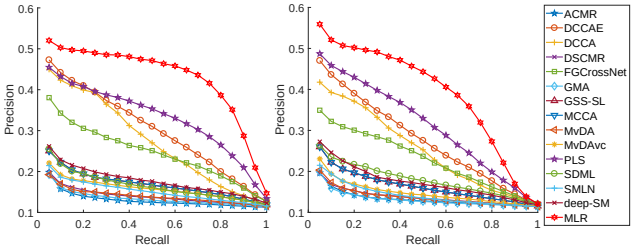
Supplementary Material for Learning Cross-Modal Retrieval with Noisy Labels

Peng Hu[1,2] Xi Peng[1*] Hongyuan Zhu[2] Liangli Zhen[3] Jie Lin[2]
[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

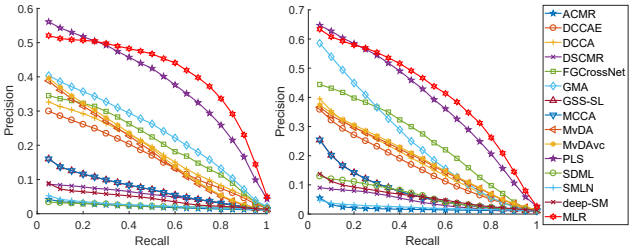
1. Precision-Recall Comparison

To visually investigate the effectivenss of the proposed method, we plot the precision-recall curves compared to the state-of-the-art methods as shown in Figures 1 to 4. From the experimental resutls, one can see that our MLR is superior to all baselines, which is consisted with the MAP scores of cross-modal retrieval.



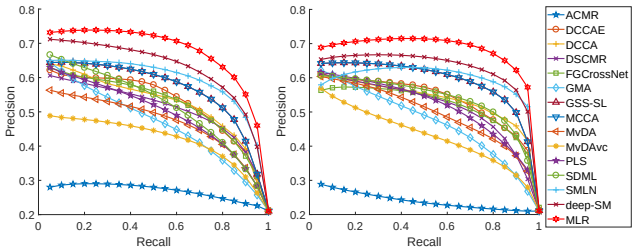
(a) Image → Text (b) Text → Image

Figure 1: The precision-recall curves on the Wikipedia dataset. The noise rate is 0.8.



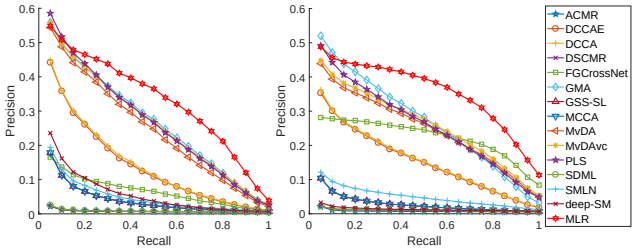
(a) Image → Text (b) Text → Image

Figure 2: The precision-recall curves on the INRIA-Websearch dataset. The noise rate is 0.8.



(a) Image → Text (b) Text → Image

Figure 3: The precision-recall curves on the NUS-WIDE dataset. The noise rate is 0.8.



(a) Image → Text (b) Text → Image

Figure 4: The precision-recall curves on the XMediaNet dataset. The noise rate is 0.8.

2. Convergence Analysis

The Convergence curves of our MRL plot the values of the loss function (*i.e.*, $\mathcal{L} = \beta\mathcal{L}_r + (1-\beta)\mathcal{L}_c$) versus different number of epochs on the INRIA-Websearch dataset as shown in Figure 5. From these figures, we can see that our MRL can fast converge between 50 and 100 epochs. Comparing to cross-entropy, our MRL is more robust and stable in the training process, indicating that our method is robust to the noisy labels, which is consistent with the experimental results.

*Corresponding author: Xi Peng (pengx.gm@gmail.com).

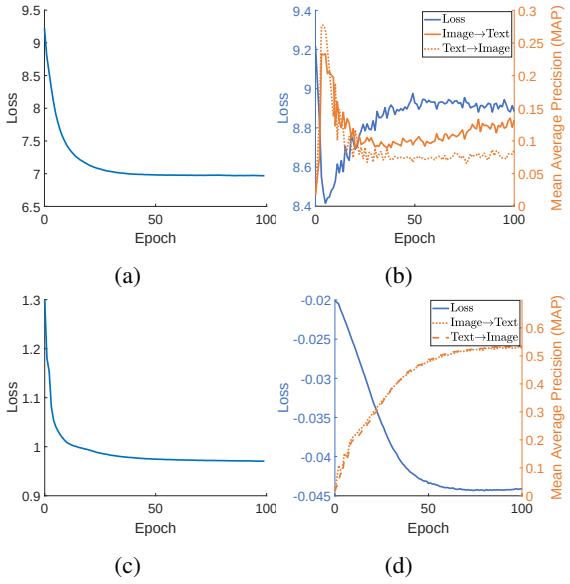
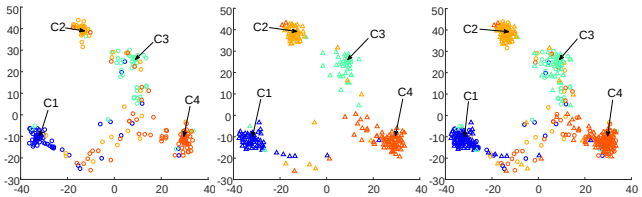


Figure 5: Convergence Analysis on INRIA-Websearch with 80% noise. (a) Loss *vs*. epoch on the training set for Cross-Entropy/deep-SM. (b) Loss/MAP *vs*. epoch on the validation set for Cross-Entropy/deep-SM. (c) Loss *vs*. epoch on the training set for our MRL. (c) Loss/MAP *vs*. epoch on the validation set for our MRL.

2.1. Visualization of the Learned Representation

To visually investigate the discrimination of common representations learned by different cross-modal methods, we adopt the t-SNE approach [1] to embed the samples from the Wikipedia dataset into a two-dimensional space as shown in Table 1. From these figures, we could see that the learned representations of most supervised methods (*i.e*., deep-SM, FGCrossNet, and SDML) from different modalities cannot overlap with each other like input data, indicating that the noisy labels interfere this supervised methods to learn the common space. For the unsupervised methods (*e.g*., DCCAE), although it can narrow the heterogeneous gap, it cannot make the different classes sufficient scattered and the same ones sufficient compact, indicating that unsupervised methods cannot push enough discrimination into the common space. Although DSCMR uses an unsupervised item to mitigate the cross-modal discrepancy, it will be interfered by the noisy labels and cannot learn the discrimination, *i.e*., each classes are randomly scattered in the common space as shown in Table 1. On the contrary, our MRL can make the different classes more scattered and the same ones more compact. With our proposed techniques, our MRL could learn discrimination from the noisy labels while narrowing the heterogeneous gap, which is consistent with the retrieval experiments. Moreover, we have added some visualization results with the symmetric noise rate of 0.6. The evaluation is conducted on four randomly selected classes of the Wikipedia dataset. From the figures, one could see that most of samples are clustered to their corresponding clustering centers (*i.e*., C1–4) that are the learned clusters (*i.e*., C). In other words, our method can automatically cluster the samples based on their semantics.



(a) Image (b) Text (c) Image & Text

Figure 6: The visualization for the clustering. In the figure, the shape of marker represents a given view, and the color indicates the class of a given point. Moreover, the solid and hollow points denote the clustering center and sample, respectively.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2, 3

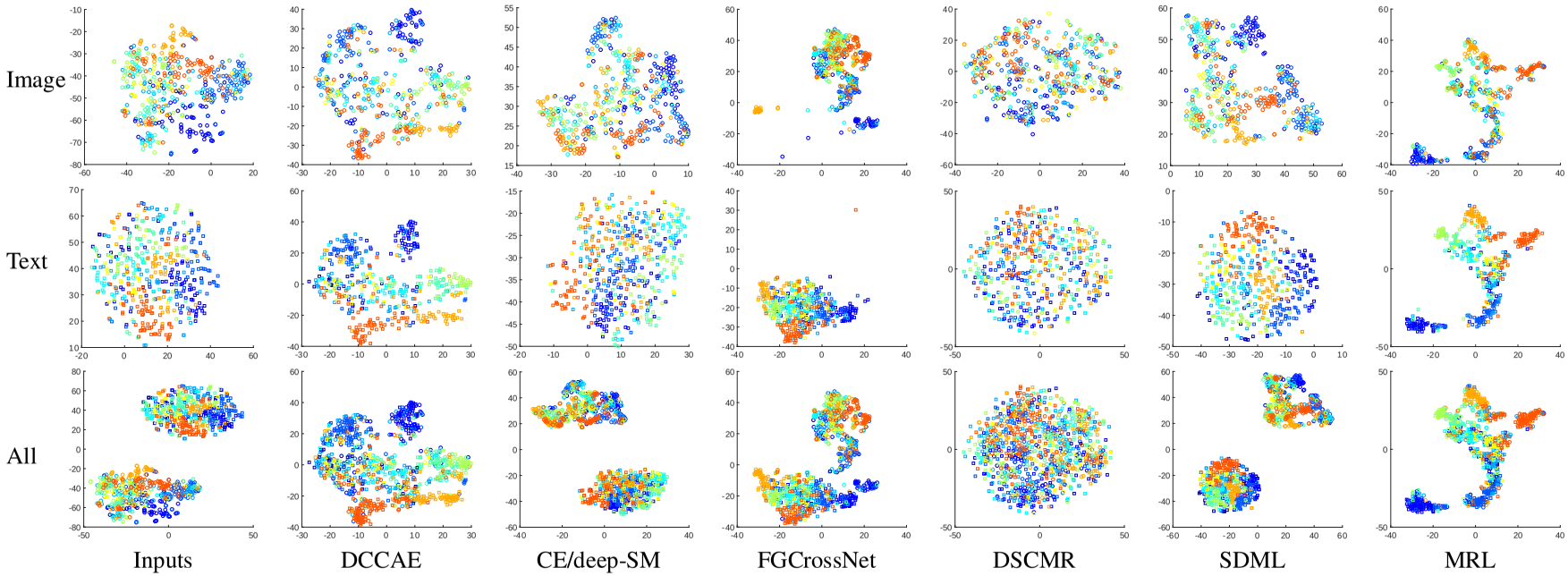


Table 1: The visualization for the test data on the Wikipedia dataset by using the t-SNE method [1]. In this figure, the different shape of markers represents its corresponding view, and the different colors denote their corresponding classes, respectively.