

Linearity of Sinusoidal Positional Encodings

Tucker Berckmann

Brown University

1. Overview of Claim

Vazwani et. al.¹ claim that for positional encodings defined by:

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

the model can easily learn relative positions, since "for any fixed offset k , PE_{pos+k} can be represented as a linear function of PE_{pos} ."

2. Examination of Claim

Let us consider a model where $d_{model} = n$ and n is even. For clarity of the notation, we will treat the definition of the entire positional encoding vector $PE_{(pos)} \in \mathfrak{R}^n$ as follows:

$$PE_{(pos)} = \Pi(p) = \begin{bmatrix} \sin c_1 p \\ \cos c_1 p \\ \dots \\ \sin c_{\frac{n}{2}} p \\ \cos c_{\frac{n}{2}} p \end{bmatrix}$$

where p is the word position and $\Pi(p)$ is the encoding vector in that word position.

Given this definition, it is claimed that $\Pi(p+k)$ can be represented as a linear function of $\Pi(p)$. In other words:

- 1 There exists a vector-valued function T such that $T(\Pi(p)) = \Pi(p+k)$
- 2 T is linear

3. Proof of Claim

3.1. Definition of T

Given the definition of $\Pi(p)$, we know that:

$$\Pi(p+k) = \begin{bmatrix} \sin c_1(p+k) \\ \cos c_1(p+k) \\ \dots \\ \sin c_{\frac{n}{2}}(p+k) \\ \cos c_{\frac{n}{2}}(p+k) \end{bmatrix} = \begin{bmatrix} \sin c_1 p \cos c_1 k + \cos c_1 p \sin c_1 k \\ \cos c_1 p \cos c_1 k - \sin c_1 p \sin c_1 k \\ \dots \\ \sin c_{\frac{n}{2}} p \cos c_{\frac{n}{2}} k + \cos c_{\frac{n}{2}} p \sin c_{\frac{n}{2}} k \\ \cos c_{\frac{n}{2}} p \cos c_{\frac{n}{2}} k - \sin c_{\frac{n}{2}} p \sin c_{\frac{n}{2}} k \end{bmatrix}$$

by the identities:

$$\sin(A+B) = \sin A \cos B + \cos A \sin B$$

$$\cos(A+B) = \cos A \cos B - \sin A \sin B$$

If we then let $\Pi(p) = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_{n-1} & \pi_n \end{bmatrix}$ and if we also let $a_i = \sin c_i k$ and $b_i = \cos c_i k$, then:

$$\Pi(p+k) = T(\Pi(p)) = T \begin{pmatrix} \pi_1 \\ \pi_2 \\ \cdots \\ \pi_{n-1} \\ \pi_n \end{pmatrix} = \begin{bmatrix} b_1 \pi_1 + a_1 \pi_2 \\ -a_1 \pi_1 + b_1 \pi_2 \\ \cdots \\ b_{\frac{n}{2}} \pi_{n-1} + a_{\frac{n}{2}} \pi_n \\ -a_{\frac{n}{2}} \pi_{n-1} + b_{\frac{n}{2}} \pi_n \end{bmatrix}$$

3.2. Linearity of T

Since T corresponds to multiplying by the matrix $\mathbf{T}$, then

$$\Pi(p+k) = \mathbf{T} \cdot \Pi(p) = \begin{bmatrix} b_1 & a_1 & \cdots & 0 & 0 \\ -a_1 & b_1 & \cdots & 0 & 0 \\ \cdots & \cdots & & \cdots & \cdots \\ 0 & 0 & \cdots & b_{\frac{n}{2}} & a_{\frac{n}{2}} \\ 0 & 0 & \cdots & -a_{\frac{n}{2}} & b_{\frac{n}{2}} \end{bmatrix} \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \\ \cdots \\ \pi_{n-1} \\ \pi_n \end{bmatrix}$$

and T is a linear function, since any function represented as matrix multiplication is linear.² Thus the claim is valid.

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin., "Attention is all you need," *Advances in neural information processing systems*, pp. 5998-6008 (2017).
2. Charles Cullen, *Linear Algebra with Applications, 2nd Edition*, p. 235, Addison-Wesley, Reading, PA (1997).