Unifying Transformers and Convolutional Neural Processes

Lakee Sivaraya (ls914)

June 14, 2023

Emmanuel College

Contents

1 Transformers

1.1 Introduction

The Transformer is a deep learning architecture introduced in [?]. 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 [?] 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 understand 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 $\boldsymbol{x}_i \in \mathbb{R}^{1\times d}$, 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 $\boldsymbol{p}^{(i)} \in \mathbb{R}^{1\times d}$ to the embedding $\boldsymbol{x}_i$. 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 [?] is given by:

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

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

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

These are all stacked together to form the input matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, where $\boldsymbol{X}_{i:} = \boldsymbol{x_i}$ 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 $\boldsymbol{p}^{(i)} \in \mathbb{R}^{1\times d}$ to the embedding. Alternatively to achieve translation equivariance, we can use *Relative Positional Encoding* [?, ?]

1.3 Attention

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

2 Neural Processes

This is an introduction. We will use the notation from [?]. As a reminder, we have the following notation:

2.1 Notation

References

[Garnelo et al., 2018] Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M. A., and Teh, Y. W. (2018). Neural processes.

[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.