# Consensus-Aware Visual-Semantic Embedding for Image-Text Matching

Haoran Wang[1*†], Ying Zhang[2†], Zhong Ji[1‡], Yanwei Pang[1], and Lin Ma[2]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
{haoranwang,Jizhong,pyw}@tju.edu.cn
[2] Tencent AI Lab, Shenzhen, China
yinggzhang@tencent.com  forest.linma@gmail.com

**Abstract.** Image-text matching plays a central role in bridging vision and language. Most existing approaches only rely on the image-text instance pair to learn their representations, thereby exploiting their matching relationships and making the corresponding alignments. Such approaches only exploit the superficial associations contained in the instance pairwise data, with no consideration of any external commonsense knowledge, which may hinder their capabilities to reason the higher-level relationships between image and text. In this paper, we propose a Consensus-aware Visual-Semantic Embedding (CVSE) model to incorporate the consensus information, namely the commonsense knowledge shared between both modalities, into image-text matching. Specifically, the consensus information is exploited by computing the statistical co-occurrence correlations between the semantic concepts from the image captioning corpus and deploying the constructed concept correlation graph to yield the consensus-aware concept (CAC) representations. Afterwards, CVSE learns the associations and alignments between image and text based on the exploited consensus as well as the instance-level representations for both modalities. Extensive experiments conducted on two public datasets verify that the exploited consensus makes significant contributions to constructing more meaningful visual-semantic embeddings, with the superior performances over the state-of-the-art approaches on the bidirectional image and text retrieval task. Our code of this paper is available at: https://github.com/BruceW91/CVSE.

**Keywords:** Image-text matching, visual-semantic embedding, consensus
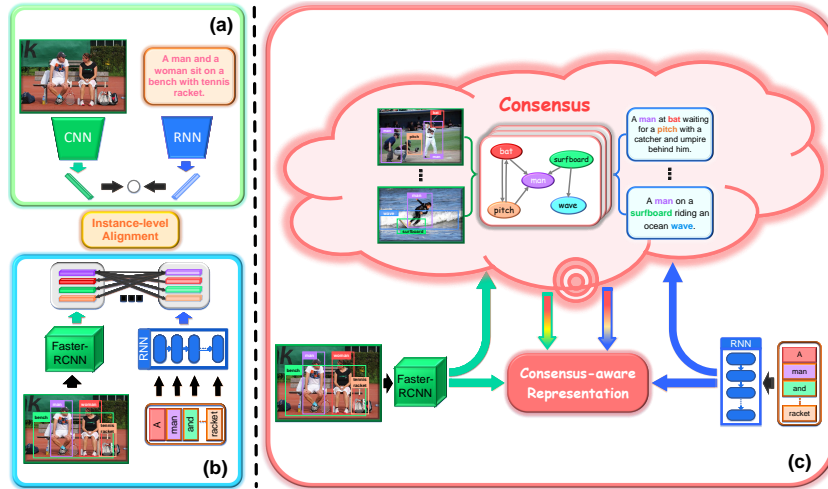
## 1 Introduction

Vision and language understanding plays a fundamental role for human to perceive the real world, which has recently made tremendous progresses thanks to

---

*Work done while Haoran Wang was a Research Intern with Tencent AI Lab.
†indicates equal contribution.
‡Corresponding author

**Fig. 1.** The conceptual comparison between our proposed consensus-aware visual-semantic embedding (CVSE) approach and existing instance-level alignment based approaches. (a) instance-level alignment based on image and text global representation; (b) instance-level alignment exploiting the complicated fragment-level image-text matching; (c) our proposed CVSE approach.

the rapid development of deep learning. To delve into multi-modal data comprehending, this paper focuses on addressing the problem of image-text matching [27], which benefits a series of downstream applications, such as visual question answering [2, 28], visual grounding [4, 35, 47], visual captioning [41, 42, 48], and scene graph generation [5]. Specifically, it aims to retrieve the texts (images) that describe the most relevant contents for a given image (text) query. Although thrilling progresses have been made, this task is still challenging due to the semantic discrepancy between image and text, which separately resides in heterogeneous representation spaces.

To tackle this problem, the current mainstream solution is to project the image and text into a unified joint embedding space. As shown in Figure 1 (a), a surge of methods [10, 21, 30, 43] employ the deep neural networks to extract the global representations of both images and texts, based on which their similarities are measured. However, these approaches failed to explore the relationships between image objects and sentence segments, leading to limited matching accuracy. Another thread of work [18, 23] performs the fragment-level matching and aggregates their similarities to measure their relevance, as shown in Figure 1 (b). Although complicated cross-modal correlations can be characterized, yielding satisfactory bidirectional image-text retrieval results, these existing approaches only rely on employing the image-text instance pair to perform cross-modal retrieval, which we name as instance-level alignment in this paper.

For human beings, besides the image-text instance pair, we have the capability to leverage our commonsense knowledge, expressed by the fundamental semantic concepts as well as their associations, to represent and align both images and texts. Take one sentence "`A man on a surfboard riding on a ocean wave`" along with its semantically-related image, shown in Figure 1 (c), as an example. When "`surfboard`" appears, the word "`wave`" will incline to appear with a high probability in both image and text. As such, the co-occurrence of "`surfboard`" and "`wave`" as well as other co-occurred concepts, constitute the commonsense knowledge, which we refer to as *consensus*. However, such consensus information has not been studied and exploited for the image-text matching task. In this paper, motivated by this cognition ability of human beings, we propose to incorporate the consensus to learn visual-semantic embedding for image-text matching. In particular, we not only mine the cross-modal relationships between the image-text instance pairs, but also exploit the consensus from large-scale external knowledge to represent and align both modalities for further visual-textual similarity reasoning.

In this paper, we propose one Consensus-aware Visual-Semantic Embedding (CVSE) architecture for image-text matching, as depicted in Figure 1 (c). Specifically, we first make the consensus exploitation by computing statistical co-occurrence correlations between the semantic concepts from the image captioning corpus and constructing the concept correlation graph to learn the consensus-aware concept (CAC) representations. Afterwards, based on the learned CAC representations, both images and texts can be represented at the consensus level. Finally, the consensus-aware representation learning integrates the instance-level and consensus-level representations together, which thereby serves to make the cross-modal alignment. Experiment results on public datasets demonstrate that the proposed CVSE model is capable of learning discriminative representations for image-text matching, and thereby boost the bidirectional image and sentence retrieval performances. Our contributions lie in three-fold.

- We make the first attempt to exploit the consensus information for image-text matching. As a departure from existing instance-level alignment based methods, our model leverages one external corpus to learn consensus-aware concept representations expressing the commonsense knowledge for further strengthening the semantic relationships between image and text.
- We propose a novel Consensus-aware Visual-Semantic Embedding (CVSE) model that unifies the representations of both modalities at the consensus level. And the consensus-aware concept representations are learned with one graph convolutional network, which captures the relationship between semantic concepts for more discriminative embedding learning.
- The extensive experimental results on two benchmark datasets demonstrate that our approach not only outperforms state-of-the-art methods for traditional image-text retrieval, but also exhibits superior generalization ability for cross-domain transferring.

## 2   Related Work

### 2.1   Knowledge Based Deep Learning

There has been growing interest in incorporating external knowledge to improve the data-driven neural network. For example, knowledge representation has been employed for image classification [31] and object recognition [7]. In the community of vision-language understanding,it has been explored in several contexts, including VQA [44] and scene graph generation [12]. In contrast, our CVSE leverages consensus knowledge to generate homogeneous high-level cross-modal representations and achieves visual-semantic alignment.
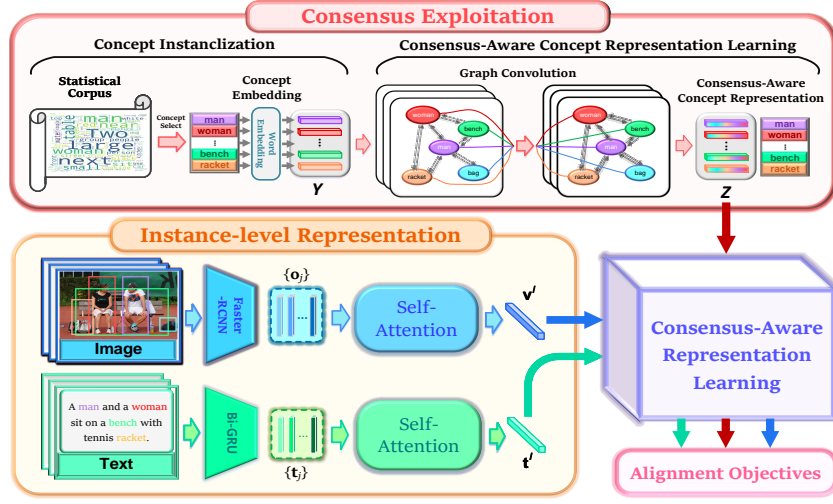
### 2.2   Image-Text Matching

Recently, there have been a rich line of studies proposed for addressing the problem of image-text matching. They mostly deploy the two-branch deep architecture to obtain the global [10, 21, 26, 27, 30, 43] or local [17, 18, 23] representations and align both modalities in the joint semantic space. Mao *et al.* [30] adopted CNN and Recurrent Neural Network (RNN) to represent images and texts, followed by employing bidirectional triplet ranking loss to learn a joint visual-semantic embedding space. For fragment-level alignment, Karpathy *et al.* [18] measured global cross-modal similarity by accumulating local ones among all region-words pairs. Moreover, several attention-based methods [16, 23, 32, 39] have been introduced to capture more fine-grained cross-modal interactions. To sum up, they mostly adhere to model the superficial statistical associations at instance level, whilst the lack of structured commonsense knowledge impairs their reasoning and inference capabilities for multi-modal data.

In contrast to previous studies, our CVSE incorporates the commonsense knowledge into the consensus-aware representations, thereby extracting the high-level semantics shared between image and text. The most relevant existing work to ours is [38], which enhances image representation by employing image scene graph as external knowledge to expand the visual concepts.Unlike [38], our CVSE is capable of exploiting the learned consensus-aware concept representations to uniformly represent and align both modalities at the consensus level. Doing so allows us to measure cross-modal similarity via disentangling higher-level semantics for both image and text, which further improves its interpretability.

## 3   Consensus-Aware Visual-Semantic Embedding

In this section, we elaborate on our Consensus-aware Visual-Semantic Embedding (CVSE) architecture for image-text matching (see Figure 2). Different from instance-level representation based approaches, we first introduce a novel Consensus Exploitation module that leverages commonsense knowledge to capture the semantic associations among concepts. Then, we illustrate how to employ the Consensus Exploitation module to generate the consensus-level representation and combine it with the instance-level representation to represent both modalities. Lastly, the alignment objectives and inference method are represented.

**Fig. 2.** The proposed CVSE model for image-text matching. Taking the fragment-level features of both modalities as input, it not only adopts the dual self-attention to generate the instance-level representations $\mathbf{v}^I$ and $\mathbf{t}^I$, but also leverages the consensus exploitation module to learn the consensus-level representations.

### 3.1 Exploit Consensus Knowledge to Enhance Concept Representations

As aforementioned, capturing the intrinsic associations among the concepts, which serves as the commonsense knowledge in human reasoning, can analogously supply high-level semantics for more accurate image-text matching. To achieve this, we construct a Consensus Exploitation (CE) module (see Figure 2), which adopts graph convolution to propagate the semantic correlations among various concepts based on a correlation graph preserving their inter dependencies, which contributes to injecting more commonsense knowledge into the concept representation learning. It involves three key steps: (1) Concept instantiation, (2) Concept correlation graph building and, (3) Consensus-aware concept representation learning. The concrete details will be presented in the following.

**Concept Instantiation.** We rely on the image captioning corpus of the natural sentences to exploit the commonsense knowledge, which is represented as the semantic concepts and their correlations. Specifically, all the words in the corpus can serves as the candidate of semantic concepts. Due to the large scale of word vocabulary and the existence of some meaningless words, we follow [9, 15] to remove the rarely appeared words from the word vocabulary. In particular, we select the words with top-$q$ appearing frequencies in the concept vocabulary, which are roughly categorized into three types, *i.e.*, *Object*, *Motion*, and *Property*. For more detailed division principle, we refer readers to [13]. Moreover, according to the statistical frequency of the concepts with same type over the whole dataset,

we restrict the ratio of the concepts with type of (*Object, Motion, Property*) to be (7:2:1). After that, we employ the glove [33] technique to instantiate these selected concepts, which is denoted as $\mathbf{Y}$.

**Concept Correlation Graph Building.** With the instantiated concepts, their co-occurrence relationship are examined to build one correlation graph and thereby exploit the commonsense knowledge. To be more specific, we construct a conditional probability matrix $\mathbf{P}$ to model the correlation between different concepts, with each element $\mathbf{P}_{ij}$ denoting the appearance probability of concept $C_i$ when concept $C_j$ appears:

$$\mathbf{P}_{ij} = \frac{\mathbf{E}_{ij}}{N_i} \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{q \times q}$ is the concept co-occurrence matrix, $\mathbf{E}_{ij}$ represents the co-occurrence times of $C_i$ and $C_j$, and $N_i$ is the occurrence times of $C_i$ in the corpus. It is worth noting that $\mathbf{P}$ is an asymmetrical matrix, which allows us to capture the reasonable inter dependencies among various concepts rather than simple co-occurrence frequency.

Although the matrix $\mathbf{P}$ is able to capture the intrinsic correlation among the concepts, it suffers from several shortages. Firstly, it is produced by adopting the statistics of co-occurrence relationship of semantic concepts from the image captioning corpus, which may deviate from the data distribution of real scenario and further jeopardize its generalization ability. Secondly, the statistical patterns derived from co-occurrence frequency between concepts can be easily affected by the long-tail distribution, leading to biased correlation graph. To alleviate the above issues, we design a novel scale function, dubbed Confidence Scaling (CS) function, to rescale the matrix $\mathbf{E}$:

$$\mathbf{B}_{ij} = f_{CS}(\mathbf{E}_{ij}) = s^{\mathbf{E}_{ij}-u} - s^{-u}, \tag{2}$$

where $s$ and $u$ are two pre-defined parameters to determine the the amplifying/shrinking rate for rescaling the elements of $\mathbf{E}$. Afterwards, to further prevent the correlation matrix from being over-fitted to the training data and improve its generalization ability, we also follow [6] to apply binary operation to the rescaled matrix $\mathbf{B}$:

$$\mathbf{G}_{ij} = \begin{cases} 0, & if \ \ \mathbf{B}_{ij} < \ \epsilon, \\ 1, & if \ \ \mathbf{B}_{ij} \geq \ \epsilon, \end{cases} \tag{3}$$

where $\mathbf{G}$ is the binarized matrix $\mathbf{B}$. $\epsilon$ denotes a threshold parameter filters noisy edges. Such scaling strategy not only assists us to focus on the more reliable co-occurrence relationship among the concepts, but also contributes to depressing the noise contained in the long-tailed data.

**Consensus-Aware Concept Representation.** Graph Convolutional Network (GCN) [3, 20] is a multilayer neural network that operates on a graph

and update the embedding representations of the nodes via propagating information based on their neighborhoods. Distinct from the conventional convolution operation that are implemented on images with Euclidean structures, GCN can learn the mapping function on graph-structured data. In this section, we employ the multiple stacked GCN layers to learn the concept representations (dubbed CGCN module), which introduces higher order neighborhoods information among the concepts to model their inter dependencies. More formally, given the instantiated concept representations $\mathbf{Y}$ and the concept correlation graph $\mathbf{G}$, the embedding feature of the $l$-th layer is calculated as

$$\mathbf{H}^{(l+1)} = \rho(\tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{4}$$

where $\mathbf{H}^{(0)} = \mathbf{Y}$, $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}$ denotes the normalized symmetric matrix and $\mathbf{W}^l$ represents the learnable weight matrix. $\rho$ is a non-linear activation function, $e.g.$, ReLU function [22].

We take output of the last layer from GCN to acquire the final concept representations $\mathbf{Z} \in \mathbb{R}^{q \times d}$ with $\mathbf{z}_i$ denoting the generated embedding representation of concept $C_i$, and $d$ indicating the dimensionality of the joint embedding space. Specifically, the $i$-th row vector of matrix $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_q\}$, $i.e.$ $\mathbf{z}_i$, represents the embedding representation for the $i$-th element of the concept vocabulary. For clarity, we name $\mathbf{Z}$ as consensus-aware concept (CAC) representations, which is capable of exploiting the commonsense knowledge to capture underlying interactions among various semantic concepts.

### 3.2   Consensus-Aware Representation Learning

In this section, we would like to incorporate the exploited consensus to generate the consensus-aware representation of image and text.

**Instance-level Image and Text Representations.** As aforementioned, conventional image-text matching only rely on the individual image/text instance to yield the correponding representations for matching, as illustrated in Figure 2. Specifically, given an input image, we utilize a pre-trained Faster-RCNN [1, 36] followed by a fully-connected (FC) layer to represent it by $M$ region-level visual features $\mathbf{O} = \{\mathbf{o}_1, ..., \mathbf{o}_M\}$, whose elements are all $F$-dimensional vector. Given a sentence with $L$ words, the word embedding is sequentially fed into a bi-directional GRU [37]. After that, we can obtain the word-level textual features $\{\mathbf{t}_1, ..., \mathbf{t}_L\}$ by performing mean pooling to aggregate the forward and backward hidden state vectors at each time step.

Afterwards, the self-attention mechanism [40] is used to concentrate on the informative portion of the fragment-level features to enhance latent embeddings for both modalities. Note that here we only describe the attention generation procedure of the visual branch, as it goes the same for the textual one. The region-level visual features $\{\mathbf{o}_1, ..., \mathbf{o}_M\}$ is used as the key and value items, while the global visual feature vector $\bar{\mathbf{O}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{o}_m$ is adopted as the query item for the attention strategy. As such, the self-attention mechanism refines the instance-level
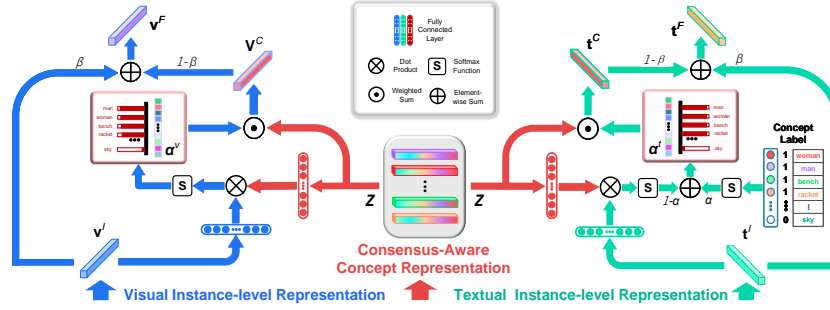
**Fig. 3.** Illustration of the consensus-level representation learning and the fusion between it and instance-level representation.

visual representation as $\mathbf{v}^I$. With the same process on the word-level textual features $\{\mathbf{t}_1, ..., \mathbf{t}_L\}$, the instance-level textual representation is refined as $\mathbf{t}^I$.

**Consensus-level Image and Text Representations.** In order to incorporate the exploited consensus, as shown in Figure 3, we take the instance-level visual and textual representations ($\mathbf{v}^I$ and $\mathbf{t}^I$) as input to query from the CAC representations. The generated significance scores for different semantic concepts allow us to uniformly utilize the linear combination of the CAC representations to represent both modalities. Mathematically, the visual consensus-level representation $\mathbf{v}^C$ can be calculated as follows:

$$
\begin{aligned}
\mathbf{a}_i^v &= \frac{exp(\lambda \mathbf{v}^I \mathbf{W}^v \mathbf{z}_i^\mathsf{T})}{\sum_{i=1}^q exp(\lambda \mathbf{v}^I \mathbf{W}^v \mathbf{z}_i^\mathsf{T})} \ , \\
\mathbf{v}^C &= \sum_{i=1}^q \mathbf{a}_i^v \cdot \mathbf{z}_i,
\end{aligned}
\tag{5}
$$

where $\mathbf{W}^v \in \mathbb{R}^{d \times d}$ is the learnable parameter matrix, $\mathbf{a}_i^v$ denotes the significance score corresponding to the semantic concept $\mathbf{z}_i$, and $\lambda$ controls the smoothness of the softmax function.

For the text, due to the semantic concepts are instantiated from the textual statistics, we can annotate any given image-text pair via employing a set of concepts that appears in its corresponding descriptions. Formally, we refer to this multi-label tagging as concept label $\mathbf{L}^t \in \mathbb{R}^{q \times 1}$. Considering the consensus knowledge is explored from the textual statistics, we argue that it's reasonable to leverage the concept label as prior information to guide the consensus-level representation learning and alignment. Specifically, we compute the predicted concept scores $\mathbf{a}_i^t$ and consensus-level representation $\mathbf{t}^C$ as follows:

$$
\begin{aligned}
\mathbf{a}_j^t &= \alpha \ \frac{exp(\lambda \mathbf{L}_j^t)}{\sum_{j=1}^q exp(\lambda \mathbf{L}_j^t)} \ + (1 - \alpha) \ \frac{exp(\lambda \mathbf{t}^I \mathbf{W}^t \mathbf{z}_j^\mathsf{T})}{\sum_{j=1}^q exp(\lambda \mathbf{t}^I \mathbf{W}^t \mathbf{z}_j^\mathsf{T})} \ , \\
\mathbf{t}^C &= \sum_{j=1}^q \mathbf{a}_j^t \cdot \mathbf{z}_j,
\end{aligned}
\tag{6}
$$

where $\mathbf{W}^t \in \mathbb{R}^{d \times d}$ denotes the learnable parameter matrix. $\alpha \in [0, 1]$ controls the proportion of the concept label to generate the textual predicted concept scores $\mathbf{a}_j^t$. We empirically find that incorporating the concept label into the textual consensus-level representation learning can significantly boost the performances.

**Fusing Consensus-level and Instance-level Representations.** We integrate the instance-level representations $\mathbf{v}^I(\mathbf{t}^I)$ and consensus-level representation $\mathbf{v}^C(\mathbf{t}^C)$ to comprehensively characterizing the semantic meanings of the visual and textual modalities. Empirically, we find that the simple weighted sum operation can achieve satisfactory results, which is defined as:

$$\begin{aligned}
\mathbf{v}^F &= \beta \mathbf{v}^I + (1 - \beta) \mathbf{v}^C, \\
\mathbf{t}^F &= \beta \mathbf{t}^I + (1 - \beta) \mathbf{t}^C,
\end{aligned} \tag{7}$$

where $\beta$ is a tuning parameter controlling the ratio of two types of representations. And $\mathbf{v}^F$ and $\mathbf{t}^F$ respectively denote the combined visual and textual representations, dubbed consensus-aware representations.

### 3.3 Training and Inference

**Training.** During the training, we deploy the widely adopted bidirectional triplet ranking loss [10, 11, 21] to align the image and text:

$$\begin{aligned}
\mathcal{L}_{rank}(\mathbf{v}, \mathbf{t}) = \sum_{(\mathbf{v}, \mathbf{t})} &\{\max[0, \gamma - s(\mathbf{v}, \mathbf{t}) + s(\mathbf{v}, \mathbf{t}^-)] \\
&+ \max[0, \gamma - s(\mathbf{t}, \mathbf{v}) + s(\mathbf{t}, \mathbf{v}^-)]\},
\end{aligned} \tag{8}$$

where $\gamma$ is a predefined margin parameter, $s(\cdot, \cdot)$ denotes cosine distance function. Given the representations for a matched image-text pair $(\mathbf{v}, \mathbf{t})$, its corresponding negative pairs are denoted as $(\mathbf{t}, \mathbf{v}^-)$ and $(\mathbf{v}, \mathbf{t}^-)$, respectively. The bidirectional ranking objectives are imposed on all three types of representations, including instance-level, consensus-level, and consensus-aware representations.

Considering that a matched image-text pair usually contains similar semantic concepts, we impose the Kullback Leibler (KL) divergence on the visual and textual predicted concept scores to further regularize the alignment:

$$\mathcal{D}_{KL}(\mathbf{a}^t \parallel \mathbf{a}^v) = \sum_{i=1}^q \mathbf{a}_i^t \; log(\frac{\mathbf{a}_i^t}{\mathbf{a}_i^v}), \tag{9}$$

In summary, the final training objectives of our CVSE model is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rank}(\mathbf{v^F}, \mathbf{t^F}) + \lambda_2 \mathcal{L}_{rank}(\mathbf{v^I}, \mathbf{t^I}) + \lambda_3 \mathcal{L}_{rank}(\mathbf{v^C}, \mathbf{t^C}) + \lambda_4 \mathcal{D}_{KL}, \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ aim to balance the weight of different loss functions.

**Inference.** During inference, we only deploy the consensus-aware representations $\mathbf{v}^F (\mathbf{t}^F)$ and utilize cosine distance to measure their cross-modal similarity.

## 4   Experiments

### 4.1   Dataset and Settings

**Datasets.** Flickr30k [34] is an image-caption dataset containing 31,783 images, with each image annotated with five sentences. Following the protocol of [30], we split the dataset into 29,783 training, 1000 validation, and 1000 testing images. We report the performance evaluation of image-text retrieval on 1000 testing set. MSCOCO [24] is another image-caption dataset, totally including 123,287 images with each image roughly annotated with five sentence-level descriptions. We follow the public dataset split of [18], including 113,287 training images, 1000 validation images, and 5000 testing images. We report the experimental results on both 1K testing set and 5K testing set, respectively.

**Evaluation Metrics.** We employ the widely-used R@K as evaluation metric [10, 21], which measures the the fraction of queries for which the matched item is found among the top $k$ retrieved results. We also report the "mR" criterion that averages all six recall rates of R@K, which provides a more comprehensive evaluation to testify the overall performance.

### 4.2   Implementation Details

All our experiments are implemented in PyTorch toolkit with a single NVIDIA Tesla P40 GPU. For representing visual modality, the amount of detected regions in each image is $M = 36$, and the dimensionality of region representation vectors is $F = 2048$. The dimensionality of word embedding space is set to 300. The dimensionality of joint space $d$ is set to 1024. For the consensus exploitation, we adopt 300-dim GloVe [33] trained on the Wikipedia dataset to initialize the the semantic concepts. The size of the semantic concept vocabulary is $q = 300$. And two graph convolution layers are used, with the embedding dimensionality are set to 512 and 1024, respectively. Regarding the correlation matrix $\mathbf{G}$, we set $s = 5$ and $u = 0.02$ in Eq. (2), and $\epsilon = 0.3$ in Eq. (3). For image and text representation learning, we set $\lambda = 10$ in Eq. (5) and $\alpha = 0.35$ in Eq. (6), respectively. For the training objective, we empirically set $\beta = 0.75$ in Eq. (7) , $\gamma = 0.2$ in Eq. (8) and $\lambda_1, \lambda_2, \lambda_3, \lambda_4 = 3, 5, 1, 2$ in Eq. (10). Our CVSE model is trained by Adam optimizer [19] with mini-batch size of 128. The learning rate is set to be 0.0002 for the first 15 epochs and 0.00002 for the next 15 epochs. The dropout is also employed with a dropout rate of 0.4. Our code is available [3].

### 4.3   Comparison to State-of-the-art

The experimental results on the MSCOCO dataset are shown in Table 1. From Table 1, we can observe that our CVSE is obviously superior to the competitors in most evaluation metrics, which yield a result of 78.6% and 66.3% on

---

[3] https://github.com/BruceW91/CVSE

**Table 1.** Comparisons of experimental results on MSCOCO 1K test set and Flickr30k test set.

| Approach | MSCOCO dataset | | | | | | | Flickr30k dataset | | | | | | |
| | Text retrieval | | | Image Retrieval | | | mR | Text retrieval | | | Image Retrieval | | | mR |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| DVSA [18] | 38.4 | 69.9 | 80.5 | 27.4 | 60.2 | 74.8 | 39.2 | 22.2 | 48.2 | 61.4 | 15.2 | 37.7 | 50.5 | 58.5 |
| m-RNN [30] | 41.0 | 73.0 | 83.5 | 29.0 | 42.2 | 77.0 | 57.6 | 35.4 | 63.8 | 73.7 | 22.8 | 50.7 | 63.1 | 51.6 |
| DSPE [43] | 50.1 | 79.7 | 89.2 | 39.6 | 75.2 | 86.9 | 70.1 | 40.3 | 68.9 | 79.9 | 29.7 | 60.1 | 72.1 | 58.5 |
| CMPM [49] | 56.1 | 86.3 | 92.9 | 44.6 | 78.8 | 89 | 74.6 | 49.6 | 76.8 | 86.1 | 37.3 | 65.7 | 75.5 | 65.2 |
| VSE++ [10] | 64.7 | - | 95.9 | 52.0 | - | 92.0 | - | 52.9 | - | 87.2 | 39.6 | - | 79.5 | - |
| PVSE [39] | 69.2 | 91.6 | 96.6 | 55.2 | 86.5 | 93.7 | - | - | - | - | - | - | - | - |
| SCAN [23] | 72.7 | 94.8 | 98.4 | 58.8 | 88.4 | 94.8 | 83.6 | 67.4 | 90.3 | 95.8 | 48.6 | 77.7 | 85.2 | 77.5 |
| CAMP [45] | 72.3 | 94.8 | 98.3 | 58.5 | 87.9 | 95.0 | 84.5 | 68.1 | 89.7 | 95.2 | 51.5 | 77.1 | 85.3 | 77.8 |
| LIWE [46] | 73.2 | 95.5 | 98.2 | 57.9 | 88.3 | 94.5 | 84.6 | 69.6 | 90.3 | 95.6 | 51.2 | 80.4 | 87.2 | 79.1 |
| RDAN [14] | 74.6 | **96.2** | **98.7** | 61.6 | 89.2 | 94.7 | 85.8 | 68.1 | **91.0** | **95.9** | 54.1 | 80.9 | 87.2 | 79.5 |
| CVSE | **78.6** | 95.0 | 97.5 | **66.3** | **91.8** | **96.3** | **87.6** | **73.6** | 90.4 | 94.4 | **56.1** | **83.2** | **90.0** | **81.3** |

R@1 for text retrieval and image retrieval, respectively. In particular, compared with the second best RDAN method, we achieve absolute boost (5.3%, 2.6%, 1.6%) on (R@1, R@5, R@10) for image retrieval. In contrast, we also find the performance of CVSE is slightly inferior to best method on R@5 and R@10 for text retrieval. However, as the most persuasive criteria, the mR metric of our CVSE still markedly exceeds other algorithms. Besides, some methods partially surpassed ours, such as SCAN [23] and RDAN [14] both exhaustively aggregating the local similarities over the visual and textual fragments, which leads to slow inference speed. By contrast, our CVSE just employs combined global representations so that substantially speeds up the inference stage. Therefore, considering the balance between effectiveness and efficiency, our CVSE still has distinct advantages over them.

The results on the Flickr30K dataset are presented in Table 1. It can be seen that our CVSE arrives at 81.3% on the criteria of "mR", which also outperforms all the state-of-the-art methods. Especially for image retrieval, the CVSE model surpasses the previous best method by (2.0%, 2.3%, 2.8%) on (R@1, R@5, R@10), respectively. The above results substantially demonstrate the effectiveness and necessity of exploiting the consensus between both modalities to align the visual and textual representations.

### 4.4   Ablation Studies

In this section, we perform several ablation studies to systematically explore the impacts of different components in our CVSE model. Unless otherwise specified, we validate the performance on the 1K test set of MSCOCO dataset.

**Different Configuration of Consensus Exploitation.** To start with, we explore how the different configurations of consensus exploitation module affects the performance of CVSE model. As shown in Table 2, it is observed that although we only adopt the glove word embedding as the CAC representations, the

**Table 2.** Effect of different configurations of CGCN module on MSCOCO Dataset.

| Approaches | CGCN | | | Text Retrieval | | | Image Retrieval | | |
|---|---|---|---|---|---|---|---|---|---|
| | Graph Embedding | CS Function | Concept Label | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| CVSE$_{full}$ | ✓ | ✓ | ✓ | **78.6** | **95.0** | 97.5 | **66.3** | 91.8 | **96.3** |
| CVSE$_{wo/GE}$ | | ✓ | ✓ | 74.2 | 92.6 | 96.2 | 62.5 | 90.3 | 94.6 |
| CVSE$_{wo/CS}$ | ✓ | | ✓ | 77.4 | 93.8 | 97.0 | 65.2 | **91.9** | 95.8 |
| CVSE$_{wo/CL}$ | ✓ | ✓ | | 72.5 | 93.5 | **97.7** | 57.2 | 87.4 | 94.1 |

**Table 3.** Effect of different configurations of objective and inference scheme on MSCOCO Dataset.

| Approaches | Objective | | Inference Scheme | | | Text retrieval | | | Image Retrieval | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Separate Constraint | KL | Instance | Consensus | Fused | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| CVSE$_{wo/SC}$ | | ✓ | | | ✓ | 74.9 | 91.1 | 95.7 | 63.6 | 90.8 | 94.2 |
| CVSE$_{wo/KL}$ | ✓ | | | | ✓ | **77.3** | 93.4 | 96.9 | **65.6** | **92.8** | **95.9** |
| CVSE ($\beta = 1$) | ✓ | ✓ | ✓ | | | 71.2 | **93.8** | **97.4** | 54.8 | 87.0 | 92.2 |
| CVSE ($\beta = 0$) | ✓ | ✓ | | ✓ | | 57.0 | 67.2 | 80.2 | 57.4 | 86.5 | 92.9 |

model (CVSE$_{wo/GC}$) can still achieve the comparable performance in comparison to the current leading methods. It indicates the semantic information contained in word embedding technique is still capable of providing weak consensus information to benefit image-text matching. Compared to the model (CVSE$_{wo/CS}$) where CS function in Eq. (2) is excluded, the CVSE model can obtain 1.2% and 1.1% performance gain on R@1 for text retrieval and image retrieval, respectively. Besides, we also find that if the concept label $\mathbf{L}^t$ is excluded, the performance of model (CVSE$_{wo/CL}$) evidently drops. We conjecture that this result is attributed to the consensus knowledge is collected from textual statistics, thus the textual prior information contained in concept label substantially contributes to enhancing the textual consensus-level representation so as to achieve more precise cross-modal alignment.

**Different Configurations of Training Objective and Inference Strategy.** We further explore how the different alignment objectives affect our performance. First, as shown in Table 3, when the separate ranking loss, *i.e.* $\mathcal{L}_{rank-I}$ and $\mathcal{L}_{rank-C}$ are both removed, the CVSE$_{wo/SC}$ model performs worse than our CVSE model, which validates the effectiveness of the two terms. Secondly, we find that the CVSE$_{wo/KL}$ produces inferior retrieval results, indicating the importance of $\mathcal{D}_{KL}$ for regularizing the distribution discrepancy of predicted concept scores between image and text, which again provides more interpretability that pariwise heterogeneous data should correspond to the approximate semantic concepts. Finally, we explore the relationship between instance-level features and consensus-level features for representing both modalities. Specifically, the CVSE ($\beta = 1$) denotes the CVSE model with $\beta = 1$ in Eq. (7), which employs instance-level representations alone. Similarly, the CVSE ($\beta = 0$) model refers to the CVSE that only adopts the consensus-level representations. Interestingly, we observe that deploying the representations from any single semantic level alone

**Table 4.** Comparison results on cross-dataset generalization from MSCOCO to Flickr30k.

| Approaches | Text retrieval | | | Image Retrieval | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| RRF-Net [25] | 28.8 | 53.8 | 66.4 | 21.3 | 42.7 | 53.7 |
| VSE++ [10] | 40.5 | 67.3 | 77.7 | 28.4 | 55.4 | 66.6 |
| LVSE [8] | 46.5 | 72.0 | 82.2 | 34.9 | 62.4 | 73.5 |
| SCAN [23] | 49.8 | 77.8 | 86.0 | 38.4 | 65.0 | 74.4 |
| $\text{CVSE}_{wo/consensus}$ | 49.1 | 75.5 | 84.3 | 36.4 | 63.7 | 73.3 |
| CVSE | **57.8** | **81.4** | **87.2** | **44.8** | **71.8** | **81.1** |

will yields inferior results compared to their combination. It substantially verifies the semantical complementarity between the instance-level and consensus-level representations is critical for achieving significant performance improvements.

### 4.5  Further Analysis

**Consensus Exploitation for Domain Adaptation.** To further verify the capacity of consensus knowledge, we test its generalization ability by conducting cross-dataset experiments, which was seldom investigated in previous studies whilst meaningful for evaluating the cross-modal retrieval performance in real scenario. Specifically, we conduct the experiment by directly transferring our model trained on MS-COCO to Flickr30k dataset. For comparison, except for two existing work [8, 25] that provide the corresponding results, we additionally re-implement two previous studies [10, 23] based on their open-released code. From Table 4, it's obvious that our CVSE outperforms all the competitors by a large margin. Moreover, compared to the baseline that only employs instance-level alignment ($\text{CVSE}_{wo/consensus}$), CVSE achieves compelling improvements. These results implicate the learned consensus knowledge can be shared between cross-domain heterogeneous data, which leads to significant performance boost.

**The Visualization of Confidence Score for Concepts.** In Figure 4, we visualize the confidence score of concepts predicted by our CVSE. It can be seen that the prediction results are considerably reliable. In particular, some informative concepts that are not involved in the image-text pair can even be captured. For example, from Figure 4(a), the associated concepts of "`traffic`" and "`buildings`" are also pinpointed for enhancing semantic representations.

**The Visualization of Consensus-Aware Concept Representations.** In Figure 5, we adopt the t-SNE [29] to visualize the CAC representations. In contrast to the word2vec [33] based embedding features, the distribution of our CAC representations is more consistent with our common sense. For instance, the concepts with POS of *Motion*, such as "`riding`", is closely related to the concept of "`person`". Similarly, the concept of "`plate`" is closely associated
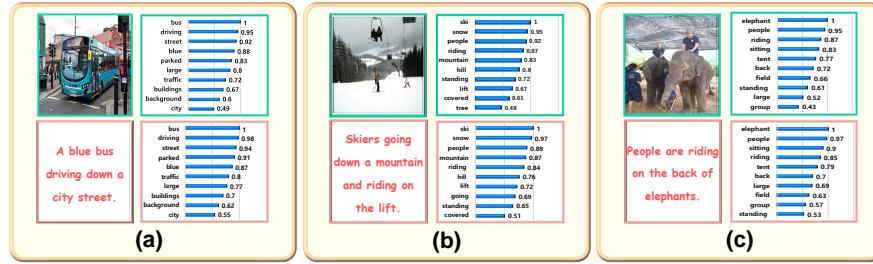
**Fig. 4.** The visualization results of predicted scores for top-10 concepts.
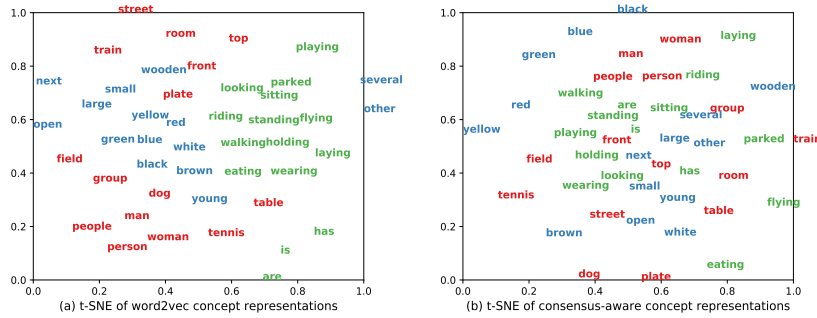


**Fig. 5.** The t-SNE results of word2vec based concept representations and our consensus-aware concepts representations. We randomly select 15 concepts per POS for performing visualization and annotate each POS with the same color.

with "`eating`". These results further verify the effectiveness of our consensus exploitation module in capturing the semantic associations among the concepts.

## 5    Conclusions

The ambiguous understanding of multi-modal data severely impairs the ability of machine to precisely associate images with texts. In this work, we proposed a Consensus-Aware Visual-Semantic Embedding (CVSE) model that integrates commonsense knowledge into the multi-modal representation learning for visual-semantic embedding. Our main contribution is exploiting the consensus knowledge to simultaneously pinpoint the high-level concepts and generate the unified consensus-aware concept representations for both image and text. We demonstrated the superiority of our CVSE for image-text retrieval by outperforming state-of-the-art models on widely used MSCOCO and Flickr30k datasets.

## 6    Acknowledgments

# References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa (2018), CVPR
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence, Z.C., Parikh, D.: Vqa: Visual question answering (2015), ICCV
3. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs (2013), ICLR
4. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video (2018)
5. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding (2018), CVPR
6. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks (2019), CVPR
7. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs (2014), ECCV
8. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization (2018), CVPR
9. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back (2015), CVPR
10. Fartash, F., Fleet, D., Kiros, J., Fidler, S.: Vse++: improved visual-semantic embeddings (2018), BMVC
11. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model (2013), NIPS
12. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction (2019), CVPR
13. Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y.: Joint syntax representation learning and visual cue translation for video captioning (2019), ICCV
14. Hu, Z., Luo, Y., Lin, J., Yan, Y., Chen, J.: Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching (2019), IJCAI
15. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching (2018), CVPR
16. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. ICCV (2019)
17. Karpathy, A., Joulin, A., Li, F.F.: Deep fragment embeddings for bidirectional image sentence mapping (2014), NIPS
18. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions (2015), CVPR
19. Kingma, D., Ba, J.: Adam: Amethod for stochastic optimization (2014), ICLR
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016), ICLR
21. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models (2014), NIPS Workshop
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks (2012), NIPS
23. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching (2018), ECCV

24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context (2014), ECCV
25. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching (2017), ICCV
26. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence (2015), ICCV
27. Ma, L., Jiang, W., Jie, Z., Jiang, Y., Liu, W.: Matching image and sentence with multi-faceted representations. IEEE Transactions on Circuits and Systems for Video Technology **30**(7), 2250–2261 (2020)
28. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network (2016)
29. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2605 (2008)
30. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn) (2015), ICLR
31. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification (2017), CVPR
32. Nam, H., Ha, J., Kim, J.: Dual attention networks for multimodal reasoning and matching (2017), CVPR
33. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation (2014), EMNLP
34. Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models (2015), ICCV
35. Plummer, B., Mallya, A., Cervantes, C., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues (2017), ICCV
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015), NIPS
37. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing **45**(11), 2673–2681 (1997)
38. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching (2019), IJCAI
39. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval (2019), CVPR
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Lukasz, K., Polosukhin, I.: Attention is all you need (2017), NIPS
41. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning (2018)
42. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning (2018)
43. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings (2016), CVPR
44. Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2413–2427 (2018)
45. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval (2019), ICCV
46. Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings (2019), ICCV

47. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamtic modulation for temporal sentence grounding in videos (2019)
48. Zhang, W., Wang, B., Ma, L., Liu, W.: Reconstruct and represent video contents for captioning via reinforcement learning (2019). https://doi.org/10.1109/TPAMI.2019.2920899
49. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching (2018), ECCV