Unifying Position Embeddings, Neural Tangent Kernel, and Fourier Features

Anonymous authors
Paper under double-blind review

Abstract

This paper explores the intricate connections between position embeddings, the Neural Tangent Kernel (NTK), and Fourier features in deep learning models. We demonstrate how these concepts are fundamentally linked through the lens of harmonic analysis and kernel methods. Our analysis reveals that position embeddings can be viewed as approximations of the NTK, which in turn can be expressed using Fourier features. We show how Euler's formula provides a unifying framework for understanding these relationships, offering insights into model behavior and guiding the design of more effective architectures.

1 Introduction

Position embeddings have become an essential component in many deep learning architectures, particularly in sequence modeling tasks. Concurrently, the Neural Tangent Kernel (NTK) has emerged as a powerful tool for analyzing neural network behavior, while Fourier features have shown promise in improving model performance. This paper explores the deep connections between these concepts, revealing a unifying perspective that enhances our understanding of modern deep learning models.

2 Theoretical Foundations

2.1 Position Embeddings and Fourier Features

Position embeddings, particularly sinusoidal embeddings used in Transformer models (**?**), can be viewed as a form of Fourier feature mapping. Consider the sinusoidal position embedding function $\phi(x)$ for a 1D input x:

$$\phi(x) = [\sin(\omega_1 x), \cos(\omega_1 x), \sin(\omega_2 x), \cos(\omega_2 x), ..., \sin(\omega_d x), \cos(\omega_d x)] \tag{1}$$

where $\omega_i = 1/10000^{2i/d}$ for $i = 0, 1, ..., d/2 - 1$.

This embedding is essentially a Fourier feature mapping with a specific choice of frequencies. The connection becomes even clearer when we consider Euler's formula:

$$e^{i\omega x} = \cos(\omega x) + i \sin(\omega x) \tag{2}$$

Using Euler's formula, we can express the position embedding in complex form:

$$\phi(x) = [e^{i\omega_1 x}, e^{-i\omega_1 x}, e^{i\omega_2 x}, e^{-i\omega_2 x}, ..., e^{i\omega_d x}, e^{-i\omega_d x}] \tag{3}$$

This representation highlights the connection to the Fourier transform and provides a bridge to the Neural Tangent Kernel.

2.2 Neural Tangent Kernel and Fourier Features

The Neural Tangent Kernel is a theoretical framework that describes the behavior of wide neural networks during training (**?**). Interestingly, the NTK can be expressed in terms of Fourier features.

For a neural network with activation function σ, the NTK can be written as:

$$K(x, x') = \int_{\mathbb{R}} \hat{\sigma}(\omega)\hat{\sigma}(\omega)e^{i\omega(x-x')}d\omega \tag{4}$$

where $\hat{\sigma}$ is the Fourier transform of σ. This formulation reveals that the NTK is essentially a convolution in the frequency domain, which can be approximated using Fourier features.

2.3 Unifying Perspective

The connections between position embeddings, NTK, 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.

2.3.1 Position Embeddings as Fourier Feature Mapping

Theorem 1. *Sinusoidal position embeddings are equivalent to a Fourier feature mapping.*

Proof. Recall the sinusoidal position embedding function $\phi(x)$:

$$\phi(x) = [\sin(\omega_1 x), \cos(\omega_1 x), \sin(\omega_2 x), \cos(\omega_2 x), ..., \sin(\omega_d x), \cos(\omega_d x)] \tag{5}$$

Using Euler's formula, we can express this in complex form:

$$\phi(x) = [e^{i\omega_1 x}, e^{-i\omega_1 x}, e^{i\omega_2 x}, e^{-i\omega_2 x}, ..., e^{i\omega_d x}, e^{-i\omega_d x}] \tag{6}$$

This is precisely the form of a Fourier feature mapping $\psi(x) = [e^{i\omega_1 x}, e^{i\omega_2 x}, ..., e^{i\omega_d x}]$ with the addition of complex conjugates. Therefore, sinusoidal position embeddings are equivalent to a real-valued Fourier feature mapping. □

2.3.2 NTK as an Integral over Fourier Features

Theorem 2. *The Neural Tangent Kernel can be expressed as an integral over Fourier features.*

Proof. For a neural network with activation function σ, the NTK is given by:

$$K(x, x') = \int_{\mathbb{R}} \hat{\sigma}(\omega)\hat{\sigma}(\omega)e^{i\omega(x-x')}d\omega \tag{7}$$

where $\hat{\sigma}$ is the Fourier transform of σ. We can rewrite this as:

$$K(x, x') = \int_{\mathbb{R}} |\hat{\sigma}(\omega)|^2 e^{i\omega(x-x')}d\omega \tag{8}$$

This is a continuous version of a Fourier feature mapping, where $|\hat{\sigma}(\omega)|^2$ acts as a weighting function over the frequencies. Thus, the NTK is indeed an integral over Fourier features. □

2.3.3 Position Embeddings as NTK Approximation

Theorem 3. *For certain activation functions, position embeddings can approximate the NTK.*

Proof. Consider a neural network with a periodic activation function σ that has a sparse Fourier transform $\hat{\sigma}(\omega)$. The NTK for this network can be approximated by:

$$K(x, x') \approx \sum_{k=1}^{d} c_k \cos(\omega_k (x - x')) \tag{9}$$

where c_k are coefficients derived from $|\hat{\sigma}(\omega_k)|^2$. This approximation has the same form as the inner product of two position embeddings:

$$\langle \phi(x), \phi(x') \rangle = \sum_{k=1}^{d} [\sin(\omega_k x) \sin(\omega_k x') + \cos(\omega_k x) \cos(\omega_k x')] \tag{10}$$

Using the trigonometric identity $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, we can see that the position embedding inner product approximates the NTK, with the coefficients c_k absorbed into the choice of frequencies ω_k. □

These proofs formalize the connections between position embeddings, NTK, and Fourier features. This unifying perspective provides several insights:

1. The effectiveness of position embeddings can be partially explained by their ability to approximate the NTK, which governs the learning dynamics of wide neural networks.
2. The choice of frequencies in position embeddings can be guided by analysis of the NTK in the Fourier domain, potentially leading to more effective embedding designs.
3. The connection to Fourier features explains why position embeddings help models capture periodic patterns and overcome spectral bias, as they explicitly introduce a range of frequencies into the model's representation.
4. This perspective suggests that position embeddings might be optimized by considering the spectral properties of the data and the desired kernel for a given task.

Furthermore, this unification opens up new avenues for research:

1. Developing new position embedding schemes based on NTK analysis for specific architectures or tasks.
2. Investigating the relationship between the NTK and the inductive biases introduced by different position embedding designs.
3. Exploring how the interplay between position embeddings and the NTK affects model generalization and extrapolation capabilities.

3 Practical Implications

Understanding the connections between position embeddings, NTK, and Fourier features has several practical implications:

1. Architecture design: The frequency spectrum of the NTK can guide the design of more effective position embedding schemes.
2. Initialization strategies: Insights from the NTK can inform better initialization of position embeddings for improved training dynamics.

3. Transfer learning: The connection to Fourier features suggests that models with well-designed position embeddings may transfer more effectively across tasks with different input distributions.
4. Interpretability: Analyzing position embeddings in the Fourier domain can provide insights into what patterns the model has learned to recognize.

4 Conclusion

By exploring the connections between position embeddings, the Neural Tangent Kernel, and Fourier features through the lens of Euler's formula, we gain a deeper understanding of the theoretical underpinnings of modern deep learning architectures. This unifying 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 connections to develop novel architectures and training strategies that leverage these insights.

References