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

Yang Yang[1*], **Chubing Zhang**[1], **Yi-Chu Xu**[2], **Dianhai Yu**[3], **De-Chuan Zhan**[2] and **Jian Yang**[1]
[1]Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology,
[2]Nanjing University,
[3]Baidu Inc
{yyang, chubingzhang, csjyang}@njust.edu.cn, yudianhai@baidu.com, xuyc@lambda.nju.edu.cn, zhandc@nju.edu.cn

Abstract

The main challenge of cross-modal retrieval is to learn the consistent embedding for heterogeneous modalities. To solve this problem, traditional label-wise cross-modal approaches usually constrain the inter-modal and intra-modal embedding consistency relying on the label ground-truths. However, the experiments reveal that different modal networks actually have various generalization capacities, thereby end-to-end joint training with consistency loss usually leads to sub-optimal uni-modal model, which in turn affects the learning of consistent embedding. Therefore, in this paper, we argue that what really needed for supervised cross-modal retrieval is a good shared classification model. In other words, we learn the consistent embedding by ensuring the classification performance of each modality on the shared model, without the consistency loss. Specifically, we consider a technique called Semantic Sharing, which directly trains the two modalities interactively by adopting a shared self-attention based classification model. We evaluate the proposed approach on three representative datasets. The results validate that the proposed semantic sharing can consistently boost the performance under NDCG metric.

1 Introduction

Cross-modal learning constructs the aligning or mapping function among different modalities [Baltrusaitis *et al.*, 2019]. An important task is the cross-modal retrieval, which aims to search examples in one modality (for example, image) that have similar semantic representations to the query from another modality (for example, text), rather than performs a similarity search within the same modality. For example, Baidu or Google can provide corresponding image display based on user's text description, and present relevant introduction according to the user's image input. Comparing to single modal search, the difficulty of cross-modal retrieval is the semantic gap of heterogeneous data, which needs to constrain the consistency between two modal semantic representations effectively.

To solve this problem, many cross-modal retrieval approaches are designed with different losses or regularizations for minimizing the heterogeneity. [Wang *et al.*, 2016b] divided these methods into real-valued and hash based approaches according to the output embedding form. On the other side, in this paper, we rely on the usage of ground-truths, i.e., the instance labels or cross-modal alignments, to partition these approaches into two categories: 1) Label-wise methods [Wang *et al.*, 2013; Wang *et al.*, 2016a; Li *et al.*, 2018; Zhen *et al.*, 2019]. These methods aim to retrieve cross-modal instances of the same category, they usually utilize the label information to construct the similarity matrix, which can constrain the cross-modal inter-class and intra-class distance for learning consistent embedding. In result, the cross-modal embeddings are similar for the same class, and dissimilar for different classes. 2) Alignment-wise methods [Wang *et al.*, 2019; Lee *et al.*, 2018; Zhang and Lu, 2018; Yu *et al.*, 2020]. These approaches are designed to retrieve accurately aligned cross-modal instances, rather than the instances of same category. Therefore, they usually adopt triplet loss or hard triplet loss using the alignment ground-truths. In result, the instance is only similar to the aligned cross-modal instance. In this paper, *we concentrate on the label-wise style methods*, which always devote to develop a multi-modal neural network, which processes an embedding based consistency loss for jointly optimizing the two modal embedding networks. However, from the results shown in Figure 1, we find that joint training with consistency loss does not significantly improve the retrieval performance, and may even reduce the performance.

Upon inspection, the problem can be attributed to the classification performance decrease of different modalities influenced by the consistency loss. In fact, different modalities contain imbalanced intrinsic information, which leads to differences in classification capabilities, i.e., there exist "weak" and "strong" modalities [Yang *et al.*, 2015]. However, the joint training consistency loss may reduce the capability of

*Contact Author

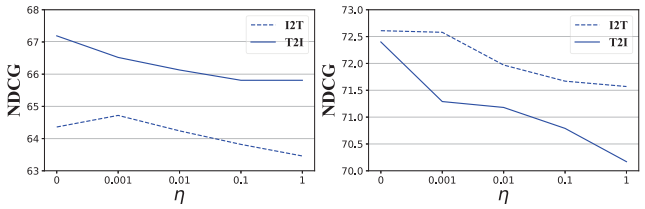
"strong" modality. On the other hand, the learning of cross-modal consistent embedding is related to the classification performance of each modality considering the learning target, so training jointly is sub-optimal for learning consistent embedding. Then how to avoid the negative effects of modal joint training with consistency loss? We rethink the cross-modal retrieval from a simple and straightforward aspect: semantic sharing, i.e., training with a shared classification network. We directly shares a unified multi-head attention classification network for the two modalities, and therefore make better use of the region structural information shared by the cross-modal instance itself. In summary, our contribution includes: 1) rethink the role of different losses in traditional label-wise methods. 2) remove the interference term (i.e., embedding consistency loss), and attempt to adopt the shared model instead of the structure consistency for learning semantically common embeddings for the two modalities.

2 Related Work

Traditional survey [Wang *et al.*, 2016b] partitioned the cross-modal retrieval approaches into real-valued and hash based methods according to their output form. The difference is that the hash methods are more efficient with hash encoding.

From another perspective, these methods can also be summarized as label-wise and alignment-wise categories according to their usage of annotations. Label-wise methods usually adopt label annotations to learn consistent embedding, which ensures the similarity of cross-modal instances in the same category. For example, [Zhen *et al.*, 2019] developed a deep supervised cross-modal retrieval method, which minimizes the discrimination loss in both label and representation space to supervise the discriminative feature learning; [Wang *et al.*, 2013] proposed a novel coupled linear regression framework, which learns two projection matrices to map multimodal data into a common feature space; [Wang *et al.*, 2016a] learned projection matrices to map multi-modal data into a common subspace, which measures the similarity between different modalities of data. Besides, alignment-wise methods turn to utilize the alignment annotations to learn consistent embedding, which minimizes the embedding differences of aligned instances with triplet loss. For example, [Lee *et al.*, 2018] presented stacked cross attention to discover the full latent alignments, using both image regions and words in a sentence as context; [Faghri *et al.*, 2018] incorporated hard negatives in the loss function, which is equivalent to minimize a modified non-transparent loss function. The motivations and evaluation metrics of these two types of methods are different, so we usually don't compare them with each other. However, it is not difficult to find that the label-wise methods are closely related to the classification performance of each modality, whereas ignore the impact of consistency loss on the classification performance.

Our work is related to previous research on multi-modal networks for cross-modal retrieval [Wang *et al.*, 2016b], which uses the joint training with the consistency loss. On the other hand, the primary task of our work is to learn the consistent embedding, which is also related to the semantic representation sharing. Some researches have just come up



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

Figure 1: Influence of consistency regularization for cross-modal retrieval. η represents the weight for consistency loss term, and NDCG is the retrieval metric. The retrieval performance on the two data sets may decreases with the consistency loss weight η becomes larger.

to use transformer encoder for learning shared cross-modal semantic representation. For example, [Lu *et al.*, 2019] extended the popular BERT architecture to process both visual and textual inputs in separate streams that interact through co-attentional transformer layers; [Li *et al.*, 2020] fed both visual and linguistic contents into a multi-layer Transformer for the cross-modal pre-training. Nevertheless, these approaches belong to the alignment-wise methods, which place emphasis on inputting the region segmentations of the aligned image-text pair to the transformer encoder, then using the traditional triple loss to learn the consistent embedding. They are different from the viewpoint stated in this paper that the consistency may affects the performance of label-wise methods.

3 The Proposed Method

3.1 Background

Without any loss of generality, the training set can be denoted as $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{v}_i \in \mathcal{R}^{d_v}$ denotes the $i-$th image instance, $\mathbf{w}_i \in \mathcal{R}^{d_w}$ represents the $i-$th sentence instance, and $\mathbf{y}_i \in \mathcal{R}^{C}$ denotes the instance label, $\mathbf{y}_{i,j} = 1$ if $i-$th instance belongs to the $j-$th label, otherwise is 0. Obviously, it is difficult to compare the two modal raw features for cross-modal retrieval considering that they lie in heterogeneous feature spaces and have various physical properties [Wang *et al.*, 2016b]. Therefore, traditional methods aim to design the inter-modal consistency loss ℓ_{con} with the label information for joint training, which can learn two embedding functions, i.e., $\mathbf{z}_{i_v} = f_v(\mathbf{v}_i)$ and $\mathbf{z}_{i_w} = f_w(\mathbf{w}_i)$ for two modalities, and $\mathbf{z}_{i_v}, \mathbf{z}_{i_w} \in \mathcal{R}^d$ are d-dimensional embedding in common space. To develop ℓ_{con}, current approaches [Wang *et al.*, 2013; Wang *et al.*, 2016a; Li *et al.*, 2018; Zhen *et al.*, 2019] always employ the discrimination loss of all examples from both modalities in the common representation space:

$$\begin{aligned}\ell_{con} = &-\sum_{i,j}^{N}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) \\ &-\sum_{m\in\mathbf{v},\mathbf{w}}\sum_{i,j}^{N}(S_{ij}^{m}\Theta_{ij}^{m} - \log(1+e^{\Theta_{ij}^{m}}))\end{aligned} \tag{1}$$

where S_{ij}/S_{ij}^{m} denotes the inter-modal/intra-modal similarity matrix. $S_{ij} = 1$ if $\mathbf{z}_{i_v}$ and $\mathbf{z}_{j_w}$ are the representations

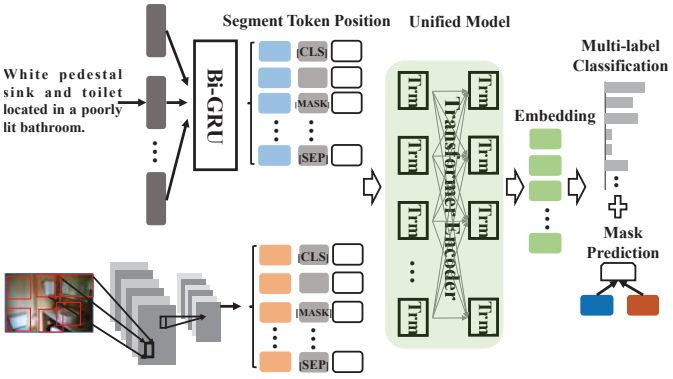


Figure 2: Illustration of the proposed semantic sharing. It has two modules: 1) modal independent segmentation model, which extracts multi-instance representations for each example; 2) unified model, which adopts multi-layer transformer encoder that map different modal segmentations into common representation space with semantic-preserving.

of intra-class samples, otherwise 0. Similarly, $S_{ij}^{m} = 1$ if $\{\mathbf{z}_{i_m}, \mathbf{z}_{j_m}, m \in \{\mathbf{v}, \mathbf{w}\}\}$ are the representations of intra-class samples in the same modality, otherwise 0. $\Theta_{ij} = \frac{1}{2}\cos(\mathbf{z}_{i_v}, \mathbf{z}_{j_w})$ denotes the inter-modal similarity, and $\Theta_{ij}^{m} = \frac{1}{2}\cos(\mathbf{z}_{i_m}, \mathbf{z}_{j_m}), m \in \{\mathbf{v}, \mathbf{w}\}$ represents the intra-modal similarity, cos is the cosine function. Therefore, the first term in Eq. 1 denotes the inter-modal consistency, while the second term implies the intra-modal consistency. In detail, a larger similarity value (cosine similarity $\cos(\mathbf{z}_{i_v}, \mathbf{z}_{j_w})$) means that $\mathbf{v}_i, \mathbf{w}_j$ should be classified as similar, and vice versa. Identically, the second term measures the intra-modal similarities. Therefore, the overall loss is:

$$L = \sum_{i=1}^{N} \ell(\mathbf{z}_{i_v}, \mathbf{z}_{i_w}, \mathbf{y}_i) + \eta \ell_{con} \tag{2}$$

where the first term is the traditional classification loss (e.g., the cross-entropy loss for each modality), then we jointly optimize the overall loss to train the two modal embedding network for learning common representations. However, the results in Figure 1 reveal that the consistency loss does not significantly improve the performance, and even reduces the performance when η increases. We attribute this phenomenon to the affection of consistency loss on every single modal classification ability, which will impact the learning of cross-modal consistent embedding inversely. To solve this problem, we instead use the shared classification model to iteratively train image and text modalities, and learn consistent semantic embeddings by ensuring that the classification performances of the two modalities are improved with the shared model.

3.2 Cross-Modal Retrieval with Semantic Sharing

Modal Independent Embedding Models

This module aims to extract initial features for each modalities. Compared to global feature embedding, inspired from [Karpathy and Fei-Fei, 2017], the semantic relationships of different regions between various modality are similar, for example, the correlation between the embeddings of "sweetveld" and "elephants" in the text is similar to that of corresponding regions in the image. Therefore, region representations are more suitable than the global features as the input for learning comparable embedding.

Based on this idea, we implement f_v and f_w as two multi-instance deep neural networks. Specifically, for image modality, we utilize the pre-trained Faster R-CNN [Lee *et al.*, 2018], which extracts visual regions with pooled ROI embeddings, i.e., the 1024-dimensional feature vector from fc7 layer, denoted as $\{\hat{\mathbf{v}}_i^t\}_{t=1}^{T_i}$ for $i-$th instance, t is the index, T_i is fixed as 36 for all image instance as [Lee *et al.*, 2018] for better performance. Meanwhile, to perform segmentation learning for text, we tokenize each instance using Bi-GRU [Bahdanau *et al.*, 2015] with the Word2Vec representations [Mikolov *et al.*, 2013] as input, and obtain the output as $\{\hat{\mathbf{w}}_j^t\}_{t=1}^{T_j}$, T_j is the length of $j-$th input. In result, each image or text example can be denoted as a bag of instances.

Shared Model

With the multi-instance representations of each modality, the shared model aims to take the shareable semantic relationships into feature encoding for learning more comparably common embeddings. Inspired from the transformer [Vaswani *et al.*, 2017], which encode the relationships by utilizing the multi-head attention mechanism. We enforce the two sub-networks, i.e., f_v and f_w, to share a unified model f for learning final embeddings. Therefore, it is intuitive to build a general model which maps different modal outputs into a common subspace with semantic-preserving, and thus we can generate correlated embeddings for image and text examples from the same category.

Specifically, with the multi-instance outputs $\{\hat{\mathbf{v}}_i^t\}_{t=1}^{T_i}$ and $\{\hat{\mathbf{w}}_j^t\}_{t=1}^{T_j}$, we build f, using the encoder as a multi-layer transformer [Vaswani *et al.*, 2017]. Each modal output is passed through L encoder-style transformer blocks to produce final common embeddings $\mathbf{z}_{i_v}$ and $\mathbf{z}_{i_w}$, using image modality as an example:

$$\begin{aligned} \mathbf{z}_{i_v} &= \mathcal{T}(\{\mathbf{e}_i^t\}_{t=1}^{T_i}) \\ \mathbf{e}_i^t &= \phi(\psi_1(\hat{\mathbf{v}}_i^t) + \psi_2(l_i^t)) \end{aligned} \tag{3}$$

where $\mathbf{z}_{i_v}$ denotes the final embeddings of $i-$th image instance, $\mathbf{e}_i^t$ represents the t-th segmentation input of $i-$th instance. Both the raw segmentation representation $\hat{\mathbf{v}}_i^t$ and position features l_i^t are fed through fully connected (FC) layers ψ_1 and ψ_2, i.e., we use one fully connected network that projects them into the same embedding space, $\psi_1(\hat{\mathbf{v}}_i^t) \in \mathcal{R}^d, \psi_2(l_i^t) \in \mathcal{R}^d$. Then, we sum the two features and use another non-linear mapping ϕ to obtain $\mathbf{e}_i^t$. For image modality, we design the position features according to [Li *et al.*, 2019a], i.e., each region position is represented with a $5-$D vector, $p = (\frac{a_1}{W}, \frac{b_1}{H}, \frac{a_2}{W}, \frac{b_2}{H}, \frac{(b_2-b_1)(a_2-a_1)}{W \cdot H})$, where (a_1, b_1) and (a_2, b_2) denote the coordinate of top-left and bottom-right corner, W, H are the width and height of the input image, and the last value represents the fraction of image covered. The position features of the text modality is designed according to original method [Vaswani *et al.*, 2017]. $\mathcal{T}$ denotes a single encoder-style transformer block with multi-head attention

block, wrapped in residual adds [Vaswani *et al.*, 2017]. Suppose H^l corresponds to the intermediate representations after l-th layer, it can be used to compute three matrices: Q, K, and V corresponding to queries, keys, and values that drive the multi-head attention block. The dot-product similarity between queries and keys determines the attention distribution of values. Then weight-averaged values form the output of the attention block.

Overall Function

We aim to learn the common embedding for two modalities. To achieve this goal, we propose to directly use the shared model to optimize the prediction loss of each modality, while use the label information to learn the cross-modal semantic-preserving embedding. The objective contains: 1) Label prediction, which utilizes the final embedding for classification; and 2) Context prediction, which tries to predict the identity of each masked segment based on all context segments.

First, the final semantic embeddings obtained by the transformer can be used for classification task. In detail, a shared classifier is connected on the top of transformer encoder network, and takes the final embeddings as input to generate a C-dimensional prediction. The following objective function can be expressed as:

$$\begin{aligned} L_1 = \sum_{i=1}^{N} \log(1+\exp(-\mathbf{y}_i{}^\top g(\mathbf{z}_{i_v}))) + \\ \log(1+\exp(-\mathbf{y}_i{}^\top g(\mathbf{z}_{i_w}))) \end{aligned} \tag{4}$$

where g denotes the shared classifier using a one layer fully connected network with softmax operator. The loss function can be any convex loss operator, and we utilize the common multi-label loss as [Zhang and Zhou, 2014].

Inspired from Bert [Devlin *et al.*, 2019], which utilizes the masked model to learn high-level representations and capture rich relationships between segments. We also consider two masked models to further improve the learning ability of encoder, i.e., masked language and object classification models. In detail, we consider predicting the identity of different modal masked segment based on all context segments:

$$\begin{aligned} L_2 &= \sum_{i=1}^{N} \ell_v(\mathbf{v}_i) + \ell_w(\mathbf{w}_i) \\ \ell_v(\mathbf{v}_i) &= -logP_{g_v}(\mathbf{v}_i^m|\mathbf{v}_i^{\setminus m}) \\ \ell_w(\mathbf{w}_i) &= -logP_{g_w}(\mathbf{w}_i^m|\mathbf{w}_i^{\setminus m}) \end{aligned} \tag{5}$$

where ℓ_v, ℓ_w can be any convex loss function, and we utilize the cross-entropy for simplicity here. g_v is the linear classifier. g_w is the trainable model for text modality. For masked language model, the mask indices are $m \in \mathcal{N}^M$. We randomly mask input segment with probability of 15% as [Li *et al.*, 2019b] for image and text modalities, and replace the masked ones $\mathbf{v}_i^m$ and $\mathbf{w}_j^m$ with special token $[MASK]$. Then, the goal is to predict these masked segments, based on their surrounding segments $\mathbf{v}_i^{\setminus m}$, $\mathbf{w}_i^{\setminus m}$, via minimizing the negative log-likelihood.

In result, the overall formulation can be represented as:

$$L = L_1 + \lambda L_2 \tag{6}$$

Algorithm 1 The pseudo code

Input: Dataset: $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i)\}_{i=1}^N$;
Parameter:λ; maxIter: T, learning rate: l_r
Output: $\mathcal{T}, g$

```
while stop condition is not triggered do
    for mini-batch smapled from D do
        Calculate label prediction loss with Eq. 4;
        Calculate context prediction loss with Eq. 5;
        Calculate overall loss with Eq. 6;
        Update model parameters using gradient descent;
    end for
end while
```

For optimization, we randomly sample a mini-batch including image and text modalities at each iteration to iteratively train $\mathcal{T}, g, g_v, g_w$. With the learned model, we conduct inductive cross-modal retrieval. The procedure of training model is summarized in Algorithm 1.

4 Experiments

In this section, we conduct extensive experiments on three real-world datasets to demonstrate the effectiveness.

4.1 Datasets

We experiment on three public datasets, i.e., FLICKR25K [Huiskes and Lew, 2008], NUS-WIDE [Chua *et al.*, 2009] and MSCOCO [Lin *et al.*, 2014]: 1) FLICKR25K consists of 31,783 images collected from Flickr website. Each image is associated with several textual descriptions. Each example is manually annotated with 24 labels. The dataset is split into 29,783 training images, 1,000 validation images and 1,000 testing images following [Karpathy and Fei-Fei, 2017]. 2) NUS-WIDE contains 260,648 web images. Each image is also associated with several textual descriptions. Each point is annotated with 81 concept labels. We select the 21 most frequent concepts as [Yang *et al.*, 2019; Jiang and Li, 2017] and keep the corresponding 195,834 text-image pairs. The dataset is split into 189,834 training images, 5,000 validation images, and 1,000 testing images; 3) MSCOCO consists of 123,287 images, and each image contains roughly five textual descriptions. We follow the data split as [Faghri *et al.*, 2018], which left out 30,504 images that were originally in the validation set. Thus, the dataset is split into 82,783 training images, 5,000 validation images, and two testing sets with 1000/5000 images.

4.2 Baselines and Evaluation Protocol

Considering that our proposed method focuses on label-wise cross-modal retrieval, we firstly compare it with six linear and deep methods: CCA [Hotelling, 1992], LCFS [Wang *et al.*, 2013], JFSSL [Wang *et al.*, 2016a], DCCA [Andrew *et al.*, 2013], DSCMR [Zhen *et al.*, 2019], SCML [Song and Tan, 2019], ViLBert [Lu *et al.*, 2019]. Note that LCFS and JFSSL belong to the best linear methods. For a fair comparison, we pre-extract deep embeddings for linear methods, CCA, LCFS and JFSSL, as the input. For deep models, we carefully implement the approaches as their released code.

	FLICKR25K			NUS-WIDE			COCO1K			COCO5K		
	I2T	T2I	Ave	I2T	T2I	Ave	I2T	T2I	Ave	I2T	T2I	Ave
CCA	30.3	30.3	30.3	51.3	50.5	50.9	49.4	60.0	54.7	47.4	49.0	47.7
LCFS	35.8	32.4	34.1	57.3	58.0	57.6	58.8	68.4	63.6	52.5	59.1	55.8
JFSSL	34.8	29.7	32.3	57.3	53.6	55.4	62.9	72.1	67.5	55.9	63.3	59.6
DCCA	52.6	52.9	52.8	58.4	58.0	58.2	51.2	60.8	56.0	51.2	50.2	50.7
DSCMR	64.7	66.5	65.6	72.5	71.2	71.9	74.5	85.6	80.1	72.0	77.7	74.8
SCML	54.0	42.0	48.0	62.4	63.4	62.9	-	-	-	-	-	-
ViLBert	53.7	60.0	56.9	63.6	63.5	63.6	62.2	79.2	70.6	47.3	67.3	57.3
ST	65.9	65.8	65.9	74.6	58.1	66.3	62.1	82.9	72.5	44.7	71.2	58.0
IMC	38.8	37.2	38.0	31.8	24.1	28.0	23.7	54.9	39.3	14.7	31.2	23.1
w/o R	64.3	67.1	65.7	72.6	72.4	72.5	75.0	86.1	80.6	72.1	79.2	75.6
w/o T	65.1	66.0	65.5	74.2	74.0	74.1	75.6	89.4	82.5	69.8	80.0	74.9
w/o C	65.9	69.2	67.5	74.9	75.2	75.0	76.1	89.3	82.7	70.2	80.5	75.3
w/o L	40.1	42.7	41.4	33.4	30.9	32.2	22.5	40.9	31.7	14.5	31.9	23.2
Ours+Con	65.6	68.8	67.2	73.5	72.4	72.9	75.4	88.8	82.1	70.6	80.9	75.7
Ours	**66.3**	**71.3**	**68.8**	**75.2**	**75.3**	**75.3**	**77.2**	**90.0**	**83.6**	**72.1**	**82.7**	**77.4**

Table 1: Performance comparison in terms of NDCG score on three datasets. The best results in testing are highlighted in bold.

We also adopt ablation study to verify the effectiveness of each module: 1) Separate Training (ST) adopts the label information to train two modal models separately (with same dimensional feature embeddings), then conduct retrieval; 2) Inter-Modal and Intra-Modal Consistency (IMC) directly adopts the Eq. 1 for training; 3) w/o R replaces region segmentations with global embedding for the input, and utilize one fully connected network as a shared model, note that w/o R equals to the method using Eq. 2 with no ℓ_{con} for training; 4) w/o T replaces the multi-head attention network with one fully connected network as a shared model; 5) w/o C calculates the loss without mask prediction; 6) w/o L calculates the loss without label prediction; 7) Ours+Con adds the triplet consistency loss.

According to traditional settings, we perform two tasks: 1) Image vs. Text (I2T). 2) Text vs. Image (T2I). Ave denotes the average score. Considering three datasets are all multi-label datasets, we adopt Normalized Discounted Cumulative Gain metric as [Song and Tan, 2019].

4.3 Retrieval Results

Table 1 presents the quantitative comparison results with both sate-of-the-art label-wise methods and baselines. "-" indicates that the original paper did not provide the information. The results reveal that: 1) Supervised methods perform better than unsupervised methods, i.e., LCFS and JFSSL perform better than CCA, and DSCMR performs superior to DCCA. This indicates that label information helps to learn consistent cross-modal embedding. 2) Deep methods perform better than linear methods, i.e., DSCMR achieves the best performance in comparing methods. This indicates that deep networks benefit the learning of discriminative embedding. 3) The performance of DSCMR on FLICKR25K is even worse than ST (i.e., I2T). A possible explanation that the classification abilities of strong and weak modalities on FLICKR25K differ widely, and the consistency loss may greatly reduce the strong modal classification performance, thereby affecting retrieval performance. 4) The w/o R achieves better performance than DSCMR on most setting except I2T on FLICKR25K dataset, which validates the conclusion that consistency may cause negative effect for retrieval. 5) Our method is superior to w/o R on all metrics, which reveals that the transformer based encoder can significantly promote the learning of consistent embedding. 6) Our proposed method achieves the best performance in all datasets on various metrics. This reveals that semantic sharing can effectively mitigate the classification degradation problem caused by consistency, which is conducive to learning semantically consistent embedding. Table 1 also delivers the ablation studies. The results reveal that: 1) ST performs worse than our method and joint training methods, for the reason that independent training cannot effectively resolve the gap of modal heterogeneity; 2) IMC performs worse than our method, which indicates that direct inter-modal consistency affects the learning of consistent cross-modal representation; 3) w/o R and w/o T are worse than our method, thereby the self-attention network based on the region segmentation is conducive to learning more discriminative features; 4) w/o C performs superior to w/o L, and the performance degradation of w/o L is serious, because label prediction is critical for learning semantic embedding; 5) our method achieves better performance than Ours+Con, which indicates that consistency loss has truly little effect; and 6) our method achieves the best performances in all datasets on various metrics, this indicates that the transformer based sharing classification model and two prediction tasks benefit the learning of consistent embedding.

4.4 Case Study

Figure 3 shows the qualitative results of sentence retrieval given image queries. Different from alignment-wise methods, we consider that the retrieval results are correct as long as there exist a shared label between retrieval result and the

(a)

A Caucasian hand holding a black cell phone. √
A hand holding a smartphone with a small screen . √
A man holding his Sprint cell phone with the words Upstage across the screen . √
A person holding a cell phone in their hand . √
A woman in blue sweater holding two cellphones while wearing headphones. √

(b)

A couple of big slabs of meat with foot on top of them. √
A sandwich with various toppings next to a sweet potato . √
A plate filled with frensh toast sitting next to a drink . √
A plate with a sandwich and a pickle. √
Wooden spoons laid out across a kitchen table. √

(c)

A person hitting a tennis ball with a tennis racket on a tennis court . √
A toddler hitting the ball with a baseball bat in his backyard . √
A child playing with a plastic bat and ball in a yard next to a garage . √
The tennis player is about to hit a ball with his racket . √
Cars are seen on the street outside a tall building . ×

(d)

A women with a tennis racquet in one hand and a towel in the other . √
A woman is holding a tennis racket and a towel . √
A small black dog standing over a plate of food . ×
A full view of a working office with computers . ×
A woman in a tennis outfit walks on a court holding a towel and a racket . √

Figure 3: (Best viewed in color when zoomed in.) Qualitative results of text retrieval given image queries. For each image query we show the top-5 ranked sentences. We observe that our method retrieves the correct results in the top-ranked sentences.

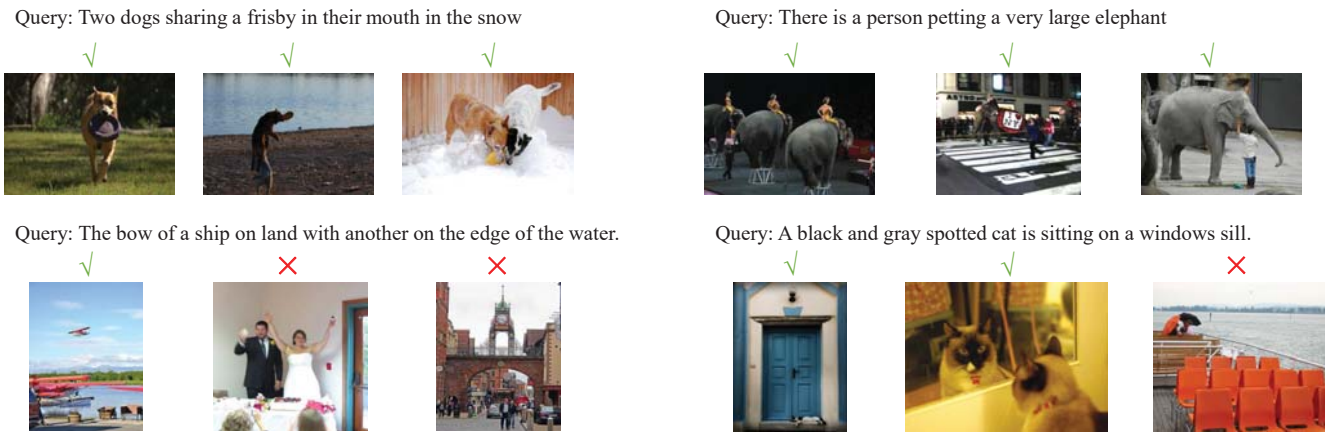


Figure 4: (Best viewed in color.) Qualitative results of image retrieval given sentence queries. For each sentence query, we show the top-3 ranked images, ranking from left to right.

query. Most of the retrieved sentences are correct (showed as green tick). On the other hand, there are semantic incorrect outputs such as c (5), d (3) and (4), possibly due to occasionally poor knowledge. Figure 4 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. In these examples, our model retrieves the ground-truth image successfully.

5 Conclusion

Traditional label-wise cross-modal approaches usually constrain the inter-modal and intra-modal embedding consistency relying on the label ground-truths. However, in this paper, we verify that a direct good shared classification model is better. That is, different modalities can use this shared model to acquire the consistent semantic embedding by directly enhancing classification performance. Specifically, we discard the previous cross-modal consistency loss, and train two modalities interactively by developing a shared transformer encoder to enhance each modal classification performance. The results validate the effectiveness of the proposed semantic sharing.

Acknowledgements

Yang Yang is also with Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education. This research was supported by NSFC (62006118,61773198,61632004), Natural Science Foundation of Jiangsu Province of China under Grant (BK20200460). CCF- Baidu Open Fund (CCF-BAIDU OF2020011), Baidu TIC Open Fund, Southeast University Open Fund (K93-9-2020-01).

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Baltrusaitis *et al.*, 2019] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2):423–443, 2019.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Faghri *et al.*, 2018] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12, 2018.

[Hotelling, 1992] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. 1992.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *ACMMM*, pages 39–43, 2008.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *CVPR*, pages 3270–3278, 2017.

[Karpathy and Fei-Fei, 2017] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *TPAMI*, 39(4):664–676, 2017.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 212–228, 2018.

[Li *et al.*, 2018] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, pages 4242–4251, 2018.

[Li *et al.*, 2019a] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, 2019.

[Li *et al.*, 2019b] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, 2019.

[Li *et al.*, 2020] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[Song and Tan, 2019] Ge Song and Xiaoyang Tan. Sequential learning for cross-modal retrieval. In *ICCV Workshops*, pages 4531–4539, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2013] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, pages 2088–2095, 2013.

[Wang *et al.*, 2016a] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *TPAMI*, 38(10):2010–2023, 2016.

[Wang *et al.*, 2016b] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *CoRR*, 2016.

[Wang *et al.*, 2019] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. CAMP: cross-modal adaptive message passing for text-image retrieval. In *ICCV*, pages 5763–5772, 2019.

[Yang *et al.*, 2015] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *IJCAI*, pages 1033–1039, 2015.

[Yang *et al.*, 2019] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, pages 4092–4098, 2019.

[Yu *et al.*, 2020] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *CoRR*, 2020.

[Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 707–723, 2018.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *TKDE*, 26(8):1819–1837, 2014.

[Zhen *et al.*, 2019] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, pages 10394–10403, 2019.