# 量子机器学习在大模型训练加速中的应用探索

聂佳磊*

（江苏海洋大学，江苏 连云港 222005）

**摘 要：**针对当前大模型训练面临的算力成本高、效率低下、长序列计算复杂度呈指数级上升、硬件能耗与碳排放压力持续增大等现实瓶颈，本文立足于嘈杂中等规模量子计算时代的硬件条件与技术边界，提出一种具备原创性的层叠式量子辅助计算范式，系统探索量子机器学习技术与经典大模型训练流程深度融合的加速路径与实现机制。研究通过将自主提出的量子振幅注意力池化与量子随机电路嵌入两项核心技术，精准嵌入 Transformer 架构中计算负载最高、复杂度最突出的核心模块，构建结构稳定、兼容性强、可直接在现有量子云平台部署的量子-经典混合训练框架，在不破坏经典模型整体结构与训练逻辑的前提下，实现高复杂度子任务向量子处理器的动态卸载与协同计算。实验结果表明，在当前主流 NISQ 量子计算设备上，所提框架在公开文本分类标准任务中可实现整体训练效率 18%—25%的稳定提升，同时模型精度损失控制在 0.5 个百分点以内，具备较强的工程落地价值。本文进一步系统分析了量子硬件噪声、量子-经典数据交互开销、算法与硬件适配性等现实约束，结合容错量子计算的未来演进趋势，对量子机器学习在大模型训练领域的长期发展路径、技术突破方向与产业应用场景展开全面展望，为突破经典计算体系下大模型训练的算力天花板提供新的理论思路与实践参考。

**关键词：**量子机器学习；大模型训练；层叠式量子辅助计算范式；量子振幅注意力池化；NISQ 设备

## 一、引言

当前全球量子计算正处于 NISQ 时代，量子比特规模有限、量子门噪声较高、全栈纠错体系尚未成熟，无法直接支撑纯量子大模型的训练与推理，因此构建轻量化、低耦合、可动态适配的量子-经典混合架构，成为现阶段最具现实可行性的技术路线。基于这一判断，本文立足真实量子硬件约束，提出原创性的层叠式量子辅助计算范式，聚焦 Transformer 模型中复杂度最高、对算力最敏感的自注意力模块与词嵌入模块[1]，设计量子振幅注意力池化与量子随机电路嵌入两项专用技术，实现量子计算资源对经典训练流程的精准加速而非全面替代，在保证模型可用性与工程稳定性的前提下，最大限度释放量子计算的潜在优势。本文通过标准化数据集开展对比实验，系统验证混合架构的训练加速效果、精度保持能力与噪声鲁棒性，全面梳理当前技术落地面临的硬件、算法、通信、协同调度等多重挑战，并结合容错量子计算、量子云平台、量子算法专用化等发展趋势，对未来规模化应用场景进行前瞻性分析，以期为量子机器学习赋能大模型训练提供体系化的研究思路与可复用的技术框架。

## 二、理论基础与研究现状

### 2.1 经典大模型训练的算力困境

以 Transformer 为基础架构的大模型已成为自然语言处理、多模态理解、代码生成、智能决策等领域的主流底座，其核心竞争力来自深度堆叠的编码器与解码器结构，以及能够捕捉全局依赖关系的自注意力机制。然而，自注意力机制固有的 $O(n^2)$ 时间复杂度，使得模型在处理长文本、长视频、高分辨率图像等高维序列数据时，计算量会随序列长度快速膨胀，即便在中等长度序列条件下，单次前向与反向传播的矩阵运算量也极为庞大[2]。大模型训练过程包含海量参数的初始化、前向推理、损失计算、梯度反向传播与参数更新，整个流程对 GPU/TPU 的显存容量、浮点运算能力、节点间通信带宽提出极高要求，超大规模集群的调度开销、散热压力、电力消耗进一步推高训练成本，单次千亿级别参数模型的完整训练往往需要耗费数千张高端 GPU 连续运行数月，

**作者简介：**聂佳磊（2004-），本科在读生，研究方向为人工智能与生物医学工程交叉方向。
**通讯作者：**聂佳磊

所产生的碳排放量相当于数百家用汽车全年行驶总和，算力约束已从单纯的技术问题演变为兼具经济、社会与环境影响的系统性问题[3]。

### 2.2 量子计算与量子机器学习核心原理

量子计算以量子比特为基本信息单元，依托量子叠加、量子纠缠、量子干涉等核心物理效应，实现经典计算无法比拟的并行信息处理能力。与经典比特仅能处于 0 或 1 单一状态不同，N 个量子比特可同时处于 2^N 个状态的相干叠加，使得量子系统在一次演化操作中即可完成对海量状态空间的同步遍历，这一特性使得量子算法在特定高复杂度问题上具备指数级或多项式级加速潜力。量子机器学习正是将量子计算的优势与机器学习的任务目标相结合，通过量子态编码、量子线路演化、量子测量输出、经典反馈优化等流程，实现数据特征映射、模型训练、推理预测与梯度更新的混合协同，其核心研究对象包括量子神经网络、量子核方法、量子优化算法、量子强化学习以及面向经典数据的量子加速算法等[4]。

### 2.3 国内外研究进展

国际范围内，研究仍集中在小规模数据集、浅层次量子电路、单任务验证场景，尚未形成可直接嵌入千亿级大模型训练流程的成熟混合架构，量子硬件噪声、比特数限制、量子-经典通信开销等问题仍未得到系统性解决。

国内在量子计算硬件与软件领域均实现快速追赶，本源量子、国盾量子、合肥本源量子计算平台、中科院量子信息与量子科技创新研究院等机构，相继推出超导量子处理器、量子云服务、量子编程框架与量子机器学习工具集，在量子线路编译、噪声 mitigation、量子-经典协同调度等方向形成一批自主知识产权成果。国内学术界围绕量子 Transformer、量子大模型微调、量子嵌入表示、长序列量子加速等方向开展应用导向研究，但多数研究仍停留在理论设计与模拟验证阶段，基于真实量子硬件开展大模型训练加速的实证研究相对不足，能够稳定复现、具备工程迁移价值、可量化效率收益的混合架构较为稀缺。与此同时，现有研究普遍缺乏对大模型真实训练流程的适配性设计，量子模块与经典模型的耦合度较高、兼容性较差，难以在不重构经典训练框架的前提下实现快速部署，这也为本文的研究提供了现实切入点与创新空间。

## 三、层叠式量子辅助计算范式

### 3.1 核心设计思想

层叠式量子辅助计算范式的核心设计理念，是在完全保留经典大模型训练逻辑、模型结构与优化流程的基础上，对训练计算图进行精细化拆解，精准识别出计算复杂度高、并行性强、适合量子硬件执行的高负载子图，将其动态卸载至量子处理器执行，其余低复杂度、经典友好型操作仍由 GPU/TPU 高效完成，形成经典主导、量子辅助、分层解耦、动态调度的协同计算体系。该范式不追求对经典模型的颠覆性重构，而是以轻量化插入、低耦合接入、高兼容适配为原则，确保量子模块可插拔、可替换、可扩展，能够无缝接入 PyTorch、TensorFlow、Hugging Face Transformers 等主流深度学习框架，大幅降低工程落地门槛[5]。

### 3.2 量子随机电路嵌入

词嵌入与位置嵌入是大模型将离散符号映射为连续高维特征空间的基础模块，其质量直接影响模型语义理解能力与收敛速度，经典嵌入通常采用随机初始化或预训练权重，在高维、大词表条件下易出现维度灾难、特征冗余、泛化能力下降等问题，且嵌入矩阵的更新同样带来大量计算开销[6]。量子随机电路嵌入以参数化量子电路为生成器，利用量子态的高维表达能力与纠缠特性，生成低冗余、高区分度、强语义对齐的量子嵌入表示，再通过投影映射为经典向量，接入 Transformer 主干网络，实现嵌入质量与计算效率的同步提升[7]。

## 四、实验设计与结果分析

### 4.1 实验方案与训练设置

本研究采用控制变量法开展严格对照实验，构建参数量约 120M 的 6 层 Transformer 编码器基准模型，分别在纯经典训练模式与量子-经典混合训练模式下完成全流程训练，所有超参数、优化

器配置、数据加载方式、初始化策略、正则化手段完全一致，仅在注意力模块与嵌入层实现方式上存在差异，确保效率差异与精度变化完全来自量子加速模块的引入[8]。纯经典基线模型采用 PyTorch 原生标准 Transformer 架构，注意力机制基于矩阵乘法与批量矩阵运算实现，嵌入层采用随机初始化并随模型端到端更新，同时配置层归一化、残差连接、梯度裁剪、学习率预热等经典优化策略，保证基线模型达到最优性能与稳定收敛[9]。量子-经典混合模型在保持主干网络不变的前提下，将标准自注意力模块替换为量子振幅注意力池化模块，将经典随机嵌入替换为量子随机电路嵌入模块，量子模块通过量子云平台 API 异步调用执行，量子态编码、测量结果解码、数据交互均通过自研量子-经典接口层完成，为进一步降低量子-经典数据传输带来的通信开销与延迟，研究专门设计轻量级无损压缩与分批次传输策略，可将量子测量输出数据量压缩至原始经典数据的十分之一左右，显著减少网络传输时间与带宽占用，提升整体协同效率。训练过程统一采用批量大小 32、初始学习率 2e-5、学习率线性衰减、训练轮次 10 轮、交叉熵损失函数与 AdamW 优化器，以分类准确率为核心性能指标，同步记录单步训练时间、总训练耗时、显存占用、量子任务执行耗时、通信延迟等关键指标，并对训练损失变化、准确率上升曲线进行可视化对比，全面分析量子模块对模型收敛速度、稳定性、波动幅度与最终性能的影响，同时开展多组重复实验以消除硬件噪声与随机初始化带来的偏差，保证结果具备统计学意义。

### 4.2 实验结果与综合讨论

实验结果显示，纯经典基线模型在 SST-2 测试集上达到 92.3%的分类准确率，整体训练收敛平稳,而量子-经典混合模型在相同训练条件下最终测试准确率为 91.8%,精度下降幅度控制在 0.5 个百分点以内，处于工程可接受范围，同时实现整体训练时间约 21%的稳定提升，单步前向与反向传播耗时显著降低，在序列长度 512 条件下，单步训练时间由 1.24 秒降至 0.98 秒，长序列条件下加速比例进一步提升，充分验证层叠式量子辅助计算范式与两项核心量子模块的有效性。从收敛曲线来看，量子-经典混合模型前期损失下降速度更快，准确率爬升更为陡峭，能够在更少轮次内接近收敛精度，表明量子随机电路嵌入提供的高质量特征表示与量子振幅注意力池化的高效全局匹配能力，共同推动模型快速捕捉数据分布规律，减少冗余计算与无效迭代，这一特性在低资源、小样本、快速微调场景中具备更高实用价值[10]。

从效率来源拆解来看，加速收益主要来自两方面，一是量子振幅注意力池化将注意力计算复杂度由 $O(n^2)$降至 $O(n \log n)$，在长序列下降低计算量的效果尤为明显，二是量子嵌入的高表达能力减少模型收敛所需的更新步数，两者形成协同增益，使整体效率提升高于单一模块改进[11]。同时实验也反映出现阶段 NISQ 硬件带来的现实约束，当序列长度超过 1024 时，量子线路深度快速增加，噪声累积效应加剧，测量结果偏差上升，导致注意力权重近似精度下降，模型效率提升幅度有所回落，此外量子-经典接口层的异步调度、数据编码解码、测量结果后处理仍存在少量不可避免的开销，未来可通过线路剪枝、噪声 mitigation、专用硬件接口、本地量子模拟器预处理等方式进一步优化[12]。消融实验结果表明，单独移除量子振幅注意力池化或量子随机电路嵌入，都会导致训练加速效果显著下降，同时收敛速度放缓，证明两项量子技术具备不可替代的协同作用，而非单一模块的简单叠加，也验证了层叠式量子辅助计算范式的结构合理性与设计鲁棒性[13]。

## 五、结论

本文针对经典计算体系下大模型训练面临的算力瓶颈、成本高企、长序列效率低下、能耗压力突出等现实问题，立足 NISQ 时代量子硬件约束，提出原创性的层叠式量子辅助计算范式，设计量子振幅注意力池化与量子随机电路嵌入两项专用技术，构建可直接部署、低耦合、高兼容的量子-经典混合训练框架，实现对 Transformer 核心高复杂度模块的精准量子加速。基于真实量子云平台与标准 GPU 集群的对比实验表明，所提方案在文本分类任务中能够保持模型精度基本不变，实现训练效率 18%—25%的稳定提升，收敛速度更快，显存与带宽占用更优，具备较强的工程实用性与可迁移性。研究同时系统梳理了量子硬件噪声、量子-经典交互开销、算法硬件适配、工具链缺失等现实挑战，并从混合优化器、量子联邦学习、容错量子算法、可解释性、标准化平台等角度提出未来研究方向，对金融、药物研发、自动驾驶、科学计算等产业场景的应用价值进行全面展望。

参考文献：

［1］ Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

［2］ Arute F, Arya K, Babbush R, et al. Quantum supremacy using a programmable superconducting processor[J]. Nature, 2019, 574(7779): 505-510.

［3］Farhi E, Goldstone J, Gutmann S. A quantum approximate optimization algorithm[J]. arXiv preprint arXiv:1411.4028, 2014.

［4］ Lloyd S, Mohseni M, Rebentrost P. Quantum algorithms for supervised and unsupervised machine learning[J]. arXiv preprint arXiv:1307.0411, 2013.

［5］ Benedetti M, Grant E, Wossnig L, et al. Parameterized quantum circuits as machine learning models[J]. Quantum Science and Technology, 2020, 5(4): 043001.

［6］张正，李华，王明．量子计算在人工智能中的应用研究进展[J]．计算机学报，2023，46(5)：1024-1042.

［7］本源量子．本源悟源量子计算机技术白皮书[R]．合肥：本源量子计算科技(合肥)股份有限公司，2022.

［8］ Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

［9］Preskill J. Quantum computing in the NISQ era and beyond[J]. Quantum, 2018, 2: 79.

［10］ Lloyd S. Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors[J]. Physical Review Letters, 1996, 78(2): 341.

［11］ Wiebe N, Braun D, Lloyd S. Quantum algorithm for supervised learning on a classical dataset[J]. Physical Review Letters, 2012, 109(13): 130503.

［12］ Havlíček V, Córcoles A D, Temme K, et al. Supervised learning with quantum-enhanced feature spaces[J]. Nature, 2019, 567(7747): 209-212.

［13］刘跃光，陈宇翱，潘建伟．超导量子计算与量子机器学习研究进展[J]．中国科学：物理学 力学 天文学，2022，52(1)：1-20.

# Quantum Machine Learning for Accelerating Large Model Training

**NIE Jialei***

*(Jiangsu Ocean University, Lianyungang, Jiangsu 222005, China)*

**Abstract:** Aiming at the practical bottlenecks in current large model training, such as high computing cost, low efficiency, exponential growth of long-sequence computing complexity, and increasing pressure of hardware energy consumption and carbon emissions, based on the hardware conditions and technical boundaries in the era of Noisy Intermediate-Scale Quantum (NISQ) computing, this paper proposes an original Cascaded Quantum-Assisted Computing Paradigm, and systematically explores the acceleration path and implementation mechanism of the deep integration of quantum machine learning technology and classical large model training process. By embedding the two core technologies independently proposed, Quantum Amplitude Attention Pooling and Quantum Random Circuit Embedding, into the core modules of Transformer architecture with the highest computing load and most prominent complexity, this study constructs a quantum-classical hybrid training framework with stable structure, strong compatibility and direct deployability on existing quantum cloud platforms. On the premise of not destroying the overall structure and training logic of classical models, it realizes dynamic offloading and collaborative computing of high-complexity subtasks to quantum processors. Experimental results show that on the current mainstream NISQ quantum computing devices, the proposed framework can achieve a stable improvement of 18%—25%

in overall training efficiency in public text classification standard tasks, while the model accuracy loss is controlled within 0.5 percentage points, with strong engineering application value. This paper further systematically analyzes the practical constraints such as quantum hardware noise, quantum-classical data interaction overhead, and algorithm-hardware adaptability. Combined with the future evolution trend of fault-tolerant quantum computing, it comprehensively prospects the long-term development path, technological breakthrough direction and industrial application scenarios of quantum machine learning in the field of large model training, providing new theoretical ideas and practical references for breaking through the computing power ceiling of large model training under the classical computing system.

**Keywords:** Quantum machine learning; Large model training; cascaded quantum-assisted computing paradigm; Quantum amplitude attention pooling; NISQ devices