Visual-Linguistic Dependency Encoding for Image-Text Retrieval

Wenxin Guo[1], **Lei Zhang**[1], **Kun Zhang**[1], **Yi Liu**[2], **Zhendong Mao**[1*]
[1]University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Communication Content Cognition, Beijing, China
{wxguo, kkzhang}@mail.ustc.edu.cn, {leizh23, zdmao}@ustc.edu.cn
gavin1332@gmail.com

Abstract

Image-text retrieval is a fundamental task to bridge the semantic gap between natural language and vision. Recent works primarily focus on aligning textual meanings with visual appearance. However, they often overlook the semantic discrepancy caused by syntactic structure in natural language expressions and relationships among visual entities. This oversight would lead to sub-optimal alignment and degraded retrieval performance, since the underlying semantic dependencies and object interactions remain inadequately encoded in both textual and visual embeddings. In this paper, we propose a novel Visual-Linguistic Dependency Encoding (VL-DE) framework, which explicitly models the dependency information among textual words and interaction patterns between image regions, improving the discriminative power of cross-modal representations for more accurate image-text retrieval. Specifically, VL-DE enhances textual representations by considering syntactic relationships and dependency types, and visual representations by attending to its spatially neighboring regions. Cross-attention mechanism is then introduced to aggregate aligned region-word pairs into image-text similarities. Analysis on Winoground, a dataset specially designed to measure vision-linguistic compositional structure reasoning, shows that VL-DE outperforms existing methods, demonstrating its effectiveness at this task. Comprehensive experiments on two benchmarks, Flickr30K and MS-COCO, further validates the competitiveness of our approach. Our code is available at https://github.com/USTC-gwx/VL-DE.

Keywords: Linguistic Dependency Information, Cross-Modal Alignment, Image-Text Retrieval

1. Introduction

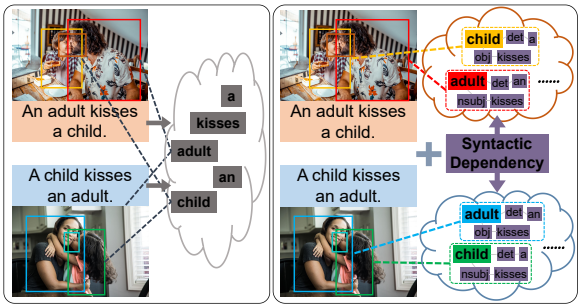
Vision and language are the two primary modalities for humans to obtain and communicate information about our physical world. Image-text retrieval focuses on establishing semantic alignment between images and descriptive texts, which is a fundamental task in the fields of natural language processing (NLP) and computer vision (CV). Due to more comprehensive language understanding with visual context and cues, this task contributes significantly to various practical applications, such as visual question answering (Yu et al., 2020; Su et al., 2021) and image captioning (Yan et al., 2021). The key challenge in image-text retrieval lies in effectively representing visual and linguistic information, and accurately aligning cross-modal data when they are semantically related.

Existing researches can be roughly categorized into global based approaches and local based approaches. Global based approaches (Faghri et al., 2017; Chen et al., 2021; Fu et al., 2023) map the entire image and text in a common latent embedding space to infer the aligning similarity. Local based approaches (Zeng et al., 2021; Zhang et al., 2022a,b; Pan et al., 2023) focus on local-level alignments between local fragments, *e.g.* salient regions in images and words in texts, establishing fine-grained correspondence between the two modalities. SCAN (Lee et al., 2018), a representative local based method, proposes a stacked cross attention network to learn semantic alignments between regions and words, and inspires a line of following works recently. Pan et al. (2023) elaborates fine-grained aligning process by mining informative region-word pairs and eliminating redundant or irrelevant alignments. Ge et al. (2023) explores intra- and inter-modal semantic correlations between objects and words based on spatial and semantic scene graph reasoning. In general, these approaches mainly focus on improving the discriminative power of visual and textual representations to capture shared semantics and establish precise region-word alignment.

However, when establishing fine-grained alignment, existing methods primarily attempt to explore correspondence between the semantic meaning of textual words and the visual appearance of image regions, ignoring the semantic discrepancy arising from syntactic divergences in linguistic expressions and relational differences among visual entities. While sentences in text often share similar constituents, their associated visual scenes can differ significantly. Without considering the semantic dependencies between words, it would be challenging to learn word representations that possess adequate discriminability, leading to inaccurate alignment with visual concepts. As shown in figure 1(a), two texts containing the same words in different orders may correspond to two visually distinct images. Due to lack of dependency information in the tex-

* Corresponding authors

LREC-COLING 2024, pages 17384–17396



(a) Existing Method (b) Our Method

Figure 1: Illustration of our motivation. Two visually distinct images can be described by texts using identical words in different orders. During the cross-modal alignment process, the text words are treated as independent fragments and aligned to image regions individually. (a) Neglecting semantic dependencies presents a challenge in learning sufficiently discriminative word representations. Here, the word *"adult"* in both texts can be aligned to regions depicting an adult in both images. (b) Our method explicitly models syntactic dependency for each word and provides distinct representations for identical words based on their different syntactic role in sentence structures, allowing for more accurate region-word alignment.

tual representations, the word *"adult"* in both texts can be semantically aligned to respective regions depicting an adult in both images, reducing retrieval performance. Therefore, it is crucial to construct a more sophisticated understanding of compositional structures in language and their corresponding visual content. By explicitly modeling the syntactic features of words, as in Figure 1(b), identical words can be represented distinctly based on their different syntactic role in sentence structures. This allows for accurate region-word alignments and enables distinguishing between semantically similar descriptions of visually disparate scenes.

To this end, we propose a novel Visual-Linguistic Dependency Encoding (VL-DE) framework, which explicitly accounts for the dependency information among textual words within a sentence and the object interactions within an image. Different from existing methods that typically align each image region with the most semantically relevant words, VL-DE captures the complex relationships between objects for both visual and textual representations, enabling more meaningful cross-modal alignments. Specifically, VL-DE adaptively learns enriched representations for textual words by hierarchically modeling the syntax relationships and dependency type within a sentence, to emphasize their relationships and syntactic functions, and enhance the textual contextual understanding. Meanwhile, for visual regions, VL-DE incorporates the natural spatial adjacencies and visual semantics from neighboring regions, which enables more detailed interpretation of visual contents. Moreover, we aggregate local alignments to an overall similarity of image-text pair by utilizing cross-attention mechanism, to capture more subtle correspondence across modalities. In this way, the alignment resolution in VL-DE can be adaptively refined, achieving more comprehensive correspondence between elaborate visual concepts and textual depictions.

Our contributions can be summarized as follows:

- We propose a novel Visual-Linguistic Dependency Encoding (VL-DE) framework, which is the first to, for the best of our knowledge, explicitly account for the semantic dependency in both modality, to construct a more comprehensive understanding of sophisticated visual and textual semantics for image-text retrieval.
- We propose a semantic dependency encoding method, which incorporate syntactic dependencies among words and spatially neighboring relationships of regions, to model the complexity and richness of semantic information for local fragments, yielding more meaningful alignments across modalities.
- We validate the effectiveness of our approach on Winoground dataset, demonstrating its competitive performance in comprehending the complex interplay between visual and textual fragments. Extensive experiments on two widely used benchmarks, *i.e.* Flickr30K and MS-COCO, further confirm the superiority and validity of our VL-DE.

2. Related Work

Image-text retrieval has aroused considerable attention in multimedia communities as it aims to bridge the semantic gap between natural language and vision. There are two main categories of existing approaches: global based and local based approaches. Additionally, a series of approaches have emerged that leverage external information to enhance the alignment process.

Global Based Approaches. A line of works (Faghri et al., 2017; Shi et al., 2018, 2019; Chen et al., 2021; Li et al., 2022b) learn global alignments between the entire image and text. A common approach is to map both the images and text into a shared embedding space, where the similarity between the two modalities can be directly measured with a distance metric. For instance, Li et al. (2022b) leverage Graph Convolutional Networks with GRU to generate enhanced visual representations to perform both local and global semantic reasoning. Fu et al. (2023) introduces a novel three-stage module for instance-level interactions, which

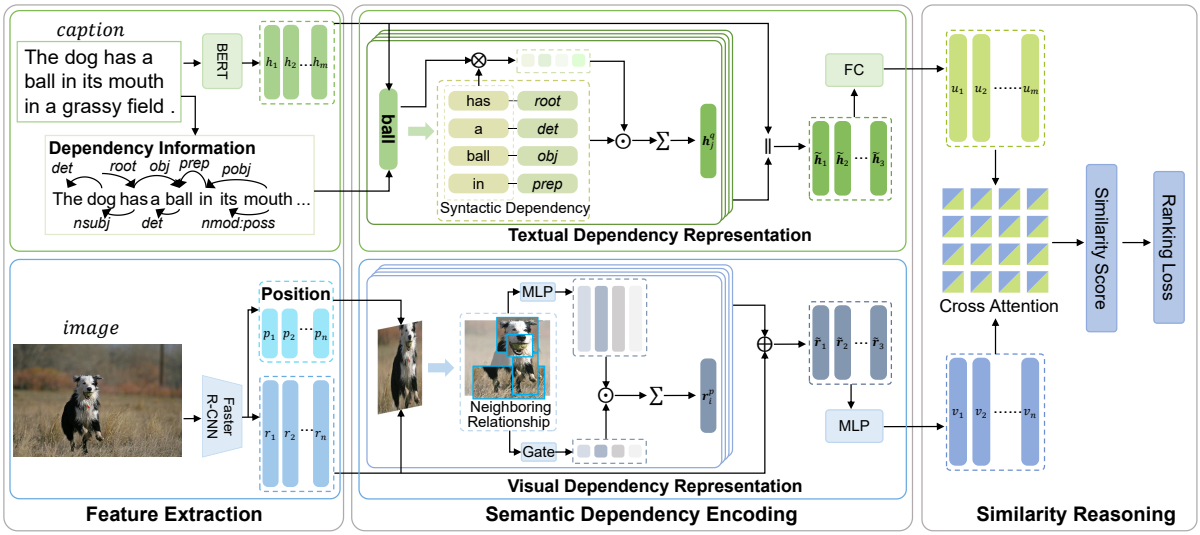


Figure 2: An overview of the proposed VL-DE framework, containing three major modules: feature extraction, semantic dependency encoding, and cross-modal similarity reasoning. VL-DE adaptively learns richer representations for local textual and visual fragments by incorporating syntactic dependencies among words and natural-neighboring relationships of regions, which enables a more comprehensive understanding of the intricate interactions among multiple entities.

employs fine-grained word-region correspondence promote model to learn instance-level relationships.

Local Based Approaches. This branch of approaches (Karpathy and Fei-Fei, 2015; Nam et al., 2017; Huang et al., 2017; Ji et al., 2019) focus on mining fine-grained alignments between local image and text fragments, and has gained popularity in image-text retrieval. One widely known method is attention-based SCAN (Lee et al., 2018), which selectively attends to specific visual and textual fragments by attending to regions and words with each other as context. Inspired by SCAN, plenty of recent studies (Wu et al., 2019; Hu et al., 2019; Chen et al., 2020a; Zhang et al., 2022b; Pan et al., 2023) have focused on designing cross-modal aligning mechanisms to enhance vision-language interactions and improve relevance measuring. Zhang et al. (2022b) introduces a unique negative mining strategy to better identify subtle mismatches across modalities, which enables more accurate multimodal aligning.

External Information Enhanced Approaches. Some works (Zhang et al., 2020; Wang et al., 2020b,a; Zhang et al., 2021a; Li et al., 2021; Cheng et al., 2022; Long et al., 2022) have centered on deriving external semantic information from both images and sentences to improve the retrieval performance. Wang et al. (2019) integrate position information of regions to enhance the correspondence learning. Zeng et al. (2021) construct scene graphs and syntactical tuple graphs to represent semantic information in images and sentences. Although high-level semantic information has greatly improved cross-modal retrieval, these methods typically focus on accurately aligning visual regions with textual contents that represent identical objects. Even though some approaches (Liu et al., 2020; Wei et al., 2020; Diao et al., 2021; Ge et al., 2023) implement intra-modal interactions between visual or textual fragments, they still tend to learn alignments between regions and the most matched words. In contrast, our VL-DE approach explicitly accounts for the semantic dependency in vision and language, and captures multi-level relationships in image-text representation by integrating position relationships of regions and dependency information among words, modeling rich semantics to promote a more comprehensive image-text retrieval.

3. Method

The overall framework of our proposed VL-DE is depicted in Figure 2, which consists of three modules. In section 3.1 and 3.2, we first extract features from both images and texts, as well as additional semantic information. Then, we introduce our approach for integrating complex dependency relationships to learn a enriched semantic representation. We further describe the segment-wise cross modal similarity reasoning method and objective function in section 3.3 and 3.4, respectively.

3.1. Textual Semantic Dependency Encoding

Feature Extraction. Given a text T containing m words, we utilize pre-trained BERT (Devlin et al., 2018) model to capture the semantic meaning of

each word. The extracted word-level textual representations can be denoted as $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m\}, \boldsymbol{h}_j \in \mathbb{R}^{D_t}$, and D_t is the dimension of textual feature representation.

Then, we parse the semantic dependency via an off-the-shell toolkit Stanford CoreNLP (Manning et al., 2014) to obtain all word-word relations and their relation types in a sentence. For each word $\boldsymbol{h}_j$, we extract the dependency word set $\{\boldsymbol{w}_{j,1}, \cdots, \boldsymbol{w}_{j,l_j}\}$ from the parsed results, comprising the l_j words that $\boldsymbol{h}_j$ depends on and the word itself. We also collect the corresponding dependency types as syntactic instances for the dependency words, denoted as $\{\boldsymbol{e}_{j,1}, \cdots, \boldsymbol{e}_{j,l_j}\}$. For example, as illustrated in Figure 2, consider the given text *"The dog has a ball in its mouth in a grassy field"*, the word *"ball"* has a determiner *"a"* and serves as the head of the preposition phrase introduced by *"in"*. It is also the object of the *root* word *"has"*. According to the above relationships, its dependency word set is $\{has, a, ball, mouth\}$, and the corresponding syntactic instance set is $\{has\text{-}root, a\text{-}det, ball\text{-}obj, in\text{-}prep\}$. Each word and syntactic instance are embedded as dependency feature vectors, *i.e.* $\boldsymbol{w}_{j,l}, \boldsymbol{e}_{j,l} \in \mathbb{R}^{D_t}$.

Textual Dependency Representation. The proposed dependency encoding method aims to learn enriched semantic representations of textual fragments by incorporating additional dependency information, which enhances the alignment with visual objects and the contextual understanding across modalities.

We adopt BERT as our text encoder, in which the word embeddings have already captured rich semantic information for each word. However, recent works have revealed that transformers in vision-language models, *e.g.* BERT, often struggle in encoding compositional relationships and are insensitive to word order in similar descriptions (Thrush et al., 2022; Lin et al., 2023). To address this limitation, we introduce syntactic dependencies to enhance word representations. In terms of syntax, different words serve different roles in a sentence. Content words, such as nouns (*e.g. "dog", "ball"*) that describe objects or entities, actively contribute to the meaning of the sentence, and are critical in establishing relationships to visual concepts because of their rich semantic information. Function words, on the other hand, such as prepositions (*e.g. "in"*) that indicate relationship between other words, have little semantic content themselves, and are less visually grounded for meaningful correspondences to image regions. To further enhance the textual representations, we aggregate the word embeddings with syntactic information, which explicitly emphasizes the dependency relationships and syntactic functions of words, enabling more subtle alignments with image regions.

To be specific, for a word $\boldsymbol{h}_j$, we consider its dependency word set $\{\boldsymbol{w}_{j,1}, \cdots, \boldsymbol{w}_{j,l_j}\}$ and the corresponding syntactic instances $\{\boldsymbol{e}_{j,1}, \cdots, \boldsymbol{e}_{j,l_j}\}$, which explicitly represent syntactical relationships between semantically related words. We leverage this syntactic information to enhance the representation of the target word:

$$q_{j,l} = \frac{\exp\left(\boldsymbol{h}_j \cdot \boldsymbol{w}_{j,l}\right)}{\sum_{k=1}^{l_j} \exp\left(\boldsymbol{h}_j \cdot \boldsymbol{w}_{j,k}\right)}, \tag{1}$$

$$\tilde{\boldsymbol{h}_j} = \boldsymbol{h}_j \parallel \left[\sum_{k=1}^{l_j} q_{j,k} \cdot \boldsymbol{e}_{j,k}\right], \tag{2}$$

where $q_{j,l}$ represents the importance of the relationship between the j-th and l-th words, enabling better distinction and utilization of semantically meaningful dependency information, while filtering out noise from the parsing results. By incorporating dependency information, we aim to capture not only the semantic context of language, but also its syntactic role and function, which allows us to establish more comprehensive conceptual correspondences between the textual and visual representations.

Further, we project the enhanced textual representation into a D-dimensional common embedding space followed by L_2 normalization:

$$\boldsymbol{u}_j = \left\| W^u \cdot \tilde{\boldsymbol{h}_j} + b^v \right\|_2, \tag{3}$$

where W^u, b^u are learnable parameters of the fully-connected layer.

3.2. Visual Semantic Dependency Encoding

Feature Extraction. Given an image I, we represent it as a set of region features using bottom-up attention mechanism (Anderson et al., 2018). Specifically, we detect n $(n = 36)$ salient regions in the image by applying Faster R-CNN (Ren et al., 2015) model, an objective detector pre-trained on Visual Genome (Krishna et al., 2017). The detected regions are then feed into pre-trained ResNet-101 (He et al., 2016) to extract mean-pooled convolutional features, denoted as $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}, \boldsymbol{r}_i \in \mathbb{R}^{D_v}$, and D_v is the dimension of visual feature representation. Further, we obtain the spatial properties for each of these n regions. The visual feature $\boldsymbol{r}_i$ extracted from Faster R-CNN is conditioned with geometric information about the region's bounding box, referred to as $\boldsymbol{p}_i \in \mathbb{R}^5$, *i.e.* the coordinates of the top left and bottom right corner as well as the area, normalized by the width/height of the image.

Visual Dependency Representation. By extracting salient regions from images, we can leverage important visual information more effectively. However, aligning visually complex regions can be

a challenge, as the rich semantics like intricate details and visual interactions between objects within a region may not be fully captured in the region embedding. To gain more detailed understanding of the semantics in complex regions, we integrate additional visual dependencies for each region, assuming a region's surrounding neighbors are prone to contain relevant semantic information. Specifically, for a region $\boldsymbol{r}_i$, we first calculate either the Euclidean distances between the centers of other regions and $\boldsymbol{r}_i$, or their Intersection over Union (IoU) scores, based on the position coordinates of their bounding boxes. These scores are then used to rank their semantic relevance to the target region. We extract the regions with the highest ranks from various directions, which are located within or closely around the target region, to form its neighbor set, denoted as $\mathcal{N}_i$. By carefully selecting only the closely neighboring regions, the semantic context is faithfully representative of the objects and their interactions within $\boldsymbol{r}_i$.

We then adopt a gated mechanism to aggregate each region with its relevant semantics along with their spatial relationships, where we select K regions (including $\boldsymbol{r}_i$) from $\mathcal{N}_i$, to serve as the visual semantic context:

$$g_{i_k} = \sigma\left(W_2^g \cdot \left(\phi\left(W_1^g \cdot [\boldsymbol{r}_{i_k} \parallel \boldsymbol{p}_{i_k}] + b_1^g\right)\right) + b_2^g\right), \tag{4}$$

$$\boldsymbol{c}_{i_k} = W_2^c \cdot \left(\phi\left(W_1^c \cdot [\boldsymbol{r}_{i_k} \parallel \boldsymbol{p}_{i_k}] + b_1^c\right)\right) + b_2^c, \tag{5}$$

$$\tilde{\boldsymbol{r}}_i = \boldsymbol{r}_i + \sum_{i_k \in \mathcal{N}_i} g_{i_k} \cdot \boldsymbol{c}_{i_k}, \tag{6}$$

where $\phi(\cdot)$ indicates the ReLU function, $\sigma(\cdot)$ indicates the $\mathrm{sigmoid}$ function, "$\parallel$" indicates concatenation, $i_k \in \mathcal{N}_i$, W^g, W^c, b^g, b^c are learnable parameters of MLP with two fully-connected layers. $\boldsymbol{c}_{i_k}$ represents the semantic dependency information implied in the relevant neighbor region along with its spatial position. The gate g_{i_k} is effective in adaptively selecting the most salient semantics from natural-neighboring regions. By this process, we enrich the semantic information of plain regions, and in the meanwhile provide a more detailed interpretation for complex regions.

Then, we project the enhanced visual representation into a D-dimensional common embedding space followed by L_2 normalization:

$$\boldsymbol{v}_i = \left\| W_2^v \cdot (\phi(W_1^v \cdot \tilde{\boldsymbol{r}}_i + b_1^v)) + b_2^v \right\|_2, \tag{7}$$

where W^v, b^v are learnable parameters of MLP with two fully-connected layers.

3.3. Cross-Modal Similarity Reasoning

To capture sophisticated correspondence between the visual and textual semantics, we reason the cross-modal similarities at a higher-level granularity. Specifically, given the enhanced region representations $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and word representations $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, we first aggregate the region features following the attention mechanism from SCAN (Lee et al., 2018), to obtain the visual context

$$\boldsymbol{a}_j^v = \sum_{i=1}^{n} \alpha_{ij} \boldsymbol{v}_i, \tag{8}$$

which is attended by the word $\boldsymbol{u}_j$, where $\alpha_{ij} = \frac{\exp(\lambda \bar{s}_{i,j})}{\sum_{i=1}^{n} \exp(\lambda \bar{s}_{i,j})}$, $\bar{s}_{i,j} = [s_{i,j}]_+ / \sqrt{\sum_{j=1}^{m} [s_{i,j}]_+^2}$. $s_{i,j}$ is cosine similarity between region $\boldsymbol{v}_i$ and word $\boldsymbol{u}_j$.

To better model the subtle relationships across modalities, we adopt a segment-wise approach that inspects finer-grained correspondences from multiple perspectives, and can obtain a more comprehensive measurement through the enhanced semantic representation. To be specific, we divide the representations of the attended image vector and the word feature into t segments:

$$\begin{aligned} \boldsymbol{a}_j^v &= [\boldsymbol{a}_{j1}^v \parallel \cdots \parallel \boldsymbol{a}_{jt}^v], \\ \boldsymbol{u}_j &= [\boldsymbol{u}_{j1} \parallel \cdots \parallel \boldsymbol{u}_{jt}], \end{aligned} \tag{9}$$

Then, we infer the relevance between word $\boldsymbol{u}_j$ and the image as a matching score:

$$s_j = W_2^s \cdot (\phi\left(W_1^s \cdot [s_{j1} \parallel \cdots \parallel s_{jt}] + b_1^s\right)) + b_2^s, \tag{10}$$

where $s_{jt} = cos(\boldsymbol{u}_{jt}, \boldsymbol{a}_{jt}^v)$ is the distinct similarity score for each segment-level pairing, W^s, b^s are learnable parameters of MLP with two fully-connected layers.

The final similarity score of the image-text pair is summarized by average pooling:

$$S(I,T) = \tanh\left(\frac{1}{m}\sum_{j=1}^{m} s_j\right), \tag{11}$$

where the $\tanh$ function compresses the similarity scores into the range $[-1, 1]$.

3.4. Objective Function

Following previous methods (Lee et al., 2018; Liu et al., 2019; Zhang et al., 2022b), we employ the triplet ranking loss as the objective function, which forces the similarity between matched image-text pairs to be higher than that between unmatched pairs by some fixed margin α. Furthermore, we focus on the hardest negatives, *i.e.* the unmatched pairs with maximum similarity scores in each mini-batch. Given the ground-truth image-text pair (I, T), the objective function is written as:

$$\begin{aligned} L(I,T) = &\left[\alpha - S(I,T) + S(I,\hat{T})\right]_+ \\ &+ \left[\alpha - S(I,T) + S(\hat{I},T)\right]_+ \end{aligned} \tag{12}$$

where $[x]_+ \equiv \max(x, 0)$, $\hat{I} = \arg\max_{V \neq I} S(V, T)$ and $\hat{T} = \arg\max_{U \neq T} S(I, U)$ are the hardest negative samples, α is margin hyperparameter.

4. Experiments

4.1. Datasets and Implementation Details

4.1.1. Datasets

We evaluate our method by performing extensive experiments on two benchmark datasets, Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014), where each image is annotated with 5 sentences. Flickr30K totally has 31,000 images and 155,000 sentences, and is split into 1,000 test images, 1,000 validation images, and 29,000 training images. MS-COCO contains 123,287 images and 616,435 sentences, where 113,287 images for training, 5,000 for validation, and 5,000 for testing. The results on MS-COCO are tested by averaging over five folds of 1K test images and also testing on the full 5K test images.

To further evaluate the ability of our method to model dependency relationships in both the visual and textual modality, we conduct experiments on Winoground (Thrush et al., 2022) dataset. The hand-crafted test set consists of 800 image-caption pairs, comprising a total of 400 *examples*. Each *example* contains two image-caption pairs, where the words are identical but their sequential order differs between the captions.

4.1.2. Evaluation Metrics

The performance on Winoground is evaluated based on three metrics, where higher values indicate better performance. **Text score** measures the percentage of *examples* where the given image and the ground-truth caption have higher similarity compared to the alternative caption, and this holds for the other image-caption pair too within the same *example*. **Image score** assesses the percentage of *examples* where the given caption and the ground-truth image exhibit higher similarity compared to the alternative image and this holds for the other image-caption pair too in the *example*. Lastly, **group score** measures the rate at which a model satisfies both text and image goals simultaneously.

To evaluate retrieval performance, we adopt the widely used Recall at K (**R@K**, K=1,5,10) metric, which is defined as the percentage of instances in the ground truth set that appear in the top-K retrieved results, and a higher R@K value indicates better performance. In addition, we calculate the **rSum** metric, which is the sum of all R@K values in both image-to-text and text-to-image directions, reflecting an overall assessment of the performance.

4.1.3. Implementation Details

All experiments are conducted using PyTorch, and trained on NVIDIA GeForce RTX 3090Ti GPU. The Adam optimizer is employed for model optimization, with a mini-batch size 64. Learning rate is set to 0.00002 initially with a decay rate of 0.1 every 20 and 15 epochs for Flickr30K and MS-COCO, respectively, and the maximum epoch number is 40 and 30. The dimension of visual feature D_v is 2048, and that of textual feature D_t is 768. The dimension of the common embedding space D is set to 1024. As for region features, we select $K = 5$ regions from the neighbor set of each region, including the region itself and its neighboring visual context regions. If any region has a neighbor set containing fewer than K regions, we include the region itself as supplementary to compensate for the lack of neighbors. At the similarity reasoning, we divide the feature vectors into 16 segments, each of which are 64-dimensional. The margin α in the triplet ranking loss function is empirically set to 0.2.

4.2. Comparison Results

Results on Winoground. As shown in Table 1, our proposed VL-DE achieves better performance across all metrics compared to recent state-of-the-art models:

- VSE++ (Faghri et al., 2017): This method independently encodes images and sentences into a holistic embedding space, and introduces triplet loss to emphasize hard negative mining.
- VSRN (Li et al., 2019a): This method utilizes GCN with GRU to generate enhanced visual representations to perform both local and global semantic reasoning.
- VSE∞ (Chen et al., 2021): This method presents a generalized pooling function to project local features into global embedding.
- NAAF (Zhang et al., 2022b): This method focuses on the positive effects of matched word-region pairs and the negative effects of mismatched pairs to jointly infer the image-text matching scores.
- CHAN (Pan et al., 2023): This method elaborates on a fine-grained aligning process by mining informative region-word pairs and eliminating redundant or irrelevant alignments.

The pre-trained models on both Flickr30K and MS-COCO are directly obtained from their official GitHub. When pretrained on Flickr30K, VL-DE outperforms existing approaches by a large margin of 5.75% in text score, 3.75% in image score and 3% in group score. When pretrained

Table 1: Results on the Winoground dataset across the text, image and group score metrics. The pre-trained models for comparison on both Flickr30K and MS-COCO are directly obtained from their official GitHub repositories, which are openly available resources. The best results are highlighted in bold.

Model	Text	Image	Group
	Pre-trained on Flickr30K		
VSE++	20.00	5.00	2.75
VSRN	20.00	5.00	3.50
VSE∞	23.75	9.00	4.25
NAAF	26.75	10.50	7.75
CHAN	29.75	12.00	8.75
VL-DE (ours)	**35.50**	**15.75**	**11.75**
	Pre-trained on MS-COCO		
VSE++	22.75	8.00	4.00
VSRN	17.50	7.00	3.75
VSE∞	26.00	8.50	5.50
NAAF	30.00	11.75	8.00
CHAN	31.50	16.50	10.75
VL-DE (ours)	**36.25**	**17.50**	**13.25**
	Richer Features		
VinVL	37.75	17.75	14.50
UNITER$_{large}$	38.00	14.50	10.50
UNITER$_{base}$	32.25	13.25	10.00
ViLLA$_{large}$	37.00	13.25	11.00
ViLLA$_{base}$	30.00	12.00	8.00
VisualBERT$_{base}$	15.50	2.50	1.50
ViLT (ViT-B/32)	34.75	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT$_{base}$	23.75	7.25	4.75
UniT$_{ITMfinetuned}$	19.50	6.25	4.00
CLIP (ViT-B/32)	30.75	10.50	8.00

on the larger COCO dataset, VL-DE also demonstrates significant improvements, obtaining 4.75% higher text score, 1% higher image score, and 2.5% higher group score than the best compared method. We also present the experimental results of some large-scale pre-training models, *i.e.* VinVL (Zhang et al., 2021b), UNITER (Chen et al., 2020b), ViLLA (Gan et al., 2020), VisualBERT (Li et al., 2019b), ViLT (Kim et al., 2021), LXMERT(Tan and Bansal, 2019), ViLBERT (Lu et al., 2019), UniT (Hu and Singh, 2021) and CLIP (Radford et al., 2021). Remarkably, our method achieves comparable results to large-scale pre-training models and even outperforms some of them, despite not utilizing large-scale pre-trained data. This suggests VL-DE is better at capturing compositional structures in *examples* by hierarchically modeling intricate dependencies between visual and linguistic fragments, allowing to successfully distinguish between semantically similar descriptions of visually disparate scenes.

Results on Flickr30K. We compare our VL-DE with a series of recent state-of-the-art methods: CAMERA (Qu et al., 2020), DSRAN (Wen et al., 2020), TERAN (Messina et al., 2021), MEMBER (Li et al., 2021), DIME (Qu et al., 2021), VSE∞ (Chen et al., 2021) VSRN++ (Li et al., 2022b), NAAF (Zhang et al., 2022b), AME (Li et al., 2022a), CHAN (Pan et al., 2023), CMSEI (Ge et al., 2023). The experimental results are directly referenced from respective papers. We report ensemble results by calculating the average similarity of two models, *i.e.* selecting neighboring regions based on distance or IoU scores.

In Table 2, we present the quantitative results of our proposed method on the Flickr30K dataset. Our VL-DE achieves better performance in terms of most evaluation metrics for both sentence and image retrieval tasks, compared to the existing approaches. Specifically, VL-DE obtains 83.7% and 96.7% in terms of text retrieval R@1 and R@5. For image retrieval, VL-DE also achieves the best 65.3% and 88.8% in R@1 and R@5, surpassing other state-of-the-art models, which validates the effectiveness of our approach on image-text retrieval.

Results on MS-COCO. The experimental results on the larger and more complex MS-COCO 1K dataset are shown in Table 2. It can be observed that our VL-DE method outperforms the state-of-the-art methods in most evaluation metrics. Specifically, VL-DE exceeds all previous model with an R@1 score of 82.2% and 68.1% for text retrieval and image retrieval, respectively. As shown in Table 3,for the full 5K test dataset, VL-DE outperforms other methods with 1.9% for text retrieval and 1.4% for image retrieval, respectively. The superior performance of VL-DE demonstrates its ability to model compositional language structures and their correspondence to visual content, and further shows that capturing the intricate dependency relationships within both the textual and visual modalities enables to obtain more accurate and comprehensive cross-modal alignments.

4.3. Ablation Study

To verify the effectiveness of semantic dependency encoding in cross-modal retrieval, we conduct extensive ablation studies on Flickr30K and MS-COCO datasets. In Table 4, the 'VL-DE-Full (*)' denotes the ensemble results of two models, and others are all single models. We compare our full IoU and distance model with four models that remove some sub-modules, and the two full single models outperform all other models: 1) VL-DE-Baseline, which omits both visual and textual dependency encoding. The performance is obviously degraded without considering dependency semantics. 2) VL-DE-w/o Neighbor, which omits visual dependency encoding with neighboring regions and only performs textual dependency encoding. The model

Table 2: Comparison with the state-of-the-art methods on Flickr30K and MS-COCO 1K test set. " * " denotes an ensemble model, and the best results are in bold.

Methods	Flickr30K							MS-COCO 1K						
	Text Retrieval			Image Retrieval				Text Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
CAMERA*$_{2020}$	78.0	95.1	97.9	60.3	85.9	91.7	508.9	77.5	96.3	98.8	63.4	90.9	95.8	522.7
DSRAN*$_{2020}$	77.8	95.1	97.6	59.2	86.0	91.9	507.6	78.3	95.7	98.4	64.5	90.8	95.8	523.5
TERAN*$_{2021}$	79.2	94.4	96.8	63.1	87.3	92.6	513.4	80.2	96.6	**99.0**	67.0	92.2	**96.9**	531.9
MEMBER*$_{2021}$	77.5	94.7	97.3	59.5	84.8	91.0	504.8	78.5	96.8	98.5	63.7	90.7	95.6	523.8
DIME*$_{2021}$	81.0	95.9	98.4	63.6	88.1	93.0	520.0	78.8	96.3	98.7	64.8	91.5	96.5	526.6
VSE∞_{2021}	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++*$_{2022}$	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
NAAF*$_{2022}$	81.9	96.1	98.3	61.0	85.3	90.6	513.2	80.5	96.5	98.8	64.1	90.7	96.5	527.2
AME*$_{2022}$	81.9	95.9	98.5	64.6	88.7	**93.2**	522.8	79.4	96.7	98.9	65.4	91.2	96.1	527.7
CHAN$_{2023}$	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	**96.9**	98.9	66.5	92.1	96.7	532.6
CMSEI*$_{2023}$	82.3	96.4	98.6	64.1	87.3	92.6	521.3	81.4	96.6	98.8	65.8	91.8	96.8	531.1
VL-DE* (ours)	**83.7**	**96.7**	**99.0**	**65.3**	**88.8**	93.1	**526.6**	**82.2**	**96.9**	**99.0**	**68.1**	**92.5**	96.8	**535.6**

Table 3: Comparison with the state-of-the-art methods on MS-COCO 5K test set. " * " denotes an ensemble model, and the best results are in bold.

Methods	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CAMERA*$_{2020}$	55.1	82.9	91.2	40.5	71.7	82.5
DSRAN*$_{2020}$	55.3	83.5	90.9	41.7	72.7	82.8
TERAN*$_{2021}$	59.3	85.8	92.4	45.1	74.6	84.4
MEMBER*$_{2021}$	54.5	82.3	90.1	40.9	71.0	81.8
DIME*$_{2021}$	59.3	85.4	91.9	43.1	73.0	83.1
VSE∞_{2021}	58.3	85.3	82.3	42.4	72.7	83.2
VSRN++*$_{2022}$	54.7	82.9	90.9	42.0	72.2	82.7
NAAF*$_{2022}$	58.9	85.2	92.0	42.5	70.9	81.4
AME*$_{2022}$	59.9	85.2	92.3	43.6	72.6	82.7
CHAN$_{2023}$	59.8	87.2	93.3	44.9	74.5	84.2
CMSEI*$_{2023}$	61.5	86.3	92.7	44.0	73.4	83.4
VL-DE* (ours)	**63.4**	**87.6**	**93.7**	**46.5**	**75.3**	**84.9**

Table 4: Ablation studies about the model design, which are obtained on the Flickr30K and MS-COCO 1K test set. " * " denotes the ensemble results of two models, and others are all single models.

Methods	Text Retr.		Image Retr.	
	R@1	R@5	R@1	R@5
Flickr30K				
VL-DE-Baseline	78.8	94.8	60.2	84.6
VL-DE-w/o Neighbor	79.9	94.9	60.7	85.4
VL-DE-w/o WordDep	79.9	95.3	62.3	85.7
VL-DE-w/o Position	78.7	95.1	61.8	85.9
VL-DE-Full (IoU)	80.4	95.3	61.9	86.7
VL-DE-Full (Distance)	80.8	95.2	63.2	86.5
VL-DE-Full (*)	**83.7**	**96.7**	**65.3**	**88.8**
MS-COCO 1K				
VL-DE-Baseline	79.6	96.6	64.6	91.0
VL-DE-w/o Neighbor	79.7	96.7	64.7	90.9
VL-DE-w/o WordDep	80.1	96.5	66.0	91.4
VL-DE-w/o Position	80.0	96.2	65.7	91.3
VL-DE-Full (IoU)	80.8	96.7	65.5	91.5
VL-DE-Full (Distance)	81.1	96.5	66.1	91.4
VL-DE-Full (*)	**82.2**	**96.9**	**68.1**	**92.5**

cannot achieve high performance since it fail to understand detailed semantics in visually complex regions. 3) VL-DE-w/o WordDep, which omits textual dependency encoding with word dependency and only performs visual dependency encoding. The degradation in performance shows that emphasizing dependency relationships and syntactic roles of words enables to further enhance word embeddings from BERT. 4) VL-DE-w/o Position, which only performs visual dependency encoding but omits the position information of bounding boxes. The performance slightly decreases, verifying that spatial relationships also contribute to incorporating relevant semantics from neighboring regions.

We also conduct ablation experiments on Winoground dataset to verify the effectiveness of semantic dependency encoding in our method. The results are shown in Table 5, where the models are pre-trained on Flickr30K dataset. The results show that both visual and textual dependency encoding contribute to the model's understanding of the compositional structures in language and their corresponding visual content.

4.4. Qualitative Analysis

To obtain a more comprehensive evaluation of our VL-DE's ability to comprehend complex relationships in vision and language, we visualize the matching scores between words and images on Winoground dataset in Figure 3. The green image-caption pair and blue image-caption pair represent a Winoground *example*. The values and

Table 5: Ablation studies about the model design, which are obtained on the Winoground dataset. The models are pre-trained on Flickr30K dataset.

Methods	Text	Image	Group
VL-DE-Baseline	27.75	12.25	9.25
VL-DE-w/o Neighbor	33.25	14.50	12.00
VL-DE-w/o WordDep	32.00	13.25	9.75
VL-DE-w/o Position	29.25	14.00	9.50
VL-DE-Full (IoU)	**35.5**	**15.75**	**11.75**

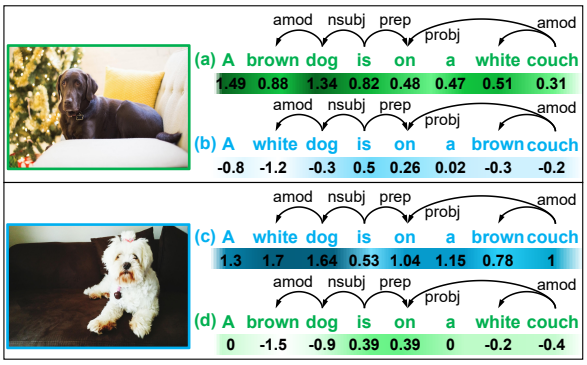


Figure 3: Visualization of the matching scores between words and the image on Winoground. The green image-caption pair and the blue image-caption pair comprise a Winoground *example*. The values and color shadings of the scores reflect the relative importance or saliency of word-image associations, with higher scores and darker shades indicating a stronger correlation between linguistic and visual fragments.

color shadings indicates the relative importance or saliency of the word-image alignments. Higher values and darker shades suggest a stronger correlation or relevance between the words and the corresponding image, while lower values and lighter shades indicate a weaker association. Given the green image above as a query, key words such as "dog" and "couch" in caption (a) are explicitly enhanced by incorporating their adjectival modifiers, namely "brown" and "white" respectively, achieving more precise alignments with the corresponding visual concepts in the query image. As a result, in our VL-DE, these words obtain significantly higher matching scores compared to the words in caption (b). These observations indicate that VL-DE effectively encodes dependency relationships and syntactic funtions of words, enabling it to capture more meaningful and subtle alignments with visual concepts. By explicitly incorporating syntactic information from text, the model is capable of capturing the intricate semantic dependencies for text and vision, encompassing interactions between objects as well as their actions and attributes, even when the linguistic constituents are identical or the visual scene are similar. This highlights the model's advantages in connecting visual scene representations with compositional linguistic structures.

5. Conclusion

In this paper, we propose a novel Visual-Linguistic Dependency Encoding (VL-DE) framework for image-text retrieval. Different from previous methods, VL-DE models complex dependency relationships within textual constituents and visual scene by incorporating syntactic dependencies of words and neighboring relationships between regions, achieving a more comprehensive understanding of local textual and visual fragments. Moreover, segment-wise cross attention mechanism is adopted to capture subtle correspondence across modalities, enabling more accurate and informative cross-modal alignment. Extensive experimental results demonstrate the superiority of our VL-DE.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62222212, 62336001.

Limitations

One limitation is the size of the Winoground dataset used to evaluate the effectiveness of our VL-DE framework in representing compositional relationships. With only 800 image-caption pairs, the dataset is relatively small, which may impact the generalizability of our model. A larger and more diverse dataset could allow for a more robust evaluation and further validate the performance of VL-DE in capturing vision-linguistic compositional structures. Another limitation to consider is that the performance of VL-DE may be influenced by the quality and accuracy of the dependency parsing tool employed during preprocessing. We incorporate the dependency information through a weighted approach that considers the significance of their relationships, thereby mitigating any impact caused by noisy parsing results to some extent. However, advances in dependency parsing itself could further improve the reliability of the structured linguistic inputs for our model.

Ethical Considerations

Image-text retrieval entails the handling of personal and sensitive information, necessitating the implementation of appropriate privacy protection measures during data collection and processing to safeguard against the disclosure of individuals' identities and sensitive information. Additionally, it is

important to address the potential bias present in datasets used for training the model. Without proper curation, datasets poses a risk of reflecting societal biases related to race, gender, culture etc. The model trained on such biased data can perpetuate and amplify those biases. Research teams should proactively evaluate datasets for skew or harms, and the model should not be trained using scraped data of unclear origins. If the model is deployed in public-facing products, it is crucial to establish safeguards ensuring user privacy protection and preventing unauthorized access to personal photos or texts.

Bibliographical References

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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120. Springer.

Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1218–1226.

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*.

Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Xuri Ge, Fuhai Chen, Songpei Xu, Fuxiang Tao, and Joemon M Jose. 2023. Cross-modal semantic enhanced interaction for image-sentence retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1022–1031.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449.

Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. 2019. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 789–795.

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

Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2019. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5754–5763.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

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*, pages 201–216.

Jiangtong Li, Liu Liu, Li Niu, and Liqing Zhang. 2021. Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. *IEEE Transactions on Image Processing*, 30:9193–9207.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022a. Action-aware embedding enhancement for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1323–1331.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019a. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2022b. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):641–656.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Visualgptscore: Visio-linguistic reasoning with multimodal generative pre-training scores. *arXiv preprint arXiv:2306.01879*.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930.

Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. 2022. Gradual: Graph-based dual-modal representation for image-text matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3459–3468.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(4):1–23.

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

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19275–19284.

Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-aware multi-view summarization network for image-text matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1047–1055.

Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. 2019. Knowledge aware semantic concept expansion for image-text matching. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 1, page 2.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3715–3727.

Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H Hsu. 2021. End-to-end video question-answer generation with generator-pretester network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4497–4507.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020a. Consensus-aware visual-semantic embedding for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 18–34. Springer.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020b. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517.

Yaxiong 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*.

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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950.

Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2020. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2866–2879.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096.

Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. 2021. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):43–51.

Jing Yu, Weifeng Zhang, Yuhang Lu, Zengchang Qin, Yue Hu, Jianlong Tan, and Qi Wu. 2020. Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(12):3196–3209.

Pengpeng Zeng, Lianli Gao, Xinyu Lyu, Shuaiqi Jing, and Jingkuan Song. 2021. Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2205–2213.

Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. 2020. Learning to represent image and text with denotation graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 823–839.

Bowen Zhang, Hexiang Hu, Linlu Qiu, Peter Shaw, and Fei Sha. 2021a. Visually grounded concept composition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 201–215.

Huatian Zhang, Zhendong Mao, Kun Zhang, and Yongdong Zhang. 2022a. Show your faith: Cross-modal confidence-aware network for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3262–3270.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022b. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 15661–15670.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

7. Language Resource References

Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. 2014. *Microsoft COCO: Common Objects in Context*. Springer.

Tristan Thrush and Ryan Jiang and Max Bartolo and Amanpreet Singh and Adina Williams and Douwe Kiela and Candace Ross. 2022. *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality*.

Young, Peter and Lai, Alice and Hodosh, Micah and Hockenmaier, Julia. 2014. *From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions*.