Notations and Definitions

The architectures used in the experiments of this thesis are based on the Transformer architecture developed by Vaswani et al. [1]. For the Transformer blocks, we use the same same structure for both image and text. As mentioned in (TODO: cite data preparation), text is tokenized into subwords using the GPT-2 byte-pair encoder also used in Data2Vec [2], [3]. Before being passed into the Transformer, a start-of-sequence token [`T_CLS`] is added to the beginning of the sequence, and an end-of-sequence token [`T_SEP`] is added to the end of the sequence. Then, the sequence is embedded into 768-dimensional vectors, and a positional encoding is added to the embeddings. In this thesis, we define a text sequence as follows:

$$\boldsymbol{H}^{s}_{w,l} = \left[\boldsymbol{h}^{s}_{w,l,[\texttt{T\_CLS}]}, \boldsymbol{h}^{s}_{w,l,1}, ..., \boldsymbol{h}^{s}_{w,l,M}, \boldsymbol{h}^{s}_{w,l,[\texttt{T\_SEP}]}\right] \tag{1}$$

Because we use KD in some parts, representations will be superscripted with s or t, for a student and teacher representation, respectively.

Bibliography

[1] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[2] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv abs/2202.03555*, 2022.

[3] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language." 2022.