Structured Multi-modal Feature Embedding and Alignment for Image-Sentence Retrieval

Xuri Ge¹, Fuhai Chen³, Joemon M. Jose¹, Zhilong Ji², Zhongqin Wu², Xiao Liu²
¹School of Computing Science, University of Glasgow. ²TAL Education Group. ³National University of Singapore
x.ge.2@research.gla.ac.uk, cfh3c@nus.edu.sg, joemon.jose@glasgow.ac.uk, [jizhilong, wuzhongqin, liuxiao15]@tal.com

ABSTRACT
The current state-of-the-art image-sentence retrieval methods implicitly align the visual-textual fragments, like regions in images and words in sentences, and adopt attention modules to highlight the relevance of cross-modal semantic correspondences. However, the retrieval performance remains unsatisfactory due to a lack of consistent representation in both semantics and structural spaces. In this work, we propose to address the above issue from two aspects: (i) constructing intrinsic structure (along with relations) among the fragments of respective modalities, e.g., “dog → play → ball” in semantic structure for an image, and (ii) seeking explicit inter-modal structural and semantic correspondence between the visual and textual modalities.

In this paper, we propose a novel Structured Multi-modal Feature Embedding and Alignment (SMFEA) model for image-sentence retrieval. In order to jointly and explicitly learn the visual-textual embedding and the cross-modal alignment, SMFEA creates a novel multi-modal structured module with a shared context-aware referral tree. In particular, the relations of the visual and textual fragments are modeled by constructing Visual Context-aware Structured Tree encoder (VCS-Tree) and Textual Context-aware Structured Tree encoder (TCS-Tree) with shared labels, from which visual and textual features can be jointly learned and optimized. We utilize the multi-modal tree structure to explicitly align the heterogeneous image-sentence data by maximizing the semantic and structural similarity between corresponding inter-modal tree nodes. Extensive experiments on Microsoft COCO and Flickr30K benchmarks demonstrate the superiority of the proposed model in comparison to the state-of-the-art methods.

CCS CONCEPTS
• Information systems → Novelty in information retrieval.

KEYWORDS
Multimodal Retrieval; Image-Sentence Retrieval; Context-aware Structured Trees; Semantics and Structural Consistency

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Xuri Ge¹, Fuhai Chen³, Joemon M. Jose¹, Zhilong Ji², Zhongqin Wu², Xiao Liu². 2021. Structured Multi-modal Feature Embedding and Alignment for Image-Sentence Retrieval. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475634

1 INTRODUCTION
Cross-modal retrieval, a.k.a. image-sentence retrieval, plays an important role in real-world multimedia applications, e.g., queries by images in recommendation systems, or image-sentence retrieval in search engines. Image-sentence retrieval aims at retrieving the most relevant images (or sentences) given a query sentence (or image), and has attracted increasing research attention recently [7, 8, 10, 14, 16, 19, 20, 30, 32]. Its main challenge lies in capturing the effective alignment (both in semantics and structural spaces) between the visual and textual modalities.

Typically, traditional approaches [7, 8, 32] model the cross-modal alignment on an instance level by directly extracting the global instance-level features of the visual and the textual modalities via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), respectively, and estimate the visual-textual similarities based on the global features, as shown in Figure 1 (a). However, as argued in [8], cross-modal semantic gap is harder to bridge with solely the global characteristics of images and sentences. To address this issue, recent works [10, 14, 18] extract the features of the visual and textual fragments, i.e., object regions in images and words in sentences, and align the visual and the textual fragment features via a soft attention mechanism, as shown in Figure 1 (b). However, there are two key defects with the above fragment-level alignment approaches. On one hand, these approaches neglect the intra-modal contextual semantic and structural relations of the fragments, thus failing to capture the semantics of the images or the sentences effectively. On the other hand, these approaches make the inter-modal fragment alignment implicitly with the many-to-many matching across the visual and textual modalities and with this, it is difficult to improve the consistency of semantic and structural representation between modalities.

In this paper, we argue that the key issues in image-sentence retrieval can be addressed by: (i) constructing the intra-modal context relations of the visual/textual fragments with a structured embedding module; and (ii) aligning the inter-modal fragments and their relations explicitly using a shared semantic structure, as shown in Figure 1 (c). We propose a novel structured multi-modal feature embedding and alignment model with visual and textual context-aware tree encoders (VCS-Tree and TCS-Tree) for image-sentence retrieval, termed SMFEA. On one hand, the context-aware
we use the KL-divergence between two spaces to optimize the uni-

terminal issue in image-sentence retrieval task is to measure the 

2 RELATED WORK

2.1 Image-Sentence Retrieval

The key issue in image-sentence retrieval task is to measure the 

2.2 Structured Feature Embedding

In terms of structured feature embedding, exiting works for multi-

The contributions of this paper are as follows:

• We propose two context-aware structured tree encoders (VCS-Tree and TCS-Tree) to parse the intrinsic (within modality) relations among the fragments of respective modalities. Thus this leads to effective semantic representation for pairwise alignment of image and sentence.

• We mine the explicit semantic and structural consistency of inter-modality corresponding tree nodes in visual and textual tree structures to align the heterogeneous cross-modality features.

• The proposed SMFEA outperforms the state-of-the-art approaches for image-sentence retrieval on two benchmarks, i.e., Flickr30K and Microsoft COCO.

2 RELATED WORK

In contrast to previous studies, SMFEA models the relation struc-

ture of intra-modal fragments/words by the use of a fixed contextual 

Figure 1: Illustration of the different schemes: (a) the traditional instance-level alignment methods, (b) the recent fragment-level alignment methods, and (c) our SMFEA method. Compared with (a) and (b), our SMFEA in (c) exploits intra-modal relations of visual/textual fragments via a tree encoder and aligns them explicitly in the corresponding nodes in two modal trees.
in each modality. However unlike [30], SMFEA approaches this in a novel way by exploiting the learned multi-modal semantic trees to enhance the structured embedding of the visual and textual modalities. By aligning the inter-modal semantics and structure consistently, the joint embedding space is obtained to reduce the heterogeneous (inter-modality) semantic gap. Doing so allows us to provide more robustness than [30], which also improves the interpretability of the model.

3 SMFEA APPROACH

The overview of SMFEA is illustrated in Figure 2. We will first describe the multi-modal feature extractors (① in Figure 2) in our work in Section 3.1. Then, the context-aware representation module is introduced in detail in Section 3.2 with context-aware structured tree encoders (② in Figure 2) and the consensus-aware concept (CAC) representation learning module (③ in Figure 2). Finally, the objective function is discussed in Section 3.3.

3.1 Multi-modal Feature Extractors

Our multi-modal feature extractors include two components to encode the region-level visual representations and word-level textual representations into the instance-level multi-modal features.

3.1.1 Visual representations. To better represent the salient entities and attributes in images, we take advantage of bottom-up-attention network [1] to embed the extracted sub-regions in an image. Specifically, given an image \( I \), we extract a set of image fragment-level sub-region features \( V = \{v_1, \ldots, v_K\} \), \( v_j \in \mathbb{R}^{2048} \), where \( K \) is the number of selected sub-regions, from the average pooling layer in Faster-RCNN [25].

Furthermore, we employ the self-attention mechanism [28] to refine the instance-level latent embeddings of sub-region features for each image, thus concentrating on the salient information exploited by the fragment-level features. In particular, following [28], the fragment visual features \( V = \{v_1, \ldots, v_K\} \) are used as the key and value items. And the initialization of instance-level features \( V \), embedded by the mean of region features, serves as the query item to fuse the important fragment features with different learning weights \( \alpha \) as new instance-level visual representation \( V^D \). These can be formulated as:

\[
\bar{V} = \frac{1}{K} \sum_{i=1}^{K} v_i
\]

\[
\alpha_i = \frac{\exp(\bar{V}v_i)}{\sum_{i=1}^{K} \exp(\bar{V}v_i)}
\]

\[
V^D = \sum_{i=1}^{K} \alpha_i v_i
\]

3.1.2 Word-level textual representations. For sentences, word-level textual representations are encoded by a bi-directional GRU network [26]. In particular, we first represent each word \( w_j \) in sentence \( S = \{w_1, \ldots, w_N\} \) with length \( N \) as a one-hot vector being the cardinality of the \( D_{v} \)-length vocabulary dictionary. The one-hot vector of \( w_j \) is projected into a fixed dimensional space \( e_j = W_f w_j \) (\( W_f \) denotes the mapping parameter) and then sequentially fed into the bi-directional GRU. The final hidden representation for each word is the average of the hidden vectors in both directions as follows:

\[
w_{j}^{f} = \frac{\text{GRU}(e_j) + \text{GRU}(e_j)}{2}
\]

where \( j \in [1, N] \). Similar to the procedure in visual branch, we finally get the refined instance-level textual representation \( S^D \) of a sentence based on the word-level textual features.

3.2 Context-Aware Representation

Our aim is to construct the intrinsic relations among the fragments of the visual/textual modality. Hence, we construct two novel context-aware structured trees from instance-level visual and textual features, with the help of a shared referral tree. To facilitate the inter-modal semantics and structure correspondence, with the aim to bridge the heterogeneous (i.e., between modalities) semantic gap, our model aligns semantic categories of the corresponding modality nodes.
3.2.1 Shared referral tree encoder. During the training we construct for each of the modalities, context-aware structured trees of three-layer tree structures, supervised by shared labels (called shared referral tree). The shared referral tree is constructed by Stanford Parser [27] from sentence, and the post-tag tool and lemmatizer tool in NLTK [21] are applied to whiten the source sentences to reduce the irrelevant words and noise configurations. As shown in the middle of Figure 2, it is a fixed-structure, three-layer binary tree, which only contains nouns (or noun pair, adjective-noun pair), verbs, coverbs, prepositions, and conjunctions. “Null” in the referral tree means the ignorable node or the unknown category (not in the entity or relation dictionaries). Only nouns are regarded as fragments and used as leaf nodes in the subsequent training. Correct semantic content can be represented by the shared referral tree in in-order traversal way. A referral tree is created for each sentence and the corresponding image pair.

3.2.2 Context-aware structured tree encoders. We construct a visual context-aware structured tree (VCS-Tree) and a textual context-aware structured tree (TCS-Tree) to parse the intra-modal structural relations of the respective fragments/words. Moreover, the VCS-Tree and TCS-Tree are utilized to align the inter-modal nodes between the images and sentences. As shown in Figure 3, the tree structure of two modalities is the same, where each modality tree parses the instance-level features $V^D/S^D$ into a three-layer architecture with seven nodes (same as referral tree), of which four leaf nodes are used to parse fragments in the $1^{st}$ layer and three parent nodes to parse relations in the $2^{nd}$ and $3^{rd}$ layers, to organise the semantic and structural relations of an image or a sentence. There are two main reasons why we adopt this fixed structure: (i) inspired by [2, 3], the tree with seven nodes can express the main semantic content of each image-sentence pair; and (ii) it is suitable for improving the consistency of coarse semantics and structural representation between modalities, thereby improving the robustness and interpretability of the model. For simplicity, we will only introduce the detailed structure of VCS-Tree and do not repeat the details for the TCS-Tree.

As shown in the top branch of Figure 3, instance-level visual feature $V^D$ is first mapped into different semantic spaces by a linear mapping function with the parameter $W^o \in \mathbb{R}^{2048 \times D_o}$, which serve as the inputs to different layers in VCS-Tree:

$$
\hat{v}_i^D = V^D W_i^o, i \in \{1, 2, ..., 7\},
$$

For simplicity, we do not explicitly represent the bias terms in our paper.

For the VCS-Tree, we broadcast the context information between different layer nodes in a novel LSTM-based ternary tree encoder with a fixed structure. It can get final structured tree embedding of the image supervised by the shared referral tree. In particular, we describe the updating of a parent node $t$ in VCS-Tree, where the detailed computation is described in Eq.(6-11). $T(t)$ denotes the set of children of node $t$. The process can be formulated as:

$$
i_t = \sigma(W^i \hat{v}^D_t + U^i h_t),$$
$$f_t = \sigma(W^f \hat{v}^D_t + U^f h_t),$$
$$o_t = \sigma(W^o \hat{v}^D_t + U^o h_t),$$
$$\tilde{c}_t = \tanh(W^c \hat{v}^D_t + U^c h_t),$$
$$c_t = i_t \odot \tilde{c}_t + f_t \odot \sum (f_k \odot c_k),$$
$$h_t = o_t \odot \tanh(c_t),$$

where $i_t, f_t, o_t$ denote the input gate, forget gate and output gate, $\tilde{c}_t, c_t, h_t$ are the candidate cell value, cell state and hidden state of tree node $t$, $\sigma$ is the sigmoid function, $\odot$ is the element-wise multiplication, all $W^*$ and $U^*$ are learning weight matrices, $h_t$ is the summing of hidden states of children nodes $T(i)$, and $T(k)$ are sub-trees of $T(t)$ in Eq.(10). In this way, the features of the parent nodes in higher layers can contain the rich context-aware semantic information by the LSTM-based attention mechanism, which combine the children nodes information as well as the leaf nodes. Finally, each node is classified into the fragment/relation category by the Softmax classifier. And the sum of all node hidden states in the tree in in-order traversal manner, which are mapped into the same dimension with original visual features as the structured tree enhancement embedding $V^T$ as follows:

$$y^V_{i,1} = \text{Softmax}(W_{e} h^V_{i,1}), i^1 \in \{1, 3, 5, 7\},$$
$$y^V_{i,2,3} = \text{Softmax}(W_{e} h^V_{i,2,3}), i^2 \in \{2, 6\}, i^3 \in \{4\},$$
$$V^T = \sum (W_i h^V_i), i \in \{1, 2, ..., 7\},$$

where $y^V_{i,1}$ and $y^V_{i,2,3}$ denote the predicted scores of the fragment categories for $1^{st}$ layer and relation categories for $2^{nd}$ or $3^{rd}$ layers. $W_{e}$ and $W_{r}$ denote the mapping parameters for the fragment and relation categories according to these dictionaries [3], respectively. $W_i$ denotes the mapping parameters.
Likewise, our TCS-Tree with seven node structures takes the mapped instance-level textual feature $S^D$ as the inputs. The structure of TCS-Tree is same with VCS-Tree and the final structured textual embedding $S^T$ is obtained by the sum of the hidden states $h^S$ of the TCS-Tree and the original instance-level feature $S^D$. Furthermore, the predicted probability vectors for seven nodes in different layers of textual fragment categories $y^S_{i,t}$, and relation categories $y^S_{i,t+1}$ are obtained.

We capture the intra-modal context relations of the visual/textual fragments by minimizing the loss of the category classification. It can guarantee the correct semantic representation for the content of the image and corresponding sentence. Furthermore, we narrow the inter-modal distance of images and sentences by the minimizing the loss of the Kullback Leibler (KL) divergence for both modality tree nodes probability distributions. Details are given in the Section 3.3.

### 3.2.3 CAC representation learning module

Following [30], we also exploit the commonsense knowledge to capture the underlying interactions among various semantic concepts by learning the dual modalities consensus-aware concept (CAC) representations $V^C/S^C$, which can improve the fine-grained semantic information of our context-aware representations to a certain extent. Due to space restrictions, we are not repeating the process in [30].

#### 3.2.4 Multiple representations fusing module

to comprehensively characterize the semantic and structured expression for both the modalities, we combine the instance-level representations $V^D/S^D$, the context-aware structured enhancement features $V^T/S^T$ and CAC representations $V^C/S^C$ into fusing modalities representations $V^F/S^F$ with simple weighted sum operation, as following:

$$V^F = \beta_d V^D + \beta_t V^T + \beta_c V^C$$  \hspace{1cm} (15)

$$S^F = \beta_d S^D + \beta_t S^T + \beta_c S^C$$  \hspace{1cm} (16)

where $\beta_d, \beta_t, \beta_c$ are the tuning parameters for balancing. This allows the SMFEA model to get rich semantic and structure representation for each modalities and also keep cross-modal consistency of structure and semantics between the modalities.

### 3.3 Objective Function

In the above training process, all the parameters can be simultaneously optimized by minimizing a bidirectional triplet ranking loss [7], where we exploit positive and negative samples and as follows:

$$L_{\text{rank}}(I, S) = \sum_{(I,S)} \left[ \nabla - \cos(I, S) + \cos(I, \bar{S}) \right]_+$$

$$+ \sum_{(I,S)} \left[ \nabla - \cos(I, S) + \cos(I, \bar{S}) \right]_+$$  \hspace{1cm} (17)

where $\nabla$ is a margin constraint, $\cos(\cdot, \cdot)$ indicates cosine similarity function, and $\left[ . \right]_+ = \max(0, \cdot)$. Note that, $(I, S)$ denotes the given matched image-sentence pair and its corresponding negative samples are denoted as $I$ and $S$, respectively.

Moreover, we minimize the loss of the node category classification on both visual and textual context-aware structured tree encoders to improve the structured semantic referring ability, using a cross-entropy loss as follows:

$$L_{\text{CE}}(V^D, S^D) = - \sum_{i=1}^{M} \left( \text{CE}(y^V_i, z^V_i) + \text{CE}(y^S_i, z^S_i) \right)$$  \hspace{1cm} (18)

where $y^V_i$ and $y^S_i$ indicate the predicted fragment/relation categories of the $i$-th node in three layers of VCS-Tree and TCS-Tree with $M$ nodes, respectively. $z^V$ and $z^S$ are category labels of the nodes, as detailed in Section 3.2.1. And to further narrow the semantic gap between modalities, we employ the Kullback Leibler (KL) divergence to regularize the probability distributions on visual and textual predicted fragment/relation category scores, which is defined as:

$$D_{\text{KL}}(P^V \parallel P^S) = \sum_{i=1}^{M} p^V_i \log(p^V_i / p^S_i)$$  \hspace{1cm} (19)

where $p^V_i$ and $p^S_i$ denote the predicted probability distributions of cross-modal corresponding tree nodes.

In this way, we utilize a shared referral tree to modal the intra-modal embedding explicitly and employ the fixed cross-modal tree alignment to guarantee the inter-modal consistency of the structure and semantics between images and sentences. Finally, the joint loss of the SMFEA model is defined as:

$$L = L_{\text{rank}}(V^F, S^F) + L_{\text{CE}}(V^D, S^D) + D_{\text{KL}}(P^V \parallel P^S)$$  \hspace{1cm} (20)

Note that, we use the final fusing features $V^F$ and $S^F$ to calculate the similarity scores during inference process.

### 4 EXPERIMENTS

In this section, we report the results of experiments to evaluate the proposed approach, SMFEA. We will introduce the dataset and experimental settings first. Then, SMFEA is compared with the state-of-the-art image-sentence retrieval approaches quantitatively. Finally, we qualitatively analyze the results in detail.

#### 4.1 Dataset and Evaluation Metrics

##### 4.1.1 Dataset

To verify the effectiveness of our proposed approach, we choose the popular Flickr30k [35] and MS-COCO [17] datasets. Flickr30k contains over 31,000 images with 29,000 images for the training, 1,000 images for the testing, and 1,014 images for the validation. There are over 123,000 images in MS-COCO with 82,738 images for training, 5,000 images for the testing, and 5,000 images for the validation. Each image in these two benchmarks is given five corresponding sentences by different AMT workers.

##### 4.1.2 Evaluation metrics

Quantitative performances of all methods are evaluated by employing the widely-used [7, 10, 14, 15] recall metric, R@K (K=1, 5, 10) evaluation metric, which denotes the percentage of ground-truth being matched at top K results. Moreover, as in the literature, we report the "rsum" criterion that sums all six recall rates of R@K, which provides more comprehensive evaluation to testify the overall performance.

### 4.2 Implementation Details

Our model is trained on a single NVIDIA 2080Ti GPU with 11 GB memory. The whole network except the Faster-RCNN model is trained from scratch with the default initializer of PyTorch using ADAM optimizer [13]. The learning rate is set to 0.0002 initially with a decay rate of 0.1 every 25 epochs. The maximum epoch
| Method          | Sentence Retrieval | Image Retrieval | rSum  |
|-----------------|--------------------|-----------------|-------|
|                 | R@1    | R@5    | R@10  | R@1   | R@5   | R@10  |     |
| VSE++ [7]       | 52.9   | 79.1   | 87.2  | 39.6  | 69.6  | 79.5  | 407.9|
| SCAN* [14]      | 67.4   | 90.3   | 95.8  | 48.6  | 77.7  | 85.2  | 465.0|
| PFAN [33]       | 70.0   | 91.8   | 95.0  | 50.4  | 78.7  | 86.1  | 472.0|
| VSRN* [15]      | 71.3   | 90.6   | 96.0  | 54.7  | 81.8  | 88.2  | 482.6|
| CAAN [36]       | 70.1   | 91.6   | 97.2  | 52.8  | 79.0  | 87.9  | 478.6|
| CVSE [30]       | 73.5   | 92.1   | 95.8  | 54.7  | 82.1  | 88.4  | 482.4|
| SMFEA(ours)     | 73.7   | 92.5   | 96.1  | 54.7  | 82.1  | 88.4  | 487.5|

Table 2: Comparisons of experimental results on MS-COCO 1K test set. * indicates the performance of an ensemble model.

| Method          | Sentence Retrieval | Image Retrieval | rSum  |
|-----------------|--------------------|-----------------|-------|
|                 | R@1    | R@5    | R@10  | R@1   | R@5   | R@10  |     |
| VSE++ [7]       | 64.7   | -      | 95.9  | 52.0  | -     | 92.0  | 304.6|
| SCO [10]        | 69.9   | 92.9   | 97.5  | 56.7  | 87.5  | 94.8  | 499.3|
| SCAN* [14]      | 72.7   | 94.8   | 98.4  | 58.8  | 88.4  | 94.8  | 507.9|
| VSRN* [15]      | 76.2   | 94.8   | 98.2  | 62.8  | 97.7  | 95.1  | 516.8|
| MMCA [34]       | 74.8   | 95.6   | 97.7  | 61.6  | 89.8  | 95.2  | 514.7|
| IMRAM* [4]      | 76.7   | 95.6   | 98.5  | 61.7  | 89.1  | 95.0  | 516.6|
| CAAN [36]       | 75.5   | 95.4   | 98.5  | 61.3  | 89.7  | 95.2  | 515.6|
| CVSE [30]       | 74.8   | 95.1   | 98.3  | 59.9  | 89.4  | 95.2  | 512.7|
| SMFEA(ours)     | 75.1   | 95.4   | 98.3  | 62.5  | 90.1  | 96.2  | 517.6|

4.3 Comparison with State-of-the-art Methods

As in MIR literature, we follow the standard protocols for running the evaluation on the Flickr30K and MS-COCO datasets and hence for comparison purposes report the results of the baseline methods in Table 1 and Table 2, including (1) early works, i.e., VSE++ [7], SCO [10], SCAN* [14], and (2) state-of-the-art methods, i.e., PFAN [33], VSRN* [15], IMRAM* [4], MMCA [34], CAAN [36] and CVSE [30]. Note that, the ensemble models with * are further improved due to the complementarity between multiple models. The best and second best results are shown using bold and underline, respectively.

4.3.1 Quantitative comparison on Flickr30K. Quantitative results on Flickr30K 1K test set are shown in Table 1, where the proposed approach SMFEA outperforms the state-of-the-art methods with impressive margins for rSum. Though for a few recall metrics slight variations in performance exists, overall SMFEA shows steady improvements over all baselines. SMFEA achieves 3.6%, 1.9%, and 8.9% improvements in terms of R@1 on sentence retrieval, R@1 on image retrieval, and rSum, respectively, compared with the state-of-the-art method CAAN [36]. Furthermore, compared with some ensemble methods, e.g. VSRN [15], our SMFEA achieves the best performance on most evaluation metrics.

4.3.2 Quantitative comparison on MS-COCO. Quantitative results on MS-COCO 1K test set are shown in Table 2, where the proposed approach SMFEA outperforms the state-of-the-art methods with impressive margins for rSum. Though for a few recall metrics slight variations in performance exists, overall SMFEA shows steady improvements over all baselines. SMFEA achieves 3.6%, 1.9%, and 8.9% improvements in terms of R@1 on sentence retrieval, R@1 on image retrieval, and rSum, respectively, compared with the state-of-the-art method CAAN [36]. Furthermore, compared with some ensemble methods, e.g. VSRN [15], our SMFEA achieves the best performance on most evaluation metrics.
We perform detailed ablation studies on Flickr30K to investigate the effectiveness of each component of our SMFEA.

4.4.1 Effects of different configurations of context-aware tree encoders. Table 5 shows the comparing between SMFEA and its corresponding baselines. SMFEA decreases absolutely by 2.0% and 3.4% in terms of R@1 for sentence and image retrieval on Flickr30K when removing the multi-modal context-aware structure tree encoders (indicated by w/o trees in Table 5). More detailed, comparison shows that removing $D_{KL}$ or $L_{CE}$ makes absolute 3.1% and 2.2% drop in terms of R@1-Sum (summing R@1 for image retrieval and sentence retrieval) on Flickr30K, respectively. It has shown that the context-aware structure tree encoders with joint $D_{KL}$ or $L_{CE}$ objectives can slightly improve the effectiveness. Please note that our SMFEA without tree encoders (indicated by w/o trees) is reproduced by using the official codes of CVSE [30] with slightly different parameters, which may result in different performances compared with [30]. In addition, to better understand how the proposed SMFEA model learns the cross-modal fragments/relations, we visualize the learned relation and fragment categories of nodes in VCS-Tree and TCS-Tree in Figure 4. The proposed VCS-Tree and TCS-Tree capture the intrinsic context semantic relation among the fragments in image and sentences in the in-order traversal manner. Also, the explicit consistency of the inter-modal tree nodes is fully excavated.

4.4.2 Effects of different embedding structures of SMFEA. As shown in Table 7, SMFEA decreases absolutely 1.92% in terms of the average of all metrics on Flickr30k when replacing context-aware tree structure by a chain-based approach [9]. In addition, the linear-based tree [3] degrades the average score by 1.55% compared with our SMFEA. These observations suggest that our context-aware tree encoders can improve the semantic and structural context consistency mining effectiveness between visual and textual features.
To better understand the effectiveness of our proposed model, we visualize matching results of the sentence retrieval and image retrieval, shown in Figure 5, we show the top 3 ranked images for each text modal consistence of tree nodes correspondence. Performance boost for both modalities retrieval, which validates that the explicitly inter-modal semantic and structural correspondence between images and sentences with the visual/textual inner context-aware structured tree encoder (VCS-Tree/TCS-Tree) capturing. We proposed a novel structured multi-modal feature embedding and alignment (SMFEA) model, which contains a VCS-Tree and a TCS-Tree to enhance the intrinsic context-aware structured semantic information for image and sentence, respectively. Furthermore, the consistency estimation of the corresponding inter-modal tree nodes is maximized to narrow the cross-modal pair-wise distance. Extensive quantitative comparisons demonstrate that our SMFEA can achieve state-of-the-art performance across popular standard benchmarks, MS-COCO and Flickr30K, under various evaluation metrics.

Figure 5: Visual comparisons of image retrieval between our SMFEA and CVSE [30] on Flickr30K (best viewed in color).

Figure 6: Visual comparisons of sentence retrieval examples between SMFEA and CVSE [30] on Flickr30K (best viewed in color).

5 CONCLUSION AND FUTURE WORK

In this paper, we exploit image-sentence retrieval with structured multi-modal feature embedding and cross-modal alignment. Our work serves as the first to narrow the cross-modal heterogeneous gap by aligning the explicitly inter-modal semantic and structure correspondence between images and sentences with the visual/textual inner context-aware structured tree encoder (VCS-Tree/TCS-Tree) capturing. We proposed a novel structured multi-modal feature embedding and alignment (SMFEA) model, which contains a VCS-Tree and a TCS-Tree to enhance the intrinsic context-aware structured semantic information for image and sentence, respectively. Furthermore, the consistency estimation of the corresponding inter-modal tree nodes is maximized to narrow the cross-modal pair-wise distance. Extensive quantitative comparisons demonstrate that our SMFEA can achieve state-of-the-art performance across popular standard benchmarks, MS-COCO and Flickr30K, under various evaluation metrics.

Future work includes the exploration of fine-grained category expansion of fragment/relation in the context-aware structured tree encoders, improving the accuracy and fine-grained representation of the referral tree, and so on.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China, under Grant No. 2020AAA0104500.
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. 6077–6086.

[2] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baoschang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. 2019. Variational Structured Semantic Inference for Diverse Image Captioning. In Advances in Neural Information Processing Systems. 1931–1941.

[3] Fuhai Chen, Rongrong Ji, Jinsong Su, Yongjian Wu, and Yunsheng Wu. 2017. Structcap: Structured semantic embedding for image captioning. In Proceedings of the 25th ACM International Conference on Multimedia. 46–54.

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jundong Han. 2020. Innam: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 12655–12663.

[5] Tianlang Chen and Jiabo Luo. 2020. Expressing Objects just like Words: Recurrent Visual Embedding for Image-Text Matching. arXiv preprint arXiv:2002.08510 (2020).

[6] Martin Engelcke, Louis Chevalier, Patrick Pérez, and Matthieu Cord. 2018. Finding beans in burgers: Deep semantic-visual embedding with localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3984–3993.

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

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems. 2121–2129.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation 9, 8 (1997), 1735–1780.

[10] 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. 636–641.

[11] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for cross-modal reasoning for image-text matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.

[12] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

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

[15] Kunpeng Li, Yulun Zhang, Kai Li, Xiaodong He, and Yun Fu. 2020. Graph Structured Network for Image-Text Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 10921–10930.

[16] Hong Liu, Mingbao Lin, Shengchuan Zhang, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2018. Dense auto-encoder hashing for robust cross-modality retrieval. In Proceedings of the 26th ACM International Conference on Multimedia. 1589–1597.

[17] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. arXiv preprint cs/0205008 (2002).

[18] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014).

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

[20] Chenxiao Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding. In Proceedings of the IEEE International Conference on Computer Vision. 1881–1889.

[21] Shaoqiang 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.

[22] Richard Socher, Cliff Chiang-Yu Lin, Andrew Y Ng, and Christopher D Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the IEEE Conference on Machine Learning.

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

[24] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. arXiv preprint arXiv:1511.08361 (2015).

[25] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In European Conference on Computer Vision. Springer, 18–34.

[26] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2 (2018), 394–407.

[27] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5005–5013.

[28] Xiaoxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748 (2019).

[29] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 10941–10950.

[30] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2 (2014), 67–78.

[31] Qi Zhang, Shun Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-Aware Attention Network for Image-Text Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3536–3545.