

Figures and Visualizations

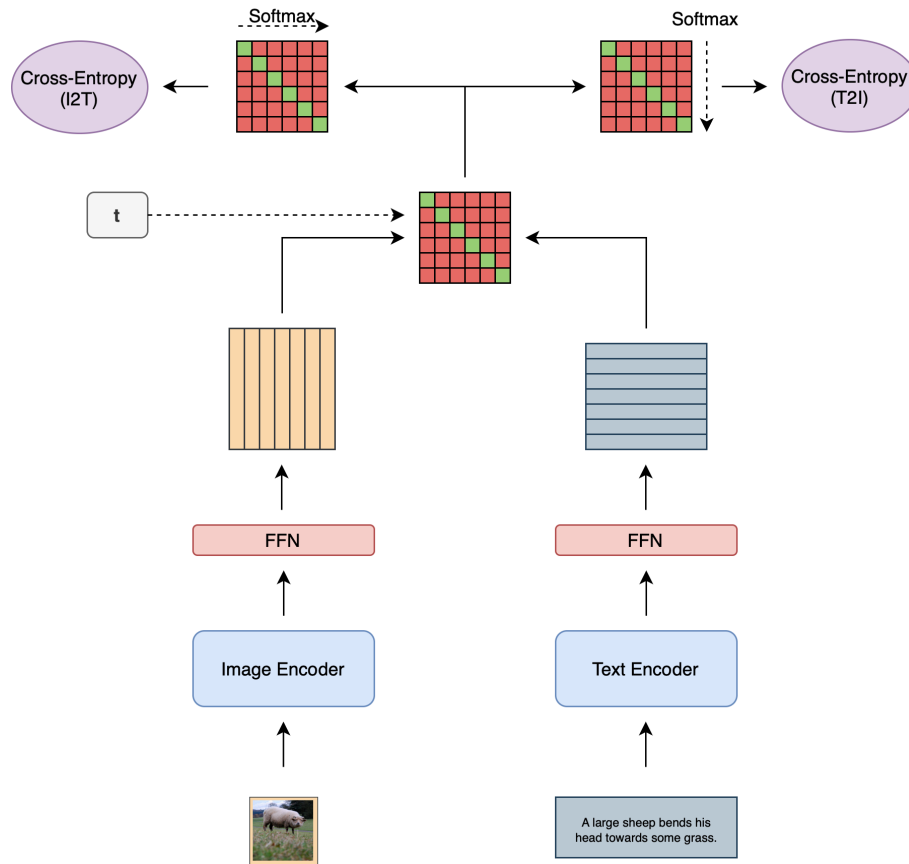


Figure 1: Illustration of CLIP training. A batch of image-text pairs is passed through the model and embedded into a shared latent space. The cosine similarity between all pairs is computed and softmax-normalized to calculate the image-to-text and text-to-image loss. The final loss is the mean of both losses [1]. 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 [2], and do not refer to the contrastive loss of 6 pairs at the top of the figure. They are merely indicators of the input to the model.

Bibliography

- [1] 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.
- [2] 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.