Fine-grained Image-text Matching by Cross-modal Hard Aligning Network

Zhengxin Pan[1]
[1]Zhejiang University
panzx@zju.edu.cn

Fangyu Wu[2*]
[2]Xian Jiaotong-liverpool University
fangyu.wu02@xjtlu.edu.cn

Bailing Zhang[3]
[3]NingboTech University
bailing.zhang@nit.zju.edu.cn

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01847

Abstract

Current state-of-the-art image-text matching methods implicitly align the visual-semantic fragments, like regions in images and words in sentences, and adopt cross-attention mechanism to discover fine-grained cross-modal semantic correspondence. However, the cross-attention mechanism may bring redundant or irrelevant region-word alignments, degenerating retrieval accuracy and limiting efficiency. Although many researchers have made progress in mining meaningful alignments and thus improving accuracy, the problem of poor efficiency remains unresolved. In this work, we propose to learn fine-grained image-text matching from the perspective of information coding. Specifically, we suggest a coding framework to explain the fragments aligning process, which provides a novel view to reexamine the cross-attention mechanism and analyze the problem of redundant alignments. Based on this framework, a Cross-modal Hard Aligning Network (CHAN) is designed, which comprehensively exploits the most relevant region-word pairs and eliminates all other alignments. Extensive experiments conducted on two public datasets, MS-COCO and Flickr30K, verify that the relevance of the most associated word-region pairs is discriminative enough as an indicator of the image-text similarity, with superior accuracy and efficiency over the state-of-the-art approaches on the bidirectional image and text retrieval tasks. Our code will be available at https://github.com/ppanzx/CHAN.

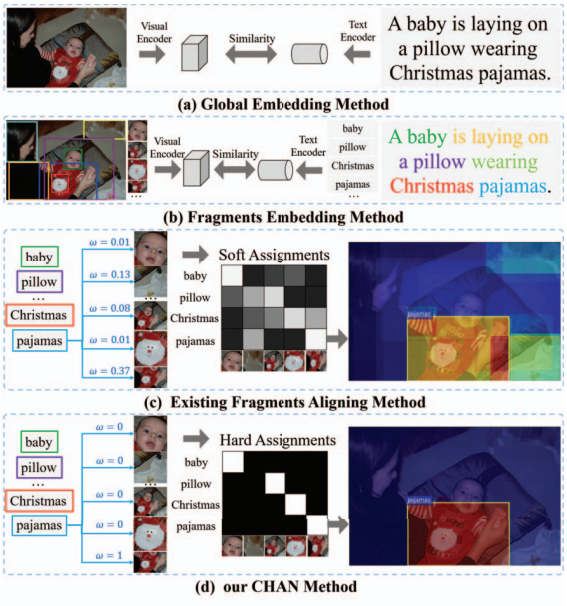

Figure 1. Illustration of different semantic corresponding methods: (a) Global Embedding methods, (b) Fragment Embedding methods, (c) existing Fragment Aligning methods, and (d) our CHAN method. Here ω in (c) and (d) is the attention weight/assignment between the word "pajamas" and the image region, where the region with the maximum attention weight is outlined in yellow below. Compared to existing Fragment Aligning methods which bring redundant alignments, we improve them by attending to the most relevant region while neglecting all of the misalignments.

1. Introduction

With the rapid development of information technology, multi-modal data, like texts, audio, images, and video, has become ubiquitous in our daily life. It is of great value to study multi-modal learning to give computers the ability to process and relate information from multiple modalities. Among the tasks of multi-modal learning, image-text retrieval is the most fundamental one, which paves the way for more general cross-modal retrieval, namely, implementing a retrieval task across different modalities, such as video-text and audio-text. Image-text retrieval has attracted broad attention in recent years [12,21,23]; yet, the key challenges, i.e., bridging the inter-modality gap and achieving the semantic correspondence across modalities, are far from being resolved. A good alignment directly links to correctly measuring the similarity between images and texts.

Early works usually adopt the intuitive idea of global embedding to find the semantic correspondence between a whole picture and the complete sentence [12]. By projecting the overall image and text into a common embedding space, the similarity between heterogeneous samples is measured for the subsequent matching of the two modalities, as

*Corresponding author

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01847

shown in Figure 1(a). However, such a global embedding method often induces background noise, which impedes the correct image-text matching. Recent works have focused on essential fragments [4, 16, 22], such as salient objects in images and keywords in texts, aiming to reduce contributions of uninterested regions as well as irrelevant conjunctions. By introducing the self-attention mechanism, the representation of holistic inputs is replaced by a weighted sum of the local fragments, thereby easing the matching obstacles caused by the noise parts, as shown in Figure 1(b). However, these fragment embedding methods do not explicitly implement fine-grained aligning, as they only focus on the complex aggregation of fragments in a single modality without taking account of correctly learning granular cross-modal semantic consistency.

Based on the consensus that overall image-text similarity is a complex aggregation of the local similarities calculated between the cross-modal fragments [19], the fragments aligning method emphasizes the aggregation of the local similarities rather than the aggregation of the local representations. SCAN [21] and its variants [10, 25, 43, 46] are the representatives of this school of thought, which align image regions and sentence words by locally associating visual semantics, and integrate the semantic similarities between relevant region-word pairs to measure the overall image-text relevance. Specifically, with the core idea of the cross-attention mechanism, they attend to the fragments related to each query fragment from another modality, thus making the semantically consistent granular pairs significantly contribute to the final image-text similarity, and at the same time eliminating or weakening the influence of inconsistent pairs.

However, there are two problems associated with the previous fragments aligning methods: (1) redundant alignments are detrimental to retrieval accuracy. Selecting semantically consistent region-word alignments and rejecting inconsistent alignments is the key to realizing fine-grained image-text matching. However, though semantically consistent alignments can be discovered by the cross-attention mechanism, it is far from enough to achieve an accurate retrieval because these meaningful alignments will be more or less disturbed by other attended fragments irrelevant to the shared semantics. As illustrated in Figure 1(c), given a text fragment "pajamas," current cross-attention-based methods not only attend to the most matched image region but also refer to other regions not exactly relevant, like "cat" and "towel," which will incorrectly estimate the affinity between "pajamas" and irrelevant regions while training. As a result, semantically inconsistent region-word pairs will eventually overwhelm those matched ones, thus compromising the effect from the most matched pairs and degenerating the final performance; (2) caching cross-attention weights is with a massive cost of memory and time. When the cross-attention mechanism is applied to fragments aligning, it is inevitable to calculate the affinities between all the cross-modal fragments, because a query needs to be reconstructed with the attention weights derived from the affinities, which incurs huge memory consumption to store the attention weights. In fact, due to the limited memory, the matching process between each query text/image and the whole image/text set requires a large number of iterations, resulting in a long retrieval time and thus compromising the practical applications of the fragments aligning method.

Inspired by the coding idea widely adopted in content-based image retrieval tasks [14, 17, 32], we propose a coding framework to explain the aligning process and rethink cross-attention-based methods from the view of soft assignment coding. Specifically, we regard each word in a sentence as a query and represent the salient regions in an image as a codebook. Therefore, the aligning of fragments is expressed as an adjustment of the measure of the relationship between query words and visual codewords. The overall image-text similarity is the aggregation of similarities between all queries and all codewords. In this view, the definition of attention weights in a cross-attention mechanism is almost the same as assignments in soft assignment coding [14] scheme, and thus the cross-attention mechanism can be explained as a kind of soft assignment coding method. Based on the assumption that there must exist a sub-region in an image which can best describe every given word in the semantically consistent sentence [19], we deem it unnecessary to consider all or even a selected part of codewords since most of them do not bring benefit for better describing the query words but lowering the efficiency. This insight inspires switching the methodology from soft assignment coding to hard assignment coding [27], with attention to the most relevant word-region/query-codeword pair which is a more accurate indication of semantic consistency between a word and an image, as shown in Figure 1(d). We further propose a novel Cross-modal Hard Aligning Network (CHAN) for fine-grained image-text matching. Our scheme not only discards redundant alignments and better discovers the granular semantic correspondence, but also relieves the costly dense cross-attention matching, thus significantly improving cross-attention baselines both in accuracy and efficiency. Our main contributions can be summarized as follows:

- We propose a coding framework to explain fragments aligning for image-text retrieval and subsequently elaborate on the aligning process of cross-attention mechanism. This elaboration allows us to pinpoint the deficiencies, and propose an improved hard assignment coding scheme.
- With the hard assignment coding scheme, we propose a novel Cross-modal Hard Aligning Network (CHAN), which can accurately discover the shared semantics of im-

age and text by mining the informative region-word pairs and rejecting the redundant or irrelevant alignments.

- Extensive experiments on two benchmarks, i.e., Flickr30K [45] and MS-COCO [5], showing the superiority of CHAN in both accuracy and efficiency compared with state-of-the-art methods.

2. Related Works

Visual Semantic Embedding. Visual Semantic Embedding (VSE) [12] is a general solution for image-text matching, with the core idea of associating the correspondence globally by separately projecting image and text into a common space using two separate networks. The subsequent works improve VSE by seeking better representative common subspace [8, 24, 33, 37], designing more appropriate similarity metrics [11, 36, 38, 40, 44] and proposing Vision Language Pre-training methods [2, 18, 30, 39]. Recent works try to exploit the intrinsic information within each modality and aggregate fine-grained information into VSE in order to produce semantically more consistent embedding for representing images and texts. For example, some works [16, 41, 43] take advantage of self-attention mechanism to focus on essential fragments; VSRN [22] and similar works [6, 23, 26] introduce Graph Convolutional Networks [42] to generate global features with local relationships; VSE∞ [4] demonstrates that aggregating local features by a learnable pooling operation outperform these complex aggregation models mentioned above.

Cross-modal Fragments Aligning. In contrast to the embedding-based methods' poor interpretability of granular semantic consistency, fragments aligning methods directly learn the semantic alignments between image regions and text words. Karpathy et al. [19] make the first attempt to infer finer-level alignments between textual segments and visual regions. They calculate the global image-text similarity by summing up all the region-word similarities. While not all fragments are equally contributive, the following methods are devoted to mining the substantial alignments. SCAN [21] is the representative work in this direction which has attracted great attention. It introduces cross-attention mechanism to concentrate on significant alignments aiming to minimize the misalignments. IMRAM [3] extends SCAN by combining a cross-modal attention unit with a memory distillation unit to refine the cross-modal attention core iteratively. Unlike methods devoted to irrelevant alignments removal, NAAF [46] explores the clues about the disparate fragments and thus discriminating subtle mismatched ones across modalities toward more accurate image-text matching. However, due to intrinsic property of the cross-attention mechanism, above methods obtain a higher accuracy with sacrifice of efficiency, which is vital for retrieval tasks.

Relation to Coding. Our intuition is activated by the famous bag-of-feature (BoF) [32] image coding scheme, which quantizes local invariant descriptors into a set of visual words for efficient image representation. The coding process of BoF approach sheds light on fine-grained image-text matching as we can compare the local regionword alignment and high-level global alignment with the essential steps of BoF, namely, (1) Coding and (2) Pooling. Such a similarity between BoF and image-text aligning inspires our viewpoint of treating cross-modal aligning using a unified coding framework. Further, the cross-modal matching can be expounded as a special case of soft assignment coding [14, 27, 34], and the approach of mining the most relevant visual codeword for a query is consistent with a hard assignment coding method [17, 32].

3. Cross-modal Hard Aligning Network

3.1. Coding Framework for Fragment Alignment

We tackle the granular semantic matching problem with coding framework. Formally, for a set of text features $\boldsymbol{T} = \{\boldsymbol{t}_i \mid i \in [1, \cdots, L],\ \boldsymbol{t}_i \in \mathbb{R}^d\}$, each text feature $\boldsymbol{t}_i$ encodes a i-th word in a sentence, where L is the length of a sentence; for a set of visual features $\boldsymbol{V} = \{\boldsymbol{v}_j \mid j \in [1, \cdots, K],\ \boldsymbol{v}_j \in \mathbb{R}^d\}$, each visual feature $\boldsymbol{v}_j$ encodes a salient j-th region in an image, where K is the number of salient regions in an image; d is the dimension of common embedding space. The semantic relevance between a sentence $\mathcal{T}$ and an image $\mathcal{V}$ can be scrutinized with the information coding framework via two processes, namely, coding and pooling, as expounded below.

The calculation of the similarity between word $\boldsymbol{t}_i$ and an image $\mathcal{V}$ can be approached by appropriate information encoding process. Concretely, let a word $\boldsymbol{t}_i$ in a sentence be the query, and the image $\mathcal{V}$ can be represented by a codebook, where every region $\boldsymbol{v}_j$ in $\mathcal{V}$ is treated as a codeword. The similarity of $\boldsymbol{t}_i$ and $\mathcal{V}$ is thus transformed to be the reconstruction error between $\boldsymbol{t}_i$ and $\hat{\boldsymbol{t}}_i$ obtained using codebook $\boldsymbol{V} = \{\boldsymbol{v}_j\}_{j=1}^K$, formally as:

$$\boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V}) = \mathcal{S}(\boldsymbol{t}_i, \hat{\boldsymbol{t}}_i) \tag{1}$$

where $\mathcal{S}$ denotes the similarity metric function. In contrast to the euclidean metric widely used in BoF methods [14,34], $\mathcal{S}$ in cross-modal retrieval tasks is usually adopted as the cosine metric function, that is, $\mathcal{S}(\boldsymbol{t}_i, \hat{\boldsymbol{t}}_i) = \frac{\boldsymbol{t}_i^\top \hat{\boldsymbol{t}}_i}{\|\boldsymbol{t}_i\| \cdot \|\hat{\boldsymbol{t}}_i\|}$. And $\hat{\boldsymbol{t}}_i$ in Eq. 1 indicates the attended version of the query $\boldsymbol{t}_i$ relative to codebook $\boldsymbol{V} = \{\boldsymbol{v}_j\}_{j=1}^K$, which is defined as:

$$\hat{\boldsymbol{t}}_i = \sum_{j=1}^{K} \omega_{ij} \boldsymbol{v}_j \tag{2}$$

where ω_{ij} is the weighting factor of $\boldsymbol{v}_j$. By defining $\boldsymbol{s}_{ij} =$

$\mathcal{S}(\boldsymbol{t}_i, \boldsymbol{v}_j) = \frac{\boldsymbol{t}_i^\top \boldsymbol{v}_j}{\|\boldsymbol{t}_i\| \cdot \|\boldsymbol{v}_j\|}$ as the similarity between query $\boldsymbol{t}_i$ and codeword $\boldsymbol{v}_j$, ω_{ij} is positively correlated with $\boldsymbol{s}_{ij}$ generally.

The final similarity score between sentence $\mathcal{T}$ and image $\mathcal{V}$ is obtained by a proper pooling operation, which combines all of the word-image scores $\boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V}),\ \forall i$. Taking LogSumExp pooling (LSE-Pooling) [21] as an example, the overall similarity can be summarised as:

$$\boldsymbol{s}(\mathcal{T}, \mathcal{V}) = \frac{1}{\lambda} \log \sum_{i=1}^{L} \exp(\lambda \boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V})) \tag{3}$$

where λ is a scaling factor that determines how much to magnify the importance of the most relevant word-image pair.

Particularly, Eq. 2 expresses a cross-attention mechanism for image-text matching, where the weighting factor ω_{ij} is linked with $\boldsymbol{s}_{ij}$ with a Gaussian kernel function under the assumption that the similarity between a query and a codeword can be described by a normal distribution [14,34], that is, $\omega_{ij} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{\boldsymbol{s}_{ij}}{2\sigma^2}\right)$, where $\boldsymbol{s}_{ij}$ represents the similarity and σ determines the size of the kernel. After normalization, ω_{ij} is represented as:

$$\omega_{ij} = \frac{\exp(\boldsymbol{s}_{ij}/\tau)}{\sum_{j=1}^{K} \exp(\boldsymbol{s}_{ij}/\tau)} \tag{4}$$

where $\sum_{j=1}^{K} \exp(\boldsymbol{s}_{ij}/\tau)$ is the normalization factor and τ is the a smooth parameter [7]. It should be noted that there is a slight difference between Eq. 4 and the definition of ω_{ij} in [21] where the similarities are empirically thresholded at zero and $\boldsymbol{s}_{ij}$ is normalized. We argue the intuition of soft assignment coding proposed in [14] is not suitable for cross-modal retrieval tasks because there is always a suitable codeword in the vocabulary appropriately representative for a word in the matched sentence.

3.2. Hard Assignment Coding

Our insight is that if a sentence is semantically consistent with an image, then every word can be representative of an appropriate region of the image, while most of the other regions are much more irrelevant. In other words, the similarity $\boldsymbol{s}_{ik}$ between the query word $\boldsymbol{t}_i$ and its semantically corresponding codeword $\boldsymbol{v}_k$, where $k = \underset{j=1\cdots K}{arg\ max}(\boldsymbol{s}_{ij})$, is much larger than $\boldsymbol{s}_{ij,j\neq k}$, which means that τ in Eq. 4 should be very small to describe such a distribution.

We extend τ in Eq. 4 to be approaching 0 and derive the Hard Assignment Coding, in which the weighting factor ω_{ij} is redefined as:

$$\omega_{ij} = \begin{cases} 1\ ,\ if\ j = \underset{j'=1\cdots K}{arg\ max}(\boldsymbol{s}_{ij'}); \\ 0\ ,\ otherwise. \end{cases} \tag{5}$$

We combine Eq. 1, Eq. 2 and Eq. 5 then rewrite the similarity between $\boldsymbol{t}_i$ and $\mathcal{V}$ as:

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V}) &= \frac{\boldsymbol{t}_i^\top \hat{\boldsymbol{t}}_i}{\|\boldsymbol{t}_i\| \cdot \|\hat{\boldsymbol{t}}_i\|} = \frac{\boldsymbol{t}_i^\top \boldsymbol{v}_k}{\|\boldsymbol{t}_i\| \cdot \|\boldsymbol{v}_k\|} \\ &= \boldsymbol{s}_{ik} = \underset{j=1\cdots K}{max}(\boldsymbol{s}_{ij}) \end{aligned} \tag{6}$$

where k has been defined above as the index of the codeword $\boldsymbol{v}_k$ most similar to query $\boldsymbol{t}_i$. From Eq. 6 we can learn that the word-image similarity $\boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V})$ is represented as the maximum word-region similarity. In this way, the hard assignment coding method avoids caching the attention weights $\{\omega_{ij}\}_{j=1}^{K}$, which will greatly reduce the time and space complexity. Besides, it skips the procedure of constructing the attended query $\hat{\boldsymbol{t}}_i$ using unnecessary codewords but only preserves the most informative one to indicate whether the semantic contained in a word is included in an image.

Discussions about the effectiveness. Following [27], the mechanism of the hard assignment coding can be interpreted probabilistically. Put it in a nutshell, the hard assignment coding models the joint probability of the semantic co-occurrence between a word and an image to learn the discriminative representation by maximizing its lower bound embodied in the most informative query-codeword relevance. Now let $P(\boldsymbol{t}_i, \mathcal{V})$ denote the probability of the semantic consistency of query $\boldsymbol{t}_i$ and a codebook $\mathcal{V}$ and let $P(\boldsymbol{t}_i, \boldsymbol{v}_j)$ denote the probability of the semantic consistency of query $\boldsymbol{t}_i$ and a codeword $\boldsymbol{v}_j$. Without loss of generality, we define $P(\boldsymbol{t}_i, \mathcal{V})$ being proportional to the word-image similarity $\boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V})$, i.e., $P(\boldsymbol{t}_i, \mathcal{V}) \propto \boldsymbol{s}(\boldsymbol{t}_i, \mathcal{V})$, so is $P(\boldsymbol{t}_i, \boldsymbol{v}_j)$, i.e., $P(\boldsymbol{t}_i, \boldsymbol{v}_j) \propto \boldsymbol{s}_{ij}$. Firstly, let us sample a subset of codewords $\{\boldsymbol{v}_j\}_{j=1}^{R}$ including $\boldsymbol{v}_k$ in a codebook $\mathcal{V}$ that all of these codewords are independent of each other. In this case, $P(\boldsymbol{t}_i, \mathcal{V})$ can be defined using $P(\boldsymbol{t}_i, \boldsymbol{v}_j)$ as:

$$\begin{aligned} P(\boldsymbol{t}_i, \mathcal{V}) &= 1 - \prod_{j=1}^{R} (1 - P(\boldsymbol{t}_i, \boldsymbol{v}_j)) \\ &\geq 1 - (1 - P(\boldsymbol{t}_i, \boldsymbol{v}_k)) = P(\boldsymbol{t}_i, \boldsymbol{v}_k) \end{aligned} \tag{7}$$

That is, the semantic consistency between query $\boldsymbol{t}_i$ and its most relevant codeword $\boldsymbol{v}_k$ is a lower bound of the probability of the presence of a word in an image. However, it is intractable for soft assignment coding to measure the relationship between $P(\boldsymbol{t}_i, \mathcal{V})$ and $P(\boldsymbol{t}_i, \boldsymbol{v}_j)$ because of the dependency of some codewords, which means that the soft assignment coding is not as effective in indicating granular correspondence as hard assignment coding. Furthermore, the above analysis provides an intriguing interpretation of hard assignment coding for cross-modal tasks. We can consider the words in a sentence as a collection of "semantic

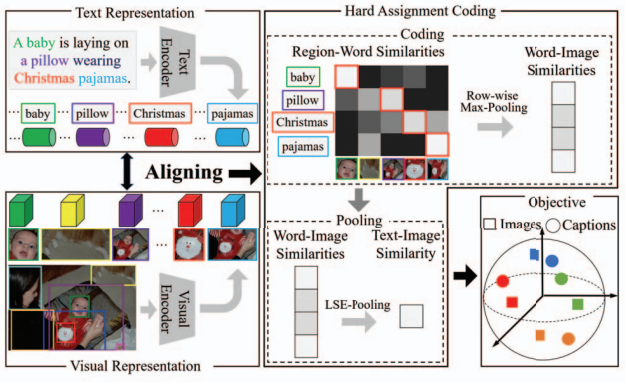

Figure 2. An overview of the proposed CHAN network. It consists of four modules: Visual representation, Text representation, Hard assignment coding and Objective function. The final form of Hard assignment coding module is obtained by performing row-wise max-pooling and LSE-pooling over the similarity matrix.

detector," and the coding process as the execution of these detectors on different locations within an image. The best response of each detector is then recorded by its highest coding coefficients. From this interpretation, the denser the sampled local features are, the more reliable the responses are.

Discussions about the efficiency. Consider a visual feature set $\boldsymbol{V} \in \mathbb{R}^{B_1 \times K \times d}$ and a text feature set $\boldsymbol{T} \in \mathbb{R}^{B_2 \times L \times d}$, where B_1 and B_2 denote the number of images and captions, respectively. To obtain the final image-text similarity, both hard assignment coding and soft assignment coding require the calculation of the assignment matrix $\boldsymbol{A} \in \mathbb{R}^{B_1 \times B_2 \times K \times L}$, resulting in the same time complexity of $\mathcal{O}(B_1 B_2 K L d)$. However, hard assignment coding has a linearly better efficiency compared to soft assignment coding under the condition of infinite memory, as it no longer needs to calculate the attended version of the text feature set $\hat{\boldsymbol{T}} \in \mathbb{R}^{B_1 \times B_2 \times L \times d}$, as shown in Eq. 6. Furthermore, due to the fact that $K \ll d$, the spatial complexity of hard assignment coding ($\mathcal{O}(B_1 B_2 K L)$) is significantly lower than that of soft assignment coding ($\mathcal{O}(B_1 B_2 L d)$), which inherently suffers from the issue of high memory consumption. This makes hard assignment coding much more efficient than soft assignment coding without the need for iterations.

3.3. Cross-modal Hard Alignment Network

As illustrated in Figure 2, our proposed CHAN is composed of four modules with more details elaborated below.

Visual representation. For each input image $\mathcal{V}$, we follow [21] to extract top-K region-level features, with the Faster R-CNN [31] model pre-trained on Visual Genomes [20] using bottom-up and top-down attention (BUTD) [1]. We utilize a fully-connected layer to embed them into d-dimensional vectors. Thereafter, like [13], we add a self-attention layer [35] to inject the contextual information for each local region feature, and subsequently constitute a discriminative codebook with K visual codewords $\{\boldsymbol{v}_j\}_{j=1}^{K}$.

Text representation. We define two formulations for text representation, based on bi-direction gated recurrent unit (BiGRU) or pre-trained Bert [9]. For BiGRU-based formulation, each sentence $\mathcal{T}$ is tokenized to several words. We embed every word using a pre-trained Glove vector [29] like [8, 46] and feed all vectors into a BiGRU to obtain text queries $\{\boldsymbol{t}_i\}_{i=1}^{L}$ by averaging the forward and backward hidden states at each time step. For Bert-based formulation, we obtain word-level vectors from the last layer of pre-trained Bert, then leverage a fully-connected layer to embed them into d-dimensional vectors.

Hard assignment coding. For a given set of text queries $\boldsymbol{T} = \{\boldsymbol{t}_i\}_{i=1}^{L}$ and a set of visual codewords $\boldsymbol{V} = \{\boldsymbol{v}_j\}_{j=1}^{K}$ obtained above, we first normalize each item of them with ℓ_2-norm, then calculate the cosine similarity matrix $\boldsymbol{S} \in \mathbb{R}^{L \times K}$ between all query-codeword pairs by matrix multiplication, i.e., $\boldsymbol{S} = \boldsymbol{T}\boldsymbol{V}^{\top}$. According to Eq. 6, we implement hard assignment coding by performing row-wise max-pooling over $\boldsymbol{S}$ to align every word with an image. Then, we implement LSE-pooling to aggregate all word-image similarities with respect to each word. On the whole, the hard assignment coding for fine-grained cross-modal matching can be summarized as:

$$\boldsymbol{s}(\mathcal{T}, \mathcal{V}) = \frac{1}{\lambda} \log \sum_{i=1}^{L} \exp(\lambda \underset{j=1 \cdots K}{max} \boldsymbol{S}) \tag{8}$$

Objective function. Following existing approaches [10, 21, 46], we minimize the hinge-based bi-direction triplet ranking loss with online hard negative mining proposed by VSE++ [11], to cluster together the word and its most relevant image region in a matched image-sentence pair while guaranteeing the word is far apart from its most relevant region in a mismatched pair. The objective function is written as:

$$\begin{aligned} \mathcal{L} = \sum_{(\mathcal{T},\mathcal{V}) \sim \mathcal{D}} & [\alpha + \boldsymbol{s}(\mathcal{T}, \hat{\mathcal{V}}) - \boldsymbol{s}(\mathcal{T}, \mathcal{V})]_+ \\ & + [\alpha + \boldsymbol{s}(\hat{\mathcal{T}}, \mathcal{V}) - \boldsymbol{s}(\mathcal{T}, \mathcal{V})]_+ \end{aligned} \tag{9}$$

where α is the margin parameter. $(\mathcal{T}, \mathcal{V})$ is a matched image-sentence pair in dataset $\mathcal{D}$ and $[x]_+ \equiv max(x, 0)$. $\hat{\mathcal{V}} = \underset{\mathcal{V}' \neq \mathcal{V}}{arg\,max}\ \boldsymbol{s}(\mathcal{T}, \hat{\mathcal{V}})$ and $\hat{\mathcal{T}} = \underset{\mathcal{T}' \neq \mathcal{T}}{arg\,max}\ \boldsymbol{s}(\hat{\mathcal{T}}, \mathcal{V})$ denote the hardest image and the hardest sentence within a training mini-batch, respectively.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on Flickr30K [45] and MS-COCO [5] datasets. The MS-COCO dataset contains 123,287 images, and each image is annotated with 5 annotated captions. We use the data split practice of [11, 21, 34] where there are 113,287 images for training, 5,000 images for validation and 5,000 images for testing. We report results by averaging over 5 folds of 1K test images and testing on the full 5K images. The Flickr30K dataset contains 31,783 images collected from the Flickr website with 5 corresponding captions each. Following the split in [11], we use 1,014 images for validation, 1,000 images for testing and the rest for training.

Evaluation Metrics. As a common practice in information retrieval, we measure the performance by $R@K$, defined as the percentage of queries correctly matched in the closest K retrieved instances. A higher $R@K$ indicates better performance. To show the overall matching performance, we sum up all recall values as RSUM at both image-to-text and text-to-image directions.

4.2. Comparison Results

Quantitative Comparison. We compare CHAN with recent state-of-the-art methods on the two benchmarks. In contrast to methods [10, 21, 28, 34, 46] which boost their performance by averaging the similarities from two separate models, we do not leverage ensemble approaches but only report our single-model retrieval results. For a fairer comparison, we divide the methods according to their feature extraction backbones.

Table 2 shows the quantitative results of our CHAN approach on Flickr30K test set. We can observe that CHAN outperforms all other methods with RSUM= 507.8 for BiGRU-based CHAN and RSUM= 518.5 for Bert-based CHAN. Compared with the baseline model SCAN, our BiGRU-based CHAN achieves over 12.3% and 11.6% improvement at $R@1$ for two-direction retrieval. Our Bert-based CHAN also outperforms other state-of-the-art methods with a large margin of over 5% at RSUM.

The quantitative comparison results on the larger and more complicated MS-COCO are shown in Table 1. Our BiGRU-based CHAN performs much better than other counterparts such as SGRAF [10] and NAAF [46] on both COCO 5-fold 1K and COCO 5K test sets. For Bert-based models, it can be seen from the bottom of Table 1 that our CHAN achieves slightly better results than the ensemble TERAN [28]. The improved accuracy of our proposed CHAN demonstrates that hard assignment coding is able to effectively uncover the common semantic from image and text while eliminating the influence of irrelevant region-word pairs.

Inference Efficiency Analysis. In addition to the improvements in accuracy, CHAN also outperforms recent state-of-the-art fragments aligning methods in terms of efficiency. In Figure 3, we present the comparison in RSUM relative to total inference time on COCO 5K, COCO 1K, and Flickr30K test sets with recent methods with publicly available source code. To compare more fairly, we reimplement SCAN by merely replacing hard assignment coding in our BiGRU-based CHAN with soft assignment coding (denoted as SCAN(ours)). Regarding the total inference time, our methods (CHAN(BiGRU), CHAN(Bert)) are over 10 times faster than other recent methods and 3 times faster than SCAN(ours) while obtaining the best accuracy on three test sets.

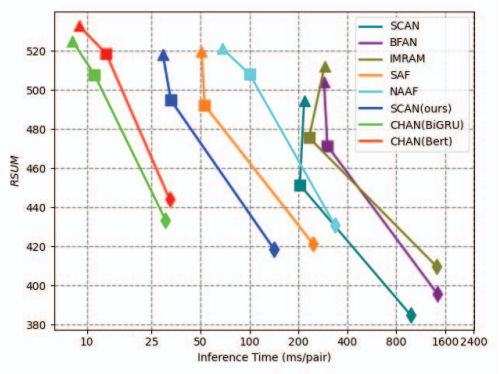

Figure 3. Performance comparison of accuracy (RSUM) and efficiency (ms/pair) between different methods on three test sets, where △, □ and ◊ represent results on COCO 5-fold 1K, Flickr30K and COCO 5K, respectively.

4.3. Ablation Study

We conduct detailed ablation studies on COCO 5K test set to investigate the effectiveness of each component of our CHAN. Without additional notation, we use the BiGRU-based CHAN as our baseline.

Effects of Network Structure. In Table 3 we investigate the effectiveness of different coding structures in our CHAN:

- Coding Types. We first compare the coding types for cross-modal retrieval, e.g., hard assignment coding with text query and visual codebook (denoted as Visual Codebook), with visual query and text codebook (Text Codebook) and soft assignment coding with the cross-attention mechanism (Cross-Attention). Compared with cross-attention-based CHAN, our CHAN baseline using visual codebook achieves an improvement of 4.1% at RSUM, which verifies the advantage of accuracy improvement by attending to the most relevant fragment rather than

Table 1. Image-Text Retrieval Results of CHAN method on COCO 5K and COCO 5-fold 1K test set, using different visual and text backbones (denoted by **bold section title**). ⋆: Ensemble results of two models. The best (in RSUM) are marked **bold** in **red**. Global, Fragment and Aligning refer to global embedding method, fragment embedding method and fragment aligning method mentioned in § 1.

METHOD	TYPE	COCO 5-fold 1K Test [5]							COCO 5KTest [5]						
		IMG → TEXT			TEXT → IMG			RSUM	IMG → TEXT			TEXT → IMG			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
ResNet-152 [15] + BiGRU															
VSE++ $[11]_{2017}$	Global	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
VSE∞ $[4]_{2021}$	Global	76.5	95.3	98.5	62.9	90.6	95.8	519.6	55.1	81.9	89.9	40.9	70.6	81.5	419.9
BUTD [1] + BiGRU															
VSRN⋆ $[22]_{2019}$	Fragment	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
VSE∞ $[4]_{2021}$	Fragment	78.5	96.0	98.7	61.7	90.3	95.6	520.8	56.6	83.6	91.4	39.3	69.9	81.1	421.9
SCAN⋆ $[21]_{2018}$	Aligning	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
IMRAM⋆ $[3]_{2020}$	Aligning	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF⋆ $[10]_{2021}$	Aligning	79.3	96.7	98.3	64.5	90.0	95.8	524.6	55.8	83.0	91.0	42.0	72.4	82.1	426.3
CGMN $[6]_{2022}$	Aligning	76.8	95.4	98.3	63.8	90.7	95.7	520.7	53.4	81.3	89.6	41.2	71.9	82.4	419.8
NAAF $[46]_{2022}$	Aligning	78.1	96.1	98.6	63.5	89.6	95.3	521.2	58.9	85.2	92.0	42.5	70.9	81.4	430.9
CHAN (ours)	Aligning	79.7	96.7	98.7	63.8	90.4	95.8	**525.0**	60.2	85.9	92.4	41.7	71.5	81.7	**433.4**
BUTD [1] + BERT [9]															
MMCA $[41]_{2020}$	Aligning	74.8	95.6	97.7	61.6	89.8	95.2	514.7	54.0	82.5	90.7	38.7	69.7	80.8	416.4
VSE∞ $[4]_{2021}$	Aligning	79.7	96.4	98.9	64.8	91.4	96.3	527.5	58.3	85.3	92.3	42.4	72.7	83.2	434.3
TERAN⋆ $[28]_{2021}$	Aligning	80.2	96.6	99.0	67.0	92.2	96.9	531.9	59.3	85.8	92.4	45.1	76.4	84.4	443.4
VSRN++⋆ $[23]_{2022}$	Aligning	77.9	96.0	98.5	64.1	91.0	96.1	523.6	54.7	82.9	90.9	42.0	72.2	82.7	425.4
CHAN (ours)	Aligning	81.4	96.9	98.9	66.5	92.1	96.7	**532.6**	59.8	87.2	93.3	44.9	74.5	84.2	**443.9**

Table 2. Image-Text Retrieval Results of CHAN method on Flickr30K test set.

METHOD	IMG → TEXT			TEXT → IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
ResNet-152 [15] + BiGRU							
VSE++ $[11]_{2017}$	52.9	80.5	87.2	39.6	70.1	79.5	409.8
VSE∞ $[4]_{2021}$	77.1	94.5	97.1	58.5	84.1	89.6	500.9
BUTD [1] + BiGRU							
SCAN⋆ $[21]_{2018}$	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN⋆ $[22]_{2019}$	71.3	90.6	96.0	54.7	81.8	88.2	482.6
BFAN⋆ $[22]_{2019}$	68.1	91.4	96.2	50.8	78.4	86.0	470.9
IMRAM⋆ $[3]_{2020}$	74.1	93.0	96.6	53.9	79.4	87.2	484.2
VSE∞ $[4]_{2021}$	76.5	94.2	97.7	56.4	83.4	89.9	498.1
SGRAF⋆ $[10]_{2021}$	78.4	94.6	97.5	58.2	83.0	89.1	500.8
CGMN $[6]_{2022}$	77.9	93.8	96.8	59.9	85.1	90.6	504.1
NAAF $[46]_{2022}$	79.6	96.3	98.3	59.3	83.9	90.2	507.6
CHAN (ours)	79.7	94.5	97.3	60.2	85.3	90.7	**507.8**
BUTD [1] + BERT [9]							
MMCA $[41]_{2020}$	74.2	92.8	96.4	54.8	81.4	87.8	487.4
VSE∞ $[4]_{2021}$	81.7	95.4	97.6	61.4	85.9	91.5	513.5
TERAN⋆ $[28]_{2021}$	79.2	94.4	96.8	63.1	87.3	92.6	513.4
VSRN++⋆ $[23]_{2022}$	79.2	94.6	97.5	60.6	85.6	91.4	508.9
CHAN (ours)	80.6	96.1	97.8	63.9	87.5	92.6	**518.5**

maximally mining the shared semantics. While text-codebook-based CHAN causes degradation of 10.4% at RSUM, which may be attributed by the cross-modal heterogeneity that a natural image is too detailed to be described by several words in a sentence.

- Pooling Types. Pooling is important for eliminating the effect of less informative query words. It can be seen that LSE-Pooling performs the best while Max-Pooling performs the worst. It's notable that Sum-Pooling adopted in [28] performs not better than Average-Pooling in our setting since the length of a sentence is stochastic, and the similarity between an image and different sentences cannot be compared fairly.

Table 3. Ablation studies on COCO 5K test set about the network structure.

METHOD	IMG → TEXT			TEXT → IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
Coding Types							
Cross-Attention	54.8	83.7	91.2	39.6	69.0	80.3	418.6
Visual Codebook	60.2	85.9	92.4	41.7	71.5	81.7	**433.4**
Textual Codebook	48.8	80.1	88.9	35.2	66.6	78.4	398.0
Pooling Types							
Max-Pooling	34.8	65.1	76.7	20.7	50.1	64.2	311.7
Average-Pooling	58.8	85.4	91.9	42.4	71.5	81.8	431.9
Sum-Pooling	58.4	85.1	92.1	41.3	70.5	80.7	428.1
Softmax-Pooling	54.7	83.0	91.3	40.3	70.0	81.0	420.5
LSE-Pooling	60.2	85.9	92.4	41.7	71.5	81.7	**433.4**

Effects of the Size of Codebook K. We visualize the retrieval accuracy relative to the total inference time under different sizes of codebook in Figure 4. It is evident that as the number of codewords K gets larger, the accuracy-efficiency curve shifts upper-right, demonstrating that an incremental K benefits the accuracy of hard assignment coding with sacrifices of the efficiency. This is consistent with our discussion at the end of § 3.2 but differs from the result in [21], where $K = 36$ yields the best results while the performance drops by introducing noisy information after

K becomes larger than 36. We attribute this discrepancy to the property that hard assignment coding can mine the most informative region and preserve the most shared semantic, thus performing better with a larger K.

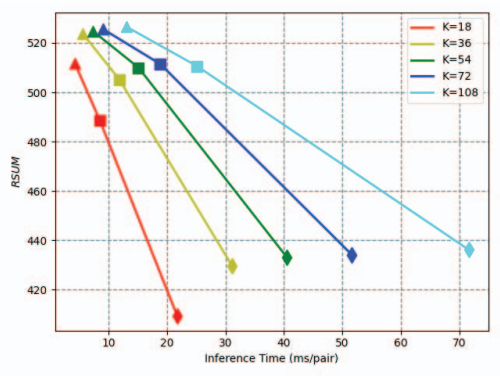

Figure 4. Performance comparison of accuracy (RSUM) and efficiency (ms/pair) between different sizes of codebook K.

4.4. Visualization and Case Study

In order to better understand the intuitive difference between our CHAN and existing cross-attention-based methods and verify our effectiveness, we visualize the coding attention weights/assignments $\{\omega_{ij}\}$ between the given word in a sentence and the visual codebook. As shown in Figure 5, for the final attention map, the attention score at each pixel location is obtained by adding up scores of all regions it belongs to, and the region with the maximum attention is outlined in yellow. We can see that cross-attention-based methods are either unable to detect the matched alignments that lead to the deviation between the highlighted/red region and the ideally semantic matching region, or incompetent to eliminate the meaningless alignments thus causing the outlined region far from the ground-truth. As a contrast, our CHAN solves these problems almost perfectly by introducing hard assignment coding. Regarding specific cases, CHAN is capable of learning the relevant region that best represents the given words, such as "shirt" in Q1 and "log" in Q4. When it comes to plural nouns like "men" in Q2 and words related to relationships such as "drawing" in Q3, CHAN tends to represent an appropriate sub-region by uniting the objects, whereas current methods use a series of sub-regions. These cases demonstrates that both methods are reasonable for situations involving multiple objects.

5. Conclusion

In this paper, we re-examine existing fine-grained cross-modal aligning methods, and propose a coding framework to explore the alignments between salient regions in an image and words in a sentence. Based on the coding framework, we introduce the hard assignment coding scheme

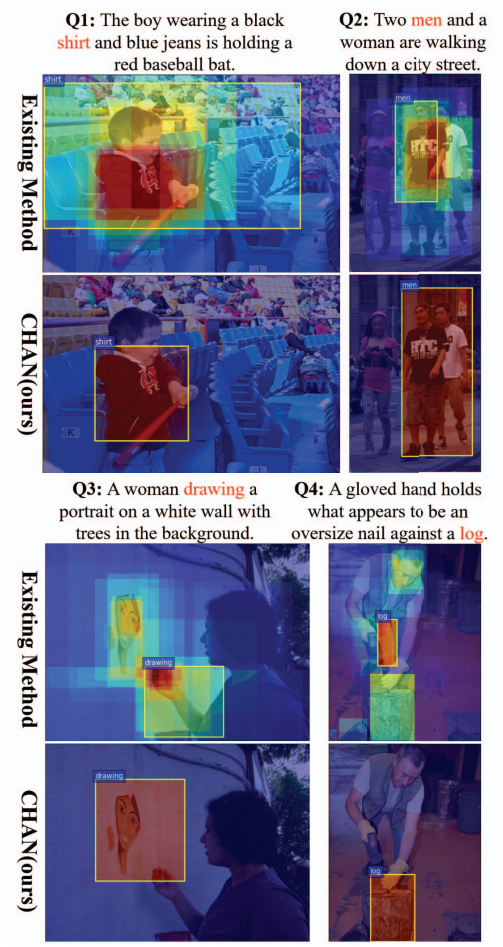

Figure 5. Visual comparison between CHAN and existing cross-attention-based method. The original image is colorized according to the attention score at each pixel location; the larger the score, the warmer the color. The most relevant region with the maximum attention score is outlined in yellow.

and develop our CHAN model to improve cross-attention-based approaches. Extensive experiments on MS-COCO and Fliker30K datasets demonstrate the resulting model consistently outperforms the state-of-the-art methods, both in accuracy and efficiency. Ablation study further validates the theoretical effectiveness of our CHAN model. Our further researches include extending our works along the line of maximizing the mutual information between an image and a text from the view of the information theory.

Acknowledgements

The research was supported by the Zhejiang Provincial Natural Science Foundation of China No. LY23F020014; Ningbo 2025 Key Scientific Research Programs No. 2019B10128; National Natural Science Foundation of China No. 62172356.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 5, 7

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020. 3, 7

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021. 2, 3, 7

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 6, 7

[6] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *TOMM*, 18(4):1–23, 2022. 3, 7

[7] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *NeurIPS*, 28, 2015. 4

[8] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424, 2021. 3, 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5, 7

[10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. Technical report, AAAI, 2021. 2, 5, 6, 7

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++:improving visual-semantic embeddings with hard negatives. In *BMVC*, 2017. 3, 5, 6, 7

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013. 1, 3

[13] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *ACMMM*, pages 5185–5193, 2021. 5

[14] Jan C van Gemert, Jan-Mark Geusebroek, Cor J Veenman, and Arnold WM Smeulders. Kernel codebooks for scene categorization. In *ECCV*, pages 696–709. Springer, 2008. 2, 3, 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[16] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, pages 2310–2318, 2017. 2, 3

[17] Yongzhen Huang, Zifeng Wu, Liang Wang, and Tieniu Tan. Feature coding in image classification: A comprehensive study. *PAMI*, 36(3):493–506, 2013. 2, 3

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 3

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2, 3

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 5

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 1, 2, 3, 4, 5, 6, 7

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *CVPR*, pages 4654–4662, 2019. 2, 3, 7

[23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Image-text embedding learning via visual and textual semantic reasoning. *PAMI*, 2022. 1, 3, 7

[24] Alex Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. In *ACL*, pages 3013–3035, 2022. 3

[25] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACMMM*, pages 3–11, 2019. 2

[26] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020. 3

[27] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *ICCV*, pages 2486–2493. IEEE, 2011. 2, 3, 4

[28] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *TOMM*, 17(4):1–23, 2021. 6, 7

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 5

[32] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 3, pages 1470–1470, 2003. 2, 3

[33] Thomas Theodoridis, Theocharis Chatzis, Vassilios Solachidis, Kosmas Dimitropoulos, and Petros Daras. Cross-modal variational alignment of latent spaces. In *CVPRW*, pages 960–961, 2020. 3

[34] Jan C Van Gemert, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *PAMI*, 32(7):1271–1283, 2009. 3, 4, 6

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 5

[36] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 3

[37] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACMMM*, pages 154–162, 2017. 3

[38] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *PAMI*, 41(2):394–407, 2018. 3

[39] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 3

[40] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *PAMI*, 2021. 3

[41] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *CVPR*, pages 10941–10950, 2020. 3, 7

[42] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016. 3

[43] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACMMM*, pages 2088–2096, 2019. 2, 3

[44] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, pages 15671–15680, 2022. 3

[45] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 3, 6

[46] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *CVPR*, pages 15661–15670, 2022. 2, 3, 5, 6, 7