Multimedia Tools and Applications (2024) 83:12209–12233
https://doi.org/10.1007/s11042-023-15869-x

Stacked cross-modal feature consolidation attention networks for image captioning

Mozhgan Pourkeshavarz[1] · Shahabedin Nabavi[1] · Mohsen Ebrahimi Moghaddam[1] · Mehrnoush Shamsfard[1]

Received: 2 January 2023 / Revised: 7 May 2023 / Accepted: 15 May 2023 /
Published online: 23 June 2023

Abstract

The attention-enriched encoder-decoder framework has recently aroused great interest in image captioning due to its overwhelming progress. Many visual attention models directly leverage meaningful regions to generate image descriptions. However, seeking a direct transition from visual space to text is not enough to generate fine-grained captions. This paper exploits a feature-compounding approach to bring together high-level semantic concepts and visual information regarding the contextual environment fully end-to-end. Thus, we propose a stacked cross-modal feature consolidation (SCFC) attention network for image captioning in which we simultaneously consolidate cross-modal features through a novel compounding function in a multi-step reasoning fashion. Besides, we jointly employ spatial information and context-aware attributes (CAA) as the principal components in our proposed compounding function, where our CAA provides a concise context-sensitive semantic representation. To better use consolidated features potential, we propose an SCFC-LSTM as the caption generator, which can leverage discriminative semantic information through the caption generation process. The experimental results indicate that our proposed SCFC can outperform various state-of-the-art image captioning benchmarks in terms of popular metrics on the MSCOCO and Flickr30K datasets.

Keywords Contextual representation · Cross-modal feature fusion · Image captioning · Stacked attention network · Visual and semantic information

1 Introduction

Automatically describing the content of images, known as image captioning, is a significant task of artificial intelligence, which combines the field of computer vision (CV) with natural language processing (NLP). Image captioning has several applications for image indexing, social media platforms, visually impaired people, etc. Although this task seems easy for

✉ Mohsen Ebrahimi Moghaddam
m_moghadam@sbu.ac.ir

[1] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

humans, it is complicated for machines. Machines should solve the problem of identifying which objects and attributes are present in the image, and their interactions must be expressed in natural language. The recent progress in deep neural networks has taken the latest significant step towards a reliable solution in generating descriptions for images.

In particular, deep image captioning architectures have shown impressive results in discovering the mapping between visual features and their correspondences in natural language. The well-known encoder-decoder framework is used to perform the task. It contains a convolutional neural network (CNN) for feature extraction and long short-term memory (LSTM) to generate a sentence based on the static overall image feature vector [23, 24, 26, 31, 33]. Although the advancements in these techniques are encouraging, a bottleneck facing the mentioned framework is that it is troublesome to mine all the visual information essential to construct a caption that accurately describes the image.

Inspired by the presence of attention in the human visual system that tends to focus on particular parts of the whole visual space [5, 7] visual attention has been proposed. Specifically, rather than encoding an image into a single static vector, visual attention encourages the model to selectively focus on salient areas of the image and use these areas to generate captions [50]. Besides, some models focus only on the salient regions without scanning the entire image, which cannot capture the context. Although that approach is interesting, it suffers from two main drawbacks, which motivate further significant research. The first is that the model generates visual words rather than high-level semantic words. The other problem is that they lack textual information, which leads to an inaccurate understanding of image context. Several studies have shown that attribute-based methods aim to generate more advanced semantic details to boost image captioning performance [49, 53, 54]. However, the downside of this effective approach is that incorporating all the existing attributes in the image into the recurrent neural network (RNN) is unnecessary or even misleading. More recent evidence highlights that there is a need for both visual and semantic information due to their complementary nature [16, 43]. Despite the most significant benefit of their work, they do not consider the interrelationship between visual and semantic information and combine them to construct an abstract representation regardless of the image context.

Generally, generating a subjective sentence to describe the salient points in the image requires more abstract words. In many cases, inferring these words needs to consider more than one region in addition to high-level semantic concepts with an awareness of contextual information. For example, in the caption, "soccer fans cheer their team and celebrate the goal in a full stadium with open-air", it is surprising that none of the words can be classified only from a bounding box visual region. Intuitively, when predicting words like "fans" and "goal", the model must make inferences based on visual and semantic information under the umbrella of contextual representation. Therefore, using the visual regions to generate fine-grained captions is not enough. On the other side, involving high-level semantic concepts and contextual information along with the region of interest (ROI) needs an advanced feature fusion approach.

This paper sheds new light on cross-modal feature consolidation for the image captioning task. Many previously proposed models of visual attention directly use meaningful regions of images to generate descriptions. However, looking for a direct transition from visual space to text is not enough to generate fine-grained captions. This paper leverages a feature-compounding approach to gather high-level semantic concepts and visual information regarding the contextual environment in a fully end-to-end manner. Specifically, we form our novel compounding function in the proposed stacked cross-modal feature consolidation (SCFC) attention network. In particular, with progressive reasoning via multiple CFC layers, the SCFC can gradually consolidate cross-modal knowledge to generate a rich representation through the caption generation process in every time step. We also construct our cross-modal features as principal components in the compounding function to boost the captioning result. Precisely, spatial-visual information, high-level semantic concepts, and contextual information are all considered for preparing an abstract and richer representation of the given image. To better use consolidated features potential, we further offer an SCFC-LSTM as the caption generator, which can leverage discriminative semantic information through the caption generation process. Besides, our model is more attractive from the modelling perspective because it can be trained fully end-to-end.

The contributions of this study are as follows:

1. A novel SCFC attention network is proposed. A compounding function is formed, which can perform multistep reasoning on cross-modal features to promote the generation of discriminative semantic features.
2. Our multi-aspect features as principal components in the compounding function contain (1) spatial-visual information, (2) high-level semantic concepts, and (3) contextual information to represent various aspects of the given image. Furthermore, there is no need for an independent stage to extract these features since the proposed model can be trained fully end-to-end with a single optimization level.
3. We provide SCFC-LSTM as the caption generator, which can use discriminatory semantic information through the caption generation process, thereby generating fine-grained captions.
4. We verify the effectiveness of our method on the benchmark MS-COCO and Flickr30K datasets. Experimental results show that the proposed method can achieve competitive results with state-of-the-art methods.

2 Related works

A large number of articles on image captioning have been published to date [57]. Several articles have used classical encoder-decoder frameworks, and some newer studies have used transformers for image captioning [19, 30, 40, 41]. Specifically, we are interested in the visual attention approaches used in the classical encoder-decoder framework, which has attracted considerable interest due to its outstanding performance. Previous works combine a CNN to encode an image into a single static visual feature map and then feed it into an RNN as a decoder [23, 34, 46]. However, this static representation can lead to losing local information.

Inspired by the attention mechanism introduced in machine translation [3, 50] firstly proposed spatial attention in image captioning, which integrates the hidden state of the last step and visual features of patches to calculate the attention weights over different patches in images. A weighted sum of all patches obtains the soft attention feature, and the hard attention feature corresponds to visual information of the most important patch. Although this has made great progress, it involves many meaningless patches, which leads to high computational complexity and increased visual interference. [11] proposed scene-specific contexts and employs selective search [44] to generate region proposals. Besides, [36] proposed an area-based attention mechanism which allows a direct association between caption words and image regions by modelling the dependencies between image regions, caption words, and the hidden states. Furthermore, [32] highlights that the decoder may predict non-visual words with little visual information from the image and introduce a visual sentinel to decide whether to participate in the visual part or language model. Therefore, by introducing adaptive attention, [32] indicates the need to consider processing non-visual words.

Different types of attention mechanisms have been proposed that can use weekly supervised multiple instance learning (MIL) to learn advanced concepts and combine them into sentences to solve the problem of generating non-visual words. [54] introduced a visual attribute classifier to generate semantic concepts in which image features are a vector of attribute classifier confidence. [53] has developed a novel attribute-guided model by integrating inter-attribute correlations into MIL to add the high-level semantic attributes into an RNN-based encoder-decoder framework to achieve better performance. They have constructed various architectures to feed these features to find the best way to incorporate semantic attributes. However, all attribute words are considered equally essential and incorporated into the RNN at every time step. Involving all of the probable words in the image may lead to generating some unrelated words to the image context in the final caption.

All in all, existing methods have tended to focus on either visual or attribute information. Due to their complementarity, they lead to insufficient knowledge of the given image. The other problem is that a few works consider the importance of preserving contextual information in the caption generation process. The generation of a subjective sentence to point out the salient event is strongly influenced by the scene in which the image appears.

3 The proposed method

This section introduces details of the proposed SCFC attention network for image captioning. The overview of the proposed framework is illustrated in Fig.1. The framework comprises three components: (1) the coupled visual detector and attributes predictor, (2) the SCFC attention network, and (3) captioning with SCFC-LSTM. First, we extract visual regions $V = \{v_1, \dots, v_n\}, v_i \in \mathbb{R}^h, V \in \mathbb{R}^{n \times h}$ and semantic attributes $A = \{a_1, \dots, a_c\}, a_i \in \mathbb{R}^{|\Sigma|}, A \in \mathbb{R}^{|\Sigma| \times c}$ from the given image in a coupled manner. Then, they are fed to the SCFC module. After constructing cross-modal features as principal components, the proposed CFC

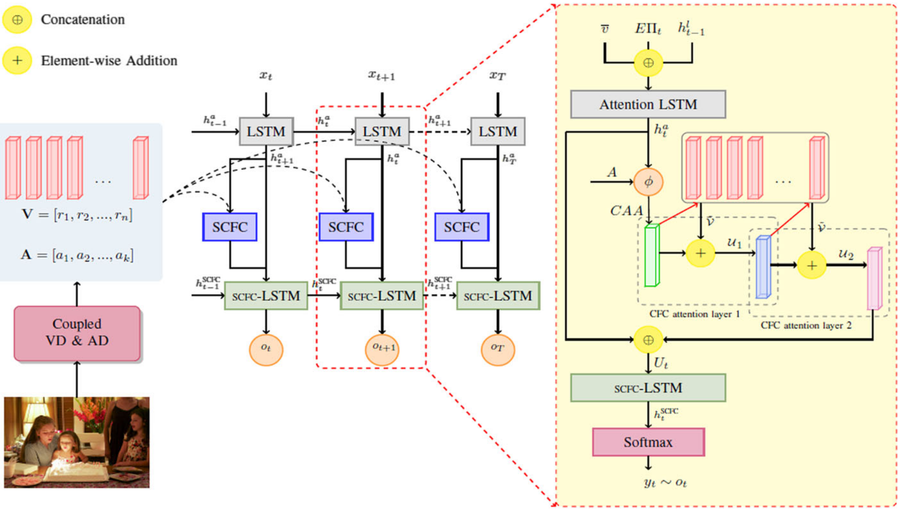

Fig. 1 The overview of the proposed SCFC for image captioning. Region proposals and attributes are extracted at the first step and then fed to the SCFC cell in each time step to consolidate cross-modal features through the caption generation process

attention network is triggered to form a compounding function inspired by the multi-step reasoning idea. Finally, the consolidated semantic features are forwarded to the SCFC-LSTM to generate semantically fine-grained image captions.

3.1 Coupled visual detector and attribute predictor

Visual and semantic features are complementary to each other. With this in mind, we develop a model that can leverage both to enhance the generation of visual and non-visual words like ”helping” and ”sitting”. In traditional works, predicting attributes is treated as an independent task and depends on a standalone stage, increasing the overall number of model parameters. Inspired by the end-to-end attribute detection in [18], we adopt an attribute predictor (AP) that can be trained jointly with the whole captioning network. Different from previous studies [2, 18], our visual detector (VD) can also be trained with the entire captioning network. In particular, we argue that training feature extractors within the whole system can boost the model to extract more task context-related features. We first detect a set of salient regions of the image, and then on the top of the visual detector, we construct our attribute predictor. The architecture of the coupled VD and AP is given in Fig. 2.

3.1.1 Visual detector

To extract the deep information of the salient candidate regions, we use ResNet-101 [17] architecture to obtain feature maps of the input image. We employ a modified region proposal network (RPN) [38] to detect regions in the given image with a set of rectangular region proposals and corresponding confidence scores.

In the proposed architecture, after the final convolutional layer of ResNet-101 [17], a 3 × 3 sliding window moves across the feature map and maps it to a lower dimension (e.g., 256-d). Multiple possible regions based on k fixed-ratio translation-invariant anchor boxes are generated for each 3 × 3 window of the feature map. Thus, the regression layer generates $4k$ output representing the bounding boxes of the regions, and the classification layer generates $2k$ outputs representing the softmax probability of each of the k bounding boxes as a confidence score. We set the value of k as 9, which includes 3 scales and 3 aspect ratios for each scale. Finally, the bilinear interpolation [20, 22] is used to enhance the nearest neighbour interpolation method in the original ROI pooling layer in [38] so that our model can extract a fixed-sized representation $V = \{v_1, \ldots, v_n\}, v_i \in \mathbb{R}^h, V \in \mathbb{R}^{n \times h}$ smoothly from each region. Moreover, bilinear interpolation allows end-to-end backpropagation through the region proposals.

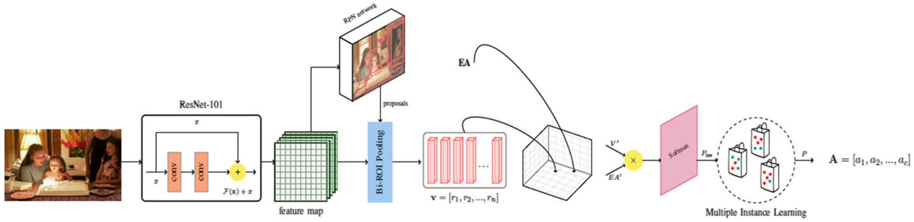
RpN network
ResNet-101
x
conv
conv
+
$\mathcal{F}(\mathbf{x})+x$
feature map
proposals
Bi-ROI Pooling
$\mathbf{v} = [r_1, r_2, ..., r_n]$
EA
V'
EA'
×
Softmax
P_{raw}
P
Multiple Instance Learning
$\mathbf{A} = [a_1, a_2, ..., a_c]$

Fig. 2 The architecture of the coupled VD and AP in our approach. Given an image, the figure shows the process of detecting visual regions V and attributes A

3.1.2 Attribute predictor

The proposed AP uses the extracted salient regions and the attribute embedding EA generated by embedding LSTM as a by-product to model the similarity between object features and attributes during the detection process. As shown in Fig. 3, the probability that a given image contains an attribute is predicted in two steps.

In the first step, we map the attribute embedding and the object features to the same space using two fully connected layers. Then, these are combined using matrix multiplication to measure the similarity. The output of these steps feeds to a softmax layer to generate the raw probability matrix $P_{raw} \in \mathbb{R}^{1000\times k}$, where P_{raw}^{ij} stands the raw probability that the j^{th} region contains the i^{th} attribute a_i. P_{raw} is obtained based on (1).

$$P_{\text{raw}} = \text{sigmoid}\left((W_{\text{AP}}EA)^T \otimes W_v V^T\right) \tag{1}$$

where $W_{AP} \in \mathbb{R}^{d\times e}$ and $W_v \in \mathbb{R}^{d\times h}$ are trainable parameters. $E \in \mathbb{R}^{e\times|\Sigma|}$ represents the embedding of all the words in the vocabulary Σ with embedding size e, $A \in \mathbb{R}^{|\Sigma|\times c}$ is the one-hot index matrix of the c attributes, $\otimes$ denotes the matrix multiplication, and the superscript T is the transpose operation.

In the second step, the probability values in each row of P_{raw} are combined using the noisy-OR multiple instance learning (MIL) [10] to generate the final probability P_i (2) that

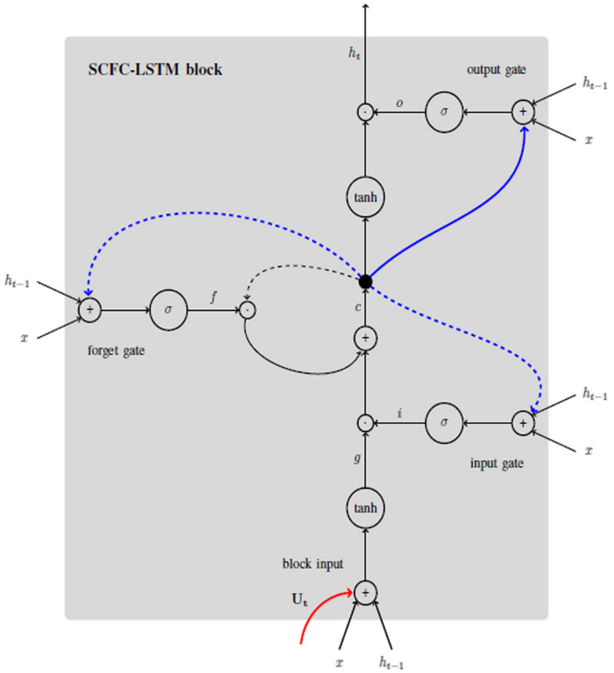

Fig. 3 The structure of our proposed SCFC-LSTM. Peephole connections and our consolidated input are shown with red and blue lines, respectively

the input image contains the i^{th} attribute a_i.

$$p_i = 1 - \prod_{j=0}^{n} \left(1 - P_{raw}^{ij}\right) \tag{2}$$

We face two imbalanced training set problems for training the coupled visual detector and attribute predictor. The number of regions proposed in the RPN network could be as high as several hundred thousand, most of which are negative examples since there is no object inside. Only a fixed number of samples with a fixed object/not-object score is sampled in classical training to overcome the class imbalance problem. Besides, the ground truth attribute vectors are sparse, as a few attributes appear in the ground truth captions.

Focal loss [28] (3) is leveraged to defeat this problem, in which all pre-located concrete anchors are taken for training with a dynamically cross-entropy loss. Although all anchors are considered, overwhelming the detector is prevented by weighting the losses of easy samples. Likewise, this modification applies to the attribute predictor, where we treat the non-existent attributes in the ground truth captions as negative examples in RPN.

$$FL(p) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), & \text{y=1} \\ -(1-\alpha)p^{\gamma}\log(1-p), & \text{otherwise} \end{cases} \tag{3}$$

We define the visual detector loss $\mathcal{L}_{VD}$ and attribute predictor loss $\mathcal{L}_{AP}$ as (4) and (5), where $\mathcal{L}_{reg}$ is the smooth $\mathcal{L}1$ loss used in the regression layer. t_i^* and p_i^* are the regression target and object/not-object labels. λ is a balancing weight, N_{cls} and N_{reg} are the normalization terms, and N_{pos} is the number of positive attributes.

$$\mathcal{L}_{VD} = \frac{1}{N_{cls}} \sum_i FL(p_i) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* \mathcal{L}_{reg}(t_i, t_i^*) \tag{4}$$

$$\mathcal{L}_{AP} = \frac{1}{N_{pos}} \sum_{j=1}^{c} FL(p_j) \tag{5}$$

It should be noted that $y = 1$ in (3) means the anchor contains an object, and the attribute a_j exists in the ground truth captions when we calculate $\mathcal{L}_{VD}$ and $\mathcal{L}_{AP}$ losses, respectively. Further, α, γ and λ hyper-parameters are empirically set to $0.3|0.95$, $2|2$, and $10|-$ for visual detector and attribute predictor losses, respectively. Finally, the $\mathcal{L}_{VDAP}$ loss is calculated using (6).

$$\mathcal{L}_{VDAP} = \mathcal{L}_{VD} + 0.5\mathcal{L}_{AP} \tag{6}$$

3.2 SCFC attention networks

At the heart of our proposed method is a compounding function to consolidate cross-modal features so that we can guide the language model in the caption generation process. Our compounding function aims to combine cross-modal features in a multi-step fashion and enable the model to attend to the constructed discriminative features in generating all semantic-level words. In the initial step, we form our principal components to feed into the proposed compounding function, which comes from two different modalities: textual $\mathcal{H}$ and visual $\mathcal{V}$. Then, we define our novel recursive function to compound cross-modal features through the attention networks.

3.2.1 The principal components

We define our principal components from two modalities: textual $\mathcal{H}$ and visual $\mathcal{V}$. The visual element is provided from the output of the visual detector. In this case, we have a set of regional feature maps to participate in the compounding function. As the second component, we leverage semantic attributes regarding the contextual environment. Previous studies employ semantic attention to involve high-level semantic concepts in the caption generation process. The kernel of semantic attention-based methods drives the model to dynamically attend to semantically essential attributes in each time step regarding the contextual information. This core has been formed by learning an attention activation state vector to calculate the weight of each attribute. Existing methods meet this goal by adding elementwise the hidden state vector of the language LSTM from the previous time step and each attribute vector. Then, the weight of each is computed through the softmax layer, and the "soft" approach is followed to obtain the output attention by using the weighted sum of the detected attributes. Note that the output attention vector may contain irrelevant attributes, making the attention guidance vague. We provide two toy examples inspired by a comprehensive investigation conducted in [16] to find the best function for measuring the similarity in considering extreme cases to make sense better.

In the first case, consider the attribute vector A as an all-one vector, and the hidden state of the language LSTM h is an all-zero vector expressing there is no contextual information. As h is an all-zero vector, there is no correlation between the attribute vector and the context. Hence, the attention activation state vector should be an all-zero vector meaning no association between the context and the attribute vector. In the second case, imagine the attribute vector A is an all-zero, and the hidden state of the language LSTM h is an all-one vector, which indicates there are no high-level concepts in the given image. Thus, the corresponding attention activation state vector is supposed to be an all-zero vector, which shows no correlation between the context and the attribute vector. Although the expected result in both examples is an all-zero vector, the attention activation state vector taken by the traditional semantic attention mechanism is an all-one vector leading to the inaccurate weights added to the detected attributes. From the examples, we can find that in some cases, the "soft" attention mechanism may lack an appropriate measure of the correlation between contextual information and high-level semantic concepts. The Context-Aware Attribute CAA suggested in this paper seeks to address this issue by formulating a function measuring the correlation between the predicted attributes and contextual information. Besides, we use the Attention LSTM [2] to represent the contextual information rather than take the hidden state of the language LSTM leading to a proper dynamic representation of the image context condition in the current linguistic context.

As illustrated in (7), attention LSTM takes the mean pooled image feature $\bar{v} = \frac{1}{N} \sum_i v_i$, the previous hidden state of the SCFC LSTM and encoding of the previously generated word as the inputs. These inputs provide the attention LSTM with maximum context regarding the state of the language LSTM, the overall content of the image, and the partial caption output generated so far, respectively.

$$h_t^{att} = LSTM[(h_{t-1}^{SCFC} \oplus \bar{v} \oplus E\Pi_t), h_{t-1}^{att}] \tag{7}$$

where E stands for the word embedding matrix shared with the attribute detector module, and Π_t is the one-hot encoding of the input word at time step t. Further, we choose linear functions to measure the correlation between the predicted attributes and contextual information as low-cost functions in terms of computational efficiency and simplicity. Due to our refinement purpose, we define Φ_{cr} as the element-wise multiplication. Thus, the semantic attention distribution β_t^j over each attribute a_t^j can be calculated using (8) and (9).

$$b_t^j = \Phi_{cr}(h_t^{att}, EA_t^j) \tag{8}$$

$$\beta_t^j = softmax(b_t^j) \tag{9}$$

where b_t^j denotes the attention activation state vector of each attribute. Then, we can construct the context-aware attributes CAA_t by multiplying the embedding of each attribute by its weight based on (10).

$$CAA_t = \sum_{j=1}^{c}(\beta_t^j \cdot EA_t^j) \tag{10}$$

It should be noted that the proposed CAA vector is supposed to represent those attributes that are associated with contextual information. Intuitively, the coexistence of the final attribute set can preserve the context dynamically in each time step through the caption generation process.

In summary, a set of regional feature maps, semantically embedding of the attributes, and contextual information are all considered to form our principal components of the compounding function. From this information, we provide visual element $V = \{v_1, \ldots, v_n\}$, $v_i \in \mathbb{R}^h$, $V \in \mathbb{R}^{n \times h}$ as the first component, and our suggested context-aware attribute CAA as the second component $\mathcal{H}$ coming from textual modality.

3.2.2 Compounding function

Given the principal components $\mathcal{V}$ and $\mathcal{H}$, we propose a new recursive function to consolidate cross-modal features in a multi-step fashion. In particular, we define an SCFC Attention Network to perform the consolidation. We then set up a stacked network with more SCFCs to operate together to simulate multi-step reasoning. Firstly, we introduce CFC, a fully functional operating standalone, and in the rest of this section, we look at how the stacked network works as a recursive compounding function.

In the SCFC, we first measure the inter-modality relations between principal component pairs to determine the relevance degree of the cross-modal features. Like the previous part, we use Φ_{cr} for calculating the relevance degree as below:

$$\alpha_{i,t} = \tanh(\Phi_{cr}(W_{\mathcal{V},\alpha} v_i, W_{\mathcal{H},\alpha}\mathcal{H}_t)) \tag{11}$$

$$\mathcal{D}_{i,t} = softmax(W_{\alpha,\mathcal{D}}^T \alpha_{i,t}) \tag{12}$$

where $W_{\mathcal{V},\alpha} \in \mathbb{R}^{d \times h}$, $W_{\mathcal{H},\alpha} \in \mathbb{R}^{d \times e}$ and $W_{\alpha,\mathcal{D}}^T \in \mathbb{R}^h$ are learned parameters. By determining the relevance degree $\mathcal{D}_{i,t}$ between textual component $\mathcal{H}$ and visual sub-components v_i, we calculate attended visual component $\mathcal{V}$ as follows:

$$\tilde{\mathcal{V}}_t^I = \sum_i \mathcal{D}_{i,t}\mathcal{V}_i \tag{13}$$

The calculated $\tilde{\mathcal{V}}_t^I$ represents the visual element, which is strained with the textual component at each time step t. To solidify the impact of the textual information, we define the compact cross-modal representation $\mathcal{U}_t$ as a second-order combination as follows:

$$\mathcal{U}_t = \tilde{\mathcal{V}}_t^I + \mathcal{H}_t \tag{14}$$

The output $\mathcal{U}_t$ is an informative representation in which the principal components $\mathcal{V}$ and $\mathcal{H}$ of a given image are encoded. Compared with the models that only adopt combined visual and semantic attention, our model constructs a richer $\mathcal{U}_t$ by imposing higher weights on the visual elements that are more relevant to the context-aware attributes. However, for abstract semantic words, a single CFC is insufficient for making inferences on the cross-modal features and generating an all-round representation. Therefore, we stacked SCFCs to iterate the above procedure imitating the multi-step reasoning approach in multi-modal tasks. Mathematically, for the s^{th} CFC layer, the SCFC takes the following formulas:

$$\alpha_{i,t}^{s} = \tanh(\Phi_{cr}(W_{\mathcal{V},\alpha}^{s} v_i, W_{\mathcal{H},\alpha}^{s} \mathcal{H}_t^{s-1})) \tag{15}$$

$$\mathcal{D}_{i,t}^{s} = softmax(W_{\alpha,\mathcal{D}}^{s,T} \alpha_{i,t}^{s}) \tag{16}$$

where $\mathcal{H}_t^{s-1}$ is the textual element from the previous SCFC layer at each time step t. It should be noted that we initialize H_t^0 with the context-aware attributes vector. In each reasoning step, the attended visual component $\mathcal{V}$ and the cross-modal representation $\mathcal{U}_t$ are obtained as below:

$$\tilde{\mathcal{V}}_t^{I,s} = \sum_i \mathcal{D}_{i,t}^{s} \mathcal{V}_i \tag{17}$$

$$\mathcal{U}_t^{s} = \tilde{\mathcal{V}}_t^{I,s} + \mathcal{H}_t^{s-1} \tag{18}$$

We repeat this algorithm S times and then use the final $\mathcal{U}_t^S$ concatenated with the contextual representation h_t^{att} as (7) to serve it to the SCFC-LSTM to generate semantically fine-grained sentences.

$$U_t = \mathcal{U}_t^S \oplus h_t^{att} \tag{19}$$

3.3 Captioning with SCFC-LSTM

There are several LSTM variants. In our work, we adopt the peephole LSTM [13, 15] model as our caption generator, which is expressed as follows:

$$\begin{aligned} i_t &= \sigma(W_i x_t + R_i h_{t-1} + P_i c_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + R_f h_{t-1} + P_f c_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + R_o h_{t-1} + P_o c_t + b_o) \\ g_t &= \tanh(W_g x_t + R_g h_{t-1} + b_g) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{20}$$

where i_t, f_t, o_t, c_t and h_t are the input gate, forget gate, output gate, memory cell, and hidden state of the peephole LSTM, respectively. σ indicates the sigmoid function, x_t shows the word input at the time step t. $P_* \in \mathbb{R}^N$, $W_* \in \mathbb{R}^{N\times M}$, $R_* \in \mathbb{R}^{N\times N}$, and $b_* \in \mathbb{R}^N$ denote peephole, input, recurrent, and bias weights, respectively, where N is the number of LSTM blocks and M is the dimension of inputs. It is worth mentioning that adding peephole connections means that we let the gate layers look at the cell state. In other words, when determining input gates, forget gates and output gates, there is a need to utilize the previous time step of the cell state c_t.

Previous studies utilize semantic features associated with either particular input locations or high-level semantic concepts as additional inputs of the language model at each time step. On the contrary, our model provides the cross-modal consolidated feature U to guide the description generation. As shown in Fig. 1, we load the constructed feature U into cell state C at each time step. Intuitively, peephole connections also allow the current time step of the gate to be aware of cross-modal informative semantic feature U in a more governable way. Thus, the fourth line in (20) must be updated as below:

$$g_t = \tanh\left(W_g x_t + W_U U_t + R_g h_{t-1} + b_g\right) \tag{21}$$

where W_U indicate a weight matrix. Finally, the probability distribution over each word in the vocabulary p_t and the word to be generated at time step t is predicted as:

$$h_t^{SCFC} = LSTM^P(h_{t-1}^{SCFC}, x_t, \mathcal{U}_t) \tag{22}$$

$$y_t \sim p_t = softmax(W^h h_t^{SCFC}) \tag{23}$$

where W indicates a weight matrix, and $LSTM^P$ denotes the LSTM with peephole connections.

3.4 Model training

The model training process consists of two rounds. In the first round, given a target ground truth sequence $y^*_{1:T}$, we optimize the model with the classical cross-entropy (XE) loss as (24), where θ stands for the captioning model parameters.

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y^*_{1:t-1})) \tag{24}$$

In the second round, we leverage deep reinforcement learning (RL) to address the exposure bias problem, which means the model has never been exposed to its predictions, resulting in accumulated errors during the inference process. From the initialization of the model trained by cross-entropy, we investigate to minimize the negative expected score corresponding to the model parameter as below:

$$\mathcal{L}_R(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})] \tag{25}$$

where r is the evaluation score metric optimized with the CIDEr-D [45] score. Considering the self-critical sequence training (SCST) [39], we approximate the gradient by the REINFORCE algorithm, given by:

$$\nabla \mathcal{L}_R(\theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y^s_{1:T}) \tag{26}$$

where $y^s_{1:T}$ is a randomly sampled caption, and $r(\hat{y}_{1:T})$ is the baseline score of the max sampled caption.

Our proposed model is trained using a single-level optimization. Specifically, the model is optimized with the overall loss function $\mathcal{L}_o$, as shown in (27). $\mathcal{L}_{VDAP}$ is defined as (6), and $\mathcal{L}_{cap}$ is cross-entropy and reinforcement learning losses in the first and second rounds, respectively.

$$\mathcal{L}_o = \mathcal{L}_{cap} + \mathcal{L}_{VDAP} \tag{27}$$

4 Experiments

4.1 Datasets

We validate the proposed model on well-known datasets, including MSCOCO [29] and Flicker30k [55]. MSCOCO contains 82,783 images for training and 40,504 images for validation. Each image has five annotated sentences. We employ the Karpathy splits [23], used widely for reporting results in prior works. This split contains 113,287 training images and 5K images for validation and testing. Flickr30k consists of 31,783 images obtained from Flickr. For a fair comparison, we use the publicly split [23] with 29,783 images for training, 1K for validation, and 1K for testing. Like MSCOCO, each image is annotated with five reference captions.

4.2 Evaluation metrics

We report the performances with the popular metrics for image captioning, including BLEU-N [35], METEOR [9], CIDEr [45], ROUGE-L [27], and SPICE [1]. We use the code published by the Microsoft COCO evaluation server to calculate all metrics. BLEU is computed by measuring the similarity of the generated sentences and the reference sentences in n-grams. METEOR is evaluated by comparing the various segments of sentences between the candidate caption and the reference caption. CIDEr is a consensus-based metric introduced specifically for image captioning tasks, which calculates consensus in image description by performing TFIDF weighting for all n-grams. ROUGE-L is used to evaluate the adequacy and fluency of machine translation, which employs the longest common subsequence between a candidate sentence and a set of reference sentences to measure their similarity at the sentence level. SPICE is determined by employing an F-score measured over tuples in the reference and candidate scene graphs, obtained through dependency parse trees.

4.3 Experimental settings

4.3.1 Data preprocessing

We adopt the standard practice and apply only minimal text preprocessing, tokenizing on white space, converting all tokens to lowercase, and discarding rare words that occur fewer than five times, resulting in the remaining words in the vocabulary of sizes 10,010 and 6,864 for MSCOCO and Flickr30k respectively. In particular, we replace less frequently occurring words with a special token $< UNK >$.

4.3.2 Implementation details

Our ground truth includes positive boxes and attribute vectors where coordinate annotations are obtained from the MSCOCO-2014 object detection dataset. To build the ground truth attribute, we select c most common words in the training caption corpus, where we set $c = 1000$ as in [12]. Then, we construct an attribute vector to determine whether each word exists in each image's description. We employ ResNet-101 with weights pre-trained on ImageNet [8] without fine-tuning in all the training phases. Our visual detector extracts regional features with a size of 36 × 2048-d according to the corresponding confidence scores as suggested in [2]. Similar to [18], attributes share the same embedding with the corresponding words, where each word is embedded in a 1000 dimensional word embedding space. The dimensions of both attention LSTM and SCFCLSTM hidden states are set to 2048-d. The Adam optimizer [25] with $\beta_2 = 0.9$, $\beta_2 = 0.999$ is utilized to minimize the loss function in (26). The basic learning rate is 1×10^{-4}. Dropout [42] and gradient clipping techniques [14] are used. In the first round of training, we adopt teacher-forced learning [48], in which we provide the ground truth words up to $t - 1$, and not the words it generated in the previous time step to train the prediction at t for sequence learning. In the testing phase, the maximum allowable sentence length is set to 20. We use the beam search strategy, and the beam size is set to 3. For more details on the implementation and source code of the proposed model, refer to https://github.com/MozhganPourKeshavarz/SCFC.

4.3.3 State-of-the-art studies

We confirm the effectiveness of our method by comparing its performance with the following state-of-the-art works:

1. NIC [46]: The first encoder-decoder framework that takes an image as input in an encoder and feeds the encoded representation into the first time step of the LSTM-based decoder to generate the corresponding description.
2. Soft-Att [50] and Hard-Att [50]: Two different spatial attention mechanisms are introduced to guide the model to selectively attend to the salient image regions in either deterministic "Soft" attention or stochastic "Hard" attention.
3. Sem-Att [54]: First, the semantic concepts are detected in the image as an independent task. Then, the global image features and the detected semantic concepts are combined, and they are progressively fed into the language model through the caption generation process.
4. LSTM-A [53]: They have suggested a novel attribute detector integrating inter-attribute correlations into multiple-instance learning (MIL) to leverage correlations between attributes. Then, they use five different forms to combine those semantically attributes into the LSTM.
5. SCA-CNN [6]: This is an improved version of visual attention that incorporates spatial, channel, and multi-layer image features to dynamically adjust the context of sentence generation.
6. SCST [39]: An advanced reinforcement learning (RL) based training method is proposed for image captioning.
7. Ada-Att [32]: Adaptive attention with a visual sentinel is proposed to determine whether to attend to the visual features or the visual sentinel.
8. RFNet [21]: They have proposed a novel recurrent fusion network (RFNet) that can exploit the interactions between the outputs of the image encoders and generate new compact and informational representations for the decoder.
9. Up-Down [2]: A combined top-down and bottom-up attention mechanism is introduced that enables attention to be determined at the level of objects and other salient image regions.
10. GCN-LSTM [52]: They have built a graph convolutional network over the detected objects in an image that integrates semantic and spatial object relationships into the image encoder in an encoder-decoder framework.
11. ARL [47]: This work considers a solution to integrate intra-regional relationships into visual content and proposes a novel visual attention mechanism to implicitly model the relationships between image regions.

12. SGAE [51]: A scene graph auto-encoder has been proposed that incorporates collocations and contextual inference into the encoder-decoder architecture as the language inductive bias by using the scene graph of the image and a trained dictionary.
13. CAVP [56]: They have proposed a new RL-based learning method and introduced a pairwise relationship learning approach in the decoder.
14. LBPF [37]: They have suggested a look back method to embed previously visual information and a predict forward strategy to look into the future to boost image captioning performance by utilizing linguistic coherence.
15. MAD+SAP [18]: This work expands semantic attention by introducing a subsequent attribute predictor module to dynamically predict a concise attribute subset at every time step to mitigate the variety of image attributes.

4.3.4 Quantitative results

Table 1 illustrates the contrastive performance comparison results on the MS-COCO dataset. The table shows that our proposed model outperforms the state-of-the-art models by a large margin in all evaluation metrics, specifically CIDEr. For a fair comparison, we also separately report the results of the ensemble models. These evaluation results indicate that the SCFC attention network boosts image captioning performance. Compared with the traditional neural image captioner NIC [46], classical visual attention Soft-Att [50], and Hard-Att [50] as benchmarks, our improvement is primarily due to the more effective collaboration of visual and semantic information. Compared with SCA-CNN [6] and LBPF [37], our model uses advanced semantic concepts to generate more diverse and semantic-enriched captions. Although Sem-Att [54] and LSTM-A [53] leverage attributes, our model can more accurately measure the correlation between attributes and contextual information, and there is no need for a separate network to train the attribute detector. Compared with methods that explore visual semantic regions by pre-training a visual detector with a large dataset (such as Visual Genome) before implementing the image captioning network (Up-Down [2], MAD+SAP [18]), we leverage the cross-modal feature consolidation layer to make up for the lack of discriminative semantic feature representation. RFNet [21] further improves image captioning performance by investigating the spatial attention mechanism.

In contrast, we employ an attention mechanism considering both semantic concepts and spatial regions. ARL [47] and GCN-LSTM [52] consider the visual relationship among regions in the image by discovering the high-level connections between regions that encode semantic concepts, thereby improving image captioning performance. At the same time, our model can capture these regions in compounding functions through the SCFC layers. CAVP [56] proposes a new RL-based learning method employed in our learning procedure to improve image captioning performance. MAD+SAP [18] utilizes visual and semantic information and focuses on the role of attributes. While keeping in mind the complementary nature of visual and semantic information, we propose a fully end-to-end model with a novel consolidation layer to efficiently generate more fine-grained captions. Unlike SGAE [51], which requires pre-training a scene graph generator and a dictionary, our model relies on the visual feature as the only input.

To further verify the effectiveness of the proposed model, we report the evaluation results of the test split of the Flickr30k dataset in Table 2. We evaluate the proposed model on the online MSCOCO test server. Table 3 reports the performance of the SCFC and other state-of-the-art works. The proposed method also achieves the best performance in most metrics.

Table 1 Performance analysis of the proposed SCFC and other state-of-the-art methods on the MSCOCO KARPATHY's test split, where B@N, M, R, C and S are the short forms of BLEU-N, METEOR, ROUGE-L, CIDER-D, and SPICE scores

	Cross-Entropy Loss						CIDEr-D Score Optimization					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
NIC [57]	66.6	24.6	-	-	-	-	-	-	-	-	-	-
Soft Att [50]	70.7	24.3	23.9	-	-	-	-	-	-	-	-	-
Hard Att [50]	71.8	25.0	23.0	-	-	-	-	-	-	-	-	-
Sem-Att [54]	70.9	30.4	24.3	-	-	-	-	-	-	-	-	-
LSTM-A [53]	73.4	54.0	100.2	18.6	78.6	35.5	27.3	56.8	118.3	20.8		
SCST:Att2in [10]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [10]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
SCA-CNN [8]	71.9	31.1	25.0	53.1	95.2	-	66.6	20.3	27.3	56.8	118.3	-
Ada-att [34]	74.2	33.2	26.6	-	108.5	-	-	-	-	-	-	-
RFNet [25]	76.4	35.8	27.4	56.5	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
Up-dOWN [11]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [42]	77.4	37.1	28.1	57.2	117.1	21.1	**80.9**	38.3	28.6	58.5	128.7	22.1
SGAE [48]	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
CAVP [6]	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
LBPF [21]	**77.8**	37.4	28.1	57.5	116.4	21.2	80.5	38.3	28.5	58.4	127.6	22.0
MAD+SAP [3]	-	37.0	28.1	57.2	117.3	**21.3**	-	38.6	28.7	58.5	128.8	**22.2**
SCFC	**77.8**	**37.5**	**28.2**	**57.6**	**119.2**	**21.3**	80.7	**38.9**	**29.0**	**58.7**	**129.9**	22.1
						Ensemble						
SCST$_{fuse}$ [10]	-	32.8	26.7	55.1	106.5	-	-	35.4	27.1	56.6	117.5	-
RFNet$_{fuse}$ [25]	77.4	37.0	27.9	57.3	116.3	21.1	80.4	37.9	28.3	58.3	125.7	21.7
SGAE$_{fuse}$ [6]	-	-	-	-	-	-	81.0	39.0	28.4	**58.9**	129.1	22.2
MAD+SAP$_{fuse}$ [3]	-	-	-	-	-	-	-	39.0	28.9	58.8	129.8	**22.3**
SCFC[Σ]	**78.0**	**37.9**	**28.2**	**57.8**	**119.9**	**21.2**	**81.2**	**39.2**	**29.1**	**58.9**	**130.8**	**22.3**

The most significant number in each column is marked in boldface

Table 2 Performance analysis of the proposed SCFC and other state-of-the-art methods on the Flicker30K publicly split using the cross-entropy loss, where B@N, M, R, C and S are the short forms of BLEU-N, METEOR, ROUGE-L, CIDER-D, and SPICE scores

	B@1	B@4	M	R	C	S
NIC [57]	66.3	18.3	-	-	-	-
Soft Att [50]	66.7	19.1	18.49	-	-	-
Hard Att [50]	66.9	19.9	18.46	-	-	-
Sem-Att [54]	64.7	23.0	18.9	-	-	-
SCA-CNN [8]	66.2	22.3	19.5	-	-	-
Ada-att [34]	67.7	25.1	20.4	-	53.1	-
ARL [14]	69.8	27.7	21.5	48.5	57.4	-
SCFC[31]	**71.2**	**27.9**	**22.2**	**49.2**	**58.1**	**15.9**

The most significant number in each column is marked in boldface

4.4 Ablation studies

4.4.1 Incrementally validation

We also conduct extensive experiments to incrementally validate our method and thoroughly show the behaviour of the proposed method. We use the following parts to ablate our model:

- Base: We construct a baseline model by following [2] to integrate an attention LSTM to the language LSTM coupled with our visual detector. Note that unlike [2], our visual detector can be trained jointly with the whole captioning network.
- Base+VDsemAtt: We inject our attribute detector through the traditional semantic attention as an extra input to the language LSTM.
- Base+CAA: To show the effectiveness of our proposed CAA, we incorporate detected attributes through the context-aware attributes module into the Base model.
- CAA+SCFC: We add the proposed stacked cross-modal feature consolidation attention network to the ablative model Base+CAA. Compared to the entire model, this variant does not adopt SCFC-LSTM.
- CAA+SCFC+SCFC-LSTM: The proposed model with CAA, SCFC attention network, and SCFC-LSTM.

The experimental results of the ablated models are reported in Table 4. The number in brackets denotes the number of stacked layers in the SCFC attention network. Our base model achieves comparable performance to [2]. The results of Base+VDsemAtt and Base+CAA demonstrate that using attributes in forming the textual component can provide better results than leveraging visual elements alone (Base). In comparison, our context-aware attributes module improves the model performance significantly. In particular, our ablative model Base+CAA can achieve 37.6 and 124.8 in the BLEU-4 and CIDEr, respectively, making the relative improvement over our baseline model with classical semantic attention by 0.6% and 2.9%, respectively. The proposed CAA performs better than traditional semantic attention on the BLEUN metric. Indicating running detected attributes through the caption generation process, regarding how much they are consistent with the contextual environment of the current step, can guide the model to attend to more context-related attributes resulting in meaningful and fluent sentences. The experiment of CAA+SCFC+ SCFC-LSTM(s), which is the full model with a different number of CFC layers, verifies the effectiveness of our stacked cross-modal feature consolidation attention network. In particular, our model CAA+SCFC+ SCFC-LSTM(3) achieves a relative improvement of 0.9%, 1.9%, 1.3%, 1.3%, 8.0%, and

Table 3 Performance analysis of the proposed SCFC and other state-of-the-art methods on the online MSCOCO test server

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40
NIC [57]	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
Hard Att [50]	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
Semantic Att [54]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
SCA-CNN [8]	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
Ada-Att [34]	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
RFNet [25]	80.4	95.0	64.9	89.3	50.1^{2}	80.1^{2}	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SCST:Att2all [10]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
LSTM-A [53]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	62.5	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [11]	80.2	95.2^{1}	64.1	88.8^{2}	49.1	79.4	36.9	98.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE [48]	-	-	-	-	-	-	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CAVP [6]	80.1	94.9	64.7	88.8^{2}	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
MAD+SAP [3]	80.5^{1}	94.9^{2}	65.1^{1}	89.1^{1}	50.4^{1}	80.0	38.4^{2}	69.4^{2}	28.6^{2}	37.7^{2}	58.7^{1}	73.3^{2}	125.1^{2}	127.0^{2}
SCFC	80.2^{2}	94.9^{2}	64.9^{2}	88.8^{2}	50.4^{1}	80.2^{1}	38.5^{1}	69.8^{1}	29.0^{1}	38.3^{1}	58.6^{2}	73.8^{1}	125.3^{1}	127.3^{1}

The superscript of each metric indicates the top 2 rankings

Table 4 The results of ablated models (Base, Base+VDsemAtt, BASE+CAA, CAA+SCFC) and our entire model CAA+SCFC+SCFC-LSTM{N} on the MSCOCO test split in CIDER-D score optimization training in terms of BLEU-N(B@N), METEOR(M), ROUGE-L(R), CIDER-D(C), and SPICE(S)

	B@1	B@4	M	R	C	S
Base	79.8	37.0	27.7	57.4	121.9	21.4
Base+ADsemAtt	80.0	37.4	28.1	57.7	123.4	21.5
Base+CAA	80.2	37.6	28.3	57.9	124.8	21.7
CAA+SCFC {1}	80.5	37.9	28.6	58.3	126.1	21.9
CAA+SCFC+SCFC-LSTM {1}	80.6	38.1	28.7	58.4	127.2	22.0
CAA+SCFC+SCFC-LSTM {2}	**80.7**	38.6	28.9	58.5	128.7	**22.2**
CAA+SCFC+SCFC-LSTM {3}	**80.7**	**38.9**	**29.0**	**58.7**	**129.9**	22.1

{N} indicates the different number of CFC layers used

0.7% in terms of BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE, respectively. Besides, our model CAA+SCFC+SCFC-LSTM(1) improves results compared with the CAA+SCFC(1) ablative models proofing the positive influence of adopting SCFC-LSTM rather than the standard LSTM.

These evaluation results indicate the efficacy of the SCFC layer. By comparing our entire model with its variants, it is not difficult to determine whether the proposed SCFC can combine complementary features to dynamically modulate a richer semantic representation of a given image at each time step. The evaluation results of models with more than one CFC layer further confirm this assertion. The entire model with three stacked CFC layers achieves a relative improvement of 8.5% in the CIDEr metric, which is remarkable in generating semantic-enriched descriptions.

4.4.2 Context-aware attributes analysis

Figure 4 shows the impact of semantic attention weights through our proposed context-aware attribute module at each step in our grown model. We only present the attention weights of three probable attributes from the attribute detector output for visual simplicity. While sentences are being generated, variations of the corresponding distinct time-varying weights are adjusted to the current context.

To better understand our model, we preserve three scenarios in CAA weights' changes and attributes' contribution in the generated captions. First, those high probability attributes obtain high weights used in the generated caption. This case almost happens when an attribute can be paired with a well-defined shape region or words describing an action. For example, in Fig. 4(c), the words "man", "playing", and "frisbee" are picked at the exact time step meaning our CAA can attend accurately to these categories of words. Second, those attributes with a high probability not only obtain high weights but also do not participate in the generated captions. This case takes place when attributes are not consistent with contextual information. For instance, in Fig. 4(b), although the word "kitchen" is predicted with a high probability (0.97%), it does not achieve a high weight. Besides, it orients the semantic tendency of the model when words like "decorating" and "cake" are generated, which is the power of the CAA module. Third, those high-probability attributes do not reach high weights through the CAA. Nevertheless, they take a seat in the generated caption. This case shows the ability of the CAA to leverage common catchwords in either spoken or written language. The term

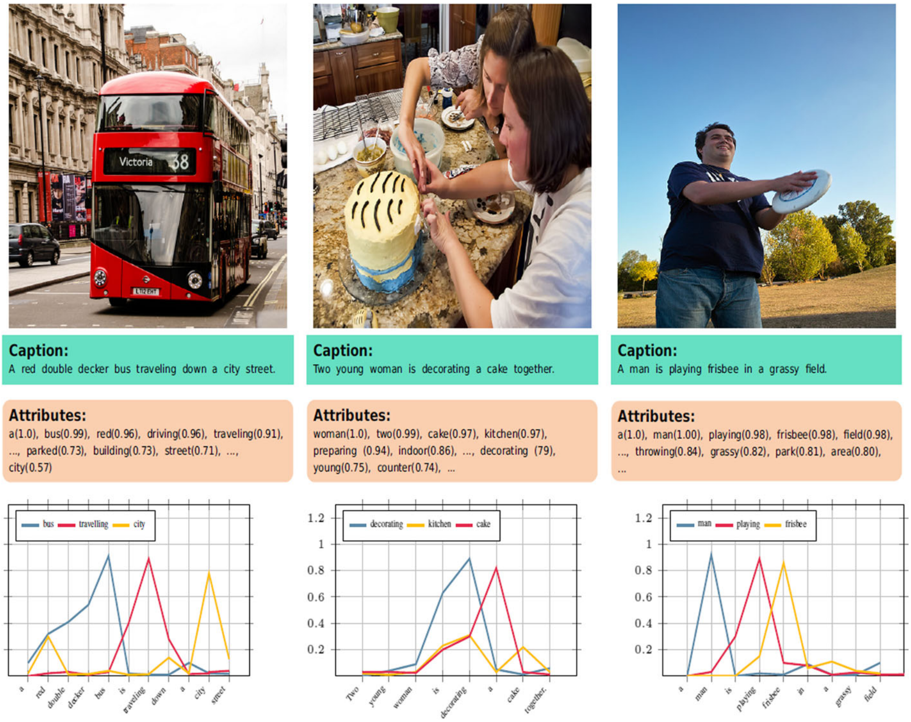

Fig. 4 Examples of attention weights' changes along with the generation of captions

"city street" in Fig. 4(a) is a good example indicating the attention to the word "city", although almost a low probability predicts it.

4.4.3 The role of SCFC attention layers

Regarding the generation of visual and non-visual words, feature compounding is beneficial to image captioning. Comparing sentences generated by the SCFC attention network and those generated by the Base model reveals this assertion. In particular, the compounding function in the SCFC layers aims to bring together semantic attributes and visual regions, leading to the generation of fine-grained captions.

We compare the result captions generated by the various number of CFC layers against the baseline model, as shown in Fig. 5. The figure shows that our SCFC attention network enables the model to perform reasoning on the input image. In particular, the captions generated by more CFC layers contain high-level words inferred by some regions and their association with some attributes that encode a semantic concept. For example, the terms "washing" in the lower-left part and "catch" in the upper-right part of Fig. 5 are such high-level words that our model can generate accurately. Furthermore, the phrase "working in a computer lab" in the upper-right part of the figure is an excellent example of generating a conceptual expression through multiple reasoning steps.

It is worth noting that captions with one CFC layer contain both visual and non-visual words regarding contextual information, indicating the strength of a standalone CFC layer. For instance, in the second example of the first row in Fig. 5, the words "standing" and

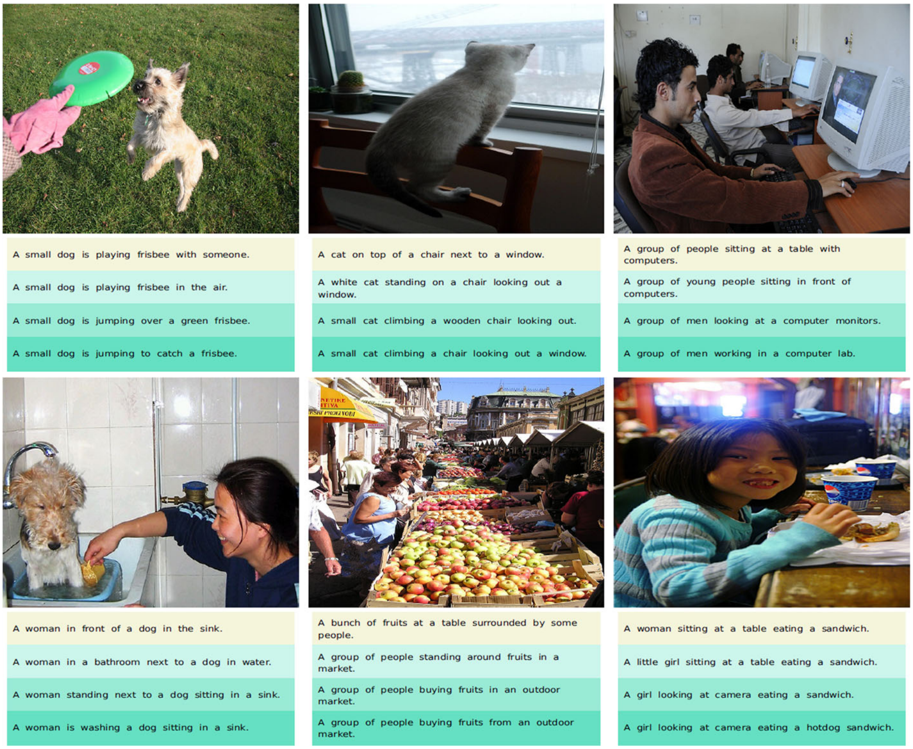
A small dog is playing frisbee with someone.
A small dog is playing frisbee in the air.
A small dog is jumping over a green frisbee.
A small dog is jumping to catch a frisbee.
A cat on top of a chair next to a window.
A white cat standing on a chair looking out a window.
A small cat climbing a wooden chair looking out.
A small cat climbing a chair looking out a window.
A group of people sitting at a table with computers.
A group of young people sitting in front of computers.
A group of men looking at a computer monitors.
A group of men working in a computer lab.
A woman in front of a dog in the sink.
A woman in a bathroom next to a dog in water.
A woman standing next to a dog sitting in a sink.
A woman is washing a dog sitting in a sink.
A bunch of fruits at a table surrounded by some people.
A group of people standing around fruits in a market.
A group of people buying fruits in an outdoor market.
A group of people buying fruits from an outdoor market.
A woman sitting at a table eating a sandwich.
A little girl sitting at a table eating a sandwich.
A girl looking at camera eating a sandwich.
A girl looking at camera eating a hotdog sandwich.

Fig. 5 Examples of caption results. The captions are generated by the Base model, SCFC, with one/two/three CFC layers in the yellow box to the solid green box, respectively

"looking" as non-visual words, and the words "sink" and "window" as visual words show that our model can selectively attend to both visual and non-visual words resulting better captions.

4.4.4 Model ensemble

Ensembles [4] have been a straightforward and effective way to enhance machine learning systems' performance for a long time. In deep architectures, one only needs to train multiple models independently on the same task, potentially varying some training circumstances and aggregating their predictions to make the outcome at inference time. For a fair comparison, we also report the ensemble performance of three SCFC models with different stacked SCFC layers in the range of 1 to 3. The result is presented in Table 1 as SCFC[Σ]. Compared with each model's results, the ensemble model boosts the performance significantly, which means various semantic-level words need a different number of reasoning steps. This achievement hints at further research on the dynamic number of stacked SCFC layers.

4.4.5 Parameter size analysis

Increasing the number of stacked SCFC layers can imitate the multi-step reasoning procedure leading to the generation of fine-grained captions. Besides, deepening the SCFC increases

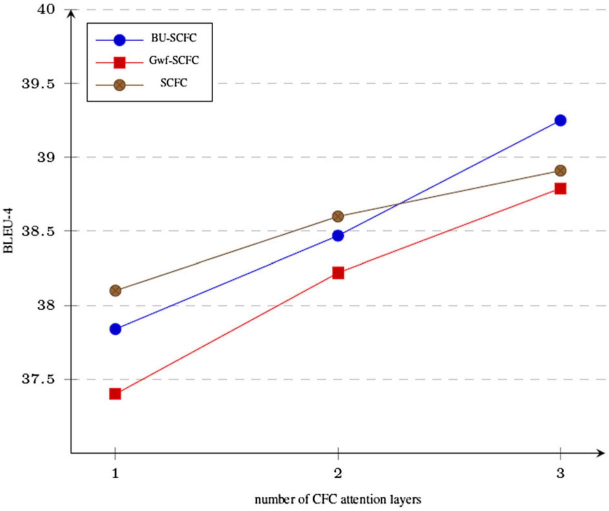

Fig. 6 The results of ablated models (BU-SCFC[s], Gwf-SCFC[s]) and our entire model SCFC[s] where s denotes the different number of CFC layers on the MSCOCO test split in CIDEr-D score optimization training in terms of BLEU-4

the overall parameter size, which can slow down the speed of convergence. To further verify the influence of growing SCFC, we conduct two experiments as follows:

- BU-SCFC: Replace our visual detector with the pre-trained bottom-up attention.
- Gwf-SCFC: Replace our word embedding with the pre-trained Glove without fine-tuning.

The experimental results are shown in Fig. 6. The figure shows that increasing the number of SCFCs in the ablative models can further enhance the captioning performance. These improvements significantly occur in the model with three SCFC layers. In particular, the ablative model BU-SCFC achieves a relative improvement of 0.34% in terms of BLEU-4 compared with the SCFC model with the same stacked SCFC numbers. Although this improvement is remarkable, it comes at the cost of a lack of end-to-end training capability. It should be noted that we do not use these parameter reduction tricks in any of the other comparison results.

4.4.6 Influence of beam size k

There are two standard procedures in the test phase to predict the next token in sentence generation. The first one is the greedy method, in which the token with the highest score at each time step is determined and used to predict the next token until it grasps the end flag or the maximum length of the caption. The second one is to use beam search to choose the best *k* subsequences at each time step and use them as applicants to generate the best *k* subsequences in the next time step. To examine the influence of the beam size in the testing step, we analyze the performances of our grown model with the various numbers of CFC layers with different beam sizes in the common range of 1 to 5 on the MSCOCO dataset. Figure 7 shows the results obtained by the normalization step according to the highest score of each evaluation metric. From the figure, we can find that as the beam size *k* increases, BLUE-4 will be enhanced in all model variations meaning we can generate sentences fluently

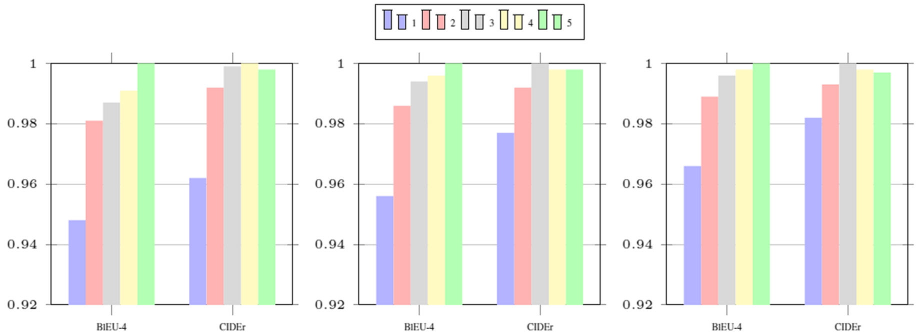

Fig. 7 The effect of beam size k on BLEU and CIDEr in our proposed SCFC with one/two/three CFC layers from left to right

by increasing the beam size in the range set in our experiment. In contrast, when the beam size k is 3, the CIDEr metric will reach its peak, except in the model with one CFC layer where the result with the beam size of 4 is slightly ahead of the result with the beam size of 3.

5 Conclusion

This paper presents an SCFC attention network for image captioning to make reasoning in multiple steps on the cross-modal features, thereby achieving a compact and informative representation in the caption generation process. We demonstrate that compounding multimodal information can boost the generation of discriminative features to generate all-level semantic words. We first form the textual component via context-aware attributes, which can jointly train through the captioning network training with the visual detector. We then propose an SCFC attention network to consolidate cross-modal features multiple times, imitating a multi-step reasoning procedure. Besides, we suggest the SCFC-LSTM encourage the model to look at the consolidated representation in a more controlled way. Our SCFC can generate more fine-grained captions. To validate the effectiveness of the proposed method, we conduct extensive experiments. Experimental results show that our method outperforms state-of-the-art models trained by both RL-based and cross-entropy losses. We also plan to design a dynamic SCFC attention network for image captioning tasks when the number of steps parameter depends on the word that should be generated rather than a predefined fixed number.

Data availability statement All datasets used in this study are well-known benchmarks freely available on the Internet.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Anderson P, Fernando B, Johnson M, Gould S (2016) “Spice: Semantic propositional image caption evaluation.” In: European conference on computer vision, Springer, pp 382–398
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (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, pp 6077–6086
3. Bahdanau D, Cho K, Bengio Y (2014) “Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473
4. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
5. Buschman TJ, Miller EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science 315:1860–1862
6. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) “Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5659–5667
7. Corbetta M, Shulman GL (2002) “Control of goal-directed and stimulus-driven attention in the brain”. Nat Rev Neurosci 3:201–215
8. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) “Imagenet: A large-scale hierarchical image database.” In: 2009 IEEE conference on computer vision and pattern recognition. Ieee, pp 248–255
9. Denkowski M, Lavie A (2014) “Meteor universal: Language-specific translation evaluation for any target language.” In: Proceedings of the ninth workshop on statistical machine translation, pp 376–380
10. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC et al (2015)“From captions to visual concepts and back.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
11. Fu K, Jin J, Cui R, Sha F, Zhang C (2017) Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE Trans Pattern Anal Mach Intell 39:2321–2334
12. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2017) “Semantic compositional networks for visual captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5630–5639
13. Gers FA, Schmidhuber J (2000) “Recurrent nets that time and count.” In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, vol. 3. IEEE, pp 189–194
14. Graves A (2013) “Generating sequences with recurrent neural networks.” arXiv preprint arXiv:1308.0850
15. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2016) Lstm: A search space odyssey. IEEE Trans Neural Netw Learn Syst 28:2222–2232
16. He C, Hu H (2019) Image captioning with text-based visual attention. Neural Process Lett 49:177–185
17. He K, Zhang X, Ren S, Sun J (2016) ”Deep residual learning for image recognition.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
18. Huang Y, Chen J, Ouyang W, Wan W, Xue Y (2020) Image captioning with end-to-end attribute detection and subsequent attributes prediction. IEEE Trans Image Process 29:4013–4026
19. Hu X, Gan Z, Wang J, Yang Z, Liu Z, Lu Y et al (2022) “Scaling up vision-language pre-training for image captioning.” In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 17980–17989
20. Jaderberg M, Simonyan K, Zisserman A et al (2015) “Spatial transformer networks.” In: Advances in neural information processing systems, pp 2017–2025
21. Jiang W, Ma L, Jiang Y-G, Liu W, Zhang T (2018) “Recurrent fusion network for image captioning.” In: Proceedings of the European conference on computer vision (ECCV), pp 499–515
22. Johnson J, Karpathy A, Fei-Fei L (2016) “Densecap: Fully convolutional localization networks for dense captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition
23. Karpathy A, Fei-Fei L (2015) “Deep visual-semantic alignments for generating image descriptions.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
24. Karpathy A, Joulin A, Fei-Fei LF (2014) “Deep fragment embeddings for bidirectional image sentence mapping.” In: Advances in neural information processing systems, pp 1889–1897
25. Kingma DP, Ba J (2014) “Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980
26. Kiros R, Salakhutdinov R, Zemel RS (2014) “Unifying visual-semantic embeddings with multimodal neural language models.” arXiv preprint arXiv:1411.2539
27. Lin C-Y (2004) “Rouge: A package for automatic evaluation of summaries.” In: Text summarization branches out, pp 74 81
28. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017)“Focal loss for dense object detection.” In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

29. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) “Microsoft coco: Common objects in context.” In: European conference on computer vision, Springer, pp 740–755
30. Li Y, Pan Y, Yao T, Mei T (2022) “Comprehending and ordering semantics for image captioning.” In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 17990–17999
31. Liu H, Yang Y, Shen F, Duan L, Shen HT (2016) “Recurrent image captioner: Describing images with spatial-invariant transformation and attention filtering.” arXiv preprint arXiv:1612.04949
32. Lu J, Xiong C, Parikh D, Socher R (2017) “Knowing when to look: Adaptive attention via a visual sentinel for image captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 375–383
33. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2015) “Deep captioning with multimodal recurrent neural networks (m-rnn).” arXiv preprint arXiv:1412.6632
34. Mao J, Xu W, Yang Y, Wang J, Yuille AL (2014)“Explain images with multimodal recurrent neural networks.” arXiv preprint arXiv:1410.1090
35. Papineni K, Roukos S, Ward T, Zhu W-J (2002) “Bleu: a method for automatic evaluation of machine translation.” In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics pp 311–318
36. Pedersoli M, Lucas T, Schmid C, Verbeek J (2017) ”Areas of attention for image captioning.” In: Proceedings of the IEEE international conference on computer vision, pp 1242–1250
37. Qin Y, Du J, Zhang Y, Lu H (2019) “Look back and predict forward in image captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8367–8375
38. Ren S, He K, Girshick R, Sun J (2015) “Faster r-cnn: Towards real-time object detection with region proposal networks.” In: Advances in neural information processing systems, pp 91–99
39. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) “Selfcritical sequence training for image captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7008–7024
40. Shao Z, Han J, Debattista K, Pang Y (2023) “Textual context-aware dense captioning with diverse words.” IEEE Trans Multimedia
41. Shao Z, Han J, Marnerides D, Debattista K (2022) “Region-object relation-aware dense captioning via transformer.” IEEE Trans Neural Netw Learn Syst
42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
43. Su Y, Li Y, Xu N, Liu A-A (2019) “Hierarchical deep neural network for image captioning.” Neural Process Lett 1–11
44. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104:154–171
45. Vedantam R, Lawrence Zitnick C, Parikh D (2015)“Cider: Consensus-based image description evaluation.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4566–4575
46. Vinyals O, Toshev A, Bengio S, Erhan D (2015) “Show and tell: A neural image caption generator.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
47. Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T (2020) Learning visual relationship and context-aware attention for image captioning. Pattern Recogn 98:107075
48. Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1:270–280
49. Wu Q, Shen C, Liu L, Dick A, Van Den Hengel A (2016) “What value do explicit high-level concepts have in vision to language problems?”. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 203–212
50. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) “Show, attend and tell: Neural image caption generation with visual attention.” In: International conference on machine learning, pp 2048–2057
51. Yang X, Tang K, Zhang H, Cai J (2019) “Auto-encoding scene graphs for image captioning.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10 685–10 694
52. Yao T, Pan Y, Li Y, Mei T (2018) “Exploring visual relationship for image captioning.” In: Proceedings of the European conference on computer vision (ECCV), pp 684–699
53. Yao T, Pan Y, Li Y, Qiu Z, Mei T (2017)“Boosting image captioning with attributes.” In: Proceedings of the IEEE international conference on computer vision, pp 4894–4902
54. You Q, Jin H, Wang Z, Fang C, Luo J (2016) “Image captioning with semantic attention.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4651–4659
55. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist 2:67–78

56. Zha Z-J, Liu D, Zhang H, Zhang Y, Wu F (2019) “Context-aware visual policy network for fine-grained image captioning.” IEEE Trans Pattern Anal Mach Intell
57. Zohourianshahzadi Z, Kalita JK (2022) Neural attention for image captioning: review of outstanding methods. Artif Intell Rev 55:3833–3862

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.