Extending Context Length in Attention Mechanisms with Fourier Features

Anonymous authors
Paper under double-blind review

Abstract

This paper explores the use of Fourier features to extend the context length in attention mechanisms within deep learning models. By leveraging the properties of Fourier transforms, we provide a framework for understanding and implementing context length extension through linear interpolation and change of base techniques. This approach offers insights into model behavior and guides the design of more effective architectures.

1 Introduction

Fourier features have shown promise in improving model performance, particularly in the context of extending the effective context length in attention mechanisms. This paper explores how Fourier features can be utilized to achieve context length extension, revealing a perspective that enhances our understanding of modern deep learning models.

2 Theoretical Foundations

2.1 Fourier Features for Context Length Extension

Fourier features provide a powerful tool for extending the effective context length in attention mechanisms. By leveraging the properties of Fourier transforms, we can perform linear interpolation and change of base (nonlinear) to adaptively adjust the context length. This section explores how these techniques can be applied to enhance attention mechanisms.

2.2 Context Length Extension with Fourier Features

Fourier features provide a powerful tool for extending the effective context length in attention mechanisms. By leveraging the properties of Fourier transforms, we can perform linear interpolation and change of base (nonlinear) to adaptively adjust the context length.

2.3 Unifying Perspective

The connections between position embeddings and Fourier features form a unifying perspective that deepens our understanding of these concepts. We will now provide formal proofs and expanded explanations for these connections.

These techniques provide several insights:

1. Fourier features enable models to capture a broader range of frequencies, which is crucial for extending context length in attention mechanisms.
2. Linear interpolation using Fourier features allows for smooth transitions between different context lengths, enhancing model flexibility.
3. Nonlinear change of base techniques can be employed to dynamically adjust context length based on input characteristics, improving model adaptability.
4. This perspective suggests that attention mechanisms can be optimized by considering the spectral properties of the data and the desired context length for a given task.

Furthermore, this approach opens up new avenues for research:

1. Developing new attention mechanisms that leverage Fourier features for context length extension.
2. Investigating the relationship between Fourier features and the inductive biases introduced by different attention designs.
3. Exploring how the interplay between Fourier features and attention mechanisms affects model generalization and extrapolation capabilities.

3 Practical Implications

Understanding the use of Fourier features for context length extension has several practical implications:

1. Architecture design: Fourier features can guide the design of more effective attention mechanisms for context length extension.
2. Initialization strategies: Insights from Fourier features can inform better initialization of attention mechanisms for improved training dynamics.
3. Transfer learning: The use of Fourier features suggests that models with well-designed attention mechanisms may transfer more effectively across tasks with different input distributions.
4. Interpretability: Analyzing attention mechanisms in the Fourier domain can provide insights into what patterns the model has learned to recognize.

4 Conclusion

By exploring the use of Fourier features for context length extension in attention mechanisms, we gain a deeper understanding of the theoretical underpinnings of modern deep learning architectures. This perspective not only enhances our theoretical understanding but also provides practical guidance for designing more effective and interpretable models. Future work may further explore these techniques to develop novel architectures and training strategies that leverage these insights.

References