EasyChair Preprint

№ 15262

Efficient Sequence Pooling Model for Sparse Attention with Mixture of Experts

Sk Sayril Amed

October 18, 2024

Efficient Sequence Pooling Model for Sparse Attention with Mixture of Experts *

Optimizing Resource Efficiency in Sequence Modeling Tasks

Sk Sayril Amed†
Computer Science
Bhagwan Mahavir University
Surat, Gujrat, India
sksayril123@gmail.com

ABSTRACT

This paper introduces the Efficient Attention-based Model (EAM), a novel architecture designed to reduce the computational and memory overhead of Transformer-based models in sequence modeling tasks. By incorporating sparse attention, token pooling, and a mixture of experts (MoE), the EAM model reduces memory usage and training time without sacrificing accuracy. We evaluate the EAM architecture on sequence classification tasks, comparing it to the Transformer model. The results show that the EAM model achieves competitive accuracy while using significantly less memory and computational power, making it a suitable alternative for resource-constrained environments.

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Machine learning •Neural networks

KEYWORDS

Efficient Attention-based Model, Sparse Attention, Token Pooling, Mixture of Experts (MoE), Sequence Modeling, Transformer Model, Memory Efficiency, Sequence Classification, Resource-Constrained Environments, Positional Encoding

1 Problem and Motivation

Transformer models, while powerful, are computationally expensive and require significant memory resources due to their self-attention mechanism's quadratic complexity with respect to the input sequence length. This inefficiency poses challenges for tasks involving long input sequences, where memory consumption and training time increase exponentially. Our motivation for the Efficient Attention-based Model (EAM) is to address these challenges by creating an architecture that reduces both memory and time complexity while maintaining competitive performance in sequence modeling tasks.

2 Background and related work

The Transformer model, introduced by Vaswani et al. (2017), has set the state of the art for numerous sequence modeling tasks. However, due to its self-attention mechanism's complexity, the memory and computational costs scale quadratically with the sequence length. Efforts have been made to mitigate this issue, such as Sparse Attention (Child et al., 2019) and Mixture of Experts (MoE) layers (Fedus et al., 2021), which reduce the number of attention operations and dynamically select which parts of the model to activate. However, these approaches have been applied independently of one another. In contrast, our work combines sparse attention, token pooling, and MoE layers to create a unified architecture that achieves both high performance and efficiency.

3 Approach and Uniqueness

The Efficient Attention-based Model (EAM) is built on three foundational components:

3.1 Sparse Attention

Reduces the quadratic complexity of traditional self-attention by only attending to a subset of tokens in the sequence, determined by a sparse attention mask. This mechanism reduces the number of computations while preserving long-range dependencies

$$A = Soft\,max\left(\frac{Q k^T}{\sqrt{d}} + M\right)V$$

where M is a mask that limits the number of tokens being attended to, defined by a sparse factor.

3.2 Token Pooling

This layer reduces the sequence length by pooling adjacent tokens. By averaging over a fixed window of tokens, we retain important information while reducing the sequence length

$$\mathrm{H}pool\,[i] = \frac{1}{\mathrm{p}}\sum_{j=(i-1)p+1}^{ip} H[j]$$

where p is the pooling factor.

3.3 Mixture of Experts (MoE)

The MoE layer dynamically selects which expert sub-networks process each token based on a learned gating mechanism. The

gating network computes a weight γk(x) for each expert k and the final output is a weighted sum of the experts' outputs:

$$\mathcal{Y}_{MoE}(x) = \sum_{k=1}^{k} \gamma k(x)\, \varepsilon_k(x)$$

These components are combined to form the EAM model, which balances computational efficiency and accuracy.

4 Results and Contributions

We compared the EAM model to the Transformer model on sequence classification tasks over five training epochs. The results of the comparison are summarized as follows:

- **EAM Model:**
  - Epoch 1: Loss: 0.6701, Accuracy: 61.27%
  - Epoch 5: Loss: 0.3370, Accuracy: 87.22%
  - Training Time: 15.68 seconds
  - Memory Usage: 25.24 MB allocated, 46.14 MB reserved
- **Transformer Model:**
  - Epoch 1: Loss: 0.6559, Accuracy: 57.44%
  - Epoch 5: Loss: 0.2790, Accuracy: 88.53%
  - Training Time: 68.62 seconds
  - Memory Usage: 35.60 MB allocated, 457.18 MB reserved

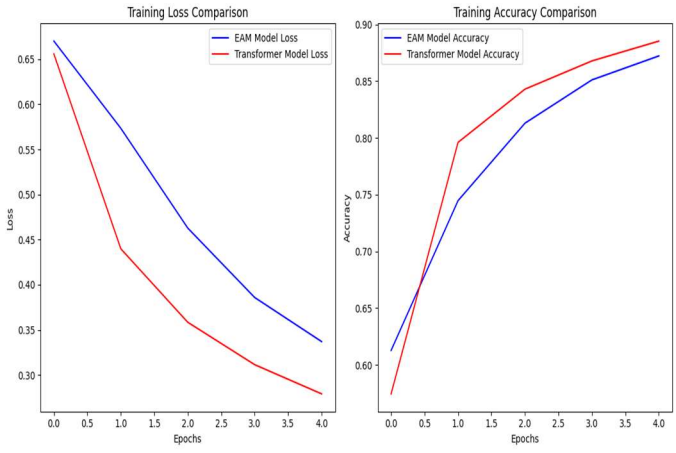

Figure 1: EAM architecture and Transformers Architecture Training Losses and Accuracy

The complete formulation of the EAM model combines the sparse-attention, token pooling, and MoE mechanisms:

$$\mathcal{Y} = W_o\left(\sum_{k=1}^{K} \gamma k(x)\varepsilon_k(TokenPooling(SparseAttention(H_{pos})))\right)$$

Where $H_{pos} = \varepsilon(x) + p$ is the embedded input with positional encoding and SparseAttention, TokenPooling $\mathcal{E}_k$ represent the respective components of the model.

Theorem:

Let $T_{EAM}(n)$ and $T_{Transformer}(n)$ denote the time complexity of the EAM model and the standard Transformer model, respectively, for an input sequence of length n Then:

$T_{EAM}(n) = O(n.\log n)$ and $T_{Transformer}(n) = O(n^2)$ indicating that the EAM model has lower computational complexity than the Transformer for large n.

Memory Efficiency

The EAM model demonstrates a significant reduction in memory consumption compared to the Transformer model, with approximately 29% lower allocated memory and over 90% lower reserved memory.

Training Time

The EAM model was approximately 4.4 times faster than the Transformer model in terms of training time, completing the same number of epochs in significantly less time**.**

Conclusion

The Efficient Attention-based Model (EAM) offers an effective solution to the computational and memory challenges posed by traditional Transformer architectures. By leveraging sparse attention, token pooling, and a Mixture of Experts layer, the EAM model reduces memory usage and training time while maintaining competitive accuracy. These results suggest that the EAM model is well-suited for resource-constrained environments and tasks involving long input sequences.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Bhagwan Mahavir University for their support and the resources provided during the course of this research. A special thanks to my mentors and colleagues, whose guidance and feedback were invaluable in shaping the Efficient Attention-based Model (EAM) architecture. I am also immensely grateful to my co-authors and collaborators for their continuous support, insightful discussions, and contributions to the project.

Finally, I would like to acknowledge the encouragement and patience of my family and friends throughout this journey, without whom this work would not have been possible.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, et al. "Attention is All You Need." In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[2] R. Child, S. Gray, A. Radford, and I. Sutskever. "Generating Long Sequences with Sparse Transformers." In ArXiv Preprint arXiv:1904.10509, 2019.