applied sciences

MDPI

Article

Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning

Tian Xie, Weiping Ding, Jinbao Zhang, Xusen Wan and Jiehua Wang *

School of Information Science and Technology, Nantong University, Nantong 226019, China; xietian@stmail.ntu.edu.cn (T.X.); ding.wp@ntu.edu.cn (W.D.); kingbao@ntu.edu.cn (J.Z.); wanxusen1997@163.com (X.W.)

* Correspondence: wang.jh@ntu.edu.cn; Tel.: +86-139-6295-5885

Abstract: The discipline of automatic image captioning represents an integration of two pivotal branches of artificial intelligence, namely computer vision (CV) and natural language processing (NLP). The principal functionality of this technology lies in transmuting the extracted visual features into semantic information of a higher order. The bidirectional long short-term memory (Bi-LSTM) has garnered wide acceptance in executing image captioning tasks. Of late, scholarly attention has been focused on modifying suitable models for innovative and precise subtitle captions, although tuning the parameters of the model does not invariably yield optimal outcomes. Given this, the current research proposes a model that effectively employs the bidirectional LSTM and attention mechanism (Bi-LS-AttM) for image captioning endeavors. This model exploits the contextual comprehension from both anterior and posterior aspects of the input data, synergistically with the attention mechanism, thereby augmenting the precision of visual language interpretation. The distinctiveness of this research is embodied in its incorporation of Bi-LSTM and the attention mechanism to engender sentences that are both structurally innovative and accurately reflective of the image content. To enhance temporal efficiency and accuracy, this study substitutes convolutional neural networks (CNNs) with fast region-based convolutional networks (Fast RCNNs). Additionally, it refines the process of generation and evaluation of common space, thus fostering improved efficiency. Our model was tested for its performance on Flickr30k and MSCOCO datasets (80 object categories). Comparative analyses of performance metrics reveal that our model, leveraging the Bi-LS-AttM, surpasses unidirectional and Bi-LSTM models. When applied to caption generation and image-sentence retrieval tasks, our model manifests time economies of approximately 36.5% and 26.3% vis-a-vis the Bi-LSTM model and the deep Bi-LSTM model, respectively.

Keywords: image captioning; bidirectional long short-term memory; attention mechanism; fast region-based convolutional network; common space

Citation: Xie, T.; Ding, W.; Zhang, J.; Wan, X.; Wang, J. Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning. *Appl. Sci.* **2023**, *13*, 7916. https://doi.org/10.3390/app13137916

Academic Editors: Mourad Oussalah and Rachid Jennane

Received: 14 June 2023
Revised: 28 June 2023
Accepted: 3 July 2023
Published: 6 July 2023

1. Introduction

Image captioning is a hot topic involving several fields such as computer vision (CV) and natural language processing (NLP), known as image semantic description or "talking about pictures" [1–6]. Image captioning technology not only needs to recognize the entity object information in the image and the relationship between objects but also needs to learn how to integrate them into the ability to describe reasonable sentence descriptions. Traditional methods use models based on visual space search, sentence template usage, and the most matching sentence in the dataset to accomplish the tasks of image captioning. The disadvantage of these methods is the low efficiency of generating real and accurate sentences and the poor ability to generate structurally novel sentences. In recent research [7–11], visual and language information has been embedded into a common space via recurrent neural networks (RNNs) initially. Convolutional neural networks

(CNNs) were then embedded within the visual space and combined with long short-term memory (LSTM) to produce more effective results.

Most models extract image features by embedding the CNN into visual space. While this method can achieve good results, the extracted image features are not highly accurate and efficient, wasting a lot of time. Many models embed LSTM and Bi-LSTM into language space to generate sentences, but the results are not accurate enough. Therefore, it is challenging for subtitling models to perform novel subtitling tasks with accurate and efficient image-sentence retrieval.

To address these issues, we propose a model leveraging a bidirectional LSTM coupled with an attention mechanism (Bi-LS-AttM). This innovative model substitutes the region convolutional neural network (RCNN)—commonly used for feature extraction—with a more efficient fast region convolutional neural network (Fast RCNN). This adjustment enhances the identification and extraction of features within the image's regions of interest (RoIs). The optimized model is then applied to refine the LSTM network's performance. By juxtaposing forward and backward outcomes and incorporating the attention boost, the Bi-LS-AttM is able to predict word vectors with greater precision and generate more fitting image captions.

Why do we use the model? We employed the model to break through the boundaries of the traditional Bi-LSTM model, which is not focused enough on the comparison of historical and future word results. In the traditional LSTM cells, the prediction of the next word x_t using the visible context V and historical context $x_{1:t-1}$ is performed by estimating $\log P(x_t|V, x_{1:t-1})$. However, in the Bi-LS-AttM, the prediction of the word x_t depends on the forward and backward results of separately maximizing $\log P(x_t|V, x_{1:t-1})$ and $\log P(x_t|V, x_{t+1:t})$ at time t. By combining the Bi-LSTM with the attention model, the model focuses increasingly on comparing historical and future word results and using their dependencies to predict and generate appropriate image captions. Figure 1 shows the example image of the Bi-LS-AttM model generating a sentence that supports our hypothesis that the Bi-LS-AttM model can generate more complementary and focused captions.

(**a**) Caption generation on Flickr30K (**b**) Caption generation on MSCOCO

Figure 1. Example captions generated by the model. (**a**) Caption generation (by the unidirectional model (**upper**) and by our model (**lower**)) on Flickr30K. (**b**) Caption generation (by the unidirectional model (**upper**) and by our model (**lower**)) on MSCOCO.

We tested the efficiency of our model on the datasets Flickr30K and MSCOCO and performed a qualitative analysis. The analysis showed that the method performs efficiently, and the proposed Bi-LS-AttM model outperforms other published models. The principal contributions of this paper are threefold:

- We proposed a trainable model incorporating a bidirectional LSTM and attention mechanism. This model embeds image captions and scores into a region by capitalizing on the long-term forward and backward context.
- We upgraded the feature extraction mechanism, replacing the conventional CNN and RCNN with a Fast RCNN. This improvement enhances the model's ability to rapidly detect and extract features from items within an image's regions of interest.
- We verified the efficiency of the framework on two datasets Flickr30K and MSCOCO. The evaluation demonstrated that the Bi-LSTM and attention mechanism model

achieved highly competitive performance results relative to current techniques in the tasks of generating captions and image sentence retrieval.

2. Related Works

Initially, researchers utilized computers to analyze identified content in image captioning, which was the original task for image recognition [12–14]. Later, they introduced additional requirements such as processing and determining object properties, identifying object relationships, and describing image content in natural language. Since then, numerous image captioning techniques have been introduced, broadly categorized into three groups: template-based, retrieval-based, and deep-learning-based methods.

Template-based methods, which utilize fixed templates for sentence generation, identify image elements such as objects, actions, and scenes based on visual dependency grammar. For instance, Farhadi [15] used a support vector machine (SVM) [16,17] to detect image items and pre-established templates for sentence descriptions. However, the limitations of datasets and template algorithms impeded their performance. Similarly, Li [18] employed Web-scale N-grams for phrase extraction linked to objects, actions, and relationships in 2011. Later, Kulkarni [19] used a conditional random field (CRF) [20,21] for data extraction from a large pool of visual descriptive text, thereby improving computer vision recognition and sentence generation. Despite these efforts, the performance of these methods was suboptimal due to the inherent restrictions of template-based approaches.

The retrieval-based method stores all image descriptions in a collection. The image to be described is then compared to the training set and filtered to find similar images. Using a similar image description to the one found, the candidate description is modified accordingly. Kuznetsova [22] proposed to search for images with attached titles on the Internet and obtain expressive phrases as tree fragments from the test images. Then, new descriptions are composed by filtering and merging the fragments of the extraction tree. Mason [23] proposed a nonparametric density estimation (NDE) technique that estimates the frequency of visual content words of the image to be detected and transforms caption generation into an extractive summarization problem. Sun [24] proposed a concept automatic recognition method that uses parallel text and visual corpora. It can filter out text terms by matching the visual characteristics of similar images in the image library and the image to be described. Retrieval-based methods can be more natural language-like, although relying heavily on the capacity of the database makes it difficult to generate sentences for specific images.

In recent years, with the continued advancement of deep learning, neural networks have been extensively used in image caption tasks. Kiros [25] first used deep neural networks and LSTM to construct two different multimodal neural network models in 2014, continuously integrating semantic information to generate words. For the encoding part, they applied an RNN to convert vocabulary into D-dimensional word vectors. The sentence described can be written as matrix $V \; x \; D$, where V is the number of words in a sentence, and D is the size of the word vector. Finally, they used a decoder consisting of LSTM cells to generate the final picture subtitle result word by word with the combination of image features and the language model sentence by sentence. In subsequent research, Xu et al. [26] incorporated attention mechanisms into the encoder and decoder structural models to describe images. By establishing an attention matrix, they can automatically focus on different areas when predicting different words at different times to enhance the description effectiveness of the model. Bo [27] used generative adversarial networks to generate diversified descriptions by controlling random noise vectors.

In contemporary research, Ayoub [28] deployed the Bahdanau attention mechanism and transfer learning techniques for image caption generation. They incorporated a pre-trained image feature extractor alongside the attention mechanism, thus improving captioning quality and precision. Muhammad [29] proposed a model blending the attention mechanism and object features for image captioning, enhancing the model's ability to leverage extracted object features from images. Chun [30] demonstrated an advanced deep

learning approach for image captioning that combined CNN for image feature extraction and RNN for caption generation, enhanced by the attention mechanism. This innovative method facilitated the automated creation of comprehensive bridge damage descriptions. Lastly, Wu [31] addressed the challenge of describing novel objects through a switchable attention mechanism and multimodal fusion approach, resulting in the generation of accurate and meaningful descriptions.

Wang [8] used a Bi-LSTM model to perform image caption. Wang has developed a deep Bi-LSTM model based on this and has achieved good results. Fazin [9] simplified Wang's model by reducing many parameters and improving the efficiency of the model. Unlike the above models, the mapping relationship between vision and language in our Bi-LS-AttM model is reverse-crossed, and the forward and backward attention of visual language are dynamically learned. As shown in Section 4, this has been demonstrated to be extremely beneficial for picture caption and image sentence retrieval.

3. Methodology

This section outlines our proposed model, an enhanced version of the deep Bi-LSTM model for image captioning as proposed by Wang [8] and Fazin [9]. In our design, we replace the RCNN used in Wang and Vahid's model with Fast RCNN to expedite the feature extraction process. Furthermore, we substitute the Bi-LSTM with the Bi-LS-AttM, representing our unique contribution to this study.

Our model framework comprises three components: a Fast RCNN for detecting objects within images; a Bi-LSTM paired with an attention mechanism to provide attentional representation for each word; and a common space to compile all sentences and their respective final scores. The specifics of each module will be elaborated in the subsequent sections.

3.1. Detect Object by Fast RCNN

In this section, we adopt the method proposed by Girshick [32] for feature extraction and recognition. The selective search algorithm is utilized to extract candidate regions from the input image. These regions are then mapped to the final convolutional feature layer based on their spatial positional relationship.

For each candidate region on the convolutional feature layer, RoI pooling is performed to secure fixed-dimensional features. These extracted features are then fed into the fully connected layer for subsequent classification and regression tasks.

Fast RCNN outputs the probability of each category for each candidate region, as well as the calculated position of each candidate box through regression. For each candidate region, the following loss function is calculated:

$$L(p, u, t^u, v) = L_{cls}(p, u)(1 + \lambda[u \geq 1]) \quad (1)$$

where p is the probability of each category belonging to the candidate region and u is the ground truth category. t is the predicted position for each category, and v is the ground truth position for the candidate field. Compared with the previous version of RCNN, Fast RCNN improves the calculation speed, saving the time and cost of object detection.

Fast RCNN combines classification and regression into a common network, enabling consistent training. In particular, its main enhancement on RCNN is that it eliminates the practice of using separate SVM classifiers and bounding regressors, which greatly improves speed.

3.2. Long Short-Term Memory

The LSTM cells form the basis of this work. They are a unique form of RNN able to memorize long-term associations. Figure 2 shows that an LSTM cell is made up of four important components: a memory cell g and three gate circuits (i is the gate of the input, f is the gate of forget, and o is the gate of the output) [9].

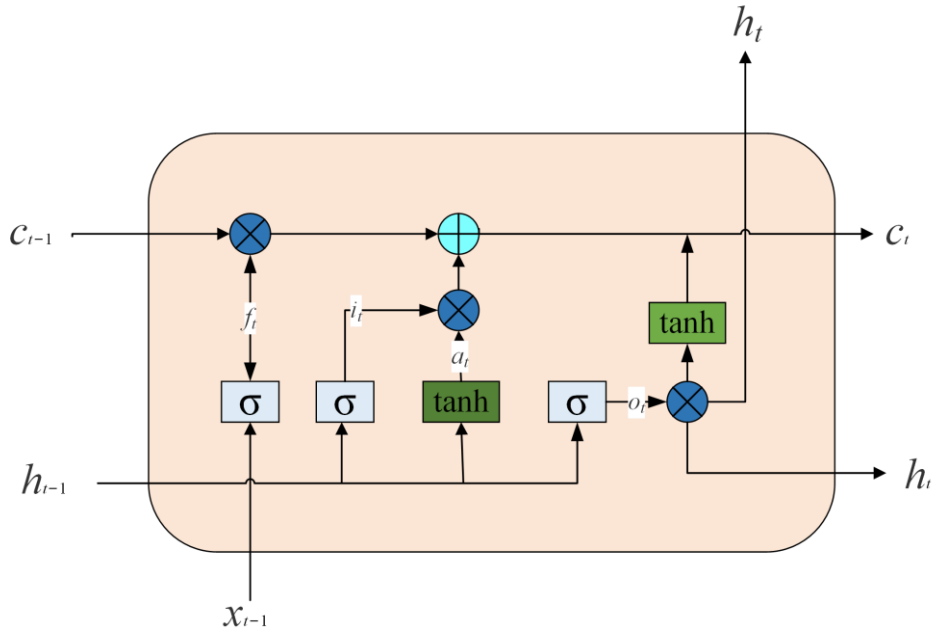


Figure 2. LSTM cell structure.

In the formula below, $f(t)$, $i(t)$, and $o(t)$ are the values of forget, input, and output at time t, respectively. $a(t)$ is the intermediate feature extract result of h_{t-1} and x_t at time t:

$$a(t) = \tanh(W_a h_{t-1} + U_a x_t + b_a) \tag{2}$$

$$i(t) = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{3}$$

$$f(t) = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{4}$$

$$o(t) = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{5}$$

where x_t is the entrance, and h_{t-1} is the hidden state value at time $t-1$. The results calculated by forgetting and inputting operate on the cell state, expressed as the formula below:

$$c(t) = c(t-1) \odot f(t) + i(t) \odot a(t) \tag{6}$$

where $\odot$ represents the Hadamard product. Finally, the hidden state is at t. $h(t)$ is obtained by multiplying the gate output $o(t)$ and the current cell state $c(t)$ using the Hadamard product:

$$h(t) = o(t) \odot tanh(c(t)) \tag{7}$$

The following equation uses the parameters W_s and b_S to predict the next word:

$$F(p_{ti}; W_s, b_S) = \frac{\exp(W_s h_{ti} + b_S)}{\sum_{j=1}^{K} \exp\left(W_s h_{tj} + b_S\right)} \tag{8}$$

where p_{ti} is the tipping probability of the forecast value.

3.3. Bi-LSTM

Both RNN and LSTM units leverage past temporal information to predict forthcoming outputs. However, in some instances, the desired output is associated with not just the previous state but also the future state. For instance, predicting a missing word within a textual context requires comprehension of both the preceding and succeeding context. This dual-

directional context analysis provides a more comprehensive and accurate interpretation, thereby achieving a genuine contextual understanding and decision-making process.

In the traditional LSTM, the forecast of the word x_t using the optical context V and historical context $x_{1:t-1}$ is performed by estimating $\log P(x_t|V, x_{1:t-1})$. However, in the Bi-LSTM with attention, the prediction of the word x_t depends on the forward and backward results of separately maximizing $\log P(x_t|V, x_{1:t-1})$ and $\log P(x_t|V, x_{t+1:t})$ at time t.

In the Bi-LSTM cell structure, the input sequence is processed in both forward and backward directions by two distinct LSTM cells to extract features. As illustrated in Figure 3, the output vectors generated are amalgamated to form the final word representation. The core concept behind the Bi-LSTM cell is to facilitate the capture of features at any given time point, encompassing information from both preceding and succeeding time steps. It is worth noting that the two LSTM units within the Bi-LSTM cell operate with independent parameters while sharing a common word-embedding vector space.

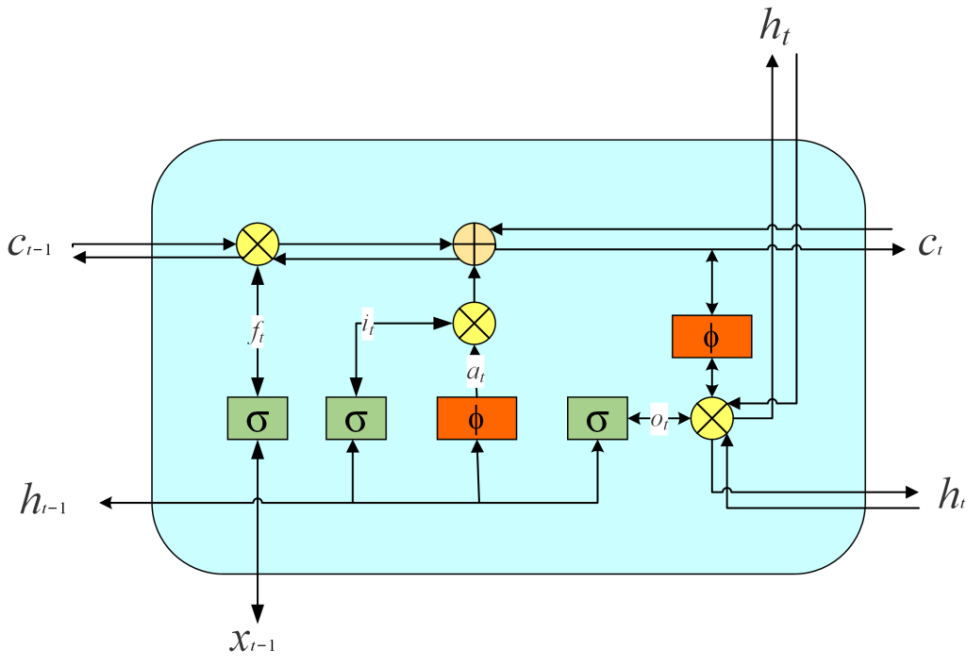


Figure 3. Bi-LSTM cell structure.

3.4. Architecture Model

The general layout of the model is illustrated in Figure 4. It is mainly composed of three modules: Fast RCNN for encoding image input, Bi-LS-AttM for encoding sentence input, and embedding picture and caption into common space and decoding it into image captions and evaluation scores.

The Bi-LS-AttM generates word vectors by comparing similarity using the context information from the frontend and the backend. More accurate words are selected after passing by attention. In our work, the model calculates the front hidden vector $\overrightarrow{h}$ and the back hidden vector $\overleftarrow{h}$. The front cell starts at $t = 1$, while the back cell starts at $t = T$. Our framework works such that for an initial input frame I, the encoding is performed as follows:

$$I_t = F(I; \theta_m) \tag{9}$$

$$\overrightarrow{h_t^1} = B\left(\overrightarrow{M}; \theta_n\right) \tag{10}$$

$$\overleftarrow{h_t^1} = B\left(\overleftarrow{M}; \theta_n\right) \tag{11}$$

where F and B represent Fast RCNN and Bi-LSTM, respectively. θ_m and θ_n are their corresponding weight coefficients. $\overrightarrow{M}$ and $\overleftarrow{M}$ are forward and backward vectors learned from the neural network, respectively. Afterward, the obtained vectors $\overrightarrow{h}$ and $\overleftarrow{h}$ are input into attention. The bilinear scoring procedure is applied to calculate the correlation between the query q and $\overrightarrow{h}$ and $\overleftarrow{h}$. Next, a SoftMax is applied to these scores to normalize them and obtain the attention distribution $a = [a_1, a_2, \ldots, a_t]$. The bilinear scoring function and SoftMax are defined as follows:

$$s\left(\overrightarrow{h_t^1}, \overleftarrow{h_t^1}, q\right) = \overrightarrow{h_t^1}^T \overleftarrow{h_t^1}^T W q \tag{12}$$

$$a_t = \frac{\exp\left(s\left(\overrightarrow{h_t^1}, \overleftarrow{h_t^1}, q\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(\overrightarrow{h_j^1}, \overleftarrow{h_j^1}, q\right)\right)} \tag{13}$$

where W is a trainable parameter matrix. s is a bilinear function.

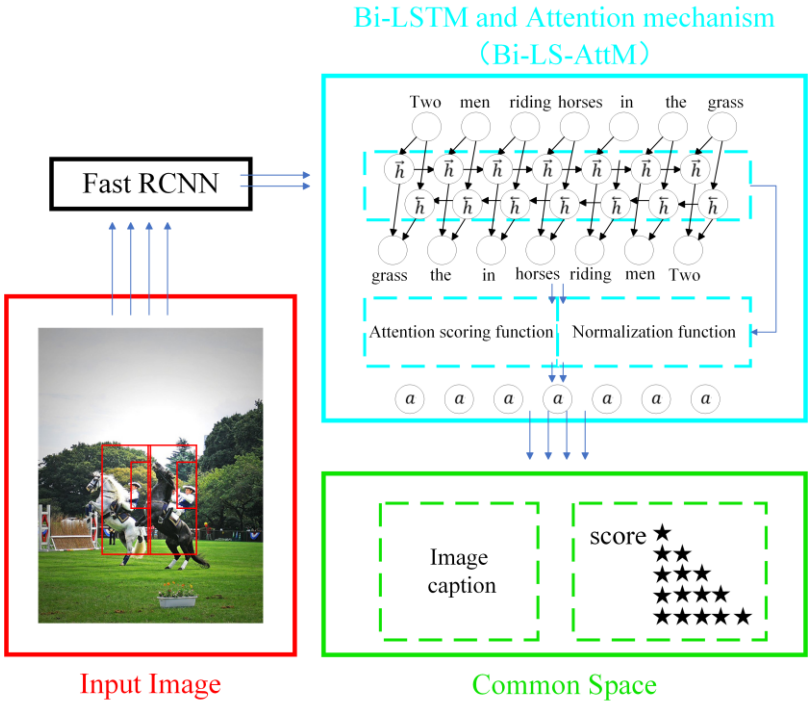


Figure 4. Proposed model architecture.

After training the model, it can predict the word x_t with a given image context V and forward word context $x_{1:t-1}$, predicted either in a forward direction using $P(x_t|x_{1:t-1}, V)$ or in a backward direction using $P(x_t|x_{t+1:t}, V)$. For both forward and backward directions, $x_1 = x_T = 0$ is set at the starting point. Finally, for sentences generated from both directions, the last sentence of the given image $P(x_{1:T}|V)$ is determined by the sum of all words' probabilities in the caption:

$$p(x_{1:T} \mid V) = max\left(\sum_{t=1}^{T}\left(\overrightarrow{p}(x_t \mid V)\right), \sum_{t=1}^{T}\left(\overleftarrow{p}(x_t \mid V)\right)\right) \tag{14}$$

$$\overrightarrow{p}(x_t \mid V) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}, V) \tag{15}$$

$$\overleftarrow{p}(x_t \mid V) = \prod_{t=1}^{T} p(x_t \mid x_{t+1}, x_{t+2}, \ldots, x_T, V) \tag{16}$$

The Bi-LSTM module and its training parameters are similar to those presented in Wang [8]. The difference is that an attention mechanism is added to it. It can focus more on comparing the forward and backward context information to obtain the attention distribution. When extracting features, the Fast RCNN is more efficient and saves time.

4. Experiments

4.1. Dataset

Experiments were performed to validate the effectiveness, generality, and robustness of the model compared to other methods on two datasets, Flickr30K [33] and MSCOCO [34]:

Flickr30K. This is an extended version of Flickr8K. The dataset can be accessed via the following link: http://shannon.cs.illinois.edu/DenotationGraph (accessed 2 May 2023). It contains 31,783 images, each with 5 captions. The dataset does not explicitly categorize the images into different types or categories. We followed the dataset partitioning proposed by Karpathy [4]. In this split of the dataset, 29,000/1000/1000 pictures were utilized for training, validation, and testing, respectively.

MSCOCO. The dataset can be accessed via the following link: https://cocodataset.org (accessed 2 May 2023). This dataset, published several years ago, includes 82,783 training, 40,504 validation, and 40,775 test images. The dataset contains 80 different object categories. Five sentences are annotated for each frame. The focus is on describing all important parts of the scene rather than unimportant details. In the absence of a standard classification, we follow the classification of Wang et al. [8], which uses 80,000 images to train and 5000 to validate and test.

4.2. Evaluation Metrics

The evaluation methods of machine translation can be referred to as the evaluation criteria, which match the generated sentences with human descriptions to obtain a similarity score to measure the accuracy of the task.

For caption generation, the previous work is continued, and the BLEU-N score [35] is used:

$$U(d, s) = b(d, s) \exp\left[\sum_{z=1}^{Z} k \log p(d, s)\right] \tag{17}$$

where d represents the candidate description, which is the reference description, b is the penalty function, k represents the probability of selecting a specific caption, and p represents the accuracy measurement function. Comparing the results of METEOR [36] and CIDEr [37], METEOR can overcome the inherent deficiencies of the BLUE standard, while CIDEr computes the closeness of reference and modeled descriptions as the evaluation standard.

In the retrieval of an image-sentence, we use R@K and Medr as assessment scores. R@K is the recall rate of the top captions. Medr is the average score of the first basic fact image and caption retrieved.

4.3. Implementation Details

During the image encoding process, we utilize the VggNet model [38] for pre-training and employ Fast RCNN to obtain the features from the final fully connected layer. This allows Fast RCNN to share features and parameters in the feature extraction and RoI pooling stages, thereby enhancing processing efficiency. The Bi-LS-AttM is deployed for training the language module. In addition, we selected the widely used and enhanced VggNet [38] and GoogleNet [39] models for our experiments. We tested our model on two datasets specifically designed for image captioning: Flickr30k and MSCOCO.

The server hardware configuration was as shown below: Intel(R) Core (TM) i5-6200U 2.30 GHz CPU, NVIDIA RTX3080Ti GPU, and Win10 OS. The respective version levels

needed for the software are Python 3.9, Torch 1.13.1, Scipy 1.2.1, H5py, and Tqdm. The parameters set for models are shown in Tables 1–3.

Table 1. Model parameters.

Parameters	Descriptions	Value
Emb-dim	Dimension of word embeddings.	256
Attention-dim	Dimension of Bi-LSTM attention linear layers.	256
Frcnn-dim	Dimension of fast RCNN.	128
Dropout	The phenomenon of learning to adapt can be considerably reduced by training batches.	0.5
Cudnn-benchmark	Set to true only if inputs to model are fixed size; otherwise, lots of computational overhead.	True

Table 2. Bi-LSTM attention parameters.

Parameters	Descriptions	Value
Lay-Num	Number of layers	3
Time-fore	Time of data used to make the forecast.	5
Hidden-size	The size of the featured area in the hidden status.	8
Epoch	Displaying the number of forward and backward calculations that could be performed.	120
Batch-size	The amount of data transferred to the trainer.	0.5
Learning-rate	Adjustment of the network weighting rate by the loss gradient.	0.002
Optimizer	Training and parameter tuning to reduce the loss of function.	Adam

Table 3. Training parameters.

Parameters	Descriptions	Value
Lay-Num	Number of layers.	3
Hidden-dim	Size of the featured area in the hidden status.	8
Epoch	Displaying the number of forward and backward calculations that could be performed.	120
Batch-size	The amount of data transferred to the trainer.	64
Learning-rate	Adjustment of the network weighting rate by the loss gradient.	0.004
Optimizer	Training and parameter tuning to reduce the loss function.	Adam
Grad-clip	Clipping to threshold, and gradient to update weights.	5
Alpha-c	Regularization parameter for attention mechanisms.	1

All words in the caption are taken from the vector used to generate the caption. Words appearing less than five times in the training set are marked and removed. A vocabulary of 7200 and 8600 words is provided for the Flickr30K and MSCOCO datasets, respectively. Additionally, 048 Bi-LSTM hidden units are used, and the initialization range of the weight coefficients is set to $[-0.06, 0.06]$.

4.4. Experimental Results on the Generated Image Caption

Our image captioning model's efficacy was evaluated through comparative experiments utilizing the BLUE-N metric, with the resultant data exhibited in Table 4. The additional attention layer implemented within our model contributed significantly to its strong performance on both evaluated datasets. Substituting AlexNet with VggNet [40] resulted in substantial performance improvements across all BLUE metrics. Our model ranks predominantly within the top two positions across these metrics. While our model lags marginally behind the top-rated Hard-attention [24] model in the B-1 metric, it surpasses the performance of both the Bi-LSTM and deep Bi-LSTM models in all other assessed metrics.

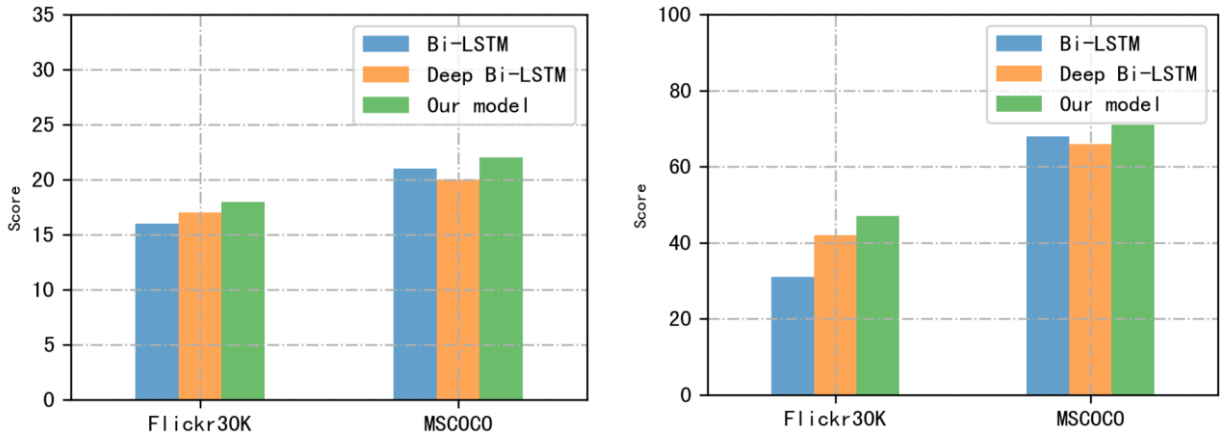
Figure 5 illustrates the comparison of our model with others on the METEOR and CIDEr metrics. We compared our model with the Bi-LSTM [9] and deep Bi-LSTM [8] models. As shown in the figure, the Bi-LS-AttM outperformed the leading-edge methods on the metrics. In the Flickr30K dataset, we improved the METEOR and CIDEr scores by about 8.0% and 12.5%, respectively. In the MSCOCO dataset, our model improved the

METEOR and CIDEr scores by about 6.8% and 15.6%, respectively. We can also speculate that it can give better results on larger datasets.

Table 4. Compare the BLEU score of each model on Flickr30K and MSCOCO.

Model	Flickr30K				MSCOCO			
	B-1	**B-2**	**B-3**	**B-4**	**B-1**	**B-2**	**B-3**	**B-4**
Deep VS [41]	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0
m-RNN(AlexNet) [40]	54.0	36.0	23.0	15.0	-	-	-	-
m-RNN(VGGNet) [40]	60.0	41.0	28.0	19.0	67.0	49.0	35.0	25.0
Hard-attention [26]	**66.9**	43.9	29.6	19.9	**71.8**	50.4	35.7	25.0
Bi-LSTM [9]	62.1	42.6	28.1	19.3	67.2	49.2	35.2	24.4
Deep Bi-LSTM [8]	63.1	44.2	29.7	20.0	67.5	49.0	35.5	24.9
Our model (Bi-LSTM and attention mechanism)	64.5	**44.6**	**29.8**	**20.2**	68.8	**51.0**	**35.9**	**25.2**

The results shown in bold type are the best values.



(**a**) METEOR (**b**) CIDEr

Figure 5. (**a**) Comparison of METEOR scores of three models on two benchmark datasets; (**b**) comparison of CIDEr scores of three models on two benchmark datasets.

We performed a comparative evaluation of our model against other advanced models on METEOR and CIDEr metrics, as depicted in Table 5. This analysis reveals that our Bi-LS-AttM model demonstrates robust competitiveness within the MSCOCO Karpathy split dataset. Particularly, on the METEOR score, our model trails slightly behind Wu's model, potentially due to their superior parameter optimization techniques and the dataset's compatibility with their switchable novel object captioner. This calls for further investigations into novel datasets. However, our model excels in the CIDEr score. A considerable score variation is evident when comparing the performance of our model with and without the attention mechanism.

Table 5. Comparison of the METEOR and CIDEr scores between our and the state-of-the-art models.

Model	METEOR	CIDEr
Muhammad's model [29]	16.3	39.0
Wu's model (SNOC) [31]	**21.9**	39.5
Our model (Bi-LSTM)	17.8	34.5
Our model (Bi-LS-AttM)	21.5	**41.2**

The results shown in bold type are the best values.

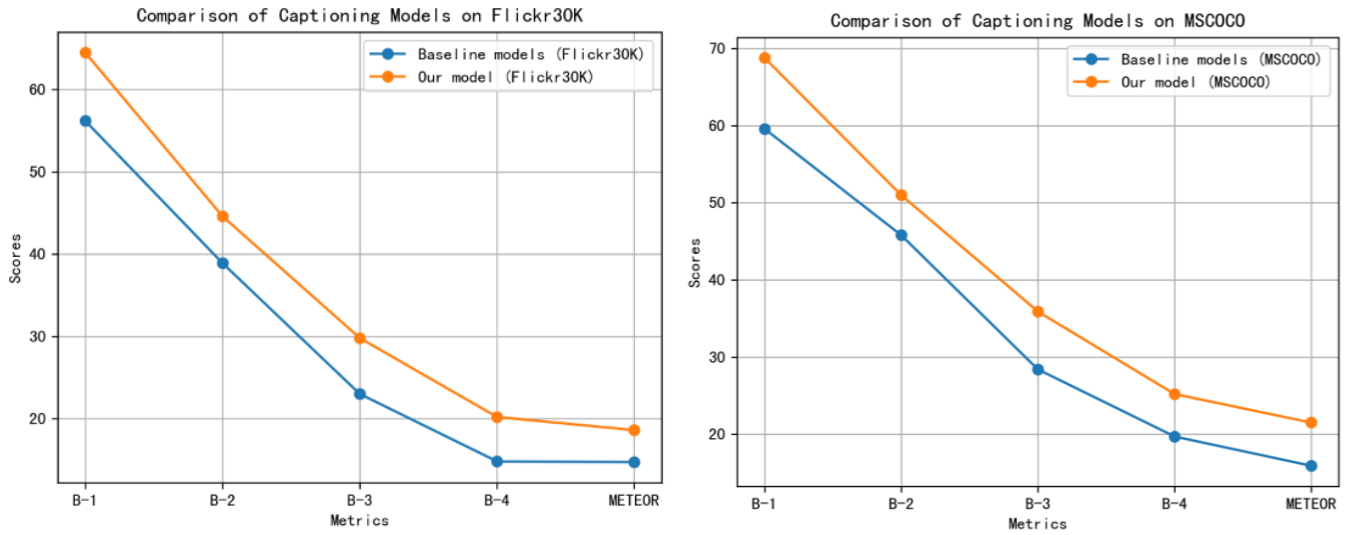
In a similar vein, we conducted a thorough comparison between the performance scores of the baseline model, which lacks an attention mechanism, and our proposed model, employing diverse evaluation metrics. The results presented in Table 6 clearly depict the

relative performance of the baseline model in contrast to our model. Remarkably, our model consistently surpasses the baseline model in terms of performance metrics on both the MSCOCO and Flickr30k test sets. Furthermore, the line graph depicted in Figure 6 visually demonstrates the competitive advantage of our model across various evaluation metrics. It is noteworthy that our model exhibits a more pronounced competitive edge, particularly on the MSCOCO test set. We attribute this observation to the larger scale of the MSCOCO test set, enabling a more comprehensive evaluation and, subsequently, yielding higher performance scores.

Table 6. Performance scores of the baseline model and our model across various metrics.

Model	Flickr30K					MSCOCO				
	B-1	**B-2**	**B-3**	**B-4**	**METEOR**	**B-1**	**B-2**	**B-3**	**B-4**	**METEOR**
Baseline models (without attention)	56.2	38.9	23.0	14.8	14.7	59.6	45.8	28.4	19.7	15.9
Our model (Bi-LS-AttM)	**64.5**	**44.6**	**29.8**	**20.2**	**18.6**	**68.8**	**51.0**	**35.9**	**25.2**	**21.5**

The results shown in bold type are the best values.



(**a**) Flickr30 (**b**) MSCOCO

Figure 6. Figure 6. (**a**) Comparison of Captioning Models on Flickr30K; (**b**) Comparison of Captioning Models on MSCOCO.

4.5. Experimental Results on the Retrieval of Image-Sentence

In assessing image-sentence retrieval, we primarily focused on retrieval scores. Table 7 presents the R@K and Medr scores from our model's image-sentence retrieval across various datasets. Generally, our model surpasses previous methodologies on most metrics, with a particularly strong performance on the MSCOCO dataset. Notably, the Bi-LS-AttM outstrips advanced models in both image-to-sentence and sentence-to-image retrieval tasks. However, some metrics reveal suboptimal performance; for instance, the Mind's Eye model [42], which effectively integrates image and text features, outperforms our model on the Flickr30K dataset. We posit that incorporating an adaptive attention mechanism could enhance efficiency in image-sentence retrieval tasks, a hypothesis we aim to investigate in future research.

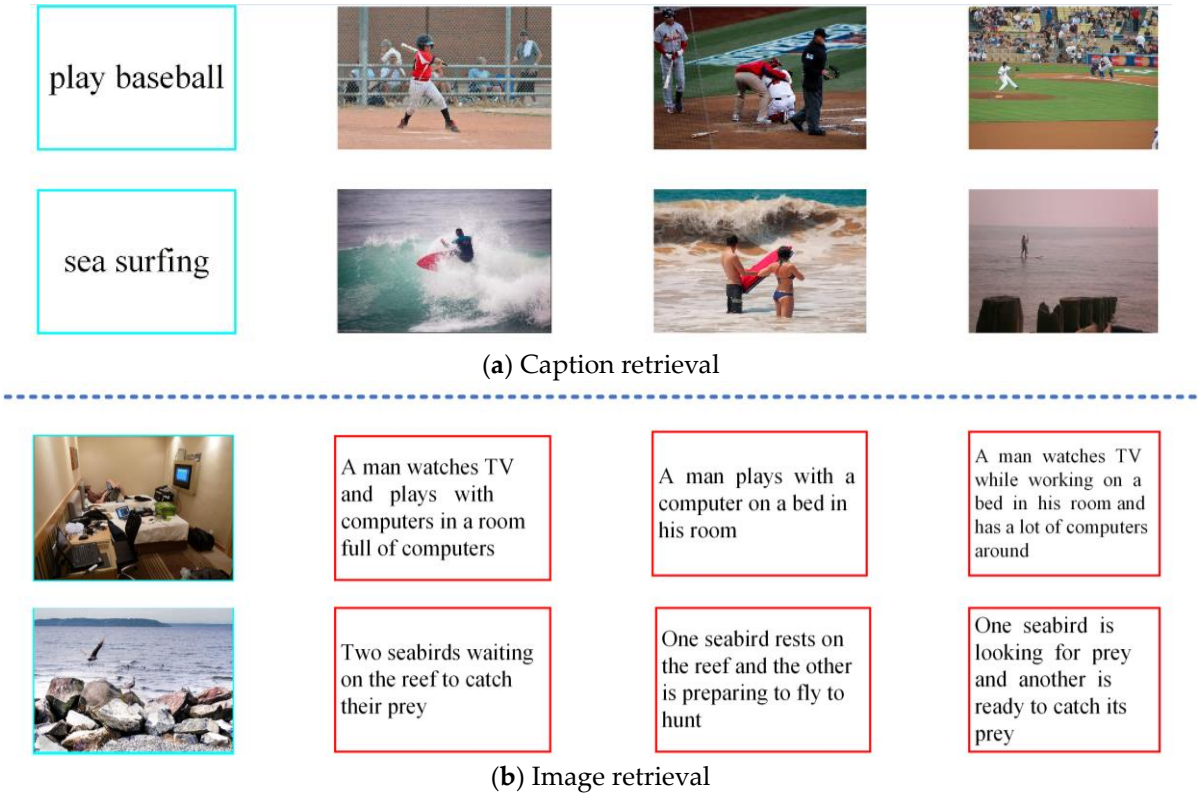
We conducted additional experiments to authenticate our model's performance in image-sentence retrieval tasks. Figure 7 presents examples from several retrieval experiments on the MSCOCO validation set. In each caption query, the model retrieves visually congruent images and captions, illustrating its proficiency in discerning the visual–textual association in image caption rankings. The upper dashed line represents image

retrieval predicated on keywords, while the lower one symbolizes sentence retrieval based on images.

Table 7. R@K (a high score is good) and Medr (a low score is good) comparison of each model on Flickr30K and MSCOCO.

Dataset	Model	Image to Sentence				Sentence to Image			
		R@1	R@2	R@3	Medr	R@1	R@2	R@3	Medr
Flickr30k	Deep VS [41]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
	m-RNN(AlexNet) [40]	18.4	40.2	50.9	10.0	12.6	31.2	41.5	16.0
	Mind's Eye [42]	18.5	45.7	58.1	7.0	16.6	42.5	**58.9**	8.0
	Bi-LSTM [9]	28.1	53.1	64.2	4.0	19.6	43.8	55.8	7.0
	Deep Bi-LSTM [8]	29.2	**54.0**	64.9	3.8	20.8	44.5	56.7	6.7
	Our model (Bi-LSTM and attention mechanism)	**29.5**	53.8	**65.0**	**3.6**	**21.0**	**44.7**	57.3	**6.5**
MSCOCO	Deep VS [41]	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0
	m-RNN(AlexNet) [40]	12.4	29.3	48.6	15.0	9.5	25.4	38.2	18.0
	Mind's Eye [42]	12.8	35.6	50.1	11.0	11.6	33.7	48.5	10.0
	Bi-LSTM [9]	13.4	33.1	44.7	13.0	9.4	26.5	37.7	19.0
	Deep Bi-LSTM [8]	16.6	39.4	52.4	9.0	11.6	30.9	43.4	13.0
	Our model (Bi-LSTM and attention mechanism)	**17.2**	**40.0**	**53.6**	**8.0**	**12.2**	**35.6**	**49.8**	**11.0**

The results shown in bold type are the best values.



(**a**) Caption retrieval

(**b**) Image retrieval

Figure 7. Example of using our model for image retrieval and caption retrieval on the MSCOCO validation set. (**a**) To search for three images using captions. (**b**) To search for three captions using images.

In the sentence-image retrieval task, we generated three images that mirror the keywords and sentences and then selected the matching image based on its similarity. For the image-sentence retrieval task, we produced three appropriate captions grounded on the im-

ages and then chose the generated sentences based on their high scores in the shared space. The given examples underscore the efficiency of our model in executing image-sentence retrieval tasks.

4.6. Discussion

Effect of Bi-LS-AttM: To gauge the influence of the Bi-LS-AttM, we compared the computational time of the Bi-LS-AttM model with the Bi-LSTM and deep Bi-LSTM models for caption creation and image-to-sentence retrieval tasks. Table 8 delineates the computational durations of these models for the respective tasks. We randomly selected 20 images from the Flickr30K validation set and assessed each model ten times for caption creation and image-to-sentence retrieval. The table provides the average time duration across the ten trials, excluding model initialization and training time.

Table 8. The cost of checking 10 images on Flickr30K.

Task	Bi-LSTM	Deep Bi-LSTM	Our Models
Caption creation	1.04s	1.07 s	**0.67 s**
Image-Sentence Retrieval	5.78 s	5.81 s	**4.28 s**

The results shown in bold type are the best values.

The caption generation time expenses encompass the extraction of image features, bidirectional caption content sampling, computation of the final caption result, and caption accuracy evaluation. Conversely, the retrieval time accounts for the computation of the image-to-sentence retrieval score, image and sentence query, and sorting operations in descending order. By employing the Fast R-CNN framework and fine-tuning the relevant parameters, our model demonstrates significant time savings in accomplishing the given task. From Table 8, we can see that our model saves about 36.5% and 26.3% of the time compared to the Bi-LSTM and deep Bi-LSTM models, respectively, in solving image captioning and image-sentence retrieval tasks. We have verified the efficiency of the Bi-LS-AttM.

Effect of image caption: We used the Bi-LS-AttM model to generate real, accurate, and novel image descriptions. Figure 8 shows the comparison of the baseline model and our model in generating captions on the datasets. We evaluated generated captions from various perspectives. In some descriptions, the relationships between objects are well expressed (e.g., "A hot dog and a red bottle of drink are on the table"). In the example above, the objects 'hot dog' and 'table' are accurately identified, and the relationship between them is established. Finally, the image is described accurately and in a novel way (e.g., "A boy dressed in black surfs the sea with a red surfboard."). However, the baseline model solely provides descriptive accounts of the images, lacking the generation of novel and expressive sentences to depict them. From the perspective of object detection, the object recognized in the baseline model is "bread" rather than the more accurate "hot dog". The results of the studies show our model has good efficiency. Our model can achieve a balance between performance and efficiency.

Examples of failed experiments: Figure 9 depicts a notable number of anomalies arising from our experimental approach. It is pertinent to note that these inaccuracies primarily originate from the Flickr30K validation set, which we hypothesize may be due to the limited range and diversity present in the training dataset of this source. For instance, in the preceding images, our model exhibits imprecision in object identification (i.e., identifying "white clothes" when the man is not clothed). Another example demonstrates an illogical caption suggesting a man cycling on water. We surmise that these limitations could be ameliorated through improvements in our visual feature extraction aspect.

Despite these anomalies, we see them as potential avenues for further research rather than setbacks. Nonetheless, it is crucial to emphasize that a substantial number of remaining cases were accurately represented, as exemplified in Figure 8.

Baseline model: A giraffe is resting at a table.
Our model: Three giraffes bend their heads eating feed in the zoo.

Baseline model: A child holds a plank at sea.
Our model : A boy dressed in black surfs the sea with a red surfboard.

Baseline model: A man taking a photo with a woman and a child.
Our model: A family of three wearing ski suits taking pictures at the ski resort.

Baseline model: Three girls are playing computer.
Our model: Three girls are happily learning information in the computer.

Baseline model: A man in a hat is playing on the grass.
Our model: A man in colorful clothes plays a frisbee on the grass.

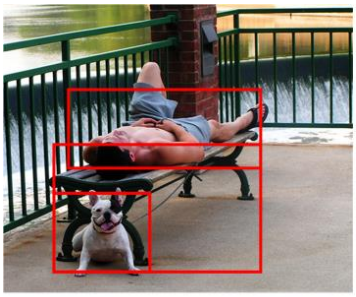
Baseline model: A glass of orange juice and bread on the table.
Our model: A hot dog and a red bottle of drink are on the table.

Baseline model: A group of people dancing at night.
Our model: A group of people walking in the park in the evening.

Baseline model: A brown dog running on grass.
Our model: A black and white puppy runs on a fenced grass field.

Baseline model: A child in a red dress is playing a game console.
Our model: A child concentrates on playing a game console in an amusement park.

Figure 8. Examples of image captioning for the baseline and our model on the datasets. The captions generated by the baseline model are above, while the captions generated by our model are below.

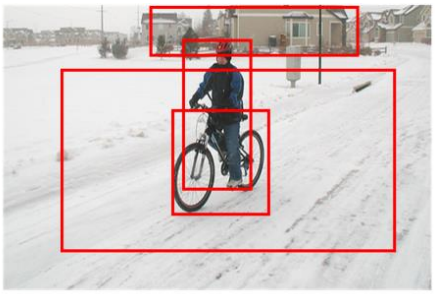


A man in white clothes sleeps with a puppy in bed

(**a**) Feature extraction error

Two little kids sitting on a stool and playing games outside

(**b**) Image representation error



A well-groomed man rides a bike on the water

(**c**) Caption logic error

Figure 9. Examples of failed experiments: (**a**) feature extraction error, (**b**) Image representation error, (**c**) caption logic error. We mark the extracted features on the image with red boxes and use blue fonts to distinguish errors.

5. Conclusions

In this study, we have introduced a model that leverages the capabilities of the Bi-LS-AttM approach to generate captions that are precise, inventive, and context-sensitive. This was accomplished by incorporating bidirectional information and an attention mechanism. For the dual purposes of feature extraction and time optimization, we utilized the Fast RCNN. Additionally, to provide a comprehensive understanding of the proposed model's structure, we generated a detailed visualization outlining the word generation process at consecutive timesteps. The model's robustness and stability were thoroughly assessed across various datasets pertinent to image captioning and image-sentence retrieval tasks.

In terms of future work, we intend to delve into more intricate domains of image captioning, including those related to remote sensing and medical imaging. We anticipate broadening the application scope of our model to encapsulate other forms of captioning tasks such as video captioning. Furthermore, we plan to explore the integration of multi-task learning methodologies with an aim to enhance the model's general applicability.

Author Contributions: Conceptualization, T.X. and J.W.; methodology, J.W. and T.X.; software, T.X.; validation, T.X., W.D. and J.W.; formal analysis, J.W.; resources, J.W.; data curation, T.X. and X.W.; writing—original draft preparation, T.X.; writing—review and editing, T.X.; visualization, W.D. and J.W.; supervision, J.Z. and J.W.; project administration, T.X; funding acquisition, W.D. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61976120, and the Basic Science Research Project of Nantong, grant number JC2020143.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and required datasets of the experiments can be obtained upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The abbreviations used in this manuscript are listed in the following tables:

CV	Computer Vision
NPL	Natural Language Processing
Bi-LSTM	Bidirectional Long Short-Term Memory
LSTM	Long Short-Term Memory
Bi-LS-AttM	Bidirectional LSTM and Attention Mechanism
Fast RCNN	Fast Region-based Convolutional Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
RoI	Region of Interest
SVM	Support Vector Machine
CRF	Conditional Random Field
NDE	Nonparametric Density Estimation
VggNet	Visual geometry group Net
Deep VS	Deep Visual Semantic
m-RNN	Multimodal Recurrent Neural Network
BULE	Bilingual Evaluation Understudy
MSCOCO	Microsoft Common Objects in Context
METEOR	Metric for Evaluation of Translation with Explicit Ordering
CIDEr	Consensus-based Image Description Evaluation
R@K	Recall@K

References

1. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250.
2. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural Baby Talk. In Proceedings of the Name of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6566–7296.
3. Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; Li, L.-J. Deep reinforcement learning-based captioning with embedding reward. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1151–1159.
4. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H.A. Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* **2019**, *51*, 118. [CrossRef]
5. Yu, N.; Hu, X.; Song, B.; Yang, J.; Zhang, J. Topic-Oriented Image Captioning Based on Order-Embedding. *IEEE Trans. Image Process.* **2019**, *28*, 2743–2754. [CrossRef]
6. Jiang, W.; Zhu, M.; Fang, Y.; Shi, G.; Zhao, X.; Liu, Y. Visual Cluster Grounding for Image Captioning. *IEEE Trans. Image Process.* **2022**, *31*, 3920–3934. [CrossRef]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
8. Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 988–997.
9. Vahid, C.; Fadaeieslam, J.; Yaghmaee, F. Improvement of image description using bidirectional LSTM. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 147–155.
10. Ahmed, S.; Saif, A.; Hanif, M.; Shakil, M.; Jaman, M.; Haque, M.; Shawkat, S.; Hasan, J.; Sonok, B.; Rahman, F. Att-BiL-SL: Attention-Based Bi-LSTM and Sequential LSTM for Describing Video in the Textual Formation. *Appl. Sci.* **2022**, *12*, 317. [CrossRef]
11. Cho, S.; Oh, H. Generalized Image Captioning for Multilingual Support. *Appl. Sci.* **2023**, *13*, 2446. [CrossRef]
12. Guo, L.; Liu, J.; Lu, S.; Lu, H. Show, Tell, and Polish: Ruminant Decoding for Image Captioning. *IEEE Trans. Multimed.* **2020**, *22*, 2149–2162. [CrossRef]
13. Zhang, L.; Zhang, Y.; Zhao, X.; Zou, Z. Image captioning via proximal policy optimization. *Image Vis. Comput.* **2021**, *108*, 104126. [CrossRef]
14. Wang, J.; Xu, W.; Wang, Q.; Chan, A. On Distinctive Image Captioning via Comparing and Reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2088–2103. [CrossRef]
15. Farhadi, A.; Hejrati, S.; Sadeghi, M.; Young, P.; Forsyth, D. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 15–29.
16. Chen, C.; Lin, C.; Schlkopf, B. A tutorial on ν-support vector machines. *Appl. Stoch. Model. Bus. Ind.* **2005**, *21*, 111–136. [CrossRef]
17. Chang, C.; Lin, C. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
18. Li, S.; Kulkarni, G.; Berg, T.; Berg, A.; Choi, Y. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland Oregon, OR, USA, 23–24 June 2011; pp. 220–228.
19. Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Berg, T. Baby talk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef]
20. Yuan, Q.; Szummer, M.; Minka, T. Bayesian conditional random fields. In Proceedings of the 10th International Conference on Artificial Intelligence and Statistics, Bridgetown, Barbados, 6–8 January 2005.
21. Sutton, C.; Rohanimanesh, K.; McCallum, A. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *J. Mach. Learn. Res.* **2007**, *8*, 693–723.
22. Kuznetsova, P.; Ordonez, V.; Berg, T.; Choi, Y. Treetalk: Composition and compression of trees for image descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 351–362. [CrossRef]
23. Mason, R.; Charniak, E. Nonparametric Method for Data-driven Image Captioning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MA, USA, 23–24 June 2014; pp. 592–598.
24. Sun, C.; Gan, C.; Nevatia, R. Automatic concept discovery from parallel text and visual corpora. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2596–2604.
25. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal Neural Language Models. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.
26. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
27. Bo, D.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2989–2998.
28. Ayoub, S.; Gulzar, Y.; Reegu, F.A.; Turaev, S. Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. *Symmetry* **2022**, *14*, 2681. [CrossRef]

29. Muhammad, A.; Jafar, A.; Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. *J. Big Data* **2022**, *9*, 1–16.
30. Chun, J.; Yamane, T.; Maemura, Y. A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 1387–1401. [CrossRef]
31. Wu, Y.; Jiang, L.; Yang, Y. Switchable Novel Object Captioner. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1162–1173. [CrossRef]
32. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
33. Plummer, B.; Wang, L.; Cervantes, C.; Caicedo, J.; Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
34. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef]
35. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
36. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the 9th Workshop on statistical machine translation (WMT 2014), Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
37. Vedantam, R.; Zitnick, C.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
38. Ke, H.; Chen, D.; Li, X.; Tang, Y.; Shah, T.; Ranjan, R. Towards brain big data classification: Epileptic eeg identification with a lightweight vggnet on global mic. *IEEE Access* **2018**, *6*, 14722–14733. [CrossRef]
39. Muthiah, M.; Logashamugam, E.; Nandhitha, N. Performance evaluation of googlenet, squeezenet, and resnet50 in the classification of herbal images. *Int. J. Eng. Trends Technol.* **2021**, *69*, 229–232.
40. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv* **2014**, arXiv:1412.6632.
41. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
42. Chen, X.; Zitnick, C. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2422–2431.