

Image-Text Retrieval

The goal of image-text retrieval (ITR) is to find the matching caption for a given image in a set of captions, and likewise, finding the matching image for a given caption in a set of images. The process begins with embedding and normalizing a set of samples, which become a set of keys. For some normalized candidate representation, called the query, the most similar key is retrieved among the set of keys is the retrieved sample. This is exactly what is learned through contrastive learning, where we try to maximize the similarity between an image or caption (query) and its paired caption or image among other samples (keys), respectively. For that, we can use the same batch-wise computation introduced in the previous section about the contrastive loss. The similarity is computed by the cosine similarity, which is, again, computed by matrix multiplication of the normalized embeddings.

Image-Text retrieval can be viewed as a form of semantic search, which has significant practical relevance in areas like recommendation systems, e.g. to find fitting images based on a given text query. This is precisely what is learned through multimodal contrastive learning.

Image-Text retrieval is a simple 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 of keys, in order for the retrieval to be considered correct. 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 this thesis, we use the 5K test set of MSCOCO [1], and the 1K test set of Flickr30k [2] for benchmarking, which are 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 for each image [1], and Flickr30k contains 1K images with 5 captions each [2]. For both datasets, all images and all texts are embedded and normalized, so that each image and each text is represented by the respective [CLS] token returned by the model. Then, matrix multiplication between all images and all 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).

For each image, $R@1$, $R@5$, and $R@10$ are computed. The mean of $R@1$, $R@5$, and $R@10$ over all images are then called text-retrieval of the respective metrics (e.g. $R@1$ -text-retrieval). We call this text-retrieval, because we are trying to retrieve the correct caption for a given image. The same is done for each caption, resulting in image-retrieval of the respective metrics (e.g. $R@1$ -image-retrieval). For each dataset, we have 6 metrics in total: $R@1$, $R@5$, and $R@10$ for text-retrieval and image-retrieval, respectively. We will report the results of image-text retrieval in the format seen in Table 1.

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image \rightarrow Text			Text \rightarrow Image			Image \rightarrow Text			Text \rightarrow Image		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
FLAVA [3]	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP [4]	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
BEiT-3 [6]	84.8	96.5	98.3	67.2	87.7	92.8	98.0	100.0	100.0	90.3	98.7	99.5

Table 1: Benchmarks of different vision-language models on the MSCOCO and Flickr30K datasets for image-text retrieval.

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.