摘要

视觉 Transformer 凭借其全局建模与特征表达能力，已迅速成为计算机视觉领域的主流范式。然而，其对视觉信号固有的稠密处理方式导致了明显的效率瓶颈。该瓶颈主要由稠密表征引起，进而制约了 Transformer 模型在视觉识别与生成两大重要场景中的高效部署。首先，在视觉表征层面，现有方法通常将视觉信号机械地映射为稠密特征序列，不可避免地产生大量冗余特征，造成庞大的计算开销。其次，在视觉识别任务的训练阶段，稠密的监督目标容易引入背景噪声，稀释有效梯度，阻碍了模型对高层语义的高效学习。第三，对于视觉生成任务，传统的自回归稠密解码机制串行生成视觉词符，造成了严重的推理延迟与吞吐瓶颈。因此，本文从模型表征、视觉识别训练、视觉生成推理三个层面研究了视觉 Transformer 稀疏优化方法，并进一步构建多模态识别与生成系统，旨在提升视觉 Transformer 在视觉识别和生成任务中的部署效率，并验证稀疏优化方法在统一交互场景中的应用可行性。

在模型表征层面，针对稠密特征表示引发特征冗余问题，本文提出了一种基于特征融合的稀疏视觉表征方法。该方法引入可学习的元词符作为语义锚点，自适应地挖掘并定位视觉内容中的核心语义线索，将原始稠密特征序列动态聚合为多组语义相关的特征集合。在聚合过程中，高度相关的冗余特征可以被压缩为更紧凑的稀疏表示；同时，通过可学习的门控机制，模型可以根据任务目标调控不同网络层的特征融合比例与稀疏粒度。此外，本文面向视频任务对视频信号中的空间结构与时序动态信息进行解耦，针对性地剔除了跨帧的冗余背景并保留运动轨迹信息，从而实现了时空维度上的联合稀疏化。该方法在保留关键空间结构与动态信息的同时，通过表征稀疏优化缓解了计算压力。实验结果显示，该方法能够在分类性能基本保持的前提下，减少高达 95% 的特征词符数量，实现 60% 以上的计算加速。

在视觉识别训练层面，针对稠密监督信号阻碍高层语义学习问题，本文提出了一种基于可进化掩码的稀疏训练方法。现有自监督预训练范式多采用稠密重建目标，易导致优化过程陷入对底层纹理细节的过度拟合，损害识别训练效率与模型的泛化能力。对此，本文构建了一种基于可进化掩码的训练机制，实现了训练难度与模型能力的协同演进。不同于传统方法在稠密特征上进行无差别的拟合，该机制利用模型当前的特征解析能力，将稠密视觉信号解析为稀疏的语义集合，并据此生成稀疏的训练掩码作为训练目标。随着训练周期的推进，掩码策略遵循由易到难的演进规律，从初期掩盖低层纹理细节逐步过渡到后期掩盖更复杂的高层语义实体。这种动态演进策略减少了背景噪声对训练梯度的稀释，并促使网络从局部像素重建转向全局语义关系推断。实验结果显示，该策略在分割、检测等识别下游任务中带来一定泛化收益，并在保证性

能的前提下，将基线模型的训练效率提升了 50% 以上。

在视觉生成推理层面，针对稠密串行解码造成速度瓶颈和质量退化问题，本文提出了一种基于分集预测与反馈校正的稀疏推理方法。首先，为突破逐词符串行预测的范式限制，本研究设计了一种下一集合预测策略。该策略将原始稠密序列稀疏化为多个语义相关的紧凑子集，支持子集内的词符进行并行预测，从而提升计算吞吐量。同时，本文引入连续域解码机制，替代传统自回归生成所使用的离散编码方案，缓解量化误差，并利用连续域表示改善视觉生成质量。此外，为进一步应对并行生成可能引发的质量退化与误差累积问题，本文提出了一种基于质量评估反馈的局部校正机制。模型能够自动评估生成内容，并仅对低质量的局部区域进行迭代优化，避免重复冗余计算。实验结果显示，该方法将传统自回归方法的推理速度提升了 5 倍以上，并使 FID、FVD 等关键生成质量指标达到或优于部分前沿自回归生成方案，在视觉生成任务中取得了较好的速度-质量平衡。

本文进一步构建了面向图像理解、文本问答和图像生成的多模态系统。该系统将稀疏化识别训练方法应用到视觉编码器中，将稀疏化生成推理方法应用到自回归模型中，从而在统一框架下观察本文方法对理解能力和生成效率的影响。实验结果表明，本文方法在视觉识别问答任务中对模型的局部细粒度理解能力带来一定改善，并在图像生成任务中降低了端到端生成延迟。

综上，本文构建了视觉 Transformer 高效计算框架，改善了视觉识别和视觉生成模型的计算效率和泛化能力，并通过多模态识别与生成系统初步验证了其实际应用潜力，为资源受限环境下的视觉大模型部署提供了技术支撑。

关键词：Transformer 模型，稀疏化方法，计算效率，视觉感知，视觉生成

Efficient Sparse Optimization Methods for Visual Transformers

Zhanzhou Feng (Computer Application Technology)
Directed by Prof. Shiliang Zhang

ABSTRACT

Vision Transformers, with their exceptional global modeling capabilities and feature expression potential, have rapidly become the mainstream paradigm in computer vision. However, their inherently dense processing of visual signals leads to severe efficiency bottlenecks. This bottleneck, originating from dense representation, significantly restricts the efficient deployment of Transformer models in two crucial application scenarios: visual recognition and visual generation. First, at the visual representation level, existing methods typically map visual signals mechanically into dense feature sequences, inevitably generating a massive amount of redundant features and causing substantial computational overhead. Second, during the training phase of visual recognition tasks, dense supervision targets tend to introduce background noise and dilute effective gradients, hindering the model's efficient learning of high-level semantics. Third, for visual generation tasks, the traditional autoregressive dense decoding mechanism generates visual tokens serially, resulting in severe inference latency and throughput bottlenecks. Therefore, this paper investigates sparse optimization methods for Vision Transformers from three perspectives: model representation, visual recognition training, and visual generation inference, and further builds a multimodal recognition and generation system to validate the proposed methods in a unified interactive scenario.

To address the problem of feature redundancy caused by dense feature representation in visual representation, this dissertation proposes a sparse representation method based on feature fusion. This method introduces learnable meta-tokens as semantic anchors to adaptively mine and locate core semantic clues in visual content, dynamically aggregating the original dense sequence into multiple sets of semantically related features. During the aggregation process, highly correlated redundant backgrounds can be compressed into a more compact sparse representation. Meanwhile, through a learnable gating mechanism, the model adaptively regulates the feature fusion ratio and sparsity granularity across different network layers according to task objectives. Furthermore, this work not only achieves feature sparse optimization for im-

ages in the spatial dimension but also extends to video tasks by decoupling the spatial structure and temporal dynamic information of video signals. It specifically eliminates cross-frame redundant backgrounds while preserving motion trajectory information, thereby achieving joint sparsification in the spatial-temporal dimension. While preserving key spatial structures and dynamic information, this method effectively alleviates the quadratic computational pressure brought by the self-attention mechanism through representation sparse optimization, achieving a dynamic match between feature representation and computational capacity. Experiments demonstrate that this method can reduce the number of feature tokens by up to 95% and achieve over 60% computational acceleration, all while maintaining nearly lossless classification performance.

To address the issues of dense supervisory signals hindering high-level semantic learning in visual recognition training, this dissertation proposes a sparse training method based on evolvable masks. Existing self-supervised pretraining works mostly adopt fixed dense reconstruction targets, easily causing the optimization process to fall into the overfitting bottleneck of low-level texture details and harming generalization. In response, this dissertation constructs a training mechanism based on evolvable masks, achieving a synergistic evolution between training difficulty and model capacity. Unlike traditional methods that perform indiscriminate fitting on dense features, this mechanism utilizes the model's current feature parsing capability to dynamically parse the input signal into sparse semantic sets, and subsequently generates sparse training masks as the training target. As the training epochs progress, the masking strategy follows an easy-to-hard evolutionary rule, gradually transitioning from masking low-level texture details in the early stages to reconstructing sparse high-level complex semantics in the later stages. This dynamic evolution strategy not only avoids the dilution of training gradients by background noise but also forces the network to accomplish an capability evolution from local pixel reconstruction to global semantic relationship inference, significantly accelerating pre-training convergence. Experimental results show that this strategy not only enhances the model's generalization ability and robustness in downstream recognition tasks such as segmentation and detection but also improves the training efficiency of baseline models by over 50% while ensuring performance.

To address the issues of speed bottlenecks and quality degradation caused by dense serial decoding in autoregressive visual generation, this dissertation proposes a sparse inference method based on set prediction and feedback calibration. First, to break through the paradigm limitation of token-by-token serial prediction, this research designs a next-set prediction strategy. This strategy sparsifies the original dense sequence into multiple semantically related

compact subsets, supporting the synchronous parallel prediction of tokens within the subsets, thereby multiplying the computational throughput. Meanwhile, this dissertation introduces a continuous-domain decoding mechanism to replace the discrete encoding scheme used in traditional autoregressive generation, eliminating quantization errors and fundamentally improving visual generation quality from the feature space. Furthermore, to overcome the issues of quality degradation and error accumulation that parallel generation may induce, this dissertation proposes a local calibration mechanism based on quality evaluation feedback. The model can autonomously evaluate the generated content and perform iterative optimization only on low-quality local regions, avoiding repetitive and redundant computations on global dense features. Experiments prove that this mechanism not only accelerates the inference speed of traditional autoregressive methods by more than 5 times but also ensures that key generation quality metrics such as FID and FVD outperform frontier autoregressive generation schemes, achieving a dual guarantee of quality and efficiency in visual generation tasks.

Based on the above methods, this dissertation further builds an integrated multimodal system for image understanding, text-based interaction, and image generation. The system deploys the sparse recognition training method on the visual encoder and the sparse generation inference method on the autoregressive generation model, thereby validating the proposed sparse optimization methods in a unified multimodal framework. Experimental results show that the system improves fine-grained visual understanding in visual question answering and substantially reduces end-to-end latency in image generation.

Overall, this dissertation constructs an efficient computational framework for Vision Transformers. It improves the computational efficiency and generalization capabilities of visual recognition and visual generation models, and validates their practical potential through a multimodal recognition and generation system. This provides technical support for the deployment of large vision models in resource-constrained environments.

KEY WORDS: Transformer, Sparsification, Efficiency, Visual perception, Visual generation