Image-Text Retrieval

The goal of image-text retrieval (ITR) is to find the matching (most similar) caption for a given image, and vice versa. The process begins with embedding and normalizing a set of samples, such as images or captions, which become a set of keys. For some candidate image or text, called the query, the most similar key is retrieved, after the query is also embedded and normalized. Similar to contrastive learning, cosine similarity is used to compute the similarity between the query and all keys, which is, again, computed by matrix multiplication of the normalized embeddings. The similarities between the query and keys are then ranked, and the key with the highest similarity to the query is the retrieved sample. This method can be viewed as a form of semantic search, which has significant practical relevance in areas like recommendation systems, e.g. to find images based on a given text query. This is precisely what is learned through multimodal contrastive learning.

Image-Text Retrieval is a cheap and efficient way to benchmark the quality of the learned representations of a vision-language model, as it does not require any finetuning, just the embeddings produced by the model. The metric used for benchmarking is Rank@K (R@K), where K determines at which rank the paired/correct sample has to be in the ranking in order to be considered as a correct retrieval. We use R@1, R@5, and R@10, where R@1 is the normal accuracy, i.e., the paired sample has to be the most similar one. R@5 means that the paired sample has to be in the top 5 most similar samples, and for R@10, it has to be in the top 10 most similar samples, in order for the retrieval to be considered correct.

In this thesis, we use the 5K test set of MSCOCO [1], and the 1K test set of Flickr30k [2] for benchmarking, which is the standard benchmarking dataset for multimodal models like FLAVA [3], CLIP [4], VLMo [5], and BEiT-3 [6]. MSCOCO contains 5K images with 5 captions each [1], and Flickr30k contains 1K images with 5 captions each [2]. For both datasets, the all images and all texts are embedded and normalized, so that each image and each text is represented by the cls token that was returned by the model. Then, matrix multiplication between the images and captions of a dataset is performed, resulting in a matrix of shape (N, M), where N is the number of images and M is the number of captions in the dataset. So for MSCOCO, the matrix is of shape (5K, 25K), and for Flickr30k, the matrix is of shape (1K, 5K).

Bibliography

[1] 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.

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[3] A. Singh *et al.*, “FLAVA: A foundational language and vision alignment model,” *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2112.04482

[4] 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.

[5] H. Bao *et al.*, “VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts,” in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://openreview.net/forum?id=bydKs84JEyw

[6] W. Wang *et al.*, "Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19175–19186. doi: 10.1109/CVPR52729.2023.01838.