

Advanced Data Analysis Homework 8

Information Science and Technology
Mathematical Informatics, M2
48226203 Kota Isayama *

July 31, 2023

Homework 1

We implement scaled dot-product attention with multi-headed version. The source code is here¹.

Homework 2

We reference the paper that proposed[1] the transformer and explain the positional encoding scheme used in this paper.

In this original paper, they add the positional encodings to the input embeddings at the bottom of the encoder and decoder stacks. They use sin and cos functions to define the positional encodings as follows:

$$\begin{aligned} \text{PE}_{(pos, 2i)} &= \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \\ \text{PE}_{(pos, 2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \end{aligned}$$

where pos is the position and i is the dimension. They chose this scheme because they hypothesized that the model can deal with relative positions more easily since $\text{PE}_{(pos+k, 2i)}$ can be represented as a linear function of $\text{PE}_{(pos, \cdot)}$, i.e. addition theorem of trigonometric functions.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

*isakou2208@g.ecc.u-tokyo.ac.jp

¹<https://github.com/Kota-Isayama/Advanced-Data-Analysis/blob/main/8/homework1.py>