

去中心化深度学习的隐私保护研究

熊凯亚

Jinan University

May 14, 2018

Outline

引言

相关工作

攻击

未来工作

引言

最近看了一些关于在去中心化的深度学习场景下的隐私保护的文章。去中心化的深度学习主要有两个方向：

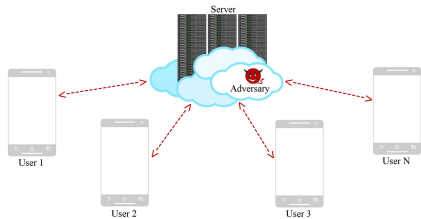
- ▶ CCS'15 上 Shokri 等人 [1] 提出的隐私保护多方参与的协作深度学习模型 (Collaborative Deep Learning)¹
- ▶ Google 提出的针对移动设备 (Android) 的多方参与的联合学习 (Federated Learning)²

¹R. 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.

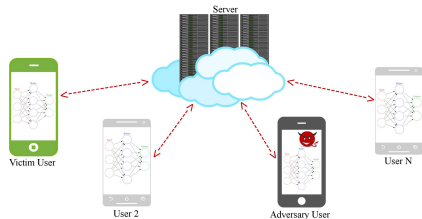
²联合学习: <https://research.googleblog.com/2017/04/federated-learning-collaborative.html>

引言

协作深度学习



(a) 传统的中心化学习



(b) 去中心化学习

Figure: 传统中心化学习和去中心化学习对比

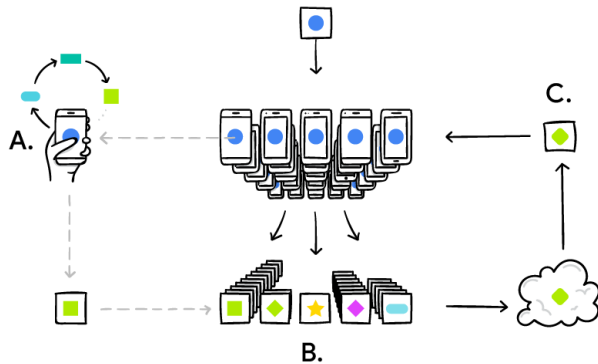


Figure: 联合学习流程图: A 根据手机的使用情况, 对本地模型进行个性化设置并训练上传模型更新; B 许多用户的模型更新被聚集在一起; C 优化共享模型, 然后很多用户又对新的模型进行下载、训练, 并重复这个过程。

相关工作

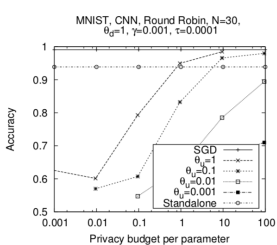
差分隐私

1. 在 [1] 提出的协作学习中³，选择性分享模型参数，保证用户的数据不被泄露，在可用性和隐私之间 tradeoff。
2. 参数分享过程可能会有非常少的一部分数据会被泄露，因此使用差分隐私的方式对分享的参数加入噪声 (record-level)，使得即使参数泄露也不会对用户的隐私数据造成威胁。
3. 当差分隐私中的 ϵ 越小时（更强的隐私保护）将会导致模型的 Accuracy 降低，因此这种方式是在模型的准确性（Accuracy）和隐私之间 tradeoff。

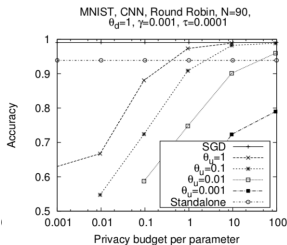
³R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, ACM, 2015, pp. 1310–1321.

相关工作

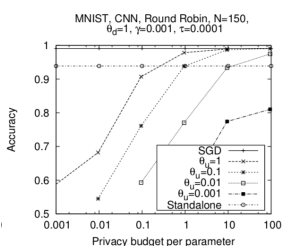
差分隐私



(a) 参与者为 30 人



(b) 参与者为 90 人



(c) 参与者为 150 人

Figure: 在 MNIST 数据集上, 对 CNN 网络的含有差分隐私的协作学习。其中横轴为 ϵ , 表示差分隐私的强度 ϵ 越大差分隐私保护添加的噪声越弱。N 表示参与者的数量, θ_d 表示参数下载的比例, θ_u 表示参数上传到比例。从图中可以看出, 1: 条件相同时参与人数越多准确性越高; 2: ϵ 越大 (噪声越小) 准确性越高; 3: θ_u 越大准确性越高。

相关工作

同态加密

1. 2017 年, Phong 等人在 TIFS'17 上发表了文章 [2]⁴, 对 [1] 提出的协作学习模型的安全及隐私性进行分析, 并提出使用同态加密的方法来加密上传给服务器的梯度 (gradients)。
2. 为了进一步确保在上传给服务器的同态加密密文的完整性, 在于服务器的通讯中采用 TLS/SSL 安全信道。
3. 一方面, 对上传的梯度使用基于 LWE 或者 Paillier 的同态加密会增加参与者与服务器之间的通讯开销, 但实验表明增加的通讯开销量还是可接受的;
4. 另一方面, 相比于 [1] 中使用差分隐私减少了一部分准确性来说, 计算量的增加带来效率的降低还是可以接受的, 因为毕竟在深度学习领域准确性比效率更重要。

⁴L. T. Phong, Y. Aono, T. Hayashi, *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1333–1345, 2017.

相关工作

同态加密

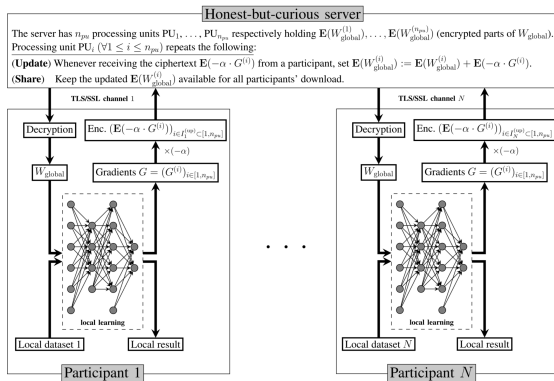


Figure: 使用同态加密的协作学习模型

攻击

介绍

- ▶ [2] 考虑的是服务器是恶意的，有可能窃取参与者的训练数据。
- ▶ Hitaj 等人在 CCS'17 上发表了一篇关于协作学习中参与者隐私泄露的文章 [3]，考虑的是参与者有可能是 active 的 Adversary，从而获取到其他参与者的训练数据。
- ▶ 文章 [3] 对 [1] 中提出的协作学习进行了攻击，通过使用对抗生成网络（GAN）成功地从其中某个参与者中获得了训练数据⁵。
- ▶ 进一步地，即使在将参数上传至服务器之前使用差分隐私在参数中加入噪声，只要差分隐私的粒度是 record-level 的，攻击仍然可以成功。

⁵B. 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.

攻击

介绍

- ▶ 黑盒攻击：在模型训练完后对模型进行攻击（无需考虑训练过程）：
 1. Model extraction attacks
 2. Model inversion attacks
 3. Membership inference attacks
- ▶ 这是一种白盒攻击：在模型训练阶段攻击，攻击者像其他参与者一样参与到模型的训练中，可以看到并使用模型的参数。
- ▶ 另一方面，攻击者会偷偷地使用其他参与者共享的参数使用 GAN 网络生成与受害者训练数据类似的数据，每一次迭代攻击者都可以获取到更多受害者的训练数据，最后攻击者可以完整的还原出受害者的训练数据，攻击成功。
- ▶ 在这种情况下，只要参与者的本地模型精确度足够高，攻击者都可以成功获取到受害者的原始训练数据。

攻击

介绍

值得一提的是，[3] 提出的这种攻击方法对于 Google 的联合学习模型同样有用。

- ▶ 和 [1] 使用差分隐私来保护分享的参数不同，联合学习中采用了一种安全聚集协议 (secure aggregation protocol)[4]⁶。
- ▶ 通过使用安全多方计算 (MPC) 计算模型参数的平均权重，依据权重每个移动设备的更新 (updates) 都可以被安全地聚集起来，而且这也使得只有在多个用户的参与下 Google 才能对模型参数进行解密，这样就可以防止 Google 窃取用户的模型参数。
- ▶ 由于 [4] 的安全模型中也是只考虑了 Google 可能是 Adversary，并没有考虑参与者的任何一个人都可能成为 Adversary，进而去攻击其他参与者以获取其训练数据。因此此攻击对于联合学习仍然有效。

⁶K. Bonawitz, V. Ivanov, B. Kreuter, *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2017, pp. 1175–1191.

攻击

威胁模型

- ▶ 设 V 为受害者并且有标签 $[a, b]$, A 为敌手且有标签 $[b, c]$ 。
- ▶ A 的目标就是尽可能推断出更多关于 a 的信息。 A 会使用 GAN 生成非常类似 a 的样本, 然后把它标记成 c , 并在本地训练, 并将其模型参数上传至服务器。
- ▶ 这样, V 需要去在 a 和 c 中辨别真假, 就会暴露更多关于 a 的信息。

详细步骤如下:

攻击

威胁模型

1. 假设 A 和 V 双方提前对学习结构和目标达成了一致。
2. V 有标签 $[a, b]$, A 有标签 $[b, c]$ 。
3. 执行协作深度学习协议若干轮, 当且仅当参数服务器上的模型和本地的模型都达到特定的准确度才停止迭代。
4. V 开始训练:
 - 4.1 V 从参数服务器上下载部分参数 (θ_d), 并更新本地模型。
 - 4.2 V 在 $[a, b]$ 上训练本地模型。
 - 4.3 V 上传部分本地模型的参数到服务器上。
5. A 开始训练:
 - 5.1 A 从参数服务器上下载部分参数 (θ_d), 并更新本地模型。
 - 5.2 A 训练本地的 GAN 来模仿 V 的 a 。
 - 5.3 A 的 GAN 生成样本, 并把它们标记成 c 。
 - 5.4 A 在 $[b, c]$ 上训练本地模型。
 - 5.5 A 上传部分本地模型的参数到服务器上。
6. 重复步骤 4 和 5, 直到模型收敛。

攻击

威胁模型

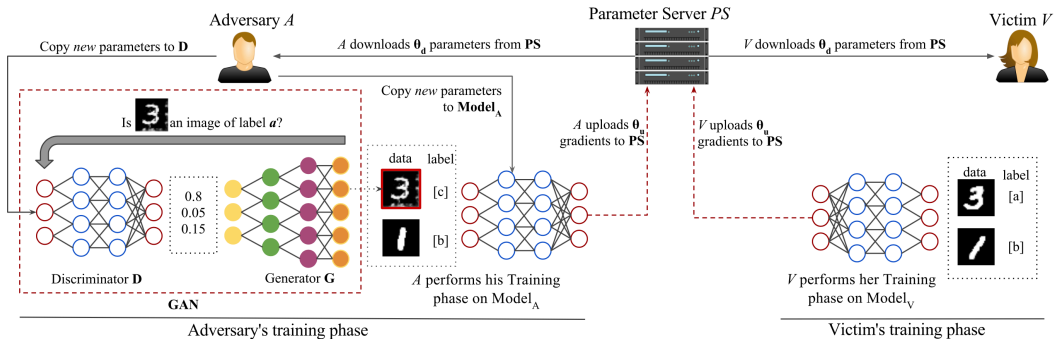


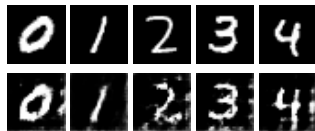
Figure: 使用 GAN 攻击正在训练 MNIST 数据集的协作深度学习模型。

攻击

实验结果



(a) $\theta_u = 1, \theta_d = 1$



(b) $\theta_u = 0.1, \theta_d = 1$



(c) $\theta_u = 0.1, \theta_d = 0.1$

Figure: 使用 GAN 在 MNIST 数据集上攻击两个参与者场景下的结果。图中下面一行是 GAN 生成的样本，上面一行是受害者训练集中的数据。其中 (a) 图的参数上传下载比例都是 100%，(b) 图为 10%，100%，(c) 图为 10%，10%。

攻击

实验结果

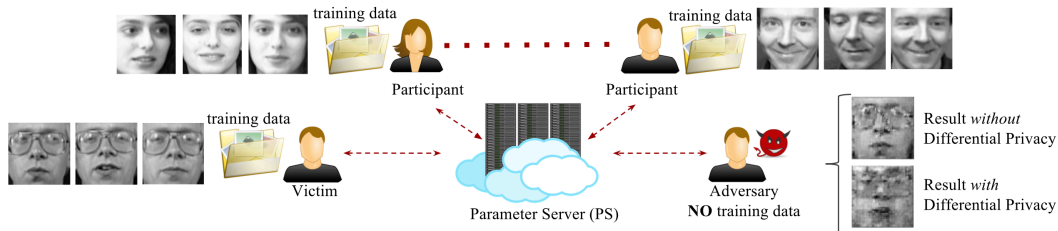






Figure: 使用 GAN 在 AT&T 的 Olivetti 人脸数据集上攻击的结果。诚实的参与者们互相独立地训练模型，且敌手 A 并没有本地的训练数据，通过使用 GAN 攻击，A 可以重新构建出受害者 V 的训练数据。即使在使用差分隐私的情况下，这种攻击也能成功。

未来工作

因为刚接触深度学习及其隐私问题不久，可能对其了解不够深刻。我想以此工作作为切入点，继续再次详细地看一看本文涉及的 paper。一方面持续关注深度学习领域的隐私保护问题，另一方面还要看看目前比较热门的基因技术和机器学习技术的结合点及其隐私保护问题，或者是否可以根据基因数据的特点，将现有的机器学习的隐私保护方案用到基因数据的隐私保护方面。

References

-  R. 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 (cit. on pp. 3, 6, 8, 10, 12).
-  L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, “Privacy-preserving deep learning via additively homomorphic encryption,” *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1333–1345, 2017 (cit. on pp. 8, 10).
-  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 pp. 10, 12).
-  K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2017, pp. 1175–1191 (cit. on p. 12).