

Private Aggregation of Teacher Ensembles

用半监督知识迁移解决深度学习中训练数据隐私问题

熊凯亚

Jinan University

May 17, 2018

Outline

PATE

- 背景
- 威胁模型
- 教师模型
- 学生模型

PATE-G

实验

- 教师模型准确率
对比

Google 在 ICLR'17 上发表的文章 [1]¹用半监督知识迁移解决深度学习中训练数据隐私问题，为了解决这个问题这篇文章提出了 PATE(Private Aggregation of Teacher Ensembles) 教师全体的隐私聚合的概念。

¹N. Papernot, M. Abadi, Ú. Erlingsson, *et al.*, "Semi-supervised knowledge transfer for deep learning from private training data," *CoRR*, vol. abs/1610.05755, 2016.

对于训练一般人脸识别模型例子：

1. 2015 年的一项研究发现²，通过模型的预测结果，可以反过来重建模型训练时使用的人脸数据 (model inversion attacks)。
2. 2016 年另一项研究发现³，同样可以根据模型的预测结果，来推理出模型训练数据中是否包含了某个具体的训练点 (training point)，这种攻击称为会员推理攻击 (membership inference attacks)。

²M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2015, pp. 1322–1333.

³R. Shokri, M. Stronati, C. Song, *et al.*, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*, IEEE, 2017, pp. 3–18.

威胁模型

两种攻击类型：

1. model querying, 模型查询：黑盒攻击，攻击者通过查询来观察模型。对于攻击者而言模型是一个黑盒，攻击者可以挑选输入值来观察模型的预测结果。
2. model inspection, 模型检验：属于白盒攻击。攻击者知道模型的结构和参数，例如 [4]⁴协作学习中的参与者对模型进行的白盒攻击。

攻击假设：

- ▶ 攻击者可以进行潜在的无限多次的查询（黑盒）。
- ▶ 攻击者能够进入到模型的内部组件（白盒）。

⁴B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2017, pp. 603–618.

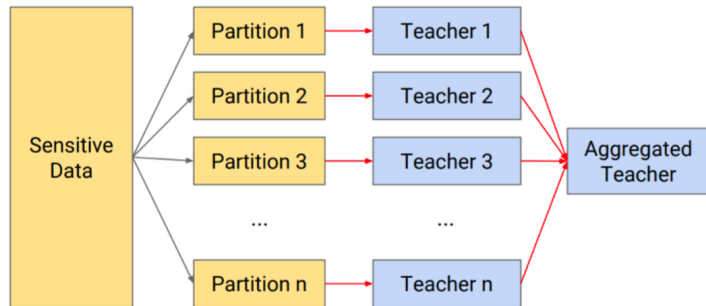
教师模型

Approach: 在不相交的子集上训练教师模型，然后在使用另外的未标记的非敏感数据对学生模型进行训练。

1. 将待训练的敏感数据分为互斥的 N 份不同数据集，分别独立训练不同的模型，得到 N 个教师模型。
2. 部署训练好的教师模型时，需要记录每一个教师模型对于查询的预测结果，选取票数最高的那个预测结果，并将预测结果聚合起来。
3. 在统计票数之后引入拉普拉斯噪声，将票数的统计情况打乱（票数会泄露隐私），从而保护了隐私。

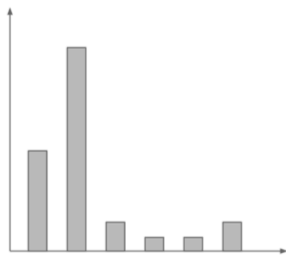
教师模型的训练及聚集过程如图 1所示。

教师模型

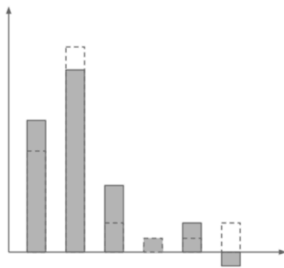


图：教师模型的聚集过程

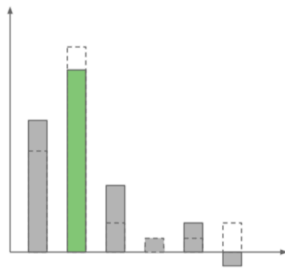
教师模型



Count votes



Add Laplacian noise



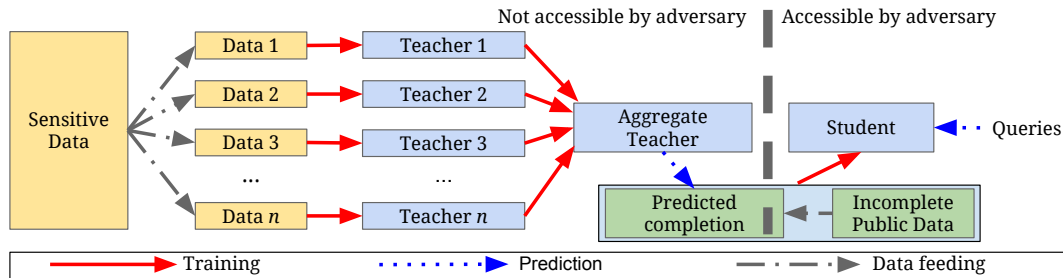
Take maximum

- ▶ 如果大部分教师模型都同意某一个预测结果，即不依赖于具体的分散数据集，隐私成本很小。
- ▶ 如果有两类预测结果有相近的票数，那么这种不一致，或许会泄露隐私信息。

学生模型

- ▶ 聚合教师模型可以看作是一个差分隐私 API，用户提交输入值，模型就会返回相应的标签，同时又能保护隐私。
- ▶ 但是，如果训练一个模型，部署到用户设备上直接运行模型得出预测结果，这样的话肯定会更好。为了训练学生模型，需要聚合教师模型以隐私保护的方式，来给公共数据进行标注，传递知识。

学生模型



图：学生模型的训练

使用未标记的公开数据训练学生模型，部分数据使用对教师模型的查询结果进行标记。

学生模型

学生模型的必要性：

1. 每次查询聚合教师模型，都会增加隐私成本。训练学生模型后，只能对聚合教师模型进行固定数量的查询，隐私成本就会被固定下来。
2. 需要防范攻击者探取模型底层函数库。教师模型是由隐私数据训练的，学生模型是由公共数据（非隐私数据）训练的，带有隐私保护的标注。即使攻击者获取到学生模型的训练数据，也只能得到带有隐私保护的标签信息。

总体上：

- ▶ 教师模型用来防御白盒攻击
- ▶ 学生模型用来防御黑盒攻击

PATE-G

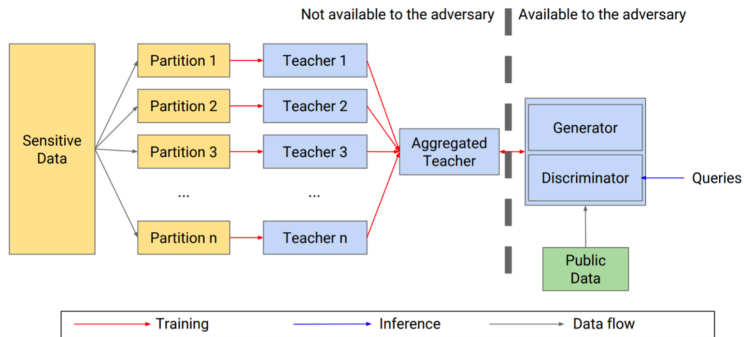


图: 使用 GAN 训练学生模型

- ▶ 将原本二元分类的判别器扩展至一个多类别的分类器，区分：已标注的真实样本，未标注真实样本，以及生成样本。
- ▶ 训练之后只使用判别器来处理查询。

实验

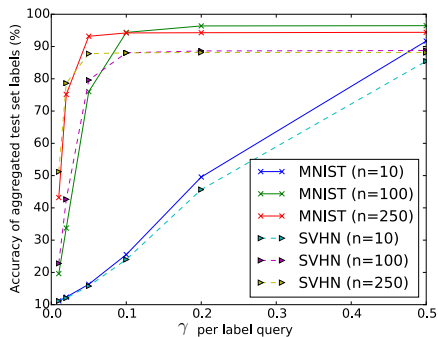
使用了四个数据集：MNIST、SVHN、UCI Adult 和 UCI Diabetes。如图所示：

Dataset	Teacher Model	Student Model	Student Public Data	Testing Data
MNIST	2 conv + 1 relu	GANs (6 fc layers)	test[:1000]	test[1000:]
SVHN	2 conv + 2 relu	GANs (7 conv + 2 NIN)	test[:1000]	test[1000:]
UCI Adult	RF (100 trees)	RF (100 trees)	test[:500]	test[500:]
UCI Diabetes	RF (100 trees)	RF (100 trees)	test[:500]	test[500:]

图：实验数据集

教师模型准确率

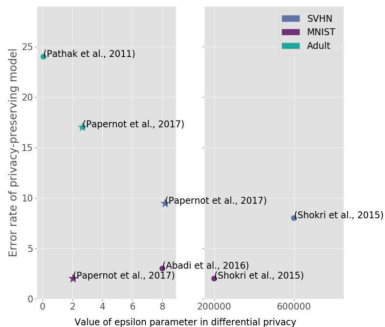
图5描绘了聚合教师模型的准确率。在训练学生模型之前，需要考虑了每一个标签的隐私。横轴是每一个标签查询的 ϵ 值，纵轴是预测结果的平均准确率。



图：聚合教师模型准确率

紫色的线代表 10 个聚合教师模型 ($n = 10$)。逐渐降低 ϵ 值，准确率也很快下降。但如图中绿线和红线的部分，分别包含 100 个和 250 个聚合教师模型 ($n = 100, n = 250$)，在较低 ϵ 值时，仍可以保持较高的准确率。

对比








图：不同 Approach 的对比

Papernot, M. Abadi, Ú. Erlingsson, et al., “Semi-supervised knowledge transfer for deep learning from private training data,” CoRR, vol. abs/1610.05755, 2016.

Abadi, A. Chu, I. Goodfellow, et al., “Deep learning with differential privacy,” in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016, pp. 308–318.

Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, ACM, 2015, pp. 1310–1321.

References

-  N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, “Semi-supervised knowledge transfer for deep learning from private training data,” *CoRR*, vol. abs/1610.05755, 2016 (cit. on p. 3).
-  M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2015, pp. 1322–1333 (cit. on p. 4).
-  R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *Security and Privacy (SP), 2017 IEEE Symposium on*, IEEE, 2017, pp. 3–18 (cit. on p. 4).
-  B. Hitaj, G. Ateniese, and F. Pérez-Cruz, “Deep models under the gan: Information leakage from collaborative deep learning,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2017, pp. 603–618 (cit. on p. 5).
-  M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proceedings of the 2016*