DYNAMIC MULTI-RESOLUTION TRANSFORMERS: EFFICIENT LONG TEXT PROCESSING WITH ADAPTIVE GRANULARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce a dynamic multi-resolution transformer designed to enhance the processing of long text sequences by adapting the granularity of attention based on text length and complexity. Traditional transformers struggle with the quadratic complexity of self-attention when dealing with long texts, making it difficult to balance detailed local understanding and global context. Our approach incorporates multiple resolution layers that dynamically adjust during processing, ensuring both fine-grained and coarse-grained text representations. We validate our method through extensive experiments on benchmarks for document summarization, long text comprehension, and long-range dependencies, demonstrating significant improvements over baseline models in efficiency and performance.

1 INTRODUCTION

Processing long text sequences effectively is a critical challenge in natural language processing (NLP), impacting applications such as document summarization, long text comprehension, and reasoning. Traditional transformer models struggle with long texts due to their quadratic complexity related to sequence length, necessitating the exploration of more efficient solutions.

Handling long texts is challenging as it requires capturing both fine-grained details and overarching themes. The complexity and length of texts demand models to dynamically adjust granularity levels, ensuring detailed local and comprehensive global representations. Existing models often fail to balance these aspects, resulting in suboptimal performance in tasks needing both detailed and holistic comprehension.

We introduce a multi-resolution processing mechanism within the transformer architecture that operates at varying levels of granularity. This involves adding multiple resolution layers that dynamically adjust based on text complexity and length, combining fine-grained and coarse-grained information. Our approach enhances long text retrieval, understanding, and reasoning by ensuring both detailed and holistic representations.

To evaluate our approach, we conduct extensive experiments using well-established benchmarks for document summarization, long text comprehension, and long-range dependency tasks. Our results show significant improvements over baseline transformer models, highlighting the benefits of adaptive multi-granularity mechanisms.

Our contributions are:

- Proposing a multi-resolution processing mechanism for transformer models operating at various granularity levels.
- Dynamically adjusting the number of resolution layers based on text complexity and length.
- Demonstrating the selective combination of fine-grained and coarse-grained information to enhance long text retrieval, understanding, and reasoning.
- Validating our approach through extensive experiments, showing significant improvements over baseline models in tasks like document summarization, long text comprehension, and long-range dependency resolution.

Future work includes optimizing the multi-resolution approach further and extending its application to other complex NLP tasks.

2 Related Work

The goal of this section is to place our dynamic multi-resolution transformer within the context of existing literature, specifically focusing on traditional transformers, long text processing strategies, and hierarchical/multi-resolution approaches. We aim to highlight the unique aspects and advantages of our method.

Traditional transformer architectures have set a high standard in NLP performance. However, their quadratic complexity with respect to sequence length poses significant challenges when processing long texts. This complexity arises from the self-attention mechanism, which necessitates pairwise comparisons between all tokens. Our multi-resolution approach mitigates this by introducing resolution layers that adjust their granularity dynamically based on text length and complexity, thus reducing computational overhead while maintaining high performance.

Several novel strategies have emerged to tackle the issue of long text processing. The Longformer (**?**) and Big Bird (**?**) introduce attention mechanisms that span fixed-size windows and sparse patterns to handle longer sequences more effectively. Similarly, the Reformer model by **?** employs locality-sensitive hashing (LSH) to bring down the self-attention complexity. While these methods provide significant improvements over traditional transformers by addressing quadratic complexity, they generally employ static strategies that do not adapt to the varying complexities and lengths of the input text. Our multi-resolution transformer, on the other hand, dynamically adjusts the attention granularity in real-time, allowing for a more nuanced balance between local and global context processing.

Hierarchical attention and multi-resolution approaches in NLP and computer vision have also been explored for their efficiency and effectiveness. **?** achieved improvements in document classification by employing hierarchical attention networks, which process information at different levels of abstraction. These methods highlight the potential benefits of multi-resolution techniques; however, they are often not integrated within transformer architectures or do not offer dynamic adaptability. Our approach builds upon these insights by integrating hierarchical, adaptive resolution layers within the transformer model, enhancing its capability to process long texts more efficiently.

In summary, while traditional transformers and other innovative models like Longformer, Big Bird, and Reformer offer valuable strategies for long text processing, they lack the dynamic adaptability intrinsic to our model. By introducing a multi-resolution mechanism that adjusts layer granularity in response to text length and complexity, our method offers a more flexible, efficient, and effective solution for long text processing in NLP.

3 Background

The evolution of transformer models has marked a significant milestone in natural language processing (NLP) by achieving superior performance across various tasks. Transformers utilize self-attention mechanisms enabling more effective context understanding compared to their predecessors, such as recurrent neural networks (RNNs). However, despite their success, transformers encounter efficiency issues when dealing with long sequences due to their quadratic complexity in relation to sequence length.

Processing long texts remains challenging because the self-attention mechanism in traditional transformers requires computations that grow quadratically with the sequence length. This leads to substantial computational and memory overhead, making them inefficient for handling long texts. Therefore, there is a critical need to explore alternative methods that can alleviate these constraints while retaining model performance.

Multi-resolution techniques have been explored in various domains such as computer vision and NLP. These techniques aim to balance the granularity and holistic understanding of data by processing it at different levels of detail. They have demonstrated potential in efficiently capturing fine-grained details and broader context.

3.1 Problem Setting

Our primary goal is to enhance the transformer architecture for effective long text processing by incorporating adaptive multi-resolution mechanisms. We define a long text T as a sequence of tokens $(t_1, t_2, \dots, t_n)$, where n is large. The key challenge is to develop a model that can dynamically adjust its resolution to capture both detailed and global information efficiently.

We rely on the assumption that the complexity of the text is quantifiable, enabling the dynamic adjustment of resolution layers. While this assumption is crucial for designing adaptive mechanisms, it may introduce limitations, particularly when dealing with highly unstructured text inputs.

4 Method

In this section, we present our proposed multi-resolution transformer architecture, designed to address the challenges of processing long text sequences. We aim to achieve this by dynamically adjusting attention granularity, ensuring efficient and comprehensive text representation. This method builds on the problem formalism introduced earlier and aligns with the advancements discussed in the background.

Our method addresses the inherent limitations of traditional transformers when dealing with long texts due to their quadratic complexity. By incorporating adaptive multi-resolution mechanisms, our model balances fine-grained detail with the capture of global context.

4.1 Architecture Overview

Our multi-resolution transformer comprises multiple layers, each designed to process text at varying granularities. Low-resolution layers handle broader chunks of text for a general understanding, while high-resolution layers focus on smaller segments for detailed comprehension. This hierarchical setup facilitates a detailed and holistic understanding of the content.

4.2 Dynamic Adjustment Mechanism

The core of our approach is the dynamic adjustment mechanism that adapts the number of resolution layers in real-time based on text complexity. This mechanism evaluates the complexity of the text sequence T and activates more layers for intricate and lengthy texts, while conserving resources for simpler texts.

Formally, consider a text sequence $T = (t_1, t_2, \dots, t_n)$, where n is the length. The processing employs a set of resolution layers $R = \{r_1, r_2, \dots, r_k\}$ with different granularities. The text complexity $C(T)$ guides our mechanism to determine the optimal subset of layers $R^* \subseteq R$:

$$R^* = \arg\min_{R'} \sum_{r \in R'} \text{Cost}(r, T) \quad \text{subject to} \quad \text{Quality}(R', T) \geq \text{Threshold}$$

where Cost(r, T) represents the computational cost of layer r for text T, and Quality(R', T) denotes the representation quality with layers R'.

4.3 Training and Optimization

The training process focuses on optimizing the trade-off between resolution layer count and text representation quality. We introduce a loss function that integrates representation accuracy and computational efficiency, employing gradient descent for parameter optimization.

4.4 Implementation

Our model extends current transformer frameworks by incorporating additional multi-resolution layers. Techniques such as layer normalization and dropout are utilized to enhance stability and prevent overfitting. The dynamic adjustment mechanism is integrated into the forward pass, ensuring adaptability during inference.

In summary, the multi-resolution transformer we propose offers a flexible and efficient solution for long text processing. By dynamically adjusting the resolution layers, our model optimizes for both detailed and comprehensive text representations, improving performance across various NLP tasks, including document summarization, long text comprehension, and long-range dependency resolution.