Sparse Low-Rank Approximation for Efficient Attention Mechanisms

GTP-4o prompted by Forrest T.

June 2024

Abstract

This paper proposes a novel approach to optimize the computational efficiency of attention mechanisms in Transformer models by combining low-rank approximations with sparsity constraints. By leveraging these techniques, we aim to reduce memory usage and computational complexity while maintaining model performance.

Introduction

Attention mechanisms are critical components of Transformer models, enabling them to capture dependencies across input sequences. However, their computational cost, particularly the quadratic complexity of the dot-product attention, poses challenges for large-scale applications. We propose a Sparse Low-Rank Approximation method to address this issue.

Methodology

Low-Rank Approximation

We begin by approximating the matrices Q, K, and V using Singular Value Decomposition (SVD). For a matrix A, its low-rank approximation $\hat{A}$ is:

$$A \approx \hat{A} = U_r \Sigma_r V_r^T$$

where U_r and V_r are matrices containing the first r singular vectors, and Σ_r is a diagonal matrix containing the first r singular values.

Applying SVD to Q, K, and V:

$$Q \approx \hat{Q} = U_Q \Sigma_Q V_Q^T$$

$$K \approx \hat{K} = U_K \Sigma_K V_K^T$$

$$V \approx \hat{V} = U_V \Sigma_V V_V^T$$

Sparsity Constraint

To enforce sparsity, we apply L1 regularization to the low-rank matrices:

$$\|\hat{Q}\|_1 = \sum_{i,j} |\hat{Q}_{ij}|$$

$$\|\hat{K}\|_1 = \sum_{i,j} |\hat{K}_{ij}|$$

$$\|\hat{V}\|_1 = \sum_{i,j} |\hat{V}_{ij}|$$

Combined Optimization

The objective function combines task-specific loss with sparsity constraints:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_Q \|\hat{Q}\|_1 + \lambda_K \|\hat{K}\|_1 + \lambda_V \|\hat{V}\|_1$$

where $\mathcal{L}_{task}$ is the original loss function, and λ_Q, λ_K, and λ_V are regularization parameters.

Attention Mechanism

The attention mechanism is computed using the approximated matrices:

$$Attention(Q, K, V) \approx softmax\left(\frac{\hat{Q}\hat{K}^T}{\sqrt{d_k}}\right)\hat{V}$$

Conclusion

By integrating low-rank approximations with sparsity constraints, this method aims to significantly reduce the computational cost of attention mechanisms while preserving model accuracy. This approach holds promise for efficient large-scale applications of Transformer models.

Future Work

Further research will involve empirical validation of this approach on benchmark datasets and exploration of dynamic adaptations based on input data characteristics.