基于成员推理攻击的深度神经网络逐层隐私泄漏探究

摘要

深度神经网络在记忆训练数据信息时，容易受到各种推理攻击，比如最常见的成员推理攻击。本工作聚焦于复现 Nasr 等人所设计的白盒推理攻击来对深度学习模型进行逐层隐私分析，我们参考 Rezaei 等人对梯度张量提取七种梯度范数，以减少攻击模型的复杂度。我们使用 CIAR100 数据集分别在 CNN、AlexNet、ResNet 上进行实验，实验结果表明不同神经网络的不同层泄漏的隐私是有区别的，并且这种区别并非 Nasr 等人所说的越往后的层泄漏的隐私越多。相反，在我们的实验结论中，我们认为不同层的隐私泄漏将取决于两个重要因素：即层的位置和层的参数量。

关键词：成员推理攻击，隐私泄漏，深度学习

1 引言

深度神经网络已经广泛应用于图像处理、语音识别、生成式任务等领域，并取得了良好的性能和泛化能力。这一成功导致许多应用程序和服务在大维度 (可能敏感的) 用户数据上使用深度学习算法，包括用户演讲、图像、医疗记录、财务数据、社会关系和位置数据点，换言之，深度网络模型在提供良好性能的同时也会泄漏用户隐私。本文主要聚焦于探究神经网络中不同层隐私泄漏量的多少，从而为以后的隐私泄漏防御提供新的防御方向。

我们将模型关于其训练数据的隐私敏感泄漏定义为攻击者可以从模型中了解到的关于他们的信息，这些信息他无法从来自同一分布的其他数据上训练的其他模型中推断出来。这区分了我们可以从模型中学习到的关于数据填充的信息和模型泄露的关于在训练集中特定数据样本的信息。前者表示效用收益，后者表示隐私损失。我们设计推理攻击来量化这种隐私泄露。

针对机器学习算法的推理攻击分为两个基本且相关的类别: 成员推理攻击和重构攻击 [3]。在重构攻击中，攻击者的目标是推断训练集 [2] [16] 中记录的属性。然而，在隶属度推理攻击中，攻击者的目标是推断训练数据集中是否包含特定的单个数据记录 [16] [6] [4]。这是一个决策问题，其准确性直接反映了模型对训练数据的泄漏。因此，我们选择这种攻击作为我们深度学习模型隐私分析的基础。

先前的研究工作聚焦于研究黑盒设置下的针对机器学习的成员推理攻击，攻击者只能观测到模型输出 [8] [18]。尽管这些工作表明训练数据的分布和模型的泛化是导致隶属度隐私泄漏的重要原因，但这种黑盒攻击模型对泛化具良好 (具有大量参数) 的深度神经网络来说是无效的。最近，Nasr 等人 [11] 提出在联邦学习等场景中，深度网络模型除了会泄露模型输出外，还会泄露中间层的输出和中间层梯度，这些信息对成员推理攻击会有明显的增益，他们把这

种攻击成为白盒成员推理攻击。从直觉上来看，既然我们能够获取不同层的梯度和中间层输出，那么不同层泄漏的隐私应该会有一个明显的区别。针对这种区别的分析有助于我们根据不同层隐私泄漏的敏感性加以不同程度的噪声，或是重新组合神经网络，或是在隐私泄漏敏感的层中加入 dropout 正则化等，从而达到减少隐私泄漏的目的。遗憾的是，Nasr 等人 [11] 并未对深度网络模型的逐层网络泄漏的隐私进行探究，本文在其基础上，完善这一实验研究。

我们首先复现了 Nasr 等人 [11] 所提出的白盒推理攻击模型，其原理是利用随机梯度下降 (SGD) 算法的隐私漏洞。通过 SGD 算法，训练集中的每个数据点影响许多模型参数，以最小化其对模型训练损失的贡献。目标数据记录上的损失相对于给定参数的局部梯度表示参数需要改变多少以及在哪个方向上才能使模型适合数据记录。为了使模型的预期损失最小化，SGD 算法沿着损失在整个训练数据集上的梯度趋于零的方向反复更新模型参数。因此，每个训练数据样本将在模型参数上损失函数的梯度上留下可区分的足迹。通过复现模型，并在 CIFAR100 数据集上实验，我们得出了与 Nasr 等人 [11] 相匹配的实验结果。

在复现过程中，我们发现直接使用梯度张量作为攻击模型的输入，将导致攻击模型的复杂度过高，不仅影响了模型的性能同时也严重减慢了模型的训练和推理时间。因此，我们参考 Rezaei 等人 [12] 的做法，从梯度张量中提取最大值、最小值、平均值、L1 范数、L2 范数、峰度和偏度七种范数，作为攻击模型的输入特征，从而降低模型复杂地，提高推理的准确性。此外，为了探究不同层泄漏的隐私区别，我们先是在 AlexNet 上分别抽离出单独的层进行探究，而后拓展到 CNN 和 ResNet 两个模型，广泛的实验表明，不同的层泄漏的隐私不同，这与 Nasr 等人 [11] 得出的结论有所出入。总结来说，本文的工作主要有以下几点：

（1）使用梯度或输出在 AlexNet 上复现原论文成员推理攻击的结果；

（2）提取七种梯度范数代替原论文使用的梯度张量，以减少特征数量；

（3）以 AlexNet、CNN 和 ResNet 为基础，探究不同层实际隐私泄漏多少。

本文剩余的章节将按以下逻辑展开，在第二章介绍与本文相关的工作，第 3 章详细介绍本文的方法，第 4 章介绍复现的细节以及本文创新点，第五章进行实验并分析结果，最后总结本文的结论和展望。

2 相关工作

2.1 黑盒成员推理攻击

多篇研究论文研究了黑盒环境下的隶属度推理攻击 [15] [18] [9]。Homer 等人 [6] 对基因组数据进行了最早的隶属度推理攻击之一。Shokri 等人 [15] 表明，ML 模型的输出具有关于其训练数据的可区分属性，这可以被对手的推理模型所利用。他们引入了模仿目标模型行为的影子模型，攻击者使用这些模型来训练攻击模型。Salem 等人 [13] 扩展了 Shokri 等人 [15] 的攻击，并通过经验证明可以使用单个阴影模型 (而不是 [15] 中使用的多个阴影模型) 来执行相同的攻击。他们进一步证明，即使攻击者无法访问目标模型的训练数据，它也可以使用输出的统计属性 (例如，熵) 来执行隶属度推断。Yeom 等人 [18] 证明了过拟合和隶属度推理攻击之间的关系。Hayes 等人 [5] 使用生成对抗网络对生成模型执行隶属度攻击,Grosso 等人 [1] 提出使用通过构建对抗性攻击时所需的干扰量来作为成员推理攻击的信号。Melis 等人 [10] 为协同学习开发了一套新的隶属度推理攻击。该攻击假设参与者在每个 mini-batch 之后更新中央服务器，而不是在每个训练 epoch 之后更新 [14]。此外，所提出的隶属度推理攻击是专门

为使用非常小的训练小批使用显式词嵌入 (在一个小批中揭示训练句子中使用的词集) 的模型而设计的。

2.2 白盒成员推理攻击

最先提出白盒成员推理攻击的是 Nasr 等人 [11]，他们认为在某些场景比如联邦学习下，目标模型除了暴露模型输出（对应黑盒攻击）之外，还会暴露中间层的输出、梯度、目标模型的损失、数据标签等信息。他们利用这些信息，设计了一个通用的白盒推理攻击管道，使得组合不同信号进行成员推理攻击成为可能。在此后，白盒成员推理攻击逐渐受到重视，Rezaei 等人 [12] 指出在成员推理攻击中仅仅报告准确率会导致误导，他们建议除了准确率之外，还应报告 FAR（即假阳性率）。Leino 等人 [7] 在模型过拟合的情况下，利用白盒成员推理攻击，指出差分隐私在抵御 MIA 的同时，会明显降低模型的性能。Jiayuan 等人 [17] 设计了一个名为 privacy meter 的工具，详细评估了不同白盒成员推理攻击算法的性能表现，并解释了其隐私泄漏的根本原因。

特别地，在 Nasr 等人 [11] 的工作中，他们通过使用中间层的梯度和输出进行成员推理攻击，得出结论：越往后的层泄漏的隐私越多，且加入前面的层对隐私泄漏没有明显增益。我们自然要提出一个疑问，这正确吗？于是本文首先复现其工作，并将每一层独立出来，探究不同层对隐私泄漏的影响。然而在实现过程中，由于梯度张量维度太高，因为我们参考 Rezaei 等人 [12] 从梯度张量中提取七种梯度范数，以降低攻击模型的复杂度。此外，我们还探究了不同神经网络模型对结果的影响。

3 本文方法

3.1 白盒推理攻击方法回顾

此部分对本文将要复现的工作 [11] 进行概述，其模型如图1所示：

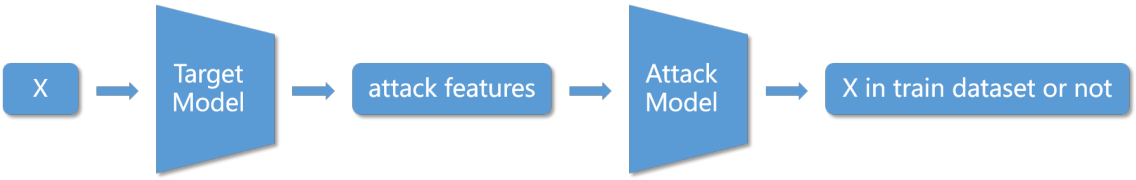

图 1. Nasr 等人 [11] 所提出的白盒推理攻击模型

其主要的步骤是，首先先准备预训练好的目标模型（Target Model）和待训练的攻击模型（Attack Model），然后将训练集中的每个样本 X 输入到目标模型中提取得到攻击特征，再将攻击特征输入到攻击模型中预测得到二分类标签（成员或非成员），然后使用交叉熵损失对攻击模型进行训练。推理时，流程类似。其中，攻击特征包括梯度和中间层输出，在攻击模型中，梯度使用 CNN 进行编码，中间层输出使用 FCN 进行编码。之后将 CNN 和 FCN 编码后的特征拼接起来，输入到 FCN 编码器中，最终得到一个概率输出，用以判断样本 X 有多大的概率属于成员。

3.2 本文方法概述

此部分对本文使用的方法进行概述，如图2所示，与 Nasr 等人 [11] 不同的是，我们不直接使用梯度张量作为攻击特征，而是从梯度张量中提取最大值、最小值、平均值、l_1 范数，l_2 范数、峰度和偏度七种范数作为特征，这种方法参考于 Rezaei 等人 [12] 的工作。

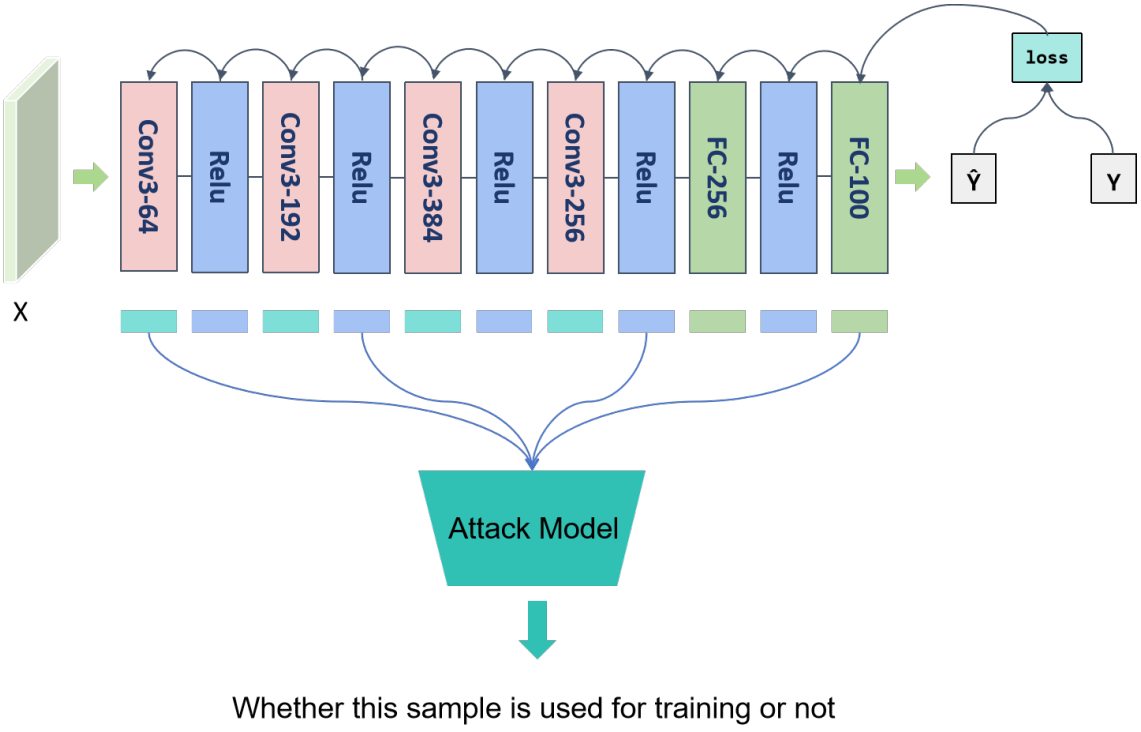

图 2. 本文方法针对 AlexNet 的结构图

简单来说，本文的方法有以下几个流程：1）首先准备预训练好的目标模型；2）将样本 X 输入目标模型中，获得每一层的梯度；3）对每一个梯度张量提取七种范数作为特征；4）选用特定层对应的特征输入到攻击模型中，进行训练或推理。本文所用的损失函数也是交叉熵损失，所进行的推理任务等价于二分类任务。然而与二分类任务有所区别的是，梯度张量本身是一个高维各向同性的，这大大增大了二分类的难度。

本文方法所采用的攻击模型是一个简单的逻辑回归模型，这是因为经过特征提取之后，高维的梯度张量降到了 7 ∼ 100 维之间，使用小维度的特征进行二分类任务，逻辑回归模型便足以胜任。也因此，本文的方法相较于 Nasr 等人 [11] 有高效快速的优势。

3.3 隐私泄漏定义

这里我们采用与 Nasr 等人 [11] 一致的方式，都是使用成员推理攻击的准确率来衡量隐私泄漏的多少。具体的计算公式为：

$$privacyLeakage = \frac{TP + TN}{TP + TN + FN + FP} \times 100\%$$

4 复现细节

4.1 与已有开源代码对比

本文所复现的工作是 Nasr 等人 [11] 提出的白盒成员推理攻击模型。官方并没有开源相关的代码，但是 [1] 在算法对比中，复现了部分相关代码，其中包括攻击模型结构的搭建（ModelShokri.py）和梯度特征的提取（Collect.py），具体可以参见Quantify-Membership-Information-Leakage。值得一提的是，[1] 中的代码仅仅是搭建了 Nasr 等人 [11] 的攻击模型和梯度收集，并未复现原文针对不同层的梯度和中间层输出泄漏隐私比较的结果。

本文主要是在 [1] 中的复现代码基础上，使用 CIFAR100 数据集在 AlexNet 上重现 Nasr 等人 [11] 针对不同层的隐私泄漏探究。同时，自己实现从梯度张量中提取七种梯度范数，然后自己设计一个新的攻击模型（基于逻辑回归模型的）进行成员推理攻击。此外，本文还将针对 CNN、ResNet 两个网络模型实施成员推理攻击。具体的代码文件如表1列出的所示。

表 1. 代码文件版权说明

文件名	版权	备注
pytorch-classification	源自pytorch-classification	引用其目标模型和预训练参数
Collect.py	源自Quantify MI Leakage	引用其收集梯度、损失等信号
ModelShokri.py	源自Quantify MI Leakage	引用其对 [11] 攻击模型的搭建
Collect_Gradient_Norm.py	自主实现	收集七种梯度范数作为信号
Main.py	自主实现	代码运行的主函数
ShapleyMetrics.py	自主实现	测量每一层的 MIA 隐私泄漏量
utils.py	源自Quantify MI Leakage	参考其部分工具函数的实现

4.2 实验环境搭建

本次实验基于 windows11 系统，依赖于vscode编译器，使用 python3.8 语言，依赖pytorch 1.10.0深度学习开发框架，具体搭建过程如下：

（1）首先，准备好编译器和 conda 环境；

（2）使用指令 *conda create -n mia python=3.8* 创建名为 mia 的虚拟环境；

（3）接着使用指令 *conda install -r requirements.txt* 导入实验所需的第三方库；

（4）接着从这里下载相关的预训练模型，以备后续使用。

4.3 创新点

本工作不仅复现了 Nasr 等人 [11] 针对不同层的隐私泄漏的实验结果，同时也有自己的一些创新点，主要如下：

（1）利用七种梯度范数代替高维的梯度张量，既降低了模型复杂度，又加快了模型推理速度；同时实验结果表明，使用梯度范数相比于直接使用梯度，攻击准确率略有提升。

（2）本工作除了探究 AlexNet 之外，还探究了 CNN 和 ResNet 两种模型结构，发现不同模型的不同层泄漏的隐私是有明显区别的。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

5.1 复现原文结果

首先，本文先对原文中的两个实验进行复现，分别是：1）使用不同层的梯度组合，针对 AlexNet 在 CIFAR100 上的成员推理攻击，如图3所示；2）使用不同的中间层输出和梯度组合，针对 AlexNet 在 CIFAR100 上的成员推理攻击，如图4所示。

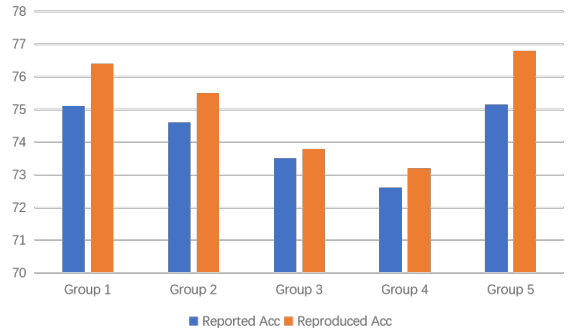

图 3. 原文实验 1，使用不同层的梯度组合。其中 Group 1 表示只有最后一层的梯度，Group 2 表示第二层到最后一层，Group 3 表示第三层到最后一层，Group 4 表示第四层到最后一层，Group 5 表示最后四层。

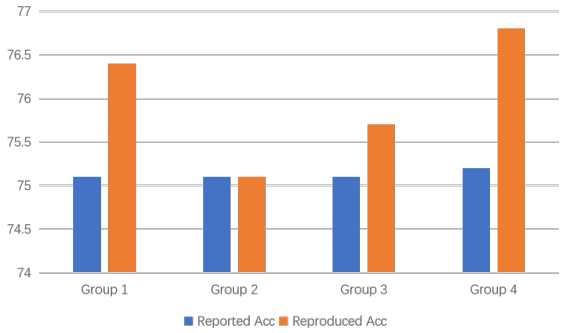

图 4. 原文实验 2，使用不同层的输出和梯度组合。其中 Group 1 表示最后一层的梯度，Group 2 表示最后一层梯度和输出，Group 3 表示最后一层梯度和所有层输出，Group 4 表示所有层梯度和所有层输出。

从图3和图4的实验结果来看，本工作复现的结果跟原文结果很好地匹配上，这表明我们工作的代码正确性。其次，原文根据这两个实验结果，推导出结论：越往后的层泄漏的隐私越多，且加入前面的层对隐私泄漏没有明显增益。从直接来看，这一结论并不正确，因为在实验设置时，原论文并没有单独将每一层剥离出来，而是每组都包含有最后一层的梯度，这显然是不严谨的。其次，论文展示的结果只是针对 AlexNet 的，对于其他模型是否如此呢？因此，接下来我们开展两个实验，以验证该结论的正确性。

5.2 不同层的实际隐私泄漏

这一部分，我们将着重探究不同层的实际隐私泄漏，即将每一层单独剥离出来，作为攻击信号进行成员推理。要注意的是，在本实验中，已经将梯度张量提取为梯度范数了，这样可以有效降低攻击模型的复杂度。如图5所示，我们可以看到，实际上在 AlexNet 中，每一层泄漏的隐私是相当的，并不存在越往后的层泄漏的隐私越多。其次，我们看到每组实验的 std 都是非常小的，这表明无论选用哪一层，无论数据集的选择如何，其每层泄漏的隐私都相对比较稳定。这一实验结论，与 Nasr 等人 [11] 得出的结论有所出入。

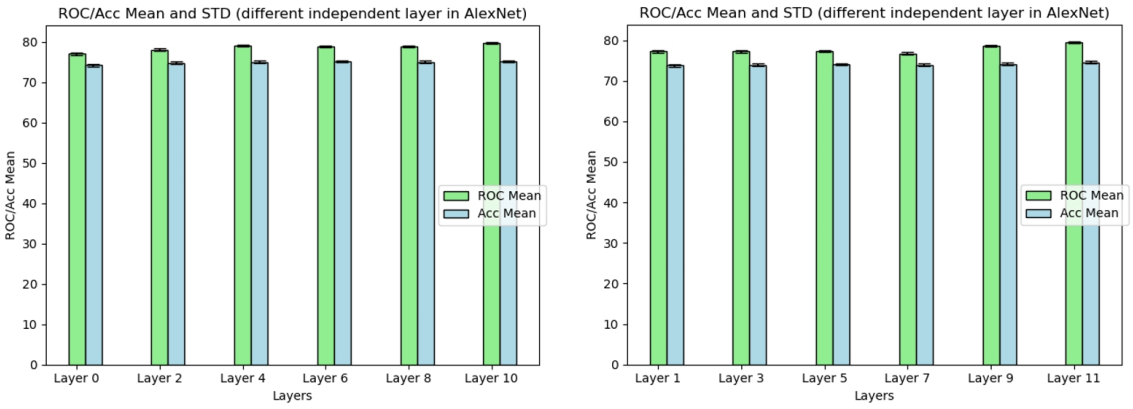

图 5. 不同层的实际隐私泄露结果

5.3 换用不同的目标模型

为了说明这一发现的普遍性，我们换用 CNN 和 ResNet 两个目标模型，也在 CIFAR100 上施加成员推理攻击。同时，为了更加公平地评估每一层的隐私泄漏，排除层与层之间的干扰，我们使用 shapley value 作为衡量指标。所谓 shapley value，其实是博弈论中的一个概念，它通过计算游戏中不同玩家在不同玩家组合中的平均边际贡献，以公平的方式量化每个玩家对整体合作效益的贡献，其计算公式如下：

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right)$$

其中：S 是不包含参与者 i 的任何参与者子集；$|S|$ 表示集合 S 中的元素数量；$|N|$ 是总参与者数量；$\frac{|S|!(|N|-|S|-1)!}{|N|!}$ 是组合 S 出现的权重，它反映了所有可能的参与者排序中 S 出现的频率；$v(S \cup \{i\}) - v(S)$ 是参与者 i 加入集合 S 带来的边际贡献。

实验结果如图6所示，可以看到前面的层与后面的层泄漏的隐私相当。但是在 CNN 中，不同层之间依然有所区别，特别是第二层与其他层相比，明显泄漏的隐私占多。

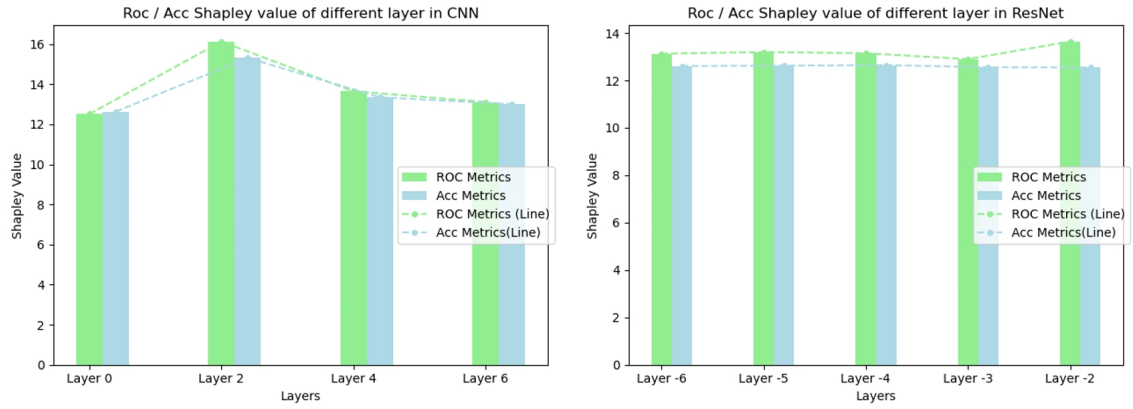

图 6. 不同层的实际隐私泄露结果

通过分析网络结构，我们发现在 CNN 中，第二层的参数量是最多的，因此其泄漏的信息自然就多一些，最终得到的 shapley value 值也会大一些。但对于 ResNet 和 AlexNet 来说，

尽管最后一层的参数量不是最多的，但是其位置比较靠后，融合了前面层的特征和信息，因此也会泄漏更多的信息，所以在这两个模型中，最后一层对应的 shapley value 也会大一些。

从整体的实验结果来看，本工作认为不同层泄漏隐私多少，将取决于两个条件：1）该层的参数量，量多的泄漏的信息多；2）该层的位置，越往后的层，泄漏的隐私越多。这实际上，也印证了 Nasr 等人 [11] 得出的结论，只是结论比其更加严谨。

6 总结与展望

本次工作聚焦于复现 Nasr 等人所设计的白盒推理攻击来对深度学习模型进行逐层隐私分析，我们从梯度张量中提取七种梯度范数，以降低攻击模型的复杂度。同时，本次工作进一步完善了 Nasr 等人的实验设置，得出更加严谨的结论，即不同层泄漏隐私多少，将取决于两个条件：1）该层的参数量，量多的泄漏的信息多；2）该层的位置，越往后的层，泄漏的隐私越多。

然而，限制于时间关系，本次工作只在 CNN、AlexNet 和 ResNet 上，使用 CIFAR100 数据集进行实验，尚未开展广泛的实验，得出的结果还不够普遍化。其次，对于本工作得出的结论，尚处于启发式阶段，如果有具体的理论证明，将会更好地为深度学习模型的隐私保护提供参考。

参考文献

[1] Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Leveraging adversarial examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10399–10409, 2022.

[2] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[3] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[4] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[5] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: evaluating privacy leakage of generative models using generative adversarial networks. *arXiv preprint arXiv:1705.07663*, pages 506–519, 2017.

[6] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig.

Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[7] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020.

[8] Lan Liu, Yi Wang, Gaoyang Liu, Kai Peng, and Chen Wang. Membership inference attacks against machine learning models via prediction sensitivity. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[9] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

[10] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[11] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pages 1–15, 2018.

[12] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021.

[13] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[14] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[15] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[16] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 534–544, 2009.

[17] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

[18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.