Multimodal Transformer with Multi-View Visual Representation for Image Captioning

Jun Yu, *Member, IEEE*, Jing Li, Zhou Yu, *Member, IEEE*, Qingming Huang, *Fellow, IEEE*

***Abstract*—Image captioning aims to automatically generate a natural language description of a given image, and most state-of-the-art models have adopted an encoder-decoder framework. The framework consists of a convolution neural network (CNN)-based image encoder that extracts region-based visual features from the input image, and an recurrent neural network (RNN) based caption decoder that generates the output caption words based on the visual features with the attention mechanism. Despite the success of existing studies, current methods only model the co-attention that characterizes the inter-modal interactions while neglecting the self-attention that characterizes the intra-modal interactions. Inspired by the success of the Transformer model in machine translation, here we extend it to a Multimodal Transformer (MT) model for image captioning. Compared to existing image captioning approaches, the MT model simultaneously captures intra- and inter-modal interactions in a unified attention block. Due to the in-depth modular composition of such attention blocks, the MT model can perform complex multimodal reasoning and output accurate captions. Moreover, to further improve the image captioning performance, multi-view visual features are seamlessly introduced into the MT model. We quantitatively and qualitatively evaluate our approach using the benchmark MSCOCO image captioning dataset and conduct extensive ablation studies to investigate the reasons behind its effectiveness. The experimental results show that our method significantly outperforms the previous state-of-the-art methods. With an ensemble of seven models, our solution ranks the 1st place on the real-time leaderboard of the MSCOCO image captioning challenge at the time of the writing of this paper.**

***Index Terms*—Image captioning, multi-view learning, deep learning.**

I. Introduction

Recent advances in deep learning have resulted in great progress in both the computer vision and natural language processing communities. These achievements make it possible to connect vision and language, and facilitate multi-modal learning tasks such as image-text matching [1], visual question answering [2][3][4], visual grounding [5] and image captioning [6][7][8][9][10].

Image captioning aims to automatically describe an image's content using a natural language sentence. The task is challenging since it requires one to recognize key objects in an image, and to understand their relationships with each other. Most successful image captioning approaches adopt the encoder-decoder framework, which is inspired by the sequence-to-sequence model for machine translation [11]. The framework consists of a convolutional neural network (CNN)-based image encoder that extracts region-based visual features from an input image, and an recurrent neural network (RNN) based caption decoder that iteratively generates the output caption words based on the visual features. The encoder-decoder model is usually trained in an end-to-end manner to minimize the cross-entropy loss. Based on the framework, plenty of improvements have been made by recent works to further improve image captioning performance further. For instance, to establish the fine-grained connections of caption words and their related image regions, an attention mechanism can be seamlessly inserted into the framework [7]. To provide a better understanding of the objects in the image, region-based bottom-up-attention features can be extracted from a pre-trained object detector to replace the traditional CNN convolutional features [6]. To address the exposure bias of generated captions by using the cross-entropy loss, reinforcement learning (RL)-based algorithms are designed to directly optimize the non-differentiable evaluation metrics (*e.g.,* BLEU [12] and CIDEr [13]) [10].

Despite the success that existing approaches have achieved, they have the following limitations: 1) the current attention mechanism in image captioning only models the *co-attention* that characterizes inter-modal interactions (*i.e.*, object-to-word) while neglecting the *self-attention* that characterizes intra-modal interactions (*i.e.*, word-to-word and object-to-object); 2) current image captioning models are usually *shallow* and may fail to fully understand the complex relationships among visual objects; and 3) the region-based visual features may fail to cover all objects in the image, leading to insufficient visual representations for generating accurate captions.

To address the first and second limitations, we extend the Transformer model for machine translation [14] to a Multimodal Transformer (MT) model for image captioning. Different from the CNN-RNN captioning models, the MT model does not use RNN and instead relies entirely on an attention mechanism to assess the global dependencies between the input and output. By properly stacking such attention blocks in depth, MT forms a deep encoder-decoder model that simultaneously captures the self-attention within each modality and the co-attention across different modalities. To address the last limitation, we introduce multi-view feature learning into the MT model to adapt both the aligned and unaligned multi-view visual features.

This work was supported in part by National Natural Science Foundation of China under Grant 61702143 and Grant 61836002. (Corresponding author: Zhou Yu.)

J. Yu, J. Li and Z. Yu are with Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China (e-mail: yujun@hdu.edu.cn; jingli@hdu.edu.cn; yuz@hdu.edu.cn).

Q. Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (email: qmhuang@ucas.ac.cn).

To summarize, the main contributions of this study are three-fold:

- The joint modeling of the self-attention and the co-attention interactions for image captioning is first proposed in the MT model. The MT model is capable of modeling three types of relationships using a modular attention block, *i.e.,* word-to-word, object-to-object, and word-to-object. By stacking such attention blocks in depth, the deep MT model significantly outperforms the state-of-the-art models, thereby highlighting the importance of deep reasoning for image captioning.
- Multi-view learning on the image is introduced in conjunction with the MT model to provide more diverse and discriminative visual representations. We introduce two alternative strategies to handle aligned and unaligned multi-view features, respectively.
- Extensive experiments on the benchmark MSCOCO image captioning dataset are conducted to quantitatively and qualitatively prove the effectiveness of the proposed models. The experimental results show that the MT significantly outperforms previous state-of-the-art approaches with a single model. Furthermore, our solution ranks the 1st place on the real-time leaderboard of the MSCOCO image captioning challenge with an ensemble of the proposed MT models.
- The proposed multi-view image representation strategy can be easily applied to other tasks like VQA and visual grounding. We conduct experiments on the benchmark VQA and visual grounding datasets and obtain significant improvement over the existing state-of-the-art methods with single-view features.

The rest of the paper is organized as follows: In section II, we review the related work of image captioning approaches, especially the ones introducing attention mechanisms. In section III, we revisit the basic Transformer model and then propose the Multimodal Transformer model for image captioning. In section IV, we introduce multi-view image representation into the MT model to increase the visual representation capacity, and the quality of the generated captions. In section V, we introduce our extensive experimental results for algorithm evaluation and use the benchmark MSCOCO image captioning dataset to evaluate our proposed approaches. Finally, we conclude this work in section VI.

II. Related Work

In this section, we briefly review the most relevant research on image captioning, especially those studies that introduce attention models.

A. Image Captioning

The research on image captioning can be categorized into the following three classes: template-based approaches [15][16][17], retrieval-based approaches [18][19][20], and generation-based approaches [10][9][21][6][22].

The template-based approaches address the task using a two-stage strategy: 1) align the sentence fragments (*e.g.*, subject, object, and verb) with the predicted labels from the image; and 2) generate the sentence from the segments using pre-defined language templates. Kulkarni *et al.* use the conditional random field (CRF) model to predict labels based on the detected objects, attributes, and prepositions, and then generate caption sentences with a template by filling in the blanks with the most likely labels [15]. Yang *et al.* employ the HMM model to select the best objects, verbs, and prepositions with respect to the log-likelihood for segments generation [17]. Intuitively, the captions that are generated by the template-based approaches highly depend on the quality of the templates and usually follow the syntactical structures. However, the diversity of the generated captions is severely restricted.

To ease the diversity problem, retrieval-based approaches are proposed to *search* the most relevant captions from a large-scale caption database with respect to their cross-modal similarities to the given image. Karpathy *et al.* propose a deep fragment embedding approach to match the image-caption pairs based on the alignment of visual segments (the detected objects) and caption segments (subjects, objects, and verbs) [18]. In the testing stage, the cross-modal matching over the whole caption database (usually the captions from the training set) is performed to generate the caption for one image. Other methods such as [19][20] use different metrics or loss functions to learn the cross-modal matching model. However, the retrieval efficiency becomes a bottleneck for these approaches when the caption database is large and restricting the size of the database may reduce the caption diversity. Moreover, retrieval-based approaches cannot generate novel captions beyond the database, which means the diversity problem has not been completely resolved.

Different from template-based and retrieval-based models, generation-based models aim to learn a language model that can generate novel captions with more flexible syntactical structures. With this purpose, recent works explore this direction by introducing the neural networks for image captioning. Vinyals *et al.* propose an encoder-decoder architecture by utilizing the GoogLeNet [23] and LSTM networks [24] as its backbones. Similar architectures are also proposed by Donahue *et al.* [25] and Karpathy *et al.* [26]. Due to the flexibility and excellent performance, generation-based models have become the mainstream for image captioning.

B. Attention Mechanism

Within the encoder-decoder framework, one of the most important improvements for generation-based models is the attention mechanism. Xu *et al.* introduce the soft and hard attention models to mimic the human eye focusing on different regions in an image when generating different caption words. The attention model is a *pluggable* module that can be seamlessly inserted into previous approaches to remarkably improve the caption quality. The attention model is further improved in [6][27][9][10]. Chen *et al.* propose a spatial- and channel-wise attention model to attend to visual features [27]. Lu *et al.* present an adaptive attention encoder-decoder model for automatically deciding when to rely on visual or language signals [9]. Rennie *et al.* design a FC model and an Att2in model that achieve good performance [10]. Anderson *et al.*

introduce a bottom-up module that uses a pre-trained object detector to extract region-based image features containing potential objects, and a top-down module that utilizes soft attention to dynamically attend to these object [6]. Compared to the commonly used grid-based convolutional features in image captioning, replacing them with the region-based detector features can bring significant performance improvement. The bottom-up module has become a de-facto component in the subsequent image captioning researches [22][28] and other related tasks that requires fine-grained image understanding [5][3][29].

Beyond the image captioning tasks, attention mechanisms are widely used in other multi-modal learning tasks such as visual question answering (VQA). Lu *et al.* propose a co-attention learning framework to alternately learn the image attention and question attention [30]. Yu *et al.* reduce the co-attention method into two steps, self-attention for a question embedding and the question-conditioned attention for a visual embedding [31]. Nam *et al.* propose a multi-stage co-attention learning model to refine the attentions based on the memory of previous attentions [32]. However, these co-attention models learn separate attention distributions for each modality (image or question) and neglect the dense interaction between each question word and each image region, which becomes a bottleneck for understanding the fine-grained relationships of multimodal features. To address this issue, dense co-attention models have been proposed, which establish the complete interaction between each question word and each image region [33][3]. Compared to the previous co-attention models with coarse interactions, the dense co-attention models deliver significantly better VQA performance.

III. Multimodal Transformer

In this section, we first briefly describe the preliminary knowledge of the Transformer model [14]. Then, we introduce the proposed Multimodal Transformer (MT) framework for image captioning, which consists of an *image encoder* and a *caption decoder*. The image encoder learns the deep image representation in a self-attention manner, and then, the caption decoder uses the attended image representations to generate textual captions.

Before presenting the MT model, we first introduce its basic components, the multi-head attention (MHA) and the feed-forward networks (FFN), which was first proposed in the Transformer model [14] for machine translation

A. MHA and FFN

Multi-head attention is a natural extension of the *scaled dot-product attention*, a general attention model depicts the interactions among a group of queries, keys, and values.

The input of the scaled dot-product attention consists of a query $q \in \mathbb{R}^d$, a set of keys $k_t \in \mathbb{R}^d$ and values $v_t \in \mathbb{R}^d$, where $t \in \{1, 2, ..., n\}$ is the number of key-value pairs and d is the common dimensionality of all the inputs features. We calculate the dot products of query with all keys, divide each by $\sqrt{d}$ and apply a softmax function to obtain the attention weights on the values. In practice, we pack all the keys and values into matrices $K = [k_1, ..., k_n] \in \mathbb{R}^{n\times d}$ and $V = [v_1, ..., v_n] \in \mathbb{R}^{n\times d}$ respectively. The attention function on a set of queries $Q = [q_1, ..., q_m] \in \mathbb{R}^{m\times d}$ can be computed in parallel as follows:

$$F = A(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where $F \in \mathbb{R}^{m\times d}$ correspond to the attended features of the queries Q.

Instead of performing a single attention function for the queries, multi-head attention (MHA) extends the scale-dot-product attention model and introduce multiple attention functions in parallel to model diverse information from different representation subspaces. The multi-head attention contains h parallel 'heads' with each head corresponding to an independent scaled dot-product attention function. The attended features F of the multi-head attention functions is given as follows:

$$F = \text{MHA}(Q, K, V) = \text{Concat}(h_1, ..., h_h)W^o \tag{2}$$

$$h_i = A(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_h}$ are the projection matrices of the i-th head. $W^O \in \mathbb{R}^{h*d_h\times d}$ is the output projection matrix that aggregates the information from different heads. d_h is the dimensionality of the output features of each head. To prevent the model from becoming too large, we set $d_h = d/h$.

In addition to the MHA that performs linear transformations, another basic component feed-forward networks (FFN) is complemented to increase the nonlinearity of the Transformer model. FFN takes the output features from MHA as its input and further transform them using two fully-connected layers with the ReLU and dropout layers in between as follows:

$$\text{FFN}(x) = \text{FC}\left(\text{Dropout}\left(\text{ReLU}\left(\text{FC}(x, 4d)\right), 0.1\right), d\right) \tag{4}$$

where the input feature $x \in \mathbb{R}^d$ is first transformed to $4d$-dimensional and then transformed to d-dimensional again. The dropout ratio is set to 0.1.

To summarize, the MHA module learns the attended features that consider the pairwise interactions between two input features, and the FFN module further nonlinearly transforms the attended features. By a modular composition of the two modules, we attain the attention blocks that are can be stacked in depth to construct the MT model for image captioning.

B. Multimodal Transformer for Image Captioning

Based on the preliminary information about the Transformer above, we describe the Multimodal Transformer (MT) architecture for image captioning, which is a deep end-to-end architecture that stacks attention blocks to form an encoder-decoder strategy. It consists of an image encoder and a textual decoder. The image encoder takes an image as its input and uses a pre-trained Faster-RCNN model [34] to extract region-based visual features. The visual features are then fed into the encoder to obtain the attended visual representation with self-attention learning. The decoder takes the attended visual features and the previous word to predict the next word

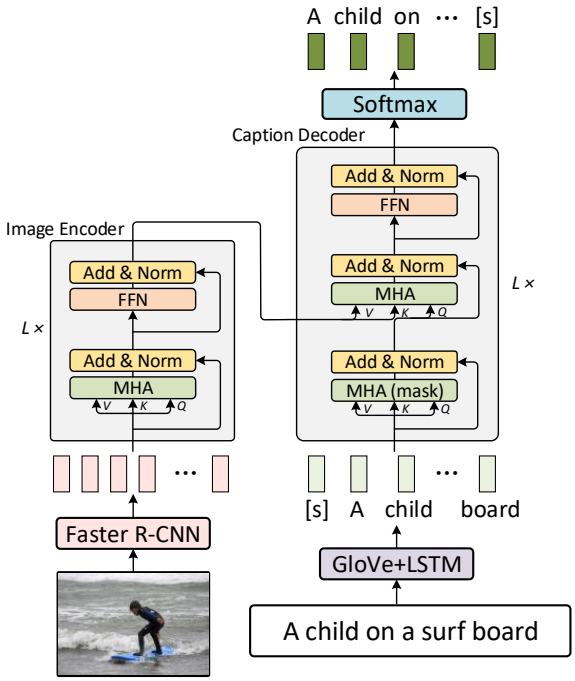


Fig. 1: Multimodal Transformer (MT) model for image captioning. It consists of an image encoder to learn self-attended visual features, and a caption decoder to generate the caption from the attended visual features. [s] is a delimiter that indicates the start or the end of the caption.

recursively. The flowchart of the MT architecture is shown in Fig. 1.

Image Encoder. The input image is represented as a group of visual features that are extracted from a pre-trained object detector [6]. Specifically, the detector is a Faster-RCNN model [34] that is pre-trained on the Visual Genome dataset [35]. We sort the detected objects w.r.t. their confidence scores in descending order and keep the top-m objects. Each object is represented as a feature vector $x_i \in \mathbb{R}^{d_x}$ by mean-pooling the convolutional feature from its detected region. Finally, the image is represented as a feature matrix $X \in \mathbb{R}^{m \times d_x}$.

The visual features X is first fed into a fully-connected layer to adapt the feature dimensionality to the encoder. The projected features (denote as $X^{(0)}$) are then fed into the encoder with L attention blocks $[A^1_{\text{enc}}, A^2_{\text{enc}}, ..., A^L_{\text{enc}}]$. The i-th attention block A^l_{enc} takes the output features X^{l-1} from the $i-1_{th}$ attention block, and output their attended features X^l in a recursive manner.

$$X^l = A^l_{\text{enc}}(X^{l-1}) \tag{5}$$

Each $A_{\text{enc}}(X)$ consists of a MHA module and a FFN module. The MHA module characterizes the self-attentions within X that the queries Q, keys K and values V in Eq. (2) all refer to the same input features X:

$$X' = \text{LayerNorm}(X + \text{MHA}(X, X, X)) \tag{6}$$

$$A_{\text{enc}}(X) = \text{LayerNorm}(\text{FFN}(X') + X') \tag{7}$$

where residual connections [36] and layer normalizations [37] are applied after the MHA and FFN modules.

Caption Decoder. Based on the visual representations from the encoder, the textual decoder generates captions for the image. The input caption is first tokenized into words and trimmed to a maximum length of n words. Each word in the caption is first represented as a word vector $y_i \in \mathbb{R}^{300}$ by using the 300-D GloVe word embedding [38] pre-trained on a large-scale corpus. We use a feature matrix $Y \in \mathbb{R}^{n \times 300}$ to represent a caption sentence. For the captions that are shorter than 16 words, we use zero-padding to fill them to the maximum size. To model the temporal information of the captions, the word embeddings are then pass through a one-layer LSTM network [24] with d_y hidden units, resulting in caption representations $Y = [y_1, ..., y_n] \in \mathbb{R}^{n \times d_y}$.

In the training stage, the caption decoder takes the inputs from both the image encoder and caption representations. Given the attended image features X^L and the caption input features Y, the caption decoder with L attention blocks ($[A^1_{\text{dec}}, A^2_{\text{dec}}, ..., A^L_{\text{dec}}]$) learns to predict the attended word features in an analogous manner to the strategy in the encoder.

$$Y^l = A^l_{\text{dec}}(X^L, Y^{l-1}) \tag{8}$$

Each $A_{\text{dec}}(X, Y)$ consists of two MHA modules and one FFN module to :

$$Y' = \text{LayerNorm}(Y + \text{MHA}(Y, Y, Y)) \tag{9}$$

$$Y'' = \text{LayerNorm}(Y' + \text{MHA}(Y', X, X)) \tag{10}$$

$$A_{\text{dec}}(X, Y) = \text{LayerNorm}(\text{FFN}(Y'') + Y'') \tag{11}$$

where the first MHA module models the self-attentions within Y and the second MHA module learns the guided-attention on Y guided by X. It is worth noting that in the self-attention learning for Y (*i.e.*, the first MHA module), each word is only allowed to attend to the words at earlier positions in the output sequence. This is implemented by masking subsequent positions (setting them to -∞) before the softmax step in the self-attention calculation, thereby resulting in a triangular mask matrix $M \in \mathbb{R}^{n \times n}$.

The output features $Y^L = [y^L_1, y^L_1, ..., y^L_n]$ are fed into a linear word embedding layer to transform the features to a d_v-dimensional space, where d_v is the vocabulary size. Subquently, softmax cross-entropy loss is performed on each word to predict the probability of its next word.

In the testing stage, the caption is generated word-by-word in a sequential manner. When generating the t_{th} word, the input features are represented as $Y_{\leq t} = [y_1, y_2, ...y_{t-1}, \mathbf{0}, ..., \mathbf{0}] \in \mathbb{R}^{n \times d_y}$, where $\mathbf{0} \in \mathbb{R}^{d_y}$ corresponds a zero-padded feature. The input caption features along with the image features are fed forward the model to obtain the word with the largest probability among the whole word vocabulary. The predicted word is then integrated into the inputs to recursively generate the new inputs $Y_{\leq t+1}$. To improve the diversity of generated captions, we also introduce the beam search strategy during the testing stage.

IV. Image Encoder with Multi-view Visual Representation

In this section, we introduce multi-view image representations and modify the the image encoder in section III-B

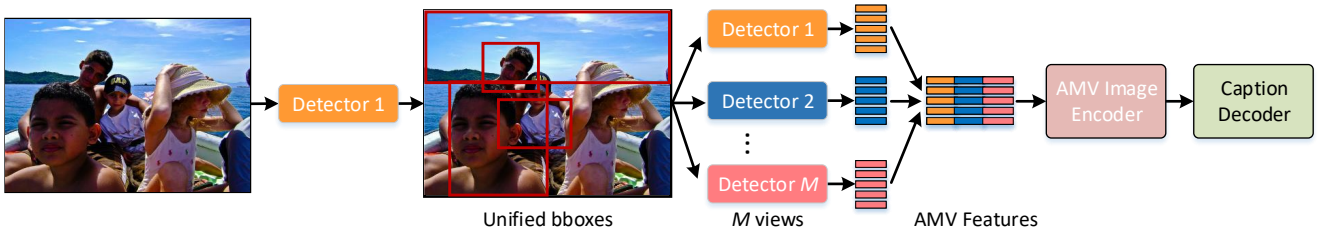


Fig. 2: The flowchart of the aligned multi-view (AMV) image encoder model. Given an image, different object detectors are regarded as the multiple views. To obtain the aligned multi-view features, we choose one of the M detectors to predict the unified bounding boxes for objects, and then use these bboxes to extract aligned multi-view features. The aligned multi-view features are fed into the AMV image encoder (which is exactly the same as the one for single-view features introduced in section III-B).

to multi-view image encoder to facilitate the representation capacity of the MT model. Though it has been intensively investigated by previous works [39][40],[41], existing multi-view learning approaches focus on integrating the *global* multi-view features (*e.g.*, color histogram or GIST descriptor) from the whole image. However, such global multi-view features may fail to preserve the fine-grained semantics of the image, thus leading to incorrect caption. In contrast, we extract region-based *local* multi-view features from different object detector to represent the image. Each object detector here is regarded as one single view and we adopt the Faster R-CNN models with different backbones (*e.g.*, ResNet-101, ResNet-152 or ResNeXt-101) to form the multi-view features.

Note that the objects extracted from different detectors are naturally unaligned, thereby making it challenging to learn the correspondence across different views. To address this problem, we extend the proposed image encoder model in section III-B, and introduce two multi-view image encoder models, namely, the *Aligned Multi-View* (AMV) image encoder and the *Unaligned Multi-View* (UMV) image encoder, respectively.

A. Aligned Multi-View Image Encoder

The AMV model uses a simple strategy to obtain the aligned multi-view features from different object detectors. Rather than extracting the object bounding boxes and corresponding features for each view, we propose a two-stage feature extraction framework. Given M pre-trained Faster R-CNN models, we first select one detector as the primary model to generate the unified bounding boxes for all views. The choices of different primary models has little influence on the quality of the generated features, and we simply choose the model with the highest detection performance. Subsequently, the unified bounding boxes are used to extract features from different Faster R-CNN models. Specifically, the Faster R-CNN models degenerate to their Fast R-CNN versions [42] that take the pre-computed bounding boxes as inputs. The resulting multi-view features are aligned such that each paired multi-view features correspond to one object in the image.

Assuming that we generate m unified bounding boxes, the extracted features from the i-th view ($i \in \{1, 2, ..., M\}$) can be represented as $X_{(i)} \in \mathbb{R}^{m \times d_i}$, where d_i is the dimensionality of the features. By simply concatenating the features in columns, we obtain the multi-view features $X = [X_{(1)}, X_{(2)}, ..., X_{(M)}] \in \mathbb{R}^{m \times (d_1 + d_2 + ..., d_M)}$. These aligned multi-view features can replace the aforementioned single-view feature, and be seamlessly fed into the image encoder. The overall flowchart of the AMV model is shown in Fig. 2.

To align the multi-view features, the AMV model uses the unified bounding boxes. However, we argue that this strategy may harm the diversity of multi-view features, leading to a limited representation capacity of the encoded image features. Moreover, the AMV model implicitly constrains the object detector for each view to be a Faster R-CNN model, which can either take the pre-computed proposals as inputs or generate the proposals using the built-in Region Proposal Networks (RPN) [34]. This constraint limits the usage of one-stage object detectors, *e.g.*, RetinaNet [43] and YOLO [44].

B. Unaligned Multi-View Image Encoder

To address the limitations of the AMV encoder model, we propose a more generalized unaligned multi-view (UMV) image encoder model that can directly integrate the unaligned multi-view features from different object detectors (see the flowchart in Fig. 3).

The extracted visual features for the i-th view can be represented as $X_{(i)} \in \mathbb{R}^{m_i \times d_i}$, where the number of features m_i and the feature dimensionality d_i can be different across multiple views. The unaligned multi-view features are fed into an encoder to be aligned and fused simultaneously. Specifically, we choose one view as the primary view and use its features to learn the guided-attention for other views. The attended features from other views are then integrated into the features in the primary view to output the output features.

Given the multi-view features $X_{(1)}, X_{(2)}, ..., X_{(M)}$, they are first linearly projected into a common d-dimensional space to obtain their transformed representations $F_{(1)}, F_{(2)}, ..., F_{(M)}$. Assuming that $F_{(1)}$ corresponds to the features of the primary view, we have $M - 1$ MHA modules in total to model the interactions between $F_{(1)}$ and $F_{(i)}$ with $i \in \{2, 3, ..., M\}$.

$$\tilde{F}_{(i)} = \text{MHA}_{(i)}(F_{(1)}, F_{(i)}, F_{(i)}) \tag{12}$$

where $\tilde{F}_{(i)} \in \mathbb{R}^{m_1 \times d}$ is the attended output features for the i-th view. The obtained features $\tilde{F}_{(2)}, \tilde{F}_{(3)}, ..., \tilde{F}_{(M)}$ have the same shape as $F_{(1)}$, and so they can be integrated with $F_{(1)}$ via an element-wise summation. The MHA modules here can

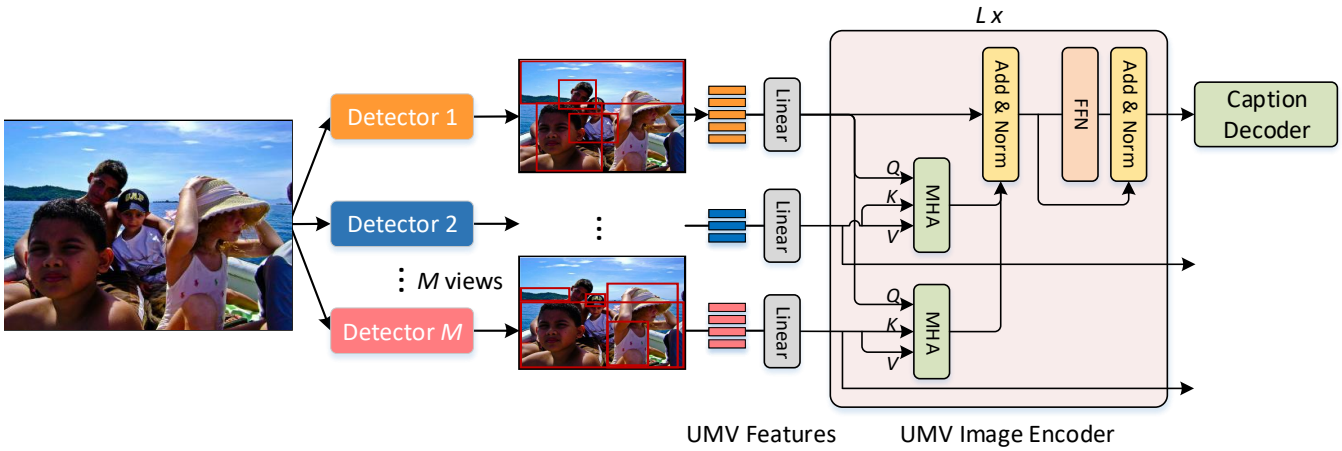


Fig. 3: The flowchart of the unaligned multi-view (UMV) image encoder model. Given an image, unaligned multi-view features are extracted from different object detectors in parallel. The unaligned multi-view features are fed into the UMV model to output the attended features with adaptive alignment learning.

be understood as learning the image-guided attention over the image features from other views.

$$\tilde{F}_{(1)} = F_{(1)} + \tilde{F}_{(2)} + \tilde{F}_{(3)}, ..., + \tilde{F}_{(M)} \tag{13}$$

Following the image encoder model in section III-B, the integrated features $\tilde{F}_{(1)}$ that are followed by layer normalization [37] are then fed forward through the FFN module to obtain the transformed representations. It is worth noting that the UMV model can also be stacked in depth to learn more accurate interactions across different views, thus resulting in more discriminative output visual features for generating captions.

V. Experiments

In this section, we conduct experiments and evaluate the proposed MT models on MSCOCO 2015 image captioning dataset [45]. Additionally, we use the Visual Genome dataset [35] to pre-train the object detectors that are further used to extract the bottom-up-attention visual features [6].

A. Datasets

MSCOCO is a benchmark dataset for various computer vision tasks, including object detection, instance segmentation, and image captioning [45]. It contains 83k training images, 40k validation images, and 81k test images. Each image is associated with five captions. Similar to [6], we use the *Karpathy* splits [26] that have been extensively used for reporting results in prior works. These splits merge the images from the original train and val splits, resulting in 121k images in total. After that, the 123k images are split into 113k/5k/5k images for training/validation/testing, respectively. The trained models are ensembled to obtain the predictions that are submitted to the official MSCOCO test server. To evaluate the caption quality, we use four automatic evaluation metrics, namely, BLEU [12], ROUGE-L [46], METEOR [47] and CIDEr [13].

Flickr30k is a dataset that is widely used in caption-image retrieval and image caption generation tasks [48]. It contains 31k images with five captions annotated for each image. We use the publicly available split where the validation set and testing set each have 1k images, and the remaining 29k images are used for training. Similar to previous works, we adopt BLEU [12] and METEOR [47] as the evaluation metrics.

Visual Genome is a large-scale dataset to evaluate the interactions between objects in the images. It contains 108k images with densely annotated objects, attributes, and relationships. Following the strategies in [6], we use the object and attribute annotations to pre-train the bottom-up-attention models. All the images are split into training (98k images), validation (5k images) and testing (5k images). Since part of images in Visual Genome are also found in the MSCOCO captioning dataset, we perform careful checking to avoid affecting the MSCOCO validation and testing splits. Similar to [6], we perform extensive cleaning and filtering of the training data to select 1,600 object classes and 400 attributes. This cleaned dataset is used for training our object detection models.

B. Implementation Details

For the captions, we perform the pre-processing as follows. All the caption sentences are converted to lower case and tokenized into words with white space. The rare words that occur less than 5 times or do not exist in the pre-trained GloVe vocabulary [38] are discarded, resulting in a vocabulary of 9,343 words. Each word in the caption is represented as word embedding vector by looking-up the GloVe word vocabulary. The out-of-vocabulary words are represented as all-zero vectors.

For the images, we use the pre-trained bottom-up-attention models to detect the objects and extract visual features for the detected objects. For multi-view image representation, we trained up to three Faster R-CNN [34] models (*i.e.*, number of views M=3) with different backbones, namely ResNet-101 [36], ResNet-152 [36] and ResNeXt-101 [49], respectively. For each model, we select the top-100 objects with the highest confidence scores to represent the image, where each object is

(a) **Caption Representations:** Scores of the MT_{sv} models (ResNet-101 backbone) with different caption representations. The reference model uses randomly initialized word embeddings and then fine-tuned. PE denotes the positional encoding to model the temporal information of the caption [14]. GloVe_{pt} and $\text{GloVe}_{\text{pt+ft}}$ mean the word embeddings are pre-trained with GloVe, while $\text{GloVe}_{\text{pt+ft}}$ is additionally fine-tuned along with the model.

Model	Cross-Entropy Loss			Self-Critical Loss		
	B@1	M	C	B@1	M	C
Rand_{ft} + PE	76.0	28.2	115.9	80.4	28.9	129.2
GloVe_{pt} + PE	76.2	28.0	116.6	80.5	29.0	129.3
GloVe_{pt} + LSTM	**76.2**	**28.3**	**117.1**	80.8	**29.1**	130.8
$\text{GloVe}_{\text{pt+ft}}$ + LSTM	**76.2**	**28.3**	**117.1**	**81.2**	**29.1**	**130.9**

(b) **Number of Attention Blocks:** Scores of the MT_{sv} models with different number of attention blocks $L \in \{1, 2, 4, 6, 8\}$. For each model, we also report its corresponding number of model parameters.

L	#Params ($\times 10^6$)	Cross-Entropy Loss			Self-Critical Loss		
		B@1	M	C	B@1	M	C
1	17.1	76.3	27.9	113.7	79.4	28.3	124.3
2	25.1	76.4	28.3	116.6	80.1	28.6	127.2
4	39.8	**76.5**	**28.4**	**117.1**	80.4	29.0	129.6
6	54.5	76.2	28.3	117.0	**80.8**	**29.1**	**130.9**
8	69.2	76.4	28.1	116.5	80.7	29.0	130.4

(c) **Single-view *vs.* Multi-view:** Scores of the MT models with single-view feature (MT_{sv}), aligned multi-view features (MT_{amv}) or unaligned multi-view features (MT_{umv}).

Model	Views	Cross-Entropy Loss			Self-Critical Loss		
		B@1	M	C	B@1	M	C
MT_{sv}	R-101	76.2	28.3	117.1	80.8	29.1	130.9
	R-152	76.4	28.4	117.5	81.0	29.3	131.2
MT_{amv}	R-101 and R-152	77.0	**28.6**	119.4	81.2	29.4	132.7
MT_{umv}	R-101 and R-152	**77.1**	**28.6**	**119.5**	**81.6**	**29.5**	**133.4**

(d) **Number of Views:** Scores of the MT_{umv} models with different number of views $M \in \{2, 3\}$.

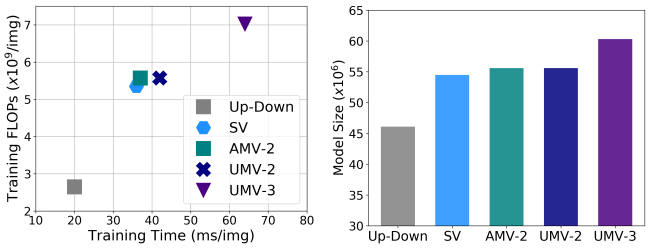
M	Views	Cross-Entropy Loss			Self-Critical Loss		
		B@1	M	C	B@1	M	C
2	R-101 and R-152	77.1	28.6	119.5	81.6	29.5	133.4
	R-101 and X-101	76.7	28.4	118.4	81.4	29.4	133.0
	R-152 and X-101	76.7	28.5	118.8	81.5	29.4	133.2
3	R-101, R-152 and X-101	**77.3**	**28.7**	**119.6**	**81.9**	**29.5**	**134.1**

TABLE I: Ablations of the proposed MT models evaluated on the MSCOCO Karpathy test split. B@1, M, and C correspond to the BLEU@1, METEOR and CIDEr scores, respectively. For each model, we report the results optimized with either the cross-entropy loss or the self-critical loss [10]. R-101, R-152, X-101 denote the object detector with ResNet-101, ResNet-152 and ResNeXt-101 backbones, respectively. All models use pre-trained Faster R-CNN models to obtain input visual features and all results are obtained with beam search in the testing stage. The best result for each evaluation metric is bolded.

represented as a vector by mean-pooling the last convolutional feature from its detected region.

The hyper-parameters of the MT models that are used in the experiments are listed as follows. The dimensionality of input image features d_x, and the input caption features d_y are 2048 and 512, respectively. According to the recommendation in [14], the latent dimensionality d in the MHA module is 512, the number of heads h is 8, and the latent dimensionality for each head $d_h = d/h = 64$. The number of attention blocks L in the encoder and decoder ranges in $\{1, 2, 4, 6, 8\}$.

To train the MT models, we use the Adam solver [50] with a batch size of 10. The base learning rate is set to $\min(1te^{-4}, 3e^{-4})$, where t is the current epoch number that starts at 1. After 6 epochs, the learning rate is decayed by 1/2 after every 3 epochs. All models are first trained for 15 epochs using the cross-entropy loss and then are further trained for additional 10 epochs using the self-critical loss to alleviate the exposure bias during cross-entropy optimization [10].



(a) FLOPs *vs.* Training Time (b) Model Sizes

Fig. 4: Ablations of the computational costs of the MT models with different visual features. SV, AMV-2, UMV-2, UMV-3 denotes the the single-view features (R-101), aligned multi-view features (R-101 and R-152), unaligned multi-view features (R-101 and R-152), unaligned multi-view features (R-101, R-152 and X-101), respectively. For each model, we report the average training FLOPs vs. training time per one image in (a) and its corresponding model size in (b). Up-Down [6] is a strong reference image captioning model.

C. Ablation Studies

We run a number of ablation experiments on MSCOCO image captioning dataset to explore the effectiveness of the single-view MT models (MT_{sv}) with different hyper-parameters, as well as its multi-view variants with aligned multi-view image encoder MT_{amv} and unaligned multi-view image encoder MT_{umv}. The results shown in Table I are discussed in detail below.

Caption Representations: Table I(a) summarizes the ablation experiments on different caption representations for MT_{sv} with the number of attention blocks L=6. Compared with the reference model that uses randomly initialized word embeddings and positional encoding [14], we can see that using the word embeddings that are pre-trained by GloVe [38] brings significant improvements. In addition, introducing other tricks such as replacing PE with an LSTM network to model the temporal information, or fine-tuning the GloVe word embeddings along with the MT model can slightly improve the performance further. Note that the GloVe_{pt}+LSTM model and the $\text{GloVe}_{\text{pt+ft}}$+LSTM model report the same performance in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2947482, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II: **Single-model** image captioning performance on the MSCOCO Karpathy test split. The methods marked with * denote using pre-trained Faster R-CNN models to obtain input visual features. R, D, I-v3, I-v4 and IR-v2 denotes the ResNet, DenseNet, Inception-v3, Inception-v4 and Inception-ResNet-v2 model, respectively.

Model	Backbone	Cross-Entropy Loss					Self-Critical Loss				
		B@1	B@4	M	R	C	B@1	B@4	M	R	C
SCST [10]	R-101	-	30.0	25.9	53.4	99.4	-	34.2	26.7	55.7	114.0
ADP-ATT [9]	R-101	74.2	33.2	26.6	-	108.5	-	-	-	-	-
LSTM-A [21]	R-101	75.4	35.2	26.9	55.8	108.8	78.6	35.5	27.3	56.8	118.3
Up-Down [6]	R-101*	77.2	36.2	27.0	56.4	113.5	79.8	36.3	27.7	56.9	120.1
RFNet [51]	R, D, I-v3, I-v4 and IR-v2	76.4	35.8	27.4	56.5	112.5	79.1	36.5	27.7	57.3	121.9
GCN-LSTM [22]	R-101*	**77.4**	37.1	28.1	57.2	117.1	80.9	38.3	28.6	58.5	128.7
MT_{sv}	R-101*	76.2	36.6	28.3	56.8	117.1	80.8	39.8	29.1	59.1	130.9
MT_{umv}	R-101, R-152 and X-101*	77.3	**37.4**	**28.7**	**57.4**	**119.6**	**81.9**	**40.7**	**29.5**	**59.7**	**134.1**

TABLE III: **Real-time leaderboard** of the state-of-the-art solutions on the online MSCOCO test server (April 21st, 2019). The first split shows the published solutions while the second split shows the unpublished ones. All the published solutions use the model ensembling strategy.

Solution	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC [52]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
M-RNN [53]	71.6	89.0	54.5	79.8	40.4	68.7	29.9	57.5	24.2	32.5	52.1	66.6	91.7	93.5
LRCN [25]	71.8	89.5	54.8	80.4	40.9	69.5	30.6	58.5	24.7	33.5	52.8	67.8	92.1	93.4
ADP-ATT [9]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
LSTM-A [21]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
SCST [10]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	65.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [6]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [51]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [22]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SRCB-ML-Lab	81.1	95.4	66.0	89.8	51.5	81.3	39.7	71.3	28.4	37.3	58.5	73.1	125.3	126.7
h-p-hl	80.5	95.0	65.3	89.6	50.9	81.1	39.0	70.9	28.7	38.2	58.6	74.1	125.0	127.2
TecentAI.v2	81.1	95.5	65.7	90.0	50.8	80.9	38.6	70.1	28.6	37.7	58.7	73.7	125.4	127.8
lun	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
MIL-HDU (ours)	**81.7**	**95.6**	**66.8**	**90.5**	**52.4**	**82.4**	**40.4**	**72.2**	**29.4**	**38.9**	**59.6**	**75.0**	**130.0**	**130.9**

the cross-entropy loss stage, as the fine-tuning is performed only in the self-critical loss stage. Directly fine-tuning the GloVe embedding from scratch (*i.e.*, from the cross-entropy loss) leads to inferior performance. This result can be explained as the word embeddings being sensitive to the captioning performance, and training from scratch may degenerate their representation capacity.

Number of Attention Blocks: Table I(b) shows the performance of the MT_{sv} models with different number of attention blocks $L \in \{1, 2, 4, 6, 8\}$. We can see that the model size grows linearly as L increases. Regarding the performance, we have two observations as follows: 1) as increasing L, the model's performance gradually improves and is saturated at a certain number. This can be explained as a deeper model capturing more complex relationships among objects, providing a more accurate understanding of the image contents. In addition, a deeper model has a larger representation capacity and has a larger risk to overfit the training set, and 2) the optimal model is achieved at different L that are trained with different losses, *i.e.*, L=4 for the cross-entropy loss and L=6 for the self-critical loss. The reinforcement learning-based self-critical loss provides a more diverse exploration of the hypothesis space to avoid overfitting, and thus it can better utilize the potential of large models.

Single-view *vs.* Multi-view: Next, we compare the MT model with single-view or multi-view features in Table I(c). We use two Faster R-CNN models with different backbones (ResNet-101 or ResNet-152) to extract the multi-view features. For MT_{amv}, the unified object boxes are extracted from the detector with the ResNet-152 backbone. From the results, we can see following that: 1) the representation capacity of the object detectors may slightly influence the image captioning performance. The MT_{sv} model with the ResNet-152 backbone steadily outperforms the counterpart with the ResNet-101 backbone; and 2) introducing multi-view features significantly improves the captioning performance over the single-view models. MT_{umv} slightly outperforms MT_{amv}, thus highlighting the effect of using diverse multi-view features with unaligned objects.

Number of Views: In Table I(d), we show the performance of the MT_{umv} models with different number of views M. We can see that the performance of MT_{umv} with M=3 has little improvement when compared to the obtained best results with M=2 (*i.e.*, backbones of R-101 and R-152). All the

other metrics except the CIDEr score has only 0.1 0.3 point improvement. On the other hand, increasing M will linearly increase the model size, computational cost, and memory usage. To make a trade-off between efficiency and efficacy, we terminate at M=3 and do not introduce more views to the image encoder.

Computational Costs and Model Sizes: In Fig. 4, we show the computational costs in terms of the average training FLOPs and training time per one image , as well as the model sizes of the MT models with different visual features. We also introduce a strong reference model Up-Down [6] for comparison. From the results, we can see that: 1) the model size, FLOPs, and training time are positively correlated with each other, therefore we use the FLOPs metric to measure the computation cost in the following ; 2) when the number of views M=2, FLOPs of MT_{amv}, and MT_{umv} are nearly identical and are only about 10% higher than that of MT_{sv}; 3) when M=3, FLOPs of MT_{umv} are about 30% higher than all the counterparts, which is sublinear with respect to M. This can be explained that there are M-1 MHA modules but only one FFN module in each of the attention block in UMV image encoder. Since FLOPs of one FFN module is much higher than one MHA module, the cost of the FFN module counteracts the costs of multiple MHA modules in MT_{umv}; and 4) with up to a 3× increase of FLOPs, our best MT_{umv} model obtains 12-point improvement over the reference Up-Down model in terms of the CIDEr score (see Table II).

D. Comparative Results on MSCOCO

By taking the ablation results into account, we compare our best single-view and multi-view MT models to the state-of-the-art approaches on MSCOCO image captioning dataset.

Results on the Karpathy test split: In Table II, we report the comparative results of our approaches along with the SCST [10], ADP-ATT [9], LSTM-A [21], Up-Down [6] and GCN-LSTM [22] on the Karpathy test split. Note that all the compared methods use the same ResNet-101 backbone. With single-view features, the MT_{sv} model outperforms most state-of-the-art methods, especially when it is optimized using the self-critical loss. When equipped with multi-view features, the MT_{umv} model (trained with the self-critical loss) achieves the new state-of-the-art single-model performance for this split in terms of all evaluation metrics. Note that the RFNet [51] also incorporates multi-view features, and they introduce more views than our approach (4 vs. 2). However, its performance is inferior to MT_{umv}, which suggests that the strategy to fuse multi-view features, rather than the number of views, is the key to the captioning performance.

Results on the official test server: We also submitted the results obtained from an ensemble of seven MT models to the official MSCOCO test server[1] and compare with the state-of-the-art in Table III. The ensemble consists of two MT_{sv} models, two MT_{amv} models, and three MT_{umv} models with different random seeds. The ensemble strategy is performed during the prediction of each caption word. Given one testing image, the input visual features are fed forward through the seven models in parallel to predict the word probabilities over the vocabulary. The predicted word distributions from different models are then integrated to vote the word with the largest probability. By doing so in a recursive manner, we finally obtain the caption sentence for the image. Table III demonstrates the results of the comparison to the state-of-the-art solutions on the leaderboard including the published ones (in the first split) and the unpublished ones (in the second split). C5 (or c40) denotes the official test settings with 5 (or 40) ground-truth captions, respectively. Compared to all the top performing solutions on the leaderboard, our solution significantly outperforms all the other solutions in terms of all reported evaluation metrics at the time of submission (April 21st, 2019).

[1]https://competitions.codalab.org/competitions/3221#results

TABLE IV: Comparison to the state-of-the-art approaches on Flickr30k datasets.

Method	B@1	B@2	B@3	B@4	M
Google NIC [52]	66.3	42.3	27.7	18.3	-
Soft-Att [54]	66.7	43.4	28.8	19.1	-
Hard-Att [54]	66.9	43.9	29.6	19.9	-
SAS-RE [55]	66.3	44.3	30.5	21.1	18.6
NeuralTalk2-T-oe [56]	64.6	43.8	31.9	22.4	19.2
Att-RegionCNN+LSTM [57]	73.0	55.0	40.0	28.0	-
MT_{sv} (R-101)	74.4	57.5	43.4	32.5	23.6
MT_{sv} (R-152)	74.6	57.7	43.6	32.8	24.1
MT_{amv} (R-101, R-152)	75.5	58.5	44.0	33.2	24.2
MT_{umv} (R-101, R-152)	75.6	58.6	44.3	**33.3**	24.3
MT_{umv} (R-101, R-152, X-101)	**75.8**	**58.7**	**44.5**	**33.3**	**24.5**

E. Comparative Results on Flickr30k

In Table IV, we compare our MT models to the state-of-the-arts on Flickr30k dataset. From the results, we have similar observations to those on the MSCOCO dataset: 1) the MT_{sv} model with single-view visual features has significantly outperformed existing state-of-the-art approaches on this dataset; 2) by replacing R-101 with R-152 backbone, the model performance is slightly improved; 3) modeling multi-view features in the MT model brings prominent improvement over the models with single-view features; and 4) MT_{umv} steadily outperforms MT_{amv} due to its capacity in modeling latent correlations among aligned objects from different views.

F. Qualitative Analysis

To better understand the effectiveness of the proposed approach, we adopt the trained model on MSCOCO and visualize the learned attentions of MT_{sv} in Fig. 5 and MT_{umv} in Fig. 6, respectively. Due to space limitations, we only show one example for each model and visualize the attention maps from typical attention blocks. From the demonstrated results, we have the following observations.

Attentions of the MT_{sv} Encoder: The self-attentions (SA) of the 1st and 6th blocks in the image encoder that are in Fig. 5 reflect the pairwise similarity of the visual objects. From the results, we can see that the following: 1) in Enc(SA)-1, the largest attention values almost appear on the diagonal line, indicating that the pairwise interactions are not learned in the first block; and 2) the largest values in Enc(SA)-6 form vertical

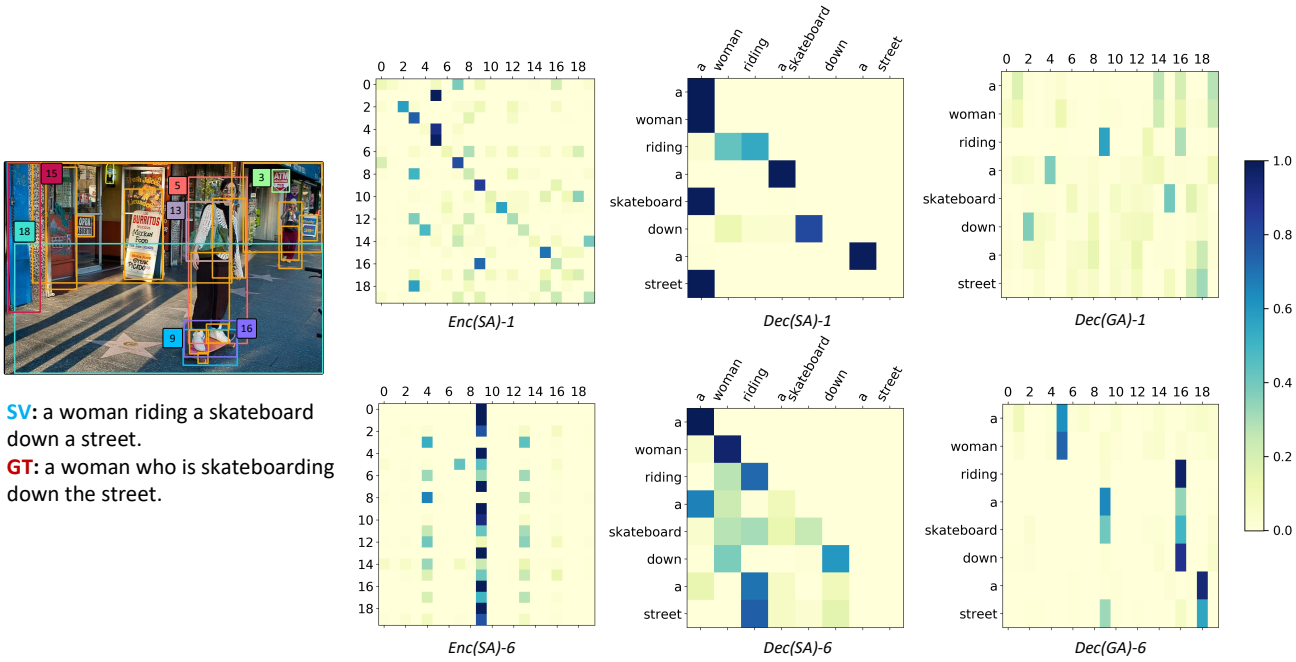


Fig. 5: Visualizations of the 1st and 6th attention maps ($\mathrm{softmax}(QK/\sqrt{d})$) of the $\mathrm{MT_{sv}}$ model with R-101 backbone. Enc(SA) denotes the self-attention in the image encoder; Dec(SA) and Dec(GA) denote the self-attention and guided-attention in the caption decoder, respectively. GT denotes the one of the five ground-truth captions provided by MSCOCO. The index within [0-19] shown on the axes of the attention maps corresponds to each object in the image (20 objects in total) . For better visualization effect, we highlight some objects in the image that receive large attention values.

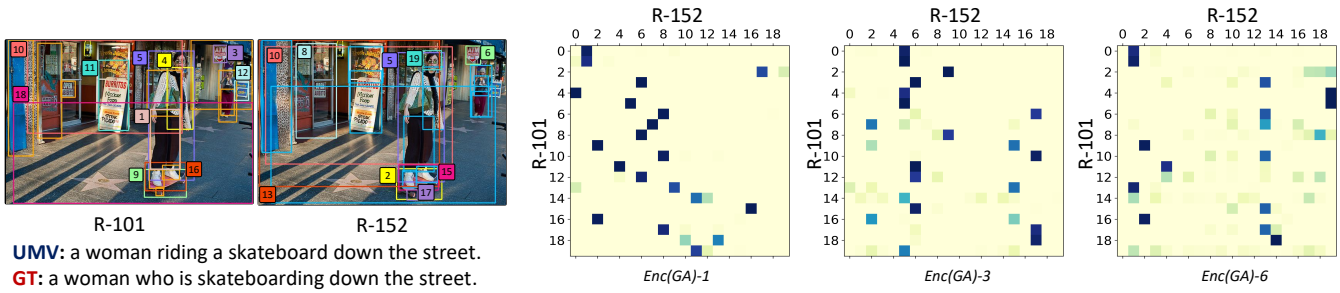


Fig. 6: Visualizations of the 1st, 3rd and 6th attention maps of the $\mathrm{MT_{umv}}$ model with R-101 and R-152 backbones. Enc(GA) denotes the guided-attention in the UMV image encoder (*i.e.*, the cross-view attention within the MHA module in Fig. 3).

lines (*e.g.*, the 4th, 9th and 13th columns), which correspond to the key objects of the image (*e.g.*, the girl and the skateboard). This result reveals that all the attended features tend to use the features of these key objects for the representation.

Attentions of the $\mathrm{MT_{sv}}$ Decoder: The self-attention the 1st and 6th blocks of the caption decoder that are shown in Fig. 5 reflects the similarity of paired words. The largest attention values in Dec(SA)-1 almost appear on the diagonal line, which is similar to those in the Enc(SA)-1. In Dec(SA)-6, the word importance and pairwise word similarities are simultaneously represented. For example, the columns of '*woman*' and '*riding*' obtain focused attention weights, and the relationship between '*woman*' and '*skateboard*' is highlighted.

The guided-attention (GA) reflects the multimodal relationships between word-object pairs. In Dec(GA)-1, the learned attentions are not concentrated, and some word-object similarities are incorrect (*e.g.*, the 15th object is not related to the word '*skateboard*'). In contrast, the attention in Dec(GA)-6 has clearer meanings. The co-attention of key objects along with their word-object relationships are highlighted accordingly.

Attentions of the $\mathrm{MT_{umv}}$ Encoder: In Fig 6, we visualize the 1st, 3rd and 6th guided-attention (GA) blocks in the multi-view image encoder. In Enc(GA)-1, the unaligned objects from different views are adaptively aligned (*e.g.*, the 5th object in R-101 and the 5-th object in R-152, and the 3rd object in R-101 and the 6th object in R-152). In Enc(GA)-3, the contextual relationships are also involved (*e.g.*, the 5th object in R-152 has large attention values to the 1st and the 4th objects in R-101, which correspond to different parts of the woman's body). In Enc(GA)-6, the modeled contextual relationships cover specific objects and contain background scenes (*e.g.*, the 13th object in R-152 and the 10-th object in R-101). These ob-

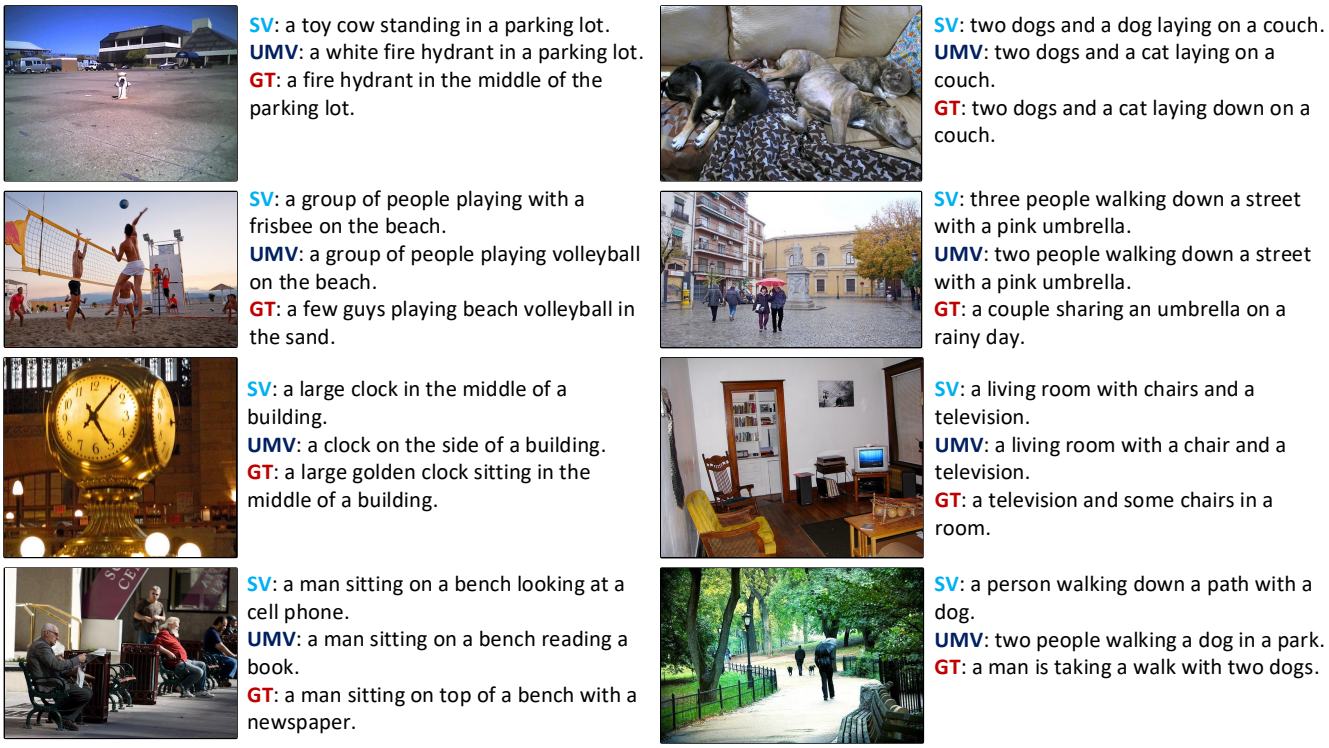


Fig. 7: Examples generated by the MT_{sv} and MT_{umv} models on MSCOCO validation set. GT denotes one of the five ground-truth captions. The first two rows show four examples that MT_{umv} outperforms MT_{sv}, and the third row shows two examples that MT_{sv} outperforms MT_{umv}. The last row shows two examples that both models generate incorrect captions.

servations reveal that the UMV image encoder learns to align the objects and explores more complex interactions across multi-view features to provide a fine-grained understanding of the image content.

Predicted Caption Examples: We show some predicted captioning examples in Fig 7. The first two rows show four examples where MT_{umv} outperforms MT_{sv}, and the third row shows two examples where MT_{sv} outperforms MT_{umv}. The last row shows two examples where both models generate incorrect captions. From the demonstrated results, we can see the following that: 1) although MT_{umv} quantitatively outperforms MT_{sv}, the performance gap is not qualitatively different and they have their advantages in different cases. This results in a diverse ensemble when they are integrated together; 2) the incorrect captions are caused by small objects (*e.g.*, the newspaper or the second person). Moreover, by analyzing the strengths and weaknesses of MT_{umv}, we can see that: compared to the single-view MT model, the multi-view MT model has an advantage in accurately describing the objects whereas has weakness in understanding the background. From the ensemble learning point of view, if every base learner is strong (*i.e.*, the visual features for each view can well describe the semantics of objects), their ensemble is able to achieve better performance than any of the base learner; on the contrary, if all base learners are weak (*i.e.*, the visual features for all views cannot understand the semantics of the background), their ensemble may result in even worse performance compared to any of the base learner.

TABLE V: Accuracies (%) on the test-dev split of VQA-v2 [58] to compare with the state-of-the-art VQA methods. All models are trained on the train+val+vg splits, where vg indicates the augmented training samples from Visual Genome. The first split shows the state-of-the-art results with single-view visual features and the second split shows the MCAN model with different AMV features.

Method	Backbone	All	Y/N	Num	Other
Bottom-Up [6]	R-101	65.32	81.82	44.21	56.05
MFH+CoAtt [31]		68.76	84.27	49.56	59.89
BAN [3]		69.52	85.31	50.93	60.26
BAN+Counter [3]		70.04	85.42	54.04	60.52
MCAN [29]		70.63	86.82	53.26	60.72
MCAN_{amv}	R-101, R-152	71.92	**87.92**	55.43	62.00
MCAN_{amv}	R-101, R-152, X-101	**72.01**	87.88	**55.73**	**62.12**

TABLE VI: Accuracies (%) on RefCOCO [59] to compare with the state-of-the-art visual grounding methods.

Method	Backbone	TestA	TestB	Val
MAttNet [60]	R-101	80.4	69.3	76.4
DDPN [5]		80.1	72.4	76.8
DDPN_{amv}	R-101, R-152	82.0	75.7	80.3
DDPN_{amv}	R-101, R-152, X-101	**82.2**	**76.1**	**80.6**

G. AMV Representation Beyond Image Captioning

Although the MT model is specifically designed for the image captioning task, the proposed aligned multi-view image representation (AMV) framework is generalized that can be simply applied to other tasks that take region-based visual

features such as VQA and visual grounding. Here we conduct experiments on one VQA dataset VQA-v2 [58] and one visual grounding dataset RefCOCO [59]. For each benchmark dataset, we choose the state-of-the-art method on this dataset and replace its single-view visual features by the AMV ones. Specifically, we adopt the MCAN model [29] as reference VQA model and the DDPN model [5] as the reference visual grounding model.

The comparative results on VQA-v2 and RefCOCO are shown in Table V and Table 3, respectively. For the results, we can see that: 1) introducing the AMV visual features can significantly improve the performance of the state-of-the-art approaches for VQA and visual grounding; and 2) the obtained improvement is getting smaller when increasing the number of views is from two to three.

VI. Conclusions

In this paper, we present a novel Multimodal Transformer (MT) framework for image captioning. The MT consists of an image encoder that generates visual representations via deep self-attention learning, and a caption decoder to transform the encoder's visual features to textual captions. To further facilitate the capacity of visual features, we introduce multi-view learning into the image encoder and propose two MT variants, $\mathrm{MT}_{\mathrm{amv}}$ and $\mathrm{MT}_{\mathrm{umv}}$, to model the aligned multi-view features and unaligned multi-view features, respectively. We quantitatively and qualitatively evaluate the proposed MT models on the benchmark MSCOCO image captioning dataset and conduct extensive ablation studies to explore the reasons behind the MT' s effectiveness. Experimental results show that our method significantly outperforms existing approaches, and an ensemble of seven models achieves the best performance on the real-time leaderboard of the MSCOCO image captioning challenge. Finally, we extend the proposed multi-view image representation strategy to other tasks like VQA and visual grounding and obtain significant improvement over the existing state-of-the-art methods with single-view features.

References

[1] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *ACM Special Interest Group on Information Retrieval (SIGIR)*, 2014, pp. 395–404.

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[3] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[4] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1839–1848, 2017.

[5] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, "Rethinking diversified and discriminative proposal generation for visual grounding," *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1114–1120, 2018.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 451–466.

[8] N. Xu, A.-A. Liu, Y. Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli, "Dual-stream recurrent neural network for video captioning," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2018.

[9] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 375–383.

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7008–7024.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 3104–3112.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2002, pp. 311–318.

[13] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.

[15] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 12, pp. 2891–2903, 2013.

[16] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III, "Midge: Generating image descriptions from computer vision detections," in *Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2012, pp. 747–756.

[17] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 444–454.

[18] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 1889–1897.

[19] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 15–29.

[20] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," *arXiv preprint arXiv:1505.01809*, 2015.

[21] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4894–4902.

[22] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 684–699.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[26] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[27] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5659–5667.

[28] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step training for image captioning," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.

[30] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in neural information processing systems (NIPS)*, 2016, pp. 289–297.

[31] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[32] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," *arXiv preprint arXiv:1611.00471*, 2016.

[33] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6087–6096, 2018.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision (IJCV)*, vol. 123, no. 1, pp. 32–73, 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[37] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[38] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[39] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multi-view stochastic learning in image classification," *IEEE Transactions On Cybernetics (CYB)*, vol. 44, no. 12, pp. 2431–2442, 2014.

[40] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multi-modality multi-stream cnn for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2018.

[41] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 28, no. 10, pp. 2657–2666, 2018.

[42] R. Girshick, "Fast r-cnn," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1440–1448.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE international conference on computer vision (ICCV)*, 2017, pp. 2980–2988.

[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[45] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," *arXiv preprint arXiv:1405.0312*, 2014.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *The 42nd Annual Meeting of the Association for Computational Linguistics (ACL) Workshop*, 2004, p. 10.

[47] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL) Workshop*, 2005, pp. 65–72.

[48] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[49] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1492–1500.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 499–515.

[52] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[53] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *International Conference on Learning Representations (ICLR)*, 2015.

[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[55] L. Zhou, Y. Zhang, Y. Jiang, T. Zhang, and W. Fan, "Re-caption: Saliency-enhanced image captioning through two-phase learning," *IEEE Transactions on Image Processing*, 2019.

[56] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2743–2754, 2018.

[57] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1367–1381, 2017.

[58] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 69–85.

[60] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

Jun Yu (M'13) received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China. He was an Associate Professor with the School of Information Science and Technology, Xiamen University, Xiamen, China. From 2009 to 2011, he was with Nanyang Technological University, Singapore. From 2012 to 2013, he was a Visiting Researcher at Microsoft Research Asia (MSRA). He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. He has authored or coauthored more than 100 scientific articles. Over the past years, his research interests have included multimedia analysis, machine learning, and image processing. He is the associate editor of IEEE Trans. on CSVT and Pattern Recognition, and the reviewer of various international journals including IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, etc. In 2017, he received the IEEE SPS Best Paper Award. Dr. Yu has (co-)chaired several special sessions, invited sessions, and workshops. He served as a program committee member or reviewer of top conferences and prestigious journals. He is a Professional Member of the Association for Computing Machinery and the China Computer Federation.

Jing Li received the B.Eng. degree from the School of Management, Hangzhou Dianzi University, Hangzhou, China, in 2017. He is currently pursuing the M.Eng. degree with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His current research interests include multimodal analysis, computer vision and machine learning.

Zhou Yu received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China, in 2010 and 2015, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His research interests includes multimodal analysis, computer vision, machine learning and deep learning. He has served as reviewers or program committee members of prestigious journals and top conferences including IEEE Trans. on CSVT, IEEE Trans. on Multimedia, IEEE Trans. on Image Processing, IJCAI and AAAI, etc.

Qingming Huang (F'18) is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia computing, image processing, computer vision and pattern recognition. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. He is the associate editor of IEEE Trans. on CSVT and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, PSIVT, etc.