深度学习中的生成对抗网络攻击设计与实现

1. 前言

深度学习算法由于可以从训练集中自动学习其中的特征，近年来，深度神经网络在诸多领域受到了广泛的关注。比起目前为止的机器学习算法，它大大提高了这些领域的表达能力。尤其是卷积神经网络和循环神经网络，这两个神经网络算法得到了广泛的应用。另一方面，在这个大数据的时代，在服务器上进行机器学习的研究已不是什么难事，有大量的学者致力于此，而对这些用于机器学习而上传的个人用户数据，其安全性如何亟待考虑。考虑到集中型学习中，由于对服务器的攻击会导致用户上传到服务器的个人信息泄露，现在有学者提出了一种协同学习的方法，允许用户共享训练后参数集而不是数据本身。然而看似安全的这种学习方式其实蕴藏了极大的风险。本研究提出一种用生成对抗网络模型（GAN）去模拟攻击在这种深度学习用场景下的用户，诱使受害者泄露更多的个人信息，而通过利用GAN的这种攻击手法，对于协同型学习场景中的每个用户来说，他们都有可能成为攻击者来获取其他用户的个人信息。本研究旨在给予协同型深度学习的用户以警示。

2.研究现状

2.1.深度学习

2006年，多伦多大学的教授—Geoffrey Hinton在《Science》上发表了一篇文章《Reducing the Dimensionality of Date with Neural Networks》[15]，该篇文章提出了深度学习的概念：机器学习中一种基于对数据进行特征学习的方法。并且指出：1.多隐层的人工神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类；2.深度神经网络在训练上的难度，可以通过“逐层初始化”（layer-wise pre-training）来有效克服。在这篇文章中，逐层初始化是通过无监督学习实现的。这篇文章在当时引发了深度学习在学术界以及工业界的浪潮，使得人们开始关注深度学习。

2012年，Hinton教授的学生Krizhevsky[16]利用深度卷积神经网络(Convolution Neural Network)在2012年国际大规模视觉识别大赛(ImageNet Large Scale Visual Recogonition Chanllenge;简称：ILSVRC)中以较大的优势胜过了传统的目标检测算法。东京大学的松尾贵教授说道, “之前的人工智能的世界是以精度一点点向上提高，错误率一点点下降作为目的，进行着机器学习的研究。而这一次，据ILSVRC，这次多伦多大学开发的SuperVision将错误率一次性从２６％降到了１５％”。

深度学习神经网络与传统的浅层神经网络的最大区别在于前者具有更多的 层(通常在 3 层以上)，从而能够提取隐藏在数据中的某些深层次的模式。卷积神经网络与传统的 BP 神经网络最大的区别除了层数 外，前者采用了针对二维图像的处理结构，即卷积层和池化层。其中，卷积层 实现特征的提取，池化层实现特征的量化。池化操作的目的是为了对卷积提取 的特征进行空间的量化，这种量化操作能够在一定的条件下实现平移、旋转和 缩放等形变的不变性。通过若干层卷积层和池化层的操作，一个卷积神经网络 能够提取输入图像多层次的特征，并且高层特征较底层特征更稳定，也越适合 用来做模式识别和目标检测等任务。

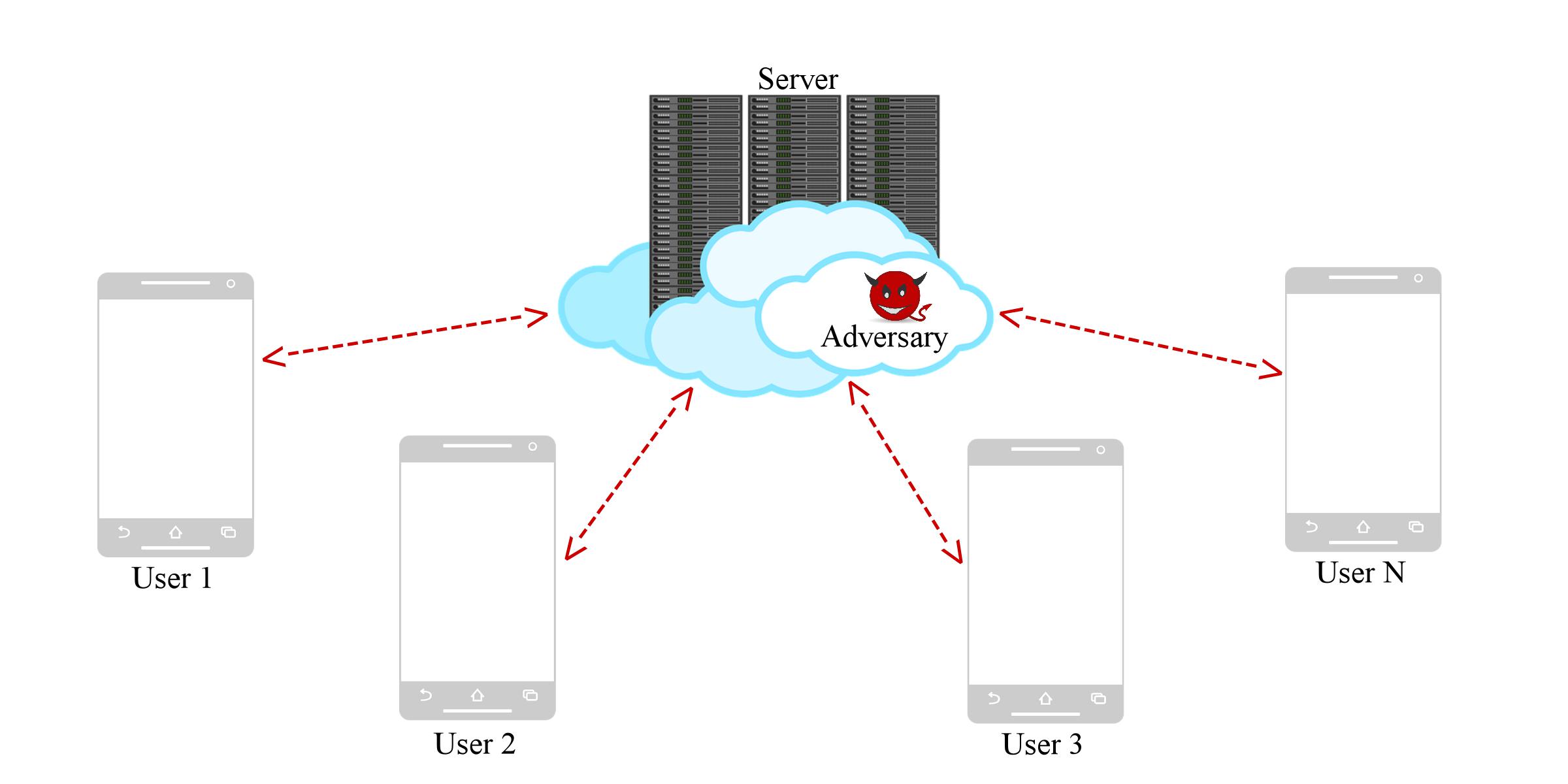
2.2生成对抗网络（GAN）

生成对抗网络由两部分构成，判别模型和生成模型。生成模型G（generator），他的目的是生成一张尽量真实的图片。于此同时我们有一个图像判别的模型D（discriminator）,他的目的是区分这张图片是否是真实的。而生成模型与判别模型均由一种机器学习网络构建，不断的调整生成模型G和分类模型D，一直到分类模型不能将生成的图片和真实的图片区分出来为止，那么我们就可以认为，生成模型生成了一张真实的图片。而由此生成的图片具有足以以假乱真的效果，可以生成与原图极其相似的图片，但如果我们将这些新生成的图片重新投入到模型中训练的话，会得到不同的分类结果。GAN模型从数学层面上看，一方面generator从训练数据中产生相同分布的samples，discriminator 则还是采用传统的监督学习的方法，最终达到所谓的纳什均衡，即对生成器来说目前采取的策略生成的图片为最接近原图的，同时对辨别器来说此时也是识别出生成器生成的图片的真假时的最优策略，达到全局最优。

2.3深度学习中的隐私问题

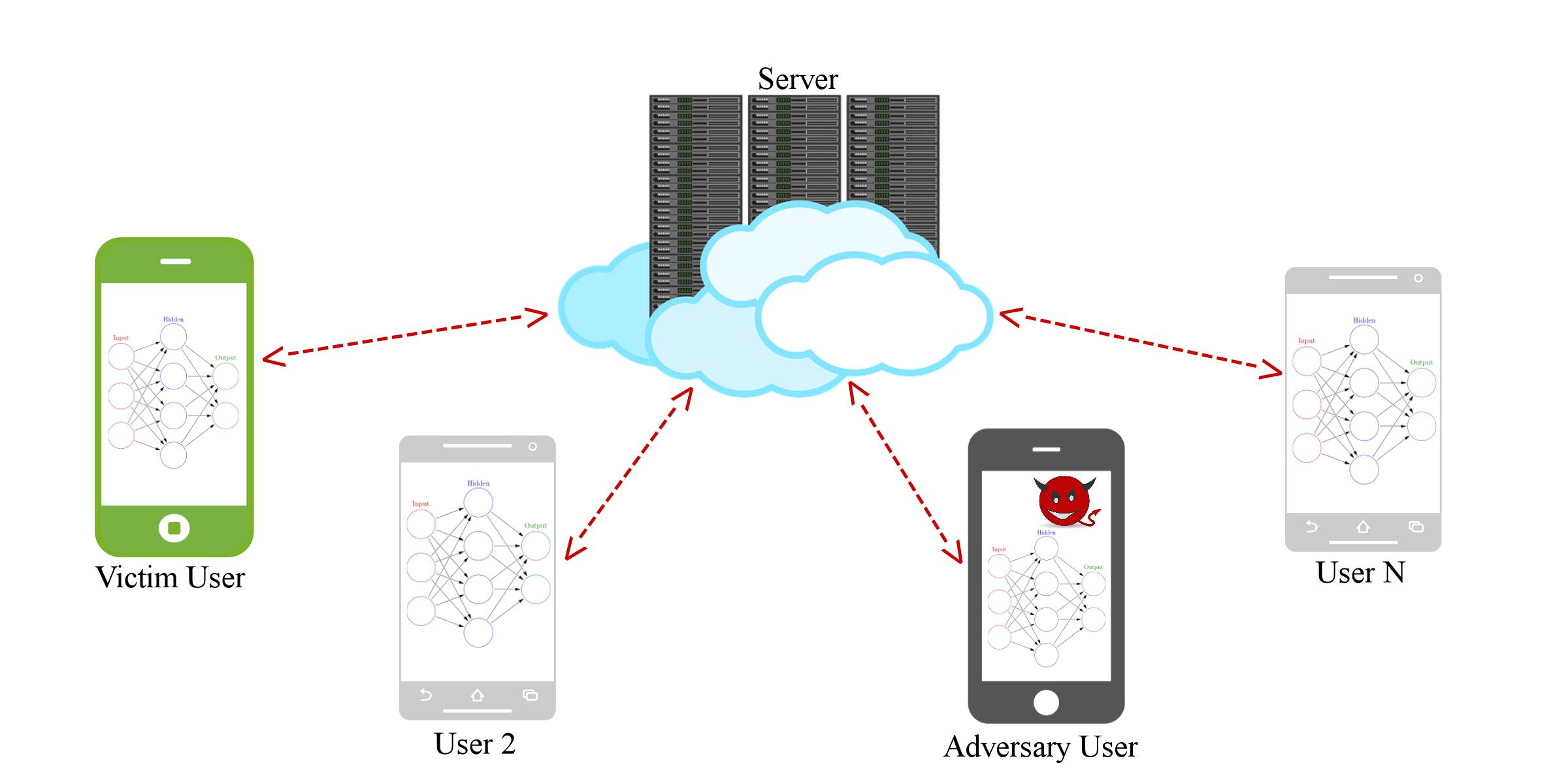
2.3.1集中型深度学习的问题

研究人员考虑到深度学习的隐私影响。模型通常以集中的方式进行训练，即集中型深度学习，所有数据都传给中央服务器，用相同的训练算法进行处理。但考虑到如果数据是一个收集用户的私人数据，包括生活习惯、个人图片、地理位置等，如果中央服务器遭到攻击，则攻击者可以很容易地访问这些对于用户来说敏感的信息，最终造成个人信息的泄露。另一方面，许多数据拥有者，例如医疗机构可能由于数据隐私和保密问题，不能通过共享数据进行大规模的深度学习，而且当数据范围过于狭小时，训练过程容易出现过拟合的问题。就为解决这个问题，研究人员最近提出了协同型深度学习模式。



2.3.2协同型深度学习的问题

考虑到集中型深度学习的弊端，由研究人员提出了如下的协同型学习模型，各用户在自己的设备端进行模型的学习，然后通过上传自己的模型参数的一部分达到共享参数，协同学习的目的。而其他用户就可以从中央服务器下载这些参数进行模型的学习。通过这种方式具有敏感度的数据着保存在用户自己的设备中，无需上传到中央服务器，保护了隐私。同时通过共享训练参数，其他用户可以进行大规模的学习与训练，得到具有更加精确结果的训练模型，在很大程度上防止了过拟合的问题。然而令人担忧的是，本论文提出的这种策略，通过在用户端构建一个GAN模型可以轻易地盗取其他任何一个用户设备中的隐私信息，即使他们并没有把这些数据统一上传到中央服务器。



3.研究内容

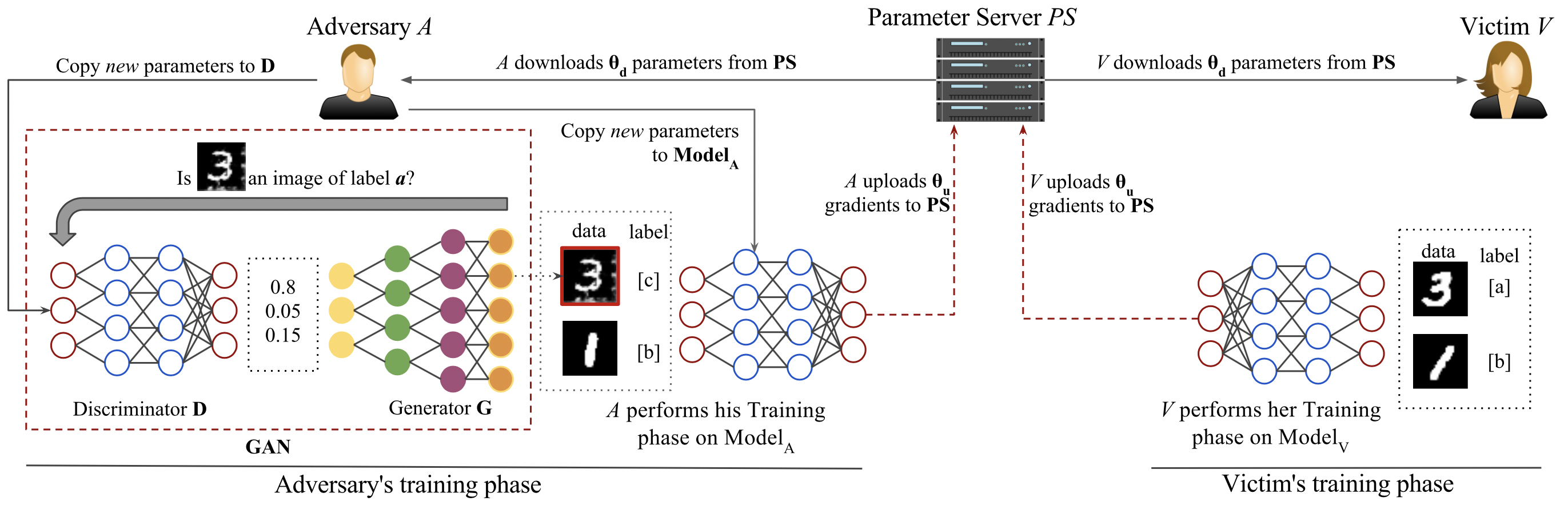
3.1预期的攻击方法

所有参与者都预先商定了一个共同的学习目标，这意味着他们就神经网络结构的类型和培训的标签达成一致。让V是一个参与者（受害者），声明标签[ A，B ]。攻击者A 声明标签[ B，C ]。因此，虽然B是共同的，A没有关于A类的信息。攻击者的目标是尽可能多地推断关于A中的元素的有用信息。攻击者使用了一个GAN来生成类似于来自受害者A类样本的实例。攻击者将这些假冒样本申明成标签 [C], 注入协同型学习过程中。这样，受害者就需要更努力地区分A和C类，因此将比最初预想的更多地揭示A类的更多信息。因此，攻击者通过模仿A的样本，并利用受害者不断改善他在训练前忽视的关于A的信息。

为了简单起见，我们这里只考虑两个参与者的情况（对手和受害者），然后扩展我们的攻击策略来适应多个用户参与的情况。每个参与者可以声明任意数量的标签，而标签之间并不需要重叠。

3.2实际攻击场景搭建

实际搭建时预期使用lua编写多方协同学习场景，GAN及部分学习网络的实现由python和tensorflow搭建。这里假设上文提到的协同学习场景中只有两个参与者（A，V），其中A为攻击者，V为受害者。假设受害者V有两类图片“3”，“1”，分别申明标签 [a], [b]，攻击者A只有数字“1”，标签为[b]。同时在攻击者一侧有GAN的网络模型，用以从中央服务器下载下来的参数中生成和受害者V的图片“3”以假乱真的图片，并将假冒的图片“3”申明为标签 [c]。攻击者的最终目的是获取受害者 V 设备中的有关于标签[a],数字”3”的信息。



4.总结

在多方学习的场景中，各个协同方训练一个本地深度模型，而云端服务器通过聚合本地模型共享的特征或参数得到全局深度模型并最终发布给各协同方。多方协同系统的搭建可以为地理上分布式的大量终端设备提供大规模智能化深度学习环境，在保护终端隐私数据的基础上协调云端和终端的决策。但是，多方协同深度学习系统的训练过程中有可能会发生隐私数据泄露。本研究中利用的生成对抗网络模型，通过生成以假乱真的图片，并将训练用的参数和特征值上传至网络，致使受害者不断地泄露相关的个人隐私情报，由此可预想到大数据时代信息安全的重要性，本研究内容对深度学习中的安全问题有很好的启发作用。

**参考文献**

[1]the Edge and End Devices[C] // Proceedings of 37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017, Atlanta, GA, USA, June 5-8, 2017. 2017: 328–339.

[2] Shokri R, Shmatikov V. Privacy-Preserving Deep Learning[C] // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS 2015. 2015: 1310–1321.

[3] Song C, Ristenpart T, Shmatikov V. Machine Learning Models that Remember Too Much[C] // Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017. 2017: 587–601.

[4] Phong L T, Aono Y, Hayashi T, et al. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption[J]. IEEE Trans. Information Forensics and Security, 2018, 13(5): 1333–1345.

[5] Hitaj B, Ateniese G, Pérez-Cruz F. Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning[C] // Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017. 2017: 603–618.

[6]原田達也：画像認識と画像復元；情報処理Vol.56 No.7(2015)

[7]松尾豊著（2016）『人工知能は人間を超えるか』KADOKAWA

[8]松尾豊:人工知能の未来とデータ共有の新しい可能性;情報管理vol. 58 no. 8 (2015)

[9]小山　俊隆、鈴木　崇啓、別所　史浩、原田　達也、國吉　康夫：移動体の自己位置情報を利用した画像からの異常領域検出；情報処理学会創立５０周年記念（第７２回）全国大会（2010）

[10]Ray Kurzuweil著、盛杨燕訳 (2016)『How to create a mind』浙江人民出版社

[11]Ray Kurzuweil 著、胡晓姣张温卓玛吴纯洁訳(2016)『The age of spiritual machines』中信出版集团

[12]Jeff Hawkins著、贺俊杰 李若子 杨倩訳（2006）『人工智能的未来』陕西科学技术出版社

[13]李承昊 ，蔡晨阳 ，李 锐 :基于神经网络的图像识别系统的研究;《科技传播》(2015)

[14]李义 :RESEARCH ON IMAGE RECOGNITION METHOD BASED ON CORRELATION LEARNING NEURAL NETWORK ;哈尔滨工业大学 (2015)

[15]河嵜光毅 吉川大弘 古橋武 :時系列データの異常検出を目的とした深層学習における再構築誤差の利用可能性に関する検討;The 29th Annual Conference of the Japanese Society for Arti cial Intelligence(2015)

[16] CS 224N / Ling 284, Neural Networks for Named Entity Recognition

[17]LeCun et al. Gradient-Based Learning Applied to Document Recognition 1998

[18]Tom M. Mitchell, "机器学习", 曾华军等译, 机械工业出版社

[19]Jake Bouvrie, Notes on Convolutional Neural Networks, 2006

[20]Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016