The Supplementary of Rethinking Label-Wise Cross-Modal Retrieval from A Semantic Sharing Perspective

In supplementary materials, we mainly focus on the impact of hyperparameters on model performance.

1 The Discussion

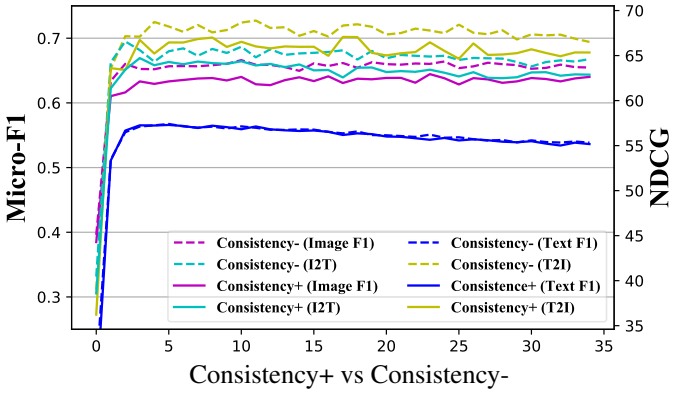


Figure 1: Discussion of consistency regularization. Each modal classification performance and cross-modal retrieval performance using consistency+ method (i.e., traditional method with consistency loss and $\eta = 0.1$) vs consistency- method (i.e., traditional method without consistency loss). The abscissa axis denote the epoch.

To better understand the effect of consistent loss, we conduct the experiments on classification-retrieval comparison using FLICKR25K dataset. Consistency+ denotes the Eq. 2 in the main body, and the Consistency- represents the Eq. 2 without ℓ_{con}. Figure 1 records the results: 1) we find that the purple solid line (i.e., with consistency loss) places lower than the purple dotted line (i.e., without consistency loss) after stable, and the performance of blue solid and dotted lines are coincident. In other words, the joint training with consistency loss actually reduces the classification capacity of "strong" modality (i.e., image modality), but doesn't improve the performance of "weak" modality (i.e., text modality). 2) we find that the retrieval performance increases with the decreasing of classification error under two settings (i.e., with or without consistency loss), and the green and yellow solid lines place lower than their corresponding green and yellow dotted lines after stable. In other word, the performance of cross-modal retrieval decrease after using the consistency loss. To summary, we consider that *consistency loss affects the classification performance of each modality, and then impact the retrieval performance.* From the learning formulation of label-wise methods, we find that the main purpose of these approaches is to make the cross-modal intra-class examples closer, and we seperate the inter-class examples with the help of similarity matrix S and S^m that constructed in advance using label information. From another perspective, the embedding representations of different modalities obtained by joint training should be conducive to classification. However, different modalities have various classification capabilities, i.e., there exist "weak" modality (e.g., the text modality in Figure 1) and "strong" modality (e.g., the image modality in Figure 1). Therefore, the joint optimization with consistency loss may overfit and pull down the classification capability of "strong" modality.

2 Implements

The proposed method has two sub-networks, one for image modality and the other for text modality. For the image part, we use fixed 36 RoIs with detection scores higher than 0.5 for each image. If eligible RoIs are less than 36, we simply select the top-36 RoIs, regardless of the detection score threshold. Each RoI is represented with a 1024-dimensional vector. For text part, we directly tokenize each input text word, and represent them by 300-dimensional vectors with Word2Vec. Besides, the shared encoder has 6 layers of Transformer blocks, where each block has 8 self-attention heads. The maximum sequence length is set as 150. The parameters are initialized from BERT-base, which is pre-trained on text data only. In all experiments, the batch size is set to 64, and the parameter λ is tuned in $\{0.001, 0.01, \cdots, 1, 10\}$. For training, we employ the ADAM [Kingma and Ba, 2015] optimizer with a learning rate of 10^{-4} and set the dropout ratio as 0.1, the maximal number of epochs as 45. The entire network is trained on a Nvidia TITAN X GPU.

3 Classification Performance

To evaluate the classification performance using the shared model, we conduct more experiments: 1) Independent, we train the image and text model independently. 2) DSCMR and Ours+Con belong to the best deep models. The results in Table 1 reveal that our shared model and interactively training can improve the classification performance of each modality

Table 1: Classification performance comparison in terms of Micro-F1 score. The best results in testing are highlighted in bold.

Methods	FLICKR25K		NUS-WIDE		COCO1K		COCO5K	
	Image	Text	Image	Text	Image	Text	Image	Text
Independent	0.672	0.53	0.644	0.529	0.628	0.623	0.624	0.617
DSCMR	0.634	0.536	0.611	0.570	0.627	0.614	0.538	0.538
Ours+Con	0.686	0.549	0.651	0.596	0.683	0.632	0.685	0.63
Ours	**0.690**	**0.556**	**0.651**	**0.597**	**0.692**	**0.635**	**0.689**	**0.633**

Figure 2: (Best viewed in color when zoomed in.) Failure results of text retrieval given image queries and image retrieval given sentence queries. For each image query we show the top-5 ranked sentences. For each sentence query, we show the top-3 ranked images, ranking from left to right.

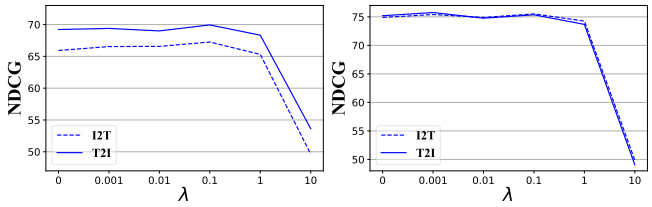
comparing with the methods including consistency loss (i.e., DSCMR and Ours+Con), and even superior to the results of independent training (i.e., Independent).

4 Failure Case Study

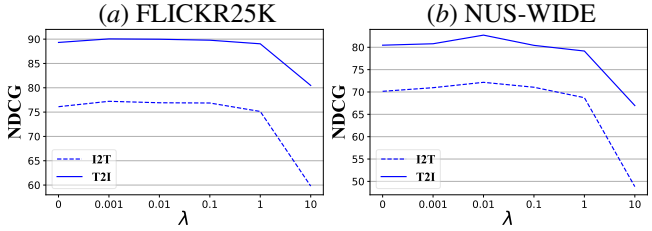
We give more cases about cross-modal retrieval. The top row of Figure 2 shows the qualitative results of sentence retrieval given image queries. The bottom row of Figure 2 illustrates the qualitative results of image retrieval given sentence queries. Each sentence corresponds to a ground-truth image. For each sentence query, we show the top-3 retrieved images, ranking from left to right. We attribute the reasons for incorrect results to two categories: 1) Semantic confusion. Image content and sentence description are difficult to distinguish, for example, in the case (a) of I2T, the shape of the spoon is similar to a ball, and "ball" also has different semantic meanings in different contexts; in the first case of T2I, traffic signs and traffic lights usually appear together, so it's difficult to separate them. 2) Super-class confusion. The labels of the search results may include or relate to the query, for example, in the case (b) of I2T, the super-class of banana is fruit, while the vegetable is closely relate to the fruit.

5 Sensitivity to Parameters

The main parameter in classification prediction is the λ in Eq. 6 of the main body. We vary the parameter in $\{0, 0.001, 0.01, 0.1, 1, 10\}$ to study its sensitivity for retrieval performance, and record the NDCG results in Figure 3. The results indicate that the performance increases with the increasing of λ at first, and maintains stability with small λ. However, it drops sharply when the mask prediction ratio is large, which also validate the phenomenon mentioned in the main body that label prediction is critical for learning consistent embedding. Therefore, we set $\lambda = 0.01$ for COCO



(*a*) FLICKR25K (*b*) NUS-WIDE



(*a*) COCO-1K (*b*) COCO-5K

Figure 3: Parameter sensitivity of λ. Solid line denotes the T2I, and dotted line denotes the I2T.

dataset, and $\lambda = 0.1$ for other two datasets.

References

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, San Diego, CA, 2015.