LEARNING ASYMMETRIC VISUAL SEMANTIC EMBEDDING FOR IMAGE-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning visual semantic similarity is the key challenge to bridge the correspondences between images and texts. However, there are many inherent variations between vision and language data, such as information density, i.e., images can contain textual information from multiple different views, which makes it difficult to accurately compute the similarity between these two modality data. In the mainstream methods, global-level methods cannot effectively handle the above problem, while local-level methods need complicated mechanism, which significantly affects the retrieval efficiency. In this paper, we propose Asymmetric Visual Semantic Embedding (AVSE), which aims to design a novel model to learn visual semantic similarity by explicitly considering the difference in information density between the two modalities and eschew the prohibitive computations. Specifically, to keep the information density of images, AVSE exploits the large spatial redundancy of image regions to capture and concatenate multi-view features as image embedding. It also has a novel module to efficiently calculate the visual semantic similarity of asymmetric image embedding and text embedding via dividing embeddings into many semantic blocks with the same dimension and compute the similarity by finding the optimal match between these semantic blocks. Extensive experiments on large-scale MS-COCO and Flickr30K datasets verify the superiority of our proposed AVSE compared with recent state-of-the-art methods. Compared to the recent NAAF method, our AVSE inference is 1000 times faster on the 1K test set and more accurately on the widely used benchmarks.

1 INTRODUCTION

Understanding the correspondence between the visual world and human language is one of the fundamental capabilities of artificial intelligence (Karpathy & Fei-Fei, 2015). It motivates much research on vision-and-language tasks. As a foundation task in the vision-and-language domain, image-text matching devotes to bridging the semantic gap between these two different modalities, which aims to search images for a given textual description or vice versa. The key challenge of image-text matching is to measure the semantic similarity between images and texts.

To accurately bridge the correspondences between images and texts, mainstream methods follow a common way to first encode images and texts into a shared embedding space, then measure the similarity between them, and optimize the model via a triplet loss that encourages the similarity of the matched image-text pair to be greater than that of unmatched pairs. For global-level matching methods, most works (Faghri et al., 2018; Li et al., 2019; Chen et al., 2021) calculate the visual semantic similarity via the inner product (cosine similarity), others (Vendrov et al., 2016; Gu et al., 2018) introduce ordered representations to measure antisymmetric visual-semantic hierarchy. For local-level matching methods, many works (Lee et al., 2018; Chen et al., 2020a; Liu et al., 2020) adopt cross-modal attention mechanism to compute the visual semantic similarity.

Although the above methods have slightly different similarity calculation modules, these methods ignore the difference between the two modal data when calculating the similarity. Actually, there are many inherent variations between vision and language data, such as the **information density** (He et al., 2022). In reality, vision is the real world that humans see, while language is a description of the part of interest in vision. This intuition is favorable to gain the precise correspondence between images and texts. As shown in Figure 1, an image can be described from multiple different views

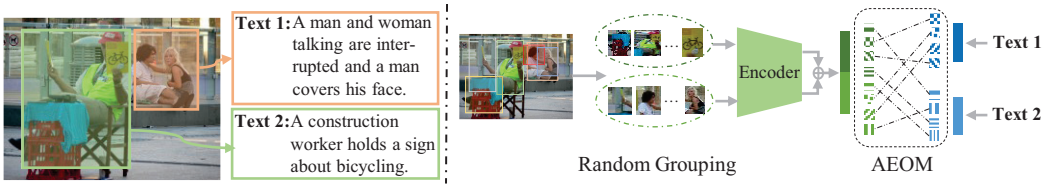

Figure 1: (Left) Information density is different in vision and language data, e.g., an image can be described from multiple different views using language. (Right) The conceptual diagram of our proposed AVSE, which can dynamically calculate the visual semantic similarity between images and matched texts with different views.

using human language, so an image text embedding of the same size could contain unequal amounts of information. However, mainstream methods directly compute visual semantic similarity using holistic visual and text embedding without considering the difference of information density in two modality data, which inevitably hinders the retrieval accuracy. The local-level matching methods implicitly handle such a problem by dynamically computing the similarity between images and text using a cross-modal attention mechanism, where the key idea is to compute all word-region similarities by attending to relevant fragments with respect to each query fragment from another modality. Such methods implicitly find the corresponding view for each sentence in the image at the word-region level, and thus obtains better retrieval results than the global-level matching methods. However, the prohibitive cost of computing the cross-modal attention (Lee et al., 2018) limits its practical usage in real-world applications.

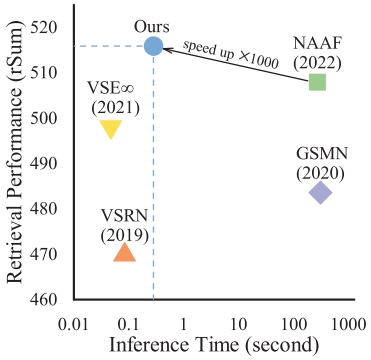

Figure 2: Accuracy-speed tradeoff of single model on Flickr30K test set. Our AVSE performs best.

Motivated by this, we aim to design a novel model to learn visual semantic similarity by explicitly considering the difference in information density between the two modalities and simultaneously achieve computational efficiency. We propose Asymmetric Visual Semantic Embedding (AVSE), which measures the visual semantic similarity by dynamically finding the most similar perspective for different texts in an image at a block-level embedding. By doing so, we use simple consine similarity to alleviate the computational cost caused by the cross-modal attention and achieve results even better than the state-of-the-art methods (See Figure 2). Specifically, in order for image features to contain information from multiple different views, AVSE exploits the large spatial redundancy of images, randomly groups regions, and uses a shared encoder to extract image embeddings of different image regions. Then these embeddings are concatenated as a larger image embedding than text embedding. Considering the different dimensions of text embedding and image embedding, we design a novel Asymmetric Embedding Optimal Matching module to effectively compute the visual semantic similarity by simulating the way humans describe images with language. We divide the visual embedding and text embedding into many semantic blocks with the same dimension and calculate the visual semantic similarity by finding the most similar image semantic block for each text semantic block via the Sinkhorn algorithm. We also propose a new loss function to regularize the image embeddings of different views to facilitate the model to find the optimal match between visual and text blocks.

To sum up, our main contributions of this paper are summarized as follows:

- We propose Asymmetric Visual Semantic Embedding (AVSE) for image-text retrieval to compute visual semantic similarity by explicitly considering the difference in information density between the two modalities and eschew the prohibitive computations.
- Asymmetric Embedding Optimal Matching (AEOM) module is proposed to efficiently calculate the visual semantic similarity and a dimension-wise regularization loss is developed to regularize the embeddings of different views for further improving the similarity score calculated by AEOM.
- Experimental results on Flickr30K and MS-COCO datasets demonstrate the superiority of our proposed AVSE compared with recent state-of-the-art baselines. Our model enjoys the benefits of both global-level methods and local-level methods, i.e. faster speed and higher accuracy.

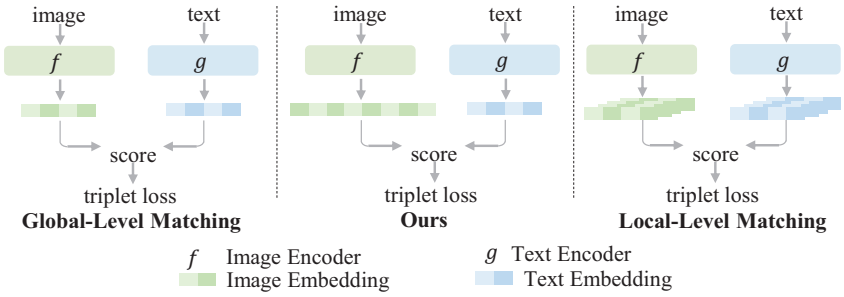

Figure 3: Illustration of the two different matching methods, and our proposed AVSE is a variant of global-level matching method.

2 Related Work

As a hot research topic to bridge the vision and language domains, the key issue of image-text retrieval is to measure the visual semantic similarity between a image and a text. According to the granularity of the matching similarity, we roughly divide mainstream works into two groups: global-level matching methods and local-level matching methods. As illustrate in Figure 3, the difference between them is whether holistic embedding is used to calculate the similarity, and our AVSE is also belongs to global-level matching.

Global-level matching methods. Global-level matching methods embed the holistic images and sentences into a shared embedding space, the visual semantic similarity can be calculated by simply inner product (cosine similarity) Kiros et al. (2014); Faghri et al. (2018) or other distance functions (Vendrov et al., 2016; Gu et al., 2018). Due to the lack of interaction between images and texts, such methods rely on high-quality image text embeddings to compute visual semantic similarity. Existing related works commonly utilize the graph convolutional network (Li et al., 2019; Wang et al., 2020; 2022; Cheng et al., 2022), self attention mechanism (Wu et al., 2019) or special pooling function (Chen et al., 2021; Li et al., 2022). In addition, many works focusing on model optimization. Wang et al. (2016) consider the external constraint loss that preserves the neighborhood structure in a single modality. Chen et al. (2020b) propose the adaptive offline quintuplet loss to improve the triplet loss effectively. Liu et al. (2022) utilize intra-modal contrastive loss to regularize the shared embedding space to gain the high-quality embeddings.

Local-level matching methods. To learn latent fine-grained correspondence between images and texts, local-level matching methods calculate visual semantic similarity by aligning the sub-fragments, i.e., regions in images and words in sentences. With the success of bottom-up attention (Anderson et al., 2018) in image captioning and VQA, Lee et al. (2018) proposed a stacked cross attention network to attend to image regions with respect to each word in sentences and versa. Recently, there are many follow-up works (Liu et al., 2019; Wang et al., 2019; Chen et al., 2020a; Liu et al., 2020; Diao et al., 2021; Zhang et al., 2022a;b) that use such stack attention mechanism to learn more region-word correspondences. Chen et al. (2020a) introduce an iterative matching scheme to capture correspondences between images and texts with multiple steps of alignments. Liu et al. (2020) construct visual and textual graph, and learn fine-grained correspondence by node-level matching and structure-level matching. Diao et al. (2021) introduce a vector similarity function to compute a similarity representation, constructing a similarity graph to reason the similarity and adopting attention filtration to eliminate the less-meaning alignments. Zhang et al. (2022a) infer the confidence of matched region-word pairs from the global perspective to refine the imagetext relevance measurement. Zhang et al. (2022b) measure the accurate similarity/dissimilarity degrees via a two-branch matching mechanism to jointly infer the overall image-text similarity.

3 Asymmetric Visual Semantic Embedding

In this section, we formally present our Asymmetric Visual Semantic Embedding (AVSE) model. Specifically, given an image-text pair, the model aims to calculate the visual semantic similarity by considering the difference in information density in two modalities. To better exploit the information density difference between the two modality data, AVSE learns an asymmetric visual semantic

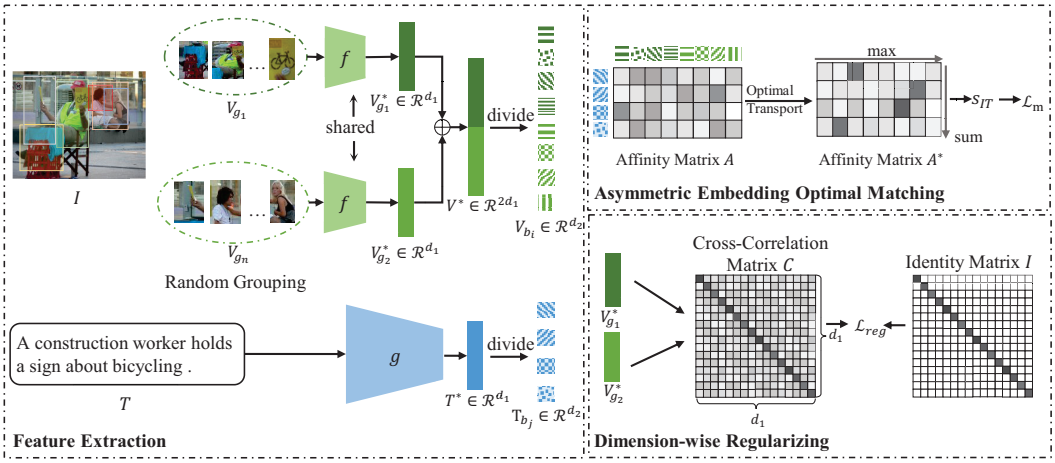

Figure 4: An overview of Asymmetric Visual Semantic Embedding (AVSE). *Asymmetric Feature Extraction* exploits the inherent differences to capture asymmetric embedding for images and texts. *Asymmetric Embedding Optimal Matching* takes full advantage of the inherent differences to find the optimal match between visual semantic blocks and textual semantic blocks to compute similarity by the affinity of block-level embeddings. *Dimension-wise Regularizing Loss* regularizes the embeddings of different image views to help AEOM to calculate more accuracy block-level affinity.

embedding and calculates the visual semantic similarity by a novel similarity learning module at a block-level. The overall framework of our proposed AVSE is depicted in Figure 4.

3.1 Asymmetric Feature Extraction

Textual Representation. Given a text T which consisted of m words, we aim to obtain a holistic textual embedding $\boldsymbol{t} \in \mathcal{R}^{d_1}$. First, each word is represented as a one-hot encoding, and embed into a pre-trained GloVe (Pennington et al., 2014) vector. Then, these vectors are fed into a bi-directional GRU to gain a set of word features. Finally, generalized pooling operator (Chen et al., 2021) is adopted as an aggregation function to encode the word vectors into the holistic text embedding $\boldsymbol{t}$.

Visual Representation With Multiple Views. Due to information density being different in images and text, visual embedding should contain more information than textual embedding.

Random Grouping. Given an image I, we aim to extract a holistic visual embedding containing different views. First, we follow (Anderson et al., 2018) to extract K salient regions with the Faster R-CNN (Ren et al., 2016) pre-trained on Visual Genomes (Krishna et al., 2017). We divide these K regions into two groups by random sampling to simulate two different views of the image I. Effect of different number of views is describe in Tab.3.

Shared-Weight Image Encoder. Then, we feed each region in all groups into a fully connected layer to transform into a d_1-dimensional vector, and a generalized pooling operator is used to aggregate the region features in each group into a holistic visual embedding $\boldsymbol{v}_{g_n} \in \mathcal{R}^{d_1}$. Finally, we concatenate all group embeddings as the visual embedding $\boldsymbol{v} \in \mathcal{R}^{2*d_1}$.

3.2 Asymmetric Embedding Optimal Matching

As the different dimensions of image embedding $\boldsymbol{v}$ and text embedding $\boldsymbol{t}$, the visual semantic similarity cannot be calculated using the conventional method. Motivated by the way humans describe images with language, we design a Asymmetric Embedding Optimal Matching (AEOM) module to calculate the visual semantic similarity efficiently. In contrast to the global-level matching approaches (Faghri et al., 2018; Chen et al., 2021) that use holistic embedding to compute similarity, we present a more fine-grained way to compute block-wise similarity, which aims to find the most similar visual block for each textual block. By doing so, we simulate the process of finding the most similar view in an image for a given sentence.

Concretely, we first split the visual embedding $\boldsymbol{v}$ and text embedding $\boldsymbol{t}$ into a set of d_2-dimensional blocks, represented as $\boldsymbol{v}_b = [\boldsymbol{v}_{b_1}, \boldsymbol{v}_{b_2}, ..., \boldsymbol{v}_{b_p}]$ and $\boldsymbol{t}_b = [\boldsymbol{t}_{b_1}, \boldsymbol{t}_{b_2}, ..., \boldsymbol{t}_{b_q}]$, respectively, where $p =$

nd_1/d_2 and $q = d_1/d_2$. Then we compute the affinity matrix $\boldsymbol{A}$ of visual blocks $\boldsymbol{v}_b$ and text blocks $\boldsymbol{t}_b$ as follows:

$$A_{i,j} = \frac{\boldsymbol{v}_{b_i}^\top \boldsymbol{t}_{b_j}}{\|\boldsymbol{v}_{b_i}\| \cdot \|\boldsymbol{t}_{b_j}\|}, i \in [1,p], j \in [1,q] \tag{1}$$

where $A_{i,j}$ indicate the affinity score of visual block $\boldsymbol{v}_{b_i}$ and text block $\boldsymbol{t}_{b_j}$. Since text blocks might not be associated with any visual blocks and vice versa, we augment the affinity matrix $\boldsymbol{A}$ with a dustbin so that unmatched blocks are explicitly assigned to it. Specifically, we augment the affinity matrix $\boldsymbol{A}$ by adding a new row and column at the end of the matrix to produce a new $\bar{\boldsymbol{A}} \in \mathcal{R}^{(p+1)\times(q+1)}$, the new row and column are initialed the new row and column are initialed by filling a single learnable parameter z. While each visual block or text block will be assigned to a single block in another modal or to the dustbin, the maximum capacity that can be assigned to each bin is the number of visual blocks or text blocks. We denote $\boldsymbol{a} = [\mathbf{1}_p^\top, q]^\top$ and $\boldsymbol{b} = [\mathbf{1}_p^\top, p]^\top$ as the number of expected matched blocks for each block and dustbin in $\boldsymbol{v}_{b_1}$ and $\boldsymbol{t}_{b_1}$. Further, note that the augmented $\bar{\boldsymbol{A}}$ is subject to the following constraints: $\bar{\boldsymbol{A}}\mathbf{1}_{q+1} = \boldsymbol{a}$ and $\bar{\boldsymbol{A}}^\top \mathbf{1}_{p+1} = \boldsymbol{b}$. where $\mathbf{1}$ is a vector of ones. The constraints of the above optimization problem are similar to solving for Optimal Transport (Peyré et al., 2019). We leverage the Sinkhorn Normalization (Adams & Zemel, 2011; Cuturi, 2013; Sinkhorn & Knopp, 1967) to automatically satisfy above constraints during training and testing. Sinkhorn algorithm is a differentiable version of the Hungarian algorithm Munkres (1957), which is used to solve an efficient yet simple approximate solution for generalized linear assignment by repeatedly normalizing rows and columns. After c iterations of Sinkhorn algorithm, $\bar{\boldsymbol{A}}$ is approximately converted to a doubly-stochastic matrix, and we obtain the final affinity matrix by $\boldsymbol{A}^* = \bar{\boldsymbol{A}}_{1:p,1:q}$.

At this point, we more precisely compute the block-wise correspondence between visual embedding and text embedding. Inspired by (Karpathy & Fei-Fei, 2015; Lee et al., 2018), we employ the max-sum pooling to caculate the visual semantic similarity between image I and text T by computing the max over the columns and then summing:

$$S(I,T) = \sum_{j=1}^{q} \max_{i \in [1,p]} \boldsymbol{A}^*_{i,j} \tag{2}$$

3.3 Dimension-wise Regularizing for Different Views

Considering that to find the best matches between text blocks $\boldsymbol{t}_b$ and visual blocks containing different views $\boldsymbol{v}_b$, we argue that image embeddings of diferent views should satisfy the same feature distribution, i.e., each channel in the embedding $\boldsymbol{v}_{g_1}$ and $\boldsymbol{v}_{g_2}$ represents the same semantics. Specifically, we first calculate the cross-correlation matrix $\boldsymbol{C}$ between $\boldsymbol{v}_{g_1}$ and $\boldsymbol{v}_{g_2}$ along the batch dimension as follows:

$$\boldsymbol{C}_{ij} = \frac{\sum_b \boldsymbol{v}^A_{b,i} \boldsymbol{v}^B_{b,i}}{\sqrt{\sum_b (\boldsymbol{v}^A_{b,i})^2} \sqrt{\sum_b (\boldsymbol{v}^B_{b,i})^2}} \tag{3}$$

where b indicates the index of batch sample and i,j indicate the vector dimension of embedding, $\boldsymbol{C}$ is a matrix with size of $b \times b$. Then, we regularize the embeddings of different views by trying to equate the diagonal elements of the cross-correlation matrix to 1, and equate the off-diagonal elements of the cross-correlation matrix to 0. We define the loss as:

$$\mathcal{L}_{reg} = \sum_i (1 - \boldsymbol{C}_{ii})^2 + \lambda_1 \sum_i \sum_{j \neq i} \boldsymbol{C}_{ij}^2 \tag{4}$$

where λ_1 is a positive constant to balance the weight of two terms of the loss function.

3.4 Objective Function

To align images and texts, we adopt hinge-based triplet loss (Kiros et al., 2014) using the hardest negative samples (Faghri et al., 2018; Lee et al., 2018; Chen et al., 2021). The loss function is defined as follows:

$$\mathcal{L}_m = \sum_{(I,T) \in \mathcal{D}} \Big[\alpha - S(I,T) + S(I,\hat{T})\Big]^+ + \Big[\alpha - S(I,T) + S(\hat{I},T)\Big] \tag{5}$$

Table 1: Comparisons of experimental results on MS-COCO and Flickr30k datasets. Methods are divided into two categories: Global-level Matching Methods and Local-level Matching Methods. '*' indicates ensemble methods. The best performances are in **bold**.

Method	MS-COCO 5-fold 1K Test							Flickr30k 1K Test						
	Text Retrieval			Image Retrieval			rSum	Text Retrieval			Image Retrieval			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	
Global-level Matching Methods														
LIWE (Wehrmann et al., 2019)	73.2	95.5	98.2	57.9	88.3	94.5	507.6	69.6	90.3	95.6	51.2	80.4	87.2	474.3
VSRN* (Li et al., 2019)	76.2	95.6	98.5	61.7	89.1	95.0	516.1	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CVSE (Wang et al., 2020)	74.8	95.1	98.3	59.9	89.4	95.2	512.7	73.5	92.1	95.8	52.9	80.4	87.8	482.5
VSE∞ (Chen et al., 2021)	78.5	96.0	98.7	61.7	90.3	95.6	520.8	76.5	94.2	97.7	56.4	83.4	89.9	498.1
MV-VSE (Li et al., 2022)	78.7	95.7	98.7	62.7	90.4	95.7	521.9	79.0	94.9	97.7	59.1	84.6	90.6	505.8
ConVSE* (Liu et al., 2022)	78.6	96.3	**98.8**	64.3	**92.4**	**96.9**	526.2	81.6	95.5	97.8	61.3	86.7	92.0	514.9
Local-level Matching Methods														
SCAN* (Lee et al., 2018)	72.7	94.8	98.4	58.8	88.4	94.8	507.9	67.4	90.3	95.2	48.6	77.7	85.2	464.4
IMRAM (Chen et al., 2020a)	76.7	95.6	98.5	61.7	89.1	95.0	516.6	74.1	93.0	96.6	53.9	79.4	87.2	484.2
SHAN (Ji et al., 2021)	76.8	96.3	98.7	62.6	89.6	95.8	519.8	74.6	93.5	96.9	55.3	81.3	88.4	490.0
GSMN* (Liu et al., 2020)	78.4	96.4	98.6	63.3	90.1	95.7	522.5	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SGRAF* (Diao et al., 2021)	79.6	96.2	98.5	63.2	90.7	96.1	524.3	77.8	94.1	97.4	58.5	83.0	88.8	499.6
NAAF* (Zhang et al., 2022b)	80.5	96.5	**98.8**	64.1	90.7	96.5	527.2	81.9	96.1	98.3	61.0	85.3	90.6	513.2
Ours: AVSE	78.5	96.3	98.7	63.3	91.0	96.1	524.0	81.5	96.2	98.0	60.9	86.8	92.0	515.5
Ours: AVSE*	**81.4**	**96.9**	**98.8**	**65.4**	91.5	96.4	**530.4**	**82.4**	**96.3**	**98.3**	**62.6**	**87.9**	**92.8**	**520.4**

where α serves as a margin parameter and $[x]^+ \equiv \max(x, 0)$. In dataset $\mathcal{D}$, the visual semantic similarity in a positive pair $S(I, T)$ should be higher than that in the hardest negative pairs $S(\hat{I}, T)$ and $S(I, \hat{T})$ by a margin α.

In summary, the final loss function of our model is defined as follows to perform joint optimization of the two objectives.

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_{reg} \tag{6}$$

4 Experiments

4.1 Dataset and Settings

Datasets. Following previous works (Faghri et al., 2018), we use two widely used benchmark datasets MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) for our experiment. MS-COCO is a dataset that contains 123287 images and five text captions are annotated for each image. Following Faghri et al. (2018), all data are split into training set, validation set and testing set which contains 113287, 5000, 5000 images respectively. Flickr30K is composed of 31783 images and each image has 5 corresponding descriptions. We followed the split in Faghri et al. (2018), using 29000 images for training, 1000 images for validation, and 1000 images for testing.

Evaluation Metrics. The commonly used evaluation metrics for image-text matching are Recall@K (K=1,5,10), denoted as R@1, R@5, and R@10, which depict the percentage of ground truth being retrieved at top 1, 5, 10 results, respectively. Specifically, we also follow previous works (Chen et al., 2021; Li et al., 2022) to use sum of all the Recall values (rSum) to evaluate the performance.

4.2 Implementation Details

All of our experiments are trained on a single Tesla P100 GPU. We train the proposed model for 25 epochs with set the mini-batch size as 128 using AdamW (Loshchilov & Hutter, 2019) optimizer. The learning rate is set as 0.0005 for the first epochs and then decreased to 0.00005 for the rest 10 epochs. We train the model on the training set, and validate it at each epoch on validation set, finally we select the best model to test on the test set. The best model is model that has the highest rSum on the validation set. For feature extractions, we set the dimension of the shared embedding space d_1 as 1024. For text encoder, we adopt the pre-trained Glove (Pennington et al., 2014) word embedding instead of the randomly initial word embedding, project it to the shared embedding space through a 1-layer Bi-GRU and finally aggregates the word embeddings by a generalized pooling operator (Chen et al., 2021) to get a holistic text embedding $\boldsymbol{t}$. For the image encoder, we follow Lee et al. (2018) implement the bottom-up attention with a Faster R-CNN (Girshick et al., 2014) model using

ResNet-101 as the backbone, which is pre-trained on the Visual Genomes (Krishna et al., 2017) dataset by Anderson *et al.* (Anderson et al., 2018). We choose the top $K = 36$ ROIs with the highest class detection confidence scores as the region features, and the dimension of each region is 2048 (Lee et al., 2018)[1]. Different from exist works, we sample these regions into n=2 groups to capture 2 different views of the image, and each groups contains 75% of the selected K fragments for training and 90% for testing. The two groups of region features is then transformed into a 1024-dimensional visual representation by a shared-weight fully-connected layer and aggregatee by a generalized pooling operator (Chen et al., 2021) to get a holistic visual embedding $\boldsymbol{v}$. For asymmetric embedding optimal matching module, we set the dimension of blocks d_2 as 512, and we set the iteration of Sinkhorn normalization as 20. For objective function, we set λ_1 in Eq. 4 as 0.0051 to balane two terms of $\mathcal{L}_{reg}$ and α in Eq. 5 is set to 0.2 as margin parameter.

4.3 Comparisons with Serveral Recent State-of-the-arts

We compare AVSE with the most recent state-of-the-art approaches on two widely used datasets, *i.e.*, Flickr30K and MS-COCO. It is noted that many state-of-the-arts methods adopt the ensemble strategy (Lee et al., 2018; Li et al., 2019; Diao et al., 2021). For a fair comparison, we also provide the ensemble version to compare with other methods. When implementing the ensemble scheme, we average the similarity scores of two already trained models for the final ranking process.

Table 2: Retrieval results on MS-COCO 5K test set. '*' indicates ensemble methods.

Method	Text Retrieval			Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
Global-level Matching Methods							
VSRN*	53.0	81.1	89.4	40.5	70.6	81.1	415.7
VSE∞	56.6	83.6	91.4	39.3	69.9	81.1	421.9
MV-VSE	56.7	84.1	91.4	40.3	70.6	81.6	424.6
Local-level Matching Methods							
SCAN*	50.4	82.2	90.0	38.6	69.3	80.4	410.9
IMRAM	53.7	83.2	91.0	39.7	69.1	79.8	415.5
SGRAF*	57.8	-	91.6	41.9	-	81.3	272.6
NAAF*	58.9	85.2	92.0	42.5	70.9	81.4	430.9
Ours	59.1	86.1	93.2	42.3	72.5	82.6	435.9
Ours*	**61.8**	**86.7**	**93.3**	**43.5**	**73.2**	**83.4**	**442.0**

Results on Flickr30K. The quantitative results of our AVSE approach on Flickr30K are shown in Tab.1. Our AVSE outperforms state-of-the-arts significantly on all evaluation metrics. Specifically, compared with the current best method NAAF (Zhang et al., 2022b), AVSE obtain 7.2% improvement on rSum, where the R@1 gains 0.5% and 1.6% improvement at text retrieval and image retrieval, respectively. Moreover, compared with the typical VSE∞ (Chen et al., 2021) which our model builds on its basis, our proposed AVSE gains 5.9% and 6.2% on R@1 at two directions, respectively, and largely improves rSum by 14.6%.

Results on MS-COCO. The experimental results of the larger and complicated MS-COCO 5-fold 1K test set and MS-COCO 5K test set are shown in Tab.1 and Tab.2, respectively. For 5-fold 1K test set, it is obvious that our AVSE outperforms state-of-the-arts in terms of most evaluation metrics. Compared with MV-VSE and SGRAF, our AVSE gains relative improvements of 8.5% and 6.1% on rSum, respectively, and we can achieve competitive results with the current best method NAAF, getting 3.2% improvements on rSum. For 5K test set, it is obvious that our method achieves the new state-of-the-art, with 61.8% R@1 and on 46.2% R@1 on text retrieval and image retrieval, respectively, which is consistent with results on Flickr dataset. Compared with the current best model NAAF, AVSE gains relative improvements of 11.1% on rSum, which again demonstrates the effectiveness of learning the visual-semantic similarities.

Inference efficiency analysis. In addition to the accuracy of caption or image retrieval, we also argue that the efficiency at the inference stage is important when evaluating the model's performance. This is crucial especially when the model is used in search engines for a large-scale database towards image or text queries. However, recent local-level matching methods (Lee et al., 2018; Chen et al., 2020a; Liu et al., 2020; Zhang et al., 2022b) usually rely on complex cross-modal attention mechanism, which significantly harm the inference speed of the model. In contrast, global-level matching methods (Chen et al., 2021; Li et al., 2019; Liu et al., 2022; Li et al., 2022) have extremely fast inference speed, but due to inability to learn the multi-view information to dynamically calculate the visual semantic similar-

[1]The region features can be downloaded from https://github.com/kuanghuei/SCAN

ity between image and matched texts with different views, therefore, the retrieval accuracy of the global-level matching methods are inferior to that of the local-level matching methods. Our AVSE also belongs to the global-level matching methods, but by learning multi-view image features and dynamically computing the visual semantic similarity, we well balance the retrieval speed and retrieval accuracy to obtain the best performance at the expense of limited retrieval speed. As shown in Figure 5, we compare inference time for image-text retrieval on single GPU with recent state-of-the-arts. It is obvious that global-level matching methods (VSRN and VSE∞) are much faster than local-level matching methods (NAAF and GSMN). When there are only 10 candidate images, the global level matching method is more than 100 times faster than the local level matching method, and when the number of candidate images increases to 1000, the time cost difference increases to 10,000 times. However, our AVSE retrieval time increases only 0.3 seconds than traditional global-level matching methods when there are 1000 candidate images, which is 1000$\times$ faster than NAAF and GSMN.

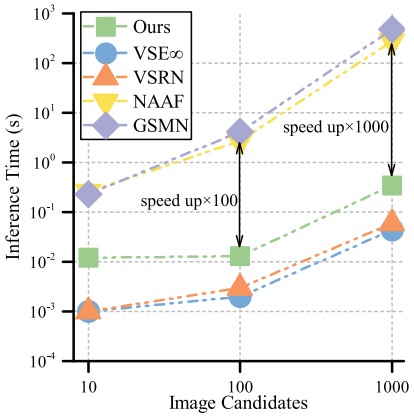

Figure 5: Inference time for image-text retrieval on GPU (lower the better).

4.4 Ablation Study

To verify the effectiveness of each component of our AVSE, we conduct extensive ablation studies on the larger and complicated Flickr30K dataset.

The impact of AEOM. The AEOM is a critial module in our AVSE, to validate the superiority of AEOM, we first compare it with the conventional similarity functions, i.e., Consine similarity, we also extract multi-view image features and feed it into a mean pooling to have multi-view information. As shown in Tab.3, we list the results of 1,2,4 views of using Consine similarity. It prove that multi-view information can directly improve the retrieval accuracy. And it is clearly to see that our AEOM achieves the best performance on all metrics when using 2 view image features. Although increasing views does not improve retrieval performance, our AEOM outperforms conventional methods by a large margin of 2.2% on R@1.

Table 3: Ablation Studies about different parameters on Flickr30K Dataset.

Method	**TR**		**IR**	
	R@1	R@10	R@1	R@10
Asymmetric Embedding Optimal Matching				
w/o AEOM (1 views)	77.2	97.4	57.6	90.6
w/o AEOM (2 views)	78.9	97.8	59.5	91.3
w/o AEOM (4 views)	79.3	97.7	58.7	91.1
w/ AEOM (2 views)	**81.5**	**98.0**	**60.9**	**92.0**
w/ AEOM (4 views)	**81.5**	97.9	60.2	91.3
Dimension of Block d_2				
initial $d_2 = 64$	75.9	97.1	55.8	90.1
initial $d_2 = 128$	77.5	97.5	56.1	89.9
initial $d_2 = 256$	80.1	**98.1**	57.4	91.0
initial $d_2 = 512$	**81.5**	98.0	**60.9**	**92.0**
Sinkhorn Normalization				
w/o Sinkhorn	79.1	97.8	59.6	91.6
w/ Sinkhorn $c = 10$	**81.9**	**98.1**	60.1	91.7
w/ Sinkhorn $c = 20$	81.5	98.0	**60.9**	**92.0**
w/ Sinkhorn $c = 30$	81.0	98.0	60.2	91.4
Dimension-wise Regularizing Loss				
w/o $\mathcal{L}_{reg}$	80.1	97.7	59.5	91.0
w/ $\mathcal{L}_{reg}$	**81.5**	**98.0**	**60.9**	**92.0**

The impact of different block dimension. The dimension of the visual and text semantic blocks d_2 is a sensitive parameter in AEOM, which determines the ability to learn the visual-semantic similarities. Hence, an appropriate parameter is important in our proposed model. To validate the impact of the blocks' dimension d_2, we conduct extensive experiments on Flickr30K dataset. Here, we investigate the matching performance by setting d_2 as 64, 128, 256 and 512, the results are reported in Tab.3. It is obvious that when blocks' dimension d_2 setting with 512, the model yields the best performance on both image retrieval and text retrieval.

The impact of Sinkhorn normalization. Sinkhorn Normalization is the most important component in AEOM, to verify the effectiveness of Sinkhorn Normalization, we perform ablation studies on Sinkhorn Normalization. To verify the effectiveness of Sinkhorn Normalization, we perform ablation studies on Sinkhorn Normalization. As shown in Tab.3, the results show that without using the Sinkhorn Normalization the retrieval performance decreases in all metrics. Then we validate

Figure 6: Result visualization of our proposed AVSE. We show the original image and two attention maps from different views in (a) and (b) to show what AVSE learned from two views. In (c),(d),(e),(f), we list qualitative results of image-text retrieval. For each query, we show the top-3 ranked results.

the effect of iteration c of Sinkhorn Normalization. The value of c controls the quality of the normalization of the affinity matrix, which is with significant importance. The ablation studies show that a suitable value of $c = 20$ is critical in maintaining good performance, though our method with different c is all with superior results.

The impact of dimension-wise regularizing loss. We validate the positive effect of this dimension-wise regularizing loss on our proposed method. We can clearly find that by using $\mathcal{L}_{reg}$, the retrieval performance is improved in all metrics. Especially for R@1, our method obtain relative 1.4% relative improvement both for image retrieval and text retrieval.

4.5 Visualization

We visualize the retrieval process in Figure 6. From the attention maps of two views, it is clear from the figures that our model can capture different views feature of images, which is crucial for dynamically calculate the visual semantic similarity. For example, in Figure 6 (a), the first view mainly focus on the child while another view mainly focus on the chairs. So retrieval performance can benefit in such multi-view information. And we also show some retrieval results in Flickr30K. In Figure 6 (c) and (d), we show some results of image-to-text retrieval, and in Figure 6 (e) and (f), we visualize the text-to-image retrieval. These show our approach always retrieves the ground truth with a high rank. In addition, our approach is able to learn detail for image-text correspondence. For example, in Figure 6 (f) , our network can accurately find women with baby.

5 Conclusion

In this paper, we propose a novel Asymmetric Visual Semantic Embedding (AVSE) for efficient image-text matching. The key insight is that the difference information density between vision and language is crucial for image-text retrieval. To better exploit the information density difference between the two modality data, let the model learn an asymmetric visual semantic embedding and make full use of the information density difference by a novel similarity learning module. Moreover, we propose a new loss function to regularize the image embeddings of different views to better help the model to find the optimal match between visual and text blocks. Comprehensive experiments on two widely-used benchmark datasets validate the effectiveness of the proposed method, leading to state-of-the-art performance.

REFERENCES

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *stat*, 1050:14, 2011.

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*, pp. 6077–6086, 2018.

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*, pp. 12655–12663, 2020a.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pp. 15789–15798, 2021.

Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, pp. 549–565. Springer, 2020b.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS*, 26, 2013.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, volume 35, pp. 1218–1226, 2021.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587, 2014.

Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pp. 7181–7189, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. In *IJCAI*, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

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

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.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pp. 201–216, 2018.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pp. 4654–4662, 2019.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. In *IJCAI*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

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 *ACM MM*, pp. 3–11, 2019.

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

Yang Liu, Hong Liu, Huaqiu Wang, and Mengyuan Liu. Regularizing visual semantic embedding with contrastive learning for image-text matching. *IEEE Signal Processing Letters*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

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

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2016.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, pp. 18–34, 2020.

Haoran Wang, Dongliang He, Wenhao Wu, Boyang Xia, Min Yang, Fu Li, Yunlong Yu, Zhong Ji, Errui Ding, and Jingdong Wang. Coder: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *ECCV*, 2022.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pp. 5005–5013, 2016.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *IJCAI*, pp. 3792–3798, 2019.

Jonatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. Language-agnostic visual-semantic embeddings. In *ICCV*, pp. 5804–5813, 2019.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM MM*, pp. 2088–2096, 2019.

Huatian Zhang, Zhendong Mao, Kun Zhang, and Yongdong Zhang. Show your faith: Cross-modal confidence-aware network for image-text matching. In *AAAI*, 2022a.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *CVPR*, pp. 15661–15670, 2022b.