Composing Kernel Models with Self-Attention

Carey Bunks
Queen Mary University in London
London, UK
c.bunks@qmul.ac.uk

Simon Dixon
Queen Mary University in London
London, UK

Bruno Di Giorgi
Apple
London, UK

Abstract

This study identifies self-attention using rotary positional embeddings (RoPE) as an instance of Nadaraya-Watson (NW) kernel regression. Leveraging the properties of kernels, we modify the attention model to replace RoPE with a novel, explicitly designed bank of decaying periodic kernels. Experiments are conducted using a GPT architecture, a character-based tokenization strategy, and a 13-million-character corpus. The results from the new model significantly outperform the baseline RoPE implementation, as measured by mean cross-entropy loss.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Machine learning*; *Artificial intelligence*; *Machine learning*; • **Theory of computation** → Models of computation.

Keywords

Transformer, Self-Attention, Kernel Methods

ACM Reference Format:
Carey Bunks, Simon Dixon, and Bruno Di Giorgi. 2025. Composing Kernel Models with Self-Attention. In *Proceedings of (UKAIRS).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The Transformer model [17] has revolutionized artificial intelligence, and has become a key foundational architecture across diverse domains such as NLP [10], computer vision [7, 11], speech recognition [6], computational biology [20], and more. Nevertheless, Transformers remain more of a heuristic than a formal scientific framework. An underlying theory explaining not just how, but why they work has remained elusive, but such a theory is, arguably, essential for predicting safety, reliability, and alignment [4]. Theoretical models are useful at several levels. They provide intuition, but more importantly, they provide a foundation for analysis when analyzing errors, and they are a springboard for inventing improved models. The objective of this work is to develop a modified, more explanatory model for self-attention and to evaluate its ability to improve performance when used in a GPT architecture.

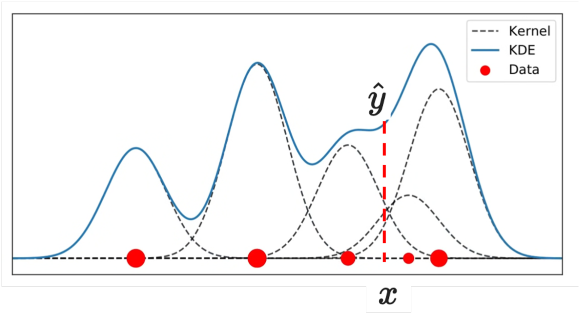

Figure 1: Illustrating Nadaraya-Watson regression. The resulting regression function, shown as a blue curve, is the weighted sum of shifted kernel functions, each shown as a black, dashed curve. The data locations, x_i, are represented by the red dots on the horizontal axis, and their values, y_i are represented by the size of the dots.

2 Methodology

2.1 Theory

The methodology used in this work is based on Nadaraya-Watson (NW) regression [12, 18], which uses a set of observed points, $\{x_i, y_i\}$, $i = 1, \ldots, N$, and a kernel function, $K_h(x - x_i)$, to estimate the value of y at any new point, x:

$$y = \mathrm{NW}(x) = \sum_{i=1}^{N} \left[\frac{K_h(x - x_i)}{\sum_{i=1}^{N} K_h(x - x_i)} \right] y_i \qquad (1)$$

In this expression, the K_h function is centered around each of the x_i and weighted by the corresponding y_i. This shifted and normalized weighted sum forms the regression function. The shape of K_h is typically a symmetric, Gaussian-like curve whose width is controlled by a parameter h. Figure 1 illustrates a simple example.

When self-attention is implemented using rotary positional embeddings (RoPE) [15], its form is the same as NW regression

$$\mathrm{Attention}(x_n) = \sum_{i=1}^{N} \left[\frac{\exp\left(\frac{x_n^T Q^T \Theta^T \Theta K x_i}{\sqrt{d}}\right)}{\sum_{i=1}^{N} \exp\left(\frac{x_n^T Q^T \Theta^T \Theta K x_i}{\sqrt{d}}\right)} \right] V x_i \qquad (2)$$

RoPE is implemented as a sparse matrix, Θ, operating on the query and key vectors. The structure of Θ is block diagonal, where each block is a 2D rotation matrix. The angles of rotation increase as a function of index and position. One of the key characteristics of RoPE is that $\Theta^T \Theta$ is a function of the indicial distance between embeddings. Attention, and NW regression, both form normalized weighted sums dependent on relative distances. Thus, attention can be interpreted as a proper kernel function centered around each

UKAIRS, Newcastle upon Tyne, UK

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

x_i [16]. Although attention is not symmetric, asymmetric kernels have been formalized in both theoretical frameworks and practical applications, and are useful for modeling conditional probabilities and directed graphs [8, 9, 19].

2.2 Kernel Modeling

Kernel functions can be combined through summation or multiplication while remaining valid kernels [1], and this characteristic makes them useful for modeling. RoPE is thought to implicitly embody decaying periodic features that occur in the structure of language [3], and the goal of this section, is to redesign attention, replacing RoPE with kernel functions designed to explicitly model these features. We begin by defining two kernel functions, one for periodicity, P_k, and another for exponential decay, D_k:

$$P_k(x_n, x_i) = \exp\left\{-2\alpha_k^2 \sin^2\left(\frac{|n-i|}{\tau_k}\right)\right\} \tag{3}$$

$$D_k(x_n, x_i) = \sigma_k^2 \exp\left\{-\frac{|n-i|}{l_k}\right\} \tag{4}$$

Each kernel is an explicit function of the indicial distance between x_n and x_i. P_k is a function of two learnable parameters α_k and τ_k, where the former controls amplitude and the latter wavelength. D_k depends on the learnable parameters σ_k and l_k, where the former is the strength of the term and the latter is a time constant or decay width parameter. The two kernels can be multiplied, $G_k(x_n, x_i) = D_k(x_n, x_i)P_k(x_n, x_i)$, to model decaying periodicity, and summed to create a bank of M such kernels:

$$G(x_n, x_i) = \sum_{k=1}^{M} D_k(x_n, x_i)P_k(x_n, x_i) \tag{5}$$

Finally, the expression in Equation 5 can be combined with that for attention from Equation 2 without, however, the $\Theta^T\Theta$ terms from RoPE:

$$\mathrm{GPA}(x_n) = \sum_{i=1}^{N}\left[\frac{G(x_n, x_i)\exp(\frac{x_n^T Q^T K x_i}{\sqrt{d}})}{\sum_{i=1}^{N} G(x_n, x_i)\exp(\frac{x_n^T Q^T K x_i}{\sqrt{d}})}\right] V x_i \tag{6}$$

We call this model Gaussian process attention (GPA), and by comparison with static RoPE, it contains $4M$ additional learnable parameters.

2.3 Experimental Setup

The experiments are based on a standard GPT Transformer, initially using RoPE as the baseline. The data for the experiment is the collected works of Charles Dickens, obtained from Project Gutenberg [5], and we use a character-based tokenization strategy [2]. The corpus contains 13m characters, with a total vocabulary of 93 tokens. The baseline Transformer architecture consists of four blocks, each containing four attention heads and a feedforward layer. The model also includes a layer norm and the usual embedding and unembedding layers. The context window was set to 256, and the embedding dimension 512. Finally, the performance metric used to evaluate results is the mean cross-entropy loss of the validation data.

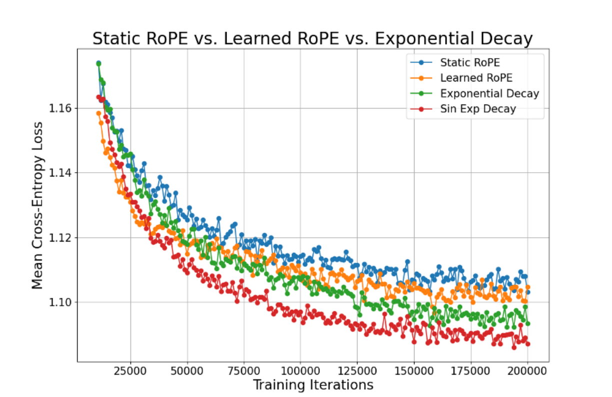

Figure 2: Comparison of four experiments: blue is static RoPE, orange is learned RoPE, green is exponential decaying kernel, and red is decaying periodic kernel.

3 Results

Four experiments were run to evaluate the kernel modeling ideas of the previous sections, and the results are shown in Figure 2. Each curve in the figure represents the mean cross-entropy (MCE) loss during training. The experiments are run for 200,000 iterations with a batch size of 256, which is the equivalent of running for 4 epochs. The data was split into a training set, with 10% set aside as a validation set. The blue curve in the figure is the MCE loss of the validation data using the baseline RoPE implementation of the GPT architecture as specified in Equation 2. The red curve is the MCE loss for the Gaussian process attention (GPA) kernel described by Equation 6 and using a bank of $M = 8$ decaying periodic kernels. It is the best performing model, showing a considerable improvement over the RoPE experiment. To test ideas, two additional models were run. The orange curve is a modification of RoPE, where the angular rotations, normally static values, are trained as learnable parameters. The legend refers to this experiment as *Learned RoPE*. The final experiment removes the periodic kernel component of the GPA expression, using only the exponential decaying part from Equation 4. This performs better than either the static or learned RoPE experiments, but not as well as the decaying sinusoidal kernel.

4 Future Work

The results from the previous section seem promising, but they are for a small corpus with a simple tokenization scheme. Experiments with a larger corpus (for example, an English Wikipedia dump), and a more sophisticated tokenization strategy (such as WordPiece [13] or byte-pair encoding [14]) need to be explored to ensure that the results carry over to more realistic problems. The use of kernels as a modeling methodology seems to hold significant potential, and suggests many new avenues of scientific enquiry, exploring kernel compositional structures, parameterization, and model bank order. Testing these models in downstream applications would provide additional insight into their strengths, weaknesses, and capabilities.

Acknowledgments

The first author gratefully acknowledges the support of the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, under grant number EP/S022694/1.

References

[1] Nachman Aronszajn. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society* 68, 3 (1950), 337–404.

[2] Nikolay Banar, Walter Daelemans, and Mike Kestemont. 2020. Character-level transformer-based neural machine translation. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 149–156.

[3] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2024. Round and Round We Go! What makes Rotary Positional Encodings useful? 10 pages. arXiv:2410.06205

[4] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety–A Review. 55 pages. arXiv:2404.14082

[5] Charles Dickens. 2009. The Works of Charles Dickens. [Online]. Available: https://www.gutenberg.org/ebooks/139.

[6] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition.

[7] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.

[8] Fan He, Mingzhen He, Lei Shi, Xiaolin Huang, and Johan AK Suykens. 2023. Enhancing Kernel Flexibility via Learning Asymmetric Locally-Adaptive Kernels.

[9] Mingzhen He, Fan He, Lei Shi, Xiaolin Huang, and Johan AK Suykens. 2023. Learning with asymmetric kernels: Least squares and feature interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10044–10054.

[10] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing.

[11] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.

[12] Elizbar A Nadaraya. 1964. On estimating regression. *Theory of Probability & Its Applications* 9, 1 (1964), 141–142.

[13] Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Institute of Electrical and Electronics Engineers, Piscataway, NJ, USA, 5149–5152.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units.

[15] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Transformer dissection: a unified understanding of transformer's attention via the lens of kernel.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008.

[18] Geoffrey S Watson. 1964. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 26, 4 (1964), 359–372.

[19] Wei Wu, Jun Xu, Hang Li, and Satoshi Oyama. 2010. Asymmetric kernel learning.

[20] Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. 2023. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances* 3, 1 (2023), vbad001.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009