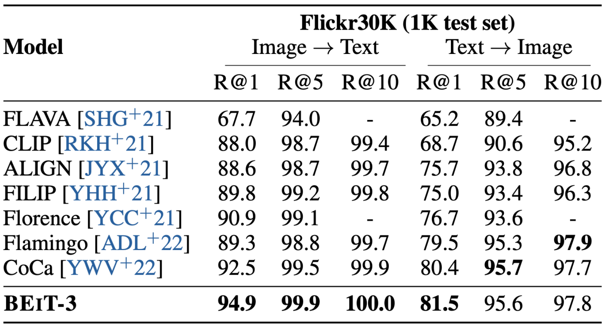
Ablation Study: Removing ITC

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 observed that contrastive learning is highly dependent on the granularity of the data (TODO: cite), e.g. the quality of captions, and the number of negative examples (TODO: cite). While there are viable options to overcome the latter, they are not always feasible, nor efficient, and the right hardware is required. Therefore, we identify Image-Text Contrastive Learning (ITC) as a weak point of our approach. The best approach would 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.

As an intersting side note, 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 [1] after fine-tuning on retrieval tasks MSCOCO [2] and Flickr30K [3]. Even without fine-tuning, BEiT-3 achives competitive results on Flickr30K, and even outperforms models trained using contrastive learning (see Figure 1).



Model	**Flickr30K (1K test set)**					
	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA [SHG+21]	67.7	94.0	-	65.2	89.4	-
CLIP [RKH+21]	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [JYX+21]	88.6	98.7	99.7	75.7	93.8	96.8
FILIP [YHH+21]	89.8	99.2	99.8	75.0	93.4	96.3
Florence [YCC+21]	90.9	99.1	-	76.7	93.6	-
Flamingo [ADL+22]	89.3	98.8	99.7	79.5	95.3	**97.9**
CoCa [YWV+22]	92.5	99.5	99.9	80.4	**95.7**	97.7
BEiT-3	**94.9**	**99.9**	**100.0**	**81.5**	95.6	97.8

Figure 1: Excerpt from the BEiT-3 paper [1], showing that BEiT-3 outperforms models trained with contrastive learning on the Flickr30K dataset without fine-tuning. As the results are based on the pre-trained model without fine-tuning, and the model has not been pre-trained with contrastive learning, nor was the Flickr30K test set used during pre-training, the retrieval application becomes **truly zero-shot**. However, it is important to note that the performance reported in the paper are based on the vision Transformer giant (ViT-G/14) architecture, which a patch size of 14x14 [1], [4], and naturally leads to better performance than smaller architectures.

Since we are removing ITC, we can also remove the additional projection introduced in (TODO: cite), which was added to seperate the regression of the [I_CLS] token of BEiT-2 (our teacher model) from the contrastive learning, and lead to superior performance. Consequently, the output cls token of our shared Transformer block will be used for the regression of the [I_CLS] token of BEiT-2 directly.

The advantage a feature-based Knowledge Distillation approach has in this case, is that even if we do not use contrastive learning, the representations will still be, to some extend, aligned. This is because we are regressing the [I_CLS] token of BEiT-2, not a probability distribution over the classes, as in SHRe [5], described in (TODO: cite). Therefore, if the model would reach a loss (MSE) of 0, then the representation between the cls token output for a caption by the student model $[\texttt{T\_CLS}]_s$, and the $[\texttt{I\_CLS}]_t$ token of BEiT-2 for the fitting image would be exactly the same. If the same holds for the cls token output for the same image by the student model $[\texttt{I\_CLS}]_s$, then $[\texttt{T\_CLS}]_s$ and $[\texttt{I\_CLS}]_t$ would be same, and aligned.

$$\begin{gathered} \mathrm{MSE}([\texttt{T\_CLS}]_s, [\texttt{I\_CLS}]_t) = 0 \land \mathrm{MSE}([\texttt{I\_CLS}]_s, [\texttt{I\_CLS}]_t) = 0 \\ \implies [\texttt{T\_CLS}]_s = [\texttt{I\_CLS}]_t \end{gathered} \tag{1}$$

The aforementioned does not hold in response-based Knowledge Distillation, as done in SHRe [5], since the probability distribution over classes is regressed. This ensures alignment on the level of categories, but not on the level of features: The cls token output would not need to be the same for an image-text pair, as long as the probability distribution over the classes is the same [5].

It follows from Equation 1, that we will use $[\texttt{T\_CLS}]_{\mathrm{s}}$ and $[\texttt{I\_CLS}]_{\mathrm{s}}$ for the retrieval tasks, we will use the cosine similarity between those as the similarity measure (as done before).

When we first reproduced the architecure of "See, Hear, and Read: Deep Aligned Representations" (SHRe) in a supervised setting (TODO: cite), we directly used ITC as defined in e.g. VLMo [6], instead of using the alignment approach actually used in SHRe. The alignment approach of SHRe is based on a pairwise cosine similarity between an image and text. The goal is to maximize the cosine similarity between a matching image-text pair, and minimize the cosine similarity between a non-matching image-text pair. What differentiates it from ITC, is that we are only intersted in the cosine similiarity between one image and one text, meaning the score/similarity is not softmax-normalized over the cosine similarity between the same image and a set of negative (non-matching) texts. This makes the approach independent of the number of negative examples: There is only ever one image and one text, which are compared. Following the cosine embedding loss of pytorch [7], we define the loss as:

$$\operatorname{loss}(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \operatorname{margin}), & \text{if } y = -1 \end{cases}$$

Figure 2: [7]

Where $y = 1$ denotes a matching pair, and $y = -1$ denotes a non-matching pair. Finding positive ($y = 1$) pairs is easy, as we can use the image-text pairs from the dataset. However, finding negative ($y = -1$) pairs is not as straight forward, and we need to find a suitable strategy. The best approach would be to use hard-negative mining, as done in VLMo for Image-Text Matching (ITM) [6], but VLMo selects hard-negatives using the cosine similarities from their ITC approach, which we do not have [6]. Since we do not have any labels, there is no other way to find negative pairs than to just randomly select them. Therefore, we resort to finding a random negative text j for each image i in the current batch, with $i \neq j$. Apart from this being the only option, it also has the advantage of being simple and fast. In total, we have N positive pairs and N negative pairs per batch, with N being the batch size. The total alignment loss is defined in Equation 2.

$$\begin{gathered} \mathcal{L}_{\mathrm{cos}}^{\mathrm{Sx3HRe}} = \frac{1}{2} * \left(\mathcal{L}_{\mathrm{cos}}([\texttt{T\_CLS}]_{\mathrm{s}}, [\texttt{I\_CLS}]_{\mathrm{s}}) + \mathcal{L}_{\mathrm{cos}}([\texttt{T\_CLS}]_{\mathrm{s}}, [\texttt{I\_CLS}]_{\mathrm{s}})\right), \\ j \sim \operatorname{Uniform}(\{1, ..., N\} \setminus \{i\}) \end{gathered} \tag{2}$$

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 [8]	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP [9]	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$_{\mathrm{ITC}}$	41.36	71.16	82.0	30.2	59.46	72.54	9.5	35.68	50.18	8.38	37.54	49.88
Sx3HRe$_{\mathrm{\sim ITC}}$	33.52	59.34	70.14	11.26	29.12	40.40	8.86	29.88	41.98	4.45	18.42	26.82
Sx3HRe$_{\mathrm{COS}}$	35.48	60.74	71.44	12.46	31.74	43.71	9.44	31.18	43.50	4.8	20.98	30.08

Table 1: ~ITC is full zero-shot on Flickr30K, and task zero-shot on MSCOCO. COS is task zero-shot on both datasets.

Bibliography

[1] 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.

[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.

[3] 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.

[4] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, “Scaling Vision Transformers,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1204–1213. doi: 10.1109/CVPR52688.2022.01179.

[5] 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

[6] 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

[7] PyTorch Contributors, “CosineEmbeddingLoss.” 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html

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

[9] 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.