Cross-Modal Late Interaction (CMLI)

- currently [T_CLS] and [I_CLS] are used for global text and image features respectively in contrastive learning
- has the disadvantage that only global information are used for contrastive learning, and fine-grained, timestep specific, information is not considered
- can be a problem, if the real-world concepts described by image and text differ in small, yet important, details
- makes it difficult for the model to differentiate between similar concepts
- to address this, authors of FILIP introduces Cross-Modal Late Interaction (CMLI) for fine-grained interaction between text and image in contrastive learning [1]
- as shown in Figure 1, there is no cosine similarity between [T_CLS] and [I_CLS] computed, but instead the cosine similarity between all image patches $\left[\boldsymbol{v}_l^k\right]_{1 \leq k \leq N}$ and text tokens $\left[\boldsymbol{w}_l^j\right]_{1 \leq j \leq M}$
- other special tokens such as the end-of-sequence token [EOS] and padding token [PAD] are also excluded, as they do not carry any semantic information, so cosine similarity is only computed between the actual text tokens and image patches
- for each image patch $\left[\boldsymbol{v}_l^k\right]_{1 \leq k \leq N}$ we now have the cosine similarity with all text tokens $\left[\boldsymbol{w}_l^j\right]_{1 \leq j \leq M}$, and vice versa
- for an image patch k, we now get the text token with the maximum cosine similarity to this image patch

$$m_k^{\mathrm{i2t}} = \underset{1 \leq j \leq M}{\operatorname{argmax}} \left[\boldsymbol{v}^k\right] \left[\boldsymbol{w}^j\right]^T \tag{1}$$

and for each text token j, we get the image patch with the maximum cosine similarity to this text token

$$m_j^{\mathrm{t2i}} = \underset{1 \leq k \leq N}{\operatorname{argmax}} \left[\boldsymbol{v}^k\right] \left[\boldsymbol{w}^j\right]^T \tag{2}$$

- the result can be seen in (2) of Figure 1
- with this approach, we achive an association between individual image patches and text tokens, which allows the model to find, fine-grained, matching patterns
- the actual similarity between an image and text is then the average of the maximum cosine similarity between the associated tokens, which can be used for image-text contrastive learning and image-text retrieval

$$s_{v,w}^{\mathrm{i2t}} = \frac{1}{N} \sum_{k=1}^{N} \left[\boldsymbol{v}_l^k\right] \left[\boldsymbol{w}_l^{m_k^{\mathrm{i2t}}}\right]^T \tag{3}$$

$$s_{v,w}^{\mathrm{t2i}} = \frac{1}{M} \sum_{j=1}^{M} \left[\boldsymbol{v}_l^{m_j^{\mathrm{t2i}}}\right] \left[\boldsymbol{w}_l^j\right]^T \tag{4}$$

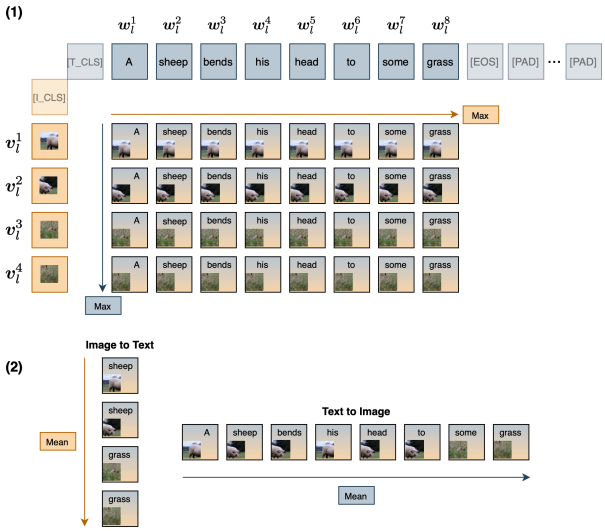

Figure 1: For a token/patch, CMLI finds the semantic timestep with the highest match from the other modality. This enables the model to associate small details of image and text with each other. Notice how through the max-operation patches containing grass are always associated with the word "grass", and the words "sheep" and "head" are matched with the head of the sheep (associations created through max are shown in (2)). The cosine similarity is then the average of all associations between an image-text pair. Figure inspired and adapted from [1].

- fine-grained alignment offers the opportunity to test image-language reasoning, an application non-referecing model previously were deemed unsuited for
- we identify the option to combine CMLI with vanilla ITC, and test the mean of both as a similarity measure

Bibliography

[1] L. Yao *et al.*, "FILIP: Fine-grained Interactive Language-Image Pre-Training," *CoRR*, 2021.