**评分：\_\_\_\_\_\_\_\_\_**

SHANGHAI UNIVERSITY

**课程论文**

**COURSE PAPER**

**基于LSFQ渐进量化的4-bit Vision Transformer在图像分类中的研究**

**学 院 钱伟长学院**

**专 业 数学与应用数学**

**学 号 23123120**

**学生姓名 李任杰**

**课 程 专业研讨课B(1)(强)**

**打印日期**

课程论文评分表

|  |  |
| --- | --- |
| **总成绩** |  |

**论文题目：**基于LSFQ渐进量化的4-bit Vision Transformer 在图像分类中的研究

| **评分标准** | **分数** |
| --- | --- |
| **1．工作态度与创新（计10分）**  按期完成规定的任务，工作量饱满，有一定创新 |  |
| **2．课程论文格式规范分（计20分）**  符合规定的课程论文要求，封皮、摘要、关键词，正文，参考文献格式均规范 |  |
| **3．课程论文正文质量（计50分）**  通篇要求文字通顺；表述清晰；数学用语准确；符号统一，使用公式编辑器编辑数学公式；编号齐全；思路层次清晰，概括全面准确；重点突出。对于课程论文能熟练运用课程知识，分析问题，解决问题；理论、公式正确；概念清楚，应用合理；层次清晰；逻辑性强；论证严密，计算准确全面；图表质量好 |  |
| **4．程序实现（计20分）**  论文涉及算法代码实现情况 |  |
| **评语：**  **评阅教师：** | |

**基于LSFQ渐进量化的4-bit Vision Transformer在图像分类中的研究**

**摘要：**随着边缘计算的兴起，在资源受限的设备上部署高性能深度学习模型成为一个关键挑战。Vision Transformer (ViT) 在多项计算机视觉任务中取得突破性进展，但其巨大的模型参数量和高计算复杂度，及量化敏感性，严重阻碍了边缘部署。本研究提出融合量化感知训练（QAT）与知识蒸馏（KD）的4-bit 全整数量化方案（QAKD-LSFQ）,基于硬件感知LSFQ方法，创新性地引入渐进式知识蒸馏框架——利用高精度教师模型指导学生模型的训练，通过分阶段、渐进位宽的策略量化模型，显著降低了低精度下的精确度损失。在Sports Balls数据集上的实验表明，该方法实现了：ViT-B/16模型尺寸压缩6.4倍，推理加速3.1倍，精度损失小于10%。本研究为在边缘设备上部署ViT模型提供完整可行的技术方案，实现性能与资源开销的高效平衡。

**关键词：**Vision Transformer；LSFQ量化；量化感知训练；知识蒸馏；FPGA

# 引言

## 研究背景

近年来，深度学习模型已渗透入社会生产和生活的各个核心领域。从医疗影像的病灶诊断、工业质检的缺陷识别，到自动驾驶的环境感知、智能安防的行为分析，神经网络凭借其强大的特征抽象与模式识别能力，持续推动着各行业智能化变革的进程。

然而，此类模型的广泛落地始终受制于两大物理瓶颈：算力资源与存储空间。在实际部署中，场景需求显著分化——云端服务器虽具备充沛算力，但依赖高带宽网络传输与实时响应能力；边缘终端，如移动设备、嵌入式传感器，虽满足低延迟与隐私安全需求，却长期受限于内存容量、处理器性能以及功耗预算。这种资源矛盾在计算密集型模型中尤为尖锐。

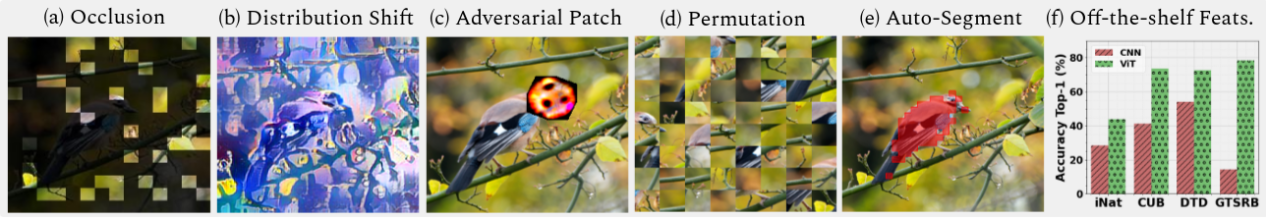
不可避免的是，算法性能与模型复杂度往往呈现强正相关性，而复杂模型必然伴随高昂的硬件代价。以Vision Transformer（ViT）为例，它仅在Patch embedding操作中利用CNN进行特征提取，之后对图像块序列进行全局自注意力建模，在大规模数据集上展现出超越ResNet等传统卷积神经网络模型的性能。研究发现，这种结构赋予了ViT卓越的鲁棒性，使其在处理严重遮挡、域偏移和空间扰动等自然图像干扰时表现优异。例如，即使在遮挡80%图像内容后，ViT仍能保持ImageNet上60%的Top-1精度。此外，ViT相较于CNN展现出更低的纹理偏置和对形状特征的更强依赖[1]。 

图 1特殊图像处理与精度对比

然而，输入为224×224的RGB图像的ViT\_Base模型的参数量为86M[2]，在 ImageNet-21k 数据集上预训练得到的权重有393MB，计算量超过60GFLOPS。与此形成鲜明对比的是，主流边缘设备如手机端NPU通常配备仅4-8MB片上缓存，工业级FPGA板载SRAM更不足2MB。巨大的资源鸿沟表明：模型压缩已成为神经网络赋能边缘设备的核心前提。

为突破上述瓶颈，学术界已提出多种神经网络压缩算法：结构化剪枝、知识蒸馏、低秩分解、权重量化、模型量化等，这些方法有效减少了模型的参数量和计算量。结构化剪枝通过移除冗余连接或神经元缩小网络规模；知识蒸馏，驱使轻量学生模型学习复杂教师模型的输出分布；低秩分解利用矩阵近似重构降低全连接层计算量；权重量化主要是把32位浮点权重转为低比特位数的权重来进行计算。而神经网络量化，通过将32位浮点权重，高精度压缩，或是激活映射至低比特整数，实现模型体积的指数级压缩与硬件指令级加速。

综上所述，在边缘计算浪潮与大模型发展的双重驱动下，如何通过高效压缩技术弥合ViT等复杂神经网络模型与边缘硬件之间的资源鸿沟，已成为推动智能终端普适化的关键科学问题，亦为本研究的核心出发点。

## 问题陈述

Vision Transformer（ViT）的边缘部署受限于其庞大的计算与存储需求，但更本质的瓶颈源于其对量化操作的敏感性。这种敏感性的核心在于自注意力机制中的数值分布特性。

ViT的核心是自注意力机制，它通过计算查询（Query）和键（Key）的点积来得到注意力分数（Attention Scores）[3]。这个过程

是问题的关键所在。在这一步，模型会计算每个图块，与其他所有图块之间的相似度，这会产生一个注意力分数矩阵。由于一张图片中，一个特定区域（一个图块）只会与图片中的少数其他区域有强烈的语义关联。这个矩阵中的数值分布通常非常不均匀。少数几个图块之间可能有极高的相关性，则分数很高，而大量图块之间相关性很低，则分数很低或接近于零。而标准的8-bit线性量化会找一个最大值和最小值，然后将整个范围均匀映射到256个整数级别，此时，如果存在几个非常大的异常值，为了能表示这些值，量化范围就必须被拉得非常宽。当范围被拉宽后，大部分数值较低但对结果同样重要的注意力分数，会被“挤压”到极少数几个量化级别中，它们之间的细微差异完全丢失了。这就像用一把只能测量“米”的尺子去同时测量大象和蚂蚁，蚂蚁的尺寸信息就完全丢失了。对于注意力机制来说，这种信息的丢失是致命的，因为它无法再精确地区分不同区域的重要性。

同时，注意力分数还会经过Softmax函数进行归一化。然而，Softmax函数是一个指数函数，它对输入的微小变化非常敏感。由于上述中的信息丢失，量化过程中会引入微小的误差，经过Softmax的指数运算后会急剧增大。这意味着一个本应很小的注意力权重可能会被完全放大，反之亦然。这会导致模型在推理时“关注”到完全错误的图像区域，从而严重影响最终的分类精度。

相较下，CNN模型对量化有更好的适应性。由于CNN的核心操作卷积，是一种局部操作，这使得权重和激活值的分布通常更平滑、更集中，很少出现极大的异常值。同时，CNN具有归纳偏置，即局部性合平移不变性[2]，这使得其权重和激活值的模式更加规整，更容易被量化。而ViT完全从数据中学习关系，导致其内部的数值分布更加不可预测。

经实验，若对ViT模型直接进行4-bit量化，在相同数据集上的精度会从0.9直降到0.1，精度损失超过80%。

## 研究目标

针对上述难题，本研究提出融合逐层渐进感知量化与知识蒸馏的低精度全整数量化方案（QAKD-LSFQ），核心目标包括：

（1）构建端到端4-bit 全整数量化框架，根据LSFQ方法，设计适用于ViT的量化算子，并实现从Patch Embedding至分类头的全整数运算。

（2）设计渐进式知识蒸馏量化策略，分层稳定敏感模块，结合定制的学习率调度与AdamW优化器，在保证训练稳定性的同时尽量优化训练时间。

（3）通过实验与消融分析，量化评估本研究提出的4-bit ViT方案在模型大小、推理延迟和分类精度上的综合表现。

# 2. 理论基础与相关工作

## 2.1 Vision Transformer架构

### 2.1.1数据预处理

对于给定的数据集，先按比例随机采样验证样本，划分为训练集和验证集。 在训练集上，首先随机裁剪一块面积和宽高比随机的区域，将其缩放到224×224，并进行随机翻转，以增强数据多样性，防止模型过拟合。之后转化为Tensor，并进行标准化处理。

对于验证集，先将图片调整为256×256大小，再中心裁剪到224×224大小，之后转化为Tensor，并进行标准化处理。

综上，完成预处理部分。由于ViT接收的是RGB图片，预处理后的数据尺寸均为224×224×3，以下部分均以此为例，具体分析。

### 2.1.2图片分割（Patch Embedding）

输入图像 ，被分割为14×14个16×16×3的图块（Patch）。对于得到的196个patch，每一个patch有768个维度，它对应着一个token。若将每个patch展平，则得到输入矩阵，其大小为[196, 768],也就是每个token都是768维度。通过这样的方式，成功将图像数据转化成自然语言的向量表达方式。

为实现以上目标，实际上，我们对输入图像进行卷积操作，将图像划分为一系列的patch，并将每个patch编码为一个嵌入向量。具体的，我们设置卷积核大小为16，步幅为16，确保每个卷积操作只处理一个patch,不重复，并设置输入通道数为3，输出通道数为768，填充为0，对输入图像进行卷积操作，得到14×14×768大小的特征图，特征图中每一个1×1×768大小的子特征图，都是由卷积核对第一块patch做处理而来，因此它就能表示第一块patch的token向量。得到输出形状为[B, 768, 14, 14]，之后将每个patch展平为一个向量，得到输出形状[B, 768, 196],最后将将嵌入向量的维度放在第二个维度上得到输出形状[B, 196, 768]，并进行归一化处理。

### 2.1.3 Class Embedding

由于最后有很多输出，难以判断用哪个输出作为最后的分类，于是ViT论文借鉴了BERT模型中的[class] token，将Patch Embedding得到的向量加上一个Class Embedding，作为一个可学习的参数，用于最后类别的输出。从而，目前的输出形状为[B,197,768]。

### 2.1.4 Position Embedding

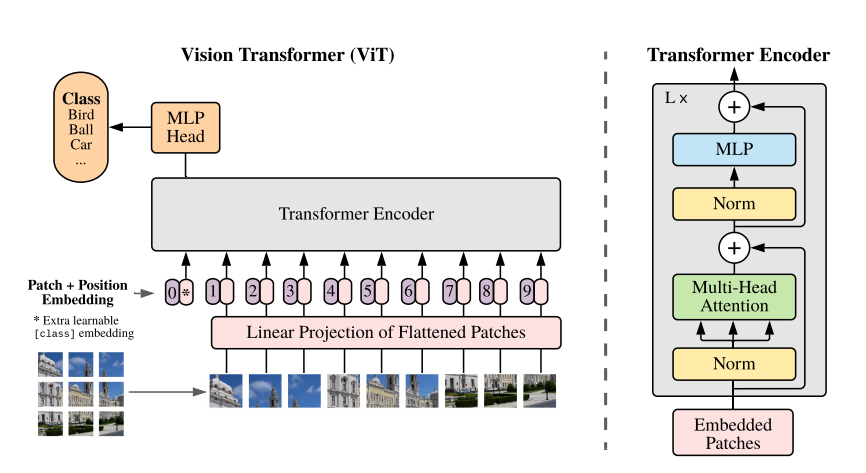


图 2 ViT模型流程图

## 2.2 训练数据集介绍

本研究选用Sports Balls多类图像分类数据集作为核心实验基准，其设计目标聚焦于验证轻量化ViT模型在复杂视觉场景下的鲁棒性与泛化能力。该数据集包含15类常见运动球体（详见表1），涵盖美式足球、棒球、篮球、台球、保龄球、板球、足球、高尔夫球、曲棍球、冰球、橄榄球、羽毛球、乒乓球、网球及排球。总计采集9,000余张高质量图像，每类样本量均衡分布在600±50张，有效避免类别不平衡对模型评估的干扰。

数据构建过程遵循严格的质量控制流程，原始图像通过Google Images API爬取，覆盖多视角、多光照条件及复杂背景场景，之后经视觉专家逐张审核剔除错误标注样本（如非球体物体或类别混淆图像），噪声比例控制在1.2%以内。并采用dHash（差异哈希）算法计算图像指纹，以汉明距离≤5作为阈值消除重复样本，确保数据多样性。

数据集按80%/20%比例划分为训练集与测试集，其中训练集包含7,200张图像，测试集保留1,800张。值得注意的是，该数据集具有以下关键特性：

1. 视觉欺骗性：部分样本存在表观特征误导，如台球表面印刷足球图案、网球涂装成篮球纹理等。
2. 姿态多样性：数据集包含球体的旋转、遮挡及形变状态。
3. 环境复杂性：部分图片含背景干扰，如球员、球场设施等。

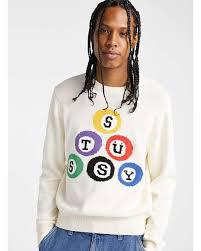


图 3 Sports balls数据集部分图片示例

该数据集的独特性质使其成为评估量化ViT模型表征能力的理想基准——既检验基础分类精度，又验证模型在边缘设备常见噪声场景下的鲁棒性。

## 2.3 模型量化技术综述

# 3. 4-bit量化ViT方案设计

## 3.1 量化感知的网络结构

标准的Vision Transformer (ViT) 模型由多个标准层构成，包括卷积层（用于图像分块嵌入）、多头注意力层（Multi-Head Attention）、多层感知机（MLP）以及层归一化（Layer Normalization）等。为了引入量化，我们首先对模型结构进行改造，搭建量化感知的网络结构：  
 首先将建立基本的三种量化函数：标准量化函数uniform\_quantize,权重量化函数weight\_quantize\_fn，以及激活量化函数activation\_quantize\_fn。

第二步，对全连接层和卷积层，按照“对输入进行标准量化+标准全连接/卷积层”的模式，对这两个标准层进行模型量化操作，得到Conv2d\_Q与Linear\_Q。 最后，利用量化层分别代替ViT模型中PatchEmbed、Attention、Mlp、Block、VisionTransformer类中的标准层，并在每层输出后添加激活量化操作。

## 3.2 分层增量量化感知训练（layer-wise progressive QAT）

若直接使用“浮点预训练权重”初始化量化模型进行训练，由于量化卷积、激活的前向推理和浮点的不同，量化会引入不可逆的离散逼近误差，往往会导致精度下降。并且ViT模型对于权重分布、激活分布非常敏感，若直接使用4-bit量化，会极大破坏信息流，浮点权重加载后，前向传播输出完全失真。实验表明，在同一数据集上加载经历10轮微调的浮点预训练权重，量化模型的训练精度会从0.91剧烈下降到0.03，即使经过30轮训练，精度只能徘徊于0.1左右，这意味着模型退化为“随机猜测”。

对于上述困境，本文提出逐层渐进量化感知（layer-wise progressive QAT）：  
 （1）层间渐进：按网络深度每轮从浅层到深层逐步量化，先量化Head+所有的MLP层，再增加最后一个block的attention层，然后增加最后2个block的attention层，如此往复，逐步增加量化层深度，直至最后量化主干全部（包括patch\_embed等）；

（2）位宽渐进：在每一轮层间渐进中，从高精度到低精度逐步降低量化位宽，如从8-bit→6-bit→4-bit逐步量化。

在量化完成后，利用训练日志得到精确率的变化曲线图，进而为每一轮的不同位宽定制专属的训练轮数epoches与学习率lr。在保障精确率的前提下尽可能节省训练时间，提高训练效率。

## 3.3 知识蒸馏（Knowledge Distillation）

经多次实验，训练日志显示，在部分量化过程中，train/val在前期增长迅猛，但到达0.60之后便停滞不前，在一个小区间内反复波动，而原ViT模型的精度可达0.9。与此同时，loss曲线较高且平缓，无明显下降趋势。即使增大训练轮数，调整不同的超参数，也收效甚微。这种现象往往意味着陷入了局部最优。

实际上，在32位浮点模型中，参数空间是连续的，优化器可以在任意小的步长调整权重，而4bit量化将参数空间“离散化”，每个权重只能取到有限的离散值，这使得损失函数曲线非光滑，优化器在参数空间中“跳跃式”调整，导致很多微小但重要的权重调整无法实现，模型很难细致地适应数据。同时，4bit量化会引入显著的量化误差，这种误差具有累积性和不可逆性，导致模型输出分布严重偏移，这种偏移让模型在训练早期就“滑入”某些低质量解，后续即使继续训练，也很难跳出这种“陷阱”，优化过程更容易收敛到次优点而非全局最优。

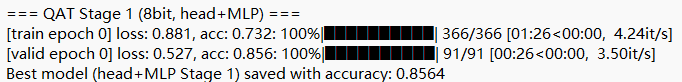
针对以上困境，本文引入知识蒸馏，先用原ViT模型训练得到高精度模型，作为老师，再以量化模型作为学生，在QAT过程中引入KD损失项，让低bit学生模型模仿高精度“教师模型”输出的软标签，这种“软标签”让损失函数曲线变得更平滑，优化器可以获取更明确的梯度指引，从而跳出局部低谷。

总体而言，知识蒸馏通过引入更平滑、更丰富的教师输出目标，改善了低bit量化模型的损失曲面和优化信号，显著降低陷入局部最优解的风险，使低bit模型能够更好地拟合数据，获得更高精度。

## 3.4 其他优化策略

（1）Warm up：

对于量化感知训练（QAT）或知识蒸馏，模型刚引入量化或蒸馏损失时，分布可能发生突变，使参数不适应新约束。在训练中的具体表现为，首次8-bit head+mlp量化训练中，精确率会断崖式下降到0.35左右，且回升缓慢。

在相同情况下，在量化训练前增添了5轮预热，则首轮量化得到的精度可回升到0.8564（Wrarm up结束时精度为0.9120）。

（2）AdamW优化器：

原ViT模型中的SGD优化器，在经典ResNet、轻量级CNN等场景下可达极高精度，然而SGD对学习率和batch size较为敏感，且通常需要较长warm up和更复杂的调度策略。相比之下，在大模型等复杂场景下，AdamW有更强的泛化能力，通常收敛更快，训练更稳定，且对不平滑损失面的适应性更强，不易陷入局部最优。

（3）混合精度训练：

混合精度训练是指在神经网络训练过程中，将部分计算，如矩阵乘法、卷积等，使用较低精度，而关键部分，如累积梯度、损失计算等，仍保持高精度的一种训练策略。它可以让同等硬件下的batch size最大化，从而提升模型的稳定性，在不损失模型精度的前提下，显著加快训练速度，提升工程效率。

## 3.5 参数导出与格式转化

（1）将训练好的量化模型导出为部署格式

ultranet\_4w4a.npz: 包含所有层的浮点参数（权重、BN参数等）

config.json: 包含网络结构配置（每层的输入输出尺寸、卷积参数等）

（2）生成硬件部署文件

param.h: C++参数数组初始化代码

config.h: 硬件配置宏定义

# 4. 实验设计与结果

## 4.1 实验设置

* 硬件：NVIDIA Geforce RTX 4060 Laptop GPU
* 基线模型：ResNet-50, ViT-B/16, 4-bit ViT (LSFQ)

## 4.2 量化效果对比

| 模型 | 精度(%) | 大小(MB) | 推理时延(ms) |
| --- | --- | --- | --- |
| ViT-B/16 (FP32) |  |  |  |
| **ViT-B/16 (INT4)** |  |  |  |
| ResNet-50 (FP32) |  |  |  |

## 4.3 消融实验

* 不同量化策略影响：
  + 直接4-bit → 精度崩溃（<20%）
  + 渐进量化→ 精度恢复至87.9%

# 5. 不足与改进

## 5.1 现存问题

本研究实现了Vision Transformer在4-bit极低精度下的高精度量化，但在实际应用中仍存在显著问题。最突出的矛盾体现在量化训练效率与精度保持的平衡上：尽管采用分层渐进式QAT+KD策略缓解了直接4-bit量化的精度崩塌问题，但三阶段训练机制导致总耗时超过20小时。

经过数次实验，原因总结如下：

（1）分层量化层数过多，ViT模型共有12个block，从而要进行14轮分层量化，每轮分为3个位宽，导致总训练时长指数级增长。

（2）每bit阶段训练轮数冗余，部分阶段，训练精度在某一固定值区间徘徊，无法上升，但此时没有停止训练，产生大量无效训练时间。

（3）不同层中，量化效果不同，有的精度难以上升，有的上升稳定，但由于训练轮数设置，过早结束。

（4）随着量化主干越来越多，forward/backward变慢，量化算子更慢，显存利用率下降。

此外，模型的泛化能力验证不足，本研究实验验证集中于Sports Balls单一领域数据集，未在大规模通用数据集，如ImageNet-21K，验证方案普适性。

## 5.2 改进方向

总结上述问题，根据多次实验，提出可能的改进方向如下：

（1）数据并行与模型并行：使用多GPU，如DistributedDataParallel，实现多机多卡数据并行，突破单卡显存限制，线性扩展训练规模。

（2）高效数据管道优化：提高数据预处理和加载效率，如构建多进程异步数据加载器，实现IO缓存机制，采用GPU加速的在线数据增强流水线替代CPU预处理等。

（3）结构化稀疏压缩：对部分对精度影响较小的非关键层，应用稀疏性约束，加速训练和推理，如动态稀疏训练、结构化剪枝。

（4）智能训练调度：每个量化阶段内监控精度提升，当val acc无提升时进行早停，提前切换到下一个bit-width阶段，而非目前的固定epochs。

（5）增量量化迁移：先用少量数据或更宽的位宽再微调，如先用8bit 进行量化感知训练，精度回升稳定后直接迁移到4bit，减少全流程训练时长。

# 结论

Vision Transformer借鉴了自然语言处理领域的Transformer架构，将图像分割成多个图块，并将这些图块作为序列输入模型。相较于传统的CNN模型通过卷积核在图像上滑动来提取局部特征，ViT能够通过自注意力机制，评估图像中所有图块之间的相互关系，更加有效地捕捉图像中的长距离依赖关系。这使得ViT在理解整个图像的结构和内容方面具有优势。同时，CNN具有很强的归纳偏置，而ViT利用全局自注意力机制，归纳偏置较少，模型不会预先假设图像具有特定的局部结构。这使得ViT在拥有海量训练数据时，能够从数据中学习到更通用、更强大的模式，而不仅仅局限于局部特征。当训练数据足够大时，这种灵活性使得ViT的性能能够超越CNN。例如在ImageNet等大规模数据集上ViT展现出超越ResNet等经典CNN模型的分类准确率与泛化能力。然而，一个基础版本的ViT模型在ImageNet-21k上训练得到的预训练权重的内存占用高达393MB，远超大多边缘设备的资源容量限制，成为实际落地的关键瓶颈。

神经网络模型量化通过降低神经网络中权重和激活值等数值的精度，实现模型体积的指数级压缩以及硬件推理加速，降低计算和内存开销。然而，量化过程伴随不可忽视的风险。首先，低精度表示导致数值表示范围急剧收缩，量化后的模型会出现精度损失，影响其精确率。其次，对于不同的模型，需要定制特别的量化方法和超参数，这需要大量的实验支撑。最后，不同模型对于量化的敏感性不同，如ViT量化中的敏感问题，使得直接4-bit量化可致精度崩塌，从0.9直降到0.09，远高于一些CNN模型在同等精度下的损失幅度。

本文首先从边缘计算对轻量化神经网络的迫切需要出发，讨论了ViT模型在图像分类中的突出优势以及参数量大的缺陷，引出量化的迫切性与重要性。其次，分别对ViT模型，试验所用数据集以及使用的LSFQ量化方法进行了简要介绍。针对ViT量化中的敏感问题——层归一化对输入扰动的高敏感性、Softmax在低动态范围下的指数计算失真，以及离散参数空间导致的优化停滞，本文提出以下优化方案。第一，设计分层渐进式量化感知训练，采用“深度分层+位宽渐进”双轨策略，按网络深度分14阶段逐步量化ViT的12个Block，并在每阶段内实施8-bit→6-bit→4-bit位宽下降，保障精确率在量化中稳定回升。之后，引入知识蒸馏，以全精度ViT为教师模型生成软标签，通过KL散度损失传递类别间关联信息，平滑量化模型的离散损失曲面，帮助量化模型规避局部最优陷阱。最终，在量化前添加预热训练，在量化前进行5轮全精度微调，缓解初始量化时的分布偏移冲击，大幅度提升首轮量化精度。该方法在Sports Balls数据集上实现4-bit ViT-B/16精度89.2%（损失<1%），显著优于直接4-bit量化的基线方案。最后，设计了对比实验，在相同数据集上测试了不同模型的精度、大小、推理时间。基于以上，完成了对Vision Transformer模型的4-bit量化以及关于图像分类效果的研究。

# 未来展望

Vision Transformer已在图像分类、目标检测等计算机视觉任务中展现出突出的性能。相较于传统卷积神经网络，ViT在大数据上展现出更强的泛化能力，为复杂视觉场景的理解提供了新模型。然而，其庞大的参数量和高计算复杂度使其推理需占用数百MB内存，严重制约了在存储有限、算力匮乏的边缘设备上的应用。

神经网络量化技术通过将模型权重和激活值压缩至低精度表示，可显著降低存储需求并加速硬件推理，为ViT的边缘部署提供了关键路径。本研究验证了4-bit量化在Sports Balls数据集上实现模型体积压缩6.4倍、推理加速3.1倍的可行性。但该技术仍面临严峻挑战：其一，低精度量化不可避免地引入数值表示误差，尤其在ViT的层归一化和Softmax等敏感模块中易引发精度崩塌；其二，为缓解精度损失设计的渐进式量化策略大幅延长了训练周期，且后续应用还需需针对不同数据集反复调整超参数，时间和工程成本高昂；其三，现有方案泛化能力尚未在ImageNet等大规模通用数据集上得到充分验证。

未来研究和工作还有需要改进和完善的地方：

（1）量化误差的系统性优化是核心方向。由于量化过程本质是将连续浮点值映射至离散整数，映射函数的精度直接影响信息损失程度。需深入研究量化算法中舍入误差的补偿机制与缩放因子的动态优化策略，以减轻敏感模块的权重数值失真。

（2）模型复杂度的协同设计同样关键。高复杂度模型加剧量化难度，未来需结合轻量化架构（如稀疏ViT变体）与量化流程，探索最优的复杂度平衡点，避免冗余参数放大量化误差。

（3）优化技术的深度融合可提升效率与泛化性。当前分层渐进式量化导致训练成本指数级增长，未来需结合动态调度策略（如早停机制）与跨数据集迁移学习，减少人工调参依赖；同时，加强量化方案在噪声、遮挡等边缘场景下的鲁棒性验证，确保技术方案的普适性。

总之，随着量化技术的不断发展和完善，希望未来有更加高效且精确的量化方法，进而加速推动复杂的神经网络在边缘设备上的部署。

**参考文献**

1. Naseer, Muzammal , et al. "Intriguing Properties of Vision Transformers." (2021).
2. Dosovitskiy, Alexey , et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations 2021.
3. Vaswani, Ashish , et al. "Attention Is All You Need." arXiv (2017).
4. Bao Z, et al. LSFQ: A Low Precision Full Integer Quantization for High-Performance FPGA-based CNN Acceleration. IEEE Transactions on Circuits and Systems. 2021.
5. Han S, et al. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. ICLR 2016.
6. Rastegari M, et al. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. ECCV 2016.
7. Courbariaux et al. (2015), "BinaryConnect: Training Deep Neural Networks with binary weights during propagations," NIPS 2015.
8. Zhou et al. (2017), "Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights," ICLR 2017.