

基于数据聚类的CSI反馈Transformer网络简化实现方法

还冬锐¹, 张逸帆¹, 姜明^{1,2}

(1. 东南大学移动通信全国重点实验室, 南京 210096; 2. 紫金山实验室, 南京 211111)

摘要: 为应对大规模多输入多输出 (Multiple-input multiple-output, MIMO) 系统中信道状态信息 (Channel state information, CSI) 反馈开销的日益增长, 基于深度学习的 CSI 反馈网络 (如 Transformer 网络) 受到了广泛的关注, 是一种非常有应用前景的智能传输技术。为此, 本文提出了一种基于数据聚类的 CSI 反馈 Transformer 网络的简化方法, 采用基于聚类的近似矩阵乘法 (Approximate matrix multiplication, AMM) 技术, 以降低反馈过程中 Transformer 网络的计算复杂度。本文主要对 Transformer 网络的全连接层计算 (等效为矩阵乘法), 应用乘积量化 (Product quantization, PQ) 和 MADDNESS 等简化方法, 分析了它们对计算复杂度和系统性能的影响, 并针对神经网络数据的特点进行了算法优化。仿真结果表明, 在适当的参数调整下, 基于 MADDNESS 方法的 CSI 反馈网络性能接近精确矩阵乘法方法, 同时可大幅降低计算复杂度。

关键词: 信道状态信息反馈; 多输入多输出; 神经网络; 近似矩阵乘法; 聚类计算

中图分类号: TN92

文献标志码: A

A Simplified Implementation Method of CSI Feedback Transformer Network Based on Data Clustering

HUAN Dongrui¹, ZHANG Yifan¹, JIANG Ming^{1,2}

(1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China; 2. Purple Mountain Laboratories, Nanjing 211111, China)

Abstract: In order to cope with the increasing overhead of channel state information (CSI) feedback in massive multiple-input multiple-output (MIMO) systems, deep learning-based CSI feedback networks (such as Transformer) have received extensive attention and become very promising intelligent transmission technologies. To this end, this paper proposes a simplification method of CSI feedback Transformer network based on data clustering, which uses clustering-based approximate matrix multiplication (AMM) to reduce the computational complexity of the Transformer network in the feedback process. In this paper, we focus on the computation of the fully connected layer in the Transformer network (equivalent to matrix multiplication), adopt the simplification methods such as product quantization (PQ) and MADDNESS, analyze their influence on the computational complexity and system performance, and optimize the algorithm according to the characteristics of neural network data. Simulation results show that the performance of the CSI feedback network based on the MADDNESS method is close to that of the exact matrix multiplication method with an appropriate parameter adjustment, and the computational complexity can be greatly reduced.

基金项目: 国家自然科学基金 (62271137)。

收稿日期: 2024-01-12; **修订日期:** 2024-04-30

Key words: channel state information (CSI) feedback; multiple-input multiple-output (MIMO); neural network; approximate matrix multiplication (AMM); clustering calculation

引言

随着第5代以及更先进的移动通信技术的发展,大规模多输入多输出(Multiple-input multiple-output, MIMO)系统的研究得到了广泛关注^[1-2]。通过充分利用信道状态信息(Channel state information, CSI),MIMO可以进行复杂的波束成形、预编码和干扰抑制,从而显著提高网络吞吐量^[3-4]。用户设备(User equipment, UE)接收到基站(Base station, BS)发送的导频信号后,将其估计的CSI反馈给基站。CSI反馈在频分双工(Frequency division duplex, FDD)模式下比在时分双工(Time division duplex, TDD)模式下更不可或缺,因为TDD模式的上下行传输在相同的频率资源中进行,利用信道互易性可大幅降低CSI反馈开销。而FDD模式的上下行链路使用不同的频段,无法利用信道互易性^[5]。此外,由于天线数、接收机数和子载波数的增加,不断升高的反馈开销使得高效准确的CSI反馈难以实现,阻碍了大规模MIMO技术在实际系统中的应用。

为了节省反馈开销,大量已有的研究在基站端对下行CSI进行高效压缩并发送到反馈链路,之后在基站端尽可能完整准确地恢复CSI。文献[6]首先采用了压缩感知(Compressive sensing, CS)方法以降低大规模MIMO系统中的CSI反馈负载。反馈编码可以通过简单的随机投影进行,并且可以在低压缩比的情况下实现高精度的CSI恢复。同时,许多算法被尝试与CS方法结合以获得更好的反馈性能,如最小绝对值收敛和选择算子(Least absolute shrinkage and selection operator, LASSO) ℓ_1 -solver^[7],近似消息传递(Approximate message passing, AMP)^[8],基于增广拉格朗日和交替方向算法的全变分最小化(Total variation minimization by augmented Lagrangian and alternating direction algorithms, TVAL3)^[9]和三维块匹配(Block matching 3D, BM3D)-AMP^[10]。但是,CS方法严重依赖于CSI矩阵的稀疏性。在实际应用中,信道通常不满足稀疏特性,这给CS方法的建模带来了困难。此外,大多数基于CS的方法采用迭代算法来恢复CSI,导致其计算效率较低。

为应对该问题,深度学习被引入到CSI反馈中,其可以自适应地捕获实际场景中的信道稀疏性,从而提高反馈性能。文献[11]在闭环MIMO系统中采用深度神经网络(Deep neural network, DNN)进行CSI编码。仿真结果表明,在所有测试信噪比下,基于DNN的编码器性能都超过了基于奇异值分解(Singular value decomposition, SVD)的MIMO空间复用系统的性能。之后,文献[12]开创性地提出CsiNet,使用基于卷积神经网络(Convolutional neural network, CNN)的自动编码器来实现CSI反馈任务,并展示其相对传统CS方法有明确的性能优势,尤其是在高压压缩比的场景中,性能增益更为显著。然而,CsiNet无法捕捉输入CNN的数据样本之间的长期依赖关系。由于Transformer网络^[13]使用自注意力机制表征数据样本之间的短距离和长距离依赖,在语义特征提取、远距离特征捕捉和综合特征提取方面都显著优于传统的CNN和循环神经网络(Recurrent neural network, RNN),文献[14]提出了一种基于Transformer网络的CSI压缩反馈方法CsiTransformer。

在CsiTransformer网络中,存在多层可等效为矩阵乘法的全连接层计算。经估算,以浮点运算数(Floating point operations, FLOPs)计,在典型的1/4压缩率下,CsiTransformer网络的全连接层计算复杂度约占完整网络计算复杂度的73%。因此,如能大幅度降低全连接层的计算复杂度,CSI压缩反馈的计算开销可大幅降低。矩阵乘法是机器学习和科学计算中最基本的运算之一,目前已有大量关于加速矩阵乘法的研究。文献[15-18]通过特定的硬件设计来加速神经网络的乘法运算;文献[19-21]提出了分布式矩阵乘法的加速方法,其目的在于解决分布式计算中的“落后者效应”,即由少数慢速处理器造成的延迟问题。在上述两类方法中,前者在硬件层面设计了加速方法,后者是专为加速分布式矩阵乘法而设计。

除以上方法外,利用数据聚类的思想对矩阵乘法进行近似计算,可在有限损失精度的前提下大幅加速计算过程,此类方法主要从软件层面进行加速,且应用场景不局限于分布式矩阵乘法。已有不同种类的基于数据聚类的近似矩阵乘法(Approximate matrix multiplication, AMM)方法被提出,包括乘积量化(Product quantization, PQ)^[22]以及MADDNESS(Multiply-ADDitioN-IESS)^[23]。两者均利用向量量化(Vector quantization, VQ)^[24]的思想进行聚类。PQ最初被设计用于最近邻搜索中近似计算向量之间的距离,结果表明PQ提供了欧氏距离的精确近似值,并实现了高效率的搜索。PQ将空间分解为低维子空间的笛卡尔乘积,并分别利用k-means聚类算法量化每个子空间。向量由其子空间量化索引组成的短代码表示,两个向量之间的欧氏距离可以利用它们的代码进行估计。PQ方法可以轻易地推广为AMM方法,即利用矩阵中向量的量化结果预先离线计算乘积查找表(Lookup table, LUT),在线计算时只需确定对应索引后查表即可。但是,PQ方法对输入数据的维度有较高要求,对此MADDNESS方法引入了局部敏感哈希作为聚类函数,降低了对矩阵维度的要求,同时降低了AMM的计算复杂度。对来自不同领域的数百个矩阵进行实验,结果表明,MADDNESS方法的运行速度比精确矩阵乘法快100倍,比之前的AMM方法快10倍。

为降低CSI反馈神经网络的计算复杂度,本文考虑将AMM方法与CsiTransformer网络相结合,替换全连接网络层可以节省大量的矩阵乘法,加快算法运算速度。本文主要的研究工作如下。

(1) 全连接层占据了CsiTransformer网络的大部分计算复杂度,本文将两类基于VQ数据聚类的AMM方法PQ和MADDNESS应用于该网络的全连接层简化计算。CsiTransformer网络所采用的数据集与文献[14]中相同,AMM的学习数据集由神经网络各全连接层的输入数据构成。根据简化前后计算复杂度被降低的程度以及CSI反馈性能的变化幅度,选出适用于本场景的AMM方法。

(2) MADDNESS最初为图像分类、滤波等图像处理应用设计,本文受其启发将其应用至通信场景,并遇到了新的问题:在CSI反馈神经网络中,部分全连接层的输入数据由于经过非线性激活函数,其含零比例较高,原始的MADDNESS方法未能将零值与其他值较好地剥离,影响了最终聚类效果,本文改进了剥离方式。此外,由于神经网络数据的巨大规模,原始MADDNESS方法的衡量聚类效果的中间数据精度较低,会导致较大误差,从而引发错误聚类,因此本文提高了相应数据的精度并获得正确的聚类结果。此外,本文还对MADDNESS方法中岭回归的存储消耗过大的问题进行了优化。

(3) 仿真结果表明,MADDNESS方法适用于CsiTransformer网络的全连接层简化。在本文所测试的4种CSI压缩比下,当向量量化的子向量长度为1时,基于MADDNESS简化的CSI反馈网络性能非常接近采用精确矩阵乘法的CSI反馈性能,且全连接层的计算复杂度降幅可达84%~85%、整个神经网络的计算复杂度降幅可达43%~62%。在其他条件相同的情况下,增大子向量长度并适当调整VQ参数,可实现计算复杂度和CSI反馈性能的灵活调整。

1 系统模型与理论基础

1.1 CsiTransformer 网络

1.1.1 应用场景

考虑单小区下行链路大规模MIMO系统,其中CSI反馈由单天线用户设备发送到配备 N_t 个天线的基站。系统以配备 $\tilde{N}_c$ 个子载波的FDD模式运行,空间频域的CSI矩阵维度为 $\tilde{N}_c \times N_t$,记为 $\tilde{H}$ 。在角延迟域中稀疏化 $\tilde{H}$,即对 $\tilde{H}$ 进行二维离散傅里叶变换,得到 H' 。 H' 是一个复矩阵,可将不可忽略的前 N_c 行合并成一个大小为 $2N_c \times N_t$ 的实值矩阵 H 。

1.1.2 基本结构

CsiTransformer网络包括编码器 f_e 和解码器 f_d ,它们分别负责CSI编码和恢复。具体地说,编码器将信道矩阵 H 转换为反馈码字

$$\mathbf{s} = f_c(\mathbf{H}) \quad (1)$$

式中 $\mathbf{s} \in \mathbf{R}^{L \times 1}$ 且 $L < 2N_c N_t$ 。因此,压缩比为 $\eta = L/(2N_c N_t)$ 。

然后,码字 $\mathbf{s}$ 被发送到基站。假设 $\mathbf{s}$ 被完美接收,则基站使用以下解码器恢复原始信道矩阵

$$\hat{\mathbf{H}} = f_d(\mathbf{s}) \quad (2)$$

基于上述定义,CsiTransformer网络解决的优化问题为

$$\operatorname{argmin}_{f_c, f_d \in \mathcal{F}} \|\mathbf{H} - \hat{\mathbf{H}}\|^2 \quad (3)$$

式中 $\mathcal{F}$ 为定义在 $\mathbf{R}$ 上的闭函数的集合。

1.1.3 评估指标

CsiTransformer采用端到端训练来训练包含编码器和解码器中的完整网络。网络参数通过AdamW算法更新,该算法使用归一化均方误差(Normalized mean squared error, NMSE)作为损失函数^[12]

$$\text{NMSE} = E \left\{ \frac{\|\mathbf{H} - \hat{\mathbf{H}}\|^2}{\|\mathbf{H}\|^2} \right\} \quad (4)$$

NMSE可以衡量CSI矩阵压缩反馈前后的差异程度。除此之外,反馈的CSI用作波束成型矢量,本文使用文献[12]中定义的余弦相似性来评估波束成型矢量的质量

$$\rho = E \left\{ \frac{1}{N_c} \sum_{n=1}^{N_c} \frac{|\hat{\mathbf{h}}_n^H \tilde{\mathbf{h}}_n|}{\|\hat{\mathbf{h}}_n\| \|\tilde{\mathbf{h}}_n\|} \right\} \quad (5)$$

式中 $\hat{\mathbf{h}}_n$ 为第 n 个子载波恢复的信道向量。

1.2 近似矩阵乘法

1.2.1 近似矩阵乘法的优化目标

记 $\mathbf{A} \in \mathbf{R}^{N \times D}$ 和 $\mathbf{B} \in \mathbf{R}^{D \times M}$ 为两个待乘矩阵,其中 $\mathbf{A}$ 为可变矩阵, $\mathbf{B}$ 为固定矩阵, $N \gg D \geq M$ 。AMM的任务为构建3个函数 $g(\cdot)$, $h(\cdot)$ 和 $f(\cdot)$ 以及常量 α 和 β ,使得

$$\|\alpha f(g(\mathbf{A}), h(\mathbf{B})) + \beta - \mathbf{AB}\|_F < \epsilon(\tau) \|\mathbf{AB}\|_F \quad (6)$$

给定时间预算 τ ,需使得误差 $\epsilon(\tau)$ 尽可能地小^[23]。

1.2.2 乘积量化方法

PQ是基于VQ的一种AMM方法,与VQ对应的方法为标量量化(Scalar quantization, SQ)。在SQ中,矩阵的元素不再以向量为最小单位进行量化,而是每个元素均单独以比特数 n_c 进行线性量化,以使得最小和最大的元素分别映射到0和 $2^{n_c} - 1$, 或 -2^{n_c-1} 和 $2^{n_c-1} - 1$ 。由于SQ方法仅进行量化而未进行乘法简化,本文将仅测试基于VQ的AMM方法。

PQ将矩阵 $\mathbf{A}$ 的行看作 N 个长度为 D 的子向量,处于相同列的子向量可以形成一个向量空间。乘积量化中的“乘积”指笛卡尔积^[22]。PQ将向量空间分解为几个低维向量空间的笛卡尔积,并分别对分解后的低维向量空间进行量化。这样,每个向量都可以由低维空间中多个量化质心的组合来表示。假设 $\mathbf{a}$ 表示 $\mathbf{A}$ 的一个行向量, $\mathbf{b}$ 表示 $\mathbf{B}$ 的一个列向量,那么有 $\mathbf{a}^T \mathbf{b} \approx \hat{\mathbf{a}}^T \mathbf{b}$, 其中 $\hat{\mathbf{a}}$ 是低维空间中多个量化质心的组合。在对 $\hat{\mathbf{a}}$ 和 $\mathbf{b}$ 之间的点积进行预计算后,符合条件的 $\mathbf{a}$ 可以重用这些乘积以获得计算加速。PQ和MADNESS这类基于VQ的AMM方法的基本流程图可由图1表示,都可分为离线学习和在线计算两部分,本文在分析复杂度时一般仅需考虑在线计算的复杂度。图1中 $n_c = D/C$, 岭回归步骤为MADNESS算法特有。

PQ方法的具体步骤包括:

(1)质心学习——将 $\mathbf{A}$ 按列划分为 C 个不同的子空间,在每个子空间内单独运行K-means算法,从而聚类形成 K 个质心。

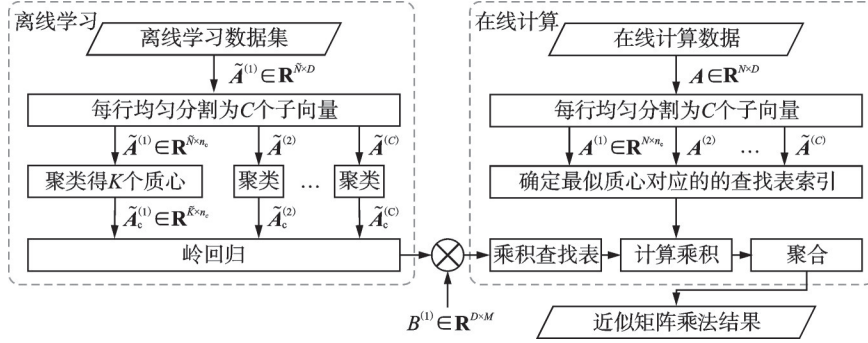


图1 基于向量量化的近似矩阵乘法方法的流程图

Fig.1 Flowchart of the AMM method based on VQ

(2)构建LUT——预计算每个质心与每个子空间对应的向量 $\mathbf{b}$ 的点积。

(3)编码函数 $g(\mathbf{a})$ ——在每个子空间中确定与 $\mathbf{a}$ 最相似的质心,并记录其索引。

(4)聚合, $f(\cdot, \cdot)$ ——对于每个子空间,根据索引和查找表找到所估计的部分乘积值,最后对所有 C 个子空间的结果求和,得到最终结果。

其中,步骤(1,2)属于离线学习步骤,步骤(3,4)属于在线计算步骤。

1.2.3 MADDNESS方法

PQ对矩阵乘法的加速效果在 $N \gg D, M \gg D$ 的情况下较为明显。同时,在PQ的编码函数中,寻找最似质心涉及包括平方运算的欧氏距离计算,码本中所有的 K 个质心都需要遍历计算一次,导致整体计算复杂度较高。即便欧氏距离计算在一些场景下可以替换为免去平方运算的曼哈顿距离计算, NCK 次的总距离计算次数仍使复杂度居高不下。

为了在 M 较小的矩阵上获得较大的加速,同时取代复杂度较高的距离计算,MADDNESS对PQ的步骤进行了优化,引入分割阈值的概念,并利用二元回归树的数据结构,使得PQ中寻找最似质心的方法被大幅简化。MADDNESS的编码函数 $g(\mathbf{a})$ 没有计算子向量 $\mathbf{a}^{(c_i)}$ 和每个原型之间的欧氏距离,而是使用基于平衡二元回归树的局部敏感哈希算法进行 $\mathbf{a}^{(c_i)}$ 的分配。为便于描述,MADDNESS引入了桶 $\mathcal{B}_i^t$ 的概念,它是映射到二元回归树第 t 层索引为 i 的节点的向量集。树的根位于第0层, $\mathcal{B}_1^0$ 包含所有向量。局部敏感哈希算法将 $\mathbf{a}^{(c_i)}$ 哈希到 K 个桶之一,其中相似的子向量倾向于哈希到同一个桶中。通过对每个桶中所有哈希的子向量求平均,可得到原型。在MADDNESS中,原型学习中最关键的步骤是利用训练数据集建立平衡二元回归树的过程。定义与桶关联的平方误差和(Sum of squared errors, SSE),以便建立平衡二元回归树

$$\mathcal{L}(j, \mathcal{B}) \triangleq \sum_{x \in \mathcal{B}} \left(x_j - \frac{1}{|\mathcal{B}|} \sum_{x' \in \mathcal{B}} x'_j \right)^2 \quad (7)$$

$$\mathcal{L}(\mathcal{B}) \triangleq \sum_j \mathcal{L}(j, \mathcal{B}) \quad (8)$$

式中 $\mathcal{B}$ 代表节点包含的向量集合, j 代表分割索引。算法1给出了MADDNESS原型学习的具体步骤。

算法1 MADDNESS原型学习

给定训练集矩阵 $\tilde{\mathbf{A}}$,码本数 C ,每码本质心数 K ;

(1) for $c_i = 1:1:C$

(2) 用 $\tilde{\mathbf{A}}$ 的第 c_i 个码本中所有的子向量组成一个子矩阵,建立根桶

(3) for $k = 1:1:\log_2 K$

(4) 针对第 $k-1$ 次循环得到的所有桶,在每个桶中找到 SSE 最大的 4 列,分别遍历各列元素,计算以其为分割阈值时,两部分数据的 SSE 之和。选出使两部分 SSE 之和最低的元素作为分割阈值,并记录分割后性能最好的列的列号以及分割阈值。最后用分割阈值将当前桶分为两个子桶。

(5) End for

(6) 在第 c_i 个码本中对子向量进行编码,用独热编码记录当前子向量所在的桶,即遍历阈值,若子向量对应位置大于阈值则当前位记 1,否则记 0;

(7) 利用岭回归在质心(包含全零列)上求微小修正项,使质心与编码向量的积更接近输入向量。

(8) End for

(9) 输出各码本中由所求阈值构成的平衡二元回归树,以及各树的叶子结点所对应的各桶的质心。

最后,在对查找表进行数值相加时,MADDNESS 将加法指令替换为平均指令,以牺牲低位信息的较小代价来提高计算速度。

2 CsiTransformer 网络的近似矩阵乘法简化

2.1 优化的全连接层

全连接层的每个节点都与前一层的所有节点相连,可以提取数据集的特征。在 CsiTransformer 网络中有 6 层全连接层,如图 2 所示,它们分别记为 etl1、etl2、el、dl、dtl1、dtl2。

全连接层的核心是矩阵乘法,其计算公式为

$$Y = XW + b_s \quad (9)$$

式中: $X \in \mathbf{R}^{N \times D}$ 为输入矩阵, $W \in \mathbf{R}^{D \times M}$ 为权重矩阵, b_s 是全连接层的偏置。式(9)中的 X 和 W 分别对应式(6)中的 A 和 B 。为统一表述,以下正文中均用 A 和 B 表示。网络中每批次各全连接层的参数呈现在表 1 中,批次大小为 32。

2.2 近似矩阵乘法的参数

AMM 方法在每个子空间中学习到的 K 个原型即为 K 个质心,包含不同质心的 C 个子空间称为 C 个码本。为了便于描述,假设 A 只有一行,在这种情况下 A 变成向量 a 。当列数为 D 时,每个码本对应的子向量长度为 $n_c = D/C$,并将聚类的 K 个质心视作 $n_c = \log_2 K$ bit 的量化。基于上述定义,本文将带参数的基于 VQ 的 AMM 方法(PQ 及 MADDNESS)表示为 $VQ(n_c, n_e)$ 。在本文中, n_c 最小值取 1, n_e 最大值取 8,即质心至多设置 256 个。一般情况下,减小 n_c 或增大 n_e 均可提升利用 AMM 简化计算后 CSI 反馈的性能。

预计算每个质心与 B 的点积,将结果存储在一个大小为 $M \times C \times K$ 的 LUT 中。LUT 的元素可以不进行量化而存储,也可按 8 bit、16 bit 等精度进行均匀量化后存储,从而节省空间并降低计算复杂度。LUT 的量化位数记为 n_L ,不进行量化而存储时, n_L 记为 32(此时 LUT 的元素数据类型为单精度浮点数)。

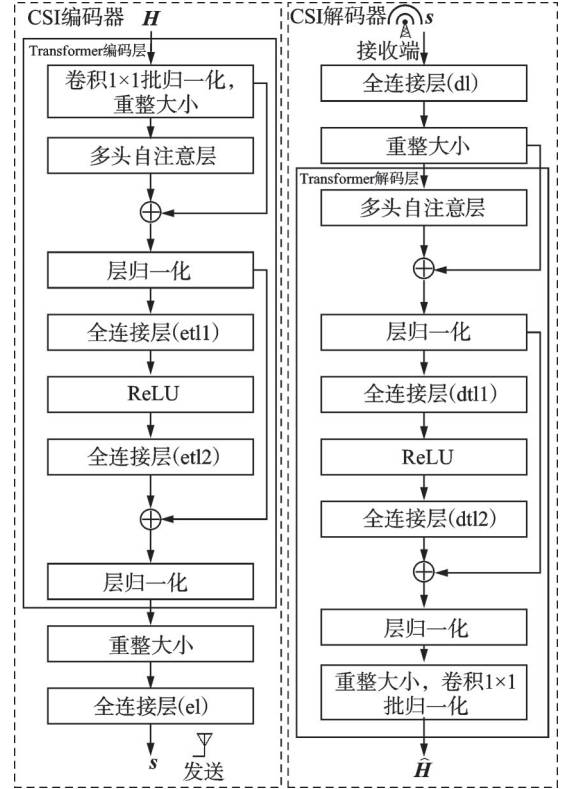


图2 CsiTransformer 网络结构及全连接层位置
Fig.2 CsiTransformer structure and position of fully connected layers

表1 每批次各全连接层对应矩阵的大小参数
Table 1 Matrix sizes for each fully connected layer every batch

参数	etl1	etl2	el	dl	dtl1	dtl2
N	1 024	1 024	32	32	1 024	1 024
D	64	512	2 048	L	64	128
M	512	64	L	2 048	128	64

2.3 针对 CsiTransformer 网络的 MADDNESS 方法优化

在原 MADDNESS 方法中, LUT 的量化位数 $n_L = 8$ 。经初步测试, 若 CsiTransformer 网络的全连接层中应用 $n_L = 8$ 的 MADDNESS 替换, 即使在参数 VQ(1, 8) 下计算简化后 CSI 反馈的性能相比未简化时的反馈性能也相去甚远。该测试说明, LUT 的量化位数过小的 AMM 不适用于简化 CsiTransformer 网络。因此, 本文提升 n_L 至 16, 在参数 VQ(1, 8) 下计算简化后 CSI 反馈的性能与原网络性能接近。

MADDNESS 方法的局部敏感哈希是基于平衡二元回归树的。当矩阵 A 包含大量的零值时, 需要修改原始局部敏感哈希^[23]的向量分配方案。以 etl2 层为例, 由于该层的输入经过了非线性激活函数线性修正单元 (Rectified linear unit, ReLU), 该层的输入矩阵具有 0.886 2 的零值比例 (以一个 1 000 批次的训练集 $\tilde{A}$ 为例)。

原始的 MADDNESS 哈希算法使用大于等于号分割桶, 如文献[23]的算法 1 第 5 行。此时, 如果分割阈值为 0, 而通过激活函数 ReLU 的值不小于 0, 则所有值都将分配给右子桶, 左子桶为空, 分割无效。所以应将“ $\geq$ ”改为“ $>$ ”, 即由 $b = x_{ij} \geq v? 1: 0$ 改为 $b = x_{ij} > v? 1: 0$, 其中“?”为三目运算符的一部分。此时 0 会被赋给左子桶, 与其他数值区分开。测试算法优化前后的性能如表 2 所示, 在码本数量为 128、不进行 LUT 量化的情况下, 修改前 CsiTransformer 网络的 etl2 层 MADDNESS 替换性能随着质心数量的增加而大幅下降, 而修改后 MADDNESS 替换的性能通常会随着质心数的增加而提高。

MADDNESS 方法中的岭回归能够让乘积更接近真实值, 可表示为

$$P = (G^T G + \lambda I)^{-1} G^T \tilde{A} \quad (10)$$

式中: $\tilde{A}$ 为训练集, G 为编码后的 $\tilde{A}$ 矩阵, P 为质心矩阵, λ 恒定为 1。

岭回归的目标是最小化 GP 与 $\tilde{A}$ 的误差, 其本质是改良的最小二乘法, 在 $G^T G$ 的对角线元素上加了 λ 。但岭回归涉及求逆等计算, 在数据量较大时 (如 1 000 批次时 $N=1\,024\,000$), 由于矩阵过大, 会爆发式地消耗训练过程的计算和存储资源。因此, 本文将训练数据均匀分为多份, 分步进行岭回归并取结果均值, 使得每次岭回归时涉及的矩阵计算所需的内存大幅降低, 且分 2 或 4 步时, 近似矩阵乘法性能几乎没有损失。

算法 1 中, 步骤 (4) 确定了当前桶 $\mathcal{B}_i^t$ 分割为两个子桶 $\mathcal{B}_{2i-1}^{t+1}$ 和 $\mathcal{B}_{2i}^{t+1}$ 的最优分割阈值。分别遍历各列元素, 计算以其为分割阈值时, 两部分数据的 SSE 之和, 记为累积 (cumulative) SSE

$$\mathcal{L}_c \triangleq \mathcal{L}(\mathcal{B}_{2i-1}^{t+1}) + \mathcal{L}(\mathcal{B}_{2i}^{t+1}) \quad (11)$$

当数据集较大时 (如 $N=1\,024\,000$), 由于原 MADDNESS 方法采用单精度浮点数定义 $\mathcal{L}_c$, 精度不够, 导致最终计算的 SSE 与真实 SSE 误差较大, 使得最优分割阈值选定错误。若将原 MADDNESS 方法定义的 $\mathcal{L}_c$ 应用于 CsiTransformer 网络的 AMM 替换, 则会导致增大质心数时 CSI 反馈性能反而下降。因此, 应当使用双精度浮点数定义 $\mathcal{L}_c$, 可使得计算精确, 从而解决性能变化趋势异常的问题。

图 3 给出了优化 SSE 精度前后 CsiTransformer 网络的 etl2 层的 MADDNESS 最优分割阈值二叉树 (部分), 测试参数设置为 $\eta = 1/4$, $C=256$, $K=256$, $N=1\,024\,000$ (1 000 批次), $n_L = 16$, 选取的码本索引为 45。由图 3(a) 可见, 优化 SSE 精度前由于 $\mathcal{L}_c$ 计算不准确, 大量分割阈值选取错误, 导致聚类后大量数据依然位于同一个桶中, 无法有效地将数据聚类为较均匀的多簇。由图 3(b) 可见, 在优化 SSE 精度后, 分割阈值二叉树的节点数值符合规律, 能够有效地进行数据聚类。

表 2 MADDNESS 哈希算法优化前后 CSI 反馈的性能

Table 2 Performance of CSI feedback before and after optimization of MADDNESS Hash algorithm

项目	C	K	NMSE	ρ
优化前	128	16	0.017 4	0.992
	128	256	0.774 0	0.705
优化后	128	16	0.011 5	0.994
	128	256	0.010 8	0.995

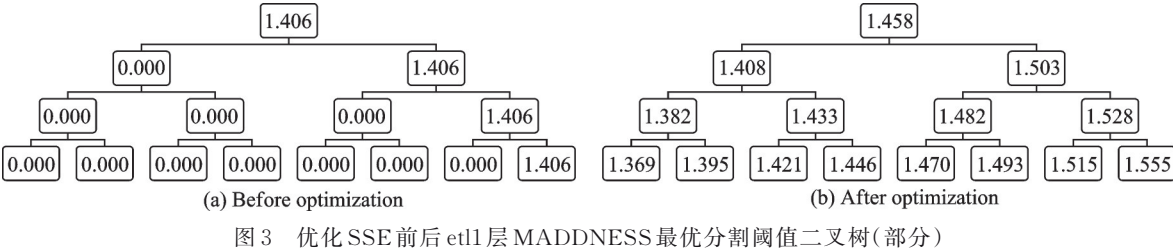


图3 优化SSE前后 etl1层MADDNESS最优分割阈值二叉树(部分)

Fig.3 Part of MADDNESS optimal threshold binary tree of etl1 layer before and after SSE optimization

3 应用近似矩阵乘法后的CSI反馈性能

以下仿真结果中,计算复杂度参考表3中的指令周期和进行计算,表中取 $N=1$,参考指令集为SSE2(Streaming SIMD Extensions 2)^[25]。AMM方法的计算复杂度占精确乘法比例即为AMM的指令周期和除以精确乘法的指令周期和。存储复杂度主要为LUT的存储空间占用,在MADDNESS方法中还需存储分割阈值二叉树的节点值。设置仿真压缩比 $\eta=1/4, 1/8, 1/16, 1/32$,基站天线数 $N_t=32$,用户天线数 $N_r=1$,子载波数 $\tilde{N}_c=256$,每个码本对应的子向量长度 $n_c=1, 2, 4, 6$,质心数等效量化比特数 $n_e=4, 5, 6, 7, 8$,LUT的量化位数 $n_L=16$ 。AMM方法可变矩阵 A 的离线学习数据集由神经网络各全连接层输入数据的前 $N_b=1\,000$ 个批次构成,故各层的学习数据集样本数为 N_bN , N 值可参考表1。为直观展示运用AMM后CSI反馈的性能变化情况,运用AMM后CSI反馈的性能一般与采用精确矩阵乘法的反馈性能对比,即原始CsiTransformer的性能。

表4估算了CsiTransformer网络各层FLOPs以及所有全连接层FLOPs占完整神经网络的比例(未计入部分归一化和激活层),可见无论压缩比取4个值中的任意值,全连接层的FLOPs均在60%以上,

表3 各矩阵乘法方法计算复杂度

Table 3 Calculation complexity of each matrix multiplication method

运算	指令周期	精确乘法	PQ	MADDNESS
浮点加法	4	DM	$2DK - CK$	
浮点乘法	4	DM	DK	
定点加法	1		CM	CM
浮点比较	4		CK	$C\log_2 K$
哈希查表			$O(MC)$	$O(MC)$
LUT大小			(M,C,K)	(M,C,K)
指令周期和		$8DM$	$12DK + CM$	$CM + 4C\log_2 K$

表4 CsiTransformer网络各层浮点运算数

Table 4 Number of FLOPs in each layer of CsiTransformer

压缩比	1/4	1/8	1/16	1/32
CSI编码器,卷积,批归一化	10 240	10 240	10 240	10 240
CSI编码器,多头自注意力层	1 326 849	1 326 849	1 326 849	1 326 849
CSI编码器,全连接层 etl1	2 097 152	2 097 152	2 097 152	2 097 152
CSI编码器,全连接层 etl2	2 097 152	2 097 152	2 097 152	2 097 152
CSI编码器,全连接层 el	1 048 576	524 288	262 144	131 072
CSI解码器,全连接层 dl	1 048 576	524 288	262 144	131 072
CSI解码器,多头自注意力层	1 326 849	1 326 849	1 326 849	1 326 849
CSI解码器,全连接层 dtl1	524 288	524 288	524 288	524 288
CSI解码器,全连接层 dtl2	524 288	524 288	524 288	524 288
卷积,批归一化	7 168	7 168	7 168	7 168
所有全连接层FLOPs比例/%	73.32	70.20	68.35	67.33

超过整个网络计算量的一半,故简化全连接层计算对于降低 CsiTransformer 网络计算复杂度有很高的实际应用价值。

3.1 各全连接层输入矩阵的数据相关性分析

本文所采用的基于 VQ 的 AMM 方法本质是用被称作质心的固定向量替代可变向量,相当于用含较少的信息量的质心表示含较多信息量的原始数据,可将其视作类似有损压缩的过程,网络的输入输出是未被压缩的数据,信息冗余较多,而反馈向量是压缩后的数据,信息冗余较少。用 Pearson 相关系数度量数据间的相关性,数据存在的信息冗余较多,其相关性较强,用聚类后质心代替的效果较好。测试 CsiTransformer 网络各全连接层矩阵行向量之间的 Pearson 相关系数并取绝对值,可得各层数据的相关性热力图,如图 4 所示。图 4 所示的热力图中,每个点的横纵坐标代表用于计算相关系数的两行索引。数值越大,颜色越深,代表该点所对应的两行的相关性越强。如图 4 所示,etl1、etl2、el 层与 dtl2、dtl1、dl 层呈对称关系,数据自 etl1~el 层的变化或自 dtl2~dl 层的变化逐渐与压缩后的反馈向量接近,并且 el、dl 层直接与压缩完成的反馈向量相连。因此,当全连接层逐渐接近网络输入输出端时,其数据相关性变化趋势增强。在图 4 中存在一个例外,即 etl2 的输入矩阵相关性强于 etl1 层,这是因为 etl1 层的输出经过了 ReLU,使得 etl2 层的输入大部分为 0。根据以上推理,从 dtl2 层的输入到输出数据应属于“解压缩”的过程之一,其相关性应该增强,图 4 可印证。综上可推测,替换更接近网络输入输出端的全连接层性能较好。

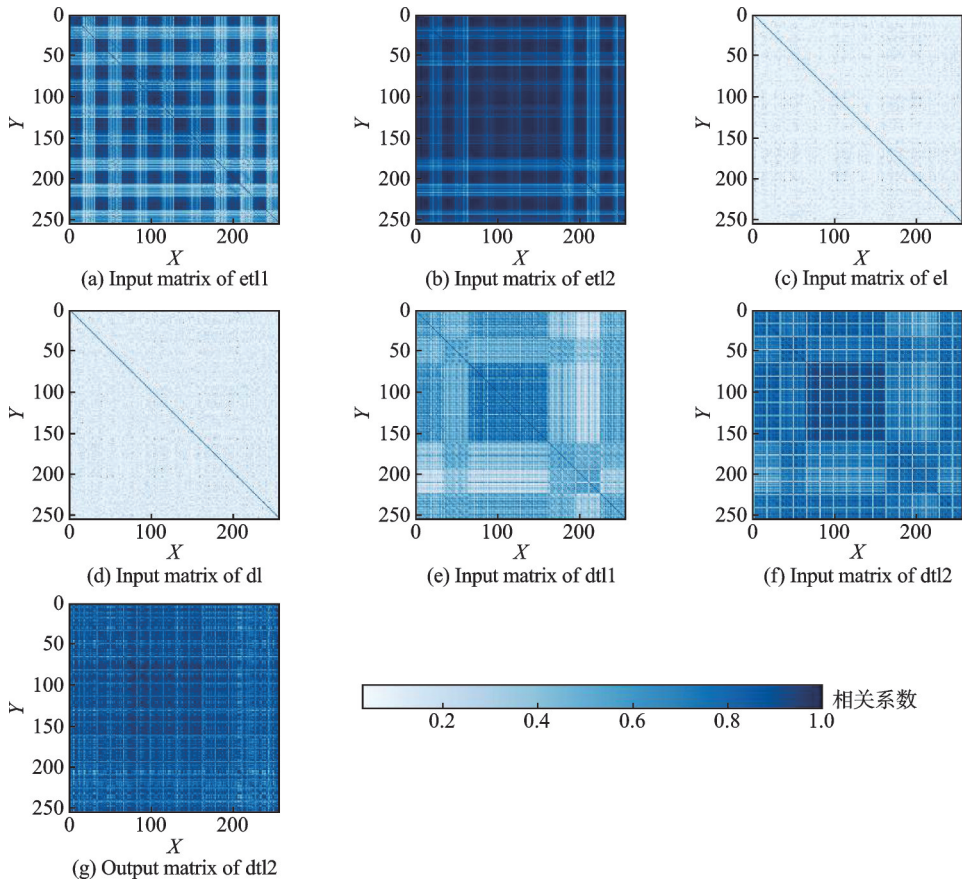


图 4 各全连接层矩阵行向量间相关系数热力图

Fig.4 Heat map of correlation coefficient between row vectors of matrices of each fully connected layer

3.2 PQ与MADDNESS方法的适用性比较

本文于1.2节指出PQ和MADDNESS方法针对不同场景的加速效果与复杂度存在差异。为测试两种AMM方法在CsiTransformer网络中的适用性,分别将两种方法应用于全部全连接层的计算简化。测试参数压缩比 $\eta=1/4$,其他参数与本文第3节初始所列相同,给出CSI反馈简化计算的NMSE性能、计算复杂度及存储复杂度的关系,如图5所示。压缩比为1/4时的性能曲线均标注了VQ(1,8)下的坐标值。结果表明,在VQ(n_c, n_e)相同的情况下,基于MADDNESS(图5~13中简称为MAD)的CSI反馈简化计算的性能总是优于基于PQ的方案,同时PQ方法的计算复杂度是MADDNESS的2.15倍~112.93倍不等。PQ的指令周期和为 $12DK+CM$,因此PQ的计算复杂度大小受质心数的影响很大,在质心数不少于128时,PQ方法的计算复杂度甚至会超过精确矩阵乘法。因为使用的LUT类似,它们的存储复杂度相当。

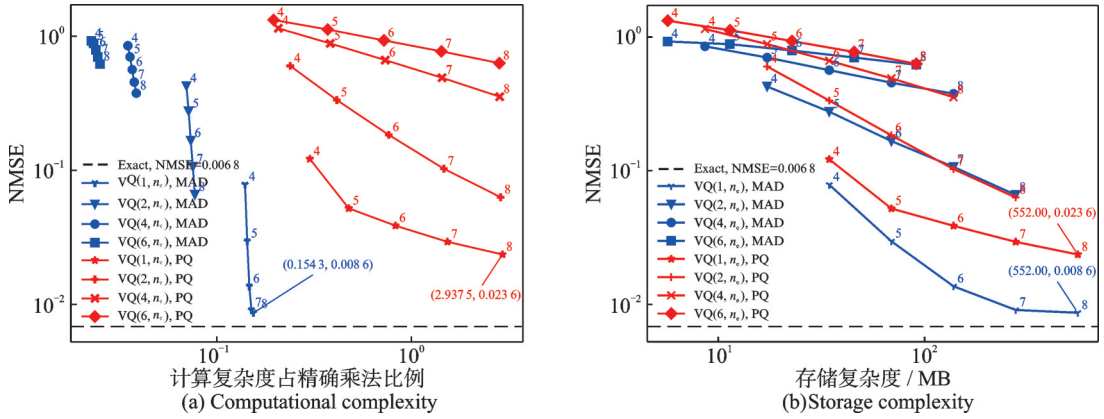


图5 基于PQ或MADDNESS的简化CSI反馈网络的NMSE与AMM复杂度

Fig.5 NMSE versus AMM complexity for simplified CSI feedback network based on PQ or MADDNESS

定义概念向量量化收益,即:在其他条件相同的情况下,增大 n_c ,出现计算复杂度减小且性能保持不变或获得提升的现象,则称该现象为向量量化收益。在基于PQ方法的简化CSI反馈网络中,未发现向量量化收益;在基于MADDNESS方法的简化CSI反馈网络中,VQ(2,8)相对于VQ(1,4)取得向量量化收益,VQ(4,8)相对于VQ(2,4)取得向量量化收益,VQ(6,6)~VQ(6,8)相对于VQ(4,4)取得向量量化收益。

综上所述,在纳入考虑的两种AMM方法中,仅MADDNESS方法表现出较好的性能,适用于CsiTransformer网络的计算简化。后续CSI反馈简化计算性能比较,VQ方法均采用MADDNESS。

3.3 基于MADDNESS的CSI反馈简化计算性能

压缩比 $\eta=1/4$ 时,CSI反馈的NMSE及余弦相似性的性能表现如图6、7所示。图6(a,b)分别给出了每种VQ(n_c, n_e)的计算复杂度占精确矩阵乘法计算复杂度的比例以及存储复杂度,其中VQ方法均采用MADDNESS算法。本次仿真分别测试了将全部6层全连接层进行简化计算后的性能,以及对etl1、etl2、dtl1、dtl2这4层全连接层进行简化计算后的性能。从4幅图中可以总结出以下要点。

(1) 在简化全部6层全连接层的情况下,VQ(1,8)时,CSI反馈的NMSE=0.0086, $\rho=0.9957$;在简化4层全连接层的情况下,VQ(1,8)时,CSI反馈的NMSE=0.0082, $\rho=0.9960$ 。它们均十分接近精确矩阵乘法的CSI反馈性能:NMSE=0.0068, $\rho=0.9965$,且AMM计算复杂度占各自对应全连接层精确矩阵乘法复杂度的比例分别为15.43%、16.25%。

(2) 无论用NMSE还是 ρ 衡量,在相同VQ(n_c, n_e)条件下,简化4层全连接层的性能总是优于简化6层全连接层的性能。由于 $\rho>0.9$ 时CSI反馈较为有价值,而简化全部6层全连接层后大部分仿真测试

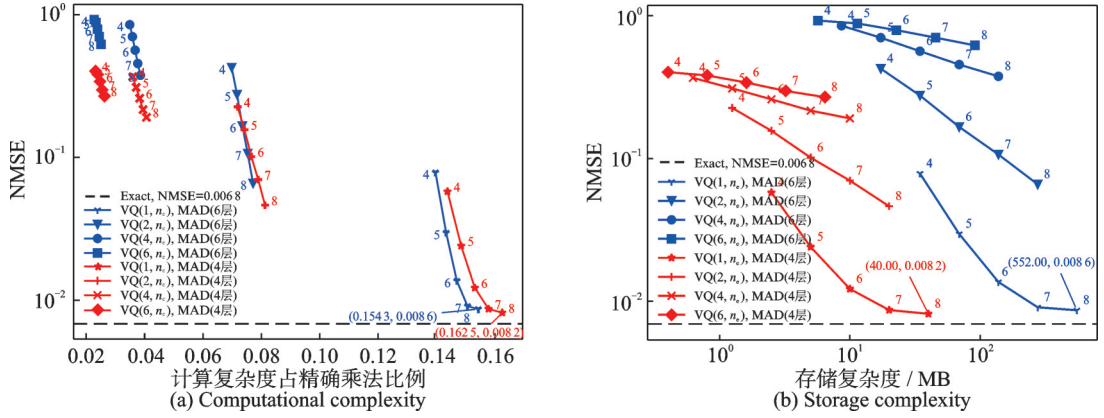


图6 压缩比为1/4时CSI反馈的NMSE与MADDNESS复杂度

Fig.6 NMSE versus MADDNESS complexity for CSI feedback at a compression ratio of 1/4

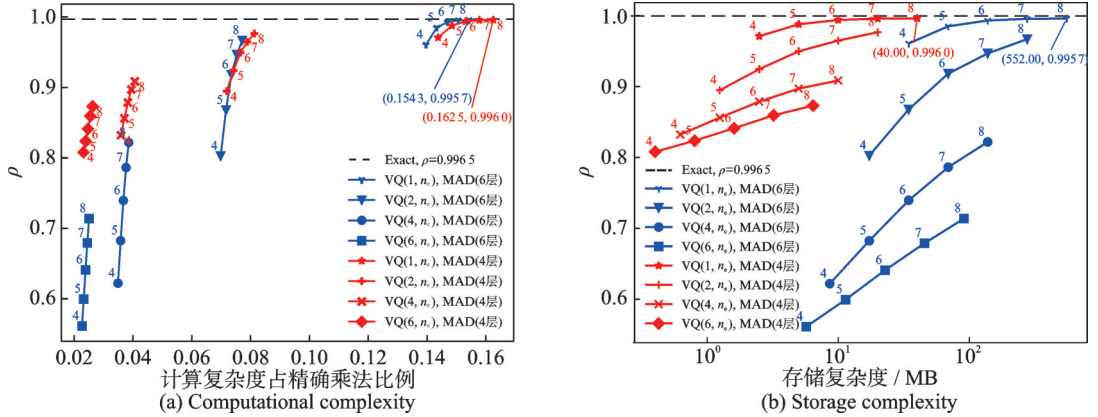


图7 压缩比为1/4时CSI反馈的余弦相似性与MADDNESS复杂度

Fig.7 Cosine similarity versus MADDNESS complexity for CSI feedback at a compression ratio of 1/4

结果均无法满足上述条件,因此简化4层全连接层可能更适用于实际应用。该仿真结果也符合3.1节中的推测。

(3) 若关注存储复杂度,可发现简化6层全连接层时所需存储空间在6~600 MB范围内,而简化4层全连接层时所需存储空间的范围仅为0.4~40 MB,后者仅为前者的6.67%。究其原因,因为简化6层全连接层时,el层输入特征维度较大且dl层神经元数目较多,待乘矩阵的尺寸(D 、 M)较大,导致LUT的占用空间过大。综合要点(2,3),本文更为推荐针对CsiTransformer网络简化4层全连接层。通过表4进行估算,以性能最优点VQ(1,8)为例,简化6层全连接层后整个网络的计算复杂度下降到原来的约38%,简化4层全连接层后整个网络的计算复杂度下降到原来的约57%。

(4) 以简化4层全连接层为例,VQ(2,8)相对于VQ(1,4)取得向量量化收益,VQ(4,7)和VQ(4,8)相对于VQ(2,4)取得向量量化收益,VQ(6,6)~VQ(6,8)相对于VQ(4,4)取得向量量化收益。简化6层全连接层的情况下的向量量化收益已于3.2节总结,此处不再赘述。

(5) 向量量化参数的选择涉及了简化计算CSI反馈网络的性能与计算复杂度及存储复杂度间的取舍。若追求接近精确矩阵乘法的CSI反馈性能,应在参数VQ(1, n_c)下简化CSI反馈网络;若追求向量量化收益带来的较低的复杂度,应在参数VQ($>1, n_c$)下简化CSI反馈网络。

图8~13给出压缩比 $\eta = 1/8, 1/16, 1/32$ 时CSI反馈性能与MADDNESS方法复杂度的关系图。所有图的性能最优点均在VQ(1,8)取得,并且在这3种压缩比下,VQ(1,8)的CSI反馈性能可以达到与

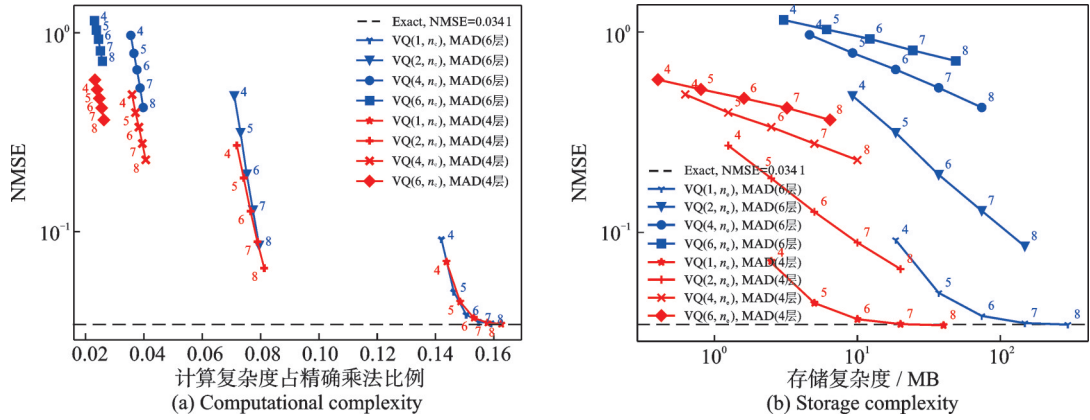


图8 压缩比为1/8时CSI反馈的NMSE与MADDNESS复杂度

Fig.8 NMSE versus MADDNESS complexity for CSI feedback at a compression ratio of 1/8

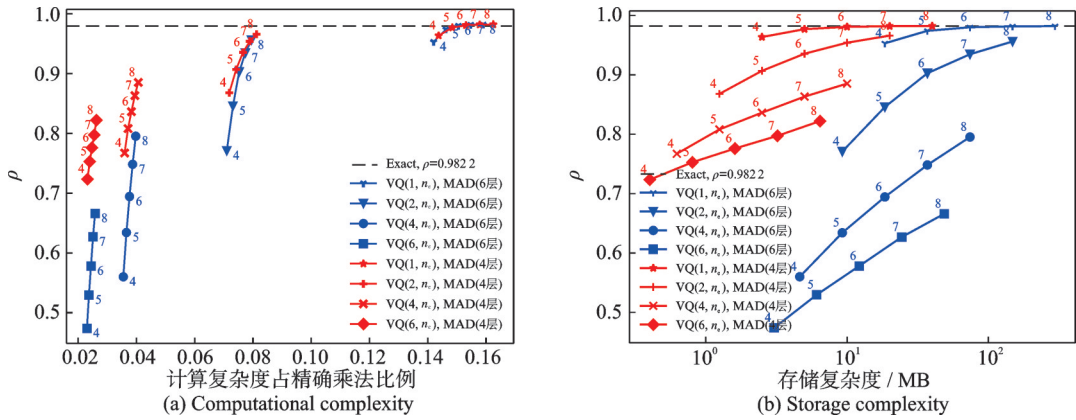


图9 压缩比为1/8时CSI反馈的余弦相似性与MADDNESS复杂度

Fig.9 Cosine similarity versus MADDNESS complexity for CSI feedback at a compression ratio of 1/8

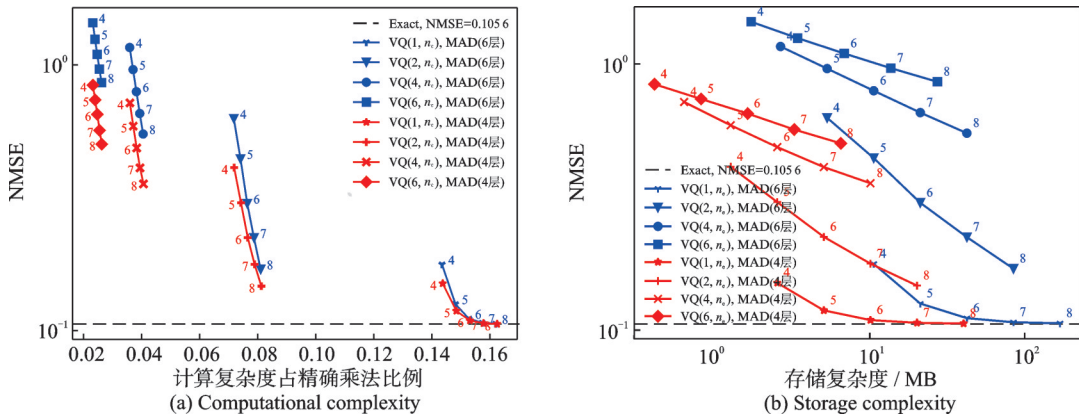


图10 压缩比为1/16时CSI反馈的NMSE与MADDNESS复杂度

Fig.10 NMSE versus MADDNESS complexity for CSI feedback at a compression ratio of 1/16

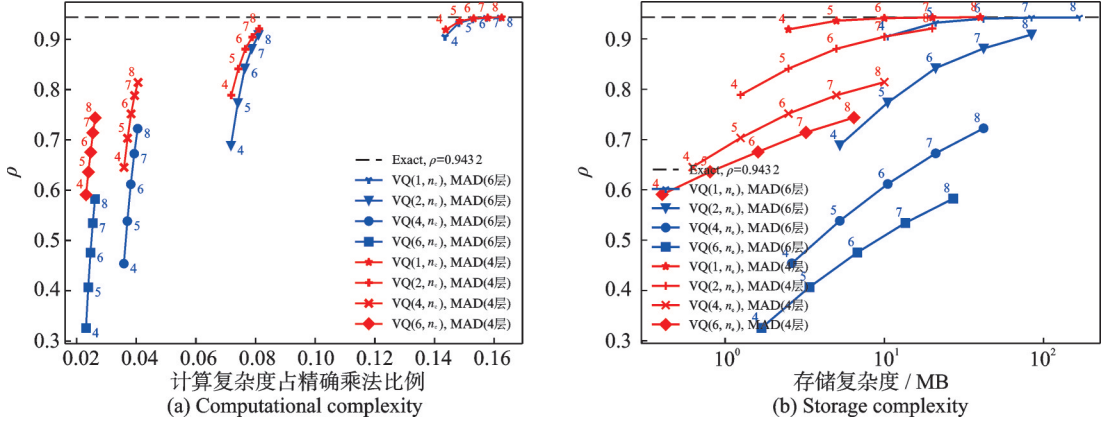


图 11 压缩比为 1/16 时 CSI 反馈的余弦相似性与 MADDNESS 复杂度

Fig.11 Cosine similarity versus MADDNESS complexity for CSI feedback at a compression ratio of 1/16

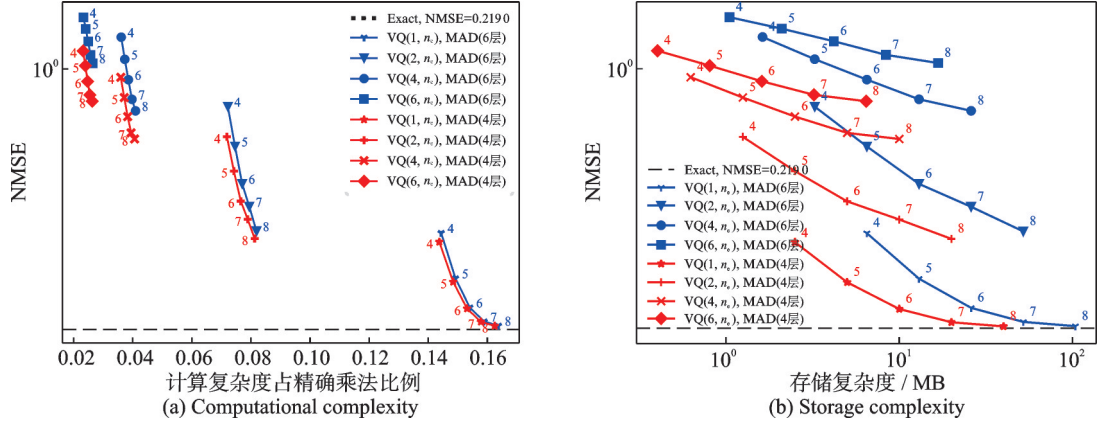


图 12 压缩比为 1/32 时 CSI 反馈的 NMSE 与 MADDNESS 复杂度

Fig.12 NMSE versus MADDNESS complexity for CSI feedback at a compression ratio of 1/32

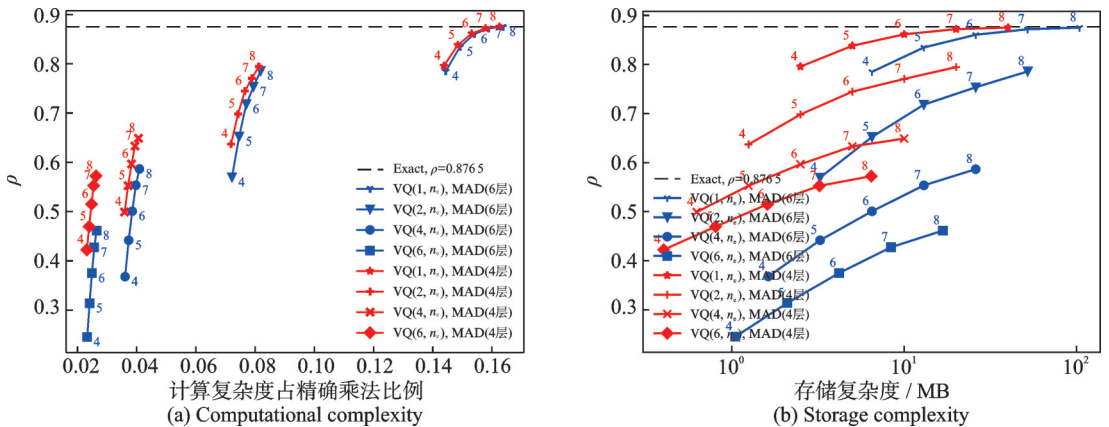


图 13 压缩比为 1/32 时 CSI 反馈的余弦相似性与 MADDNESS 复杂度

Fig.13 Cosine similarity versus MADDNESS complexity for CSI feedback at a compression ratio of 1/32

精确矩阵乘法相同的水平,即此时可在性能不下降的前提下降低精确矩阵乘法复杂度至17%以下。表5估算了简化全连接层后,不同压缩比下整个网络的计算复杂度占简化前计算复杂度的百分比。从表5中可以发现,随着压缩比的减小,针对6层全连接层的计算复杂度简化收益越来越小,简化4层的收益越来越大。观察结果,应用AMM至少将原复杂度降低了约43%,至多降低约62%。此外,从图8~13中依然可以发现与要点(4)类似的向量量化收益。

4 结束语

本文提出了一种新型的基于数据聚类处理的CSI反馈Transformer网络简化实现方法,显著降低了大规模MIMO系统中CSI反馈的复杂度开销。将经过针对性优化的基于数据聚类的AMM方法如MADNESS应用于神经网络后,不仅降低了计算复杂度,亦保持了CSI反馈Transformer网络的高性能。仿真结果及复杂度分析表明,在简化4层全连接层且压缩比为1/4时,CSI反馈的NMSE及余弦相似性(0.008 2、0.996 0)均十分接近简化前的性能(0.006 8、0.996 5),同时可将被简化层的计算复杂度降低约83%,整个网络的计算复杂度降低约43%,以较小的存储开销显著降低了计算复杂度。此外,Csi-Transformer网络中多头自注意力层存在一部分相较全连接层更复杂的矩阵乘法形式,如何简化这类矩阵乘法是未来的研究方向之一。

参考文献:

- [1] LU L, LI G Y, SWINDLEHURST A L, et al. An overview of massive MIMO: Benefits and challenges[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2014, 8(5): 742-758.
- [2] LARSSON E G, EDFORS O, TUFVESSON F, et al. Massive MIMO for next generation wireless systems[J]. *IEEE Communications Magazine*, 2014, 52(2): 186-195.
- [3] NGO H Q, LARSSON E G, MARZETTA T L. Energy and spectral efficiency of very large multiuser MIMO systems[J]. *IEEE Transactions on Communications*, 2013, 61(4): 1436-1449.
- [4] XU Q, JIANG C, HAN Y, et al. Waveforming: An overview with beamforming[J]. *IEEE Communications Surveys & Tutorials*, 2017, 20(1): 132-149.
- [5] BJÖRNSON E, LARSSON E G, MARZETTA T L. Massive MIMO: Ten myths and one critical question[J]. *IEEE Communications Magazine*, 2016, 54(2): 114-123.
- [6] KUO P H, KUNG H T, TING P A. Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays[C]//*Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*. Paris, France: IEEE, 2012: 492-497.
- [7] DAUBECHIES I, DEFRISE M, DE MOL C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint[J]. *Communications on Pure and Applied Mathematics*, 2004, 57(11): 1413-1457.
- [8] DONOHO D L, MALEKI A, MONTANARI A. Message-passing algorithms for compressed sensing[J]. *Proceedings of the National Academy of Sciences*, 2009, 106(45): 18914-18919.
- [9] LI C, YIN W, ZHANG Y. User's guide for TVAL3: TV minimization by augmented Lagrangian and alternating direction algorithms[J]. *CAAM Report*, 2009, 20(46/47): 4.
- [10] METZLER C A, MALEKI A, BARANIUK R G. From denoising to compressed sensing[J]. *IEEE Transactions on Information Theory*, 2016, 62(9): 5117-5144.
- [11] O'SHEA T J, ERPEK T, CLANCY T C. Deep learning based MIMO communications[EB/OL]. (2017-07-25)[2024-01-05].

表5 简化全连接层后整个网络的计算复杂度占简化前计算复杂度的百分比

	压缩比	1/4	1/8	1/16	1/32
简化6层全连接层		37.99	40.96	42.73	43.69
简化4层全连接层		56.14	51.01	47.96	46.30

<https://arxiv.org/abs/1707.07980>.

- [12] WEN C K, SHIH W T, JIN S. Deep learning for massive MIMO CSI feedback[J]. IEEE Wireless Communications Letters, 2018, 7(5): 748-751.
- [13] VASWANI A, BENGIO S, BREVDO E, et al. Tensor2Tensor for neural machine translation[EB/OL]. (2018-05-16)[2024-01-05]. <https://arxiv.org/abs/1803.07416>.
- [14] XU Y, YUAN M, PUN M O. Transformer empowered CSI feedback for massive MIMO systems[C]//Proceedings of Wireless and Optical Communications Conference (WOCC). Taipei, China: IEEE, 2021: 157-161.
- [15] HAN S, LIU X, MAO H, et al. EIE: Efficient inference engine on compressed deep neural network[J]. ACM SIGARCH Computer Architecture News, 2016, 44(3): 243-254.
- [16] HANIF M A, KHALID F, SHAFIQUE M. CANN: Curable approximations for high-performance deep neural network accelerators[C]//Proceedings of the 56th Annual Design Automation Conference 2019. New York, NY, USA: Association for Computing Machinery, 2019: 1-6.
- [17] TASOULAS Z G, ZERVAKIS G, ANAGNOSTOPOULOS I, et al. Weight-oriented approximation for energy-efficient neural network inference accelerators[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2020, 67(12): 4670-4683.
- [18] HAMMAD I, LI L, EL-SANKARY K, et al. CNN inference using a preprocessing precision controller and approximate multipliers with various precisions[J]. IEEE Access, 2021, 9: 7220-7232.
- [19] YU Q, MADDAH-ALI M A, AVESTIMEHR A S. Straggler mitigation in distributed matrix multiplication: Fundamental limits and optimal coding[J]. IEEE Transactions on Information Theory, 2020, 66(3): 1920-1933.
- [20] JIA Z, JAFAR S A. Cross subspace alignment codes for coded distributed batch computation[J]. IEEE Transactions on Information Theory, 2021, 67(5): 2821-2846.
- [21] DAS A B, RAMAMOORTHY A. Coded sparse matrix computation schemes that leverage partial stragglers[J]. IEEE Transactions on Information Theory, 2022, 68(6): 4156-4181.
- [22] JEGOU H, DOUZE M, SCHMID C. Product quantization for nearest neighbor search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(1): 117-128.
- [23] BLALOCK D, GUTTAG J. Multiplying matrices without multiplying[C]//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 2021: 992-1004.
- [24] GRAY R M, NEUHOFF D L. Quantization[J]. IEEE Transactions on Information Theory, 1998, 44(6): 2325-2383.
- [25] INTEL. Intel® Intrinsic Guide[EB/OL]. (2023-07-12)[2024-01-05]. <https://www.intel.com/content/www/us/en/docs/intrinsics-guide/index.html>.

作者简介:



还冬锐(1998-),男,硕士研究生,研究方向:通信信号处理和人工智能辅助通信, E-mail: drhuan@seu.edu.cn。



张逸帆(1999-),男,硕士研究生,研究方向:编码和人工智能辅助通信, E-mail: zhangyifan@seu.edu.cn。



姜明(1976-),通信作者,男,副研究员,博士生导师,研究方向:编码和调制技术, E-mail: jiang_ming@seu.edu.cn。

(编辑:陈璐)