
SPHERICAL POSITION ENCODING FOR TRANSFORMERS

Eren Unlu
Datategy SAS
Paris, France
`eren.unlu@datategy.fr`

ABSTRACT

Position encoding is the primary mechanism which induces notion of sequential order for input tokens in transformer architectures. Even though this formulation in the original transformer paper has yielded plausible performance for general purpose language understanding and generation, several new frameworks such as Rotary Position Embedding (RoPE) are proposed for further enhancement. In this paper, we introduce the notion of "geotokens" which are input elements for transformer architectures, each representing an information related to a geological location. Unlike the natural language the sequential position is not important for the model but the geographical coordinates are. In order to induce the concept of relative position for such a setting and maintain the proportion between the physical distance and distance on embedding space, we formulate a position encoding mechanism based on RoPE architecture which is adjusted for spherical coordinates.

Keywords Deep Learning · Transformers · Position Information Encoding

1 Introduction

Transformer architecture proposed by [1] has proven its efficiency and robustness and has become the ultimate backbone of numerous revolutionizing natural language generation applications and even other modalities of generative artificial intelligence [2][3]. Unlike their predecessors Recurrent Neural Networks (RNNs), these architectures do not encode the sequential positions inherently, which is a byproduct of their parallel processing for efficiency with self-attention [4][5]. As sequential order is one of the most impactful aspect of natural language, authors of [1] have proposed to use position embeddings as an intuitive and effective solution. For natural language understanding and generation this has proven to be highly robust and productive as many generative models igniting the emergence of generative AI era are based on its slight variants.

Though position encoding proposed in the original transformer architecture has shown sufficient capabilities in terms of sequential order understanding [1][6], several attempts have been made to further refine and enhance this mechanism [7]. [8] and [9] propose to encode relative position of tokens, also interacting with query, key and value matrices or introducing new types of neural layers. While these methods are effective, they typically incorporate position information into the context representation, making them incompatible with the linear self-attention architecture as noted by [10].

The recent paper on the novel Rotary Position Embedding (RoPE) presents an innovative approach to integrating positional information into the transformer architecture [10]. This novel embedding, distinguishing itself from traditional methods, employs a rotation matrix to encode absolute positions while concurrently embedding explicit relative position dependencies in the self-attention mechanism. They demonstrate theoretically that relative position can be formulated as a vector multiplication in self-attention as absolute position is encoded through a rotation matrix. The proposed method in this study can be considered as a generalization of RoPE mechanism for three dimensional spherical space, which has crucial importance for geographical data representation.

Firstly, we introduce the notion of "geotokens" and "cartographical transformer architecture". Encoding geographical entities properly has immense potential for forthcoming generative AI age, where physical coordinates are well represented and notions of relative distance and hierarchical nature are well preserved. Therefore a transformer based

framework is proposed where each input token isn't just a piece of textual piece, but a "geotoken" representing a geographical entity. In its simplest formulation as presented in this paper, the geotokens are not sequence dependent intuitively but are based on spatial relationships. The pivotal idea is that the significance of a geotoken is not derived from its position in a sequence, as is the case with typical natural language tokens, but rather from its geographical coordinates and its relative positioning to other geotokens. In order to encode this geographical position we propose a three dimensional extension of RoPE mechanism on spherical coordinates.

2 Geotokens and Cartographical Transformer

Being able to encode geographical entities have tremendous potential, especially as we venture into an era where data is not just textual but spatial, and where the insights derived from such data can be transformative for a myriad of applications, ranging from urban planning and environmental monitoring to navigation and tourism. Though one can propose to represent the spatial encoding within natural language as in [11], the necessity to encode geographical coordinates with a more efficient and robust method is evident.

For this purpose, in this paper we conceptualize a regular transformer based architecture where the input isn't traditional textual tokens but "geotokens". A geotoken encapsulates both the semantic meaning and the spatial information of a geographical entity. These geotokens can represent anything from specific landmarks to broader regions or zones like a desert or an urban area. For the sake of simplicity only punctual locations are considered represented by a latitude and longitude. It is assumed that each data point has a pre-embedded vector retaining valuable information about the location itself, which may have been encoded by any type of neural architecture or mechanism, such as a natural language model processing its verbal description or a CNN extracting its visual features.

The term "geotoken" in this context refers to a tokenized representation of a geographical entity, which could range from specific landmarks and places to broader regions. Unlike standard tokens, which represent words or characters in a text, geotokens embody the spatial attributes, the semantics, and the context of geographical entities. It is straightforward to think that a transformer model processing these geotokens shall not encode sequential order as position but their geographical coordinates. Hence, a new mechanism for position encoding is necessary for such a setting.

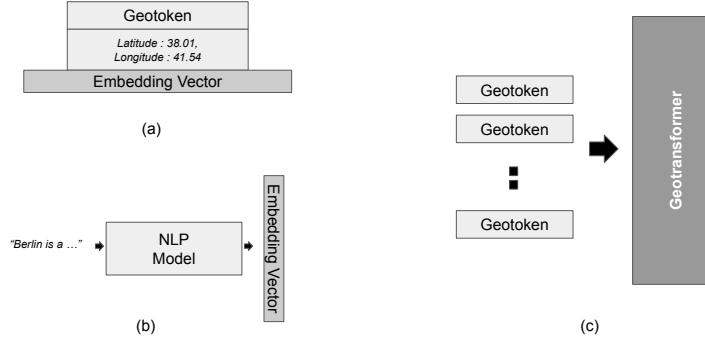


Figure 1: (a) A simple geotoken is defined with a position (latitude, longitude). (b) Its latent representative features may have been encoded with any kind of pre-trained neural model, in this case NLP processed textual description. (c) Geotransformer architecture processing geotokens.

3 Regular Position Encoding

As mentioned previously, proposed method to encode sequential positions in the original transformer architecture proposal has been proven to be quite plausible despite its few drawbacks [10]. Following the same notation in [10], let $\mathbb{S} = \{w_i\}_{i=1}^N$, for N input tokens with input embedding vectors of $\mathbb{E} = \{x_i\}_{i=1}^N$. The projected query, key, value vectors in self-attention are as follows respectively :

$$\begin{aligned} q_m &= f_q(x_m, m) \\ k_n &= f_k(x_m, n) \\ v_n &= f_v(x_m, n) \end{aligned} \quad (1)$$

m and n denoting the respective positions in vectors.

The authors of [10] propose to define a non-trainable additive matrices into input embeddings defined as :

$$\begin{aligned} p_{i,2t} &= \sin(k/10000^{2t,d}) \\ p_{i,2t+1} &= \cos(k/10000^{2t,d}) \end{aligned} \quad (2)$$

where sine and cosine functions of absolute token position is encoded in even and odd numbered indexes of positional vector of same size with the d dimensional input embeddings. The intuitive theory behind this is that through trigonometric identity functions the relative positional distances can be represented by linear algebraic multiplications of the encodings.

4 Rotary Position Embedding (RoPE)

[9] made an assertion regarding how the relative positions of two tokens should be modeled. [10] firstly formulate the position encoding as a function g applied on inner products of query and key vectors :

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n) \quad (3)$$

$m - n$ representing the relative positions. Via this type of a basis formulation authors prove the relative position can be encoded as a rotation matrix as :

$$f_{\{q,k\}}(x_m, m) = \mathbf{R}_{\Theta,m}^d \mathbf{W} x_m \quad (4)$$

$\mathbf{R}_{\Theta,m}^d$ being the proposed rotation matrix as :

$$\mathbf{R}_{\Theta,m}^d = \begin{bmatrix} \cos(m\theta_1) & -\sin(m\theta_1) & 0 & 0 & \dots & 0 & 0 \\ \sin(m\theta_1) & \cos(m\theta_1) & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos(m\theta_2) & -\sin(m\theta_2) & \dots & 0 & 0 \\ 0 & 0 & \sin(m\theta_2) & \cos(m\theta_2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos(m\theta_{d/2}) & -\sin(m\theta_{d/2}) \\ 0 & 0 & 0 & 0 & \dots & \sin(m\theta_{d/2}) & \cos(m\theta_{d/2}) \end{bmatrix} \quad (5)$$

where inner sections are taken from regular two dimensional matrix rotation. As it can be seen from the rotation matrix, the embedding space is bisected evenly and the positional index are encoded inside, somehow taking inspiration from the original absolute position encoding, θ representing the similar angle function :

$$\theta_i = 10000^{-(2i-1)/d} \quad (6)$$

5 Spherical Position Encoding

As mentioned previously for such a framework where geotokens needed to be position encoded according to their global coordinates we need a mechanism to handle the spherical space. For this purpose, we propose to extend the RoPE method in spherical coordinates. Let us define the longitude and latitude of any arbitrary position as θ and ϕ respectively. For the sake of simplicity, without loss of generalization and omitting the fractional errors let us assume that globe is a perfect sphere with constant radius R .

Three dimensional Euler angles can be used to define rotation matrix in this fixed coordinate system. The general form of the rotation matrix in this setting is as follows :

$$\begin{bmatrix} \cos(\psi)\cos(\theta) & -\cos(\phi)\sin(\theta) + \sin(\phi)\sin(\psi)\cos(\theta) & \sin(\phi)\sin(\theta) + \cos(\phi)\sin(\psi)\cos(\theta) \\ \cos(\psi)\sin(\theta) & \cos(\phi)\cos(\theta) + \sin(\phi)\sin(\psi)\sin(\theta) & -\sin(\phi)\cos(\theta) + \cos(\phi)\sin(\psi)\sin(\theta) \\ -\sin(\psi) & \sin(\phi)\cos(\psi) & \cos(\phi)\cos(\psi) \end{bmatrix} \quad (7)$$

ϕ, ψ, θ denoting rotation along x, y, z axes respectively. Assuming longitude and latitude variation defined as a rotation on a sphere along x and z axes respectively with setting the angles ϕ and θ , intuitively we need to keep the y -axis rotation constant by equating the angle $\psi = 0$.

Therefore the rotation matrix in this case can be written as :

$$\begin{bmatrix} \cos(\theta) & -\cos(\phi)\sin(\theta) & \sin(\phi)\sin(\theta) \\ \sin(\theta) & \cos(\phi)\cos(\theta) & -\sin(\phi)\cos(\theta) \\ 0 & \sin(\phi) & \cos(\phi) \end{bmatrix} \quad (8)$$

Basing on this particular rotation matrix above, taking inspiration from the RoPE architecture [10], we propose to encode the rotational position encoding matrix as follows, assuming a multiple of 3 :

$\mathbf{R}_{\Theta,m}^d$ being the proposed rotation matrix as :

$$\begin{bmatrix} \cos(\theta_1) & -\cos(\phi_1)\sin(\theta_1) & \sin(\phi_1)\sin(\theta_1) & 0 & 0 & 0 & \dots & 0 & 0 \\ \sin(\theta_1) & -\cos(\phi_1)\cos(\theta_1) & -\sin(\phi_1)\cos(\theta_1) & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sin(\phi_1) & \cos(\phi_1) & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \cos(\theta_2) & -\cos(\phi_2)\sin(\theta_2) & \sin(\phi_2)\sin(m\theta_2) & \dots & 0 & 0 \\ 0 & 0 & 0 & \sin(\theta_2) & -\cos(\phi_2)\cos(\theta_2) & -\sin(\phi_2)\cos(m\theta_2) & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \sin(\phi_2) & \cos(\phi_2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \cos(\theta_{d/3}) & -\cos(\phi_{d/3})\sin(\theta_{d/3}) & \sin(\phi_{d/3})\sin(\theta_{d/3}) \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \sin(\theta_{d/3}) & -\cos(\phi_{d/3})\cos(\theta_{d/3}) & -\sin(\phi_{d/3})\cos(\theta_{d/3}) \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & \sin(\phi_{d/3}) & \cos(\phi_{d/3}) \end{bmatrix} \quad (9)$$

where ϕ and θ correspond to longitude and latitude in radial values of a given geotoken. Note that we do not need to calculate an auxillary angle value as in original absolute position encoding or RoPE as geographical coordinates are inherently angular. For the sake of simplicity, the embedding dimension is a multiple of three due to natural requirements, however this choice might be inconvenient as many embedders of different modalities might not adhere to this constraint. The possible circumvention to this issue is out of scope of this paper, such as possibly adding padding indices. In addition, further possible challenges such as proper scaling are kept out of scope as well, where in case one training the architecture with limited geolocations, rather than whole globe.

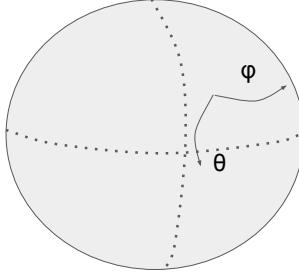


Figure 2: The repositioning on world can be formulated as a rotation along longitude and latitude axes, keeping one of Euler general angles constant.

6 Conclusion

In this work, we have presented a novel concept for the transformer architecture, integrating "geotokens" as a representation of geographical entities. This framework allows representation of any pointwise geographical location with their representative feature vectors, which might be encoded hypothetically with other pre-trained neural architectures

of various modalities. This integration is not just a semantic enhancement but also introduces a unique challenge – encoding the geographical coordinates rather than the traditional sequential positions. Recognizing the limitations of traditional position embeddings in this spatial context, we employed and extended the Rotary Position Embedding (RoPE) mechanism to accommodate spherical coordinates, aligning the transformer architecture to work seamlessly with geographical data.

References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)
- [2] Maurcio, J., Domingues, I. & Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*. **13**, 5521 (2023)
- [3] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. & Others Emergent abilities of large language models. *ArXiv Preprint arXiv:2206.07682*. (2022)
- [4] Brasoveanu, A. & Andonie, R. Visualizing transformers for nlp: a brief survey. *2020 24th International Conference Information Visualisation (IV)*. pp. 270-279 (2020)
- [5] Schmidt, R. Recurrent neural networks (rnns): A gentle introduction and overview. *ArXiv Preprint arXiv:1912.05911*. (2019)
- [6] Yun, C., Bhajanapalli, S., Rawat, A., Reddi, S. & Kumar, S. Are transformers universal approximators of sequence-to-sequence functions?. *ArXiv Preprint arXiv:1912.10077*. (2019)
- [7] Ke, G., He, D. & Liu, T. Rethinking positional encoding in language pre-training. *ArXiv Preprint arXiv:2006.15595*. (2020)
- [8] Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. *ArXiv Preprint arXiv:1803.02155*. (2018)
- [9] He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv Preprint arXiv:2006.03654*. (2020)
- [10] Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B. & Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *ArXiv Preprint arXiv:2104.09864*. (2021)
- [11] Unlu, E. Chatmap : Large Language Model Interaction with Cartographic Data. (2023)