

RESEARCH ARTICLE

Context-Adaptive-Based Image Captioning by Bi-CARU

SIO-KEI IM^{1,2}, (Member, IEEE), AND KA-HOU CHAN^{1,2,3}, (Member, IEEE)¹Faculty of Applied Sciences, Macao Polytechnic University, Macau, China²Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic University, Macau, China³Faculty of Science, The Hong Kong Polytechnic University, Hong Kong, China

Corresponding author: Ka-Hou Chan (chankahou@mpu.edu.mo)

This work was supported by the Macao Polytechnic University Research Project under Grant RP/FCA-06/2023.

ABSTRACT Image captions are abstract expressions of content representations using text sentences, helping readers to better understand and analyse information between different media. With the advantage of encoder-decoder neural networks, captions can provide a rational structure for tasks such as image coding and caption prediction. This work introduces a Convolutional Neural Network (CNN) to Bidirectional Content-Adaptive Recurrent Unit (Bi-CARU) (CNN-to-Bi-CARU) model that performs bidirectional structure to consider contextual features and captures major feature from image. The encoded feature coded from image is respectively passed into the forward and backward layer of CARU to refine the word prediction, providing contextual text output for captioning. An attention layer is also introduced to collect the feature produced by the context-adaptive gate in CARU, aiming to compute the weighting information for relationship extraction and determination. In experiments, the proposed CNN-to-Bi-CARU model outperforms other advanced models in the field, achieving better extraction of contextual information and detailed representation of image captions. The model obtains a score of 41.28 on BLEU@4, 31.23 on METEOR, 61.07 on ROUGE-L, and 133.20 on CIDEr-D, making it competitive in the image captioning of MSCOCO dataset.

INDEX TERMS CNN, RNN, NLP, image captioning, Bi-CARU, context-adaptive, attention mechanism.

I. INTRODUCTION

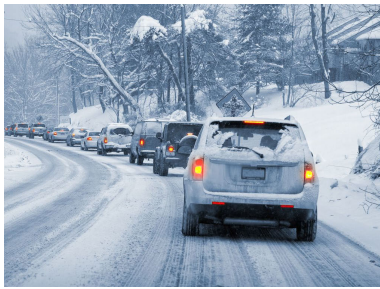


Image captioning is an assistive process that helps people understand media information and highlights the most important features that the sender wants to present in an image [1]. Generally, a complete image captioning task consists of two parts: computer vision and Natural Language Processing (NLP) [2]. The computer vision performs image encoding, which investigates the information within an image to determine the objects in a frame and their mutual correspondence and relationship [3]. The encoded feature is then passed to an NLP model to decode the information into a text-based sentence [4]. The purpose of image captioning conducts to generate natural language captions for input images that accurately describe these elements [5]. The model performs dynamic multimodal analysis and inference on the visual

content and generated words during the caption word generation process [6]. A major challenge is handling the two different media on the encode and decode side. This challenge is addressed by the encoder-decoder approach, which primarily examines the image's global region while generating the image captioning [7]. Additionally, the attention mechanism refines the determination of interested objects by normalising the extracted visual features into a set of attention weight or trainable parameters for neural training [8]. In recent years, visual content and semantic attention have been shown to be superior in such domain, improving the model's interpretability [9]. Table 1 illustrated that these images may contain multiple objects depending on the captioning task. In order to connect their relationship, this work proposes a method to generate captions according to the attention mechanism, aiming to produce a refined caption from a single image.

In current years, deep learning first made major breakthroughs in image captioning. Researchers quickly applied

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

TABLE 1. Image captioning under different targets.

Target			
Objects	cars	a train and cars	a cruise ship
Scene	on the road	tracks and highway	Ocean and Coast
Weather	it is snowing a lot	twilight is descending	it is a bright day
Overall	cars drive down the road on a snowy day	a train is running on the tracks next to the highway	a cruise ships moored at the shore

the encoder-decoder framework to the video captioning domain, with various results [10]. Image captioning tasks aim to generate a sentence that summarises the main content of an image in a natural way. Since most images contain multiple objects with related background information or overlapping action-oriented activities, expressing all the complex content of an image in a sentence is challenge. To address this, image captioning tasks have used neural network technology that integrates Natural Language Processing (NLP) to describe all the important events and details, resulting in more natural and accurate descriptions for Artificial Intelligence (AI). However, using a linear Recurrent Neural Network (RNN) that only performs feedforward data representation at the decoding side leads to a limited understanding of the impact of sequence orientation on prediction [11]. This is especially true when dealing with different relationships between multiple objects in a scene, which can lead to contextual inadequacies in sentence prediction and potential misinterpretation of contextual information [12], [13], [14]. In this work, an advanced RNN layer-based Bidirectional Content-Adaptive Recurrent Unit (Bi-CARU) structure is introduced to alleviate this problem [15], [16]. It consists of two CARU neural layers that perform forward and backward coding together to achieve a context-adaptive approach and more accurate predictions, as well as a deeper understanding of sequence orientation. By incorporating contextual information, this structure provides a more suitable understanding of the context for sentence prediction. The forward network processes data in a traditional and unidirectional pattern, while the backward network processes data in a reverse manner, taking into account the natural language forms of reading comprehension simultaneously [17]. In addition, the proposed Bi-CARU structure also employs an adaptive layer to collect the RNN features, which are then combined by summing the forward and backward outputs. This allows a more comprehensive content extraction from both the forward and backward CARU networks.

Furthermore, CARU can outperform other RNN units in NLP tasks due to the context-adaptive gate used as a decoder [18], [19]. By applying forward and backward RNN approaches to a sentence, the captions generated by CARU can achieve a complete representation. In practice, visual content and semantic attention should coincide at the currently encoded feature(s) [20], [21]. In fact, it is found that the current feature in the backward CARU layer does not provide effective context information, while the same feature can be produced in the forward order at the same time [22], [23]. In turn, the forward method cannot provide adequate context information if the same features are simultaneously generated in the backward order. To address the problem of asynchrony between forward and backward directions while making the most of contextual information, this work proposed an attention mechanism that aims to collect the hidden feature produced by the context-adaptive gate in the middle of the CARU layer, and also connects to the forward and backward CARU layers to refine the prediction of semantic information [24]. This approach unifies semantic information to produce complementary outputs. It overcomes the limitation that forward and backward simultaneous semantics cannot be generated due to incompatibility, resulting in more accurate sentence prediction [25], [26].

Inspired by the above studies, we introduce a CNN-to-Bi-CARU model to achieve the encoder-decoder approach, and our main contributions can be summarised as follows:

- The Convolutional Neural Network (CNN) first extracts the main regions and determines the objects of interest according to the attention mechanism. We propose a weighting processing to collect these hidden features generated by each CNN layer, with the aim of refining the detection results and discarding noise and uninteresting objects.
- We employ Bi-CARU as a decoder for feature extraction in both directions. With the advantage of CARU, its context-adaptive gate is able to produce the

context information, which can be further applied to the attention approach to discover the relation between the current state and the entire sequence.

- To refine the accuracy of the similarity module in the output, it is important to ensure that the forward and backward hidden states can be adjusted/tuned. We achieve this by combining the features extracted from forward and backward attention, and introduce an additional procedure to refine the prediction by considering the hidden states and attention. This results in complementary and fine-grained sentences.

The organisation of this article is briefly described below: Section II provides an overview of related work on image annotation approaches and their integration with neural network technology. Section III describes the architecture of the proposed CNN-to-Bi-CARU model, including its core technical work in detail. Section IV presents an introduction to the dataset used, the implementation and configuration of the experiment, and comparisons with other methods. Finally, the article is concluded and future work is discussed in Section V.

II. RELATED WORK

Image captioning produces simple text sentences that describe the behaviour of the interested objects in a captured shot. A typical image is usually formed by combining various elements (including objects, events, backgrounds, and other noise) within a scene. As a result, early approaches to image captioning were more or less prompted by pre-processing tasks and had to provide rough input before deciding on the generated caption [27], [28]. Thanks to powerful neural network technology, especially the encoder-decoder model structure based on deep learning [7], [29], [30] has been introduced. The encoding side uses the Very Deep Convolutional Networks (VGG) network to process the features and discover the characteristics of the media data [31], and also uses the optical flow layer as a filter to produce hidden states [32]. The filtered states are then passed to the decoding side, and a textual description is generated by a RNN network, such as Long Short-Term Memory (LSTM) [33], Gated recurrent unit (GRU) [34], and CARU [35]. However, due to the context of behaviour and relationship in one shot image, these methods do not perform well in reflecting multiple object information in the image, and the produced text caption is not detailed enough for description. Therefore, many works use advanced CNN module as an encoder for the feature extraction of image, which employ the ConvNet adds optical flow features on the basis of multi-label and multi-attribute, and then coded from the ConvNet into the LSTM and represent the sequence of words [36], [37]. Moreover, dense relational image captioning has provided a new perspective of the image captioning task [38]. Based on the image-to-text approach, the original descriptive text is encoded and extended as one of the input features based on the regions of the object of interest for text caption prediction, which refines the comprehensiveness and diversity of the descriptions and improves the accuracy. In addition, the well-trained CNN parameters

are further applied to COCO caption dataset is used as initial to the ConvNet used, which further improves the convergence and performance of image feature extraction [39]. For such tasks, a caption dataset of ImageNet-Captions is also proposed, which provides a high-level view of the field of image captions [40]. Next, [41] proposed the concept of multilingual captioning according to various target, combining multiple description text into one sentence to achieve a competitive text caption, and its decoding part is divided into two modules: sentence collection and caption composition. In these studies, the description texts can be considered as references in the dataset and further used as inputs for training the neural network, which contributed and inspired the study of image captioning tasks to learn the association information between sentences [42].

In NLP tasks, RNNs outperform general image captioning, and most models use the ability of LSTMs to remember long sequences as decoders to produce reasonably descriptive text. In [43], an extension of the LSTM approach was proposed, where the cross-entropy loss during the recurrent step was investigated, and a correlation loss was introduced to allow the model to learn both semantic relationships and visual content, providing fully associated sentences used as references with visual features [44], [45], [46]. In turn, an advanced boundary-aware encoding model also uses the RNN as the encoding part and proposes recurrent image schemes [47], which provide a new way to explore and investigate the hierarchical structure in the media data and improve the relationship matching between multiple objects in a scene. Recently, many researchers tend to apply the attention mechanism to the field of image captioning and achieve good performance in NLP. In [48], an attention-oriented transformer was introduced for image captioning, which aims to give more attention weight to refine the attention weight distribution. Also, [49] makes use of an attention weight α to calculate the major features, obtaining high attention to important information in the image. These attention approaches are able to discard most noise of unimportant information [50]. In practice, since RNN structures are challenging to train in parallel, the transformer framework introduces a global connection based on an attention mechanism [51] and discovers such relationships based on the semantic information of the reference samples describing the sentences in the dataset [52]. Also, [29] extended the transformer model with one encoder corresponding to two decoders, allowing the image to be encoded according to different tasks that can be decoded into multiple descriptions separately. Similarly, [8] and [53] proposed the multiple decoding method to enhance the convergence and speed up the training process in parallel, thus the semantic relationship in reference sentences through multiple transformer ways in parallel [54]. These works reflect that the transformer framework provides potential performance and supports a dynamic relationship module to interpret the global features of the image, taking into account the accuracy and diversity of text descriptions.

According to the above-mentioned study, the use of the transformer framework is suitable for feature extraction of region relations. It is also effective in discovering connections with implicitly related semantics. This method produces more accurate descriptions, and the dynamic search of related visuals between multiple regions also performs well. Although experts have made improvements and extensions to address the problems of insufficient use of image features and insufficient correlation between media and text relationship, the complexity of bidirectional structure makes it more comprehensive and accurate text expressions for image abstraction. In this work, we use the transformer idea to refine the proposed Bi-CARU by using an attention mechanism to collect the hidden state produced by the context-adaptive gate. This design can effectively extract context information from decoding and provide more accurate prediction results. Therefore, making better use of BI-CARU features in the image remains a research challenge in the field of image captioning.

III. PROPOSED METHODOLOGY

This work introduces a Bidirectional CARU (Bi-CARU) as a decoder in image captioning tasks. Considering the connections between interesting objects in an image, this model also applies the attention mechanism to determine their relationship and features for output encoding. This efficiently extracts contextual information while aligning \vec{h}_t with \overleftarrow{h}_t via context-adaptive attention. Such two CARU layers can produce the feature of semantics and thus the complementary output more accurately. Fig. 1 illustrates the proposed model in detail. The context-adaptive is able to discover the part-of-speech that helps to align the hidden state extracted by the forward and backward CARU layers, respectively.

A. CNN ENCODER

For the encoding side, we recommend to use the CNN encoder attention approach for image encoding. This technique is popular because it is able to capture the most important local features of the input image and encode the media data into a fixed size vector. With respect to Fig. 1, the input image is first convolved into extracted features v_t , each of which represents the local pattern in an object. These features are then passed through an attention layer to determine the weighted features $weight_t$ of interest and discover the most important ones that represent the local features r_t for interested objects. The attention mechanism is then applied to these features to produce a fixed size vector containing the most important local features. This helps in the prediction work by assigning a weight to each feature map, with the training parameters weighting the more important features for the relationship connection. Similar to the attention mechanism, the weighted feature maps are also summed to produce the fixed-size vector for producing the hidden state.

To encode an image, its visual information must be enhanced and noise must be discarded. In our proposed model, we incorporate a soft attention layer to achieve this

task. The convolved image features v_t extracted from each CNN layer can be expressed as follows:

$$weight_y = \mathbf{W}_t \tanh \left(\sum_i^N \mathbf{W}_i v_i \right) \quad (1)$$

Here, the first \mathbf{W}_t denotes the training parameter aimed at enhancing the $weight_t$ for attention. The $\mathbf{W}_{i=1,2,\dots,N}$ represents the CNN internal training parameter, where N is the number of convolutional layers used to implement the CNN framework. In fact, we collect all the features produced by each convolutional layer, because the early layer may contain many global shapes features but poor for local information. In turn, the later layer can represent the local pattern with global features. Besides, the features produced by different layers can be used for various tasks. For instance, the global feature is suitable for weather detection or scenes addressing, while the local feature is good for object detection and pattern recognition, etc. Therefore, this $weight_t$ in (1) needs to dynamically account for attention by applying these convolved features v_i from images in response to changes in the visual context like:

$$a_t = \text{Softmax} (weight_1, weight_2, \dots, weight_t)$$

This Softmax is used to project these $weight_t$ into a probability domain. The visual output r_t is thus connected from the same global image features to changing local features of the image, and the v_t can be directly obtained from the results of the t -th CNN layer. As a result, the weight distribution between image and attentional features can be obtained by follows:

$$r_t = a_t v_t$$

B. BI-CARU DECODER

The linear RNN is known to be effective for NLP decoding tasks, but it is also associated with the long-term dependency problem and poor convergence, often due to gradient vanishing. Many studies have proposed various architectures of RNN units, such as LSTM and GRU, to address these issues. In this work, an advanced RNN unit called CARU is employed to alleviate such problems. Compared to GRU, CARU introduces two gates, the context-adaptive gate and the update gate, which contain fewer parameters to handle the data flow than other well-known RNN units. The advantage of CARU is that the context-adaptive gate is able to produce the weight of the current input x_t similar to the reset gate in GRU, but is based only on the current feature instead of the entire sequence. The product of such a gate and the memory information of the previous hidden state h_{t-1} achieves the purpose of weight combination. In practice, it can be considered as a tagging task that connects the relationship between the weight and the parts-of-speech, which allows filtering the noise and enhancing the major feature in the current input

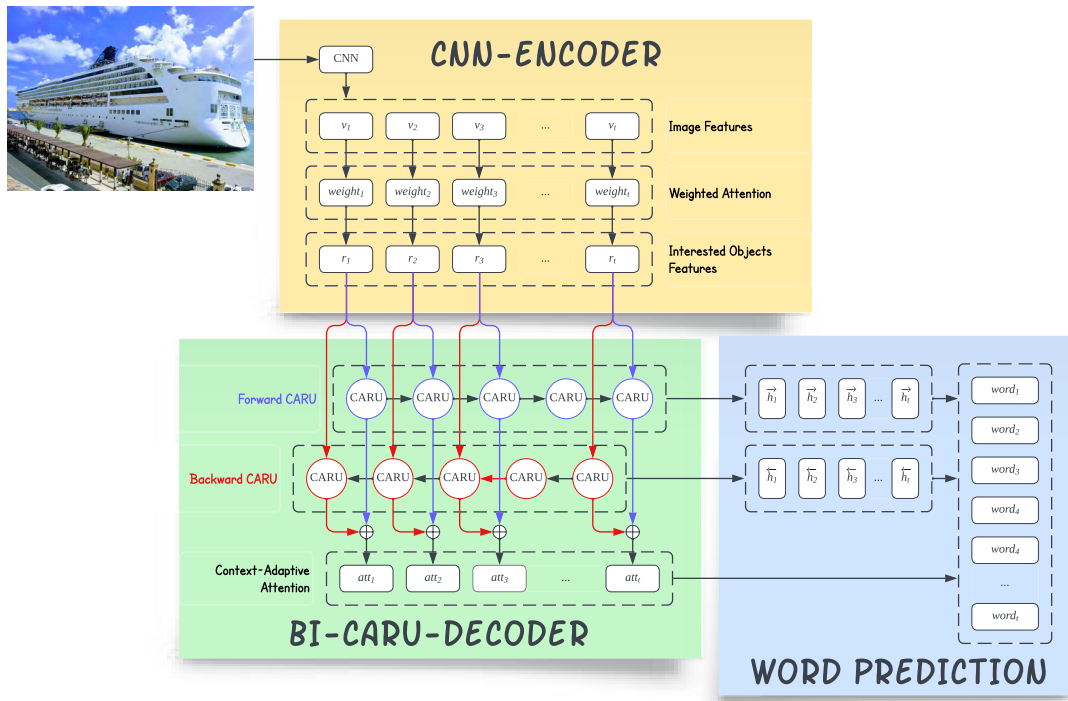


FIGURE 1. The proposed image captioning network model employs CNN as an encoder to extract object features from images, and considers weighted attention to refine the selection of object features. Besides, we make use of an advanced Bi-CARU as a decoder to produce captions. This decoder utilises the forward and backward features \vec{h}_t and \overleftarrow{h}_t , as well as an additional attention layer that collects hidden states from context-adapted gates to determine more accurate predictions.

feature. The complete procedure is as follows:

$$x_t = \mathbf{W}_{vn}v_t + \mathbf{B}_{vn} \tag{2a}$$

$$n_t = \tanh(\mathbf{W}_{hn}h_{t-1} + \mathbf{B}_{hn} + x_t) \tag{2b}$$

$$z_t = \sigma(\mathbf{W}_{hz}h_{t-1} + \mathbf{B}_{hz} + \mathbf{W}_{vz}v_t + \mathbf{B}_{vz}) \tag{2c}$$

$$l_t = \sigma(x_t) \odot z_t \tag{2d}$$

$$h_t = (1 - l_t) \odot h_{t-1} + l_t \odot n_t \tag{2e}$$

It can be found that CARU processes the data flow in a similar way to GRU, but dispatches word weights to the proposed gates and multiplies them by the content weights. In such procedure, it enables the content-adaptive gates to consider both words and content, with each step briefly as follows:

- (2a) A linear layer is first used to apply the current input. This result is used for the next hidden state and is passed to the content adaptive gate. Noted that this result will be assigned to h_t directly if h_{t-1} is not received in the current step.
- (2b) The previous hidden state h_{t-1} is also applied to another linear layer, which is then summed with the result of (2a) and then passed through the tanh activation function in order to extract the integrated information.
- (2c) It performs hidden state transitions like the update gate in GRU, taking into account the current input and combining it with the previous hidden state. This part allows to discover relationships over

content information, but has a long-term dependency problem.

- (2d) To alleviate the long-term dependence problem, this step investigates the features of the current input, which can be considered as a tapping process. It is then multiplied by the z_t conduction to dynamically enhance or dilute the long-term dependence to obtain accurate predictions during the RNN decoding process.

- (2e) The new hidden state output generated by using linear interpolation.

Compared with traditional RNN structures that predict the output of the next state based only on the current information with historical sequences, using a bidirectional structure can improve context awareness by considering the previous and next hidden states. In addition, with the benefit of CARU, it can effectively predict individual words in a sentence while accurately analysing its content based on the current part-of-speech, resulting in more accurate contextual sentence prediction.

C. CONTEXT-ADAPTIVE FOR ATTENTION MECHANISM

Taking advantage of the attention mechanism, it has the ability to discover the (weighted) connection between each state and the entire sequence. This capability can also be applied to the context-adaptive part of CARU. As mentioned in Section III-B, l_t can be seen as a tapping process that is

potentially used in the input of the attention mechanism. Therefore, let \vec{l}_t and \overleftarrow{l}_t denote the state produced by forward CARU and backward CARU, respectively, and the final states passed to the attention mechanism can be obtained by:

$$l_t = \vec{l}_t \oplus \overleftarrow{l}_t$$

Moreover, an attention processing is presented to discover the relationships between the decoded information. We modified the attention structure in the decoder side to model this l_t and calculate its weight of current sequence $I \in \{l_1, l_2, \dots, l_t\}$. The proposed attention formula is expressed as:

$$att_t = L2 \left(\sqrt{\exp \left(\frac{(\mathbf{W}_Q I) (\mathbf{W}_K I)^T}{\sqrt{d}} \right)} \right) (\mathbf{W}_V I) \quad (3)$$

Here \mathbf{W}_Q , \mathbf{W}_K and \mathbf{W}_V are training parameters similar to the attention approach [55], and d denotes the variance of I . It is obvious that we make use of a novel normalisation function $L2(\sqrt{\exp(*)})$ [56], [57], which can enhance convergence during the training process and also maintain the same performance as the Softmax function. Besides, we also design a trainable procedure for adaptively adjust caption information to predict the upcoming decoded states $h_t = \vec{h}_t \oplus \overleftarrow{h}_t$, as follows:

$$H = att_t \times \text{Softplus}(\sigma(\mathbf{W}h_t) \odot \tanh(\mathbf{W}h_t)) \quad (4)$$

Because the hidden state h_t contains contextual information, this formula indicates the trend of the current feature before entering the decoding RNN. In our design, the tanh function determines how the relation to the previous or next state should be considered, and the σ function weights the current state against the entire sequence. The att_t contains both attention weights, allowing the hidden state to be adjusted for accurate word prediction. To the end, we obtain the probability output as follows:

$$\text{Prob}(word_t) = \text{FFN}(H)$$

here we use FFN as a feed-forward network to update the vocabulary area and see that each update can be broken down into sub-updates corresponding to individual FFN parameter vectors. Each of these vectors promotes concepts that are often easy for humans to understand. Note that the model is trained with word-level cross-entropy loss, and thanks to the optimiser of Adam [58].

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section describes the dataset we used and presents the configuration and environment we applied for our development work. It also discusses the experimental results and compares them to previous methods in the field.

A. BENCHMARK DATASETS

To validate our work in the area of image captioning, we used the MSCOCO [59] benchmark dataset,¹ which contains images from websites covering a variety of topics such

as people, animals, vehicles, and focal objects in captured images. Its captions were written by human annotators as a ground truth reference, providing a detailed description of the content of each image: the MSCOCO dataset provides over 330k images for the study of image captions, each with five different captions. This dataset provides researchers with rich descriptions to develop and study their own methods. This dataset has been widely used for the study of image captions with the goal of automatically generating natural language descriptions of media data. Researchers have used this dataset to train and evaluate machine learning models and to develop new algorithms and techniques for image captioning.

For pre-processing, the ‘‘Karpathy’’ segmentation setting is recommended for the MSCOCO datasets to obtain fair comparison results [60]. It selected about 113k training images, 5k validation images, and 5k test images for the MSCOCO. Moreover, we adopt the text pre-processing in [61], the truecase, tokenisation, and cleaning symbols must be completed, and the start mark <BOS> and end mark <EOS> must be inserted at the beginning and end of a sentence, respectively. Since the vocabulary is limited and some word pairs are low frequency or inaccurately described, we discard words with a frequency of less than five and replace them uniformly with a token <UNK>, which can be ignored and is not considered part of the vocabulary. In practice, the vocabulary size can be reduced to about 10k words. To ensure that the decoder receives input from the beginning, the new token <BOS> must be inserted as the first token in a sentence. The sentence is generated exactly as written until the unique end token <EOS> is encountered.

B. TRAINING STRATEGY AND IMPLEMENTATION

In terms of hardware conditions, there is a deep learning workstation that provides four NVIDIA Quadro RTX A4000 with 16.0 GB of memory per GPU, for a total of 64.0 GB of device memory. For the training environment, these experiments were set up on the Ubuntu 22.04 operating system, and the proposed model is developed on the neural network engine of PyTorch [62]. More specifically, we apply the advanced training strategy where a scheduler is used to adjust the learning rate during the training process, aiming to reduce the learning rate adaptively. Besides, the learning rate is initially set to $1e-3$, and a warm-up update function is required to alleviate divergence issue in the early stages of training. Moreover, half-precision floating-point and Distributed Data Parallel (DDP) are enabled to reduce memory consumption and accelerate computation, respectively.

For pre-processing, we only applied data enhancement to improve performance by randomly cropping 90.0% of the original images and erasing 50.0% of the original images during online test server submission. For training coding, we used Faster R-CNN [69] with VGGreNet [31] on the ImageNet dataset to pre-train image classification, and further refined it on the selected dataset used in this work. The VGGreNet arranges the 4,096 feature sizes fully associated

¹<https://cocodataset.org/#download>

TABLE 2. The performance of our model and other advance methods for various metrics of MSCOCO. All values and their error ranges are also reported.

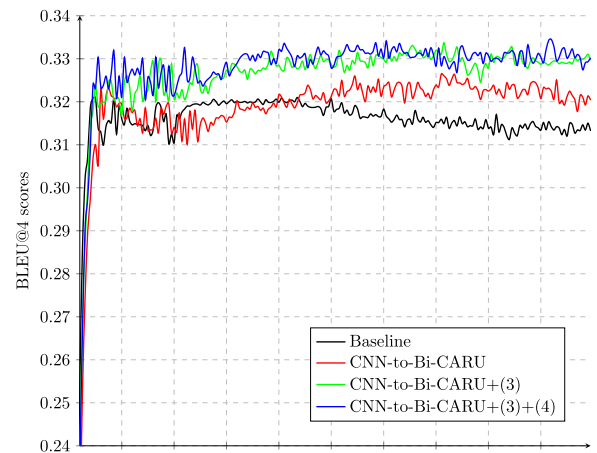
Model		BLEU@4	METEOR	ROUGE-L	CIDEr-D
Baseline	RDN [63]	37.32 ± 0.17	28.16 ± 0.42	57.48 ± 0.63	125.2 ± 1.76
	Show-Attend-Tell [64]	29.68 ± 0.24	25.26 ± 0.37	52.65 ± 0.31	94.08 ± 3.06
Advanced	pLSTM [65]	34.65 ± 0.15	27.63 ± 0.29	58.29 ± 0.61	113.56 ± 1.91
	CNN-BiLSTM-s [66]	37.51 ± 0.24	28.91 ± 0.42	58.19 ± 0.34	121.37 ± 0.96
	M2 [67]	39.74 ± 0.16	29.49 ± 0.29	59.24 ± 0.42	129.35 ± 2.01
	GRIT [68]	42.47 ± 0.25	30.61 ± 0.30	60.76 ± 0.56	144.26 ± 1.24
Proposed	CNN-to-Bi-CARU	39.32 ± 0.12	29.93 ± 0.26	60.16 ± 0.33	128.28 ± 2.21
	CNN-to-Bi-CARU+(3)	40.53 ± 0.10	30.95 ± 0.32	60.51 ± 0.46	131.66 ± 1.18
	CNN-to-Bi-CARU+(3)+(4)	41.28 ± 0.09	31.23 ± 0.23	61.07 ± 0.48	133.20 ± 1.55

with the image features v_t in Fig. 1, while Faster R- CNN implements the extraction of local features, such as face detection. For the remaining image regions, we ranked them by their confidence scores from high to low and applied a threshold of 0.3 to discard the low regions. Each region has a feature size of 2,048, which is the global average pooling result of the encoding side. The results produced by the encoder are processed by Dropout before passing to the decoder. To cover the vocabulary size (about 10k words), we set the word embedding vector and the hidden size in each CARU layer to 1k, and the size of the weighted layer $weight_t$ and the attention layer att_t are configured to 512, respectively. In practice, we tuned the trade-off parameter λ on the “Karpathy” validation split to obtain the best performance by setting it to 0.02. Note that the gradient is mainly contributed by the cross-entropy loss, and the Adam optimiser was used to train up to 100 epochs until there was no improvement for 20 consecutive times.

C. EXPERIMENTAL RESULTS AND DISCUSSION

Compare our CNN-to-Bi-CARU with other state-of-the-art works, two baseline methods RDN [63], Show-Attend-Tell [64] and four advanced methods pLSTM [65], CNN-BiLSTM-s [66], M2 [67], GRIT [68] are compared in our experiment. We also used evaluation metrics commonly used in image captioning to evaluate and investigate the quality of predicted word sequences from automatic machine production. BLEU@4 and ROUGE-L were originally designed to evaluate machine translation, while CIDEr-D was specifically designed to evaluate the accuracy of image descriptions against reference sentences. METEOR is particularly effective in capturing the semantic aspect of captions, as it identifies all possible matches by extracting precise and synonymous matches using the WordNet database, and computes sentence-level similarity scores for matching loads. To quantitatively evaluate the performance of our approach, Table 2 indicates all the metrics of each selected method and our proposed method.

According to the data presented in Table 2, the proposed method has achieved competitive performance compared to others in terms of BLEU@4, METEOR, ROUGE-L, and CIDEr-D scores, with percentages of 41.28, 31.23, 61.07, and 134.20, respectively. In practice, the METEOR and

**FIGURE 2.** BLEU@4 scores obtained by applying the baseline and proposed models on the MSCOCO dataset.

ROUGE-L metrics focus on the assessment of appropriateness and contextual relevance, while the CIDEr-D focuses on the grammar and fluency of the captions produced. These results demonstrate the benefits of the context-adaptive gate in Bi-CARU, which can obtain the best scores for the METEOR and ROUGE-L metrics, but poor fluency and relevance. It can also be found that we had given the three cases of with and without (3) and (4) approaches in the proposed CNN-to-Bi-CARU. This additional work provides a complete view of each procedure we propose in a variety of scenarios. We found that these results trained using the att_t layer outperformed all baseline and some advanced models. Fig. 2 also illustrates the training process of BLEU@4 over the validation set. It can reach the best score faster when (3) is activated. Therefore, the combination of (3) and (4) in this work can further outperform the most advanced model in terms of METEOR and ROUGE-L scores. With more technology used in the last three rows of Table 2, it can be seen that the improvement is not significant, but its error range is smaller than the others. This is as expected in Section III-C. The purpose of (4) is to design a trainable procedure to predict image captions by adaptively considering caption information, aiming to refine the content adaptation of word prediction in decoding, and (3) performs to enhance their convergence and make hidden states more discriminative.

TABLE 3. Images of the MSCOCO dataset, and the corresponding captions predicted by various selected models.




Captions			
Reference	a man riding a skateboard down the side of a ramp	a group of young children playing a game of soccer	a large train stopped in the train station
RDN [63]	a man on a skateboard doing a trick	a soccer team playing a game of soccer	a train station with a train on the tracks
M2 [67]	a man riding a skateboard up the side of a ramp	children playing soccer on a field	a train station with a train on the tracks
GRIT [68]	there is a man that is doing a trick on a skateboard	several children are playing soccer in a field with a crowd of people	there is a train that is sitting on the tracks in the station
Proposed	a performer riding a skateboard up the side of a ramp	a team of children playing a game of soccer	a train is pulling into a train station in the evening

Table 3 demonstrates the images from the MSCOCO dataset and their captions as predicted by the various selected models. It clearly shows that the captions describe the content of the images well, but still differ in details. In our practice, the advantage of the proposed model is that it is better able to discover interesting objects and their relationships. Our results are able to represent and enhance the details between objects through contextual descriptions, while refining the weak dependence on visual features. For instance, the proposed model decodes the interested object as a “performer” rather than a “man”, which is better understood in the first image. Next, our method can identify relationships in the second image even when the interesting objects are not clear. It can further describe the “team” relationship of these children and soccer in the predicted text “a team of children playing a game of soccer”. It extracts the main objects in the image better than the baseline model thanks to CARU, which provides contextual adaptation to determine the relationships between image objects. Similarly, the third image presents a “train” in the station, but it may not accurately describe the image content if its behaviour (stopped or be pulling into) is ambiguous and independent, resulting in the caption produced by the proposed model being logically correct but not accurately describing the reference of the image content. Moreover, our method first adaptively models the context to refine the representation in captions that contain semantic relationships between decoded words. We then proceed to measure the importance of appearance features in the detected object. With a weak dependence on appearance, it can be found that our method is able to decode some background information such as the “evening” in the third image. In summary, our approach has demonstrated its ability to comprehend image content by exploiting enhanced image comprehension capabilities.

However, the performance of the proposed model is slightly lower when evaluated using the BLEU@4 and CIDEr-D metrics compared to the advanced model of GRIT. The GRIT has an advantage because it employs dual visual features to train captioners and achieves a superior transformer design in terms of BLEU@4 and CIDEr-D due to the superiority of its decoding model over the attention mechanism in the encoder transformer results. In contrast, this work focuses on improving the decoding side, which may result in slightly lower performance in encoding, but reduces the complexity of the model and yields a competitive lightweight model. As a result, our model has a clear advantage in understanding the correlation between objects of interest. It is reasonable to expect that incorporating additional information, such as appearance, motion, and attribute features, can further improve the performance of our proposed approach.

V. CONCLUSION

This work presents the CNN-to-Bi-CARU model, which uses a bidirectional structure based on the attention mechanism to better extract contextual information from the context-adaptive feature provided by the context-adaptive gate in CARU, aiming to promote comprehensive understanding and guidance of image features and apply it to tasks such as image captioning. The proposed CNN-to-Bi-CARU encodes data from media information by using a forward and backward layer of CARU to encode sentence context. The CNN extracts regions and identifies objects using an attention mechanism. A weighting process is used to refine the detection results and remove noise and uninteresting objects by collecting hidden features from each CNN layer. The produced hidden states are then passed to the proposed layer, which adopts the attention mechanism to obtain a weighted feature to compute the similarity between the

hidden states. In addition, a novel normalisation is employed to further enhance the convergence and make the hidden states more discriminative. The Bi-CARU is employed as a feature extractor in both directions. CARU's context-adaptive gate produces context information, which is then used in the attention approach to understand the relationship between the current state and the entire sequence. Experimental results on MSCOCO demonstrate that our proposed method can achieve more accurate extraction of contextual and relational information, outperforming the baseline and being competitive with advanced models on various metrics. Future work will investigate the feasibility of applying this approach to video, with a focus on improving the encoding side to produce fine-grained features from a global sequence perspective.

In future work, we have the potential to extend the proposed model to video captioning tasks. Video captioning involves identifying and describing the visual and auditory content of a video, which can be achieved by combining audio and image recognition technologies. However, video captioning is generally more complex than image captioning due to the time dimension, which requires more advanced algorithms to handle the behaviour of objects in a scene. Therefore, further investigation is required to overcome the sequence of frames/images for input in our proposed work. This will aim to improve the understanding and representation of video content, thus achieving a native sentence for video captioning.

REFERENCES

- [1] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023.
- [2] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.
- [3] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.
- [4] L. K. Allen, S. D. Creer, and M. C. Poulos, "Natural language processing as a technique for conducting text-based research," *Lang. Linguistics Compass*, vol. 15, no. 7, Jul. 2021, Art. no. e12433.
- [5] A. M. Rinaldi, C. Russo, and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," *Results Eng.*, vol. 18, Jun. 2023, Art. no. 101107.
- [6] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1372–1383, May 2020.
- [7] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2942–2956, Nov. 2019.
- [8] Z. Zohourianshahzadi and J. K. Kalita, "Neural attention for image captioning: Review of outstanding methods," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3833–3862, Nov. 2021.
- [9] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 4, pp. 297–312, Aug. 2019.
- [10] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.
- [11] D. Bacciu, A. Carta, and A. Sperduti, "Linear memory networks," in *Artificial Neural Networks and Machine Learning—ICANN 2019: Theoretical Neural Computation*. Munich, Germany: Springer, 2019, pp. 513–525.
- [12] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5694–5702.
- [13] Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du, and X. Lan, "A review of object detection based on deep learning," *Multimedia Tools Appl.*, vol. 79, nos. 33–34, pp. 23729–23791, Jun. 2020.
- [14] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.
- [15] K.-H. Chan and S.-K. Im, "BI-CARU feature extraction for semantic analysis," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICOIAC)*, Aug. 2022, pp. 183–187.
- [16] K.-H. Chan and S.-K. Im, "Sentiment analysis using bi-CARU with recurrent CNN models," in *Proc. 8th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Jun. 2023, pp. 1–5.
- [17] M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, and F. P. de Lange, "A hierarchy of linguistic predictions during natural language comprehension," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 32, Aug. 2022, Art. no. e2201968119.
- [18] K.-H. Chan, W. Ke, and S.-K. Im, "CARU: A content-adaptive recurrent unit for the transition of hidden state in NLP," in *Neural Information Processing*. Bangkok, Thailand: Springer, 2020, pp. 693–703.
- [19] S.-K. Im and K.-H. Chan, "Multilayer CARU model for text summarization," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2023, pp. 399–402.
- [20] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Exploring region relationships implicitly: Image captioning with visual relationship attention," *Image Vis. Comput.*, vol. 109, May 2021, Art. no. 104146.
- [21] Y. Zhu and S. Jiang, "Attention-based densely connected LSTM for video captioning," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 802–810.
- [22] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.
- [23] S. Chen, X. Zhong, L. Li, W. Liu, C. Gu, and L. Zhong, "Adaptively converting auxiliary attributes and textual embedding for video captioning based on BiLSTM," *Neural Process. Lett.*, vol. 52, no. 3, pp. 2353–2369, Sep. 2020.
- [24] K.-H. Chan and S.-K. Im, "Partial attention modeling for sentiment analysis of big data," in *Proc. 7th Int. Conf. Frontiers Signal Process. (ICFSP)*, Sep. 2022, pp. 199–203.
- [25] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2585–2594.
- [26] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3391–3402, Sep. 2021.
- [27] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 290–298.
- [28] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 353–369.
- [29] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image captioning: A comprehensive survey," in *Proc. Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, Feb. 2020, pp. 325–328.
- [30] R. Ramos and B. Martins, "Using neural encoder–decoder models with continuous outputs for remote sensing image captioning," *IEEE Access*, vol. 10, pp. 24852–24863, 2022.
- [31] K.-H. Chan, S.-K. Im, and W. Ke, "VGGreNet: A light-weight VGGNet with reused convolutional set," in *Proc. IEEE/ACM 13th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2020, pp. 434–439.
- [32] S. Abbasi and M. Rezaeian, "Visual object tracking using similarity transformation and adaptive optical flow," *Multimedia Tools Appl.*, vol. 80, no. 24, pp. 33455–33473, Aug. 2021.
- [33] H. Sharma and A. S. Jalal, "Incorporating external knowledge for image captioning using CNN and LSTM," *Mod. Phys. Lett. B*, vol. 34, no. 28, Jul. 2020, Art. no. 2050315.

- [34] R. A. Ahmad, M. Azhar, and H. Sattar, "An image captioning algorithm based on the hybrid deep learning technique (CNN+GRU)," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2022, pp. 124–129.
- [35] X. Huang, W. Ke, and H. Sheng, "Enhancing efficiency and quality of image caption generation with CARU," in *Wireless Algorithms, Systems, and Applications*. Cham, Switzerland: Springer, 2022, pp. 450–459.
- [36] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Feb. 2019.
- [37] Y. Liu, Y. Guo, and M. S. Lew, "What Convnets make for image captioning?" in *MultiMedia Modeling*. Reykjavik, Iceland: Springer, Dec. 2016, pp. 416–428.
- [38] D.-J. Kim, T.-H. Oh, J. Choi, and I. S. Kweon, "Dense relational image captioning via multi-task triple-stream networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7348–7362, Nov. 2022.
- [39] N. Gupta and A. S. Jalal, "Integration of textual cues for fine-grained image captioning using deep CNN and LSTM," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 17899–17908, Oct. 2019.
- [40] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, "Deep image captioning: A review of methods, trends and future challenges," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126287.
- [41] S. Cho and H. Oh, "Generalized image captioning for multilingual support," *Appl. Sci.*, vol. 13, no. 4, p. 2446, Feb. 2023.
- [42] M. Aydođan and A. Karci, "Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification," *Phys. A, Stat. Mech. Appl.*, vol. 541, Mar. 2020, Art. no. 123288.
- [43] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2021.
- [44] C. Avci, B. Tekinerdogan, and C. Catal, "Analyzing the performance of long short-term memory architectures for malware detection models," *Concurrency Comput., Pract. Exp.*, vol. 35, no. 6, p. 1, Jan. 2023.
- [45] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4257–4268, Aug. 2023.
- [46] R. Castro, I. Pineda, W. Lim, and M. E. Morochó-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022.
- [47] L. Zhao, S.-P. Lu, T. Chen, Z. Yang, and A. Shamir, "Deep symmetric network for underexposed image enhancement with recurrent attentional learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12075–12084.
- [48] Z. Fei, "Attention-aligned transformer for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 1, pp. 607–615.
- [49] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.
- [50] C. He and H. Hu, "Image captioning with visual-semantic double attention," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1, pp. 1–16, Jan. 2019.
- [51] Y. Wei, C. Wu, G. Li, and H. Shi, "Sequential transformer via an outside-in attention for image captioning," *Eng. Appl. Artif. Intell.*, vol. 108, Feb. 2022, Art. no. 104574.
- [52] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognit. Lett.*, vol. 143, pp. 43–49, Mar. 2021.
- [53] S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty, and S. Saha, "Image captioning using inception V3 transfer learning model," in *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jul. 2021, pp. 1103–1108.
- [54] L. Lou, K. Lu, and J. Xue, "Improved transformer with parallel encoders for image captioning," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 4072–4075.
- [55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
- [56] K.-H. Chan, S.-K. Im, and W. Ke, "Multiple classifier for concatenate-designed neural network," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1359–1372, Sep. 2021.
- [57] S.-K. Im and K.-H. Chan, "Vector quantization using k -means clustering neural network," *Electron. Lett.*, vol. 59, no. 7, Mar. 2023, Art. no. e12758.
- [58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [59] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.
- [60] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
- [61] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [62] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
- [63] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8888–8897.
- [64] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, F. Bach and D. Blei, Eds., Lille, France, Jul. 2015, pp. 2048–2057. [Online]. Available: <https://proceedings.mlr.press/v37/xuc15.html>
- [65] J. Zhang, K. Li, and Z. Wang, "Parallel-fusion LSTM with synchronous semantic and visual information for image captioning," *J. Vis. Commun. Image Represent.*, vol. 75, Feb. 2021, Art. no. 103044.
- [66] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "Image caption generation using contextual information fusion with bi-LSTM-s," *IEEE Access*, vol. 11, pp. 134–143, 2023.
- [67] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
- [68] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "GRIT: Faster and better image captioning transformer using dual visual features," in *Computer Vision—ECCV 2022 (Lecture Notes in Computer Science)*. Cham, Switzerland: Springer, 2022, pp. 167–184.
- [69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.



SIO-KEI IM (Member, IEEE) received the degree in computer science and the master's degree in enterprise information system from the King's College, University of London, U.K., in 1998 and 1999, respectively, and the Ph.D. degree in electrical engineering from the Queen Mary University of London (QMUL), U.K., in 2007. He was a Lecturer with the Computing Program, Macao Polytechnic Institute (MPI), in 2001. In 2005, he became the Operations Manager of the MPI-QMUL Information Systems Research Center, jointly operated by MPI and QMUL, where he carried out signal processing work. He was promoted to a Professor with the Macao Polytechnic Institute, in 2015. He was a Visiting Scholar with the School of Engineering, The University of California, Los Angeles (UCLA), and an Honorary Professor of The Open University of Hong Kong.



KA-HOU CHAN (Member, IEEE) received the bachelor's degree in computer science from the Macau University of Science and Technology, in 2009, the master's degree from the Faculty of Science and Technology, University of Macau, in 2015, and the Ph.D. degree from the Faculty of Applied Sciences, Macao Polytechnic University, Macau, China, in 2023. His research interests include algorithm analysis and optimization of video coding, image processing, parallel computing, neural networks, and computer graphics.