High-Order Semantic Alignment for Unsupervised Fine-Grained Image-Text Retrieval

Rui Gao, Miaomiao Cheng, Xu Han, Wei Song
Information Engineering College, Capital Normal University, Beijing, China
{rgao, miaomiao, hanxu, wsong}@cnu.edu.cn

Abstract

Cross-modal retrieval is an important yet challenging task due to the semantic discrepancy between visual content and language. To measure the correlation between images and text, most existing research mainly focuses on learning global or local correspondence, failing to explore fine-grained local-global alignment. To infer more accurate similarity scores, we introduce a novel High-Order Semantic Alignment (**HOSA**) model that can provide complementary and comprehensive semantic clues. Specifically, to jointly learn global and local alignment and emphasize local-global interaction, we employ the tensor-product (t-product) operation (Misha et al., 2011) to reconstruct one modal's representation based on another modal's information in a common semantic space. Such a cross-modal reconstruction strategy would significantly enhance inter-modal correlation learning in a fine-grained manner. Extensive experiments on two benchmark datasets validate that our model significantly outperforms several state-of-the-art baselines, especially in retrieving the most relevant results. The code used for the experiments is publicly available on GitHub at https://github.com/cnunlp/HOSA.

Keywords: fine-grained cross-modal retrieval, high-order semantic alignment, image-text matching, t-product

1. INTRODUCTION

Multimedia data such as texts, images, audio, and video, have been ubiquitous in our daily life. The rich content of multimedia data has sparked people's need to obtain more comprehensive and enriched information on the same event or topic from different modalities. In this study, we focus on the task of image-text cross-modal retrieval, which aims to retrieve images given text queries or find matching textual descriptions given image queries. It is a challenging task to measure the relevance between images and texts due to the semantic gap across different modalities.

To learn the correlations between modalities, early works (Long et al., 2016; Zhang et al., 2018; Shen et al., 2017; Radford et al., 2021) primarily focus on identifying a common embedding space for the overall image and text, to find the semantic correspondence between a whole picture and a complete sentence. However, such coarse-grained global alignment learning methods often induce background noise while failing to effectively capture the sophisticated interactions between modalities, which impedes the correct image-text alignment. For example, when users submit an image of "Slaty-backed Gull" as the query, these methods treat it as "Bird" and may return textual descriptions including other bird species like "Herring Gull".

To acquire more accurate similarities, recent works (Chen et al., 2021; Li et al., 2019) focus on detecting fine-grained region-word correspondences. Although improving the performance of image-text retrieval to some extent, they are still one-sided due to the neglect of global contexts. Based on this observation, exploring both global correspondence and local correspondence to measure cross-modal similarity is increasingly favored by researchers. These approaches (Ji et al., 2020a; Messina et al., 2021a,b; Wei and Zhou, 2020; Liu et al., 2022; Wang et al., 2023) first utilize cross-attention to obtain the local alignment of regions and words, and then compute scalar-based cosine distances between region-word pairs to reflect the similarities of overall image-sentence.

Nevertheless, these methods ignore that a region or a word may contain different semantics in different global contexts. For example, the appearance of the region "life preserver" shown on the left part of Figure1 is similar to that of "tire", but they are mismatched semantically in global contextual information, i.e., "The bicycle has a clock as a tire" and "The blue boat themed bathroom with a life preserver on the wall". Besides, the interaction between the "blue boat", "life preserver", and "wall" areas corresponding to the "bathroom" can also provide richer semantic information.

Motivated by these, we propose a novel High-Order Semantic Alignment (**HOSA**) model for unsupervised fine-grained image-text retrieval, to explore the high-order correlations across modalities from multiple perspectives, i.e., local correspondence, local-global interaction, and global correspondence. Its framework is shown in Figure1. Specifically, region-level features of images are extracted by the bottom-up attention model based on Faster-RCNN (Ren et al., 2015), and word-level features of the text are extracted by the pre-trained

BERT (Devlin et al., 2019) model, respectively. To determine the high-order correlations across fragments among different modalities, we employ t-product (Misha et al., 2011) based on the circular convolution operation to reconstruct one modal with another modal information. In this way, the reconstruction coefficients can characterize the local and global correspondence between regions and words/sentences under multiple instances between different modalities, modeling the high-order semantic alignment for image and text. For the sake of more comprehensive semantic alignment, the max-over-regions sum-over-words ($M_r S_w$) pooling strategy is adopted for aggregating the cosine similarities between reconstructed regions and words.

To our knowledge, we are the first to adopt a modal reconstruction strategy using the t-product (Misha et al., 2011) operation to explore high-order semantic consistency in image-text alignment. We perform comprehensive experiments, demonstrating the effectiveness of our method for fine-grained image-text retrieval. Our HOSA model shows remarkable superiority compared with the most advanced approaches in two evaluation metrics $R@1$ and $R@5$. We also conduct an in-depth analysis to investigate the ways of modal reconstruction.

2. RELATED WORK

In the literature, many approaches have been proposed for image-text retrieval, which can be roughly divided into three groups.

Global alignment methods focus on learning cross-modal similarity directly for the entire images and sentences, by projecting them into a common latent space (Zhang et al., 2018; Long et al., 2016) or exploiting visual-semantic embeddings (Faghri et al., 2018; Chen et al., 2021; Zheng et al., 2020; Radford et al., 2021). As a result, they often fall short of deeply mining the intricate relationships between visual objects and textual terms. Therefore, when confronted with natural scenes involving multiple objects and more complex descriptions, their performance may not meet expectations.

Local alignment methods seek to explore local correlations between image regions and sentence words for more accurate cross-modal alignment. Karpathy and Fei-Fei (2015) pioneered aligning local image regions detected by multi-modal RNN with words in the sentence. Subsequently, Lee et al. (2018) utilized stacked cross attention to align salient regions and keywords, underscoring the effectiveness of region-word alignment and inspiring subsequent works. FPAN (Wang et al., 2019) was proposed to emphasize the importance of different positions within each region. CAMP (Wang et al., 2020) innovatively introduced an adaptive regulation of the information flow in cross-modal message transmission. Although these methods improve the cross-modal retrieval performance, they fail to thoroughly mine the intra-modal correlations in the context of these fine-grained fragments. Unlike this, we model adaptively multi-level correspondences and comprehensively explore fine-grained visual-semantic similarity for more complete alignment.

Multi-order alignment methods aim to exploit both global and local correspondence to achieve more accurate cross-modal matching. Ji et al. (2020a) employed the attention mechanism to locate semantically meaningful portions for local alignment, and used memory networks to capture long-term contextual knowledge for global alignment. Wei and Zhou (2020) combined adversarial networks for local alignment and utilized attention mechanisms for global alignment. Qu et al. (2020) designed a gating self-attention mechanism for context modeling and a multi-view summarization module for asymmetry matching, to obtain local and global correspondence. Messina et al. (2021b,a) achieved multi-order reasoning within the same modality for regions and words by employing the Transformer Encoder Reasoning Network. Wang et al. (2023) leveraged infrequent textual content to mitigate the long-tail effect in image-text matching for local alignment, and then utilized the attention mechanism to achieve global alignment. They align image regions and text words by locally associating visual semantics and mechanically aggregate the semantic similarity between matched region-word pairs to measure the overall image-text correlation.

Despite the superior performance for cross-modal retrieval, the above approaches cannot implement local-global interaction, which also plays an important role in semantic alignment. Generally, each region-word pair may be inconsistent from the global perspective image-text. The main reason is that individual regions or words may have different semantics from global contexts. Thus, in this paper, we focus on a novel high-order semantic alignment to comprehensively explore the correlations among fragments and the entire context.

3. Methodology

In order to comprehensively capture the high-order correlations across modalities, we propose a novel High-Order Semantic Alignment (**HOSA**) model for fine-grained image-text retrieval. Figure 1 illustrates the proposed framework, which contains three main modules: *feature representation* with embedding modal-specific segments, *high-order semantic alignment* by exploring global and local correspondence and local-global interaction, and *cross-modal relevance measuring* by aggregating local similarities. Next, we will first define general

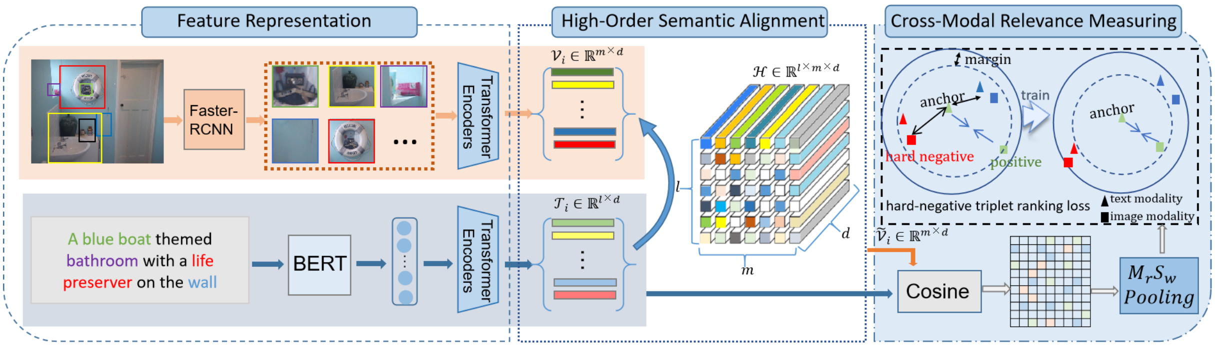

Feature Representation
High-Order Semantic Alignment
Cross-Modal Relevance Measuring
Faster-RCNN
Transformer Encoders
$\mathcal{V}_i \in \mathbb{R}^{m \times d}$
$\mathcal{H} \in \mathbb{R}^{l \times m \times d}$
l
d
m
margin
anchor
train
anchor
hard negative
positive
hard-negative triplet ranking loss
text modality
image modality
A blue boat themed bathroom with a life preserver on the wall
BERT
Transformer Encoders
$\mathcal{T}_i \in \mathbb{R}^{l \times d}$
$\widetilde{\mathcal{V}}_i \in \mathbb{R}^{m \times d}$
Cosine
$M_r S_w$ Pooling

Figure 1: The framework of the **HOSA** model. An image with m regions and a text with l words are encoded in a $d-$dimensional common space via a stack of transformer layers, and $\mathcal{H} \in \mathbb{R}^{l \times m \times d}$ is the reconstruction coefficient tensor, characterizing the high-order semantic alignment across modalities.

notations and then describe each module in detail.

3.1. Notations

Generally, tensors and matrices are denoted by bold calligraphy letters, e.g., $\mathcal{A}$, and bold upper case letters, e.g., $\mathbf{A}$, respectively. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{a}$, and scalars are denoted by lowercase letters, e.g., a. For a three-order tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, using Matlab notations, $\mathcal{A}_i = \mathcal{A}(i,:,:)$, $\mathcal{A}(:,i,:)$, $\mathcal{A}^{(i)} = \mathcal{A}(:,:,i)$ corresponds respectively to the i-th horizontal, lateral and frontal slice. A tube fiber of tensor $\mathcal{A}$ is defined by holding the first two indices fixed and varying the third, e.g., $\mathcal{A}(i,j,:)$ is the ij-th tube of $\mathcal{A}$. The Frobenius(F) norm of a matrix $\mathbf{A}$ and a tensor $\mathcal{A}$ is defined as $\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2$ and $\|\mathcal{A}\|_F^2 = \sum_{i,j,k} \mathcal{A}(i,j,k)^2$, respectively. $\mathcal{A}^T \in \mathbb{R}^{n_2 \times n_1 \times n_3}$ is the transpose of tensor $\mathcal{A}$.

$\overline{\mathcal{A}}$ is a tensor obtained by taking the Fourier transform along the third mode of $\mathcal{A}$, i.e, $\overline{\mathcal{A}}(i,j,:) = fft(\mathcal{A}(i,j,:))$. In Matlab notation, $\overline{\mathcal{A}} = fft(\mathcal{A}, [], 3)$, and one can also compute $\mathcal{A}$ from $\overline{\mathcal{A}}$ via $\mathcal{A} = ifft(\overline{\mathcal{A}}, [], 3)$.

3.2. Feature Representation

Image Representation: Considering multiple local features are more accurate to describe an image than a single global feature, we aim to represent each input image with a set of features $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_m\}$, $\mathbf{r}_i \in \mathbb{R}^{d_v}, i \in [1, m]$. Specifically, the representation $\mathbf{r}_i$ denotes the $i-$th salient region or object in an image. Following (Anderson et al., 2018), we utilize the pre-trained Faster-RCNN (Ren et al., 2015) model on Visual Genomes (Krishna et al., 2017) by bottom-up and top-down attention to select and extract features for salient regions.

Each representation $\mathbf{r}_i$ is defined as the mean-pooled convolutional feature for the $i-$th region or object. In our experiment, the Faster-RCNN feature is 2048 dimensional and $m = 36$; the top 36 regions with the most information related to the geometry are selected from the bounding boxes. Then, we utilize the Transformer Encoders followed in TERAN (Messina et al., 2021a) to obtain initial region representations with a common embedding space. Specifically, four transformer encoder layers are adopted with sequences or sets of entities as input, reasoning upon these entities without considering their intrinsic nature. A linear projection layer is added between transform encoder layers to transform the embeddings into d-dimensional vectors $\{\mathbf{v}_i \in \mathbb{R}^d\}_{i=1}^m$.

Text Representation: Following the recent trends in Natural Language Processing, we use a pre-trained BERT (Devlin et al., 2019) model for sentence texts to extract world-level textual representations. Similar to image representation, we also add a liner projection layer between the four transformer encoder layers to transform them into a d dimensional embedding space, denoted as $\{\mathbf{t}_j \in \mathbb{R}^d\}_{j=1}^l$. Specifically, $\mathbf{t}_j$ encodes the j-th word and l is the number of words.

3.3. High-Order Semantic Alignment

To facilitate the comprehensive semantic understanding and final cross-modal similarity calculation, we devise a High-Order Semantic Alignment (HOSA) module to capture the complicated semantic relationships hidden in image-text pairs.

Given a visual feature set $\mathcal{V} \in \mathbb{R}^{n \times m \times d}$, and a text feature set $\mathcal{T} \in \mathbb{R}^{n \times l \times d}$, where $\mathcal{V}_q = \{\mathbf{v}_q\}_{q=1}^m \in \mathbb{R}^{m \times d}$ denotes the $q-$th image with m regions, $\mathcal{T}_j = \{\mathbf{t}_j\}_{j=1}^l \in \mathbb{R}^{l \times d}$ represents the j-th sentence text with l words, and n denotes the number of image-text pairs. To minimize modal discrepancy and mitigate the semantic gap between images and texts, we introduce a mapping function $\mathcal{H} \in \mathbb{R}^{l \times m \times d}$ to align visual and text representations. Additionally, considering the high-order relationship between region-word and region-sentence, we assume that each region in an image can be represented linearly by words of the correspondence sentence. In other words, vision can be well reconstructed through textual descriptions in a common semantic space. It can be formulated as

$$\min_{\mathcal{H}} \|\mathcal{V} - \mathcal{T} * \mathcal{H}\|_F^2, \tag{1}$$

where $\mathcal{T} * \mathcal{H}$ is the t-product (Misha et al., 2011) of two three-order tensors. It can be calculated as

$$\mathcal{T} * \mathcal{H} = fold(bcirc(\mathcal{T})\, bvec(\mathcal{H})), \tag{2}$$

here $bcirc(\mathcal{T})$ is the block circulant matrix with the size of $nl \times ld$ obtained from d frontal slices of tensor $\mathcal{T}$ as

$$bcirc(\mathcal{T}) = \begin{bmatrix} \mathcal{T}^{(1)} & \mathcal{T}^{(d)} & \cdots & \mathcal{T}^{(2)} \\ \mathcal{T}^{(2)} & \mathcal{T}^{(1)} & \cdots & \mathcal{T}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}^{(d)} & \mathcal{T}^{(d-1)} & \cdots & \mathcal{T}^{(1)} \end{bmatrix}. \tag{3}$$

Here, $bvec(\mathcal{H})$ is to vectorize the tensor $\mathcal{H}$ along its frontal slices, i.e.,

$$bvec(\mathcal{H}) = \begin{bmatrix} \mathcal{H}^{(1)} \\ \mathcal{H}^{(2)} \\ \vdots \\ \mathcal{H}^{(d)} \end{bmatrix}, \tag{4}$$

and $fold(bvec(\mathcal{H})) = \mathcal{H}$ takes $bvec(\mathcal{H})$ back to tensor form.

Obviously, Eq.(1) is for representing visual information with textual descriptions through a linear combination of circulant, and reveals local correspondence of region-word, i.e.,$\mathcal{V}(i,j,:) = \mathcal{T}(i,k,:) * \mathcal{H}(k,j,:)$, local-global (region-sentence/ image-word) interaction, i.e.,$\mathcal{V}(i,j,:) = \mathcal{T}_i * \mathcal{H}(:,j,:)$, $\mathcal{V}_i = \mathcal{T}(i,j,:) * \mathcal{H}_j$, and global correspondence of image-sentence, i.e., $\mathcal{V}_i = \mathcal{T}_i * \mathcal{H}$.

Instead of directly using the fragment embeddings (Ji et al., 2020a; Messina et al., 2021b,a; Wei and Zhou, 2020; Qu et al., 2020), we consider utilizing the global inter-modal interactions to obtain fine-grained features. It allows the fragment features to contain rich semantic information across modalities. Meanwhile, benefiting from the t-product operation, the obtained mapping function $\mathcal{H}$ can characterize both local and global structures among the fragments hidden in multiple instances.

It should be noted that the employment of the t-product ($*$) based on the circular convolution operation not only facilitates a more comprehensive understanding of semantic relationships but also ensures computational efficiency. Specifically, Compared to the complexity of using dot product operation to achieve circular convolution,i.e., $O(nl^2dm)$, our model is more efficient, reducing computational costs to $O((nld)log(nld) + nld)$ for using Fast Fourier Transform (FFT) (Rojo and Rojo, 2006), which is illustrated in Theorem 9.1 in the Appendices.

3.4. Cross-modal Relevance Measuring

Once we achieve fine-grained features with high-order relationship alignment information, we identify the semantic relevance between an image $\mathcal{V}_i$ and sentence $\mathcal{T}_j$.

Let a region $\mathbf{v}_i$ in the q-th image ($\mathcal{V}(q,i,:)$) be the query, and its representation can be reconstructed by textual information with the complicated region-sentence interaction, i.e., $\tilde{\mathcal{V}}(q,i,:) = \mathcal{T}_q * \mathcal{H}(:,i,:)$. The similarity between the region $\mathbf{v}_i$ in the q-th image and the word $\mathbf{t}_j$ of the r-th sentence is computed as

$$\mathcal{S}(q,r,i,j) = cos(\tilde{\mathbf{v}}_i, \mathbf{t}_j) = \frac{\tilde{\mathcal{V}}(q,i,:)\mathcal{T}(r,j,:)^T}{\|\tilde{\mathcal{V}}(q,i,:)\|\|\mathcal{T}(r,j,:)^T\|}, \tag{5}$$

where $\mathcal{S}$ is the region-word cosine similarity tensor with $n \times n \times m \times l$.

Following the max-over-regions sum-over-words ($M_r S_w$) pooling function used in (Messina et al., 2021a; Lee et al., 2018), the global similarity $\mathbf{S}_{qr}$ between the q-th image and the r-th sentence is computed as

$$\mathbf{S}_{qr} = \sum_{j=1}^{l} \max_{i=1,2,\ldots,m} \mathcal{S}(q,r,i,j). \tag{6}$$

Such computing similarity strategy means finding the top relevant region for each word, and then summing up these top relevant similarity scores.

For the matching part, we follow (Messina et al., 2021a; Faghri et al., 2018; Lee et al., 2018) to adopt a hinge-based triplet ranking loss, focusing the attention on hard negatives, i.e., the negatives closest to each training query. The loss function can be formulated as

$$L_{qr} = \max_{r'}[\alpha + \mathbf{S}_{qr'} - \mathbf{S}_{qr}]_+ + \max_{q'}[\alpha + \mathbf{S}_{q'r} - \mathbf{S}_{qr}]_+, \tag{7}$$

where $[x]_+ \equiv max(0,x)$, and α is a margin parameter. To improve computational efficiency, hard negatives are found in each minibatch, instead of the entire training set.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

To validate the effectiveness of our method, all the experiments are conducted on the MSCOCO (Chen et al., 2015) dataset and the Flickr30K (Young et al., 2014) dataset. In both datasets, we follow the split proposed by (Messina et al., 2021a). The detailed information is shown as follows.

- Flickr30k is an image-captioning dataset, including $31,783$ images, and each image is also given with five caption sentences. Among them, $28,000$ images are reserved for training, $1,000$ for validation, and $1,000$ for testing.

- MSCOCO is another image-captioning dataset consisting of $123,287$ images, each with five textual descriptions. In particular, it contains $113,287$ images for training, $5,000$ images for validation, and $5,000$ images for testing. At test time, the retrieval results are reported by testing on full $5k$ test split and averaging over 5-fold of $1k$ test images.

4.1.2. Implementation Details

Our method HOSA is implemented in a conda environment with TensorFlow 2.11.0 and Python 3.8, and all the experiments are conducted on a Linux server with a GeForce RTX 2080 Ti (12GB memory). In feature representation, for each image, the region vector extracted by a bottom-up attention (Anderson et al., 2018) is $2,048$-dimensional. As for the textual data, the word vector extracted by the pre-trained BERT model (Devlin et al., 2019) is 768-dimensional. For simplicity, the weights of the BERT model implemented by HuggingFace[1] are fixed during the training stage.

Following previous approaches (Messina et al., 2021b,a), we separately transform each region and word feature vector into a $1,024$-dimensional common space with Transformer Encoders for image-text alignment. The model is trained for 30 epochs with the Adam optimizer (Kingma and Ba, 2014). The learning rate is initialized as $1e-5$ for the first 20 epochs and then decayed by 10 times for the remaining 10 epochs. The mini-batch size is set to 40, and the margin parameter in Eq.(7) is set to 0.02, respectively.

4.1.3. Evaluation Metrics

Following previous work (Ji et al., 2020b; Lee et al., 2018; Messina et al., 2021a), we measure the performance with $R@K$ (recall at K), defined as the percentage of queries whose ground truth is ranked within the top K results. It can be calculated as

$$R@K = \frac{1}{N}\sum_{i=1}^{N} RL_K^{(i)}, \tag{8}$$

where N is the number of instances in the testing set. For the i-th test instance, $RL_K^{(i)}$ is set to 1 if the top K retrieved objects have the ground-truth result, otherwise, it is 0.$R@\{1,5,10\}$ are adopted in the evaluation.

4.2. Performance Comparison

To demonstrate that HOSA can embody the advantage of local-global interaction, we compared it with thirteen state-of-the-art models on two datasets. These compared models can be roughly divided into global alignment, local alignment, and multi-order alignment learning methods. The global alignment ones, i.e., VSE++ (Faghri et al., 2018), DPC (Zheng et al., 2020) and CLIP(Radford et al., 2021), explore the semantic correlations between the entire image and text. The local alignment ones,i.e., SCAN (Lee et al., 2018) and VSRN (Li et al., 2019), explore region-word correspondence to identify the relationships between image and text.

The multi-order alignment ones, i.e., SMAN (Ji et al., 2020b), AAMEL (Wei and Zhou, 2020), RESG (Liu et al., 2022), M3A-Net (Ji et al., 2020a), TERN (Messina et al., 2021b), TERAN (Messina et al., 2021a), RAAN (Wang et al., 2023) and MSG-CNN (Yu et al., 2023), integrate global and local correspondences to further align image and text. Note that, we directly quoted the results from their original papers. For each metric, the best result is in bold and the second one is underlined.

Next, we will conduct qualitative and ablation studies to investigate how the HOSA can detect complicated semantic relationships between modalities and then improve fine-grained image-text retrieval task performance.

4.3. Ablation Study

Based on the comparison results in Table 1-Table 3, we have the following observations:

- Compared with the global alignment approaches, the local alignment ones obtain better results, indicating that fine-grained correspondence information is beneficial to learning more accurate semantic alignment.

- Compared with global or local alignment models, multi-order alignment ones achieve better performance, demonstrating the important role of global or local correspondence in cross-modal semantic alignment.

- Comparing to the multi-order alignment approaches, i.e., M3A-Net (Ji et al., 2020a), AAMEL (Wei and Zhou, 2020), RESG (Liu et al., 2022), CAMERA (Qu et al., 2020) and TERN (Messina et al., 2021b), our HOSA improves the performance of image retrieval and sentence retrieval tasks on two datasets. The

[1] https://huggingface.co/bert-base-uncased

Methods	Image Retrieval			Sentence Retrieval		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
Global Alignment						
VSE++ (Faghri et al., 2018)	52.0	84.3	92.0	64.6	90.0	95.7
DPC (Zheng et al., 2020)	47.1	79.9	90.0	65.6	89.8	95.5
CLIP (Radford et al., 2021)	52.5	85.5	93.2	65.3	91.2	96.3
Local Alignment						
SCAN (Lee et al., 2018)	58.8	88.4	94.8	72.7	94.8	98.4
VSRN (Li et al., 2019)	60.8	88.4	94.1	74.0	94.3	97.8
Multi-order Alignment						
M3A-Net (Ji et al., 2020a)	58.4	87.1	94.0	70.4	91.7	96.8
AAMEL (Wei and Zhou, 2020)	59.9	89.0	95.1	74.3	95.4	98.2
RESG (Liu et al., 2022)	64.1	90.5	96.0	78.1	96.2	98.0
CAMERA (Qu et al., 2020)	62.3	90.1	95.2	75.9	95.5	98.6
TERN (Messina et al., 2021b)	54.5	86.9	94.2	65.5	91.0	96.5
TERAN (Messina et al., 2021a)	65.0	91.2	96.4	77.7	95.9	98.6
RAAN (Wang et al., 2023)	61.8	89.5	95.8	76.8	96.4	98.3
MSG-CNN (Yu et al., 2023)	62.8	90.0	95.2	78.7	95.8	98.8
HOSA	**67.0**	**92.7**	**96.9**	**79.7**	**97.0**	**99.2**

Table 1: Performance comparison of HOSA with the state-of-the-art baselines on MSCOCO 1k test set.

Methods	Image Retrieval			Sentence Retrieval		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
Global Alignment						
VSE++ (Faghri et al., 2018)	30.3	59.4	72.4	41.3	71.1	81.2
DPC (Zheng et al., 2020)	25.3	53.4	66.4	41.2	70.5	81.1
CLIP (Radford et al., 2021)	26.1	64.6	81.2	48.0	77.5	88.2
Local Alignment						
SCAN (Lee et al., 2018)	38.6	69.3	80.4	50.4	82.2	90.0
VSRN (Li et al., 2019)	37.9	68.5	79.4	50.3	79.6	87.9
Multi-order Alignment						
M3A-Net (Ji et al., 2020a)	38.3	65.7	76.9	48.9	75.2	84.4
AAMEL (Wei and Zhou, 2020)	39.9	71.3	81.7	51.9	84.2	91.2
RESG (Liu et al., 2022)	41.8	72.7	82.0	55.1	82.5	90.3
CAMERA (Qu et al., 2020)	39.0	70.5	81.5	53.1	81.3	89.8
TERN (Messina et al., 2021b)	31.4	62.5	75.3	40.2	71.1	81.9
TERAN (Messina et al., 2021a)	42.6	72.5	82.9	55.6	83.9	91.6
RAAN (Wang et al., 2023)	39.6	65.4	74.6	**64.5**	**88.5**	92.3
MSG-CNN (Yu et al., 2023)	42.5	71.2	**84.3**	57.0	85.4	**93.2**
HOSA	**43.0**	**72.9**	83.9	56.9	83.9	91.4

Table 2: Performance comparison of HOSA with the state-of-the-art baselines on MSCOCO 5k test set.

main reason is that they focus on the relationship between fragments within each modality, ignoring the semantic relationship between regions and words in each instance.

- When comparing with the multi-order alignment approaches considering both local and global correspondence, i.e., TERAN (Messina et al., 2021a), RAAN (Wang et al., 2023) and MSG-CNN (Yu et al., 2023), our HOSA also achieves a relative improvement for image retrieval or sentence retrieval task. With regard to the image retrieval task, our method consistently outperforms all baselines on different datasets. As for the sentence retrieval task, our method obtains more than $1.0\%/2.3\%$ relative gain against the best baseline (MSG-CNN/TERAN) in terms of $R@1$ on MSCOCO 1k and Flickr30K.

In general, our method obtains significant improvements in $R@1$ and $R@5$, indicating its strong ability to retrieve the most relevant images or documents. This highlights the effectiveness of HOSA in leveraging high-order semantic relationships between the image and text modalities.

Methods	Image Retrieval			Sentence Retrieval		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
Global Alignment						
VSE++ (Faghri et al., 2018)	39.6	70.1	79.5	52.9	80.5	87.2
DPC (Zheng et al., 2020)	39.1	59.2	80.9	55.6	81.9	89.5
CLIP (Radford et al., 2021)	36.0	71.9	83.4	55.8	80.7	88.3
Local Alignment						
SCAN (Lee et al., 2018)	48.6	77.7	85.2	67.4	90.3	95.8
VSRN (Li et al., 2019)	53.0	77.9	85.7	70.4	89.2	93.7
Multi-order Alignment						
SMAN (Ji et al., 2020b)	43.4	73.7	83.4	57.3	85.3	92.2
AAMEL (Wei and Zhou, 2020)	49.7	79.2	86.4	68.5	91.2	95.9
RESG (Liu et al., 2022)	57.2	82.4	89.1	74.8	94.0	**97.3**
CAMERA (Qu et al., 2020)	58.9	84.7	90.2	76.5	**95.1**	97.2
TERN (Messina et al., 2021b)	41.1	71.9	81.2	53.2	79.4	86.0
TERAN (Messina et al., 2021a)	59.5	84.9	90.6	75.8	93.2	96.7
RAAN (Wang et al., 2023)	56.0	82.4	89.1	74.5	93.6	95.8
MSG-CNN (Yu et al., 2023)	57.2	82.4	89.1	74.8	94.0	**97.3**
HOSA	**60.3**	**85.7**	**91.4**	**78.1**	92.4	96.3

Table 3: Performance comparison between our HOSA and the state-of-the-art baselines on Flickr30K.

Model	Image Retrieval			Sentence Retrieval		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
MSCOCO 1k						
TERAN (Messina et al., 2021a)	65.0	89.5	96.4	77.7	95.9	98.6
$\text{Text}_{rec\odot}$	64.8	90.5	96.8	77.5	96.1	98.8
$\text{Image}_{rec\odot}$	65.8	92.0	96.7	78.0	96.5	99.0
Text_{rec*}	66.2	92.6	96.7	79.5	96.7	**99.3**
Image_{rec*}	67.0	**92.7**	**96.9**	79.7	**97.0**	99.2
$\text{Image}_{rec_*} + \text{Text}_{rec_*}$	**67.1**	92.4	96.8	**80.5**	96.0	99.1
MSCOCO 5k						
TERAN (Messina et al., 2021a)	42.6	72.5	82.9	55.6	83.9	91.6
$\text{Text}_{rec\odot}$	42.8	72.4	83.0	55.8	83.5	91.3
$\text{Image}_{rec\odot}$	42.4	72.2	82.3	55.4	83.2	91.2
Text_{rec*}	42.6	72.8	83.0	56.3	84.5	91.6
Image_{rec*}	43.0	**72.9**	**83.9**	56.9	83.9	91.4
$\text{Image}_{rec_*} + \text{Text}_{rec_*}$	**43.2**	72.7	83.0	**57.2**	**84.6**	**91.7**
Flickr30k						
TERAN (Messina et al., 2021a)	59.5	84.9	90.6	75.8	**93.2**	**96.7**
$\text{Text}_{rec\odot}$	59.6	84.0	90.6	75.6	92.8	96.2
$\text{Image}_{rec\odot}$	59.8	84.8	90.8	76.2	92.6	96.5
Text_{rec*}	**60.5**	85.2	91.0	75.8	93.0	96.4
Image_{rec*}	60.3	**85.7**	**91.4**	**78.1**	92.4	96.3
$\text{Image}_{rec_*} + \text{Text}_{rec_*}$	60.0	85.0	90.9	76.0	92.9	96.3

Table 4: Analysis of the effects of one modality's reconstruction with another and the applied Hadamard product ($\odot$) and t-product ($*$) operations.

4.4. Case Study

In this subsection, we conduct a qualitative analysis to demonstrate the effectiveness of the proposed model for capturing the high-order semantic alignment between modalities. We compare HOSA with the top-performing baseline TERAN (Messina et al., 2021a). Figure 2 shows the top retrieved sentences given image queries on the MSCOCO 1K test set.

We can see TERAN mistakenly matched the words such as "bike", "clock", and "tire" with the specific regions such as "preserver" in the queried image, as their appearances are similar, while the retrieved sentences are semantically inconsistent with the whole image queries. The key reason for this is that TERAN focuses on capturing local-to-local correspondence and ignores that a word or a region may have different semantics in differ-

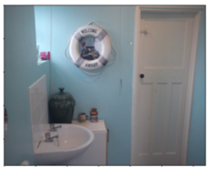
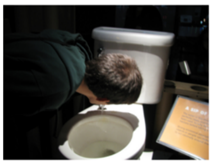
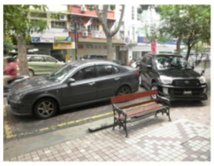
Query	TERAN	HOSA
	**Rank1:** The bike has a clock as a tire. ✗ **Rank2:** There is a GOL plane taking off in a partly cloudy sky. ✗	**Rank1:** A blue boat themed bathroom with a life preserver on the wall. ✓ **Rank2:** Blue and white color scheme in a small bathroom. ✓
	**Rank1:** This is an advanced toilet with a sink and control panel. ✗ **Rank2:** A cat is drinking water out of a toilet. ✗	**Rank1:** A man getting a drink from a water fountain that is a toilet. ✓ **Rank2:** A young man drinking from a water fountain in the shape of a toilet. ✓
	**Rank1:** Two husky dogs ride in a car with their heads hanging out windows. ✗ **Rank2:** City street with parked cars and a bench. ✓	**Rank1:** two cars parked on the sidewalk on the street. ✓ **Rank2:** A couple of cars parked in a busy street sidewalk. ✓

Figure 2: Demonstration of sentence retrieval results ontained by HOSA and TERAN on MSCOCO 1K. For easier reference, we have highlighted some objects and attributes in blue within the textual sentences, while verbs are indicated in red.

ent global contexts, i.e., the local-global (region-sentence and image-word) interactions.

In contrast, HOSA is capable of accurately retrieving relevant results in response to image queries, because it has the ability to determine the complicated inter-modal correlation, not only encompassing local-to-local and global-to-global correspondence but also capturing local-global interaction. This attests to the value of modality reconstruction assisted by t-product in determining high-order semantic correlations for cross-modal alignment.

We perform a series of ablation experiments on the two benchmark datasets to evaluate the contribution of the key components of HOSA.

Specifically, we emphasize on exploring two aspects. First, we analyze the ways of modality reconstruction, including image reconstruction with textual information, text reconstruction with visual information, and a unified approach. Second, we compare the applied tensor operations of the reconstruction process, including the Hadamard product (element-wise multiplication) and the t-product. We use the following notations to indicate specific settings.

- $\mathrm{Image}_{rec\odot}$ and Image_{rec*} denote reconstructing images with textual information using the Hadamard product ($\odot$) and the t-product ($*$), respectively.
- $\mathrm{Text}_{rec\odot}$ and Text_{rec*} denote reconstructing texts with visual information using the Hadamard product and the t-product, respectively.
- Image$_{rec_*}$+Text$_{rec_*}$ denotes unifying image reconstruction and text reconstruction into one framework with t-product, i.e., $\|\mathcal{V}-\mathcal{T}*\mathcal{H}\|_F^2+\|\mathcal{T}-\mathcal{V}*\mathcal{H}\|_F^2$.

From the comparison results shown in Table 4, we can have the following observations:

- When employing the t-product rather than using the Hadamard product, all modal reconstruction strategies yield better performance. The improvements are significant and consistent in $R@1$ and $R@5$. The results demonstrate that using the t-product based on the circular convolution operation is beneficial to explore the high-order semantic relationship.
- From the perspective of modal reconstruction, reconstructing the item based on the queries can achieve better performance. Specifically, if it is to retrieve images given text queries, the image reconstruction strategy obtains better results. If returning textual descriptions given image queries, the text reconstruction strategy plays a more important role. Besides, the unified approach (Image$_{rec_*}$+Text$_{rec_*}$) achieves better performance than the text reconstruction strategy Text$_{rec*}$, while slightly worse than the image reconstruction strategy Image$_{rec_*}$ in most cases. So our study suggests that image reconstruction is a better choice. The reason may be that text features are better at describing relevant topics, which is conducive to identifying more accurate semantic relationships between modalities.

- Compared with TERAN, the image reconstruction strategy with either the t-product or the Hadamard product achieves better performance, verifying that such an alignment strategy indeed benefits from capturing semantic relevance.

5. Conclusion

In this paper, we present a novel High-Order Semantic Alignment (HOSA) model for unsupervised fine-grained image-text retrieval. Our main idea is to construct one modal using another modal's information with the linear combination of circulation in a common latent space. It can simultaneously capture local correspondences, global correspondences, and local-global correspondences across different modalities, thereby identifying comprehensive semantic alignment for subsequent retrieval tasks. Both qualitative and quantitative experiments conducted on two standard datasets demonstrate the superiority of the proposed HOSA compared with the state-of-the-art methods. Ablation studies further validate the theoretical effectiveness of our model. Future works include integrating attention-aware learning with HOSA to identify discriminative inter-modal semantic relationships from multiple perspectives.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62306188, No.62376166). Miaomiao Cheng and Wei Song are the corresponding authors.

7. Bibliographical References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and vqa. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

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

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the Association for Computational Linguistics*, pages 4171–4186.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proc. of the British Machine Vision Conference*, page 12.

Zhong Ji, Zhigang Lin, Haoran Wang, and Yuqing He. 2020a. Multi-modal memory enhancem- ent attention network for image-text matching. *IEEE Access*, 8:38438–38447.

Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2020b. Sman: Stacked multimodal attention network for cross-modal image-text retrieval. *IEEE Transactions on Cybernetics*, 52(2):1086–1097.

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

Misha E. Kilmer, Karen Braman, Ning Hao, and Randy C. Hoover. 2013. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proc. of the European Conference on Computer Vision*, pages 201–216.

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

Xuancheng Liu, Yuanming He, Yui Man Cheung, Xiaowei Xu, and Nianhua Wang. 2022. Learning relationship-enhanced semantic graph for fine-grained image-text matching. *IEEE Transactions on Cybernetics*, 54(2):948–961.

Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S. Yu. 2016. Composite correlation quantization for efficient multimodal retrieval. In *Proc. of the ACM Special Interest Group on Information Retrieval*, pages 579–588.

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

Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2021b. Transformer reasoning network for image-text matching and retrieval. In *Proc. of the International Conference on Pattern Recognition*, pages 5222–5229.

E. Misha, D. Kilmer, Carla, and Martin. 2011. Factorization strategies for third-order tensors - sciencedirect. *Linear Algebra and its Applications*, 435(3):641–658.

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

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of the Advances in Neural Information Processing Systems*, pages 1137–1149.

Oscar Rojo and Héctor Rojo. 2006. Some results on symmetric circulant matrices and on symmetric centrosymmetric matrices. *Linear Algebra and Its Applications*, 392:211–233.

Xiaobo Shen, Fumin Shen, Quan-Sen Sun, Yang Yang, Yun-Hao Yuan, and Heng Tao Shen. 2017. Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval. *IEEE Transactions on Cybernetics*, 47(12):4275–4288.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. In *Proc. of the International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3792–3798.

Ying Wang, Yulei Su, Wenjing Li, and et al. 2023. Rare-aware attention network for image-text matching. *Information Processing and Management*, 60(3):103280.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2020. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773.

Kai-Ya Wei and Zhibo Zhou. 2020. Adversarial attentive multi-modal embedding learning for image-text matching. *IEEE Access*, 8:96237–96248.

Runde Yu, Fusheng Jin, Zhuang Qiao, Ye Yuan, and Guoren Wang. 2023. Multi-scale image–text matching network for scene and spatio-temporal images. *Future Generation Computer Systems*, 142:292–300.

Jian Zhang, Yuxin Peng, and Mingkuan Yuan. 2018. Unsupervised generative adversarial cross-modal hashing. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 539–546.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23.

8. Language Resource References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.

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

9. Appendices

In this section, we provide a detailed discussion of Theorem 9.1 (Kilmer et al., 2013), which plays a crucial role in understanding the computational efficiency of the tensor-product (t-product) operation used in our High-Order Semantic Alignment (HOSA) model.

Theorem 9.1. *For a tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, its block-circulant matrix can be block-diagonalized by*

$$(\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) bcirc(\mathcal{A}) (\mathbf{F}_{n_3}^{-1} \otimes \mathbf{I}_{n_2}) = \overline{\mathcal{A}} = bdiag(\overline{\mathcal{A}}), \tag{9}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{F}_{n_3}$ is the $n_3 \times n_3$ Discrete Fourier Transform (DFT) matrix, $\mathbf{I}_{n_1}$ and $\mathbf{I}_{n_2}$ denote $n_1 \times n_1$ and $n_2 \times n_2$ identity matrices, respectively. $bdiag(\overline{\mathcal{A}})$ is denoted as the following form:

$$bdiag(\overline{\mathcal{A}}) = \begin{bmatrix} \overline{\mathcal{A}}^{(1)} & & & \\ & \overline{\mathcal{A}}^{(2)} & & \\ & & \ddots & \\ & & & \overline{\mathcal{A}}^{(n_3)} \end{bmatrix}. \quad (10)$$

According to the relation between the circular convolution and the Discrete Fourier Transform, we note that $\mathcal{T} * \mathcal{H} \Leftrightarrow bdiag(\overline{\mathcal{T}})bvec(\overline{\mathcal{H}})$. Thus the optimization solution of our model (Eq.(1)) can be obtained by solving d independent optimization problems. For the i-th ($i = 1, 2, \dots, d$) subproblem, we have

$$arg \min_{\overline{\mathcal{H}}^{(i)}} \|\overline{\mathcal{V}}^{(i)} - \overline{\mathcal{T}}^{(i)}\overline{\mathcal{H}}^{(i)}\|_F^2. \quad (11)$$

After solving each frontal slice of $\overline{\mathcal{H}}$, we could get the solution of $\mathcal{H}$ via inverse Frourier transformation, i.e., $\mathcal{H} = ifft(\overline{\mathcal{H}}, [], 3)$.