Multimodal Knowledge Distillation

Seperate Self-Attention

Baseline

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: