**摘要**

深度学习最近在机器学习方面非常受欢迎，因为它能够解决端到端学习系统，其中功能和分类器同时被学习，在高度结构化和大型数据库的存在下显着提高分类准确性。

它的成功归功于最近的算法突破，越来越强大的计算机以及大量数据。

研究人员还考虑了深度学习对隐私的影响。通常以集中方式训练模型，所有数据由同一个的训练算法处理。如果数据是用户私人数据的集合，包括习惯，个人图片，地理位置，兴趣等，则集中式服务器将可以访问可能被错误处理的敏感信息。为了解决这个问题，最近提出了协作深度学习模型，其中各方在本地训练他们的深度学习结构并且仅共享参数的子集以试图保持他们各自的训练集私有。参数也可以通过差异隐私（DP）进行模糊处理，以使信息提取更具挑战性，正如Shokri和Shmatikov在CCS'15提出的那样。

不幸的是，我们表**明任何保护隐私的协作深度学习都容易受到我们在本文中设计的强大攻击。特别是，我们表明，分布式，联合或分散的深度学习方法从根本上被打破，并不保护诚实参与者的培训集**。我们开发的攻击利用了学习过程的实时性，允许对手训练生成对抗网络（GAN），**该网络生成旨在私有的目标训练集的原型样本**（由GAN生成的样本是意图来自与训练数据相同的分布）。有趣的是，我们表明应用于模型的共享参数的记录级差异隐私，如先前工作中所建议吗那样，是无效的（即，记录级DP不是为解决我们的攻击而设计的）。

**关键词：**协作学习; 安全; 隐私; 深度学习

**1引言**

深度学习是机器学习的一个新分支，利用神经网络，这个概念可以追溯到1943年[49]，为各种复杂的任务找到解决方案。神经网络的灵感来自人类大脑学习的方式，即分布式人工神经网络也可以学习非平凡的任务，即使当前的架构和学习过程远非大脑般的行为。

算法的突破，收集大量数据的可行性以及增加的计算能力促成了当前神经网络的普及，特别是多个（深层）隐藏层，确实已经开始超越以前的先进技术机器学习技术[6,29​​,75]。与传统的机器学习方法不同，深度学习不需要输入的特征工程[45]，因为模型本身自己提取相关特征并定义哪些特征与每个问题相关[29,45]。

深度学习模型对相关数据的表现非常好，这有助于计算机视觉[47]，图像处理，视频处理，人脸识别[82]，语音识别[34]，文本到语音系统[64]和自然语言处理[2,15,90]。深度学习也被用作更复杂系统的组成部分，能够玩游戏[33,42,57,60]或诊断和分类疾病[16,18,26]。

然而，由于训练有素的模型包含有关训练集的基本信息，因此深度学习会产生严重的隐私影响。从模型中提取敏感信息相对简单[4,27,28]。

考虑图1所示的以下情况，其中N个用户在其各自的设备上存储私人信息的本地数据集，并希望合作建立一个常见的判别机器。我们可以通过将所有数据集上传到单个位置（例如，云）来构建分类器，如图1（a）所示。服务运营商在组合数据集上训练模型。这种集中式方法非常有效，因为模型可以访问所有数据，但由于操作员可以直接访问敏感信息，因此不保护隐私。我们还可以采用协作学习算法，如图1（b）所示，其中每个参与者在他的设备上训练本地模型，并与其他用户共享模型的一小部分参数。通过收集和交换这些参数，服务运营商可以创建一个训练有素的模型，该模型几乎与使用集中式方法构建的模型一样准确。分散方法被认为更加隐私友好，因为数据集不是直接暴露的。此外，即使在仅共享一小部分模型参数和/或通过差异隐私来截断和/或混淆参数的情况下，也实验性地示出了收敛[77]。但它需要通过数据进行多次培训，用户在每个时期更新参数。

深度学习社区最近提出了生成性对抗网络（GAN）[30,70,72]，这些网络仍在深入开发[3,9,32,43,56]。 GAN的目标不是将图像分类到不同的类别，而是生成看起来与训练集中的样本相似的样本（理想情况下具有相同的分布）。更重要的是，GAN生成这些样本而无需访问原始样本。 GAN仅与判别性深度神经网络交互以学习数据的分布。

在本文中，我们使用GAN设计了一种针对协作深度学习的强大攻击。攻击的结果是任何充当内部人员的用户都可以从受害者的设备中推断出敏感信息。攻击者只需运行协作学习算法并重建存储在受害者设备上的敏感信息。攻击者还能够影响学习过程并欺骗受害者发布更详细的信息。攻击可以在不影响服务运营商的情况下工作，即使模型参数通过差异隐私进行模糊处理也是如此。如图1（a）所示，集中式服务器是唯一损害数据隐私的参与者。在图1（b）中，我们显示任何用户都可以故意破坏任何其他用户，使得分布式设置更加不受欢迎。

 我们的主要贡献是在协作环境中提出并实现一类新的深度神经网络主动推理攻击。我们的方法比现有的黑盒或白盒信息提取机制更有效。

即，我们的贡献是：

（1）我们针对分布式深度学习设计了一种基于GAN的新攻击。 GAN通常用于隐式密

度估计，据我们所知，这是第一个恶意使用GAN的应用程序。

（2）我们的攻击比当前的信息提取机制更通用和有效。特别是，我们的方法可以用于卷积神经网络（CNN），这对于模型反演攻击来说是众所周知的难[78]。

（3）我们在协作学习中引入了欺骗的概念，其中对手欺骗受害者发布关于敏感数据的更准确信息。

（4）当通过差异隐私对参数进行模糊处理时，我们设计的攻击也是有效的。我们强调，它不是针对差异隐私的攻击，而是仅针对差异隐私在协作深度学习中的建议用途。在实践中，我们表明在[77]和[1]（示例/记录级差异隐私）中应用的差分隐私训练在我们的隐私概念下的协作学习环境中是无效的。

**2备注**

我们设计了一种比当前信息提取机制更通用和有效的新攻击。它基于生成对抗网络（GAN），它被提议用于隐式密度估计[30]。如第5节所述，GAN通过将生成性深层神经网络与判定性深度神经网络进行对比，生成似乎来自训练集的样本。只要判别模型无法确定样本是来自GAN还是训练集，生成学习就会成功。重要的是要认识到判别和生成网络相互影响，因为判别算法试图将GAN生成的样本与实际样本分开，而GAN试图生成更逼真的样本（理想情况下来自原始的相同分布）数据）。 GAN从未看到实际的训练集，它只依赖于存在于判别模型中的信息。该过程类似于警察用来识别嫌疑人的面部复合成像，其中复合艺术家根据对嫌疑人面部的目击者判别描述生成草图。虽然复合艺术家（GAN）从未见过真实的面孔，但最终的图像是基于目击者的反馈。

我们以新的方式使用GAN，因为它们用于在协作的深度学习框架中从诚实的受害者中提取信息。 GAN创建一个应该是私有的类的实例。我们的基于GAN的方法仅在协作深度学习的训练阶段起作用。我们的攻击甚至对于众所周知难以反转的卷积神经网络也是有效的[78]，或者当通过在记录级别设置粒度的差异隐私对参数进行模糊处理时（如[77]和[1]中所提出的））。它在白盒访问模型中工作，攻击者可以看到并使用模型的内部参数。这与黑盒访问相反，黑盒访问攻击者只能看到每个特定输入的模型输出。这不是我们程序的限制，因为协作学习的目的是共享参数，即使只是一小部分。

一旦分布式学习过程结束，参与者总是可以对训练模型应用模型反演或类似攻击。这并不奇怪。我们在本文中展示的是，恶意参与者可以看到模型如何演变并影响其他诚实参与者，并迫使他们发布有关其私有数据集的相关信息。这种欺骗诚实用户的能力是我们攻击的独特之处。此外，截断或混淆共享参数将无济于事，因为只要本地模型的准确性足够高，我们的攻击就是有效的。

但我们强调，我们的攻击不会违反差异隐私（DP），后者被定义为保护数据库。问题在于，在协作深度学习中，DP正在应用于模型的参数，并且在记录/示例级别设置粒度。然而，一旦模型变得准确，最终必须包含添加到学习参数的噪声。只要模型能够准确地对类进行分类，我们的攻击就会起作用，并且将生成该类的代表。在[77]和[1]中应用DP的方式最多可以防止与在学习阶段确实使用的标签相关的特定元素的恢复。我们的攻击结果可能会或可能不会被视为侵犯隐私权。请考虑以下示例：

（1）受害者的设备包含标准医疗记录。 GAN将生成看起来像通用医疗记录的元素，即来自训练集中相同分布。在这种情况下，攻击者可能没有学习到感兴趣的事情，也没有隐私侵犯。但是，如果受害者的设备包含患有癌症的患者的记录，那么攻击者可能会看到不存在的患者，但都患有癌症。根据具体情况，这可能被视为侵犯隐私。

（2）受害者的设备包含色情图片。 GAN将生成类似的场景。虽然它们可能看似模拟，但泄露给对手的信息非常重要。在其他情况下，我们的攻击可能对作为对手的执法官员有用。例如，当受害者的设备包含恐怖分子的儿童色情图片或培训材料时。

（3）受害者的设备包含语音记录。 GAN将产生bab呀声，带有许多虚构的类似单词的声音（当没有文本序列训练网络时，可与WaveNet [64]相比），因此没有隐私侵犯。然而，可以推断所使用的语言（例如，英语或中文）或者说话者是男性还是女性，并且这种泄露的信息可能构成隐私侵犯。

（4）受害者的设备包含Alice的图像。 GAN将生成类似于Alice的面孔，就像复合艺术家生成目击者记忆中Alice的草图一样。在我们的攻击框架中，对手也将收集Alice的所有这些图画并错误地声称它们是Eve的。这将迫使受害者设备内的本地模型释放关于爱丽丝脸部的更多相关和独特的细节，从而加剧泄漏。然而，虽然许多人认为这是侵犯隐私，但其他人可能不同意，因为对手可能无法恢复Alice的确切面孔而只是重建（见图2）。另一方面，如果Alice戴着眼镜或有棕色头发，那么这些信息将被泄露并且可能构成隐私违规，具体取决于上下文。图3给出了另一个例子，其中DCGAN在CIFAR-10数据集上运行[41]同时针对一个由大约6,000个包含各种马的图像组成的类。请注意，该类可以标记为“jj3h221f”，并且没有明确提及马匹。 GAN制作的图像会告诉对手，“jj3h221f”类不包含汽车或飞机，而是包含动物（可能是马）。

**协作学习中的差异隐私旨在保护培训期间使用的特定元素的恢复**。即，攻击者无法判断训练集中是否包含某个X（达到某个阈值）。我们通过生成一个与X无法区分的X'来规避这种保护。在图2中，我们展示了一个面部X和X'的真实例子，即GAN生成的图像。即使X'不是，它们看起来也很相似X 。**虽然这不违反DP，但在许多情况下，它显然会导致严重的隐私侵犯**。我们的观点是，在这种情况下，示例/记录级别的DP是不够的，就像在选择明文攻击（CPA）中的安全加密在活跃的对抗环境中是不够的。 DP本身没有任何问题（因为CPA安全加密没有任何问题）;显然，DP提供了信息理论保护，但重要的是要正确设置其粒度级别。在纪录的水平上，仅仅保护协作学习中的敏感信息对抗活跃的对手是不够的。可以考虑不同粒度的DP（例如，在用户或设备级别），但这不是[77]中提出的。研究人员可以继续争论DP的正确使用或DP应该保护什么[40,53,54,58]，但最终，在这项工作的背景下，人们应该问：我是否会使用一个系统让临时用户恢复与我的图片文件夹中的图像无法区分的图像？

关键是，隐私的协作学习不如集中式学习方法更理想，它应该改进：在集中式学习中，只有服务提供商可以侵犯用户的隐私，但在协作学习中，任何用户都可能违反其他用户的隐私在系统中，不涉及服务提供者（见图1）。

**3影响**

Google采用集中式方法，将Android设备的使用信息收集到一个集中式数据库中，并在其上运行机器学习算法。谷歌最近推出了联邦学习[50,51]，使移动设备能够协作学习共享预测模型，同时保持所有训练数据的本地化。设备从Google服务器下载当前模型，并通过学习本地数据来改进它。

联合学习似乎与协作学习相同，我们的攻击应该同样有效。最后，每台设备都会从Google服务器下载经过培训的模型，只要本地模型正在学习，GAN就能成功运行。

在联合学习中，可以保护单个模型更新。谷歌建议使用安全聚合协议，而不是像[77]中那样使用差异隐私。通过利用安全多方计算（MPC）计算模型参数的加权平均值[8]，可以安全地聚合各个用户设备的更新，以便Google服务器只有在多个用户参与时才能解密结果。我们认为，正如他们的论文所述，这种机制对我们的攻击架构是无效的，因为我们只是依赖于本地模型已成功学习的事实。他们的安全模型仅考虑Google是审查个人更新的对手的情况。因此，他们没有考虑我们在本文中提出的观点，即临时用户可以攻击其他用户。这使得联邦学习可能比它应该取代的集中学习更加危险，至少在目前的形式下。实际上，我们的评估是基于公告和两篇研究论文中的描述。我们还无法访问系统的实际实施，产品往往会随着时间的推移而显着改善。

据说Apple在未来版本的iOS [35]中在众包学习框架中应用差异隐私。虽然我们不了解细节，但我们希望我们的论文可以作为对在协作深度学习中不正确地应用差异隐私的风险的警告。我们的对手不必为服务提供商工作，但他是针对其他用户（例如名人或政治家）的常规用户。

**4相关工作**

事实证明，深度学习在计算机科学的各个领域都取得了成功。从大量数据中学习，处理和生成相关信息的能力使深度学习成为网络安全领域的一个很好的选择。然而，出现了新的和独特的攻击，这些攻击对正在处理的信息的隐私构成严重威胁。

**4.1对机器学习模型的攻击**

据我们所知，第一项涉及从训练模型中提取意外信息的工作是Ateniese等人的工作。 [4]（2011年发布，2013年发布于arXiv [4,12]）。在那里，作者设计了一个元分类器，该元分类器经过训练，可以入侵其他机器学习分类器，从训练集中推断敏感信息或模式。例如，他们能够从训练有素的语音识别系统中提取种族或性别信息。

Fredrikson等人后来扩展了这项工作[27,28]，他们通过利用模型揭示的置信度信息提出了对机器学习算法的模型反演攻击。例如，当应用于面部识别系统时，它们表明可以重建关于对手已知的特定标签的图像。

 最近，Tramèr等人的工作[83]表明，只考虑模型提供的预测，窃取机器学习模型是可能的。 会员推理攻击是由Shokri等人开发的[78]。 在这里，对手被赋予对模型的黑盒访问权，并且可以推断出某个记录最初是否在训练集中。

麦克弗森等人[52]使用深度学习来推断和揭示模糊图像背后的主体的身份。在他们的工作中，Papernot等人[66]表明，对手制作的输入可以被输入深度学习模型并使它们容易出错，即使模型对输入进行错误分类，从而产生不正确的输出。例如，道路上的STOP标志可以巧妙地修改为看起来与人眼相同，但是被训练的模型归类为另一个标志。这项工作在[36,44,65,87]中得到了延伸。

**4.2隐私保护机器学习**

Shokri和Shmatikov [77]设计了针对强大对手的防御机制。作者介绍了分布式深度学习作为保护训练数据隐私的一种方法[85]。在该模型中，多个实体通过参数服务器彼此共享其各个模型的梯度来协作地训练模型。 [17,51,59,80,89,91]也考虑了分布式学习.Mohasselet等[61]提供了训练神经网络的解决方案，同时保留了参与者的隐私。但是，它在双服务器模型中部署安全的多方计算，其中客户端将计算外包给两个不可信但非串通的服务器。然而，Shokri和Shmatikov [77]是第一个考虑隐私保护措施的人，目的是寻找昂贵的多方计算（MPC）技术的实际替代方案。

谷歌开发了直接在智能手机上训练模型的技术，而无需将敏感数据传输到公司的数据中心[8,51]。微软开发了CryptoNets [20]来对加密数据进行深度学习，并为用户提供加密输出[86]。 Ohrimenko等人[63]开发了在可信处理器上训练的数据不经意的机器学习算法。差异隐私在深度学习中起着重要作用，如[1,39,77,79]所示。

**4.3差异隐私**

DifferentialDrivacy（DP）由Dwork [21]引入. 它的目的是为数据库记录提供可证明的隐私保证，而不会显着降低查询准确性。Dwork等人考虑了大数据的差异隐私[23]。一些作品采用DP作为有效的防御机制[5,7,11,13,19,24,25,38,55,62,67,74,88]。

Shokri和Shmatikov [77]提出的协作深度学习使用DP来模糊共享参数，而Abadi等人[1]建议在训练期间将DP应用于参数。 DP在[69]中用于深度自动编码器。

然而，如Haeberlen等人的着作所示，隐蔽通道可用于击败受DP保护的数据库。 [37]。一般而言，如果对手可以访问辅助信息（DP模型之外），则无法保证隐私[22]。在NDSS'16，Liu等人展示了[48]某些粒度的DP在现实生活场景中无效，其中社交数据，移动数据或医疗记录等数据彼此之间具有很强的相关性。请注意，这是设置DP粒度正确并且DP完全没有被违反的问题。

**4.4保护隐私的协作式深度学习**

深度学习的集中方法迫使多个参与者将他们的数据集汇集到一个大的中央训练集中，在该训练集上可以训练模型。正如Shokri和Shmatikov [77]所指出的那样，这构成了严重的隐私威胁，并且不信任的参与者可能不愿意合作。

考虑到上述安全和隐私问题，Shokri和Shmatikov [77]引入了一种新的协作学习方法，允许参与者训练他们的模型，而无需明确地共享他们的训练数据。他们利用了优化算法，例如随机梯度下降（SGD），可以并行化并异步执行。他们的方法包括在SGD期间结合本地参数更新的选择性参数共享过程。参与者通过参数服务器（PS）仅共享其局部模型梯度的一小部分。每个参与者轮流上传并下载一定百分比的最新渐变，以避免陷入局部最小值。只有参与者事先就网络架构达成一致[77]，此过程才有效。

可以以各种方式模糊与PS共享的参数。除了仅上传所有梯度的一小部分之外，参与者还可以选择高于阈值，在特定范围内或与差异隐私过程一致的噪声的某些参数。

**5背景**

监督机器学习算法采用标记数据并生成分类器（或回归器），使其能够准确地预测之前未见过的新实例的标签。机器学习算法遵循归纳学习原则[84]，其中它们从一组示例到一般规则，该规则适用于来自与训练集相同的分布的任何数据。给定来自p（x，y）的独立且相同分布的（iid）样本，即，其中∈和∈{1,2，..}，它们求解以下优化问题找到一个准确的分类器：



其中 表示学习机，即，对于任何输入x，它提供类标签y的估计。 L（w，y）是一种损失函数，用于测量用w错误分类y的误差。 Ω（θ）是一个正则化器（独立于训练数据），可避免过度拟合。监督学习算法，如支持向量机（SVM）[76]，随机森林[10]，高斯过程（GPs）[71]，当然还有深度神经网络[29]，可以通过这个通用框架来描述。

深度神经网络正在成为解决具有高维强相关输入的大型数据库的机器学习问题的首选武器，因为它们能够提供显着的准确性增益。他们的改进是基于另外学习分类器中的功能。在深度学习之前，在处理高维强相关输入（例如，图像或语音）的问题中，为减少维度和相关性而构建的人工设计特征被馈送到选择的分类器。深度神经网络革命表明，这些特征不应该是人工设计，而是从数据中学习，因为手工编码的特征缺少相关信息，无法为可用数据产生最佳结果。深度神经网络学习对每个问题有意义的有用特征，而不是依赖于最佳猜测。深度神经网络结构旨在利用输入中的相关性来学习理想的最佳分类特征。在提供顶层的准确预测之前，需要深层结构来分阶段提取这些特征，从较低层的局部特征移动到较高层的全局特征。当数据集的大小和丰富度增加时，这些结果变得不言而喻。

学习机f（x;θ）在估计参数 中总结训练数据库。从学习机器及其估计的参数，可以恢复训练数据库的相关特征，如果不是完整的训练示例。因此，想要从原始训练数据中学习特征的对手如果能够访问学习机，则可以这样做。例如，SVM存储来自中的每个类的原型示例，并且GP存储所有训练点，因此对手在那些分类器中学习每个类的原型示例是没有挑战的。对于深度神经网络，与D中训练点之间的关系更为微妙，因此研究人员试图证明隐私在这些网络中是可能的[77]。但是模型反演攻击[27,28]已经证明我们可以恢复看起来类似于训练集中的输入（例如图像），向对手泄漏关于每个类看起来如何的信息。并且随着深度神经网络被训练有未处理的输入，这些攻击恢复了原始输入的原型示例。

重要的是要强调这是任何机器学习算法的固有属性。如果算法已经学习并且它提供了准确的分类，则可以访问模型的对手可以从类中获取信息。如果攻击者可以访问模型，它可以从每个类中恢复原型示例。如果分类器需要敏感或私人信息以最佳地执行，则学习机可能潜在地将该信息泄露给对手。我们不能两种方式都实现，学习机器成功学习，或数据保密。

**5.1模型反转攻击的局限性**

模型反演攻击以一种简单的方式工作[27,28]：一旦网络被训练，我们就可以遵循用于调整网络权重的梯度，并获得网络中所有代表类的反向工程示例。对于那些我们没有先验信息的类，我们仍然可以恢复原型示例。这种攻击表明，任何准确的深度学习机，无论如何训练，都可能泄漏有关它可以区分的不同类的信息。

此外，模型反转攻击可能只恢复与定义该类的实际数据几乎没有相似性的原型示例。这是由于深度学习机的丰富结构，其中输入空间的广泛区域被高度精确地分类，但遗漏了其他东西[31,81]。如果是这种情况，对手可能会认为他在获取无意义的信息时已经恢复了该类的敏感信息。例如，我们请读者参考[81]中的图5，其中六个训练图像的校车，鸟，寺庙，肥皂分配器，螳螂和狗已被略微调整被归类为鸵鸟（Struthio camelus） ），虽然他们仍然看起来像原始图像。在[31]中，作者在图5中显示了类似于模型反演攻击的程序。随机生成的图像，加上来自深度信念网络的梯度信息，产生一个被归类为飞机的随机图像。深度神经网络的结构是如此之大和灵活，即使图像对人类看起来不像它，它也可以被愚弄成一个准确的标签。

因此，任何模型反演攻击都可以从训练有素的深度神经网络中获取私人信息，但是它可以落在输入空间的无代表部分，看起来不像每个类定义的真实输入。 ML社区的广泛研究表明，GAN生成的样本与训练数据非常相似，因此来自我们攻击的结果揭示了与平均样本或模型反演预期的聚合信息相比模型反转型攻击能获取更多关于训练数据的敏感信息。

**5.2生成性对抗网络**

解决[31,81]中突出显示的问题的一种方法是生成更多的训练图像，以便覆盖更大部分的空间。这可以通过Generative Adversarial Networks（GANs）[30]来实现。

GAN程序针对生成性深度学习网络建立了一个判别式深度学习网络。在原始论文[30]中，训练判别网络以区分来自原始数据库的图像和由GAN生成的图像。首先用随机噪声初始化生成网络，并且在每次迭代时，训练它以模仿判别网络的训练集中的图像。由GAN过程解决的优化问题可以概括为



其中是来自原始数据的图像，是随机生成的图像（例如，均匀地分布在0和255之间的每个像素）。设是一个判别性的深度神经网络，给定一个图像，产生一个类标签，让表示它的参数。设为生成深度神经网络，给出随机输入产生图像。

训练程序的工作原理如下。首先，我们计算上的梯度，以最大化辨别深度神经网络的性能。因此，能够区分来自原始数据的样本，即，以及从生成结构生成的样本，即。其次，我们计算上的梯度，因此从生成的样本看起来像原始数据的完美复制品。

当判别网络不能区分原始数据库的样本和生成网络生成的样本时，该过程结束。论文的作者[30]证明了以下定理：

**Theorem5.1**。当且仅当时，实现（2）中的虚拟训练标准的全局最小值。

该定理表明，当GAN生成看似来自原始数据集的图像时，对抗性游戏结束。

在[32]中，作者表明，在无限样本限制中，生成网络将从原始训练分布中抽取样本。但它也承认GAN程序不会收敛。在最近的一篇论文[72]中，作者大大改进了GAN的训练，包括改进密度模型收敛的新特征。

**6威胁模型**

我们的威胁模型遵循[77]，但依赖于积极的内部人员。

 对手假装是协作深度学习协议的诚实参与者，但试图提取有关他不拥有的一类数据的信息。对手也会秘密地影响学习过程，以欺骗受害者发布有关目标类的更多细节。这种对抗性影响使我们的攻击更有效，例如，仅对最终训练的模型应用模型反演攻击[27]。此外，我们的攻击适用于更一般的学习模型（可以实现GAN的模型），包括模型反转攻击众所周知无效的模型（例如，卷积神经网络）。

具体来说，我们考虑以下场景：

•对手在隐私保护协作深度学习协议中充当内部人员。

•对手的目标是推断出他不拥有的标签的有意义的信息。

•攻击者不会破坏向参与者收集和分发参数的中央参数服务器（PS）。也就是说，参数服务器或我们示例中的服务提供者不受对手的控制。在我们的实际例子中，对手是一个成熟的内部人员，不必为服务提供商工作。

•对手处于活动状态，因为他直接操纵值并在本地构建GAN。与此同时，他遵循受害者所看到的协议规范。特别地，对手轮流，遵循参数选择过程，上传和下载预先商定的正确量的梯度，并且根据协作学习过程的要求对上传的参数进行模糊处理。

•如 [77]，它认为所有参与者都对共同的学习目标有所了解。这意味着对手知道模型结构，特别是其他参与者的数据标签。

•与模型反演[27]中的静态对手不同，我们的对手可以自适应并在学习过程中实时工作。对手将能够通过共享特制的渐变来影响其他参与者，并诱使参与者泄露有关其本地数据的更多信息。这是可能的，因为分布式学习过程需要在成功之前运行几轮。

**7建议攻击**

对手A参与协作深度学习协议。所有参与者都提前同意共同的学习目标[77]，这意味着他们就神经网络架构的类型以及进行培训的标签达成一致。

设V是声明标签[a，b]的另一个参与者（受害者）。对手A声明了标签[b，c]。因此，虽然b是共同的，但A没有关于类a的信息。对手的目标是尽可能多地推断出有关元素的有用信息。

我们的内部人员使用GAN来生成看起来像受害者的a类样本的实例。内部人员将来自a的这些假样本作为c类注入分布式学习过程。通过这种方式，受害者需要更加努力地区分类别a和c，因此将揭示关于类别a的更多信息而不是最初的预期。因此，内部人员模仿来自a的样本并使用受害者来丰富他在训练前忽略的类别的知识。 GAN最初设计用于密度估计，因此我们可以从a分类器的输出中学习数据的分布，而无需直接查看数据。在这种情况下，我们使用此属性来欺骗受害者，以提供有关内部人员未知的类的更多信息。

为简单起见，我们考虑前两个参与者（对手和受害者），然后扩展我们的攻击策略以考虑多个用户。每个玩家都可以声明任意数量的标签，并且不需要重叠这些类。

（1）假设两个参与者A和V.建立并商定共同的学习结构和目标。

（2）V声明标签[a，b]和A标签[b，c]。

（3）针对多个时期运行协作深度学习协议，并且仅当参数服务器（PS）处的模型和两个本地模型达到高于特定阈值的准确度时停止。

（4）首先，受害者训练网络：

（a）V从PS下载一定百分比的参数并更新其本地模型。

（b）V的本地模型在[a，b]上训练。

（c）V将其本地模型的参数选择上传到PS。

（5）其次，对手训练网络：

（a）从PS下载一定百分比的参数并更新他的本地模型。

（b）A训练他当地的生成对抗网络（受害者不知道），以模仿受害者的a类 。

（c）A从GAN生成样本并将其标记为c类。

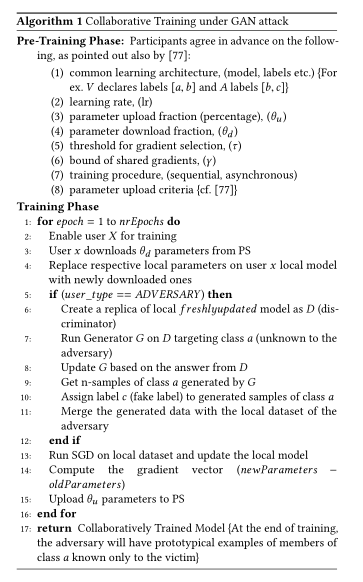
（d）A的本地模型在[b，c]上训练。

（e）A将其本地模型的参数选择上传到PS。

（6）在4）和5）之间迭代直到收敛。

上面5b）和5c）中突出显示的步骤表示对手执行的额外工作，以尽可能多地学习目标标签a的元素。该过程如图4所示。算法1报告了对多个用户的攻击概括。

只要A的本地模型随着时间的推移提高其准确性，GAN攻击就会起作用。另一个重要的一点是，即使采用差分隐私或其他混淆技术，GAN攻击仍然有效。它不是对差异隐私的攻击，而是对协作深度学习的建议用途。虽然结果质量可能会下降，但我们的实验表明，只要模型正在学习，GAN也可以改进和学习。当然，可能总是存在可能阻止攻击的设置。这可以通过设置更强的隐私保证，释放更少的参数或建立更严格的阈值来实现。然而，正如[77]中的结果所示，这些措施导致模型无法学习或比集中数据训练的模型表现更差。最后，即使在部署差异隐私时攻击也是有效的，因为生成 - 判别协同学习的成功仅依赖于判别模型的准确性而不依赖于其实际梯度值。



**8实验设置**

[77]的作者为我们提供了实现完整的分布式协作学习系统的源代码。我们的攻击是通过实施差异隐私来实现的。

**8.1数据集**

我们在两个众所周知的数据集上进行了实验，即MNIST [46]和面部的AT＆T数据集[73]（a.k.a.Olivetti面部数据集）。，

8.1.1 MNIST图像数据集。 MNIST是几种深度学习应用中首选的基准数据集。它由手写的灰度图像组成，数字范围从0到9.每个图像都是32×32像素并居中。该数据集包括60,000条训练数据记录和10,000条记录作为测试数据。

8.1.2面部的AT＆T数据集（Olivetti数据集）。先前在[27]的工作中使用的AT＆T数据集包括在不同位置拍摄的几个人的面部的灰度图像。我们实验中使用的版本包含400张64×64像素的图像。 数据集包含40个不同人的图像，即每人10张图像。

对于这些实验，我们没有对数据进行任何预处理。对数据执行的唯一处理是将每个图像缩放到[ - 1，+ 1]范围，类似于[70]。这样做是为了采用[70]的最先进的发生器模型，它在其最后一层具有双曲正切tanh激活函数，因此也在[-1，+ 1]范围内输出结果。

**8.2框架**

我们在Torch7科学计算框架上构建我们的实验。Torch是最广泛使用的深度学习框架之一。感谢LuaJIT 4，它是一种基于Lua 5的脚本语言，它提供了快速有效的深度学习模型构建。

**8.3系统架构**

我们在MNIST和AT＆T的实验中使用了基于卷积神经网络（CNN）的架构。网络的层基于nn.Sequential（）容器顺序地彼此连接，使得层处于前馈完全连接的方式。

对于MNIST（图15），模型由两个卷积层组成，nn.SpatialConvolutionMM（），其中tanh函数在转发到最大池层之前应用于每个层的输出，nn.SpatialMaxPoolinд（ ）。第一个卷积层有一个大小为5×5的卷积核，它需要一个输入平面，它产生32个输出平面。而第二个卷积层需要32个输入平面并产生64个输出平面，它有一个大小为5×5的卷积核。在最后一个最大池化层之后，数据在256的张量上重新整形，在该张量上进行线性变换。应用将尺寸为256的张量作为输入并输出大小为200的张量。然后将tanh激活函数应用于输出，然后进行另一个线性变换，其将尺寸为200的张量作为输入并输出张量大小为11.我们将输出层从10修改为11，其中第11个输出是对手用G生成的结果训练的地方。和Goodfellow等人一样[30]，第11类是放置“假”图像的类。更多细节在第9节中提供。模型的最后一层是LogSoftMax层，nn. LogSoftMax（）。

面部的AT＆T数据集中的图像较大（64×64）。因此，我们构建了一个卷积神经网络（图17），由三个卷积层和三个最大池层组成，最后是完全连接的层。与MNIST架构一样，tanh用作激活函数。该模型具有大小为41的输出层，即40为人的真实数据，1为对手为其感兴趣的类别提供重建的类。由于面部比数字更难重建，因此我们以不同的方式实现了算法1。对于这种情况，发生器G每个时期更多次查询鉴别器D（对手的训练数据的大小除以批量大小）以更快地改进。

用于MNIST相关实验的发生器（G）架构，图16，由4个卷积层组成，对应于来自火炬'nn'库的nn.SpatialFullConvolution（）。批量标准化，nn.SpatialBatchNormalization（），应用于除最后一层之外的所有层的输出。激活功能是经过整流的线性单位函数nn.ReLU（）。模型的最后一层是双曲正切函数tanh，用于将G的输出设置为[-1，+ 1]范围。由于AT＆T图像较大（64x64），因此G具有附加（第5）卷积层。使用[68]中的技术自动计算所需的卷积层数。 G将100维均匀分布[14,70]作为输入，并将其转换为MNIST的32x32图像或AT＆T的64x64图像。如[14]所示，我们用0均值和0.02标准差初始化了发电机的权重。虽然[70]将此初始化函数应用于D和G，但我们仅对G执行，因为D是在所有参与者之间共享的模型。

上面描述的两种体系结构在图16和18中表示为由Torch7打印出来的。

有关Torch7提供的体系结构的更多详细信息，请参阅附录A.

**8.4超参数设置**

对于与MNIST相关的实验，我们将协同训练模型和鉴别器模型的学习率设置为1 e - 3，学习率衰减1 e - 7，动量0和批量大小64。

对于AT＆T相关实验，我们将学习率设置为0。然而，对于关于多参与者场景的AT＆T实验，我们使用批量大小为1.我们将其余的超参数保持类似于MNIST案例。学习率为0。 02工作得更好，因为它允许更多的随机性，从而允许模型更快地收敛。

DCGAN [70]的作者使用Adam optimize r，学习率为0。在DCGAN [14]的火炬实施中提供了0002和动量项β1为0.5。我们修改了使用随机梯度下降（SGD）的过程，对于这种配置，发电机的学习率为0.02，效果更好。

**9实验**

我们现在评估我们的GAN程序如何从其他参与者恢复记录。我们将实验重点放在包含图像的MNIST和AT＆T数据集上。然而，原则上，我们的对抗策略可以扩展到其他类型的数据，例如音乐，音频，医疗记录等。我们首先在传统环境中的GAN攻击与模型反演进行比较。如前所述，模型反演有一些局限性，可能对某些类型的神经网络无效。虽然从理论角度来看这可能是明确的，但我们也在第一个实验中提供了这种说法的实验证据。

在第二组实验中，我们展示了GAN攻击如何在分布式设置中工作，其中对手无视某些或所有标签的内容，参见图7。

在第三组实验中，我们表明在将深度神经网络的参数上传到参数服务器之前将噪声添加到它们中并不能防止我们的GAN攻击。通常，部署记录级差异隐私以模糊模型参数对我们的攻击是无效的。 GAN的功效仅受判别器的准确性限制。

**9.1 MI攻击与GAN攻击**

在第一个例子中，我们比较了模型反演（MI）和GAN攻击，并为它们提供了所有数据。对手可以使用训练有素的模型。

对于MI攻击，我们在MNIST数据集的所有60,000个训练样本上训练卷积神经网络。一旦深度神经网络被训练，我们在[27]中应用模型反演攻击。然而，我们没有像[27]中那样逼近导数，而是收集了模型根据给定的输入和感兴趣的标签（类）计算出的精确梯度。结果显示在图5中.MI适用于MLP网络但显然无法通过CNN。这与作者获得类似结果的工作[78]一致。在处理更复杂的学习结构时，MI似乎无效。虽然相关信息在网络中，但梯度可能会将我们带到输入空间的一个区域，该区域不代表我们尝试恢复的数据。

对于GAN方法，我们采用[70]中的DCGAN架构，以及[14]中的火炬实现。该模型由鉴别器（D）与DCGAN发生器（G）组合而成。我们使生成器模型与MNIST类型的图像兼容，并使用[68]中提出的方法，以便我们的代码可以自动计算所需的卷积层数。有关架构的更多详细信息，请参阅8.3节。我们进行了10次实验（MNIST数据集中每个类别一次），我们让模型进行训练，直到D达到的精度达到97％以上。我们在图5中显示结果。

**注意显着差异**：在GAN攻击中，生成模型与判别模型一起训练，而在MI中，判别模型仅在训练阶段结束时访问。但是，这种对模型的实时访问是使我们的攻击适用于协作深度学习的原因。

**9.2 GAN攻击无差异隐私的协作学习**

现在我们在一个像[77]中提出的协作环境中设置GAN攻击。我们使用第7节中描述的模型，如图4所示。

9.2.1对MNIST的实验。我们不是为每个用户使用两个标签，而是为第一个用户使用五个标签，为第二个用户使用六个标签。第一个用户可以访问0到4的图像（标签为1到5），第二个用户（对手）可以访问5到9的图像（标签6到10）。攻击者使用其第六类来提取第一个用户的一个标签上的信息。

结果显示在图6中。对于每个检索到的图像（底行），我们在其上方放置来自第一个用户的实际训练图像（我们显示最接近L1范数的图像）。我们用三种不同的参数设置重复了实验。在（a）中，用户上传并下载整个模型。在（b）中，用户下载完整模型，但仅在每个时期上传10％的参数。最后，在（c）中，上传和下载仅为10％。

9.2.2 AT＆T的实验。我们对AT＆T数据集进行了类似的实验，该数据集由来自40个不同人的面孔组成。最初，我们测试了两个参与者的场景，其中一个是受害者，另一个是对手。我们将前20个班级分配给第一个用户，剩下的20个班级分配给了eadversary。额外的班级给予对手以影响培训过程。我们运行了几种具有不同上传速率的配置，参见图8.结果显示，攻击者可以对目标人脸进行相当好的重建。有些图像比其他图像噪声更大，但鉴于该特定数据集的模型精度往往保持较低，因此很难证明这一点。

我们还实现了一个多参与者场景，见图7，有41个参与者，其中40个是诚实的，1个是对抗性的。每个诚实的参与者拥有与一个类相关的图像作为训练数据，而对手没有他自己的训练数据。即，对手仅训练由发生器（G）产生的图像。即使启用了差分隐私，结果（θu= 1，θd= 1）也非常好（图7）。

**9.3 GAN攻击，无影响与对协作学习的影响**

人们可能想知道假标签对协作学习的影响。回想一下，由生成模型生成的图像被放入一个人工类中，以欺骗受害者释放目标类的更精细细节。我们测量了对抗性影响的影响，我们通过实验证实其效果非常显着：学习速度更快，而且对手获取的信息也明显更好。我们进行了实验，直到测试集上模型的准确度高于97％，协同训练CNN模型。对手和受害者的数据集彼此分开，没有共同的标签。

在图9和图10中，当我们试图分别从第一个用户恢复0和3时，我们显示了第7节中提出的标准GAN攻击的被动GAN攻击的结果。在第一行中，我们显示来自被动攻击的图像而没有影响，并且在底行中显示来自标准过程的图像以及人工类的影响。对抗性影响的效果是显而易见的，即使在每个参与者仅50个时期之后，图像看起来更清晰和更清晰。在我们的实验中，我们注意到一旦模型的准确度达到80％，G就会开始产生良好的结果。

**9.4 GAN攻击差分隐私协作学习**

在[77]中已经提出，差异隐私可用于向深度学习模型的参数添加噪声“以确保参数更新不会泄漏关于训练数据集中任何单个点的过多信息。”（引自[77]。）作者只考虑被动对手，并依靠差异隐私来减少可能来自参数更新的泄漏。他们强调了两种潜在泄漏案例：（i）如何进行梯度选择和（ii）为了解决这两个问题，[77]中的方法依赖于稀疏向量技术[23]。对于协作学习过程的每个时期（迭代），它们为每个定义了一个完整的隐私预算ε。参与者。这个预算被分成c部分，其中c是每个时期可以共享的梯度总数。梯度的一部分被随机选择，使得它们高于阈值（τ）。他们专注于8/9的预算。 / C的参数的选择，并使用剩余的1/9以释放值。他们依靠拉普拉斯机制在选择期间添加噪声以及共享参数，与分配的隐私预算一致。

为了证明记录级别的差异隐私对于活跃的对手是无效的，我们在两个参与者（A和V）之间运行了协作学习过程，启用了差异隐私。我们保持参与者的数据集不同：在MNIST实验中，V只有0到4级的记录，A有5到9级的记录加上A引入的人工类。对于AT＆T实验，V有数据集中前20个类的记录，接下来20个类中的A加上9.2小节中的人工类。在我们的实验中，我们将下载速率（θd）固定在100％，阈值（τ）固定在0.0001，范围（γ）固定在0.001，类似于[77]。在图11和14中，我们提供了每个参数（εc）100和10的隐私预算以及不同的上传速率（θu）的结果。尽管模型在差异隐私约束下收敛需要更长的时间，但我们的结果证明了我们的主张，即只要训练过程成功且模型收敛，G就可以产生良好的结果。

关于ε值。我们观察到[77]中的ε非常大，差异隐私的影响可能是有问题的。然而，对于小ε，本地模型无法学习，协作学习完全失败。这与[77]中报道的一致。实际上，我们在更严格的隐私限制下进行了实验。发电机未能产生良好的结果，但因为本地模型根本无法学习。在图12和图13中，我们展示了一个示例，其中我们设置了更严格的隐私限制，这转化为更强的差异隐私保证，并且GAN无效。同时，这是预期的，因为本地模型和参数服务器中的模型无法学习并且协作学习没有发生。可以使用[1]中的技术将ε降低到一位数值。但是，我们再次强调，我们的攻击与使用的任何记录级DP实现无关。只要鉴别器正在学习，GAN就会产生良好的样本（见图2）。

**10结论**

在这项工作中，我们在协作环境中提出并实现了一类新的深度神经网络主动推理攻击。我们的方法依赖于生成对抗网络（GAN），并且比现有的信息提取机制更有效和更通用。我们相信，我们的工作将对现实世界产生重大影响，因为主要公司正在考虑分发，联合或分散的深度学习方法来保护用户的隐私。

我们研究的主要观点是，协作学习不如它应该取代的集中学习方法更令人满意。在协作学习中，任何用户都可能违反系统中其他用户的隐私而不涉及服务提供商。

最后，我们无法针对我们的攻击制定有效的对策。解决方案可能依赖于安全的多方计算或（完全）同态加密。然而：（1）引入了隐私保护协作学习作为避免这些昂贵的加密原语的方法[77]，以及（2）我们基于它们探索的解决方案仍然容易受到某些形式的攻击。另一种方法是考虑不同粒度的差异隐私。用户或设备级DP可以防止本文中设计的攻击。然而，目前尚不清楚如何利用设备，类或用户级DP构建用于协作学习的真实系统（例如，用户以不可预测的方式表现和共享数据）。因此，我们将此主题留待将来工作。