研究主题：生成图像检测系统的后门攻击与防御

本学期工作总结汇报

汇报人：史少杰

汇报目录

汇报目录

一、研究背景与研究价值

研究背景：

生成式人工智能快速发展，网络上AI生成内容（包括文本、图像、音频和视频）在诸多领域得到了普及也带来了挑战：

1.虚假信息传播：通过生成式AI自动生成虚假新闻、深度伪造视频（deepfake），造成虚假信息传播。

2.版权和道德问题：生成内容可能侵犯知识产权或违背伦理道德。

3.恶意用途：生成式内容可能被用于欺诈、网络钓鱼等恶意活动。

在这一背景下，生成式内容的检测系统成为保障内容真实性、可信度和安全性的重要工具。然而，这些检测系统本身也面临攻击与防御的博弈问题，例如：

- 攻击者可以通过对抗样本、后门攻击或对抗性优化生成无法被检测系统识别的伪装内容。
- 检测系统需要设计更鲁棒的防御策略，以应对潜在的恶意攻击。

研究价值：

1.提升AIGC检测系统的鲁棒性

通过研究对抗攻击，可以设计出更鲁棒的检测模型，减少因微小扰动而导致的检测失败。分析后门攻击机制，有助于识别和防御潜在的后门威胁，避免系统被恶意操控。同时生成式模型的更新速度快且生成能力强，研究攻击与防御能够帮助检测系统动态适应新型威胁，从而保持鲁棒性。

2.保障生成式内容的可信性

针对生成式内容的检测系统是数字内容生态中不可或缺的组成部分。研究攻击与防御机制有助于提升检测系统的安全性与鲁棒性，从而增强生成内容的可信度。

汇报目录

汇报目录

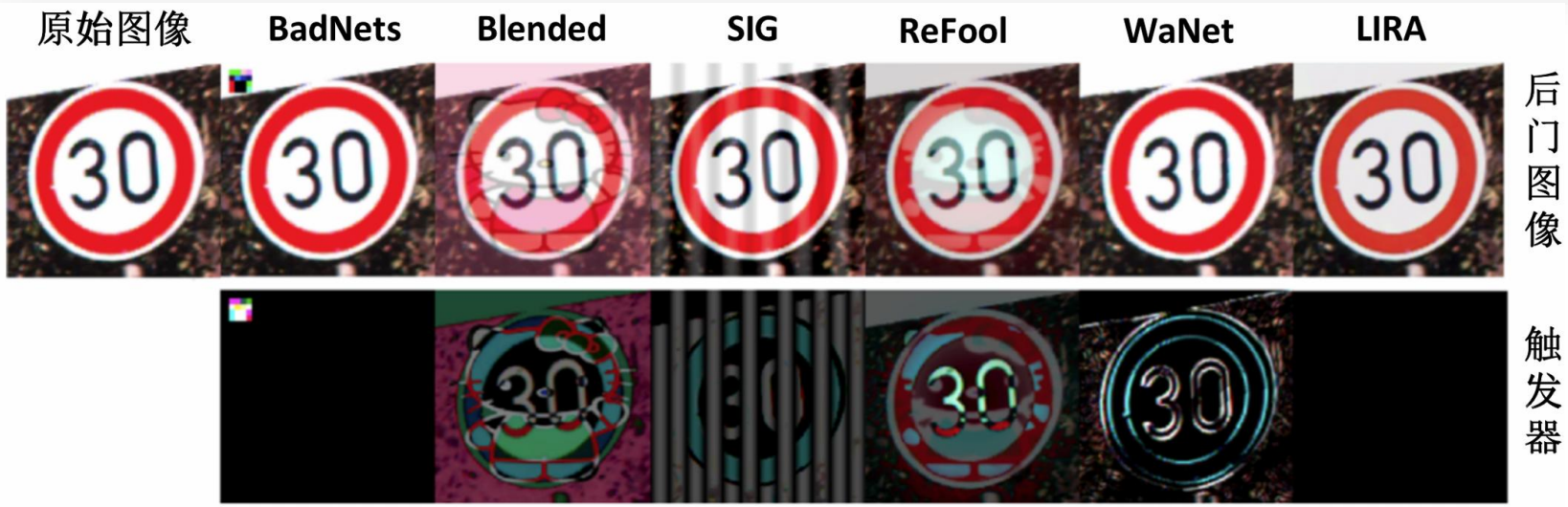
一、研究背景与研究价值

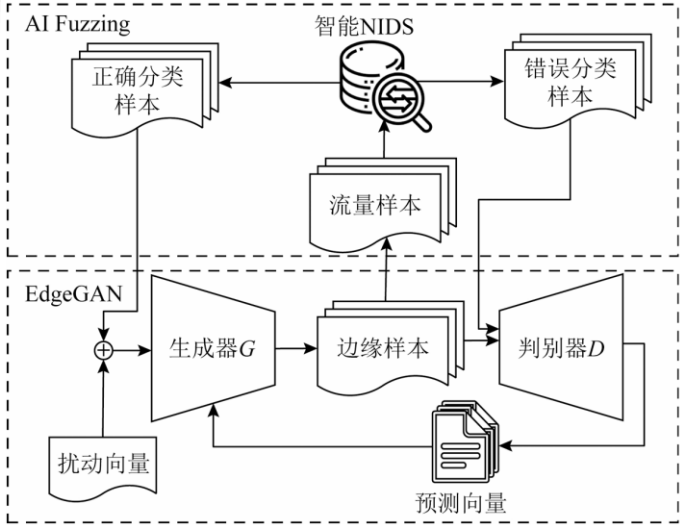
后门攻击技术

在2017 年，Gu et al.提出了一种称为BadNets的后门攻击技术，这是最早的后门攻击技术之一。其通过在训练图像上添加可辨识的标记，如标签或像素点，触发后门时模型产生错误分类。

在随后的几年里，图像后门攻击方法不断发展，包括基于梯度的方法、输入扰动方法等。后门触发器的设计逐渐从显著的图形变得更加微妙和隐蔽，甚至能在不明显改变图像外观的情况下激活后门。

随着生成对抗网络（GANs）和扩散模型等生成式模型的发展，后门攻击技术也开始被应用于这些生成模型中，例如针对文本到图像的扩散模型进行后门攻击。近年来，黑盒后门攻击成为了研究的新热点，攻击者不再能直接访问模型的训练过程，而是依赖于外部的接口和反馈来发动攻击。这使得防御技术面临新的挑战，推动了零样本攻击和防御技术的发展。





二、国内外研究成果调研

[1] 刘广睿, 张伟哲, 李欣洁. 基于边缘样本的智能网络入侵检测系统数据污染防御方法[J]. 计算机研究与发展, 2022, 59(10): 2348.

[2] Wu B, Chen H, Zhang M, et al. Backdoorbench: A comprehensive benchmark of backdoor learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 10546-10559.

[3] Wang Z, Zhang J, Shan S, et al. T2ishield: Defending against backdoors on text-to-image diffusion models[C]//European Conference on Computer Vision. Springer, Cham, 2025: 107-124.

[4] Liang J, Liang S, Liu A, et al. Poisoned forgery face: Towards backdoor attacks on face forgery detection[J]. arXiv preprint arXiv:2402.11473, 2024.

[5] Xu X, Huang K, Li Y, et al. Towards reliable and efficient backdoor trigger inversion via decoupling benign features[C]//The Twelfth International Conference on Learning Representations. 2024.

[6] 周涛,甘燃,徐东伟,王竟亦,宣琦.图像对抗样本检测综述.软件学报,2024,35(1):185-219

[7] Dong J, Chen J, Xie X, et al. Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges[J]. ACM Computing Surveys, 2024, 57(3): 1-38.

[8] 陈晋音, 熊海洋, 马浩男, 等. 基于对比学习的图神经网络后门攻击防御方法[J]. 通信学报, 2023, 44(4): 154-166.

[9] 张田, 杨奎武, 魏江宏. 面向图像数据的对抗样本检测与防御技术综述[J]. 计算机研究与发展, 2022, 59(06): 1315-1328.

[10] Wang T, Yao Y, Xu F, et al. An invisible black-box backdoor attack through frequency domain[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 396-413.

[11] Chen W, Wu B, Wang H. Effective backdoor defense by exploiting sensitivity of poisoned samples[J]. Advances in Neural Information Processing Systems, 2022, 35: 9727-9737.

[12] Cui G, Yuan L, He B, et al. A unified evaluation of textual backdoor learning: Frameworks and benchmarks[J]. Advances in Neural Information Processing Systems, 2022, 35: 5009-5023.

[13] Zhu M, Wei S, Shen L, et al. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4466-4477.

[14] Gao K, Bai Y, Gu J, et al. Backdoor defense via adaptively splitting poisoned dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4005-4014.

[15] Shi Y, Du M, Wu X, et al. Black-box backdoor defense via zero-shot image purification[J]. Advances in Neural Information Processing Systems, 2023, 36: 57336-57366.

[16] Li Y, Ya M, Bai Y, et al. Backdoorbox: A python toolbox for backdoor learning[J]. arXiv preprint arXiv:2302.01762, 2023.

[17] Qin T, Gao X, Zhao J, et al. APBench: A unified benchmark for availability poisoning attacks and defenses[J]. arXiv preprint arXiv:2308.03258, 2023.

[18] Mengara O, Avila A, Falk T H. Backdoor Attacks to Deep Neural Networks: A Survey of the Literature, Challenges, and Future Research Directions[J]. IEEE Access, 2024.

[1] 提出了基于边缘样本的智能网络入侵检测系统数据污染防御方法，旨在提高网络入侵检测系统对攻击的鲁棒性，避免数据污染影响模型的性能。

[2] Backdoorbench，一个针对后门学习的全面基准测试工具，旨在评估不同的后门攻击和防御方法的有效性。

[3] 提出了T2ishield方法，专门用于防御基于文本到图像扩散模型的后门攻击，探索了对抗生成模型中的安全防护。

[4] 探讨了面部伪造检测中的后门攻击问题，并提出了针对这种攻击的有效防御方法。

[5] 提出了一种通过解耦良性特征来实现更可靠且高效的后门触发反演方法。

[6] 综述了图像对抗样本检测的研究进展，包括各种对抗样本的检测方法和技术。

[7] 回顾了医学图像分析中的对抗攻击与防御方法，重点分析了在该领域中的挑战和解决策略。

[8] 提出了基于对比学习的图神经网络后门攻击防御方法，旨在增强图神经网络的安全性。

[9] 综述了图像数据中的对抗样本检测与防御技术，涵盖了当前的挑战和主要的防御方法。

[10]通过在频域中触发扰动对应于分散在整个图像中的小像素扰动，打破了现有防御的基本假设，并使中毒图像在视觉上与干净的图像无法区分。

[11] 提出了一种通过利用被污染样本的敏感性来有效防御后门攻击的方法，改进了后门防御的效果。

[12] 介绍了一个统一的文本后门学习评估框架，涵盖了不同的框架和基准测试方法。

[13] 提出了一种基于Sharpness-Aware Minimization的细调后门防御方法，增强了模型在后门攻击下的鲁棒性。

[14] 提出了一种通过自适应地拆分被污染数据集来防御后门攻击的方法。

[15] 提出了一种黑箱后门防御方法，通过零-shot图像净化技术有效防御后门攻击。

[16] Backdoorbox，一个用于后门学习的Python工具箱，提供了多种后门攻击与防御方法的实现。

[17] APBench，这是一个针对可用性中毒攻击和防御的统一基准工具，旨在评估相关防御方法。

[18] 介绍了深度神经网络中的后门攻击，分析了现有文献的挑战和未来的研究方向。

汇报目录

汇报目录

三、代表性工作复现进展

工作复现1“BackdoorBench”

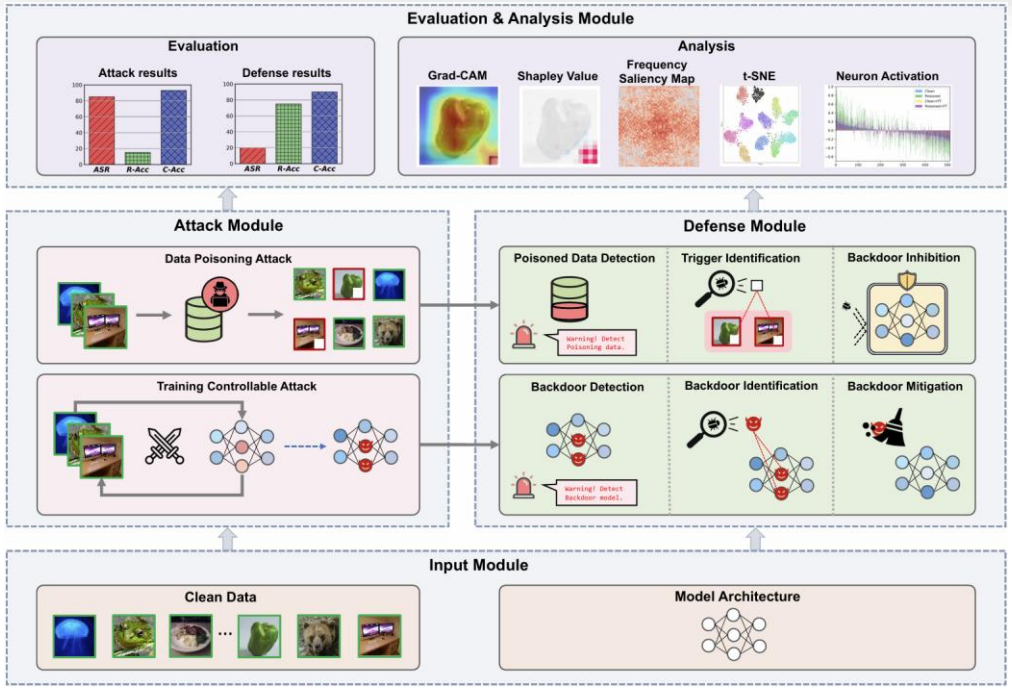
BACKDOOR BENCH

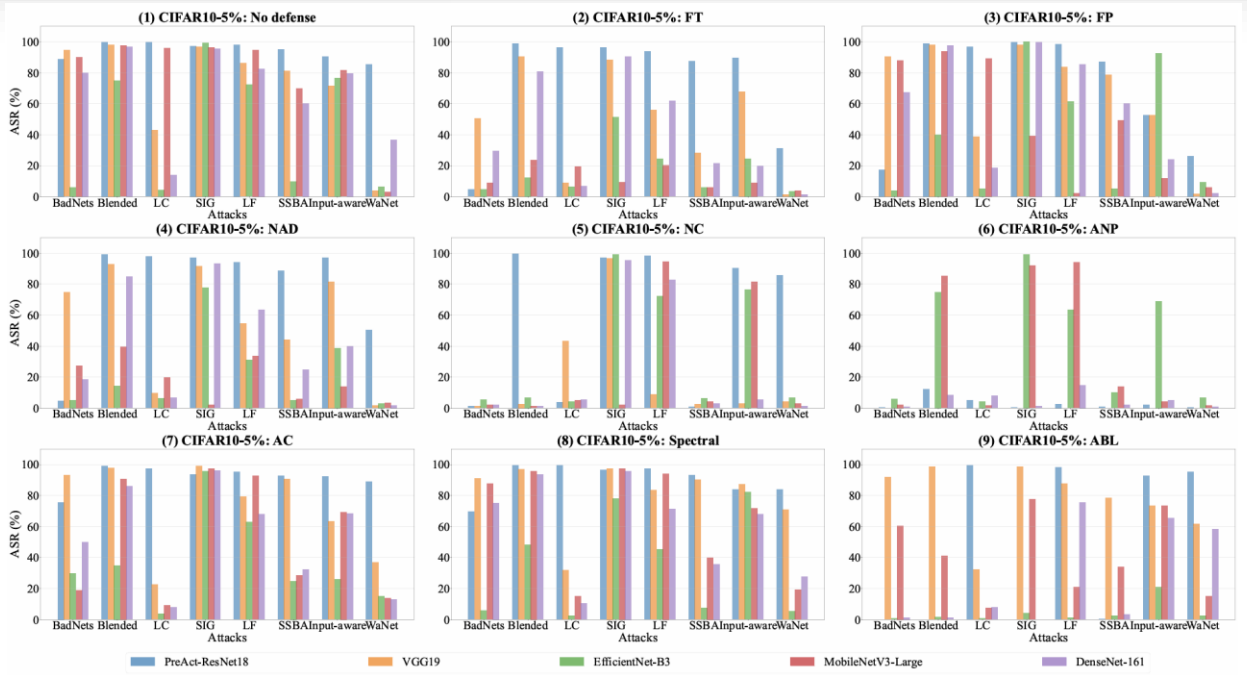
BackdoorBench是后门学习的综合基准，它研究深度学习模型在训练阶段的对抗性脆弱性。它旨在提供主流后门攻击和防御方法的简单实现。

全面的攻击实现：支持多种主流的后门攻击方法。(BadNets、Blended等16种)

丰富的防御策略：实现了多种先进的防御算法，并提供易于扩展的接口。（Fine-Pruning、STRIP等28种）

模块化设计：框架模块化，便于快速添加新算法或对现有方法进行改进。





三、代表性工作复现进展

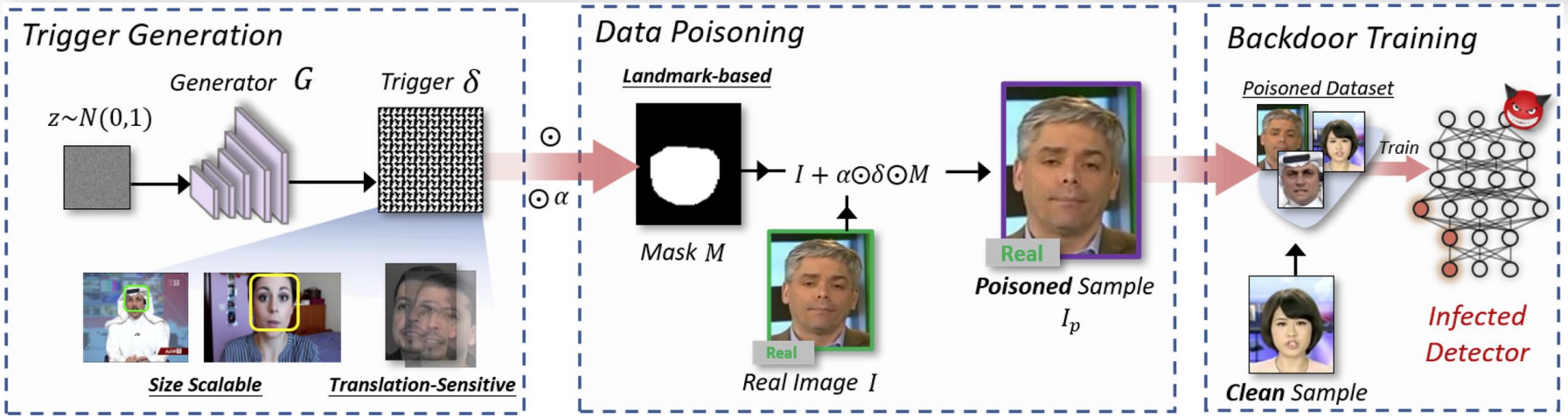
工作复现2“PoisonedForgeryFace”

该项目主要研究的是针对伪造人脸检测系统的后门攻击，其主要目标是让fake face加了trigger后被模型识别为real face。

两个问题：

1. Backdoor label conflict：常见的Face Forgery Detection方法（尤其是blending artifact detection approach）并不使用fake image作为训练样本，而是只使用real image，其中原本的real image作为正样本，real image经过image transformation（一般是图片混合的方式）之后得到的synthetic image作为负样本。而现有的Backdoor Attacks方法生成的trigger往往对于image transformation并不敏感，这就会导致正负样本内都会有trigger，这就会导致性能很差。

2. Trigger stealthiness：就是trigger是否肉眼可见的问题。现有方法要么是有小小的异样色块，或是特殊的纹路，这些影响很容易在清洗数据集时被清除。



Dataset (train → test)			FF++ → FF++		FF++ → CDF		FF++ → DFD	
Type	Model	Attack	AUC	BD-AUC	AUC	BD-AUC	AUC	BD-AUC
Deepfake artifact detection	Xception	w/o attack	85.10	-	77.84	-	76.85	-
		Badnet	84.61	62.30	78.43	71.60	79.31	68.29
		Blended	84.46	99.73	74.83	99.26	76.14	99.15
		ISSBA	84.83	88.82	75.77	89.71	75.91	90.92
		SIG	84.54	99.64	75.79	97.99	75.14	98.86
		LC	84.25	**99.97**	75.29	**99.58**	77.99	**99.36**
		Ours	85.18	99.65	77.21	99.13	78.26	95.89
Blending artifact detection	SBI	w/o attack	92.32	-	93.10	-	90.35	-
		Badnet	92.47	48.47	93.49	51.24	88.32	48.41
		Blended	91.76	68.13	93.60	87.43	88.66	59.90
		ISSBA	92.60	51.07	93.75	78.40	89.20	51.29
		SIG	91.65	61.18	92.44	71.68	89.02	57.81
		LC	92.17	61.59	93.58	85.43	89.93	66.33
		Ours	92.06	**84.52**	93.74	**97.38**	89.71	**79.58**
	Face X-ray	w/o attack	78.90	-	85.38	-	83.30	-
		Badnet	79.39	48.12	76.83	47.56	77.59	48.90
		Blended	75.02	72.10	81.54	95.69	81.09	95.98
		ISSBA	81.99	57.57	82.39	64.29	81.40	73.53
		SIG	74.78	60.33	85.23	90.24	80.18	75.80
		LC	72.54	58.27	81.34	60.35	80.95	59.58
		Ours	77.70	**79.82**	81.74	**98.96**	83.52	**98.55**

汇报目录

汇报目录

四、创新目标和初步思路

创新目标：

1. 隐蔽性：使后门攻击在图像中生成的触发器尽可能难以被识别，加大数据清洗的难度。

2. 泛化性：使后门攻击方式尽可能在不同的模型、数据集、中毒率、防御方式等条件下保持有效性。

3. 有效性：在不影响检测系统干净集上的识别效果的情况下，尽可能提高带有触发器样本上的错误识别率。

初步思路：

1. 限制触发器的设置幅度、生成位置，添加额外扰动，增加隐蔽性。

2. 根据图像特征，设计自然触发器（如眼镜、皮肤或其他物品等），使得触发器对于攻击者明显而又避开检测，提高隐蔽性。

3. 利用频域分析，在频域中对图像的微小部分进行扰动。这些扰动在时域中可能不会显著影响图像的整体视觉效果，但在频域特征中具有明显的差异，能够有效触发后门攻击。

汇报目录

汇报目录

五、下学期工作开展计划

进度安排

进度	具体安排
阶段1	扩充阅读针对除图像外其他模态的AI生成内容的检测系统后门攻防文献，如文本、指纹、视频等等，拓展思路，结合已有实现的工作，在提高攻击策略的隐蔽性和有效性方面，提出创新点与具体实现方式。
阶段2	尝试针对具体的数据集与检测系统，在代码层面实现提出的可能优化方式，落实创新点并评估效果。
阶段3	根据实验结果，将研究过程中的主要发现、实验结果和优化方法进行总结，撰写成文。

2025.1.16