Multi-Scale Visual Semantics Aggregation with Self-Attention for End-to-End Image-Text Matching

Zhuobin Zheng[1,2] † ZHENGZB16@MAILS.TSINGHUA.EDU.CN

Youcheng Ben[1,2] † BYC16@MAILS.TSINGHUA.EDU.CN

Chun Yuan[2,3] * YUANC@SZ.TSINGHUA.EDU.CN

[1] *Department of Computer Science and Technologies, Tsinghua University, Beijing, China*

[2] *Graduate School at Shenzhen, Tsinghua University, Shenzhen, China*

[3] *Peng Cheng Laboratory, Shenzhen, China*

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

Image-text matching has been a hot research endeavor recently. One promising direction is to infer fine-grained correspondences between visual instances and textual concepts, which makes learning instance-level visual features fundamental to this task. Detection-based approaches extract visual features directly from region proposals, but they are neither end-to-end learnable requiring extensive annotations nor adaptive to unseen instances. Attention-based approaches sequentially attend to different visual semantics in fixed time steps with global context as reference, but they are not flexible to handle situations when varying number of instances exist in different images. In this paper, we propose **S**elf-**A**ttention **V**isual-semantic **E**mbeddings (SAVE), which aggregates instance-level semantics from all potential positions of the image in an end-to-end manner. Specifically, feature maps with spatial size $k \times k$ are first divided into k^2 instance candidates. For each instance candidate, we explore two variants of self-attention mechanisms to model its correlation with others and aggregate similar semantics, which exploits flexible spatial dependencies between distant regions. Furthermore, a multi-scale feature fusion technique is utilized to obtain different levels of semantic concepts for richer information from different representation scales. We evaluate our model on two benchmark datasets: MS-COCO and Flickr30K, which demonstrates both effectiveness and applicability of our method with favorably competitive performance as the state-of-the-art approaches.

Keywords: visual-semantic embeddings, end-to-end, self-attention, multi-scale

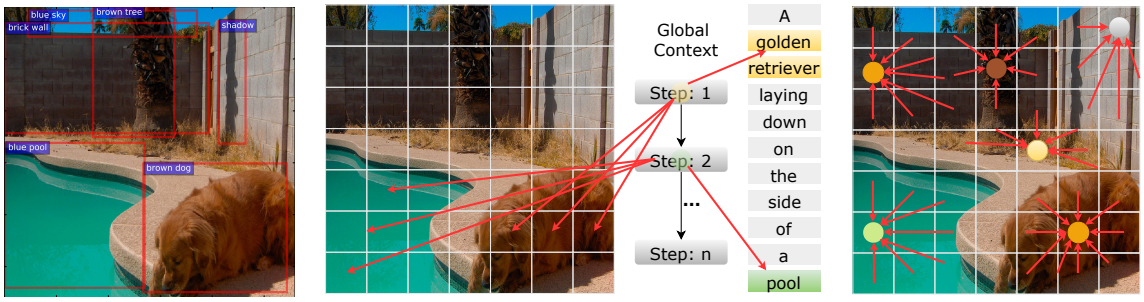
1. Introduction

Research at the intersection of vision and language has been popular in recent years. In this paper we attend to the problem of image-text matching, which is to search images for given sentences with visual descriptions or to retrieve sentences that are relevant to given image queries.

The primary challenge of this task lies in the heterogeneity of data since images and texts are of different modalities. Given that the essence of image-text retrieval is to measure

† The first two authors contribute equally to this work.

* Corresponding author: Chun Yuan.



(a) Region proposals (b) Global context as reference (c) Self-attention (**Ours**)

Figure 1: A comparison of different ways to capture instance-level image features.

the visual-semantic similarity between pairwise images and sentences, it is crucial to map multi-modal data into a shared semantic space in the first place. The question is how to conduct this mapping appropriately.

Some previous approaches [Frome et al. (2013); Kiros et al. (2014); Vendrov et al. (2016); Yan and Mikolajczyk (2015); Wang et al. (2016); Zheng et al. (2017)] attempt to map *global* representations (*i.e.,* features extracted from whole images and full sentences) into the joint embedding space. However, high-level semantics entailed in the global context are not sufficient for the retrieval of images or sentences with well-aligned local details, such as objects in images and phrases in sentences. Recently, [Gu et al. (2018); Huang et al. (2018)] incorporate generative processes into the cross-modal embedding in order to improve the global representations, but training together with generative models such as GAN Goodfellow et al. (2014) makes the overall system complicated and possibly unstable.

In contrast, several approaches [Karpathy et al. (2014); Niu et al. (2017); Huang et al. (2017); Nam et al. (2017); Lee et al. (2018)] have shown the benefits to infer *local* correspondences between visual objects and textual words, which is essential to more interpretable image-text matching. To this end, the model should capture fine-grained representations from salient objects or stuff in images (namely *instance-level* features), and then align them with word or phrase embeddings in the shared semantic space. In this case, *learning instance-level image features is fundamental to the retrieval task.*

To obtain instance-level image features, [Karpathy et al. (2014); Niu et al. (2017); Lee et al. (2018)] attend to salient regions explicitly predicted by the pre-trained object detectors. Despite preciseness, these approaches are not end-to-end trainable, which makes it hard to apply them to different contexts due to inconsistent optimization objectives for detection and retrieval. In addition, training object detectors [Shaoqing et al. (2015); Ross et al. (2014)] usually requires expensive human annotations. To circumvent this situation, [Nam et al. (2017); Huang et al. (2017)] harness global context as the reference and attend to different regions for instance-level features sequentially in a pre-defined number of steps, despite the number of visual instances varys in different images. In terms of the retrieval performance, these approaches are not satisfactory compared to detection-based ones.

In this paper, we strive to capture instance-level visual features from all potential instances in an end-to-end manner. Inspired by detection-based approaches [Lee et al. (2018); Niu et al. (2017)], we attempt to replace region proposals with self-attention mechanism (Figure 1) and investigate its effectiveness in aggregating instance-level semantics. The

method we proposed, namely SAVE, applies self-attention to multi-scale feature maps for different levels of visual semantics. Specifically, features maps of spatial size $k \times k$ are evenly split into k^2 instance candidates, corresponding to k^2 regions in the input image. For each instance candidate, we explore two different ways, Deterministic Self-Attention (DSA) and Adaptive Self-Attention (ASA), to estimate its correlation with *all candidates* as the attention, and aggregate instance-level semantics as a weighted sum of *all candidates* from the feature maps.

Besides, multi-scale fusion is applied to these features for both spatial details and global semantics. To evaluate the performance of our approach, we perform a series of experiments on two benchmark datasets: MS-COCO Lin et al. (2014) and Flickr30K Plummer et al. (2015). In summary, our contributions are four-fold:

1. Two variants of self-attention are explored to aggregate instance-level visual semantics from all potential regions and capture spatially long-range dependencies.

2. SAVE simultaneously infers all possible alignments between images and sentences in an end-to-end manner, which improves both flexibility and adaptability.

3. Multi-scale feature fusion is applied to further enhance the semantics entailed in instance-level features, which incorporate both low-level spatial details and high-level semantic concepts for the retrieval.

4. Our model achieves results on par with or better than the state-of-the-art approaches on both datasets. However, it neither involves additional generative processes nor uses object detectors pre-trained on large datasets with expensive human annotations.

2. Related Work

2.1. Visual-Semantic Embeddings Learning

A few works directly captured global representations for image-text matching. Frome et al. (2013) proposed the first visual-semantic embedding model, with CNN Alex et al. (2012) and Skip-Gram Mikolov et al. (2013) to extract features for images and labels respectively. A hinge-based triplet ranking loss was optimized to ensure the matched image-label pairs have smaller distances than mismatched pairs. Kiros et al. (2014) proposed a similar framework, which replaced Skip-Gram with LSTM Hochreiter and Schmidhuber (1997) as the sentence encoder. Vendrov et al. (2016) introduced an improved objective which can preserve the partial order structure of visual-semantic hierarchy. Wang et al. (2016) additionally considered within-view constraints to capture structure-preserving representations. Similarly, Zheng et al. (2017) proposed the intra-modal instance loss to learn more discriminative embeddings. Faghri et al. (2018) optimized the ranking objective with hardest negatives. Gu et al. (2018) incorporated generative processes to learn local grounded features, while Huang et al. (2018) used a context-gated sentence generation scheme to enhance semantics in image representations.

Other works explored the alignment of visual objects and textual words. Karpathy and Fei-Fei (2015) performed the learning of local similarities between pairwise image instances (detected by R-CNN Ross et al. (2014)) and words. Niu et al. (2017) adopted a

Figure 2: The proposed SAVE captures instance-level features by aggregating semantics from visually similar regions rather than local regions of fixed shape. In each case, the first image shows three representative query locations with red bounding boxes. The other three images are attention maps for those query locations, with the transparency summarizing the most-attended regions.

tree-structured LSTM Tai et al. (2015) to embed phrases and model hierarchical interactions between visual objects and phrases. Most recently, Lee et al. (2018) obtained image instances with Faster R-CNN Shaoqing et al. (2015) and proposed Stacked Cross Attention, similar to bi-directional attention flow Minjoon et al. (2017), to aggregate local similarities attentively. Instead of aggregating instance-level image features with object detectors, Huang et al. (2017) proposed a context-modulated attention theme to selectively attend to a pair of instances which appear in both the image and sentence. Likewise, Nam et al. (2017) proposed Dual Attention Network to capture fine-grained interplay between vision and language through multiple steps. Our approach also infers local correspondences between image-sentence pairs. However, rather than object detectors or multi-modal global context, we aggregate instance-level image features with self-attention mechanisms.

2.2. Self-Attention Models

Self-attention [Cheng et al. (2016); Parikh et al. (2016)] computes the response at a position in a sequence by attending to all positions within the same sequence. Vaswani et al. (2017) demonstrated that machine translation models could achieve state-of-the-art results by solely using a self-attention model. Wang et al. (2018) formalized self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Zhang et al. (2018) integrated GAN with self-attention to allow long-range dependency modeling for image generation. In spite of this progress, self-attention has not been well explored in the context of cross-modal retrieval. Rather than applying self-attention as non-local blocks between convolutional layers to model long-range dependency, we purely impose it on feature maps to aggregate similar semantics and obtain instance-level image features for the retrieval task.

3. The Proposed Method

In this section, we present the details of the proposed SAVE method. We begin with deep convolutional neural networks to capture image features as instance candidates. Meanwhile, recurrent neural networks are leveraged to embed words in sentences. Following that, we impose self-attention on candidate image features to aggregate instance-level semantics. Multi-scale feature fusion is explored to further enhance the semantics. Finally, Stacked Cross Attention Lee et al. (2018) is adopted to infer the image-sentence similarity by aggregating local similarities between image instances and words.

3.1. Input Representation

3.1.1. Image Representation

Given an input image I, we aim to represent it with a set of instance-level features $V = \{v_1, \cdots, v_k\}, v_i \in \mathbb{R}^D$, such that each feature attends to a salient object or region in the image. However, it is difficult to obtain these features directly, since the visual content is arbitrary where the instances could appear in any location with various scales. Different from [Karpathy and Fei-Fei (2015); Niu et al. (2017); Lee et al. (2018)], which capture image features from region proposals predicted by object detectors [Ross et al. (2014); Shaoqing et al. (2015)], we *directly* encode images with plain neural networks, *e.g.*, VGG Simonyan and Zisserman (2015) or ResNet He et al. (2016).

To obtain instance-level visual semantics, we first define candidate semantics based on feature maps as follows. For a feature layer of spatial size $M \times N$ with C channels, we concatenate values at each spatial location along the channel dimension in order to obtain $k = M \times N$ feature vectors $U = \{u_1, \cdots, u_k\}, u_i \in \mathbb{R}^C$. These features encoding specific regions in the raw input image are thus regarded as candidate semantics for those regions. We then impose self-attention on these features in order to aggregate instance-level visual semantics V. More details are introduced in Section 3.2.

3.1.2. Sentence Representation

For a sentence $S = \{w_1, \cdots, w_n\}$, the underlying instances mostly exist in the word level or phrase level. Therefore, the goal here is to learn word (or phrase) embeddings. We first tokenize and split the sentence into words, and then employ a bi-directional GRU (BGRU) Bahdanau et al. (2015) to embed the words along with the sentence context.

For the i-th word w_i in the sentence, we first represent it with an one-hot vector $\mathbb{I}_i$ to indicate its position in the vocabulary, and then embed $\mathbb{I}_i$ into a d-dimensional vector x_i with an embedding matrix W_x, *i.e.*, $x_i = W_x\mathbb{I}_i, i \in [1, n]$. These word embeddings are sequentially taken as the input for a bi-directional GRU and we average the hidden states in both directions at the same timestep as the final embedding $e_i \in \mathbb{R}^D$ for each word.

3.2. Self-Attention Mechanisms

We aggregate instance-level visual semantics based on the correlations between candidate features, which can be interpreted as a kind of self-attention. Two variants are proposed to estimate this attention, namely the Deterministic Self-Attention (DSA) and the Adaptive

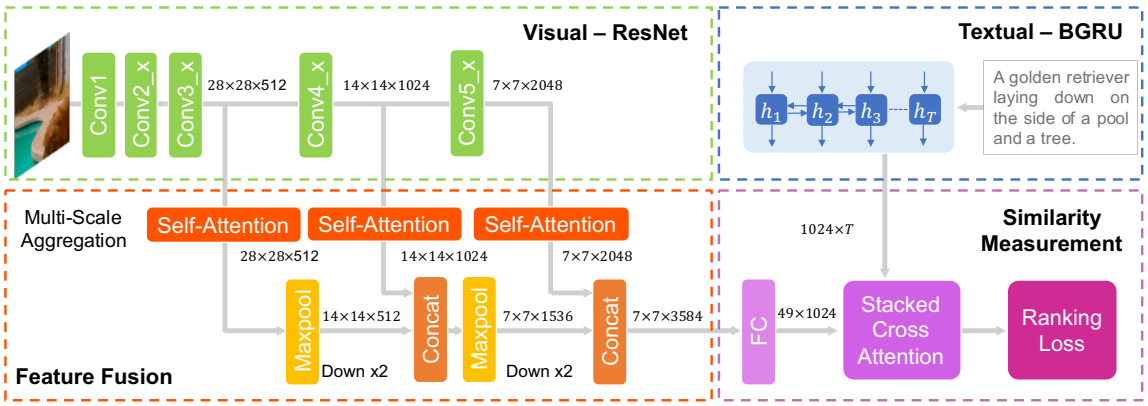


Figure 3: An overview of the proposed SAVE method. 1) ResNet He et al. (2016) and BGRU Bahdanau et al. (2015) are leveraged to capture visual and textual features respectively. 2) Self-Attention is imposed on different scales of feature maps in order to aggregate different levels of visual semantics. 3) Feature Fusion is conducted sequentially by down-sampling and concatenation to reduce the number of image instances. 4) Stacked Cross Attention Lee et al. (2018) is applied to measure the final image-text similarity. 5) The overall model is optimized end-to-end by minimizing the ranking loss.

Self-Attention (ASA). DSA is an intuitive solution, which calculates the attention maps directly from cosine similarities between all pairs of instance candidates, while ASA adopts the non-local operation Wang et al. (2018) to model the correlations between them adaptively. Instance-level features are then aggregated from candidates according to the attention maps.

3.2.1. Deterministic Self-Attention

For an input image I, we have defined a set of instance candidates $U = \{u_1, \cdots, u_k\}, u_i \in \mathbb{R}^C$. To aggregate instance-level features $V = \{v_1, \cdots, v_k\}, v_i \in \mathbb{R}^D$, we first compute the cosine similarity matrix for all possible candidate pairs, *i.e.*,

$$s_{ij} = \frac{{u_i}^T u_j}{||u_i|| ||u_j||}, i, j \in [1, k], \tag{1}$$

where s_{ij} represents the similarity between the i-th candidate u_i and the j-candidate u_j. We follow [Lee et al. (2018); Karpathy et al. (2014)] to threshold s_{ij} at zero and normalize the similarity matrix as $\bar{s}_{ij} = [s_{ij}]_+ / \sqrt{\sum_{i=1}^{k} [s_{ij}]_+^2}$, where $[x]_+ \equiv \max(x, 0)$.

To gather correlative candidate semantics with respect to i-th instance candidate u_i, we define a weighted combination of all the candidates as the attended instance-level features

$$a_i = \sum_{j=1}^{k} \alpha_{ij} u_j, \tag{2}$$

where

$$\alpha_{ij} = \frac{\exp(\bar{s}_{ij})}{\sum_{j=1}^{k} \exp(\bar{s}_{ij})}. \tag{3}$$

We also add a fully-connected layer to transform a_i into a D-dimensional feature vector

$$v_i = W_v a_i + b_v. \tag{4}$$

Therefore, the final representation of an image I is a set of instance-level embedding vectors $V \subset \mathbb{R}^D$.

As is shown in Figure 2, attention maps computed by DSA on 7×7 feature maps to some extent comply with the shape of visual instances in the input image. In terms of the convolution operation, the similarity of candidate features at a given location is high in local neighbourhood and in visually similar regions of the image. Therefore, *from appropriate query locations*, we can aggregate meaningful instance-level visual semantics.

3.2.2. Adaptive Self-Attention

In fact, we can adopt the non-local operation Wang et al. (2018) to model the correlations between all pairs of instance candidates and adaptively learn the attention maps.

Given image feature $x \in \mathbb{R}^{C \times M \times N}$, we first reshape it into a two-dimensional matrix $u \in \mathbb{R}^{C \times k}, k = M \times N$, which represents k instance candidates as explained in Section 3.1.1. To obtain the attention map $\alpha \in \mathbb{R}^{k \times k}$, u is mapped into two feature spaces θ, ϕ, where $\theta(u) = W_\theta u$, $\phi(u) = W_\phi u$, and

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{j=1}^{k} \exp(s_{ij})}, \tag{5}$$

where

$$s_{ij} = \theta(u_i)^T \phi(u_j), i, j \in [1, k]. \tag{6}$$

α_{ij} indicates the extent to which the model attends to the j-th candidate when aggregating semantics for the i-th candidate. Then the output of the attention layer is $a = (a_1, a_2, \cdots, a_k) \in \mathbb{R}^{C \times k}$, where

$$a_i = \sum_{j=1}^{k} \alpha_{ij} g(u_j). \tag{7}$$

In the above formulation, $g(u_j) = W_g u_j$, and $W_\theta \in \mathbb{R}^{\overline{C} \times C}$, $W_\phi \in \mathbb{R}^{\overline{C} \times C}$, $W_g \in \mathbb{R}^{C \times C}$ are the learned weight matrices, which are implemented as 1×1 convolutions. We use $\overline{C} = \frac{C}{2}$ in all our experiments.

Besides, we follow [Wang et al. (2018); Zhang et al. (2018)] to further map the output of the attention layer by a scale parameter γ and add back the input feature maps. Therefore, the final output is given by

$$y_i = \gamma a_i + u_i, \tag{8}$$

where γ is initialized as 0 for identity mapping [Goyal et al. (2017); He et al. (2016)]. Similar to Equation (4), we leverage a fully-connected layer to transform y_i into a D-dimensional vector v_i.

Method	Sentence Retrieval				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
SM-LSTM (VGG) Huang et al. (2017)	42.5	71.9	81.5	2	30.2	60.4	72.3	3	358.8
DAN (ResNet) Nam et al. (2017)	55.0	81.8	89.0	1	39.4	69.2	79.1	2	413.5
VSE++ (ResNet) Faghri et al. (2018)	52.9	80.5	87.2	1	39.6	70.1	79.5	2	409.8
DPC (ResNet) Zheng et al. (2017)	55.6	81.9	89.5	1	39.1	69.2	80.9	2	416.2
SCO (ResNet) Huang et al. (2018)	55.5	82.0	89.3	-	41.1	70.5	80.1	-	418.5
DVSA (R-CNN, AlexNet) Karpathy and Fei-Fei (2015)	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2	235.2
HM-LSTM (R-CNN, AlexNet) Niu et al. (2017)	38.1	-	76.5	3	27.7	-	68.8	4	-
SCAN (Faster R-CNN, ResNet) Lee et al. (2018)	67.4	90.3	95.8	-	48.6	77.7	85.2	-	465.0
Ours: (ResNet)									
SAVE (DSA)	65.2	88.6	93.7	1	49.0	77.1	85.3	2	458.9
SAVE (ASA)	64.9	88.5	92.4	1	48.7	77.4	85.5	2	457.4
SAVE (ASA + ms2)	66.6	89.0	94.0	1	49.2	78.1	86.2	2	463.1
SAVE (ASA + ms3)	67.2	88.3	94.2	1	**49.8**	**78.7**	**86.2**	2	464.4

Table 1: Comparison of the cross-modal retrieval results in terms of Recall@K(R@K) on Flickr30K.

3.2.3. Summary

The main purpose of proposing DSA and ASA is to show that it is feasible to aggregate instance-level visual semantics with self-attention, regardless of concrete implementations. Both DSA and ASA are general and beneficial to the retrieval task.

Specifically, DSA directly calculates the cosine similarities between candidate features, which is straightforward but effective to model correlations among high-level semantics (7×7 feature maps). However, when it comes to lower-level feature maps (14×14, 28×28), adaptively modeling the correlations with learnable parameters (ASA) can obtain more promising performance. Ablation studies in Section 4.4 have demonstrated this difference.

3.3. The Overall Architecture

Figure 3 illustrates the pipeline of the proposed SAVE method. We introduce the details of multi-scale feature fusion and similarity measurement as follows.

3.3.1. Multi-Scale Feature Fusion

As is depicted in Figure 3, we apply self-attention to different scales of feature maps and then fuse them together by sequentially down-sampling larger-scale feature maps. Low-level features are typically interpreted as fine-grained spatial details while high-level features are abstract semantic concepts. By combining both of them, we expect learned features to contain both local details and global semantics. Note that up-sampling smaller-scale feature maps for the fusion is not a wise choice, as large-scale feature maps may yield excessive image instance candidates for the retrieval task.

Importantly, different from Wang et al. (2018) adding self-attention to different stages of ResNet in order to model long-range dependency, we directly impose self-attention on multi-scale feature maps to aggregate instance-level features and obtain different levels of semantics. Extensive experiments in Section 4 demonstrate the effectiveness of our approach.

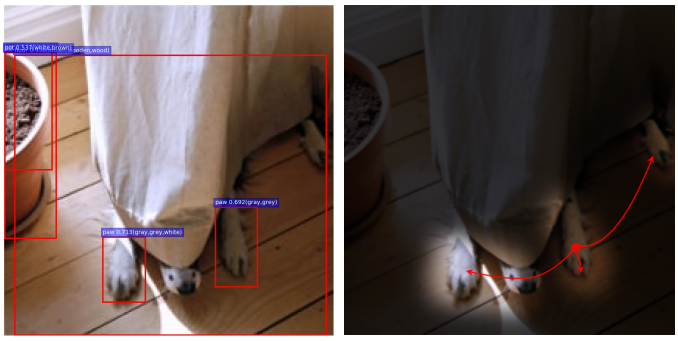


Figure 4: An example illustrating the difference between object detectors (left) and self-attention (right) in capturing visual semantics from occluded instances. Detection-based methods are stuck in occluded cases caused by only recognizing decentralized and uncorrelated instances (*e.g.,* body parts of the *dog*), while SAVE can capture visual semantics with correlations by modeling spatial dependencies, which further facilitates retrieval.

3.3.2. SIMILARITY AGGREGATION

For a whole image I and a full sentence S, we have already obtained instance-level image features $V = \{v_1, \cdots, v_k\}, v_i \in \mathbb{R}^D$ and context-aware word embeddings $E = \{e_1, \cdots, e_n\}, e_i \in \mathbb{R}^D$. To obtain a similarity score between I and S, we adopt Stacked Cross Attention Lee et al. (2018) with two directions, *i.e., Image-Text (i-t)* and *Text-Image (t-i). i-t* attends to words given each image instance, while *t-i* attends to instances given each word.

Take the *i-t* direction for example, given the i-th image instance v_i, the attended textual vector a_i^t is obtained as the weighted sum of all word embeddings $\{e_1, \cdots, e_n\}$, where the importance of each word e_j is calculated by its similarity to the image instance v_i.

Local similarity $R(v_i, a_i^t)$ is computed as the cosine similarity between v_i and a_i^t. Global similarity between I and S is then aggregated by either LogSumExp pooling (LSE)

$$S_{LSE}(I, S) = \log(\sum_{i=1}^{k} \exp(\lambda_2 R(v_i, a_i^t)))^{(1/\lambda_2)}, \tag{9}$$

which approximates to $\max_i R(v_i, a_i^t)$ as $\lambda_2 \to \infty$, or average pooling (AVG)

$$S_{AVG}(I, S) = \frac{1}{k} \sum_{i=1}^{k} R(v_i, a_i^t). \tag{10}$$

3.4. Learning Objective

Our model can be trained with a hinge-based triplet ranking loss which encourages the similarity scores of matched image-sentence pairs to be larger than those of mismatched pairs. We follow Faghri et al. (2018) to emphasize the hardest negatives in a mini-batch for training. For a matched pair (i, s), the hardest negatives are defined as $i_h = \arg\max_{x \neq i} S(x, s)$ and $s_h = \arg\max_{y \neq s} S(i, y)$. The hinge-based ranking loss is then measured by

$$\mathcal{L}_r = [m - S(i, s) + S(i_h, s)]_+ + [m - S(i, s) + S(i, s_h)]_+, \tag{11}$$

where m is the margin and $[x]_+ \equiv \max(x, 0)$.

3.5. Why Self-Attention for Visual Semantics

3.5.1. Instance-Level Aggregation

In terms of attention, recent research [Huang et al. (2017); Nam et al. (2017)] leverages global context to aggregate instance-level semantics by sequentially attending to different regions simply with high cosine similarities. However, semantics of different visual instances are intricately entangled in the global context, which may reduce its correlation with features from specific instances.

In contrast, self-attention simultaneously aggregates instance-level semantics by automatically capturing disentangled representations and accumulating correlated semantics from positions with high responses. This allows us to preserve relatively disentangled features and exploit richer information of specific instances, which facilitates more precise instance-wise retrieval.

3.5.2. Spatial Dependency

Aggregating semantics from regions with spatial dependencies is particularly important for difficult situations (*e.g.,* occluded or unseen objects). As shown in Figure 4, detection-based methods can only capture features from decentralized parts without recognizing the occluded *dog* instance by uncorrelated semantics (*e.g.,* body parts).

Intuitively, our self-attention simulates long-range dependency Vaswani et al. (2017) differently in the spatial dimension. In this case, SAVE aggregates semantics from any distant regions by spatial dependencies and bridges communication among different correlated semantics of the same instance. It yields large-range dependencies and correlations which enhances informative representation ability for the inference in retrieval tasks.

3.5.3. End-to-End

To obtain instance-level visual semantics, detection based approaches [Karpathy and Fei-Fei (2015); Niu et al. (2017); Lee et al. (2018)] require additional annotations to pre-train the object detectors. Besides, features captured from region proposals are frozen for subsequent training of the rest of the retrieval model, which limits the flexibility of these approaches, especially when deployed in novel tasks. In contrast, SAVE aggregates visual semantics based on internal correlations among instance candidates in an end-to-end manner without any task-specific supervision, which is more adaptive and applicable.

4. Experiments

4.1. Datasets

We evaluate the SAVE method on MS-COCO Lin et al. (2014) and Flickr30K Plummer et al. (2015) datasets. Flickr30K contains $31,783$ images collected from Flickr website, and each image is associated with 5 text descriptions. Following the split in Kiros et al. (2014), we use $29,000$ images for training, $1,014$ images for validation and $1,000$ images for testing. MS-COCO contains $123,287$ images, each of which is annotated with five captions. According to Kiros et al. (2014), the dataset is split into $82,783$ training images, $5,000$ validation images and $5,000$ test images. We follow [Faghri et al. (2018); Lee et al. (2018)]

Method	Sentence Retrieval				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
					1K Test Images				
SM-LSTM (VGG) Huang et al. (2017)	53.2	83.1	91.5	1	40.7	75.8	87.4	2	431.8
2WayNet (VGG) Eisenschtat and Wolf (2017)	55.8	75.2	-	-	39.7	63.3	-	-	-
VSE++ (ResNet) Faghri et al. (2018)	64.6	90.0	95.7	1	52.0	84.3	92.0	1	478.3
DPC (ResNet) Zheng et al. (2017)	65.6	89.8	95.5	1	47.1	79.9	90.0	2	467.9
GXN (ResNet) Gu et al. (2018)	68.5	-	97.9	1	56.6	-	94.5	1	-
SCO (ResNet) Huang et al. (2018)	69.9	92.9	97.5	-	56.7	87.5	94.8	-	499.3
DVSA (R-CNN, AlexNet) Karpathy and Fei-Fei (2015)	38.4	69.9	80.5	-	27.4	60.2	74.8	-	351.2
HM-LSTM (R-CNN, AlexNet) Niu et al. (2017)	43.9	-	87.8	2	36.1	-	86.7	3	-
SCAN (Faster R-CNN, ResNet) Lee et al. (2018)	72.7	94.8	98.4	-	58.8	88.4	94.8	-	507.9
SCAN (Faster R-CNN, ResNet) †	71.3	94.4	98.2	1	57.3	88.1	94.4	1	503.7
Ours: (ResNet)									
SAVE (DSA)	69.4	92.1	97.0	1	56.0	87.1	94.0	1	495.6
SAVE (ASA)	70.1	92.6	97.1	1	56.2	87.1	94.1	1	497.2
SAVE (ASA + ms2)	70.8	93.2	97.6	1	56.9	87.6	94.4	1	500.5
					5K Test Images				
VSE++ (ResNet) Faghri et al. (2018)	41.3	71.1	81.2	2	30.3	59.4	72.4	4	355.7
DPC (ResNet) Zheng et al. (2017)	41.2	70.5	81.1	2	25.3	53.4	66.4	5	337.9
GXN (ResNet) Gu et al. (2018)	42.0	-	84.7	2	31.7	-	74.6	3	-
SCO (ResNet) Huang et al. (2018)	42.8	72.3	83.0	-	33.1	62.9	75.5	-	369.6
SCAN (Faster R-CNN, ResNet) Lee et al. (2018)	50.4	82.2	90.0	-	38.6	69.3	80.4	-	410.9
SCAN (Faster R-CNN, ResNet) †	46.7	78.6	88.1	2	34.2	65.3	77.3	3	390.2
Ours: (ResNet)									
SAVE (DSA)	44.9	75.0	84.7	2	33.3	63.6	75.9	3	377.4
SAVE (ASA)	45.5	75.1	85.0	2	33.3	63.8	76.1	3	378.8
SAVE (ASA + ms2)	46.7	76.3	86.1	2	34.0	64.8	77.0	3	384.9

Table 2: Comparison of the cross-modal retrieval results in terms of Recall@K(R@K) on MS-COCO. Note that † denotes the reproduced best result of SCAN Lee et al. (2018) under exactly the same experimental settings.

to use additional 30,504 images that were originally in the validation set of MS-COCO but have been left out in this split for training (113,287 training images in total). The results are reported by either averaging over 5 folds of 1K test images or testing on the full 5K test images. Note that some early works such as [Kiros et al. (2014); Huang et al. (2017)] only use a training set containing 82,783 images.

4.2. Results on Flickr30K

Table 1 lists quantitative results[1] on Flickr30K, where all formulations of our approach outperform other end-to-end trainable models (listed in the first section) in all measures. Our best result is achieved by adaptive self-attention (ASA) with 3 scales of feature fusion, which is comparable to the state-of-the-art approach, *i.e.,* SCAN. However, our model does not require additional images or human annotations to train object detectors.

4.3. Results on MS-COCO

Table 2 presents experimental results[1]on MS-COCO 1K and 5K test sets, where our best result is achieved by adaptive self-attention (ASA) with 2 scales of feature fusion, which outperforms other end-to-end trainable models (listed in the first and fourth section) in all

[1] For each formulation, we follow Lee et al. (2018) to average similarity scores obtained in two directions (i-t and t-i) and report final retrieval results.

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
i-t LSE:						
visual genome	**42.2**	**74.8**	**85.3**	**32.1**	**62.5**	**74.9**
pascal voc	29.9	62.2	74.9	21.1	48.6	61.8
t-i LSE:						
visual genome	**43.1**	**75.1**	**85.7**	**29.9**	**60.3**	**73.2**
pascal voc	31.6	62.1	74.6	20.8	47.3	60.9

Table 3: Effect of visual features extracted by object detectors pre-trained on different datasets (Visual Genome Krishna et al. (2017) and Pascal VOC Everingham et al. (2010)) and evaluated on MS-COCO $5K$ test set. Results are reported in terms of Recall@K(R@K).

measures. However, there are gaps in results between our method and SCAN, especially in the $5K$ test set.

Several factors may account for this phenomenon. 1) The object detector used by SCAN is pre-trained on Visual Genome Krishna et al. (2017) with additional images ($51,208$ new images for MS-COCO) and annotations ($1,600$ object classes and 400 attribute classes), which leads to better generalizability of features. 2) Table 3 has shown the importance of a good detector to SCAN. The retrieval performance drops dramatically as we pre-train the same object detector on PASCAL VOC Everingham et al. (2010) rather than Visual Genome dataset, which reveals the weakness of these non end-to-end approaches when applied to different contexts. 3) We reproduce the results reported by SCAN Lee et al. (2018) after multiple trials under exactly the same experimental settings[2] though with non-negligible gaps. We ascribe this problem to the unstable performance of detection which should be improved by tuning tricks. We provide the best results reproduced by SCAN in Table 2 for comparison.

4.4. Ablation Studies

To systematically evaluate the proposed method, we perform comprehensive ablation studies in terms of the following three aspects:

1) **Effect of self-attention.** Three variants of SAVE are compared in Table 4 to evaluate its effect on 7×7 feature maps. Specifically, **no-SA** directly measures the similarity between instance candidates and word embeddings, while **DSA** and **ASA** additionally apply deterministic self-attention and adaptive self-attention to instance candidates for instance-level semantics aggregation.

Results on two datasets under two different settings have shown that self-attention can stably improve the retrieval performance. Note that **DSA** achieves better results than **ASA** in nearly all measures, which verifies the effectiveness of deterministic self-attention in modeling the correlations among high-level candidates.

2) **Effect of multi-scale instance-level features.** For convenience, we conduct experiments with at most 3 scales of feature maps (28×28, 14×14, 7×7) denoted as *c3*, *c4* and *c5* respectively. Two variants of SAVE are proposed, where **ms2** fuses 2 scales of feature maps (*c4* + *c5*) and **ms3** fuses 3 scales of feature maps (*c3* + *c4* + *c5*). To compare with approaches where self-attention is sequentially added to different stages of ResNet as

[2] Project: https://github.com/kuanghuei/SCAN

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO 5K Test Images						
i-t AVG:						
no-SA	38.4	68.6	80.4	27.2	57.6	71.3
DSA	**42.3**	**72.7**	**83.5**	**31.2**	**61.3**	**73.6**
ASA	42.2	72.5	83.2	30.7	60.8	73.2
t-i LSE:						
no-SA	39.7	70.6	81.9	28.6	58.7	71.8
DSA	**40.8**	**72.0**	**83.0**	**30.8**	**61.0**	**73.5**
ASA	40.2	71.2	82.3	29.0	59.0	71.9
Flickr30K Test Images						
i-t AVG:						
no-SA	60.0	85.4	**91.8**	42.1	74.0	83.0
DSA	59.8	**86.0**	91.7	45.5	**75.4**	**83.6**
ASA	**61.7**	85.7	90.6	**45.6**	74.5	83.1
t-i LSE:						
no-SA	57.6	85.0	90.4	43.6	74.0	82.6
DSA	**62.8**	**87.3**	**93.0**	**47.2**	**75.8**	**84.2**
ASA	60.9	84.4	90.0	44.1	73.9	82.2

Table 4: Effect of self-attention on Flickr30K and MS-COCO 5K test set. Results are reported in terms of Recall@K(R@K).

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
i-t AVG, Flickr30K:						
DSA	59.8	86.0	91.7	45.5	75.4	83.6
DSA-ms2	58.8	83.4	91.0	45.0	74.4	83.0
ASA	61.7	85.7	90.6	45.6	74.5	83.1
ASA-ms2	61.3	**88.6**	**93.8**	46.8	75.9	84.3
ASA-ms3	**64.2**	86.3	93.0	**47.4**	**76.4**	**84.5**
NL-ms2	57.2	84.7	91.5	44.8	75.0	83.4
i-t AVG, MS-COCO 5K:						
DSA	42.3	72.7	83.5	31.2	61.3	73.6
DSA-ms2	42.4	72.7	83.4	31.2	61.5	74.1
ASA	41.6	72.1	82.7	30.6	60.3	73.0
ASA-ms2	**44.2**	**73.5**	**83.8**	**33.0**	**62.6**	**74.9**
NL-ms2	40.4	69.8	80.8	31.5	61.6	74.3

Table 5: Effect of multi-scale self-attention under i-t AVG setting on Flickr30K and MS-COCO 5K test set. Results are reported in terms of Recall@K(R@K).

described in Wang et al. (2018), we denote **NL-m2** as the variant where self-attention is applied to 2 scales of feature maps (*c4*, *c5*) without feature fusion.

Table 5 shows retrieval results on two datasets under the same setting. For adaptive self-attention (ASA), improvements derived from feature fusion are noticeable, which verifies the effectiveness of ASA in modeling the correlations among candidate features from different semantic levels with learnable parameters. For deterministic self-attention (DSA), no obvious improvement is observed in all measures, which means correlations among lower-level candidate features are not easily established by cosine similarities. The performance of **NL-m2** is inferior to **ASA-m2** in all measures, which demonstrates that our way of feature fusion illustrated in Figure 3 is more effective for the retrieval task.

3) **Effect of multi-scale fusion without self-attention.** We are also interested in the effect of multi-scale fusion of features without instance-level semantics. **no-SA-ms2** represents the variant of SAVE which fuses 2 scales of feature maps (14×14, 7×7) without

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
i-t LSE:						
no-SA	**58.3**	**83.8**	**90.6**	43.5	**73.9**	**83.2**
no-SA-ms2	56.5	83.2	90.3	**43.7**	73.7	83.0
i-t AVG:						
no-SA	**60.0**	**85.4**	**91.8**	**42.1**	**74.0**	**83.0**
no-SA-ms2	58.4	83.6	90.7	41.4	73.4	82.9
t-i LSE:						
no-SA	**57.6**	**85.0**	**90.4**	**43.6**	**74.0**	**82.6**
no-SA-ms2	48.4	76.0	85.4	37.1	66.7	76.0
t-i AVG:						
no-SA	**60.5**	84.4	90.8	**46.4**	**74.5**	**83.5**
no-SA-ms2	60.2	**85.5**	**91.7**	45.0	74.4	82.6

Table 6: Effect of multi-scale features without applying self-attention on Flickr30K test set. Results are reported in terms of Recall@K(R@K).

self-attention. In comparison, **no-SA** directly uses feature maps of scale 7×7 without self-attention for cross-modal retrieval.

Table 6 presents results on Flickr30K. Compared to **no-SA**, the retrieval performance drops clearly for **no-SA-ms2** under four settings especially *t-i* LSE, which means multi-scale fusion provides redundant information for the retrieval task. Therefore, simply applying multi-scale fusion to visual features can not enhance the semantics. It is necessary to aggregate instance-level features with self-attention beforehand.

5. Conclusion

In this paper, we attempt to aggregate instance-level visual semantics from all potential image instances in an end-to-end manner. Specifically, two variants of self-attention (DSA and ASA) are explored to capture spatially long-range dependencies and model the correlations between all pairs of instance candidates, which are utilized as weights to aggregate instance-level semantics. Besides, our model is capable of inferring all possible alignments between images and sentences simultaneously, which is adaptable and applicable to retrieval tasks. Furthermore, we exploit multi-scale fusion on these features to incorporate different levels of semantics for richer information from different representation scales. Results on two public datasets, MS-COCO and Flickr30K, demonstrate the effectiveness and flexibility of our method, which is competitive with several state-of-the-art approaches.

Acknowledgments

This work is supported by NSFC project Grant No. U1833101, Shenzhen Science and Technologies project under Grant No. JCYJ20160428182137473 and the Joint Research Center of Tencent & Tsinghua.

References

Krizhevsky Alex, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105. 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *EMNLP*, pages 551–561, 2016.

Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010.

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

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680. 2014.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018.

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

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897. 2014.

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, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

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

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*, pages 740–755, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Seo Minjoon, Kembhavi Aniruddha, Farhadi Ali, and Hajishirzi Hananneh. Bidirectional attention flow for machine comprehension. In *ICLR*, 2017.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, 2017.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, pages 2249–2255, 2016.

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*, 2015.

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

Ren Shaoqing, He Kaiming, Girshick Ross, and Sun Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, pages 1556–1566, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008. 2017.

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

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

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.