

# 1 Transformers

## 1.1 Introduction

The Transformer is a deep learning architecture introduced in [Vaswani et al., 2017]. It is a sequence-to-sequence model that uses attention to learn the relationships between the input and output sequences.

In this report we will follow the notation that is used in [Vaswani et al., 2017] where embeddings are represented as rows  $\in \mathbb{R}^{1 \times d}$  and all matrix multiplications are right multiplications.

## 1.2 Embedding

A machine is not capable of understanding wordings, hence we need to transform it into a vector representation called an *embedding*. Let us denote the embedding of the  $i$ -th word in the input sequence as  $\mathbf{x}^{(i)} \in \mathbb{R}^{1 \times d}$ , where  $d$  is the feature dimension. The transformer is able to process these embeddings in **parallel** so we need a way to encode the position of the word in the sequence. We do this by adding a positional encoding  $\mathbf{p}^{(i)} \in \mathbb{R}^{1 \times d}$  to the embedding  $\mathbf{x}_i$ . The positional encoding is a vector that is unique to the position of the word in the sequence. The positional encoding that is used in [Vaswani et al., 2017] is given by:

$$\mathbf{p}_j^{(i)} = \begin{cases} \sin\left(\frac{j}{10000^{2i/d}}\right) & \text{if } i \text{ is even} \\ \cos\left(\frac{j}{10000^{2i/d}}\right) & \text{if } i \text{ is odd} \end{cases} \quad (1)$$

where  $j$  is the dimension of the positional encoding. The positional encoding is added to the embedding as follows:

$$\mathbf{x}_i = \mathbf{x}_i + \mathbf{p}^{(i)} \in \mathbb{R}^{1 \times d} \quad (2)$$

These are all stacked together to form the input matrix  $\mathbf{X} \in \mathbb{R}^{N \times d}$ , where  $\mathbf{X}_i = \mathbf{x}^{(i)}$  where  $N$  is the number of words in the input sequence.

### Alternative Positional Encoding

There are many ways to go about positional encoding. Another way is to use a learned positional encoding. This is done by adding a learnable vector  $\mathbf{p}^{(i)} \in \mathbb{R}^{1 \times d}$  to the embedding. Alternatively to achieve translation equivariance, we can use *Relative Positional Encoding* [Shaw et al., 2018, Wu et al., 2021]. These positional encodings schemes will be useful when trying to build in equivariance into the Transformer model.

## 1.3 Attention

The attention mechanism is a way to learn the relationships between the input and output sequences.

## References

- [Shaw et al., 2018] Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- [Wu et al., 2021] Wu, K., Peng, H., Chen, M., Fu, J., and Chao, H. (2021). Rethinking and improving relative position encoding for vision transformer.