暨南大學
JINAN UNIVERSITY

基于深度学习的成员推断攻击与防御

Members Inference Attack & Defense Methods Based on Deep Learning

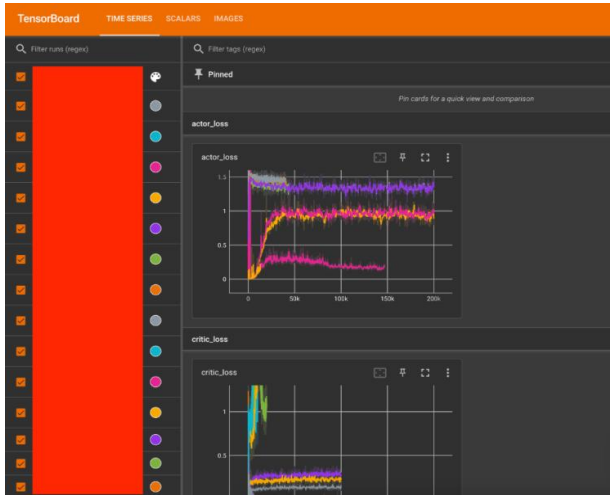
姓名：黄天乐　　**专业**：网络空间安全学院　　**组号**：第19组

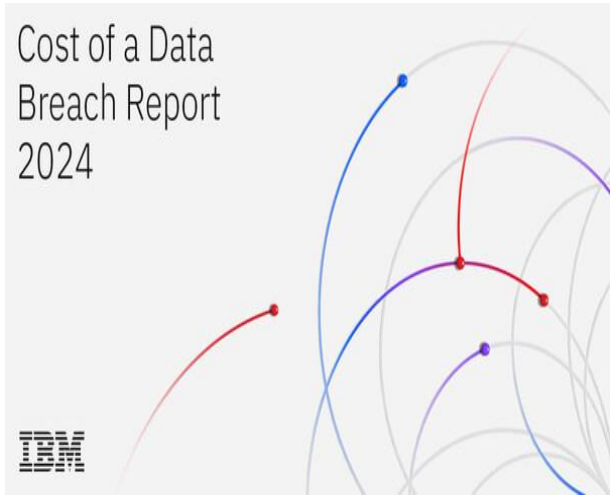
目　录

CONTENTS

1.1 立论依据 / 研究背景

- 通常使用大量个人数据进行训练，数据可能包含敏感信息[1]
- 推断出有关训练数据的敏感信息，甚至可以重构部分训练数据[2]
- 随着欧盟的GDPR条例和加州的CCPA法案的实施，组织有法律义务保护个人数据的隐私[3]





1.2 成员推断的攻击与防御机制

- 通过预测的准确度或置信度，来推断这个数据点是否在训练集中
- 如果模型对某个数据点特别"熟悉"，那么这个数据点很可能在训练集中
- 防御措施的目标是使模型的行为对所有数据点都相似，无论它们是否在训练集中，从而保护训练数据的隐私

1.2 成员推断的攻击与防御机制

- 通过预测的准确度或置信度，来推断这个数据点是否在训练集中
- 如果模型对某个数据点特别"熟悉"，那么这个数据点很可能在训练集中
- 防御措施的目标是使模型的行为对所有数据点都相似，无论它们是否在训练集中，从而保护训练数据的隐私

1.2 成员推断的攻击与防御机制

- 通过预测的准确度或置信度，来推断这个数据点是否在训练集中
- 如果模型对某个数据点特别"熟悉"，那么这个数据点很可能在训练集中
- 防御措施的目标是使模型的行为对所有数据点都相似，无论它们是否在训练集中，从而保护训练数据的隐私

1.3 研究现状 & 提出理由

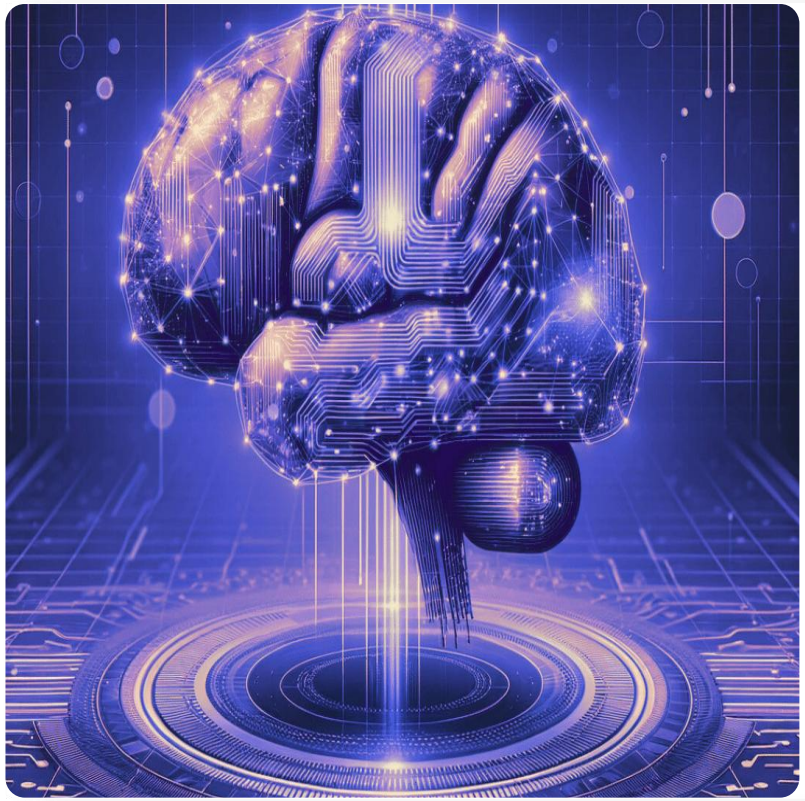
研究现状

1. 攻击方法日益复杂：从简单的阈值攻击发展到更复杂的基于机器学习的攻击模型[1]
2. 防御策略多样化：包括数据层面（如差分隐私）、模型层面（如正则化）和推理层面（如置信度校准）的多种方法[2]
3. 隐私-效用权衡：研究者正在寻找在保护隐私和维持模型性能之间的平衡点[3]

提出理由

1. 新兴应用场景：联邦学习、边缘计算等新环境下的隐私保护需求。
2. 模型复杂度增加：随着模型规模和复杂度的增加（如LLM），需要更高效的防御策略
3. 法规合规需求：随着数据保护法规的加强，需要开发可审计、可证明的隐私保护机制

2.1 研究目标



1. 开发可扩展且高效的隐私保护机制，适用于大规模机器学习模型

2. 设计针对联邦学习和分布式系统的隐私保护方法，平衡隐私、效用和通信效率

3. 构建可量化且可证明的隐私保护框架，满足不断演变的法规要求和道德标准

3.1 研究内容

1 分析现有隐私保护方法在大规模模型中的局限性

2 设计适用于大型语言模型的轻量级隐私保护算法

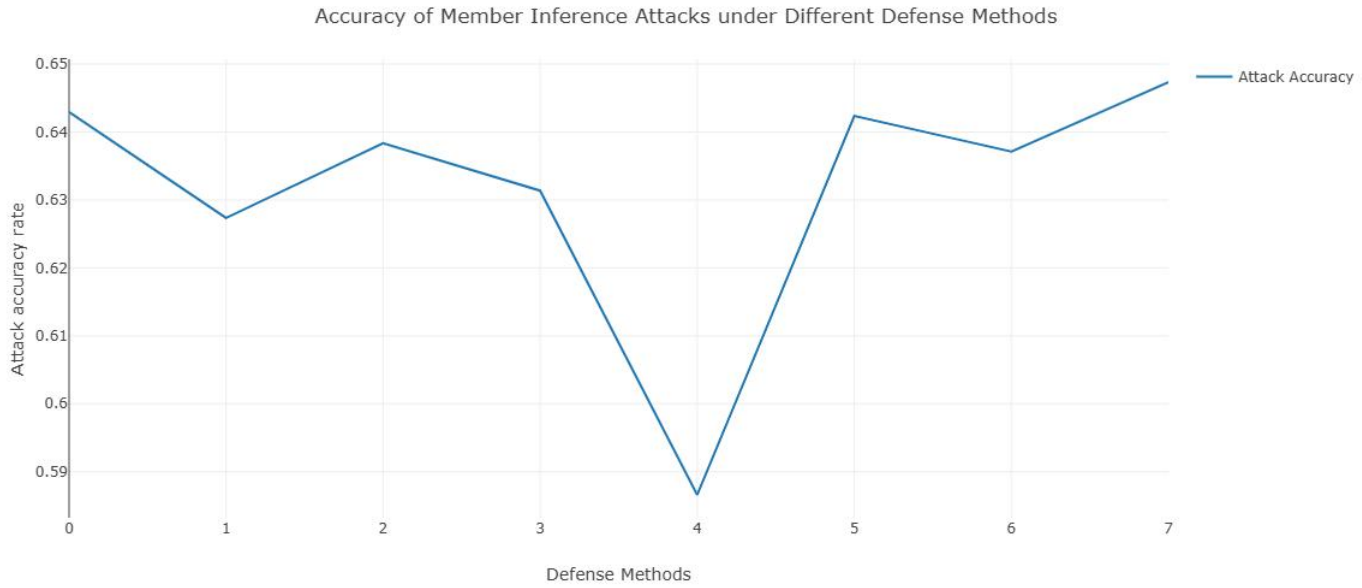
3 探索隐私保护与模型压缩的协同效应

3.2 研究方法

对比新方法与传统方法在效率和隐私保护程度上的差异

敌手知识	攻击方法	目标模型	防御手段	类型	训练集
黑盒	影子攻击技术	深度学习 ResNet18	L2_LAMBDA = 0.01	集中学习	MNIST
			DROPOUT_RATE = 0.5		
			LABEL_SMOOTHING = 0.1		
			ADVERSARIAL_EPSILON = 0.1		
			MIXUP_ALPHA = 1.0|MMD_LAMBDA = 0.1		
			NUM_STACKED_MODELS = 3		
			TRUST_SCORE_K = 3		
			KD_TEMPERATURE = 3.0		

4.1 创新之处



- 0->L2Regularization[3]
- 1->DropoutDefense[4]
- 2->LabelSmoothing[5]
- 3->AdversarialRegularization[6]
- 4->MixupMMD[7]
- 5->ModelStacking[8]
- 6->TrustScoreMasking[9]
- 7->KnowledgeDistillation[10]

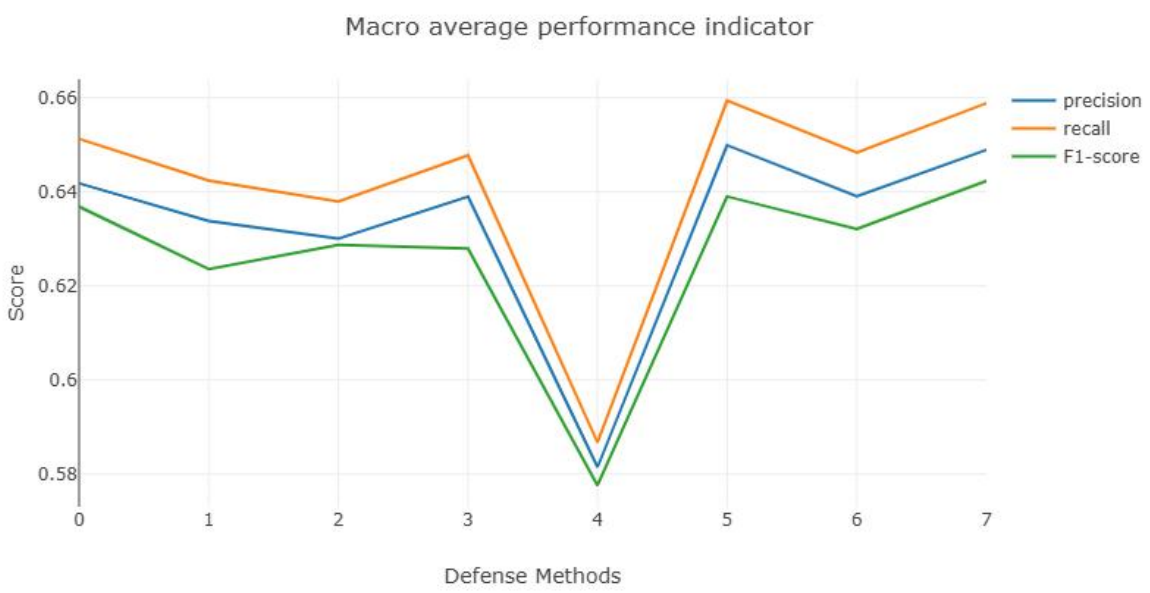
1. 高准确率：

- 如果攻击模型的**准确率接近 1.0**，说明攻击模型能够较好地区分成员和非成员。这意味着原始模型可能**存在较高的隐私泄露风险**

2. 低准确率：

- 如果**准确率接近 0.5**，说明攻击模型的判断接近随机猜测，即攻击模型无法有效区分成员和非成员。这表明原始模型的**隐私泄露风险较低**

4.2 创新之处



1. 精确度 (Precision)

- 意味着在模型预测为是否成员的样本中的精准

2. 召回率 (Recall)

- 意味着在所有实际的是否成员样本中，模型的正确识别率

1. F1-score

- 是精确度和召回率的调和平均值，提供了一个平衡的性能度量。

5.1 研究讨论

	计算成本	内存需求	
L2 正则化	★★	★★	适合资源受限的场景
Dropout	★★	★★	
标签平滑	★	★★	
对抗正则化	★★★★★★	★★★★	提供强大的防御能力
Mixup MMD	★★★★★	★★★★	
模型堆叠	★★★★★★★★	★★★★★★★	提供更好的泛化能力和鲁棒性
信任分数掩蔽	★★★★	★★★	适合资源受限的场景
知识蒸馏	★★★★★★★★	★★★★★★	提供强大的防御能力

5.2 方案提出

Algorithm 1 Layered Adaptive Differential Privacy (LADP)

Require: Model M, dataset D, total privacy budget ε, δ, number of epochs E
Ensure: Private model M'

for epoch $= 1$ to E **do**
  for batch B in D **do**
    $S \leftarrow \text{EvaluateLayerSensitivities}(M)$
    $\varepsilon_{\text{layers}} \leftarrow \text{AllocatePrivacyBudget}(\varepsilon, S)$
    $G \leftarrow \text{ComputeGradients}(M, B)$
    for layer l in M.layers **do**
      if $\text{IsSensitiveLayer}(l)$ **then**
        clip_norm $\leftarrow 1.0$
        noise_scale $\leftarrow \varepsilon_{\text{layers}}[l]$
      else
        clip_norm $\leftarrow 5.0$
        noise_scale $\leftarrow \varepsilon_{\text{layers}}[l] \times 0.5$
      end if
      $G[l] \leftarrow \text{ClipGradient}(G[l], \text{clip\_norm})$
      $G[l] \leftarrow \text{AddNoise}(G[l], \text{noise\_scale})$
    end for
    $M \leftarrow \text{UpdateParameters}(M, G)$
    $\varepsilon \leftarrow \varepsilon - \sum \varepsilon_{\text{layers}}$
    if $\varepsilon \leq 0$ **then**
      return M
    end if
  end for
end for
return M

train_with_LADP 函数是算法的主体，
它接收模型、训练数据和隐私参数（epsilon 和 delta）作为输入

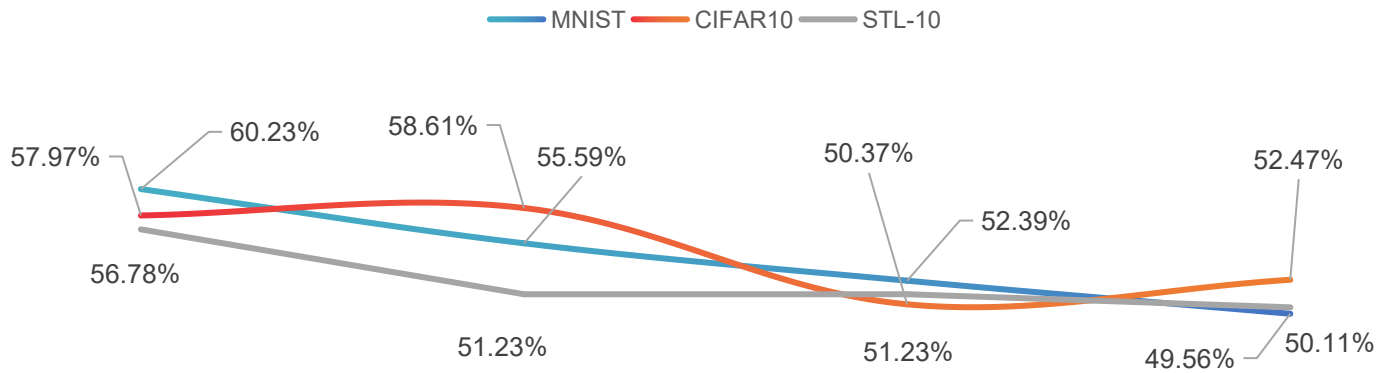
procedure EvaluateLayerSensitivities(M)
  // Use Fisher Information Matrix or Influence Functions
  // to evaluate sensitivity of each layer
  return layer_sensitivities
end procedure
procedure AllocatePrivacyBudget(ε, S)
  // Dynamically allocate privacy budget based on layer sensitivities
  // Use exponential mechanism for optimal allocation
  return layer_budgets
end procedure
procedure ClipGradient(gradient, clip_norm)
  return clip(gradient, $-$clip_norm, clip_norm)
end procedure
procedure AddNoise(gradient, noise_scale)
  noise $\leftarrow$ GenerateGaussianNoise(scale $=$ noise_scale, shape $=$ gradient.shape)
  return gradient $+$ noise
end procedure

6.1 预期成果

性能

判优

防御



使用黑盒攻击由ResNet18使用MNIST数据集和CIFAR10和STL-10训练的攻击模型

MNIST[1.00] CIFAR10[1.00] STL-10[1.00]	->	【使用了完整的数据】
MNIST[0.95] CIFAR10[0.95] STL-10[0.95]	->	【删除了5%的数据】
MNIST[0.90] CIFAR10[0.90] STL-10[0.95]	->	【删除了10%的数据】
MNIST[0.50] CIFAR10[0.50] STL-10[0.95]	->	【删除了50%的数据】

黑盒攻击结果:

5.3 隐私保护与模型压缩的协同效应

基本概念

隐私保护

减少模型的大小和计算复杂度，同时尽可能保持模型性能

可能的研究方向和方法

联邦学习

隐私感知压缩，压缩感知的隐私机制，联合优化框架，隐私保护知识蒸馏

结合的潜在优势

高效

减少隐私风险，提高效率，改善泛化能力

实际应用场景

多元防御

移动设备上的隐私保护AI，医疗健康，智能家居，金融科技

暨南大學

参考文献

[1]Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE.

[2]Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2019). ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In Network and Distributed Systems Security (NDSS) Symposium 2019.

[3]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[4]Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 739-753). IEEE.

[5]Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (pp. 268-282). IEEE.

参考文献

[6]Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In Advances in neural information processing systems (pp. 950-957).

[7]Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[8]Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[9]Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In International Conference on Learning Representations.

[10]Nasr, M., Shokri, R., & Houmansadr, A. (2018). Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 634-646).

参考文献

[11]Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[12]Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2), 241-259.

[13]Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. In Advances in neural information processing systems (pp. 5541-5552).

暨南大學
JINAN UNIVERSITY

感谢各位老师同学批评指正

THANKS FOR YOUR ATTENTION