VHENN:基于环上零知识证明协议的可验证同态加密神经网络推理方案

杨文梯[1)] 何朝阳[1)] 李 萌[2)] 张子剑[3)] 关志涛[1)] 祝烈煌[3)]

[1)](华北电力大学控制与计算机工程学院 北京 102206)

[2)](合肥工业大学计算机与信息学院 合肥 230601)

[3)](北京理工大学网络空间安全学院 北京 100081)

摘 要 近年来,诸如ChatGPT、DeepSeek等神经网络推理服务的发展,使得小微企业及个人等不具备海量数据或充足算力的用户也能受益于神经网络强大的表征能力。然而,随着人们对隐私泄露问题的关注,神经网络推理服务中的两个关键问题亟待解决:(1)如何在推理过程中保护用户的数据和推理结果不被泄露;(2)如何在保证模型隐私不被泄露的前提下,实现用户对模型和推理结果的可验证性。虽然目前已有部分研究分别基于同态加密、安全多方计算等密码学技术实现对用户数据和推理结果的隐私保护,基于零知识证明实现在保护模型隐私的前提下的推理可验证性,但这些研究均未能同时解决上述两个问题。因此,本文结合同态加密和零知识证明,提出了一种可验证同态加密神经网络推理方案-VHENN。为了解决同态加密与零知识证明结合过程中存在的各种挑战,本方案首先基于Rinocchio,一种用于环上电路的零知识简洁非交互知识论证,以适应基于环多项式构造的同态加密方案,实现同态加密计算的可验证性。随后,将可验证同态加密方案与神经网络推理相结合,实现满足模型、推理数据、推理结果隐私保护以及模型真实性和推理正确性可验证的神经网络推理方案。实验结果表明,得益于同态加密可以采用单指令多数据操作的特性,本方案在零知识证明的构造过程中显著减少了约束数量,降低幅度达到1至3个数量级。相比于对比方案,本方案在可信设置、证明生成和验证等环节的计算时间缩短了超过4个数量级。

关键词 神经网络推理;隐私保护;可验证;同态加密;零知识证明

中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2025.01458

VHENN: A Verifiable Homomorphic Encrypted Neural Network Inference Scheme Based on Zero-Knowledge Proof for Ring Computations

YANG Wen-Ti[1)] HE Zhao-Yang[1)] LI Meng[2)] ZHANG Zi-Jian[3)]

GUAN Zhi-Tao[1)] ZHU Lie-Huang[3)]

[1)](School of Control and Computer Engineering, North China Electric Power University, Beijing 102206)

[2)](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601)

[3)](School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081)

Abstract In recent years, neural network inference services such as ChatGPT and DeepSeek have provided small and medium-sized enterprises and individuals with access to advanced AI capabilities without requiring massive datasets or extensive computational power. These services have made it possible to harness the representation power of neural networks across a wide range

收稿日期:2024-09-19;在线发布日期:2025-03-21。本课题得到国家自然科学基金面上项目(62372173, 62372149)、国家自然科学基金重点项目(U23A20303)资助。 **杨文梯**,博士研究生,主要研究方向为隐私保护机器学习、隐私计算。E-mail:yangwt1020@qq.com。**何朝阳**,硕士研究生,主要研究方向为隐私保护机器学习。**李 萌**,博士,副教授,主要研究方向为数据安全、隐私保护、应用密码学、区块链、TEE和车联网。**张子剑**,博士,教授,主要研究方向为身份认证与密钥协商协议的设计以及实体行为与偏好分析。**关志涛**(通信作者),博士,教授,主要研究方向为人工智能安全、隐私计算、系统安全、区块链与应用密码。E-mail:guan@ncepu.edu.cn。**祝烈煌**,博士,教授,主要研究方向为密码学、网络与信息安全。

of applications, from natural language processing to image recognition, enabling users to achieve sophisticated results with minimal technical expertise. However, the widespread adoption of these services has raised significant privacy concerns, particularly in scenarios where sensitive user data is involved. Two critical issues arise in neural network inference services that must be addressed: (1) ensuring that users′ data and inference results are protected from potential leaks during the inference process, and (2) providing a mechanism for verifying the authenticity of models and correctness of inference results while preserving the privacy of the model itself. To address these challenges, cryptographic techniques such as Homomorphic Encryption (HE) and Secure Multi-party Computation (MPC) have been explored to safeguard user data and inference results, enabling computations on encrypted or shared data without exposing sensitive information. However, despite these advances, neither HE nor MPC alone can address the dual requirements of privacy preservation and verifiability in neural network inference. Zero-Knowledge Proofs (ZKPs) have been introduced to ensure the verifiability of models and inference results without revealing sensitive model details, but integrating these cryptographic tools into a single, cohesive framework that addresses both privacy and verifiability remains an open challenge. In this paper, we propose VHENN (Verifiable Homomorphic Encrypted Neural Network Inference Scheme), a novel scheme that combines homomorphic encryption and zero-knowledge proofs to provide a solution for both privacy and verifiability in neural network inference. Our approach is built on Rinocchio, a Zero-Knowledge Succinct Non-Interactive Argument of Knowledge (zk-SNARK) protocol, which is tailored for ring circuits. Rinocchio is particularly well-suited for verifiable with homomorphic encryption schemes due to its compatibility with schemes based on ring polynomials. By leveraging Rinocchio, we achieve verifiability of homomorphically encrypted computations, allowing us to confirm the verifiability of encrypted computations without revealing the underlying computed data. The core innovation of VHENN lies in its ability to integrate verifiable homomorphic encryption with neural network inference. This integration ensures that user data, models and inference results are fully protected during the inference process, while also providing verifiable guarantees of model authenticity and result correctness. Furthermore, the scheme addresses the efficiency challenges associated with combining homomorphic encryption and zero-knowledge proofs. Specifically, our approach takes advantage of the Single Instruction, Multiple Data (SIMD) feature of homomorphic encryption, which allows multiple operations to be performed simultaneously on encrypted data. This significantly reduces the number of constraints in the construction of zero-knowledge proofs, cutting them by 1 to 3 orders of magnitude compared to non-SIMD solutions. Experimental results demonstrate the effectiveness of VHENN in reducing computational overhead. Compared to other privacy-preserving inference schemes, VHENN achieves substantial improvements in the computation time required for trusted setup, proof generation, and verification—by more than 4 orders of magnitude.

Keywords　neural network inference; privacy-preservation; verification; homomorphic encryption; zero-knowledge proofs

1　引　言

近年来，随着计算机硬件设施性能不断提升、大数据持续积累、深度学习关键技术取得重大突破，神经网络得到了空前的发展。为了满足小微企业及个人等不具备海量数据或充足算力的用户对深度学习的需求，诸如ChatGPT、DeepSeek等神经网络推理服务得到广泛应用。在这种服务模式下，服务方提供具有良好性能的神经网络模型，用户根据需求将

数据发送给服务方以获得推理服务。然而，随着对数据安全与隐私问题关注度的提升，神经网络推理服务中存在的两个问题引发了研究人员的关注：

（1）隐私泄露：在神经网络推理服务中，用户需要向服务方提供私有数据，由于服务方对于用户并不是完全可信的，数据中包含的敏感信息可能会被泄露或滥用，推理结果对于服务方也是完全公开的。因此现有的神经网络推理模式中面临着较为严峻的隐私泄露问题[1-2]。

（2）可验证性：在神经网络推理服务中，存在一些问题影响推理结果。首先，模型所有者可能在推理服务中提供错误或与最初所宣称的不一致的模型（后文称为模型真实性）。其次，模型所有者可能会为了降低计算开销、遭受攻击等原因，返回错误的推理结果（后文称为推理正确性）。而模型本身属于服务方的重要资产，通过公开模型以验证推理结果的方式难以实现。因此，为了在不泄露模型隐私信息的前提下，保证模型的真实性和推理结果的正确性，实现神经网络推理的可验证至关重要[3]。

以如下场景为例：在AI医生的场景中，患者希望能够在不泄露个人隐私信息和诊断结果的情况下获取诊断服务。此外，患者也希望能够验证AI医生模型的真实性以及诊断结果的准确性，而服务提供方则希望保护模型参数等核心信息不被泄露。因此，在此背景下，实现诊断过程的隐私保护与结果的可验证性，对于患者和服务提供方而言都是关键需求。

为了解决神经网络推理中数据隐私泄露的问题，已有部分相关工作提出基于安全多方计算（Multi-Party Computation，MPC）[4-6]和同态加密（Homomorphic Encryption，HE）[7-9]的神经网络安全推理，通过多方合作或加密推理的形式保证模型、推理数据和推理结果的隐私。为了防止恶意服务器破坏推理正确性，研究者提出了具有恶意安全的多方安全推理方案。然而当前方案满足不诚实大多数的恶意安全模型，仅容忍1[10-11]个或$t<n/2$[12]个恶意方，其中t为腐败阈值，n为参与方的数量。或仅支持$t<n-1$个[13]恶意、静态敌手，只能防止协议启动前的破坏。因此，这些方案只能在恶意方的数量不超过t时，才能保证计算的正确性。部分工作[14-15]采用敏感样本生成、混合检查等方法来验证推理结果的完整性或正确性，这些方案中的验证方法类似于抽样检查，其可靠性和有效性是概率性的，并且只能支持批数据推理的验证。此外，上述安全推理方案均未实现对模型真实性的验证。

现有基于零知识证明（Zero-Knowledge Proof，ZKP）的可验证神经网络推理方案中，ZKP的零知识性保证了神经网络模型的机密性，这些方案主要包括：（1）基于交互式ZKP的可验证推理方案[16-19]，要求在生成证明时验证者与证明者进行交互；（2）基于非交互式ZKP的可验证推理方案，主要采用零知识简洁非交互知识论证（Zero-Knowledge Succinct Non-Interactive Argument of Knowledge，zk-SNARKs）作为底层技术[20-24]。虽然zk-SNARKs在证明生成过程对内存的要求较高，但它不要求验证者与证明者的交互，因此在神经网络推理中仍然具有较大的应用潜力。然而，上述工作并未考虑推理数据和推理结果的隐私保护问题。

因此，为了同时满足模型、推理数据、推理结果的隐私保护以及模型真实性和推理正确性的可验证，直观方法是将安全多方计算或同态加密与零知识证明进行结合，实现可验证的多方计算或加密计算，并将其应用于神经网络推理中。由于基于安全多方计算需要多方参与，并且最多只能抵抗n-1个参与方的合谋攻击，为了防止n个参与方合谋攻击导致的隐私泄露，需要数据所有者和模型所有者也作为参与方执行多方推理过程，无法很好地适用于资源受限的用户。因此，本文研究集中于如何基于zk-SNARKs协议构造可验证同态加密方案，并将其应用于神经网络推理中。

在构造可验证神经网络加密推理方案时，存在一些挑战亟待解决：（1）构造可验证同态加密方案。在基于同态加密的计算中，复杂的同态操作会导致计算电路的改变，原有的底层加法和乘法需要由基于环多项式的同态加法和同态乘法实现，并且需要复杂的重线性化、密钥交换等操作。而zk-SNARKs通常支持基于有限域的算术电路或布尔电路，因此基于zk-SNARKs为基于环多项式算术电路的同态加密计算生成零知识证明成为一大挑战。（2）可验证同态加密方案与神经网络推理的结合。由于神经网络推理中存在一些非线性运算，无法将其直接构造为算术电路，因此将可验证同态加密方法用于神经网络推理过程中面临着计算不兼容的问题，需要对神经网络加密推理过程进行适应性的优化。（3）计算效率问题。由于zk-SNARKs非交互式的特点，在计算针对神经网络推理过程的零知识证明时，需要对推理过程中所有的计算生成约束并包含在证明中。而神经网络推理过程又包含了大量的基础运算，因此证明中会包含大量约束，造成较大的计算开销。这是当前基

于zk-SNARKs方案的可验证神经网络推理普遍面临的问题，而结合同态加密后，每个基础运算都会扩展为同态运算，计算效率问题会更为突出。

为了解决上述挑战，本文提出了一种可验证的同态加密神经网络推理方案VHENN。首先基于Rinocchio[25]，一种用于环上电路的简洁非交互知识论证（Succinct Non-Interactive Argument of Knowledges，SNARKs），构造可验证同态加密方案。Rinocchio提出了二次环程序（Quadratic Ring Program，QRP），通过将环上的算术电路转换为QRP，并将其用于zk-SNARKs协议的构造，从而实现同态加密环上电路的零知识证明。随后，将可验证同态加密方案与神经网络推理相结合，以构造可验证的安全推理方案。本文主要贡献总结如下：

（1）首先，本文基于环上电路的zk-SNARKs协议，提出了同态加密方案中的乘法、位分解等重要运算到QRP的转换方法，从而构造可验证同态加密方案。

（2）其次，本文将可验证同态加密方案结合到神经网络加密推理中，并对推理中的非线性计算进行了适应性的调整，实现了满足模型、推理数据、推理结果隐私保护以及模型真实性和推理正确性可验证的神经网络推理方案VHENN。

（3）最后，本文对所提方案进行了实验评估及对比。通过采用单指令多数据（Single Instruction，Multiple Data，SIMD）操作，降低zk-SNARKs证明系统的约束数量，从而提升可信设置、证明生成和验证的效率。在实验示例中，相比于未采用SIMD技术时的计算，本文方案约束数量降低了1至3个数量级，相比于对比方案pvCNN[26]，本文方案在各环节的计算时间降低超过4个数量级。

2　相关工作

本文方案的重点在于保证神经网络推理的隐私保护和可验证性，而据我们所知，目前尚未有研究基于可验证同态加密实现神经网络的推理方案。因此本章从神经网络安全推理、可验证神经网络推理、可验证神经网络安全推理以及可验证同态加密四个方面对相关工作进行介绍。

2.1　神经网络安全推理

机器学习中通用的隐私保护技术的研究当前主要包括：以安全多方计算和同态加密为代表的加密技术，以差分隐私（Differential Privacy，DP）为代表的扰动技术，以及以可信执行环境（Trusted Execution Environment，TEE）为代表的基于硬件的隐私保护技术。其中，基于MPC和HE的隐私保护方法具有较高的安全性，能够满足机器学习中多种隐私保护需求，由于通信和计算开销较大，目前研究主要集中于对效率的优化[27]。基于DP的隐私保护方法由于其依赖于统计分析中的数据扰动机制，无法满足单个数据点的隐私保护需求，从而无法在神经网络推理中使用以保护推理数据的隐私[28]。基于TEE的隐私保护方法的安全性高度依赖硬件环境，对所依赖的硬件设施有较高的要求[29]。总的来说，若能有效降低通信与计算开销，基于MPC和HE的隐私保护技术能够更广泛地适用于神经网络推理场景中，并提供更高的安全性。

基于MPC的神经网络安全推理方案的研究主要集中于：（1）神经网络线性层（如全连接层、卷积层）和非线性层（如激活函数）MPC协议的转换；（2）为了兼容神经网络推理计算与安全多方计算而衍生出的定点数浮点数转换问题、密码学方案友好的激活函数构造等；（3）为了提高安全性，实现支持恶意安全模型的方案；（4）提高多方交互时的通信效率等方面[4, 30-32]。此外，随着大语言模型的发展，部分工作[5, 33-35]针对Transformer模型和大语言模型的安全多方推理进行了研究。由于安全多方计算需要多方进行交互才能实现安全的计算，因此基于安全多方计算的神经网络安全推理目前面临的主要问题就是通信效率问题。特别是面对神经网络推理中大量的乘法运算，以及为了兼容神经网络所需要的协议转换、安全截断等带来的通信开销。

在基于HE的神经网络安全推理方案中，通过将推理中的明文运算转换为同态加密运算，保证推理数据或模型的隐私。Dowlin等人[7]提出CryptoNets，首个神经网络加密推理方案，其采用有限级数同态加密实现加密数据的神经网络安全推理，并采用SIMD技术增加同态加密的计算效率。目前基于同态加密的研究主要集中于降低计算开销、近似函数构造等方面[36-39]。为了避免激活函数的多项式近似，也有部分研究针对算术同态加密，如BGV（Brakerski-Gentry-Vaikuntanathan，BGV）、CKKS（Cheon-Kim-Kim-Song，CKKS），和布尔同态加密，如（Fully Homomorphic Encryption over the Torus，TFHE），的转换[40]，但由于转换效率问题，目前未得到广泛研究与应用。由于有限级数同态加密限制了加密计算的深度，同态加密难以在深度神经网络中

应用，而采用全同态加密会导致更高的计算开销。因此，虽然同态加密避免了多方交互，并且能够提供更高的安全性，适用范围更加广泛，但其效率问题仍然是亟待解决的一大挑战。

2.2 可验证神经网络推理

为了在保证模型隐私的前提下，实现模型真实性或结果正确性的验证，部分研究基于零知识证明构造可验证神经网络推理方案。当前基于零知识证明的神经网络推理方案主要分为两类[3]，一类基于交互式零知识证明协议[16-19]，另一类则基于常见的非交互式零知识证明协议 zk-SNARKs[20-23, 26]。交互式可验证神经网络推理方案在证明生成时间与公共参数大小方面具有较大的优势，但要求验证者（多为推理服务的用户）与证明者（多为模型所有者）进行交互，且验证开销较大，因此对用户具有一定的要求。非交互式可验证神经网络推理由于其非交互式、易验证等特性，对用户更加友好。但其证明生成计算开销较大且具有较大的公共参数。

上述方案解决了神经网络推理中的可验证问题，在保护模型不被泄露的前提下，实现用户对模型真实性和结果正确性的验证。然而，未考虑推理数据和推理结果的隐私保护问题。

2.3 可验证神经网络安全推理

部分工作尝试实现可验证神经网络安全推理方案，但这些方案或仅针对简单机器学习模型，如支持向量机、线性回归等[41-42]；或采用敏感样品生成、混合和检验等方法[14-15]，这些方案的可验证性本质上是概率性的，并且这些可验证方案仅验证计算结果的正确性，均未考虑对模型真实性的验证。

Weng 等人[26]提出了 pvCNN，通过同态加密和 zk-SNARKs 来实现卷积神经网络推理中的隐私保护和可验证性。他们将模型分为 PriorNet 和 LaterNet，其中 PriorNet 保持私有，而 LaterNet 为非隐私部分，委托给服务器进行计算。然而，这会导致委托模型缺乏足够的隐私保护。此外，pvCNN 是在不同的阶段使用 HE 和 zkSNARKs，其可验证性仅在服务器端有效。

2.4 可验证同态加密

当前已有针对同态加密可验证的研究，主要包括基于消息认证码（Message Authentication Code，MAC）、可信执行环境（Trusted Execution Environment，TEE）、零知识证明等三类底层技术的方法。MAC 通常被用于验证传统对称加密的完整性。在同态加密中，需要采用同态 MAC，使得服务器能够在对密文进行同态操作时，将输入密文的有效 MAC 转换为输出密文的有效 MAC。然而，现有基于 MAC 的解决方案仅支持半同态加密或部分同态加密（如仅支持一次同态乘法操作），尚不清楚目前是否存在完全支持常用的全同态加密操作及密文维护操作（如重新线化）的同态 MAC 方法[43]。基于 TEE 的安全性对执行硬件有较强的依赖性，而同态加密较高的计算复杂度，特别是在包含大量同态操作的神经网络推理中，相比于不可信的底层硬件，对 TEE 在内存和计算能力等方面具有更高的要求[44]。零知识证明作为实现可验证同态加密的潜在解决方案，近年来得到了研究者的关注。然而，现有基于 ZKP 的可验证同态加密方案仅支持加法同态加密如 Paillier 同态加密[45]或简单的有限级数同态加密如 BV（Brakerski-Vaikuntanathan）同态加密[46]，无法很好地应用于神经网络推理中。

3 预备知识

3.1 多项式环

环多项式是指定义在代数结构“环”上的多项式。给定一个环 R，一个环多项式可以表示为：$P(x)=a_0+a_1x+\cdots+a_nx^n$，其中 $a_i\in R$ 为环多项式的系数。多项式环 $R[X]$ 则表示所有属于环多项式的集合。多项式环满足封闭性、交换律、结合律和分配律。多项式环可以通过模一个不可约多项式 $f(x)$ 得到商环，即 $R[X]/f(x)$，表示所有多项式在多项式 $f(x)$ 下按模运算的等价类集合。商环的同构性使其在密码学方案的构造中起到了重要作用。

3.2 BGV 同态加密

目前神经网络加密推理方案中，通常采用 BGV[47]或 CKKS[48]作为底层同态加密方案。CKKS 支持浮点数运算，在神经网络中的应用更为广泛。然而，为了实现浮点数运算，CKKS 引入了缩放因子，将实数编码为整数进行加密和运算。由于舍入误差和噪声积累，解密结果通常为近似值。目前传统 zk-SNARKs 协议主要针对有限域、环等精确元素进行验证，二者结合存在较多困难。因此，本文选择 BGV 作为底层同态加密方案。

本文采用基于环上误差学习（Ring Learning with Errors，RLWE）的 BGV 方案。定义 λ 为安全参数，在 RLWE 中，首先定义多项式环 $R[X]=\mathbb{Z}[X]/(X^d+1)$，其中 $d=d(\lambda)$ 选定为 2 的幂，$\mathbb{Z}$

为整数环，即整数集合。$R_q[X]=\mathbb{Z}_q[X]/(X^d+1)$为密文空间，$q=q(\lambda)$为密文模数。$R_t[X]=\mathbb{Z}_t[X]/(X^d+1)$为明文空间，$t=t(\lambda)$为明文模数，$t\ll q$，$\mathbb{Z}_q[X]$，$\mathbb{Z}_t[X]$分别表示系数在模$q$和模$t$的整数环$\mathbb{Z}_q$和$\mathbb{Z}_t$上的多项式环。$\chi=\chi(\lambda)$定义为环$R$上的一个噪音分布。定义$params=(d,q,t,\chi)$。算法的具体描述如下：

(1)$Setup(1^\lambda,1^L)\rightarrow(params)$：以安全参数$\lambda$和级数$L$为输入。首先定义模数链$\{p_0,p_1,\cdots,p_L\}$，对于每个层级$0\leqslant l\leqslant L$，密文模数$q_l=p_0\cdot p_1\cdot\cdots\cdot p_l$，初始密文模数$q=q_L=p_0\cdot p_1\cdot\cdots\cdot p_L$。记$params=(q_L,\cdots q_0,d,t,\chi)$。

(2)$KeyGen(params)\rightarrow(pk,sk)$：选择小系数多项式$s\leftarrow R_q$，使得$s$的系数属于$\{-1,0,1\}^d$，记作$s\in\{-1,0,1\}^d$。随机选择参数$a\leftarrow R_q$，$e\leftarrow\chi$，计算$b=[a\cdot s+t\cdot e]_q$，其中$[\cdot]_q$表示对$q$取模。记$pk=(b,a)$，$sk=(1,s)$。

(3)$Enc(params,pk,PT)\rightarrow CT$：对于明文$PT\in R_t$，选择小系数多项式$v\leftarrow R_q$使得$v\in\{-1,0,1\}^d$，随机选择$e_0\leftarrow\chi$，$e_1\leftarrow\chi$，计算$c_0=[b\cdot v+te_0+PT]_q$，$c_1=[-a\cdot v+te_1]_q$。记$CT=(c_0,c_1)$。

(4)$Dec(params,sk,CT)\rightarrow PT$：计算$PT=[[c_0+c_1s]_{q_l}]_t$，其中$q_l$为密文$CT$所属层级$l$的模数。

(5)$Add(CT_1,CT_2)\rightarrow CT_3$：对两个相同层级下的密文$CT_1=(c_{10},c_{11})$，$CT_2=(c_{20},c_{21})$执行加法运算，得到$c_{30}=c_{10}+c_{20}$，$c_{31}=c_{11}+c_{21}$，$CT_3=(c_{30},c_{31})$。由于加法不会导致密文规模增大，也不会造成过大的噪音增加，因此直接在密文各自的分量上执行加法即可。

(6)$Mul_{PT}(PT,CT)\rightarrow CT_1$：对密文$CT=(c_0,c_1)$，$PT=p$，计算$c_{10}=p\cdot c_0$，$c_{11}=p\cdot c_1$，$CT_1=(c_{10},c_{11})$。

(7)$Mul_{CT}(CT_1,CT_2)\rightarrow CT_3$：对两个相同层级下的密文$CT_1=(c_{10},c_{11})$，$CT_2=(c_{20},c_{21})$执行乘法运算，有$<CT_1,sk>\cdot<CT_2,sk>=<CT_1\otimes CT_2,sk\otimes sk>$，该等式是由克罗内克积的性质得到。因此$CT_1$与$CT_2$的乘法可表示为$CT_1$与$CT_2$的张量积：$CT_3=CT_1\otimes CT_2$，对应密钥$sk'=sk\otimes sk$。记$d_0=c_{10}\cdot c_{20}$，$d_1=c_{10}\cdot c_{21}+c_{11}\cdot c_{20}$，$d_2=c_{11}\cdot c_{21}$，$CT_3=(d_0,d_1,d_2)$，$sk'=(1,s,s^2)$。在执行密文乘法后，密文和密钥扩张为对应的张量积，需要采用密钥交换方法，将扩张后的密文和对应的密钥转换为降低规模后的密文和密钥。另外，为了控制乘法带来的噪声增长，需要执行模切换操作。将乘法后需执行的上述操作合并为$Refresh$算法，具体描述见下文。

(8)$SwitchKeyGen(T,sk,sk')\rightarrow ek$：在执行密钥交换前，需要生成交换密钥。首先，选择$T$，一个较小的质数或2的幂，作为位分解的基。随后，对于$i=0,\cdots,\lceil\log_T q\rceil-1$，随机选择$a_i\leftarrow R_q$，$e_i\leftarrow\chi$，$b_i=[a_i\cdot s+t\cdot e_i+T^i\cdot s^2]_q$，其本质是用$s$加密$s^2$。记$ek=(T,\{a_i,b_i\}_0^{\lceil\log_T q\rceil-1})$。

(9)$Refresh(CT_3,ek)\rightarrow CT_5$：本算法对执行密文乘法后扩张的密文执行密钥交换与模切换操作，以恢复密文密钥带来的扩张并控制乘法操作带来的噪声增长。

①$SwitchKey(CT_3,l,ek)\rightarrow CT_4$：对于$l$层级的密文$CT_3=(d_0,d_1,d_2)$，对$d_2$进行分解使得：

$$d_2=\sum_{i=0}^{\lceil\log_T q_l\rceil-1}d_{2,i}\cdot T^i \tag{1}$$

随后计算：

$$c_{40}=[d_0+\sum_{i=0}^{\lceil\log_T q_l\rceil-1}d_{2,i}\cdot b_i]_{q_l} \tag{2}$$

$$c_{41}=[d_1+\sum_{i=0}^{\lceil\log_T q_l\rceil-1}d_{2,i}\cdot a_i]_{q_l} \tag{3}$$

记$CT_4=(c_{40},c_{41})$。

②$ModSwitch(CT_4,l)\rightarrow CT_5$：为了控制噪声的相对大小，对于$l$层级的密文$CT_4$，需要通过模切换将密文模数$q_l$降低为$q_{l-1}$，得到$l$-1层级的密文$CT_5=(c_{50},c_{51})$。

执行上述步骤完成密钥交换与模交换，密文恢复原有规模，级数降低，当级数降低为0时，便无法继续进行乘法操作。

BGV的安全性基于RLWE问题的决策困难性假设，决策RLWE问题描述如下：

给定环$R_q[X]=\mathbb{Z}_q[x]/f(x)$，模数$q$，以及噪声分布$\chi$。随机选择$a\leftarrow R_q$，以及小系数多项式$s\leftarrow R_q$，噪声多项式$e\leftarrow\chi$。给定$(a,b)$使得$b=a\cdot s+e$，攻击者无法在计算上区分$b$是由$b=a\cdot s+e$构造的还是从$R_q$中随机选择的。

在随机密钥下，BGV满足语义安全性（选择明文攻击下的不可区分性，IND-CPA），即对于任意明文PT_0、PT_1，加密后密文CT_1、CT_2的分布对任何多项式时间的攻击者是不可区分的。决策RLWE假设确保密文中隐藏的密钥s和噪声e无法通过密文的统计性质泄露。

3.3　同态加密SIMD操作

BGV同态加密方案支持基于SIMD的高效批

量同态操作，通过将多个明文编码并加密到一个密文中，每个明文对应密文中的一个槽，从而实现在两个不同密文的多个槽之间并行地执行同态操作。由于神经网络的计算中包含大量基础运算，并行操作可以大大降低计算电路大小，从而降低证明系统的约束数量，提高证明计算效率。神经网络推理计算中，线性层操作可以表示为向量点积运算，基于SIMD操作的环多项式向量点积的计算过程如图1所示，其中$\odot$表示ct_1和ct_2间的分量乘积，$ct_4=Rot_{L,1}(ct_3)$，$ct_5=Rot_{L,2}(ct_3)$，$ct_6=Rot_{L,3}(ct_3)$，$Rot_{L/R,i}(ct)$表示对密文ct的旋转操作，L表示向左旋转，R表示向右旋转，i为旋转的步长。

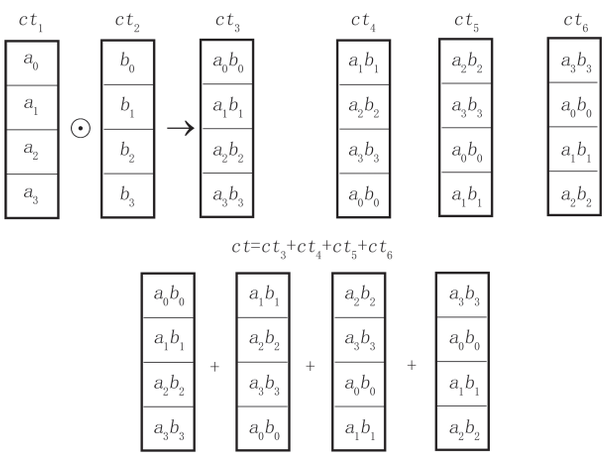

图1 基于SIMD操作的环多项式向量点积计算过程

3.4 二次环程序

在传统的zk-SNARKs协议中，通常采用二次算术程序(Quadratic Arithmetic Programs，QAP)表示基于有限域的算术电路中的约束，通过将电路验证问题转化为多项式的点值验证问题，从而通过代数运算验证计算的正确性。Ganesh等人[25]首次提出二次环程序(Quadratic Ring Program，QRP)，其与QAP具有类似的特征，不同的是其可以满足基于环的算术电路的验证问题。一个基于有限交换环R上的QRP关系Q包括三组多项式：$U=\{u_k(X):k\in[0,m]\}$，$V=\{v_k(X):k\in[0,m]\}$，$W=\{w_k(X):k\in[0,m]\}$，以及一个目标多项式$t(X)$，这些多项式均属于环多项式$R(X)$。设$\mathcal{C}$是一个具有n输入和n'个输出的环上算术电路，若满足下列条件，则称Q为电路C的QRP关系：

存在一组有效的输入/输出变量$a_1,a_2,\cdots,a_n,a_{m-n'+1},a_{m-n'+2},\cdots,a_m\in R^{n+n'}$，使得存在变量$a_{n+1},a_{n+2},\cdots,a_{m-n'}\in R^{m-n-n'}$，满足$t(X)$整除$p(X)$，其中，

$$p(X)=U(X)\cdot V(X)-W(X) \tag{4}$$

$$U(X)=(u_0(X)+\sum_{k=1}^{m}a_k\cdot u_k(X)) \tag{5}$$

$$V(X)=(v_0(X)+\sum_{k=1}^{m}a_k\cdot v_k(X)) \tag{6}$$

$$W(X)=(w_0(X)+\sum_{k=1}^{m}a_k\cdot w_k(X)) \tag{7}$$

定义Q的大小为m，目标多项式$t(X)$的度为m。上述多项式$U(X)$、$V(X)$、$W(X)\in R[X]$以及对应的变量赋值组成电路$\mathcal{C}$的QRP关系。

为了构造电路$\mathcal{C}$的QRP，首先选择一个特殊集A，其中的元素两两之间的差值都是可逆的。为每个乘法门$g\in\mathcal{C}$选择元素$r_g\in A$，并定义目标多项式：

$$t(X)=\prod_{g\in C}(X-r_g) \tag{8}$$

根据中国剩余定理(Chinese Remainder Theorem，CRT)，多项式$u_k(X)$、$v_k(X)$、$w_k(X)$可以通过对$r_g\in A$进行多项式插值得到。对于定义为$I_g=x-r_g$的$I_1,I_2,\cdots,I_{\deg(t(X))}$，由于A是一个特殊集，$I_1,I_2,\cdots,I_{\deg(t(X))}$必然是互素的。对于$p(X)\equiv p(r_g)\bmod(X-r_g)$，满足：

$$\phi:R[X]/t(X)\simeq R[X]/I_1\times R[X]/I_2\times\cdots\times R[X]/I_{\deg(t(X))} \tag{9}$$

$$p(X)\rightarrow(p(r_1),p(r_2),\cdots,p(r_{\deg(t(X))})) \tag{10}$$

上述同构表明，当且仅当$p(r_g)=0$时，才可以满足$t(X)$整除$p(X)$，记$t(X)$除$p(X)$得到的商多项式为$h(X)$，则满足$p(X)=t(X)h(X)$。

3.5 环上zk-SNARKs协议

Groth16是一种基于椭圆曲线群构造的zk-SNARKs协议，其在证明大小和验证时间方面具有较大的优势，能够很好地适用于神经网络推理中资源受限的用户。本文采用了Rinocchio[25]中构造的具有Groth16类似结构的环上zk-SNARKs协议(后续称为R-Groth16)。首先将QRP多项式由在一个秘密点处计算的多项式编码表示，该编码在基于环的计算中具有加同态特性。将编码方案表示为算法$(Gen,Encode)$，具体过程如下：

(1)$Gen(1^\lambda)\rightarrow(pk,sk)$：该算法为密钥生成算法，以安全参数$\lambda$为输入，输出为公钥$pk$和私钥$sk$。

(2)$Encode(a,sk)\rightarrow E(a)$：该算法为一种概率编码算法，将一个环元素$a\in R$映射到编码空间$S$，使得集合$\{\{E(a)\}:a\in R\}$成为编码空间$S$的一个划分。

在R-Groth16的构造中，$(Gen,Encode)$表示仅

线性编码(Linear-only encodings)，其线性操作在编码中保持一致性，包括加法和标量乘法的一致性。具体的，对于$\alpha\in R^*, x, y\in R$，有$E(x)+E(y)=E(x+y)$，$\alpha E(x)=E(\alpha x)$，其中R^*为环R的单位群，该单位群是由R中所有存在乘法逆元的元素组成的集合，$A^*\in R^*$表示一个特殊集。

假设一个环R上的电路C具有m个线和n个乘法门，电路C的QRP关系表示为$Q=(\{u_k(X), v_k(X), w_k(X)\}_{k=0}^m, t(X))$。使$I_s=1,2,\cdots,l$与电路的声明值相对应，声明(Statement)是公开的，表示需要被证明的公开信息，l为电路中表示声明的线的数量。$I_w=l+1, l+2,\cdots,m$与电路的见证值相对应，见证(Witness)是与声明相关的秘密信息，证明者需要基于这个信息生成证明，该信息对验证者保密。R-Groth16的方案构造如下：

(1)$Setup(1^\lambda, Q)\rightarrow(CRS, vk)$：该算法由可信第三方执行，以安全参数$\lambda$和关系$Q$为输入，输出为公共参考字符串(Common Reference String，CRS)和验证密钥vk。首先由可信第三方执行$Gen(1^\lambda)$算法，生成公私钥对(pk, sk)。随后随机选择$\alpha,\beta,\gamma,\delta\leftarrow R^*, \varepsilon\leftarrow A^*$，并计算CRS：

$$CRS=\begin{pmatrix} pk,\ \{E(\varepsilon^j)\}_{j=0}^{n-1}, E(\alpha), E(\beta), \\ \left\{E(\gamma^{-1}(\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)))\right\}_{k\in I_s}, \\ \left\{E(\delta^{-1}(\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)))\right\}_{k\in I_w}, \\ \left\{E(\delta^{-1}(\varepsilon^j t(\varepsilon)))\right\}_{j=0}^{n-1} \end{pmatrix} \tag{11}$$

$$vk=(sk, CRS, \varepsilon, \gamma, \delta) \tag{12}$$

在上述参数中，CRS用于生成证明，被发送给证明方。其中，$\left\{E(\gamma^{-1}(\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)))\right\}_{k\in I_s}$在证明生成中作用于与声明相对应的参数，$\left\{E(\delta^{-1}(\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)))\right\}_{k\in I_w}$在证明生成中作用于与见证相对应的参数。$vk$用于验证，被发送给验证方。

(2)$Prove(Q, CRS, x, w)\rightarrow\pi$：该算法由证明方执行，以QRP关系$\left(Q=\left\{u_k(X), v_k(X), w_k(X)\right\}_{k=0}^m, t(X)\right)$，公共参考字符串$CRS$，声明$x=(a_1, a_2,\cdots, a_l)$，见证$w=(a_{l+1}, a_{l+2},\cdots, a_m)$为输入，输出为与关系$Q$对应的电路的证明$\pi$。设$a_0=1$，$u_w(\varepsilon)=\sum_{k=l+1}^m a_k u_k(\varepsilon)$，$v_w(\varepsilon)=\sum_{k=l+1}^m a_k v_k(\varepsilon)$，$w_w(\varepsilon)=\sum_{k=l+1}^m a_k w_k(\varepsilon)$。证明方计算：

$$A=E(A_u)=E\left(\alpha+\sum_{k=0}^m a_k u_k(\varepsilon)\right) \tag{13}$$

$$B=E(B_v)=E\left(\beta+\sum_{k=0}^m a_k v_k(\varepsilon)\right) \tag{14}$$

$$C=E(C_w)=E\left(\frac{\beta u_w(\varepsilon)+\alpha v_w(\varepsilon)+w_w(\varepsilon)+h(\varepsilon)t(\varepsilon)}{\delta}\right) \tag{15}$$

最后，将证明$\pi=(A, B, C)$发送给验证方。

(3)$Verify(Q, vk, x, \pi)\rightarrow\{0,1\}$：该算法由验证方执行，以QRP关系$Q$，验证密钥$vk$，声明$x$，证明$\pi$为输入，如果验证成功，则返回1。若验证不成功，则返回0。首先验证方计算

$$f_s=\frac{(\beta u_s(\varepsilon)+\alpha v_s(\varepsilon)+w_s(\varepsilon))}{\gamma} \tag{16}$$

$$F=E(f_s) \tag{17}$$

其中，$u_s(\varepsilon)=\sum_{k=0}^l a_k u_k(\varepsilon)$，$v_s(\varepsilon)=\sum_{k=0}^l a_k v_k(\varepsilon)$，$w_s(\varepsilon)=\sum_{k=0}^l a_k w_k(\varepsilon)$。随后，验证下列等式是否成立：

$$AB=E(\alpha)E(\beta)+\gamma F+\delta C \tag{18}$$

若成立则返回1，若不成立则返回0。

R-Groth16协议满足以下安全性：

定义1. 完备性(Completeness). 持有声明和见证的模型所有者可以生成一个证明。在验证这个证明时，验证者输出0的概率$negl(\lambda)$是可以忽略不计的，其中λ是安全参数。

定义2. 知识合理性(Knowledge Soundness). 对于计算能力有限且不持有见证的敌手，存在一个计算能力有限的提取器ε，它可以完全访问敌手的状态。每当敌手生成一个有效的证明时，提取器ε就可以计算出一个相应的见证，使得一组满足QRP关系Q的声明x和见证w，即$(x, w)\in Q$，和证明π说服验证者的概率可以忽略不计。

定义3. 零知识性，(Zero-Knowledge). 存在一个模拟器S，可以在不依赖见证的情况下生成与真实证明无法区分的模拟证明。即对于一个能力有限的敌手，它可以以忽略不计的概率$negl(\lambda)$区分真实证明和模拟证明。

4　方案概述

4.1　系统模型

本文采用Commit-and-Prove证明系统，首先由

证明方通过承诺协议对所持有的“秘密”进行承诺，在神经网络推理场景下，该“秘密”特指神经网络参数。随后，证明某结果是由承诺的参数经过执行特定计算得到的。也就是，证明推理结果是由所承诺的神经网络经过前向传播算法所得。本文从神经网络加密推理和证明系统的构造两方面进行描述，主要由三类实体组成，包括可信第三方、模型所有者、数据所有者。对每个实体的描述如下：

（1）可信第三方：可信第三方在初始化阶段生成零知识证明协议的公共参考字符串（Common Reference String，CRS）和承诺协议的结构化参考字符串（Structured Reference String，SRS）。用于生成CRS和SRS的密钥必须保密，并在初始化过程结束后销毁。

（2）模型所有者：模型所有者首先对其持有的模型参数进行承诺，并负责生成神经网络推理计算和承诺计算对应的QRP关系。在接收到数据所有者发送的加密数据后，执行加密推理和证明生成过程。

（3）数据所有者：将持有的待推理数据利用同态加密方案进行加密，并发送给模型所有者。在接收到加密的推理结果及其对应的证明后，对推理结果进行验证，并进行解密。

4.2 威胁和安全模型

在传统基于同态加密的神经网络安全推理中，通常设置执行加密推理的服务器为被动敌手模型（也称半诚实敌手模型），在这个假设中，服务器虽然会试图挖掘用户的隐私信息，但也会按照既定的协议执行加密推理过程，并返回正确的计算结果。然而，在实际应用中，可能存在服务器为恶意或受到恶意攻击的情况，从而破坏推理结果的正确性或带来安全威胁。因此，在基于同态加密的神经网络加密推理中，同样也面临着推理结果不可验证的问题。实现加密推理的可验证性，可以使得服务器满足主动敌手模型（也称恶意敌手模型），对于解决下列挑战至关重要：

（1）正确性：在基于同态加密的神经网络推理中，服务器可能会返回不正确的推理结果。推理结果的正确性主要包括两个方面，一方面，恶意服务器没有正确地执行推理过程或在推理中采用了不一致的模型，从而导致错误的推理结果；另一方面，服务器由于计算失误（如超出预期的噪声溢出）产生了错误的密文，导致无法正确解密。这种错误可以通过按照计算电路合理地设置安全参数来进行规避。

（2）机密性：恶意服务器可能会利用针对同态加密的密钥恢复攻击[49-50]破坏用户数据的机密性。这超出了现有基于同态加密的神经网络安全推理方案中假定的被动敌手模型设置的范围。密钥恢复攻击通过构造错误的特殊密文，并利用用户对特殊密文的解密失败后做出的反应（例如，通过请求重新运行计算或终止进一步交互）恢复部分或全部用户私钥，从而造成用户数据的隐私泄露。

本方案设定模型所有者为主动敌手模型，它不仅会试图挖掘用户推理数据和推理结果中的隐私信息，还可能执行错误的协议，从而得到错误的推理结果。另一方面，本方案为数据所有者设定被动敌手模型。他们可能试图从神经网络模型或推理的数据中提取隐私信息，但会遵循既定的协议，不会主动参与操纵、干扰或修改数据。

本方案主要考虑以下两方面的安全性：

（1）在加密推理阶段，推理数据和推理结果的机密性依赖于BGV同态加密方案的安全性，其安全性具体描述见3.2节。

（2）在加密推理证明阶段，模型的机密性与验证的安全性依赖于所基于的零知识证明协议的安全性，包括完备性、知识合理性和零知识性，具体描述见3.5节。

4.3 方案流程

为了实现可验证的神经网络加密推理，本方案首先基于环上运算的zk-SNARK方案，构造可验证的BGV同态加密方案。随后，基于可验证BGV，完成VHENN的方案设计。如图2所示，VHENN主要包括：模型承诺、初始化、加密推理、加密推理的证明生成、验证与解密。每个步骤的说明如下：

（1）模型承诺：模型所有者对所持有的模型进行承诺，并将生成的承诺公开。该步骤保证在后续的推理服务中，模型所有者始终采用了同一模型。

（2）初始化：在初始化阶段，首先构造加密推理的QRP，具体的，将神经网络前向传播中的每个基础操作转换为同态操作，并为其构造子电路，生成对应的QRP。然后将模型参数承诺的计算过程生成QRP。最后，将这些子电路的QRP进行连接，最终生成整个电路的QRP。随后执行R-Groth16协议和BGV同态加密的设置阶段，生成必要的参数如R-Groth16的CRS和BGV的公共参数。其中R-Groth16的设置为可信设置，需要由可信第三方执行，BGV的设置则无需额外设置可信第三方。

（3）加密推理：数据所有者将推理数据加密并发送给模型所有者。模型所有者完成对加密数据的

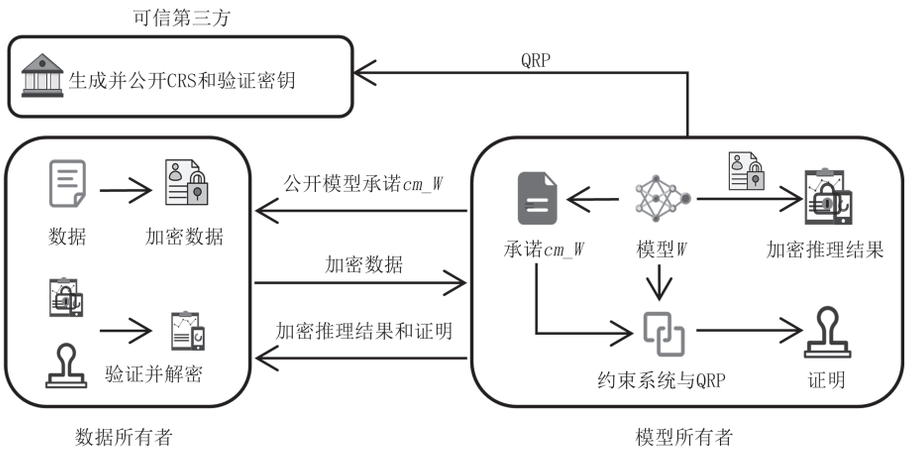

图2　VHENN方案流程

神经网络推理过程，得到加密的推理结果并发送给数据所有者。该过程保护了数据所有者的数据及推理结果的机密性。

（4）加密推理的证明生成：基于神经网络加密推理的QRP关系，完成加密推理的证明生成，并将证明发送给数据所有者。

（5）推理结果的验证与解密：收到加密推理结果和证明后，用户对其进行验证及解密。如果验证失败，则表明模型所有者没有遵守约定的计算协议，或者在推理过程中使用了不正确的模型参数。

5　可验证BGV

采用zk-SNARK协议构造可验证BGV，要点在于如何将基于BGV的密文计算电路转换为合适的代数表示形式，从而实现简洁、高效的证明系统。现有包括QAP、R1CS（Rank-1 Constraint System）等电路的代数表示形式通常适用于基于有限域的算术或布尔电路。在基于BGV的加密计算中，明文包括一个多项式环元素，密文包括两个多项式环元素，加密计算组成的环上算术电路中包括的环多项式加法和环多项式乘法主要由BGV的密文加法、密文与明文乘法以及密文乘法等运算组成。由于QAP等用于有限域电路的表达形式难以直接用于环上算术电路，因此我们采用Rinocchio中提出的QRP作为BGV密文计算电路的代数表示形式，以构造可验证的BGV方案。

首先，对BGV中同态操作的QRP构造与证明生成过程进行描述。记BGV中的明文为$PT\in R_t$，密文为$CT\in R_q$，BGV中的同态操作主要包含密文加法$Add(CT_1,CT_2)\to CT_3$，密文与明文乘法$Mul_{PT}(PT,CT)\to CT_1$和密文乘法$Mul_{CT}(CT_1,CT_2)\to CT_3$。对于密文加法和密文与明文乘法，只需要将密文中对应的组分分别执行环多项式加法和环多项式乘法即可实现。具体的，在密文加法中，对$CT_1=(c_{10},c_{11})\in R_q^2$和$CT_2=(c_{20},c_{21})\in R_q^2$，计算$c_{30}=c_{10}+c_{20}$，$c_{31}=c_{11}+c_{21}$，$CT_3=(c_{30},c_{31})$，相当于计算两个度至多为$d-1$的多项式的加法运算，将各项系数相加即可。在密文与明文乘法中，对$CT=(c_0,c_1)\in R_q^2$，$PT=p\in R_t$，计算$c_{10}=p\cdot c_0$，$c_{11}=p\cdot c_1$，$CT_1=(c_{10},c_{11})$，只需进行两次环多项式乘法。

同态密文乘法相较于上述两种操作较为复杂，因此实现可验证性BGV，主要工作集中于密文乘法操作的QRP构造。

如3.2节所述，执行密文乘法中的复杂操作主要包括：

（1）$CT_3=CT_1\otimes CT_2$，张量积操作由环多项式乘法实现。

（2）密钥交换中，对密文$CT_3=(d_0,d_1,d_2)$中的d_2进行位分解：

$$d_2=\sum_{i=0}^{\lceil \log_T q_l\rceil-1} d_{2,i}\cdot T^i \tag{19}$$

（3）高维向量（$\lceil \log_T q_l\rceil$）点积操作（包括环多项式乘法、环多项式加法）。

（4）模切换中常数与环多项式乘法操作。

下面主要介绍多项式乘法和位分解的QRP构造方法，以实现BGV中同态操作的QRP构造。

5.1　环多项式乘法QRP构造

直观上，多项式乘法可以简单地由多项式各项相乘得到，然而，对两个次数为n的多项式直接相乘时间复杂度为$O(n^2)$，在处理大规模数据时，传统多项式乘法计算会变得非常昂贵。数论变换（Number

Theoretic Transform，NTT)提供了一种快速计算多项式乘法的方法，可以将时间复杂度降低为$O(n\log n)$。通过NTT对环多项式乘法进行优化，可以将多项式乘法转换为一系列加法门、乘法门。

首先，根据3.5节乘法门的QRP构造方法实现对单个乘法门的QRP构造。随后，对于两个输出线和输入线首尾相连的算术电路$\mathcal{C}_1$和$\mathcal{C}_2$，I_1和I_2为电路中线的索引，$I_1\cap I_2$表示$\mathcal{C}_1$中输出线作为$\mathcal{C}_2$中输入线的部分。对于$i\in\{1,2\}$，$k\in I_i$：

$$Q_i=\begin{pmatrix}U^{(i)}=\{u_k^{(i)}(X)\\V^{(i)}=\{v_k^{(i)}(X)\\W^{(i)}=\{w_k^{(i)}(X)\\t^{(i)}(X)\end{pmatrix} \tag{20}$$

为相应的QRP。那么，$Q=Q_2\circ Q_1$表示为电路$\mathcal{C}=\mathcal{C}_1\circ\mathcal{C}_2$的QRP，其中，$U=\{u_k(X):k\in I_1\cup I_2\}$，$V=\{v_k(X):k\in I_1\cup I_2\}$，$W=\{w_k(X):k\in I_1\cup I_2\}$。对于目标多项式，首先将其表示为$t(X)=t^{(1)}(X)\cdot t^{(2)}(X)$，随后，对于所有的电路线索引$\tilde{k}\in I_2\backslash I_1$，设置$Q_1$中的对应多项式为$u_{\tilde{k}}^{(1)}(X)=v_{\tilde{k}}^{(1)}(X)=w_{\tilde{k}}^{(1)}(X)=0$。同样的，对于$\tilde{k}\in I_1\backslash I_2$，设置$Q_2$中的对应多项式为$u_{\tilde{k}}^{(2)}(X)=v_{\tilde{k}}^{(2)}(X)=w_{\tilde{k}}^{(2)}(X)=0$。对于$k\in I_1\cup I_2$，只要目标多项式没有共同的根，就可以满足下列模的等价性设置：

$$\begin{aligned}u_k(X)&\equiv u_k^{(i)}(X)\bmod t^{(i)}(x)\\v_k(X)&\equiv v_k^{(i)}(X)\bmod t^{(i)}(x)\\w_k(X)&\equiv w_k^{(i)}(X)\bmod t^{(i)}(x)\end{aligned} \tag{21}$$

通过上述方法，即可通过单个乘法门的QRP构造出环多项式乘法对应的整个电路的QRP。

5.2 位分解QRP构造

在位分解电路中，假设有l层级的密文$CT=(d_0,d_1,d_2)$，对于输入$d_2\in R_q$，通过基为T的位分解操作得到d_2的位表示$d_{2,0},d_{2,1},\cdots,d_{2,\lceil\log_T q_l\rceil-1}$，使得$d_2=\sum_{i=0}^{\lceil\log_T q_l\rceil-1}d_{2,i}\cdot T^i$。首先，将输出线标记为$0,1,\cdots,\lceil\log_T q_l\rceil-1$，输入线标记为$\lceil\log_T q_l\rceil$。设：

$$t(X)=(X-r)\prod_{i=0}^{\lceil\log_T q_l\rceil-1}(X-r_i) \tag{22}$$

其中，$r,r_0,\cdots,r_{\lceil\log_T q_l\rceil-1}\in A$，$A$为特殊集。

对于$0\leqslant i\leqslant\lceil\log_T q_l\rceil-1$，设：

$$\begin{aligned}&u_0(r)=0,u_i(r)=T^i,u_{\lceil\log_T q_l\rceil}(r)=0\\&v_0(r)=1,v_i(r)=0,v_{\lceil\log_T q_l\rceil}(r)=0\\&w_0(r)=0,w_i(r)=0,w_{\lceil\log_T q_l\rceil}(r)=1\end{aligned} \tag{23}$$

对于$0\leqslant j\leqslant\lceil\log_T q_l\rceil-1$，设

$$\forall i\neq j,u_j(r_j)=1,u_i(r_j)=0 \tag{24}$$

$$\forall i\neq 0,i\neq j,v_0(r_j)=1,v_j(r_j)=-1,v_i(r_j)=0 \tag{25}$$

$$\forall i,w_i(r_j)=0 \tag{26}$$

若

$$\begin{aligned}&(u_0(X)+\textstyle\sum_{k=0}^{\lceil\log q_l\rceil-1}d_{2,k}\cdot u_k(X))\cdot\\&(v_0(X)+\textstyle\sum_{k=0}^{\lceil\log q_l\rceil-1}d_{2,k}\cdot v_k(X))-\\&(w_0(X)+\textstyle\sum_{k=0}^{\lceil\log q_l\rceil-1}d_{2,k}\cdot w_k(X))\end{aligned} \tag{27}$$

可以被$t(X)$整除，则该式在r点处的值一定为0，因此公式(23)可以保证$d_2=\sum_{i=0}^{\lceil\log_T q_l\rceil-1}d_{2,i}\cdot T^i$。公式(24)-(26)可以保证每个$r_i$都是多项式$t(X)$的一个根，使得$d_{2,j}(1-d_{2,j})=0$，因此保证$d_{2,0},d_{2,1},\cdots,d_{2,\lceil\log_T q_l\rceil-1}$是$d_2$的位分解。

6 VHENN方案设计

本章主要基于可验证BGV设计VHENN方案，首先为神经网络加密推理过程生成QRP关系；随后基于BGV实现神经网络加密推理，基于R-Groth16和可验证BGV实现证明的生成；最后是推理结果的验证与解密。

6.1 神经网络参数承诺与QRP构造

模型所有者首先对持有的模型进行承诺，并将承诺公开，并且在后续的计算中，将承诺的运算加入到证明的约束系统中，从而保证在后续的推理过程中，采用的模型始终为最初所承诺的模型。在基于环多项式的加密推理中，神经网络的前向传播计算为加密计算，其中的算术运算是由同态操作(如同态加法、同态乘法)实现的。每个同态乘法需要由若干复杂的环多项式操作表示。因此，首先需要将神经网络前向传播中的每个加法转换为多项式加法，乘法转换为由环多项式乘法、位分解等操作组成的子电路$C_\times$，生成对应的QRP。随后，将这些子电路的QRP按照5.1节所述的方法进行连接，最终生成整个电路的QRP。

在神经网络加密推理服务中，模型所有者为数据所有者提供推理服务，加密推理计算由模型所有者执行，输入的加密推理数据、模型的承诺值以及输出的加密推理结果对数据持有者(即验证者)是公开的，因此被指定为声明(Statement)。剩余部分包括模型权重、偏置、加密的中间值等信息，对数据所有者是保密的，因此被指定为见证(Witness)。最终，我们可以得到一个QRP关系，表示为：

$$Q=(R_q,l,\{u_k(X),v_k(X),w_k(X)\}_{k=0}^{m},t(X)) \quad (28)$$

该关系满足下列条件：

(1) $\{u_k(X),v_k(X),w_k(X)\}_{k=0}^{m}$是度为$n-1$的QRP多项式，其中$n$为约束数量。QRP多项式与神经网络模型的加密推理过程和模型参数的承诺计算过程相关联。

(2) $t(X)$是度为n的目标多项式。

(3) $x=(a_1,a_2,\cdots,a_l)$为声明，包括输入的加密数据$[\![d]\!]$，模型W的承诺值cm_W以及加密的推理结果$[\![y]\!]$。其中，$[\![\cdot]\!]$表示加密数据，$[\![d]\!]$，$[\![y]\!]\in R_i$，$i\in[0,L]$为密文的层级，承诺值$cm_W\in R_t$，R_t为明文空间。

(4) $w=(a_{l+1},a_{l+2},\cdots,a_m)$为见证，包括模型$W$的参数，如权重和偏置，以及在推理过程中生成的加密的中间结果，即每层网络的输出。其中，W的参数属于R_t，加密的中间结果属于R_t。

(5) 对于$(x,w)\in Q$，$p(X)=\sum_{k=0}^{m}a_ku_k(X)\cdot\sum_{k=0}^{m}a_kv_k(X)-\sum_{k=0}^{m}a_kw_k(X)$，其中$a_0=1$，$p(X)$被$t(X)$整除。

通过上述过程，可以得到一个满足相应声明和见证的QRP关系Q。模型所有者可以基于关系Q生成证明，以说服数据所有者，他拥有一个神经网络模型W，加密的推理结果$[\![y]\!]$是通过对数据所有者的加密数据$[\![d]\!]$在模型W上进行推理得到的。

上述QRP生成过程，假设神经网络推理加密计算的电路中仅包含乘法和加法的线性操作。但在实际应用中，神经网络计算还包括激活函数等非线性运算。本文主要针对卷积神经网络的计算进行描述，其中线性运算，包括卷积层、全连接层、平均池化层等可以直接转换为包括乘法门和加法门的算术电路，最后生成对应的QRP。下面主要对非线性运算的处理进行描述。

卷积神经网络模型中的非线性运算主要存在于激活函数中，主要的处理方式包括多项式近似、参数辅助和查找表三种。具体需要选择的处理方式需要与加密推理时采用的方法相对应。若加密推理阶段采用多项式近似方法，在QRP构造过程中也要选择对应的多项式进行构造，具体描述见6.2节。若加密推理阶段保持激活函数形式不变，则需要为相应的激活函数寻找适合的QRP表示或其他代数表达方式来进行构造。

对于ReLU等需要执行比较操作的激活函数，可以通过添加辅助参数，生成QRP关系。在推理过程中，除了输出ReLU的结果外，增加一个布尔值的输出用于表示所比较的参数与0的大小关系，从而将比较运算转换为算术运算。例如，假设密文$[\![x]\!]$为ReLU函数的输入，在推理过程中，若$x>0$，则输出$[\![x]\!]$及$tmp=1$。若$x\leqslant 0$，则输出$[\![0]\!]$及$tmp=0$。在构造QRP时，将运算表示为$[\![x']\!]=tmp\cdot[\![x]\!]$，即可以保证比较运算的正确性。该方式可以保证证明系统的高效运行，但引入额外的辅助参数的可能会带来一些安全问题。

目前，有研究尝试通过查找表的方式实现非线性操作的零知识证明[18-19, 23]。然而，这种方法主要针对交互式零知识证明协议，在针对算术电路和布尔电路的零知识证明协议之间进行转换。由于算术电路与布尔电路之间的非交互式零知识证明协议的转换仍处于研究的初期阶段，同时本文采用的环上算术电路的QRP代数表示形式也尚处于探索阶段，因此本方案暂未涉及对该实现方法的研究。

6.2　加密推理

在神经网络加密推理中，主要的运算包括密文加法、明文与密文乘法、密文与密文乘法。我们将遵循同态加密运算的协议表示为Π，则基于同态加密的加密推理方案可表示为$\Pi(f[\![d]\!],W)\rightarrow[\![y]\!]$，其中函数$f$表示前向传播算法，$[\![d]\!]$表示加密的数据，$W$表示模型参数，$[\![y]\!]$表示加密的推理结果。

在神经网络的线性计算层，如卷积层和全连接层，基础运算主要包括向量和矩阵点积，哈达玛(Hadamard)积等。这些基础运算均可以由加法和乘法实现。因此线性层的计算在加密推理中可以直接转换为同态加法和乘法操作。而对于神经网络的非线性计算，如激活函数的计算，则无法直接采用BGV同态加密进行实现。为了实现神经网络非线性层的计算，可以采用两种方式。第一种是多项式近似方法，将非线性激活函数近似为线性的多项式。第二种保持激活函数形式不变，通过在非线性计算环节转换为安全两方计算方案，采用混淆电路、布尔共享等方式实现安全计算。

6.2.1　多项式近似

多项式近似方法是基于同态加密的神经网络安全推理方案中最常用的非线性运算处理方式，也是本文主要采用的方式。在多项式近似方法中，采用高阶多项式近似可以最大程度上降低近似误差，从

而降低模型准确率下降的问题。但由于同态加密的计算效率较低，且存在级数限制，因此阶数越高的多项式计算效率越低，需要在准确率和效率之间进行权衡。对于Softmax等平滑函数，逼近误差通常较小，可以很容易地采用泰勒级数等方法得到近似多项式。对于ReLU等含有比较操作的激活函数，本文采用最小最大逼近(Minimax Approximation)的多项式组合方法[51]。该方法以符号函数$\mathrm{sgn}(x)$作为其他分段比较函数的基础，并且相较于直接逼近ReLU函数，在精度和计算复杂度之间可以达到更好的平衡。

首先寻找符号函数：

$$\mathrm{sgn}(x)=\begin{cases}1, & \text{if } x>0\\ 0, & \text{if } x=0\\ -1, & \text{if } x<0\end{cases} \tag{29}$$

的最佳多项式阶数组合。其目标是找到一组多项式阶数组合$\{d_1,d_2,\cdots,d_k\}$使得组合多项式$p_k(p_{k-1}(\cdots p_1(x)))$：

(1)满足给定的逼近误差τ，使得$\|p(x)-\mathrm{sgn}(x)\|_\infty\leqslant\tau$。

(2) 最小化同态加密中的非标量乘法次数(计算代价)。

(3) 最小化同态加密的级数消耗。

通过切比雪夫(Chebyshev)多项式作为基础函数来构造每个多项式$p_i(x)$，并通过动态规划找到满足上述优化目标的多项式阶数组合。记$\mathrm{sgn}(x)\approx p_k(p_{k-1}(\cdots p_1(x)))$，ReLU函数可以表示为

$$\mathrm{ReLU}(x)=\frac{x+x\cdot\mathrm{sgn}(x)}{2} \tag{30}$$

从而得到满足同态加密计算效率需求，且误差较小的ReLU函数的多项式近似表示。随后采用同态加密对该多项式进行计算。

6.2.2 方案转换

另外一种可行的方法是在非线性计算部分转换安全计算方案。由于本方案研究集中于可验证同态加密方案，因此主要采用多项式近似方法，这里仅对方案转换的方法进行简单介绍。假设通过线性层的计算得到了一个加密的中间值$[\![z]\!]$，服务器首先选择一个随机数r，使用同态加密的公钥pk加密随机数r，得到$[\![r]\!]$。随后，服务器计算$[\![z]\!]-[\![r]\!]$并将结果发送给数据所有者。数据所有者解密$[\![z]\!]-[\![r]\!]$得到明文$z\text{-}r$，可以将$z\text{-}r$和r视为z的两个加法秘密份额。因此数据所有者和模型所有者可以基于混淆电路或布尔共享交互完成ReLU函数的比较操作。线性计算与非线性计算交替进行，最终完成神经网络推理过程。

6.3 证明生成

模型所有者在执行加密推理之外，还需要生成加密推理的零知识证明。利用5.1节中生成的QRP关系，记为$Q=(R_q,l,\{u_k(X),v_k(X),w_k(X)\}_{k=0}^{m},t(X))$。假设电路C具有$m$个线和$n$个乘法门，使$I_s=1,2,\cdots,l$对应满足该QRP关系的声明$x=(a_1,a_2,\cdots,a_l)$，包括推理数据$[\![d]\!]$，模型承诺，以及加密的输出结果$[\![y]\!]$；$I_w=l+1,l+2,\cdots,m$对应满足该QRP关系的见证$w=(a_{l+1},a_{l+2},\cdots,a_m)$，包括模型参数W和电路的中间值。因此，对于加密推理计算$\Pi(f)$，其证明需要满足$Q(x,w)=1$，即满足QRP关系。在证明生成和验证中，模型所有者持有声明$x=(a_1,a_2,\cdots,a_l)$与见证$w=(a_{l+1},a_{l+2},\cdots,a_m)$，其负责生成对应QRP关系$Q$明$x=(a_1,a_2,\cdots,a_l)$，负责对生成的证明$\pi$进行验证。

6.4 验证与解密

数据所有者收到加密的推理结果$[\![y]\!]$及证明π后，首先执行$Verify(Q,vk,x,\pi)\to\{0,1\}$，其中声明$x=(a_1,a_2,\cdots,a_l)$，包括推理数据$[\![d]\!]$，模型承诺，以及加密的输出结果$[\![y]\!]$。若验证成功，则返回1。随后解密$[\![y]\!]$得到明文推理结果$y$。若验证不成功，则意味着模型所有者没有遵守约定的计算协议，或者在推理过程中使用了不正确的模型参数。

7 安全性分析

由于VHENN在加密推理阶段的安全性完全依赖于同态加密方案的安全性，因此在本章不再论述。在加密推理证明阶段，VHENN的安全性依赖于R-Groth16的安全性。由于在Rinocchio[25]中仅提供R-Groth16的构造作为对QRP应用的示例，并未给出形式化的安全性分析，因此本章给出了基于R-Groth16的加密推理证明生成协议Γ：$(Setup,Prove,Verify)$在完备性(Completeness)、知识合理性(Knowledge Soundness)和零知识性(Zero-Knowledge)三个方面的安全性分析。

定理1. 协议Γ是完备的，如果模型所有者为$(x,w)\in Q$生成了一个证明π，该证明能够以概率Pr

说服一个诚实的验证者：

$$Pr\begin{bmatrix}(CRS,vk)\leftarrow Setup(1^{\lambda},Q)\\ \pi\leftarrow Prove(Q,CRS,x,w):\\ 1\leftarrow Verify(Q,vk,x,\pi)\end{bmatrix}\geqslant 1-negl(\lambda)\tag{31}$$

证明. 对于$(x,w)\in Q$的证明$\pi=(A,B,C)$，只有满足等式$AB=E(\alpha)E(\beta)+\gamma E(f_s)+\delta C$，其中

$$f_s=\frac{(\beta u_s(\varepsilon)+\alpha v_s(\varepsilon)+w_s(\varepsilon))}{\gamma}\tag{32}$$

该证明才会被接受。设

$$u(\varepsilon)=\sum_{k=0}^{m}a_k u_k(\varepsilon)=u_s(\varepsilon)+u_w(\varepsilon)\tag{33}$$

$$v(\varepsilon)=\sum_{k=0}^{m}a_k v_k(\varepsilon)=v_s(\varepsilon)+v_w(\varepsilon)\tag{34}$$

$$w(\varepsilon)=\sum_{k=0}^{n}a_k w_k(\varepsilon)=w_s(\varepsilon)+w_w(\varepsilon)\tag{35}$$

由于Encode算法$E()$具有加法同态性，上述等式左侧和右侧的计算分别如下：

$$\begin{aligned}A\cdot B&=E(\alpha+u(\varepsilon))\cdot E(\beta+v(\varepsilon))=\\&\big(E(\alpha)+E(u(\varepsilon))\big)\cdot\big(E(\beta)+E(v(\varepsilon))\big)=\\&E(u(\varepsilon))E(v(\varepsilon))+E(\alpha)E(\beta)+\\&E(\beta)E(u(\varepsilon))+E(\alpha)E(v(\varepsilon))=\\&\boldsymbol{E(u(\varepsilon))E(v(\varepsilon))}+Z\end{aligned}$$

$$\begin{aligned}&E(\alpha)E(\beta)+\gamma E(f_{io})+\delta C=\\&E(\alpha)E(\beta)+\gamma E\left(\frac{\beta u_s(\varepsilon)+\alpha v_s(\varepsilon)+w_s(\varepsilon)}{\gamma}\right)+\delta C=\\&E(\alpha)E(\beta)+E(\beta u(\varepsilon)+\alpha v(\varepsilon)+\\&w(\varepsilon))+E(h(\varepsilon)t(\varepsilon))=\\&\boldsymbol{E(w(\varepsilon))+E(h(\varepsilon)t(\varepsilon))}+Z\end{aligned}$$

因此，如果验证者在验证阶段对一个正确的证明π输出0，则意味着$E(u(\varepsilon))E(v(\varepsilon))\neq E(w(\varepsilon))+E(h(\varepsilon)t(\varepsilon))$，也就是$u(\varepsilon)v(\varepsilon)\neq w(\varepsilon)+h(\varepsilon)t(\varepsilon)$，与QRP关系$Q$条件不符。因此，协议$\Gamma$：$(Setup,Prove,Verify)$是完备的。

定理2. 协议Γ是知识合理的，如果对于计算受限的敌手$\mathcal{A}$，和一个计算受限且可以完全访问敌手$\mathcal{A}$状态的提取器E，满足：

$$Pr\begin{bmatrix}(CRS,vk)\leftarrow Setup(1^{\lambda},Q)\\ (x,\pi',w)\leftarrow(\mathrm{A}\,|\,E)(Q,CRS):\\ (x,w)\notin Q\wedge\\ 1\leftarrow Verify(Q,vk,x,\pi')\end{bmatrix}\leqslant negl(\lambda)\tag{36}$$

证明. 假设有一个模拟器可以生成一系列系数$A_\alpha,A_\beta,A_\gamma,A_\delta,\{A_k\}_{k=0}^{m}$和两个多项式$A(x),A_h(x)$，则证明$\pi$中的元素$A,B,C$对应的$A_u$可以表示为：

$$\begin{aligned}A_u=&A_\alpha\alpha+A_\beta\beta+A_\gamma\gamma+A(\varepsilon)+\\&\sum_{k=0}^{l}A_k\frac{\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)}{\gamma}+\\&\sum_{k=l+1}^{m}A_k\frac{\beta u_k(\varepsilon)+\alpha v_k(\varepsilon)+w_k(\varepsilon)}{\delta}+\\&A_h(\varepsilon)\frac{t(\varepsilon)}{\delta}\end{aligned}$$

B_v、C_w也可以以相同的方式构造，该构造方式使得A_u、B_v、C_w包含了验证阶段的验证等式中包含的所有项。

对于验证阶段的等式$AB=E(\alpha)E(\beta)+\gamma F+\delta C$，可以将其视为一个多元Laurent多项式的一个等式。根据Rinocchio中定义的环上的Laurent多项式Schwartz－Zippel引理，敌手$\mathcal{A}$可以以忽略不计的优势构造出满足等式的A_u、B_v、C_w，且提取器E无法从A_u、B_v、C_w计算出一个见证w'使得$(x,w')\in Q$。因此概率Pr可以忽略不计，协议Γ是知识合理的。

定理3. 协议Γ是零知识的，如果一个计算受限的敌手$\mathcal{A}$无法从诚实的模型所有者生成的证明中得到任何关于见证的信息。正式的，对于所有$(x,w)\in Q$和一个计算受限的敌手$\mathcal{A}$，满足：

$$Pr\begin{bmatrix}(CRS,vk)\leftarrow Setup(1^{\lambda},Q)\\ \pi\leftarrow Prove(Q,CRS,x,w):\\ 1\leftarrow\mathcal{A}(Q,vk,\pi)\end{bmatrix}=Pr\begin{bmatrix}(CRS,vk)\leftarrow Setup(1^{\lambda},Q)\\ \pi'\leftarrow Sim(Q,vk,x):\\ 1\leftarrow\mathcal{A}(Q,vk,\pi)\end{bmatrix}\tag{37}$$

证明. 假设存在一个模拟器，可以利用陷门$vk=(sk,CRS,\varepsilon,\gamma,\delta)$执行$\pi\leftarrow Sim(Q,vk,x)$，以获得模拟证明$\pi'$，计算过程如下：

模拟器$\mathcal{B}$随机选择$A_u,B_v\in R_q$，且有$A=E(A_u)$，$B=E(B_v)$，随后计算：

$$C=E(C_w)=\frac{A\cdot B-E(\alpha)E(\beta)-E(\beta u_s(\varepsilon)+\alpha v_s(\varepsilon)+w_s(\varepsilon))}{\delta}$$

因此，模拟器$\mathcal{B}$可以利用陷门生成模拟证明$\pi'=(A,B,C)$，使得验证等式$AB=E(\alpha)E(\beta)+\gamma F+\delta C$满足，该证明同样可以被诚实的验证者所接受。由正确的见证生成的证明π与模拟证明π'对

于验证者是不可区分的，而模拟证明π'中，不包含任何与见证相关的信息。因此，如果敌手可以从证明π中提取到见证的信息，则无法满足π与π'不可区分的特性。因此，计算受限的敌手$\mathcal{A}$无法从一个证明中获得任何关于见证份额的信息，协议Γ满足零知识性。

8 实验评估

8.1 实验设置

实验配置：本文实验是在一台配备了英特尔酷睿(TM)i7-10700k@3.8 GHz CPU、128 GB内存的Ubuntu 20.04操作系统电脑上进行的。本文实验采用C++语言，零知识证明基于libsnark[①]和ringSNARK[②]实现，同态加密基于SEAL库[③]实现。同态加密采用SIMD操作，可并行处理同态操作。

本文实验首先测试了环上基础运算的证明生成与验证开销，包括环上乘法、同态加密密文乘法，同态加密密文与明文乘法、位分解等。随后在三个具有不同规模的分类任务的数据集和不同模型上测试加密推理的证明与验证时间。采用的数据集包括：WINE(Wine Data Set)、MNIST(Modified National Institute of Standards and Technology database)、MedMNIST、CIFAR10 (Canadian Institute for Advanced Research, 10 classes)。其中，WINE为结构化数据集，用于小规模、高维特征的分类问题，典型代表传统机器学习任务。MNIST、MedMNIST与CIFAR10为图像数据集，MNIST为二维数据集被广泛用于深度学习模型的基线评估，CIFAR10为三维数据集，相比于MNIST更复杂。MedMNIST为医学领域的数据集，可以展示本方案在潜在的应用场景中的应用潜力。相应的，在上述数据集上采用的模型包括：全连接神经网络模型(ShallowNet)、卷积神经网络模型(LeNet)，测试具有不同数量级参数和计算量的模型推理性能。另外，由于本文方案采用了同态加密技术，其计算深度受到加密噪声增长的限制，且计算效率较低。这些因素使得基于同态加密的神经网络推理方案目前在深层次神经网络中的应用面临较大的性能瓶颈[8, 38]。为了在现有技术和资源条件下展示本文方案的实际效果，实验评估主要针对较为简单的浅层全连接和卷积神经网络和小型数据集进行验证。数据集和模型架构详情见表1和表2。

表1 数据集详情

数据集	类别	特征数	样本数
WINE	3	13	178
MNIST	10	28*28*1	70 000
MedMNIST	11	28*28*1	58,830
CIFAR10	10	32*32*3	60 000

表2 模型架构

数据集模型	WINE	MNIST	MedMNIST	CIFAR10
卷积层1	/	/	卷积核5×5 通道数4	卷积核5×5 通道数6
池化层	/	/	卷积核2×2	卷积核2×2
卷积层2	/	/	卷积核5×5 通道数12	卷积核5×5 通道数16
池化层	/	/	卷积核2×2	卷积核2×2
卷积层3	/	/	卷积核4×4 通道数64	卷积核4×4 通道数120
全连接层1	13×32	784×128	256×84	480×84
全连接层2	32×3	128×10	84×11	84×10

8.2 实验结果

8.2.1 环上基础运算开销

本节测试了环上基础运算的可信设置、证明生成、验证的计算开销，以及约束数量。可信设置由可信第三方执行，相同电路的多次运算仅需要一次可信设置。如表3所示，展示了环上乘法(明文乘法(Mul-PP))、同态加密密文与明文乘法(Mul-CP)、同态加密密文乘法(不包含位分解)(Mul-CC)、位分解(Decom)等计算的开销。由于约束系统是由乘法电路转换而来，加法电路无需额外添加约束，因此本节未给出对加法运算的开销展示。位分解为BGV同态加密执行密文乘法后，密钥交换所需要的基础操作。实验结果表明，随着约束数量的增长，可信设置、证明生成和验证的计算开销也在逐渐增长。因此在神经网络的实际应用中，计算复杂度会随着输入规模和模型复杂度的增加而增加。降低计算开销的重点在于减少相关高开销基础操作数量。

表3 基础操作计算开销

操作	Mul-PP	Mul-CP	Mul-CC	Decom
可信设置(s)	0.157	0.167	0.176	0.590
证明生成(s)	0.027	0.035	0.095	1.610
验证(s)	0.039	0.039	0.041	0.360
约束数量	1	2	4	34

① libsnark: a C++ library for zkSNARK proofs, https://github.com/scipr-lab/libsnark

② ringSNARK - A library for zkSNARKs over rings, https://github.com/zkFHE/ringSNARK/tree/main

③ Microsoft SEAL, https://github.com/microsoft/SEAL

8.2.2　神经网络证明与验证开销

本节测试了四个数据集在不同神经网络模型下，实现加密推理可验证所需的各阶段操作的开销，包括可信设置、证明生成及验证。对于相同网络结构的推理验证，可信设置仅需执行一次。证明生成过程由模型所有者完成，变量包括加密后的数据、权重、计算中间值及推理结果等，对于验证者，权重和中间值为秘密值，加密数据和推理结果为公开值。验证过程由验证者完成。如表4所示，随着模型和数据集的复杂度增长，可信设置、证明生成、验证时间及约束数量均随之增长。其中，主要开销为证明生成过程。造成开销增长的主要原因在于，本文方案所基于的R-Groth16方案，可信设置与电路直接相关，每个电路都需要进行一次可信设置。较为复杂的神经网络模型与数据集会导致推理计算的电路复杂度增加，电路转换为QRP时约束数量也相应的增长。因此可信设置、证明生成、验证及约束数量都对应增加。若想降低计算开销，主要可以从两方面入手，一方面优化约束构造方式，降低约束数量。另一方面降低约束的计算时间，从而提升整体效率。本文实验主要通过采用SIMD操作并行执行推理任务，从而降低推理计算的约束数量。

表4　神经网络推理验证的计算开销

模型	ShallowNet-WINE	ShallowNet-MNIST	LeNet-MEDMNIST	LeNet-Cifar10
可信设置(s)	40	563	916	1743
证明生成(s)	358	5359	8755	16 501
验证(s)	129	1950	3134	5125
约束数量	646	2580	3302	4540

8.2.3　约束数量及对比

本节对采用SIMD技术对加密推理过程进行优化后构造约束系统的约束数量，以及未采用SIMD技术，对直接进行加密推理计算的电路构造约束系统的约束数量进行了对比。如表5所示，采用SIMD操作后单个推理任务的约束数量大大减少。主要原因在于对密文乘法等同态操作的并行运算使得推理计算电路门数量大大降低，例如，在全连接层，假设输入数据的维度为m，神经元数量为n，即该数据向量需要与n个m维的权重向量进行点积运算，每个点积运算需要$m\times m$次乘法，因此每个全连接层在明文状态下进行计算大约会产生$n\times m\times m$个约束。当点积运算中包含密文时，密文与明文或密文乘法操作以及带来的额外位分解等操作会导致更多的约束数量。在SIMD操作下执行加密推理运算时，一个或多个m维数据（根据设定的参数，若数据维度较大，也可以编码一个数据中的部分元素）可以被编码并加密到同一个密文中，同样的权重向量也可以被编码到同一个明文中，这样多个乘法约束被压缩为1个乘法约束，从而大大降低约束数量。为了使对比更清晰，本文选择一次编码单个数据。在实际操作中，对于WINE、MNIST等小型数据集，每次可以编码多个数据进行并行处理，从而得每个推理任务会具有更低的平均约束数量。

表5　约束数量

模型	ShallowNet-WINE	ShallowNet-MNIST	LeNet-MEDMNIST	LeNet-Cifar10
SIMD	646	2580	3302	4540
无SIMD	1609	205 598	1 219 650	1 764 610

8.2.4　存储要求

本节对方案的内存占用、证明密钥、验证密钥及证明大小进行了测试。如表6所示，随着模型复杂度的增加，内存占用与证明密钥的大小也随之增加，而验证密钥与证明大小保持不变。本文方案的内存占用主要是在可信设置与证明生成过程，用户（即验证者）无需承担该过程的计算，因此仍然可以满足资源受限的用户的需求。证明密钥主要用于证明生成过程，该过程由模型所有者执行，因此该过程也不涉及用户。验证密钥与证明大小保持不变，且只需要较小的存储空间，因此即便发送给用户执行验证过程，也不需要对用户的资源做过高的要求。

表6　内存占用及CRS和证明大小

模型	ShallowNet-WINE	ShallowNet-MNIST	LeNet-MEDMNIST	LeNet-Cifar10
内存占用(GB)	5.27	19.44	24.83	35.16
CRS(MB)	1457.25	5349.82	6887.27	9368.81
验证密钥(MB)	0.48	0.48	0.48	0.48
证明大小(B)	2.04	2.04	2.04	2.04

8.2.5　准确率对比

本节对比分析了采用ReLU激活函数的原始神经网络推理方案与本文基于多项式近似的加密神经网络推理方案的准确率。我们对ReLU不同阶数的近似多项式进行了测试，并选择了9阶和5阶多项式进行对比。其中9阶多项式需要5次密文乘法，消耗5层密文深度。5阶多项式需要3次密文乘法，消耗3层密文深度。如表7所示，相较于采用ReLU激活函数的原始神经网络推理准确率，本文方案采用

表7 准确率对比

模型	ShallowNet-WINE	ShallowNet-MNIST	LeNet-MEDMNIST	LeNet-Cifar10
ReLU	100%	97.96%	82.42%	63.44%
9阶多项式	100%	97.32%	81.18%	63.42%
5阶多项式	100%	95.61%	77.88%	61.89%

9阶多项式的加密推理准确率下降幅度整体在1.3%以内，并未对准确率造成过大的影响。采用5阶多项式时，准确率降低幅度相比于9阶多项式较大，但采用5阶多项式计算开销与消耗深度更低，在实际应用中可以根据需求进行选择。

8.2.6 方案对比

本节将VHENN与ZKML[23]和pvCNN[26]进行了对比。ZKML基于一种ZK-SNARK协议Halo2设计了可以用于大型机器学习模型的可验证框架，提出了一种低级操作的高效约束gadgets，并开发了一个优化器用于高效布置电路中的gadgets。ZKML采用的Halo2协议相比于Groth16和R-Groth16无需可信设置。但是ZKML没有关注推理数据和推理结果的隐私保护问题。pvCNN与本文方案更为接近，同样采用同态加密和zk-SNARKs实现神经网络推理中的隐私保护和可验证性。然而，该方案并未将同态加密和zk-SNARKs进行结合，而是将模型拆分为PriorNet和LaterNet，其中PriorNet保持私有，LaterNet被设定为非隐私部分，委托给服务器进行计算。在PriorNet部分采用同态加密保护数据隐私，在LaterNet部分采用基于二次矩阵程序的LegoSNARK协议实现对服务器运算结果的可验证性。由于pvCNN并非在整个神经网络模型上采用zk-SNARKs，为了保持相同的对比基准，本节给出了输入维度为32*32*3，卷积核为5×5，通道数为6的一层卷积层中，执行前向传播时，构造证明系统所需的(可信)设置(ZKML无需可信设置)、证明生成、验证所需的时间，内存占用，CRS大小以及证明大小等方面的对比。如表8所示，得益于同态加密可以采用SIMD的特性，本方案可以同时处理多个数据，使得对于单个输入的前向传播电路构造的约束数量大大降低，因此在各方面的表现均优于pvCNN，在大多数性能的表现中优于ZKML。此外，pvCNN所需的验证时间过长，对验证者的要求过高，并不适用于神经网络推理服务的场景。更重要的，本方案结合了同态加密与zk-SNARKs，相较于pvCNN的隐私保护和可验证分别针对部分模型，ZKML仅关注可验证，本方案可以同时实现神经网络推理的隐私保护与可验证。

表8 与pvCNN的对比

方案	ZKML	pvCNN	VHENN
可信设置(s)	14.58	4650.41	0.35
证明生成(s)	51.05	3997.82	0.56
验证(s)	0.02	123 963.61	0.15
内存占用	1.12 GB	17.30 GB	0.63 GB
CRS大小	600.14 MB	14.29 GB	44.18 MB
证明大小	9.06 KB	4.76 GB	491.52 KB

9 总结与展望

本文提出了一种可验证的神经网络加密推理方案-VHENN。首先结合环上zk-SNARKs协议与基于环多项式的同态加密方案-BGV，实现了可验证同态加密方案。随后将可验证同态加密方案与基于BGV的神经网络加密推理方案相结合，实现了满足模型、推理数据、推理结果隐私保护以及模型真实性和推理正确性可验证的神经网络推理方案。最后，本文进行了实验评估，以展示VHENN的性能，并与相关方案进行了对比分析。

本方案首次将可验证同态加密方案应用于神经网络推理中，实现可验证的神经网络加密推理方案。然而，受限于同态加密的效率，当前对神经网络加密推理的研究与应用发展较慢，结合零知识证明会进一步地降低整体方案的效率，且需要较大的内存开销，因此仍然需要进一步的研究与优化。未来研究重点计划集中于两个方面：一、研究如何提高同态加密的计算效率以及优化同态加密与零知识证明结合的方法，使得可验证神经网络加密推理的效率进一步提高，使其能够在实际应用中部署。二、研究QRP的代数表示形式与针对布尔电路的代数表示形式(如二次扩展程序(Quadratic Span Programs，QSP))之间的转换方法，以避免当前采用多项式近似处理非线性操作而导致的准确率下降、适用范围受限等问题。

参 考 文 献

[1] Liu Jun-Xu, Meng Xiao-Feng. A review of privacy protection research in machine learning. Journal of Computer Research and Development, 2020, 57(2): 346-362 (in Chinese)
(刘俊旭，孟小峰. 机器学习的隐私保护研究综述. 计算机研究与发展, 2020, 57(2): 346-362)

[2] Ng L K L, Chow S S M. SoK: Cryptographic neural-network

computation//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2023:497-514

[3] Xing Z, Zhang Z, Liu J, et al. Zero-knowledge proof meets machine learning in verifiability: A survey. arXiv preprint arXiv: 231014848, 2023: 1-23

[4] Mohassel P, Zhang Y. Secureml: A system for scalable privacy-preserving machine learning//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017:19-38

[5] Liu Wei-Xin, Guan Ye-Wei, Huo Jia-Rong, et al. A fast and secure transformer inference scheme with secure multi-party computation. Journal of Computer Research and Development, 2024, 61(5): 1218-1229 (in Chinese)
(刘伟欣，管晔玮，霍嘉荣，等. 一种基于安全多方计算的快速Transformer安全推理方案. 计算机研究与发展，2024, 61(5): 1218-1229)

[6] Guo Juan-Juan, Wang Qiong-Xiao, Xu Xin, et al. Secure multiparty computation and application in machine learning. Journal of Computer Research and Development, 2021, 58(10): 2163-2186 (in Chinese)
(郭娟娟，王琼霄，许新，等. 安全多方计算及其在机器学习中的应用. 计算机研究与发展，2021, 58(10): 2163-2186)

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

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

[9] Ren Yan-Li, Yu Ling-Zan, He Gang, et al. A scheme of privacy-preserving convolutional neural network prediction. Journal of Computer Research and Development, 2023, 46(08): 1606-1619 (in Chinese)
(任艳丽，余凌赞，何港，等. 一种隐私保护的卷积神经网络预测方案. 计算机学报，2023, 46(08): 1606-1619)

[10] Chaudhari H, Rachuri R, Suresh A. Trident: Efficient 4pc framework for privacy preserving machine learning//Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2020. DiegoSan, USA, 2019:1-18

[11] Wagh S, Tople S, Benhamouda F, et al. Falcon: Honest-majority maliciously secure framework for private deep learning//Proceedings of the Proceedings on Privacy Enhancing Technologies. On the Internet, 2021:188-208

[12] Dalskov A, Escudero D, Keller M. Fantastic four: Honest-majority four-party secure computation with malicious security//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21. Virtual, 2021:2183-2200

[13] Yuan B, Yang S, Zhang Y, et al. MD-ML: Super fast privacy-preserving machine learning for malicious security with a dishonest majority//Proceedings of the 33rd USENIX Security Symposium. Philadelphia, USA, 2024:2227-2244

[14] Xu G, Li H, Ren H, et al. Secure and verifiable inference in deep neural networks//Proceedings of the Annual Computer Security Applications Conference. Austin USA, 2020:784-797

[15] Dong C, Weng J, Liu J-N, et al. Fusion: Efficient and secure inference resilient to malicious servers//Proceedings of the Network and Distributed System Security (NDSS) Symposium 2023. DiegoSan, USA, 2023:1-18

[16] Ghodsi Z, Gu T, Garg S. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud//Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017. LongBeach, USA, 2017: 1-10

[17] Liu T, Xie X, Zhang Y. ZkCNN: Zero knowledge proofs for convolutional neural network predictions and accuracy//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual, Republic of Korea, 2021: 2968-2985

[18] Weng C, Yang K, Xie X, et al. Mystique: Efficient conversions for zero-knowledge proofs with applications to machine learning//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21. Virtual, 2021:501-518

[19] Hao M, Chen H, Li H, et al. Scalable zero-knowledge proofs for non-linear functions in machine learning//Proceedings of the 33rd USENIX Security Symposium. Philadelphia, USA, 2024: 3819-3836

[20] Lee S, Ko H, Kim J, et al. vCNN: Verifiable convolutional neural network based on zk-snarks. IEEE Transactions on Dependable and Secure Computing, 2024, 21(4): 4254 - 4270

[21] Feng B, Qin L, Zhang Z, et al. ZEN: An optimizing compiler for verifiable, zero-knowledge neural network inferences. Cryptology ePrint Archive, 2021: 1-25

[22] Zhao L, Wang Q, Wang et al. VeriML: Enabling integrity assurances and fair payments for machine learning as a service. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(10): 2524-2540

[23] Chen B-J, Waiwitlikhit S, Stoica I, et al. ZKML: An optimizing system for ML inference in zero-knowledge proofs//Proceedings of the 19th European Conference on Computer Systems. Athens, Greece, 2024:560-574

[24] Fan Y, Xu B, Zhang L, et al. psvCNN: A zero-knowledge CNN prediction integrity verification strategy. IEEE Transactions on Cloud Computing, 2024, 12(2): 359-369

[25] Ganesh C, Nitulescu A, Soria-Vazquez E. Rinocchio: SNARKs for ring arithmetic. Journal of Cryptology, 2023, 36(4): 1-50

[26] Weng J, Weng J, Tang G, et al. pvCNN: Privacy-preserving and verifiable convolutional neural network testing. IEEE Transactions on Information Forensics and Security, 2023, 18: 2218-2233

[27] Han Wei-Li, Song Lu-Bin, Ruan Wen-Qiang, et al. Secure multi-party learning: From secure computation to secure learning. Chinese Journal of Computers, 2023, 46(7):1494-1512 (in Chinese)
(韩伟力，宋鲁杉，阮雯强，等. 安全多方学习:从安全计算到安全学习. 计算机学报，2023, 46(7):1494-1512)

[28] Hu Ao-Ting, Hu Ai-Qun, HU Yun, et al. Differentially private data sharing and publishing in machine learning: Techniques, applications, and challenges. Journal of Cyber Security, 2022, 7(4):1-16 (in Chinese)

(胡奥婷，胡爱群，胡韵，等. 机器学习中差分隐私的数据共享及发布:技术，应用和挑战. 信息安全学报，2022，7(4):1-16)

[29] Ma Jun-Ming, Wu Bing-Zhe, Yu Chao-Fan, et al. S3ML: Secure serving system for machine learning inference. Journal of Software, 2022, 33(9):3312－3330(in Chinese)
(马俊明，吴秉哲，余超凡，等. S3ML:一种安全的机器学习推理服务系统. 软件学报，2022，33(9): 3312－3330)

[30] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada, 2018:35-52

[31] Chandran N, Gupta D, Obbattu S L B, et al. SIMC: ML inference secure against malicious clients at semi-honest cost//Proceedings of the 31st USENIX Security Symposium (USENIX Security 22. Boston, USA, 2022:1361-1378

[32] Koti N, Patra A, Rachuri R, et al. Tetrad: Actively secure 4pc for secure training and inference//Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2022. DiegoSan, CA, USA, 2021:1-18

[33] Hou X, Liu J, Li J, et al. Ciphergpt: Secure two-party gpt inference. Cryptology ePrint Archive, 2023: 1-16

[34] Dong Y, Lu W-j, Zheng Y, et al. Puma: Secure inference of llama-7b in five minutes. arXiv preprint arXiv: 230712533, 2023: 1-13

[35] Gupta K, Jawalkar N, Mukherjee A, et al. Sigma: Secure gpt inference with function secret sharing//Proceedings of the Privacy Enhancing technologies Symposium (PETS) 2024. Bristol, UK, 2024:1-19

[36] Hesamifard E, Takabi H, Ghasemi M. CryptoDL: Deep neural networks over encrypted data. arXiv preprint arXiv: 171105189, 2017: 1-21

[37] Juvekar C, Vaikuntanathan V, Chandrakasan A. GAZELLE: A low latency framework for secure neural network inference//Proceedings of the 27th USENIX Security Symposium (USENIX Security 18. Baltimore, USA, 2018:1651-1669

[38] Dathathri R, Saarikivi O, Chen H, et al. CHET: An optimizing compiler for fully-homomorphic neural-network inferencing//Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. Phoenix USA, 2019: 142-156

[39] Zhang Q, Xin C, Wu H. GALA: Greedy computation for linear algebra in privacy-preserved neural networks//Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2021. Virtual, 2021:1-16

[40] Lu W-j, Huang Z, Hong C, et al. PEGASUS: Bridging polynomial and non-polynomial evaluations in homomorphic encryption//Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2021: 1057-1073

[41] Niu C, Wu F, Tang S, et al. Toward verifiable and privacy preserving machine learning prediction. IEEE Transactions on Dependable and Secure Computing, 2020, 19(3): 1703-1721

[42] Li X, He J, Vijayakumar P, Zhang X, et al. A verifiable privacy-preserving machine learning prediction scheme for edge-enhanced HCPSs. IEEE Transactions on Industrial Informatics, 2021, 18(8): 5494-5503

[43] Chatel S, Knabenhans C, Pyrgelis A, et al. Verifiable encodings for secure homomorphic analytics. arXiv preprint arXiv:220714071, 2022: 1-26

[44] Natarajan D, Loveless A, Dai W, et al. Chex-mix: Combining homomorphic encryption with trusted execution environments for two-party oblivious inference in the cloud. Cryptology ePrint Archive, 2021: 1-19

[45] Gong B, Lau W F, Au M H, et al. Efficient zero-knowledge arguments For paillier cryptosystem//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA, 2024:96-96

[46] Fiore D, Nitulescu A, Pointcheval D. Boosting verifiable computation on encrypted data//Proceedings of the Public-Key Cryptography – PKC 2020: 23rd IACR International Conference on Practice and Theory of Public-Key Cryptography. Edinburgh, UK, 2020:124-154

[47] Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory (TOCT), 2014, 6(3): 1-36

[48] Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers//Proceedings of the Advances in Cryptology-ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security. Hong Kong, China, 2017:409-437

[49] Zhang Z, Plantard T, Susilo W. Reaction attack on outsourced computing with fully homomorphic encryption schemes//Proceedings of the Information Security and Cryptology-ICISC 2011: 14th International Conference. Seoul, Republic of Korea, 2012: 419-436

[50] Chaturvedi B, Chakraborty A, Chatterjee A, et al. A practical full key recovery attack on tfhe and fhew by inducing decryption errors. Cryptology ePrint Archive, 2022: 1-18

[51] Lee E, Lee J W, No J S, et al. Minimax approximation of sign function by composite polynomial for homomorphic comparison. IEEE Transactions on Dependable and Secure Computing, 2022, 19(6): 3711-3727

YANG Wen-Ti, Ph. D. candidate. Her research interests include privacy-preserving machine learning, privacy computing.

HE Zhao-Yang, M. S. candidate. His research interest is privacy-preserving machine learning.

LI Meng, Ph. D., associate professor. His research interests include data security, privacy preservation, applied cryptography, blockchain, TEE, and the Internet of Vehicles.

ZHANG Zi-Jian, Ph. D., professor. His research interests include design of authentication and key agreement protocol and analysis of entity behavior and preference.

GUAN Zhi-Tao, Ph. D., professor. His research interests include AI security, privacy computing, system security, blockchain and applied cryptography.

ZHU Lie-Huang, Ph. D., professor. His research interests include cryptography, network and information security.

Background

This paper addresses a key issue in the field of privacy-preserving neural network inference, specifically focusing on how to protect the confidentiality of user data, models and inference results while ensuring the verifiability of inference results. With the increasing adoption of AI services such as ChatGPT and ERNIE Bot, privacy concerns have become a critical focus, especially for small and medium-sized enterprises and individual users. These services require secure inference processes that not only protect sensitive data but also guarantee the verification of the inference process.

Some current approaches are being explored to tackle this problem. Cryptographic methods like homomorphic encryption (HE) and secure multi-party computation (MPC) are commonly used to ensure data privacy. Homomorphic encryption allows operations on encrypted data without revealing the underlying information, while MPC enables multiple parties to collaborate on computations while keeping their inputs private. However, these methods do not address the verifiability of the inference results. To solve this, zero-knowledge proofs (ZKPs) are being utilized to provide proof of computation correctness without disclosing sensitive information. Despite progress, a comprehensive solution that meet both privacy and verifiability requirements is still lacking.

This paper advances the state of the field by proposing VHENN (Verifiable Homomorphic Encrypted Neural Network Inference), a scheme that combines homomorphic encryption with zero-knowledge proofs. VHENN protects the privacy of user data, models, and inference results while offering verifiable guarantees of model authenticity and inference correctness. Although VHENN has not yet reached the efficiency required for practical applications due to the computational overhead of homomorphic encryption and zero-knowledge proofs, it still serves as a pioneering effort that lays a foundation for future advancements in achieving both privacy preservation and verifiable neural network inference.

This work was supported by the General Program of National Natural Science Foundation of China (No. 62372173), the General Program of National Natural Science Foundation of China (No. 62372149), and the Key Program of National Natural Science Foundation of China (No. U23A20303). Our research group has a solid foundation in this field, having developed innovative methods using homomorphic encryption and secure multi-party computation for machine learning systems.