applied sciences

MDPI

Article

Video Description Model Based on Temporal-Spatial and Channel Multi-Attention Mechanisms

Jie Xu *, Haoliang Wei, Linke Li, Qiuru Fu and Jinhong Guo *

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; chooperliang@gmail.com (H.W.); linkeli1992@163.com (L.L.); fuqr@uestc.edu.cn (Q.F.)

* Correspondence: xuj@uestc.edu.cn (J.X.); guojinhong@uestc.edu.cn (J.G.)

Received: 24 April 2020; Accepted: 18 June 2020; Published: 23 June 2020

Featured Application: This work can be widely used in advanced intelligent systems, including smart city, smart transportation and smart home, etc.

Abstract: Video description plays an important role in the field of intelligent imaging technology. Attention perception mechanisms are extensively applied in video description models based on deep learning. Most existing models use a temporal-spatial attention mechanism to enhance the accuracy of models. Temporal attention mechanisms can obtain the global features of a video, whereas spatial attention mechanisms obtain local features. Nevertheless, because each channel of the convolutional neural network (CNN) feature maps has certain spatial semantic information, it is insufficient to merely divide the CNN features into regions and then apply a spatial attention mechanism. In this paper, we propose a temporal-spatial and channel attention mechanism that enables the model to take advantage of various video features and ensures the consistency of visual features between sentence descriptions to enhance the effect of the model. Meanwhile, in order to prove the effectiveness of the attention mechanism, this paper proposes a video visualization model based on the video description. Experimental results show that, our model has achieved good performance on the Microsoft Video Description (MSVD) dataset and a certain improvement on the Microsoft Research-Video to Text (MSR-VTT) dataset.

Keywords: intelligent imaging technology; deep learning; video description; multi-attention perception mechanism; consistency of visual features; visualization model

1. Introduction

Video description is widely used in advanced intelligent technology, including smart city, smart transportation and smart home [1–5]. Video description technology is a part of computer vision and natural language processing, which has attracted much attention in recent years [6–9]. In 2014, Venugopalan [10,11] proposed a video description model based on the framework of "encoding-decoding." The encoding method in his model extracted features from a single frame of video by using CNN, then adopted the mean pooling and time series encoding models, respectively. Although the Venugopalan model proposed has been applied successfully in video description, there are also some problems with this model.

The first problem is that video features are not utilized effectively. Video features are only used in the first decoding and not used subsequently, thus reducing the ability of video features to predict words when time-series increase [10,11]. Therefore, the capability of sentence generation decreases. The second problem is the consistency of visual content features and sentence descriptions. In the first problem, application of the temporal attention mechanism increases the utilization of video

features. However, this approach does not model the relationship between video features and sentence descriptions [12,13]. Therefore, it brings about the second problem that is how to ensure the consistency of visual content features and sentence descriptions.

For these problems, one solution is to add the video feature each time. However, the video feature consists of multiple images. If we still use the pooling encoding method to send the video feature into the decoding model each time, then the video feature will not be utilized effectively. Xu [12] proposed an image description model based on attention mechanism. The model weighted every region of each image by using the attention mechanism before each word-predicting process, making the feature used in each prediction different. Based on this idea, Yao [13] proposed a video description model based on a temporal attention mechanism. Their model weighted the features of all video frames and summed them whenever making word prediction. Experimental results showed that the video feature was utilized effectively.

In this paper, we propose a video description model based on temporal-spatial and channel attention to solve the abovementioned problems. At present, the most effective way to extract image features is the convolutional neural network (CNN) [14,15]. For an image with a size of w × h (w represents the width and h represents the height), by processing it with CNN, we obtain a new coding feature with a size of w × h × c (c represents the new feature obtained by CNN). The convolution kernel can detect different features of an image. In general, the convolution kernel at lower layers can detect information such as edge texture and the convolution kernel at higher layers can detect features with semantic information. Therefore, the feature map obtained from an image through CNN contains spatial and semantic information. However, most existing models only focus on the attention mechanism of time or space [10]. Thus, the substantive feature of the CNN network is not fully utilized. However, our multi-attention video description model introduces the channel attention mechanism on the foundation of a traditional temporal and spatial attention mechanism. Besides, this model makes a stronger combination of visual features and sentence descriptions so that the accuracy of the model is increased. The model is experimented on the datasets Microsoft Video Description (MSVD) and Microsoft Research-Video to Text (MSR-VTT). Further, BLEU@4 [16], CIDEr [17], METEOR [18], and ROUGE_L [19] are adopted as evaluation indexes. Experimental results verified the effectiveness of our model. Then a video visualization model is proposed based on video description. In this video visualization model, we made a visual analysis of our attention mechanism and proved the accuracy of the model intuitively.

2. Attention Mechanism

The video description has achieved a breakthrough with the combination of deep learning [20–23]. Meanwhile, the technique based on visual attention mechanism has also been successfully applied in the video description model and it solves the first problem mentioned above.

The visual attention mechanism is widely applied in image and video description tasks because human vision does not process the entire visual input at once. Instead, human vision only focuses on the information of crucial parts. Based on this reasonable hypothesis, current description models usually do not use static coding features of an image or a video. Instead, they use an attention mechanism and sentence context information to extract image features. Therefore, visual attention is an encoding mechanism that extracts features dynamically based on context information over the whole time-series. At present, the attention mechanism is mainly based on time and space. We would improve our model by utilizing these two factors and take advantage of CNN. Then, we will introduce the idea of a channel attention mechanism.

2.1. Temporal Attention

Yao [13] was the first person to propose a video description model based on a temporal attention mechanism. In a video, each frame contains information at a different time, so the visual feature of each frame is a relatively complete semantic information expression. It expresses global time-series

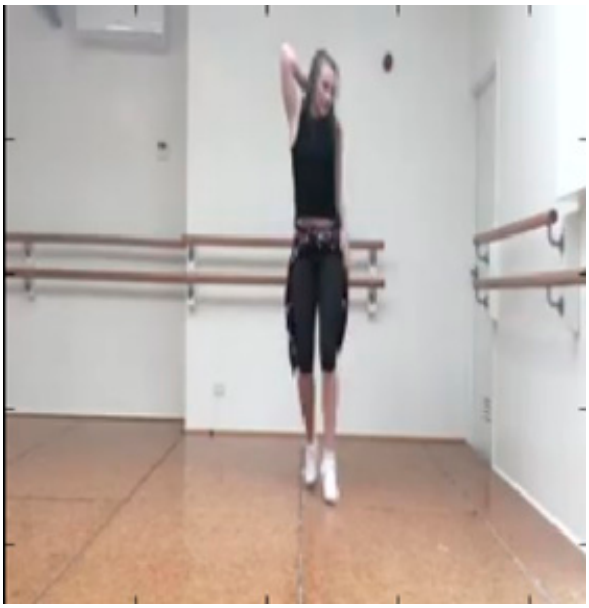
information, such as the order in which objects, actions, scenes, and characters appear in a video. This kind of information may exist throughout a video. A good video description model should focus on important moments in a video sequence. With the help of temporal attention mechanism, we can decide the importance of different frames.

2.2. Spatial Attention

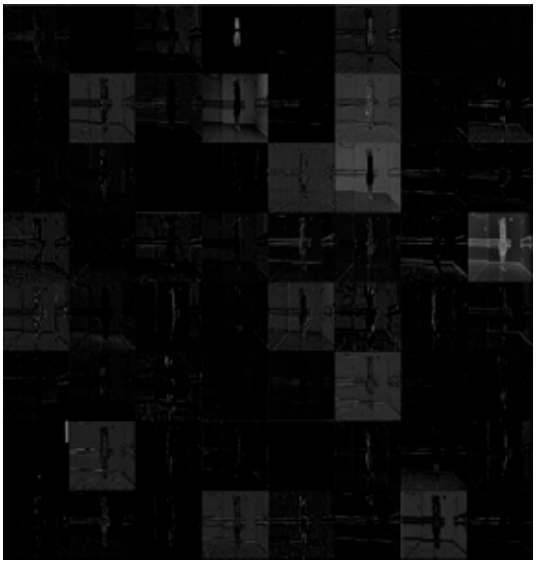
Apparently, for a video, it has not only global temporal features, but also local temporal features. These local temporal features are usually the representations of the fine-grained features of actions in a video, such as standing up, calling someone, etc. These features are also the fine-grained features for identifying important regions of an image, such as the region of the face in an image. These features only appear in certain regions of a certain time-series in a video generally. If we do global pooling to all video frames, then the ability to capture local information will decrease. The spatial attention mechanism enables the model to acquire local features. Xu [12] proposed the idea of applying a spatial attention mechanism in the video description.

2.3. Spatial Attention

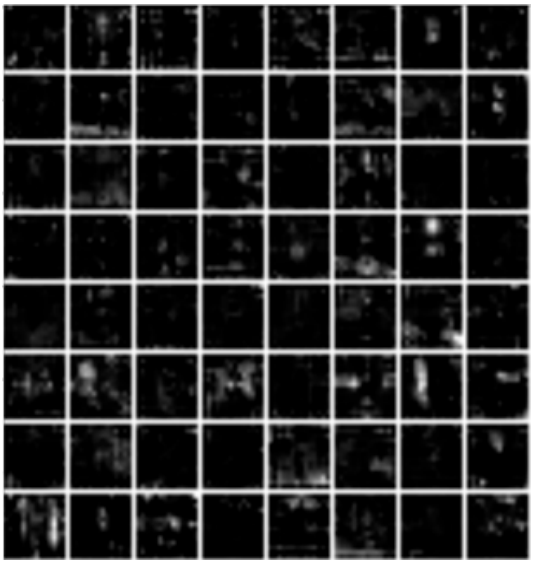
After an image is processed with the CNN, each layer of the CNN will produce different numbers of feature graphs, i.e., channel features. The experiment done by Zeiler [24] showed that the feature map of different layers of the CNN can show different semantic information. Specifically, low-level feature maps show low-level visual features, such as texture and color, whereas the high-level feature maps show high-level semantic features such as objects with different spatial features. Because the CNN contains many convolution kernels that can detect different features, we can obtain different feature maps. Figure 1 is the visualization result of a part of low-level and high-level feature maps after an image entered the CNN. In this figure, the low-level features focus on some edge contour information of the image (Figure 1b), whereas the high-level features prefer to express semantic information of the image (Figure 1c). As shown in Figure 1, the brightness of the feature represents the response value to the original image. While it is brighter, the response value is higher. We can see that the Figure 1c shows the response of the feature channel in the same layer after the image is input into the CNN network. It can be seen from the figure that the response positions of different feature maps are different and that these different response positions express different semantic information prominently. If the response position of some feature maps is on the person, then the region with people will have a larger regional response value. Some feature maps are responses to other objects. Inspired by this visualization result, for obtaining video feature information consistent with description words of a video in video descriptions more effectively, this paper proposes a channel attention mechanism. We can enhance the ability of the model to focus on the feature maps in need by using channel attention mechanism and then improve the description capability of the model.



(**a**)



(**b**)



(**c**)

Figure 1. Visualization result of low-level and high-level feature maps. (**a**) Original image; (**b**) low-level feature maps; (**c**) high-level feature maps.

3. Multi-Attention Video Description Model

3.1. Network Architecture

In the previous section, we introduced temporal, spatial, and channel attention mechanisms and their corresponding extraction abilities for different features in a video. To make full use of capabilities of attention mechanisms, we propose a video description model based on the temporal-spatial and channel attention mechanisms in this section. The fundamental network architecture is shown in Figure 2. This model utilizes different attention mechanisms. For example, when predicting the words in need, this model will combine the features effectively in the video by the attention mechanism

automatically, then generate the appropriate corresponding words; thus, it is a solution to the problem related to the consistency of visual content features and sentence descriptions.

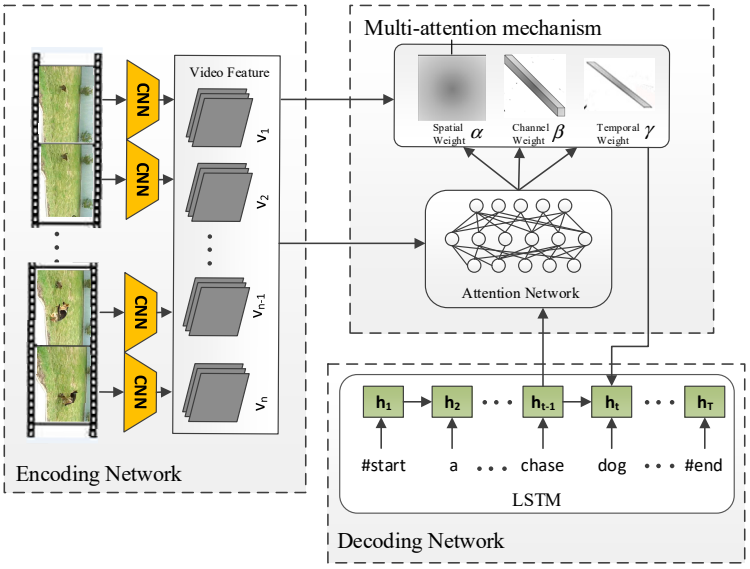


Figure 2. Video description model based on temporal-spatial and channel attention.

As shown in Figure 2, the multi-attention model proposed in this paper consists of three parts: a video encoding network, a multi-attention network, and a decoding network. The video encoding network mainly uses the CNN to extract features of video frames. The multi-attention network comprises two parts: an attention network and a feature fusion. Therefore, the attention network can utilize the video features and the previous output to calculate the attention weight of the video feature in time, space, and channel respectively, then recalculate the fused features by the obtained weights and the video features. In this way, we can get more useful features. Finally, the fused features will be exploited by the decoding network for encoding output to obtain more consistent description with video content.

By extracting visual features effectively in time, space, and channel, the multi-attention description model proposed in this paper enhances the representation ability of the model. The three attention models we propose are relatively independent. We can combine three models randomly to research the different effects of various combinations. An attention model named S-C-T (spatial-channel-temporal) is shown in Figure 2.

First, from Video I, we can extract the network features after operating it in CNN:

$$V = f_{CNN}(I) \tag{1}$$

$V = \{v_1, v_2, \ldots, v_n\}$ indicates that the video has n frames with a dimension of $K * K * C$, and the video feature of every frame is $v_i \in R^{K*K*C}$. At time t, we use the decode unit, LSTM (Long Short-Term Memory), to get the output h_{t-1} at the previous time. With known values of h_{t-1} and V, the unknown attention weight factors (α, β, γ) of time, space, and channel can be calculated by Equation (2):

$$(\alpha, \beta, \gamma) = \Phi(h_{t-1}, V) \tag{2}$$

Φ covers a series of equations in the Section 3.2. We can obtain (α, β, γ) from h_{t-1} and V, where $\alpha \epsilon R^{K*K}$, $\beta \epsilon R^{C}$, $\gamma \epsilon R^{N}$. The corresponding formula is presented in detail in the next section. Then, we apply the three weight factors to V and obtain the features needed to be input into the LSTM:

$$z = f(V, \alpha, \beta, \gamma) \tag{3}$$

where f represents the output of three attentional operations on V. Section 3.2 is the detailed introduction to Equation (3). Then, we can start to predict the next word:

$$h_t = f_{LSTM}(h_{t-1}, z, w_{t-1}) \tag{4}$$

$$y_t \sim p_t = \text{softmax}(W_e h_t + b) \tag{5}$$

where w_{t-1} represents the word vector of the corresponding word y_{t-1} and $p_t \in R^{|D|}$ is the probability distribution of words. In this model, there is a word vector space $W_D \in R^{D*S}$, where D is the number of words and S is the dimension of the word vector. Supposing we have M videos, the number of words in the sentence is T, and the loss function of the whole model is Equation (6), whose mathematical meaning is a maximum likelihood estimation:

$$L_y = \frac{1}{M}\sum_{d=1}^{M}\sum_{t=1}^{T} -\log P_{it}(y_t | y_1, y_2, \ldots\ldots, y_{t-1}, V_i, \Omega) \tag{6}$$

$$L_\lambda = \lambda \|\theta\|_2^2 \tag{7}$$

$$L = L_y + L_\lambda \tag{8}$$

where Ω represents the parameters that can be trained, including word vectors. According to the formula, the loss of the model consists of two parts. L_y is the loss of predicting the consistent word. Due to the high complexity of the model, the model is prone to overfitting. λ is the regularization coefficient, and θ is the training model data. L_λ is the loss of regular function. To prevent the model from overfitting and control the complexity of the model, we add a second loss to regularize the model parameters.

In the training process of the whole model, our primary objective is its optimization. By using the backpropagation algorithm to update the parameters and minimize the loss function, we can achieve an optimum model.

3.2. Attention Calculation

From the previous section, we know that this paper has three attentions: temporal attention, spatial attention, and channel attention. The three attentions are relatively independent, so we can change the order in which they act on the video features. As shown in Figure 2, we apply the spatial attention on video features first and then apply the channel attention. It will be introduced in detail in the following section.

3.2.1. Spatial Attention

From a video, CNN can extract the video features $V = \{v_1, v_2, \ldots\ldots, v_n\}$, where v_i is the feature of each frame with a dimension of $K * K * C$. We can represent the video feature v_i as $v_i = \{r_{i1}, r_{i2}, \ldots\ldots, r_{ik^2}\}$, where r_{ij} represents the j region feature of the i frame with a dimension of C. The spatial attention for the ij region at time t can be formulated as follows:

$$e_{ij}^{(t)} = w_{att-s}^T \tan h\left(W_{att-s} h_{t-1} + U_{att-s} r_{ij} + b_{att-s}\right) \tag{9}$$

$$\alpha_{ij}^{(t)} = \exp\left\{e_{ij}^{(t)}\right\} / \sum_{l=1}^{k^2} \exp\left\{e_{il}^{(t)}\right\} \tag{10}$$

$$v_{i-s}^{(t)} = v_i * \alpha_i^t \tag{11}$$

$$X_{att-s}^{(t)} = f_{att-s}\left(V, \alpha^{(t)}\right) \tag{12}$$

where W, U, and b are network weight parameters that can be learned. $\alpha_i^{(t)} = \left\{\alpha_{i1}^{(t)}, \alpha_{i2}^{(t)}, \ldots, \alpha_{ik^2}^{(t)}\right\}$. f_{att-s} means applying a special attention operation on all video frames as in (9)–(11); finally, the video features through spatial attention are obtained: $X_{att-s}^{t} = \{v_{1-s}, v_{2-s}, \ldots, v_{n-s}\}$, whose dimension is the same as V.

3.2.2. Channel Attention

Through the space attention, we obtain new video features: $X_{att-s}^{(t)} = \{v_{1-s}, v_{2-s}, \ldots v_{n-s}\}$. We will convert $X_{att-s}^{(t)}$ to $U = \{u_i, \ldots u_c\}$, where $u_i \in R^{K*K*N}$ and C denotes the number of feature channels. Then we can adopt the average pooling operation and obtain the channel feature vector of the video: $c = [c_1, \ldots, c_i]$, $c \in R^C$, where the scalar c_i represents the mean value of the vector u_i, which denotes the channel feature value. The attention can be calculated by Equations (13)–(16):

$$\mathbf{b} = \tan \mathrm{h}((W_{att-c} \otimes c + b_c) \oplus W_{hc} h_{t-1}) \tag{13}$$

$$\beta = \mathrm{softmax}\left(W' \mathbf{b} + b_i'\right) \tag{14}$$

where $\otimes$ is the symbol of the outer product and $\oplus$ denotes the addition of matrices and vectors. The dimension of β is C and we apply β on v_{i-s}:

$$v_{i-c} = v_{i-s} * \beta \tag{15}$$

$$X_{att-c}^{(t)} = f_{att-c}\left(X_{att-s}, \beta^{(t)}\right) \tag{16}$$

f_{att-c} indicates applying the channel attention operation on all v_{i-s}, as in Equations (13)–(15) and we obtain the video features after channel attention operation: $X_{att-c}^{(t)} = \{v_{1-c}, v_{2-c} \ldots v_{n-c}\}$.

3.2.3. Temporal Attention

After the feature processing of the spatial attention and channel attention, the video feature is $X_{att-c}^{(t)} = \{v_{1-c}, v_{2-c} \ldots v_{n-c}\}$ with a dimension of $K * K * C$. Then, we adopt a pooling operation on it. Finally, we get $V' = \left\{v_1', v_2', \ldots, v_n'\right\}$ with a dimension of C. After obtaining the features of the video sequence, we can weigh the features by temporal attention to obtain the whole features expression of the whole video. Suppose $\gamma^{(t)} = \left\{\gamma_1^{(t)}, \gamma_2^{(t)}, \ldots, \gamma_n^{(t)}\right\}$ and $\gamma_i^{(t)}$ is the weight of the video feature of the frame i when the model predicts words at time t, which satisfies $\sum_{i=1}^{n} \gamma_i^{(t)} = 1$. $\gamma_i^{(t)}$ can be computed from the output of the previous time h_{t-1} and the video feature V:

$$e_i^t = w_{att-t}^T \tan \mathrm{h}\left(W_{att-t} h_{t-1} + U_{att-t} v_i' + b_{att-t}\right) \tag{17}$$

$$\gamma_i^{(t)} = \exp\left\{e_i^{(t)}\right\} / \sum_{g=1}^{n} \exp\left\{e_g^{(t)}\right\} \tag{18}$$

Then, for the i frame, the result is:

$$v_{i-t} = v_{i-c} * \gamma_i^{(t)} \tag{19}$$

$$X_{att-t}^{(t)} = f_{att-t}\left(X_{att-c}^{(t)}, \gamma^{(t)}\right) \tag{20}$$

f_{att-t} means the operation of temporal attention for all frames as in Equations (17)–(19) and we obtain the video features after the temporal attention operation: $X_{att-t}^{(t)} = \{v_{1-t}, v_{2-t} \ldots v_{n-t}\}$.

After the operations in spatial, channel, and temporal domain, we get the feature $X_{att-t}^{(t)}$. After weighting the $X_{att-t}^{(t)}$, the feature z that needs to be sent to the decoding network is obtained, where the pool indicates the space pooling operations:

$$z = \frac{1}{n}\sum_{i}^{n} \text{pool}(v_{i-t}) \tag{21}$$

Furthermore, the dimensions of $X_{att-t}^{(t)}$, $X_{att-c}^{(t)}$, $X_{att-s}^{(t)}$, and V are the same. $X_{att-t}^{(t)}$, $X_{att-c}^{(t)}$, and $X_{att-s}^{(t)}$ represent the processing results of temporal, channel, and spatial attention to V, respectively. Here, we calculate the spatial (S) first and then the channel (C), and finally the temporal (T). Obviously, the logical order of the calculations can be replaced; that is, the calculation order of the formula 12/16/20 can be changed and the corresponding video features can be used. Equation (3) expresses the calculation of the entire attention feature. Here we mainly calculate the spatial (S), then calculate the channel (C), and finally calculate the temporal (T), which is the S-C-T model. Similarly, in order to calculate the attention characteristics of the T-S-C model, we just need to calculate the above three characteristics of $X_{att-t}^{(t)}$, $X_{att-c}^{(t)}$, and $X_{att-s}^{(t)}$ according to the corresponding calculation order. For exploring the influence of the different logical orders of attention operation on the model, we propose eight models of T-S-C, T-C-S, S-T-C, S-C-T, C-T-S, C-S-T, ST-C, and C-ST. The first six models are combinations of different logical orders of the three attention mechanisms. In the video features, the temporal-spatial features are usually regarded as a holistic feature, so the ST-C and C-ST models represent the temporal-spatial features as a whole for modeling, and the calculation principle is the same with the first six models. We will analyze the results of these models in the experiment.

3.3. Attention Visualization

This paper attempts to adopt a top-down analysis on the model with the foundation of existing video description models and proposes significance analysis based on the visual video description. In detail, in an existing video description model, when we get a video and the corresponding description, we want to use the model directly to establish the correspondence between the input object and the words in the sentence, then output the saliency map of the video. The basic "encoding–decoding" model is used in this paper. As shown in Figure 3, the front h_1 to h_m is the encoding network, which is the encoding network of the attention model proposed in the paper. Its main principle is to measure the significance level by figuring out the loss of information when using local visual inputs to approximate the entire input sequence.

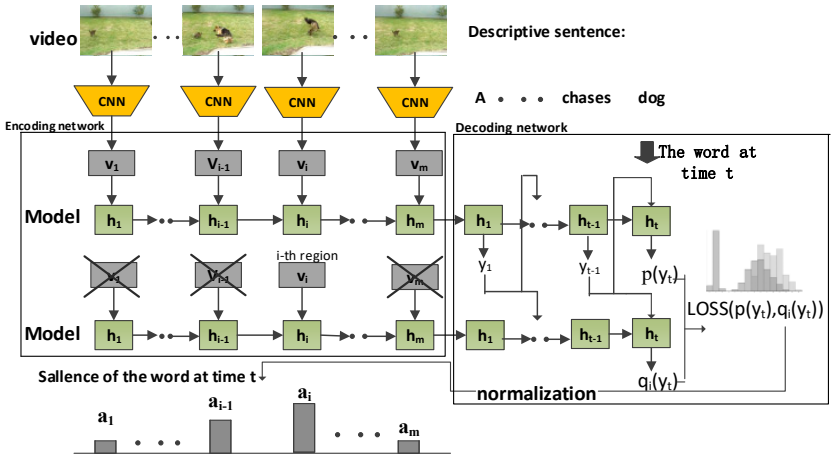


Figure 3. Saliency analysis flowchart based on the video description model.

Suppose that the probability distribution of words $p(y_t)$ is predicted by LSTM at each moment in the prediction process. We assume that the probability distribution is a "true" distribution. To measure how much information the i-th local description feature of the video brings to the word at time t, we only use the i-th local description feature input encoder and calculate the probability distribution $q_i(y_t)$ at time t after encoding and decoding. Then, we calculate its KL divergence as a measure of information loss:

$$p(y_t) = P(y_t | y_{1:t-1}, v_{1:m}) \tag{22}$$

$$q_i(y_t) = P(y_t | y_{1:t-1}, v_i) \tag{23}$$

$$Loss(t, i) = D_{KL}(p(y_t) || q_i(y_t)) \tag{24}$$

In the input sentence S, because what we input is the true distribution of words, the probability distribution of words should satisfy "one-hot" distribution. The above equation can be simplified as follows:

$$Loss(t, i, w) = \sum_{k \epsilon W} p(y_t = k) \log \frac{p(y_t = k)}{q_i(y_t = k)} = \log \frac{1}{q_i(y_t = w)} \tag{25}$$

Therefore, we get the information loss of the predicted word w in the i-th local feature description of the video at time t, and this value represents the connection between the visual features and the words. Hence, we can use it as a saliency value:

$$e_{ti} = scale(-Loss(t, i, w)) \tag{26}$$

The range of the value is 0–1. When the value is higher, the saliency is stronger.

4. Experiment

4.1. Datasets and Evaluation Metrics

The model is experimented on the datasets Microsoft Video Description (MSVD) [25] and Microsoft Research-Video to Text (MSR-VTT) [26]. The MSVD dataset contains 1970 short videos, 1200 videos for the training set, 100 videos for validation set, and 670 videos for the testing set. The dataset MSR-VTT contains 6513 videos for the training set, 497 videos for validation set, and 2990 videos for the testing set. BLEU@4 [16], CIDEr [17], METEOR [18], and ROUGE_L [19] are adopted as evaluation indexes.

4.2. Experiment Setting

In our experiment, we sampled uniformly and set 26 frames as an interval group in a video. If a video was less than 26 frames, then we would fill the video with "zero-pad". For the corpus in the dataset, we used the NLTK (Natural Language Toolkit) [27] to tokenize and process the sentence. When word frequency is too low, it is difficult for the model to learn. At the same time, according to our statistics, the proportion of words with a word frequency of less than 4 is relatively small in the entire data set. For example, this proportion is only 1.69% in the MST-VTT data set. There are still some misspelled words or words with no specific meaning in the dataset, which are all incorrect words. The presence of these incorrect words will make it difficult for the model to describe the content of the video. Thus, words with a frequency less than four or incorrect would be eliminated. This is also a general processing method in video description [28]. Meanwhile, three specific words (<start>, <end>, and <unknown>) are added to represent the beginning of the model, the end of the model, and the unknown identifier, respectively. Finally, the word dictionary is obtained. For each description sentence, we would fix the length of the sentence to 20. If the length of the sentence were less than 20, then we would fill it with ending characters.

The Inception-V3 [29] served as a CNN for extracting video features in the model. We used the output of "pool3" in the Inception-V3 as the video features with a dimension of $8 \times 8 \times 2048$. We only used the Inception-V3 to extract video features, and we did not train the model parameters whose

input was 229 × 229. The word representation used in our model is word embedding. In this model, the dimension of the hidden layer of LSTM was 1024, the word vector dimension of words was 512, and the time steps of LSTM were 20. The Adam [30] algorithm is utilized to optimize the model training process with an initial learning rate of 0.001.

4.3. Experimental Result

4.3.1. Analysis of Different Attention Combinations

The multi-attention video description model we proposed in this paper is shown in Figure 3 and contains temporal, spatial, and channel attention mechanisms represented by T (temporal), S (spatial), and C (channel), respectively. We can combine the three attention mechanisms to build six basic models: T-S-C, T-C-S, S-T-C, S-C-T, C-T-S, and C-S-T. Besides, because the information in the time domain and the space domain are often correlated in a video, we consider the space and the time as a whole so that we build two models: ST-C, and C-ST. Therefore, our model has eight different combinations in total. Venugopalan et al. [10] proposed the Base Model. The experimental results on MSVD are shown in Table 1. The experimental result is the statistical value of multiple experiments. All the test results were obtained by a simple greedy search method.

Table 1. Test results of different combinations on Microsoft Video Description (MSVD).

Structure	BLEU-4	CIDEr	METEOR	ROUGE_L
Base Model	41.2	64.8	31.4	67.3
T-S-C	**44.9**	76.6	32.1	68.9
T-C-S	44.3	74.5	32.2	68.4
S-T-C	43.5	74.9	32.2	68.1
S-C-T	44.6	**78.0**	**32.6**	**68.9**
C-T-S	44.4	76.1	32.4	68.4
C-S-T	43.5	73.1	32.1	68.0
C-ST	43.6	72.6	32.1	67.9
ST-C	**44.9**	75.5	32.4	68.6

The results show that the S-C-T model has the best results on CIDEr, METEOR and ROUGE_L evaluation indexes, and the ST-C model has the best result on BLEU-4 evaluation index. Therefore, the S-C-T model was the best one in the first six models while the C-S-T model performed worst. This difference indicated that the model focused more on certain spatial region features of video frames in the video description model, and improving the effectiveness of the model by obtaining effective spatial features. Similarly, when we considered the space and the time as a whole, the results of ST-C performed well on certain performance indexes. According to the results, if the model initially focused on the channel attention, then the results were relatively poor among all the models. Because the channel attention focused on certain semantics of the local features in the video, which indicated that focusing on the global feature initially, the local features helped improve the performance of the model.

Table 2 shows the results of the S-C-T model on Greedy Search and Beam Search. For the Beam Search algorithm, the test results of K from 2 to 10 are shown in this table. The experimental result is the statistical value of multiple experiments It can be seen that the Beam Search algorithm has a significant improvement on BLEU-4, CIDEr, and METEOR, and a small improvement on ROUGE_L. This shows that the Beam Search algorithm has better results. Because compared to the Greedy Search algorithm, the Beam Search algorithm has more search paths and explores more possible sentence sequences. However, as the K increases, the calculation time of Beam Search algorithm increases, and meanwhile, the result tends to be stable. Therefore, it is necessary to select an appropriate K to achieve a balance in time and performance, and usually the K is selected to be 6 or less. As Table 2, when K = 6, the S-C-T model can obtain the best results in this paper.

Table 2. Test results of the spatial-channel-temporal (S-C-T) model on Greedy Search and Beam Search.

Test Method	BLEU-4	CIDEr	METEOR	ROUGE_L
Greedy Search	44.6	78.0	32.6	68.9
Beam Search (2)	48.3	80.2	33.2	69.6
Beam Search (3)	48.4	80.3	32.8	68.9
Beam Search (4)	47.8	80.0	33.0	68.7
Beam Search (5)	48.4	80.4	33.1	68.6
Beam Search (6)	**48.8**	**82.0**	**33.4**	**69.7**
Beam Search (7)	**48.8**	79.8	33.3	68.5
Beam Search (8)	48.3	78.8	33.2	68.3
Beam Search (9)	48.3	78.3	33.1	68.2
Beam Search (10)	48.5	78.2	33.2	68.2

4.3.2. Comparison with Methods in Other Papers

To verify the effectiveness of the model, we compared our methods with other methods in existing relevant articles, including basic Seq-Seq (S2VT) [11], joint embedded model (LSTM-E) [31], temporal attention model (TA) [13], hierarchical model (HRNE-SA) [32], paragraph description model (P-RNN) [33], task-driven dynamic model (TDDF) [34], multi-level attention model based RNN (MAM-RNN) [35], LSTM-GAN model [7], video captioning model with tube features (VCTF) [36], VD-SVOs model [37], and spatial-temporal attention mechanism (STAT) [38]. Among them, the joint embedded model, HRNE-SA, and P-RNN used the average pooling, spatial attention, and temporal-spatial attention, respectively. To compare with other models effectively, this model utilized the Beam Search algorithm, where K is 6.

Table 3 shows that our S-C-T model was the best on CIDEr and METEOR, and consistent with the best method TDDF on the ROUGE-L. Compared with STAT, although our S-C-T model is weaker than STAT on BLEU-4, it is much stronger than STAT on CIDEr index, and slightly better on METEOR. However, STAT uses image features, motion features and local features, while we only use image features, which makes our S-C-T model easy to implement. Furthermore, when STAT only employ image features (STAT_GlobFeat), our model is superior to STAT on CIDEr and METEOR, and almost consistent in BLEU-4. One prominent progress in our experiment was the significant improvement of CIDEr compared with others. Meanwhile, different from other indexes, CIDEr is exclusively used in image video description index and it refers to the punctuation of words, the accuracy of word order, the accuracy of semantic and content descriptions, and fluency synthetically. Therefore, this experiment verified the effectiveness of our multi-attention model on taking advantage of the visual features and solving the problem of consistency between visual features and the semantic description.

Table 3. Results of different models on MSVD, where (-) indicates unknown scores.

Model	BLEU-4	CIDEr	METEOR	ROUGE-L
S2VT [11]	-	-	29.8	-
LSTM-E(VGG+C3D) [31]	45.3	-	31.0	-
TA [13]	-	51.7	29.6	-
HRNE-SA [32]	43.8	-	33.1	-
P-RNN-VGG [33]	44.3	62.1	31.1	-
P-RNN-C3D [33]	47.4	53.6	30.3	-
TDDF(VGG+3D) [34]	45.8	73.0	33.3	**69.7**
MAM-RNN [35]	41.3	53.9	32.2	68.8
LSTM-GAN [7]	42.9	-	30.4	-
VCTF [36]	43.8	52.2	32.6	69.3
VD-SVOs [37]	40.5	58.4	30.2	-
STAT_GlobFeat [38]	48.9	67.1	32.6	-
STAT [38]	**52.0**	73.8	33.3	-
Our S-C-T (K = 6)	48.8	**82.0**	**33.4**	**69.7**

To analyze the model on a larger dataset, we evaluated the model on the MSR-VTT [25] dataset. The result is shown in Table 4. Among the models in Table 3, only TDDF, LSTM-GAN and STAT were tested on MSR-VTT, thus our paper only compares with these models. As Table 4, our model has an improvement on BLEU-4 and METEOR, and has a little improvement on ROUGE-L. However, the results of CIDEr show significant differences on the MSVD and MSR-VTT datasets. For the MSVD dataset, the CIDEr score of our model is significantly higher than TDDF and STAT, but lower than them on MSR-VTT dataset. This result can be analyzed from the dataset. The MSVD dataset has fewer videos than the MSR-VTT dataset, but each video has 40 description statements, whereas the MSR-VTT dataset has only 20 description statements. Our model mainly obtains the effective features of the video from the temporal, spatial, and channel attention, and the single video (which has more descriptions) helps the feature acquisition ability of the attention mechanism. Therefore, our model has better performance for more diverse descriptions. At the same time, the base feature of our model only use image features, whereas TDDF uses both image features and motion features and STAT uses image features, motion features and local features, which is beneficial for datasets with more action categories like MSR-VTT.

Table 4. Results of different models on Microsoft Research-Video to Text (MSR-VTT).

Model	BLEU-4	METEOR	CIDEr	ROUGE-L
TDDF(VGG+3D)	37.3	27.8	43.8	59.2
LSTM-GAN	36.0	26.1	-	-
STAT_GlobFeat	37.1	25.9	41.0	-
STAT	**37.9**	26.8	**44.0**	-
Our S-C-T	**37.9**	**28.4**	40.6	**59.3**

Some results of video descriptions are shown in Figure 4. We can see that for the first two scenarios, the model gives a very appropriate statement to describe the content of the video objectively. A bad example is also shown in this figure. The statement description is "a man is singing a a a", which shows that the statement has an obvious error. The model cannot give the correct word after the article "a," which may be a word similar like "song." This may be due to the lack of attention mechanism for the ability to model abstract nouns such as ""song" that cannot express specifically.

Figure 4. Examples of video description result of our model.

4.3.3. Visual Analysis and Validation of Attention Mechanism

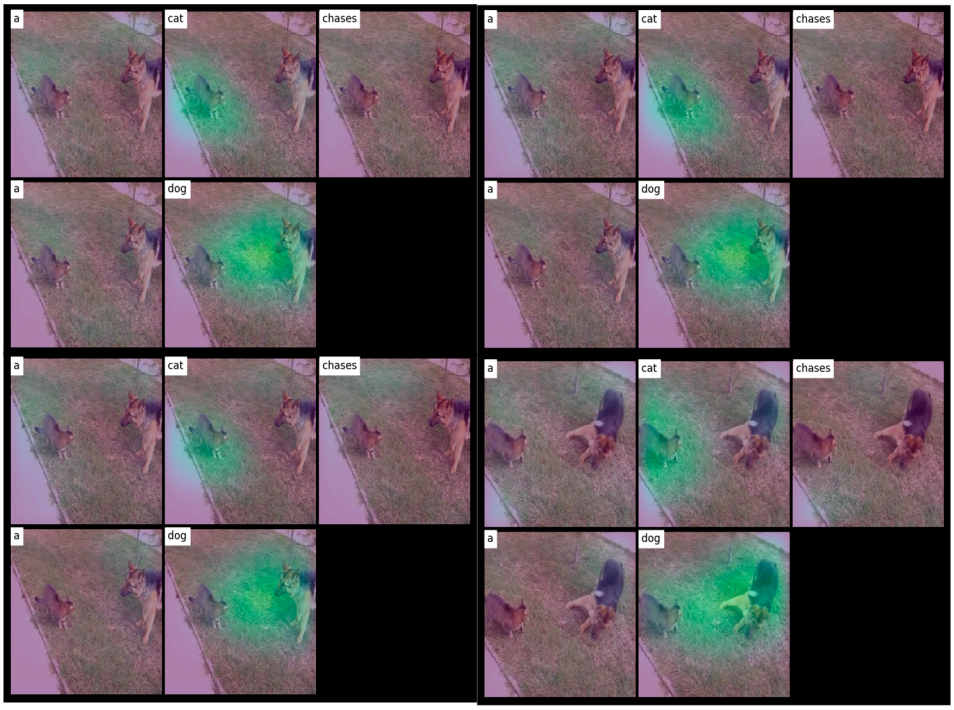
To understand the correlation between visual features and semantic features intuitively, we explored the relationship between the video description and the video content and attempted to analyze their inner links visually. Finally, we proposed a visualization model based on the video description.

We selected a video in the MSVD and used the method introduced in the previous sections to obtain the salient region of each word in the sentence description model. In Figure 5, it was divided into

three parts: the beginning of the video, the body of the video, and the end of the video. Additionally, each part displayed four images. The descriptive sentence was "a cat chases a dog." Every picture had the same five video frames with a descriptive word representing the salience response of the descriptive word in the image over time in the top left corner. Furthermore, the luminosity of the salient region had a positive relationship with the intensity of salience response in that region.

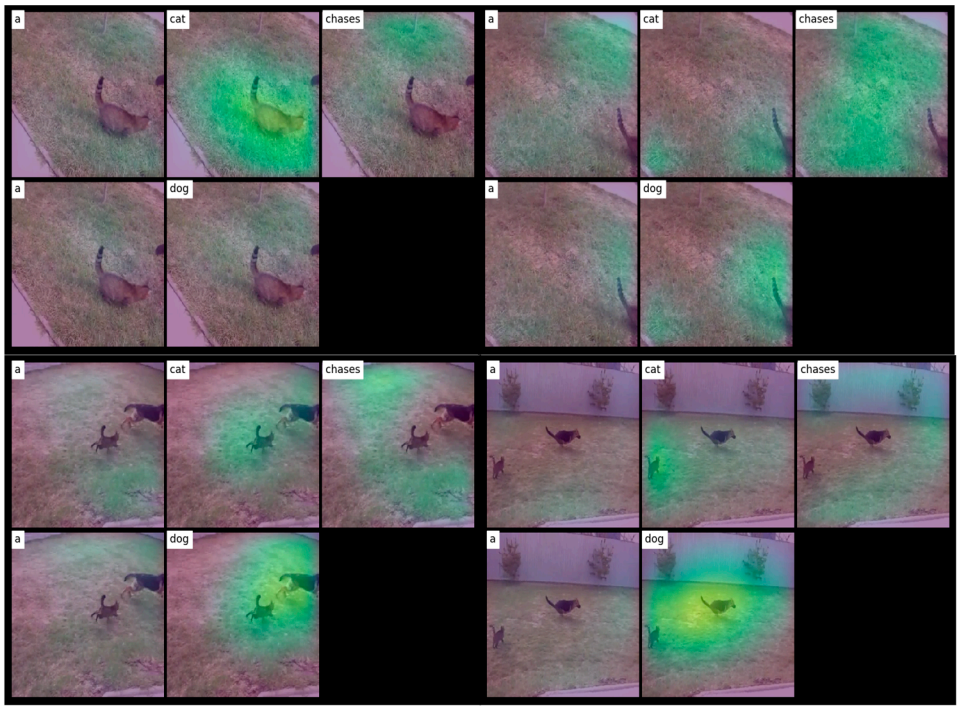
In Figure 5, our model responded effectively to nouns such as "cat" and "dog". Therefore, our model can differentiate nouns well. Additionally, the strongest salience responses to word "cat" were concentrated in the cat. Likely, the salience response to word "dog" is accurate, even with the video running. For articles like "a", because it is not very relevant to the vision, all regions in the video are equally treated, so there would be no salient regions. Similarly, verbs like "'chases", there was no region of salience response initially. However, with the movement of the cat and the dog in the video, some regions of the video frame began to receive responses. Therefore, our model can capture the verbs with temporal-spatial continuity to a degree.

We selected representative images from the dataset. In Figure 6, the figure displays the salience response of the model to human, animal, and objects, respectively. In this picture, it can be seen that our model realized excellent salience responses to different categories and the model can distinguish between different categories. For example, when multiple categories appear in a figure, this model can generate the corresponding response properly.

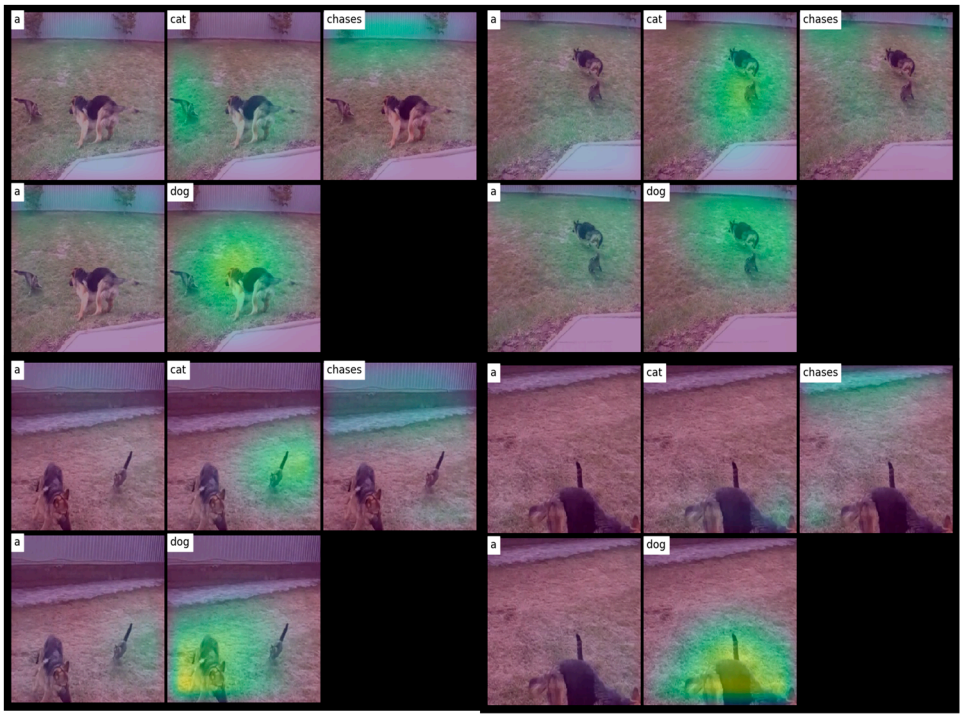


(**a**)

Figure 5. *Cont.*



(**b**)



(**c**)

Figure 5. The visual salience diagram based on video description. (**a**) The beginning of video; (**b**) the middle of video; and (**c**) the end of video.

Figure 6. Salience response to some images.

Meanwhile, in Figure 7, the image contains examples of verbs. For example, there is a verb "play" in Figure 7, Row 1. We can see the response to "play" is mainly concentrated around the guitar, indicating that the model has a certain response ability to the verb.

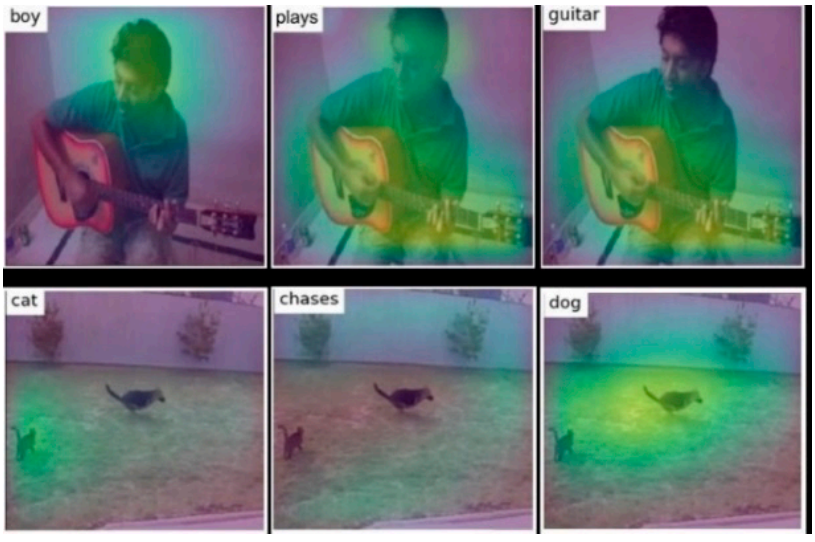


Figure 7. Salience response to the verb.

The salience visualization analysis of the video description illustrates the correlation between visual features and the sentence description. Furthermore, it validated the consistency of visual features and sentence descriptions, and established a corresponding effective model.

5. Conclusions

In this paper, we proposed a video description model based on temporal-spatial and channel attention. In detail, we fully utilized the essential characteristics of CNN and added channel features into the attention mechanism of the model. Therefore, the model can use visual features more effectively and ensure the consistency of visual features and sentence descriptions to enhance the effect of our model. Moreover, our experimental results show that the model has achieved good performance on MSVD dataset. To analyze the model on a larger dataset, we evaluated the model on the MSR-VTT dataset. The results show that our model also has a certain improvement on the MSR-VTT dataset. We also proposed a video visualization model based on the video description and visually demonstrated its effectiveness.

Author Contributions: Conceptualization, J.X. and H.W.; methodology, J.X. and L.L.; validation, J.X. and L.L.; formal analysis, H.W. and L.L.; data curation, L.L. and Q.F.; writing—original draft preparation, J.X., H.W. and L.L.; writing—review and editing, Q.F. and J.G.; visualization, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key Research and Development Program (Grant No. 2016YFB0800105), Sichuan Province Scientific and Technological Support Project (Grant Nos. 2018GZ0255, 2019YFG0191).

Acknowledgments: The authors would like to thank the National Key Research and Development Program and the Sichuan Province Scientific and Technological Support Project for financially supporting this research.

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

References

1. Zeng, Z.; Li, Z.; Cheng, D.; Zhang, H.; Zhan, K.; Yang, Y. Two-stream multi-rate recurrent neural network for video-based pedestrian re-identification. *IEEE Trans. Ind. Inform.* **2017**, *14*, 3179–3186. [CrossRef]
2. Park, J.S.; Rohrbach, M.; Darrell, T.; Rohrbach, A. Adversarial Inference for Multi-Sentence Video Description. *arXiv* **2018**, arXiv:1812.05634.
3. Kim, B.; Shin, S.; Jung, H. Variational Autoencoder-Based Multiple Image Captioning Using a Caption Attention Map. *Appl. Sci.* **2019**, *9*, 2699. [CrossRef]
4. Zhang, J.; Peng, Y. Object-aware Aggregation with Bidirectional Temporal Graph for Video Captioning. *arXiv* **2019**, arXiv:1906.04375.
5. Arachchi, S.P.K.; Shih, T.K.; Hakim, N.L. Modelling a Spatial-Motion Deep Learning Framework to Classify Dynamic Patterns of Videos. *Appl. Sci.* **2020**, *10*, 1479. [CrossRef]
6. Jin, T.; Li, Y.; Zhang, Z. Recurrent convolutional video captioning with global and local attention. *Neurocomputing* **2019**, *370*, 118–127. [CrossRef]
7. Yang, Y.; Zhou, J.; Ai, J.; Bin, Y.; Hanjalic, A.; Shen, H.T. Video captioning by adversarial LSTM. *IEEE Trans. Image Process.* **2018**, *27*, 5600–5611. [CrossRef] [PubMed]
8. Zhao, B.; Li, X.; Lu, X. CAM-RNN: Co-Attention Model based RNN for Video Captioning. *IEEE Trans. Image Process.* **2019**, *28*, 5552–5565. [CrossRef] [PubMed]
9. Nabati, M.; Behrad, A. Video captioning using boosted and parallel Long Short-Term Memory networks. *Comput. Vis. Image Underst.* **2020**, *190*, 102840. [CrossRef]
10. Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; Saenko, K. Translating videos to natural language using deep recurrent neural networks. *arXiv* **2014**, arXiv:1412.4729.
11. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Saenko, K. Sequence to sequence-video to text. In Proceedings of the IEEE Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4534–4542.
12. Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhtdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
13. Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; Courville, A. Describing videos by exploiting temporal structure. In Proceedings of the IEEE Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4507–4515.

14. Hossain, M.S.; Al-Hammadi, M.; Muhammade, G. Automatic fruits classification using deep learning for industrial applications. *IEEE Trans. Ind. Inform.* **2018**, *15*, 1027–1034. [CrossRef]
15. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
16. Papineni, K. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, In Proceedings of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–13 July 2002; pp. 311–318.
17. Vedantam, R.; Lawrence, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
18. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
19. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out: ACL-04 Workshop, Barcelona, Spain, 25–26 July 2004.
20. Zhou, L.; Kalantidis, Y.; Chen, X.; Corso, J.J.; Rohrbach, M. Grounded Video Description. *arXiv* **2018**, arXiv:1812.06587.
21. Pei, W.; Zhang, J.; Wang, X.; Ke, L.; Shen, X.; Tai, Y.W. Memory-Attended Recurrent Network for Video Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8347–8356.
22. Wu, A.; Han, Y.; Yang, Y.; Hu, Q.; Wu, F. Convolutional Reconstruction-to-Sequence for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
23. Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; Courville, A. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv* **2015**, arXiv:1502.08029.
24. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer: Berlin, Germany, 2014; pp. 818–833.
25. Chen, D.L.; Dolan, W.B. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 190–200.
26. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
27. Bird, S.; Loper, E. NLTK: The Natural Language Toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, Barcelona, Spain, 21–26 July 2004; pp. 214–217.
28. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
29. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
30. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Pan, Y.; Mei, T.; Yao, T.; Li, H.; Rui, Y. Jointly modeling embedding and translation to bridge video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4594–4602.
32. Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; Zhuang, Y. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1029–1038.
33. Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; Xu, W. Video paragraph caption using hierarchical recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4584–4593.

34. Zhang, X.; Gao, K.; Zhang, Y.; Zhang, D.; Li, J.; Tian, Q. Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, HI, USA, 21–26 July 2017; pp. 3713–3721.
35. Li, X.; Zhao, B.; Lu, X. MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2208–2214.
36. Zhao, B.; Li, X.; Lu, X. Video Captioning with Tube Features. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 1177–1183.
37. Yue, W.; Jinlai, L.; Xiaojie, W. Video description with subject, verb and object supervision. *J. China Univ. Posts Telecommun.* **2019**, *26*, 52–58.
38. Yan, C.; Tu, Y.; Wang, X.; Zhang, Y.; Hao, X.; Zhang, Y.; Dai, Q. STAT: Spatial-temporal attention mechanism for video captioning. *IEEE Trans. Multimed.* **2020**, *22*, 229–241. [CrossRef]