面向深度学习的高效安全推理研究综述

胡鹏[1,2]，孙磊[1,3]，胡翠云[1,3]，郭松[1,3]，王晶雯[1]，王志鸿[1]，姚敬怡[1]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 69016部队，新疆 乌鲁木齐 830001；
3. 河南省信息安全重点实验室，河南 郑州 450001）

摘　要：基于同态加密、安全多方计算等隐私保护密码技术实现深度学习模型安全推理的同时，也引入了巨大的计算和通信开销。针对如何加速安全推理，对现有研究成果进行了总结与梳理。首先，对实现安全推理的2个关键环节——安全协议和推理模型的技术路线和优化方法进行了系统性对比分析。针对安全协议，区分底层的线性和非线性运算，对不同密码原语方案的性能和效率进行对比分析；针对推理模型，就如何平衡安全推理的性能和效率，讨论分析了现有的主要优化方法。其次，增加了对安全推理方案构建成本和主流隐私保护框架的讨论分析，从实际应用角度出发进一步充实了高效安全推理研究的关注范畴。最后，通过分析安全推理面临的问题挑战，面向实际应用需求提出未来深度学习安全推理的探索方向。

关键词：深度学习；隐私保护；安全推理；同态加密；安全多方计算
中图分类号：TP309.2
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025086

Survey of research on efficient secure inference for deep learning

HU Peng[1,2], SUN Lei[1,3], HU Cuiyun[1,3], GUO Song[1,3], WANG Jingwen[1], WANG Zhihong[1], YAO Jingyi[1]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2. 69016 Troops of PLA, Urumqi 830001, China
3. Henan Key Laboratory of Information Security, Zhengzhou 450001, China

Abstract: Secure inference in deep learning with homomorphic encryption and secure multi-party computation has brought substantial computational and communication overhead. To address this challenge, a systematic review of existing research on accelerating secure inference was provided. First, a comprehensive comparison was conducted for the technical approaches and optimization methods of two critical components: security protocols and inference models. For secure protocols, the performance and efficiency of different cryptographic primitive solutions were analyzed by distinguishing between linear and nonlinear operations. For inference models, major optimization strategies were evaluated to balance accuracy and computational efficiency. Furthermore, the costs associated with existing schemes were analyzed, and the core framework of secure inference was outlined, expanding the practical considerations for efficient secure inference. Finally, by discussing the challenges faced by secure inference, potential future research directions aligned with practical application needs are proposed.

Keywords: deep learning, privacy-preserving, secure inference, homomorphic encryption, secure multi-party computation

0　引言

深度学习作为一种重要的人工智能技术已经被广泛用于金融、医疗乃至军事等各个领域[1]，其功能实现主要基于大量的数据训练出的强大模型，因

收稿日期：2025–02–06；**修回日期：**2025–04–25
通信作者：孙磊，sl20210221@163.com
基金项目：国家自然科学基金资助项目（No.62176265）
Foundation Item：The National Natural Science Foundation of China (No.62176265)

此数据量的爆发式增长推动了深度学习的高速发展，而大模型的出现更是成为深度学习发展的里程碑[2]。

以大模型为代表的深度学习模型快速发展的背后是数据和算力的支持，而缺乏专业技术与资源储备的企业和个人用户难以承受模型训练对海量数据和算力的需求，使得以云服务商搭建模型供用户使用的深度学习即服务模式（DLaaS, deep learning as a service）成为当前深度学习最为广泛的应用场景[3]。

但随着DLaaS应用范围不断拓展，其在数据隐私安全方面的潜在问题也愈发凸显[4]。数据与模型属于不同的参与方，使用深度学习服务时数据必须交由其他参与方执行运算，这就导致用户敏感数据完全暴露给模型，存在严重的隐私泄露风险。

将同态加密（HE, homomorphic encryption）[5]、安全多方计算（MPC, secure multiparty computation）[6]等基于密码学的隐私保护技术与深度学习结合，构建满足隐私保护的安全推理方案，解决了数据使用和数据隐私间的矛盾，实现了深度学习模型的安全推理，使得隐私保护的DLaaS在学术和工业界都得到了广泛关注和研究。但采用密码学方法引入的巨大计算和通信开销大大降低了模型的推理速度，严重影响了安全推理方案的实际应用。

目前，已有大量研究者针对如何加速安全推理展开了研究。从实现角度看，将密码技术与推理模型结合才能构建完整的安全推理解决方案，但一些研究者主要针对基于密码技术的安全协议或推理模型中的某些痛点问题进行了研究与优化，本文将这些研究成果也纳入安全推理解决方案的范畴，统称为安全推理方案。对现有方案进行系统归纳和总结对于安全推理在深度学习中的研究和应用具有重要意义。

文献[7-8]分别从基于密码技术实现深度学习隐私保护、针对深度学习的安全攻击与防御技术等角度进行分类和对比分析，为深度学习隐私保护面临的挑战和发展方向提供了综合视角，但这些文献未专门针对深度学习安全推理研究进行总结梳理。文献[9]首次回顾了隐私保护的机器学习即服务相关研究，通过提出的多角度分类方法，对隐私保护深度学习中的威胁模型、隐私保护解决方案以及面临的挑战等问题进行了综述。文献[10]针对高效的隐私保护机器学习即服务相关研究进行了综述，就安全计算效率低下的主要原因进行了讨论，并分析了2种优化该问题的典型策略。文献[11]针对机器学习安全推理现有研究，从安全假设角度对现有研究进行分类和总结，给出现有安全推理方案在计算效率、安全保护能力、可扩展性以及实际应用场景适应性方面的局限性。文献[12]针对深度学习即服务模式下的隐私保护研究成果进行了总结，该文献根据设计安全推理方案时遵循的思路，将现有工作区分线性层和非线性层进行分类梳理，在同一类别中按照时间轴总结了不同密码协议实现安全推理的发展脉络，同时也归纳了各安全推理方案的技术路线，为隐私保护深度学习的研究提供了全面的分析。

现有综述已对深度学习隐私保护研究成果进行多角度的总结，对可采用的技术有了清晰的概括。但鲜有从优化效率这一目标出发，站在全局角度对安全推理不同维度的技术路线和优化方法进行对比综述。效率问题一直是安全推理方向研究的关键问题，但针对不同底层运算、选择不同的密码原语以及采用不同模型优化技术，对安全推理的效率提升效果存在显著差异，且在设计安全推理优化方案时需要考虑对模型性能这一重要指标的影响。

为尽可能囊括安全推理效率优化的相关技术路线和方法，本文在现有综述分类的基础上，面向深度学习安全推理具体过程，首先从目标导向出发，对安全推理方案的关键要素进行模块化分层分类，从整体视角理清了构建安全推理方案的各个环节。然后分别从实现安全推理方案的2个核心模块：安全协议和推理模型，以及影响安全推理方案构建的2个重要因素：方案构建成本和隐私保护框架出发，对安全推理优化的现有研究成果进行了系统总结和对比分析，本文的层次分类框架如图1所示。

具体来说，本文将安全推理分为运算层、实现层和目标层3个层次。运算层由线性运算、非线性运算2种基本运算模块组成，是安全推理的底层基础。实现层由安全协议和推理模型2个模块组成，分别代表了安全和推理这2个核心功能的实现，协同作用共同构成了安全推理方案。安全协议主要基于同态加密和安全多方计算等密码原语实现，从现有安全推理方案来看，综合利用不同密码原语的优势设计混合密码协议能够更好地满足实际应用需

求，因此被广泛采用。推理模型在深度学习中已经得到了广泛研究，但在安全推理场景下，模型的推理速度大大降低，模型准确率等其他指标也可能会受到影响。因此，针对深度学习安全推理下的模型研究主要是对模型进行优化，以更好地满足安全推理的目标需求。从现有研究成果看，安全推理主要采用的模型优化技术有近似、剪枝、量化和蒸馏。

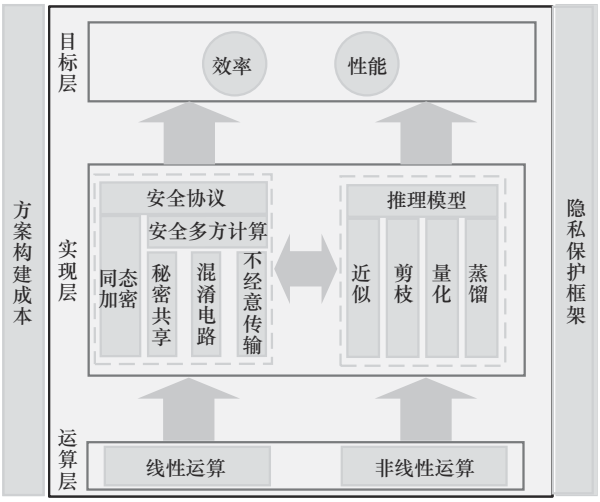


图1　高效安全推理相关研究层次分类框架

目标层则由安全推理主要关注的推理效率、模型性能等目标项组成，代表了安全推理方案的设计目标，也是对推理方案的评价指标。此外，本文加入了2个与安全推理方案构建密切相关的重要因素的总结分析，一是方案构建成本。在安全推理方案构建时引入的模型训练等成本也是实际应用需要考虑的重要因素，而这在现有研究中却往往被忽略。本文对不同方案的构建成本进行了分析讨论，力求更加全面客观地开展方案间的对比研究。二是隐私保护框架，与深度学习中的Pytorch、Tensflow等框架类似，一些成熟的隐私保护框架为安全推理方案的实现和应用提供了基础平台，也是方案易用性和扩展性的重要体现，因此本文对主流的隐私保护框架进行了对比分析，从实现环境的角度为深度学习安全推理的研究设计提供了更多参考。

本文将安全推理领域现有的各种技术路线纳入层次分类框架中，能够以更加全面、实用的视角对实现高效安全推理相关研究进行梳理和总结。与其他安全推理相关的文献综述相比，本文主要贡献包括以下几个方面。

1) 站在更高视角从目标层、实现层和运算层3个层次对现有研究成果进行了分类梳理，能够继承现有基于密码技术、运算类型等主要安全推理方案分类方式，从全局把握深度学习高效安全推理的研究现状。

2) 针对安全协议和推理模型这2个安全推理的关键环节，区分底层运算进行归纳总结，对不同密码原语方案和模型优化方法在效率、性能这2个关键指标上进行了详细对比分析。

3) 本文增加了对安全推理方案构建成本和可依托隐私保护框架的总结和讨论，这对安全推理方案的实现与应用非常重要，据笔者所知，这是首次在安全推理综述中关注成本和易用性问题。

4) 在总结现有研究成果的基础上，本文对安全推理研究方向面临的挑战进行了分析，并对未来可能的发展方向给出了合理预测，为推进安全推理领域的研究发展提供了思路和建议。

1　深度学习安全推理概述

1.1　深度学习模型概述

深度学习的强大能力主要依托深度神经网络模型实现。从内部实现机理来看，深度学习系统本质上由线性运算和非线性运算交替组合构建而成，这种层级化结构赋予其处理各类复杂任务的能力。当前典型的应用领域包括图像识别和自然语言处理，其中卷积神经网络（CNN, convolutional neural network）和Transformer分别作为2类任务中的代表性模型架构，在各自领域取得了显著成效，并被广泛应用于跨领域任务。

1) CNN

CNN是深度学习中应用最广泛的模型之一，特别是在图像处理、目标检测、视频分析等任务中表现优异[13]。CNN的核心思想是通过卷积算子对输入数据进行局部处理，并通过池化操作逐步提取高级特征，最后通过全连接网络对图像的特征表达进行分类。典型的CNN结构如图2所示，下面分别就线性和非线性算子对CNN的底层运算进行介绍。

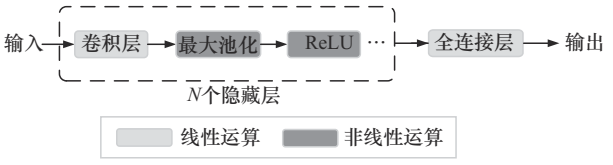


图2　典型的CNN结构

① 线性算子

卷积操作：CNN 中最核心的线性算子是卷积操作，卷积层通过卷积核在输入数据上滑动，进行局部的加权求和，从而提取局部特征。卷积操作可表示为

$$\boldsymbol{Y} = \boldsymbol{X} * \boldsymbol{W} + b \tag{1}$$

其中，$\boldsymbol{Y}$表示输出特征图，$\boldsymbol{X}$为输入特征图，$\boldsymbol{W}$为卷积核，b为偏置项，$*$代表卷积。CNN 中包含的主要线性操作就是卷积和全连接，而所有的卷积操作都能转化为全连接的形式。除此之外，还有一些 CNN 中常见的线性操作或可以直接转化为线性的操作，如平均池化、批归一化（BatchNorm）、dropout 等。

② 非线性算子

神经网络中非线性特性的实现主要通过激活函数来完成[14]，激活函数本身的定义是一种将输入映射到输出的函数，但是在神经网络中主要通过其提供的非线性变换实现对任意复杂函数的拟合，从而学习更加复杂的模式，因此通常就将非线性函数表述为激活函数。

CNN 中隐藏层最常用的非线性激活函数是 ReLU 函数。

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

输出层的非线性算子主要用于分类等任务，神经网络中常用的输出层激活函数是 Sigmoid 函数和 Softmax 函数，分别用于二分类和多分类任务中。

$$\mathrm{Sigmoid}(x) = \frac{1}{1 - \mathrm{e}^{-x}} \tag{3}$$

除此之外，CNN 中常用的非线性算子还包括 Tanh 函数、最大池化等。通过不同线性和非线性算子的良好组合，大量经典的 CNN 模型结构被提出，如 AlexNet[15]、VGG[16]、ResNet[17]等。

2) Transformer

Transformer[18]是目前大语言模型的基础架构，因为引入了注意力机制，大大提高了对长距离依赖的处理能力，所以被广泛应用于自然语言处理等任务并取得了出色的效果。典型的 Transformer 模型编码器结构如图 3 所示。可以看出，与 CNN 相比，Transformer 包含的算子类型更多、结构更加复杂，下面对 Transformer 的主要算子进行介绍。

① 线性算子

Transformer 的线性运算中除了常见的全连接线性变换外，最主要的线性算子是矩阵乘法，用于实现自注意力机制。自注意力运算表达式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$均为来源于输入序列$\boldsymbol{X}$的矩阵，d_k表示键向量$\boldsymbol{K}$的维度，因此自注意力运算的计算复杂度与输入序列长度的平方成正比，当输入序列较长时，矩阵乘法开销将显著增加。

② 非线性算子

从模型内部组成结构来看，Transformer 中的非线性算子主要包括 Softmax 函数、GeLU 函数以及层归一化。

$$\mathrm{Softmax}(x_i) = \frac{\mathrm{e}^{x_i}}{\sum_{i=1}^{n} \mathrm{e}^{x_i}} \tag{5}$$

Transformer 中还包含有一个 2 层的前馈网络，该模块中常使用的激活函数为 GeLU，有时为简化计算，也采用 ReLU 函数直接进行替代。

为提升训练速度，Transformer 模型中利用层归一化（LayerNorm）对层输出进行规范化处理，与 CNN 中常采用的 BatchNorm 对每个批次的输入数据进行规范化不同，LayerNorm 是对层中的所有样本输出进行规范化，计算式为

$$\mathrm{LayerNorm}(x) = \frac{x - \mathrm{E}(x)}{\sqrt{\mathrm{Var}(x) + \varepsilon}}\gamma + \beta \tag{6}$$

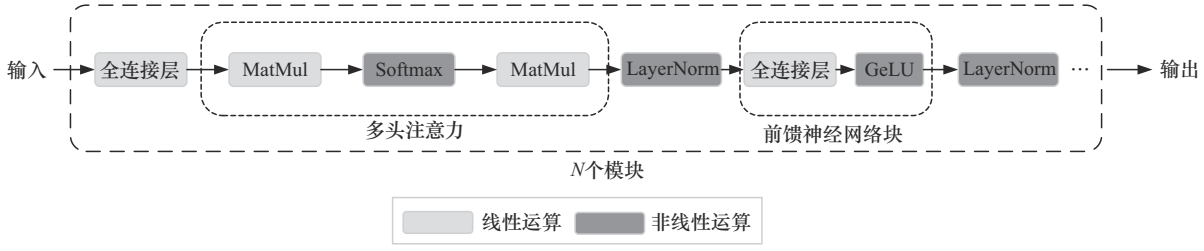


图 3　典型的 Transformer 模型编码器结构

其中，$\mathrm{E}(x)$、$\mathrm{Var}(x)$分别表示输入x的均值和方差，ε是很小的常数，以避免分母为0，γ和β为可学习的参数，用于对归一化输出进行缩放。LayerNorm的非线性主要来源于方差计算，由于各层输出随着输入数据的不同而随时变化，无法提前计算方差。而BatchNorm是对数据样本进行统计，对不同的数据集可以采用整体数据集的统计方差近似替代，因此BatchNorm可以转化为线性运算。

1.2 同态加密技术

同态加密是一种允许用户在密文上直接进行计算的加密技术，其计算结果解密后与对明文直接计算的结果一致。这一概念最早由Rivest等[19]在20世纪70年代末提出。然而，直到2009年，Gentry[20]才提出了第一个全同态加密（FHE, fully homomorphic encryption）方案，使得不受信任的第三方能够在加密数据上执行任意计算，而无须了解数据内容。

FHE方案的关键特性在于它允许在加密数据上执行同态加法和乘法运算。相比之下，早期的HE方案仅支持这2种操作中的一种，因此被认为是部分同态的。例如，RSA（Rivest Shamir Adleman）公钥密码系统[21]支持在加密数据上进行任意数量的模乘法，而Paillier密码系统[22]则支持任意数量的模加法。

现有FHE方案中，通常采用多项式环。实践中，通过选择二次幂循环多项式作为不可约多项式，以便利用快速傅里叶变换（FFT, fast Fourier transform）[23]进行高效的算术运算。然而，当前FHE方案的密文本质上包含一定量的噪声，这种噪声在同态操作过程中会增加。如果噪声过大，即使使用合法的解密密钥，也无法正确解密。为了实现无限次操作，一种方法是以同态方式执行重新加密过程，即自举。然而，自举在实践中的计算成本很高。

随着研究的发展，FHE方案已经发展到第四代[24-26]，整体性能变得更加实用，并且方案支持浮点数运算等更多功能。此外，为避免自举操作，一些研究者采用层次型全同态加密（LFHE, leveled fully homomorphic encryption）方案[27]，该方案主要通过预期的电路深度来调整参数，能够在预定深度下执行任意复杂度的计算而无须自举操作。

1.3 安全多方计算技术

安全多方计算技术源于姚期智提出的百万富翁问题，旨在解决一组互不信任的参与方如何在保持隐私的同时进行协同计算的问题[28]。MPC通过一系列基础密码工具，如不经意传输（OT, oblivious transfer）、混淆电路（GC, garbled circuits）、秘密共享（SS, secret sharing），构建安全协议来实现安全运算。

1) OT协议

OT协议最早由Rabin[29]提出，是一个两方计算协议，涉及发送方和接收方。接收方获得部分信息，而发送方无法得知接收方获得了哪些消息。在恶意敌手模型下，MPC所需的OT次数可能达到数百万次。例如，在计算隐私集合求交电路时，可能需要执行2^{30}次OT计算，这使得OT成为两方计算的瓶颈[30]。为了提高效率，研究者们致力于减少OT调用次数或开发OT扩展技术，以少量的基本OT协议实现大量OT实例[31]。

2) GC协议

GC协议由Yao[32]提出，旨在解决安全两方计算问题。Lindell等[33]提供了安全性证明，Bellare等[34]给出了GC的标准化定义。GC协议以常数轮交互为特点，适用于任意大小的电路，但通信量较高。Micali等[35]提出GMW（Goldreich Micali Wigderson）协议，是另一种基于GC协议的通用且高效的安全多方计算协议。与姚氏GC协议类似，GMW协议需要将函数描述为布尔电路，不同之处在于，GMW协议在评估电路的每一层布尔门时都需要一轮交互。与姚氏GC协议相比，GMW协议需要更少的数据通信。如果仅考虑在线成本，GMW协议中的大部分计算和通信可以转移到预处理阶段，使得在线阶段非常高效。近年来，GC协议的研究集中在降低通信开销和提升计算效率上，以实现更高效、更实用的解决方案。

3) SS协议

SS协议由Shamir[36]和Blakley基于拉格朗日插值和线性几何投影理论独立提出，典型的SS协议包括Shamir秘密共享协议、Blakley秘密共享协议和中国剩余定理等。Shamir[36]的门限秘密共享协议将秘密信息拆分为n个份额，需至少t个份额才能恢复秘密。从研究现状来看，针对SS协议当前的研究重点在于优化份额大小和简化整数规划问题。

1.4　安全推理与威胁模型

1) 安全推理的定义

现有研究成果中，对安全推理的概念并没有统一的定义。从已有文献的表述来看，尽管采用的术语不同，但它们的核心目标是一致的，即通过技术手段保护神经网络推理过程中数据和模型的机密性与隐私性。需要注意的是，现有安全推理中的隐私保护技术并非完全等同。差分隐私[12]等基于混淆的技术虽然能够在一定程度上保护隐私信息，但它们无法提供完全的机密性防护。差分隐私方法通常基于噪声来模糊数据，从而限制攻击者通过推理还原数据的能力，但在某些情况下，攻击者仍然可以通过多轮推理进行信息推断。

本文聚焦于基于密码学的隐私保护技术，如同态加密、安全多方计算等，这些技术能够实现对神经网络推理中数据和模型的完全机密性保护。因此，本文采用安全推理这一术语，以明确标识基于密码学方法实现模型推理中的高安全性隐私保护。与之相对，传统的明文推理则指在没有任何隐私保护机制的情况下，直接使用未加密的模型对数据进行推理的过程。

2) 安全推理的目标

DLaaS 模式下，客户端首先向服务器发送想要处理的数据或指令，服务器根据收到的数据执行推理过程，推理结束后将结果发送回客户端，推理过程如图 4 所示。从推理过程来看，安全推理希望保护的隐私信息，即安全推理的目标应当包括：①客户端发送的数据或指令不能被泄露；②推理结果只能被客户端获知；③属于服务器核心资产的模型参数不能被泄露。

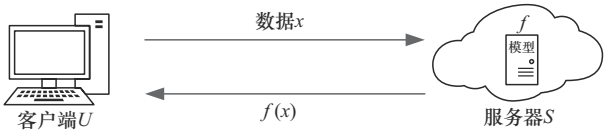


图 4　DLaaS 推理过程

3) 威胁模型

神经网络安全推理需要在一定安全假设下考虑安全问题，即威胁模型，按照敌手力量主要可分为半诚实（SH, semi-honest）模型和恶意（MA, malicious）模型。2 种安全威胁模型代表了不同的安全假设，适合不同的应用场景。半诚实模型安全性更弱，但在该威胁模型下，能够建立更加高效的推理协议，目前半诚实模型在神经网络安全推理方案中使用更加广泛。

2　安全推理协议优化方法

目前很多研究者从不同角度针对安全推理协议进行了设计和优化，本文从同态加密和安全多方计算 2 种主要的技术路线出发，对现有方案中的线性和非线性安全运算协议构建方法进行了总结分析，理清了安全推理协议的整体发展脉络。

2.1　线性安全运算协议

从同态加密和安全多方计算技术自身特点来看，都能够支持线性运算。从现有针对线性运算的安全协议相关研究来看，主要采用的密码原语以同态加密和秘密共享为主，只有少部分文献结合了混淆电路或不经意传输来实现线性层的安全推理。下面，分别从同态加密和安全多方计算 2 个技术路线出发，对线性安全运算协议进行总结综述，其中对于采用同态加密和秘密共享结合的方案，线性运算协议主要基于秘密共享完成，而同态加密仅用于辅助运算，因此将该类方法归于安全多方计算中。

1) 基于同态加密的线性安全运算协议

同态加密在加法和乘法运算上的同态性质使得其非常适合进行线性运算，但从实际效果来看，基于同态加密的安全推理方案时延很高，而高时延的主要原因是巨大的计算开销。分析计算开销的主要来源，主要包括：采用的同态加密算法、加解密过程和密文运算过程。具体来说，不同的同态加密算法在处理不同数据时的效率不同，例如，BFV（Brakerski Fan Vercauteren）同态加密算法更适合整数计算，而 CKKS（Cheon Kim Kim Song）同态加密算法可以处理浮点数运算，在数据量较大时，采用 BFV 算法比 CKKS 算法的效率更高。同态加密算法对数据执行加解密过程需要较多的计算时间，且加密后的密文通常比明文大得多，这样会增加数据的存储开销，而密文在传输处理中，这些存储开销也将带来额外的推理时延。密文运算中，常见的线性计算有加法、乘法和旋转位移操作，其中开销最大的是旋转位移操作。此外，密文运算过程中需要针对噪声增长进行控制，或为支持更深层次的运算采用自举技术，这些操作都给基于同态加密的安全运算带来了巨大的计算开销。

表1从模型、同态加密方案、威胁模型、优化对象、关键技术以及性能和效率等角度对基于同态加密实现线性安全运算的主要方案进行了对比分析。其中目标项仅考虑该文献中线性安全运算优化对性能和效率的影响程度，B表示同类技术路线的最早方法，以该方法为基准，↑、↑↑、↑↑↑分别表示与基准方法相比，提升效果一般、较大、显著；↓、↓↓、↓↓↓则分别表示与基准方法相比，降低程度一般、较大、显著。从现有研究成果来看，缓解该类方法开销问题的主要技术路线包括：针对运算过程采用降低同态加密运算中的昂贵操作、选择设计合适的同态加密参数或方案等。

文献[37]首次将全同态加密技术与神经网络结合，同时通过单指令多数据流（SIMD，single instruction multiple data）批处理技术降低单次运算开销。针对开销大的问题，研究者针对卷积、全连接等主要的线性运算协议进行了研究[49]。文献[38]为矩阵−向量乘法和卷积运算设计了优化的同态运算内核，能够快速实现加密运算协议。文献[41]尝试在频域卷积神经网络内优化卷积运算，通过傅里叶变换简化线性层并减少旋转操作，从而提高运算效率。

根据开销的主要来源，一些研究者针对如何降低运算中的昂贵操作进行了广泛的研究。文献[50]将线性运算分解为一系列同态加法乘法和排列操作，采用先加后排列的方法。文献[51]则设计了高效的乘积累加算法。文献[42]通过观察线性运算过程，对多项式运算进行了精心编码设计，由此完全避免了线性运算中的旋转操作。但特殊编码后的多项式无法直接应用SIMD批处理技术。

为综合各运算优化技术的优势，一些方案将减少昂贵操作与多种技术相结合提高效率。文献[48]通过合理安排运算顺序减少数论变换（NTT, number-theoretic transform）操作以提高运算速度。文献[46]采用降低或消除旋转操作对数据进行编码，而后采用优化的打包技术提高推理效率。

综合分析，同态加密由于其与线性运算的天然契合，目前依旧是实现线性层安全协议的一个主流方向。采用同态加密设计线性运算协议的主要优势在于：数据全过程处于加密状态，安全性能够得到保证，且运算中无须进行交互，节省通信开销。主要可采用针对线性运算本身就同态加密方式下的数据编码等进行优化，例如降低昂贵运算优化线性运算算子、设计打包编码方式提高并行效率等；也可从同态密码算法入手选择设计合适的密码方案。目前，采用优化编码减少昂贵操作的同时结合具体线性运算优化策略以提高运算效率，其推理效果相比于仅完全消除旋转等昂贵操作的方案更优，因此结合各技术路线的综合方法将是未来的主要研究方向。

表1　　基于同态加密的线性安全运算协议的主要方案

方案	推理模型	同态加密方案	威胁模型	优化方法	关键技术	目标	
						性能	效率
文献[37]	CNN	YASHE	SH	—	打包	—	B ↑（摊销）
文献[38]	CNN	AHE	SH	乘法、卷积	设计同态线性运算内核	—	↑↑
文献[39]	CNN	Paillier	SH	同态加密方案	选择轻量级同态算法	—	↑
文献[40]	CNN	BFV	SH	加密方式	加密整层	—	↑↑
文献[41]	CNN	FV	SH	线性运算	利用FFT简化线性运算	↓	↑↑
文献[42]	CNN	BFV	SH	旋转操作	精心设计的数据编码方式	—	↑↑↑
文献[43]	CNN	BFV-CKKS	SH	旋转操作、卷积	一次乘法实现批量卷积	—	↑↑↑
文献[44]	Transformer	CKKS	SH	—	—	—	B
文献[45]	Transformer	BFV	SH	矩阵乘法	紧凑打包技术	—	↑
文献[46]	Transformer	BFV	SH	矩阵乘法、旋转操作	优化的打包技术，大小步策略	—	↑↑
文献[47]	CNN	MKHE	SH	同态算法	简化重线性化	—	—
文献[48]	CNN	BFV	SH	旋转操作、卷积	乘积累加设计和部分和累积策略	—	↑↑↑

2) 基于安全多方计算的线性安全运算协议

安全多方计算各密码原语中，秘密共享采用的线性分片方式使得其非常适合线性运算，但运算中的频繁交互导致通信开销较大，进而造成了高时延。分析其通信开销，主要来源于秘密份额交换，而这与参与方的数量、秘密共享构造方案及实现的运算息息相关，例如加法运算可以由各自参与方在本地完成而无须交互，但乘法操作无法单独完成，需要实施份额交换以完成交叉项的运算，因此乘法运算的实现带来了大量的通信开销。混淆电路和不经意传输虽然也能实现安全的线性运算，但与秘密共享相比，其在大多数线性运算的计算和通信开销上并没有优势，仅对二值神经网络等特殊运算开销更小，因此单独采用混淆电路和不经意传输实现线性运算的研究较少。

针对以上安全多方计算各密码原语在线性运算中的主要特点和开销来源，现有基于安全多方计算实现线性运算的方案大多采用秘密共享为主，研究设计如何降低通信开销。而为提高协议中某些操作的效率，会将其他密码原语与秘密共享结合使用。表 2 从模型、密码原语、威胁模型、参与方数、优化方法以及性能和效率等角度对基于安全多方计算实现线性安全运算的代表性方案进行了对比分析。

文献[53]是最早通过秘密共享构建的安全推理协议，针对乘法运算引入了点积三元组生成协议，并将部分操作置于离线阶段完成，与文献[37]和文献[52]方案相比，显著降低了在线预测阶段的时延和通信开销。由于生成大量的乘法三元组会带来额外的开销，一些研究者致力于降低生成秘密共享所需乘法三元组的开销和成本。文献[55]通过优化协议减少通信轮数，并引入一个半诚实第三方（STP, semi-honest trust party）用于预先生成三元组。文献[56]首次在三方场景中基于秘密共享实现安全线性运算协议，此外提出新的近似乘法协议，消除了离线阶段三元组的生成步骤。

秘密共享很容易扩展到三方和多方，因此出现了大量在不同参与方间基于秘密共享的线性协议。典型的方案如文献[60]采用 2-out-of-3 复制秘密共享，相比文献[56]，文献[60]在私有推理上快约 8 倍，在私有训练上快约 6 倍，通信效率高出 16 到 200 倍。

一些方案面向安全性更高或其他特定场景，基于秘密共享设计了针对线性运算推理协议，在该场景下达到了更高效的安全推理。Lehmkuhl 等[59]揭示了半诚实模型在安全性上的缺陷并提出能够防御恶意客户端的线性安全运算协议。文献[61]通过引入“可分离”和“条件可分离”的概念，允许卷积神经网络的不同层在多个服务器上协同计算，从而显著减少了线性层的计算开销。

随着大模型的出现，模型中的线性运算更加复杂，一些研究者将秘密共享的乘法运算推广到矩阵乘法，设计了高效的矩阵乘法方案[63]，例如，文献[62]通过多项式编码和密文打包技术，大大降低了模型中矩阵乘法的运算和通信开销。

表 2　基于安全多方计算技术实现线性安全运算协议的代表性方案

方案	推理模型	密码原语	威胁模型	参与方数/个	优化方法	目标	
						性能	效率
文献[52]	CNN	SS+OT	SH	2	整数运算及乘法截断	↓	B
文献[53]	CNN	SS	SH	2	离线生成三元组	—	↑
文献[54]	CNN	GC	SH	2	数据和模型预处理	—	↑
文献[55]	CNN	SS	SH	3	离线运算卸载	—	↑
文献[56]	CNN	SS	SH	3	第三方生成三元组	—	↑
文献[57]	CNN	SS	SH+MA	3	高效的转换协议	—	↑↑
文献[58]	CNN	SS	SH	2	离线使用 LHE 层次化同态加密（LHE）共享份额	—	↑↑
文献[59]	CNN	SS+OT	MA	2	改进 Beaver 三元组生成	—	↑↑
文献[60]	CNN	SS+OT	SH+MA	3	复制秘密共享	—	↑↑↑
文献[61]	CNN	SS	SH	*N*	多服务器协同卷积运算	—	↑↑↑
文献[62]	Transformer	SS	SH	2	多项式编码和密文打包	—	↑↑

综合分析，基于安全多方计算实现线性安全运算协议时，秘密共享是首选，针对频繁交互通信开销大的问题，可以设计不同方案尽可能将运算置于离线完成，也可根据不同场景需求，针对不同参与方或复杂的线性运算，选择复制秘密共享等新的密码方案，或结合其他密码原语设计更加高效的线性运算协议，这也是该类方法未来研究的主要方向。

2.2 非线性安全运算协议

从密码技术自身特性来看，安全多方计算技术能够实现非线性安全运算，而同态加密并不支持该类运算，但为利用同态加密的优势，一些研究者将非线性函数近似为线性运算，由此构建完全基于同态加密的安全推理方案。由于该类方法与同态加密紧密相关，因此将采用近似方法实现基于同态加密的非线性运算在非线性安全运算协议部分进行介绍。下面，分别对基于同态加密和安全多方计算的非线性安全运算协议进行总结综述。

1) 基于同态加密的非线性安全运算协议

大多数单独基于同态加密实现非线性安全运算协议的方案其本质是近似的，因为同态加密本身无法支持非线性运算。早期一些研究[37]直接采用平方函数替代模型中的ReLU激活函数，达到了实现完全基于同态加密构建安全推理方案的目标。该方法在较小规模的网络中可以达到与ReLU函数近似的效果。但是随着模型规模的不断扩大，直接采用平方函数的近似方法，精度损失过大。因此，后续研究主要针对模型精度和效率的平衡展开研究。表3从模型、优化方法、性能和效率等角度对基于同态加密实现非线性安全运算的代表性方案进行了对比分析。

文献[64]将多项式系数限制为2的次方，首先找到最小最大近似多项式，然后将其系数取整，最终得到满足约束的最优多项式。文献[66]采用三次多项式近似激活函数。文献[67]通过将输入先进行平均分块，最后再进行组合的做法，减少了客户端和服务端进行交互的时间开销。文献[68]进一步提高了多项式次数，采用了一种优化的10次多项式来近似ReLU激活函数，在输入范围内达到了10位的精度。

一些研究者在离散化的神经网络上通过环上友好的同态加密方案设计协议，避免了对非线性激活函数进行近似。文献[65]利用二进制同态算法实现高效的ReLU激活和最大池化，通过对数量化减少了计算开销且利用移位操作替换运算成本高的乘法，显著提高了推理效率。

文献[44]采用了另一种方式：利用客户端辅助模型来计算激活函数，在遇到非线性激活层时，服务器将加密值发送给客户端，客户端解密后计算非线性激活函数再加密返回给服务器。通过该方案执行非线性运算速度更快，但带来更多的通信开销，且对用户不友好，客户端进行明文运算也存在泄露隐私的风险。

综合来看，单纯基于同态加密实现非线性运算协议的方法中，多项式近似方案能够满足完全基于同态加密构建安全推理方案，且低次多项式能大幅提高运算效率。但随着网络深度的增加，精度的损失会逐渐变得不可接受，为此如何实现推理精度与效率间的平衡是设计基于同态加密的非线性安全运算协议需要重点考虑的问题。而不同场景对精度和

表3 基于同态加密实现非线性安全运算协议的代表性方案

方案	推理模型	优化方法	目标	
			性能	效率
文献[37]	CNN	—	↓↓↓	B
文献[43]	CNN	高精度近似	↓	↓↓
文献[44]	Transformer	近似为ReLU，而后发回客户端运算	↓↓	↓↓↓
文献[46]	Transformer	训练得到最优近似多项式	↓	↓
文献[64]	CNN	近似多项式优化	↓↓	↓
文献[65]	CNN	采用二进制友好同态算法	—	↓↓↓
文献[66]	CNN	三次多项式	↓↓	↓
文献[67]	CNN	降低乘法深度	↓↓	↓
文献[68]	CNN	训练包含低次多项式激活函数的模型	↓	↓

效率的需求不同，针对特定任务采用调整精度和效率的平衡方案在该类方法未来的研究中仍值得借鉴。

2) 基于安全多方计算的非线性安全运算协议

安全多方计算技术中，混淆电路能够将运算转换为门电路，因而能够方便实现比较函数，非常适合非线性函数的运算，因此混淆电路在非线性运算中的应用非常广泛。虽然通信轮数与混淆电路深度无关，但一些复杂的非线性运算需要庞大的电路，使得通信开销依然很高。一些研究者采用缩小电路大小[51]、模型参数二值化[69]等方式对基于混淆电路实现非线性函数的方法进行优化。

布尔秘密共享能够实现比较协议，而ReLU等一些非线性函数能转化为由比较函数构成，结合秘密共享能够离线预先计算部分运算的优势，基于秘密共享的非线性运算逐渐成为热门研究方向。但算数秘密共享和布尔秘密共享存在相互转换问题，因此针对该问题的研究也受到了一些学者的重点关注。

典型方案如文献[70]首次使用函数秘密共享（FSS, function secret sharing）设计比较协议，极大减少了通信轮次。文献[71]则首次将该类方案用于Transformer的推理，比现有的针对Transformer的安全推理方案效率提高了11.5~19.4倍。

采用不经意传输实现非线性运算的研究较少，主要在于其适用范围有限，在各非线性运算中，均有更优的密码原语进行替代，因此与线性安全运算类似，不经意传输在非线性运算中也主要用于实现协议中的特定运算[56,60]。

由于GC和SS都存在通信开销大的问题，一些研究者综合利用各类密码原语的优势，设计了基于混合密码原语的非线性函数协议。文献[72]利用SS和OT开发了低交互的安全比较协议，并设计了安全指数和除法协议实现安全归一化层。通过这些优化，显著降低了通信轮次和开销。文献[73]利用不同密码原语设计了底层非线性运算协议，通过使用查找表和混合位宽法降低运算的通信开销。表4从模型、密码原语、优化对象、性能和效率等角度对基于安全多方计算实现非线性安全运算协议的代表性方案进行了对比分析。

综合分析，从安全推理协议研究现状来看，同态加密和安全多方计算都各有优缺点，同态加密非常适合线性运算，加密后的运算过程无须数据方参与，无须构建复杂的安全协议且安全强度高，能够适应更多的场景。但密文计算的巨大开销是同态加密的主要瓶颈，也是该类技术路线研究解决的关键问题。秘密共享能够很好地支持线性运算，且针对通信开销大的问题，能够采取离线预先计算以及综合其他密码原语设计更加高效的协议，一些研究基于良好的协议设计已经将推理时延降低到可实用。但该方法需要数据方参与运算，频繁的中间结果交互可能降低协议的安全强度，且不同密码原语的转换可能带来额外开销，使得协议在实际应用中面临着更多潜在问题。因此，同态加密和安全多方计算目前不存在绝对的优势，未来基于2种密码技术实现安全推理协议依然是2个主要的研究方向。

表4　基于安全多方计算实现非线性安全运算协议的代表性方案

方案	推理模型	密码原语	优化方法	目标	
				性能	效率
文献[60]	CNN	OT+SS	避免除法和指数的Softmax近似	↓	↑↑
文献[69]	CNN	GC	模型参数二值化	—	↑↑
文献[70]	CNN	FSS	ReLU、最大池化和BatchNorm协议	↓	↑↑↑
文献[71]	Transformer	FSS	Softmax、GeLU协议	↓	↑↑↑
文献[72]	CNN	SS+OT	指数和除法协议	—	↑↑
文献[73]	CNN	SS+GC	查表	↓	↑↑
文献[74]	CNN	SS+GC	ReLU、最大池化和除法的协议	—	↑
文献[75]	CNN	FSS	基于FSS的ReLU协议	↓	↑↑↑
文献[76]	Transformer	SS+GC	Softmax和层归一化的安全协议	↓	↑↑
文献[77]	Transformer	SS+GC	Softmax、GeLU和LayerNorm协议	↓	↑↑

3 安全推理模型优化方法

目前安全推理模型的相关研究中，主要针对不同模型的线性运算和非线性运算进行优化。优化方法与模型类型紧密相关，从现有研究来看，安全推理模型的主要研究对象包括CNN和Transformer。从2种模型安全推理过程各算子的开销占比来看，传统CNN中，非线性运算占据整体运算开销的大部分。而Transformer中，非线性算子更加多样复杂，其开销依然占主导地位，但由于注意力机制中矩阵乘法的存在，线性运算开销大大增加。下面，分别对安全推理模型中线性和非线性运算的主要优化方法进行总结分析。

3.1 线性运算优化方法

模型分解为具体算子，从线性算子的开销占比来看，卷积与矩阵乘法运算是开销的主要来源。因此现有针对安全推理模型中的线性运算优化的研究中，主要围绕这2类线性运算设计可行的优化方案，表5对主要的优化方案进行了对比分析。

具体来看，部分研究结合了同态加密中的旋转、打包等技术与模型的线性运算进行联合优化。文献[81]采用张量框架及相关操作符，利用密文SIMD特性和交错分块打包技术，优化了二维卷积操作的效率，减少了旋转和乘法操作。文献[48]通过重新排序操作并结合改进的矩阵编码，降低了NTT操作数量，并在大矩阵计算时引入分块处理技术，显著提高了在线处理速度，减少了运行时间和通信开销。文献[82]则将常规的矩阵−向量乘法转换为HE友好的一维卷积，并结合二阶信息时延感知公式和层融合技术，有效降低推理时延。

以上优化方案主要考虑CNN模型中的线性运算，而随着大模型的出现和应用，一些研究者开始研究实现面向Transformer模型的安全推理。文献[44]首次利用同态加密实现了Transformer模型下的隐私保护推理，但没有针对各线性算子进行优化。文献[45]利用紧凑打包技术改进矩阵乘法，将多个矩阵行编码为一个密文，大大降低了线性运算的通信开销。文献[46]设计了相比文献[45]更加紧凑的打包技术，避免了因稀疏打包而导致的额外通信开销。文献[83]将FHE用于多项式操作，同时提出了词元Tokens打包方法替代基于特征的密文打包方法，减少了同态旋转，降低了离线和在线开销。文献[62]采用批量矩阵乘法和动态压缩策略，有效平衡计算与通信开销。文献[80]通过将特定层归为线性类型，结合加法同态算法及对角方法，在2轮内执行密文矩阵乘法，减少了密文向量数量，提升了线性运算效率。

从优化效果来看，虽然对模型线性层进行量化与降维减少了推理时延，但相较于同类方法提升效果有限，主要原因在于基于MPC的Transformer模型推理运算中，线性算子开销占比并不高。相较于明文模型，改变模型算子对安全推理模型性能都有一定影响。但相较于优化前，对矩阵乘法的优化能够将推理速度提高数倍甚至数10倍，优化效果明显。总之，对模型线性运算进行优化能够提高模型推理效率，但需要研究解决模型性能下降的问题。

3.2 非线性运算优化方法

对安全推理的时延进行分析，无论采用何种模型，非线性运算均占据了主要的开销。因此，针对

表5 安全推理模型线性运算优化方案对比

方案	推理模型	安全协议	优化方法	关键技术	目标	
					性能	效率
文献[46]	Transformer	HE	矩阵乘法优化	运算打包优化	↓	↑↑
文献[48]	CNN	HE	卷积运算优化	高效的乘累加设计	↓↓	↑↑
文献[62]	Transformer	MPC	矩阵乘法优化	—	↓	↑↑↑
文献[78]	Transformer	MPC	量化	将量化位数加入损失函数进行训练	↓	↑
文献[79]	Vision Transformer	MPC	降维	提出一种降维归一化注意力架构	↓	↑
文献[80]	Transformer	HE	矩阵乘法优化	—	↓↓	↑↑
文献[81]	CNN	HE	卷积运算优化	运算打包优化	↓↓	↑
文献[82]	CNN	HE	层合并	—	↓	↑↑

非线性运算进行优化一直是该领域的研究热点。在具体实施的方法上，主要包括剪枝、近似和蒸馏。剪枝是通过移除模型中不重要的非线性算子来降低模型整体运算开销；近似是使用对于安全推理计算开销更小的算子替代模型中的非线性算子；蒸馏则是通过将复杂模型中的知识转移到简单模型，实现保持性能的同时大幅减少计算复杂度。针对不同的模型和算子，上述几种优化方法产生的优化效果各不相同，下面分别对几种非线性运算优化方法进行总结与分析。

1) 剪枝

从基于剪枝方法的研究成果来看，目前主要针对 CNN 中的 ReLU 函数进行剪枝。从关键技术来看，现有的剪枝方法大致可分为 3 类：一是结构化剪枝，该方法主要针对层、通道等较大结构单元进行整体性的剪枝操作；二是非结构化剪枝，该方法则从更为细粒度的神经元层面入手，进行精细的剪枝处理；三是通过重新设计神经网络模型架构，直接构建具有较低 ReLU 数量的网络模型。典型方案对比如表 6 所示。

综合分析，从实际效果来看，结构化剪枝虽然操作简单，但粗粒度的剪枝对模型性能影响较大。采用重新设计低 ReLU 数量的神经网络模型架构方法，能够在设定 ReLU 预算下，通过训练自动寻找最优剪枝对象，在相同预算下，能够更好保持模型性能。但重新训练网络将带来巨大的额外开销，这将在下文进行具体介绍。非结构化的细粒度剪枝方法能够综合以上 2 种方法的优势，但目前研究较少，且仅有文献[91]能够在不重新训练网络的情况下，实现对模型的深度剪枝，达到模型性能和效率的更优平衡，该方法可在未来研究中重点关注。

2) 近似

从现有研究来看，基于近似的安全推理模型优化方法在各类模型中均被广泛采用。CNN 中主要针对 ReLU 进行近似，而 Transformer 模型中主要针对 Softmax 函数设计近似策略，对激活函数 GeLU，研究者们倾向于采用 ReLU 或二次多项式进行直接替代。就实现近似的具体方法而言，该领域的研究已经由最初的固定近似函数逐步发展到基于梯度的自动化搜索方法，以及更加灵活的自适应多项式近似方法。此外，针对注意力机制的近似替代，近年来已成为该领域的研究热点。表 7 从模型、优化对象、具体优化方法、性能和效率等角度对基于近似的安全推理模型非线性运算优化方案进行了对比。

综合分析，近似方法在提升模型推理效率方面展现出了显著效果，但这些方法大都对模型性能有所影响，原因可能是多项式函数的表达能力相对有限，无法完全捕捉到数据中的复杂特征和关系，从而导致模型性能的降低，因此很多研究者针对如何恢复模型性能展开研究。从模型优化效果来看，与剪枝方法相似，采用自动化搜索方式进行选择和近似替代，能够最大限度地保持模型精度并获得更优效率，但可能带来巨大的重训练开销，特别是在处理如 Transformer 这类复杂模型时，这一问题尤为突出。未来研究探索降低重训练开销也是一个可行的研究思路，如开发更高效的训练算法或利用预训练模型的迁移学习能力，以降低重训练开销，推动近似方法在复杂模型中的广泛应用。

表 6　　基于剪枝的安全推理模型非线性运算优化方案对比

方案	推理模型	安全协议	优化方法	关键技术	目标	
					性能	效率
文献[84]	CNN	MPC	重新设计模型	构建跳跃连接的网络，并搜索最优的网络架构	↓	↑
文献[85]	CNN	MPC	结构化	基于 ReLU 重要性的结构化剪枝	↓↓↓	↑
文献[86]	CNN	MPC	结构化	神经网络架构搜索进行结构化剪枝	↓↓↓	↑
文献[87]	CNN	MPC	非结构化	基于梯度进行选择	↓	↑
文献[88]	CNN	MPC	重新设计模型	神经网络架构搜索重新设计网络	↓	↑↑
文献[89]	CNN	MPC	重新设计模型	重新设计 ReLU 低的网络架构	↓	↑↑
文献[90]	CNN	MPC	重新设计模型	神经网络架构搜索重新设计网络	↓	↑↑↑
文献[91]	CNN	MPC	结构化+非结构化	基于神经元输出态进行结构化和非结构化剪枝	↓↓	↑↑

表7 基于近似的安全推理模型非线性运算优化方案对比

方案	推理模型	安全协议	优化方法	关键技术	目标	
					性能	效率
文献[92]	CNN	MPC	ReLU	具有多度数选项的通道级激活近似	↓	↑
文献[93]	CNN	MPC	ReLU	近似的符号测试，并引入新的截断方法	↓	↑↑
文献[94]	Transformer	MPC	SoftmaxGeLU	提出2Quad-Softmax近似注意力机制	↓↓	↑
文献[95]	Vision Transformer	MPC	Softmax	提出一种可学习的2Quad-Softmax方法	↓	↑↑
文献[96]	Transformer	MPC	Softmax GeLU	基于函数特殊性质设计了高质量的近似	↓	↑↑
文献[97]	Vision Transformer	MPC	Softmax GeLU	MPC感知的神经网络架构搜索	↓	↑↑
文献[98]	CNN	MPC	ReLU	自动化ReLU和多项式函数的选择，并引入了分布感知多项式近似	↓	↑↑

3) 蒸馏

除了以上提到的剪枝和量化方法外，在安全推理模型非线性函数优化中，蒸馏技术也被广泛使用[94,97,99]。

从蒸馏技术在安全推理模型优化中的具体应用来看，该技术不单独作为非线性函数的优化方法，而是在采用其他模型优化方法后，模型性能出现较大幅度下降的情况下，作为模型性能进行快速恢复的方法。一些研究通过知识蒸馏技术对模型训练或微调中的训练数据进行隐私保护[100-101]，防止通过推理结果推断出隐私训练数据，这类方法与本文关注的安全推理模型优化并不属于同一领域。目前，针对蒸馏技术在明文模型训练中已有广泛研究，用于包括训练规模更小性能更优的模型等场景[102]。但安全推理下基于蒸馏技术进行模型优化的研究还停留在简单作为性能恢复技术进行使用，缺乏针对蒸馏过程的条件参数、优化策略等研究，在实际应用中也存在蒸馏过程开销大、性能恢复效果不好等问题。

综合分析，针对模型中非线性算子的优化方法，近似方法能够最大化平衡模型性能和精度，能够更好地满足安全推理的目标需求，因此一直是模型优化的主要方法。但从近似方法的研究过程来看，需要解决设计更优替换算子、搜索最佳替换单元以及重新训练模型等问题。剪枝方法主要研究模型结构和性能关系，通过总结提出的理论规律能够实现快速高效的模型优化。但这也导致该方法与具体的模型紧密相关，适用性和可扩展性相对欠缺。目前，已有学者结合不同方法的优势，研究设计具备理论基础、便于实践应用的安全推理模型综合优化方法，这也是未来针对安全推理模型优化的一个可行的研究思路。

4 安全推理优化方案的构建成本

随着大模型的广泛应用，深度学习模型的训练成本愈发高昂。现有深度学习安全推理方案大都关注于推理过程中的性能和效率，忽略了构建安全推理方案时可能带来的训练开销等额外成本。例如一些研究者通过对网络结构进行修改，设计更加高效的安全推理模型，但需要重新对模型进行训练，由此带来的训练开销往往是巨大的。这些隐性成本虽未出现在安全推理过程中，但对该类方案的实际应用带来了巨大挑战。因此，本文增加对现有研究成果中与构建成本相关的方案进行总结梳理，主要包括对主动考虑或实际降低构建成本的安全推理方案进行对比分析。

表8中列出了现有安全推理方案中与方案构建成本相关的典型研究成果，从方案构建成本优化的角度看，主要有2条技术路线。一是针对采用神经网络架构搜索（NAS, neural network architecture search）实现安全推理模型优化的方法中，研究搜索空间压缩来降低方案构建成本；二是设计避免或降低重训练开销的安全推理方案。

综合分析，当前针对安全推理方案构建成本的研究还相对缺乏，主要研究方向在基于NAS实现安全推理模型优化时，考虑针对NAS搜索空间的优化策略。但通过设计无须重训练的策略实现低成本构建安全推理方案的文献还相对较少，针对该问

表 8　　构建成本相关的安全推理方案对比

方案	推理模型	安全协议	目标	关键技术	模型精度影响	成本降低效果
文献[58]	CNN	HE+SS+GC	降低NAS搜索开销	NAS模式可选	—	—
文献[84]	CNN	SS+GC	降低NAS搜索开销	搜索过程解耦	—	↑
文献[85]	CNN	SS+GC	无须重训练	基于ReLU重要性分布	—	↑
文献[91]	CNN	SS+GC	无须重训练	基于神经元输出态	↓	↑↑
文献[103]	CNN	HE+GC	降低NAS搜索开销	模型参数复用	↓↓↓	↑↑↑
文献[104]	CNN	—	降低重训练开销	基于ReLU重要性分布	↓↓	↑↑↑

题的研究目前还停留在对ReLU等特定非线性函数在特定网络中的优化，应用范围较窄，对安全推理的效率提升效果有限。但随着大模型对训练成本的愈发关注，将会更加重视针对安全推理方案构建成本的研究。

5　面向安全推理的隐私保护框架

从隐私保护深度学习这个研究方向看，需要同时掌握MPC等密码学和深度学习的相关知识，导致针对该方向的研究和实现面临着巨大挑战。为此，一些研究者遵循Pytorch、Tensflow等深度学习框架的设计思路，提出了能够屏蔽底层技术细节的隐私保护框架。将隐私保护框架用于安全推理，能够促进研究者或用户更加友好地构建和实现深度学习安全推理方案。

典型的早期工作包括EzPc[105]、ABY[106]、MP-SPDZ[107]、SoK[108]等，虽然这些研究提出了支持各类运算的混合安全协议或设计了面向隐私保护深度学习领域的MPC编译器，大大降低了开发难度。但这些框架主要面向密码协议进行设计，缺乏通用性，与现有的深度学习框架难以兼容，基于这些框架实现深度学习的安全推理依然非常复杂。

为与现有的主流深度学习框架保持一致，文献[109]提出了CrypTen框架，该框架模仿Pytorch接口设计，为MPC实现提供了通用的接口，屏蔽了底层的密码实现细节。用户实现模型的安全推理时，只需要按照Pytorch代码风格编写，就能快速构建安全推理模型。CrypTen框架底层的安全协议实现参考了SPDZ和ABY协议并进行了优化，具体采用算数秘密共享和布尔秘密共享以及二者之间的转换，实现了模型中主要的安全运算。

虽然CrypTen框架大大提高了通用性，但依然需要手动针对现有明文模型进行修改替换。为提供更加友好的转换方式，文献[110]提出了隐私保护机器学习框架隐语安全处理单元SecretFlow-SPU，该框架不限定代码基于的机器学习框架，开发者能够使用主流深度学习框架编写代码，而后仅修改少量代码就能实现向深度学习安全推理的迁移。SecretFlow-SPU框架对底层的安全协议也进行了优化，较MP-SPDZ等传统框架推理速度快了数倍，且支持ABY3、Cheetah、semi2K等多种高效的安全协议，大大提高了用户构建满足不同场景需求的安全推理方案的效率。

现有安全推理方案中，SecretFlow-SPU和CrypTen框架已经被广泛使用，2种框架对比如表9所示。从基于2种隐私保护框架实现的研究成果来看，主要集中在安全推理模型优化的相关研究中，原因在于隐私保护框架通常将底层的密码协议进行封装，仅提供部分调用接口，因此很难对协议进行修改和优化。但对安全推理模型的优化，基于隐私保护框架能够快速验证优化策略，且方便不同方案间进行统一对比。因此，未来针对安全推理模型优化的研究，可以更多考虑基于成熟的隐私保护框架展开，以更加关注策略本身和优化效果而不是底层密码协议实现。

表 9　　2种典型的隐私保护框架对比

方案	安全假设	语言（前端）	是否开源	通用机器学习	安全协议	易用性	基于该框架的典型安全推理方案
文献[109]	半诚实模型	Python	√	√	自定义（算数SS+布尔SS）	较强	文献[79,94-95,98]
文献[110]	适用各种威胁模型	Python	√	√	ABY3、Cheetah、semi2K	强	文献[96-97,111]

6 挑战与展望

从深度学习隐私保护近年来的研究现状来看，基于现代密码学实现安全推理已经取得了巨大进步，一些研究者开始将模型压缩、数据预处理、图形处理器（GPU）硬件加速、端到端编译环境等机器学习优化方法用于构建安全推理方案，并取得了突破性进展。但从目前该领域研究进展与深度学习的发展现状相比，可行性和实用性方面还远远滞后，特别是大数据、大模型时代下，对数据安全和隐私保护提出了新的要求和挑战，具体主要体现以下几个方面。

首先，现有的安全推理协议优化方法普遍采用固定的安全假设，主要关注于高效的密码协议设计。例如半诚实模型是目前采用最多的安全设定，而针对恶意模型下的安全推理优化研究进展则存在较大差距[11]。相同的安全配置虽然便于不同方案间的对比，但实际应用中不同场景对安全强度的要求不同，安全设置不够灵活，使得现有协议优化方法在安全性和效率间存在较大的不平衡性。此外，现有方案仅能保证推理过程的隐私，很少考虑安全推理面临的其他安全威胁。其次，模型优化技术过于依赖已有的明文模型优化方法，技术迁移过程中针对不同模型运算算子开销进行优化关注多，但基于不同密码原语实现安全推理，模型算子的开销也存在显著差异。安全运算与明文运算底层原理的不同也导致了模型优化方法在性能和效率方面难以平衡。再次，一些安全推理优化方案在提高推理效率的同时，方案本身却引入了大量额外的开销，例如重复训练等，这大大降低了大模型场景下优化方案的实用性。此外，针对影响安全推理性能和效率的理论研究还相对缺乏，优化方法缺乏可解释性。针对上述深度学习安全推理优化研究面临的主要问题，对下一步值得深入研究探索的方向进行总结和展望。

6.1 研究安全设置与场景需要相适应的安全推理协议

目前，已经有研究者开始关注安全协议的安全性与效率平衡问题。文献[112]从攻击角度进行数据隐私性评估，而后根据隐私强度设置加密层和明文层的边界，仅对部分运算层设计安全协议，大大提高了模型的安全推理效率。实现安全性的灵活设置，首先需要进行隐私量化，当前针对该问题的研究，更多关注于特定领域，其应用范围有限。加之隐私泄露涉及因素众多，目前尚未形成统一的评价模型及体系。因此，研究建立统一的隐私泄露衡量标准，结合信息熵、互信息量等指标，对推理过程中数据、模型参数及中间结果的潜在泄露风险进行建模分析，根据数据隐私保护强度设计更加轻量的安全推理协议是未来有待进一步深入研究的方向。

此外，现有安全协议主要考虑推理过程中的数据隐私问题，对模型逆向、成员推理等攻击手段的防御能力不足，需研究在协议中加入噪声、特征混淆等防御机制提高抵抗多种攻击能力。例如，文献[113]针对基于安全多方计算的安全协议仅能保护数据隐私，无法抵御通过结果对模型实施推理攻击的局限性，通过加入差分隐私噪声，在不显著增加开销的情况下提高了安全协议的安全防御能力。文献[114]通过在可信执行环境中实施安全推理，能够提供更多的安全防护，并可实现模型低通信成本的高效执行。未来可从不同攻击手段出发研究构建安全性与可用性相适应的安全推理协议，解决安全目标不明确，安全假设单一等问题。

6.2 构建面向安全推理的神经网络模型和训练方法

当前，大模型呈现井喷式发展，各类模型层出不穷，通过对明文模型进行修改和优化以适应安全推理的传统方法已难以满足快速变化的需求，需要面向安全推理研究更加适配的模型设计方法。目前，虽然已经有部分研究考虑通过神经网络架构搜索来重新构建安全推理友好的模型，但相关研究还较少。在神经网络构建时，架构的选择、参数的数量、模型的训练方式和后处理方法都可能会对推理阶段的性能产生重大影响。需重构隐私感知的训练范式，在模型训练和优化阶段显式引入安全计算约束。文献[97]将安全运算的实际时延加入损失函数进行模型训练，构建的MPC感知的Transformer模型更加适用于安全推理场景。也可利用知识蒸馏技术将大模型能力迁移至专为安全推理设计的小模型中，避免大模型的效率瓶颈以及在大模型上执行模型优化对精度的影响。因此，未来针对安全推理领域的独特性，研究构建专用的深度学习模型、特定的训练方法以及标准化的安全算子库，从而端到端实现安全计算更加友好、推理时延更低的安全推理模型。

6.3　设计包含性能、效率、成本等关键要素的安全推理方案综合评价指标

当前针对安全推理方案的评估主要采用效率和性能2个指标，但不同方案采用的基准测试集、硬件环境以及对比基线的不同导致了横向对比的困难，此外，在真实环境中对部分方案进行测试也发现实验结果与真实表现存在较大差异。一个主要原因在于评价指标单一，方案设计时对场景适用性和方案易用性等考虑不足。本文对方案构建成本、可依托的隐私保护框架相关研究也进行了总结分析，目前面向实际应用进行安全推理解决方案的研究还相对欠缺。未来可从关系深度学习安全推理的关键环节出发，例如实际推理时延、模型算子、密码协议等，对这些要素进行整体衡量，研究设计综合成本、安全、性能、效率等目标的综合评价指标。基于综合评价指标，研究者也可针对安全推理框架展开研究，站在统揽安全推理全过程的视角，研究构建满足安全推理整体优化的基础范式，为安全推理的方案构建提供更加完整、统一的参考框架。

6.4　探索基于可解释性深度学习的安全推理优化机理

目前针对安全推理优化的研究还主要依赖经验性参数调优，缺乏对密码协议与模型架构交互机理的深入理解。可解释性分析方法是解决深度学习模型黑盒问题的主要方法手段，能够从全局视角给予模型构建相应的理论支撑，目前已经受到很多学者的关注和研究，但针对安全推理模型的可解释性研究还相对缺乏。未来可以从深度学习安全推理内在机理出发，研究密码协议、模型结构与推理结果间的关联关系，例如通过可视化技术定位安全推理中的速度瓶颈，对少数关键计算步骤进行重点优化；建立数学模型，量化不同加密技术与模型性能、推理时延间的联合优化指标，推导出安全推理相关的最优化方法。通过将可解释性分析等理论用于深度学习安全推理中，发现并提出相关理论规律，充实完善安全推理优化方法的理论研究。

7　结束语

随着以深度学习为代表的人工智能的普及，特别是大模型的高速发展和应用，深度学习即服务中的隐私安全问题愈发严峻。将基于密码的隐私保护技术与深度学习结合能够实现模型的安全推理，但巨大的推理时延限制了安全推理的实际应用，因此如何实现高效的安全推理得到学者们的广泛关注和研究。本文首先针对安全推理研究成果分类不统一的问题，构建了安全推理层次结构框架，从运算层、实现层和目标层3个角度理清了安全推理方案中的关键要素及关联关系。然后面向实际应用需求，从实现层出发对现有研究成果进行总结梳理，区分线性和非线性的底层实现，对安全推理协议优化和安全推理模型优化目前的技术路线、问题缺陷及发展趋势进行了总结归纳，并对典型的代表性方案进行了对比分析。此外，本文还增加了对方案构建成本、隐私保护框架这2个与优化方案可用性紧密相关问题的研究与分析，进一步充实和完善了对深度学习高效安全推理研究现状的综述成果。最后，本文分析了当前及未来深度学习安全推理可能面临的挑战，并提出了解决这些问题的思路见解，期望为该领域未来的研究和发展提供更加清晰、实用的参考和帮助。

参考文献：

[1] SURMA J. Deep learning in military applications[J]. Safety & Defense, 2024, 10(1): 1-7.

[2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[3] LIANG W, LI Y H, XU J L, et al. QoS prediction and adversarial attack protection for distributed services under DLaaS[J]. IEEE Transactions on Computers, 2024, 73(3): 669-682.

[4] CUI L Z, CHEN Z T, YANG S, et al. A secure and decentralized DLaaS platform for edge resource scheduling against adversarial attacks[J]. IEEE Transactions on Computers, 2024, 73(3): 631-644.

[5] MARCOLLA C, SUCASAS V, MANZANO M, et al. Survey on fully homomorphic encryption, theory, and applications[J]. Proceedings of the IEEE, 2022, 110(10): 1572-1609.

[6] DAMGÅRD I, ESCUDERO D, FREDERIKSEN T, et al. New primitives for actively-secure MPC over rings with applications to private machine learning[C]//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 1102-1120.

[7] ZHANG Q, XIN C S, WU H Y. Privacy-preserving deep learning based on multiparty secure computation: a survey[J]. IEEE Internet of Things Journal, 2021, 8(13): 10412-10429.

[8] 何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049-2070.
HE Y Z, HU X B, HE J W, et al. Privacy and security issues in machine learning systems: a survey[J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070.

[9] TANUWIDJAJA H C, CHOI R, BAEK S, et al. Privacy-preserving deep learning on machine learning as a service: a comprehensive survey[J].

IEEE Access, 2020, 8: 167425-167447.

[10] ZHANG Q, XIANG T, CAI Y F, et al. Privacy-preserving machine learning as a service: challenges and opportunities[J]. IEEE Network, 2023, 37(6): 214-223.

[11] 龙春, 李丽莎, 李婧, 等. 机器学习安全推理研究综述[J]. 数据与计算发展前沿(中英文), 2024, 6(5): 1-12.
LONG C, LI L S, LI J, et al. Review of research on secure inference in machine learning[J]. Frontiers of Data & Computing, 2024, 6(5): 1-12.

[12] 陈品枫, 何琨, 陈晶, 等. 隐私保护深度学习研究综述[J]. 密码学报(中英文), 2024, 11(4): 771-798.
CHEN P J, HE K, CHEN J, et al. A survey on privacy-preserving deep learning[J]. Journal of Cryptography (Chinese and English), 2024, 11(4): 771-798.

[13] LI Z W, LIU F, YANG W J, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999-7019.

[14] 张焕, 张庆, 于纪言. 激活函数的发展综述及其性质分析[J]. 西华大学学报(自然科学版), 2021, 40(4): 1-10.
ZHANG H, ZHANG Q, YU J Y. Overview of the development of activation function and its nature analysis[J]. Journal of Xihua University (Natural Science Edition), 2021, 40(4): 1-10.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556, 2014.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceeding of the 31st International Conference on Neural Information Processing Systems. California: MIT Press, 2017: 6000-6010.

[19] RIVEST R L, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms[J]. Foundations of secure computation, 1978, 4(11): 169-180.

[20] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//Proceedings of the forty-first annual ACM symposium on Theory of computing. New York: ACM Press, 2009: 169-178.

[21] RIVEST R L, SHAMIR A, ADLEMAN L. A method for obtaining digital signatures and public-key cryptosystems[J]. Communications of the ACM, 1978, 21(2): 120-126.

[22] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classess[C]//International Conference on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2007: 223-238.

[23] POLLARD J M. The fast Fourier transform in a finite field[J]. Mathematics of Computation, 1971, 25(114): 365-374.

[24] BRAKERSKI Z, GENTRY C, VAIKUNTANATHAN V. (Leveled) fully homomorphic encryption without bootstrapping[J]. ACM Transactions on Computation Theory, 2014, 6(3): 1-36.

[25] GENTRY C, SAHAI A, WATERS B. Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based[C]//Proceedings of the 33rd Annual Cryptology Conference. Berlin: Springer, 2013: 75-92.

[26] CHEON J H, KIM A, KIM M, et al. Homomorphic encryption for arithmetic of approximate numbers[C]//Proceedings of the 2017 23rd International Conference on the Theory and Applications of Cryptology and Information Security. Berlin: Springer, 2017: 409-437.

[27] WANG B, WANG X, XUE R. Leveled FHE with matrix message space[C]//Proceedings of the 13th International Conference on Information Security and Cryptology. Berlin: Springer, 2018: 260-277.

[28] YAO A C. Protocols for secure computations[C]//Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982). Piscataway: IEEE Press, 1982: 160-164.

[29] RABIN M O. How to exchange secrets with oblivious Transfer[J]. Cryptology ePrint Archive, 2005, 187: 1-26.

[30] JIANG H, XU Q L. Secure multiparty computation in cloud computing[J]. Journal of Computer Research and Development, 2016, 53(10): 2152-2162.

[31] ISHAI Y, KILIAN J, NISSIM K, et al. Extending oblivious transfers efficiently[C]//Proceedings of the 23rd Annual International Cryptology Conference. Berlin: Springer, 2003: 145-161.

[32] YAO A C. How to generate and exchange secrets[C]//Proceedings of the 27th Annual Symposium on Foundations of Computer Science (sfcs 1986). Piscataway: IEEE Press, 1986: 162-167.

[33] LINDELL Y, PINKAS B. A proof of security of Yao's protocol for two-party computation[J]. Journal of Cryptology, 2009, 22(2): 161-188.

[34] BELLARE M, HOANG V T, ROGAWAY P. Foundations of garbled circuits[C]//Proceedings of the 2012 ACM conference on Computer and communications security. New York: ACM Press, 2012: 784-796.

[35] MICALI S, GOLDREICH O, WIGDERSON A. How to play any mental game[C]//Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC. New York: ACM Press, 1987: 218-229.

[36] SHAMIR A. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612-613.

[37] Gilad-Bachrach R, Dowlin N, Laine K, et al. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy[C]//Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, 2016: 201-210.

[38] JUVEKAR C, VAIKUNTANATHAN V, CHANDRAKASAN A. Gazelle: a low latency framework for secure neural network inference[J]. arXiv Preprint, arXiv: 1801.05507, 2018.

[39] ZHANG Q, WANG C, WU H Y, et al. GELU-Net: a globally encrypted, locally unencrypted deep neural network for privacy-preserved learning[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2018: 3933-3939.

[40] BRUTZKUS A, GILAD-BACHRACH R, ELISHA O. Low latency privacy preserving inference[C]//Proceedings of the 36th International Conference on Machine Learning. New York: PMLR, 2019: 812-821.

[41] LI S H, XUE K P, ZHU B, et al. FALCON: a Fourier transform based approach for fast and secure convolutional neural network predictions[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8702-8711.

[42] SAIT S M, MEHTA P, GÜRSES D, et al. Cheetah optimization algorithm for optimum design of heat exchangers[J]. Materials Testing,

2023, 65(8): 1230-1236.

[43] KIM D, GUYOT C. Optimized privacy-preserving CNN inference with fully homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2175-2187.

[44] CHEN T, BAO H, HUANG S, et al. The-x: privacy-preserving transformer inference with homomorphic encryption[J]. arXiv Preprint, arXiv: 2206.00216, 2022.

[45] HAO M, LI H, CHEN H, et al. Iron: private inference on transformers[J]. Advances in Neural Information Processing Systems, 2022, 35: 15718-15731.

[46] PANG Q, ZHU J H, MÖLLERING H, et al. BOLT: privacy-preserving, accurate and efficient inference for transformers[C]//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2024: 4753-4771.

[47] CHEN H, DAI W, KIM M, et al. Efficient multi-key homomorphic encryption with packed ciphertexts with application to oblivious neural network inference[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 395-412.

[48] BALLA S, KOUSHANFAR F. HELiKs: HE Linear Algebra Kernels for Secure Inference[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2023: 2306-2320.

[49] 李瑞琪, 易琴, 黄艺璇, 等. 基于同态密文转换的隐私保护卷积神经网络推理方案[J]. 通信学报, 2024, 45(S1): 12-23.
LI R Q, YI Q, HUANG Y X, et al. Privacy-preserving convolutional neural network inference scheme based on homomorphic ciphertext transformation[J]. Journal on Communications, 2024, 45(S1):12-23.

[50] ZHANG Q, XIN C S, WU H Y. GALA: greedy ComputAtion for linear algebra in privacy-preserved neural networks[J]. arXiv Preprint, arXiv: 2105.01827, 2021.

[51] XU G W, HAN X S, ZHANG T W, et al. SIMC 2.0: improved secure ML inference against malicious clients[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(4): 1708-1723.

[52] MOHASSEL P, ZHANG Y P. SecureML: a system for scalable privacy-preserving machine learning[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 19-38.

[53] LIU J, JUUTI M, LU Y, et al. Oblivious neural network predictions via MiniONN transformations[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 619-631.

[54] ROUHANI B D, RIAZI M S, KOUSHANFAR F. DeepSecure: scalable provably-secure deep learning[C]//Proceedings of the 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2018: 1-6.

[55] RIAZI M S, WEINERT C, TKACHENKO O, et al. Chameleon: a hybrid secure computation framework for machine learning applications[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. New York: ACM Press, 2018: 707-721.

[56] MOHASSEL P, RINDAL P. ABY3: a mixed protocol framework for machine learning[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2018: 35-52.

[57] WAGH S, GUPTA D, CHANDRAN N. SecureNN: 3-party secure computation for neural network training[J]. Proceedings on Privacy Enhancing Technologies, 2019(3): 26-49.

[58] SRINIVASAN W Z, AKSHAYARAM P, ADA P R. Delphi: a cryptographic inference service for neural networks[C]//Proceedings of the 29th USENIX Security Symposium. Berkeley: USENIX Association, 2019, 3.

[59] LEHMKUHL R, MISHRA P, SRINIVASAN A, et al. Muse: secure inference resilient to malicious clients[C]//Proceedings of the 30th USENIX Security Symposium. Berkeley: USENIX Association, 2021: 2201-2218.

[60] WAGH S, TOPLE S, BENHAMOUDA F, et al. Falcon: honest-majority maliciously secure framework for private deep learning[J]. Proceedings on Privacy Enhancing Technologies, 2021, 1: 188-208.

[61] SUN L L, LI H, PENG Y G, et al. Serpens: privacy-preserving inference through conditional separable of convolutional neural networks[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2022: 1837-1847.

[62] LU W J, HUANG Z C, GU Z, et al. BumbleBee: secure two-party inference framework for large Transformers[J]. Cryptology ePrint Archive, 2023, 1678: 1-18.

[63] 马敏, 付钰, 黄凯, 等. 基于秘密共享的轻量级隐私保护 ViT 推理框架[J]. 通信学报, 2024, 45(4): 27-38.
MA M, FU Y, HUANG K, et al. Light weighted privacy protection ViT inference framework based on secret sharing[J]. Journal on Communications, 2024, 45(4): 27-38.

[64] CHOU E, BEAL J, LEVY D, et al. Faster CryptoNets: leveraging sparsity for real-world encrypted inference[J]. arXiv Preprint, arXiv: 1811.09953, 2018.

[65] LOU Q, JIANG L. SHE: a fast and accurate deep neural network for encrypted data[J]. arXiv Preprint, arXiv: 1906.00148, 2019.

[66] LI Q F, HUANG Z C, LU W J, et al. HomoPAI: a secure collaborative machine learning platform based on homomorphic encryption[C]//Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2020: 1713-1717.

[67] MEFTAH S, TAN B H M, MUN C F, et al. DOReN: toward efficient deep convolutional neural networks with fully homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3740-3752.

[68] KIM D, PARK J, KIM J, et al. HyPHEN: a hybrid packing method and its optimizations for homomorphic encryption-based neural networks[J]. IEEE Access, 2023, 12: 3024-3038.

[69] RIAZI M, SAMRAGH M, CHEN H, et al. XONN: XNOR-based oblivious deep neural network inference[C]//Proceedings of the 28th USENIX Security Symposium (USENIX Security 19). Berkeley: USENIX Associatio, 2019: 1501-1518.

[70] RYFFEL T, THOLONIAT P, POINTCHEVAL D, et al. AriaNN: low-interaction privacy-preserving deep learning via function secret sharing[J]. arXiv Preprint, arXiv: 2006.04593, 2020.

[71] GUPTA K, JAWALKAR N, MUKHERJEE A, et al. SIGMA: secure GPT inference with function secret sharing[J]. Cryptology ePrint Archive, 2023, 1269: 1-19.

[72] BI R W, XIONG J B, LUO C Q, et al. Communication-efficient privacy-preserving neural network inference via arithmetic secret sharing[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 6722-6737.

[73] RATHEE D, RATHEE M, KIRAN GOLI R K, et al. SiRNN: a math library for secure RNN inference[C]//Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2021: 1003-1020.

[74] RATHEE D, RATHEE M, KUMAR N, et al. Cryptflow2: practical 2-party secure inference[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 325-342.

[75] DONG Y, CHEN X J, JING W Z, et al. Meteor: improved secure 3-party neural network inference with reducing online communication costs[C]//Proceedings of the ACM Web Conference 2023. New York: ACM Press, 2023: 2087-2098.

[76] DING Y C, GUO H, GUAN Y W, et al. East: efficient and accurate secure transformer framework for inference[J]. arXiv Preprint, arXiv: 2308.09923, 2023.

[77] CHEN Y T, MENG X J, SHI Z Y, et al. SecureTLM: Private inference for transformer-based large model with MPC[J]. Information Sciences, 2024, 667: 120429.

[78] WU H, FANG W, ZHENG Y, et al. Ditto: quantization-aware secure inference of transformers upon MPC[J]. arXiv Preprint, arXiv: 2405.05525, 2024.

[79] CHEN D K, ZHANG Y K, KUNDU S, et al. RNA-ViT: reduced-dimension approximate normalized attention vision transformers for latency efficient private inference[C]//Proceedings of the 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD). Piscataway: IEEE Press, 2023: 1-9.

[80] WANG W Z, KUANG Y. CipherFormer: efficient transformer private inference with low round complexity[C]//Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Piscataway: IEEE Press, 2024: 3054-3059.

[81] AHARONI E, ADIR A, BARUCH M, et al. Helayers: a tile tensors framework for large neural networks on encrypted data[J]. arXiv Preprint, arXiv: 2011.01805, 2020.

[82] XU T, WU L, WANG R, et al. PrivCirNet: efficient private inference via block circulant transformation[J]. arXiv Preprint, arXiv: 2405.14569, 2024.

[83] ZHENG M, LOU Q, JIANG L. Primer: fast private transformer inference on encrypted data[C]//Proceedings of the 2023 60th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2023: 1-6.

[84] GHODSI Z, VELDANDA A K, REAGEN B, et al. Cryptonas: private inference on a ReLU budget[J]. Advances in Neural Information Processing Systems, 2020, 33: 16961-16971.

[85] JHA N K, GHODSI Z, GARG S, et al. Deepreduce: ReLU reduction for fast private inference[C]//International Conference on Machine Learning. New York: PMLR, 2021: 4839-4849.

[86] CHO M, GHODSI Z, REAGEN B, et al. Sphynx: a deep neural network design for private inference[J]. IEEE Security & Privacy, 2022, 20(5): 22-34.

[87] CHO M, JOSHI A, REAGEN B, et al. Selective network linearization for efficient private inference[C]//International Conference on Machine Learning. New York: PMLR, 2022: 3947-3961.

[88] CHENG K, XI N, LIU X M, et al. Private inference for deep neural networks: a secure, adaptive, and efficient realization[J]. IEEE Transactions on Computers, 2023, 72(12): 3519-3531.

[89] JHA N K, REAGEN B. DeepReShape: redesigning neural networks for efficient private inference[J]. arXiv Preprint, arXiv: 2304.10593, 2023.

[90] ZENG W X, LI M, YANG H C, et al. CoPriv: network/protocol co-optimization for communication-efficient private inference[J]. Advances in Neural Information Processing Systems, 2023, 36: 78906-78925.

[91] HU P, SUN L, HU C Y, et al. DReP: Deep ReLU pruning for fast private inference[J]. Journal of Systems Architecture, 2024, 152: 103156.

[92] LOU Q, SHEN Y, JIN H, et al. SafeNet: a secure, accurate and fast neural network inference[C]//International Conference on Learning Representations. Vienna: ICLR, 2021: 1-13.

[93] GHODSI Z, JHA N K, REAGEN B, et al. Circa: stochastic ReLUs for private deep learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 2241-2252.

[94] LI D C, SHAO R L, WANG H Y, et al. MPCFormer: fast, performant and private Transformer inference with MPC[J]. arXiv Preprint, arXiv: 2211.01452, 2022.

[95] ZHANG Y K, CHEN D K, KUNDU S, et al. SAL-ViT: towards latency efficient private inference on ViT using selective attention search with a learnable softmax approximation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 5093-5102.

[96] DONG Y, LU W J, ZHENG Y C, et al. PUMA: secure inference of LLaMA-7B in five minutes[J]. arXiv Preprint, arXiv: 2307.12533, 2023.

[97] ZENG W X, LI M, XIONG W J, et al. MPCViT: searching for accurate and efficient MPC-friendly vision transformer with heterogeneous attention[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 5029-5040.

[98] PENG H W, HUANG S Y, ZHOU T, et al. AutoReP: automatic ReLU replacement for fast private network inference[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 5155-5165.

[99] DONG C Q, WENG J, LIU J N, et al. Fusion: efficient and secure inference resilient to malicious servers[J]. arXiv Preprint, arXiv: 2205.03040, 2022.

[100] TIAN Z, ZHAO Y, HUANG Z, et al. Seqpate: differentially private text generation via knowledge distillation[J]. Advances in Neural Information Processing Systems, 2022, 35: 11117-11130.

[101] WANG J, BAO W D, SUN L C, et al. Private model compression *via* knowledge distillation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1190-1197.

[102] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.
SHAO R R, LIU Y A, ZHANG W, et al. A survey of knowledge distillation in deep learning[J]. Chinese Journal of Computers, 2022, 45(8): 1638-1673.

[103] BIAN S, JIANG W, LU Q, et al. Nass: optimizing secure inference via neural architecture search[J]. arXiv Preprint, arXiv: 2001.11854, 2020.

[104] KUNDU S, LU S, ZHANG Y K, et al. Learning to linearize deep neural networks for secure and efficient private inference[J]. arXiv Preprint, arXiv: 2301.09254, 2023.

[105] CHANDRAN N, GUPTA D, RASTOGI A, et al. EzPC: programmable and efficient secure two-party computation for machine learning[C]//Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2019: 496-511.

[106] DEMMLER D, SCHNEIDER T, ZOHNER M. ABY - a framework for efficient mixed-protocol secure two-party computation[C]// Proceedings of the 2015 Network and Distributed System Security Symposium. San Diego: Internet Society. 2015:1-15.

[107] KELLER M. MP-SPDZ: a versatile framework for multi-party computation[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 1575-1590.

[108] HASTINGS M, HEMENWAY B, NOBLE D, et al. SoK: general purpose compilers for secure multi-party computation[C]//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 1220-1237.

[109] KNOTT B, VENKATARAMAN S, HANNUN A, et al. Crypten: secure multi-party computation meets machine learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 4961-4973.

[110] MA J, ZHENG Y, FENG J, et al. SecretFlow-SPU: a performant and user-friendly framework for privacy-preserving machine learning[C]// 2023 USENIX Annual Technical Conference. Berkeley: USENIX Association, 2023: 17-33.

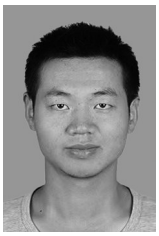
[111] DHYANI N, MO J Q, CHO M, et al. PriViT: vision transformers for fast private inference[J]. arXiv Preprint, arXiv: 2310.04604, 2023.

[112] ZHANG Y K, CHEN D K, KUNDU S, et al. C2PI: an efficient crypto-clear two-party neural network private inference[C]//Proceedings of the 2023 60th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2023: 1-6.

[113] JAYARAMAN B, WANG L, EVANS D, et al. Distributed learning without distress: privacy-preserving empirical risk minimization[J]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 6346-6357.

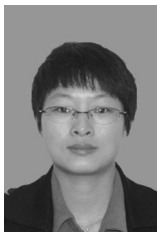
[114] NATARAJAN D, LOVELESS A, DAI W, et al. Chex-mix: combining homomorphic encryption with trusted execution environments for oblivious inference in the cloud[C]//Proceedings of the 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2023: 73-91.

[作者简介]

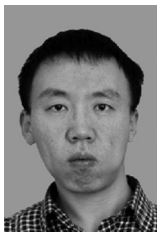


胡鹏（1988– ），男，甘肃武威人，信息工程大学博士生，主要研究方向为网络空间安全、隐私保护、人工智能安全等。

孙磊（1973– ），男，江苏靖江人，博士，信息工程大学教授、博士生导师，主要研究方向为人工智能和信息系统安全。



胡翠云（1985– ），女，河南辉县人，博士，信息工程大学讲师，主要研究方向为人工智能和数据安全。



郭松（1985– ），男，河北保定人，博士，信息工程大学讲师，主要研究方向为人工智能和信息系统安全。

王晶雯（2001– ），女，河南南召人，信息工程大学硕士生，主要研究方向为神经网络安全推理。

王志鸿（2002– ），男，江苏盐城人，信息工程大学硕士生，主要研究方向为网络空间安全、人工智能安全等。

姚敬怡（2001– ），女，河南开封人，信息工程大学硕士生，主要研究方向为网络空间安全、人工智能安全等。