Label-Only Model Inversion Attacks via Boundary Repulsion

Mostafa Kahla Virginia Tech Si Chen Virginia Tech

摘要

最近的研究表明，最先进的深度神经网络容易受到模型反转攻击的影响，其中滥用模型访问来重建任何给定目标类的私有训练数据。现有的攻击依赖于访问完整的目标模型 (白盒) 或模型的软标签 (黑盒)。然而，在更难但更实用的场景中没有做先前的工作，在这种情况下，攻击者只能访问模型的预测标签，而没有置信度度量。在本文中，我们介绍了一种算法 Boundary-Repling Model Inversion (BREP-MI)，仅使用目标模型的预测标签反转私有训练数据。我们算法的关键思想是评估模型在球体上的预测标签，然后估计到达目标类的质心的方向。使用人脸识别的例子，我们表明 BREP-MI 重建的图像成功地再现了各种数据集和目标模型架构的私有训练数据的语义。我们将 BREP-MI 与最先进的白盒和黑盒模型反演攻击进行比较，结果表明，尽管假设目标模型的知识较少，但 BREP-MI 优于黑盒攻击，并取得了与白盒攻击相当的结果。

关键词：模型反演攻击；隐私泄露

1 引言

机器学习 (ML) 算法通常在私有或敏感数据上进行训练，例如人脸图像、医疗记录和财务信息。不幸的是，由于 ML 模型倾向于记住有关训练数据的信息，即使安全地存储和处理，隐私信息仍然可以通过访问模型来暴露。事实上，先前关于隐私攻击的研究已经证明了在不同粒度暴露训练数据的可能性，范围从“粗粒度”信息，例如确定某个点是否参与训练或训练数据集是否满足某些属性，到更多的“细粒度”信息，如重建原始数据。在本文中，我们专注于模型反转 (MI) 攻击，其目标是在给定对训练模型的访问的情况下重新创建训练数据或敏感属性。由于攻击揭示的“细粒度”信息，MI 攻击会造成巨大的伤害。例如，应用于个性化医学预测模型的 MI 攻击导致个体基因组属性的泄漏。最近的研究表明，MI 攻击甚至可以成功地重建高维数据，如图像。证明了从仅给定名称的人脸识别模型中恢复一个人的图像的可能性。

2 相关工作

模型反转攻击。模型反演试图从部分到完整的训练样本重建。通常，MI 攻击可以形式化为一个优化问题，目标是找到在被攻击的模型下达到最高可能性的敏感特征值。然而，当目标模型是一个深度神经网络 (DNN) 或私有数据位于高维空间时，这种优化问题变得非凸，通过梯度下降直接求解它可能会导致攻击性能较差; 例如，在攻击人脸识别模型时，恢复的图像是模糊的，不包含太多的私人信息。最近的工作提出了一种基于 GAN 的 MI 攻击方法，该方法对 DNN 有效。特别是，他们通过 GAN 从公共数据中学习通用先验，并解决潜在空间而不是无约束环境空间的优化问题。然而，他们的攻击方法在训练 GAN 阶段没有充分利用目标模型中包含的私人信息。通过 GAN 的特殊设计显着提高了攻击性能，可以从目标模型中提取知识；因此，生成的图像与私有分布更好地对齐。他们通过确保恢

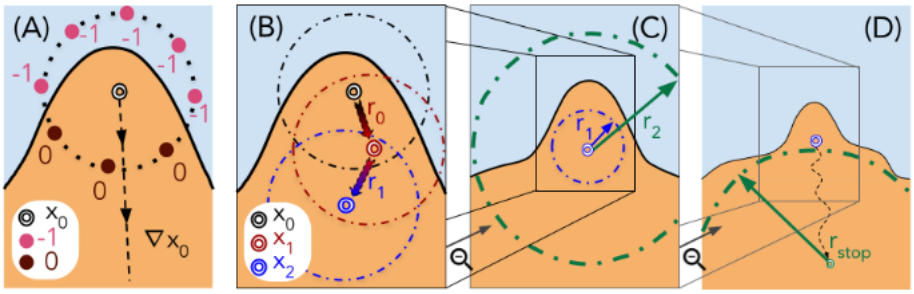

图 1: 过程示意图

复的图像及其相邻图像都具有很高的可能性来进一步提高性能。虽然在攻击各种模型和数据集方面取得了成功，但它们的攻击依赖于模型的白盒访问。在许多情况下，攻击者只能对模型进行预测查询，而不能实际下载模型，这激发了黑盒 MI 攻击的研究。分析了黑箱设置，提出了一种攻击模型，该模型交换目标模型的输入和输出向量来执行模型反演。建议同时训练 GAN 和代理模型，GAN 生成类似于私有训练数据的输入和模仿目标模型行为的代理模型。所有的黑盒攻击都假设目标模型的预测置信度被泄露给攻击者。然而，在现实世界的环境中，只有对模型查询的对手只能获得硬标签，而没有置信度分数。从这个角度来看，我们的目标是提供一种有效的 MI 攻击方法，该方法只需要访问硬标签，我们将其称为仅标签的 MI 攻击。

3 本文方法

3.1 本文方法概述

我们将首先将 MI 攻击制定为优化问题。然后，我们描述了一种仅基于预测标签估计 MI 优化目标梯度的算法。我们将严格描述线性模型特殊情况的估计梯度和真实梯度之间的对齐，并为深度非线性模型的攻击效率提供见解。显然，目标类 c□ 最具代表性的输入应该与所有其他类最可区分。因此，我们将 MI 问题转换为一个优化问题，该问题寻求实现目标类的置信度与其他类的最高置信度之间的最大差异的输入：我们算法背后的直觉是一个点离类的决策边界越远，代表这一点变成班级。因此，任何类的质心都应该是一个很好的代表。受此启发，我们设计了一种试图逐渐远离决策边界的算法。在高层次上，我们的算法首先采样球体上的点，然后查询它们的标签。直观地说，未预测到目标类的点代表了我们想要远离的方向。因此，我们对这些点取平均值，并以与平均值相反的方向移动。如果所有点都被预测到目标类中，那么我们将增加半径。然而，对于图像，x 通常位于高维连续数据空间中，对该空间进行优化很容易陷入与任何有意义的图像不对应的局部最小值。为了解决这个问题，我们利用 [4, 24, 25] 中的想法，并在语义上有意义的潜在空间中进行优化。这是通过使用公共数据集来训练 GAN 模型然后优化 GAN 生成器的输入来完成的。用 G(z) 表示公开可用的训练生成器。现在，可以更新 MI 优化问题以反映优化 z 而不是 x 的变化，如下所示：并根据方程式更新 z。请注意，如果新点 z 位于目标类之外，则恢复更新。在这种情况下，我们将重新采样球体上的点并计算新的更新。当不可能找到更大的球体时，算法将停止该算法，以便该球体上的所有样本都属于目标类。该算法的输出是一个点 (z)，其球体最大，可以适合目标类。这表明该点离边界最远的。我们将使用这个点来评估攻击。

4 复现细节

4.1 与已有开源代码对比

改进了神经网络的网络架，尝试了更多地超参数，运用自己的理解改进了代码，使代码的可读性更强。

4.2 实验环境搭建

我们对三个不同的人脸识别数据集进行了实验:CelebA、Facescrub 和 Pubfig83。与类似，我们将所有数据集的图像裁剪到中心，并将它们调整为 64x64。我们将身份拆分为公共域 (我们在上训练 GAN)，以及私有域 (我们将在上训练目标模型)。公共域和私有域之间没有重叠的身份。这意味着攻击者对私有域中的身份有零知识。然后，我们对在私有域上训练的分类器执行攻击。每个数据集的详细信息如表 1 所示。为了研究私有域和公共域之间较大的分布偏移对攻击性能的影响，我们使用 FFHQ 数据集作为我们的公共域来训练 GAN，并将上述三个数据集作为私有域。目标模型。除了评估我们对一系列数据集的攻击外，我们还使用各种架构评估我们对不同模型的攻击。为了与之前的工作保持一致的结果，我们使用了最先进的 MI 攻击中使用的相同模型架构：（1）face.evave adpated from ；(2) ResNet-152 改编自；(3) VGG16。评估协议。对于所有使用 BREP-MI 的评估，我们执行有针对性的攻击，因为与非目标攻击相比，它是一个更具挑战性的设置。我们使用攻击准确度来衡量攻击性能。攻击精度基于评估分类器，该分类器预测重建人脸图像的身份，是人类法官的代表。具体来说，攻击精度是通过重建图像的数量与重建图像总数相比被正确分类到相应目标类的比率来计算的。由于评估分类器反映了人类判断，它应该具有很高的性能。同时，它应该与被攻击的目标模型不同，以避免一些与目标模型过度拟合的语义无意义的重建图像被认为是良好的重建。超参数。我们在评估中手动微调了 BREP-MI 的超参数。我们凭经验发现最佳初始半径 R0 为 2，半径展开系数 γ 为 1.3，步长 $\alpha t=\min(R/3,3)$。我们选择 N，即球体上采样点的数量，除非另有说明，否则设置为 32。maxIters 被选择为 1000，即当超过 1000 次迭代传递给某个 R 时，BREP-MI 终止，而没有将球体上的所有点分类为目标类。基线。由于这是第一个为仅标签 MI 攻击提供解决方案的工作，我们选择针对攻击者在有关目标模型的额外知识方面拥有更大优势的白盒和黑盒攻击进行评估。为了确保公平比较，我们对每个数据集和相同的目标模型在同一组目标身份上应用所有基线。然后我们针对相同的评估分类器评估攻击准确性。我们的两个基线是白盒攻击，包括生成模型反转 (GMI) 10，它是第一个针对深度网络的 MI 攻击算法，以及知识丰富的分布模型反转攻击 (KeD-MI) [4]，它提供了当前最先进的白盒 MI 性能。GMI 中的 GAN 模型在我们的攻击中设置为与 GAN 相同。KeD-MI 依赖于在训练 GAN 模型时对目标模型参数的访问。但是，我们无法访问这些信息并在我们的设置中训练相同的 GAN。我们还采用了黑盒攻击，称为基于学习的模型反转 (LB-MI)，作为我们的基线之一。LB-MI 构建了一个反转模型，该模型学习从目标模型产生的软标签中重建图像。为了重建给定身份最具代表性的图像，我们在反转模型的输入处为该身份提供 one-hot 编码并接收输出。

4.3 创新点

（1）我们提出了第一个仅标签模型反转攻击算法。(2) 我们通过证明我们的算法中使用的更新与梯度对齐，并分析非线性模型的对齐误差，为线性目标模型情况下算法提供了理论依据。(3) 我们评估了一系列模型架构、数据集的攻击，并表明尽管利用了关于目标模型的信息较少，但我们的攻击大

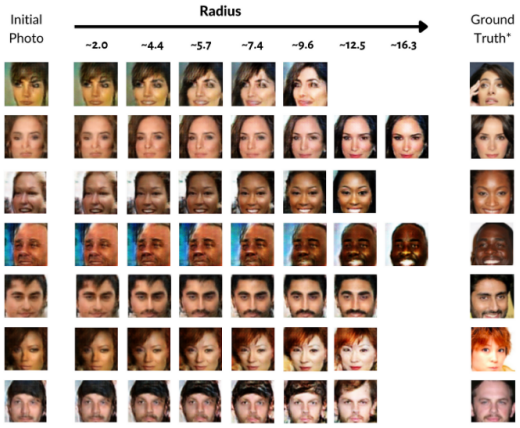

图 2: BREP-MI's progression along each radius from the first random initial point until the algorithm's termination.

大优于基于置信度的黑盒攻击，并实现了与最先进的白盒攻击相当的性能。此外，我们将发布数据、代码和模型以促进未来的研究。

5 实验结果分析

不同数据集的性能。我们将 BREPMI 与三个不同面部数据集上的白盒和黑盒方法进行比较。我们在所有数据集中使用 FaceNet64 作为目标模型。对于每个数据集，GAN 模型在其公共身份上进行训练，目标模型在私有身份上进行训练。表 2 显示我们的方法在所有数据集上都大大优于白盒 GMI 攻击和黑盒攻击。此外，我们的方法在 Pubfig83 上超越了最先进的白盒 KeD-MI 攻击，并在 Celeb 数据集上实现了接近的攻击精度。另一方面，我们在 Facecrub 数据集上落后 15。值得注意的是，该实验的结果意味着在 MI 攻击中的其他威胁模型中仍有相当大的发展潜力，特别是黑盒攻击（相对于其他威胁模型表现不佳）。即使使用白盒知识，GMI 表现不佳的原因是它优化仅合成数据点的可能性，而不考虑点的邻域。因此，优化可能会陷入不代表类的尖锐局部最大值。另一方面，BREP-MI 和 KeD-MI 都明确地找到了具有高可能性的邻域，事实证明这对于产生代表性点并提高攻击性能至关重要。值得注意的是，黑盒攻击虽然利用更多关于目标模型的知识，但始终表现最差它使用公共数据来训练反演模型，而其他攻击都在公共数据上训练 GAN 模型。结果表明，GAN 在提取公共知识方面比反演模型更有效。因此，改进黑盒攻击的一个潜在方法是通过 GAN 对合成图像进行正则化。

6 总结与展望

我们提出了一种新的算法来执行第一个仅标签的 MI 攻击。实验表明我们的方法在不同的数据集和模型架构上的有效性。有趣的是，该方法提供了与最先进的白盒攻击相当的结果，并且优于所有其他基线，尽管它们对攻击者知识做出了更强的假设。作为未来的工作，结果在仅标签攻击和最先进的白盒攻击之间的多次实验中的接近程度表明白盒攻击仍有改进的空间。类似地，黑盒基线攻击大大优于我们的仅标签攻击，尽管它可以访问比仅标签攻击更细粒度的模型输出。理论上，它应该是我们性能的上限