IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 December 2022, accepted 24 December 2022, date of publication 26 December 2022, date of current version 2 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232508

RESEARCH ARTICLE

Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s

HUAWEI ZHANG, CHENGBO MA, ZHANJUN JIANG, AND JING LIAN, (Member, IEEE)
Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730000, China

Corresponding author: Huawei Zhang (zhanghuawei@lzjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62061023, Grant 61941109, and Grant 61961037.

ABSTRACT The image caption generation algorithm necessitates the expression of image content using accurate natural language. Given the existing encoder-decoder algorithm structure, the decoder solely generates words one by one in a front-to-back order and is unable to analyze integral contextual information. This paper employs a Bi-LSTM (Bi-directional Long Short-Term Memory) structure, which not only draws on past information but also captures subsequent information, resulting in the prediction of image content subject to the context clues. The visual information is respectively fed into the F-LSTM decoder (forward LSTM decoder) and B-LSTM decoder (backward LSTM decoder) to extract semantic information, along with complementing semantic output. Specifically, the subsidiary attention mechanism S-Att acts between F-LSTM and B-LSTM, while the semantic information of B-LSTM and F-LSTM is extracted using the attention mechanism. Meanwhile, the semantic interaction is extracted pursuant to the similarity while aligning the hidden states, resulting in the output of the fused semantic information. We adopt a Bi-LSTM-s model capable of extracting contextual information and realizing finer-grained image captioning effectively. In the end, our model improved by 9.7% on the basis of the original LSTM. In addition, our model effectively solves the problem of inconsistent semantic information in the forward and backward direction of the simultaneous order, and gets a score of 37.5 on BLEU-4. The superiority of this approach is experimentally demonstrated on the MSCOCO dataset.

INDEX TERMS Bi-LSTM, image caption generation, semantic fusion, semantic similarity.

I. INTRODUCTION

Image captioning [1], [2], [3], [4] serves as a complex multi-modal scene understanding task involving two fields of study: computer vision [5], [6] and natural language processing [7], [8], whose purpose is to automatically generate proximate natural language captions for the salient visual content of input images. This task requires the model to complete the following actions: First, the model allows for comprehending the visual content in the image by identifying salient elements in the image with their mutual correspondence. Second, on the basis of these visual understandings, the model is also able to accurately describe these structured visual information word by word using natural language. Dynamic multi-modal analysis and reasoning are performed on the visual content, as well as generated words in the course of caption word generation. At present, the image captioning model is primarily based on the encoder-decoder [9], [10] approach, whose model solely examines the image's global region while generating the image caption. The encoder transforms the image as the average value of global area features, ignoring the image's local saliency. As a consequence, the attention mechanism [11] is applied to image captions, the extracted visual features are normalized into a set of weight values, and the external visual features of the encoder are corresponding to its internal semantic features, further improving the model's interpretability. In recent years, visual attention [12], [13], [14] and semantic attention [15] have proved their superiority in this domain.

The common difficulty with most approaches lies in that the deep neural network based on LSTM [16] simply considers unidirectional data input while ignoring the impact of the orientation of the sequence on prediction. Yet the prediction of a sentence is supposed to be determined by

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

the context, it hence is imperative to consider both prior and subsequent moment information. To address this issue, this paper employs the Bi-LSTM [17], [18] structure, which comprises two LSTM neural networks, featured by one forward and one backward. In contrast to the traditional unidirectional LSTM network [19], [20], [21], the Bi-LSTM structure considers the inherent laws of forward with backward data simultaneously while predicting from both the past and the future. Besides, it employs two independent hidden layers to respectively process the forward and backward semantic information. Then, the forward and backward outputs are drawn upon summation. The content is extracted from the forward and backward LSTM. As illustrated in Fig. 1, forward and backward extract semantic features about ''riding'', ''wave'', and attention mechanisms extract salient regions.

When Bi-LSTM is employed as the decoder, the image captions generated by the forward and backward generation approaches for the same image are prone to vary widely, and the semantic contents of the same time step barely match. When the current word is generated in forwarding order, the backward generation approach fails to offer effective context information synchronously; similarly. When it is generated in backward order, the forward generation approach fails to provide valid context information synchronously either. Therefore, aiming to fully utilize context information while addressing the issue of out-of-sync between forward and backward directions, this paper proposes the S-Att, which employs the subsidiary attention mechanism between F-LSTM and B-LSTM, extracting the correlation intensity of the F-LSTM and B-LSTM semantic information. As a result, the semantic information is aligned and output complementary. This method addresses the limitation that forward and backward synchronization semantics are incompatible and cannot be produced, contributing to more precise sentence predictions.

Consequently, our final model employs the CNN-Bi-LSTM-s encoder-decoder, as indicated in Fig. 2. CNN is employed to extract features and attention mechanisms to extract salient regions. Bi-LSTM is employed to extract contextual information, with S-Att raised to fuse semantics and align complementary outputs.

In summary, our main contributions are shown as follows:

- We adopt Bi-LSTM as a decoder to extract different directional features to obtain more fine-grained contextual information.
- We adopt the subsidiary attention mechanism to fix the semantic information, and align the forward and backward hidden states through the similarity module to improve the output accuracy.
- We fuse the features extracted by visual attention and subsidiary attention to obtain complementary and progressively finer grained sentences.

II. RELATED WORKS

Following various generation approaches, the current major image caption generation algorithms are split into three types[19]: module-based matching algorithms [22], [23], [24], migration-based algorithms [25], [26], and neural network-based algorithms [1], [2], [11].

The module-based matching algorithm first identifies the objects, attributes, actions, coupled with other information present in the image using multiple classifiers, and then puts the detected information into a manually designed sentence module to generate image captions. Although this algorithm is considered straightforward and intuitive, it remains difficulties to recognize more sophisticated image information and unable to generate sentences with more complicated structures given the constraints of classifiers or sentence modules [23].

The migration-based algorithm retrieves similar images in the existing database and then regards the caption of the similar image as the caption of the image to be queried. Since the sentences in the database are entirely human-generated, the migration-based algorithm produces grammatically correct sentences. However, considering that the searched image and the image to be queried are similar instead of being definitely identical, the sentences directly generated in this case may not accurately describe the content of the image to be queried.

In recent years, deep neural networks have been applied to image retrieval [27] and machine translation [28] with success. Inspired by this trend, a variety of image caption generation algorithms based on deep neural networks were proposed, followed by great breakthroughs. This type of algorithm extracts image features using CNN, and further decodes image features into fluent sentences using RNN [29]. Unlike module-based matching algorithms or migration-based algorithms, the neural network-based algorithms not only eliminate the limitation of sentence modules, but also generate novel sentences not available in current databases, which is due to the characterization capabilities of CNN combined with the efficient modeling capabilities of RNN for variable-length sequences. In a novel parallel fusion LSTM structure [30]. It adopts hidden states which are based on two parallel LSTMs to make attributes and visual image information complementary and enhanced at each time step. An innovative structure eliminates the redundancy that exists in the training set, increasing adaptive weights to increase the ability to generalize captions [31]. A more sophisticated attention mechanism [32] is employed to extract salient region features. It combines sentence-level attention models with word-level attention models to generate more accurate captions. Exploring region relationships [33] implicitly explores the relationship between related semantics and dynamically searches the related visual relationships between multiple regions, making the description of image captions more accurate. Attribute-driven image captioning model [34], which selects a specific area of the image and then decides which Attribute to focus on. This improves the coverage of visual attributes. The excellent performance of Bi-LSTM in machine translation makes many tasks try to use bidirectional LSTM. In Automatic language identification task [35], Bi-LSTM effectively extracts ''future'' speech sequences, and the effect is remarkable.

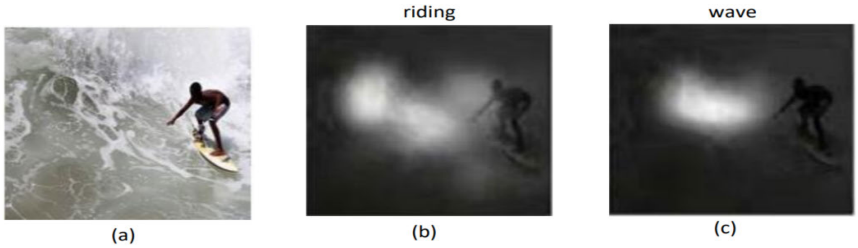


FIGURE 1. Demonstrates the features of F-LSTM and B-LSTM extracted by the decoder. (a) is the original image. (b) is the visual feature extracted by F-LSTM, and (c) is the visual feature extracted by B-LSTM.

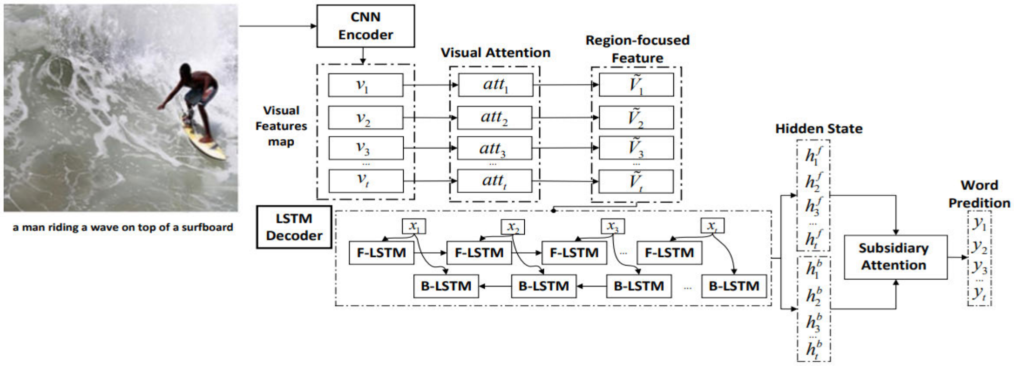


FIGURE 2. Indicates our proposed image caption attention framework. The model employs CNN as the encoder to extract the visual area features of the image, and Bi-LSTM as the decoder to extract the hidden state h_{t-1}^{f} at the previous moment with the hidden state h_{t+1}^{b} at the next moment. S-Att is introduced into Bi-LSTM to fuse and complement the hidden states h_t^f and h_t^b, thereby making further predictions.

In the sentiment analysis task [36], Bi-LSTM can effectively extract the context information and obtain more accurate prediction results. Many studies have demonstrated that features can be extracted efficiently using Bi-LSTM. In the implicit discourse relation recognition task [37], the discourse arguments are encoded by Bi-LSTM to preserve contextual information, and the final result is better than the performance of LSTM. In the event detection task [38], the algorithm adopts the Bi-LSTM model to capture contextual information, and the final result is better than the result of LSTM.

III. PROPOSED METHOD

A bidirectional LSTM is introduced as a decoder in the image caption, which efficiently extracts contextual information; meanwhile, the F-LSTM is aligned with the B-LSTM via subsidiary attention, followed by semantic complementary output. The following elaborates on our model, as presented in Fig. 3. The hidden state extracted by the fixed forward LSTM and the hidden state of the backward LSTM is semantically aligned by the similarity module.

A. ENCODER-DECODER

When an image I is given, the image caption aims to generate a sentence $Y = \{y_1, y_2 \ldots, y_T\}$ for describing the image. Thus, its purpose is to maximize the probability of the formula:

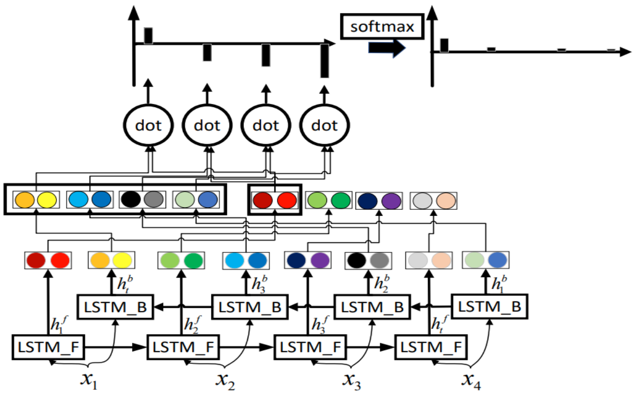


FIGURE 3. The S-Att attention model, which we propose in the middle of the bidirectional LSTM, calculates the relevance by fixing the previous moment h_{t-1}^{f} and the next moment h_{t+1}^{b}, then outputting the relevance with the softmax function before aligning.

$$\theta^* = \arg\max_{\theta} \sum_{(I,y)} \log p(Y \mid I; \theta) \tag{1}$$

θ represents the image caption parameter, typically applied in the chain rule to model the joint probabilities:

$$\log p(Y \mid I) = \sum_{t=1}^{T} \log p(y_t \mid I \ , y_{1:t-1}). \tag{2}$$

We employ a unified encoder-decoder framework to generate captions:

Encoder-CNN: The pixel value of the input image I has been fixed, and the input image I is encoded as a spatial vector using CNN, where $V = CNN(I)$ is CNN function to obtain the spatial feature $V = \{v_1, v_2, \ldots, v_z\}$. z represents the number of image space regions, and $v_i \in \mathbb{R}^D$ denotes the image space region.

Decoder-LSTM: The conventional recurrent neural network (RNN) comes with the issue of gradient vanishing while processing time-series tasks; thus, we adopt the long short-term memory (LSTM) to replace the conventional RNN as the encoder. Compared with RNN, LSTM adds three threshold units (input gate, forget gate, output gate) to control the flow of data. The forget gate connects the hidden state h_{t-1} of the previous moment with the input x_t of the current moment as the total input of the sigmoid activation function to generate the forget mask f_t. The product of f_t and the memory information c_{t-1} of the previous moment achieves the purpose of removing the previous moment's worthless information. The input gate computes the output mask i_t using the same way, and employs i_t to filter the memory information $\tilde{c}_t$ at the current moment. Once c_{t-1} and $\tilde{c}_t$ are filtered, they are further summed to obtain comprehensive memory information c_t. The output gate computes the output mask o_t following the same way as the first two threshold units, and the comprehensive memory information c_t is multiplied with o_t after going through the tanh activation function to obtain the hidden state h_t at the current moment. The computational procedure is as follows:

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\ f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\ o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\ \tilde{c}_t &= \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \\ c_t &= i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

B. VISUAL ATTENTION GUIDE

In order to make the most of semantic and visual information, we incorporate the two using soft attention [11] in LSTM. The primary task is to properly integrate semantic and visual information. Second, more focus is paid to different time steps under the two elements of information. As a result, the visual output shifts from the same global image features to changing image local features as each word is generated. Attention dynamically extract the attention from images in response to changes in the visual context. It is defined as follows:

$$z_t^i = W_z \tanh(W_v V_i + W_h h_t) \tag{4}$$

$W_z \in \mathbb{R}^{1\times k_1}$, $W_v \in \mathbb{R}^{k_1\times k_2}$ and $W_h \in \mathbb{R}^{k_1\times k_3}$ represent the trainable parameter (transition matrices), W_v is denoted as drawing the visual feature V_i into a visual feature map. W_h refers to plotting the semantic feature h_t as a semantic feature map.

$$a_t^i = softmax(z_1^i, z_2^i, \ldots, z_t^i) \tag{5}$$

It is normalized via softmax, thereby generating the attention weight distribution.

$$\bar{V}_t = \sum_n a_t^i v_i \tag{6}$$

$\bar{V}_t$ represents the generated visual attention feature.

C. BI-LSTM

The conventional LSTM simply predicts the output of the next moment based on the temporal information of the present moment. However, the output of the current moment is relevant to both the state of the previous moment and the next moment. Predicting the exact word in a sentence, for instance, should be judged not only on the prior text but also on the following content, thereby realizing proper judgments based on context.

$$h_t^f = LSTM([x_t; \tilde{V}_t], h_{t-1}^f) \tag{7}$$

$$h_t^b = LSTM([x_t; \tilde{V}_t], h_{t+1}^b) \tag{8}$$

$$p_t^f(y_t|y_1, y_2, \ldots, y_{t-1}, I) = softmax(h_t^f) \tag{9}$$

$$p_t^b(y_t|y_1, y_2, \ldots, y_{t-1}, I) = softmax(h_t^b) \tag{10}$$

The hidden state h_t^f of the F-LSTM at time t depends on the hidden state h_{t-1}^f and input x_t of the previous moment, while the hidden state of the B-LSTM at time t depends on the hidden state h_{t+1}^b and input x_t at the next moment. p_t^f and p_t^b represent the word conditional probability distribution of F-LSTM and B-LSTM at time t, respectively.

D. SUBSIDIARY ATTENTION GUIDE

The hidden state h_t at time t is obtained in Bi-LSTM by summing the h_{t-1}^f and h_{t+1}^b; the former is obtained with F-LSTM, while the latter is obtained with B-LSTM. F-LSTM and B-LSTM output semantics are inconsistent and impossible to be aligned, resulting in unsatisfactory output. Consequently, the semantics are fixed as the F-LSTM output hidden state h_t^f. The fixed semantics facilitates alignment, with the subsequent subsidiary attention mechanism extracting the semantic similarity of the forward and backward LSTMs. Our semantic similarity goal is to numerically indicate how similar h_t^b is to the individual word vectors of h_t^f. To indicate how much two vectors point in the same direction, we take a simple inner product of vectors, and hence utilize the inner product as the similarity of two vectors.

$$z_t^s = h_{t-1}^f \cdot h_{t+1}^b \tag{11}$$

$$a_t^s = softmax(z_1^s, z_2^s, \ldots z_t^s) \tag{12}$$

a_t^s represents the weight of similarity degree in the forward and backward direction at time t.

$$h_t^m = h_{t-1}^f + h_{t+1}^b \tag{13}$$

$$h_t^n = h_{t-1}^f + h_*^b \tag{14}$$

The summation of h_{t-1}^{f} and h_{t+1}^{b} indicates that Bi-LSTM extracts past and future information to obtain the hidden state h_t^m based on visual attention; the summation of h_{t-1}^{f} and h_{t-1}^{f} corresponding to each position represents the hidden state h_t^n of the context aligned by a fixed position via S-Att. Then, the final hidden state under dual attention is as follows:

$$h_t = \lambda h_t^m + (1-\lambda)h_t^n \tag{15}$$

The loss function of the bidirectional LSTM includes:

$$L_{XE}^{f}(\theta) = -\sum_{t=1}^{T} \log p_{\theta}^{f}(y_t|y_{1:t-1}) \tag{16}$$

$$L_{XE}^{b}(\theta) = -\sum_{t=1}^{T} \log p_{\theta}^{b}(y_t|y_{1:t-1}) \tag{17}$$

$$L = L_{XE}^{f} + L_{XE}^{b} \tag{18}$$

L_{XE}^{f} and L_{XE}^{b} stand for the loss functions of the F-LSTM and B-LSTM, respectively. The conventional cross-entropy error training strategy is employed, with L as the final loss function. There is a distinction between training and testing conventional image captioning models, with testing relying on words previously generated by the model. When the preceding period's results are incorrect, the errors would accumulate and succeeding words cannot be generated correctly. To address these issues, we approach image caption production as a reinforcement learning problem, directly optimizing sentence generation based on the model's evaluation metrics, with the ultimate goal of minimizing the following negative expected returns:

$$L_{RL}(\theta) = -E_{Y^s \sim p_\theta}[r(Y^s)] \tag{19}$$

$r(Y^s)$ serves as the reward obtained via CIDEr [39], BLEU [40] and other computing methods when the prediction is over; besides, LSTM updates its internal hidden state attention weights and other states.

The gradient can be approximated by the following formula:

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y^s) - r(y^*))\nabla_\theta \log p_\theta(y^s) \tag{20}$$

y^* represents the baseline score obtained at test time with beam search decoding.

IV. EXPERIMENTS

In order to demonstrate the effectiveness of the proposed bidirectional LSTM model, we perform extensive experiments to test the model while also comparing it with the advanced models. The elaborated material of the experiment is included as follows, ranging from the dataset to assessment metrics, implementation details, and testing approach.

A. DATASET

We evaluate our model on the widely-used mscoco [41] dataset, which acts as a large-scale dataset with diversified object identification, segmentation, and captioning; each image is collected from daily life, making it the primary experimental dataset for image captioning. the individual image contains a multi-entity target with five manual labels for labeling the caption. this dataset includes 91 targets, 328,000 images, and 2.5 million labels. the largest dataset with semantic segmentation provides 80 categories, over 330,000 images, 200,000 of which are annotated, and over 1.5 million individuals in the entire dataset. we adopt 110,000 photos for training, 5,000 images for validation, and 5,000 images for testing [1].

B. EVALUATION METHODS

In experiments, we adopt bleu1-4 [40], meteor [42], rouge-l [43], cider [39], and spice [3] to evaluate our model performance metrics, which are widely used in image captioning.

C. DATA PRE-PROCESSING

In this paper, we implement a bidirectional LSTM with a subsidiary attention mechanism. Our parameter settings and experimental details are as followers.

First, we replace all words in the dataset with lowercase, and truncate the sentence caption length to 16, among which the words with a frequency of less than or equal to 5 are deleted, and finally a word list with a number of 9500 is obtained.

Second, we adopt the pre-trained Resnet-101 [44] to encode the image in the encoding phase, which encodes the image into a visual feature map of size 14 × 14 with 2048 dimensions. The visual feature map is mainly applied to represent the fine-grained information of the image.

D. DECODING PHASE

We employ a Bi-LSTM-s structure to decode visual feature maps into image captions with word embedding of dimension as 512. The forward LSTM, backward LSTM, and attention dimension are set to 512.

Finally, during the training phase, we train our model with the Adam [45] optimizer under the cross-entropy loss. We fine-tune Resnet-101's last convolutional layer to adjust the appropriate training parameters. The learning rate is 1×10^{-5}, which decreases by 0.5 every six epochs. The batch size is set to 64, and the model is trained for 30 epochs. Subsequently, building on the training model, we employ reinforcement learning-based methods in order to optimize the CIDEr assessment metrics. At this phase, the learning rate is set to 5×10^{-5}, the batch size is set to 64, and the training runs for 30 epochs. During the training phase, we evaluate our model on the validation machine at the conclusion of each epoch and save the model with the best current result. Then, the next phase of training will continue on the model with the best performance from the previous phase. In terms of testing, we select the model with the greatest CIDEr score on the validation set, and we utilize beam search to produce phrases with the beam size set to 5.

If the performance fails to improve after 6 training epochs, the training would be terminated.

TABLE 1. Verifying the performance of optimized CIDEr and auxiliary attention mechanism using MSCOCO.

Loss	Method	B@1	B@2	B@3	B@4	M	R	S	C
XE	Bi-LSTM	77.9	61.4	46.9	36.4	27.5	56.1	20.1	112.5
	Bi-LSTM-s	78.9	63.4	48.3	37.0	28.5	57.3	21.9	118.6
RL	Bi-LSTM	78.3	63.1	47.5	36.9	28.3	56.9	21.4	117.9
	Bi-LSTM-s	**79.3**	**64.0**	**49.1**	**37.5**	**28.9**	**58.1**	**22.5**	**121.3**

TABLE 2. Verifying the performance of Bi-LSTM and parallel double-layer LSTM using MSCOCO.

Method	B@1	B@2	B@3	B@4	M	R	S	C
p-LSTM	76.4	60.9	45.6	35.4	26.8	54.6	19.6	107.6
Bi-LSTM	**78.3**	**63.1**	**47.5**	**36.9**	**28.3**	**56.9**	**21.4**	**117.9**

E. EXPERIMENTAL RESULTS AND ANALYSIS

We design ablation experiments to evaluate the effectiveness of our proposed model in image captioning; all metric scores are designed using the MSCOCO Karparthy test segmentation.

First, as shown in Table 1, the improvement brought by CIDEr optimization is validated on our model. XE represents training under the cross-entropy loss function, and RL indicates the result of optimizing the scoring index based on the optimal XE training model. Second, two sets of models, Bi-LSTM and Bi-LSTM-s, are set to verify the effectiveness of our auxiliary attention mechanism. Bi-LSTM solely employs the visual attention mechanism, while Bi-LSTM-s incorporates S-Att into Bi-LSTM. The following training results ensure that the parameters are consistent in order to maintain fairness.

As illustrated in the table 1-5, each represents a different ablation experiment. In Tables 1, we demonstrated that reinforcement learning improves significantly in image caption tasks. In Tables 2 and 3, we proved the superiority of the our model. In Table 4, we verified the effect of averaging and taking the maximum input on the result when semantic fusion occurs. In Table 5, it is the influence of different hyperparameters on the results of the experiment.

The following evaluation criteria: B@1, B@2, B@3, B@4, M, R, S, C, represent BLEU1-4, METEOR, ROUGE-L, SPICE, CIDEr. BLEU: It calculates the similarity and penalizes sentences of insufficient length. METEOR: It focuses on the number of co-occurrences of words and establishes a penalty mechanism based on word order changes to get scores. ROUGE-L: The similarity is measured by calculating the longest common sequence between the predicted sentence and the standard translation.

SPICE: It encodes images into objects, attributes, and relationships, and then selects the highest scoring statement based on scene graphs. CIDEr: It calculates the similarity, which is based on the frequency of the words.Firstly, as shown in Table 1. The cross-entropy loss function is compared to the optimized CIDEr score (using the CIDEr score as an example), the CIDEr score of our Bi-LSTM model climbed from 112.5 to 117.9 by 5.4; the CIDEr score of our Bi-LSTM-S model climbed from 118.6 to 121.3 by 2.7, indicating that the current leading methodologies come with a significant enhancement in optimizing CIDEr on the basis of cross-entropy error. Second, the Bi-LSTM-s improved from 112.5 to 118.6 by 6.1 in the cross-entropy loss function experiment, and 3.4 improved from 117.9 to 121.3 in the optimized CIDEr experiment. Thus, it reflects that our subsidiary attention can efficiently extract, align, and produce finer captions from the semantic relations of the forward LSTM and backward LSTM.

Secondly, as shown in Table 2. Our ablation experiments are primarily utilized to validate our model's superiority, with constant experimental training parameters. To be specific, p-LSTM denotes the simultaneous superposition of two layers of structure processing, in which the hidden state $\tilde{h}_t^1$ of the first layer of p-LSTM is learnt and $\tilde{h}_t^1$ is transferred to the second layer of LSTM; then, the input gate, forget gate, and output gate of the second layer of LSTM will all employ $\tilde{h}_t^1$ as input. The following uncovers the final hidden state:

$$\tilde{h}_t^2 = LSTM(\tilde{h}_t^1, \tilde{h}_{t-1}^2) \tag{21}$$

The final hidden state at time t is derived from the first layer's hidden state $\tilde{h}_t^1$ and the second layer's previous moment $\tilde{h}_{t-1}^2$. According to the optimized CIDEr score, the Bi-LSTM model has improved by 10.3 points. In our model, our hidden state computation h_t is related not only to the current input, but also to h_{t-1}^f and h_{t+1}^b. Bi-LSTM considers previous and future information simultaneously; thus, it truly achieves context-based output.

Aiming to demonstrate the superiority of our model, we evaluate it against eight measures and six prominent methods. As shown in Table 3. First, the foundational model is established. The most typical model, NIC, does not include an attention mechanism. The goal of Soft-Attention is to introduce a soft attention mechanism into difficult tasks. The attention mechanism is extended from spatial to channel by SCA-CNN. SCST is the application of reinforcement learning to the optimization of sentence-level rewards. Second, the pLSTM-A-2, DAIC and our model are improved on the

TABLE 3. Comparison with advanced models on MSCOCO dataset.

Model	B@1	B@2	B@3	B@4	M	R	S	C
Google NIC[2]	66.6	46.1	32.9	24.6	-	-	-	-
Soft-Attention[11]	70.7	49.2	34.4	24.3	23.9	-	-	-
SCA-CNN[20]	71.9	54,8	41,1	31,1	25.0	-	-	-
SCST[46]	-	-	-	34.2	26.7	55.7	-	114.0
pLSTM-A-2[30]	76.3	60.0	46.2	35.2	26.8	56.9	19.6	108.3
DAIC[32]	77.6	61.8	47.4	35.4	26.7	55.7	-	114.0
Ours	**79.3**	**64.0**	**49.1**	**37.5**	**28.9**	**58.1**	**22.5**	**121.3**

TABLE 4. Forward and backward LSTM hidden state fusion on MSCOCO dataset.

Method	B@1	B@2	B@3	B@4	M	R	S	C
Max	78.6	63.8	48.9	37.2	**29.3**	58.0	22.1	119.6
Mean	**79.3**	**64.0**	**49.1**	**37.5**	28.9	**58.1**	**22.5**	**121.3**

TABLE 5. Dual attention weight coefficients.

λ	B@1	B@2	B@3	B@4	M	R	S	C
0.3	76.2	58.6	44.7	35.2	27.9	55.6	19.5	110.1
0.4	**79.3**	**64.0**	**49.4**	**37.5**	**28.9**	**58.1**	**22.5**	**121.3**
0.5	79.0	63.1	48.2	36.8	28.4	57.5	21.9	119.4
0.6	77.4	60.1	45.9	35.6	28.1	56.8	21.1	116.8

basis of the above models. pLSTM-A-2 encodes images using two separate encoders (MIML and CNN) and simultaneously merges the semantic information of the two decoders, resulting in more accurate and richer captions.

DAIC extracts the encoder's image input to sentence-level and word-level attention respectively, while the final output combines sentence-level and word-level information to generate more accurate captions. Our model employs bidirectional LSTM as the decoder, accepts both past and future information simultaneously, and truly achieves prediction based on contextual information. It also employs two attention mechanisms, one of which will dynamically extract visual information accompanied by integrating visual information with semantic information; another auxiliary attention aligns the semantic information of the bidirectional LSTM, contributing to more diversified semantic information. Our model has displayed significant advantages in scoring.

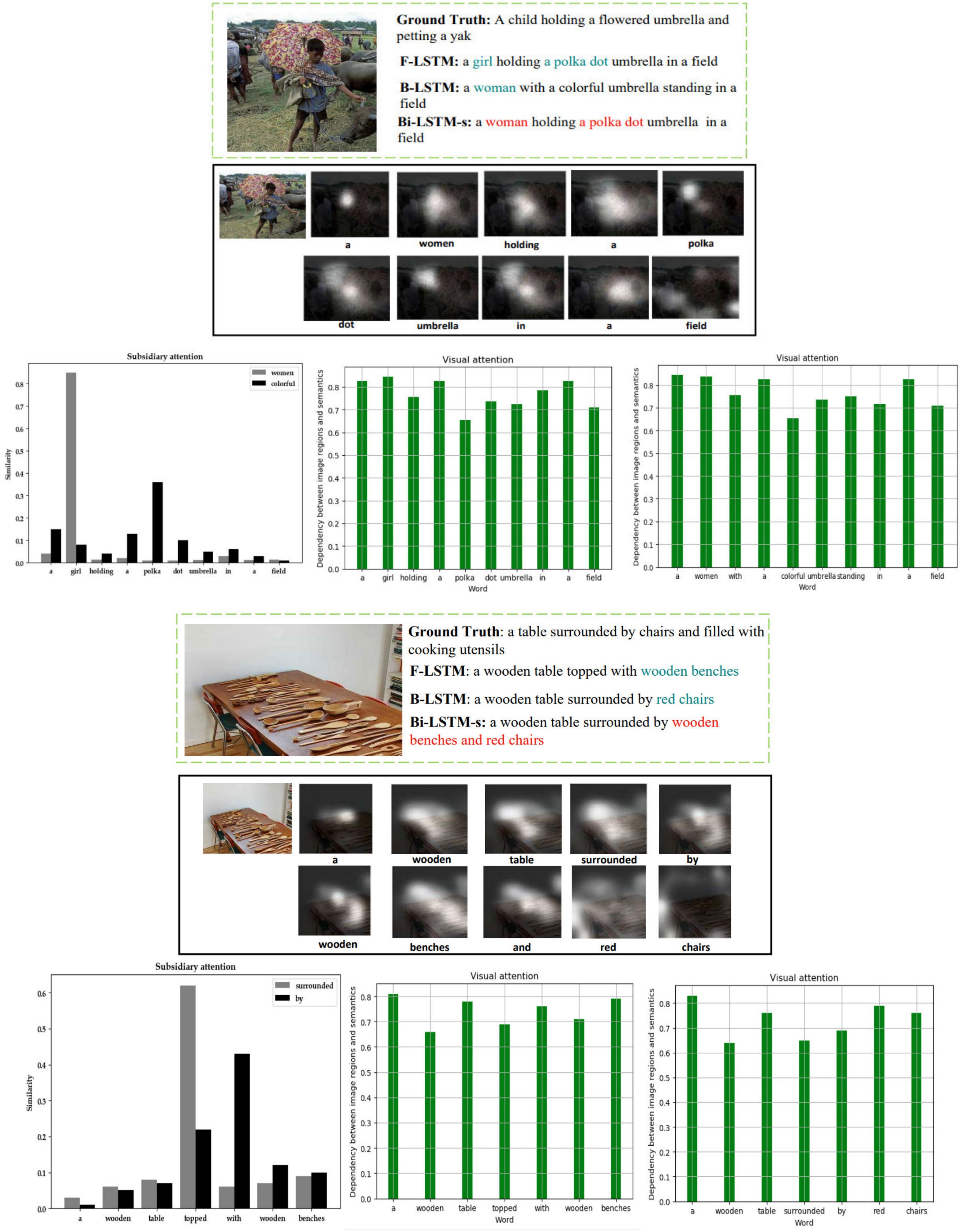
As shown in Table 4. Considering the fusion of the forward and backward hidden states: Max is the maximum value of h_{t-1}^{f} and h_{t+1}^{b}, while Mean is the weighted sum of h_{t-1}^{f} and h_{t+1}^{b}. The data reveals that the outcome of taking the average is slightly better. When fusing forward and backward semantics using Max, simply considering forward or backward to obtain a single result causes insufficient semantics and the loss of partial semantics. On the other hand, using Mean considers the shared scope of forward and backward while retaining the original semantic information, thereby achieving fused semantics.

As shown in Table 5. Under the combined effect of dual attention, an oversized selection of λ results in the extraction of unaligned semantics before-and-after, worsening the caption result. Yet, a small selection of λ leads to excessive reliance on similarity; the prior semantics over-absorb the following semantics, degrading the caption result.

F. VISUALIZATION

Fig. 4 depicts the visualization results, which allow us to better represent our proposed approach, including the ground truth. F-LSTM, B-LSTM, and Bi-LSTM-s fused with semantic features based on contextual output. It also displays the interaction among Visual-Att focusing on image key regions and text dependencies, coupled with the extraction of keywords using an auxiliary attention mechanism. Moreover, a visualization is presented at the same time. All of the image elements are derived from the MSCOCO dataset

It reveals that our model can effectively extract fine-grained information, such as "polka dot", "wooden benches", and "red chairs". F-LSTM extracts "polka dot," and the semantic features are fused into Bi-LSTM. F-LSTM extracts "wooden benches," with B-LSTM extracting "red chairs," and the one is complemented with another for output. S-Att extracts "girl" and "women," presenting a dependency of 0.85; then, they are fused to complement the output. Also, "surrounded" and "topped" have a dependency of 0.62, while the two are fused to complement the output. Fig. 5 depicts the fine-grained information extracted by our model. We set Bi-LSTM as the control groups. The subsidiary



Ground Truth: A child holding a flowered umbrella and petting a yak

F-LSTM: a girl holding a polka dot umbrella in a field

B-LSTM: a woman with a colorful umbrella standing in a field

Bi-LSTM-s: a woman holding a polka dot umbrella in a field

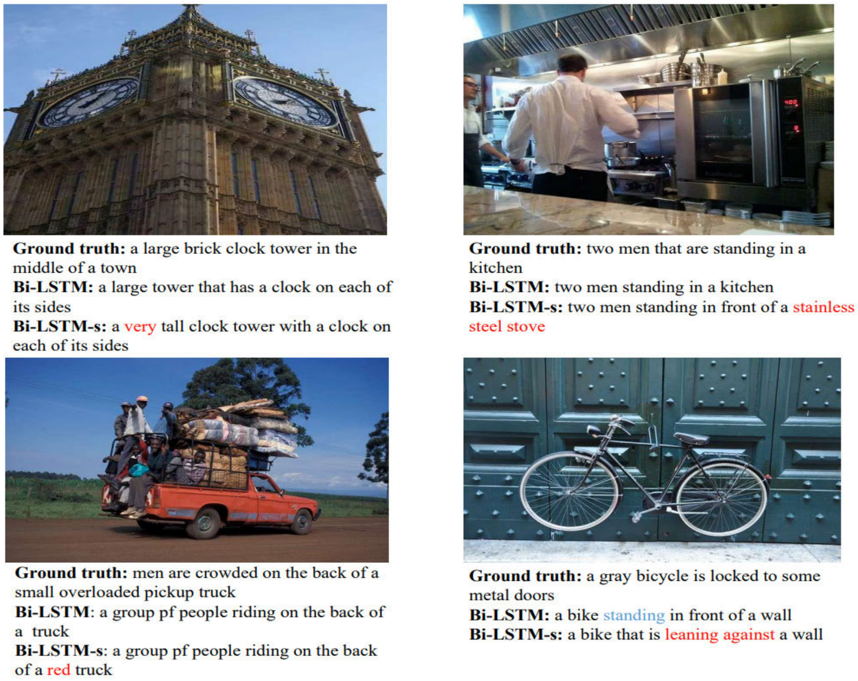
Ground Truth: a table surrounded by chairs and filled with cooking utensils

F-LSTM: a wooden table topped with wooden benches

B-LSTM: a wooden table surrounded by red chairs

Bi-LSTM-s: a wooden table surrounded by wooden benches and red chairs

FIGURE 4. Comparison with ground truth on MSCOCO.



Ground truth: a large brick clock tower in the middle of a town
Bi-LSTM: a large tower that has a clock on each of its sides
Bi-LSTM-s: a very tall clock tower with a clock on each of its sides

Ground truth: two men that are standing in a kitchen
Bi-LSTM: two men standing in a kitchen
Bi-LSTM-s: two men standing in front of a stainless steel stove

Ground truth: men are crowded on the back of a small overloaded pickup truck
Bi-LSTM: a group pf people riding on the back of a truck
Bi-LSTM-s: a group pf people riding on the back of a red truck

Ground truth: a gray bicycle is locked to some metal doors
Bi-LSTM: a bike standing in front of a wall
Bi-LSTM-s: a bike that is leaning against a wall

FIGURE 5. Supplemental experimental simulation diagrams on MSCOCO.

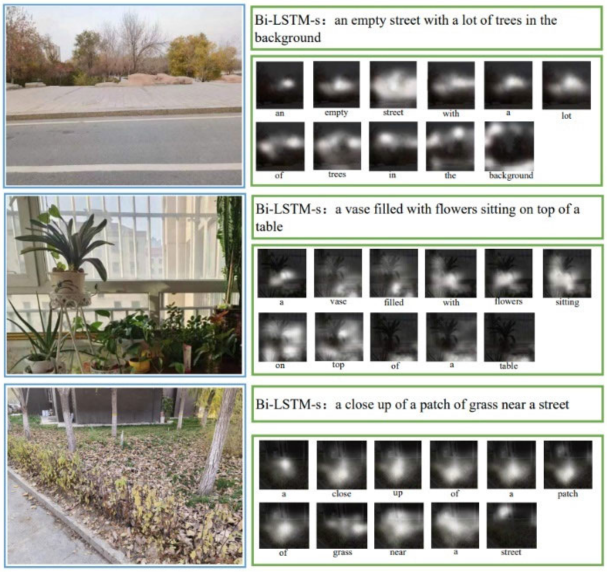

Bi-LSTM-s： an empty street with a lot of trees in the background
an
empty
street
with
a
lot
of
trees
in
the
background
Bi-LSTM-s： a vase filled with flowers sitting on top of a table
a
vase
filled
with
flowers
sitting
on
top
of
a
table
Bi-LSTM-s： a close up of a patch of grass near a street
a
close
up
of
a
patch
of
grass
near
a
street

FIGURE 6. Supplementary experimental pictures are from real photography.

attention mechanism effectively complements the forward and backward output hidden states with progressive output to obtain fuller semantics, such as ''very'', ''red'', and predicting the fine-grained information ''stainless steel stove'', the action will be more comprehensively such as ''leaning against''. Fig.6 shows that all the photos are taken from real life, and our model can extract fine-grained information.

V. CONCLUSION

At present, the existing mainstream models simply take into account the impact of the previous information on sentences. A model Bi-LSTM-s is hence created to efficiently extract past and future information in order to fully extract context information. Specifically, Bi-LSTM-s encodes the sentence context as hidden states of F-LSTM and B-LSTM, respectively. After that, S-Att obtains the word similarity between the hidden states via the attention mechanism, performing semantic alignment, semantic complementarity, as well as semantic fusion output. With extensive experimental analysis achieved on the MSCOCO dataset, our model allows us to fully extract contextual information, together with fine-grained information. Furthermore, we demonstrate the superiority of this strategy using a range of evaluation metrics.

However, bidirectional LSTM still has its limitations. First of all, bidirectional LSTM has too many parameters, which may lead to prediction time delay for real-time tasks. Secondly, two basic LSTM cells still work inside the bidirectional LSTM, and the GRU with fewer parameters can be considered to replace the LSTM during training. At present, most training features encourage the output of words with high frequency priority, which leads to the restriction of semantic information. The further study, we will focus on generating different constraints to produce fine-grained semantic information from a global perspective.

REFERENCES

[1] 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.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, ''Show and tell: A neural image caption generator,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[3] P. Anderson, B. Fernando, M. Johnson, and S. Gould, ''SPICE: Semantic propositional image caption evaluation,'' in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Berlin, Germany: Springer, Oct. 2016, pp. 382–398.

[4] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Trans. Multimedia*, vol. 23, pp. 2413–2427, 2021.

[5] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[6] L. Ruotsalainen, A. Morrison, M. Makela, J. Rantanen, and N. Sokolova, "Improving computer vision-based perception for collaborative indoor navigation," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4816–4826, Mar. 2022.

[7] S. Wu, D. Zhang, Z. Zhang, N. Yang, M. Li, and M. Zhou, "Dependency-to-dependency neural machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2132–2141, Nov. 2018.

[8] M. A. Kastner, K. Umemura, I. Ide, Y. Kawanishi, T. Hirayama, K. Doman, D. Deguchi, H. Murase, and S. Satoh, "Imageability- and length-controllable image captioning," *IEEE Access*, vol. 9, pp. 162951–162961, 2021.

[9] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[10] Y. Luo, J. Lu, X. Jiang, and B. Zhang, "Learning from architectural redundancy: Enhanced deep supervision in deep multipath encoder–decoder networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4271–4284, Sep. 2022.

[11] 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. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[12] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4894–4902.

[13] 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.

[14] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2017, pp. 4133–4139.

[15] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

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

[17] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, Apr. 2017.

[18] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[19] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2407–2415.

[20] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[21] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.

[22] A. Gupta and P. Mannem, "From image annotation to image description," in *Proc. 19th Int. Conf. Neural Inf. Process. (ICONIP)*, Doha, Qatar. Berlin, Germany: Springer, Nov. 2012, pp. 196–204.

[23] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[24] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Heraklion, Greece. Berlin, Germany: Springer, Sep. 2010, pp. 15–29.

[25] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2012, pp. 359–368.

[26] V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daumé, A. C. Berg, Y. Choi, and T. L. Berg, "Large scale retrieval and generation of image descriptions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, Aug. 2016.

[27] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017.

[28] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[29] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[30] 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.

[31] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021.

[32] H. Wei, Z. Li, C. Zhang, and H. Ma, "The synergy of double attention: Combine sentence-level and word-level attention for image captioning," *Comput. Vis. Image Understand.*, vol. 201, Dec. 2020, Art. no. 103068.

[33] 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.

[34] Y. Zhou, J. Long, S. Xu, and L. Shang, "Attribute-driven image captioning via soft-switch pointer," *Pattern Recognit. Lett.*, vol. 152, pp. 34–41, Dec. 2021.

[35] H. S. Das and P. Roy, "A CNN-BiLSTM based hybrid model for Indian language identification," *Appl. Acoust.*, vol. 182, Nov. 2021, Art. no. 108274.

[36] W. Li, L. Zhu, Y. Shi, K. Guo, and E. Cambria, "User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models," *Appl. Soft Comput.*, vol. 94, Sep. 2020, Art. no. 106435.

[37] F. Guo, R. He, and J. Dang, "Implicit discourse relation recognition via a BiLSTM-CNN architecture with dynamic chunk-based max pooling," *IEEE Access*, vol. 7, pp. 169281–169292, 2019.

[38] G. Xu, Y. Meng, X. Zhou, Z. Yu, X. Wu, and L. Zhang, "Chinese event detection based on multi-feature fusion and BiLSTM," *IEEE Access*, vol. 7, pp. 134992–135004, 2019.

[39] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2014, pp. 740–755.

[42] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[43] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

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

[46] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.