第一章 背景和相关工作

1.1对抗攻击的概念

对抗攻击是指在机器学习，特别是深度学习领域中，一种针对模型的攻击方式，旨在通过输入微小但故意的扰动来误导模型。这种攻击利用了机器学习模型，尤其是深度神经网络在处理输入数据时的内在脆弱性。

对抗攻击的核心是创建“对抗性示例”（adversarial examples）。这些示例在人类观察者看来可能与原始输入无异或仅有微小差别，但却能导致模型做出错误的判断或分类。例如，在图像识别系统中，通过对原始图像进行轻微修改，一个被正确识别为“猫”的图像可能被错误地识别为“狗”。

对抗攻击的实现通常基于对机器学习模型的输入数据进行精确而微妙的修改。这些修改是通过计算输入对模型输出的梯度（或导数）来识别的，目的是找出哪些小的改动会在模型的决策过程中产生最大的影响。例如，在图像识别系统中，算法会计算每个像素点的改变如何影响分类结果，并相应地调整像素值，以导致模型做出错误的分类。

1.2 对抗攻击的背景意义

1.2.1 对人工智能的安全与伦理影响

在当前的人工智能发展背景下，对抗攻击的安全和伦理影响成为了一个重要的研究议题。这种影响主要体现在对抗攻击揭示了机器学习模型在面对精心设计的输入时的脆弱性，这对于依赖这些技术的关键系统来说可能带来严重的安全威胁。例如，在自动驾驶汽车、医疗诊断、金融服务等领域，对抗性攻击可能导致错误的决策，从而引发生命安全和财产损失的风险。这不仅挑战了技术的可靠性，也影响了公众对这些先进技术的信任和接受度。

此外，对抗攻击在伦理层面引发了对人工智能应用责任和可信度的深入讨论。由于对抗性示例的使用可能涉及误导和欺骗，这在某种程度上挑战了人工智能系统的道德边界。例如，使用对抗性示例来欺骗面部识别系统可能引起隐私和安全的关切，而在医疗系统中的错误诊断可能导致严重的健康后果。

因此，对抗攻击的研究不仅揭示了机器学习模型的技术限制，也提出了关于如何建立和维护高度自动化、智能化系统的安全、可靠和道德标准的重要问题。这要求研究者、开发者和政策制定者共同努力，不仅要提高技术的鲁棒性，也要确保技术发展符合社会伦理和法律框架。通过这样的努力，可以更好地利用人工智能的潜力，同时最大程度地减少其潜在风险。

1.2.2 对可解释人工智能的影响

对抗攻击对于可解释人工智能（Explainable AI，简称XAI）的意义在于，它为理解和揭示复杂机器学习模型的内部工作机制提供了一个独特的视角。尽管深度学习模型在性能上取得了显著成就，但它们的“黑盒”特性使得理解模型的决策过程变得复杂。对抗攻击通过创建能够显著改变模型输出的微小输入变化，揭示了模型对特定特征的敏感性和依赖性，从而为理解模型如何处理和解释数据提供了线索。

通过分析对抗性示例对模型决策的影响，研究人员可以识别模型在进行预测时依赖的关键特征。这种分析有助于揭示模型的依赖模式，即模型在做出决策时更倾向于依赖哪些数据特征，从而提供了解释模型行为的一个窗口。对抗攻击还可以作为一种工具，用于评估和改善模型的鲁棒性。通过理解模型在面对对抗性扰动时的表现，研究人员可以更好地识别和解决模型在真实世界应用中可能遇到的问题。

可解释的人工智能强调了模型决策的透明度和可理解性。通过对抗攻击的研究，可以更清楚地识别和解释模型可能的偏见和局限性，进而引导开发更加公正和负责任的AI系统。

总之，对抗攻击的研究扩展了我们对深度学习模型决策过程的理解，为提高人工智能系统的透明度和可解释性提供了新的途径。这对于构建用户信任、确保系统公正性和负责任地应用人工智能具有重要意义。它帮助我们更好地理解和解释这些复杂系统的行为和决策过程。

1.3 对抗攻击的相关工作

Szegedy等人在2013年首次发现了对抗性样本的现象[1]，当时他们通过向人眼无法检测到的输入样本添加微小的扰动，对基于深度神经网络的图像分类系统进行了错误分类。继Szegedy之后，对对抗样本的相关研究已成为深度学习领域的研究热点。 随着对抗性攻击领域的发展，研究人员提出了许多生成对抗样本的方法，如增加攻击强度、改善模型迁移、优化计算能力等。

基于攻击者对目标模型的了解，对抗性攻击通常可以分为白盒攻击、灰盒攻击和黑盒攻击。进一步的，Serban and Poll提出对抗性攻击的策略可以细分为多种类型[2]，包括梯度攻击（包括基于优化的攻击等）、基于迁移的攻击、基于分数的攻击（包括基于决策的攻击、注意力攻击[3]）和基于几何变换的攻击等。

此外，根据攻击目标的不同，攻击策略可以分为有针对性和无针对性攻击。 针对性攻击有指定的错误分类类别，而非针对性攻击只需要模型错误分类，而针对性攻击通常更难。 此外Moosavi-Dezfooli等人提出，攻击者可以选择使用单个对抗样本，同时作用于多个目标模型，称为通用攻击[4]。

攻击成功率是衡量对抗性攻击样本质量的重要指标，此外，扰动范数也被广泛用于量化对抗性样本的质量。 以图像数据为例，扰动应用于对抗性攻击的范数包括

1），它指的是修改后的像素数据的数量; 2），扰动元素的平方和再平方根，对于图像数据，范数越小表示对抗样本更难被人眼识别; 3） ，表示扰动元素中的最大值。 各种攻击方法通常使对抗样本中的扰动成为攻击者也可以通过限制范数的值来选择使用具有扰动的单个对抗样本。

在攻击过程中，攻击者可以使用迭代或一次性方法来查找并生成扰动。 对抗性攻击中的迭代是指多次计算产生对抗性扰动，而one-shot是指仅使用一次计算产生对抗性扰动。 一般认为，迭代次数越多，攻击性能越好，而另一方面，迭代次数越多，计算资源消耗就越多。

现有的绝大多数攻击都集中在数字攻击上，这些攻击可以完全访问目标的数据输入模型。相比之下，物理攻击无法访问目标模型的数据，而是在生成模型输入的图片之前添加所有对抗性扰动。 物理攻击在自动驾驶、安防、监控系统等场景中具有重要应用。

对抗性防御用于构建足够健壮的分类器（防御模型），即使在输入攻击图像时也能正确分类。 Akhtar and Mian提出构建防御模型主要有三种方法[5],一种是训练一个更强大的分类器，这是对分类器本身的防御，通常通过对抗性学习进行训练。 第二种是在将输入攻击图像传递给分类器之前对输入攻击图像进行一些预处理，目的是尽可能减少攻击噪声[6]。 第三种是在模型中引入其他结构，例如检测器，以帮助更好地检测攻击，如Wang等人通过测试对抗性和良性样本对随机突变的敏感性来检测攻击[7]。

1.4 图像隐写的概念

图像隐写是一种信息安全领域的技术，它涉及将秘密信息嵌入到数字图像中，以隐藏这些信息，同时不引起人眼的察觉。图像隐写的目的是在图像中嵌入信息，使得这些信息对普通观察者来说是不可见的，而只有授权的人或特定的软件能够提取出这些隐藏的数据。

图像隐写的方法通常包括将秘密信息以某种方式编码并嵌入到图像的像素值中，或者通过微小的变化来隐藏信息，这些变化对于人眼来说几乎是不可察觉的。图像隐写可以用于各种目的，包括保护数字内容的版权，隐私保护，秘密通信等。然而，它也可以被滥用，用于隐藏恶意目的，如数字取证的规避或恶意软件的传播。

总之，图像隐写是一种重要的信息安全技术，它允许在数字图像中嵌入和提取隐藏的信息，同时保持图像的视觉质量和完整性。

1.5图像隐写背景

1.6图像隐写与对抗攻击

基于扩散模型的图像隐写不仅可以很好地利用扩散模型对噪声天然的鲁棒性来进行图像隐藏，也可根据基于扩散模型的图像隐写技术来实现对抗攻击和防御

具体来说通过把对抗样本使用图像隐写技术重新生成可以去除对抗攻击造成的影响，实现对抗防御；

此外通过基于扩散模型的图像隐写技术通过把私钥和公钥的提示词互相替换可以实现图像隐写生成的图片分类为提示词的类别

# 第二章 作品设计和意义

本作品所实现的程序旨在实现一个图形界面化的，用于生成对抗样本的定制化整合程序。同时我们尝试探索图像隐写技术与对抗图像样本生成之间的关联与作用，为日后的对抗攻击研究提供测试经验。

我们已经完成了对抗图像样本的整合生成程序，但碍于图像隐写的……我们未能实现将图像隐写与对抗攻击的融合程序。但是我们测试并总结了图像隐写对于对抗攻击的关联和研究意义。基于图像隐写的对抗攻击程序有待日后……的完善和进一步研究测验。

程序的对抗攻击是针对经典的Resnet18图像分类模型。Resnet18是在ImageNet数据集上预训练好的CNN模型，在性能优越的同时有着95以上的样本图像分辨准确率，因此将它作为对抗攻击程序的目标模型。

程序的定制化和整合化源于其生成对抗样本图像示例的简洁自定义流程。程序支持FGSM、BIM、PGD、Deepfool、Onepixel等经典或前沿的对抗攻击方式。部分攻击方式支持定向攻击（如BGM、PGD）。同时每一种对抗攻击方式都支持自定义参数或者默认参数训练。用户首先输入自己的示例图像并选择攻击方式，然后自定义对应的参数，即可开始生产对抗图像样本。

整体的对抗攻击的预期流程是这样的：用户的原图像是“干净的”，肉眼看上去是什么图像就会被图像分类模型判断为什么种类，比如猫咪。基于这个原图像生成的对抗图像样本期望在视觉上无法发现异常，肉眼看上去是一张与原图像类似或者看上去一模一样的图像，但是会被图像分类模型误判为其他种类，比如看上去是猫咪但会被判为飞机。

让图像分类模型误判并不难，难的是如何保证对抗图像样本的视觉质量，这里的视觉质量的评判标准也有两个层次，一种是让对抗样本单纯看上去是一张自然正常的图像，图像主体没有很大变化，与原图像整体近似。更深一种层次是希望对抗样本不仅是正常自然的，还希望与原图像细节逼近，追求与原图像的匹配度。

影响对抗图像样本视觉质量的因素有很多，首先本身选择攻击的图像分类模型如果过于复杂，对图像的特征识别较为灵敏，对抗图像样本需要变动的区域就会扩大加深，也就间接影响了对抗图像样本视觉质量。因此这也是本次项目选择Resnet18这种相对轻量化的图像分类模型原因之一。另外，用户选择的攻击方式，攻击方式的参数，都会直接影响生成的对抗图像样本。而本项目所实现的简洁整合的对抗样本生成流程是有利于用户通过多次调参观攻击方式以及参数对对抗样本的视觉影响。

另外，图像隐写技术的探索引入扩散模型的图像隐写，扩增了对抗样本图像的视觉质量提升的途径，可以实现对抗攻击和对抗防御

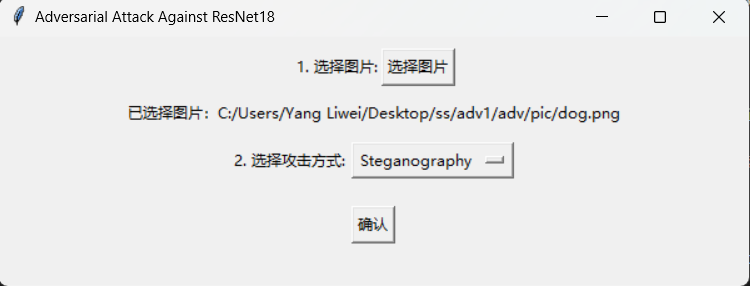
该程序项目的意义是多方面的。对于初次了解对抗攻击的学习者，该程序简洁清晰的展示了生成对抗图片样本的整体流程，让初学者生动感受对抗攻击的实际应用。同时对于研究者而言，该项目能够为他们的对抗攻击研究提供一个便捷的测试环境。让他们可以测试不同图片，不同攻击方式，不同参数所生成的对抗图像样本。

该程序项目仍有许多有待日后改进的地方，比如通过扩散模型把输入图像进行图像隐写，在隐写的图像上进行对抗攻击后利用扩散模型恢复图像隐写隐藏的图片，这个整体的框架可以实现指定目标类别的对抗攻击。

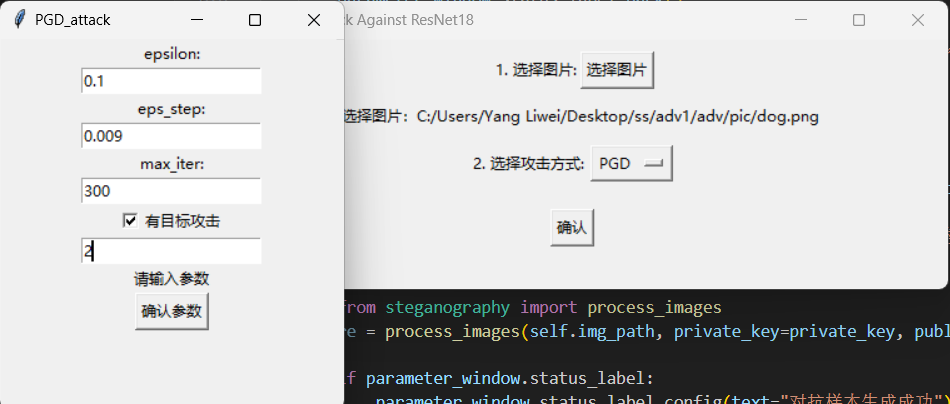
# 第三章 作品架构分析

3.1 图像界面使用

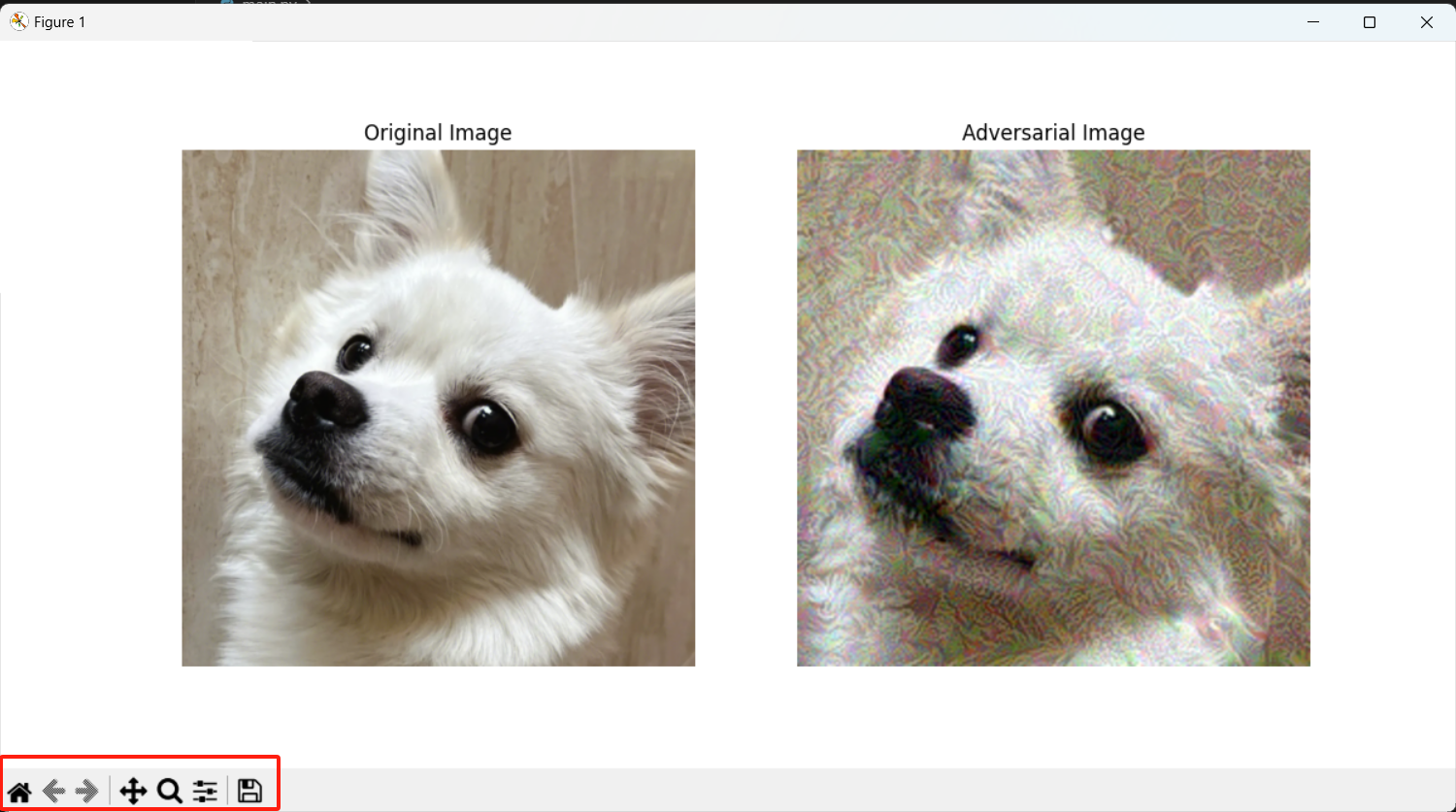
我们的图形化界面十分简洁，分为一个主界面和若干个分界面，主界面只有两个参数，用于选择图片和攻击方式。主界面如下：



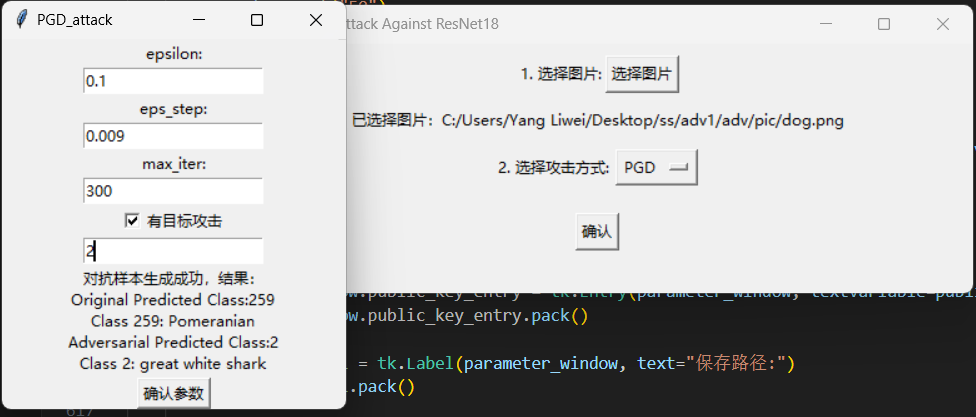
点击确认后会根据选择的攻击方式跳出对应的窗口，用来收集相关参数，如选择PGD攻击后会跳出左侧的窗口，已有的参数是我们推荐的最佳参数，用户也可以根据需求更改，对一些攻击方式，我们同时实现了有目标攻击和无目标攻击，可以根据参数设置进行何种攻击。



点击确认参数后，会调用函数进行对应的攻击，攻击完成后会显示攻击后图像与原图像，通过左下角的按钮可以处理该图像，攻击生成的对抗样本也会自动存储到默认文件夹./adving

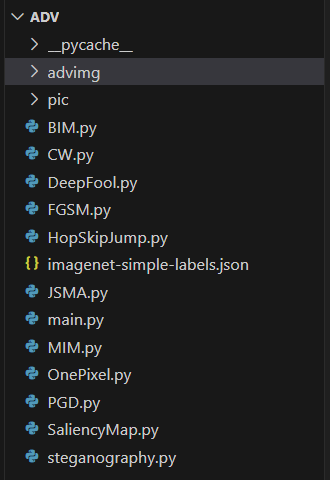


关闭该窗口后，会显示攻击的结果：

该图像中的结果表示原图被识别为第259类，第259类的名称为Pomeranian,对抗样本被识别为第2类，第2类名称为 great white shark。

3.2 代码组织方式

3.2.1整体项目文件说明



项目文件目录如上。其中main.py文件是图形化界面入口，另外十个.py文件是十种攻击方式的代码，一个json文件是ImageNet数据集的标注列表，用于根据标签寻找具体种类。Adving文件夹是存储对抗样本的文件夹。

3.2.1 main文件说明

main文件的主体是类ImageAttackGUI，该类定义了主界面和所有副界面，同时封装了所有攻击方式，通过统一的格式将所有攻击方式组织了起来。

主界面定义函数：create\_widgets() 该函数定义了主界面。

攻击方式触发函数：on\_attack\_selected() 每当用户点击确认后，触发该函数，检查用户选择的攻击方式，并调用相应的副界面定义函数。

副界面定义函数：命名格式，params\_advname(),

例如：params\_fgsm(),params\_pgd(),params\_cw()每个函数根据攻击所需参数生成相应的界面，并通过钩子函数触发相应的攻击函数。如下代码所示：

def params\_fgsm(self):  
 # 创建一个新窗口用于输入参数  
 parameter\_window = tk.Toplevel(self.root)  
 parameter\_window.title('FGSM\_attack')  
 parameter\_window.geometry("300x200")  
  
 # 在新窗口中添加参数输入框和确认按钮  
 #........添加窗口元素  
 # 钩子  
 confirm\_parameter\_button = tk.Button(parameter\_window, text="确认参数", command=lambda: self.fgsm\_attack(  
 parameter\_window, float(parameter\_window.epsilon\_entry.get())  
 ))  
 confirm\_parameter\_button.pack()

攻击函数：命名格式，advname\_attack(),每个函数都封装了相应的攻击。如下代码所示：

def fgsm\_attack(self, parameter\_window, epsilon):  
 # 更新消息  
 if parameter\_window.status\_label:  
 parameter\_window.status\_label.config(text="对抗样本生成中")  
 parameter\_window.update\_idletasks() # 立即刷新窗口  
 # 调用定义在FGSM文件中的攻击函数，传递相关参数  
 from FGSM import FGSM\_attack  
 re = FGSM\_attack(self.img\_path, epsilon, targeted=False,target\_label=0)  
# 更新消息  
 if parameter\_window.status\_label:  
 parameter\_window.status\_label.config(text="对抗样本生成成功，结果：\n" + ''.join([str(item) for item in re]))  
 parameter\_window.update\_idletasks() # 立即刷新窗口

**3.3模型封装的攻击方式**

FGSM (Fast Gradient Sign Method): 一种快速且高效的攻击方法，通过利用模型的梯度信息来生成扰动。

BIM (Basic Iterative Method): 基于迭代的攻击方法，逐步增加小的扰动以找到有效的对抗性示例。

PGD (Projected Gradient Descent): 这是一种更强大的迭代攻击方法，通过在每一步都应用投影梯度下降来制造扰动。

DeepFool: 一种高效的攻击方法，旨在最小化输入的扰动，同时使模型分类错误。

C&W (Carlini & Wagner): 一种高效且复杂的攻击方法，专注于生成几乎无法察觉的对抗性扰动。

JSMA (Jacobian-based Saliency Map Attack): 使用雅可比敏感度分析图来确定对输出影响最大的像素，然后进行修改。

MIM (Momentum Iterative Method): 在迭代攻击过程中加入动量，以提高攻击的稳定性和效果。

One-Pixel: 仅通过修改图像中的一个像素来欺骗神经网络，显示了深度学习模型的脆弱性。

HopSkipJump: 一种基于决策边界的攻击方法，不需要访问模型梯度信息。

SaliencyMap: 利用显著性图来确定最重要的像素，并对这些像素进行修改以引发分类错误。

Steganography: 不是一种传统的对抗性攻击方法，而是一种隐藏信息的技术，通过在图像中嵌入隐蔽信息。

# 第四章 创新性说明

**1实现了通过基于扩散模型的图像隐写技术分别进行对抗攻击和对抗防御**

具体来说：

1.1对抗攻击：通过对多抗样本进行图像隐写，隐藏和恢复提示词都设置为预定类别比如在对猫的图片图像隐藏和恢复提示词均写为“a tiger”，最终生成的图片别分类为a tiger cat。

1.2对抗防御：通过对多抗样本进行图像隐写，隐藏和恢复提示词均写为图像原始类别，最终生成的图像分类为正确样本而不是被对抗攻击后分类为错误样本。

**2提出了通过基于扩散模型的对抗攻击框架**

提出了通过扩散模型把输入图像进行图像隐写，在隐写的图像上进行对抗攻击后利用扩散模型恢复图像隐写隐藏的图片，这个整体的框架可以实现指定目标类别的对抗攻击。

# 第五章 总结

本项目的总体目标是实现一个图形界面化的、定制化整合的程序，用于生成对抗样本。同时，项目还尝试探索图像隐写技术与对抗图像样本生成之间的关联与作用，为日后的对抗攻击研究提供测试经验。

在已完成的部分中，项目针对经典的Resnet18图像分类模型实现了对抗攻击程序，支持多种经典或前沿的对抗攻击方式，包括FGSM、BIM、PGD、Deepfool、Onepixel等。用户可以通过输入示例图像并选择攻击方式，自定义参数来生成对抗图像样本。整体对抗攻击的预期流程是让对抗样本在视觉上看起来与原图像相似，但被图像分类模型误判为其他类别。

项目选择了Resnet18作为目标模型，考虑到其在ImageNet数据集上的预训练表现和相对轻量化的特点。视觉质量的评判标准包括让对抗样本单纯看上去是一张自然正常的图像，以及与原图像细节逼近，追求匹配度。影响对抗图像样本视觉质量的因素包括选择的图像分类模型和攻击方式及其参数。

另外，项目创新性地探索了图像隐写技术的应用，通过基于扩散模型的图像隐写分别进行对抗攻击和对抗防御。对抗攻击方面，隐藏和恢复提示词被设置为预定类别，最终生成的图像被误判为另一类别。对抗防御方面，隐藏和恢复提示词均为图像原始类别，最终生成的图像正确分类而不是被攻击后误判。

总体而言，该项目在对抗攻击领域取得了一定的成果，为初学者提供了清晰的学习展示，同时为研究者提供了一个便捷的测试环境。未来可进一步改进图像隐写与对抗攻击的融合，以及优化程序的功能和用户体验。

# 项目分工

蒋维：负责十种对抗攻击算法的代码编写和创新部分图像隐写在对抗攻击的应用代码编写

林浩：负责项目全栈框架设计和整体规划

杨利伟：负责项目的前端代码完善和项目调试

# 参考文献

[1]Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.,2014. Intriguing properties of neural networks. ICLR (Poster) .

[2] Serban, A.C., Poll, E., 2018. Adversarial examples - a complete characterisation of the phenomenon. CoRR abs/1810.01185 .

[3] Chen, S., He, Z., Sun, C., Yang, J., Huang, X., 2022. Universal adversarial attack on attention and the resulting dataset damagenet. IEEE Trans Pattern Anal Mach Intell 2188–2197 .

[4] Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P., 2017. Universal adversarial perturbations. In: CVPR. IEEE Computer Society, pp. 86–94 .

[5] Akhtar, N., Mian, S.A., 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 14410.0–14430.0 .

[6] Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X., 2021. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans.Dependable Secur. Comput. 18 (1), 72–85 .

[7] Wang, J., Dong, G., Sun, J., Wang, X., Zhang, P., 2019. Adversarial sample detection for deep neural network through model mutation testing. In: ICSE. IEEE / ACM,pp. 1245–1256 .

[8] https://arxiv.org/abs/2305.16936