Ablation Study: Removing ITC

In the previous chapters, we made efforts to improve and extend the architecture and training of the model to achieve better alignment of the modalities, which should go hand in hand with better retrieval performance. However, we had severe difficulties with the length of captions (TODO: cite), increasing the number of negative examples (TODO: cite), and the granularity of alignment (TODO: cite). This shows that the approach is very much dependent on high-quality data, and the right hardware, which is why we identify Image-Text Contrastive Learning (ITC) as a weak point of the approach. The best approach would therfore be to not use contrastive learning at all, which is why we will investigate the effects of an absence of ITC in this ablation study.

At the point of writing, the state-of-the-art (SOTA) vision-language model, BEiT-3, gives us a good reason to discard ITC. BEiT-3 pre-training is performed without contrastive learning, and the authors report SOTA results after fine-tuning on retrieval tasks MSCOCO [1] and Flickr30K [2]. Even without fine-tuning, BEiT-3 achives competitive results on Flickr30K, and even outperforms models that where trained using contrastive learning (see Figure 1).

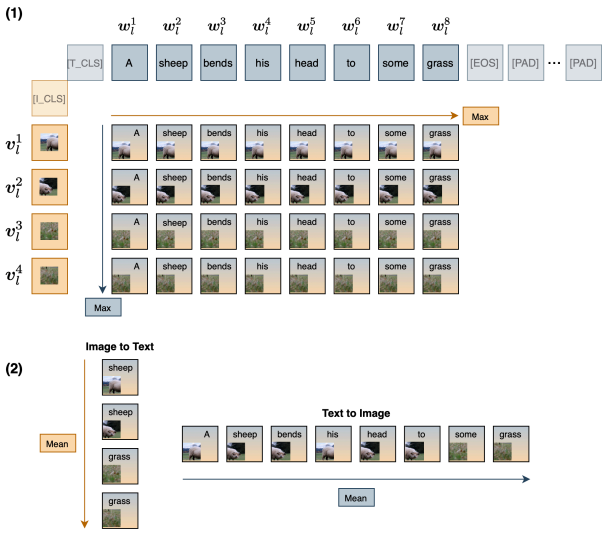

Figure 1:

- retrieval application then truly becomes zero-shot, and authors report high performance
- [3] mentions that without ITC, model outputs are not aligned, and therefore not suited for retrieval
- ablation study in (TODO cite) shows: TODO
- for current approach representations will still be aligned, because we are regressing the [I_CLS] token of BEiT-2, not a probability distribution over the classes
- so we are regressing actual features -> loss of 0 would mean representations are exactly the same and therefore aligned
- [3] does prediction of teacher outputs on the level of classes, which are not feature/representation-based
- so it is true that in the supervised case, when regressing a probability distribution the final representation, i.e. the [I_CLS] token, does not need to be aligned, only the probability distribution
- but again, we are explicitly regressing [I_CLS], so representations are aligned

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → 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 [4]	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP [5]	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
Sx3HRe$_{\text{ITC}}$	41.36	71.16	82.0	30.2	59.46	72.54	9.5	35.68	50.18	8.38	37.54	49.88
Zero-Shot												
Sx3HRe$_{\text{-ITC}}$	33.52	59.34	70.14	11.26	29.12	40.40	8.86	29.88	41.98	4.45	18.42	26.82

Table 1:

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] Y. Aytar, C. Vondrick, and A. Torralba, "See, Hear, and Read: Deep Aligned Representations," *arXiv preprint arXiv:1706.00932*, 2017, [Online]. Available: https://arxiv.org/abs/1706.00932

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

[5] 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.