Research Article

Research on the Image Description Algorithm of Double-Layer LSTM Based on Adaptive Attention Mechanism

Cifeng Qin, Wenyin Gong, and Xiang Li

School of Computer Science, China University of Geosciences, Wuhan 430078, China

Correspondence should be addressed to Xiang Li; lixiang@cug.edu.cn

Received 21 December 2021; Revised 21 April 2022; Accepted 3 May 2022; Published 21 May 2022

Academic Editor: Paolo Spagnolo

Copyright © 2022 Cifeng Qin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Image text description is a multimodal data processing problem in the computer field, which involves the research tasks of computer vision and natural language processing. At present, the research focus of image text description task is mainly on the method based on deep learning. The work of this paper is mainly focused on the imprecise description of visual words and nonvisual words in the description of image description tasks in the image text description. An adaptive attention double-layer LSTM (long short-term memory) model based on coding-decoding is proposed. Compared with the algorithm based on the adaptive attention mechanism based on the coding-decoding framework, the evaluation index BLEU-1 is improved by 1.21%. The METEOR was 0.75% higher and CIDEr was 0.55%, while the indexes of BLEU-4 and ROUGE-L were not as good as those of the original model, but the index was not different. Although it cannot surpass all the performance indicators of the original model, the description of visual words and nonvisual words is more accurate in the actual image text description.

1. Introduction

With the rapid development of the Internet in recent years, massive information is uploaded to Google, Facebook, TikTok, Sina Weibo, and other large Internet websites every day. Aiming at the process, the PB-level massive text, pictures, video, and audio information generated has become a key issue worthy of research in the era of big data. For different formats of data, there are different research methods such as for natural scene text [1], there are text recognition [2], text detection [3], automatic summarization, natural language processing, and other tasks; for image data processing, there are computer vision domain tasks such as image semantic segmentation, target recognition [4], and salient target detection; for audio data, there are speech recognition, music information retrieval [5], environmental sound recognition [6], etc. However, the algorithms mentioned above are generally based on single-mode data processing methods. At this stage, multimode data processing problems involving multiple computer fields are proposed, such as image text description [7], video text description [8], autopilot [9], and other research tasks of computer vision and natural language processing. The text description task of natural language processing images is to automatically generate the content description sentences of images. Image text description is the cornerstone of the field of multimodal data processing, and it also has extraordinary value in real life. The problem of image text description is a cross-cutting research topic. It is a comprehensive problem involving the two fields of computer vision and natural language processing. The input is an image, and the output is a text description of the content of the image. Image text description is the conversion between two different types of information. Compared with other image tasks, image text description tasks require a deeper understanding of images. It not only needs to accurately identify each object in the image but also describe multiple objects. The mutual relationship. How to make the computer quickly analyze the image data and get the semantics based on the image is of great significance. For example, for the visually impaired, the
external image information can be converted into voice text information to help them better understand the surrounding environment; in the field of video monitoring, the automatic detection of abnormal objects in the video can be used to understand, which can be used to monitor the abnormal highway conditions for real-time warning; in the field of automatic driving, the driving environment around the vehicle can be understood in real time.

Compared with the ability of a human to recognize the content in the image and the ability of logical reasoning between the content, the ability of the computer to understand the image content is far from enough. With the increasing demand for image understanding, how to make the computer understand image content has become a research hotspot in the field of artificial intelligence and machine learning. For the problem of image description, how to generate the correct natural language to describe the text content is a very challenging problem because it not only needs to classify image objects but also needs to analyze the correlation between object attributes and objects. In addition, the sentences generated based on the above content must conform to a certain description of grammatical structure and semantic correctness, which means that in addition to visual understanding, text generation model is also needed. It is the above factors that lead to the slow progress of image text description, but the rise of deep learning in recent years has brought new opportunities to the content of image text description.

In recent years, thanks to the rapid development of the deep neural network, researchers proposed an image description generation model based on deep learning architecture. Inspired by the task of machine translation, a method based on DCNN (Diffusion Convolutional Recurrent Neural Network) and RNN (Recurrent Neural Network) cascading is proposed, which is called the encoding-decoding framework. Under this framework, the encoding process is to extract the visual features of the target image through DCNN, and the decoding process is to generate descriptive sentences from the obtained visual features through RNN.

While the basic coding-decoding framework has made initial progress in the image description, researchers draw on the application of the attention mechanism in machine translation and propose a coding-decoding algorithm based on the attention mechanism to describe sentence generation tasks. The algorithm introduces the attention mechanism in the decoding stage, combines the words and local visual information of the image generated at the previous moment, dynamically focuses on different regions of the image when the language module generates each word, and generates description sentences. In 2015, Xu et al. [10] first applied the attention mechanism to image description task, including the hard attention mechanism which only uses visual information of a single local region and soft attention mechanism which uses visual information of all local regions of the image. The different results of the two mechanisms in the task are analyzed and discussed. In 2016, Yang et al. [11] proposed an image description algorithm based on review-based attention, which can perform the multistep review and enhance the stability of the model. In 2017, Pedersoli et al. [12] proposed an image description method based on region attention, which associates the image region with RNN dynamic words. In each time step of RNN, the word at the next moment and the image region associated with the word are predicted to generate description sentences. In the same year, Lu et al. [13] proposed an adaptive attention model based on visual sentinel. The words in the description sentence are divided into visual words and nonvisual words. The former depends on the entities in the image, and the latter depends on the language model. In word generation, Visual Sentry is used to select whether the word generation information comes from image visual information or language model. In 2018, Lu et al. [14] proposed a deep learning model NBT that combines template-filled sentences. The NBT model uses Fast R-CNN to extract image features to predict words. If it is a visual word, it is directly filled into the sentence template, which effectively solves the problem of the traditional template filling the sentence. In 2019, Xinwei et al. [15] proposed a new visually dense semantic attention network model VD-SAN (Visual-Densely Semantic Attention Network), which considers the relationship between image attribute prediction tasks and image surface feature extraction tasks. In the same year, Huang et al. [15] proposed the AoA (Attention on Attention) module, which expands the conventional attention mechanism by adding an additional attention to achieve better results in the image text description. Zhou et al. [16] proposed the POS-SCAN (Part-Of-Speech) model in the same year. The POS-SCAN model uses weak supervision on the basis of the SCAN model to train the attention module, which effectively improves the image text. The quality of the description. In 2020, Liu et al. [17] proposed a Chinese image description model NIC-VATP2L, which uses a visual attention mechanism to reduce the semantic deviation of Chinese and English image descriptions, and uses a semantic model to improve the accuracy and precision of text descriptions. Sex. Pan et al. [18] proposed the attention mechanism X-Linear attention for high-order feature interaction. This model can use bilinear fusion technology to mine second-order to high-order feature interaction information between different modalities to enhance cross-modal content understanding. In the same year, Guo et al. [19] proposed the NG-SAN model. The NG-SAN model improved the self-attention model and proposed two new models: normalized self-attention (NSA) and geometrically-aware self-attention (GSA). The concept makes up for the main limitation that the transformer model cannot model the geometric structure of the input object. In 2020, Sammani et al. [20] proposed a visual language training model ETN model. ETN can iteratively polish the generated description, effectively improving the readability of the description sentence. From the above research, it can be shown that the encoding-decoding framework that introduces the attention mechanism can pay attention to different regions in the image when generating words and has achieved good results.

At this stage, the task of image text description has achieved gratifying results, but there are still some problems. When evaluating the generated image description, it is not
only necessary to compare the evaluation indicators but also to evaluate the grammar, context, diversity, and completeness of sentences. When describing image content, the current image description algorithm is not accurate in describing visual words (such as “man,” “boat”) and nonvisual words (such as “is,” “the”) in sentences so that there are grammatical errors in the generated text description, and there is low readability; when describing images of complex scenes, the effect of describing the main content of the image is not accurate, and the description of the main body is incorrect, making the generated description and the original image. The content is quite different, so further research is needed.

1.1. Paper Structure. In the section Improved Adaptive Attention Double-Layer LSTM Model, we introduced the research motivation of the paper, and what kind of requirements led us to propose a double-layer LSTM model based on adaptive attention and the model structure, algorithm flow, and mathematical method of the proposed adaptive attention-based two-layer LSTM model.

In the section Experiment and Result Analysis, we introduce the experimental parameters and environment settings and introduce the comparison results of the experimental results and mainstream image description models on various indicators.

2. Improved Adaptive Attention Double-Layer LSTM Model

2.1. Research Motivation. Image text description task is essentially a problem of the integration of computer vision and natural language processing. The input of the image text description is an image, and the output is the text description of the image content. Image text description is the conversion between two different kinds of information. Compared with other image tasks, image understanding requires a deeper understanding of the image, which not only needs to accurately identify each object in the image but also needs to describe the relationship between multiple objects. At present, most of these problems are based on the attention mechanism of the encoding-decoding framework. Adding the attention mechanism to the encoding-decoding framework can improve the readability of the statement description. At present, most of the algorithms that use the encoding-decoding framework in image description algorithms use a single-layer LSTM model to process image features and text generation at the same time to generate image description sentences. The single-layer LSTM needs to capture visual attention words and describe the relationship between visual words in the decoding process, which leads to more reliance of nonvisual words on semantic information rather than visual information. In the text description generation, nonvisual words will reduce the effectiveness of visual information, resulting in inaccurate expression of visual words and nonvisual words. In order to solve this problem, this paper proposes an adaptive attention double-layer LSTM model based on the original adaptive attention mechanism algorithm. Figure 1 shows the LSTM cell structure.

2.2. Adaptive Attention Double-Layer LSTM Model. In the original coding-decoding framework based on the attention mechanism, the single-layer LSTM model is generally used to process the image context information and image description task, which makes it easy to generate the problem of insufficient expression accuracy of visual words and nonvisual words in the task of generating text descriptions. So in this paper, we design a double-layer LSTM based on the adaptive attention mechanism (AAM-DLSTM).

AAM-DLSTM uses ResNet (Residual Network) 50 to extract the global features and spatial local features of the image. The input of LSTM-1 is the global feature $F$ of the image, the previous context content vector $m_{t-1}$, and the word vector $s_{tE}$, and the output is the hidden layer state of LSTM-1 $h^1_t$ and visual attention sentinel $s_{tE}$. The input of LSTM-2 is the context content vector $m_t$ at the current moment and the hidden layer state $h^1_t$ of LSTM-1, and the output is the hidden layer state $h^2_t$ of LSTM-2, which is generated through multiple time steps. The text description of the image. The pseudo-code implementation of the AAM-DLSTM algorithm is shown in Algorithm 1.

As shown in Figure 2, first input the original image (as shown in box A in Figure 2) and extract the global feature vector and local feature vector of the image through ResNet (as shown in box B in Figure 2). The spatial local feature $V$ is sent to the adaptive attention module to obtain the context content vector $m_{t-1}$ at the current moment (as shown in box C in Figure 2). Taking into account the influence of the word $s_{t-1}$ generated at the previous moment, the word vector $s_{tE}$ of the word $s_{t-1}$ is sent to the visual information module LSTM-1 to obtain dependency information (as shown in box D in Figure 2). $s_{t-1}$ is a word in $S = \{s_{t-1}, s_0, s_1, s_2, ..., s_{LS}\}$ in the generated image description, where $LS$ indicates the length of the generated description sentence, $s_{t-1}$ indicates the start position, and $s_{LS}$ indicates the end. The word vector $s_{tE}$ can be expressed as $s_{tE} = W_{E} * s_{t-1}$, where $W_{E}$ is the word vector matrix. The input $x^1_t$ of the visual information module at the current moment is composed of the word vector $s_{tE}$, the global feature vector $F$, and the context content vector $m_{t-1}$, that is, $x^1_t = [s_{tE}, F, m_{t-1}]$. According to the LSTM principle introduced in the previous section, the hidden layer vector $h^1_t$ and the memory layer vector $c^1_t$ of the
LSTM-1 at time $t$ can be updated as shown in the following formula:

$$ h^1_t, c^1_t = \text{LSTM}_1(x^1_t, h^1_{t-1}, c^1_{t-1}). $$  \hfill (1)

Among them, $h^1_{t-1}$ and $c^1_{t-1}$, respectively, represent the state of the hidden layer and the state of the memory layer of a unit in LSTM-1 at time $t$. In order to obtain the source of information on which nonvisual words are generated, the visual sentinel needs to be output for LSTM-1. The vector $v_{st}$ (shown on the right side of box D in Figure 2) mainly depends on the memory layer state $c^1_t$ at the current moment of LSTM-1, and the calculation method of the $v_{st}$ vector is shown in the following formula:

$$ v_{st} = \sigma(W_{vsa}x^1_t + W_{vsb}h^1_{t-1}) \odot \tanh(c^1_t). $$  \hfill (2)

Among them, $W_{vsa}$ and $W_{vsb}$ are the weight parameters obtained from model training, $\sigma$ is the sigmoid function, $\odot$ is the XOR logic operator, and $\tanh$ is the hyperbolic tangent function. LSTM-2 mainly relies on the visual sentinel vector $v_{st}$ to generate nonvisual words (as shown in box G in Figure 2) so that the readability of the text can be effectively improved when the text is generated, and the update principle of LSTM-1 is used to update the LSTM at time $t$.

The hidden layer state $h^2_t$ and the memory layer state $c^2_t$ of LSTM-2 are expressed as shown in the following formula:

$$ h^2_t, c^2_t = \text{LSTM}_2(x^2_t, h^2_{t-1}, c^2_{t-1}). $$  \hfill (3)

$$ x^2_t = [m_t, h^1_t]. $$  \hfill (4)

In formula (4), $x^2_t$ is the input vector of the text generation module LSTM-2, which is composed of the context content vector $m_t$ and the hidden layer state $h^1_t$ of LSTM-1 (as shown in box F in Figure 2). Similarly, $h^2_{t-1}$ and $c^2_{t-1}$, respectively, represent the state of the hidden layer and the state of the memory layer of a unit in LSTM-2 at the previous moment at time $t$. For the context content vector $m_t$ (as shown in box E in Figure 2), its mathematical expression is shown in the following formula:

$$ m_t = \sum_{i=1}^{k^2+1} \alpha_{t,i} v_{si}. $$  \hfill (5)

In the formula, $\alpha_{t,i}$ represents the attention weight distribution at the current moment, $v_{si}$ is the visual sentinel, and the attention weight distribution is $\alpha_{t,i}$ which is specifically shown in the following formula:
Among them, softmax is the normalized exponential function; \( \text{tanh} \) is the hyperbolic tangent function; \( W_a, W_vsa, \) and \( W_vsb \) are the weight parameter matrices obtained from model training. According to the updated current state \( h_t \), the predicted probability distribution of the word \( s \) generated at the current moment is calculated (as shown in box \( G \) in the figure) as shown in the following formula:

\[
\alpha_{tj} = \text{soft max}(W_a \text{tanh}(W_{vsa}S_j + W_{vsb}h_t^j)).
\]

(6)

Among them, softmax is the normalized exponential function; \( \text{tanh} \) is the hyperbolic tangent function; \( W_a, W_vsa, \) and \( W_vsb \) are the weight parameter matrices obtained from model training. According to the updated current state \( h_t \), the predicted probability distribution of the word \( s \) generated at the current moment is calculated (as shown in box \( G \) in the figure) as shown in the following formula:

\[
p_t = \text{soft max}(W_p h_t^j).
\]

(7)

Among them, \( W_p \) is the weight parameter of the probability distribution calculation. The goal of the decoder optimization is to maximize the sum of the log-likelihoods of all training samples, as shown in the following formula:

\[
\theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|I; \theta).
\]

(8)

Among them, \( \theta \) is the total parameter of model training, \( \theta^* \) is the total parameter after model optimization, \( I \) is the image in the input training set, and \( S \) is the sentence marked with the training image. The loss function of the decoder is to minimize the cross-entropy (CE) loss function, as shown in the following formula:

\[
\log p(S|I) = -\sum_{t=0}^{LS} \log p(s_t|I, s_0, ..., s_{t-1}; \theta).
\]

(9)

In the formula, \( LS \) is the length of the sentence. When the generated sentence reaches the maximum generated sequence length, the decoding is stopped, and the complete image description sentence is finally obtained (as shown in box \( H \) in the figure).

3. Experiment and Result Analysis

3.1. Experimental Parameters and Environment Settings

In the experiment, we use MSCOCO database which is based on the challenge of image recognition and image description held by Microsoft in 2014. MSCOCO 2014 database contains images and various manual annotations of images. Aiming at the problem of image description, each image includes 5 sentences of text description of the image. For MSCOCO data set, 123287 images are divided into training set 113287, verification set 5000, and test set 5000; when annotating text, use <start> to mark the beginning statement of the image description, use <end> to mark the end statement of image description, and use <UNK> to mark the unrecognized word group in the image description. For text annotation, one-hot vector coding is used to generate a matrix with lower dimension. GeForce GTX 1080ti is used in the whole experiment. The video memory of the card is 11 g, and the memory size of the server is 32 g. The software environment is TensorFlow, an open-source framework of Python. The image size is 224 * 224, the number of iterations is 10000, and the vector dimension of the LSTM module is set to 1024; using Adam to optimize the model, the learning rate of training is 0.005. In training, the size of the data set is set to 32; in order to avoid overfitting in the experiment, dropout is set to 0.2 in the neural network. LSTM model parameter settings input dimension is set to120, hidden layer dimension is set to 1024, and layers are set to 2. Parameter settings are shown in Table 1.

4. Analysis and Comparison of Results

4.1. Comparison Test of Convolution Network

ResNet models that perform well in image classification tasks include ResNet34, ResNet50, and ResNet101. As the number of layers is higher, the model and effect are better. However, the higher number of layers will lead to higher hardware requirements and longer training time. The experiment uses the ImageNet data set to experiment with ResNet and selects BLEU-1 and BLEU-4 evaluation indicators to evaluate the experimental results. The experimental results are shown in Table 2.

The two evaluation indicators BLEU-1 and BLEU-4 evaluate the quality of the generated text. The larger the indicator data, the better the generated effect. According to the experimental data in Table 2, in the BLEU-1 and BLEU-4 indicators, the experimental data of ResNet50, ResNet101, and ResNet 152 are not much different, and the network models of ResNet101 and ResNet 152 are relatively complex, and the hardware requirements of the machine are relatively high. Training time is longer, comprehensive training cost, time-consuming and other factors, this experiment selects ResNet 50 as the image feature extractor. The comparison of model training time is shown in Table 3.

4.2. Comparative Experiment of Classical Image Description Algorithms

From the perspective of the model structure, in order to verify that modifying the model structure can affect the accuracy of the image text description, the experiment set up a total of three AAM-DLSTM model variants for ablation comparison experiments. The model for the comparison experiment is the adaptive attention model, AAM-TLSTM model and AAM-DLSTM-A. In order to ensure the effect of the comparative experiment, the experiment environment used by different variant models in the experiment is kept consistent.

Adaptive Attention Model: the first model compared in this experiment is the adaptive attention model, in which there is only one LSTM module. Compared with the AAM-DLSTM model, the original LSTM model replaces the LSTM-1 visual information module and the LSTM-2 text generation module in AAM-DLSTM, and formula (3) needs to be changed to formula (10) and (4) needs to be changed to formula (11).

\[
h_t, c_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}),
\]

(10)

\[
x_t = [s_{LE}, F, m_t],
\]

(11)

where \( h_t \) represents the hidden layer state of the LSTM module at the current moment, \( c_t \) represents the memory layer state of the LSTM module at the current moment, \( x_t \) represents the input of the single-layer LSTM module, \( h_{t-1} \)
Table 1: Experimental parameter setting.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Value</th>
<th>Parameter name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set size</td>
<td>113287</td>
<td>Validation set size</td>
<td>5000</td>
</tr>
<tr>
<td>Test set size</td>
<td>5000</td>
<td>Iterations</td>
<td>10000</td>
</tr>
<tr>
<td>Optimization method</td>
<td>Adam</td>
<td>Learning rate</td>
<td>0.005</td>
</tr>
<tr>
<td>Batch data</td>
<td>32</td>
<td>Dropout</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 2: Comparative experiment of ResNet with different layers.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet34</td>
<td>0.743</td>
<td>0.341</td>
</tr>
<tr>
<td>ResNet50</td>
<td>0.754</td>
<td>0.343</td>
</tr>
<tr>
<td>ResNet101</td>
<td>0.758</td>
<td>0.346</td>
</tr>
<tr>
<td>ResNet152</td>
<td>0.756</td>
<td>0.349</td>
</tr>
</tbody>
</table>

Table 3: Comparison of training time.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet34</td>
<td>2.2</td>
</tr>
<tr>
<td>ResNet50</td>
<td>2.7</td>
</tr>
<tr>
<td>ResNet101</td>
<td>4.0</td>
</tr>
<tr>
<td>ResNet152</td>
<td>6.2</td>
</tr>
</tbody>
</table>

represents the hidden layer state of the LSTM module at the previous moment, \( c_{t-1} \) represents the memory layer state of the LSTM module at the previous time, \( s_{tE} \) represents the word vector formed by combining the words generated at the previous time, \( F \) represents the image global feature vector, and \( m_t \) represents the context content vector at the current time.

AAM-TLSTM Model: the second algorithm model compared in this experiment is the AAM-TLSTM model. This model considers adding a layer of the LSTM module on the basis of the AAM-DLSTM model to verify the influence of the number of LSTM layers on the effect of the image text description algorithm. The LSTM model added by AAM-TLSTM on the basis of AAM-DLSTM is named LSTM-3, and formula (12) is added after formula (4), and formula (13) is as follows:

\[
\begin{align*}
  h_t^3, c_t^3 &= LSTM_3(x_t^3, h_{t-1}^3, c_{t-1}^3), \\
  x_t^3 &= [m_{t+1}, h_t^2],
\end{align*}
\]

where \( h_t^3 \) represents the hidden layer state of the LSTM-3 module, \( c_t^3 \) is the hidden layer state of the LSTM-3 module, \( x_t^3 \) is the input of the LSTM-3 module, \( m_{t+1} \) is the context content vector, and \( h_t^2 \) is the hidden layer state of the LSTM-2 module. The corresponding algorithm framework is shown in Figure 3.

AAM-DLSTM-A: the third algorithm model compared in this experiment is the AAM-DLSTM-A model. This model mainly considers the influence of the input of the text generation module LSTM-2 on the final generation of the complete sentence when the word at time \( t \) is generated. Compared with the AAM-DLSTM model, AAM-DLSTM-A modifies the input \( x_t^2 = [m_t, h_t^1] \) of the original LSTM-2 text generation module, and the modified expression is shown in the following formula:

\[
x_t^2 = [m_t],
\]

where \( x_t^2 \) represents the input of LSTM-2 and \( m_t \) represents the context content vector at the current moment. The meaning of this formula is that the text generation is only affected by the context content vector at the current moment. The algorithm framework diagram is shown in Figure 4, the input of LSTM-2 is only the context content vector \( m_t \).

Hard attention and soft attention are the first time to propose the attention mechanism for image text description tasks in 2018. Hard attention mainly uses reinforcement learning methods, and the hidden layer of the entire coding layer is not selected as input during the training process. In the input, a random process is adopted to sample and input the hidden layer. Soft attention mainly calculates coding through deterministic scores. When the attention is assigned to the probability distribution, the corresponding word is predicted for each attention value.

VD-SAN is a new type of dense semantic attention network proposed in 2019. This model mainly considers the correlation between image attributes and image feature extraction tasks and has achieved good results in image text description tasks.

AoANet is the AoANet model proposed in 2019. In the AoANet model, a module called AoA (attention on attention) is proposed. This module expands the conventional attention mechanism by adding an additional attention, making it possible to achieve a better effect in the image text description.

NBT is an image description method using “slots” proposed in 2018. NBT uses a neural network model to extract sentence templates and classifies the words generated at the current moment. When classified into visual words, the visual words are directly generated from the characteristics of the corresponding regions of the image, and when classified into text words, they are generated from the text lexicon to a certain extent. The above solves the problem of the traditional template filling sentence dull.

ETN is a vision-language training model proposed in 2020, which can iteratively polish the generated description and effectively improve the readability of the description sentence.

In order to verify the effect of the AAM-DLSTM algorithm in the task of image text description, this experiment compares the AAM-DLSTM algorithm with multiple public algorithm models and uses multiple objective quantitative scoring methods to evaluate the test results. The scoring table is shown in Table 4.

BLEU is a bilingual mutual evaluation aid tool. It is used to analyze the degree of the common occurrence of N-tuples in the candidate translation and the reference translation. BLEU-1 and BLEU-4 represent N values of 1 and 4, respectively.

ROUGE-L is calculated based on the recall rate and is the evaluation standard for automatic summarization tasks. L
refers to the longest common subsequence, and the longest common subsequence of the machine translation C and the reference translation S is used in the calculation of ROUGE-L.

METOR is a single-precision weighted harmonic mean and single-word recall metric that aims to address some of the inherent flaws in the BLEU standard.

CIDEr is a combination of BLEU and a vector space model. It regards each sentence as a document and then calculates the cosine angle of the TF-IDF vector to obtain the similarity between the candidate sentence and the reference sentence.

SPICE encodes the objects, attributes, and relationships in the caption using a graph-based semantic representation.

Comparative experiments were carried out on the three AAM-DLSTM model variants mentioned above, and the experimental results are shown in Table 4. From the experimental results, it can be seen that increasing the number of LSTM layers can improve the effect of image text description. When the number of LSTM layers is 2, the trained model has achieved good results on most evaluation indicators; when the number of LSTM layers is 1, the input information of the single-layer LSTM module in the adaptive attention model is reduced, which leads to the deterioration of the effect of image text description; when the number of LSTM layers is 3, the input of LSTM-2 in the AAM-TLSTM model contains the state of the hidden layer of LSTM-1. The input of LSTM-3 contains the state of the
hidden layer of LSTM-2. During model training, LSTM-3 has more reference information in the input when generating words, which makes the area of the current moment of attention in the image larger, but the accuracy of generating words is reduced, which leads to the description accuracy of the AAM-TLSTM model not as high as the accuracy of the AAM-DLSTM model; it can be seen that in this experiment, the stacking on the number of layers does not linearly improve the accuracy of the model. When the number of LSTM layers is 2, the information input by the text generation module is more abundant than that of the single-layer LSTM module, and the redundant information is more concise than the three-layer LSTM module, so the accuracy of the generated image text description is higher. Comparing the data of the AAM-DLSTM model and the AAM-DLSTM-A model, it is verified that when the input of the text generation module LSTM-2 is composed of the context content vector \( m_t \) and the hidden layer state \( h^l_t \) of LSTM-1, the effect of image description is better. When the input of LSTM-2 is only the context content vector, LSTM-2 cannot consider the hidden layer state in the LSTM-1 visual information module. When generating words, it only relies on the context content vector for prediction, which has a strong dependence on the spatial and local features of the image, resulting in the generated words tend to ignore the global features of the image, that is, the text description of the context image does not match the context in the image. Compared with soft attention, hard attention, NBT, and VD-SAN, the AAM-DLSTM model has improved some indicators. Compared with the adaptive attention model with unchanged structure, it has increased by 1.21% on BLEU-1, 0.75% on METEOR, 0.55% on CIDEr, and 1.96% on SPICE; the evaluation index of BLEU-4 and ROUGE-L is not much different from the evaluation index of adaptive attention. The evaluation indicators of BLEU-4 and ROUGE-L are not much different from those of adaptive attention. Compared with NBT, AAM-DLSTM is only slightly worse on BLEU-1. Other indicators are better than the NBT model, which verifies that the AAM-DLSTM model can improve the accuracy of image text description, but AAM-DLSTM is compared with the AoANet model. The ETN model has a limited effect on the improvement of image text description.

4.3. Comparison of Experimental Results. Although the evaluation indicators in the previous section can reflect the effect of the generated description sentences to a certain extent, these evaluation indicators cannot reflect the context and accuracy of the actual image text description sentences. In the experiment, the adaptive attention model, AAM-DLSTM model, and ETN model using adaptive mechanism are selected as the comparison objects of the experiment. The experiment uses the images in the MSCOCO 2014 test set to conduct a comparative experiment and shows the description effect (as shown in Figure 5).

As shown in Figure 5(a), the adaptive attention model uses the word “play” when describing nonvisual words, while the AAM-DLSTM model only describes the action of the stick, ignoring some details. The description of the main body of the image by the ETN model is more accurate, and the description of the visual word “baseball player” and the nonvisual word “swing” is more accurate; (b) in the figure, the adaptive attention model describes the dog playing on the green. The playing object and the frisbee in the mouth are not described. The AAM-DLSTM model does not capture the dog’s playing movements but only describes the dog holding the frisbee on the grass. ETN plays with both visual word “flying saucer” and nonvisual word “play”. Very good description; (c) in the figure, the adaptive attention model is less accurate when describing the gathering of people. The AAM-DLSTM model and ETN can effectively identify the situation of people. The ETN model is better than AAM in capturing details. DLSTM is more excellent and can describe the distribution of personnel well; (d) in the figure, the adaptive attention model ignores some visual words in the image such as padded jacket and sled when introducing the image; the AAM-DLSTM model is in the description, and there are similar problems. The above examples show that the AAM-DLSTM model is more accurate than the adaptive attention model, and the generated image description is more in
line with the context, but not as good as the ETN model, which can describe the details of the image more clearly. In summary, when performing text descriptions on images in the test set, the AAM-DLSTM model has improved image description accuracy compared to the adaptive attention model and has a better description of visual and nonvisual words. But the description of the details in the image is not as good as the ETN model. It also verifies that AAM-DLSTM has the problem of unclear description of some details in the image and also points out the optimization direction for subsequent research.

5. Conclusions

This section first pointed out that the existing image description algorithm of adaptive attention mechanism has a poor description effect, discussed the influence of different structure models on image description effect, and proposed a two-layer LSTM model based on adaptive attention mechanism. The image description algorithm AAM-DLSTM uses a two-layer LSTM model on the basis of the encoding-decoding framework. The two-layer LSTM model includes the visual information module LSTM-1 that can capture the visual words in the image at the current moment and the text generation module LSTM-2 that receives the context content vector and the description sentence generation. Through comparative experiments on the MSCOCO 2014 data set, it is verified that compared with some classic image description algorithms, AAM-DLSTM has improved some evaluation indicators, and the generated text description is more in line with the image context.

Data Availability

The (experimental) data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China, Research on Adaptive Integrated Evolutionary Algorithm for Multi-Root Solving of Complex Nonlinear Equations (62076225).

References

[6] Q. Huang, Research on Music Genre Classification Model Based on Convolutional Neural Network [D], Jilin University, Changchun, China, 2019.