Cross-Modal Image-Text Retrieval with Semantic Consistency

Hui Chen
School of Software; Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China
jichenhui2012@gmail.com

Guiguang Ding*
School of Software; Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China
dinggg@tsinghua.edu.cn

Zijia Lin
Microsoft Research
Beijing, China
zijlin@microsoft.com

Sicheng Zhao
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
schzhao@gmail.com

Jungong Han
WMG Data Science, University of Warwick, Coventry, UK
jungonghan77@gmail.com

ABSTRACT

Cross-modal image-text retrieval has been a long-standing challenge in the multimedia community. Existing methods explore various complicated embedding spaces to assess the semantic similarity between a given image-text pair, but consider **no/little** about the consistency across them. To remedy this situation, we introduce the idea of semantic consistency for learning various embedding spaces jointly. Specifically, similar to the previous works, we start by constructing two different embedding spaces, namely the image-grounded embedding space and the text-grounded embedding space. However, instead of learning these two embedding spaces separately, we incorporate a semantic consistency constraint in the common ranking objective function such that both embedding spaces can be learned simultaneously and benefit from each other to gain performance improvement. We conduct extensive experiments on three benchmark datasets, *i.e.,* Flickr8k, Flickr30k and MS COCO. Results show that our model outperforms the state-of-the-art models on all three datasets, which can well demonstrate the effectiveness and superiority of the introduction of semantic consistency. Our source code is released at: https://github.com/HuiChen24/SemanticConsistency.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; *Multimedia information systems.*

KEYWORDS

Cross-modal; image-text retrieval; semantic consistency

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351055

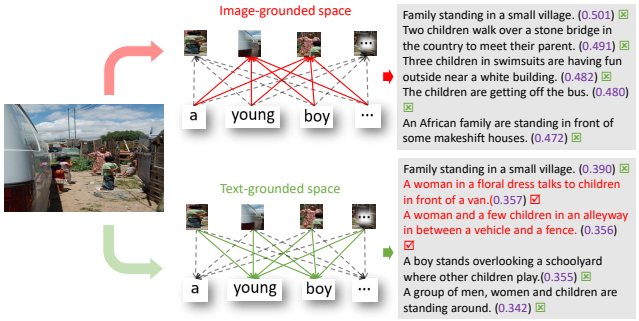

Figure 1: Illustration of the inconsistency of semantic similarity. Correct texts are colored in red. Mismatches are marked with a green cross. Numbers in purple are similarity scores between the query image and each text, calculated in the corresponding embedding spaces (*i.e.,* image-grounded, and text-grounded space).

ACM Reference Format:
Hui Chen, Guiguang Ding, Zijin Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-Modal Image-Text Retrieval with Semantic Consistency. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351055

1 INTRODUCTION

The task of image-text retrieval focuses on constructing a mutual retrieval procedure between images and texts, *i.e.,* retrieving semantically similar images for a given text, or retrieving semantically similar texts for a given image. It is challenging due to the large visual-semantic discrepancy between images and texts [15]. To bridge such a gap, the joint embedding-based method is usually employed, in which information about the images and the texts are generally mapped into a common embedding space. The similarity between images and texts are then calculated by measuring the distance between the corresponding embedding vectors.

Benefiting from the development of deep learning, much progress on the image-text retrieval has been achieved. Karpathy and Fei-Fei [17] enabled each region in a given image to align with all

words of a given text in the common space, and used a simple pooling method to calculate the image-text similarity. Nam et al. [26] learned various dual embedding spaces by combining the attention and memory mechanisms, which explored fine-grained interactions between images and texts. Lee et al. [20] proposed a stacked cross attention, where information from cross-modalities is leveraged as the context to guide each other through the cross attention mechanism, thus enabling the similarity between images and texts to be measured in two different joint embedding spaces. And it achieves the state-of-the-art cross-modal image-text retrieval performance.

Despite the thrilling success in learning different fine-grained joint embedding spaces, existing methods usually learn different spaces independently, resulting in a semantic inconsistency issue. Generally, images and texts are two different representations of scenes in the high-level semantic space. Given an image-text pair, the semantic similarity between them is essentially determined by the semantics they are sharing, which should be resistant to the changes of joint embedding space and the measurement means in the common embedding space. In other words, semantic similarity should be preserved consistently across different embedding spaces. However, existing methods pay **no/little** attention to such semantic consistency.

On the other hand, without modeling the semantic consistency, different embedding spaces may work oppositely. Figure 1 shows a retrieval example derived by models in [20], where the image-grounded embedding space and the text-grounded one are learned independently. And it can be seen that the ground-truth text "A woman and a few children in an alleyway in between a vehicle and a fence." is retrieved successfully in the text-grounded embedding space, but failed in the image-grounded one.

Inspired by the above observations, in this paper, we propose to model the semantic consistency for image-text retrieval, which aims to learn different embedding spaces jointly and consistently. Following previous works [17, 20, 26], we first construct two embedding spaces to capture the shared semantics for any input image-text pair. During training, in addition to assessing the semantic similarity neatly (*i.e.,* similar items should be kept close and dissimilar items should be pushed away in the embedding space), we further regularize that the semantic similarity should be consistent across different embedding spaces. By imposing such a constraint on semantic consistency, we can learn all the embedding spaces jointly such that they can work together to derive consistent semantic similarities for any given image-text pair in both embedding spaces.

We verify the idea of semantic consistency based on [20], which is the state-of-the-art model for image-text retrieval. Specifically, for the image-grounded embedding space, given an image-text pair, each image region in the image would derive a textual context feature via the attention mechanism on each word of the text. Then similarities in the image-grounded embedding space are calculated as the similarities between image regions and their corresponding textual context features. Likewise, for the text-grounded embedding space, each word would get a visual context feature via the attention mechanism on image regions, and similarities would be derived between words and their corresponding visual context features. In addition to the commonly used ranking objective, we learn both embedding spaces jointly under the constraint of semantic consistency, which minimizes the deviation of the corresponding similarities measured in both embedding spaces. Note that both embedding spaces are built on top of the same image features and the same word features, which can help to capture the shared semantic information between images and texts to boost the similarities assessed in both embedding spaces.

The contributions of our work are summarized as follows:

- We advocate the use of "semantic consistency" to keep the consistent semantic similarity among images and sentences in various latent embedding spaces for the task of image-text retrieval.
- We successfully incorporate the semantic consistency constraint into the ranking objective such that learning different embedding spaces (*e.g.,* image-grounded and text-grounded embedding spaces) jointly and consistently becomes possible.
- We conduct extensive experiments and analyses on three benchmark datasets, *i.e.,* Flickr8k, Flickr30k and MS COCO. Experimental results well demonstrate that the proposed semantic consistency works effectively and our method achieves the state-of-the-art performance.

2 RELATED WORK

Recently, there has been much interest in the task of image-text retrieval [9, 13, 18]. Kiros et al. [18] made the early attempt to learn cross-modality representations with a hinge-based triplet ranking loss, where deep Convolutional Neural Networks (CNNs) are used to encode images and Recurrent Neural Networks (RNNs) are adopted to encode texts. Faghri et al. [9] made a successful step toward revising the commonly used triplet loss function by leveraging hard negatives and yielded significant improvement. To enhance the cross-modality feature embedding learning, adversarial objectives are proposed in [12, 29, 33] to distinguish different modalities or instances, and reasonable performance improvement are obtained.

To explore more powerful matching models, researchers have also proposed various methods considering the latent vision-language correspondence at a more fine-grained level. Karpathy and Fei-Fei [17] adopted an R-CNN [11] to detect image regions at the object level, and meanwhile, extracted region-level image representations. Afterwards, the similarity scores over all possible region-word pairs are aggregated to infer the image-text similarity. Niu et al. [27] presented a hierarchical model that maps not only full texts and whole images but also phrases within texts and salient regions within images into a shared multi-modal embedding space.

Recently, the attention mechanism [32, 35] was introduced into the task of image-text retrieval and boosted the development of this area greatly. Huang et al. [14] proposed a multi-modal LSTMs where a context-modulated attention scheme was developed to selectively attend to a pair of instances appearing in both image and text. Nam et al. [26] proposed to capture the fine-grained interplay between image and text by performing a dual attention network through multiple steps. Lee et al. [20] proposed a stacked cross attention to discover the full latent alignments between regions in images and words in texts.

Our work is based on the model in [20], as it is the state-of-the-art for image-text retrieval. Apart from the proposed semantic consistency constraint, our work is also substantially distinguished

from [20] in the training process. Specifically, instead of learning each embedding space independently, we train solely one model to learn different embedding spaces simultaneously. We argue that, with the semantic consistency constraint, the joint learning strategy can enhance the representation of cross-modal data, and thus boost the retrieval performance.

Another line of research in this field is about the hash-based cross-modal retrieval, where hash codes are learned for the cross-modal data to enable low cost but fast retrieval [3, 7, 21, 22, 36, 38]. However, hash codes tend to hinder the information representing in the latent embedding space due to their binary features. While our work focus on the cross-modal retrieval with real value features which can better represent the cross-modal information and enhance the correlation among data. With the quantization or binarization processing, our method may be extended to remedy the hash-based cross-modal retrieval problem, which leaves for the future work.

3 METHODOLOGY

In this section, we describe the proposed model with semantic consistency for image-text retrieval. As illustrated in Figure 2, we adopt a convolutional neural network as an image encoder to convert image regions into features. Meanwhile, we utilize a recurrent neural network as a text encoder to carry out the same task for each word of the text. Given the image/text features, we construct the image-grounded embedding space and text-grounded embedding space, respectively. Finally, we engage both embedding spaces with a semantic consistency constraint in the objective function.

This section is structured as follows: we first introduce the feature encoders in Section 3.1. Then we describe how to construct the image-grounded embedding space and the text-grounded embedding space in Section 3.2. Afterwards we elaborate the proposed semantic consistency constraint over these two embedding spaces in Section 3.3. Finally the objective function for training is provided in Section 3.4.

3.1 Feature Encoders

Let $\{(I_i, S_i)\}_{i=1}^{N} \sim \mathcal{D}$ denote the dataset containing N image-text pairs, where I and S are the images and the texts, respectively. Since it is hard to connect raw images and texts directly, we leverage feature encoders to encode them into feature vectors. Then the image-grounded and text-grounded embedding spaces are constructed based on these latent features of images and texts.

Image Encoder. Benefiting from the development of deep learning, the problem of extracting effective image features has been studied thoroughly nowadays. Convolutional neural networks (CNNs), pre-trained on large-scale datasets like ImageNet [6] or Visual Genome [19], have been proved to be powerful in capturing discriminative information for images. Therefore, following [20], we adopt a Faster R-CNN [31], pretrained by [1] on Visual Genome, to extract discriminative region-level image features for given images.

Specifically, given an image I_i, Faster R-CNN first detects objects in the image using bounding boxes. For a selected region indexed by j, which corresponds to a detected bounding box, a feature vector $f_{i,j}$ is defined as the mean-pooled convolutional feature from this region. To adapt to the benchmark datasets, we transform $f_{i,j}$ to a d-dimensional vector $v_{i,j}$ by a fully-connected layer:

$$v_{i,j} = W_v f_{i,j} + b_v \tag{1}$$

where W_v and b_v are to-be-learned parameters.

Finally, we obtain a set of feature vectors for each image I_i denoted as $V_i = \{v_{i,j} | j = 1, ..., k, v_{i,j} \in \mathbb{R}^d\}$, where each $v_{i,j}$ encodes a salient region and k is the number of detected regions in I_i.

Text Encoder. Recently, a group of word embedding methods, such as word2vec [24], glove [30], fasttext [2, 16] and so on, have been proposed to represent the information of words, and exhibit impressive performance in many tasks in natural language processing domains. And usually, recurrent network networks (RNNs) are applied to encode words for enhancing the context information among words in the text. As [20], here we employ a bi-directional gated recurrent unit (GRU, a variant of RNNs) [4] as the text encoder to obtain the feature vectors for each word.

Specifically, for a text S_i, we firstly denote each word in S_i with a d'-dimensional one-hot vector $w_{i,j}$, where d' is the size of all words and in $w_{i,j}$ only the position corresponding to the word is assigned as 1, while others as 0. Then we embed the word into a contiguous space through an embedding matrix W_e: $x_{i,j} = W_e w_{i,j}, \forall j \in [1, n]$ where W_e is a to-be-learned parameter and n is the length of a text.

After that, we adopt a bi-directional GRU to summarize information from both forward and backward directions in the text S_i. We use $\overrightarrow{h}_{i,j}$ and $\overleftarrow{h}_{i,j}$ to denote the hidden states from the forward GRU and the backward GRU:

$$\begin{aligned} \overrightarrow{h}_{i,j} &= \overrightarrow{GRU}(x_{i,j}, \overrightarrow{h}_{i,j-1}); \\ \overleftarrow{h}_{i,j} &= \overleftarrow{GRU}(x_{i,j}, \overleftarrow{h}_{i,j+1}) \end{aligned} \tag{2}$$

Then the final feature vector $t_{i,j}$ for the word $w_{i,j}$ is computed by averaging both hidden states:

$$t_{i,j} = \frac{\overrightarrow{h}_{i,j} + \overleftarrow{h}_{i,j}}{2} \tag{3}$$

Finally, we obtain a set of feature vectors for each text S_i denoted as $T_i = \{t_{i,j} | j = 1, ..., n, t_{i,j} \in \mathbb{R}^d\}$, where each $t_{i,j}$ encodes a word information, and shares the same dimension as $v_{i,j}$ in Equation (1).

3.2 Image-text Matching

In this section, we describe how to assess the similarity given two sets of feature vectors, *i.e.,* image features $V_i = \{v_{i,j} | j = 1, ..., k, v_{i,j} \in \mathbb{R}^d\}$ for an image I_i and word features $T_i = \{t_{i,j} | j = 1, ..., n, t_{i,j} \in \mathbb{R}^d\}$ for a text S_i. To explore the fine-grained relatedness between V_i and T_i, we adopt two embedding spaces, *i.e.,* image-grounded embedding space and text-grounded embedding space, where image regions and words are used as the context to each other when inferring the similarity. Unlike [20], we couple both embedding spaces together to build an end-to-end system. To ease the explanation, we discard the subscript i and use I and S to denote an image and a text, respectively, and use V and T to denote their corresponding feature vectors.

Image-grounded Embedding Space. Image-grounded embedding space attempts to project the word features into the latent space grounded on the image features. First, for each image region,

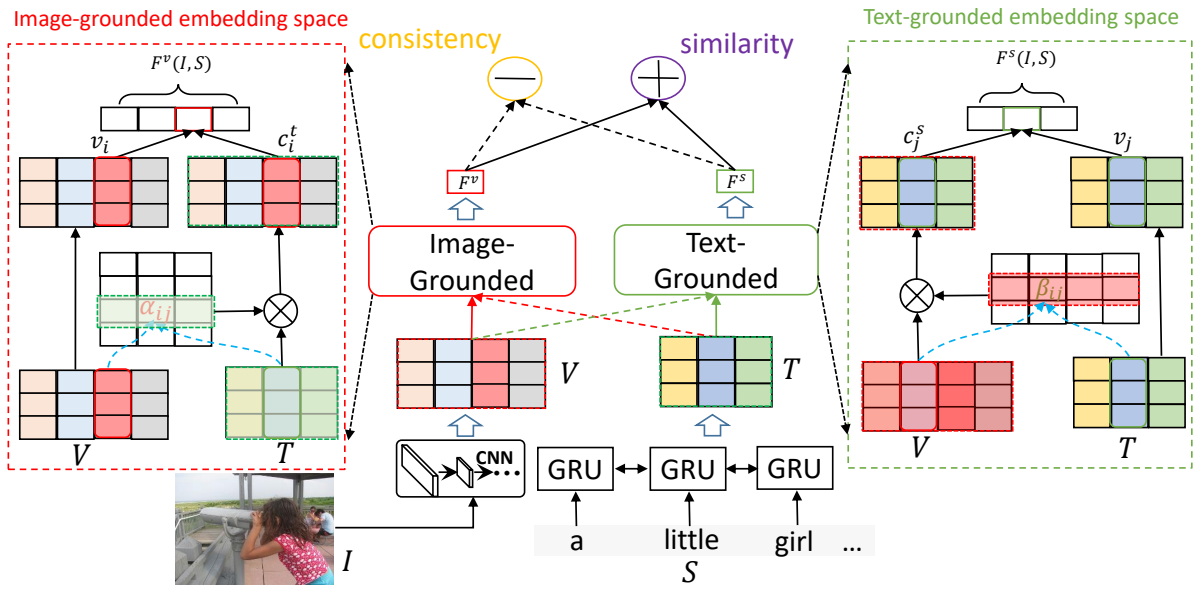

Figure 2: Overview of the proposed model with semantic consistency preserved. The input image and text are denoted as I and S, respectively. Columns in different colors indicate feature vectors w.r.t different image regions and text words. Best viewed in color.

it would attend to each word in the text via the attention mechanism to derive a weight. Then a textual context feature is composed of the weighted summation of all word features in the text. Afterwards, a cosine similarity score is computed using the feature vector of the image region and its corresponding textual context feature vector. Finally, all those similarity scores are aggregated to derive the similarity between the image and the text.

Specifically, given an image feature set V with k region-level feature vectors, w.r.t. an image I, and a text feature set T with n feature vectors, w.r.t. a text S, a cosine similarity matrix is firstly computed to discover relationships among all possible region-word pairs:

$$s_{ij} = \frac{v_i^T t_j}{||v_i|| \cdot ||t_j||}, \forall i \in [1,k], \forall j \in [1,n] \tag{4}$$

where v_i is the feature vector corresponding to the i-th region in the image I, and t_j is the feature vector of the j-th word in the text S. s_{ij} measures the similarity between the i-th region and the j-th word.

As [20], we further normalize the similarity matrix along the column dimension as: $\bar{s}_{ij} = \frac{\sigma(s_{ij})}{\sqrt{\sum_{i=1}^{k} \sigma(s_{ij})^2}}$ where σ is a non-linear function, *i.e.*, $relu()$ here.

Then we compose a textual context feature vector, c_i^t, for the region v_i by combining all word feature vectors, *i.e.*, T, through the attention mechanism: $c_i^t = \sum_{j=1}^{n} \alpha_{ij} t_j$, where

$$\alpha_{ij} = \frac{exp(\lambda_1 \bar{s}_{ij})}{\sum_{j=1}^{n} exp(\lambda_1 \bar{s}_{ij})} \tag{5}$$

where λ_1 is an inverse temperature of the softmax function [5].

After that, we compute the relevance score $R(v_i, c_i^t)$ between the region feature v_i and its corresponding context feature c_i^t by the cosine similarity: $R(v_i, c_i^t) = \frac{v_i^T c_i^t}{||v_i|| \cdot ||c_i^t||}$.

Finally, the similarity between the image I and the text S is calculated as the average of all relevance scores:

$$F^v(I,S) = \frac{1}{k} \sum_{i=1}^{k} R(v_i, c_i^t) \tag{6}$$

Text-grounded Embedding Space. Opposite to the image-grounded embedding space, the text-grounded embedding space attempts to project the image region features into the latent space grounded on the text features. Then the similarity between a given image and a given text is calculated in this embedding space. Specifically, after obtaining the cosine similarity matrix by Equation (4), where each entry s_{ij} assesses the relationship between the region feature v_i and the word feature t_j, we further normalize the similarity matrix along the row dimension, $\bar{s}'_{ij} = \frac{\sigma(s_{ij})}{\sqrt{\sum_{j=1}^{n} \sigma(s_{ij})^2}}$ where σ is $relu()$.

Then we compose a visual context feature, c_j^v for the word t_j by combining all region features, *i.e.*, V, through the attention mechanism: $c_j^v = \sum_{i=1}^{k} \beta_{ij} v_i$, where

$$\beta_{ij} = \frac{exp(\lambda_2 \bar{s}'_{ij})}{\sum_{i=1}^{k} exp(\lambda_2 \bar{s}'_{ij})} \tag{7}$$

where λ_2 is an inverse temperature of the softmax function.

After that, we compute the relevance score $R(t_j, c_j^v)$ between the word feature t_j and its corresponding visual context feature c_j^v by the cosine similarity: $R(t_j, c_j^v) = \frac{t_j^T c_j^v}{||t_j|| \cdot ||c_j^v||}$.

Finally, the similarity score between the image I and the text S is calculated as the average of all relevance scores:

$$F^s(I,S) = \frac{1}{n}\sum_{j=1}^{n} R(t_j, c_j^v) \tag{8}$$

Integrated similarity score. The integrated similarity score for a given image-text pair is defined as the summation of measured similarities in the image-grounded embedding space and the text-grounded embedding space:

$$F(I,S) = F^v(I,S) + F^s(I,S) \tag{9}$$

3.3 Learning with Semantic Consistency

Given an image and a text, the similarity between them is essentially dependent on the semantics they shared in the high-level semantic space. The portion of shared semantics can be regarded as shared properties of the image and the text, which should not vary in different embedding spaces. Therefore, we propose to jointly learn embedding spaces with semantic consistency, which requires that semantic similarity assessed by various embedding spaces should be consistent.

Specifically, given an image I and a text S, as described above, two kinds of similarity scores are defined as $F^v(I,S)$ and $F^s(I,S)$, which indicate similarities assessed in the image-grounded embedding space and the text-grounded embedding space, respectively. Then we define the dispersion $D(I,S)$ between them as:

$$D(I,S) = \delta(F^v(I,S), F^s(I,S)) \tag{10}$$

Here, the function $\delta()$ can be calculated in different ways, *e.g.,* Kullback-Leibler divergence and mean-square error (MSE). In our paper, we simply adopt MSE to define the dispersion. And we leave other calculation ways as future works. Namely, $D(I,S)$ is specified as follows in this paper.

$$D(I,S) = \left(F^v(I,S) - F^s(I,S)\right)^2 \tag{11}$$

During training, for a mini-batch, *i.e.,* $\{(I_i,S_i)\}_{i=1}^{N_b} \sim \mathcal{D}$ where N_b is the size of the mini-batch, we compute the semantic consistency loss by exploring all dispersions between any image-text pair (I_i,S_j), as shown below:

$$\mathcal{L}_{con} = \sum_{i=1}^{N_b}\sum_{j=1}^{N_b} D(I_i,S_j) \tag{12}$$

3.4 Loss Function for Training

Triplet ranking loss is commonly used for retrieval problems. Previous approaches [18] usually employed a hinge-based triplet ranking loss, in which a positive pair would get a higher similarity score than a negative pair by a margin Δ at least. In our case, given the mentioned training data, each image-text pair (I_i,S_i) is treated as a positive pair, as I_i and S_i are relevant and coupled. And we can build negative pairs as follows: 1) for I_i, we can use any $S_j(j \neq i)$ to build a negative pair (I_i,S_j); 2) for S_i, we can use any $I_j(j \neq i)$ to build another negative pair (I_j,S_i). Then we can derive the ranking loss function as follows.

$$\begin{aligned}\mathcal{L}_{rank}(I_i,S_i) = &\sum_{j\neq i}[\Delta - F(I_i,S_i) + F(I_i,S_j)]_+ \\ &+ \sum_{j\neq i}[\Delta - F(I_i,S_i) + F(I_j,S_i)]_+\end{aligned} \tag{13}$$

where $[x]_+ = max(x,0)$ and $F(I,S)$ measures the similarity between I and S as Equation (9).

In practice, for computational efficiency, instead of comparing with all negative samples in the training set, it is common to consider only the hard negatives, *i.e.,* the negatives closest to each training query [9], within a mini-batch. And in this paper, following [9], we focus on the hardest negatives in a mini-batch to compute the ranking loss for the positive pair. Specifically, during training, for any positive image-text pair *i.e.,*(I_i,S_i), we firstly search the hardest negatives given by:

$$\begin{aligned} I_i^h &= \arg\max_{I_j, j\neq i} F(I_j,S_i) \\ S_i^h &= \arg\max_{S_j, j\neq i} F(I_i,S_j)\end{aligned} \tag{14}$$

where I_i^h forms the hardest negative pair with S_i and S_i^h forms the hardest negative pair with I_i. Then the refined ranking loss $\mathcal{L}_{rank}^h(I_i,S_i)$ is given as:

$$\begin{aligned}\mathcal{L}_{rank}^h(I_i,S_i) = &[\Delta - F(I_i,S_i) + F(I_i,S_i^h)]_+ \\ &+[\Delta - F(I_i,S_i) + F(I_i^h,S_i)]_+\end{aligned} \tag{15}$$

Together with the proposed semantic consistency constraint described in Equation (12), for a mini-batch data, *i.e.,* $\{(I_i,S_i)\}_{i=1}^{N_b} \sim \mathcal{D}$, the total training objective is defined as:

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N_b}\mathcal{L}_{rank}^h(I_i,S_i) + \lambda_{con} * \sum_{i=1}^{N_b}\sum_{j=1}^{N_b} D(I_i,S_j) \\ &= \mathcal{L}_{rank}^h + \lambda_{con} * \mathcal{L}_{con}\end{aligned} \tag{16}$$

where λ_{con} is a trade-off factor that balances the ranking loss and the semantic consistency loss.

Note that unlike [20] where the image-grounded embedding space and the text-grounded embedding space are learned independently using separate models, we utilize one model to learn both embedding spaces jointly and consistently, sharing image encoders and text encoders. And besides the ranking loss, we encourage both embedding spaces to be consistent in terms of the measurement of semantic similarities, which is supposed to help preserve more shared semantics between images and texts within embedding spaces.

4 EXPERIMENT

We conduct extensive experiments and analyses on three benchmark datasets to verify the performance of the proposed method. We also compare our model with state-of-the-art models on all three datasets. In this section, we first describe the details about the datasets and the adopted evaluation metrics in Section 4.1, followed by the description on the implementation details in Section 4.2. Then we report the comparison results with state-of-the-art models on three benchmark datasets in Section 4.3. And we also conduct parameter analysis to investigate the effect of the hyper-parameters,

Table 1: Comparison of the cross-modal image-text retrieval performances in terms of Recall@K (*R@K*) on Flickr8k. * indicates the performance of an ensemble model. - means unknown results. We test SCAN [20] with codes provided by the authors, denoted as SCAN(ours) and SCAN*(ours).

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DeViSE [10]	4.8	16.5	27.3	5.9	20.1	29.6
DVSA [17]	16.5	40.6	54.2	11.8	32.1	44.7
m-CNN [23]	24.8	53.7	67.1	20.3	47.6	61.7
HM-LSTM [27]	27.7	-	68.6	24.4	-	68.1
SCAN(ours)	46.8	77.8	86.3	34.4	64.5	76.1
Ours	**53.7**	**83.0**	**90.7**	**39.6**	**68.4**	**79.2**
SCAN*(ours)	50.3	80.8	88.5	36.8	66.8	78.3
Ours*	**56.2**	**84.6**	**92.3**	**41.2**	**70.4**	**80.8**

Table 2: Comparison of the cross-modal image-text retrieval performances in terms of Recall@K (*R@K*) on Flickr30k. * indicates the performance of an ensemble model. - means unknown results.

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [17]	22.2	48.2	61.4	15.2	37.7	50.5
HM-LSTM [27]	38.1	-	76.5	27.7	-	68.8
SM-LSTM [14]	42.5	71.9	81.5	30.2	60.4	72.3
Webly [25]	47.4	-	85.9	35.2	-	74.8
2WayNet [8]	49.8	67.5	-	36.0	55.6	-
VSE++ [9]	52.9	-	87.2	39.6	-	79.5
DAN [26]	55.0	81.8	89.0	39.4	69.2	79.1
DPC [37]	55.6	81.9	89.5	39.1	69.2	80.9
SCO [15]	55.5	82.0	89.3	41.1	70.5	80.1
SCAN [20]	67.9	89.0	94.4	43.9	74.2	82.8
Ours	**68.5**	**90.8**	**95.8**	**53.0**	**79.1**	**86.1**
SCAN* [20]	67.4	90.3	95.8	48.6	77.7	85.2
Ours*	**69.7**	**91.7**	**96.4**	**54.0**	**79.7**	**87.2**

i.e., λ_{con} in Section 4.4. Finally, we present some retrieval examples to reveal the advantage of introducing the semantic consistency in Section 4.5.

4.1 Datasets and Evaluation Metric

We present experiments on three benchmark datasets for image-text retrieval, *i.e.,* Flickr8k, Flickr30k and MS COCO, to evaluate the performance of our proposed method and make comparisons with other state-of-the-art approaches.

Flickr8k. Flickr8k contains 8,000 images. And each image is provided with 5 sentences. We adopt its standard training, validation, and testing split as [23, 27].

Flickr30k. Flickr30k contains 31,000 images with totally 158,915 English sentences. Each image is annotated with 5 sentences by Amazon Mechanical Turk workers. Here, we follow the dataset split provided in [20, 25], where 29,000 images are used for training and 1,000 images for validation, while the remaining 1,000 images are used as the test set.

MS COCO. MS COCO is a large-scale image description dataset, which has about 123K images and each image comes with at least 5 sentences. As previous works [20, 25], we split the dataset into a training set, a validation set and a test set, whose numbers of images are 113,287, 5,000 and 5,000, respectively.

Evaluation Metric. We use the same evaluation metrics for the image-text retrieval task as in previous works [9, 20, 25]. Namely, we adopt Recall at K (R@K) to measure the performance of text retrieval given image query and image retrieval given text query. R@K is defined as the percentage of test samples for which the correct items are retrieved within the top-K nearest items to the query [25]. We report R@1, R@5, and R@10 for both benchmark datasets as in [20].

4.2 Implementation Details

We use Pytorch v1.0 [28] to implement our model. For fair comparisons, we adopt the same experiment settings as [20]. Specifically, for all datasets, we use a Faster R-CNN model in conjunction with ResNet-101 pre-trained on Visual Genome [19] as the image encoder. We select the top 36 regions of interest (ROIs) with the highest detection confidence scores. We use average pooling to extract salient features, ending up with a 2,048 dimensional feature vector. A bi-directional GRU with one layer is adopted as the text encoder. The dimensionality of the hidden state in the GRU is set as

Table 3: Comparison of the cross-modal image-text retrieval performances in terms of Recall@K (*R@K*) on MS COCO. * indicates performance of an ensemble model. - means unknown results.

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [17]	38.4	69.9	80.5	27.4	60.2	74.8
HM-LSTM [27]	43.9	-	87.8	36.1	-	86.7
SM-LSTM [14]	53.2	83.1	91.5	40.7	75.8	87.4
2WayNet [8]	55.8	75.2	-	39.7	63.3	-
Webly [25]	61.5	-	96.1	46.3	-	89.4
VSE++ [18]	64.6	-	95.7	52.0	-	92.0
DPC [37]	65.6	89.8	95.5	47.1	79.9	90.0
CHAIN-VSE [34]	59.4	88.0	94.2	43.5	79.8	90.2
GXN [12]	68.5	-	97.9	56.6	-	94.5
SCO [15]	69.9	92.9	97.5	56.7	87.5	**94.8**
SCAN [20]	70.9	94.5	97.8	56.4	87.0	93.9
Ours	**73.0**	**94.9**	98.2	**58.9**	**88.1**	94.4
SCAN* [20]	72.7	94.8	**98.4**	58.8	88.4	94.8
Ours*	**73.8**	**95.3**	98.3	**59.9**	**88.9**	**94.9**

1,024, the same as the dimensionality of either image-grounded or text-grounded embedding space. The dimensionality of the word embeddings that are fed into the GRU is set as 300. During training, the margin of the triplet loss, *i.e.,* Δ in Equation (15) is set as 0.2. The inverse temperature factors, *i.e.,* λ_1 in Equation (5) and λ_2 in Equation (7), are all set as 9. To avoid the gradient explosion, we set the maximum gradient norm as 2.0 for gradient clipping. All models are trained in a mini-batch with a size of 128 using the Adam optimizer. Following previous works [9, 20], we validate the performance of each retrieval model on the validation set every epoch, and save the best version on the validation set according to the summation of six recall metrics, which is later evaluated on the test set. As for the trade-off factor, *i.e.,* λ_{con} in Equation (16), we search for the best value via the grid search algorithm with predefined value ranges on the validation set. We will discuss the effect of λ_{con} in Section 4.4.

Table 4: The effect of λ_{con} in terms of Recall@K (*R@K*). SCAN [20] means NO semantic consistency constraint is introduced, and thus λ_{con} is not needed.

Dataset	λ_{con}	Text Retrieval		Image Retrieval	
		R@1	R@10	R@1	R@10
Flickr8k	0.0	52.2	90.0	38.1	78.4
	0.3	53.7	**91.5**	**39.6**	**79.2**
	0.6	**54.1**	90.2	38.7	79.0
Flickr30k	0.0	67.5	95.0	50.3	85.0
	1.0	**68.5**	**95.8**	**53.0**	**86.1**
	2.0	67.1	95.4	51.2	86.1
MSCOCO	0.0	71.1	97.8	57.6	94.0
	0.3	**73.0**	**98.2**	**58.9**	**94.4**
	0.6	72.6	98.2	58.4	94.3

4.3 Comparison with State-of-the-Art

To verify the effectiveness of the proposed method, we conduct experiments and compare our model with state-of-the-art models for the task of image-text retrieval. Note that not all compared approaches were verified on all three datasets and reported all evaluation metrics as ours in their papers. Here we directly cite reported performances from respective papers when available. And following [25], if performances of multiple models are reported in a paper, we select the best one for comparison.

Results. Comparison results are reported in Table 1, Table 2 and Table 3 for Flickr8k, Flickr30k and MS COCO , respectively. Experimental results show that the proposed method can obtain substantial performance improvement compared with state-of-the-art methods, which is attributed to the introduction of semantic consistency.

On Flickr8k, our single model can outperform all baselines on all metrics. Compared with the best baseline SCAN (*i.e.,* the one that our model is based on), our single model (*i.e.,* Ours) can achieve maximal performance improvements of 6.9% (R@1) and 5.2% (R@1) for the text retrieval and the image retrieval, respectively. In contrast to ensemble SCAN*, our ensemble model (*i.e.,* Ours*) can obtain a maximal performance improvement of 5.9% (R@1) for the text retrieval, and 4.4% (R@1) for the image retrieval.

On Flickr30k, our model can also achieve the best performance in terms of all metrics. Compared with SCAN, our single model can obtain maximal improvements of 1.8% (R@1) and 9.1% (R@1) for two retrieval tasks, respectively. While compared with SCAN*, the maximal improvements obtained by our ensemble model are 2.3% (R@1) and 5.4% (R@1), for two retrieval tasks, respectively.

On MS COCO, our model achieves superior performance than all baseline models in terms of R@1 and R@5 for both retrieval tasks, and meanwhile achieves comparable performances as the best baselines in terms of R@10. Particularly, compared with SCAN, our single model can gain maximal performance improvements of 2.1% (R@1) and 2.5% (R@1) for both tasks, respectively. And compared with the ensemble SCAN*, our ensemble model can also gain substantial performance improvement in terms of most metrics.

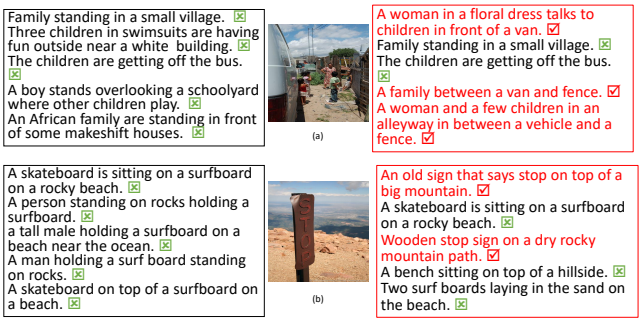

Figure 3: Examples of the text retrieval results by the best baseline model SCAN (in black boxes) and our single model (in red boxes). Results marked in red are correct, while those marked with a green cross are incorrect.

Figure 4: Examples of the image retrieval results. Due to the space limitation, for each query text, we only show top 3 images retrieved by the best baseline model SCAN (in the left) and our single model (in the right). Results marked in red are correct, while those marked in green are incorrect.

4.4 Parameter Analysis

The parameter λ_{con} in Equation (16) determines to what extent the model is encouraged to keep embedding spaces consistent. Here we conduct experiments to analyse its effects on cross-modal retrieval performances, by varying λ_{con} in predefined value ranges on all benchmark datasets. We can see that for all datasets, an appropriate λ_{con} can lead to a better performance. And the best λ_{con} is 0.3 ,1.0 and 0.3 for Flickr8k, Flickr30k and MS COCO, respectively. We can also see that models with various λ_{con} can obtain better performance than $\lambda_{con} = 0$ (*i.e.,* no semantic consistency constraint), which further demonstrates the effectiveness of the introduction of semantic consistency.

4.5 Qualitative Analysis

Figure 3 and Figure 4 show some examples of text retrieval results and image retrieval results, comparing the best baseline model SCAN and our single model. We can see that, with the introduction of the semantic consistency, the proposed method can fix some failed cases of SCAN, and thus further improve the retrieval performance.

Figure 5 shows some examples of alignment results between image regions and text words. We can see that with the proposed semantic consistency, alignments between text words and image regions in the latent embedding space can be enhanced. Compared to the baseline model, such alignments can be better captured by our model and thus our model can further gain performance improvement.

Figure 5: Examples of alignment results produced by our model (middle) and our baseline model, i.e. SCAN (right), given the raw image (left). We only show alignment results of one word (in red and bold) with respect to the image regions in the text-grounded embedding space. More results can be found at https://huichen24.github.io/SemanticConsistency/

5 CONCLUSION

In this paper, we introduce the semantic consistency for cross-modal image-text retrieval. We attempt to boost the retrieval performance by encouraging the semantic similarities assessed in different embedding spaces to be consistent. And thus we propose a semantic consistency constraint enhancing the common ranking objective function. With the semantic consistency constraint, we also enable different embedding spaces to be learned in one model simultaneously and consistently. We verify our proposed method through extensive experiments and analyses on three benchmark datasets, *i.e.,* Flickr8k, Flickr30k and MS COCO. Experimental results show that the proposed method can substantially improve the retrieval performance and outperform state-of-the-art approaches.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (No. 2018YFC0806900) and the National Natural Science Foundation of China (Nos. 61571269, 61701273).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. 2017. Transitive hashing network for heterogeneous multimedia retrieval. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[5] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*. 577–585.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. 2016. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing* 25, 11 (2016), 5427–5440.

[8] Aviv Eisenschtat and Lior Wolf. 2017. Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4601–4611.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[12] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7181–7189.

[13] Christian Henning and Ralph Ewerth. 2018. Estimating the information gap between textual and visual representations. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 43–56.

[14] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2310–2318.

[15] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6163–6171.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 427–431.

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, and Justin et.al. Johnson. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.

[20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[21] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang. 2016. Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE transactions on cybernetics* 47, 12 (2016), 4342–4355.

[22] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3864–3872.

[23] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*. 2623–2631.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[25] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E. Papalexakis, and Amit K. Roy-Chowdhury. 2018. Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. 1856–1864.

[26] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.

[27] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. 2017. Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1899–1907. https://doi.org/10.1109/ICCV.2017.208

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[29] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2017. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning. *arXiv preprint arXiv:1710.05106* (2017).

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[33] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 154–162.

[34] Jônatas Wehrmann and Rodrigo C. Barros. 2018. Bidirectional Retrieval Made Simple. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*. 2048–2057.

[36] Xin Zhao, Guiguang Ding, Yuchen Guo, Jungong Han, and Yue Gao. 2017. TUCH: turning cross-view hashing into single-view hashing via generative adversarial nets. IJCAI.

[37] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding with Instance Loss. *arXiv preprint arXiv:1711.05535* (2017).

[38] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 415–424.