摩石观察

密码是保障网络安全的核心技术和基础支撑，在维护国家安全、促进经济社会发展、保护人民群众利益中发挥着不可替代的重要作用。“云物移大智”的蓬勃发展，5G、智慧城市、互联网＋政务服务的全力推进，离不开用密码技术来保障网络安全、保护数据安全、保证网上诚信。实现网络安全需要密码学与其他学科深入合作，需要密码产业与其他产业的深度融合，需要产学研管用的真诚协作，更需要全社会共同传播密码知识与政策、研究密码应用技术、推进密码应用方案。

为激浊扬清，构建以密码技术为核心、多种技术相互融合的新网络安全体系，推进密码技术科学规范应用，长期深耕于信息安全一线的卫士通公司凝聚了一批国内顶尖的密码专家于2016年建立了摩石实验室。依托密码基础理论，探索密码创新实践，解决密码应用难题，培养密码专业人才，摩石实验室致力于让密码技术更好地服务于网络强国、数字中国和智慧社会。

本着共同的愿景，《信息安全与通信保密》杂志社与摩石实验室精诚合作，专门开辟《摩石观察》栏目。立足于密码本真，反思密码实践，《摩石观察》将以密码人的细致严谨叩问密码创新的真理之门，为广大读者了解、认识、掌握、使用密码技术提供准确规范的参考依据；同时我们期望以密码会友，与理想作伴，热忱邀请有志之士共同探索密码理论与应用的最佳实践，为推动金融等重要领域密码应用与创新发展而奋斗。

深度学习中的隐私保护技术综述

唐鹏[1]，黄征[1,2]，邱卫东[1]
（1. 上海交通大学网络空间安全学院，上海，200240）
2（卫士通摩石实验室，北京，100070）

[摘要] 如今机器学习以及深度学习在各个领域广泛应用，包括医疗领域、金融领域、网络安全领域等等。深度学习的首要任务在于数据收集，然而在数据收集的过程中就可能产生隐私泄露的风险，而隐私泄露将导致用户不再信任人工智能，将不利于人工智能的发展。本文总结了目前在深度学习中常见的隐私保护方法及研究现状，包括基于同态加密的隐私保护技术、差分隐私保护技术等等。

[关键词] 隐私保护；深度学习；同态加密；差分隐私保护
[中图分类号] TP393　　[文献标识码] A　　[文章编号] 1009-8054（2019）06-0055-08

1 研究背景

1.1 隐私泄露风险

2006 年 Hinton[1] 和他的学生在《Science》上发表了一篇名为“Reducing the Dimensionality of Data with Neural Networks”的文章，开启了深度学习在学术界和工业界的浪潮，同时也将人工智能推向了一个新的高潮。目前，在人工智能的几乎所有领域，深度学习技术已经远远超过了传统方法的性能，包括计算机视觉、音频处理、自然语言处理、大数据分析等。

推动深度在各个领域取得巨大成功主要有以下几个因素：

(1) 数据井喷，全球数据中心数据量在未来几年年均增速 40%。

(2) 计算能力突破，基于大型 GPU 集群的强大计算能力，使得训练深度神经网络的速度从 2006 年到 2016 年提升了 255 倍。

(3) 算法突破，算法突破推动 AI 技术成熟和实用化。

对于一些大型的网络企业，用户的照片、语音、视频、文本等数据可以被收集和保存，以备将来使用。正因为如此，深度学习领域中大部分成功的应用都是大型组织，它们既有大量有价值的数据，又有足够的计算能力来训练深度模型，以改进其产品和服务。

虽然深度学习带来了巨大的好处，但它需要收集大量的数据，这些数据涉及用户的隐私信息，例如用户兴趣、爱好、个人信息等，这些隐私数据的泄露会导致不可预估的财产以及生命安全问题。

1.2 隐私保护相关法律

随着互联网的快速发展以及隐私泄露问题日益严重，为减轻隐私泄露带来的负面影响，美国、欧盟、中国等国家正在不断的通过完善数据安全和隐私保护法律法规对企业以及个人进行监管。

美国是最早通过法律法规对隐私进行保护的国家，其在 1974 年通过并发布的《隐私法案》是美国最重要的一部保护个人隐私的法律法规，到 20 世纪 80 年代又先后制定和颁布了《电子通讯隐私法案》《电脑匹配与隐私权法》以及《网上儿童隐私权保护法》。1980 年，经济合作与发展组织（OECD）在《关于保护隐私和个人信息跨国流通指导原则》中揭示了个人信息保护八大原则，即收集限制原则、数据质量原则、目的明确原则、使用限制原则、安全保障原则、公开性原则、个人参与原则和问责制原则。这些指导原则对全球各国的立法产生了巨大的影响，有“已经成为制定个人信息保护文件的国际标准”之称。

我国也在多部法律中对隐私权进行保护，比如在《侵权责任法》中规定了若干种承担侵权责任的方式；《中华人民共和国宪法》第三十八条规定：“中华人民共和国公民的人格尊严不受侵犯。”等等。同时我国还在制定专门的《个人信息保护法》《中华人民共和国网络安全法》等法律对网络数据以及用户隐私进行保护。

本文总结了在使用深度学习的同时保护用户隐私信息的方法，主要包括同态加密隐私保护技术、差分隐私保护技术。同态加密和差分

隐私是密码学中常见的隐私保护手段，这两种方法被运用于深度学习过程中的隐私保护，具有显而易见的效果。

2 同态加密隐私保护技术

2.1 同态加密技术

假设存在加密函数 f，使得明文 M 加密后变成密文 M’，明文 N 加密后变成密文 N’，即 f(M)=M’，f(N)=N’，存在 f 的解密函数 f-1 能够将 f 加密后的密文解密成加密前的明文。将 M’与 N’相加得到 P’，如果解密函数 f-1 对 P’解密后的结果等于 M 和 N 相加的结果，即 f-1 (P’)= f-1 (M’+N’)=M+N，则 f 是可以进行同态加密的加密函数。

同态加密可以分为加法同态、乘法同态以及全同态。加法同态指的是加密算法满足 f(M)+f(N)=f(M+N)，乘法同态指的是加密算法满足 f(M)*f(N)=f(M*N)。而全同态加密指的是一个加密函数同时满足加法同态和乘法同态，全同态加密函数可以完成加减乘除、多项式求值、指数、对数、三角函数等运算 [22][23]。

常见的 RSA 算法对于乘法操作是同态的，Paillier 算法则是对加法同态的，Gentry 算法则是全同态的。

2.2 同态加密技术应用

2.2.1 数据处理与隐私保护

大数据时代下的海量个人信息存储与处理是隐私保护面临的关键问题，用户往往不希望将个人资料、保密文件、隐私信息存储在服务提供商中，而人工智能时代又需要对这些用户信息进行挖掘分析。同态加密是解决这一矛盾的新技术，用户可以将个人敏感信息加密后存储在服务提供商或者云端服务中，服务器可以对密文进行处理以及分析，并将密文结果返回给用户，只有用户能够解密密文结果。

2.2.2 密文检索

当越来越多的加密数据存储在服务器或者云端时，对加密数据的检索成为了一个急需解决的问题。现有的密文检索算法包括线性搜索、公钥搜索和安全索引可以解决对服务端的加密数据进行检索问题，然而这些方法需要花费较高代价并且只能运用于小规模数据集中。基于全同态加密的密文检索方法可以直接对加密数据进行检索，对密文做基本的加法和乘法能够有效降低运算复杂度，同时也不改变相应的明文顺序，既保护了用户的数据安全，又提高了密文的检索效率 [20]。

2.2.3 数字水印

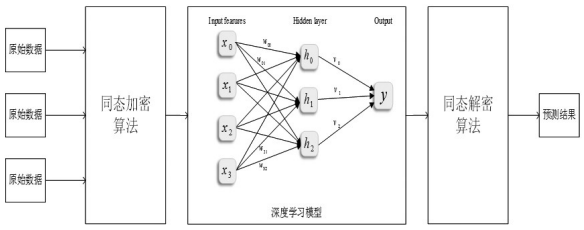
目前数字水印技术已经成熟地运用在数字产品的版权保护中，然而数字水印系统也存在安全挑战，例如对水印算法、水印密钥的恶意攻击，从而破译水印并且伪造水印。使用同态加密技术对数字水印进行加密后嵌入，在检测水印时首先需要对水印进行同态解密，该方法能够有效抵抗数字水印的非授权攻击 [24]。

2.3 基于同态加密的深度学习中隐私保护技术

同态加密的核心是能够直接在密文上做运算，运算结果解密后与明文运算结果相同，这是对用户隐私的最直接和有效的保护手段 [15]。在机器学习和深度学习过程中使用同态加密对数据加密然后分析计算，能够很好地解决许多领域要求数据保密、安全的问题。同态加密可

以确保在密文上进行计算而不进行解密，从而解密的结果与对明文执行相同计算得到的结果相同。由于目前的同态加密方案仍然存在许多局限，如只支持整数数据，需要固定的乘法深度，不能无限期地进行加法和乘法运算，全同态加密不支持比较和取最大值等操作。因此，现有的同态加密方案不能简单地应用于机器学习以及深度学习中。目前常用的解决策略有两种：①通过安全的多方计算来构造一种适合于基于同态加密的机器学习算法的协议，并通过执行该协议来完成该算法；②寻求原始机器学习算法的近似算法，使其仍然可以使用同时不依赖交互方案，并且满足同态加密方案的数据和操作要求[2][3][4]。

图 1 基于同态加密的隐私保护技术流程



在机器学习和深度学习的预测过程中，利用同态加密算法的性质对数据进行加密，然后在加密训练集上进行机器学习以及深度学习的建模训练，同样使用训练好的模型对加密的测试集进行预测，返回的预测结果也是密文，从而有效地保护用户隐私数据。在 2007 年，Orlandi[18] 就提出了利用同态加密技术结合多方安全计算使神经网络具有处理加密数据的能力，并且考虑到了神经网络本身的安全性。2011 年，Barni[19] 等人将基于同态加密的神经网络应用于心电图分类中，可以实现远程服务器对客户提供的生物医学信号进行分类，而不获取任何有关信号本身的信息和分类的最终结果。2016 年，Dowlin[5] 等人提出一种可以应用于加密数据的神经网络 CryptoNets，同时作者证明云服务能够将神经网络应用于加密数据以进行加密预测，并以加密的形式返回这些预测。这些加密的预测可以发送回能够解密它们的密钥所有者。该网络对 minist 数据集的识别精度达到了 99%。在 2017 年，Hesamifard[6] 等在训练阶段使用 Chebyshef 多项式来近似模拟激活函数，从而证明了利用加密数据训练神经网络，进行加密预测，并以加密形式返回预测是可行的和实用的，该方法比 CryptoNets 在 MNIST 数据集上的精度提高了 0.52%。为了解决基于全同态加密技术的机器学习巨大的计算开销问题，Baryalai[7] 等提出了一种非共谋双云模型 (CloudA，CloudB)，该模型使用 Paillier 密码系统提高运算速度，减少运算开销。在训练阶段，加密也可以用来保护敏感数据集。Xie[8] 等利用 Stone—Weierstrass 理论，提出 crypto-nets 在密文上做预测，利用同态加密和对激活函数的修改以及对神经网络的激活函数和训练算法的修改，并证明了所提出的加密网络的构造是可行的。该方法为在不侵犯用户隐私的情况下建立基于云的安全神经网络预测服务奠定了基础。Zhang[9] 等提出了一种基于 BGV 加密方案的保密双投影深度计算模型(PPDPDCM)，给出直接在密文上训练神经网络的解决方案。

目前，利用加密技术来保护机器学习以及深度学习中的用户敏感数据已经取得较大进展，包括在预测阶段以及训练阶段的数据加密。但是在使用同态加密的过程中存在资源消耗的问题，深度学习本身已经消耗了大量的计算资源，

结合同态加密技术将大大增加深度学习的计算量。如何在结合同态加密算法而尽量减少对深度学习性能的影响将是未来主要的研究方向。

3 差分隐私保护技术

3.1 差分隐私保护技术

差分隐私的任务是提供一种关于输出的概率分布的机制或者协议，允许用户对数据进行一定程度的修改，但不影响总体输出，从而使得攻击者无法知道数据集中关于个人的信息，达到隐私保护的作用。

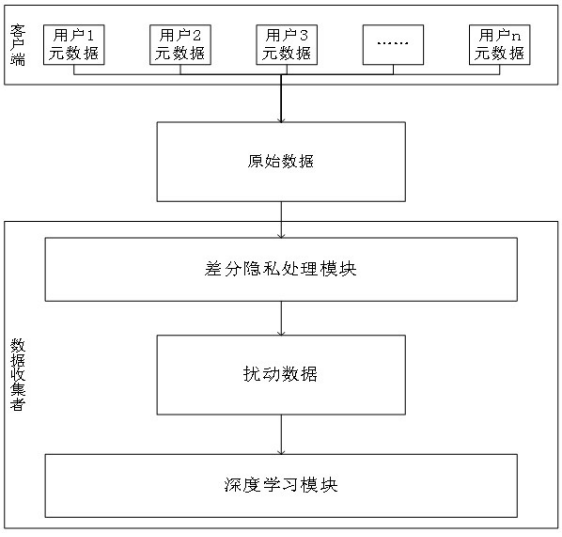
差分隐私指的是存在两个之间至多相差一条记录的数据集 D 和 D’以及一个隐私算法 A，Range(A) 为 A 的取值范围，若算法 A 在数据集 D 和 D’上任意输出结果 O（O ∈ Range(A)）满足不等 $Pr[A(D)=O] \leqslant e^{\varepsilon} \times Pr[A(D')=O]$ 时，A 满足 ε – 差分隐私。差分隐私最主要的方法是在数据集中添加噪声来实现的，常用的噪音机制包括 Laplace 机制 [21] 和指数机制，Laplace 机制适用于连续型数据集，而指数机制适用于离散型数据集。

3.2 深度学习中的差分隐私保护技术

差分隐私 (differential privacy) 是一种基于差分隐私机制的隐私保护技术，这个概念是 Dwork[10] 在 2006 年提出来的，通过在原始数据中引入噪声，达到使至多相差 1 个数据的 2 个数据集查询结果概率不可分的目的。

差分隐私保护分为集中式学习和分布式学习两大类，Abadi[11] 于 2016 年提出的基于差分隐私的深度学习算法基于集中式学习，该方法在梯度下降的过程中利用梯度增大扰动方法来报数敏感数据，并详细分析了差异化隐私的框架下的隐私成本。实验已经证实了可以在适度的隐私预算下，以可管理的软件复杂性、训练效率和模型质量成本，训练具有非凸目标的深度神经网络。2017 年，Papemot[12] 等提出用半监督知识迁移方法来解决深度学习中训练数据隐私泄露问题，以黑盒的方式将多个模型与不相交的数据集（例如来自不同用户子集的记录）相结合。由于改进了隐私分析和半监督学习，作者在 mnist 和 svhn 上实现了最先进的隐私 / 效用权衡。Ji Wang[16] 于 2018 年提出了一种对本地数据进行扰动变换的机制，该机制基于差分隐私计算方法，同时使用噪声训练方法增加云端深度神经网络对移动设备提交的扰动数据的鲁棒性，该机制解决了将数据从移动设备传输到云中心时的隐私泄露问题。

图 2 差分隐私保护模型



在联合分布式学习环境中，数据所有者分布式训练具有相同目标的神经网络模型，根据自己的数据集独立训练，但共享训练结果。Shokri 和 Shmatikov[13] 在 2015 年提出了一个共同分布式

深度学习方案来保护隐私，首次将隐私保护的概念引入深度学习。同时利用现代深度学习中使用的优化算法，即基于随机梯度下降的优化算法，可以实现异步并行和执行，在引入噪声后，每个参与者将一小部分局部梯度参数上传到中心参数服务器。每次更新本地参数时，都会从服务器下载最新的渐变参数进行更新，这样就允许参与者在自己的数据集上独立训练，并在训练期间有选择地共享模型关键参数的小子集。参与者在保留各自数据的隐私的同时，仍然受益于其他参与者的模型，从而提高他们的学习准确性，而不仅仅是通过自己的输入实现的。作者同时证明了在基准数据集上的隐私保护深度学习的准确性。在此基础上，Mohassel[17] 采用随机梯度下降法，提出了一种新的高效的线性回归、逻辑回归和神经网络训练保密机器学习协议。该协议属于双服务器模型，数据所有者将其私有数据分配给两个非协作服务器，这些服务器使用安全的双方计算（2PC）对联合数据上的各种模型进行训练。

4 前景展望

随着深度学习的兴起，人工智能在各个领域迎来新的一波发展热潮，然而在人工智能迅速发展的同时，其安全与隐私问题也引起了人们的关注，人工智能的安全和隐私的威胁已经阻碍了人工智能的发展。保护用户隐私成为人工智能发展的关键，当前基于深度学习的隐私保护的研究仍处于起步阶段，还有许多亟待解决的问题[14]，我们可以从以下几个方面进行重点研究从而找到有解决人工智能中隐私泄露的有效方法。

建立完善的评估机制与法律手段。建立一套统一的隐私泄露安全评估标准以及衡量标准，完善相关法律，从源头上制止企业和组织非法泄露用户信息信息。

（2）高效的加密算法。加密技术是最直接有效的隐私保护手段，但目前同态加密技术运算开销过大，结合本身就消耗大量计算资源的深度学习算法，将大大降低算法性能。因此，研究高效的加密方法保护用户隐私是一个重要研究问题。

参考文献

[1] Hinton, G. E. and R. R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks." Science 313: 504–507.

[2] Brakerski Z, Vaikuntanathan V. Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages[C]// Cryptology Conference. 2011.

[3] Jean-S é bastien Coron, Mandal A , Naccache D , et al. Fully Homomorphic Encryption over the Integers with Shorter Public Keys[J]. 2011.

[4] Yagisawa M . Fully Homomorphic Encryption without bootstrapping[J]. Acm Transactions on Computation Theory, 2015, 6(3):1-36.

[5] DOWLIN N，RAN G B，LAINE K，et a1 . CryptoNets：applying neural networks to encrypted data with high throughput and accu, mcy[C]//Radio and Wireless Symposium. 2016：76・78

[6] HESAMIFARD E，TAKABI H，GHASEMI M，ct a1. Privacy-preserving machine learning in cloud[C]//The 2017 on Cloud Computing Security Workshop. 2017：39-43.

[7] BARYALAI M，JANG-JACcARD J，LIU D. Towards privacy-preserving classification in neural networks[c]//IEEE Privacy,Security and Trust. 2017：392—399.

[8] Xie P, Bilenko M, Finley T, et al. Crypto-Nets: Neural Networks over Encrypted Data[J]. Computer Science, 2014.

[9] Zhang Q , Yang L T , Chen Z . Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning[J]. IEEE Transactions on Computers, 2016, 65(5):1351-1362.

[10] Dwork C , Mcsherry F , Nissim K , et al. Calibrating Noise to Sensitivity in Private Data Analysis[C]// Theory of Cryptography Conference. Springer, Berlin, Heidelberg, 2006.

[11] Abadi, Mart í n, Chu A , Goodfellow I , et al. Deep Learning with Differential Privacy[J]. 2016.

[12] Papernot N , Abadi, Mart í n, Erlingsson, Úlfar, et al. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data[J]. 2016.

[13] Shokri R, Shmatikov V. Privacy-Preserving Deep Learning[C]// Allerton Conference on Communication. 2015.

[14] 宋蕾，马春光，段广晗．机器学习安全及隐私保护研究进展 .[J]. 网络与信息安全学报，2018, 4(08):5-15.

[15] Plantard T , Susilo W , Zhang Z . Fully Homomorphic Encryption Using Hidden Ideal Lattice[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(12):2127-2137.

[16] Wang J , Zhang J , Bao W , et al. Not Just Privacy: Improving Performance of Private Deep Learning in Mobile Cloud[J]. 2018.

[17] Mohassel P , Zhang Y . [IEEE 2017 IEEE Symposium on Security and Privacy (SP) - San Jose, CA, USA (2017.5.22-2017.5.26)] 2017 IEEE Symposium on Security and Privacy (SP) - SecureML: A System for Scalable Privacy-Preserving Machine Learning[C]// Security & Privacy. IEEE, 2017:19-38.

[18] Orlandi C , Piva A , Barni M . Oblivious Neural Network Computing via Homomorphic Encryption[J]. EURASIP Journal on Information Security, 2008, 2007(1):1-11.

[19] Barni M , Failla P , Lazzeretti R , et al. Privacy-preserving ECG classification with branching programs and neural networks[J]. IEEE Transactions on Information Forensics & Security, 2011, 6(2):452-468.

[20] 史晓倩．基于全同态加密的密文搜索方案 [D].

[21] 周大力．基于 Laplace 机制的差分隐私回归分析相关优化研究 [D].

[22] 同态加密技术及其应用研究 [D]. 安徽大学，2013.

[23] 林如磊．全同态加密技术及其应用 [D]. 南京

航空航天大学 , 2012.
[24] 佚名 . 基于同态公钥加密系统的数字水印算法研究 [D]. 暨南大学 , 2018.

作者简介：

唐鹏，上海交通大学网络空间安全学院，博士研究生。主要研究方向为深度学习与网络安全、隐私保护及隐私计算。

黄征，上海交通大学网络空间安全学院，副教授。主要研究方向为机器学习、人工智能和隐私保护计算。

邱卫东，上海交通大学网络空间安全学院，教授 / 副院长。长期从事计算机取证、密码分析、人工智能安全、大数据隐私保护等方向的研究工作。

A Survey of Privacy Protection Technologies in Deep LearningPasswords in Smart Cities

Tang Peng[1],Huang Zheng[1], Qiu Wei-dong [1]
(1. School of Cyber Science and Engineering, Shanghai Jiaotong University, Shanghai, 200240
2. Westone Cryptologic Research Center, Beijing, 100070)

[Abstract] Machine learning and deep learning are now widely used in various fields, including medical, financial, and network security. The primary task of deep learning lies in data collection. However, in the process of data collection, the risk of privacy leakage may occur, and privacy leakage will cause users to no longer trust artificial intelligence, which will be detrimental to the development of artificial intelligence. This paper summarizes the current privacy protection methods and research status in deep learning, including privacy protection technology based on homomorphic encryption, differential privacy protection technology and so on.

[Keywords] Privacy Protection; Deep Learning; Homomorphic Encryption; Differential Privacy Protection