

PoPE: Legendre Orthogonal Polynomials Based Position Encoding for Large Language Models

Arpit Aggarwal

Airtel Africa Digital Labs

ARPIT.AGGARWAL@AFRICA.AIRTEL.COM

Abstract

There are several improvements proposed over the baseline Absolute Positional Encoding (APE) method used in original transformer. In this study, we aim to investigate the implications of inadequately representing positional encoding in higher dimensions on crucial aspects of the attention mechanism, the model’s capacity to learn relative positional information, and the convergence of models, all stemming from the choice of sinusoidal basis functions. Through a combination of theoretical insights and empirical analyses, we elucidate how these challenges extend beyond APEs and may adversely affect the performance of Relative Positional Encoding (RPE) methods, such as Rotatory Positional Encoding (RoPE).

Subsequently, we introduce an innovative solution termed Orthogonal Polynomial Based Positional Encoding (PoPE) to address some of the limitations associated with existing methods. The PoPE method encodes positional information by leveraging Orthogonal Legendre polynomials. Legendre polynomials as basis functions offers several desirable properties for positional encoding, including improved correlation structure, non-periodicity, orthogonality, and distinct functional forms among polynomials of varying orders. Our experimental findings demonstrate that transformer models incorporating PoPE outperform baseline transformer models on the *Multi30k* English-to-German translation task, thus establishing a new performance benchmark. Furthermore, PoPE-based transformers exhibit significantly accelerated convergence rates.

Additionally, we will present novel theoretical perspectives on position encoding based on the superior performance of PoPE.

Keywords: Large Language Models, Position Encoding, Transformers, Natural Language Processing

1 Introduction

Positional encoding was perhaps one of the two major ideas along with multi-headed attention (Vaswani et al., 2017) which revolutionized the field of language modeling. Unlike other language modeling paradigms based on recurrent architectures such as RNNs, LSTMs, and GRUs, position information of a token is not explicitly given in transformer models but rather incorporated implicitly by generally adding (Vaswani et al., 2017) or multiplying (Su et al., 2021) by some pre-defined function with token embedding, or in some cases learned as parameter (Raffel et al., 2020) of the model, along with other variations. Several explanations have been proposed to understand the role of positional encoding in attention mechanism (Chen et al., 2021; He et al., 2020; Kazemnejad et al., 2023; Raffel et al., 2020; Su et al., 2021). In this paper we shall expand some theoretical aspects of position encoding, discuss shortcoming of some existing methods and propose a new absolute positional encoding scheme based on orthogonal polynomials. We shall show with empirical, theoretical, and

experimental evidence that using such polynomials for encoding positional information has inherent advantages over some of the existing APE and RPE methods. We shall approach this paper as follows:

- Our initial focus will be on examining the limitations of prevailing methods in accurately representing positional information, particularly in high-dimensional contexts. These shortcomings lead to the provision of a biased informative prior to the model. Through empirical and theoretical analyses, we will illustrate how this bias negatively impacts the performance of certain Absolute Positional Encoding (APE) methods and minimizes the effectiveness of some Relative Positional Encoding (RPE) techniques. We will provide compelling evidence to demonstrate that the inadequacy of most APE and RPE methods in addressing this issue is primarily attributed to the selection of basis functions.
- Next, we introduce a novel positional encoding scheme based on Legendre polynomials. We aim to illustrate how this innovative approach effectively addresses prevalent challenges encountered in current encoding methodologies. We will elucidate the advantageous mathematical properties inherent in the Legendre polynomial family, including orthogonality, distinct functional forms, three-recursion relations, and superior correlation structures, particularly evident in higher dimensions. We assert that these properties position PoPE as a promising alternative to existing Absolute Positional Encoding (APE) and Relative Positional Encoding (RPE) methods.
- With experimental results, we shall demonstrate that transformer models with PoPE outperforms both *base* and *big* transformer models on *Multi30k* English to German translation task and consistently converges much faster during training routines.
- In conclusion, we will put forth theoretical explanations that delve into the core mechanisms of positional encoding methods. Our aim is to elucidate why the PoPE method exhibits superior performance, followed by the slightly enhanced performance of Relative Positional Encoding (RPE) techniques like RoPE, and the advantages inherent in alternative paradigms such as learned positional encoding.

2 Prelude to PoPE

In this section we shall lay out the mathematical structure of position encoding and linear attention mechanism, then we shall demonstrate some shortcoming of the existing approaches through empirical evidence and mathematical analysis. Application of proposed PoPE methodology is not limited to any particular model architecture, in this paper we have chosen original transformer architecture as a case study. Later, we shall show that the principles discussed here have wider scope.

2.1 Background and Framework

We shall adapt mathematical notations from Su et al. (2021) regarding attention mechanism of transformer models.

Let $\mathbb{S}_N = \{w_i\}_{i=1}^N$ be a sequence of N input tokens.

w_i the i^{th} token.

$\mathbb{E}_N = \{\mathbf{x}_i\}_{i=1}^N$ denoted the token embedding of \mathbb{S}_N ,

where $\mathbf{x}_i \in \mathbb{R}^d$ is the d -dimensional token embedding vector of token w_i .

In self attention, Key, query, and values (\mathbf{q}_n , \mathbf{k}_n , and \mathbf{v}_n) in self attention incorporate the m^{th} and n^{th} positions through transformations f_q , f_k and f_v (Equation (1)).

$$\begin{aligned} \mathbf{q}_m &= f_q(\mathbf{x}_m, m) \\ \mathbf{k}_n &= f_k(\mathbf{x}_n, n) \\ \mathbf{v}_n &= f_v(\mathbf{x}_n, n), \end{aligned} \tag{1}$$

In typical absolute positional encoding and attention scheme, positional encoding is added to the token embedding and then transformed using linear operators, such as different projection matrices for key, query, and values (Equation (2)).

$$f_{t:t \in \{q,k,v\}}(\mathbf{x}_i, i) := \mathbf{W}_{t:t \in \{q,k,v\}}(\mathbf{x}_i + \mathbf{p}_i), \tag{2}$$

Key and Query values are used to compute attention a_{mn} using inner product of \mathbf{q}_m and \mathbf{k}_n . The output \mathbf{o}_m is the sum of value representations \mathbf{v}_n , weighted by calculated attentions (Equation (3)).

$$\begin{aligned} a_{m,n} &= \frac{\exp(\frac{\mathbf{q}_m^\top \mathbf{k}_n}{\sqrt{d}})}{\sum_{j=1}^N \exp(\frac{\mathbf{q}_m^\top \mathbf{k}_j}{\sqrt{d}})} \\ \mathbf{o}_m &= \sum_{n=1}^N a_{m,n} \mathbf{v}_n \end{aligned} \tag{3}$$

The inner product of query and keys form the basis for the above attention machinery (Equation (4)).

$$\mathbf{q}_m^\top \mathbf{k}_n = \mathbf{x}_m^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{x}_n + \mathbf{x}_m^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{p}_n + \mathbf{p}_m^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{x}_n + \mathbf{p}_m^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{p}_n \tag{4}$$

2.2 Empirical Analysis

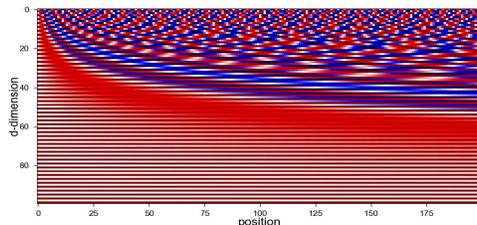
In the original transformer architecture, a sinusoidal position encoding scheme was proposed. The sinusoidal positional encoding method exhibits a fundamental limitation characterized by high mutual information at higher dimensions (Figure 1(a)). For instance, our analysis reveals a significant correlation trend between positional encoding values of a model with d_{512} dimensions at higher dimensions ($d > 356$) (see Figure 1(b)):

$$Corr(\mathbf{p}_{pos=n, d > 356}, \mathbf{p}_{pos=m, d > 356}) \tag{5}$$

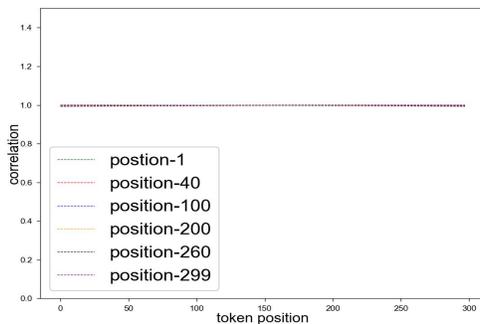
Where, $n \in \{1, 40, 100, 200, 260, 299\}$ and m varies for all d

At higher dimensions, a near-perfect correlation (>0.999 percent) is observed, even for token positions that are far apart in the sentence, such as position 299 and position 2. This observation underscores the inadequacy of sinusoidal functions in developing a dense and unique positional representation in higher dimensions when used in their original form. In the next section, we shall demonstrate mathematically that:

1. This causes the model to have a biased informative prior, causing learning overheads.
2. Failing to have a proper dense representation might lead the model to lose ability to learn relative positioning effectively.



(a) Cosine positional encoding



(b) Near perfect correlation at higher dimensions

Figure 1: Low variance at higher dimensional values of sinusoidal positional encoding (a), and near perfect correlation among encoding of different token positions (b)

2.3 Mathematical Analysis

Equation (4) admits various interpretations. Ke et al. (2020) proposed the possibility of a correlation between positional encoding and words by analyzing the cross-product terms between word embedding and positional encoding. Conversely, He et al. (2020) argued that relative positional information can only be modeled using the middle two terms, i.e., the cross-terms between words and positional encoding. In this section, we present a different interpretation, focusing our attention primarily on the fourth cross-product term between two positional encoding vectors.

We begin by isolating the effect of high correlation in higher dimensions on attention, as discussed in the previous section. Without loss of generality and for better explanatory

clarity, we assume that positional encodings are concatenated with token embeddings, denoted as $[\mathbf{x}_i|\mathbf{p}_i]$, rather than added to them. Functionally, these two operations are similar, requiring only slight adjustments to the projection matrices \mathbf{W}_k , \mathbf{W}_q , and \mathbf{W}_v . We denote these modified projection matrices as $\mathbf{W}'_q \sim \mathbf{W}_q$, $\mathbf{W}'_k \sim \mathbf{W}_k$, and $\mathbf{W}'_v \sim \mathbf{W}_v$.

Further, we shall rewrite $[\mathbf{x}_i|\mathbf{p}_i]$ as an augmented matrix of three blocks, i.e., token embedding, lower dimension values of positional encoding, \mathbf{p}_i^{l-} (below a correlation threshold of l), and higher dimensional values of positional encoding, \mathbf{p}_i^{l+} , with high degree of correlation (Equation (6)).

$$\mathbf{s}_i = [\mathbf{x}_i|\mathbf{p}_i] = [\mathbf{x}_i|\mathbf{p}_i^{l-}|\mathbf{p}_i^{l+}] \tag{6}$$

We can refactor Equation (2) in terms of augmented matrices and write the inner product self attention using query and key vectors as follows:

$$\mathbf{q}_m^\top \mathbf{k}_n \sim (\mathbf{W}'_q \mathbf{s}_m)^\top (\mathbf{W}'_k \mathbf{s}_n) \tag{7}$$

$$= \mathbf{s}_m^\top \mathbf{W}'_q{}^\top \mathbf{W}'_k \mathbf{s}_n \tag{8}$$

Using Equation (6) and writing the product $\mathbf{W}'_q{}^\top \mathbf{W}'_k$ as a single matrix of three blocks corresponding to \mathbf{s}_i .

$$\mathbf{W}'_q{}^\top \mathbf{W}'_k = [\mathbf{W}^1_{kq} | \mathbf{W}^2_{kq} | \mathbf{W}^3_{kq}]^\top \tag{9}$$

from which it follows

$$\mathbf{q}_m^\top \mathbf{k}_n \sim \mathbf{x}_m \mathbf{W}^1_{kq} \mathbf{x}_n + \mathbf{p}_m^{l-} \mathbf{W}^2_{kq} \mathbf{p}_n^{l-} + \mathbf{p}_m^{l+} \mathbf{W}^3_{kq} \mathbf{p}_n^{l+} \tag{10}$$

As discussed in the previous section, a notable degree of correlation exists between \mathbf{p}_m^{l+} and \mathbf{p}_n^{l+} , the third term representing the generalized inner product between higher-dimensional values of positional encoding. This correlation essentially introduces a biased informative prior into the self-attention mechanism. Here, we contend that this poor representation leads to training and learning overheads. Furthermore, this effect extends to cross-terms involving word embeddings and positional embeddings, thereby hindering the learning of relative positioning as proposed by some studies (Ke et al., 2020). We conjecture the following regarding the importance of a proper dense representation of positional information:

- We argue that if position encoding is represented densely by some functional form, the generalized inner product between position encoding ($\mathbf{p}_m^{l+} \mathbf{W}^3_{kq} \mathbf{p}_n^{l+}$) has sufficient mathematical structure to it to learn relative position information. This assertion holds true for the fourth term of Equation (4).
- While some existing methods, such as RoPE, are inherently multiplicative in nature, thereby avoid introducing an explicit bias, our analysis demonstrates that they still fail to mitigate the high degree of correlation at higher dimensions induced by sinusoidal functions used as multiplicative factors.

- Previous works have attempted to replace positional encoding with a scalar learnable bias Raffel et al. (2020). In the current context, this approach resembles Equation (11). Despite empirical and theoretical evidence suggesting the efficacy of this positional encoding scheme, it falls short in explicitly providing relative positioning information.
- We argue that there is no absolute positional encoding scheme that satisfies all the criteria of: 1) effectively representing positional information in a dense and well-behaved manner, even in higher dimensions, and 2) encoding relative position information effectively.

$$\mathbf{q}_m^\top \mathbf{k}_n \sim \mathbf{x}_m \mathbf{W}_{kq}^1 \mathbf{x}_n + b_{mn} \quad (11)$$

3 Proposed Method: PoPE

In this section, we propose an Orthogonal Polynomial Based Positional Encoding (PoPE) scheme. Specifically, we introduce a unique class of orthogonal polynomials referred to as Legendre polynomials. We will detail our proposed formulation of the positional encoding scheme, highlighting the desirable mathematical structure and properties of Legendre polynomials that address certain limitations observed in existing positional encoding schemes.

3.1 Legendre Polynomials

Legendre polynomials $P_n(x)$ are special category of orthogonal polynomials (Figure 2). They appear in numerous engineering and physics problems, especially related to spherical harmonics (Dassios, 2012). One of the compact forms of Legendre polynomials is given by Rodrigues’ formula (Equation (12)).

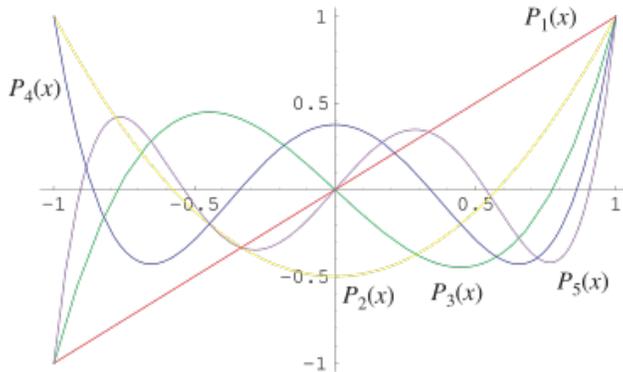


Figure 2: First four order Legendre Polynomials, credit: *Mathworld*

$$\mathbf{P}_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad (12)$$

Where, $\mathbf{P}_n(x)$ is Legendre polynomial of order n . Legendre polynomials are orthogonal functions under the restricted domain $x \in [-1, 1]$ (Equation (13)), thus, just as sinusoidal functions, can be chosen as basis functions in Fourier system.

$$\int_{-1}^1 \mathbf{P}_n(x) \mathbf{P}_m(x) dx = \delta_{mn} \quad (13)$$

3.2 Proposed Formulation

For a token of Position = pos , we propose deterministic and equidistant sampling from a Legendre polynomial $\mathbf{P}_n(x)$ ($x \in [-1, 1]$) of order $n = pos$, hence for a given d_{model} , Equation (14) defines the value of position encoding for a given position pos and index i .

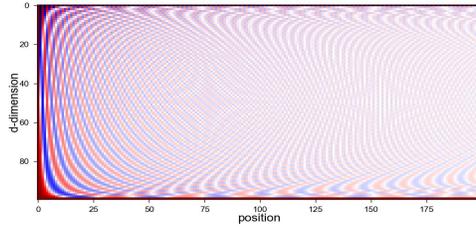
$$PE_{(pos,i)} = \mathbf{P}_{pos}(x_i) \quad (14)$$

Where, x_i are equal interval samples in $[-1, 1]$ with interval length = $2i/d_{model}$.

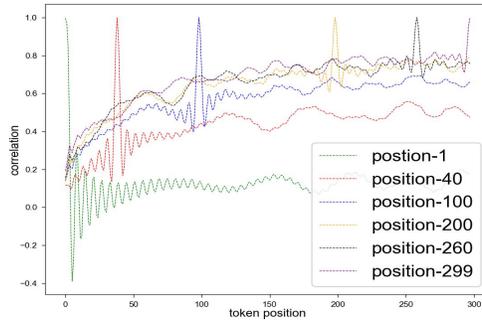
3.3 Properties of PoPE

Well behaved at higher dimensions: Similar to sinusoidal functions, Legendre polynomials exhibit orthogonality; however, they possess non-periodic characteristics, with each polynomial of order n having a distinct functional form. Drawing an analogy from signal processing, sinusoidal position encoding resembles frequency or phase modulation, where position is encoded by altering the phase or frequency (both represented as angular terms in a sinusoidal function). In contrast, Legendre polynomials can be likened to modulation in both phase and amplitude. This dual modulation results in increased entropy, allowing for the encoding of a greater amount of information even in higher dimensions. Consequently, Legendre polynomials offer a more comprehensive and efficient representation (Figure 3(a)). Improved dense representation offers two significant advantages. Firstly, the cross-product term between positional encoding will not exhibit a biased informative prior. As illustrated in Figure 3(b), the correlation is more accurately represented as a function of distance. Additionally, as previously argued, enhanced representation allows the generalized inner product term between position encoding to effectively learn relative positioning between tokens. This effect will be further pronounced due to other properties of PoPE, which we will discuss in the following section.

PoPE



(a) PoPE Scheme



(b) Much better correlation structure at higher dimensions

Figure 3: PoPE has much better representation (a), correlation among token positions are much better managed (b)

Learning relative positions: It is often considered beneficial for positional encoding to have a learnable linear structure between positions, which means for any given offset k positional encoding PE_{pos+k} can be represented as a linear function of PE_{pos} . The original transformer paper (Vaswani et al., 2017) claims this for sinusoidal positional encoding. In this section we shall show how PoPE has a well defined linear recurrent structure which helps us to define relative positions. By virtue of Favard’s theorem (Favard, 1935), all orthogonal polynomials satisfy a three-recurrence relations, Equation (15) defines a three-recurrence relation for a Legendre polynomial.

$$(2n + 1)xP_n(x) = nP_{n-1}(x) + (n + 1)P_{n+1}(x) \quad (15)$$

This recurrent relation for Legendre polynomials can be formulated as a tri-diagonal Jacobi matrix (Van Assche, 2022), refer Equation (16). The Jacobi matrix truncated to any order n has eigen values equal to the roots of Legendre polynomial of order n , this implies relation recurrence relations between Legendre polynomials contain complete infor-

mation about any $P_n(x)$. It is interesting to note that Legendre polynomial of any degree has a linear relation with one lower and higher order polynomial, which can be extended indefinitely over any order $n \in [0, \infty)$. Also to be noted that for a given model architecture, X is a constant, since it is fixed d_{model} sample values from interval $[-1, 1]$, where d_{model} corresponds to the dimension of the token embedding. This is a highly desirable property of Legendre polynomials and orthogonal polynomials in general, it provides a firm mathematical structure to each position encoding of a token in a sentence with respect to each position in both forward and backward direction.

$$\mathbf{X} \begin{bmatrix} P_0(x) \\ P_1(x) \\ P_2(x) \\ P_3(x) \\ \vdots \\ P_{n-1}(x) \\ P_n(x) \end{bmatrix} = \begin{bmatrix} b_0 & a_0 & 0 & 0 & 0 & \dots & 0 \\ c_1 & b_1 & a_1 & 0 & 0 & \dots & 0 \\ 0 & c_2 & b_2 & a_2 & 0 & \dots & 0 \\ 0 & 0 & c_3 & b_3 & a_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & c_{n-1} & b_{n-1} & a_{n-1} & 0 \\ \dots & \dots & \dots & \dots & c_n & b_n & a_n \end{bmatrix} \begin{bmatrix} P_0(x) \\ P_1(x) \\ P_2(x) \\ P_3(x) \\ \vdots \\ P_{n-1}(x) \\ P_n(x) \end{bmatrix} \quad (16)$$

Hence, any Legendre polynomial can be expressed as a linear combination of other Legendre polynomials of higher and lower orders through nested recursion relations. This indicates that the position encoding for any given token position inherently contains relative information from all other positions embedded within it, owing to the recurrence relation.

$$P_n(x) = \sum_l a_l P_{n+l}(x)$$

Where, $l \in \{[n, \infty) - \{0\}\}$

We argue that this property facilitates the injection of relative position information into the model, thereby enhancing the model’s capability to learn relative position information through generalized inner product terms between positional encodings ($\mathbf{p}_m^{l+/-} \mathbf{W}_{kq}^3 \mathbf{p}_n^{l+/-}$ and $\mathbf{p}_m^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{p}_n$). The inner product term between position encoding of m^{th} and n^{th} token diffuses across position encoding values, thereby imparting relative position information to the model.

$$\mathbf{p}_m^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{p}_n \sim \sum_l \sum_j a_i a_j \mathbf{p}_i^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{p}_j$$

Where a_i, a_j are specific to any given two position m and n .

4 Experiments and Results

We assessed the performance of the baseline transformer augmented with PoPE on the *Multi30K* English-to-German translation task. In this section, we provide details on the experimental setup, including hardware specifications, model parameters, and the comparative analysis of PoPE with previous benchmarks.

4.1 Data and task

We evaluated the performance of the base transformer model with both sinusoidal and Legendre polynomial-based positional encoding (PoPE) schemes on a bilingual (English to German) language translation task using the *Multi30K* dataset (Elliott et al., 2016).

Multi30K dataset extends the *Flickr30K* dataset with translated and independent German sentences (Young et al., 2014). Dataset contains 29,000 training and 1,014 validation sentence pairs. There are three test datasets available *test_2016_flickr* and *test_2017_flickr* with 1,000 sentence pairs, and *test_2017_mscoco* with 461 sentences. In this paper we have used all testing samples. Due to limited compute resources at disposal, *Multi30K* provided the ideal experiment data to experimentally support theoretical and empirical merits and aspects of PoPE discussed in this paper.

4.2 Model and Hardware

We carried out the experimentation using the base transformer model, as defined in the original transformer with $d_{model} = 512$, eight attention heads $h = 8$, and $d_q = d_k = d_v = d_{model}/h$, dimensions of inner layers $d_{ff} = 2048$, six encoder and decoder blocks $N = 6$. Due to compute restrictions, we did not evaluate the transformer (big) model and larger WMT 2014 dataset, this is a work in progress. We trained and evaluated the models on 1 NVIDIA RTX™ A4500 GPU. To make a one to one comparison, we trained the model for 10,000 iterations on *Multi30k* dataset, inline with Zhang et al. (2019).

4.3 Results

On *Multi30k* English to German translation task, the base transformer model with PoPE achieves a BLEU score of 40.7, which is 4-5 BLEU more than base and big transformer models with sinusoidal positional encoding, achieving SOTA performance among baseline transformer models and establishing new a new benchmark, it is to be noted that, while *Multi30k* is a multimodal dataset, we are comparing NMT text to text performance here. There are many multimodal studies which outperform text to text models on *Multi30K* English to German translation task (IKD-MMT (Peng et al., 2022) BLEU 41.28, ERNIE-UniX2 BLEU 49.3 (Shan et al., 2022)), but base transformer with PoPE was able to outperform some of the multimodal approaches (DCCN (Lin et al., 2020) BLEU 39.7, Caglayan (Sulubacak et al., 2020) BLEU 39.4, Multimodal Transformer (Yao and Wan, 2020) BLEU 38.7).

In our training routines, the proposed model (transformer + PoPE) converged at two to three times faster than original transformer model with sinusoidal positional encoding (Figure 4).

Table 1: Comparison with Transformer Based Text to Text NMT on EN-DE Task

Study	Model	Task	Dataset	Training steps	BLEU score
Vaswani et al. (2017)	Transformer (base)	EN to DE	<i>WMT</i> 2014	100,000	27.3
	Transformer (big)	EN to DE	<i>WMT</i> 2014	300,000	28.4
Zhang et al. (2019)	Transformer (base)	EN to DE	<i>Multi30k</i>	10,000	35.59
	Transformer (big)	EN to DE	<i>Multi30k</i>	10,000	36.86
This study (PoPE)	Transformer (base)	EN to DE	<i>Multi30k</i>	10,000	40.7

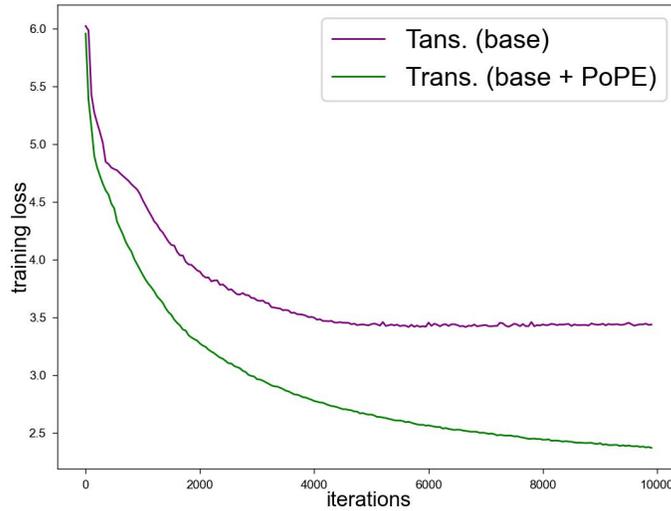


Figure 4: Training loss convergence with and without PoPE

5 Discussion

Through empirical, theoretical, and experimental analyses presented in this paper, we have demonstrated how better representation can address many issues related to positional encoding. We showcased the advantages of Orthogonal Polynomial-based Positional Encoding (PoPE), highlighting its desirable mathematical properties and superior performance in comparison to other models, with significantly improved convergence rates.

In summary, our this paper yields several key insights:

1. PoPE offers superior representation of positional encoding even at higher dimensions compared to sinusoidal positional encoding, which often exhibits highly correlated values at higher dimensions. This can be attributed to limited encoding capability of sinusoidal functions with frequency/phase modulation alone at higher dimensions, various properties of Legendre polynomials such as orthogonality, non periodicity, and different functional form of the Legendre polynomial functions of different orders within $[-1, 1]$, which proves them to be much better basis than sinusoidal functions for encoding position information effectively. It is easy to prove that even rotatory encoding scheme does not solve this problem, though multiplication avoids a bias term in attention and prove to be marginally better than APE method using sinusoidal functions.
2. From theory it is indicative that that the cross term between higher dimension values of position encoding ($\mathbf{p}_m^{l+} \mathbf{W}_{kq}^3 \mathbf{p}_n^{l+}$) behave like a biased informative prior, and is an additional learning overhead. We argue this induces bias can potentially cause slow convergence of the model. PoPE avoids this problem altogether by better representation and thus have much superior converging properties, proven by experiment

results. This observation sheds light on why RoPE-based positional encoding, which explicitly avoids a bias term, exhibits slightly faster convergence compared to baseline transformer models, a phenomenon noted but not fully addressed in the original RoPE paper (Su et al., 2021). From this, it is evident that there is some merit in avoiding a weak positional encoding altogether and explicitly learn a scalar valued bias (Raffel et al., 2020).

3. We argue that providing relative position information to the model offers little merit. We argue, with dense representation across dimensions, the generalized inner product term $(\mathbf{p}_n^{l+/-} \mathbf{W}_{kq}^3 \mathbf{p}_m^{l+/-})$ is sufficient to infer relative positioning. Methods such as RoPE and are categorized as relative positional encoding scheme, but performs much like absolute positional encoding schemes (Kazemnejad et al., 2023) with marginal benefits (Shaw et al., 2018).
4. Theoretically, PoPE-based encoding schemes can learn both absolute and relative position information effectively due to their better representation and the existence of three-recursion relations among orthogonal polynomials. The linear relation between Legendre polynomials of different orders (and hence positions) is more comprehensive than that afforded by sinusoidal functions.

6 Limitations and Future Scope

- We have tested the proposed PoPE scheme on a smaller *Multi30K* dataset. Further work is needed to gather more experimental evidence on much larger benchmark datasets and task to fully assess the capability of PoPE.
- Merits of PoPE are not restricted to the structure of the original transformer paper, it could also be used in other modeling paradigms as well, such as providing per head positional encoding, which removes rank restriction on attention matrices (Chen et al., 2021). Further work is required to explore full potential of PoPE.
- We have demonstrated that the recurrence properties of Legendre polynomials possess sufficient structure, akin to sinusoidal functions, to facilitate the linear learning of relative positions. However, the relationship becomes more intricate with orthogonal polynomials. Further research is required to comprehensively develop a mathematical understanding of how these complex relations can aid large language models such as transformers.

7 Declaration

The views expressed in this paper are solely those of the author.

References

Pu-Chin Chen, Henry Tsai, Srinadh Bhojanapalli, Hyung Won Chung, Yin-Wen Chang, and Chun-Sung Ferng. A simple and effective positional encoding for transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors,

- Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2988, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.236. URL <https://aclanthology.org/2021.emnlp-main.236>.
- George Dassios. *Ellipsoidal Harmonics: Theory and Applications*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2012.
- Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459, 2016. URL <http://arxiv.org/abs/1605.00459>.
- J. Favard. Sur les polynomes de Tchebicheff. *C. R. Acad. Sci., Paris*, 200:2052–2053, 1935. ISSN 0001-4036.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020. URL <https://arxiv.org/abs/2006.03654>.
- Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24892–24928. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4e85362c02172c0c6567ce593122d31c-Paper-Conference.pdf.
- Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *CoRR*, abs/2006.15595, 2020. URL <https://arxiv.org/abs/2006.15595>.
- Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM ’20, page 1320–1329, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413715. URL <https://doi.org/10.1145/3394171.3413715>.
- Ru Peng, Yawen Zeng, and Jake Zhao. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2379–2390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.152. URL <https://aclanthology.org/2022.emnlp-main.152>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.

- Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation, 2022.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018.
- Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021. URL <https://arxiv.org/abs/2104.09864>.
- Umut Sulubacak, Ozan Çağlayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. Multimodal machine translation through visuals and speech. *Machine Translation*, 34, 09 2020. doi: 10.1007/s10590-020-09250-0.
- Walter Van Assche. Orthogonal polynomials, toda lattices and painlevé equations. *Physica D: Nonlinear Phenomena*, 434:133214, 2022. ISSN 0167-2789. doi: <https://doi.org/10.1016/j.physd.2022.133214>. URL <https://www.sciencedirect.com/science/article/pii/S0167278922000343>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL <http://arxiv.org/abs/1706.03762>.
- Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.400. URL <https://aclanthology.org/2020.acl-main.400>.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a.00166. URL <https://aclanthology.org/Q14-1006>.
- Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2019.