This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

More Than Just Attention: Improving Cross-Modal Attentions with Contrastive Constraints for Image-Text Matching

Yuxiao Chen[1] *, Jianbo Yuan[2], Long Zhao[1], Tianlang Chen[2], Rui Luo[2], Larry Davis[2], Dimitris N. Metaxas[1]

[1]Rutgers University, [2]Amazon.com Services, Inc

Abstract

Cross-modal attention mechanisms have been widely applied to the image-text matching task. They have achieved remarkable improvements thanks to their capability of learning fine-grained relevance across different modalities. However, the cross-modal attention models of existing methods could be sub-optimal and inaccurate because there is no direct supervision provided during the training process. In this work, we propose two novel training strategies, namely Contrastive Content Re-sourcing (CCR) and Contrastive Content Swapping (CCS) constraints, to address such limitations. These constraints supervise the training of cross-modal attention models in a contrastive learning manner without requiring explicit attention annotations. They are plug-in training strategies and can be generally integrated into existing cross-modal attention models. Additionally, we introduce three metrics, including Attention Precision, Recall, and F1-Score, to quantitatively measure the quality of learned attention models. We evaluate the proposed constraints by incorporating them into four state-of-the-art cross-modal attention-based image-text matching models. Experimental results on both Flickr30k and MS-COCO datasets demonstrate that integrating these constraints generally improves the model performance in terms of both retrieval performance and attention metrics.

1. Introduction

The task of image-text matching aims to learn a model that measures the similarity between visual and textual contents. By using the learned model, users can retrieve images that visually match the context described by a text query, or retrieve texts that best describe the image query. Because of its critical role to bridge the human vision and language world, this task has emerged as an active research area [5, 17, 8, 12, 16, 1, 23].

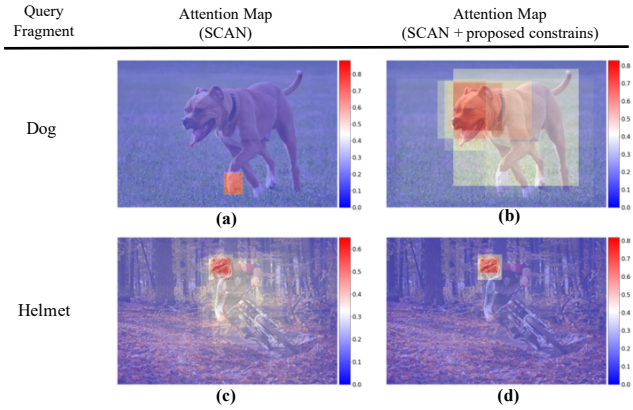


Figure 1: Visualization of the attention maps of the SCAN model learned without and with our proposed constraints.

Recently, cross-modal attention models have been widely applied to this task [16, 12, 17, 8, 7, 2, 13, 4]. These approaches have achieved remarkable improvements thanks to their ability to capture fine-grained cross-modal relevance by the cross-modal attention mechanism. Specifically, given an image description and its corresponding image, they are first represented by fragments, i.e., individual words and image regions. We refer to the fragments of the context modality as query fragments, and the fragments of the attended modality as key fragments. Given a query fragment, a cross-modal attention model first assigns an attention weight to each key fragment, each of which measures the semantic relevance between the query fragment and the corresponding key fragment. Then the attended information of the query fragment is encoded as the weighted sum of all key fragment features. The similarity between each query fragment and its attended information is thus aggregated as the similarity measurement between the query and the retrieval candidates.

In ideal cases, well-trained cross-modal attention mod-

*This work was done while Yuxiao Chen was a research intern at Amazon. Correspondence to: Yuxiao Chen (yc984@cs.rutgers.edu)

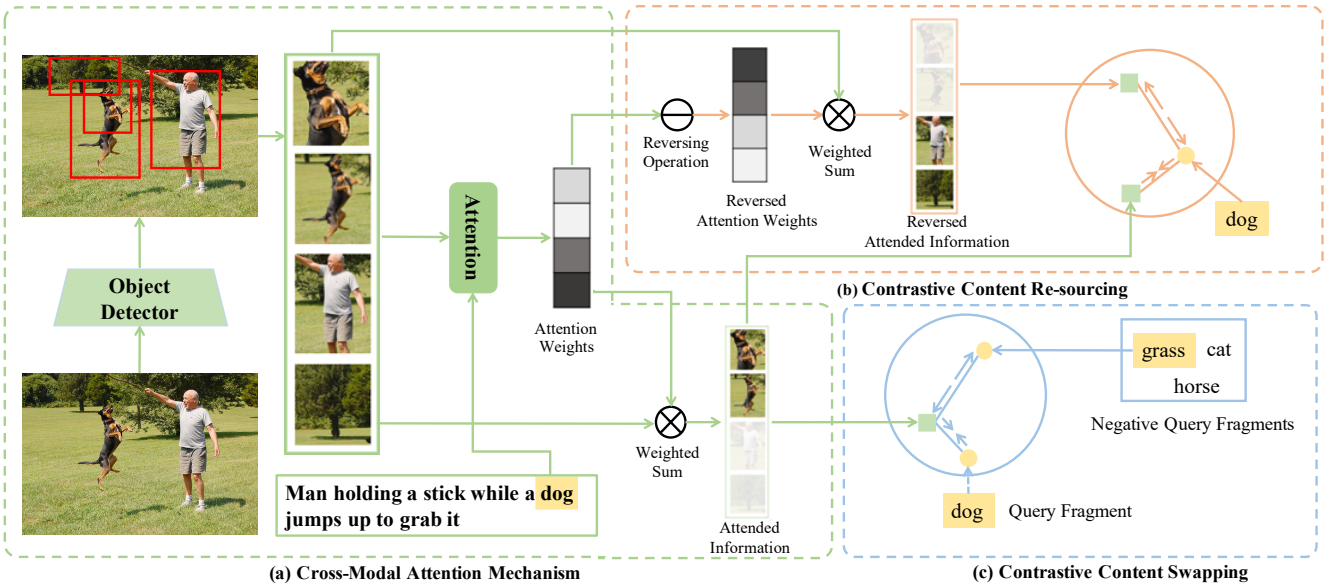


Figure 2: Overview of the training pipeline which contains **(a)** the cross-modal attention mechanism and our proposed attention constraints including **(b)** Contrastive Content Re-sourcing (CCR) and **(c)** Contrastive Content Swapping (CCS).

els will attend to the semantically relevant key fragments by assigning large attention weights to them, and ignore irrelevant fragments by producing small attention weights. Take Figure 1 (b) as an example: when "dog" is used as a query fragment, the cross-modal attention model is supposed to output large attention weights for all image regions containing the dog, and small attention weights for other irrelevant image fragments. However, since the cross-modal attention models of most existing image-text matching methods are trained in a purely data-driven manner and do not receive any explicit supervision or constraints, the learned attention models may not be able to precisely attend to the relevant contents. As shown in Figure 1 (a), the learned SCAN model [12], a state-of-the-art cross-modal attention-based image-text matching model, fails to attend to the relevant image regions containing the dog's main body when using the word "dog" as the query fragment. This example illustrates a false negative case, i.e., a low attention "recall". Additionally, a learned cross-modal attention model might also suffer from false positives (low attention "precision"). As shown in Figure 1 (c), when using "helmet" as the query fragment, the SCAN model assigns large attention weights to the irrelevant human body and background areas. A possible solution to these limitations is to rely on manually generated attention map ground truth to supervise the training process of cross-modal attention models [19, 27]. However, annotating attention distributions is an ill-defined task, and will be labor-intensive.

To this end, we propose two learning constraints, namely **Contrastive Content Re-sourcing (CCR)** and **Contrastive Content Swapping (CCS)**, to supervise the training process of cross-modal attentions. Figure 2 gives an overview of our method. CCR enforces a query fragment to be more relevant to its attended information than to the reversed attended information, which is generated by calculating the weighted sum of key fragments using reversed attention weights (details in Section 3.2). It can guide a cross-modal attention model to assign large attention weights to the relevant key fragments and small weights to irrelevant fragments. On the other hand, CCS further encourages a cross-modal attention model to ignore irrelevant key fragments by constraining the attended information to be more relevant to the corresponding query fragment than to a negative query fragment. In the example shown in Figure 2 (c), by using the word "grass" as a negative query fragment, the attention weights assigned to regions containing grass will be diminished so that a more accurate attention map is generated. The proposed constraints are plug-in training strategies that can be easily integrated into existing cross-modal attention-based image-text matching models.

We evaluate the performance of the proposed constraints by incorporating them into four state-of-the-art cross-modal attention-based image-text matching networks [12, 16, 23, 4]. Additionally, in order to quantitatively compare and measure the quality of the learned attention models, we propose three new attention metrics, namely **Attention Precision**, **Attention Recall** and **Attention F1-Score**. The experimental results on both MS-COCO [14] and Flickr30K [25] demonstrate that these constraints significantly improve image-text matching performances and

the quality of the learned attention models.

To sum up, the main contributions of this work include: (i) we propose two learning constraints to supervise the training of cross-modal attention models in a contrastive manner without requiring additional attention annotations. They are plug-in training strategies and can be easily applied to different cross-modal attention-based image-text methods; (ii) we introduce the attention metrics to quantitatively evaluate the quality of learned attention models, in terms of precision, recall, and F1-Score; (iii) we validate our approach by incorporating it into four state-of-the-art attention-based image-text matching models. Extensive experiments conducted on two publicly available datasets demonstrate its strong generality and effectiveness.

2. Related Work

Image-Text Matching. The task of image-text matching is well-explored yet challenging. Its main challenge is how to measure the similarity between texts and images. Early approaches propose to measure the similarity at the global level [10, 6, 26, 5]. Specifically, these methods first train an image encoder and a text encoder to embed the global information of images and sentences into feature vectors, and then measure the similarity between images and sentences by calculating the cosine similarity between the corresponding feature vectors. For example, by using the triplet ranking loss with hard negative samples, Faghri *et al*. [5] train a VGG-based image encoder [21] and a GRU-based text encoder [3], respectively. One major limitation of these methods is that they failed to capture fine-grained image-text relevance.

To address this limitation, recent studies propose to apply the cross-modal attention mechanism to measure the similarity between texts and images at the fragment level [16, 12, 23, 24]. Typically, given an image and a sentence, these methods first extract embeddings on object regions from the image by feeding it into an object detection model, such as Faster R-CNN [20], and embed each word of the sentence by using recurrent neural networks. Then the relevant regions of each word and the relevant words of each region are inferred by leveraging the text-to-image and image-to-text attention, respectively. The similarity between each fragment (word or image region) and its relevant information is calculated and aggregated as the final similarity score between the image and sentence. Although these methods have achieved notable results, the learning process of these cross-modal attention models could be sub-optimal due to the lack of direct supervision, as discussed in Section 1.

Supervision on Learning Cross-Modal Attention. The task of training cross-modal attention models with proper supervision has drawn growing interests. The main challenge lies in how to define and collect supervision signals. Qiao *et al*. [19] first train an attention map generator on a human annotated attention dataset and then apply the attention map predicted by the generator as weak annotations. Liu *et al*. [15] leverage human annotated alignments between words and corresponding image regions as supervision. Similar to [15], image local region descriptions and object annotations in Visual Genome [11] are used for generating attention supervision [27]. These methods obtain attention supervision from different forms of human annotations, such as word-image correspondence and image local region annotations. By contrast, we provide attention supervision by constructing pair-wise samples in a contrastive learning manner which does not require additional manual attention annotations.

3. Methodology

3.1. Cross-Modal Attention Model

Given an image-sentence pair in image-text matching, they are first represented as fragments, i.e., individual words and image regions. The fragments of the context modality are query fragments, and the fragments of the attended modality are key fragments. Each of these fragments is encoded as a vector. A cross-modal attention model takes these vectors as input, and infers the cross-modal relevance between each query fragment and all key fragments. The similarity score of the image-sentence pair is then calculated according to the obtained cross-modal relevance.

Let q_i and k_j refer to the feature representation of the i-th query and j-th key fragments, respectively. The cross-modal attention model first calculates k_j's attention weight with respect to q_i as follows:

$$\begin{aligned} e_{i,j} &= f_{att}(q_i, k_j), \\ w_{i,j} &= \frac{\exp(e_{i,j})}{\sum_{j \in K} \exp(e_{i,j})}, \end{aligned} \tag{1}$$

where f_{att} is the attention function whose output is a scalar $e_{i,j}$ that measures the cross-modal relevance between q_i and k_j; K is a set of indexes of all key fragments; $w_{i,j}$ is k_j's attention weight with respect to q_i.

q_i's **attended information** (i.e., q_i's relevant cross-modal information) is defined as the **attention feature** a_i using the weighted sum of key fragment features in the following equation:

$$a_i = \sum_{j \in K} \left(w_{i,j} \cdot k_j \right). \tag{2}$$

The similarity score between the image I and the sentence T is then defined as:

$$S(I, T) = AGG_{i \in Q}(Sim(q_i, a_i)), \tag{3}$$

where Q denotes the set of indexes of all query fragments; Sim is the similarity function; AGG is a function that aggregates similarity scores among all query fragments, such as the average pooling function [12].

The most widely used loss function for this task is the triplet ranking loss with hard negative sampling [5] defined as:

$$\ell_{rank} = [S(I,\hat{T}) - S(I,T) + \gamma_1]_+ + [S(\hat{I},T) - S(I,T) + \gamma_1]_+, \quad (4)$$

where γ_1 controls the margin of similarity difference; the matched image I and the sentence T form a positive sample pair, while $\hat{T}$ and $\hat{I}$ represent the hardest negative sentence and image for the positive sample pair as defined by [5]. ℓ_{rank} enforces the similarity between the anchor image I and its matched sentence T to be larger than the similarity between the anchor image and an unmatched sentence by a margin γ_1. Vice versa for the sentence T.

However, this loss function works at the similarity level and does not provide any supervision for connecting cross-modal contents at the attention level. In other words, learning cross-modal attentions is a pure data-driven approach and lacks supervision. As a result, the learned cross-modal attention model could be sub-optimal.

3.2. Contrastive Content Re-sourcing

A desired property of a well-learned cross-modal attention model is that, for a query fragment, the attention model should assign large attention weights to the key fragments that are relevant to the query fragment, and assign small attention weights to the key fragments that are irrelevant to the query fragment. The Contrastive Content Re-sourcing (CCR) constrain is proposed to explicitly guide attention models to learn this property. It enforces a query fragment to be more relevant to its attended information than to its reversed attention information. For example, as shown in Figure 2 (b), the query "dog" is required to be more relevant to its attended information than to the reversed attended information which contains the person and trees.

To be specific, given a query fragment q_i, its attended information is embedded as the attention feature a_i. Its reversed attention information is encoded by the vector $\hat{a}_i$, which is obtained by reversing attention weights and calculating weighted sum of key fragment features based on the reversed attention weights, as shown in Equation 5:

$$\begin{aligned} \hat{w}_{i,j} &= \frac{1 - w_{i,j}}{\sum_{j \in K}(1 - w_{i,j})}, \\ \hat{a}_i &= \sum_{j \in K} \left(\hat{w}_{i,j} \cdot k_j\right), \end{aligned} \quad (5)$$

where $\hat{w}_{i,j}$ is the reversed attention weight of the key fragment k_j with respect to the query fragment q_i.

We use the similarity function Sim to measure the relevance between the query fragment and either the attention feature or reversed one. Therefore, the loss function for CCR is defined as:

$$\ell_{CCR} = [Sim(q_i, \hat{a}_i) - Sim(q_i, a_i) + \gamma_2]_+, \quad (6)$$

where γ_2 controls the similarity difference margin.

Intuitively, in order to minimize this loss, a cross-modal attention model should assign large attention weights to relevant key fragments to increase q_i's relevant information ratio in a_i and decrease that contained in $\hat{a}_i$. The attention model will also learn to assign small attentions weights to irrelevant key fragments to diminish q_i's irrelevant information ratio in a_i and increase that in $\hat{a}_i$.

3.3. Contrastive Content Swapping

As shown in Figure 1 (c), attention models could assign large attention weights to both relevant and irrelevant key fragments. In such cases, the CCR constraint might not be able to fully address these false-positive scenarios because the query fragment can be more relevant to its attended information than to its reversed attention information. Therefore, we propose the Contrastive Content Swapping (CCS) constraint to address this problem. It constrains a query fragment's attended information to be more relevant to the query fragment than to a negative query fragment.

Specifically, given a query fragment q_i, we first sample its negative query fragment $\hat{q}_i$ from a predefined set $\hat{Q}_i$ which contain all negative query fragments with respect to q_i. The relevance between the attended information and either the query fragment or the negative query fragment is also measured by the similarity function Sim. Then the CCS constraint's loss function ℓ_{CCS} is defined as:

$$\ell_{CCS} = [Sim(\hat{q}_i, a_i) - Sim(q_i, a_i) + \gamma_3]_+, \quad (7)$$

where γ_3 is the margin parameter.

The CCS constraint will enforce the cross-modal attention model to diminish the attention weights of the key fragments that are relevant to $\hat{q}_i$. As a result, the information that is relevant to $\hat{q}_i$ but irrelevant to q_i is eliminated.

By incorporating the CCR and CCS constraints for image-text matching, we obtain the full objective function by Equation 8, where λ_{CCR} and λ_{CCS} are scalars that control the contributions of CCR and CCS, respectively:

$$\ell = \ell_{rank} + \lambda_{CCR} \cdot \ell_{CCR} + \lambda_{CCS} \cdot \ell_{CCS}. \quad (8)$$

3.4. Attention Metrics

Previous studies [12, 16] focus on qualitatively evaluating the attention models by visualizing attention maps. These approaches cannot serve as standard metrics for comparing attention correctness among different models.

Therefore, we propose Attention Precision, Attention Recall and Attention F1-Score, to quantitatively evaluate the performance of learned attention models. Attention Precision is the fraction of attended key fragments that are relevant to the correspondent query fragment, and Attention Recall is the fraction of relevant key fragments that are attended. Attention F1-Score is a combination of the Attention Precision and Attention Recall that provides an overall way to measure the attention correctness of a model.

In this paper, we only evaluate the attention models that use texts as the query fragments. This is because text encoders used in the evaluated models [12, 23, 16, 4] are GRUs [3] or Transformers [22], where defining the relevant and irrelevant key text fragments of a query region fragment could be difficult since the text fragments will be updated to include global information by the text encoder.

Given a matched image-text pair, an image fragment v is labeled as a relevant fragment of the text fragment t if the Intersection over Union (IoU)[1] between v and the correspondent region[2] of t is larger than a threshold T_{IoU}. In addition, v is regraded as an attended fragment by t if v's attention weight with respect to t is larger than a threshold T_{Att}. Let A and R be the sets of attended and relevant image fragments of t. t's Attention Precision (AP), Attention Recall (AR), and Attention F1-Score (AF) are defined as:

$$AP = \frac{|A \cap R|}{|A|}, AR = \frac{|A \cap R|}{|R|}, AF = 2 \times \frac{AP \times AR}{AP + AR}. \tag{9}$$

The annotations [18] that are used to calculate attention metrics provide the correspondence between noun phrases and image regions. A noun phrase might contain multiple words, and different words could correspond to the same image region. In order to obtain the overall attention metrics of a learned attention model, we first calculate the attention metrics at word-level, and use the maximal values within each phrase as the phrase-level metrics. The overall attention metrics are then obtained by averaging the phrase-level metrics.

4. Experiments

4.1. Datasets and Evaluations

Datasets. We evaluate our method on two public image-text matching benchmarks: Flickr30K [25] and MS-COCO [14]. Flickr30K [25] dataset contains 31K images, each of which is annotated with 5 captions. Following the setting of [16, 12], we split the dataset into 29K training images, 1K validation images, and 1K testing images. The MS-COCO dataset used for image-text matching consists of 123,287 images, each of which includes 5 human-annotated descriptions. Following [16, 12], the dataset is divided into 113,283 images for training, 5K images for validation, and 5K images for testing.

[1]Given two bounding boxes, the IoU score between them is calculated as the ratio of their joint area to their union area.

[2]The correspondent regions of a word t are the regions that contain the object described by t.

Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	rsum
SCAN [12]	67.2	90.7	94.8	48.4	77.6	84.9	463.6
+ CCR	67.8	91.1	95.0	49.4	77.6	85.3	466.2
+ CCS	**69.1**	91.1	**95.4**	50.8	78.4	85.6	470.4
+ CCR & CCS	68.8	**91.6**	95.3	**51.1**	**79.0**	**86.5**	**472.3**
PFAN [23]	69.7	90.2	94.1	50.1	78.6	86.0	468.7
+ CCR	70.3	90.5	94.7	51.9	79.4	86.7	473.5
+ CCS	70.3	90.9	95.2	51.9	79.2	86.5	474.0
+ CCR & CCS	**70.9**	**91.8**	**95.6**	**52.5**	**79.6**	**86.9**	**477.3**
BFAN [16]	70.7	92.3	**96.3**	51.8	79.3	85.9	476.3
+ CCR	71.7	92.8	96.0	**53.2**	**80.5**	**87.1**	481.3
+ CCS	71.0	93.2	96.0	52.6	79.4	86.4	478.6
+ CCR & CCS	**72.0**	**93.4**	96.2	53.1	80.3	86.9	**481.9**
SGRAF [4]	77.8	94.5	96.8	59.0	82.9	88.6	499.6
+ CCR	78.0	**95.2**	97.2	59.5	83.1	88.7	501.7
+ CCS	78.3	94.6	97.4	59.6	83.5	**89.0**	502.4
+ CCR & CCS	**79.3**	**95.2**	**98.0**	**59.8**	**83.6**	88.8	**504.7**

Table 1: Results of the sentence retrieval and image retrieval tasks on the Flickr30K test set.

Evaluation Metrics. Following [16, 12, 23], we measure the performance of both **Image Retrieval** and **Sentence Retrieval** tasks by calculating recalls at different K values (R@K, K = 1, 5, 10), which are the proportions of the queries whose top-K retrieved items contain their matched items. We also report *rsum*, which is the summation of all R@K values for a model. On the Flickr30K dataset, we report results on the 1K testing images. On the MS-COCO dataset, we report results through averaging over 5-folds 1K test images (referred to MS-COCO 1K), and testing on the full 5K test images (referred to MS-COCO 5K) following the standard evaluation protocol [12, 16, 23].

To compute the attention metrics, T_{IoU} is set as 0.4, and the results for other values of T_{IoU} can be found in the supplementary material. The possible values of T_{Att} are uniformly chosen between 0 and 0.1 with the interval of 0.01. We set the range of T_{Att} based on the experimental results that when achieving the best Attention F1-Score the $T_{Att}$is ranging from 0 to 0.1. We calculate the Attention Precision, Attention Recall and Attention F1-Score for each value of T_{Att}, and then report the precision-recall (PR) curves and the best Attention F1-Score with its correspondent Attention Precision and Attention Recall.

4.2. Baselines and Implementation Details

We evaluate the proposed constraints by incorporating them into the following state-of-the-art attention-based image-text matching models:

Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	rsum
1K Test Images							
SCAN [12]	70.6	93.8	**97.7**	54.1	86.0	93.4	495.6
+ CCR	71.4	**94.2**	**97.7**	55.6	86.7	93.8	499.4
+ CCS	71.1	94.0	**97.7**	**56.6**	87.2	**94.0**	500.6
+ CCR & CCS	**71.6**	94.0	**97.7**	56.4	**87.3**	**94.0**	**501.0**
PFAN* [23]	74.5	95.4	**98.6**	59.8	88.8	94.8	511.9
+ CCR*	74.4	95.3	98.3	60.5	**89.1**	94.8	512.4
+ CCS*	74.9	**95.8**	98.3	60.8	**89.1**	94.5	513.4
+ CCR & CCS*	**75.2**	95.6	98.2	**61.2**	88.9	**94.7**	**513.8**
BFAN [16]	75.0	95.0	98.2	58.8	88.3	94.4	509.7
+ CCR	**75.2**	95.3	**98.3**	60.1	88.7	**94.7**	512.3
+ CCS	75.1	95.3	**98.3**	59.6	88.5	94.6	511.4
+ CCR & CCS	**75.2**	**95.5**	98.1	**60.3**	**88.8**	**94.7**	**512.6**
SGRAF [4]	79.7	96.5	98.5	63.3	90.1	95.7	523.8
+ CCR	79.7	**96.8**	98.7	63.8	90.4	**95.9**	525.3
+ CCS	79.7	**96.8**	**98.8**	63.8	90.3	95.7	525.1
+ CCR & CCS	**80.2**	**96.8**	98.7	**64.3**	**90.6**	95.8	**526.4**
5K Test Images							
SCAN [12]	47.2	77.6	87.7	34.7	65.2	77.3	389.7
+ CCR	47.7	78.3	**88.2**	36.2	66.6	78.2	395.2
+ CCS	46.5	**78.5**	88.0	36.5	66.6	78.3	394.4
+ CCR & CCS	**47.9**	78.1	**88.2**	**36.9**	**66.9**	**78.4**	**396.4**
BFAN [16]	52.5	80.3	89.5	37.5	66.7	78.1	404.6
+ CCR	52.0	**81.5**	89.9	**38.7**	**67.8**	**78.8**	408.7
+ CCS	**53.8**	81.1	89.9	38.0	67.3	78.5	408.6
+ CCR & CCS	53.4	81.3	**90.1**	38.4	67.6	78.6	**409.4**
SGRAF [4]	58.3	84.8	91.9	41.8	70.9	81.2	428.9
+ CCR	59.2	84.8	92.0	42.2	71.1	81.7	431.0
+ CCS	58.6	**85.0**	**92.2**	42.2	71.2	81.6	430.8
+ CCR & CCS	**59.7**	**85.0**	92.0	**42.3**	**71.4**	**81.9**	**432.3**

Table 2: Results of the sentence retrieval and image retrieval tasks on the MS-COCO test set. *Note that since the official implementation of PFAN only provides 1K images for testing, PFAN is tested **without** 5-fold cross-validation under the setting of 1K test images, and cannot be tested under the setting of 5K test images.

- **SCAN** [12] is a stacked cross-modal attention model to infer the relevance between words and regions and calculate image-text similarity.
- **PFAN** [23] improves cross-modal attention models by integrating image region position information into them.
- **BFAN** [16] is a bidirectional cross-modality attention model which allows to attend to relevant fragments and also diverts all the attention into these relevant fragments to concentrate on them.
- **SGRAF** [4] first learns the global and local alignments between fragments by using cross-modal attention models, and then applies the graph convolutional networks [9] to infer relation-aware similarities based on the local and global alignments.

We apply the proposed constraints to one randomly sampled query fragment for each matched image-text pair, in order to reduce the computational cost. For a query word fragment, its negative query set Q_i is consisted of the other words of its correspondent sentence. For a query region fragment, its Q_i is set as the other regions of its correspondent image. The constraint loss weight factors λ_{CCR} and λ_{CCS} could be 0.1 or 1, and constraint similarity margins γ_2 and γ_3 are set to 0, 0.1 or 0.2. We train models with all possible combinations with constraint loss weight factors and similarity margins, and report the best results.

The experiments on Flickr30K and MS-COCO are conducted on the RTX8000 and A100 GPU, respectively. All the baselines are trained by their officially released codes. [3] [4] [5] [6]. All models are trained from scratch by completely following their original hyper-parameters settings such as the learning rate, batch size, model structure, and optimizer [12, 16, 23, 4]. More implementation details can be found in the supplementary materials.

4.3. Experiments on Image-Text Matching

We start by evaluating the proposed approach for image and sentence retrieval tasks on both Flickr30K and MS-COCO datasets. Table 1 and Table 2 show the results on the Flickr30K and MS-COCO datasets, respectively. We find that when the proposed CCR and CCS constraints are employed separately, they both achieve consistent performance improvements on all baselines and tasks. More importantly, all models achieve the best overall improvements (*rsum*) when we apply both constraints. These results demonstrate the strong generality of our proposed constraints for different models and datasets. We also note that using CCR or CCS alone achieve better results than using both CCR and CCS under some metrics. One possible reason is that the CCR is expected to assign large attention weights to the key fragments that contain both irrelevant and relevant information. For example, it will attend to regions containing background and described objects. CCS tends to ignore these key fragments to decrease the attention weights on irrelevant information. As a result, using CCR and CCS together might result in conflicts in some rare cases, and using CCR (or CCS) alone may achieve slightly better results under some metrics.

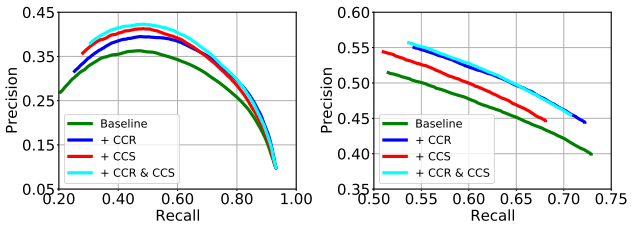
4.4. Attention Evaluation

Quantitative Analysis. We report the results on Flickr30K since it has publicly available cross-modal correspondence annotations [18] while MS-COCO does not.

[3] https://github.com/kuanghuei/SCAN
[4] https://github.com/CrossmodalGroup/BFAN
[5] https://github.com/HaoYang0123/Position-Focused-Attention-Network
[6] https://github.com/Paranioar/SGRAF

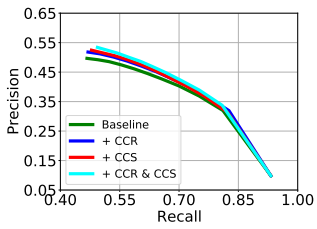
Method	Attention Precision	Attention Recall	Attention F1-Score
SCAN [12]	32.79	65.30	39.96
+ CCR	36.30	**66.80**	43.10
+ CCS	37.28	64.97	43.38
+ CCR & CCS	**38.81**	64.62	**44.44**
BFAN [16]	46.08	63.32	48.91
+ CCR	50.21	**64.20**	**51.78**
+ CCS	49.16	61.44	49.74
+ CCR & CCS	**51.13**	62.97	51.73
SGRAF [4]	44.54	61.98	47.91
+ CCR	45.22	**64.07**	49.12
+ CCS	47.43	60.41	49.20
+ CCR & CCS	**49.48**	62.12	**50.90**

Table 3: Results of Attention Precision, Attention Recall and Attention F1-Score (%) of the SCAN, BFAN, and SGRAF models trained on the Flickr30K dataset.



(a) PR curves of SCAN

(b) PR curves of BFAN



(c) PR curves of SGRAF

Figure 3: The Attention PR curves of the SCAN, BFAN, and SGRAF models trained on the Flickr30K dataset.

We note that the results of PFAN are not reported because we cannot obtain the bounding boxes of the input image regions that are correspondent to the testing data provided by its official implementation.

The attention metrics of SCAN, BFAN, and SGRAF are shown in Table 3. We can see that applying CCR and CCS individually yields higher Attention F1-Score than both baseline methods, and this is consistent to the observations in Section 4.3. More interestingly, we can find that using CCR alone improves both Attention Precision and Attention Recall; using CCS alone mainly improves Attention Precision; combining both constraints further improves Attention Precision. These results show the constraints work as intended. Note that the slight decrease in Attention Recall caused by CCS might be due to the fact that CCS enforces attention models to ignore the regions containing both foreground objects and noise background. We also present the PR curves of SCAN, BFAN, and SGRAF in Figure 3 to demonstrate the impact of different T_{Att} on Attention Precision and Attention Recall. We can observe that applying the proposed constraints yields consistently better results than both baseline methods for different T_{Att}.

We further evaluate the relation between the image-text matching performance and the quality of learned attention models by calculating the Pearson correlation coefficient between Attention F1-Score and *rsum* for each model. The obtained correlation coefficients of the SCAN, BFAN, and SGRAF models are 0.967, 0.992, and 0.941, respectively. The p-values are all less than 0.05. The results show that the image-text matching performance has strong positive correlation with the quality of learned attention models, which further demonstrate our motivation to propose the constraints.

Qualitative Analysis. We visualize the attention weights with respect to three sampled query word fragments on the Flickr30K and MS-COCO dataset. The results are shown in Figure 4 and Figure 5, respectively. More examples are provided in the supplementary material due to the space limitation. In the examples of the query word fragment "fire" and "mouse", the learned attention model of SCAN (see Column (b)) fails to assign large attention weights to the most regions containing fire or mouse. By contrast, the CCR constraint (see Column (c)) mitigates this issue by significantly increasing the attention weights assigned to the regions containing fire or mouse. The CCS constraint (see Column (d)) is less effective in these cases. In the cases of the query word fragment "infant" and "surfer", the learned attention model of SCAN (see Column (b)) assigns large attention weights to both the irrelevant and relevant regions. In this case, the CCR constraint (see Column (c)) cannot fully diminish the attention weights assigned to the regions irrelevant to "infant" and "surfer'. In contrast, as shown in Column (d), the attention weights assigned to irrelevant regions are largely diminished by the CCS constraint. In the examples of the query word "guy" and 'suitcases', they show that combining both constraints decreases the attention weights of the background regions (e.g., the surrounding areas of the "guy") more significantly than applying the them separately.

5. Conclusions

To tackle the issue of missing direct supervisions in learning cross-modal attention models for image-text matching, we introduce the constraints of CCR and CCS to supervise the learning of attention models in a contrastive

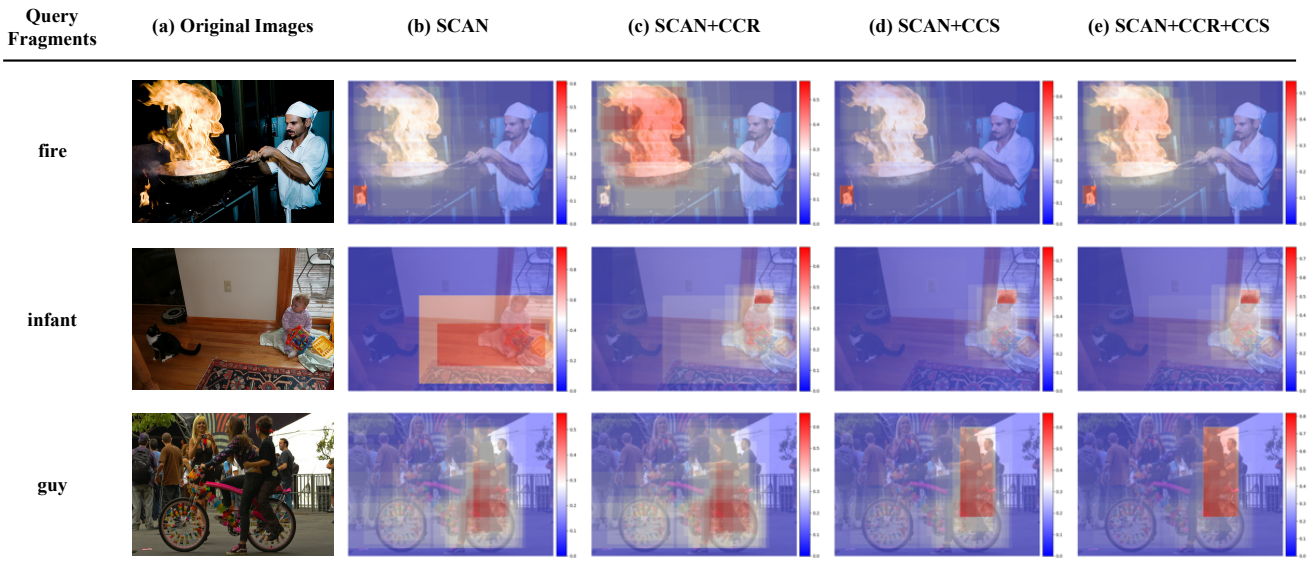


Figure 4: Examples illustrating attended image regions with respect to the given words for the SCAN model on the Flickr30K dataset.

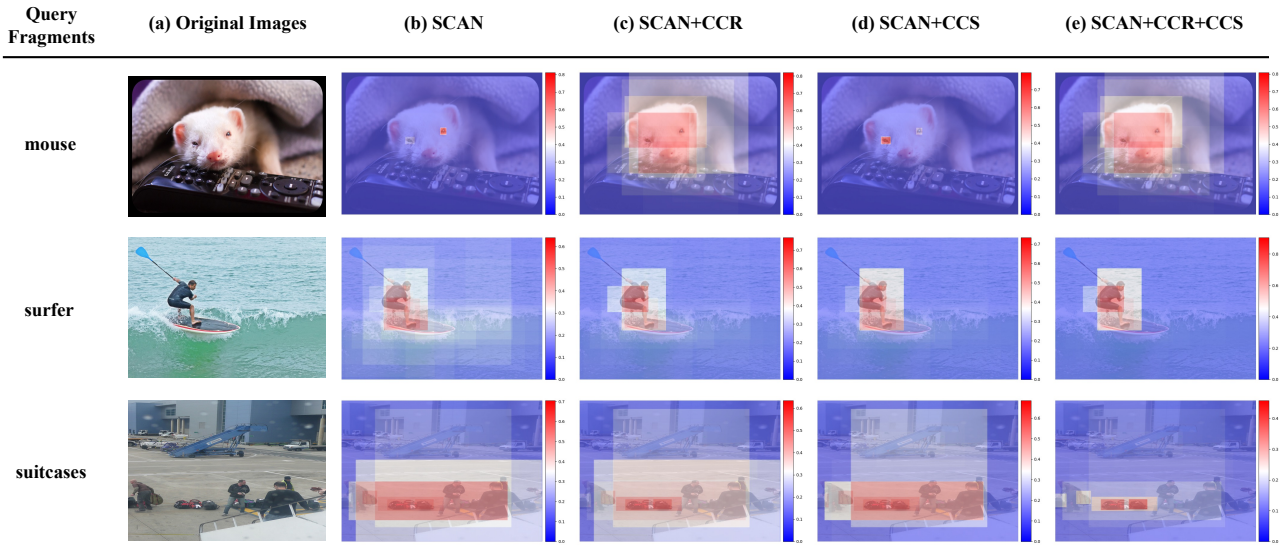


Figure 5: Examples illustrating attended image regions with respect to the given words for the SCAN model on the MS-COCO dataset.

manner without requiring additional attention annotations. Both constraints are generic learning strategies that can be generally integrated into attention models. Furthermore, in order to quantitatively measure the attention correctness, we propose three new attention metrics. The extensive experiments demonstrate that the proposed constraints manage to improve the cross-modal retrieval performance as well as the attention correctness when integrated into four state-of-the-art attention models. For future work, we will explore on how to extend the proposed constraints to other cross-modal attention models based tasks, such as Visual Question Answering (VQA) and Image Captioning.

References

[1] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. *arXiv preprint arXiv:2003.03669*, 2020.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy,

Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1218–1226, 2021.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[6] 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 *Advances in neural information processing systems*, pages 2121–2129, 2013.

[7] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, and Zhoujun Li. Bi-directional spatial-semantic attention networks for image-text matching. *IEEE Transactions on Image Processing*, 28(4):2008–2020, 2018.

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

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

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

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

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

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[14] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[16] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11, 2019.

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

[18] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[19] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[23] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019.

[24] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image-text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

[26] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701, 2018.

[27] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE, 2019.