Notations and Definitions

The architectures used in the experiments of this thesis are based on the Transformer [1], and vision Transformer [2] architecture. Therefore, both image and text are represented as sequences of embeddings, which are processed by the Transformer blocks.

Image Representation

We define an image as a 3-dimensional tensor $\boldsymbol{v} \in \mathbb{R}^{C \times H \times W}$. Because we will use the base variant of the vision Transformer, ViT-B/16 [2], the image is patchified into 14x14 patches, each being a square of size 16x16 pixels. Each image patch represents one timestep in the sequence, and the number of patches N is given by $N = H \times \frac{W}{P^2}$, with P being the number of patches per dimension, and $P = 14$. Since we use an image size of 224x224 pixels, so $\boldsymbol{v} \in \mathbb{R}^{3 \times 244 \times 244}$, we will have $N = 244 \times \frac{244}{14^2} = 196$ patches, or timesteps respectively. Each patch is flattened into a 256-dimensional vector, and then projected into a 768-dimensional vector $\boldsymbol{e}_i^v \in \mathbb{R}^{768}$, using a fully connected layer. The image sequence is prepended with a special 768-dimensional learnable $[\texttt{I\_CLS}] \in \mathbb{R}^{768}$ token, which is used to aggregate the global information/content of the image, and, following [2], also 768-dimensional. The result is a sequence of patch embeddings, which we define as $\boldsymbol{E}_v$, where v indicates an image:

$$\boldsymbol{E}_v = \left[\boldsymbol{e}^v_{[\texttt{I\_CLS}]}, \boldsymbol{e}^v_1, \boldsymbol{e}^v_2, ..., \boldsymbol{e}^v_N\right] \tag{1}$$

To give the Transformer a sense of order in the image patches/timestep, a unique positional encoding is added to each patch embedding. This can either be learned or fixed, with the latter being for example a sinusoidal positional encoding [1]. This positional encoding is also represented as a sequence of 768-dimensional vectors:

$$\boldsymbol{T}^{\text{pos}}_v = \left[0, \boldsymbol{t}^v_{\text{pos}_1}, \boldsymbol{t}^v_{\text{pos}_2}, ..., \boldsymbol{t}^v_{\text{pos}_N}\right] \tag{2}$$

Since the $[\texttt{I\_CLS}]$ token is not part of the image, the positional encoding for the $[\texttt{I\_CLS}]$ token is set to zero, so nothing is added to it. An image representation is defined as:

$$\boldsymbol{H}^s_{v,l} = \left[\boldsymbol{h}^s_{v,l,[\texttt{I\_CLS}]}, \boldsymbol{h}^s_{v,l,1}, ..., \boldsymbol{h}^s_{v,l,N}\right] \tag{3}$$

In Equation 3, l denotes the layer of the Transformer block that returned the image representation, and v indicates that the representation is an image. Since we use Knowledge Distillation (KD) in some parts of this thesis, representations will be, if neccessary, superscripted with s or t, for a student and teacher representation, respectively.

We define $l = 0$ as the input to the Transformer, and $l = L$ as the output of the Transformer, where L is the number of layers in the Transformer. Consequently, the image input to the Transformer is defined as:

$$\boldsymbol{H}^s_{v,0} = \left[\boldsymbol{h}^s_{v,0,[\texttt{I\_CLS}]}, \boldsymbol{h}^s_{v,0,1}, ..., \boldsymbol{h}^s_{v,0,N}\right] = \boldsymbol{E}_v + \boldsymbol{T}^{\text{pos}}_v \tag{4}$$

The output of the Transformer is defined as:

$$\boldsymbol{H}^s_{v,L} = \left[\boldsymbol{h}^s_{v,L,[\texttt{I\_CLS}]}, \boldsymbol{h}^s_{v,L,1}, ..., \boldsymbol{h}^s_{v,L,N}\right] \tag{5}$$

Text Representation

We define a text as a sequence of discrete tokens, which are, similiar to image patches, embedded into 768-dimensional vectors, using an embedding matrix. A single token i is represented as $\boldsymbol{e}^t_i \in \mathbb{R}^{768}$, and the sequence of tokens, representing the text, is prepended with a start-of-sequence token $[\texttt{T\_CLS}] \in \mathbb{R}^{768}$, and appended with an end-of-sequence token $[\texttt{T\_SEP}] \in \mathbb{R}^{768}$. The purpose of the $[\texttt{T\_CLS}]$ token is, as with $[\texttt{I\_CLS}]$, to aggregate the global information/content of the text. The $[\texttt{T\_SEP}]$

token is used to indicate the end of the text sequence. A text sequence consists of M tokens, and we use w to denote a text sequence:

$$\boldsymbol{E}_w = \left[\boldsymbol{e}^w_{[\texttt{T\_CLS}]}, \boldsymbol{e}^w_1, \boldsymbol{e}^w_2, ..., \boldsymbol{e}^w_M, \boldsymbol{e}^w_{[\texttt{T\_SEP}]}\right] \tag{6}$$

As with the image, a positional encoding is added to the text embeddings, to give the Transformer a sense of order in the text sequence. Since the special token [T_SEP] denotes the end of the text sequence, it is part of the sequence, and therefore has a positional encoding. The latter does not hold for the [T_CLS] token, as it is used to aggregate the global information/content of the text.

$$\boldsymbol{T}^{\text{pos}}_w = \left[0, \boldsymbol{t}^w_{\text{pos}_1}, \boldsymbol{t}^w_{\text{pos}_2}, ..., \boldsymbol{t}^w_{\text{pos}_M}, \boldsymbol{t}^w_{\text{pos}_{[\texttt{T\_SEP}]}}\right] \tag{7}$$

A text representation is defined as:

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 [3], [4]. 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:

$$\boldsymbol{H}^s_{w,l} = \left[\boldsymbol{h}^s_{w,l,[\texttt{T\_CLS}]}, \boldsymbol{h}^s_{w,l,1}, ..., \boldsymbol{h}^s_{w,l,M}, \boldsymbol{h}^s_{w,l,[\texttt{T\_SEP}]}\right] \tag{8}$$

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. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR*, 2021.

[3] 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.

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