ACMNet: Adaptive Confidence Matching Network for Human Behavior Analysis via Cross-modal Retrieval

HUI CHEN and GUIGUANG DING, Beijing National Research Center for Information Science and Technology (BNRist); School of Software, Tsinghua University, China
ZIJIA LIN, Microsoft Research, China
SICHENG ZHAO, University of California, Berkeley, USA
XIAOPENG GU, National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data(PSRPC); China Academy of Electronics and Information Technology, China
WENYUAN XU, Ubiquitous System Security Lab (USSLab), Zhejiang University, China
JUNGONG HAN, University of Warwick, UK

Cross-modality human behavior analysis has attracted much attention from both academia and industry. In this article, we focus on the cross-modality image-text retrieval problem for human behavior analysis, which can learn a common latent space for cross-modality data and thus benefit the understanding of human behavior with data from different modalities. Existing state-of-the-art cross-modality image-text retrieval models tend to be fine-grained region-word matching approaches, where they begin with measuring similarities for each image region or text word followed by aggregating them to estimate the global image-text similarity. However, it is observed that such fine-grained approaches often encounter the similarity bias problem, because they only consider matched text words for an image region or matched image regions for a text word for similarity calculation, but they totally ignore unmatched words/regions, which might still be salient enough to affect the global image-text similarity. In this article, we propose an Adaptive Confidence Matching Network (ACMNet), which is also a fine-grained matching approach, to effectively deal with such a similarity bias. Apart from calculating the local similarity for each region(/word) with its matched words(/regions), ACMNet also introduces a confidence score for the local similarity by leveraging the global text(/image) information, which is expected to help measure the semantic relatedness of the region(/word) to the whole text(/image). Moreover, ACMNet also incorporates the confidence scores together with the local similarities in estimating the global image-text similarity. To verify the effectiveness of ACMNet, we conduct extensive experiments and make comparisons with state-of-the-art methods on two benchmark datasets, i.e., Flickr30k and MS COCO. Experimental results show that the proposed ACMNet can outperform the state-of-the-art

This work was supported by the National Key R&D Program of China (Grant No. 2018YFC0806900), the National Natural Science Foundation of China (Grants No. 61571269 and No. 61701273) and the National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

Authors' addresses: H. Chen and G. Ding (corresponding author), Beijing National Research Center for Information Science and Technology (BNRist); School of Software, Tsinghua University, Beijing, 100084, China; emails: jichenhui2012@gmail.com, dinggg@tsinghua.edu.cn; Z. Lin, Microsoft Research, Beijing, China; email: zijlin@microsoft.com; S. Zhao, University of California, Berkeley, Berkeley, USA; email: schzhao@gmail.com; X. Gu, National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data(PSRPC); China Academy of Electronics and Information Technology, Beijing, China; email: gxr0212@163.com; W. Xu, Ubiquitous System Security Lab (USSLab), Zhejiang University, Zhejiang, China; email: wyxu@zju.edu.cn; J. Han, University of Warwick, Coventry, UK; email: jungonghan77@gmail.com.

1551-6857/2020/04-ART27 $15.00
https://doi.org/10.1145/3362065

methods by a clear margin, which well demonstrates the effectiveness of the proposed ACMNet in human behavior analysis and the reasonableness of tackling the mentioned similarity bias issue.

CCS Concepts: • **Information systems → Retrieval models and ranking**; *Multimedia information systems*;

Additional Key Words and Phrases: Cross-modality retrieval, human behavior analysis, image-text retrieval, adaptive confidence matching network

ACM Reference format:
Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, Xiaopeng Gu, Wenyuan Xu, and Jungong Han. 2020. ACMNet: Adaptive Confidence Matching Network for Human Behavior Analysis via Cross-modal Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 1s, Article 27 (April 2020), 21 pages.
https://doi.org/10.1145/3362065

1 INTRODUCTION

With the rapid development of the Internet, data acquired from different modalities, such as image and text, are growing at an unprecedented speed, which leads to great challenges and demands for human behavior analysis. Studying such multi-modality data can provide comprehensive understanding for human behavior. For example, when reporting a social event, various modalities of data, like images, texts, and videos, are leveraged to describe the event or express people's opinions. In spite of the success of human behavior analysis on single modality, e.g., human gesture recognition [22], video tracking [43, 58–60], image classification [11, 18], video retrieval [52], there exist many fundamental problems to be solved for cross-modality human behavior analysis, such as learning common representations of human behaviors based on data of different modalities, cross-modality retrieval for retrieving semantically related data of a different modality, and so on.

Our article focuses on the cross-modality retrieval problem mentioned above, particularly on cross-modal image-text retrieval [26, 50, 55]. In cross-modal retrieval, given a query of a modality, instances of another modality is expected to be retrieved once they are semantically similar to the query. Taking the cross-modality image-text retrieval as an example, given an image as a query, we expect semantically related texts to be retrieved to describe the content in the image. Likewise, given a text as a query, we hope that semantically related images are returned to convey the meaning of the text. Such a cross-modality image-text retrieval can provide different aspects of information from other modalities and thus can benefit the understanding of human behavior.

However, human behavior understanding via cross-modal image-text retrieval is rather challenging. It is difficult to capture the similarity between images and texts due to the heterogeneity across modalities [17]. One common solution is to map images and texts into some common embedding spaces, where similar image-text pairs are close while dissimilar ones are kept far away from each other. Then, the similarity between a given image and a given text can be directly calculated by the distance between their representations in the common spaces.

Early works on cross-modality image-text retrieval usually used the global features of images and texts to estimate their global similarities for ranking. However, these global features usually neglect details about the data, leading to unsatisfactory performance for cross-modality retrieval. Recently, researchers have dedicated themselves to exploring fine-grained matching approaches, in which the global image-text similarity generally arises from an aggregation of local similarities between image regions and text words [19]. Karpathy and Fei-Fei [24] proposed to align each region in an image with words of a given text in the common embedding space, and then aggregate the region-word similarities into an image-text similarity with a common pooling method.

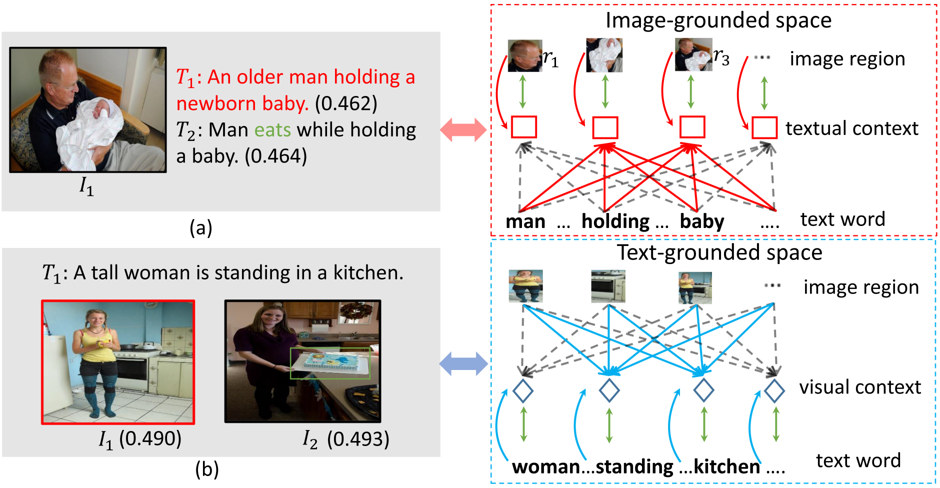

T_1: An older man holding a newborn baby. (0.462)
T_2: Man eats while holding a baby. (0.464)
I_1
(a)
Image-grounded space
r_1
r_3
...
image region
textual context
man ... holding ... baby ....
text word
T_1: A tall woman is standing in a kitchen.
I_1 (0.490)
I_2 (0.493)
(b)
Text-grounded space
...
image region
visual context
woman...standing ...kitchen ....
text word

Fig. 1. Illustrations of the similarity bias in both embedding space defined in Reference [31]. In example (a), T_1 is the ground-truth text of I_1, whereas T_2 is not. In example (b), I_1 is the ground-truth image of T_1, whereas I_2 is not. Numbers in parenthesis are similarity scores computed in respective embedding spaces. Bi-directional green arrows mean the calculation of local similarities. Best viewed in color.

Lee et al. [31] further proposed a stack cross attention network (SCAN). Specifically, in SCAN, given an image-text pair, the global similarity between them is measured by aggregating local similarities in two separate embedding spaces, i.e., the image-grounded and the text-grounded embedding spaces. In the **image-grounded embedding space**, local similarities are calculated by measuring the similarity between each image region and its corresponding textual context feature vector derived via the attention mechanism on all words of the text. Likewise, in the **text-embedding embedding space**, local similarities are calculated by measuring the similarity between each word and its corresponding visual context feature vector derived via the attention mechanism on all image regions. SCAN is quite effective, because it achieves the state-of-the-art performance for cross-modality image-text retrieval.

Despite the state-of-the-art performance obtained by those fine-grained matching approaches above, their region-word matching strategies carry out the same problem more or less, namely, when calculating local similarities, each region only considers the related words in the text and each word only considers related regions in the image, but those unmatched words/regions get ignored. For example, for both embedding spaces in Reference [31], each region is encouraged to match its related text words, and each word is encouraged to match its related image regions, via the attention mechanism, and thus matched region-word would contribute most of the similarities while unmatched words/regions would contribute none or little. However, in many cases, such unmatched words/regions can be salient in the text/image to substantially change the corresponding semantics, and then affect the semantic similarity between the image and the text. Failure to consider such unmatched but salient words/regions in similarity calculation can cause a similarity bias when estimating the image-text semantic similarity. As illustrated in Figure 1(a), both texts, i.e., "An older man holding a newborn baby" and "Man eats while holding a baby," can well describe the semantics of image I, i.e., "man," "holding," and "baby." For the mentioned fine-grained matching approaches like Reference [31], these shared semantics will be leveraged to compute local similarities in the image-grounded embedding space, resulting in similar similarity scores of T_1 and T_2 with respect to I_1. However, T_2 is not very semantically related to I_1 because of the additional semantics in it, i.e., the word "eats" highlighted in green in T_2. Similarly, in example (b), in the text-grounded embedding space, semantics of the text T_1, i.e., "woman," "standing," "kitchen," and so on, appear in both images, i.e., I_1 and I_2, resulting in similar similarity scores of I_1 and I_2 with respect to T_1. However, clearly, I_2 should not be as similar as I_1 to the text T_1 due to the unrelated but salient semantics in it, i.e., the object "cake" in the green bounding box in I_2.

To deal with the above similarity bias, we propose an **Adaptive Confidence Matching Network** (**ACMNet**) for cross-modality image-text retrieval. We base our ACMNet on Reference [31], as it is the state-of-the-art. It is intuitive that if there is less unmatched semantics between a region(/word) and the text(/image), the corresponding region(/word)-specific local similarity would be more confident. However, representing such unmatched semantics in the latent embedding space is challenging. Here, we propose to leverage the global image/text information to implicitly involve the unmatched semantics information. And to tackle the similarity bias issue described above, we adopt the gating mechanism to estimate the confidence score for each local similarity adaptively, taking both unmatched and matched regions/words into consideration. Specifically, like Reference [31], the proposed ACMNet would build both image-grounded and text-grounded embedding spaces. And in the former, we use the global text feature as well as the feature of a region to estimate the confidence score for the corresponding local similarity via a gate function, which denotes "how much the region is related to the whole semantics of the text." Likewise, in the latter, we use the global image feature together with the feature of a word to estimate the confidence score via another gate function, which denotes "how much the word is related to the whole semantics of the image." Then, we aggregate all local similarities with their corresponding confidence scores into the global similarity. Each local similarity measures the semantic similarity between a region(/word) with its matched words(/regions), and the corresponding confidence score further measures the semantic relatedness between it with the whole text(/image), which is expected to help tackle the mentioned similarity bias issue.

The contributions of our work are summarized as follows:

- We observe that the state-of-the-art fine-grained matching methods [24, 31] usually encounter a similarity bias issue, and thus propose an adaptive confidence matching network (ACMNet) to tackle it for cross-modality image-text retrieval.
- Our ACMNet can assess the confidence of each local similarity adaptively and measure the global similarity more consistently. As a result, only image-text pairs sharing full semantics with each other would obtain higher global similarities.
- We verify the proposed adaptive confidence matching network (ACMNet) by extensive experiments and analyses on benchmark datasets. Experimental results well demonstrate that the proposed method can outperform other state-of-the-art approaches.

The rest of this article is structured as follows. Section 2 provides a comprehensive review of related works on cross-modality image-text retrieval. Section 3 elaborates the architecture of the proposed adaptive confidence matching network (ACMNet), including image/text representation learning, adaptive confidence matching network grounded on both features, and loss function for training. Experimental results and analyses are presented in Section 4, followed by conclusions in Section 5.

2 RELATED WORK

Recently, there has been much interest in the task of image-text retrieval. Existing methods can be classified into two categories [19]: (1) one-to-one matching and (2) many-to-many matching.

One-to-one matching. One-to-one matching methods usually associate global representations of images and texts using structured objective [26, 49] or a canonical correlation objective [57].

Frome et al. [15] used convolutional neural networks (CNNs) and Skip-Gram [33] to extract global representations for images and texts, respectively, and then associated them with a structured objective function where the matched image-text pairs are enforced to be close to each other. Vendrov et al. [49] focused on refining the training objective so that the partial order structure of visual-semantic hierarchy could be preserved. Kiros et al. [26] employed a hinge-based triplet ranking loss for learning cross-modality representations, and achieved encouraging performances. Faghri et al. [14] further refined the hinge-based triplet loss function by leveraging hard negatives, and yielded significant improvement. Peng et al. [39] and Gu et al. [17] incorporated generative objectives to enhance the cross-modality feature learning and obtained reasonable performance improvement.

Many-to-many matching. This category of methods model the latent vision-language correspondence at a fine-grained level. Namely, local similarities are obtained by comparing pairs of instances in images and texts, i.e., image regions and text words, and then the global similarity is calculated by aggregating these local similarities. Karpathy and Fei-Fei [24] first proposed to model the global similarity through local similarities between image regions and words of texts with a structured objective. They used an R-CNN [16] to detect image regions at the object level and obtained region-level representations for given images. Plummer et al. [42] proposed to localize textual entity mentions in an image, and model region-to-phrase correspondences for instance-level image-text matching. Niu et al. [37] detected fine-grained element first, i.e., phrases within the texts and salient regions within images, then mapped them into the common embedding space, together with the global image and text information, through a hierarchical model.

Recently, the attention mechanism [48, 54, 62] has achieved great success to boost the performance for cross-modality image-text retrieval. Huang et al. [19] developed a context-modulated attention scheme by multi-modal LSTMs to selectively attend to a pair of instances appearing in both image and text. Nam et al. [36] proposed to perform a dual attention network within multiple steps and captured the fine-grained interplay between images and texts. Lee et al. [31] proposed to discover the full latent alignments between regions in images and words in texts via a stack cross attention model.

Our work lies in the category of many-to-many matching methods. Similar to the mentioned baselines above [24, 31], we compute the global similarity for the given image-text pair through aggregating local similarities. However, in the manner of calculating and aggregating local similarities, those methods can have similarity biases. Namely, with the attention mechanism, when calculating local similarities for image regions, they probably only focus on semantically related words, and ignore other unmatched but meaningful words that can affect the semantics of the sentence. Meanwhile, when calculating local similarities for words, they probably only focus on semantically related regions, and ignore others that are unmatched but salient enough to change the semantics of the image. And thus to tackle such a similarity bias issue, our work propose an adaptive confidence network through introducing confidence to determine the semantic relatedness between local similarities with the whole text(/image).

3 ACMNET: ADAPTIVE CONFIDENCE MATCHING NETWORK

In this section, we describe the proposed Adaptive Confidence Matching Network (ACMNet) for image-text retrieval. First, a convolutional neural network (CNN) is employed as an image encoder to convert image regions into region-level features. Meanwhile, a recurrent neural network (RNN) is adopted as a text encoder to encode each word of the text into word-level features. Then, given the image/text features, we construct the adaptive confidence matching network grounded on images and texts, respectively. Finally, we leverage a ranking objective to train the whole model in an end-to-end manner.

This section is structured as follows: We first introduce the image feature learning in Section 3.1, followed by description about the text feature learning in Section 3.2. Then, we describe how to construct our adaptive confidence matching network for the image-grounded embedding space and the text-grounded embedding space in Sections 3.3 and 3.4, respectively. Finally, the objective function for training is provided in Section 3.5.

3.1 Image Feature Learning

Recent researches have convincingly demonstrated that CNNs are highly capable of learning informative representations for images. CNNs were first proposed to tackle the challenge of image classification. Thrilling performance on ImageNet [8] was achieved by different CNNs, such as VGGNet [46], ResNet [18], InceptionNet [47], and so on. From then on, CNNs greatly boost the development of fields in computer vision, like object detection [44, 45], visual tracking [10, 28–30], image/video captioning [1, 5, 56], visual question answering [2, 25], and so on. Following Reference [31], we adopt a Faster R-CNN [45] to extract discriminative region-level image features for given images. Specifically, given an image I, instead of dividing the image into grids uniformly like conventional CNNs (e.g., ResNet [18]), Faster R-CNN first detects objects in the image and marks regions with an objects inside using bounding boxes. Then, for each region r_i, a feature vector f_i is obtained through a ROI pooling method [45]. To adapt to the benchmark datasets, we further map f_i to a latent space by a fully connected layer:

$$v_i = \mathbf{W}_v f_i + \mathbf{b}_v, \tag{1}$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times d_f}$ and $\mathbf{b}_v \in \mathbb{R}^d$ are to-be-learned parameters. d is the dimensionality of the latent space, and d_f is the dimensionality of f_i. Finally, for each image, we get a set of feature vectors denoted as $V = \{v_i | i = 1, \ldots, n, v_i \in \mathbb{R}^d\}$, where n is the number of detected regions in I.

We consider each feature vector v_i in V as a local region-level feature, since it is only related to a partial region of the image I. We can easily obtain a global image feature $\bar{v}$, which is supposed to represent whole semantics in the image, by the average pooling:

$$\bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i. \tag{2}$$

3.2 Text Feature Learning

Recently, researchers have explored numerous ways to learn informative representations for texts and words in them, such as word2vec [34], glove [40], fasttext [4, 23], ELMO [41], BERT [9], and so on. Such representations have exhibited impressive performance in many tasks in natural language processing domains, such as named entity recognition [21, 62], machine translation [3, 48]. Instead of directly using these pre-trained representations to measure similarities between images and texts, following [14, 26, 31], we employ a bi-directional gated recurrent unit (GRU, a variant of RNNs) [6] to extract features for the texts.

Specifically, given a text $S = \{w_1, w_2, \ldots, w_m\}$ where each word is denoted as a one-hot vector, w_j. Here w_j is a d'-dimensional vector, where d' is the size of the word vocabulary, and only the position corresponding to the word in it is set as 1, while others are set as 0. Then, we use an embedding matrix $\mathbf{W}_e$ to embed the word into a continuous embedding vector x_j:

$$x_j = \mathbf{W}_e w_j, \forall j \in [1, m], \tag{3}$$

where $\mathbf{W}_e \in \mathbb{R}^{d_e \times d'}$ is the to-be-learned embedding matrix, with d_e being the dimensionality of the embedding vector. And m is the length of a text here.

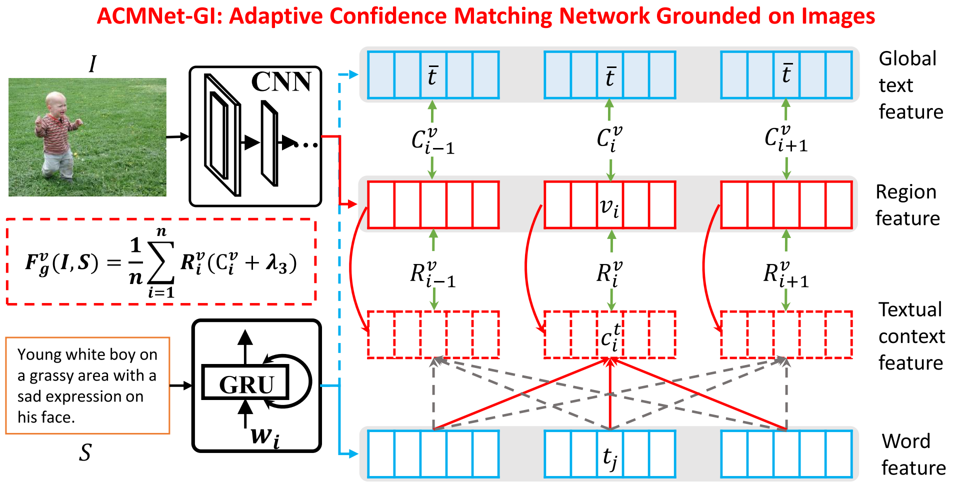

ACMNet-GI: Adaptive Confidence Matching Network Grounded on Images
I
CNN
$\bar{t}$
$\bar{t}$
$\bar{t}$
Global text feature
C_{i-1}^v
C_i^v
C_{i+1}^v
v_i
Region feature
$F_g^v(I,S) = \frac{1}{n}\sum_{i=1}^{n} R_i^v(C_i^v + \lambda_3)$
R_{i-1}^v
R_i^v
R_{i+1}^v
c_i^t
Textual context feature
Young white boy on a grassy area with a sad expression on his face.
GRU
w_i
S
t_j
Word feature

Fig. 2. Overview of the proposed adaptive confidence matching network grounded on images. For simplicity, we use R_i^v and C_i^v to denote $R(v_i, c_i^t)$ in Equation (11) and $C(v_i, \bar{t})$ in Equation (12), respectively.

After that, a bi-directional GRU is used to capture the context information from both forward and backward directions in the text S. To ease the explanation, we use $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ to denote the hidden states of the forward $\overrightarrow{\text{GRU}}$ and the backward $\overleftarrow{\text{GRU}}$, respectively:

$$\overrightarrow{h}_j = \overrightarrow{\text{GRU}}(x_j, \overrightarrow{h}_{j-1}); \quad \overleftarrow{h}_j = \overleftarrow{\text{GRU}}(x_j, \overleftarrow{h}_{j+1}). \tag{4}$$

And, we define the word-level feature t_j for word w_j as

$$t_j = \frac{\overrightarrow{h}_j + \overleftarrow{h}_j}{2}, \tag{5}$$

where $t_j \in \mathbb{R}^d$, with the same dimension as v_j in Equation (1).

We define a local word-level feature set denoted as $T = \{t_j | j = 1, \ldots, m, t_j \in \mathbb{R}^d\}$ for a text S. And the global text feature $\bar{t}$ of S is derived as follows, i.e., the last hidden states of the forward and the backward GRU.:

$$\bar{t} = \frac{\overrightarrow{h}_m + \overleftarrow{h}_0}{2}. \tag{6}$$

3.3 ACMNet-GI: Adaptive Confidence Matching Network Grounded on Images

In this section, we introduce the proposed adaptive confidence matching network (ACMNet) on the image-grounded embedding space. As illustrated in Figure 2, given an image and a text, local similarities between image regions and words of the text are first calculated through the attention mechanism on all words of the text. Then, we leverage the global text feature, i.e., $\bar{t}$ in Equation (6), to estimate the confidence for each local similarity through a gating mechanism. Finally, a global similarity between the image and the text is derived by adaptively weighting each local similarity of an image region with its corresponding confidence score.

Local similarity grounded on images. The calculation of local similarities grounded on images is decomposed into three steps. First, for each image region, we align it with all words of the given text to explore the correspondences between each region-word pair. Then, we derive a textual context feature for each region via the attention mechanism. Finally, a cosine similarity score is computed using the region feature vector and its corresponding textual context feature vector.

Specifically, given the image feature set V with n local region-level feature vectors corresponding to an image I, and the text feature set T with m local word-level feature vectors corresponding to a text S, relationships among all possible region-word pairs are discovered by a cosine similarity:

$$s_{ij} = \frac{v_i^T t_j}{||v_i|| \cdot ||t_j||}, \forall i \in [1, n], \forall j \in [1, m], \tag{7}$$

where v_i is the feature vector corresponding to the region r_i in the image I, and t_j is the feature vector of word w_j in the text S. If region r_i is semantically related to word w_j, then the cosine similarity s_{ij} is expected to be large. Otherwise, s_{ij} should be small.

Enumerating all pairs of region-word, we obtain a similarity matrix, denoted as $\mathbf{s}$, for a given image-text pair. As in References [24, 31], we further normalize it along the **column** dimension as

$$\bar{s}_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{i=1}^{n} [s_{ij}]_+^2}}, \tag{8}$$

where $[x]_+ \equiv \max(x, 0)$.

To attend to the word feature vectors, we then obtain $\boldsymbol{\alpha}$ where each row is a probability distribution over all words of being selected by an image region:

$$\alpha_{ij} = \frac{exp(\lambda_1 \bar{s}_{ij})}{\sum_{l=1}^{m} exp(\lambda_1 \bar{s}_{il})}, \tag{9}$$

where λ_1 is an inverse temperature of the softmax function [7].

A textual context feature vector, c_i^t, for region r_i is then given by a weighted combination of word feature vectors, i.e., T:

$$c_i^t = \sum_{j=1}^{m} \alpha_{ij} t_j. \tag{10}$$

After that, we define the relevance score $R(v_i, c_i^t)$ between the region feature v_i and its corresponding textual context feature c_i^t as

$$R(v_i, c_i^t) = \frac{v_i^T c_i^t}{||v_i|| \cdot ||c_i^t||}. \tag{11}$$

We can regard $R(v_i, c_i^t)$ as the local similarity associated with the image region r_i for a given image-text pair (I, S). And thus, for the input image I and S, we can derive a set of local similarities, denoted as $F_l^v(I, S) = \{R(v_i, c_i^t) | i = 1, \ldots, n\}$.

Global similarity with adaptive confidence matching. Given all local similarities, i.e., $F_l^v(I, S) = \{R(v_i, c_i^t) | i = 1, \ldots, n\}$, the global similarity is usually obtained by simply averaging these local similarities as in Reference [31]. Namely, $F_g^v(I, S) = \text{avg}[F_l^v(I, S)] = \frac{1}{n} \sum_{i=1}^{n} R(v_i, c_i^t)$, where $F_g^v(I, S)$ is the global similarity. However, as illustrated by example (a) in Figure 1, in the image-grounded space, such an aggregating strategy may improperly assign two semantically different texts with close global similarities, as long as both can fully describe the semantics of the image, though there may still exist some unmatched but salient words in the texts that can change their meanings substantially, like the word "cat" of T_2 in Figure 1. The disadvantage of the aggregating strategy above is that it just focuses on the matched semantics in the text, and ignore its unmatched semantics. Therefore, to tackle that, we introduce an adaptive aggregating strategy through the adaptive confidence matching for local similarities grounded on images, which leverages the whole semantics of the text to estimate a confidence for each local similarity and incorporates it into deriving the global similarity.

Specifically, to estimate the confidence for each image region r_i of a given image I, We leverage the global text feature, i.e., $\bar{t}$ in Equation (6) and the local region feature v_i to output a confidence score through a gate function:

$$C(v_i, \bar{t}) = \sigma\left(\mathbf{W}_c^v [v_i; \bar{t}] + \mathbf{b}_c^v\right), \tag{12}$$

where σ is a sigmoid function and [;] is a concatenation of two vectors. $\mathbf{W}_c^v \in \mathbb{R}^{2d\times 1}$ and $\mathbf{b}_c^v \in \mathbb{R}^1$ are to-be-learned parameters.

The global text feature $\bar{t}$ can fully depict the whole semantics of the text S, which involves those unmatched semantics w.r.t. the region feature r_i as well. Therefore, the less unmatched semantics is there, the more confident the corresponding local similarity will be, leading to a higher score output by Equation (12).

Finally, the global similarity score between the image I and the text S is given as

$$F_g^v(I, S) = \frac{1}{n} \sum_{i=1}^{n} R\left(v_i, c_i^t\right) [C(v_i, \bar{t}) + \lambda_3], \tag{13}$$

where λ_3 is a parameter tuning the impact of the confidence score. We empirically set λ_3 as 0.5. v_i is the region feature vector of the region r_i, and c_i^t is its corresponding textual context feature. $\bar{t}$ is the global text feature of S.

Compared with the global similarity obtained by simply averaging all local similarities in Reference [31], Equation (13) can stress more on those similarities with higher confidence, which makes the global similarity between a given image-text pair more accurate.

3.4 ACMNet-GT: Adaptive Confidence Matching Network Grounded on Texts

In this section, we introduce the proposed adaptive confidence matching network (ACMNet) on the text-grounded embedding space. Opposite to the image-grounded embedding space, local similarities are associated with each word in the text here, which are calculated through the attention mechanism on the image regions. Then, for each local similarity associated with one word of the text, we use the global image feature (i.e., $\bar{v}$ in Equation (2)) and the corresponding word feature to estimate its confidence score, similar to that in ACMNet-GI. Finally, a global similarity between the image and the text is calculated as a weighted combination of all local similarities, with the computed confidence scores as weights. Figure 3 gives an overview of the proposed ACMNet grounded on texts.

Local similarity grounded on texts. The calculation of local similarities grounded on texts is also decomposed into three steps. First, for each word of a text, we align each word with all image regions of a given image to explore the correspondences between each word-region pair. Then, we derive a visual context feature for each word of the text via the attention mechanism. Finally, a cosine similarity score is computed using the word feature vector and its corresponding visual context feature vector.

Specifically, as in the image-grounded embedding space, the similarity matrix **s** is first calculated by Equation (7), where each element s_{ij} assesses the relationship between the region feature v_i and the word feature t_j. Then, we also normalize it along the **row** dimension as [24, 31]

$$\bar{s}'_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{j=1}^{m} [s_{ij}]_+^2}}, \tag{14}$$

where $[x]_+ \equiv \max(x, 0)$.

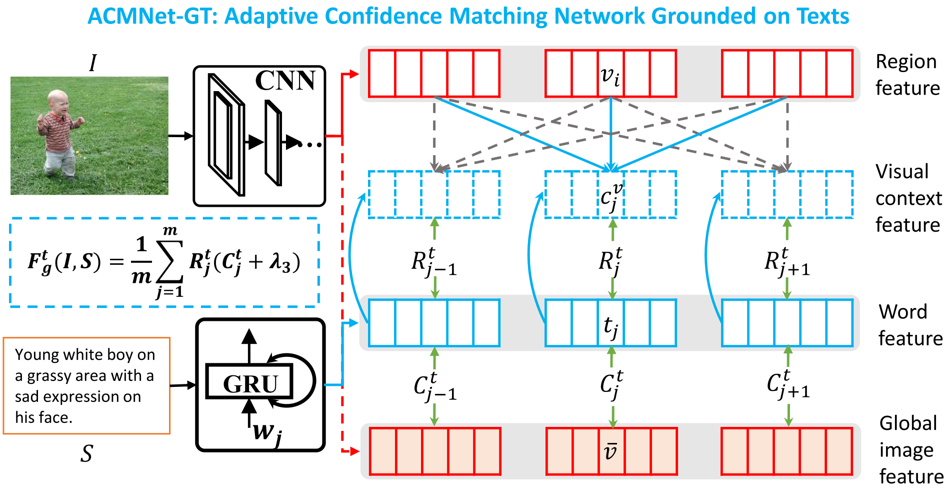

ACMNet-GT: Adaptive Confidence Matching Network Grounded on Texts
I
CNN
v_i
Region feature
c_j^v
Visual context feature
$F_g^t(I,S) = \frac{1}{m}\sum_{j=1}^{m} R_j^t(C_j^t + \lambda_3)$
R_{j-1}^t
R_j^t
R_{j+1}^t
t_j
Word feature
Young white boy on a grassy area with a sad expression on his face.
GRU
w_j
S
C_{j-1}^t
C_j^t
C_{j+1}^t
$\bar{v}$
Global image feature

Fig. 3. Overview of the proposed adaptive confidence matching network grounded on texts. For simplicity, we use R_j^t and C_j^t to denote $R(t_j, c_j^v)$ in Equation (17) and $C(t_j, \bar{v})$ in Equation (18), respectively.

To attend to the image region feature vectors, we then obtain a matrix $\boldsymbol{\beta}$ whose each column is a probability distribution over all image regions of being selected by a word of the text:

$$\beta_{ij} = \frac{exp(\lambda_2 \bar{s}'_{ij})}{\sum_{l=1}^{n} exp(\lambda_2 \bar{s}'_{lj})}, \tag{15}$$

where λ_2 is an inverse temperature of the softmax function.

A visual context feature vector, c_j^v, for the word w_j is then derived by a weighted combination of image region feature vectors, i.e., V:

$$c_j^v = \sum_{i=1}^{n} \beta_{ij} v_i. \tag{16}$$

After that, we define the relevance score $R(t_j, c_j^v)$ between the word feature t_j and its corresponding visual context feature c_j^v as

$$R\left(t_j, c_j^v\right) = \frac{t_j^T c_j^v}{||t_j|| \cdot ||c_j^v||}. \tag{17}$$

We regard $R(t_j, c_j^v)$ as the local similarity associated with the word w_j for the given image-text pair. And thus, for the input image I and S, we can derive a set of local similarities, denoted as $F_l^t(I,S) = \{R(t_j, c_j^v) | j = 1, \ldots, m\}$.

Global similarity with adaptive confidence matching. Given all local similarities, i.e., $F_l^t(I,S) = \{R(t_j, c_j^v) | j = 1, \ldots, m\}$, we can simply average them to get the global similarity as in Reference [31]. Namely, $F_g^t(I,S) = \text{avg}[F_l^t(I,S)] = \frac{1}{m}\sum_{j=1}^{m} R(t_j, c_j^v)$, where $F_g^t(I,S)$ is the global similarity. Similarly, such an aggregating strategy may improperly assign two semantically different images with close global similarities, as long as both images can fully describe the semantics of the texts, ignoring that there may still exist some unmatched but salient semantics in the image that can change their meanings substantially, like the "cat" of I_2 in Figure 1. The disadvantage of the aggregating strategy above is that it just focuses on the matched semantics in the image, and ignore its whole semantics. Therefore, to tackle that, we introduce an adaptive aggregating strategy through the adaptive confidence matching for local similarities grounded on texts, in which to estimate the confidence for each word w_j of a text S, we leverage the global image feature, i.e., $\bar{v}$ in Equation (2), and the word feature t_j to predict the related confidence through a gate function:

$$C(t_j, \bar{v}) = \sigma\left(\mathbf{W}_c^t[t_j; \bar{v}] + \mathbf{b}_c^t\right), \tag{18}$$

where σ is a sigmoid function and [;] is a concatenation of two vectors. $\mathbf{W}_c^t \in \mathbb{R}^{2d \times 1}$ and $\mathbf{b}_c^t \in \mathbb{R}^1$ are to-be-learned parameters.

Considering that the global image feature $\bar{v}$ can fully represent the whole semantics in the image I, including those unmatched semantics w.r.t. the word feature t_j. Therefore, a high confidence score output of Equation (18) indicates less unmatched semantics between the image I and the word w_j.

Finally, the global similarity score $F_g^t(I, S)$ between the image I and the text S is given by

$$F_g^t(I, S) = \sum_{j=1}^{m} R\left(t_j, c_j^v\right)[C(t_j, \bar{v}) + \lambda_3], \tag{19}$$

where λ_3 is to tune the impact of the confidence score, which is empirically set as 0.5. t_j is the word feature vector of w_j and c_j^v is its corresponding visual context feature. $\bar{v}$ is the global image feature of I.

Like Equation (13), Equation (19) can concentrate on those similarities with higher confidence, leading to a more accurate global similarity compared to the one in Reference [31].

3.5 Loss Function for Training

In this section, we describe the loss function based on a ranking loss for training. Most prior approaches utilized a triplet ranking loss as the training objective for learning the embedding spaces for visual input and textual input [14, 24, 26, 31, 35]. Generally, a hinge-based triplet ranking loss is employed to maximize the similarity between a positive image-text pair and meanwhile minimize the similarities of all negative image-text pairs.

In our case, given a training set, denoted as $\{(I_i, S_i)\}_{i=1}^N \sim \mathcal{D}$ containing N image-text pairs, each image-text pair (I_i, S_i) is treated as a positive pair, as I_i and S_i are relevant and coupled. And negative image-text pairs can be constructed as follows: 1) for an image I_i, we regard any text $S_j (j \neq i)$ as a non-matching text, ending up with a negative pair (I_i, S_j); 2) for a text S_i, we regard any image $I_j (j \neq i)$ as a non-matching image, ending up with another negative pair (I_j, S_i). Then the hinge-based triplet ranking loss function can be derived as follows:

$$\mathcal{L}_{rank}(I_i, S_i) = \sum_{j \neq i} \max[0, \Delta - F(I_i, S_i) + F(I_i, S_j)] + \sum_{j \neq i} \max[0, \Delta - F(I_i, S_i) + F(I_j, S_i)], \tag{20}$$

where $F(I, S)$ measures the similarity between I and S given by Equation (13) or (19). Note that ACMNet-GI and ACMNet-GT are two independent models that are trained separately, and for cross-model image-text retrieval, each of them can be applied. We will also compare them in our experiments.

The hinge-based triplet ranking loss encourages that a positive pair get a higher similarity score than a negative pair by a margin Δ at least. In practice, it is not efficient to compare with all negative samples in the training set. Instead, usually only the hard negatives within a mini-batch, i.e., the negatives closest to each training query [14], are considered during training. Specifically, within a mini-batch, we first search the hardest negatives for any positive image-text pair, i.e.,(I_i, S_i) as follows:

$$I_i^h = \mathop{\arg\max}_{I_j, j \neq i} F(I_j, S_i) \quad \text{or} \quad S_i^h = \mathop{\arg\max}_{S_j, j \neq i} F(I_i, S_j), \tag{21}$$

where I_i^h forms the hardest negative pair with S_i and S_i^h forms the hardest negative pair with I_i. Then the refined ranking loss $\mathcal{L}_{rank}^h(I_i, S_i)$ is given as

$$\mathcal{L}_{rank}^h(I_i, S_i) = \max[0, \Delta - F(I_i, S_i) + F(I_i, S_i^h)] + \max[0, \Delta - F(I_i, S_i) + F(I_i^h, S_i)]. \tag{22}$$

Following Reference [35], we name Equation (20) as VSE loss and Equation (22) as VSEPP loss. And in the experiments, we will utilize both functions to evaluate the proposed approach.

4 EXPERIMENT

To verify the effectiveness of the proposed adaptive confidence matching network (ACMNet), we conduct extensive experiments and analyses on two benchmark datasets. And, we also make comparisons with state-of-the-art models on both datasets. In this section, we first describe details about both benchmark datasets for cross-modality image-text retrieval in Section 4.1. Then, we elaborate on the implementation details and the adopted evaluation metrics in Section 4.2. After that, we report the performance of human-behavior-related cross-modality image-text retrieval in Section 4.3 and comparison results with state-of-the-art models on both benchmark datasets in Section 4.4, followed by model analyses in Section 4.5. Finally, we present some retrieval examples on both datasets in Section 4.6.

4.1 Datasets

We adopt two benchmark datasets, i.e., Flickr30k and MS COCO, to evaluate the proposed ACMNet and other state-of-the-art approaches.

Flickr30k. Flickr30k contains 31,000 images collected from the Flickr website. Each image is annotated with 5 English texts by Amazon Mechanical Turk workers. We use the public dataset split as [14, 31, 35], which has 29,000 images for training, 1,000 for validation, and 1,000 for testing.

MS COCO. MS COCO consists of about 123K images, each with at least 5 texts. As previous works [31, 35], we preserve 113,287 images for training, 5,000 for validation and the remaining 5,000 for testing.

4.2 Implementation Details

Evaluation Metric. Recall rates are commonly used to evaluate the performance of cross-modality image-text retrieval models. As previous works [14, 31, 35], we adopt Recall at K (R@K) to evaluate the proposed method, which is defined as the percentage of test samples for which the correct instances are retrieved within the top-K nearest to the query [35]. We report R@1, R@5, and R@10 for both Flickr30k and MS COCO as in Reference [31]. And, we also compute an additional criterion "R@sum" to evaluate the overall performance for both "text retrieval given image query" and "image retrieval given text query" as follows:

$$R@sum = \underbrace{R@1 + R@5 + R@10}_{\text{Text Retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{Image Retrieval}}. \tag{23}$$

Model Implementation. Pytorch [38] is adopted to implement our model. (1) For both datasets, we employ a Faster R-CNN model pre-trained on Visual Genome [27] as the image feature encoder. For each image, we select the top 36 regions of interest (ROIs) with the highest detection confidence scores, and then we use average pooling to extract salient features. Each image region is encoded into a 2,048-dimensional feature vector, i.e., $d_f = 2048$ in Equation (1). And thus an image is represented by a 36-by-2048 matrix. The feature vector of each region is then mapped into a

Table 1. Performance of the Proposed Models for Human-behavior-related Cross-modality Image-text Retrieval

Models	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Flickr30k-human							
SCAN i-t AVG	68.2	89.5	94.4	44.9	74.8	83.0	454.8
ACMNet-GI	**70.7**	91.4	95.5	50.3	77.9	85.3	471.1
SCAN t-i AVG	62.8	89.5	94.4	45.6	74.5	82.7	449.5
ACMNet-GT	67.8	**92.4**	**96.3**	**54.0**	**79.5**	**86.3**	**476.3**
MS COCO-human							
SCAN i-t AVG	48.6	76.8	86.5	29.5	60.4	72.7	374.5
ACMNet-GI	52.1	80.0	88.9	40.0	68.7	79.2	408.9
SCAN t-i AVG	50.1	79.3	87.7	37.2	66.3	77.8	398.4
ACMNet-GT	**54.3**	**81.5**	**89.3**	**40.8**	**69.1**	**79.8**	**414.8**

1,024-dimensional feature vector, i.e., $d = 1{,}024$ in Equation (1), resulting in a 1,024-dimensional feature vector for the global image feature. (2) We use a bi-directional GRU with one layer to learn the text representation. The dimensionality of its hidden state is set as 1,024, resulting in a dimensionality of 1,024 for the global text feature vector, i.e., $\bar{t}$. The dimensionality of the word embedding is set as 300.

Parameter Setting. During training, we set the margin of the hinge-based triplet loss, i.e., Δ in Equations (20) and (22) as 0.2, as in [31]. The inverse temperature factors, i.e., λ_1 in Equation (9) and λ_2 in Equation (15), are set via grid search in predefined value ranges on the validate set. For Flickr30k, the learning rate is set as 2e-4 in the begining. And after 15 epochs, it is set as 2e-5 for another 15 epochs. As for MS COCO, the learning rate is set as 5e-4 for the beginning 10 epochs and 5e-5 for the subsequent 10 epochs. The mini-batch size is set as 128 for all models.

Testing Setting. Following previous works [14, 31], we save the best model with the highest R@sum on the validation sets during training for our proposed ACMNet. Performance validation is performed every epoch on the validation set. After training, we evaluate the saved best models on the test set and report the result for comparisons with other methods.

4.3 Cross-modality Human Behavior Analysis via Image-Text Retrieval

To evaluate the retrieval performance of our cross-modality image-text retrieval models for human behavior analysis, we collect a fraction of human-behavior-related image-text pairs in Flickr30k and MS COCO for evaluation, denoted as Flickr30k-human and MS COCO-human, respectively. We use human-related key words, like "person," "man," "woman," "men," "women," "kid," "child," "boy," and "girl," to find image-text pairs from the test sets of both benchmark datasets, ending up with 842 images for Flickr30k-human and 2,334 images for MS COCO-human. Note that SCAN i-t AVG [31] calculates local similarity scores in the image-grounded embedding space, which is similar to ACMNet-GI. And SCAN t-i AVG [31] calculate local similarities in the text-grounded embedding space, which is similar to our ACMNet-GT. Therefore, we regard SCAN i-t AVG and SCAN t-i AVG as base models for ACMNet-GI and ACMNet-GT, respectively.

Table 1 shows the performance of our ACMNets for the human-behavior-related cross-modality image-text retrieval. We can see that our ACMNet-GI and ACMNet-GT can consistently outperform the base models on both Flickr30k-human and MS COCO-human. Specifically, compared with SCAN i-t AVG, our ACMNet-GI can achieve a maximum performance improvement of 16.3% and 34.4% in terms of R@sum for Flickr30k-human and MS COCO-human, respectively. And

Table 2. Comparison of the Cross-modality Image-text Retrieval Performances in Terms of Recall@K (*R@K*) and R@sum on Flickr30k

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [24]	22.2	48.2	61.4	15.2	37.7	50.5	235.2
HM-LSTM [37]	38.1	-	76.5	27.7	-	68.8	-
SM-LSTM [19]	42.5	71.9	81.5	30.2	60.4	72.3	358.8
Webly [35]	47.4	-	85.9	35.2	-	74.8	-
2WayNet [13]	49.8	67.5	-	36.0	55.6	-	-
VSE++ [14]	52.9	-	87.2	39.6	-	79.5	-
DAN [36]	55.0	81.8	89.0	39.4	69.2	79.1	413.5
DPC [61]	55.6	81.9	89.5	39.1	69.2	80.9	416.2
SCO [20]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
SCAN [31]	**67.9**	89.0	94.4	43.9	74.2	82.8	452.0
ACMNet	66.0	**90.7**	**95.8**	**51.6**	**78.0**	**85.8**	**467.9**

- means unreported results.

comparing our ACMNet-GT with SCAN t-i-AVG, the improvement is 26.8% and 16.4% in terms of R@sum for Flickr30k-human and MS COCO-human, respectively. Considering that the proposed ACMNet is based on SCAN, the experimental results well demonstrate that the proposed ACMNet can help to tackle the mentioned similarity bias issue and thus gain superior performance.

4.4 Comparison with State-of-the-Art on Benchmark Datasets

In this section, following References [31, 35], we compare our best model with the state-of-the-art published models for the tasks of cross-modality image-text retrieval, including DVSA [24], HM-LSTM [37], Order-embeddings [49], SM-LSTM [19], 2WayNet [13], DAN [36], VSE++ [14], DPC [61], GXN [17], CHAIN-VSE [51], SCO [20], SCAN [31], and Webly [35]. We directly cite reported performances of compared methods when available. For unreported metrics, we mark them with "–". The additional metric R@sum is calculated only when all six Recall@K metrics are provided.

Results. Tables 2 and 3 show the comparison results on Flickr30k and MS COCO, respectively. For both datasets, we denote our best single model as ACMNet. Details about our single models can be referred to Section 4.5.

Experimental results on both datasets indicate that the proposed adaptive confidence matching network (ACMNet) can obtain substantial performance improvement over state-of-the-art methods.

Specifically, on Flickr30k in Table 2, our model (i.e., ACMNet) can outperform all baselines in nearly all evaluation metrics. Particularly, compared with the best baseline SCAN, our model can achieve a maximum performance improvement of 1.7% in terms of R@5 for the task of text retrieval given image query, and 7.7% in terms of R@1 for the task of image retrieval task given text query. And as the overall performance, our model can achieve 467.9% in terms of R@sum, significantly outperforming SCAN by a large margin of 15.9%.

Moreover, on MS COCO in Table 3, compared with the state-of-the-art retrieval model, i.e., SCAN [31], our ACMNet can achieve significantly better performances in terms of all evaluation metrics. Particularly, our model can outperform SCAN by a maximum margin 1.3% in terms of R@1 for the task of text retrieval given image query. As for the task of image retrieval given text query, our model can also gain a maximum performance improvement of 2.8% in terms of R@1.

Table 3. Comparison of the Cross-modality Image-text Retrieval Performances in Terms of Recall@K ($R@K$) and R@sum on MS COCO

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [24]	38.4	69.9	80.5	27.4	60.2	74.8	351.2
HM-LSTM [37]	43.9	-	87.8	36.1	-	86.7	-
Order-embeddings [49]	46.7	-	88.9	37.9	-	85.9	-
SM-LSTM [19]	53.2	83.1	91.5	40.7	75.8	87.4	431.7
2WayNet [13]	55.8	75.2	-	39.7	63.3	-	-
Webly [35]	61.5	-	96.1	46.3	-	89.4	-
VSE++ [26]	64.6	-	95.7	52.0	-	92.0	-
DPC [61]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
CHAIN-VSE [51]	59.4	88.0	94.2	43.5	79.8	90.2	455.1
GXN [17]	68.5	-	97.9	56.6	-	94.5	-
SCO [20]	69.9	92.9	97.5	56.7	87.5	**94.8**	499.3
SCAN [31]	70.9	94.5	97.8	56.4	87.0	93.9	500.5
ACMNet	**72.1**	**95.2**	98.1	**59.2**	**88.1**	94.4	**507.1**

- means unreported results.

And in terms of R@sum, our model can achieve 507.1%, significantly outperforming SCAN by a large margin of 6.6%.

4.5 Model Analyses

In this section, we first analyze the effect of the proposed adaptive confidence matching and make comparisons with the corresponding base models in Reference [31]. Then, we analyze different factors to observe their effects on the performance, including inverse temperature factor λ_1 and λ_2, feature normalization and loss function. Results are shown in Table 4 for Flickr30k and Table 5 for MS COCO. And in Tables 4 and 5, we use ‡ to mark models with the same experiment settings as the corresponding base model of SCAN, which is the default settings. And models without ‡ only differ in the factor to be analyzed. For example, in Table 4, for ACMNet-GI, by default, λ_1=4; norm=both; loss=VSEPP. And loss=VSE means that only the training loss is changed to VSE, while other factors are the same as the default setting.

Effect of adaptive confidence matching. Under the default settings, we can verify the effectiveness and the superiority of the proposed ACMNet by directly comparing our models (i.e., ACMNet-GI and ACMNet-GT) with their corresponding base models in SCAN [31] (i.e., SCAN i-t AVG and SCAN t-i AVG). From Tables 4 and 5, we can see that under the same default experiment settings, both ACMNet-GI and ACMNet-GI (i.e., models with ‡) can obtain substantially better performance than base models at nearly all metrics. Specifically, on Flickr30k, compared with SCAN i-t AVG, ACMNet-GI can gain a maximum improvement of 2.9% in terms of R@sum. And compared with SCAN t-i AVG, ACMNet-GT can obtain a maximum improvement of 14.9% in terms of R@sum. On MS COCO, the same results can be observed. ACMNet-GI can gain a maximum improvement of 3.0% in terms of R@sum, compared with SCAN i-t AVG. And ACMNet-GT can obtain a maximum improvement of 4.0% in terms of R@sum, compared with SCAN t-i AVG. Performance improvements are attributed to the proposed adaptive confidence matching strategy for aggregating the local similarities into the global similarity.

Effect of inverse temperature factor λ_1 and λ_2. For both datasets, λ_1 for ACMNet-GI refers to the inverse temperature factor in Equation (9) and λ_2 for ACMNet-GT refers to the one in Equation (15). And, we list the best factor for both models compared with the default settings. We

Table 4. Factor Analyses of the Proposed ACMNet, in Terms of Recall@K (*R@K*) and R@sum on Flick30k

Setting	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
ACMNet-GI							
SCAN i-t AVG [31]	67.9	89.0	94.4	43.9	74.2	82.8	452.0
λ_1=1	58.7	84.8	91.2	38.3	69.5	79.5	422.0
λ_1=4‡	66.5	90.2	94.5	45.5	74.7	83.5	454.9
λ_1=6	**68.3**	**90.8**	**95.4**	**49.2**	**77.2**	**85.3**	**466.2**
norm=text	65.8	89.7	93.7	44.2	74.7	83.7	451.8
norm=image	63.7	87.5	93.9	42.4	72.7	82.6	442.8
norm=no	64.6	88.9	94.5	42.8	73.8	82.5	447.1
norm=both‡	66.5	90.2	94.5	45.5	74.7	83.5	454.9
VSE	53.6	83.8	90.9	39.8	70.1	79.9	418.1
VSEPP‡	66.5	90.2	94.5	45.5	74.7	83.5	454.9
ACMNet-GT							
SCAN t-i AVG [31]	61.8	87.5	93.7	45.8	74.4	83.0	446.2
λ_2=1	56.7	82.8	90.1	42.1	71.2	80.1	423.0
λ_2=9‡	64.8	89.8	95.6	48.9	77.2	84.8	461.1
λ_2=11	65.5	**90.9**	**95.8**	49.9	77.0	85.1	464.2
norm=text	64.4	89.4	95.1	47.0	75.8	83.7	455.4
norm=image	66.0	90.7	**95.8**	**51.6**	**78.0**	**85.8**	**467.9**
norm=no	**66.5**	90.1	95.4	49.8	77.4	84.8	464.0
norm=both‡	64.8	89.8	95.6	48.9	77.2	84.8	461.1
VSE	49.9	81.3	89.1	38.7	66.3	75.1	400.4
VSEPP‡	64.8	89.8	95.6	48.9	77.2	84.8	461.1

Models with ‡ use the same experiment settings as SCAN [31].

can see that for both datasets, with proper λ_1 and λ_2, both models can obtain superior performance than with the default λ_1 or λ_2. Specifically, for ACMNet-GI, the best performance is achieved by λ_1=6 for both Flickr30k and MS COCO. And for ACMNet-GT, the best performance is achieved by λ_2=11 for both datasets. We can also see that when this factor is deactivated, i.e., setting it as 1, the performance degrades greatly, which is consistent with the observation in Reference [31].

Effect of feature normalization. Feature normalization has been proved to be effective to improve the performance in previous works [14, 26, 31]. In Tables 4 and 5, we denote this factor as norm. Following References [14, 26, 31], norm=image means only image features are normalized. norm=text means only text features are normalized. norm=no indicates that no features are normalized. norm=both means both image features and text features are normalized. And cosine function is used as the normalization function as References [14, 26, 31], seeing Equation (7). We can see that, on both datasets, normalizing both image features and text features is the best strategy for ACMNet-GI. While for ACMNet-GT, only normalizing the image features is a better choice.

Effect of loss function. Similar to previous works [14, 31], loss functions have a great impact on both ACMNet-GI and ACMNet-GT. On both datasets, models trained with VSEPP can outperform those trained by VSE with a large margin in terms of all evaluation metrics.

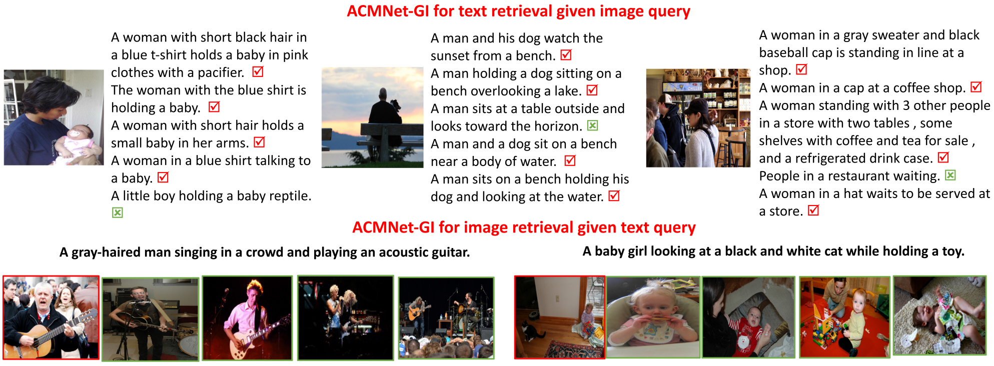
4.6 Human Behavior Analysis via Cross-modality Image-Text Retrieval

Figures 4 and 5 show some qualitative results of human-behavior-related cross-modality image-text retrieval, generated by our ACMNet-GI and ACMNet-GT, respectively. For each query, we

Table 5. Factor Analyses of the Proposed ACMNet, in Terms of Recall@K (*R@K*) and R@sum on MS COCO

Setting	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
ACMNet-GI							
SCAN i-t AVG [31]	69.2	93.2	97.5	54.4	86.0	93.6	493.9
λ_1=1	65.9	91.4	96.5	45.6	80.9	90.5	470.8
λ_1=4‡	69.8	**94.2**	97.6	55.4	86.3	93.6	496.9
λ_1=6	**71.0**	93.9	**98.0**	**57.5**	**87.2**	**93.7**	**501.3**
norm=text	68.5	92.7	97.2	49.8	83.3	91.8	483.3
norm=image	68.9	92.9	97.3	50.7	83.8	92.2	485.8
norm=no	68.5	92.0	97.2	50.1	83.5	92.0	483.3
norm=both‡	69.8	94.2	97.6	55.4	86.3	93.6	496.9
VSE	60.3	89.0	95.2	43.3	79.6	90.0	457.4
VSEPP‡	69.8	94.2	97.6	55.4	86.3	93.6	496.9
ACMNet-GT							
SCAN t-i AVG [31]	70.9	94.5	97.8	56.4	87.0	93.9	500.5
λ_2=1	66.1	91.9	96.7	52.2	84.2	92	483.1
λ_2=9‡	72.0	94.9	98.1	58.1	87.5	93.9	504.5
λ_2=11	**73.1**	94.8	98.1	58.9	87.8	94.2	506.9
norm=text	69.9	93.9	97.8	56.2	86.8	93.7	498.3
norm=image	72.1	**95.2**	**98.1**	**59.2**	**88.1**	**94.4**	**507.1**
norm=no	69.3	93.8	97.5	55.5	86.2	93.3	495.6
norm=both‡	72.0	94.9	98.1	58.1	87.5	93.9	504.5
loss=VSE	66	92.9	97.4	54.8	86.1	93.5	490.7
loss=VSEPP‡	72.0	94.9	98.1	58.1	87.5	93.9	504.5

Models with ‡ use the same experiment settings as SCAN [31].



ACMNet-GI for text retrieval given image query

A woman with short black hair in a blue t-shirt holds a baby in pink clothes with a pacifier. ☑
The woman with the blue shirt is holding a baby. ☑
A woman with short hair holds a small baby in her arms. ☑
A woman in a blue shirt talking to a baby. ☑
A little boy holding a baby reptile. ☒

A man and his dog watch the sunset from a bench. ☑
A man holding a dog sitting on a bench overlooking a lake. ☑
A man sits at a table outside and looks toward the horizon. ☒
A man and a dog sit on a bench near a body of water. ☑
A man sits on a bench holding his dog and looking at the water. ☑

A woman in a gray sweater and black baseball cap is standing in line at a shop. ☑
A woman in a cap at a coffee shop. ☑
A woman standing with 3 other people in a store with two tables , some shelves with coffee and tea for sale , and a refrigerated drink case. ☑
People in a restaurant waiting. ☒
A woman in a hat waits to be served at a store. ☑

ACMNet-GI for image retrieval given text query

A gray-haired man singing in a crowd and playing an acoustic guitar.

A baby girl looking at a black and white cat while holding a toy.

Fig. 4. Examples of the retrieval results for our ACMNet-GI on Flickr30k-human. Ground-truth matched retrieval results are marked in red, while unmatched ones are marked in green. Best viewed in color.

show the top-5 retrieved results for both models. We can see that both ACMNet-GI and ACMNet-GT can usually retrieve majorities of matched texts within top-5 for given image queries. And given a text query, both models can rank the most related images at the top. Such results qualitatively indicate the effectiveness of the proposed adaptive confidence matching network for human behavior analysis.

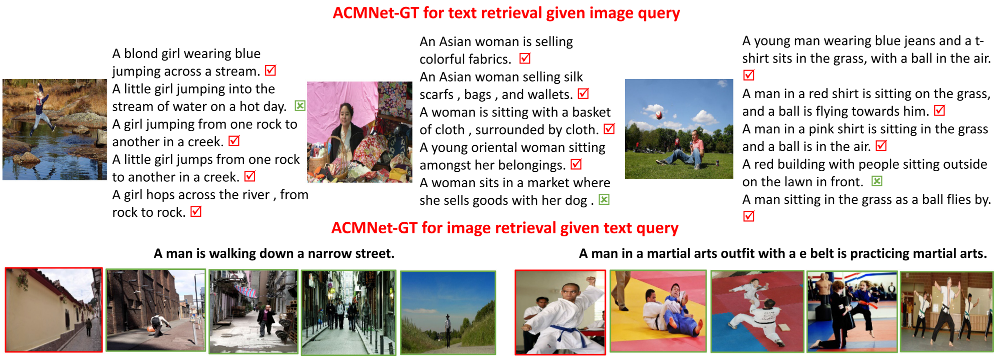

ACMNet-GT for text retrieval given image query
A blond girl wearing blue jumping across a stream. ☑
A little girl jumping into the stream of water on a hot day. ☒
A girl jumping from one rock to another in a creek. ☑
A little girl jumps from one rock to another in a creek. ☑
A girl hops across the river , from rock to rock. ☑
An Asian woman is selling colorful fabrics. ☑
An Asian woman selling silk scarfs , bags , and wallets. ☑
A woman is sitting with a basket of cloth , surrounded by cloth. ☑
A young oriental woman sitting amongst her belongings. ☑
A woman sits in a market where she sells goods with her dog . ☒
A young man wearing blue jeans and a t-shirt sits in the grass, with a ball in the air. ☑
A man in a red shirt is sitting on the grass, and a ball is flying towards him. ☑
A man in a pink shirt is sitting in the grass and a ball is in the air. ☑
A red building with people sitting outside on the lawn in front. ☒
A man sitting in the grass as a ball flies by. ☑
ACMNet-GT for image retrieval given text query
A man is walking down a narrow street.
A man in a martial arts outfit with a e belt is practicing martial arts.

Fig. 5. Examples of the retrieval results for our ACMNet-GT on Flickr30k-human. Ground-truth matched retrieval results are marked in red, while unmatched ones are marked in green. Best viewed in color.

5 CONCLUSION

In this article, we focus on the cross-modality image-text retrieval for human behavior analysis. We show a similarity bias issue for existing state-of-the-art fine-grained cross-modality matching methods. And, we propose an adaptive confidence matching network (ACMNet) to deal with it for cross-modality image-text retrieval. As in Reference [31], we first compute local similarities between image regions and text words in two grounded embedding spaces, i.e., the image-grounded and text-grounded embedding spaces. Then instead of directly aggregating these local similarities into the global similarity, we adopt a gate function to utilize the global feature of the image/text to estimate a confidence score of each local similarity, and further incorporate them into the calculation of the global similarity, to tackle the mentioned similarity bias issue. We verify the proposed ACMNet through extensive experiments and comparisons with other state-of-the-art approaches on two benchmark datasets, i.e., Flickr30k and MS COCO. Experimental results show that the proposed ACMNet can achieve state-of-the-art performance of cross-modality image-text retrieval on both datasets. In the future, we will study the possibility of our method applied to other cross-modality tasks, e.g., image-text hashing [53], video-text retrieval [12], temporal activity localization [32], and so on.

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.* 6.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision.* 2425–2433.

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

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5 (2017), 135–146.

[5] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2018. Show, observe and tell: Attribute-driven attention model for image captioning. In *Proceedings of the International Joint Conferences on Artificial Intelligence.* 606–612.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14).* 1724–1734.

[7] Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems.* MIT Press, 577–585.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *Arxiv Preprint Arxiv:1810.04805* (2018).

[10] Guiguang Ding, Wenshuo Chen, Sicheng Zhao, Jungong Han, and Qiaoyan Liu. 2018. Real-time scalable visual tracking via quadrangle kernelized correlation filters. *IEEE Trans. Intell. Transport. Syst.* 19, 1 (2018), 140–150.

[11] Guiguang Ding, Yuchen Guo, Kai Chen, Chaoqun Chu, Jungong Han, and Qionghai Dai. 2019. DECODE: Deep confidence network for robust image classification. *IEEE Transactions on Image Processing* 28, 8 (2019), 3752–3765.

[12] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19).*

[13] Aviv Eisenschtat and Lior Wolf. 2017. Linking image and text with 2-way nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4601–4611.

[14] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *Arxiv Preprint Arxiv:1707.05612* (2017).

[15] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems.* MIT Press, 2121–2129.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 580–587.

[17] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7181–7189.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[19] 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.

[20] 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.

[21] Guiguang Ding, Jianguang Lou, Yusen Zhang, Hui Chen, Zijia Lin, and Borje Karlsson. 2019. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI'19).*

[22] Feng Jiang, Shengping Zhang, Shen Wu, Yang Gao, and Debin Zhao. 2015. Multi-layered gesture recognition with kinect. *J. Mach. Learn. Res.* 16, 8 (2015), 227–254. Retrieved from http://jmlr.org/papers/v16/jiang15a.html.

[23] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* Association for Computational Linguistics, 427–431.

[24] 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.

[25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems.* MIT Press, 1571–1581.

[26] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *Arxiv Preprint Arxiv:1411.2539* (2014).

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, and Justin, et al. Johnson. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* 123, 1 (2017), 32–73.

[28] Xiangyuan Lan, Andy Jinhua Ma, Pong C. Yuen, and Rama Chellappa. 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE Trans. Image Process.* 24, 12 (2015), 5826–5841.

[29] Xiangyuan Lan, Mang Ye, Rui Shao, Bineng Zhong, Pong C. Yuen, and Huiyu Zhou. 2019. Learning modality-consistency feature templates: A robust rgb-infrared tracking system. *IEEE Trans. Industr. Electron.* 66, 12 (2019), 9887–9897. DOI: 10.1109/TIE.2019.2898618

[30] Xiangyuan Lan, Shengping Zhang, Pong C. Yuen, and Rama Chellappa. 2018. Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker. *IEEE Trans. Image Process.* 27, 4 (2018), 2022–2037.

[31] 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'18).* 201–216.

[32] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18).* ACM, New York, NY, 15–24. DOI: https://doi.org/10.1145/3209978.3210003

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR'13)*.

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. MIT Press, 3111–3119.

[35] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E. Papalexakis, and Amit K. Roy-Chowdhury. 2018. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. ACM, New York, NY, 1856–1864. DOI: https://doi.org/10.1145/3240508.3240712

[36] 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.

[37] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. 2017. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 1899–1907. DOI: https://doi.org/10.1109/ICCV.2017.208

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of Neural Information Processing Systems*.

[39] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2017. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *Arxiv Preprint Arxiv:1710.05106* (2017).

[40] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

[41] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[42] 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. In *Proceedings of the IEEE International Conference on Computer Vision*. 2641–2649.

[43] Yuankai Qi, Shengping Zhang, Lei Qin, Qingming Huang, Hongxun Yao, Jongwoo Lim, and Ming-Hsuan Yang. 2019. Hedging deep features for visual tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 5 (2019), 1116–1130.

[44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. MIT Press, 91–99.

[46] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *Arxiv Preprint Arxiv:1409.1556* (2014).

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. MIT Press, 5998–6008.

[49] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *Arxiv Preprint Arxiv:1511.06361* (2015).

[50] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5005–5013.

[51] Janatas Wehrmann and Rodrigo C. Barros. 2018. Bidirectional retrieval made simple. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*.

[52] Gengshen Wu, Jungong Han, Yuchen Guo, Li Liu, Guiguang Ding, Qiang Ni, and Ling Shao. 2018. Unsupervised deep video hashing via balanced code for large-scale video retrieval. *IEEE Trans. Image Process.* 28, 4 (2018), 1993–2007.

[53] Gengshen Wu, Jungong Han, Zijia Lin, Guiguang Ding, Baochang Zhang, and Qiang Ni. 2018. Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning. *IEEE Trans. Industr. Electron.* 66, 12 (2018), 9868–9877.

[54] 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 *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[55] Chenggang Yan, Liang Li, Chunjie Zhang, Bingtao Liu, Yongdong Zhang, and Qionghai Dai. 2019. Cross-modality bridging and knowledge transferring for image understanding. *IEEE Trans. Multimed.* 21, 10 (2019), 2675–2685.

[56] C Yan, Y. Tu, X Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai. 2020. STAT: spatial-temporal attention mechanism for video captioning. *IEEE Trans. Multimed.* 22, 1 (Jan 2020), 229–241. https://doi.org/10.1109/TMM.2019.2924576

[57] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3441–3450.

[58] Shengping Zhang, Xiangyuan Lan, Yuankai Qi, and Pong C. Yuen. 2017. Robust visual tracking via basis matching. *IEEE Trans. Circ. Syst. Video Technol.* 27, 3 (2017), 421–430.

[59] Shengping Zhang, Xiangyuan Lan, Hongxun Yao, Huiyu Zhou, Dacheng Tao, and Xuelong Li. 2017. A biologically inspired appearance model for robust visual tracking. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 10 (2017), 2357–2370.

[60] Shengping Zhang, Huiyu Zhou, Feng Jiang, and Xuelong Li. 2015. Robust visual tracking using structurally random projection and weighted least squares. *IEEE Trans. Circ. Syst. Video Technol.* 25, 11 (2015), 1749–1760.

[61] 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).

[62] Andrej Zukov-Gregoric, Yoram Bachrach, Pasha Minkovsky, Sam Coope, and Bogdan Maksak. 2017. Neural named entity recognition using a self-attention mechanism. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence.* 652–656.

Received April 2019; revised July 2019; accepted September 2019