Appendix of Dissecting Deep Metric Learning Losses for Image-Text Retrieval

1. Validation on Gradient Method

Table 1 shows the results from origin VSE++ and VSE∞ work trained with triplet loss and the results implemented with the equivalent gradient methods with combination of T^{con} and P^{con} as mentioned in Section 4.2.

	Image→ Text			Text → Image		
Method	R@1	R@5	R@10	R@1	R@5	R@10
VSE++(R152,FT)	41.3	-	81.2	30.3	-	72.4
VSE++(R152,FT) ours	41.0 ± 0.3	70.4 ± 0.4	81.3 ± 0.3	30.2 ± 0.1	60.1 ± 0.1	72.5 ± 0.1
VSE∞(BUTD)	58.3	85.3	-	42.4	72.7	-
VSE∞(BUTD) ours	58.3 ± 0.7	85.5 ± 0.4	92.6 ± 0.0	43.1 ± 0.0	73.3 ± 0.1	83.4 ± 0.2
VSE∞(WSL)	66.4	89.3	-	51.6	79.3	-
VSE∞(WSL) ours	66.2 ± 0.2	89.5 ± 0.2	94.8 ± 0.3	51.6 ± 0.3	79.3 ± 0.2	87.6 ± 0.2

Table 1. Results verification of the model trained with triplet loss function backward vs the model trained with gradient backward on three VSE methods

2. Flickr 30K test result

Similar Experiment on Flickr 30K as mentioned in Section 4.3.

VSE++ (ResNet152, fine-tuned)						
	Image→ Text			Text → Image		
	T^{con}	T^{nca}	T^{cir}	T^{con}	T^{nca}	T^{cir}
P^{con}	55.0 ± 0.5	55.2 ± 0.3	54.7 ± 0.9	40.9 ± 0.7	39.6 ± 0.1	40.3 ± 0.2
P^{lin}	55.4 ± 0.7	54.9 ± 1.1	55.6 ± 0.9	41.2 ± 0.5	40.9 ± 0.1	41.4 ± 0.4
P^{sig}	56.8 ± 0.4	57.0 ± 0.5	56.3 ± 0.6	41.5 ± 0.2	42.4 ± 0.5	42.1 ± 0.4
P^{lin-ms}	54.9 ± 0.6	56.2 ± 0.7	55.7 ± 0.5	40.6 ± 0.8	41.1 ± 0.4	41.0 ± 0.3
P^{sig-ms}	53.8 ± 0.7	56.4 ± 1.3	55.7 ± 1.1	40.3 ± 0.4	41.1 ± 0.2	40.6 ± 0.2
VSE++ (ViT-base-patch16, fine-tuned)						
P^{con}	67.3 ± 0.9	67.1 ± 0.3	68.0 ± 0.6	52.8 ± 0.2	53.6 ± 0.4	53.3 ± 0.3
P^{lin}	67.4 ± 0.6	68.1 ± 0.1	68.1 ± 0.4	52.8 ± 0.3	53.4 ± 0.2	53.5 ± 0.4
P^{sig}	68.9 ± 0.2	68.4 ± 0.5	68.9 ± 0.2	53.6 ± 0.1	54.6 ± 0.2	54.2 ± 0.2
P^{lin-ms}	68.9 ± 0.8	68.4 ± 0.6	69.1 ± 1.0	53.1 ± 0.2	53.6 ± 0.4	53.4 ± 0.2
P^{sig-ms}	68.8 ± 0.0	70.5 ± 1.4	70.1 ± 1.6	53.1 ± 0.4	54.8 ± 0.2	54.4 ± 0.3

Table 2. Result of Image→ Text and Text → Image Recall@1 on Flickr 30K test with different gradient combinations on two steps VSE++ training with ResNet152.

VSE∞(BUTD)						
	Image→ Text			Text → Image		
	T^{con}	T^{nca}	T^{cir}	T^{con}	T^{nca}	T^{cir}
P^{con}	81.1 ± 0.6	81.8 ± 0.8	81.8 ± 0.9	62.1 ± 0.5	62.3 ± 0.8	62.2 ± 0.7
P^{lin}	80.5 ± 0.5	80.9 ± 0.7	80.3 ± 0.7	62.2 ± 0.5	62.8 ± 0.2	62.3 ± 0.8
P^{sig}	80.2 ± 0.7	81.5 ± 0.7	80.7 ± 1.1	62.2 ± 1.6	62.9 ± 0.8	62.8 ± 0.4
P^{lin-ms}	80.6 ± 0.6	80.9 ± 0.9	81.7 ± 1.3	62.3 ± 0.4	62.3 ± 0.6	62.5 ± 0.9
P^{sig-ms}	80.5 ± 0.6	81.9 ± 0.8	82.3 ± 0.9	63.2 ± 0.6	63.6 ± 0.7	64.0 ± 0.6
VSE∞(WSL)						
P^{con}	87.9 ± 1.1	89.4 ± 1.3	88.4 ± 0.6	74.0 ± 0.4	74.5 ± 0.6	74.0 ± 0.2
P^{lin}	88.0 ± 0.1	88.8 ± 0.4	89.0 ± 0.6	74.1 ± 0.7	74.8 ± 0.6	74.5 ± 0.7
P^{sig}	88.5 ± 0.7	88.7 ± 0.9	89.7 ± 0.4	74.8 ± 0.7	75.4 ± 0.2	75.2 ± 0.2
P^{lin-ms}	88.0 ± 0.6	89.5 ± 0.7	89.6 ± 0.8	74.6 ± 0.7	75.6 ± 0.7	75.1 ± 0.4
P^{sig-ms}	89.4 ± 0.8	89.6 ± 0.3	90.6 ± 0.8	75.9 ± 0.6	76.0 ± 0.3	76.7 ± 0.2

Table 3. Result of Image→ Text and Text → Image Recall@1 Flickr 30K test with different gradient combinations on VSE∞(BUTD) and VSE∞(WSL).