



## 针对低比特Transformer的无约束训练后量化

江佳浩 尹鹏 王轩瀚 曾鹏鹏 宋井宽

### Constraint-free Post-training Quantization for Low-bit Transformer

JIANG Jia-Hao, YIN Peng, WANG Xuan-Han, ZENG Peng-Peng, SONG Jing-Kuan

在线阅读 View online:

<http://www.aas.net.cn/article/current/f91e80fe684cf02a02b40eb0a178ad4802cd4ec96e859642eac4f8b78892a205>

---

## 您可能感兴趣的其他文章

### 基于自注意力模态融合网络的跨模态行人再识别方法研究

Self-attention Cross-modality Fusion Network for Cross-modality Person Re-identification

自动化学报. 2022, 48(6): 1457–1468 <https://doi.org/10.16383/j.aas.c190340>

### 基于自注意力对抗的深度子空间聚类

Self-attention Adversarial Based Deep Subspace Clustering

自动化学报. 2022, 48(1): 271–281 <https://doi.org/10.16383/j.aas.c200302>

### 基于重组性高斯自注意力的视觉Transformer

Vision Transformer Based on Reconfigurable Gaussian Self-attention

自动化学报. 2023, 49(9): 1976–1988 <https://doi.org/10.16383/j.aas.c220715>

### 融合自注意力机制和相对鉴别的无监督图像翻译

Unsupervised Image-to-Image Translation With Self-Attention and Relativistic Discriminator Adversarial Networks

自动化学报. 2021, 47(9): 2226–2237 <https://doi.org/10.16383/j.aas.c190074>

### 联合深度超参数卷积和交叉关联注意力的大位移光流估计

Large Displacement Optical Flow Estimation Jointing Depthwise Over-parameterized Convolution and Cross Correlation Attention

自动化学报. 2024, 50(8): 1631–1645 <https://doi.org/10.16383/j.aas.c230049>

### 多维注意力特征聚合立体匹配算法

Multi-dimensional Attention Feature Aggregation Stereo Matching Algorithm

自动化学报. 2022, 48(7): 1805–1815 <https://doi.org/10.16383/j.aas.c200778>

# 针对低比特 Transformer 的无约束训练后量化

江佳浩<sup>1</sup> 尹鹏<sup>2</sup> 王轩瀚<sup>3</sup> 曾鹏鹏<sup>3</sup> 宋井宽<sup>3</sup>

**摘要** Transformer 模型具有庞大的参数规模和计算开销,难以直接部署于资源受限的边缘设备,限制了其在开放环境中的实际应用。低比特量化能够显著降低模型存储需求并提升推理效率,是实现 Transformer 模型轻量化部署的重要技术路径。然而,现有量化方法通常受固定量化区间约束,在处理非线性激活产生的长尾分布时,难以有效平衡裁剪误差与舍入误差,导致模型性能明显下降。为此,提出一种面向低比特 Transformer 的无约束训练后量化方法 (CFQuant),通过建模神经网络激活值的概率分布实现自适应量化。首先,CFQuant 在校准过程中自适应估计激活分布密度,并通过迭代搜索最小量化误差。其次,设计高效尺度偏移算法动态调整模型分布,以减小校准阶段与推理阶段之间的分布偏移,进一步提升量化稳定性。此外,构建基于查找表的矩阵乘法加速机制,将低比特推理中的浮点乘法转换为预计算查找操作,从而降低推理计算开销。最后,通过在多类 Transformer 模型以及视觉、语言和多模态任务上的大量实验,充分验证了所提方法的灵活性和有效性。

**关键词** 深度学习; 训练后量化; Transformer; 量化误差; 自注意力

**引用格式** 江佳浩, 尹鹏, 王轩瀚, 曾鹏鹏, 宋井宽. 针对低比特 Transformer 的无约束训练后量化. 自动化学报, xxxx, xx(x): x-xx

**DOI** 10.16383/j.aas.c250702 **CSTR** 32138.14.j.aas.c250702

## Constraint-free Post-training Quantization for Low-bit Transformer

JIANG Jia-Hao<sup>1</sup> YIN Peng<sup>2</sup> WANG Xuan-Han<sup>3</sup> ZENG Peng-Peng<sup>3</sup> SONG Jing-Kuan<sup>3</sup>

**Abstract** The massive parameter scales and high computational overhead of Transformer models make them difficult to deploy directly on resource-constrained edge devices, limiting their practical applications in open environments. Low-bit quantization can significantly reduce model storage requirements and improve inference efficiency, making it an important technical approach for lightweight deployment of Transformer models. However, existing quantization methods are usually constrained by fixed quantization intervals. When processing long-tailed distributions produced by nonlinear activations, they have difficulty balancing clipping and rounding errors effectively, leading to considerable performance degradation. To address this problem, we propose a constraint-free post-training quantization method for low-bit Transformer (CFQuant), which performs adaptive quantization by modeling the probability distributions of neural-network activations. First, CFQuant adaptively estimates activation-distribution density during calibration and minimizes quantization error through iterative search. Second, an efficient scale-shift algorithm dynamically adjusts model distributions to reduce the distribution shift between calibration and inference stages, thereby further improving quantization stability. Moreover, an acceleration mechanism based on matrix multiplication with a lookup table converts floating-point multiplications in low-bit inference into precomputed lookup operations, thereby reducing inference computational overhead. Finally, extensive experiments on multiple types of Transformer models and on vision, language, and multimodal tasks demonstrate the flexibility and effectiveness of the proposed method.

**Keywords** deep learning; post-training quantization; Transformer; quantization error; self-attention

**Citation** Jiang Jia-Hao, Yin Peng, Wang Xuan-Han, Zeng Peng-Peng, Song Jing-Kuan. Constraint-free post-training quantization for low-bit Transformer. *Acta Automatica Sinica*, xxxx, xx(x): x-xx

收稿日期 2025-12-04 录用日期 2026-06-24

Manuscript received December 4, 2025; accepted June 24, 2026  
教育部学科突破先导项目 (JYB2025XDXM103), 国家自然科学基金 (62425208, U22A2097, 62402094, 62502076) 资助

Supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103) and National Natural Science Foundation of China (62425208, U22A2097, 62402094, 62502076)

本文责任编辑 鲁继文

Recommended by Associate Editor LU Ji-Wen

1. 同济大学电子与信息工程学院 上海 201804 2. 腾讯科技 (深圳)

近年来, 基于 Transformer 的模型在多种人工智能任务中表现出优异效果, 包括视觉识别<sup>[1-2]</sup>、语言生成<sup>[3-5]</sup>乃至具身智能领域机器人的感知与决策<sup>[6-8]</sup>。

有限公司 深圳 518054 3. 同济大学计算机科学与技术学院 上海 201804

1. College of Electronic and Information Engineering, Tongji University, Shanghai 201804 2. Tencent Technology (Shenzhen) Co., Ltd., Shenzhen 518054 3. School of Computer Science and Technology, Tongji University, Shanghai 201804

然而, Transformer 模型通常需要大量的计算资源,这不仅导致内存占用高、计算功耗上升,还会增加推理延迟<sup>[9]</sup>. 比如,在具身智能场景中,基于 Transformer 的视觉语言动作模型给机器人带来强大的零样本推理能力,但是庞大的参数量会导致对推理速度和计算资源的要求更高,是制约其广泛应用所面临的严峻挑战<sup>[10]</sup>. 因此,在资源受限的边缘设备上,或在对推理性能有严苛要求的场景中,部署此类模型面临显著挑战,这也限制了人工智能技术在更广泛领域的应用与发展.

为解决这些效率瓶颈,学术界和工业界广泛研究模型压缩技术,如蒸馏<sup>[11-12]</sup>、剪枝<sup>[13-14]</sup>和量化(quantization)<sup>[15-17]</sup>. 其中,模型量化展现了巨大的优势,尤其是在低成本部署和实时推理至关重要的实际场景中. 模型量化通过将权重和激活值压缩为较少比特的表示,从而提高计算效率. 训练后量化(post-training quantization, PTQ)是最常用的量化方法之一,其优点在于仅需少量未标记的样本进行量化参数校准,而无需在庞大的训练集上重新训练量化模型,可以极大提高模型量化的效率<sup>[16, 18-20]</sup>. 训练后量化中,常见的量化方法包括均匀量化<sup>[21]</sup>和 Log2 量化<sup>[22]</sup>,这类方法简化了量化过程并提高了推理效率. 其中,均匀量化要求量化区间宽度相等,而 Log2 量化则要求量化区间满足倍增关系. 考虑到量化区间的约束,本文将这些主流量化方法统称为受约束的量化.

受约束的量化方法虽然能够对 Transformer 模型参数进行有效量化,但由于神经网络内在的部分层所计算的激活值具有长尾特性的分布,使得这些量化方法所产生的最终量化效果较差,尤其是在对 Softmax 和 GELU 层的激活值进行量化时. 具体而言,在量化过程中的裁剪误差与舍入误差的权衡困难,导致较大的量化误差(quantization error, QE). 这一问题在于受约束量化方法的表示能力有限,特别是在低比特场景下. 为构建更具灵活性和更强大的量化器,消除这些约束至关重要.

为此,本文提出一种无约束训练后量化方法(constraint-free post-training quantization, CFQuant),该方法根据输入激活的分布特征自适应地进行量化,从而显著降低量化误差. 首先,CFQuant 去除量化区间的约束,避免裁剪误差与舍入误差之间的权衡. 随后,在校准过程中,采用 Lloyd-Max 算法<sup>[23-24]</sup>通过迭代优化量化区间,不断适应输入激活值的分布特性. 在推理阶段(inference stage),CFQuant 无需进行任何迭代优化,从而避免了额外的时间开销和成本. CFQuant 使用校

准过程中获得的量化输出,并设计高效的尺度偏移算法,以减少推理阶段与校准阶段(calibration stage)之间的分布偏移. 最后,为进一步提高无约束训练后量化的推理效率,本文设计基于查找表的矩阵乘法(matrix multiplication with lookup table, MM-LUT),以加速模型推理中的量化矩阵乘法. 在模型校准后,预计算每个可能的量化权重与量化输出之间的乘法值,在推理时,通过查找表矩阵乘法将复杂的浮点乘法操作转换为查找操作,从而显著加速了模型推理过程.

本文的主要贡献如下:

1) 揭示主流量化器存在的本质局限性,特别是其受限的量化区间导致了显著的量化误差,这一问题在 Softmax 输出和 GELU 输出等非线性激活中尤为突出;

2) 为缓解这一问题,提出无约束训练后量化方法(CFQuant),该方法能够基于激活分布进行迭代优化,同时设计高效尺度偏移算法以动态调整模型分布,并利用基于查找表的矩阵乘法加速量化矩阵乘法;

3) 在多种 Transformer 主干网络以及视觉、语言和多模态任务上进行的广泛实验,充分验证了 CFQuant 的灵活性、有效性和跨架构适用性.

## 1 相关工作

### 1.1 Transformer 架构

Transformer 模型<sup>[25]</sup>凭借其强大的注意力机制,在自然语言处理领域取得了显著成功. 后续的工作如 BERT<sup>[26]</sup>和 GPT<sup>[27]</sup>等,进一步扩展了 Transformer 架构,证明了其在理解和生成任务中的通用性. 受其捕捉长程依赖能力的启发,视觉 Transformer (vision Transformer, ViT)<sup>[1]</sup>将注意力机制应用于计算机视觉领域. 在此基础上,数据高效图像 Transformer (data-efficient image Transformer, DeiT)<sup>[28]</sup>进一步提高了模型训练的数据效率. ViT 将输入图像划分为图像块(patch),将这些图像块嵌入为词元(token),随后利用 Transformer 架构进行处理. ViT 在各类下游任务中展现出优异的性能,包括图像分类、目标检测<sup>[9, 29]</sup>以及图像分割<sup>[30]</sup>. 近年来,基于 Transformer 架构,拥有数十亿参数的大语言模型 (large language models, LLMs) 展现出卓越的零样本 (zero-shot) 能力,并在众多语言生成基准测试中占据了主导地位. 尤其是 GPT-3<sup>[31]</sup>的问世,凭借其在广泛的自然语言处理任务中卓越的少样本 (few-shot) 和零样本性能,重塑了

深度学习与人工智能的发展格局. 与此同时, LLaMA<sup>[5]</sup> 推出了一系列兼具性能与效率的开源模型, 极大地推动了指令微调 and 模型蒸馏的相关研究.

## 1.2 模型量化

量化作为深度神经网络模型压缩和加速的关键技术之一, 通过降低模型参数的数值精度来优化计算效率和存储开销, 同时尽可能保持模型的性能表现. 量化将高精度的浮点数 (如 32 位浮点数) 转换为低精度的定点数 (如 8 位整数), 从而显著减少模型的内存占用和计算资源需求, 为在资源受限设备上的部署提供解决方案<sup>[17-18, 32]</sup>. 现有的量化方法主要归类为量化感知训练 (quantization-aware training, QAT) 和训练后量化 (PTQ). 训练后量化方法无需使用训练数据集进行重新训练, 因而在低成本和实际部署场景中更受青睐. 训练后量化方法通常围绕量化误差最小化展开, 相关研究<sup>[18, 33-37]</sup> 通过块级重构、量化参数校准或旋转变换等策略降低量化误差. 另外有研究<sup>[19-20, 38]</sup> 发现, 现有的量化器在处理某些不平衡分布时会导致性能次优, 并据此提出替代的量化器设计.

大语言模型通常包含数十亿甚至数千亿个参数, 导致其计算和存储开销巨大, 对内存容量、显存规模和计算资源等硬件条件提出了更高要求, 量化任务也变得愈发具有挑战性, 一些研究将训练后量化技术扩展至大语言模型. 例如, SmoothQuant<sup>[15]</sup> 在保持均匀量化器形式不变的前提下, 通过对激活与权重进行尺度重分配, 将激活中的异常值 (outliers) 转移至权重侧, 从而降低激活分布的动态范围, 使其更适配均匀量化器. 该策略的核心思想是在不改变量化器形式的条件下调整数据分布, 在 8 比特等中高比特设置下能够有效缓解激活分布不均带来的量化困难. OmniQuant<sup>[36]</sup> 则在此基础上进一步引入可学习的校准机制, 通过对量化参数进行优化, 降低跨层量化敏感性差异和误差累积问题. 这类方法仍然建立在均匀量化框架之上, 其改进主要体现在量化参数的自适应调整, 而非量化器结构本身的改变.

上述方法隐含的一个假设在于: 在给定比特宽度下, 均匀量化器具有足够的表达能力, 主要误差来源在于分布不匹配或尺度不平衡. 然而, 在低比特场景下, 量化误差的主要来源发生了变化. 此时量化分辨率显著受限, 模型性能受限的关键因素不再仅是异常值引起的动态范围扩张, 而更多体现在有限刻度对复杂分布的表达能力不足. 对于具有长尾特征和较高信息熵的 Transformer 激活分布而

言, 即使通过平滑或尺度重分配缩小动态范围, 均匀量化器仍难以充分刻画其概率密度结构, 从而导致表示能力受限.

基于这一观察, 本文在训练后量化框架内进一步探讨低比特条件下量化器表达能力的限制问题. 与通过调整数据分布以适配固定量化器的不同, 本文聚焦量化器结构本身的适配性问题, 以缓解在低比特场景下由分辨率受限带来的性能退化.

## 2 问题描述与量化误差分析

### 2.1 Transformer 结构

本文以 ViT<sup>[1]</sup> 为例介绍 Transformer 结构. 首先, 输入图像经由嵌入层 (embedding layer) 投影为长度为  $N$  的  $D$  维向量序列, 记作  $\mathbf{X}_0 \in \mathbf{R}^{N \times D}$ . 忽略批次维度, 随后,  $\mathbf{X}_0$  被送入堆叠的 Transformer 模块中, 每个模块由多头自注意力 (multi-head self-attention, MSA) 单元和多层感知机 (multilayer perceptron, MLP) 单元组成. 为简化起见, 忽略层归一化 (layer normalization, LayerNorm), Transformer 模块的计算公式如下:

$$\mathbf{Y}_{l-1} = \text{MSA}(\mathbf{X}_{l-1}) + \mathbf{X}_{l-1} \quad (1)$$

$$\mathbf{X}_l = \text{MLP}(\mathbf{Y}_{l-1}) + \mathbf{Y}_{l-1} \quad (2)$$

其中,  $l = 1, 2, \dots, L$ ,  $L$  表示 Transformer 模块的数量.

MSA 模块通过以下方式学习输入  $\mathbf{X}' \in \mathbf{R}^{N \times D}$  图像块间的相关性:

$$[\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i] = \mathbf{X}' \mathbf{W}^{\text{qkv}} + \mathbf{b}^{\text{qkv}} \quad (3)$$

$$\text{Attn}_i = \text{Softmax} \left( \frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{D_h}} \right) \mathbf{V}_i \quad (4)$$

$$\text{MSA}(\mathbf{X}') = [\text{Attn}_1, \text{Attn}_2, \dots, \text{Attn}_h] \mathbf{W}^{\text{out}} + \mathbf{b}^{\text{out}} \quad (5)$$

其中,  $i = 1, 2, \dots, h$ ,  $h$  为注意力头的数量;  $\mathbf{Q}_i$ 、 $\mathbf{K}_i$  和  $\mathbf{V}_i$  分别表示第  $i$  个注意力头的查询、键和值矩阵;  $\mathbf{W}^{\text{qkv}} \in \mathbf{R}^{D \times 3D_h}$ ;  $\mathbf{b}^{\text{qkv}} \in \mathbf{R}^{3D_h}$ ;  $\mathbf{W}^{\text{out}} \in \mathbf{R}^{hD_h \times D}$ ;  $\mathbf{b}^{\text{out}} \in \mathbf{R}^D$ ;  $D_h$  为每个注意力头的特征维度;  $\text{Attn}_i \in \mathbf{R}^{N \times D_h}$  表示第  $i$  个注意力头的输出矩阵.

MLP 模块将特征投影至更高维度的  $D'$  空间以学习特征表示. 设 MLP 模块的输入为  $\mathbf{Y}' \in \mathbf{R}^{N \times D}$ , 其计算过程如下:

$$\text{MLP}(\mathbf{Y}') = \text{GELU}(\mathbf{Y}' \mathbf{W}^{\text{fc1}} + \mathbf{b}^{\text{fc1}}) \mathbf{W}^{\text{fc2}} + \mathbf{b}^{\text{fc2}} \quad (6)$$

其中,  $\mathbf{W}^{\text{fc1}} \in \mathbf{R}^{D \times D'}$ ;  $\mathbf{b}^{\text{fc1}} \in \mathbf{R}^{D'}$ ;  $\mathbf{W}^{\text{fc2}} \in \mathbf{R}^{D' \times D}$ ;  $\mathbf{b}^{\text{fc2}} \in \mathbf{R}^D$ .

## 2.2 量化目标函数

给定一个输入分布  $X$ , 其概率密度函数为  $p(x)$ , 假设存在  $M$  个量化值  $a_k, k = 1, 2, \dots, M$ , 以及  $M+1$  个决策边界  $b_k, k = 0, 1, \dots, M$ . 如果样本值  $x \in X$  落入第  $k$  个量化区间  $[b_{k-1}, b_k)$ , 则该样本被量化为量化索引  $k$ . 通过反量化操作, 可由量化索引重构出量化值. 量化的操作如下所示:

$$\begin{cases} x_q = Q(x) = k, & x \in [b_{k-1}, b_k) \\ \hat{x}_q = Q^{-1}(x_q) = a_k, & x \in [b_{k-1}, b_k) \end{cases} \quad (7)$$

其中,  $x$  和  $x_q$  分别表示输入浮点值和量化后的整数值;  $\hat{x}_q$  表示反量化后的重构值;  $Q(\cdot)$  和  $Q^{-1}(\cdot)$  分别表示量化函数和反量化函数.

量化不可避免地会导致一定程度的信息丢失, 因为  $\hat{x}_q$  与原始输入  $x$  存在差异. 这种信息丢失可用量化误差来表征:

$$\sigma^2 = \int (\hat{x}_q - x)^2 p(x) dx = \sum_{k=1}^M \int_{b_{k-1}}^{b_k} (\hat{x}_q - x)^2 p(x) dx \quad (8)$$

其中,  $\sigma^2$  表示均方量化误差. 通常, 一个较好的模型量化方法能够在降低模型表征比特数的同时, 减小量化误差, 并且降低模型的性能损失, 训练后量化的优化目标是:

$$\arg \min_{\{a_k, b_k\}} \sum_{k=1}^M \int_{b_{k-1}}^{b_k} (\hat{x}_q - x)^2 p(x) dx \quad (9)$$

## 2.3 受限量化器及其局限性

本节对现有受限量化方法产生量化误差的原因和局限性进行分析. 在训练后量化中, 最常用的量化器是均匀量化器和 Log2 量化器. 均匀量化要求量化区间的间隔保持一致, 公式为:

$$\begin{cases} Q(x) = \text{clip}\left(\left\lfloor \frac{x}{s} \right\rfloor + z, 0, 2^b - 1\right) \\ Q^{-1}(x_q) = s(x_q - z) \end{cases} \quad (10)$$

其中,  $\text{clip}(\cdot)$  表示将输入截断至给定区间的函数;  $\lfloor \cdot \rfloor$  表示四舍五入取整函数;  $b \in \mathbf{N}$  为量化的比特宽度;  $s \in \mathbf{R}^+$  表示量化步长 (或缩放因子);  $z \in \mathbf{Z}$  表示量化零点.  $s$  和  $z$  的计算方式如下:

$$\begin{cases} s = \frac{\max(x) - \min(x)}{2^b - 1} \\ z = \left\lfloor -\frac{\min(x)}{s} \right\rfloor \end{cases} \quad (11)$$

Log2 量化要求量化区间满足二倍的关系, 公

式为:

$$\begin{cases} Q(x) = \text{clip}\left(\left\lfloor -\log_2 \frac{x}{s} \right\rfloor, 0, 2^b - 1\right) \\ Q^{-1}(x_q) = s2^{-x_q} \end{cases} \quad (12)$$

Log2 量化器通常以零点为起始, 因此此处舍弃了零点参数.

在量化后, 整数乘法和底数为 2 的幂运算可以通过整数运算和移位操作高效实现<sup>[39-40]</sup>. 为寻找这些量化器中的最优量化参数  $z$  和  $s$ , 模型量化的优化目标设定为:

$$\min_{z, s} \int (\hat{x}_q - x)^2 p(x) dx \quad \text{s.t. 式 (10) 或式 (12)} \quad (13)$$

由于均匀量化和 Log2 量化对量化区间有限制, 所有这些量化器中仅优化缩放因子  $s$  和零点  $z$ , 而非针对  $\{a_k\}$  和  $\{b_k\}$  进行调整. 虽然这种方法简化了量化过程, 并且对于类高斯分布的情况效果良好, 但未能有效量化非线性和长尾分布的激活值. 图 1 展示不同模块约束量化的可视化. 图 1 中垂直虚线表示量化值, 与高斯分布相比, Softmax 输出和 GELU 输出的激活呈现长尾分布且不平衡, 其量化误差分别为 0.20 和 0.22, 几乎是高斯分布激活值的两倍.

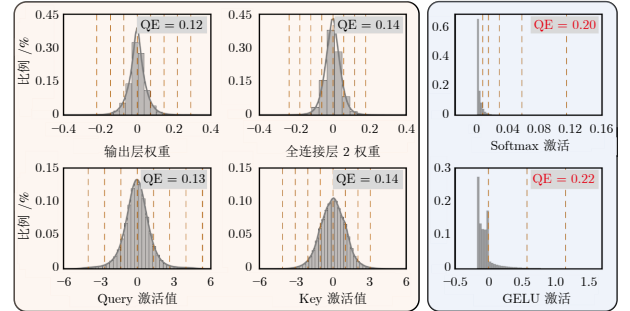


图 1 不同分布的量化误差对比

Fig. 1 Comparison of quantization errors for different distributions

根据率失真理论<sup>[41]</sup>, 在量化比特数固定的情况下, 对量化形式施加进一步的限制将提高量化误差的下界. 在均匀量化中, 假设量化值  $a_i$  和量化步长  $s$  已确定, 则  $a_j = a_i + s(j - i)$  是确定的. 然而当  $a_i$  和  $a_j$  受到量化步长的约束时, 会引发裁剪误差和舍入误差之间的权衡问题<sup>[19]</sup>. 如图 2 所示, 在 GELU 后的激活值的量化中, 由于均匀量化器的约束, 一个较大的量化步长  $s$  虽然能减小裁剪误差, 但会引入更多的舍入误差. 相反, 一个较小的  $s$  会减少舍入误差, 却会引入更多的裁剪误差. 由于均匀量化器和 Log2 量化器的约束, 舍入误差和裁剪误差难以平衡, 导致较大的量化误差.

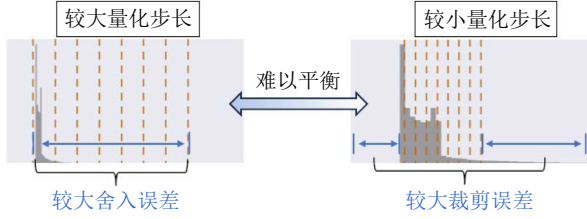


图 2 舍入误差和裁剪误差

Fig.2 Rounding error and clipping error

进一步的实验证明, 对具有较大量化误差的激活值进行量化, 会严重损害模型性能. 如图 3 所示, 将不同模块的激活保持为全精度, 并与全量化基准进行比较. 图 3 中的 Attn.Q、Attn.K 和 Attn.V 分别表示查询、键和值对应的注意力激活. 结果表明, 将 GELU 后的激活值保留为全精度带来了最显著的提升, 使 ViT 性能提高了 7.14%; 将 Softmax 后激活值保留为全精度也带来了 1.90% 的提升. 相反, 将量化误差较小的激活值保留为全精度并未带来显著的性能提升.

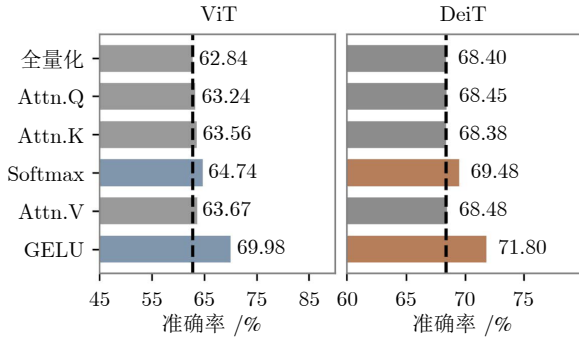


图 3 保留不同激活值为全精度的性能比较

Fig.3 Performance comparison when different activations are retained at full precision

因此, 减少这些非线性和长尾激活值的量化误差对于提高模型性能至关重要. 由于现有的量化器限制了量化区间并且未能有效量化这些激活, 本文提出一种无约束量化器, 能够根据激活值分布的密度自适应地估计和拟合激活分布, 从而使得量化后的分布保留更多原始分布的特征, 进而降低量化误差, 提高模型性能.

### 3 针对低比特 Transformer 的无约束训练后量化方法

为解决第 2 节中分析的受限量化器在处理长尾分布时的局限性, 本文提出无约束训练后量化方法 (CFQuant), 该方法主要针对 Transformer 中具有长尾分布特征的中间激活值, 如 Softmax 输出的注

意力权重和 GELU 输出的激活值. 均匀量化和 Log2 量化由于受到量化区间的约束, 导致较大的量化误差, 而 CFQuant 能够依据数据分布特征自适应调整量化步长, 在数据高频分布区间实现更细粒度的离散化表示, 从而更好地保留了激活值的分布特征, 降低量化失真. 实验表明, CFQuant 使 Softmax 后激活值的量化误差平均降低了 30%, GELU 后激活值的量化误差平均降低了 52%. 下面对无约束训练后量化方法进行介绍.

#### 3.1 无约束量化器

图 4 展示无约束训练后量化的整体框架, 对 Softmax 输出和 GELU 输出的激活值应用无约束量化. 图 4 中,  $Q$ 、 $K$  和  $V$  分别表示查询、键和值矩阵, MatMul 表示矩阵乘法, Proj 表示线性投影层, FC1 和 FC2 分别表示 MLP 模块中的第 1 个和第 2 个全连接层. 现有常见的量化方法对量化区间施加特定约束, 从而限制量化误差的优化空间, 影响模型量化的效果, 导致较大的量化误差. 本文提出的无约束训练后量化方法的核心在于优化目标的调整, 即通过移除预设的量化区间约束, 根据激活分布优化可变区间, 从而更好地保留分布特征并显著降低量化误差, 提升模型的性能. 其优化目标可表示为:

$$\min_{\{a_k, b_k\}} \sigma^2 = \sum_{k=1}^M \int_{b_{k-1}}^{b_k} (\hat{x}_q - x)^2 p(x) dx \quad (14)$$

与均匀量化和 Log2 量化不同, 该方法不对量化形式做任何限制. 为寻找最优解, 采用 Lloyd-Max 算法<sup>[23-24]</sup>进行迭代求解. Lloyd-Max 算法通过交替优化量化值  $a_k$  和量化边界  $b_k$ , 交替优化直至收敛, 最终达到最小化均方误差的效果. 无约束训练后量化方法具体求解过程如下:

首先计算  $a_k$  的偏导数:

$$\begin{aligned} \frac{\partial \sigma^2}{\partial a_k} &= 2 \int_{b_{k-1}}^{b_k} (a_k - x) p(x) dx = \\ &= 2a_k \int_{b_{k-1}}^{b_k} p(x) dx - 2 \int_{b_{k-1}}^{b_k} x p(x) dx \end{aligned} \quad (15)$$

然后, 将其设为零可得:

$$a_k = \frac{\int_{b_{k-1}}^{b_k} x p(x) dx}{\int_{b_{k-1}}^{b_k} p(x) dx} \quad (16)$$

这表明最优量化值  $a_k$  本质上是局部区间  $[b_{k-1}, b_k)$  内激活值概率密度的质心, 算法会将量化点拉向概率密度更高的区域. 对于 Softmax 或 GELU 产生的长尾分布, 这种机制能够确保在数据密集的

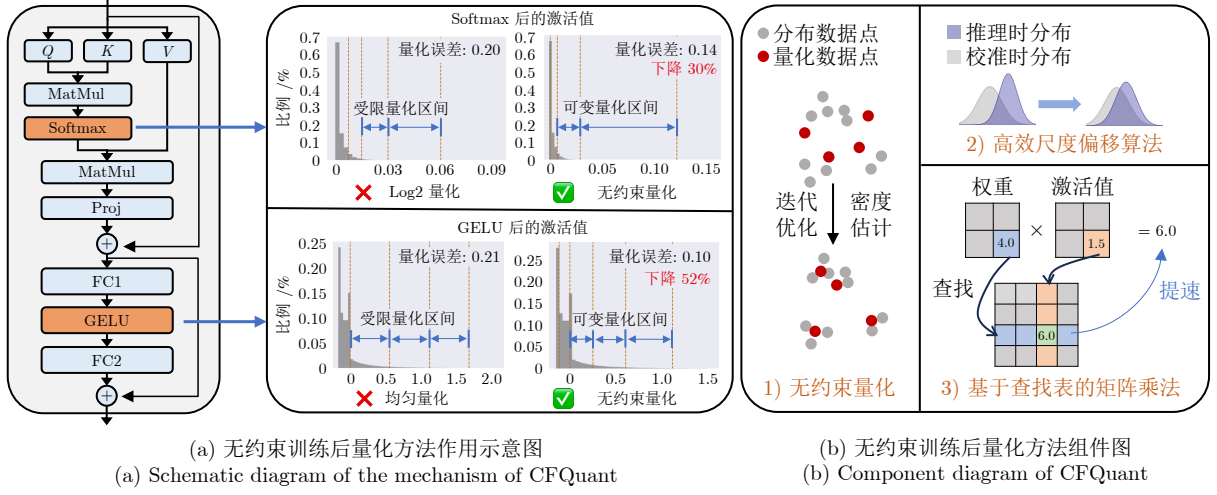


图 4 无约束训练后量化方法整体架构

Fig. 4 Overall architecture of CFQuant

区域分配更多的量化精度,而在稀疏的尾部区域分配较少,从而更有效地捕捉长尾分布中的高频信息.同理,考虑关于  $b_k$  的偏导数:

$$\frac{\partial \sigma^2}{\partial b_k} = (a_k - b_k)^2 p(b_k) - (a_{k+1} - b_k)^2 p(b_k) \quad (17)$$

将其设为零,得到

$$b_k = \frac{1}{2}(a_k + a_{k+1}) \quad (18)$$

这表明最优决策边界  $b_k$  恰好是相邻两个量化值的中点,符合最近邻原则,即输入值  $x$  应当被映射到与其欧氏距离最近的量化值  $a_k$  或  $a_{k+1}$ ,从而保证在给定量化值情况下的量化失真达到局部最小.接着,固定  $a_k$  并更新  $b_k$ ,通过式 (16) 和式 (18) 的交替迭代优化,逐步迭代求解得到最优的量化输出  $\{a_k\}$  和决策边界  $\{b_k\}$ .上述基于 Lloyd-Max 算法的迭代优化过程仅在离线校准阶段进行,收敛速度快,且不会增加在线推理阶段的计算负担.在校准阶段,假设校准数据集大小为  $\mathcal{N}$ ,量化级别为  $\mathcal{M}$ ,迭代次数为  $T$ ,Lloyd-Max 算法的单次迭代需要遍历所有样本点以更新质心和边界,则时间复杂度为  $O(T \times \mathcal{N} \times \mathcal{M})$ .由于  $\mathcal{N}$  和  $\mathcal{M}$  均为常数,迭代次数  $T$  较小,通常在 10 到 30 次迭代内即可收敛,且无需进行反向传播梯度计算,所以相比于量化感知训练或逐层重构方法,CFQuant 的校准开销较低.在推理阶段,由于量化参数  $\{a_k, b_k\}$  已被固定,模型仅需执行查表和简单的线性变换,因此推理时间复杂度不受迭代过程影响,不会引入额外的迭代计算开销.最后,根据式 (7) 对激活值进行量化与反量化.由于 CFQuant 没有约束量化区间,并且采用自适应的密度估计,所以能够适应那些非线性和长

尾分布的激活值,并保持较低的量化误差.

在推理阶段,由于数据样本和分布的差异,激活值的分布可能与校准阶段不同.同时,由于在部署中对实时延迟的要求严格,在推理过程中进行迭代优化算法是不可行的.为解决这些问题,本文提出高效尺度偏移算法 (efficient scale-shift algorithm, ESA),以动态对齐模型分布,减少校准阶段与推理阶段之间的分布偏移.设输入激活为  $\mathbf{X} \in \mathbf{R}^D$ ,反量化值为  $\hat{\mathbf{X}}_q$ .高效尺度偏移算法引入了尺度因子  $\alpha_{scale}$  和偏移因子  $\alpha_{shift}$ :

$$\alpha_{scale} = \frac{\|\mathbf{X}\|_1}{\|\hat{\mathbf{X}}_q\|_1}, \alpha_{shift} = \frac{1}{D} \sum_{i=1}^D (X^i - \hat{X}_q^i) \quad (19)$$

$$\hat{\mathbf{X}}'_q = (\hat{\mathbf{X}}_q + \beta_1 \alpha_{shift}) \beta_2 \alpha_{scale} \quad (20)$$

其中,  $X^i$  和  $\hat{X}_q^i$  分别表示输入激活向量  $\mathbf{X}$  和反量化向量  $\hat{\mathbf{X}}_q$  的第  $i$  个分量,  $i = 1, 2, \dots, D$ ;  $\hat{\mathbf{X}}'_q$  表示经 ESA 校正后的反量化激活向量;  $\beta_1$  和  $\beta_2$  为两个超参数,通过网格搜索算法对其进行优化;  $\alpha_{scale}$  利用  $L_1$  范数的比值来表征校准分布与推理分布在整体能量或幅度上的差异;  $\alpha_{shift}$  则通过均值差来捕获两者在分布中心上的系统性偏移.通过引入这两个因子,推理时的激活值分布能够动态地拉伸和平移,重新对齐到校准时的最佳量化状态,从而有效缓解因输入数据非平稳性带来的量化误差,同时也避免了在推理时进行迭代优化.尽管算法需要在运行时计算尺度因子和偏移因子,但仅会引入极少的计算开销.

### 3.2 基于查找表的矩阵乘法

在对输入激活  $\hat{\mathbf{X}}'_q \in \mathbf{R}^D$  和模型权重  $\mathbf{W}_q \in \mathbf{N}^D$

进行量化后, 矩阵乘法可表示为:

$$\begin{aligned} \mathbf{W}_q \hat{\mathbf{X}}_q' &= \beta_2 \alpha_{scale} \sum_{i=1}^D W_q^i (\hat{X}_q^i + \beta_1 \alpha_{shift}) = \\ &\beta_2 \alpha_{scale} \sum_{i=1}^D (W_q^i \hat{X}_q^i + \beta_1 W_q^i \alpha_{shift}) \end{aligned} \quad (21)$$

其中,  $W_q^i$  表示量化权重向量  $\mathbf{W}_q$  的第  $i$  个分量,  $i = 1, 2, \dots, D$ . 式 (21) 的变换核心在于实现计算解耦, 括号内的  $W_q^i \hat{X}_q^i$  项仅涉及量化整数与离散反量化值之间的乘法, 这是计算密集度最高的部分, 适合通过查找表预计算;  $\beta_2 \alpha_{scale}$  和  $\beta_1 \alpha_{shift}$  带来的修正则是浮点标量运算, 其计算开销极小且可提取至求和操作之外. 这种分解在数学上保证了等价性, 同时在物理实现上最大化了查找表的加速收益. 具体而言, 将  $\beta_1 \alpha_{shift}$  表示为二进有理数<sup>[21, 42]</sup>, 即  $\beta_1 \alpha_{shift} = m/2^d$ , 其中,  $m$  为整数,  $d$  为非负整数. 然后,  $\beta_1 W_q^i \alpha_{shift} = W_q^i m/2^d \gg d$  可以通过整数运算和移位操作高效实现. 由于  $\beta_2 \alpha_{scale} \in \mathbf{R}$  可以被合并并且只需计算一次, 因此在此处可忽略不计.

此时运算中仅剩  $W_q^i \hat{X}_q^i$ , 即量化整数与离散反量化值之间的乘法. 但是这种乘法类型在现有的图形处理器 (graphics processing unit, GPU) 硬件上支持性较差. 为解决这一问题, 本文提出了基于查找表的矩阵乘法. 查找表是由哈希表<sup>[43]</sup> 支持的一种高效操作. 假设执行  $k$ -比特训练后量化, 则  $W_q^i \in \{0, 1, \dots, 2^k - 1\}$  且  $\hat{X}_q^i \in \{a_0, a_1, \dots, a_{2^k-1}\}$ . 因此,  $W_q^i \hat{X}_q^i$  的所有可能取值共有  $2^k \times 2^k = 2^{2k}$  种. 预先计算一个大小为  $2^{2k}$  的表  $T$ , 其中, 第  $r$  行第  $c$  列的值表示对应的乘法结果  $W_q^r \hat{X}_q^c$ . 然后, 乘法运算即可高效地重构为查找操作:

$$W_q^r \hat{X}_q^c = \text{LUT}(W_q^r, \hat{X}_q^c) = T_{r,c} \quad (22)$$

其中,  $T_{r,c}$  表示查找表  $T$  中第  $r$  行、第  $c$  列的表项,  $r$  和  $c$  分别为量化权重与量化激活值对应的行、列索引. CFQuant 引入了极小的额外内存来存储查找表  $T$ , 其包含  $2^{2k}$  个元素. 在  $k = 4$  的低比特量化场景下, 这一额外内存开销微乎其微, 比如, 在 DeiT-S 模型中, 查找表只占用约 0.01 MB.

### 3.3 整体算法流程

无约束训练后量化方法的整体架构如图 4 所示, 主要包含三个核心组件: 1) CFQuant 基于激活值的分布估计以迭代方式进行模型量化, 相较于传统的均匀量化器, 它能够捕获到数据分布的特性, 从而显著降低量化误差; 2) CFQuant 为减小模型校准和推理时的分布差异, 设计高效尺度偏移算法,

从而进一步降低量化误差; 3) 提出基于查找表的矩阵乘法, 它可以加速量化矩阵乘法运算, 从而加速推理并降低延迟.

CFQuant 的算法流程如算法 1 所示. 主要包括校准阶段和推理阶段. 在校准阶段, 首先计算权重量化参数, 并随机初始化量化输出和边界参数, 然后通过迭代优化量化输出与边界, 使其收敛; 在推理阶段, 利用优化后的参数对激活值进行量化, 并采用高效尺度偏移算法计算修正后的激活值, 最终通过查找表高效计算量化输出.

#### 算法 1. 无约束训练后量化方法 (CFQuant)

输入. 激活值  $\mathbf{X}$  与权重  $\mathbf{W}$ .

输出. 量化输出  $\mathbf{W}_q \hat{\mathbf{X}}_q'$ .

/\* 校准阶段 \*/

- 1) 根据式 (10) 对权重  $\mathbf{W}$  进行逐通道均匀量化获得  $\mathbf{W}_q$ ;
- 2) 随机初始化  $a_k$  和  $b_k$ ;
- 3) **while** 未收敛 (not converged) **do**
- 4)   根据式 (16) 更新量化输出  $a_k$ ;
- 5)   根据式 (18) 更新量化边界  $b_k$ ;
- 6) **end while**
- /\* 推理阶段 \*/
- 7) 根据式 (7) 利用  $a_k$ 、 $b_k$  对激活值  $\mathbf{X}$  进行量化, 得到  $\hat{\mathbf{X}}_q$ ;
- 8) 根据式 (20) 利用高效尺度偏移算法 (ESA) 计算  $\hat{\mathbf{X}}_q'$ ;
- 9) 根据式 (21) 和式 (22) 利用查找表计算  $\mathbf{W}_q \hat{\mathbf{X}}_q'$ ;
- 10) 返回  $\mathbf{W}_q \hat{\mathbf{X}}_q'$ .

## 4 实验与结果分析

### 4.1 实验设定

#### 4.1.1 数据集

本节介绍用于评估无约束训练后量化方法的数据集. 为全面评估 CFQuant 的性能, 本文在涵盖视觉、语言和多模态基准的多种任务上开展测试. 在视觉任务中, 使用 ImageNet-1k 数据集<sup>[44]</sup> 评估视觉分类任务, 该数据集包含 1 000 个类别, 约 120 万张训练图像和 5 万张验证图像, 是计算机视觉领域最广泛使用的基准数据集之一. 同时, 采用 MS-COCO 数据集<sup>[45]</sup> 进行目标检测和实例分割任务的评估. 该数据集包括 80 个物体类别, 约 12 万张训练图像和 5 000 张验证图像, 凭借其丰富的标注信息, 成为目标检测和分割任务的重要基准. 目标检测和实例分割任务均采用平均精度 (average precision, AP) 作为评价指标, 分别记为边界框 AP 和掩码 AP. 在语言任务中, 本文在六个流行的零样本数据集和两个语言生成数据集上评估无约束训练后

量化. 零样本数据集包括: 用于物理常识推理的 PIQA<sup>[46]</sup>、用于科学问答的 ARC<sup>[47]</sup>、用于是非问答的 BoolQ<sup>[48]</sup>、用于常识推理的 HellaSwag<sup>[49]</sup> 以及用于代词消解的 WinoGrande<sup>[50]</sup>. ARC 包含 ARC-Easy (ARC-e) 和 ARC-Challenge (ARC-c) 两个子集. 语言生成数据集包括: WikiText-2<sup>[51]</sup>, 一个从维基百科高质量文章中提取的语言建模数据集, 包含约 255 万个词元; C4<sup>[52]</sup>, 一个从 Common Crawl 网页数据中提取的大规模文本数据集, 经过严格过滤和清理, 包含约 3.65 亿个英文文本段落. 此外, 为验证所提方法在视觉语言模型上的适用性, 本文进一步采用 MMBench<sup>[53]</sup> 的中英文评测集进行综合性多模态理解任务评估. 通过这些多样化的数据集上进行评估, 本文全面验证了无约束训练后量化方法在视觉、语言和多模态任务中的有效性与鲁棒性. 下文各表中的 W/A 分别表示权重 (weight) 和激活值 (activation) 的量化比特数.

#### 4.1.2 模型

本文采用多种视觉 Transformer 作为实验模型, 包括 ViT<sup>[1]</sup>、DeiT<sup>[28]</sup> 和 Swin Transformer<sup>[9]</sup>, 它们分别代表独特的视觉表征学习方法. 本文以 Swin 表示 Swin Transformer, 模型名称后缀 T、S 和 B 分别表示 Tiny、Small 和 Base 版本. 这些模型在图像分类、分割及其他视觉任务中表现出强大性能, 是评估量化影响的理想对象. 除视觉 Transformer 外, 本文还在大语言模型 (LLMs) 上评估 CFQuant, 选取主流的 LLaMA 系列模型<sup>[5]</sup>, 包括 LLaMA-7B 和 LLaMA-13B. 该系列模型在多项自然语言处理基准任务中表现出较强性能, 因而被广泛采用. 此外, 为验证 CFQuant 在新一代大语言模型架构上的通用性与鲁棒性, 本文还引入 Qwen3 系列模型<sup>[54]</sup> 作为实验对象, 包括 Qwen3-8B 和 Qwen3-14B. 进一步地, 引入 Qwen3-VL-4B<sup>[55]</sup> 作为视觉语言模型实验对象, 以评估 CFQuant 在多模态 Transformer 架构上的适用性. Qwen3 系列模型作为具有代表性的先进开源大语言模型之一, 其激活值分布特性与 LLaMA 系列存在显著差异, 对量化算法提出了更高挑战<sup>[56]</sup>.

#### 4.1.3 实现细节

所有实验均基于 PyTorch 框架实现. 遵循主流的量化方法设定<sup>[19, 57]</sup>, 对涉及矩阵乘法操作的权重和激活值均进行量化. 参照过往研究<sup>[19, 57–58]</sup>, 为保持模型的稳定性与精度, 层归一化和 Softmax 函数的计算过程保留为全精度; 对于 Softmax 输出的注意力权重和 GELU 输出的激活值, 则在其参与后续矩阵乘法前进行量化. 在量化策略方面, 对激活值采用逐层量化, 对权重采用逐通道量化, 确保实现的高效性与可扩展性. 对于视觉 Transformer, 所有预训练模型权重均调用自 timm 库<sup>[59]</sup>. 对于视觉任务, 采用 RepQ-ViT 方法<sup>[20]</sup> 作为基准方法, 并在 Softmax 输出和 GELU 输出激活值上应用无约束训练后量化. 关于校准数据集, 本文随机从训练集中抽取 32 张图像作为校准样本, 用于校准量化参数. 语言任务和多模态任务采用 SmoothQuant<sup>[19]</sup> 作为基准方法, 并对比分析 CFQuant 在自然语言处理和多模态理解任务中的性能.

### 4.2 视觉任务评估

#### 4.2.1 图像分类

实验结果如表 1 所示. 本文提出的无约束训练后量化方法在不同的视觉 Transformer 模型上均取得了最高 Top-1 准确率. 具体而言, 在 ViT-T 上, CFQuant 达到了 55.41% 的 Top-1 准确率, 显著高于 APQ-ViT<sup>[33]</sup> 方法. 在 DeiT 模型, 如 DeiT-S 上, CFQuant 相比 RepQ-ViT<sup>[20]</sup> 提升了 3.54 个百分点, 达到 72.57% 的 Top-1 准确率. 在分层视觉 Transformer 如 Swin-B 上, CFQuant 的 Top-1 准确率达到 81.84%, 与全精度模型相比仅有 3.43 个百分点的性能差距. 总体而言, 本文的无约束训练后量化方法在多个模型上优于先前的量化方法, 体现其高效性与灵活性.

#### 4.2.2 目标检测

除分类任务外, 本节还在 MS-COCO 数据集上进行目标检测与实例分割实验, 结果如表 2 所示. 总体而言, 除 Mask R-CNN 框架下 Swin-S 的边界框 AP 略低于部分对比方法外, CFQuant 在其余设置

表 1 不同方法在 ImageNet-1k 数据集上分类任务的 Top-1 准确率对比 (%)

Table 1 Top-1 accuracy comparison of different methods on classification tasks on ImageNet-1k dataset (%)

| 方法                       | 比特数 (W/A) | ViT-T        | ViT-S        | ViT-B        | DeiT-T       | DeiT-S       | DeiT-B       | Swin-T       | Swin-S       | Swin-B       |
|--------------------------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 全精度                      | 32/32     | 75.47        | 81.39        | 84.54        | 72.21        | 79.85        | 81.80        | 81.37        | 83.23        | 85.27        |
| PTQ4ViT <sup>[19]</sup>  | 4/4       | 16.53        | 42.57        | 30.69        | 36.96        | 34.08        | 64.39        | 73.06        | 76.09        | 74.02        |
| APQ-ViT <sup>[33]</sup>  | 4/4       | 17.56        | 47.95        | 41.41        | 47.94        | 43.55        | 67.48        | —            | 77.15        | 76.48        |
| RepQ-ViT <sup>[20]</sup> | 4/4       | 26.72        | 65.05        | 68.48        | 57.43        | 69.03        | 75.61        | 72.37        | 79.45        | 80.74        |
| CFQuant                  | 4/4       | <b>55.41</b> | <b>70.95</b> | <b>73.72</b> | <b>62.57</b> | <b>72.57</b> | <b>78.02</b> | <b>73.80</b> | <b>80.65</b> | <b>81.84</b> |

下均取得了具有竞争力或更优的性能. 在使用 Swin-T 模型的 Mask R-CNN 中, CFQuant 达到了 36.6 的边界框 AP 和 36.5 的掩码 AP. 在 Cascade Mask R-CNN 框架下, CFQuant 在 Swin-S 上达到了 50.1 的边界框 AP 和 43.6 的掩码 AP, 优于 RepQ-ViT<sup>[20]</sup> 的 49.3 和 43.1, CFQuant 与全精度模型之间的边界框 AP 和掩码 AP 差距分别仅为 1.8 和 1.4, 表明无

约束训练后量化方法能够有效适应密集预测任务.

### 4.3 语言与多模态任务评估

#### 4.3.1 零样本任务

除视觉任务外, 本节在自然语言处理领域评估 CFQuant, 特别是在大语言模型上, 实验结果如表 3 和表 4 所示. CFQuant 在 LLaMA-7B 和 LLaMA-

表 2 不同方法在 MS-COCO 数据集上目标检测和实例分割任务的性能对比 (%)

Table 2 Performance comparison of different methods on object detection and instance segmentation tasks on the MS-COCO dataset (%)

| 方法                       | 比特数 (W/A) | Mask R-CNN  |             |             |             | Cascade Mask R-CNN |             |             |             |
|--------------------------|-----------|-------------|-------------|-------------|-------------|--------------------|-------------|-------------|-------------|
|                          |           | Swin-T      |             | Swin-S      |             | Swin-T             |             | Swin-S      |             |
|                          |           | 边界框 AP      | 掩码 AP       | 边界框 AP      | 掩码 AP       | 边界框 AP             | 掩码 AP       | 边界框 AP      | 掩码 AP       |
| 全精度                      | 32/32     | 46.0        | 41.6        | 48.5        | 43.3        | 50.4               | 43.7        | 51.9        | 45.0        |
| PTQ4ViT <sup>[19]</sup>  | 4/4       | 6.9         | 7.0         | 26.7        | 26.6        | 14.7               | 13.5        | 0.5         | 0.5         |
| APQ-ViT <sup>[33]</sup>  | 4/4       | 23.7        | 22.6        | <b>44.7</b> | 40.1        | 27.2               | 24.4        | 47.7        | 41.1        |
| RepQ-ViT <sup>[20]</sup> | 4/4       | 36.1        | 36.0        | 44.2        | 40.2        | 47.0               | 41.4        | 49.3        | 43.1        |
| CFQuant                  | 4/4       | <b>36.6</b> | <b>36.5</b> | 43.4        | <b>40.6</b> | <b>47.7</b>        | <b>42.0</b> | <b>50.1</b> | <b>43.6</b> |

表 3 在多个零样本任务上 LLaMA 系列模型不同方法的准确率对比 (%)

Table 3 Accuracy comparison of different methods on multiple zero-shot tasks across LLaMA models (%)

| 模型        | 方法                          | 比特数 (W/A) | PIQA         | ARC-e        | WinoGrande   | BoolQ        | ARC-c        | HellaSwag    | 平均准确率        |
|-----------|-----------------------------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| LLaMA-7B  | 全精度                         | 16/16     | 77.47        | 52.48        | 67.07        | 73.08        | 41.46        | 73.00        | 64.09        |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 49.80        | 30.40        | 48.00        | 49.10        | 25.80        | 27.40        | 38.41        |
|           | OmniQuant <sup>[36]</sup>   | 4/4       | 66.15        | <b>45.20</b> | 53.43        | 63.51        | 31.14        | 56.44        | 52.65        |
|           | AffineQuant <sup>[60]</sup> | 4/4       | <b>69.37</b> | 42.55        | 55.33        | 63.73        | 31.91        | <b>57.65</b> | 53.42        |
|           | CFQuant                     | 4/4       | 67.91        | 43.90        | <b>55.36</b> | <b>64.45</b> | <b>32.19</b> | 57.14        | <b>53.49</b> |
| LLaMA-13B | 全精度                         | 16/16     | 79.10        | 59.89        | 70.31        | 68.01        | 44.45        | 76.21        | 66.33        |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 61.04        | 39.18        | 51.06        | 61.80        | 30.80        | 52.29        | 49.36        |
|           | OmniQuant <sup>[36]</sup>   | 4/4       | 69.69        | <b>47.39</b> | 55.80        | 62.84        | 33.10        | 58.96        | 54.63        |
|           | AffineQuant <sup>[60]</sup> | 4/4       | 66.32        | 43.90        | 54.70        | 64.10        | 29.61        | 56.88        | 52.58        |
|           | CFQuant                     | 4/4       | <b>71.20</b> | 45.61        | <b>56.39</b> | <b>64.17</b> | <b>33.35</b> | <b>59.51</b> | <b>55.04</b> |

表 4 在多个零样本任务上 Qwen3 系列模型不同方法的准确率对比 (%)

Table 4 Accuracy comparison of different methods on multiple zero-shot tasks across Qwen3 models (%)

| 模型        | 方法                          | 比特数 (W/A) | PIQA         | ARC-e        | WinoGrande   | BoolQ        | ARC-c        | HellaSwag    | 平均准确率        |
|-----------|-----------------------------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Qwen3-8B  | 全精度                         | 16/16     | 77.80        | 80.93        | 67.72        | 86.61        | 56.66        | 74.92        | 74.11        |
|           | SmoothQuant <sup>[15]</sup> | 4/8       | 73.42        | 71.01        | 62.16        | 79.11        | 44.32        | 51.68        | 63.62        |
|           | PrefixQuant <sup>[61]</sup> | 4/8       | 71.98        | 70.50        | 64.09        | <b>82.26</b> | 46.93        | 65.09        | 66.81        |
|           | CFQuant                     | 4/8       | <b>76.17</b> | <b>71.13</b> | <b>65.82</b> | 79.79        | <b>47.61</b> | <b>71.56</b> | <b>68.68</b> |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 50.53        | 25.51        | 52.17        | 39.58        | 25.73        | 26.72        | 36.71        |
|           | PrefixQuant <sup>[61]</sup> | 4/4       | 56.42        | 38.55        | 48.07        | 45.69        | 26.79        | 34.82        | 41.72        |
|           | CFQuant                     | 4/4       | <b>72.74</b> | <b>71.84</b> | <b>64.09</b> | <b>78.81</b> | <b>47.87</b> | <b>67.93</b> | <b>67.21</b> |
| Qwen3-14B | 全精度                         | 16/16     | 79.82        | 82.87        | 72.93        | 89.33        | 60.24        | 78.82        | 77.34        |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 51.22        | 25.81        | 50.18        | 38.47        | 26.51        | 25.79        | 36.33        |
|           | PrefixQuant <sup>[61]</sup> | 4/4       | 57.89        | 43.81        | 51.30        | 56.73        | 29.10        | 39.83        | 46.44        |
|           | CFQuant                     | 4/4       | <b>65.07</b> | <b>58.71</b> | <b>56.91</b> | <b>78.84</b> | <b>40.10</b> | <b>49.39</b> | <b>58.17</b> |

13B 模型上针对六个零样本任务的平均准确率分别为 53.49% 和 55.04%，超过了 OmniQuant<sup>[36]</sup> 和 AffineQuant<sup>[60]</sup>。这表明 CFQuant 不仅在视觉任务中有效，而且在自然语言处理任务中同样表现出色，具备跨领域适用性。在 BoolQ 数据集上，CFQuant 在 LLaMA-13B 模型上达到了 64.17% 的准确率，而全精度模型为 68.01%，将性能差距缩小至 3.84 个百分点。

为进一步验证 CFQuant 在不同模型架构以及先进模型不同比特设置下的优势，表 4 展示 Qwen3-8B 和 Qwen3-14B 模型在零样本任务上的性能对比。Qwen3 模型系列采用了更先进的预训练技术，使模型参数的冗余度更低，对量化带来的信息损失更为敏感，在低比特量化时表现出更明显的性能下降<sup>[56]</sup>。

实验结果表明，CFQuant 在 Qwen3 模型的不同量化设置下整体取得了更优的性能，尤其在低比特量化设置下优势更加显著。对于 Qwen3-8B 模型，CFQuant 在 W4/A8（权重 4 比特/激活值 8 比特）配置下取得了 68.68% 的平均准确率，超越了 PrefixQuant 的 66.81%。在极低比特的 W4/A4 配置下，SmoothQuant 和 PrefixQuant 面临严重的性能崩塌，量化精度降至 4 比特时，平均准确率分别降至 36.71% 和 41.72%；相比之下，CFQuant 依然保持了 67.21% 的平均准确率，仅比 W4/A8 下降了 1.47 个百分点，显著优于 PrefixQuant，提高了 25.49 个百分点。值得注意的是，CFQuant 在 W4/

A4 设置下的性能甚至超越了 PrefixQuant 在 W4/A8 设置下的表现，平均准确率 67.21% 优于 PrefixQuant 在 W4/A8 下的 66.81%，高出 0.40 个百分点。在更大参数规模的 Qwen3-14B 模型上，CFQuant 同样表现出优秀的性能。在 W4/A4 设置下，CFQuant 的平均准确率达到 58.17%，超过 PrefixQuant 的 46.44%。

CFQuant 的无约束量化策略在自然语言处理任务中展现出优异的性能。这些结果证明，在低比特场景下，本文提出的无约束量化范式在处理长尾分布问题方面具有显著优势。CFQuant 能够有效适应各类大语言模型中复杂的激活值分布，减少量化带来的性能损失，在先进大语言模型和低比特量化场景中具有巨大的应用潜力和价值。

#### 4.3.2 语言生成

本节评估无约束训练后量化方法在语言生成任务中的性能，结果以困惑度指标进行评估。困惑度越低，表示模型生成的文本越自然、准确。实验结果如表 5 所示，在 LLaMA 系列模型上，CFQuant 在两个数据集上均取得了较低的困惑度。以 LLaMA-7B 为例，在 WikiText-2 数据集上，OmniQuant 的困惑度为 11.26，而 CFQuant 为 10.36，效果更加优越；在 C4 数据集上，CFQuant 在量化方法中取得了最低的困惑度 13.91，整体表现优于基线方法。

在 Qwen3 系列模型上，不同方法在低比特量化条件下的表现差异更加明显，对比方法的困惑度出现较大幅度上升。以 Qwen3-8B 为例，Smooth-

表 5 在 WikiText-2 和 C4 数据集上语言生成任务的困惑度比较  
Table 5 Perplexity comparison on language generation tasks on WikiText-2 and C4 datasets

| 模型        | 方法                          | 比特数 (W/A) | WikiText-2 ↓       | C4 ↓               |
|-----------|-----------------------------|-----------|--------------------|--------------------|
| LLaMA-7B  | 全精度                         | 16/16     | 5.68               | 7.08               |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 25.25              | 32.32              |
|           | OmniQuant <sup>[36]</sup>   | 4/4       | 11.26              | 14.51              |
|           | CFQuant                     | 4/4       | <b>10.36</b>       | <b>13.91</b>       |
| LLaMA-13B | 全精度                         | 16/16     | 5.09               | 6.61               |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | 40.05              | 47.18              |
|           | OmniQuant <sup>[36]</sup>   | 4/4       | 10.87              | 13.78              |
|           | CFQuant                     | 4/4       | <b>10.14</b>       | <b>13.28</b>       |
| Qwen3-8B  | 全精度                         | 16/16     | 9.72               | 13.29              |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | $3.36 \times 10^4$ | $2.29 \times 10^4$ |
|           | PrefixQuant <sup>[61]</sup> | 4/4       | 155.78             | 119.93             |
|           | CFQuant                     | 4/4       | <b>13.78</b>       | <b>17.26</b>       |
| Qwen3-14B | 全精度                         | 16/16     | 8.64               | 12.01              |
|           | SmoothQuant <sup>[15]</sup> | 4/4       | $2.16 \times 10^5$ | $1.99 \times 10^5$ |
|           | PrefixQuant <sup>[61]</sup> | 4/4       | 187.28             | 129.05             |
|           | CFQuant                     | 4/4       | <b>32.30</b>       | <b>35.76</b>       |

Quant 在 WikiText-2 上的困惑度为  $3.36 \times 10^4$ , 表明模型预测能力明显下降; PrefixQuant 在 WikiText-2 上的困惑度为 155.78, 出现了明显的性能损失. 上述结果说明, 在激活分布存在较强异常值特征的模型上, 低比特量化对模型性能影响显著. 相比之下, CFQuant 在 Qwen3 模型上保持了相对稳定的量化性能. 在 Qwen3-8B 模型上, 在 WikiText-2 和 C4 困惑度为 13.78 和 17.26, 而相同设置下 PrefixQuant 为 155.78 和 119.93; 在 Qwen3-14B 模型上, CFQuant 仍取得相对较低的困惑度. 实验结果表明, 在低比特量化场景下, CFQuant 能够较好地维持语言建模性能, 体现出强大的稳定性与适应性, 展现了其在语言模型中的应用潜力.

#### 4.3.3 视觉语言模型上的多模态理解

本节在 MMBench-V1.1<sup>[53]</sup> 的英文评测数据集 (MMBench-EN) 和中文评测数据集 (MMBench-CN) 上开展多模态理解实验, 以进一步验证所提方法在视觉语言模型架构上的适用性. 相比纯语言模型, 视觉语言模型同时涉及视觉特征提取、跨模态对齐与语言理解, 其激活分布和特征交互形式更加复杂, 因此能够更全面地检验量化方法在多模态 Transformer 架构中的稳定性与泛化能力.

实验基于 Qwen3-VL-4B 模型进行, 结果如表 6 所示. 在相同的 W4/A8 设置下, CFQuant 在 MMBench-EN 和 MMBench-CN 上分别达到 79.41% 和 78.72% 的准确率, 均取得最优性能. 具体而言, 相比 SmoothQuant, CFQuant 在英文和中文评测数据集上分别提升 4.10 和 3.72 个百分点; 相比 PrefixQuant 分别提升 2.32 和 0.93 个百分点. 与全精度模型相比, CFQuant 量化后模型性能仅分别下降 3.72 和 2.08 个百分点, 表明其在同时降低权重和激活表示精度的情况下, 仍能够较好地保持视觉语言模型的多模态理解能力.

此外, 本文进一步引入激活感知权重量化 (activation-aware weight quantization, AWQ)<sup>[62]</sup> 作为

表 6 在 MMBench 数据集上 Qwen3-VL-4B 模型的不同方法准确率对比 (%)

Table 6 Accuracy comparison of different methods for the Qwen3-VL-4B model on the MMBench dataset (%)

| 方法                          | 比特数 (W/A) | MMBench-EN   | MMBench-CN   |
|-----------------------------|-----------|--------------|--------------|
| 全精度                         | 16/16     | 83.13        | 80.80        |
| AWQ <sup>[62]</sup>         | 4/16      | 81.58        | 79.26        |
| SmoothQuant <sup>[15]</sup> | 4/8       | 75.31        | 75.00        |
| PrefixQuant <sup>[61]</sup> | 4/8       | 77.09        | 77.79        |
| CFQuant                     | 4/8       | <b>79.41</b> | <b>78.72</b> |

代表性权重量化方法进行对比, 以展示不同量化范式在视觉语言模型上的性能表现. AWQ 属于激活感知的训练后权重量化方法, 其主要将模型权重量化为 4 比特, 而激活值仍保持 16 比特精度. 因此, AWQ 对应的是 W4/A16 设置, 而本文方法关注的是更具挑战性的权重与激活联合低比特量化场景. 实验结果表明, AWQ 在 MMBench-EN 和 MMBench-CN 上分别取得 81.58% 和 79.26% 的准确率. 相比之下, CFQuant 在更严格的 W4/A8 设置下同时量化权重和激活值, 仍能与 W4/A16 的 AWQ 保持较小性能差距, 尤其在中文评测数据集 MMBench-CN 上仅低 0.54 个百分点. 上述结果说明, CFQuant 在激活值同样低比特量化的条件下, 仍能够有效保持视觉语言模型的多模态理解能力.

#### 4.4 消融实验

##### 4.4.1 查找表的额外内存开销

尽管 CFQuant 需要利用查找表对乘法进行预计算并由此引入内存占用, 但如表 7 所示, 这些额外参数的内存开销相对较小, 可以忽略不计. 在 4 比特量化场景下, 本文比较未采用 CFQuant 与采用后的模型尺寸和性能变化. 结果显示, 查找表带来的最大增量为 0.02 MB, 仅占模型总大小的 0.1%. 在性能方面, CFQuant 在引入查找表的同时显著提升了模型精度, DeiT-S 在应用 CFQuant 后, Top-1 准确率升至 72.57%, 带来了 4.30 个百分点的提升. 在 Swin-S 上, Top-1 准确率提升至 80.65%, 性能提升了 1.62 个百分点. 这些结果表明, CFQuant 通过引入查找表来优化推理过程, 虽然增加少量内存开销, 但在性能提升上的贡献远大于这一成本.

表 7 在 ImageNet-1k 数据集上查找表额外内存开销的消融实验

Table 7 Ablation study of the additional lookup-table memory overhead on the ImageNet-1k dataset

| 模型     | CFQuant | 模型尺寸 (MB)     | Top-1 准确率 (%) |
|--------|---------|---------------|---------------|
| DeiT-S | ×       | 11.18         | 68.27         |
|        | √       | 11.19 (+0.1%) | 72.57 (+4.30) |
| Swin-S | ×       | 24.78         | 79.03         |
|        | √       | 24.80 (+0.1%) | 80.65 (+1.62) |

##### 4.4.2 量化误差与模型性能的可视化

为直观表明 CFQuant 对量化误差和模型性能的影响, 本节对 CFQuant 应用前后的效果进行可视化. 量化误差定义为 Softmax 输出和 GELU 输出激活值的平均量化误差. 如图 5 所示, CFQuant 显著减少了量化误差, 误差降低幅度接近一半. 与

此同时, 模型性能得到了显著提升, 呈现出明显的左上方趋势. 这意味着 CFQuant 不仅减少了量化误差, 还有效提高了模型准确性. 值得注意的是, 所有这些 ViT 骨干网络均展现了类似的趋势, 在多种不同的 Transformer 结构下, CFQuant 都能够减少量化误差并提高模型性能, 这表明 CFQuant 的灵活性和有效性, 体现其在多种应用场景中的广泛适用性和优化潜力.

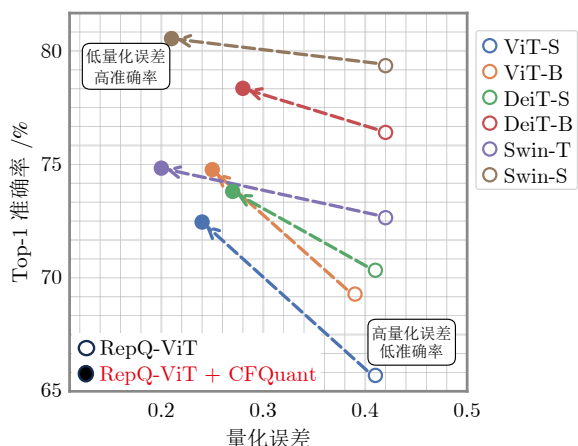


图 5 CFQuant 对量化误差和模型性能的影响  
Fig.5 Effect of CFQuant on quantization error and model performance

#### 4.4.3 Softmax 与 GELU 的消融实验

本节针对 CFQuant 在不同激活函数下的表现进行消融研究, 具体分析 CFQuant 应用于 Softmax 输出和 GELU 输出激活值时的效果. 表 8 的实验结果表明, 对于 DeiT-S, 当 CFQuant 应用于 GELU 输出激活值时, Top-1 准确率从 68.27% 提升至 71.33%. 当进一步在 Softmax 后激活值上应用 CFQuant 时, 性能得到进一步提升, Top-1 准确率从 71.33% 升至 72.57%. 在 Swin-S 上, 实验结果与 DeiT 模型相似, 同样验证了 CFQuant 在不同激活函数下的有效性. 在 Swin-S 上, CFQuant 应用于 GELU 输出激活值后使性能提高了 1.33 个百分点, 而在进一步应用于 Softmax 输出激活值后, 模型性能继续提升. 这些结果表明, CFQuant 能够有效降低具有长尾分布特征的激活值的量化误差, 从而提升模型的整体性能.

#### 4.4.4 高效尺度偏移算法的消融实验

如表 9 所示, 本节评估高效尺度偏移算法 (ESA) 的有效性. 超参数  $\beta_1$  和  $\beta_2$  通过网格搜索进行优化, 搜索间隔为 0.000 1, 迭代次数为 10 次, 以确定最优值. ESA 旨在缓解校准阶段与推理阶段之间激活值的分布差距, 从而提升模型性能. 在 DeiT-S 上,

表 8 在 ImageNet-1k 数据集上不同激活函数量化策略的消融实验

Table 8 Ablation study of quantization strategies for different activation functions on the ImageNet-1k dataset

| 模型     | Softmax | GELU | Top-1 准确率 (%) |
|--------|---------|------|---------------|
| DeiT-S | ×       | ×    | 68.27         |
|        | √       | ×    | 68.82 (+0.55) |
|        | ×       | √    | 71.33 (+3.06) |
|        | √       | √    | 72.57 (+4.30) |
| Swin-S | ×       | ×    | 79.03         |
|        | √       | ×    | 79.28 (+0.25) |
|        | ×       | √    | 80.36 (+1.33) |
|        | √       | √    | 80.65 (+1.62) |

表 9 在 ImageNet-1k 数据集上高效尺度偏移算法的消融实验

Table 9 Ablation studies of efficient scale-shift algorithm on the ImageNet-1k dataset

| 模型     | ESA | Top-1 准确率 (%) |
|--------|-----|---------------|
| DeiT-S | ×   | 71.08         |
|        | √   | 72.57 (+1.49) |
| Swin-S | ×   | 78.93         |
|        | √   | 80.65 (+1.72) |

应用 ESA 将分类 Top-1 准确率从 71.08% 提升至 72.57%, 带来了 1.49 个百分点的性能提升; 同时, 对于 Swin-S 模型, ESA 将 Top-1 准确率提升了 1.72 个百分点. 这些结果证明了 ESA 在提升模型性能方面的有效性和泛化能力.

#### 4.4.5 校准样本数量的敏感性分析

为探究校准样本数量对量化性能的影响, 并分析方法在小样本条件下的稳定性, 本节进行校准样本数敏感性分析实验. 如图 6 所示, 随着样本数量的增加, CFQuant 的 Top-1 准确率快速上升, 并在  $N = 16$  时基本趋于稳定. 对于 CFQuant, 少量样本已能够充分地表征激活分布特征, 使无约束量化器学习到稳定的量化参数. 在小样本条件下, CFQuant 已达到接近饱和的性能水平, 且方差较小, 体现出良好的统计稳定性. 相比之下, 基准方法 RepQ-ViT 在小样本下 Top-1 准确率下降更为明显, 且性能波动较大, 表明其对校准数据规模敏感. 总体而言, CFQuant 的迭代过程旨在学习激活值通用的分布形状, 而非特定样本的数值统计量. 通过结合校准阶段的迭代优化与推理阶段的 ESA 机制, 方法对校准样本数量依赖性较低, 表现出良好的数据效率与泛化能力. 当  $N \geq 32$  时, 各方法均已进入性能

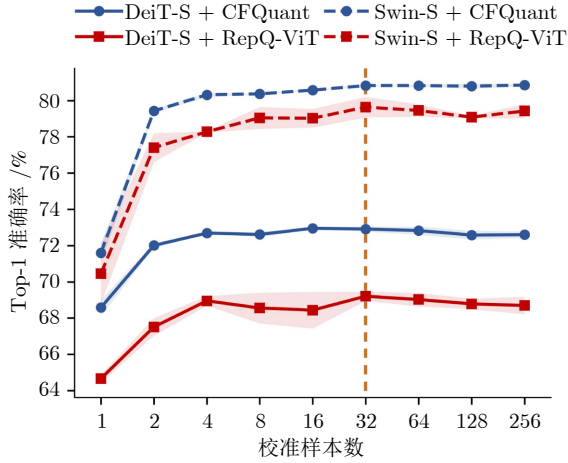


图6 不同校准样本数量下的 Top-1 准确率对比

Fig.6 Top-1 accuracy comparison with different calibration sample sizes

稳定区间, 综合效率与性能表现, 采用 32 个样本作为校准设置.

#### 4.4.6 校准时间

本节对比不同量化方法在 32 个校准样本设置下的校准总耗时, 实验结果如表 10 所示. 由于 CFQuant 无需进行模型重构<sup>[18-19, 33]</sup>, 避免梯度计算, 因此能够保持较低的校准开销. 具体而言, 在 DeiT-S 模型上, CFQuant 使用 32 个校准样本完成校准仅需 3.28 min; 在 Swin-S 模型上, CFQuant 的校准时间为 7.19 min, 相比 PTQ4ViT 的 8.46 min, 缩短约 15.01%. 与 RepQ-ViT 相比, CFQuant 由于引入 Lloyd-Max 迭代优化, 校准时间略有增加, 但在 Swin-S 上增加幅度仅约为 4.66%. 上述结果表明, 虽然 CFQuant 在校准阶段引入迭代求解过程, 但该过程仅在少量校准样本上离线执行, 且无需复杂的端到端梯度计算或逐层重构, 因此并未成为部署流程中的主要时间瓶颈. CFQuant 在提升低比特量化性能的同时, 仍保持训练后量化方法高效快捷的优势, 适合快速模型部署场景.

#### 4.5 推理效率与部署资源分析

为全面评估模型在实际部署中的性能, 本节在 DeiT-S 主干网络上开展实验, 从存储开销、计算复杂度及推理时延三个维度对 CFQuant 进行分析. 实验基于 NVIDIA RTX 3060 GPU 和 TensorRT 框架进行, 实验结果如表 11 所示.

##### 4.5.1 存储开销与占用

CFQuant 引入查找表作为额外组件, 但对于 4 比特量化, 查找表带来的额外存储开销约为 0.01 MB, 可以忽略不计. 这使 CFQuant 能够充分发挥低比

表 10 在 ImageNet-1k 数据集上不同量化方法校准时间的对比

Table 10 Comparison of calibration time for different quantization methods on the ImageNet-1k dataset

| 模型     | 方法                       | Top-1 准确率 (%) | 耗时 (min) |
|--------|--------------------------|---------------|----------|
| DeiT-S | 全精度                      | 79.85         | —        |
|        | PTQ4ViT <sup>[19]</sup>  | 34.08         | 3.61     |
|        | RepQ-ViT <sup>[20]</sup> | 69.03         | 3.12     |
|        | CFQuant                  | 72.57         | 3.28     |
| Swin-S | 全精度                      | 83.23         | —        |
|        | PTQ4ViT <sup>[19]</sup>  | 76.09         | 8.46     |
|        | RepQ-ViT <sup>[20]</sup> | 79.45         | 6.87     |
|        | CFQuant                  | 80.65         | 7.19     |

表 11 在 ImageNet-1k 数据集上 DeiT-S 模型的推理效率对比

Table 11 Comparison on inference efficiency for DeiT-S model on the ImageNet-1k dataset

| 方法                       | 模型尺寸 (MB) | 位运算量 (G) | 延迟 (s) |
|--------------------------|-----------|----------|--------|
| 全精度                      | 89.44     | 4 741    | 10.39  |
| RepQ-ViT <sup>[20]</sup> | 11.18     | 141      | 2.67   |
| CFQuant                  | 11.18     | 346      | 4.33   |
| CFQuant + MM-LUT         | 11.19     | 144      | 2.80   |

特量化的优势. 与全精度模型相比, CFQuant 将模型存储需求从 89.44 MB 大幅压缩至 11.19 MB, 实现了约 8× 的存储压缩比. 这种低比特量化大幅降低了模型权重与激活值的存储开销, 同时也能够减少推理阶段的访存带宽压力, 从而提升整体部署效率.

##### 4.5.2 计算复杂度分析

本节采用位运算量 (bit operations, BitOPs) 作为衡量计算量的指标. 实验结果表明, 在引入 MM-LUT 后, CFQuant 将位运算量从全精度模型的 4 741 G 降低至 144 G, 理论计算量约为原始模型的 3%. 这一优化效果主要得益于本文提出的 MM-LUT 机制, 将原本的浮点乘法操作转换为预计算的查找表索引操作, 有效降低量化计算的计算复杂度.

##### 4.5.3 实际推理时延

得益于计算复杂度的降低与 MM-LUT 的加速, CFQuant 实现了显著的端到端推理加速. 在 DeiT-S 模型上, 推理延迟从全精度的 10.39 s 降低至 2.80 s, 实现了约 3.7× 的端到端加速. 相比于 RepQ-ViT 方法, CFQuant 在仅增加 0.13 s 延迟 (+4.8%) 的代价下, 实现了 Top-1 准确率提升 3.54 个百分点, 证明了本文方法在精度与效率权衡方面具有优势.

## 5 结束语

本文提出一种针对低比特 Transformer 的无约

束训练后量化方法 (CFQuant), 该方法通过去除量化区间的约束, 克服了传统均匀和 Log2 量化器的局限性, 显著降低了量化误差并提升了模型量化性能. 首先, CFQuant 通过自适应估计激活分布和迭代优化, 实现了更精确的量化; 其次, 高效尺度偏移算法进一步降低了校准阶段和推理阶段之间的分布差异; 此外, 本文还设计基于查找表的矩阵乘法, 利用预计算的查找表将浮点乘法转换为更高效的查找操作, 显著提高了模型推理速度; 最后, 本文在多种 Transformer 模型以及视觉、语言和多模态任务上验证了 CFQuant 的适用性和有效性.

### 参考文献

- 1 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth  $16 \times 16$  words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.
- 2 Dai Y, Chen X J, Wang X H, Pang M H, Gao L L, Shen H T. ReSParser: Fully convolutional multiple human parsing with representative sets. *IEEE Transactions on Multimedia*, 2024, **26**: 1384–1394
- 3 Zhang S S, Roller S, Goyal N, Artetxe M, Chen M Y, Chen S H, et al. OPT: Open pre-trained Transformer language models. arXiv preprint arXiv: 2205.01068, 2022.
- 4 Li S S, Xu X, Meng W X, Song J K, Peng C, Shen H T. Mitigating hallucinations in large vision-language models via reasoning uncertainty-guided refinement. *IEEE Transactions on Multimedia*, 2025, **27**: 7380–7391
- 5 Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288, 2023.
- 6 Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proceedings of the Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 2165–2183
- 7 Li Hao-Ran, Chen Yu-Hui, Cui Wen-Bo, Liu Wei-Heng, Liu Kai, Zhou Ming-Cai, et al. Survey of vision-language-action models for embodied manipulation. *Acta Automatica Sinica*, 2026, **52**(1): 18–51  
(李浩然, 陈宇辉, 崔文博, 刘卫恒, 刘锴, 周明才, 等. 面向具身操作的视觉-语言-动作模型综述. *自动化学报*, 2026, **52**(1): 18–51)
- 8 Qu D L, Song H M, Chen Q Z, Wang D, Yao Y Q, Ye X Y, et al. SpatialVLA: Exploring spatial representations for visual-language-action model. In: Proceedings of the Robotics: Science and Systems 2025. Los Angeles, USA: University of Southern California, 2025. Article No. 011
- 9 Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9992–10002
- 10 Wang Wen-Sheng, Tan Ning, Huang Kai, Zhang Yu-Nong, Zheng Wei-Shi, Sun Fu-Chun. Embodied intelligence systems based on large models: A survey. *Acta Automatica Sinica*, 2025, **51**(1): 1–19  
(王文晟, 谭宁, 黄凯, 张雨浓, 郑伟诗, 孙富春. 基于大模型的具身智能系统综述. *自动化学报*, 2025, **51**(1): 1–19)
- 11 Gu Y X, Dong L, Wei F R, Huang M L. MiniLLM: Knowledge distillation of large language models. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 12 Chen G B, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc., 2017. 742–751
- 13 Ashkboos S, Croci M L, do Nascimento M G, Hoefler T, Hensman J. SliceGPT: Compress large language models by deleting rows and columns. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 14 Sun M J, Liu Z, Bair A, Kolter J Z. A simple and effective pruning approach for large language models. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 15 Xiao G X, Lin J, Seznec M, Wu H, Demouth J, Han S. SmoothQuant: Accurate and efficient post-training quantization for large language models. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 38087–38099
- 16 Liu Y J, Yang H R, Dong Z, Keutzer K, Du L, Zhang S H. NoisyQuant: Noisy bias-enhanced post-training activation quantization for vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 20321–20330
- 17 Yin P, Zhu X S, Song J K, Gao L L, Shen H T. SI-BiViT: Binarizing vision Transformers with spatial interaction. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 8169–8178
- 18 Li Y H, Gong R H, Tan X, Yang Y, Hu P, Zhang Q, et al. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.
- 19 Yuan Z H, Xue C H, Chen Y Q, Wu Q, Sun G Y. PTQ4ViT: Post-training quantization for vision Transformers with twin uniform quantization. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 191–207
- 20 Li Z K, Xiao J R, Yang L W, Gu Q Y. RepQ-ViT: Scale reparameterization for post-training quantization of vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 17181–17190
- 21 Jacob B, Kligys S, Chen B, Zhu M L, Tang M, Howard A, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2704–2713
- 22 Cai J Y, Takemoto M, Nakajo H. A deep look into logarithmic quantization of model parameters in neural networks. In: Proceedings of the 10th International Conference on Advances in Information Technology. Bangkok, Thailand: ACM, 2018. Article No. 6
- 23 Lloyd S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982, **28**(2): 129–137
- 24 Max J. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 1960, **6**(1): 7–12
- 25 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc., 2017. 5998–6008
- 26 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 4171–4186
- 27 Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I.

- Language models are unsupervised multitask learners [Online], available: [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf), July 3, 2026
- 28 Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image Transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 10347–10357
  - 29 Zhang Z Z, Zhang H, Zhao L, Chen T, Arik S Ö, Pfister T. Nested hierarchical Transformer: Towards accurate, data-efficient and interpretable visual understanding. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 3417–3425
  - 30 Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 7242–7252
  - 31 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2020. Article No. 159
  - 32 Sun Y X, Liu R K, Bai H L, Bao H, Zhao K, Li Y N, et al. FlatQuant: Flatness matters for LLM quantization. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 57587–57613
  - 33 Ding Y F, Qin H T, Yan Q H, Chai Z H, Liu J J, Wei X L, et al. Towards accurate post-training quantization for vision Transformer. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022. 5380–5388
  - 34 Jeon Y, Lee C, Cho E, Ro Y. Mr.BiQ: Post-training non-uniform quantization based on minimizing the reconstruction error. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 12319–12328
  - 35 Frantar E, Ashkboos S, Hoefler T, Alistarh D. OPTQ: Accurate quantization for generative pre-trained Transformers. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.
  - 36 Shao W Q, Chen M Z, Zhang Z Y, Xu P, Zhao L R, Li Z Q, et al. OmniQuant: Omnidirectionally calibrated quantization for large language models. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
  - 37 Liu Z C, Zhao C S, Fedorov I, Soran B, Choudhary D, Krishnamoorthi R, et al. SpinQuant: LLM quantization with learned rotations. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
  - 38 Oh S, Sim H, Kim J, Lee J. Non-uniform step size quantization for accurate post-training quantization. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 658–673
  - 39 Lee E H, Miyashita D, Chai E, Murmann B, Wong S S. LogNet: Energy-efficient neural networks using logarithmic computation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 5900–5904
  - 40 Lin Y, Zhang T Y, Sun P Q, Li Z, Zhou S C. FQ-ViT: Post-training quantization for fully quantized vision Transformer. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI, 2022. 1173–1179
  - 41 Berger T. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*. Hoboken: John Wiley & Sons, 2003.
  - 42 Yao Z W, Dong Z, Zheng Z C, Gholami A, Yu J L, Tan E, et al. HAWQ-V3: Dyadic neural network quantization. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 11875–11886
  - 43 Song H Y, Dharmapurikar S, Turner J, Lockwood J. Fast Hash table lookup using extended bloom filter: An aid to network processing. *ACM SIGCOMM Computer Communication Review*, 2005, **35**(4): 181–192
  - 44 Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248–255
  - 45 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 740–755
  - 46 Bisk Y, Zellers R, le Bras R, Gao J F, Choi Y. PIQA: Reasoning about physical commonsense in natural language. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 7432–7439
  - 47 Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv: 1803.05457, 2018.
  - 48 Clark C, Lee K, Chang M W, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 2924–2936
  - 49 Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. HellaSwag: Can a machine really finish your sentence? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019. 4791–4800
  - 50 Sakaguchi K, le Bras R, Bhagavatula C, Choi Y. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 2021, **64**(9): 99–106
  - 51 Merity S, Xiong C, Bradbury J, Socher R. Pointer sentinel mixture models. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.
  - 52 Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 2020, **21**(140): 1–67
  - 53 Liu Y, Duan H D, Zhang Y H, Li B, Zhang S Y, Zhao W B, et al. MMBench: Is your multi-modal model an all-around player? In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 216–233
  - 54 Yang A, Li A F, Yang B S, Zhang B C, Hui B Y, Zheng B, et al. Qwen3 technical report. arXiv preprint arXiv: 2505.09388, 2025.
  - 55 Bai S, Cai Y X, Chen R Z, Chen K Q, Chen X H, Cheng Z S, et al. Qwen3-VL technical report. arXiv preprint arXiv: 2511.21631, 2025.
  - 56 Zheng X Y, Li Y Y, Chu H R, Feng Y, Ma X D, Wang Z N, et al. An empirical study of Qwen3 quantization. *Visual Intelligence*, 2026, **4**(1): Article No. 11
  - 57 Liu Z H, Wang Y H, Han K, Zhang W, Ma S W, Gao W. Post-training quantization for vision Transformer. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. Article No. 2152
  - 58 Li R D, Wang Y, Liang F, Qin H W, Yan J J, Fan R. Fully quantized network for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2805–2814
  - 59 Wightman R. PyTorch Image Models [Online], available:

<https://github.com/huggingface/pytorch-image-models>, July 3, 2026

60 Ma Y X, Li H X, Zheng X W, Ling F, Xiao X F, Wang R, et al. AffineQuant: Affine transformation quantization for large language models. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

61 Chen M Z, Liu Y, Wang J H, Bin Y, Shao W Q, Luo P. PrefixQuant: Eliminating outliers by prefixed tokens for large language models quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: [10.1109/TPAMI.2026.3711802](https://doi.org/10.1109/TPAMI.2026.3711802)

62 Lin J, Tang J M, Tang H T, Yang S, Chen W M, Wang W C, et al. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In: Proceedings of the 7th Annual Conference on Machine Learning and Systems. Santa Clara, USA: MLSys, 2024. 87-100



**江佳浩** 同济大学电子与信息工程学院硕士研究生. 主要研究方向为具身智能和计算机视觉.  
E-mail: [jjhao@tongji.edu.cn](mailto:jjhao@tongji.edu.cn)  
(**JIANG Jia-Hao** Master student at the College of Electronic and Information Engineering, Tongji University. His research interests include embodied intelligence and computer vision.)



**尹 鹏** 腾讯科技 (深圳) 有限公司算法工程师. 2025 年获得电子科技大学硕士学位. 主要研究方向为深度学习和模型量化.  
E-mail: [raymond2yp@gmail.com](mailto:raymond2yp@gmail.com)  
(**YIN Peng** Algorithm engineer at Tencent Technology (Shenzhen) Co., Ltd. He received his master degree from University of Electronic Science and Technology of China

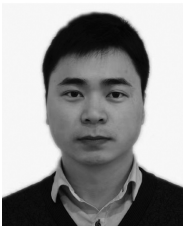
in 2025. His research interests include deep learning and model quantization.)



**王轩瀚** 同济大学计算机科学与技术学院研究员. 主要研究方向为以人为中心的视觉智能和多模态具身智能. 本文通信作者.  
E-mail: [wxuanhan@hotmail.com](mailto:wxuanhan@hotmail.com)  
(**WANG Xuan-Han** Researcher at the School of Computer Science and Technology, Tongji University. His research interests include human-centered visual intelligence and multimodal embodied intelligence. Corresponding author of this paper.)



**曾鹏鹏** 同济大学计算机科学与技术学院研究员. 主要研究方向为多模态理解, 计算机视觉和具身智能.  
E-mail: [is.pengpengzeng@gmail.com](mailto:is.pengpengzeng@gmail.com)  
(**ZENG Peng-Peng** Researcher at the School of Computer Science and Technology, Tongji University. His research interests include multimodal understanding, computer vision, and embodied intelligence.)



**宋井宽** 同济大学计算机科学与技术学院教授. 主要研究方向为多模态学习和具身智能.  
E-mail: [jingkuan.song@gmail.com](mailto:jingkuan.song@gmail.com)  
(**SONG Jing-Kuan** Professor at the School of Computer Science and Technology, Tongji University. His research interests include multimodal learning and embodied intelligence.)