Multimodal Knowledge Distillation

Seperate Self-Attention

Baseline

- currently only 6 layers, 5 out of which are modality specific, 1 is shared
- we experiment with adding one additional moadality specific layer, and one additional shared layer in another experiment

-> more difficult to align mutliple modalities, than just training one -> add one layer -> motivation for modality specific: after 5 layers information might not be high level enough so that one layer can process the information -> add one additional modality specific -> motivation for shared: after 5 layers information might be high level enough, but capturing modality agnostic information might take more than one layer -> add one additional shared

- added shared layer improves performance slightly, but adds 7 million parameters and 41 minutes to training time
- looking at the improvement in zero-shot, which increases the average Recall from 29.93% to 30.8%, this is not much of an improvement, considering the amount of parameters we add to the model

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	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
Data2Vec2	0.02	0.08	0.22	0.01	0.10	0.19	0.02	0.12	0.26	0.02	0.06	0.12
MM-D2V2 (Ours)	4.24	12.12	17.96	1.77	6.54	10.91	1.2	4.88	8.18	0.54	2.52	4.58
MM-D2V2 (Ours)†	31.72	56.78	67.9	12.42	31.05	42.5	7.7	26.18	37.6	4.08	17.01	24.26
MM-D2V2 7_2(Ours)†	32.78	58.34	69.3	12.83	31.85	43.4	8.08	27.92	38.6	4.14	17.5	24.82
MM-D2V2 7(Ours)†	30.24	56.48	67.46	11.96	30.48	41.88	7.36	26.42	36.6	3.7	16.58	23.84

Table 1: Comparison of Zero-shot Image-Text and Text-Image Retrieval of first results with FLAVA and Data2Vec2 papers. Because Data2Vec2 is a unimodal model, we embed each image with the D2V2-Image model and each text with the D2V2-Text model. This yields unusable results, as there has been no incentive for the models to learn a shared representation, as both are unimodal. This is why we had to use both the image and the text model to embed the data.
†: This version has been trained with BEiT-2 as the teacher model, not the D2V2 Image model.

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
Zero-Shot												
FLAVA	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
MM-D2V2 (Ours)	31.72	56.78	67.9	12.42	31.05	42.5	7.7	26.18	37.6	4.08	17.01	24.26
Finetune												
BEiT-3	84.8	96.5	98.3	67.2	87.7	92.8	98	100	100	90.3	98.7	99.5
VLMo	74.8	93.1	96.9	57.2	82.6	89.8	92.3	99.4	99.9	79.3	95.7	97.8

Table 2:

- looking at the validation loss of image and text seperatly, on COCO val set, we observe that the loss on images is significantly lower than the loss on text, which might be due to the fact that the teacher model is a vision model and the target, the cls token, might be biased towards the image modality, as it is unimodal
- interestingly, this bias also seem to be directly translated to the performance on image-text retrieval, as the performance on image-text retrieval is significantly higher than on text-image retrieval -> we are learning the cls token representation, and using the leared cls token as an output for the student model, to encode a modality for retrieval and other downstream task

-> suggests that the cls token is biased towards the image modality, or rather that the model is better in encoding images than text

- we can see that the performance of e.g. BEiT-3 and VLMo is also lower on text-image retrieval than on image-text retrieval, but not the the extend that we observe with our model

Model	**MSCOCO (5K test set)**		**Flickr30K (1K test set)**	
	Image → Text	Text → Image	Image → Text	Text → Image
MM-D2V2 7(Ours)†	51.39	28.11	23.46	14.71
BEiT-3	93.2	82.57	99.33	96.17
VLMo	88.27	76.53	97.2	90.93

Table 3: Average recall of image-text and text-image retrieval on MSCOCO and Flickr30K. All models continously perform better on image-text retrieval than on text-image retrieval, but the difference is more pronounced for our model.

- currently vl layer(s) (or rather mulimodal layer(s)) are randomly initialized, one option is to specifically initialize the multimodal layers with the weight of the final layers of the D2V text model -> initial state is closer closer to text modality