Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching

Chunxiao Liu
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University
of Chinese Academy of Sciences
liuchunxiao@iie.ac.cn

Zhendong Mao*
University of Science and
Technology of China
maozhendong2008@gmail.com

An-An Liu
Tianjin University
ana0422@gmail.com

Tianzhu Zhang
University of Science and
Technology of China
tzzhang10@gmail.com

Bin Wang
Xiaomi AI Lab
wangbin11@xiaomi.com

Yongdong Zhang
University of Science and
Technology of China
zhyd73@ustc.edu.cn

ABSTRACT

Learning semantic correspondence between image and text is significant as it bridges the semantic gap between vision and language. The key challenge is to accurately find and correlate shared semantics in image and text. Most existing methods achieve this goal by representing the shared semantic as a weighted combination of all the fragments (image regions or text words), where fragments relevant to the shared semantic obtain more attention, otherwise less. However, despite relevant ones contribute more to the shared semantic, irrelevant ones will more or less disturb it, and thus will lead to semantic misalignment in the correlation phase. To address this issue, we present a novel Bidirectional Focal Attention Network (BFAN), which not only allows to attend to relevant fragments but also diverts all the attention into these relevant fragments to concentrate on them. The main difference with existing works is they mostly focus on learning attention weight while our BFAN focus on eliminating irrelevant fragments from the shared semantic. The focal attention is achieved by preassigning attention based on inter-modality relation, identifying relevant fragments based on intra-modality relation and reassigning attention. Furthermore, the focal attention is jointly applied in both image-to-text and text-to-image directions, which enables to avoid preference to long text or complex image. Experiments show our simple but effective framework significantly outperforms state-of-the-art, with relative Recall@1 gains of 2.2% on both Flicr30K and MSCOCO benchmarks.

CCS CONCEPTS

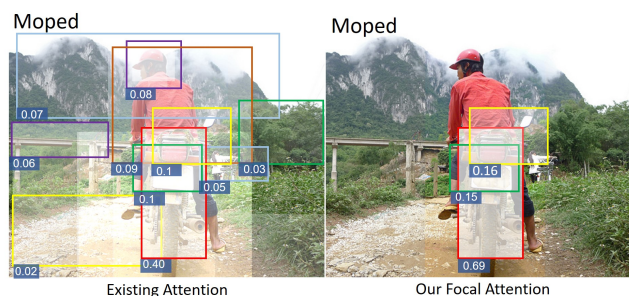• **Computing methodologies → Computer vision tasks**.

Figure 1: Existing attention vs focal attention. The attended regions are bounded by color box, where whiteness reflects attention weight. Existing attention attends to regions irrelevant to text query "moped", like road, bridge and mountain, which will lead to semantic misalignment as "moped" is learned to be similar to irrelevant regions. The focal attention avoids it by eliminating irrelevant regions.

KEYWORDS

Image-text matching, Attention

1 INTRODUCTION

There is a surge of interest in image-text matching since it bridges the semantic gap between vision and language, which has the potential to integrate multimodal information into existing applications such as the search engine, recommendation system and question answering system. The key challenge in image-text matching is to accurately find and associate shared semantics in image and text.

Existing image-text matching approaches focus on learning a neural network to find and associate shared semantics in image-text pairs. Early works [18, 21, 24] achieve this goal by projecting all the

fragments (regions and words) in image and text into a latent space without using attention mechanism. Motivated by recent progress in other cross-modal applications like visual question answering and image caption [2, 3, 19, 27, 33, 35, 36], attention has become an important component in image-text matching framework as it allows to look less on unimportant fragments and look more on important fragments in terms of specific semantic. Some attention-based approaches attend to fragments from different modalities parallelly. A typical approach is [23] that separately performs multi-step attention operation in image and text branch, in which all the shared semantics can be discovered step-by-step. Several extensions have been presented, [10] holds the insight that partial regions and words contribute to the global semantic. They propose to iteratively select important region-word pairs and learn to maximumly associate them. Despite much progress has been achieved by the above models, they neglect that the importance of regions is dynamically changed with respect to different words, and the importance of words is also dynamically changed with respect to different regions. To solve this problem, [15] proposes to attend to fragments interactively. They develop a more interpretable framework that determines the attention of fragments based on fragments from another modality. Similar works [9, 34] have been proposed motivated by the above model. Nonetheless, these approaches follow an invariant attention framework in which shared semantics are discovered by attending differentially over all the fragments.

However, despite that shared semantics can be found in previous attention mechanism, they cannot be reflected accurately. It is because many fragments are irrelevant to shared semantics, which are also attended, and thus shared semantics will be more or less disturbed. As a result, it will lead to semantic misalignment when learning to associate the shared semantics selected from image and text, i.e., irrelevant fragments from different modalities being closely correlated except for relevant fragments. As is illustrated in Figure 1, given a text fragment "moped", conventional attention methods not only attend to the target image region but also attend to its irrelevant regions like road and mountain, which will incorrectly improve relevance between "moped" and these irrelevant regions while training. Consequently, only attending to relevant fragments is crucial for learning accurate region-word correspondence.

In this paper, we propose a novel Bidirectional Focal Attention Network (BFAN) to address the semantic misalignment by only attending to relevant fragments instead of all the fragments. This is in contrast to traditional attention where the focal attention focus on irrelevant fragments removal, such that the shared semantics selected from image and text are highly relevant. The focal attention is achieved by preassigning attention, identifying relevant fragments and reassigning attention. Though it is hard to identify relevant fragments without explicit annotation, the focal attention is able to find them by learning a function that scores each fragment based on its preassigned attention relative to other fragments. Fragments that obtain higher preassigned attention than most other fragments with high confidence will be considered as relevant, otherwise irrelevant. The intuition behind this strategy is the attention distribution can roughly determine the gap in relevant and irrelevant fragments. Furthermore, we maximumly associate image-text pairs by applying the focal attention into both image-to-text and text-to-image directions as it avoids preference to long text or complex image.

The major contributions of this work can be summarized as (1) We propose a novel Bidirectional Focal Attention Network that can learn semantic alignment accurately by only focusing on relevant fragments. The focal attention is presented to score each fragment based on relative attention to all other fragments. To the best of our knowledge, it is the first work that only attends to relevant fragments and ignores irrelevant fragments in image-text matching. (2) We jointly integrate image-to-text and text-to-image matching into a unified framework, which enables to avoid the preference to long text or complex image and maximumly associate relevant image-text pairs. (3) We conduct extensive experiments on benchmarks, which demonstrate the proposed simple bidirectional focal attention network significantly outperforms state-of-the-art.

2 RELATED WORK

Many approaches have been proposed for image-text matching, which can be roughly grouped into one-to-one and many-to-many approaches. The one-to-one approaches learn correspondence between the whole image and text, and the many-to-many approaches learn correspondence between image regions and text words.

A general solution for one-to-one matching is to associate shared semantics in image-text pairs by projecting them into a common space and optimizing their relevance. Many works [37, 38] associate shared semantics by improving multimodal representations. One of the most typical works is [14] that makes the first attempt to encode image and text using Convolution Neural Networks and Recurrent Neural Networks. Similar works are proposed to associate semantic representations in common space by integrating different modules, such as identity mapping [18], character-level inception [30] and generative adversarial networks [7]. Different from them, [4, 11, 12, 28, 31, 32] learn to associate shared semantics using different objective functions. One of the most popular function is triplet ranking loss [12] that forces relevant image-text pairs being more similar than irrelevant pairs by a fixed margin. [11] goes a step further by attending to hard negatives in the ranking loss function, which significantly improves the performance. Some [28, 32] restrain the multimodal representations by preserving neighborhood structure or geometric structure. Inspired by recent progress achieved by attention mechanism, several attention-based image-text matching methods have been proposed, because it enables to attend to features differentially based on their contribution to shared semantics. A typical work is [23] that finds shared semantics by attending to specific vectors from image and text. Many extensions have also been proposed, like [6, 16].

Many-to-many approaches usually learn a latent region-word correspondence through correlating shared semantics comprised of regions and words. It is first proposed by Karpathy et al. [13] that selects shared semantics by finding most similar region-word pairs. Following this idea, [24] presents a hierarchical LSTM to jointly associate shared semantics in words and regions. Recently, attention has become the most effective method to many-to-many approach as it enables to focus more on fragments relevant to shared semantics, and less on irrelevant fragments, where fragments can be either regions or words. Many attention modules have been proposed [9, 10, 15]. sm-LSTM [10] employs attention to sequentially find all possible shared semantics and simply associates them by
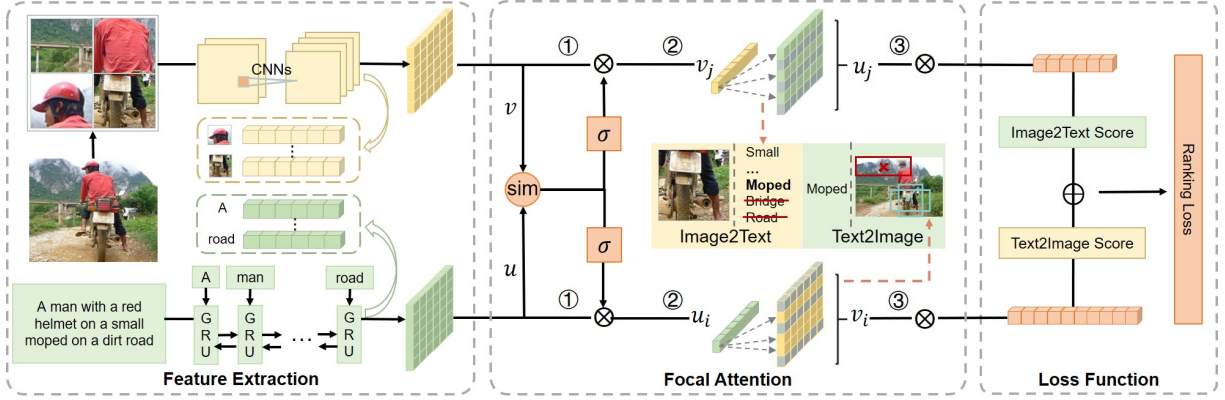
Figure 2: The overall framework of BFAN that consists of feature extraction, focal attention and loss function module. The focal attention module takes the extracted feature as input, and then attends to regions and words interactively. Specifically, ①it preassigns attention to all the fragments in one modality based on fragments from another modality; ②it identifies partial relevant parts based on internal relationship of fragments within the same modality; ③it reassigns the focal attention at image2text and text2image directions, which is then jointly integrated to be optimized by pair-wise ranking loss.

optimizing pair-wise similarities. He et.al [15] extend this idea and yield appealing performance. A stacked cross attention is designed to dynamically associate shared semantics by attending to words with respect to regions or attending to regions with respect to words, which makes many-to-many matching more interpretable as it changes the importance of target fragments based on fragments from another modality. Similar works are proposed in [9]. The attention mechanism they employ is within a fixed pattern, where the representation of shared semantics is a weighted combination over all the image regions or text words according to their contribution to shared semantics. However, only a fraction of regions or words relevant to shared semantics, integrating all of them will disturb the target semantic and thus lead to semantic misalignment. In this work, we address this issue by proposing a novel focal attention that eliminates irrelevant regions/words from shared semantics.

3 METHOD

The overall framework of our BFAN is illustrated in Figure 2. It consists of three components: feature extraction, focal attention and objective function. In this section, we first summarize the general attention framework in image-text matching, analyzing the semantic misalignment problem caused by existing framework in section 3.1. Then, we introduce our proposed focal attention and how to employ it into text-to-image and image-to-text matching, describing why and how to integrate them together in section 3.2. Last, we detail the objective function and feature extraction of our BFAN in section 3.3 and 3.4, respectively.

3.1 General Attention Framework

Without loss of generality, given an image-text pair consists of m text words and n image regions, a general image-text matching is to first project each image region and text word into a common d-dimensional space using deep neural networks, getting text representation $u \in \mathbb{R}^{m \times d}$ and image representation $v \in \mathbb{R}^{n \times d}$, and then learn to associate shared semantics in image and text using neural network blocks. The shared semantics are composed of multiple

local shared semantics, such as local regions and words, and thus the overall objective is to maximumly improve the relevance of each local shared semantic:

$$R(u, v) = \frac{1}{K} \sum_k R(S_k^u, S_k^v) \quad (1)$$

where S_k^u and S_k^v denote k-th shared semantic selected from image and text, respectively. $R(\cdot)$ denotes the relevance of shared semantics, which is computed using a similarity metric. K denotes the number of shared semantics.

Existing attention methods find shared semantics by learning the attention distribution over all the fragments, e.g. regions or words, where fragments relevant to shared semantics obtain higher attention than others. Then, all the fragments are aggregated to represent shared semantics using a weighted combination:

$$S_k^u = \sum_{i=1}^m w_{ik} u_i, S_k^v = \sum_{j=1}^n w_{jk} v_j \quad (2)$$

where w_{ik} and w_{jk} are attention distribution with respect to k-th shared semantic, u_i and v_j denote i-th word and j-th region, respectively. However, not all the fragments support the specific shared semantic as many of them are irrelevant to it, the shared semantic will be more or less disturbed by irrelevant fragments if they are aggregated. More seriously, it will lead to semantic misalignment since different semantics cannot be appropriately decoupled. Therefore, it is necessary to represent the shared semantics by integrating a subset of fragments that are relevant to the target semantic.

3.2 Our Focal Attention

To address the semantic misalignment problem caused by the general attention framework, our focal attention proposes to learn a scoring function F to identify fragments relevant to shared semantics, through which irrelevant fragments can be removed from shared semantics. Here, we set fragments with scores greater than zero as relevant, that is:

$$H(x) = \mathbb{I}(F(x) > 0) \quad (3)$$

where $\mathbb{I}(\cdot)$ is an indicator function, x can be either regions or words.

It is impractical to find a fixed margin between relevant and irrelevant fragments based on the absolute value, e.g. similarity value between fragments and shared semantics, because it depends on iteratively updated fragment representations. Some attention approaches [20] attend to local fragments by simply masking fragments based on their position, but there is no connection with semantic relevance and fragment position. Inspired by non-local blocks proposed in [29], we determine the relevance of fragments by computing the relative importance of them to other fragments. The intuition behind this operation is irrelevant fragments always obtain low importance to the shared semantic compared with other relevant fragments. The scoring function is formulated as:

$$F(x_i) = \sum_{\forall x} f(x_i, x_j)g(x_j) \tag{4}$$

The pairwise function $f(x_i, x_j)$ computes relative importance of target i-th fragment to j-th fragment, and $g(x_j)$ denotes the confidence of the fragment being compared, followed by an operation that sums up the weighted comparison results with all the other fragments. A fragment can be considered as relevant if it is similar to other relevant fragments with high confidence scores. Then, the k-th shared semantic can be simply defined as:

$$S_k^x = \sum_{\forall x} w_{ik} x_i H(x_i) \tag{5}$$

In this work, our goal is to eliminate irrelevant fragments from context, which is totally different from traditional attention methods that focus on learning attention weight. In addition, differs from hard attention [33] that estimates gradient using random sampling, the focal attention can compute gradient directly, because fragment except for irrelevant ones contribute to the forward-propagation. This allows for training the network both efficiently and effectively. We will depict how to employ the focal attention into text-to-image and image-to-text matching in 3.2.1 and 3.2.2.

3.2.1 Text-to-Image Focal Attention. ==In this work, we find shared semantics in image and text by fixing one modality and finding relevant fragments from another modality==, where fragments in fixed modality are considered as the shared semantic. For text-to-image direction, text words are fixed as the shared semantic, we need to find relevant image regions for each text word. The overall framework includes three steps: preassign attention, identify relevant regions and reassign attention. To be specific, we first preassign attention score for each region, it is implemented by computing cosine similarities between regions and words, and normalizing them using softmax activation:

$$w_{ij} = \sigma(\alpha \frac{u_i^T v_j}{\|u_i\| \|v_j\|}), i \in [1, m], j \in [1, n]. \tag{6}$$

where σ denotes softmax activation. α is a scaling factor to further increase the gap between relevant and irrelevant regions, which is set as 20 in our implementation.

Second, we identify relevant regions by scoring each region based on its allocated attention relative to other regions, regions with

scores greater than zero are relevant regions, otherwise irrelevant.

$$F(v_{ij}) = \sum_{t=1}^{n} f(v_{ij}, v_{it})g(v_{it}) \tag{7}$$

We set $f(v_{ij}, v_{it})$ as the difference of their preassigned attention to determine the relative attention of j-th region to t-th region since they are scalar value. The confidence score for t-th region being compared is set as its relevance to the i-th query word, e.g. $\sqrt{w_{it}}$. Alternatively, we also set confidence of each image region as equal, it also proves to be effective in section 4.3.

After that, relevant regions can be selected by element-wise product between each image region and function H. Third, we reassign attention weights for these selected relevant regions by renormalization. Note that irrelevant regions will not contribute to this process as their scores are zero:

$$w'_{ij} = \frac{w_{ij} H(v_{ij})}{\sum_{j=1}^{n} w_{ij} H(v_{ij})}. \tag{8}$$

The reassigned attention weights will replace conventional attention w_{ik} in equation 2, which allows to focus on most relevant regions as attention weights for irrelevant regions are zero. The shared semantic selected from the image based on i-th word is computed as $v'_i = \sum_{j=1}^{n} w'_{ij} v_j$. The global relevance of image and text is formulated as:

$$R(u, v) = \frac{1}{m} \sum_{i=1}^{m} R(u_i, v'_i). \tag{9}$$

3.2.2 Image-to-Text Focal Attention. Analogously, image regions are fixed as the shared semantic in image-to-text direction, we need to find relevant text words for each region. To this end, we first preassign attention by computing the similarity score between each image region and text word using cosine similarity, and normalize similarity scores of each word with respect to query region into [0,1] using softmax, attention on i-th word is denoted as w_{ji}. During this process, relevant words can be paid more attention, but irrelevant words also contribute to shared semantics between image region and target text. The second step is to score each word based on its preassigned attention relative to other words, that is:

$$F(u_{ji}) = \sum_{t=1}^{m} f(u_{ji}, u_{jt})g(u_{jt}) \tag{10}$$

Next, the indicator function $H(u_{ji})$ is applied to identify relevant words based on computed score, where relevant words are set as one, otherwise zero. The attention for relevant words will be reassigned

$$w'_{ji} = \frac{w_{ji} H(u_{ji})}{\sum_{i=1}^{m} w_{ji} H(u_{ji})}. \tag{11}$$

The reassigned attention will be paid to all relevant words using element-wise product with their representations in d-dimensional space. The shared semantic with j-th region is selected from the text, computed as a weighted combination of relevant words $u'_j = \sum_{i=1}^{m} w'_{ji} u_i$, where the learned focal attention determines the weight. The local relevance score $R(v_j, u'_j)$ can be computed through cosine similarity. The global relevance score for image and text is

Table 1: Comparison results with baselines on Flickr30K. Image-to-Text denotes retrieve texts using image query, and Text-to-Image denotes retrieve images using text query. The best results are in bold.

Method	Image-to-Text			Text-to-Image			
	Recall@1	Recall@5	rmean	Recall@1	Recall@5	rmean	rsum
Deep Fragment (single) [13]	16.4	40.2	28.3	10.3	31.4	20.9	98.3
HM-LSTM (single) [24]	38.1	-	38.1	27.7	-	27.7	65.8
sm-LSTM (ensemble) [10]	42.5	71.9	57.2	30.2	60.4	45.3	205.0
BSSAN (single) [9]	44.6	74.9	59.8	33.2	62.6	47.9	215.3
VSE++ (single) [5]	52.9	-	52.9	39.6	-	39.6	92.5
DANs (single) [23]	55.0	81.8	68.4	39.4	69.2	54.3	245.4
SCO (single) [11]	55.5	82.0	68.8	41.1	70.5	55.8	249.1
SCAN (single) [15]	67.9	89.0	78.5	43.9	74.2	59.1	275.0
SCAN (ensemble)	67.4	90.3	78.9	48.6	77.7	63.2	284.0
Ours:							
BFAN-prob (single)	65.5	89.4	77.5	47.9	77.6	62.8	280.4
BFAN-equal (single)	64.5	89.7	77.1	48.8	77.3	63.1	280.3
BFAN-prob+equal (ensemble)	**68.1**	**91.4**	**79.8**	**50.8**	**78.4**	**64.6**	**288.7**

Table 2: Comparison results with baselines on MSCOCO. Image-to-Text denotes retrieve texts using image query, and Text-to-Image denotes retrieve images using text query. The best results are in bold.

Method	Image-to-Text			Text-to-Image			
	Recall@1	Recall@5	rmean	Recall@1	Recall@5	rmean	rsum
HM-LSTM (single) [24]	43.9	-	43.9	36.1	-	36.1	80.0
sm-LSTM (ensemble) [10]	53.2	83.1	68.2	40.7	75.8	58.3	252.8
BSSAN (single) [9]	56.0	82.6	69.3	41.8	76.7	59.3	257.1
VSE++ (single) [5]	64.6	-	64.6	52.0	-	52.0	116.6
GXN (single) [7]	68.5	-	68.5	56.6	-	56.6	125.1
SCO (single) [11]	69.9	92.9	81.4	56.7	87.5	72.1	307.0
SCAN (single) [15]	70.9	94.5	82.7	56.4	87.0	71.7	308.8
SCAN (ensemble)	72.7	94.8	83.8	58.8	88.4	73.6	314.7
Ours:							
BFAN-prob (single)	73.0	94.8	83.9	58.0	87.6	72.8	313.4
BFAN-equal (single)	73.7	94.9	84.3	58.3	87.5	72.9	314.4
BFAN-prob+equal (ensemble)	**74.9**	**95.2**	**85.1**	**59.4**	**88.4**	**73.9**	**317.9**

calculated as the averaging of local relevance scores, that is:

$$R(v, u) = \frac{1}{n} \sum_{j=1}^{n} R(v_j, u_j^{'}). \quad (12)$$

3.2.3 Bidirectional Focal Attention. Focal attention on text-to-image and image-to-text are independent modules, where text-to-image focal attention learns to pick out a subset of image regions that semantically similar to each word, and image-to-text focal attention learns to pick out a subset of text words that semantically similar to each region. If we employ the focal attention to one direction, it will result in the preference to long text or complex image. It is because long text or complex image contains more information, and thus is more possible to get high response to query. Therefore, we present to jointly apply focal attention in two directions by combining their relevance as the overall relevance score. The bidirectional network will maximally associate image-text pairs as it considers the semantic overlap instead of intersection in image and text. Specifically, we compute global relevance score between image and text in two directions separately, and then integrate them by taking their sum as the final score of image-text pairs, such that

both directions contribute to the final relevance score:

$$S_{uv} = R(u, v) + R(v, u). \quad (13)$$

Despite that recent approach [9] also employ bidirectional attention, ours are totally different. It derives from they simultaneously restraint scores at each direction while we restraint the overall score. This can relax constraint and avoid overfitting, because similar samples (one of them is long/complex) cannot be distinguished in single direction, it is inappropriate to restraint each direction.

3.3 Objective Function

To optimize the proposed network, we employ a structured ranking loss as the objective function, which has been proven to be able to maximize relevance scores of relevant image-text pairs and minimize that of irrelevant text-image pairs. Motivated by previous approach proposed by [5], we focus on hard negatives in each mini-batch, which produces maximum relevance score over any other irrelevant pairs. Given a pair of relevant image-text, we denote their relevance score as S_{uv}, $\bar{v} = \arg\max_{t \neq v} S_{ut}$ denotes the hard negative when using the text to retrieve image, and $\bar{u} = \arg\max_{t \neq u} S_{tv}$ denotes the hard negative when using the image to retrieve text,

Table 3: Ablation studies on Flickr30K, the best results are in bold.

Method	Image-to-Text			Text-to-Image			
	Recall@1	Recall@5	rmean	Recall@1	Recall@5	rmean	rsum
BFAN-w/o-t2i	60.4	85.4	72.9	46.3	76.5	61.4	268.6
BFAN-w/o-i2t	63.0	87.2	75.1	45.9	75.0	60.5	271.1
BFAN-w/o-focal	63.2	88.8	76.0	48.7	76.9	62.8	277.6
BFAN-prob	65.5	89.4	77.5	47.9	77.6	62.8	280.4
BFAN-equal	64.5	89.7	77.1	48.8	77.3	63.1	280.3
BFAN-prob+equal	**68.1**	**91.4**	**79.8**	**50.8**	**78.4**	**64.6**	**288.7**

Table 4: Ablation studies on MSCOCO, the best results are in bold.

Method	Image-to-Text			Text-to-Image			
	Recall@1	Recall@5	rmean	Recall@1	Recall@5	rmean	rsum
BFAN-w/o-focal	65.8	91.9	78.9	43.6	79.3	61.5	280.6
BFAN-w/o-i2t	69.3	93.7	81.5	55.2	85.6	70.4	303.8
BFAN-w/o-t2i	70.3	93.9	82.1	55.6	86.5	71.1	306.3
BFAN-prob	73.0	94.8	83.9	58.0	87.6	72.8	313.4
BFAN-equal	73.7	94.9	84.3	58.3	87.5	72.9	314.4
BFAN-prob+equal	**74.9**	**95.2**	**85.1**	**59.4**	**88.4**	**73.9**	**317.9**

their relevance score with the text or image are forced to be lower than that between relevant image-text pairs by a fixed margin, i.e.

$$L = [\alpha - S_{uv} + S_{\bar{u}v}]_+ + [\alpha - S_{uv} + S_{u\bar{v}}]_+. \quad (14)$$

where, $[x]_+ \equiv max(x, 0)$, we set the loss as zero if relevance score with hard negative is not as large as that with relevant pairs. The margin α is a hyperparameter that is set as 0.2.

3.4 Feature Extraction

3.4.1 Image Feature. In many-to-many image-text matching, each image is comprised of multiple regions. We detect salient regions that contribute most to the global semantic, and encode each of them into feature vectors. In this work, we detect salient regions using a popular object detection tool Faster R-CNN [26]. The tool predicts object bounding boxes and scores them. We select top K (K=36) salient objects according to their scores, and extract mean-pooled convolutional features for these bounding boxes using pretrained ResNet-101 [8]. A fully-connected layer is applied to transform features into target d-dimensional feature vector.

3.4.2 Text Feature. Similar to the image, each text contains a set of words, we encode each word into d-dimensional feature vectors as well as image region and combine them as the global feature of the text. To this end, we employ the bidirectional GRU to integrate the feedforward and backward contextual information into word representations. Specifically, we first split a text into multiple words, and embed each word into a low-dimensional vector to decrease the computation cost of GRU, which are then fed into bidirectional GRU. After multi-step iterations, the average of forward and backward hidden state can be considered as the text representation, which contains d-dimensional features for each word in the text.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. We conduct several experiments on image-text matching benchmarks, Flickr30K [25] and MSCOCO [17]. Flickr30K

is a standard dataset for image-text matching, it contains 31,000 images and 155,000 texts in total, each image relates to five texts. Following [13, 18, 22], we split Flickr30K benchmark into 29K training images, 1K validation images and 1K testing images. MSCOCO is a large-scale benchmark that contains 123,287 images with five texts each. We use 113,287 images for training, 5,000 images for validation and 5,000 for testing follow [5, 15]. We report experimental results through averaging 5-folds on 1K test images.

4.1.2 Evaluation. We evaluate the performance of our proposed approach by reporting Recall@K (K = 1,5) values for both image-to-text and text-to-image matching task. The Recall computes the proportion of correct image or text being retrieved among top K results. In addition, we compute mean value of Recall (rmean) in each direction, and sum of Recall (rsum) to show overall performance.

4.1.3 Settings. The proposed network is implemented using PyTorch, and trained on 1 NVIDIA TITAN Xp optimized by Adam. We start training the network with learning rate 0.0002 on Flickr30K and 0.0005 on MSCOCO, and decay by 0.1 after every 10 epochs. The mini-batch size is set as 32. Our network requires to train 15 epochs on Flickr30K and 20 epochs on MSCOCO, training instances are randomly shuffled at each epoch. We set the dimensionality of image region representations as 1024. The initial one-hot vector of word embeddings are covert to 300-dimensional, and then fed into bidirectional GRU that produces 1024-dimensional representations.

4.1.4 Baselines. We select most representative works as baselines, including the first many-to-many approach Deep Fragment [13], recent works Hierarchical Multimodal LSTM (HM-LSTM) [24], Selective Multimodal LSTM (sm-LSTM) [10], Bi-Directional Spatial-Semantic Attention Networks (BSSAN) [9] and state-of-the-art Stacked Cross Attention Network (SCAN) [15]. We also make comparisons with most recent one-to-one matching works, including Dual Attention Networks (DANs) [23], visual-semantic embeddings (VSE++) [5], semantic-enhanced image and sentence matching model (SCO) [11] and generative cross-modal feature learning

Figure 3: Visualization of our focal attention and conventional attention [15] with respect to each word shown at the top left corner of each image, where brighter regions obtain more attention. Relevant and irrelevant regions are outlined in yellow and red boxes, respectively, which shows our attention always focus on relevant regions while [15] distracts attention as it attends to many irrelevant regions in addition to relevant ones.

framework (GXN) [7]. We provide two versions of focal attention implementation, including BFAN-prob and BFAN-equal, where one takes confidence of each compared fragment into account, and another one treats each fragment equally. Note that some approaches use ensemble model by averaging the global relevance score of two single models, we also provide single and ensemble model to make a fair comparison, it is achieved by averaging relevance scores calculated by single models.

4.2 Comparison Results

We conduct extensive experiments on Flickr30K and MSCOCO, respectively. Quantitative results on Flickr30K are listed in Table 1. In real application, top-1 result is more concerned by users, so improving Recall@1 is crucial to improve user experience, this is exactly advantage of focal attention. It is observed that our approach achieves more improvement on Recall@1 than other metrics. The BFAN achieve 68.1% and 50.8% Recall@1 value on image-to-text and text-to-image matching, respectively. It is the first time that Recall@1 in text-to-image matching over 50% on Flickr30K benchmark, getting a 2.2% relative improvement than state-of-the-art SCAN. Compared with VSE++, which also optimizes on hard negatives, we can obtain relative Recall@1 gains with 15.2% and 11.2%. Although BSSAN proposes similar bidirectional networks, our BFAN outperforms it with over 18% relative gains on average since relevant fragments are tightly correlated without interference of irrelevant ones. Compared with two most effective one-to-one methods, SCO and GXN, our approach not only outperforms them,

but also learns more fine-grained region-word correspondence, which is significant for real multimodal application.

Quantitative results on MSCOCO are listed in Table 2. MSCOCO is a larger image-text matching benchmark, our improvement on MSCOCO shows the proposed approach has excellent and stable capability of generalization. Our single model outperforms state-of-the-art single model with a relative 5.3% ~5.4% gain in terms of rsum, our ensemble model also outperforms the best ensemble model. Note that our BFAN achieves more improvement on Recall@1, which is significant for image-text matching.

4.3 Ablation study

Table 3 shows ablation study results on Flickr30K. Both focal attention and bidirectional version contribute towards the overall performance. To evaluate the effect of focal attention, we remove focal attention in our full model, and employ traditional attention on both image-to-text and text-to-image directions, denoted as BFAN-w/o-focal. Focal attention proves to be critical to improving overall matching performance, especially for Recall@1 as it removes most irrelevant fragments. To evaluate the effect of bidirectional focal attention, we employ focal attention in either image-to-text or text-to-image direction, referred as BFAN-w/o-t2i and BFAN-w/o-i2t. Results show that the single directional focal attention will decrease all the Recall value by nearly 2% on average compared with full single model. It derives from the single model is partial to long text and complex image as they are more likely to contain target fragments. Our bidirectional attention avoids this by considering the

proportion of relevant fragments instead of their absolute quantity. In addition, we also investigate the effect of our focal attention with different implementation, i.e. BFAN-prob and BFAN-equal. Both of them can achieve great performance, and combining them can largely improve the Recall value. It is because their combination can learn a better mode of relevant fragments selection. Ablation study results on MSCOCO benchmark is shown in Table 4. Different from the results on Flickr30K, BFAN-w/o-focal achieves better performance than BFAN-w/o-t2i and BFAN-w/o-i2t, the two full single models still outperform other models at all the evaluation metrics, which shows focal attention and bidirectional version complement to each other, and can be stably applied to different datasets.

4.4 Attention Visualization

To better understanding the difference in focal attention and conventional attention, we visualize attention weights for each image region with respect to query word in Figure.3. We make comparisons with the baseline model [15], attention weight of each bounding box (image region) released by bottom-up attention [1] is computed using BFAN and baseline, respectively. We use the brightness to visualize attention weights, and brighter regions obtain more attention. We show attention distribution in image regions with respect to nouns and verbs in the text, where the chosen word is shown at the top left corner of each image. Relevant and irrelevant regions are mainly outlined in yellow and red boxes. It is observed that BFAN learns better semantic alignment. For example, in Q1, "Standing" corresponds to the gigot by our BFAN, while baseline also aligns it with irrelevant regions, like sky and grass.

4.5 Qualitative Results

We also provide visualization for text-to-image and image-to-text matching. For text-to-image matching shown in Figure 4, we show top 3 ranked images for each text query. Images in first three columns are retrieved by our approach, and the last three columns are by baseline [15]. The correctly retrieved images are in green box. Long and short text queries can be well matched with their most relevant images. For the first text query, the correct image ranks first by our BFAN despite local regions in the second image hit some keywords, such as "black snow pants" and "wearing a black coat". Baseline model gives incorrect ranks since irrelevant regions disturb the semantic alignment. For example, they will attend to irrelevant "red coat" when matching "black coat", which will lower the response to query word, but the incorrect image will give a high response since most people wear "black coat".

We visualize image-text matching performance in Figure 5. The ground truth (GT), top-1 ranked text produced by our approach and baseline [15] are listed at the right-hand of each image query, where correct results are marked as green. As shown in the first example, baseline gets an incorrect result as it always attends to keyword "water", and thus it plays an important role while querying other objects like the person and action "flip". It further confirms the necessity of using focal attention. Results also show that our approach can capture and discriminate more detailed information. For example, for the last example, baseline gets the wrong answer since it cannot identify "cookie" and "highchair" despite most other keywords match the query image, while the BFAN performs well.
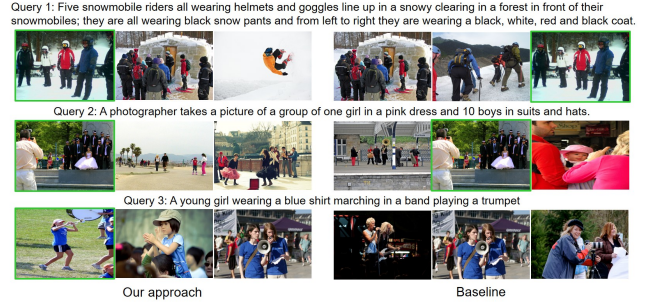

Figure 4: Text-to-image matching by our approach and baseline [15]. For each text query, we list top-3 ranked images from left to right, where correct answers are outlined as green box. The first three columns are our results and the last three columns are baseline results.
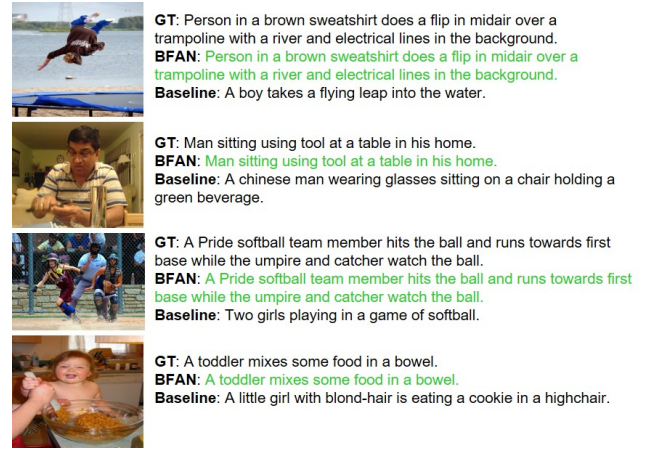

Figure 5: Image-to-text matching by our approach and baseline [15]. For each image query, we provide the ground truth (GT), top-1 ranked text by BFAN and baseline at the right-hand of the image, where correct ones are marked as green.

5 CONCLUSION

In this paper, we propose a novel bidirectional focal attention model for image-text matching. Different from conventional attention, our focal attention only attends to fragments relevant to query fragment, which can address semantic misalignment caused by existing attention methods. The directional version can also avoid the preference to long text or complex image. We conduct comprehensive experiments that demonstrate the proposed method can significantly outperform state-of-the-art. Future research directions include applying the focal attention into other cross-modal applications such as translation, image caption and visual question answering.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2018), 423–443.

[4] A. Eisenschtat and L. Wolf. 2017. Linking Image and Text with 2-Way Nets. In *CVPR*. 1855–1865.

[5] Fartash Faghri, David J. Fleet, Jamie Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*.

[6] Haoqi Fan and Jiatong Zhou. 2018. Stacked Latent Attention for Multimodal Reasoning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1072–1080.

[7] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. 2017. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. *CoRR* abs/1711.06420 (2017).

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778.

[9] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, and Zhoujun Li. 2019. Bi-Directional Spatial-Semantic Attention Networks for Image-Text Matching. *IEEE Transactions on Image Processing* 28, 4 (2019), 2008–2020.

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

[11] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6163–6171.

[12] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[13] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. 3 (2014), 1889–1897.

[14] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *31st International Conference on Machine Learning, ICML 2014* 3 (11 2014).

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[16] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-Aware Textual-Visual Matching with Latent Co-attention. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1908–1917.

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

[18] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew. 2017. Learning a Recurrent Residual Fusion Network for Multimodal Matching. In *ICCV*. 4127–4136.

[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. *CoRR* abs/1606.00061 (2016).

[20] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Computer Science* (2015).

[21] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In *ICCV*. 2623–2631.

[22] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep Captioning with Multimodal Recurrent Neural Networks. *CoRR* abs/1412.6632 (2014).

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

[24] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding. In *ICCV*. 1899–1907.

[25] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *ICCV* (2015), 2641–2649.

[26] S. Ren, K. He, R Girshick, and J. Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[27] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[28] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2019. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 394–407.

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7794–7803.

[30] Jonatas Wehrmann and Rodrigo C. Barros. 2018. Bidirectional Retrieval Made Simple. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7718–7726.

[31] Yiling Wu, Shuhui Wang, and Qingming Huang. 2017. Online asymmetric similarity learning for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4269–4278.

[32] Yiling Wu, Shuhui Wang, and Qingming Huang. 2018. Learning Semantic Structure-preserved Embeddings for Cross-modal Retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 825–833.

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[34] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[36] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4709–4717.

[37] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.

[38] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding with Instance Loss. *arXiv preprint arXiv:1711.05535* (2017).