Computer Science and Application 计算机科学与应用, 2026, 16(4), 195-204
Published Online April 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.164122

Hans汉斯

可验证外包的二值神经网络隐私推理方案

郭玉麒，岳笑含

沈阳工业大学信息科学与工程学院，辽宁 沈阳

收稿日期：2026年3月9日；录用日期：2026年4月10日；发布日期：2026年4月20日

摘　要

随着深度学习在移动终端与物联网场景中的广泛应用，资源受限设备对高效、安全模型推理的需求日益增强。二值神经网络(BNN)通过权重与激活二值化显著降低计算与存储开销，但在“机器学习即服务”模式下，推理外包至云端执行，如何兼顾输入隐私、模型机密性与结果可验证性成为关键问题。针对现有方案在效率、通信与可验证性之间难以平衡的不足，本文提出一种支持外包可验证的BNN隐私推理框架。该方案基于椭圆曲线ElGamal同态加密构建密文线性计算结构，引入“Ciphertext as Commitment”范式实现与Pedersen承诺的统一表达，并结合广义内积论证协议，实现线性层对数级通信验证。针对非线性层推理，设计同态置换与乘法掩码结合的交互式协议，实现符号激活函数的安全验证。安全性分析与实验结果表明，该方案在保障推理正确性的同时有效降低验证与通信开销，适用于资源受限环境下的安全外包推理。

关键词

二值神经网络，同态加密，可验证计算

Verifiable Privacy Inference Scheme for Binary Neural Networks Based on Outsourcing

Yuqi Guo, Xiaohan Yue

School of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning

Received: March 9, 2026; accepted: April 10, 2026; published: April 20, 2026

Abstract

With the widespread application of deep learning in mobile terminals and IoT scenarios, the demand for efficient and secure model inference on resource-constrained devices is increasing day by

文章引用: 郭玉麒, 岳笑含. 可验证外包的二值神经网络隐私推理方案[J]. 计算机科学与应用, 2026, 16(4): 195-204.
DOI: 10.12677/csa.2026.164122

day. Binary neural networks (BNN) significantly reduce computational and storage costs by binarizing weights and activations. However, in the "machine learning as a service" model, inference is outsourced to the cloud for execution. How to balance input privacy, model confidentiality, and result verifiability becomes a key issue. In response to the shortcomings of existing solutions in balancing efficiency, communication, and verifiability, this paper proposes a BNN privacy inference framework that supports outsourced verification. This scheme is built based on the elliptic curve ElGamal homomorphic encryption to construct a ciphertext linear computing structure. The "iphertext-as-Commitment" paradigm is introduced to achieve a unified expression with Pedersen commitment, and combined with the generalized inner product argument protocol, it realizes logarithmic-level communication verification for the linear layer. For non-linear layer inference, an interactive protocol combining homomorphic permutation and multiplication mask is designed to achieve secure verification of symbolic activation functions. Security analysis and experimental results show that this scheme effectively reduces verification and communication costs while ensuring the correctness of inference, and is suitable for secure outsourced inference in resource-constrained environments.

Keywords

Binary Neural Network, Homomorphic Encryption, Verifiable Computation

1. 引言

随着深度学习在计算机视觉、自然语言处理等领域的广泛应用，人工智能模型正逐步向移动终端与物联网设备延伸。然而，传统深度神经网络参数规模庞大、计算密集，难以直接部署于资源受限设备[1]。二值神经网络(BNN)通过将权重与激活函数量化为二值形式，以 XNOR-Popcount 位运算替代浮点乘累加操作，显著降低了存储与计算开销，在移动医疗[2]、IoT 异常检测[3]等实时场景中展现出良好的应用前景。

尽管 BNN 缓解了终端计算压力，在实际应用中，基于“机器学习即服务”(MLaaS)的云端外包推理模式仍然占据主流。一方面，客户端可借助云端算力完成模型推理；另一方面，服务方也可保护模型权重不被泄露。然而，该模式带来了严重的隐私与安全挑战：客户端输入数据通常具有高度敏感性，而模型参数亦属于核心商业资产。同时，在 GDPR 等[4]数据合规要求不断强化的背景下，明文推理方式已无法满足安全标准。

现有加密推理技术能够在一定程度上保护数据与模型隐私，但仍存在关键缺陷：服务器在“黑盒”环境下执行计算，客户端无法验证其是否诚实完成推理。恶意服务器可能通过跳过计算步骤、替换模型参数甚至伪造结果来降低成本。在医疗诊断、金融风控等[5]高风险场景中，缺乏计算完整性保障将直接影响决策可靠性。因此，构建同时满足数据隐私、模型机密性与计算可验证性的外包推理框架，成为当前亟待解决的重要问题。

综上，如何在资源受限环境下，构建兼顾隐私保护与计算完整性的 BNN 安全推理机制，是本文关注的核心问题。

2. 相关工作

深度学习外包推理的隐私保护与可验证性研究是当前密码学与机器学习交叉领域的研究热点，现有

研究主要围绕隐私保护计算和可验证计算两大技术方向展开，部分方案尝试融合二者特性实现隐私与可验证性的兼顾，但针对二值神经网络(BNN)的轻量化、高兼容性方案仍存在诸多不足。本文从技术路线分类梳理现有研究，并分析各类方案在 BNN 外包推理场景下的适配性问题。

2.1. 基于隐私保护计算的神经网络推理

隐私保护计算技术旨在实现密文空间下的神经网络推理，从根源上保护输入数据与模型参数的隐私，主流技术包括全同态加密(FHE)、安全多方计算(MPC)与混淆电路(GC)。

全同态加密技术支持密文域下的任意布尔运算和算术运算，是实现隐私推理的理想技术方案[6]，但该技术存在自举操作开销大、密文扩展率高的问题，即使针对轻量化的 BNN，也难以满足资源受限终端的实时性需求。安全多方计算(MPC) [7]与混淆电路(GC)则通过将计算任务拆分至多个参与方实现隐私保护，XONN [8]等方案基于 XNOR 位运算优化 BNN 的混淆电路推理过程，显著降低了计算复杂度，但此类方案需要多轮交互完成计算，通信开销居高不下，且仅实现隐私保护，未提供推理结果的可验证机制，无法抵抗恶意服务器的计算篡改行为。

整体而言，此类方案的核心短板为缺乏计算可验证性设计，且在计算效率与通信开销的平衡上难以适配资源受限设备的实际需求。

2.2. 基于可验证计算的神经网络推理

可验证计算技术通过生成计算过程的零知识证明，实现客户端对服务器计算结果的完整性验证，主流研究基于零知识证明(ZKP)构建可验证推理框架[9]。

现有可验证推理方案多面向浮点型或整数型传统神经网络，需将复杂的网络计算转化 R1CS/QAP 等形式的算术电路，再基于电路生成零知识证明[10]。该方式存在证明生成开销大、证明规模庞大的问题，尤其在处理卷积层与激活函数时，算术电路的构建成本极高，验证阶段的计算与通信开销也超出了资源受限终端的承载能力。同时，多数方案仅针对神经网络的线性层设计验证机制，缺乏对非线性激活函数在加密状态下的统一验证方案，无法实现端到端的推理可验证性。

此类方案虽解决了计算可验证性问题，但未针对 BNN 的二值化特性做轻量化优化，且隐私保护与可验证性的融合设计存在明显缺陷。

2.3. 面向 BNN 的隐私可验证推理研究现状

BNN 通过权重与激活的二值化实现了计算与存储的轻量化，成为资源受限设备的首选网络模型，但现有针对 BNN 的安全推理研究仍处于初步阶段。部分研究尝试将 FHE、MPC 与 BNN 结合实现隐私推理，部分研究探索将 ZKP 与 BNN 的位运算结合实现可验证性，但同时满足输入隐私、模型机密、计算可验证且适配资源受限环境的方案极少。

现有方案的核心技术瓶颈体现在三方面：一是难以在密文空间中保持 BNN 的位运算优势，密文计算过程大幅增加了计算开销；二是缺乏线性层与非线性层统一的轻量化验证机制，验证阶段的通信与计算成本过高；三是在双向隐私保护、推理可验证性与终端可部署性三者间难以实现平衡，无法满足 MLaaS 模式下 BNN 外包推理的实际需求。

2.4. 本文方案与现有工作的比较

针对现有研究在隐私保护、计算效率与可验证性之间难以兼顾的问题，本文提出了一种支持外包可验证的 BNN 隐私推理框架。该方案利用椭圆曲线 ElGamal 同态加密构建密文计算结构，并通过“Ciphertext-as-Commitment”范式将密文表示与 Pedersen 承诺统一，从而实现密文计算与可验证计算的

融合。此外，本文结合广义内积论证协议实现线性层计算的对数级通信验证，并针对 BNN 符号激活函数设计交互式验证协议，从而实现高效的非线性层验证。

为了更加直观地说明本文方案的特点，选取当前领域内 2 个代表性方案与本文方案进行定量对比，其中 ZKCNN [9]为基于零知识证明的通用可验证神经网络推理方案，XONN [8]为基于混淆电路的 BNN 隐私推理方案，对比维度涵盖计算复杂度、通信开销、交互轮次、安全假设与可验证性。

Table 1. Quantitative comparison of verifiable privacy inference schemes
表 1. 可验证隐私推理方案定量对比

方案	计算复杂度	通信开销	交互轮次	安全假设	可验证性
ZKCNN [9]	$O(N^2)$	高(KB~MB)	$O(N)$	离散对数假设	是
XONN [8]	$O(N\log N)$	中(KB)	$O(\log N)$	混淆电路语义安全性	否
本文方案	$O(N\log N)$	低(B)	$O(1)$	椭圆曲线离散对数假设	是

由表 1 可知，ZKCNN 虽实现了推理可验证性，但计算复杂度为平方级，且不支持 BNN，通信与交互开销较大，无法适配资源受限设备；XONN 针对 BNN 做了轻量化优化，降低了计算复杂度，但未实现可验证性，且交互轮次多、通信开销较高；相比之下，本文方案在保证输入隐私和推理可验证性的同时，通过利用 BNN 结构特性显著降低了通信与计算开销，更适用于资源受限环境下的安全推理场景。

综合来看，现有研究在实现隐私保护推理方面已取得重要进展，但在效率、通信成本与可验证性之间仍然存在一定权衡。本文通过结合二值神经网络结构特点与可验证计算协议，在保证隐私安全的同时有效降低验证开销，为资源受限环境下的安全外包推理提供了一种可行方案。

3. 方案构建

3.1. 方案系统模型

本文构建一个典型的机器学习外包推理方案如图 1 所示，系统由客户端与云服务器两类实体构成。客户端为资源受限的终端设备，持有待推理的私有输入数据；云服务器具备充足的计算与存储资源，并持有预训练完成的二值神经网络模型参数。双方在无可信第三方参与的条件下，通过密码学协议协作完成模型推理任务。

在系统初始化阶段，客户端生成基于椭圆曲线 ElGamal 体制的公私钥对，并将公钥发送至服务器。随后，客户端使用公钥对本地输入数据进行加密，并将加密后的密文上传至云端。自此，所有计算均在密文状态下进行，服务器无法获知任何明文输入信息。服务器在接收到加密数据后，依据预训练的 BNN 模型结构，在密文域内执行各层同态运算，并维护所有中间结果的 ElGamal 密文表示。计算完成后，服务器向客户端返回最终推理结果的密文形式及相应的可验证证明信息。

在职责划分上，客户端负责密钥生成、输入加密、结果验证与最终解密，不参与具体模型计算过程；服务器负责密文推理与证明生成，但既无法获取输入明文，也不向客户端公开模型参数。该职责分离确保了输入数据与模型权重的双向隐私保护。

从体系结构上看，本方案采用“同态计算与计算型承诺相结合”的设计思想。ElGamal 密文在具备语义安全性的同时，其代数结构天然具有计算绑定特性，因此可视为对底层明文值的承诺表示。服务器在密文域中执行的每一步同态运算共同构成完整的推理计算轨迹，而客户端通过验证服务器提供的证明，

确认该计算轨迹严格遵循预定义模型逻辑。由此，在不泄露输入与模型信息的前提下，实现了推理结果的可验证性与计算完整性保证。

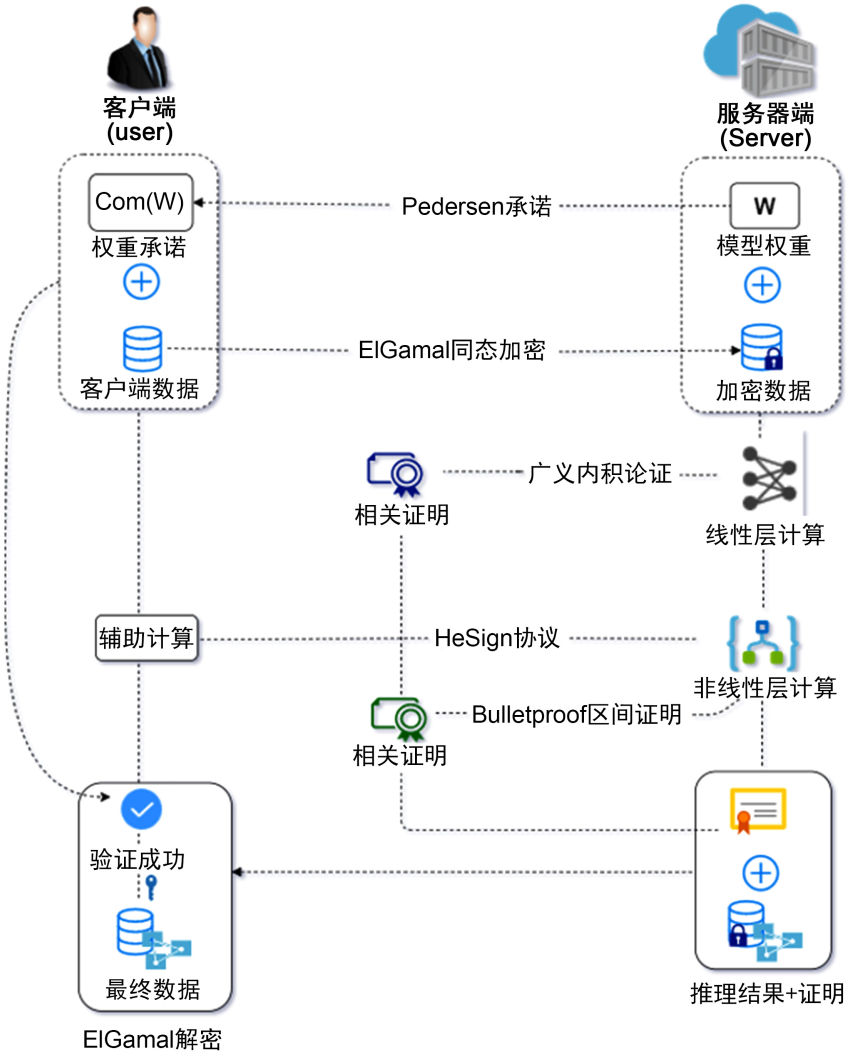

Figure 1. System model diagram
图 1. 系统模型图

3.2. 方案形式化定义

本节将本文提出的支持外包可验证的 BNN 隐私保护推理框架抽象为一个由六个阶段组成的密码学协议。基于“密文即承诺”范式，本方案定义在椭圆曲线 ElGamal 密文空间上的运算与验证关系。本方案形式化地定义一个六元组：(Setup, ModelCommit, Encrypt, Infer, Verify, Decyrpt)，并给出各阶段的具体形式化定义。

1) 初始化阶段

在初始化阶段，客户端负责建立系统的公共参数和密钥体系。

$\mathrm{Setup}\left(1^{\lambda}\right)\to pp$：输入为安全参数 λ，输出全局公开参数。

$\mathrm{KeyGen}\left(pp\right)\to\left(sk,pk\right)$：算法由客户端执行，输出用于加密数据的密钥对 $\left(sk,pk\right)$。

2) 承诺阶段

承诺阶段主要针对服务器持有的模型参数。虽然输入数据的密文本身即承诺，但模型权重在推理前是固定的，服务器必须对模型权重进行承诺，以防在推理过程中偷换模型。

$\mathrm{ModelCom}(pp,\mathbf{W})\to\mathcal{C}_{\mathbf{W}}$：该算法由服务器执行，输入公共参数和模型权重集合，生成模型承诺集合，用于后续的推理证明。

3) 数据加密阶段

在数据加密阶段，针对要被推理的原始数据，进行加密和结构调整，使得其在适配 BNN 模型推理的同时，保证不暴露客户端隐私。

$\mathrm{Enc}(pk,\mathbf{A}_{in})\to\mathbf{C}^{(0)}$：该算法由客户端执行，输入公钥和客户端原始数据，输出能被 BNN 推理的密文数据。

4) 模型推理阶段

服务器依据模型权重在密文空间中执行推理计算。该阶段是协议的核心，包含线性层的同态计算和非线性层的交互式激活两个子过程。

$\mathrm{Infer}\left(\mathbf{C}^{(l-1)},\mathbf{W}^{(l)}\right)\to\left(\mathbf{C}^{(l)},\Pi^{(l)}_{proof}\right)$：对于 BNN 的神经网络中的第 l 层$\left(l\in[1,L]\right)$，输入为上一层密文 $\mathbf{C}^{(l-1)}$，第 l 层模型权重 $\mathbf{W}^{(l)}$，输出下一层的密文运算结果，以及相应的运算证明。

5) 验证阶段

在验证阶段，当服务器推理结束后，会将运算推理结果发送给客户端，由客户端去验证每个运算的正确性，进而验证推理过程的完整性。

$\mathrm{Verify}\left(\mathbf{C}^{(0)},\mathbf{C}^{(L)},\mathcal{C}_{\mathcal{W}},\Pi_{proof}\right)\to b$：该算法由客户端执行，以初始输入密文 $\mathbf{C}^{(0)}$、最终输出密文 $\mathbf{C}^{(L)}$、模型承诺 $\mathcal{C}_{\mathcal{W}}$，各神经网络层生成的证明集 Π_{proof} 作为输入，如果验证通过输出 1，否则输出 0。

6) 解密阶段

在此阶段开始之前，假设在验证阶段，客户端已经验证了服务器的推理结果是完整且正确的。因此，该阶段算法为：

$\mathrm{Dec}\left(sk,\mathbf{C}^{(L)},b\right)\to y$：输入私钥 sk，最终输出密文 $\mathbf{C}^{(L)}$ 以及验证结果 b，客户端利用私钥进行解密得到推理结果。

3.3. 非线性层交互式协议

在二值神经网络中，激活层执行非线性的符号函数 Sign(·)，将卷积层的输出特征图二值化为$\{-1,+1\}$。由于 ElGamal 同态加密方案仅支持线性运算，无法直接在密文空间下执行比较与符号判定操作。为此，本方案设计了一种基于同态置换与乘法掩码的客户端辅助计算协议。该协议在利用客户端私钥协助解密计算符号函数的同时，严格保护了特征图的空间结构与数值分布，防止半诚实客户端通过中间结果发动模型反演攻击，导致模型隐私泄露。

为了防御模型反演攻击，本方案采用了“数值–空间”双重混淆机制。其设计动机源于对单纯数值掩码安全缺陷的修补。数值混淆是利用随机掩码向量 $\boldsymbol{r}_{mask}$ 对特征值进行盲化($z'=z\cdot r$)。然而，乘法掩码存在零值泄露隐患。若原始特征值 $z=0$ (即卷积累加和恰好为 0)，则无论掩码 r 为何值，盲化结果 z' 恒为 0。客户端若在解密后发现大量零值，即可掌握特征图的稀疏性分布。这种结构化信息极易被利用来反推模型的网络结构或权重分布，导致严重的模型结构泄露。

为了解决零值泄露引发的结构暴露问题，本方案引入了随机置换矩阵 $\boldsymbol{\Pi}$。在应用掩码之前，先打乱特征图的空间位置($\mathrm{Enc}\left(Z_{perm}\right)=\Pi\cdot\mathrm{Enc}\left(Z\right)$)。通过置换，原始特征图中的零值被随机分散到未知位置。即使客户端解密后观察到零值，由于缺乏位置映射信息，也无法将零值与特定的神经元关联。这种空间

相关性的破坏，有效地掩盖了模型的稀疏性结构，与数值掩码共同构成了完备的隐私防线。

基于上述“数值–空间”双重混淆机制，算法 3.1 给出了激活层完整的交互式计算与证明生成流程。由于涉及解密与重加密，激活层是恶意服务器进行欺诈的高风险环节。为了确保计算的端到端完整性，客户端必须对服务器的参数与执行过程进行验证。在执行恢复操作前，为防止服务器使用错误的去掩码因子 σ_r，服务器需证明公开的 σ_r 确实是秘密掩码 $\boldsymbol{r}_{mask}$ 的符号。在参数 σ_r 和置换 $\boldsymbol{\Pi}$ 被确立后，需进一步验证服务器是否确实按照声明的步骤对密文进行了线性变换。

算法 3.1：激活层交互式计算与证明生成流程

Input：$\mathcal{S}$ (服务器)输入上一层输出的密文张量 $\mathbf{C}_{in}$，$\mathcal{C}$ (客户端)输入私钥 sk，公钥 pk

Output：S 输出本层的激活输出的符号密文 $\mathbf{C}_{out}$，激活层证明集 $\left(\pi_{inj},\pi_{mask},\pi_{rec}\right)$

执行流程：

1. $\mathcal{S}$ 将 $\mathbf{C}_{in}$ 转换为对应的 Pedersen 承诺 $\mathbf{Comm}_{in}=f\left(\mathbf{C}_{in}\right)$
2. $\mathcal{S}$ 生成随机置换矩阵 $\boldsymbol{\Pi}$ 和随机乘法掩码向量 $\boldsymbol{r}_{mask}\in\left(\mathbb{Z}_q^*\right)^N$ (要求元素非零)
3. $\mathcal{S}$ 执行置换与盲化计算，生成混淆密文 $\mathbf{C}_{mask}\leftarrow\mathbf{r}_{mask}\odot\left(\boldsymbol{\Pi}\cdot\mathbf{C}_{in}\right)$
4. 针对上述变换，$\mathcal{S}$ 调用 GIPA 协议生成注入合法性证明：
$$\pi_{inj}\leftarrow\text{GIPA.Prove}\left(\mathbf{C}_{in},\boldsymbol{\Pi},\boldsymbol{r}_{mask},\mathbf{C}_{mask}\right)$$
5. **return** $\left(\mathbf{C}_{mask},\pi_{inj},\mathbf{Comm}_{in}\right)$ **to** $\mathcal{C}$
6. $\mathcal{C}$ 接收来自 $\mathcal{S}$ 的请求，首先调用算法验证注入合法性：
$$b_{inj}\leftarrow\text{GIPA.Verify}\left(\mathbf{Comm}_{in},\mathbf{C}_{mask},\pi_{inj}\right)$$
7. **if** $b_{inj}=1$ **then**
8. $\quad\mathcal{C}$ 使用私钥 sk 解密混淆后的 $\mathbf{C}_{mask}$，得到盲化特征值 $\mathbf{Z}_{mask}$
9. $\quad\mathcal{C}$ 计算符号 $\sigma'\leftarrow\text{Sign}\left(\mathbf{Z}_{mask}\right)$，并重新加密符号 $\mathbf{C}_{\sigma'}\leftarrow\text{Enc}\left(\sigma',pk\right)$
10. **return** $\mathbf{C}_{\sigma'}$ **to** $\mathcal{S}$
11. $\mathcal{S}$ 利用同态运算与逆置换恢复真实符号密文 $\mathbf{C}_{out}\leftarrow\boldsymbol{\Pi}^{-1}\cdot\left(\text{Sign}\left(\boldsymbol{r}_{mask}\right)\odot\mathbf{C}_{\sigma'}\right)$
12. $\mathcal{S}$ 生成掩码一致性证明 π_{mask} 和恢复完整性证明 π_{rec}
13. **return** $\mathbf{C}_{out}$ **and** $\left(\pi_{mask},\pi_{rec}\right)$

(1) 注入合法性证明：证明 $\mathbf{C}_{mask}$ 是由 $\mathbf{C}_{in}$ 经合法混淆得到的。这是为了防止服务器构造恶意密文诱导客户端解密，客户端在解密前必须验证此证明。

(2) 恢复完整性证明：证明 $\mathbf{C}_{out}$ 是由 $\mathbf{C}_{\sigma'}$ 经合法恢复得到的。这是为了保证最终推理结果的正确性。

上述两个变换本质上均为矩阵-向量乘法(线性变换)。因此统一复用 GIPA 协议。通过上述机制，方案在不泄露置换矩阵 $\boldsymbol{\Pi}$ 和掩码 $\boldsymbol{r}_{mask}$ 具体数值的前提下，构成了严密的验证闭环。

3.4. 方案的安全性以及功能性需求

结合终端推理外包的实际应用场景，本文在恶意服务器与半诚实客户端模型下，对所提出方案提出

如下安全性与功能性需求：

输入隐私性：在协议执行过程中，服务器仅接触客户端输入的 ElGamal 密文。在离散对数假设成立的前提下，该加密体制满足语义安全性(IND-CPA)，从而保证任意多项式时间攻击者区分不同明文输入的优势为可忽略函数，无法从密文及交互信息中获得关于输入的有效语义信息。

模型权重隐私性：客户端不应获得模型权重、网络结构或特征分布等敏感信息。服务器仅以密文计算与零知识证明形式参与交互，不暴露明文参数。协议应保证即使在多轮交互下，客户端仍无法恢复或逼近模型权重。

计算可验证性：对于任意恶意服务器，若其偏离预定义 BNN 计算逻辑(如篡改权重或伪造结果)，则客户端验证通过的概率应为可忽略函数。仅当服务器严格按模型在密文域执行计算时，证明方可通过验证，从而保证推理完整性与结果可信性。

4. 性能分析

本节进行仿真实验来测试方案的实际性能，为评估所提出可验证 BNN 隐私推理方案的实际可行性与性能开销，本文从端到端同态推理正确性与 GIPA 验证协议性能开销两个方面进行实验分析。所有实验均在统一测试平台上完成，具体配置如表 1 所示。

4.1. 实验环境说明

表 2 给出了实验运行环境，包括处理器型号、内存容量、操作系统版本、编译工具链以及所使用的密码学与并行计算库。实验在 CPU 环境下完成，未依赖 GPU 加速，以更真实地反映资源受限场景下的运行性能。

Table 2. The configuration of test platform
表 2. 测试平台配置

项目	配置
处理器	AMD Ryzen 7 6800H (8 核 16 线程, 3.2 GHz, L3 = 16 MB)
内存	7.5 GB DDR5
操作系统	Ubuntu 20.04.6 LTS (WSL2, 内核 6.6.87)
编译器	GCC 9.4.0, C++17 标准, -O2 优化
构建系统	CMake 4.1.1
密码学库	libsecp256k1 0.1 (secp256k1 椭圆曲线运算)
并行框架	OpenMP (GCC 内置)
训练框架	PyTorch 1.13.1 (CPU), NumPy 1.24.4

4.2. 密文推理正确性分析

为验证本文提出的 BNN 密文算术化重构方案在功能层面的正确性，本文对端到端同态推理结果进行系统测试。正确性指标包括：符号匹配率：密文推理输出的 margin 符号与对应明文推理结果是否一致；精确值匹配率：解密后的 margin 数值是否与明文推理结果完全相同。实验随机抽检 1536 个测试通道中的 500 个样本进行验证，统计结果如表 3 所示。

Table 3. End-to-end homomorphic inference correctness summary
表 3. 端到端同态推理正确性汇总

度量指标	结果
符号匹配率	100%
精确值匹配率	100%
测试通道数(抽检)	500/1536

实验结果表明，在所有测试样本中，密文域同态推理结果与明文推理结果一致。符号匹配率与精确值匹配率均达到 100%，未出现计算偏差。

4.3. GIPA 协议性能评估

在保证推理正确性的基础上，本节进一步分析可验证机制的性能开销。实验统计了不同向量规模下 GIPA 协议在证明生成(Prove)与验证(Verify)阶段的执行时间，以及递归折叠轮数与证明大小。实验结果如表 4 所示。

结果表明，随着向量规模的增加，密文生成与证明生成时间均呈近似线性增长趋势，验证时间同样随规模线性增加，但整体显著低于证明生成时间。在 $N = 1024$ 时，证明生成时间为 224.7 ms，验证时间为 83.9 ms，验证阶段约占证明生成阶段时间的 37%。该现象主要源于验证方无需重新计算完整内积，仅需执行递归折叠运算与最终一致性检验，从而有效降低了客户端计算负担。

Table 4. GIPA protocol performance benchmark
表 4. GIPA 协议性能基准

N	折叠轮数 k	密文生成(ms)	生成证明(ms)	验证证明(ms)	证明大小(B)
16	4	16.0	3.3	2.3	560
64	6	63.9	13.4	6.4	824
256	8	254.5	54.8	22.9	1088
512	9	512.9	112.7	43.6	1220
1024	10	1022.2	224.7	83.9	1352

从通信开销角度观察，证明大小随折叠轮数呈对数级增长。当向量规模由 16 扩展至 1024 时，证明大小仅由 560 字节增长至 1352 字节，增长幅度较为平缓，符合广义内积论证协议的理论复杂度特性。该结果说明，本方案在保证可验证性的同时，能够将通信复杂度控制在较低水平，适用于带宽受限的终端设备场景。

结合实际 BNN 单通道规模并进行估算，在向量维度对齐至 1024 的情况下，单通道证明生成时间约为 1.8 秒，验证时间约为 0.7 秒。相比客户端重新执行完整推理计算所需时间，验证开销显著更低，从而体现出验证优于重算的设计优势。整体而言，实验结果表明本文提出的可验证 BNN 隐私推理框架在保证端到端计算正确性的前提下，实现了验证时间与通信规模的有效控制，在资源受限环境中具备实际部署可行性。

5. 结论

本文提出了一种可验证的二值神经网络隐私推理框架。该方案基于椭圆曲线 ElGamal 加法同态加密构建密文线性计算结构，引入“Ciphertext as Commitment”范式并结合广义内积论证协议，实现线性层对数级通信验证，同时通过交互式机制完成符号激活函数的安全验证，从而覆盖完整 BNN 推理流程。理论分析与实验结果表明，方案在满足输入隐私性、模型机密性与计算可验证性的同时，将验证开销控制在可接受范围内，在安全性与效率之间取得了良好平衡，具有实际部署价值。

参考文献

[1] Menghani, G. (2023) Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, **55**, 1-37. https://doi.org/10.1145/3578938

[2] Fasfous, N., Vemparala, M., Frickenstein, A., Frickenstein, L., Badawy, M. and Stechele, W. (2021) Binarycop: Binary Neural Network-Based COVID-19 Face-Mask Wear and Positioning Predictor on Edge Devices. 2021 *IEEE International Parallel and Distributed Processing Symposium Workshops* (*IPDPSW*), Portland, 17-21 June 2021, 108-115. https://doi.org/10.1109/ipdpsw52791.2021.00024

[3] Qiu, H., Ma, H., Zhang, Z., Gao, Y., Zheng, Y., Fu, A., *et al*. (2023) RBNN: Memory-Efficient Reconfigurable Deep Binary Neural Network with IP Protection for Internet of Things. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **42**, 1185-1198. https://doi.org/10.1109/tcad.2022.3197499

[4] General Data Protection Regulation (GDPR) (EU) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj/

[5] Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F. (2020) Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging. *Nature Machine Intelligence*, **2**, 305-311. https://doi.org/10.1038/s42256-020-0186-1

[6] Onoufriou, G., Hanheide, M. and Leontidis, G. (2022) EDLaaS: Fully Homomorphic Encryption over Neural Network Graphs for Vision and Private Strawberry Yield Forecasting. *Sensors*, **22**, Article 8124. https://doi.org/10.3390/s22218124

[7] Berry, C. and Komninos, N. (2022) Efficient Optimisation Framework for Convolutional Neural Networks with Secure Multiparty Computation. *Computers & Security*, **117**, Article 102679. https://doi.org/10.1016/j.cose.2022.102679

[8] Riazi, M.S., Samragh, M., Chen, H., *et al*. (2019) {XONN}:{XNOR-Based} Oblivious Deep Neural Network Inference. 28*th USENIX Security Symposium* (*USENIX Security* 19), Santa Clara, 14-16 August 2019, 1501-1518.

[9] Liu, T., Xie, X. and Zhang, Y. (2021) zKCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy. *Proceedings of the* 2021 *ACM SIGSAC Conference on Computer and Communications Security*, Seoul, 15-19 November 2021, 2968-2985.

[10] Feng, B., Wang, Z., Wang, Y., Yang, S. and Ding, Y. (2024) ZENO: A Type-Based Optimization Framework for Zero Knowledge Neural Network Inference. *Proceedings of the* 29*th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, Volume 1, San Diego, 27 April-1 May 2024, 450-464.