基于神经网络的计算机翻译优化研究

李金珩

新加坡国立大学统计学院,新加坡 119077

摘要：本研究聚焦神经机器翻译（NMT）系统在边缘设备部署时的高能耗难题，从软件层面提出一种集成动态架构剪枝的优化方案。通过在 Transformer 编码器中引入熵驱动的动态退出机制，结合混合精度参数分组策略，有效降低了系统能耗。在 WMT2024 中英数据集上，单句推理能耗降至 3.1 焦耳（减少 34.8%），同时 BLEU 分数稳定在 40.5±0.3（$p < 0.05$），为低资源场景下的实时翻译提供了软件层面的可行解决方案。

关键词：神经机器翻译；动态架构修剪；软件优化；能耗降低；混合精度训练

1 研究背景与研究意义

1.1 研究背景

在全球化浪潮的推动下，跨语言交流的需求呈现出爆发式增长。据统计，全球现存超过 7000 种语言，这些语言在语法规则、词汇语义以及文化语境方面存在着巨大差异，成为跨语言交流的主要障碍。例如，在国际商务谈判中，语言障碍可能导致合同条款的歧义，进而引发商业纠纷；在学术研究领域，语言不通会造成前沿成果的传播滞后，阻碍学术交流与发展；在公共服务方面，语言沟通不畅可能造成信息传递失真，影响服务质量。

神经机器翻译（Neural Machine Translation, NMT）技术的出现为解决语言障碍带来了新的希望。基于 Transformer 架构的深度神经网络模型，能够有效捕捉跨语言语义关联，显著提升翻译质量。然而，在实际应用中，NMT 模型面临着严峻挑战。其庞大的参数规模和复杂的计算流程，使得在智能手机、物联网设备等资源受限的边缘场景部署时，产生高昂的计算能耗与存储成本。

研究表明，标准 Transformer 模型在处理简单语句时，仍需激活全部编码层，造成约 27%的无效能耗，这严重制约了其在低资源设备上的实时翻译能力。此外，传统的模型压缩方法，如知识蒸馏、剪枝等多采用静态策略，难以根据输入文本的复杂度差异进行动态调整，导致简单文本过度计算，而复杂文本翻译质量下降，进一步限制了 NMT 技术在跨境电商客服、多语言会议同传等跨语言即时交互场景中的应用。

1.2 研究意义

从应用层面来看，突破 NMT 模型在边缘设备上的能耗瓶颈具有重要的现实意义。

赋能边缘智能设备：为智能手机、可穿戴设备等提供低功耗实时翻译功能，满足国际旅行者即时沟通、跨境电商多语言客服等需求，显著提升设备用户体验。

支撑低资源场景：为资源受限的物联网设备，如智能家居终端、工业传感器等构建轻量化翻译解决方案，促进多语言环境下的设备协同与数据流通。

促进全球普惠交流：降低技术部署成本，推动翻译服务向小语种地区、发展中国家普及，有助于消除语言数字鸿沟，促进跨文化理解与合作。

从学术层面分析，本研究推动了多学科的交叉融合。

创新模型优化理论：通过深入剖析 NMT 模型的计算冗余机制，探索动态自适应的轻量化

方法，如动态层选择、语义复杂度感知压缩等，为神经翻译模型的能效优化提供全新的理论框架。

拓展跨学科研究范式：融合计算机科学中的算法优化、语言学中的语义表征理论，为复杂系统的能效-性能协同优化提供有效的方法论参考。

完善翻译技术体系：揭示文本复杂度与模型计算资源分配的内在关联，为翻译质量评估引入能耗效率指标，推动机器翻译从“质量优先”向“质量-效率双优”的范式升级。

2 研究现状

在计算机辅助翻译技术的发展历程中，早期的机器翻译主要依赖规则和词典来完成基本翻译任务。这种方法虽然简单直接，但准确性较低，难以处理自然语言中的细微差别和复杂性。随着技术的不断进步，统计机器翻译应运而生。布朗[1]等人在 1990 年提出了基于统计的方法，通过词对齐和语言模型，有效提高了翻译质量，使模型能够学习不同语言中词语之间的概率关系[2]。

近年来，深度学习技术的飞速发展彻底改变了机器翻译领域[3]。神经机器翻译（NMT）技术凭借深度神经网络学习源语言和目标语言之间的复杂映射关系[4]，展现出强大的翻译能力[5]。基于 Transformer 架构的 NMT 模型在处理长句和复杂句子时表现卓越，能够捕捉远距离的语义依赖关系，生成更加准确、流畅的翻译。例如，采用 NMT 技术的谷歌翻译在多语言对的翻译方面取得了显著进展，为行业树立了新的标杆[6]。

在 NMT 模型优化领域，研究人员进行了多方面的探索。模型压缩技术是其中的重点关注方向之一，知识蒸馏旨在将复杂教师模型的知识转移到简单学生模型中，实现模型轻量化，但它难以适应输入数据的多样性和复杂性。模型剪枝技术通过消除模型中的冗余连接或参数来减轻计算负担，但大多数现有剪枝方法为静态方式，无法根据输入实时调整模型结构。量化技术虽然可以减少内存占用和计算复杂度，但往往会导致翻译质量下降。

国内的计算机辅助翻译技术研究起步相对较晚，但在国家对人工智能和语言技术的大力支持下，发展迅速。许多科研机构和高校投入大量资源进行研究，中国学者致力于结合汉语的独特特征，如深入的语义理解和语法分析，来提高翻译准确性。然而，国内外在翻译准确性影响因素的研究方面仍存在一些局限。当前的机器翻译模型在理解和翻译比喻、双关、语义模糊等复杂语言现象时，表现仍不尽如人意。在多模态信息融合领域，虽然已经尝试将文本与图像、声音等信息结合，但如何有效整合以提高翻译准确性，仍处于探索阶段。此外，现有模型在通用性和专业性之间难以达到完美平衡，特定领域的翻译准确性还有很大的提升空间。

3 研究内容与创新

本研究聚焦于软件层面，构建了一个针对 NMT 系统高能耗问题的优化框架。在算法层面，设计了基于隐藏状态熵的动态退出机制，同时采用混合精度参数分组策略，在保证翻译质量的前提下，降低内存占用和计算负担。

本研究的主要创新点如下：

提出熵驱动的动态退出机制：该机制实现了 Transformer 编码器计算层数量的动态调整。它依据输入数据的复杂度自适应地分配计算资源，避免了不必要的计算，提高了计算效率。

采用混合精度参数分组策略：对注意力矩阵和梯度更新采用不同的精度进行存储，减少了内存占用，同时防止了数值下溢，确保翻译质量不受影响。

4 方法改进

4.1 动态计算路径优化

4.1.1 熵门控机制

为实现 Transformer 编码器中计算层数量的动态调整，本研究引入了熵门控机制。该机制基于信息熵原理，通过计算编码器每一层隐藏状态的熵来衡量输入文本的复杂度。信息熵是信息论中的一个重要概念，它可以反映信息的不确定性。在本研究中，较大的熵值表示文本的不确定性较高，需要更多的计算资源进行处理；相反，较小的熵值则表明文本较为简单，可能不需要所有编码层来进行计算。 具体而言，编码器第 t 层的退出决策函数定义为：

$$\Gamma_t = \sigma\left(W_g \cdot \text{Entropy}(\hbar_t) + b_g\right)$$

σ是 Sigmoid 函数，用于将输出值映射到区间[0,1]； W_g 和 b_g 是可学习参数；$\text{Entropy}(\hbar_t)$表示第 t 层编码器隐藏状态$\hbar_t$的熵。当满足条件 $\Gamma_t > 0.65$ 时，模型终止后续层的计算，并直接输出当前层的结果。这样，原始的 12 层编码器在处理简单文本时可以压缩到平均 5.2 层，有效减少了计算负担。

4.1.2 混合精度参数分组

在模型训练和推理过程中，内存占用和数值精度是影响计算效率和翻译质量的重要因素。为了在这两者之间找到平衡，本研究提出了混合精度参数分组策略。 对于注意力矩阵，使用半精度浮点数（FP16）进行存储。FP16 的内存占用仅为单精度浮点数(FP32)的一半，这使得注意力矩阵的内存占用减少了 42%。这不仅减少了内存访问开销，还提高了计算效率。然而，FP16 的精度相对较低，在梯度更新过程中可能会导致数值下溢，影响模型的训练效果。因此，在更新梯度时，保留 FP32 的精度以确保梯度信息的准确性并防止数值下溢问题的发生。

5 实验检查

5.1 数据集和基线

本研究使用官方的 WMT2024 中英测试集进行实验。该测试集包含 5,000 个句子对，涵盖了多种类型的文本，能够全面评估模型的翻译性能和能耗。选择标准 Transformer（FP32）作为基线模型，该模型是 NMT 领域常用且具有代表性的模型。

5.2 关键指标比较

实验主要比较不同方案的翻译质量（BLEU 分数）、能耗（每句焦耳数）和能效比（GFLOPs/焦耳）。BLEU 分数用于衡量翻译结果与参考翻译之间的相似度，值越高，翻译质量越好。能耗指标反映了模型在单句推理过程中的能耗情况。能效比综合考虑了计算负载和能耗，用于评估系统的能源利用效率。实验结果如下表所示：

方案	BLEU	能耗 (J / 句)	能效比 (GFLOPs/J)
基线模型	41.2	4.8	9.1
静态层剪枝	39.8▼	3.9▼	11.2▲
本研究方案	40.5	3.1	13.6

注：▼表示降低，▲表示增加；显著性检验 $p < 0.01$。

从结果可以看出，本研究提出的方案相比基线模型减少了 34.8%的能耗，能效比提高到 13.6 GFLOPs/J。同时，BLEU 分数保持在 40.5，与基线模型相比仅有微小差距，差异具有统计学意义（$p < 0.05$）。与周等人（2024）的静态层剪枝方案相比，本方案在翻译质量和能效比方面均具有明显优势。

5.3 消融实验分析

为了深入分析每个优化组件的贡献，进行了消融实验。通过将其与固定 6 层架构进行比较，评估了动态退出机制的贡献。实验结果表明，动态退出机制的贡献达到了 58%。这意味着动态退出机制在降低能耗方面发挥了重要作用，它可以根据输入文本的复杂度有效减少计算层的数量，避免冗余计算。 通过比较不同精度设置下的峰值内存占用，分析了混合精度训练对内存占用的影响。结果显示，混合精度训练将峰值内存占用降低至 1.6GB，减少了 39%。这表明混合精度参数分组策略在减少内存占用方面是有效的，并为模型在资源受限设备上的运行提供了更有利的条件。

6 结论

这项研究验证了仅从软件层面实施动态计算分配策略，也能够有效消除神经机器翻译中的冗余计算。通过利用熵门控机制，在算法层面实现了智能剪枝。结合混合精度参数分组策略，实现了显著的 34.8%能耗降低，同时保持了可控的质量损失。这一创新方案为在低功耗边缘计算设备上部署 NMT 系统提供了一种新颖的软件技术途径，在促进 NMT 技术在资源受限场景中的广泛应用方面展现出巨大潜力。

对于未来的研究，进一步探索更精确的动态退出机制具有巨大的潜力。这可能涉及微调基于熵的决策过程，以更好地适应不同输入文本的复杂性。此外，优化混合精度策略至关重要。通过在训练和推理过程中仔细调整模型不同部分的精度设置，可以实现内存使用、计算效率和翻译质量之间的更好平衡。本研究提出的方法也有潜力扩展到更广泛的语种对和模型架构。这种扩展有助于验证其在不同语言和计算环境中的普遍性和有效性。

参考文献

[1]Brown, P. F., Cocke, J., Della Pietra, et al. A statistical approach to machine translation[J]. Computational Linguistics,1990,16(2):79-85.

[2]Duan, R., & Duan, X. Neural machine translation based on adaptive training and discarding mechanism[J]. Computer Engineering,2023,49(10):24.

[3]张笑燕，逄磊，杜晓峰，等。基于单语优先级采样自训练神经机器翻译的研究[J].通信学报，2024,45(4):65-72.

[4]黄勃,李文超,刘进,等.探索 ChatGPT 模型能力:工业应用的前景和挑战[J].武汉大学学报(理学版),2024,70(03):267-280.

[5]黄孟钦.基于双语词典的远距离语对无监督神经机器翻译方法[J].现代电子技术,2024,47(07):161-164.

[6]刘欢,刘俊鹏,黄锴宇,等.面向低资源俄汉机器翻译的领域适应方法[J].厦门大学学报(自然科学版),2022,61(04):654-659.