Figures and Visualizations

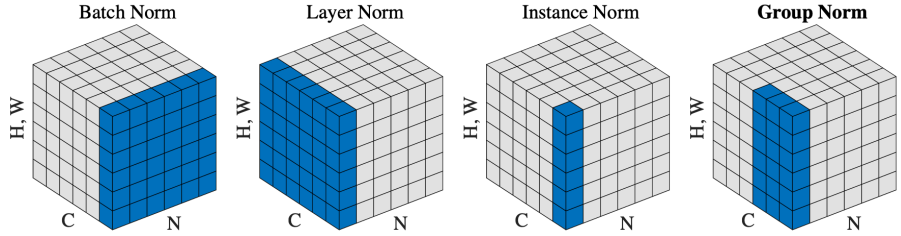

Figure 1: Comparison of different normalization operations on the example of images. Dimension "H, W" refers to the height and width of the input, "C" to the number of channels or embedding dimensions, and "N" to the number of samples, i.e. the batch dimension. Since we are working with sequences of embeddings, the height and width dimension correspond to the time steps ("H, W" -> "T"). The normalization operations work correspondingly on text sequences, where we also have time steps, so dimension "H, W" can also be replaced by "T" [1]. Please note that group norm, even though it is displayed in bold , is not used in this work. The figure merely was introduced in the paper of Group Norm [1].

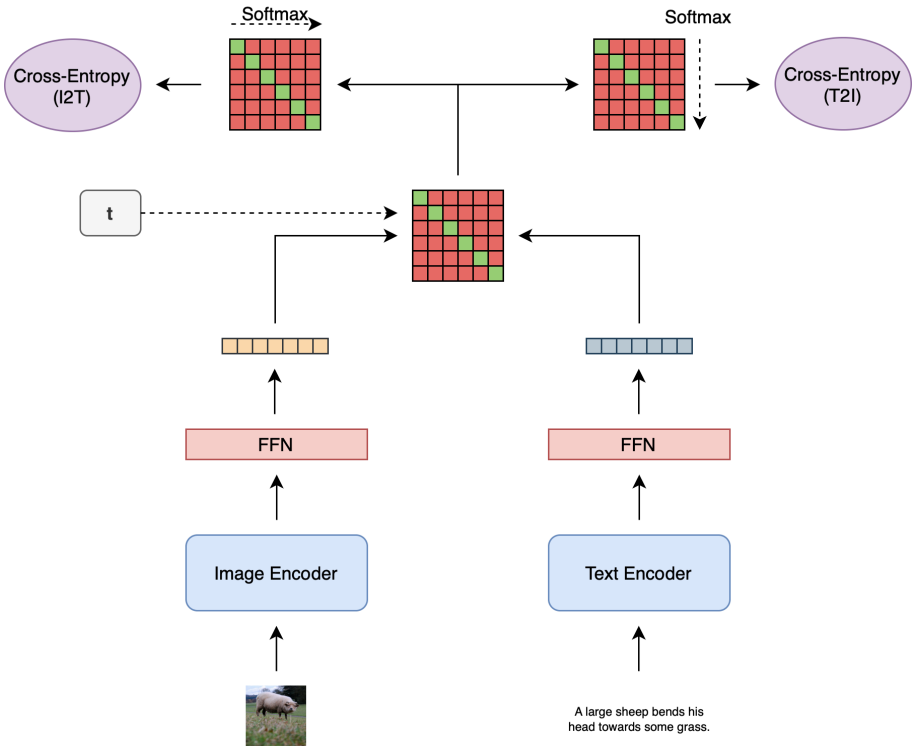

Figure 2: Illustration of CLIP training. A batch of image-text pairs is passed through the model and embedded into a shared latent space. The cosine similarity between all pairs is computed and softmax-normalized to calculate the image-to-text and text-to-image loss. The final loss is the mean of both losses [2]. The example is shown with a batch size of 6. The figure does not originate from the original paper, but is a custom visualization of the concept. Image-Text pair is taken from the MSCOCO train set [3], and do not refer to the contrastive loss of 6 pairs at the top of the figure. They are merely indicators of the intput to the model.

Bibliography

[1] Y. Wu and K. He, "Group Normalization," in *Computer Vision -– ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, Munich, Germany: Springer-Verlag, 2018, pp. 3–19. doi: 10.1007/978-3-030-01261-8_1.

[2] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[3] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.