Noname manuscript No.
(will be inserted by the editor)

Incorporating Attentive Multi-Scale Context Information for Image Captioning

Jeripothula Prudviraj · Yenduri Sravani, · C. Krishna Mohan

Received: date / Accepted: date

Abstract In this paper, we propose a novel encoding framework to learn the multi-scale context information of the visual scene for image captioning task. The devised multi-scale context information constitutes spatial, semantic, and instance level features of an input mage. We draw spatial features from early convolutional layers, and multi-scale semantic features are achieved by employing a feature pyramid network on top of deep convolutional neural networks. Then, we concatenate the spatial and multi-scale semantic features to harvest fine-to-coarse details of the visual scene. Further, the instance level features are captured by employing a bi-linear interpolation technique on fused representation to hold object-level semantics of an image. We exploit an attention mechanism on attained features to guide the caption decoding module. In addition, we explore various combinations of encoding techniques to acquire global and local features of an image. The efficacy of the proposed approaches is demonstrated on the COCO dataset.

Keywords Image captioning · visual attention · multi-scale context information · image encoding mechanism

1 Introduction

Automatically generating a natural language description for a given image termed image captioning, is one of the primary goals of scene understanding. It has a tremendous significance on robotic vision, streaming platforms, and aid visually impaired users. Image captioning [56] is a more challenging task than conventional tasks such as image classification and object detection for two reasons [46]: i) The visual analysis of the scene not only accounts for objects present in an image but also must capture and understand the relationships among the objects. ii) The semantically generated caption is notably more challenging than assigning class labels to an image. Despite these challenges, a neural encoder-decoder

Jeripothula Prudviraj
Department of Computer Science
Indian Institute of Technology Hyderabad
E-mail: cs17resch01005@iith.ac.in

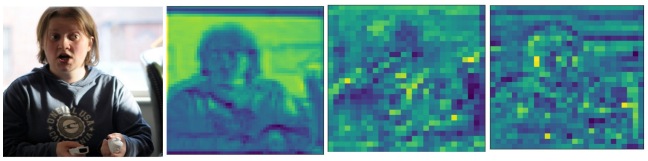

Fig. 1 Visualization of deep CNN features. Typically, earlier layers of the CNN network contain spatial features, and top layers employ semantic features. a) Input image. b) Spatial features incorporate finer details and retain higher spatial resolutions. c) Semantic features are responsible for coarse-details of an image d) Multi-scale context features own both fine-to-coarse details.

based frameworks [30,41,58] adopted from machine translation [50] are showing promising results on captioning tasks. These methods employ convolutional neural networks (CNNs) to encode the input image into a semantic feature vector, and recurrent neural networks (RNNs) are explored to decode the attained feature vector into a natural language description. Later, visual attention mechanism has been explored in image captioning tasks [61,10,32,7,46,19,58,2,39]. These mechanisms are showing an impeccable performance by learning to focus on influential image regions on each generated word. In short, most of the image captioning approaches adopt either encoder-decoder framework [56,28,24] or attention-based approaches [61,7,35,68] to generate a caption.

Existing image captioning approaches distill an entire image into semantic representation to generate a caption. However, the potential drawback of these models is that the semantic visual representation may not hold object level, instance level, and spatial information of an image. The attained feature vector of an image is limited for image captioning task due to the following reasons: i) Diminished spatial features, i.e., the spatial representation of an image diminishes as the depth of convolutional layers increases (see Figure.1b and 1c). ii) Lack of instance-level & region-level features, i.e., the model needs to incorporate fine-grained visual clues in order to generate human-like caption. iii) Static content, i.e., traditional models do not inspect various regions of an image while generating a caption. iv) Inadequate spatial localization, i.e., existing approaches interpret the visual scene using a global representation of an image instead of focusing on both local and global relevant aspects.

On the other hand, the feature pyramid network provides rich instance level [9] and object-level features [37] for semantic understanding of an image. Further, we achieve a distilled representation of the visual scene by incorporating low-level representation and interpolating multi-scale semantic features (Figure. 1d). This sort of representation is vital for image captioning tasks as it holds spatial, semantic, and region level features of a visual scene.

Motivated by the above observations, we propose attentive multi-scale context information for image captioning. The devised multi-scale contextual information learns to retain spatial, semantic, and region level features of an input image to generate a caption. In this paper, we investigate various combinations of fusing techniques to generate captions. Typically, the proposed frameworks are constructed with interpolated features of fused low-level representation and multi-scale semantic information. This is vital when the image is involved with a lot

of fine-grained objects. In addition, we incorporate the attention mechanism on combined spatial and semantic features of an image (Figure 2) to assign a weight to prominent regions.

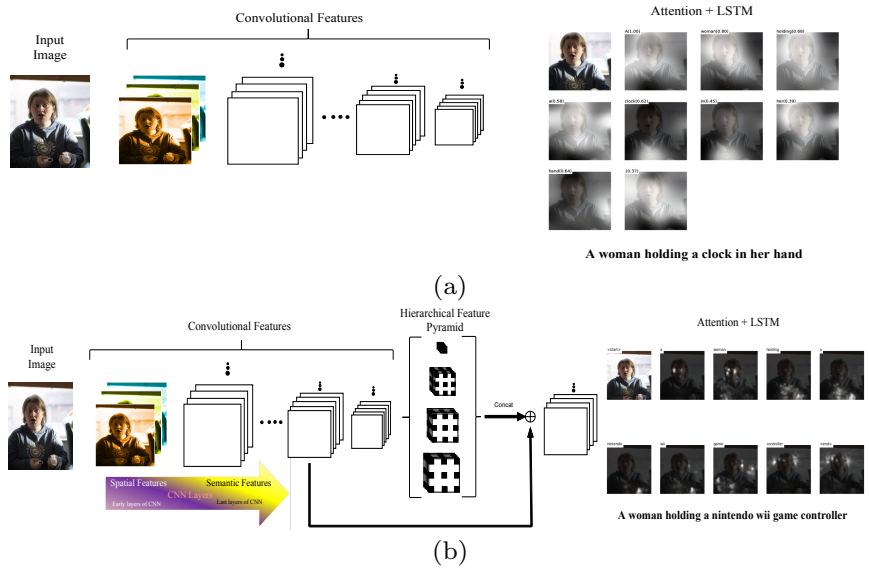

Fig. 2 Illustration of visual attention model [61] and our proposed model. (a) The visual attention model [61] employs semantic information to generate caption. (b) The proposed attentive multi-scale context information incorporates both spatial and semantic information.

At first, we explore convolutions with upsampled filters as a potential semantic representation, dubbed bottom-up and top-down encoding mechanism (BuTd, Section 4.1). The bottom-up pathway is enclosed with a conventional convolutional network, and the top-down pathway is incorporated with upsampled filters. Most of the captioning works capitalize last convolutional layer of the bottom-up pathway to encode an image into rich semantic information. However, the spatial resolution decreases with the increase of depth in the network. In other words, the representation of small objects diminishes in the top layers of convolutions. Hence, we adopt upsampling filters to restore resolution with rich semantic information in the top-down pathway.

Second, we propose a wider multi-scale context feature encoding mechanism (WMSC, Section 4.2) to investigate spatial and semantic information further. In WMSC, the context information and local information of visual features are employed by investigating various receptive fields at the top semantic layer, and then lateral connections are incorporated in between reconstructed and intermediate layers.

Third, atrous multi-scale context feature encoding mechanism (AMSC, Section 4.3) exploits atrous/ dilated convolutions to adjust filter field-of-view and procure the resolution of feature response to investigate multi-scale context information. Finally, recurrent pooling network is employed on proposed encoding frameworks

to utilize the complementary information from all encoders. This allows us to learn interactions among various viewpoints of multiple encoders and generate diverse and precise captions. The detailed framework of recurrent pooling network is discussed in Section 4.4.

Further, we make use of Xu *et al.* [61] attention-based framework to investigate multi-scale context information. The typical comparison between the visual attention model and proposed approach is depicted in Figure 2. The weights of the attention layer in the visual attention model [61] are computed on the positions of the activation grid of semantic convolutional layers. Whereas, the proposed approach attends on the local regions of holistic spatial and semantic convolutional features. The proposed framework does not explicitly depend on object detector frameworks to obtain object feature maps. Instead, it learns the latent alignment of objects present in the image. This admits that the proposed method goes beyond image level, object level, and also learns to attend instance-level concepts.

In summary, the proposed models retain spatial features by concatenating them to the last layers of CNN, and the instance level & region level features are incorporated by constructing the hierarchical feature pyramid (Figure 2). Then, attention mechanism is utilized to address dynamic nature and spatial localization for an image while generating caption. Figure 2 (b) demonstrates the adaptability of incorporated attentive multi-scale context information for image captioning task. The generated caption not only refers to dominant objects ("woman") in a scene, but also captures small objects ("nintendo wii"), object properties ("game controller"), and their interactions ("holding").

The main contributions of the proposed work are:

- We propose a novel encoding framework that accounts for object level, instance level, and spatial information of an image.
- Our encoding scheme leverages semantic information of various receptive fields at multiple scales.
- We introduce three encoding techniques to investigate visual context information.
- Unlike existing works [2,20,10,33], The proposed models do not depend on object detection frameworks, semantic tags, and external domain knowledge.
- Proposed multi-scale context information with a simple attention mechanism [61] shows comparable performance on the COCO dataset.

The remainder of the paper is organized as follows: We first present related work in Section 2. Then, Section 3 reviews the conventional image captioning framework. In section 4, we present proposed encoding mechanisms for image captioning task. The experimental results and analysis of the proposed approaches are demonstrated in Section 5. Then, we bring up various prospects of the existing models in Section 6. Finally, the conclusion is provided in Section 7.

2 Related work

The generation of captions for an image has been a challenging problem in the intersection of computer vision and natural language processing. Towards this goal, the classical approach is to use template based methods [42,31], retrieval-based models [29,43,17], and sentence generation models [16,34]. Although, these

methods bridge the gap between computer vision and natural language processing through visual elements (e.g.,objects) & language semantics (e.g., verbs), they are fixed & limited, and cannot generate natural language descriptions. To address such issues, probabilistic models like Markov chains [62] and neural networks based RNNs [20] have been explored in image caption generation.

Inspired by the success of RNNs in machine translation [27,12,50], most of the image captioning works explored encoder-decoder models [23,55,64,65,20,39, 59]. In its basic form, the convolutional network acts as encoder to attain a vectorial representation of an image and recurrent neural network serves as decoder to sequentially predict the next word in the caption. Several pioneering works [23, 55,64,65,20,39,59] rely on such combination of CNN+RNN encoder-decoder approach. For instance, Vinyals *et al.* [56] combined deep CNN with an LSTM to describe the content of an image. Similarly, Mao *et al.* [30] explored a multi-modal CNN+RNN architecture to model visual features and word embeddings. Further, the bi-directional mapping between images and their corresponding natural language sentences is explored in [11]. Zhang *et al.* [68] proposes a concept of visual keyword to align visual and semantic information for better image understanding and sentence generation. Further, Tian *et al.* [53] presents a multi-level semantic context information network to leverage the scene contextual information.

The recent surge of research interest in the image captioning task is attention based models which exploit attention mechanism in caption generation. The presence of attention is one of the aspects of the human visual system [48,13]. Inspired by this, Xu *et al.* [61] were the first to incorporate attention mechanism to the existing encoder-decoder approach for image captioning, in which attentive visual representation of an image region is utilized to update the recurrent neural network (RNN) state. At each step of RNN, attention weights for each pixel in the attention network are generated using encoded image representation and the previous hidden state of RNN. Then, the previously generated word and weighted average of encoded image representation are fed to the decoder to generate the next word in the caption. Usually, the last convolutional layer of CNNs is considered as encoded image representation, and long-short term memory networks (LSTMs) are served as a decoder. They introduced two variants of attention mechanisms in which "soft attention" formulates attention weights deterministically using the expectation of the context vector, and the "hard attention" considers a single region stochastically. These mechanisms are further extended by various approaches [2,7,10,2] to effectively focus on salient aspects of an image.

3 Review of image captioning frameworks

In this section, we first describe the conventional encoder-decoder framework for image captioning in Section 3.1, then attention based captioning is reviewed in Section 3.2

3.1 Encoder-decoder model for image captioning

This section outlines the widely adopted encoder-decoder framework [56,28] for image captioning. Suppose, we have an image I to be captioned by natural lan-

guage sentence S, where $S = \{w_1, w_2, w_3, \ldots\}$ comprised of N_s words (w). We first extract a semantic representation of an image I using top layer features of a deep CNN and then embed it through a linear projection W_I. Each word in the caption S is represented with one hot encoding and embedded with a linear embedding E, which has the same dimension as W_I. Usually, long-short term memory (LSTM) cells are utilized to generate the next word for a given image and previous words at time step t. The recursive formulation of an LSTM network is defined as:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \quad f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f),$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), c_t = i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) + f_t \odot c_{t-1},$$

$$h_t = o_t \odot tanh(c_t), z_t = W_z h_t,$$

where $x_t = E_{w_{t-1}}$ for $t > 1$ and $x_1 = W_I$. The ϕ is maxout non-linearity and σ is the sigmoid activation function. W_*, V_*, U_*, and b_* are the parameters to be learned. The distribution over the next word for obtained h_t and c_t can be formulated using softmax function, i.e.,

$$w_t = softmax(z_t). \tag{1}$$

The encoder-decoder model is proposed to learn the parameters θ by maximizing the likelihood of the observed sequence. The objective is to minimize the cross entropy loss of the model for a given image and its corresponding caption.

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t|w_1, w_2, \ldots, w_{t-1}, I)). \tag{2}$$

where $(p_\theta(w_t|w_1, w_2, \ldots, w_{t-1}, I))$ is devised from Equation 1

3.2 Attention in captioning

Noname manuscript No. (will be inserted by the editor) Incorporating Attentive Multi-Scale Context Information for Image Captioning Jeripothula Prudviraj ·Yenduri Sravani, · C. Krishna Mohan Received: date / Accepted: date Abstract In this paper, we propose a novel encoding framework to learn the multi-scale context information of the visual scene for image captioning task. The devised multi-scale context information constitutes spatial, semantic, and instance level features of an input mage. We draw spatial features from early convolutional layers, and multi-scale semantic features are achieved by employing a feature pyra- mid network on top of deep convolutional neural networks. Then, we concatenate the spatial and multi-scale semantic features to harvest fine-to-coarse details of the visual scene. Further, the instance level features are captured by employing a bi-linear interpolation technique on fused representation to hold object-level se- mantics of an image. We exploit an attention mechanism on attained features to guide the caption decoding module. In addition, we explore various combina- tions of encoding techniques to acquire global and local features of an image. The efficacy of the proposed approaches is demonstrated on the COCO dataset. Keywords Image captioning ·visual attention ·multi-scale context information · image encoding mechanism 1 Introduction Automatically generating a natural language description for a given

image termed image captioning, is one of the primary goals of scene understanding. It has a tremendous significance on robotic vision, streaming platforms, and aid visually impaired users. Image captioning [56] is a more challenging task than conven- tional tasks such as image classification and object detection for two reasons [46]: i) The visual analysis of the scene not only accounts for objects present in an image but also must capture and understand the relationships among the objects. ii) The semantically generated caption is notably more challenging than assign- ing class labels to an image. Despite these challenges, a neural encoder-decoder Jeripothula Prudviraj Department of Computer Science Indian Institute of Technology Hyderabad E-mail: cs17resch01005@iith.ac.in 2 Jeripothula Prudviraj et al. Fig. 1 Visualization of deep CNN features. Typically, earlier layers of the CNN network contain spatial features, and top layers employ semantic features. a) Input image. b) Spatial features incorporate finer details and retain higher spatial resolutions. c) Semantic features are responsible for coarse-details of an image d) Multi-scale context features own both fine-to- coarse details. based frameworks [30, 41, 58] adopted from machine translation [50] are showing promising results on captioning tasks. These methods employ convolutional neural networks (CNNs) to encode the input image into a semantic feature vector, and re- current neural networks (RNNs) are explored to decode the attained feature vector into a natural language description. Later, visual attention mechanism has been explored in image captioning tasks [61, 10, 32, 7, 46, 19, 58, 2, 39]. These mechanisms are showing an impeccable performance by learning to focus on influential image regions on each generated word. In short, most of the image captioning approaches adopt either encoder-decoder framework [56, 28, 24] or attention-based approaches [61, 7, 35, 68] to generate a caption. Existing image captioning approaches distill an entire image into semantic rep- resentation to generate a caption. However, the potential drawback of these models is that the semantic visual representation may not hold object level, instance level, and spatial information of an image. The attained feature vector of an image is limited for image captioning task due to the following reasons: i) Diminished spa- tial features, i.e., the spatial representation of an image diminishes as the depth of convolutional layers increases (see Figure.1b and 1c). ii) Lack of instance-level region-level features, i.e., the model needs to incorporate fine-grained visual clues in order to generate human-like caption. iii) Static content, i.e., traditional models do not inspect various regions of an image while generating a caption. iv) Inade- quate spatial localization, i.e., existing approaches interpret the visual scene using a global representation of an image instead of focusing on both local and global relevant aspects. On the other hand, the feature pyramid network provides rich instance level [9] and object-level features [37] for semantic understanding of an image. Further, we achieve a distilled representation of the visual scene by incorporating low- level representation and interpolating multi-scale semantic features (Figure. 1d). This sort of representation is vital for image captioning tasks as it holds spatial, semantic, and region level features of a visual scene. Motivated by the above observations, we propose attentive multi-scale con- text information for image captioning. The devised multi-scale contextual infor- mation learns to retain spatial, semantic, and region level features of an input image to generate a caption. In this paper, we investigate various combinations of fusing techniques to generate captions. Typically, the proposed frameworks are constructed with interpolated features of fused low-level representation and multi- scale semantic information. This is vital when the image

is involved with a lot Incorporating Attentive Multi-Scale Context Information for Image Captioning 3 of fine-grained objects. In addition, we incorporate the attention mechanism on combined spatial and semantic features of an image (Figure 2) to assign a weight to prominent regions.A woman holding a clock in her hand Convolutional FeaturesAttention + LSTM Input Image (a)Concat A woman holding a nintendo wii game controller Convolutional Features Hierarchical Feature Pyramid Attention + LSTM Input Image Spatial Features Early layers of CNN Semantic Features Last layers of CNNCNN Layers (b) Fig. 2 Illustration of visual attention model [61] and our proposed model. (a) The visual attention model [61] employs semantic information to generate caption. (b) The proposed attentive multi-scale context information incorporates both spatial and semantic information. At first, we explore convolutions with upsampled filters as a potential seman- tic representation, dubbed bottom-up and top-down encoding mechanism (BuTd, Section 4.1). The bottom-up pathway is enclosed with a conventional convolu- tional network, and the top-down pathway is incorporated with upsampled filters. Most of the captioning works capitalize last convolutional layer of the bottom-up pathway to encode an image into rich semantic information. However, the spatial resolution decreases with the increase of depth in the network. In other words, the representation of small objects diminishes in the top layers of convolutions. Hence, we adopt upsampling filters to restore resolution with rich semantic information in the top-down pathway. Second, we propose a wider multi-scale context feature encoding mechanism (WMSC, Section 4.2) to investigate spatial and semantic information further. In WMSC, the context information and local information of visual features are em- ployed by investigating various receptive fields at the top semantic layer, and then lateral connections are incorporated in between reconstructed and intermediate layers. Third, atrous multi-scale context feature encoding mechanism (AMSC, Section 4.3) exploits atrous/ dilated convolutions to adjust filter field-of-view and procure the resolution of feature response to investigate multi-scale context information. Finally, recurrent pooling network is employed on proposed encoding frameworks 4 Jeripothula Prudviraj et al. to utilize the complementary information from all encoders. This allows us to learn interactions among various viewpoints of multiple encoders and generate diverse and precise captions. The detailed framework of recurrent pooling network is discussed in Section 4.4. Further, we make use of Xu et al. [61] attention-based framework to investi- gate multi-scale context information. The typical comparison between the visual attention model and proposed approach is depicted in Figure 2. The weights of the attention layer in the visual attention model [61] are computed on the positions of the activation grid of semantic convolutional layers. Whereas, the proposed approach attends on the local regions of holistic spatial and semantic convolutional features. The proposed framework does not explicitly depend on object detector frameworks to obtain object feature maps. Instead, it learns the latent alignment of objects present in the image. This admits that the proposed method goes beyond image level, object level, and also learns to attend instance-level concepts. In summary, the proposed models retain spatial features by concatenating them to the last layers of CNN, and the instance level region level features are incorporated by constructing the hierarchical feature pyramid (Figure 2). Then, attention mechanism is utilized to address dynamic nature and spatial localization for an image while generating caption. Figure 2 (b) demonstrates the adaptability of incorporated attentive multi-scale context information for image captioning

task. The generated caption not only refers to dominant objects ("woman") in a scene, but also captures small objects ("nintendo wii"), object properties ("game controller"), and their interactions ("holding"). The main contributions of the proposed work are: – We propose a novel encoding framework that accounts for object level, instance level, and spatial information of an image. – Our encoding scheme leverages semantic information of various receptive fields at multiple scales. – We introduce three encoding techniques to investigate visual context informa- tion. – Unlike existing works [2, 20, 10, 33], The proposed models do not depend on object detection frameworks, semantic tags, and external domain knowledge. – Proposed multi-scale context information with a simple attention mechanism [61] shows comparable performance on the COCO dataset. The remainder of the paper is organized as follows: We first present related work in Section 2. Then, Section 3 reviews the conventional image captioning framework. In section 4, we present proposed encoding mechanisms for image captioning task. The experimental results and analysis of the proposed approaches are demonstrated in Section 5. Then, we bring up various prospects of the existing models in Section 6. Finally, the conclusion is provided in Section 7. 2 Related work The generation of captions for an image has been a challenging problem in the intersection of computer vision and natural language processing. Towards this goal, the classical approach is to use template based methods [42, 31], retrieval- based models [29, 43, 17], and sentence generation models [16, 34]. Although, these Incorporating Attentive Multi-Scale Context Information for Image Captioning 11 features from small receptive fields. The attained multi-scale semantic information is upsampled using bi-linear interpolation and then concatenated with Conv3. The new fusion feature map termed multi-scale context feature is associated with an attention mechanism to produce encoding representation of an image. Encoder-decoder framework: In our work, we leverage Vanilla LSTM net- work to generate captions, and the feature map obtained from the proposed multi- scale context feature encoding scheme serves as an encoder. Training methodology: The detailed framework of wider multi-scale context feature encoding is depicted in Figure 4. In this work, both semantic features Conv5 and spatial features Conv3 are extracted from pre-trained ResNet [21]. The semantic features are fed to the proposed wider network where several 1 $\times$1 convolutions are employed to reduce the depth of the feature map and exploited various receptive fields of size 3 $\times$3 and 5 $\times$5. And, global average pooling and 1$\times$1 convolutions are employed parallel to the standard convolutions. The standard 3 $\times$3 and 5 $\times$5 convolutions are convolved using depthwise separable convolutions. Features of all receptive fields are concatenated and convolved with a 1 $\times$1 filter to further reduce the depth of concatenated features. The obtained aggregated features are restored to a spatial resolution of Conv3 with bi-linear interpolation then concatenated to convolved features of Conv3. Finally, multi scale context features are encoded to 32 $\times$32 $\times$512 feature map and established with a 512 dimensional attention layer. The annotation vectors of attention module are fed to LSTM to generate a caption, where LSTM is initialized with the average of attention vectors and 512-dimensional hidden units. 4.3 Atrous multi-scale context feature encoding mechanism Our goal is to employ the most desirable semantic and spatial information for image captioning task. In this work, we exploit atrous convolutions by adopting multiple atrous rates on various receptive fields to construct a deep convolutional network that aggregates multi-scale contextual information without losing spatial resolution. Atrous dilated convolution is an effective tool

to explicitly regulate filter's field-of-view and controls the resolution of attained feature maps [8]. In addition, it supports the exponential expansion of the field-of-view without loss of contextual information and spatial resolution [66]. This module can be plugged into our proposed framework at various resolutions. The proposed atrous multi-scale context feature module distill image-level fea- tures by harvesting the convolutional features at multiple scales, which probes a global context. We also establish a lateral connection with the concatenated features of the atrous module to provide local context. Further, the attention mechanism is investigated on aggregated features and fed to the LSTM. Atrous/ dilated convolutions: Deep convolutional neural networks are the de facto for encoding images in the task of image captioning. However, the con- tinual use of max-pooling and striding in the layers of the network substantially reduces the spatial resolution of the output feature map. To counterbalance, we acclaim the use of "atrous convolution", originally proposed to address the com- putations of undecimated wavelet transformation [8]. Incorporating Attentive Multi-Scale Context Information for Image Captioning 13 Training methodology: We revisit a multi-scale context feature encoding method by incorporating atrous convolutions. Conv5 feature map of pre-trained ResNet is processed through multiple atrous rates like 2,4,6 on 3 ×3 receptive field along with 1 × 1 and average pool filters. The concatenated feature map produces a pyramid of semantic features. The number of channels of obtained feature map reduced to 512 using 1 ×1 convolutions. The encoded features are bi-linearly interpolated to increase spatial resolution on the feature pyramid. Then, the later connection is established using a convolved Conv3 feature map. After the concatenation, we apply a 3×3 depthwise separable convolutions with 512 channels to refine the concatenated features. The concatenated feature map 32 ×32 ×512 is broadcasted through the attention layer before feeding it to the LSTM network. 4.4 Recurrent pooling network Even though the individual encoding mechanisms encode effective information of an image, the multiple encoders characterize the complementary behaviours of various encoding schemes [26, 18]. Motivated by the above observation, we leverage a recurrent pooling network in order to combine complementary information from all our encoders. We employ multiple encoders i.e., BuTd, WMSC, AMSC, and Resnet-101baseline with multiple RNNs to generate diverse and precise captions of an image. The overview of the recurrent pooling network is illustrated in Figure 6. The proposed framework has two phases, where we employ multiple LSTMsPooling Resnet BuTd WMSC AMSC LSTM LSTM LSTM LSTM Multi-attention LSTM caption Fig. 6 Recurrent pooling network with attention mechanism on multiple encoders to generate caption. At first, each individual LSTM network computes hidden states on each encoder component. Then, we pool all hidden states and share among the all components to learn the interactions of one component with other set of components. The hidden state of qth component is computed at time step t as [hq t,cq t] = LSTMq t (Ht,Attq t(Eq,hq t1)) (4) where Ht is the pool of hidden states, Attq t is the attention module, Eq is the encoding component, and ht ct are the hidden and context vectors of LSTM network.

In this section, we summarizes the conventional attention mechanism proposed in [61] for image captioning task. The set of feature vectors of an image, referred to as annotation vectors, are captivated from the top layer of the CNN network. The convolutional layer features allow the decoder to deliberately focus on prominent regions of an image by choosing a subset of all the annotation vectors. Let a =

$\{a_1, a_2, \ldots, a_L\}$ be the set of annotation vectors. Then, a_i, i= $\{1, 2, \ldots, L\}$ is the extracted feature vector at image location i. For each location i, the attention mechanism assigns a positive weight of α_i. In other words, the significance of location i is blending with a_i's together. The attention weight α_i is computed by attention model q_{att} using annotation vector a_i and previous hidden state of LSTM cell h_{t-1}. The attention module is formulated as $b_{ti} = q_{att}(a_i, h_{t-1})$, $\alpha_{ti} = \frac{\exp(b_{ti})}{\sum_{k=1}^{L} \exp e_{tk}}$. From the attained attention weights, context vector (z_t) is updated as $z_t = \psi(a_i, \alpha_i)$, where ψ is a function that outputs a single vector from the set of annotation vectors and their corresponding attention weights.

4 Attentive Multi-Scale Context Information Model

Despite the advancements in the captioning models, they still lack accountability and explainability − i.e., there is a need for accounting fine-grain details along with coarse-grain details of an image, and modelling their relationships. Most of the image captioning approaches utilize the output of the last layer of CNN to encode the global representation of an image. This representation holds rich semantic information of objects and robust to critical appearance variations in an image. However, the extracted global representation lacks spatial resolution and suffers from the mislocalization of subtle objects in a scene. In contrast, early layers of convolutional networks contribute rich spatial information and support precise localization [41].

In this paper, we interpret multi-scale context information as an encoding representation of an image to generate captions. Further, a classical attention mechanism [61] is adopted to focus on prominent feature regions of an encoded representation. Section 4.1 presents a bottom-up and top-down encoding mechanism which restores the spatial resolutions of a convolved feature with upsampling filters. Section 4.2 introduces a wider multi-scale context feature encoding technique by concatenating both global and local features. Where, the top convolutional layer is convolved with various receptive fields, then integrated with the intermediate layers of CNN to provide spatial resolutions. Further, Section 4.3 presents atrous multi-scale context feature encoding model, which inspects atrous convolution with various field-of-view on the semantic feature map and combined with spatial features.

4.1 Bottom-up and top-down encoding mechanism

Existing captioning works generate captions by leveraging the bottom-up pathway as an encoding mechanism. In this work, our key idea is to employ a top-down pathway to the existing bottom-up pathway for encoding the image, which has high-level semantics. The objective of the proposed approach is to establish high-level information by exploiting a semantic feature map from the bottom-up pathway and low-level features from the top-down pathway. The proposed bottom-up and top-down encoding mechanism is shown in Figure 3. The proposed model typically follows the encoder-decoder framework. The encoder model takes a pre-processed image as input and outputs a unified representation of global and local

semantics of an image. Then, the decoder model leverages long-short term memory to generate a caption. We adopt the conventional attention mechanism prescribed in Section 3.2 to aim attention at significant regions. The construction of various modules is stated as follows.

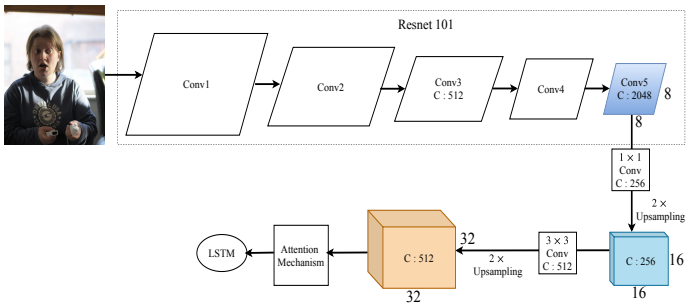

Fig. 3 Bottom-up and top-down encoder framework

Bottom-up pathway: The bottom-up pathway is a typical feed forward convolutional neural network, which formulates a feature hierarchy by convolving the input feature map at several scales with a scaling step of 2. The deeper network potentially increases the semantic value at each layer. The potential objective in the bottom-up pathway is to devise rich semantic information of an input. We utilize ResNet [21] architecture as a backbone network for our bottom-up pathway. It consists of five residual blocks which we refer as $Conv_i$ for $i = \{1, 2, 3, 4, 5\}$ with implicit convolution layers. The spatial dimension at each step is reduced to 0.5 of its original size. The output of the last layer convolution module is fed to the top-down pathway. This setting is sophisticated since the deepest layer of CNN has the substantial features.

Top-down pathway: The proposed top-down pathway takes a convolutional feature map ($Conv_5$) from the bottom-up pathway, which inherently contains rich semantic features. Then, two stage upsampling is performed at a step of 2 to boost the spatial resolution of the input feature. The depth of the feature map is reduced using convolutional filters, and spatial resolution is increased using bilinear interpolation to make a compact and unified representation of both spatial & semantic features. Further, a soft attention mechanism is employed on the upsampled feature map. The attention layer learns to focus on important spatial regions of the feature map dynamically by conditioning on hidden states of LSTM and previously generated words. Finally, LSTM generates a natural language sentence, as described in Section 3.

Training methodology: Unifying top-down pathway with a bottom-up pathway is only worthwhile when the latter approach can represent semantic concepts in deeper layers. Hence, we typically employ a pre-trained bottom-up network. Then, we construct the top-down path on the final convolutional layer of pre-trained network. Please refer to Figure 3. for a detailed framework.

We consider a $Conv_5$ feature map of pre-trained ResNet architecture, which has a spatial resolution of 8×8 with 2048 channels for a given pre-processed image. This feature map is fed to the top-down pathway. First, we reduce the dimensionality of

the attained feature map to 256 channels with 1×1 convolutions. Then the spatial resolution is doubled using bilinear interpolation. Further, receptive field, number of channels, and spatial resolution are increased by employing 3×3 convolutions with 512 channels & bilinear interpolation. The fundamental units like attention unit, hidden state, and cell state are initialized with 512-dimensional vectors.

The Bottom-up and top-down encoding mechanism constituted with rich semantic features but lacks multi-scale context features, which possess global information at multiple scales. This is essential when an image involved with a lot of small objects. Hence, we investigate multi-scale context features along with spatial features of an image in further sections.

4.2 Wider multi-scale context feature encoding mechanism

In scene understanding, both semantic and spatial information of an image play a vital role. Semantic features typically contain the context information of dominant objects, and spatial features usually hold the fine-grained information of objects in a scene. However, the reconstructed feature map of the bottom-up and top-down encoding mechanism is semantically strong but does not hold information of small objects. Most of the finer details are lost in the initial layers of CNN architecture. Hence, we need to design a framework that outputs the representation of both coarse-grain and fine-grain objects. Therefore, in this work, we incorporate both features in order to caption the image. By taking advantage of the powerful representation of convolutional networks, image classification, object detection, and image captioning have achieved impressive progress. Different from existing approaches, we incorporate various receptive fields over the final layer of CNNs architecture.

We construct a wider multi-scale network by incorporating various receptive fields on top of semantic features ($Conv_5$). The obtained multi-scale semantic features are further up-sampled to restore spatial resolutions, then concatenated with spatial feature maps ($Conv_3$) through lateral connections to equip fine-grain details. Further, each multi-scale context feature is associated with attention weights α_i to focus on both fine-grain and coarse-grain objects selectively. The detailed framework is illustrated as follows:

Multi-scale semantic features: In visual representation, multi-scale context information is crucial for two reasons. i) understanding of scene context relationship plays a prominent role in complex scene understanding, especially when a scene involved with co-occurrent visual patterns. For instance, a game controller likely to be in hand although human recognized in the scene. ii) Usually, the visual scene includes various arbitrary size objects. Certain small objects like signboards or game controllers are hard to detect while looking at the global representation of the scene, and they may be essential while captioning. In brief, most errors are entirely or partly related to both global information and the contextual relationship of objects.

Typically, the size of the receptive field in a deep neural network characterizes the context information which we use. Zhao *et al.* [10] demonstrates that the receptive field is much smaller on high-level layers of CNN (ResNet), and effective representation of small or large objects can be achieved by gathering various receptive fields information. By considering the above observations, in this work, we

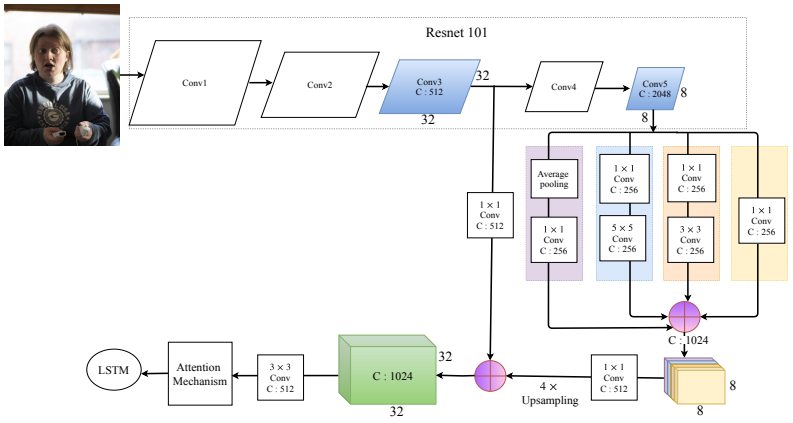

Fig. 4 Wider multi-scale context feature encoding framework

exploit the potency of global context information by aggregating various receptive field's semantic information. The multi-scale context feature module fuses semantic features under four different scales. We achieve this by introducing global average pooling and receptive fields of multiple scales on global representation. Global average pooling is widely adopted as global contextual prior in image classifications tasks [24, 11] and semantic segmentation [9]. And, semantic information from different receptive fields at multiple scales helps to attain representation of various categories without uncertainty. The construction of a multi-scale semantic feature module upon the final layer feature map (Conv5) of the deep convolutional network (ResNet) is illustrated in Figure 4. We employ various receptive fields of varied sizes i.e. 1×1, 3×3 and 5×5 along with global average pooling, inspired by the Inception framework [51]. 3×3 and 5×5 convolutional filters are employed with depthwise separable convolutions to reduce the number of parameters in each weight matrix. 1×1 convolutions are incorporated before each receptive filed to reduce the dimension of the semantic feature in order to preserve the same weight at each output feature. Then we concatenate the semantic feature and further reduced to 512-dimensional feature map. Attained feature map is upsampled via bilinear interpolation to restore spatial resolution of the feature map. Note that the number of receptive fields and the size of each filter can be modified.

Depthwise separable convolution: The classical Inception module exploits 1×1 convolutions to pay attention to cross channel correlations and then maps all correlations using regular 3×3 or 5×5 convolutions. On the other hand, depthwise separable convolutions factorize a standard convolution into a depthwise convolution, followed by a point-wise convolution. This form of convolutions highly reduces computational complexity. Typically, the spatial convolutions for each input channel are carried out by depthwise convolution, whereas the point-wise convolution is incorporated to integrate the output of the depthwise convolutions.

Semantic and Spatial Content Fusion: The proposed method builds on the prospect of the last convolutional layers of deep CNN network holds semantic abstraction, and the early layers preserve fine-grained spatial details of the input image. For instance, the area of "remote controller" in Figure 4. is about 20×40, and we could attain substantial representation only within the shallow layers of

ConvNet ($Conv_2$ or $Conv_3$). Finer details of the "remote controller" will gradually diminish on the following layers and be completely lost on the coarse, the deepest layer. Hence, we propose a fusion module to take advantage of the shallow layers to incorporate rich fine details and semantic information from the coarse layers of a deep network.

We can analyze semantic spatial content fusion from Figure 4. The early feature map ($Conv_3$) inherently has a high spatial resolution and possesses refined information of small objects. We add semantic information to these spatial features to construct semantic context information through fusion. We typically choose multi-scale semantic information rather than semantic information in order to acquire features from small receptive fields. The attained multi-scale semantic information is upsampled using bi-linear interpolation and then concatenated with Conv3. The new fusion feature map termed multi-scale context feature is associated with an attention mechanism to produce encoding representation of an image.

Encoder-decoder framework: In our work, we leverage Vanilla LSTM network to generate captions, and the feature map obtained from the proposed multi-scale context feature encoding scheme serves as an encoder.

Training methodology: The detailed framework of wider multi-scale context feature encoding is depicted in Figure 4. In this work, both semantic features $Conv_5$ and spatial features $Conv_3$ are extracted from pre-trained ResNet [21]. The semantic features are fed to the proposed wider network where several 1×1 convolutions are employed to reduce the depth of the feature map and exploited various receptive fields of size 3×3 and 5×5. And, global average pooling and 1×1 convolutions are employed parallel to the standard convolutions. The standard 3×3 and 5×5 convolutions are convolved using depthwise separable convolutions. Features of all receptive fields are concatenated and convolved with a 1×1 filter to further reduce the depth of concatenated features. The obtained aggregated features are restored to a spatial resolution of $Conv_3$ with bi-linear interpolation then concatenated to convolved features of $Conv_3$. Finally, multi scale context features are encoded to $32 \times 32 \times 512$ feature map and established with a 512 dimensional attention layer. The annotation vectors of attention module are fed to LSTM to generate a caption, where LSTM is initialized with the average of attention vectors and 512-dimensional hidden units.

4.3 Atrous multi-scale context feature encoding mechanism

Our goal is to employ the most desirable semantic and spatial information for image captioning task. In this work, we exploit atrous convolutions by adopting multiple atrous rates on various receptive fields to construct a deep convolutional network that aggregates multi-scale contextual information without losing spatial resolution. Atrous dilated convolution is an effective tool to explicitly regulate filter's field-of-view and controls the resolution of attained feature maps [8]. In addition, it supports the exponential expansion of the field-of-view without loss of contextual information and spatial resolution [66]. This module can be plugged into our proposed framework at various resolutions.

The proposed atrous multi-scale context feature module distill image-level features by harvesting the convolutional features at multiple scales, which probes a global context. We also establish a lateral connection with the concatenated

features of the atrous module to provide local context. Further, the attention mechanism is investigated on aggregated features and fed to the LSTM.

Atrous/ dilated convolutions: Deep convolutional neural networks are the de facto for encoding images in the task of image captioning. However, the continual use of max-pooling and striding in the layers of the network substantially reduces the spatial resolution of the output feature map. To counterbalance, we acclaim the use of "atrous convolution", originally proposed to address the computations of undecimated wavelet transformation [8].

Given input feature map x, convolutional filter w, and output feature map y of two-dimensional signals, the atrous convolution is applied as follows:

$$y[i] = \sum_{k} x[i + r \cdot k]w[k], \tag{3}$$

where the atrous rate r denotes the stride with which we sample the input signal. The standard convolution is defined at the rate $r = 1$. As we tune atrous rate, the atrous convolutions adaptively modify the filter's filed-of-view.

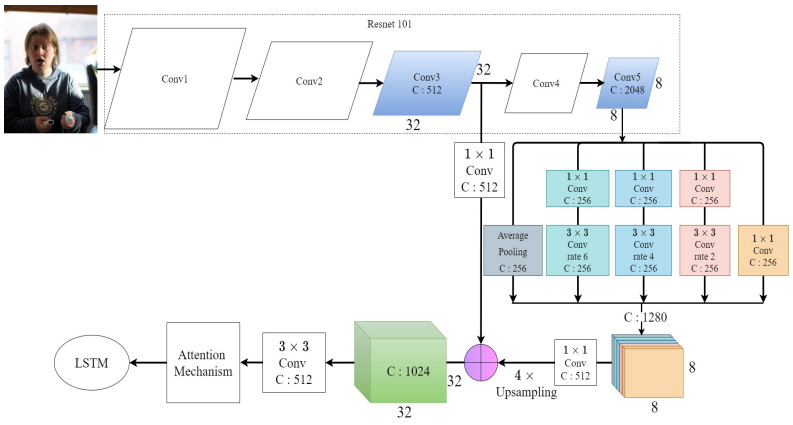

Fig. 5 Atrous multi-scale context feature encoding framework

Multi-scale context information This module presents the systematic way of aggregating the multi-scale context information by leveraging atrous/ dilated convolutions without losing spatial resolutions. We propose atrous convolutions by the fact that they support the exponential expansion of the receptive field of feature map with the change of the atrous rate r.

The context module employs atrous convolution on top of a deep convolutional network at multiple dilation rates. First, we map the output feature map to 256 channels using 1×1 convolutions, and then the atrous convolutions are incorporated by applying 3×3 convolutions with different rates. The dilation rates are 1, 6, 4, 2, 1, where rate $r = 1$ is employed for average pooling and 1×1 convolutions. Each 3×3 convolutions are computed using depthwise separable convolutions. i.e., the standard convolution with dilation in the first two dimensions followed by pointwise convolutions. Specifically, we employ 1×1 convolutions and global average pooling on the last feature map of the deep network to adopt image-level

features from the global context information. Whereas, the atrous module with different rates effectively captures multi-scale context information. To be concrete, we consider the last layer feature map of ResNet 101 (denoted as Conv5 in Figure 5) and apply parallel atrous convolutions with different dilation rates to produce convolutional features at multiple scales along with the image-level features. The attained multi-scale semantic features are upsampled by a factor of 4 with bilinear interpolation and then concatenated with the corresponding spatial features to produce a multi-scale context feature.

Training methodology: We revisit a multi-scale context feature encoding method by incorporating atrous convolutions. $Conv_5$ feature map of pre-trained ResNet is processed through multiple atrous rates like $2, 4, 6$ on 3×3 receptive field along with 1×1 and average pool filters. The concatenated feature map produces a pyramid of semantic features. The number of channels of obtained feature map reduced to 512 using 1×1 convolutions. The encoded features are bilinearly interpolated to increase spatial resolution on the feature pyramid. Then, the later connection is established using a convolved $Conv_3$ feature map. After the concatenation, we apply a 3×3 depthwise separable convolutions with 512 channels to refine the concatenated features. The concatenated feature map $32 \times 32 \times 512$ is broadcasted through the attention layer before feeding it to the LSTM network.

4.4 Recurrent pooling network

Even though the individual encoding mechanisms encode effective information of an image, the multiple encoders characterize the complementary behaviours of various encoding schemes [26,18]. Motivated by the above observation, we leverage a recurrent pooling network in order to combine complementary information from all our encoders. We employ multiple encoders i.e., BuTd, WMSC, AMSC, and Resnet-$101_{baseline}$ with multiple RNNs to generate diverse and precise captions of an image. The overview of the recurrent pooling network is illustrated in Figure 6. The proposed framework has two phases, where we employ multiple LSTMs

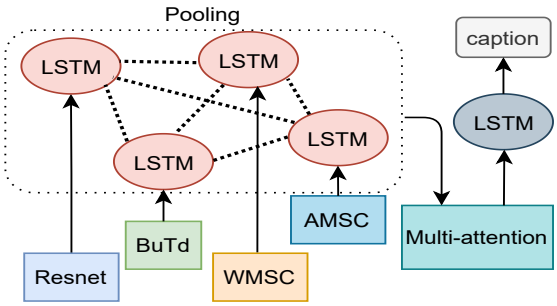

Fig. 6 Recurrent pooling network

with attention mechanism on multiple encoders to generate caption. At first, each individual LSTM network computes hidden states on each encoder component. Then, we pool all hidden states and share among the all components to learn the

interactions of one component with other set of components. The hidden state of q^{th} component is computed at time step t as

$$[h_t^q, c_t^q] = LSTM_t^q \left(H_t, Att_t^q(E^q, h_{t-1}^q) \right) \tag{4}$$

where H_t is the pool of hidden states, Att_t^q is the attention module, E^q is the encoding component, and h_t & c_t are the hidden and context vectors of LSTM network.

On achieving hidden states of each components, we employ multi-attention and compress the outputs of first phase into one hidden vector. This mechanism exploits the interactions among the hidden states and captures the complementary information. The basic intuition behind the proposed recurrent pooling network is to combine the complementary information from the multiple encoders and produce comprehensive and effective hidden states than conventional ones.

5 Experimental Results

In this section, we demonstrate the efficacy of the proposed models with quantitative and qualitative analysis on COCO dataset. First, we present the quantitative analysis on proposed models using standard evaluation metrics like BLEU-n, METEOR, ROUGE L, and CIDEr. Then, the couple of sample images are examined using attention maps and generated captions in qualitative analysis.

5.1 Dataset

We use challenging COCO captioning dataset [38] to conduct the experiments and validate the proposed models. The COCO dataset is the largest image captioning dataset, which contains $82,783$ images in training set and $40,504$ & $40,775$ images in both validation & test set. Further, each image in the dataset is provided with 5 human annotated captions. Since the ground truth captions of COCO test is not available, we adopt a widely used COCO data split [28]. This new split [28] divides the original validation set into new validation and test subsets for experiments & model selection. The new data split [28] contains $113,287$ images for training and 5000 images for validation and test each.

We evaluate generated captions using standard evaluation metrics like BLEU-n [44] (B-1, B-4), METEOR (MR) [4], ROUGE_L (RL) [36], CIDEr (Cr) [54], and SPICE [1] . In a nutshell, all these metrics evaluate the coherence between the n-gram occurrence in reference caption and generated caption. The coherence is weighted by the diversity and saliency of the ground-truth caption.

5.2 Implementation details

We mainly implement proposed captioning models on Pytorch [45] framework. The multi-scale context feature embedding, attention layer embedding, and LSTM hidden units are fixed to 512 for all proposed models. ADAM optimizer is used with a learning rate of 0.0001 and 0.0004 for encoder and decoder, respectively. All models are specified with a batch size of 32 and learned until the accuracy

on the validation set does not change for 20 epochs. The decay rate is established when the model does not improve for eight consecutive epochs.

Encoder: The input image is fed to the pre-trained ResNet [21] to extract global ($Conv_5$) and local ($Conv_3$) convolutional features. The obtained $Conv_5$ features map is further processed to boost the semantic information by constructing a feature pyramid with an atrous convolutional module and wider multi-scale context module. The constructed framework is proposed to learn by increasing the spatial resolution without losing the semantic information of an image. The attention mechanism [61] is utilized to selectively focus on the regions of multi-scaled features. In our settings, we exploit convolution filters of size 1×1, 3×3, & 5×5 and feature maps of size 512 to 2048. The dilation rates which we use in AMSC are $1, 2, 4$, and 6 as shown in AMSC architecture (Figure 4.3).

Decoder: The standard LSTM [22] cells are incorporated to generate captions. However, the input to the LSTMs includes annotation vectors obtained from the attention layer and input word at time instance t. Each annotation vector dynamically learns to focus on region with respect to previously generated words and hidden state of LSTM as described in Section 3.2.

5.3 Quantitative analysis

Recent works on image captioning task present various approaches to effectively describe the content of an image in a natural language sentence. For instance, Xinyu [60] *et al.* presented a hierarchical three-layer LSTM network to fuse visual and textual semantics to generate captions. The cross-modal circular correlation learning (C^3L) approach aims to understand the latent correlation between image and text, further realize cross model generation to produce descriptions. Lian *et al.* [70] leverage visual saliency and semantic saliencey to caption an image. Topic oriented image captioning is proposed in [67], where topic, caption, and image are preserved in the embedding space in a hierarchical structure. Lingxiang *et al.* [57] proposed a Recall network, which selectively recall image contents while generating each word by incorporating the semantic information using a GridLSTM.

The other set of works leverage attention mechanism to the existing encoder-decoder framework. Linghui *et al.* [33] introduced global and local attention to associate object-level features with image-level feature. Chen *et al.* [7] proposed a spatial and channel-wise attention that dynamically modulates the sentence generation context in multi-layer feature maps. A compact attention module is presented in [52]. Gan *et al.* [20] incorporated tag dependent weight matrix to the LSTM network. In contrast to the existing works, we propose a novel encoding approach that leverages multi-scale context information to produce a feature map at multiple field-of-view. The proposed encoding mechanism enriches the semantic and spatial features of an image. The fused representation advocates global, local, and region level features of an image. Although the proposed model is evaluated on image captioning task, it can be extended to many vision to language tasks, like visual question answering, visual commonsense reasoning, and image retrieval. Moreover, the data cleaning operations proposed in [6,5] provide better functional dependencies (fds) and relaxed functional dependencies (rfds) to boost the performance of the model.

Table 1 Performance of the proposed models and other state-of-the-art methods on the COCO dataset, where B-n, MR, RL, and Cr are short for BLEU-n, METEOR, ROUGE_L, CIDEr-D scores, and SPICE respectively.

Model	B-1	B-4	MR	RL	Cr	S
S-ATT [61]	70.7	24.3	23.9	-	-	-
SCA-CNN [7]	71.9	31.1	25.0	-	-	-
SCN [20]	72.8	33.0	25.7	-	101.2	-
C^3L [47]	68.5	28.3	23.3	-	84.3	-
VIS-SAS [70]	72.49	28.08	23.66	55.43	82.12	-
GLA [33]	72.5	31.2	24.9	53.3	96.4	-
COMIC [52]	72.9	32.8	-	-	100.1	-
DHEDN3 [60]	73.1	32.3	25.6	53.7	99.3	-
T-oe [67]	73.9	32.6	26.1	54.4	103.8	-
CTI [49]	74.3	33.9	26.2	54.8	103.6	-
EE-LSTM-P [69]	75.7	34.6	26.8	56	109.6	-
Recall [57]	75.8	33.06	24.6	-	103.7	-
AoANet [25]	80.2	38.9	29.2	58.8	129.8	22.4
M^2T [14]	80.8	39.1	29.2	58.6	**131.2**	22.6
Proposed-BuTd	76.3	33.5	26.4	55.4	114.6	19.7
Proposed-WMSC	78.9	36.3	27.9	56.9	120.2	21.2
Proposed-AMSC	79.2	37.9	28.4	57.3	121.4	21.9
Proposed-RPN	**81.2**	**39.4**	**29.8**	**58.8**	130.4	**22.7**

Our goal is to produce effective encoding representation for a given image without explicitly depending on external knowledge like semantic attributes [20] and region proposals generated by object detection framework [33]. We introduce multi-scale context features rather than semantic features obtained from the final convolution layer to represent an image. In addition, we investigate attained multi-scale context features with a simple attention mechanism proposed in [61].

Table 1 presents the performance comparison with the state-of-the-art models on the COCO dataset. From the table, we can infer that the proposed recurrent pooling network and atrous multi-scale context feature encoding mechanisms are outperforming recent state-of-the-art models. And, the BuTd and WMSC approaches are showing comparable performance with the conventional methods in all metrics. In particular, our proposed RPN approach surpasses the recent attention based encoder-decoder approach [52] and hierarchical encoder-decoder model [60]. In addition, our approach learns to align object level & semantic level features while generating natural sentences and shows superior performance over the previous state of the approaches which exploits semantic alignment embedding [69] and self-attention mechanisms [25, 14].

Even though the proposed models outperform recent works [67, 57, 60], there are still many scopes involved in the proposed approach. Primarily, we utilized conventional attention mechanism to selectively focus on prominent regions of an image. However, other hard-wired attention mechanisms [3, 40, 15] can be adopted to the proposed model. In addition, we can leverage semantic attributes, relations, visual reasoning to guide the captioning module. As in [60, 2], we can investigate a stack of LSTM networks with visual attention to effectively utilize visual features.

In addition, we present the performance of different components of recurrent pooling network in Table 2. From the table, we can observe that the proposed encoding mechanisms are constantly showing comparable performance on COCO dataset. Specifically, $\mathbf{RPN_{AMSC+WMSC}}$ is showing comparable perfor-

mance with $\mathbf{RPN_{All}}$, i.e. $\mathbf{RPN_{Resnet+BuTd+AMSC+WMSC}}$. From the above observation, we can infer that the joint AMSC and WMSC model can inherently learn deep multi-level features and complement the ResNet and BuTd models. Also, we could observe that $\mathbf{RPN_{AMSC}}$ is showing favourable performance when compared with $\mathbf{RPN_{Resnet+BuTd}}$. Although different combinations of RPNs are showing consistent performance on COCO dataset, the unified model, i.e., $\mathbf{RPN_{All}}$ is surpassing previous state-of-the-art models and reporting its superiority on COCO dataset.

Table 2 Performance of the different components on the COCO dataset

Model	B-1	B-4	MR	RL	Cr	S
$\mathbf{RPN_{Resnet}}$	75.6	33.2	25.7	53.5	111.8	19.1
$\mathbf{RPN_{BuTd}}$	76.3	33.5	26.4	55.4	114.6	19.7
$\mathbf{RPN_{WMSC}}$	78.9	36.3	27.9	56.9	120.2	21.2
$\mathbf{RPN_{AMSC}}$	79.2	37.9	28.4	57.3	121.4	21.9
$\mathbf{RPN_{Resnet+BuTd}}$	79.5	38.2	28.7	57.9	123.2	21.8
$\mathbf{RPN_{Resnet+WMSC}}$	80.2	38.3	28.9	58.1	123.4	21.9
$\mathbf{RPN_{Resnet+AMSC}}$	80.5	38.6	28.9	58.1	123.7	21.9
$\mathbf{RPN_{BuTd+WMSC}}$	80.9	38.7	29.0	58.2	124.6	22.2
$\mathbf{RPN_{BuTd+AMSC}}$	81.2	39.3	29.1	**58.8**	126.2	22.3
$\mathbf{RPN_{AMSC+WMSC}}$	**81.7**	39.1	29.3	58.5	129.7	22.5
RPN_{All}	81.2	**39.4**	**29.8**	**58.8**	**130.4**	**22.7**

5.4 Qualitative analysis

This section presents a qualitative analysis of the proposed encoding mechanisms. The attention maps and generated captions of a given input image are depicted in Figure 7. From Figure, we can examine that the captions generated by the proposed approaches are diverse and precise. In particular, the generated word *"grassy field"* validates that the proposed model retains spatial resolutions at the encoding phase. The generated word *"lush green"* and *"grass covered"* indicates that the model is accounting finer details. Besides, the words like *"couple"*, *" standing"* and *"top"* determines the semantic relationships in the generated caption. Moreover, the captions generated by BuTd, WMSC, AMSC, RPN show a successive refinement over one another. For instance, the generated words *"lush green field"* and *"standing on top of"* by RPN model are more fine-grained and divers over the generated words of BuTd, i.e., *"grassy field" "standing in"*.

Also, we can observe that the bottom-up and top-down mechanism (BuTd) is producing blobs on coarse-grain objects. And, the multi-scale context information incorporated in WMSC and AMSC approaches is learning to focus on fine-grain details. Whereas, the multi-level features incorporated using recurrent pooling network (RPN) learn to focus on both coarse-grain and fine-grain details. This implication of the proposed frameworks is inferred from the focused regions of attention maps, where large blobs indicate rich semantic information, and small blobs demonstrate multi-scale context information.

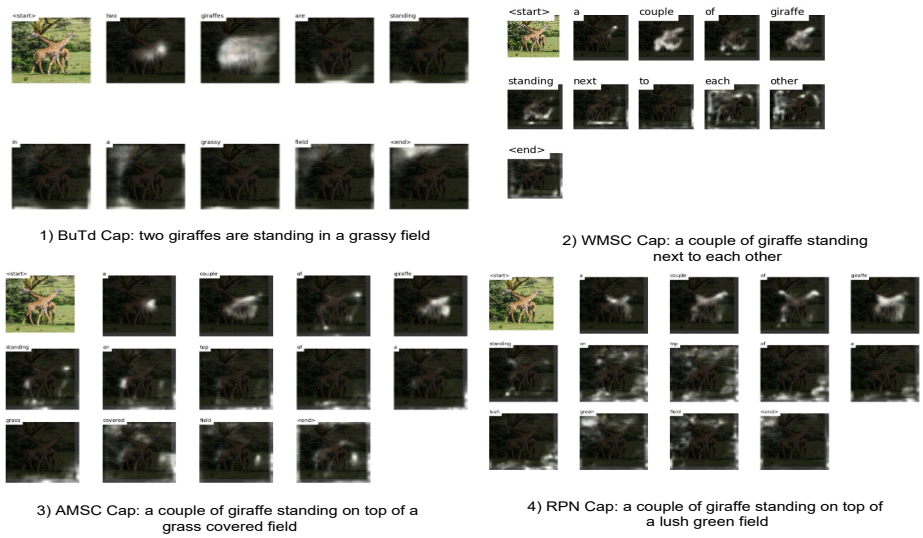

Fig. 7 Visualization of attention maps along with generated captions on the COCO dataset. The output captions are generated by 1) Bottom-up and top-down encoding mechanism, 2) Wider multi-scale context feature encoding mechanism, 3) Atrous multi-scale context feature encoding mechanism, 4) Recurrent pooling network

6 Conclusion

The state-of-the-art approaches highly depend on either object detection frameworks [33,63] or explicit knowledge like semantic tags [20] to generate human-like caption. On the other hand, simple CNN plus RNN approaches [56,28] are not showing promising results due to a lack of powerful representation of an image. In this work, our goal is to show the effectiveness of attentive multiscale context information in the context of image captioning task. For each input image to be captioned, we equip both semantic and spatial representations of an image by exploiting multi-scale context information. We investigate three encoding mechanisms to produce effective context information of an image to generate captions. All three encoding mechanisms utilize pre-trained ResNet for both semantic and spatial convolutional features. The bottom-up and top-down encoding mechanism is proposed to reconstruct the spatial resolution by utilizing the final convolutional layer of the CNN framework. Here, attention mechanism selects regions of reconstructed feature maps to generate captions. The wider multi-scale context feature encoding technique employs various receptive fields on the semantic feature map then concatenated with the spatial features obtained from the early layers of convolutions. The atrous multi-scale context feature encoding mechanism utilizes atrous/ dilated convolutions with varied filter sizes to provide multi-scale field-of-view. The obtained feature pyramid is further concatenated with lateral connections of deep CNN. Finally, we pool all encoder components with recurrent pooling network to learn complimentary information of all our encoders. Further, the proposed approaches are built on a simple attention mechanism rather than a hard-wired mechanism. The proposed encoder multi-scale context module has wide-ranging uses, and in this work, we exploit for image captioning task. In which,

the models are proposed to learn global and local context at various filed-of-view. The effectiveness of the proposed encoder mechanisms is demonstrated by comparing it with the recent works. Besides, the attention maps and generated captions signify that the proposed models learnt to focus on fine-grain details to generate captions.

Conflict of Interest Statement Jeripthula Prudiraj, Yenduri Sravani, and C. Krishna Mohan declare that they have no conflict of interest.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European conference on computer vision, pp. 382–398. Springer (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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 (2018)
3. Arik, S.O., Pfister, T.: Tabnet: Attentive interpretable tabular learning. arXiv preprint arXiv:1908.07442 (2019)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72 (2005)
5. Breve, B., Caruccio, L., Cirillo, S., Deufemia, V., Polese, G.: Dependency visualization in data stream profiling. Big Data Research **25**, 100240 (2021)
6. Caruccio, L., Cirillo, S.: Incremental discovery of imprecise functional dependencies. Journal of Data and Information Quality (JDIQ) **12**(4), 1–25 (2020)
7. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: 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 (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818 (2018)
10. Chen, S., Zhao, Q.: Boosted attention: Leveraging human attention for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 68–84 (2018)
11. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2422–2431 (2015)
12. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
13. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience **3**(3), 201–215 (2002)
14. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10578–10587 (2020)
15. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
16. Elliott, D., Keller, F.: Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1292–1302 (2013)
17. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision, pp. 15–29. Springer (2010)

18. Firat, O., Cho, K., Bengio, Y.: Multi-way, multilingual neural machine translation with a shared attention mechanism. arXiv preprint arXiv:1601.01073 (2016)
19. Fu, K., Jin, J., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2321–2334 (2016)
20. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5630–5639 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
23. Hossain, M., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR) **51**(6), 118 (2019)
24. Hsieh, H.Y., Huang, S.A., Leu, J.S.: Implementing a real-time image captioning service for scene identification using embedded system. Multimedia Tools and Applications **80**(8), 12525–12537 (2021)
25. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)
26. Jiang, W., Ma, L., Jiang, Y.G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 499–515 (2018)
27. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1700–1709 (2013)
28. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128–3137 (2015)
29. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems, pp. 1889–1897 (2014)
30. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: International conference on machine learning, pp. 595–603 (2014)
31. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2891–2903 (2013)
32. Li, L., Tang, S., Deng, L., Zhang, Y., Tian, Q.: Image caption with global-local attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
33. Li, L., Tang, S., Zhang, Y., Deng, L., Tian, Q.: Gla: Global–local attention for image description. IEEE Transactions on Multimedia **20**(3), 726–737 (2017)
34. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220–228. Association for Computational Linguistics (2011)
35. Li, Z., Li, Y., Lu, H.: Improve image captioning by self-attention. In: International Conference on Neural Information Processing, pp. 91–98. Springer (2019)
36. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74–81 (2004)
37. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125 (2017)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740–755. Springer (2014)
39. Lu, J., Xiong, C., Parikh, D., Socher, R.: 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 (2017)
40. Ma, X., Pino, J., Cross, J., Puzon, L., Gu, J.: Monotonic multihead attention. arXiv preprint arXiv:1909.12406 (2019)

41. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
42. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daumé III, H.: Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 747–756. Association for Computational Linguistics (2012)
43. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Advances in neural information processing systems, pp. 1143–1151 (2011)
44. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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 (2002)
45. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
46. Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. In: Proceedings of the IEEE international conference on computer vision, pp. 1242–1250 (2017)
47. Peng, Y., Qi, J.: Show and tell in the loop: Cross-modal circular correlation learning. IEEE Transactions on Multimedia **21**(6), 1538–1550 (2018)
48. Rensink, R.A.: The dynamic representation of scenes. Visual cognition **7**(1-3), 17–42 (2000)
49. Su, J., Tang, J., Lu, Z., Han, X., Zhang, H.: A neural image captioning model with caption-to-images semantic constructor. Neurocomputing **367**, 144–151 (2019)
50. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27**, 3104–3112 (2014)
51. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
52. Tan, J.H., Chan, C.S., Chuah, J.H.: Comic: Toward a compact image captioning model with attention. IEEE Transactions on Multimedia **21**(10), 2686–2696 (2019)
53. Tian, P., Mo, H., Jiang, L.: Image caption generation using multi-level semantic context information. Symmetry **13**(7), 1184 (2021)
54. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566–4575 (2015)
55. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5753–5761 (2017)
56. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164 (2015)
57. Wu, L., Xu, M., Wang, J., Perry, S.: Recall what you see continually using gridlstm in image captioning. IEEE Transactions on Multimedia **22**(3), 808–818 (2019)
58. Wu, Q., Shen, C., Liu, L., Dick, A., Van Den Hengel, A.: 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 (2016)
59. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1367–1381 (2017)
60. Xiao, X., Wang, L., Ding, K., Xiang, S., Pan, C.: Deep hierarchical encoder–decoder network for image captioning. IEEE Transactions on Multimedia **21**(11), 2942–2956 (2019)
61. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048–2057 (2015)
62. Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 444–454. Association for Computational Linguistics (2011)
63. Yang, Z., Zhang, Y.J., ur Rehman, S., Huang, Y.: Image captioning with object detection and localization. In: International Conference on Image and Graphics, pp. 109–118. Springer (2017)

64. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4894–4902 (2017)
65. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651–4659 (2016)
66. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
67. Yu, N., Hu, X., Song, B., Yang, J., Zhang, J.: Topic-oriented image captioning based on order-embedding. IEEE Transactions on Image Processing **28**(6), 2743–2754 (2018)
68. Zhang, S., Zhang, Y., Chen, Z., Li, Z.: Vsam-based visual keyword generation for image caption. IEEE Access **9**, 27638–27649 (2021)
69. Zhang, X., He, S., Song, X., Lau, R.W., Jiao, J., Ye, Q.: Image captioning via semantic element embedding. Neurocomputing **395**, 212–221 (2020)
70. Zhou, L., Zhang, Y., Jiang, Y.G., Zhang, T., Fan, W.: Re-caption: Saliency-enhanced image captioning through two-phase learning. IEEE Transactions on Image Processing **29**, 694–709 (2019)