applied sciences

MDPI

Article

Att-BiL-SL: Attention-Based Bi-LSTM and Sequential LSTM for Describing Video in the Textual Formation

Shakil Ahmed [1,*], A F M Saifuddin Saif [2], Md Imtiaz Hanif [1], Md Mostofa Nurannabi Shakil [3], Md Mostofa Jaman [4], Md Mazid Ul Haque [1], Siam Bin Shawkat [1], Jahid Hasan [1], Borshan Sarker Sonok [1], Farzad Rahman [1] and Hasan Muhommod Sabbir [5]

1 Department of Computer Science, American International University-Bangladesh, 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh; sakil.imtiaz@gmail.com (M.I.H.); mazid@aiub.edu (M.M.U.H.); sb.shawkat@gmail.com (S.B.S.); jahidhasansaif094@gmail.com (J.H.); sonok.sarker06@gmail.com (B.S.S.); farzadrahman59@gmail.com (F.R.)
2 Institute IR4.0 (IIR4.0), Universiti Kebangsaan Malaysia, Bangi 43600 UKM, Selangor, Malaysia; saif@ukm.edu.my
3 Skill Development for Mobile Game and Application Project, ICT Division, E-14/X, ICT Tower, Agargaon, Sher-e-Bangla Nagar, Dhaka 1207, Bangladesh; shakilcse9@gmail.com
4 Head of Learning & Development and Public Relations at Genex Infosys Limited, Adjunct Professor at Southeast University, Nitol Niloy Tower, Nikunja-2, Dhaka 1229, Bangladesh; jaman.ites@gmail.com
5 Department of Electrical and Electronics Engineering, BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh; hasan.sabbir@g.bracu.ac.bd
* Correspondence: ahmed.shakil.v3@gmail.com; Tel.: +880-1761-290093

Abstract: With the advancement of the technological field, day by day, people from around the world are having easier access to internet abled devices, and as a result, video data is growing rapidly. The increase of portable devices such as various action cameras, mobile cameras, motion cameras, etc., can also be considered for the faster growth of video data. Data from these multiple sources need more maintenance to process for various usages according to the needs. By considering these enormous amounts of video data, it cannot be navigated fully by the end-users. Throughout recent times, many research works have been done to generate descriptions from the images or visual scene recordings to address the mentioned issue. This description generation, also known as video captioning, is more complex than single image captioning. Various advanced neural networks have been used in various studies to perform video captioning. In this paper, we propose an attention-based Bi-LSTM and sequential LSTM (Att-BiL-SL) encoder-decoder model for describing the video in textual format. The model consists of two-layer attention-based bi-LSTM and one-layer sequential LSTM for video captioning. The model also extracts the universal and native temporal features from the video frames for smooth sentence generation from optical frames. This paper includes the word embedding with a soft attention mechanism and a beam search optimization algorithm to generate qualitative results. It is found that the architecture proposed in this paper performs better than various existing state of the art models.

Keywords: video captioning; Bi-directional long short-term memory; attention-mechanism; video to text; video description generation

check for updates

Citation: Ahmed, S.; Saif, A.F.M.S.; Hanif, M.I.; Shakil, M.M.N.; Jaman, M.M.; Haque, M.M.U.; Shawkat, S.B.; Hasan, J.; Sonok, B.S.; Rahman, F.; et al. Att-BiL-SL: Attention-Based Bi-LSTM and Sequential LSTM for Describing Video in the Textual Formation. *Appl. Sci.* **2022**, *12*, 317. https://doi.org/10.3390/app12010317

Academic Editor: David Megías

Received: 8 November 2021
Accepted: 21 December 2021
Published: 29 December 2021

1. Introduction

Throughout recent years, people are having easier access to a massive proportion of visual data in various online platforms, which usually contain the following—sound, visual scene, and sometimes textual data. The advancement of hand-held contraptions and client devices prepared for obtaining chronicles has been colossal. Moreover, the amount of recording videos is increasing in online platforms that are in different forms. A report showed that web traffic recordings might increment to 82% in 2021, which announced 73% of total traffic in 2016 (e.g., Netflix and YouTube) [1]. These integrated multiple-sourced

pieces of information require more analytic processing powers and also require a huge amount of storage space. By bearing in mind these gigantic measures of video information, the end-users cannot explore required data in an efficient manner. A succinct, video synopsis featuring the essential pieces of the video will help in the ordering and quicker recovery of the required information. It will likewise be valuable in the route, notice, and information investigation throughout a large video.

Researchers are very keen to find various techniques to find solutions that can address the effective access of the video data. They performed various video summarization techniques such as visual frame reductions [2–7] and described the video in a textual formation [8–35]. Frame reduction by applying deep learning approaches is a method to discard the video's useless or low attention frames. That process reduces the run-time of a video and saves some storage space. The textual-based summarization of visual recordings is much more efficient than frame reduction techniques. Nowadays, visual and textual data are integrated. Researchers performed a crucial work for developing such deep learning methods to describe a video in the textual form [11–13]. Video summarization in textual formation is addressed to the problem of producing textual data from visual contents. Visual textual formation is one of the successors of image captioning [36–42]. In past exploration, the researchers utilized different Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and so on for generating image and video descriptions [10–20]. The outcomes from the past researchers demonstrated promising results in their works. Semantic measurements from visual substance comprehension through different deep learning techniques improved by the researchers have demonstrated the productivity and the effectiveness of video captioning.

The encoder-decoder model is one of the utmost communal techniques for the textual description generation from the video. Usually, an encoder consists of CNN, where CNN remains extracting features from the video frames, and a decoder, with the RNN model responsible for the sentence generation. However, with the rapid advancements of various neural network, researchers have now incorporated RNN features in the encoder stage [10,11,13–17]. They also found promising results from their RNN based LSTM frameworks. The LSTM is suggested to exhaust the evaporating angles issue by empowering the network to realize when to overlook past hidden positions and refresh hidden positions by incorporating memory cells. Therefore, we proposed a method called "attention-based bidirectional LSTM and sequential LSTM" (Att-BiL-SL) encoder-decoder model for describing the video in textual format. While most of the existing models work with single layered bi-LSTM, LSTM, or GRU approaches, the model of this study consists of two layers of attention-based bi-LSTM and one layer of sequential LSTM for video captioning. The model also extracts the universal and native temporal features from the video frames for smooth sentence generation that was absent in some previous models. With the mentioned two-layered attention-based bi-LSTM and one-layered sequential LSTM for extraction universal and native features, this study also includes word embedding with a soft attention mechanism in the encoder section to generate qualitative results. Through rigorous analysis, it is found that the model of this research work has outperformed various existing models.

The residue of this paper is organized as follows. We review the previous video description generation methods and try to find the limitations of those models in Section 2. The proposed framework is illustrated in Section 3, accompanied by the overall process. In Section 4, we discuss the experimental procedure along with implementation details. In Section 5, the proposed model is compared with previous baseline models. Moreover, this section reveals the results of our model, which overtakes the various previous model. In Section 6, we discuss the quantitative analysis and limitations of our proposed method. Lastly, Section 7 accomplishes the paper by deliberating the model results and the possibility of future work.

2. Related Work

2.1. Template Matching

Early video textual formation strategies [43,44] generally utilized predefined formats to produce video portrayals. The conceivable semantics with visual classifiers were first identified here and were subsequently fed in the required formats into the model. It is instinctive that the consequences of format put-together strategies are exceptionally reliant concerning the predefined formats' nature. The past work was the predefined punctuation rules sent in a layout-based strategy comprised of the subject, verb, and object represented as <SVO> design [45–58]. This technique can create literary development from a video in syntax-based principles and that is simple to actualize. In any case, it neglects to perceive accurate highlights, activities, and targets through a whole video. There are some lost, misidentified, and wrong information eliminated after the video. This strategy relies upon video substance, and yield collection suffered because of the surmised limitation. Additionally, those strategies are an earlier procedure to caption a video.

2.2. End-to-End Model

In the latest years, analysts are centered on different deep learning ways to deal with video captioning. End-to-end encoder and decoder approaches are likewise becoming well known for their promising outcomes [9–11,13]. In this model, the encoder encoded the component vectors from the basic structures through a neural network. In addition, the decoder comprised of a repetitive neural network interprets the highlights and produces sentences. LSTM is a distinct type of RNN with the ability of disappearing gradient problems. To understand the temporal dynamics of skeletal sequences, several LSTM-based models were applied previously. An attention-based spatial-temporal graph convolutional LSTM network was proposed by [59] for recognizing the karate action from the videos. Using the benefits of the Delaunay technique, the authors found effective spatiotemporal structure information of actions in karate and improved the feature extraction by applying the attention-enhanced graph convolutional LSTM networks. They also constructed a small and researchable dataset for karate action and tactics recognition. Not only are LSTM-based approaches utilized in image recognition systems, but they may also be used to identify and solve a wide variety of problems. For example, they are employed in time-series forecasting [60], energy consumption forecasting [61], and metallic gear life span prediction [62].

Although an RNN is comprised of LSTM or GRU for generating video descriptions, those models likewise presented a consideration instrument for taking a gander at the semantic area in a video outline, as far as producing printed depictions [15,17,20]. In those models, CNN alludes to the semantic areas of the edges, and RNN fuses the following indispensable words to deliver sentences. Those models were exceptionally prepared to utilize the different CNN and RNN models. Gao et al. [15] presented consideration systems with LSTM [52] to distinguish notable developments of the visual scene, and the technique created sentences with unexpected semantic substances. They considered the 2D and 3D CNN portrayal and an LSTM decoder to create a particular word and subsequently utilized the multimodal word-embedding framework to produce the video depiction semantically. In any case, the issue with this model is that the model can only deliver a single line sentence from the visual substance, and different ongoing models are producing the best outcomes rather than this captioning model. The Bi-LSTM [16] was likewise utilized for video captioning. Visual information was encoded with a forward and reverse pass extracted from the VGG-16 CNN [42] architecture and merged the sentences in a sequential model. The technique can create each word recurrently in every period. The model presented just single-layer Bi-LSTM and solely integrated global temporal features.

Video Response Map (VRM) with medium-level attention to generate video descriptions was proposed in [8]. Authors utilized the VGG-16 [42] pre-trained on ImageNet [57] for separating the video features and went with the highlights through twofold LSTM to create descriptions. The shaded video frames were taken into consideration, the model

was compared among the past cutting-edge models, and it resulted in promising outcomes. Sah et al. [30] proposed a semantic content synopsis of long recordings. The highlights from the visual substance were successively removed and later passed through a decoder to deliver textual depictions. Moreover, the long captions subsequently went through another RNN to deliver the Neural Language Generation (NLG) based textual summarization. Another decent model proposed by Xu et al. [12] enhanced the audiovisual captioning superiority by joining semantic concepts with audio amplifier features. A video descriptive model, combining spatial and temporal analysis, was proposed by Danny et al. [18] that can take care of different areas in given regions, dependent on what was shown in the past regions. A technique to generate numerous sentences from the visual data has been proposed by Song et al. [14]. The strategy expanded the RNN model by past and exact next dissemination with a non-linear layer. This technique can deal with vulnerability by applying a hidden state. The model additionally neglected to catch the high-level temporal structure of the visual frames.

Another approach can be found, which is a double-stream RNN for a video description in a textual formation [19]. The method focused on the hidden state of both visual and semantic streams to produce meaningful sentences from the visual contents. The structure incorporated a semantic component extraction strategy and twisted it with word embedding. At that point, the visual and semantic stream together produced an important textual description from a visual scene. Another approach that included adaptive attention and mixed loss optimization for performing video captioning was presented by Xiao et al. [21]. The technique included a reinforced adaptive attention mechanism. The procedure was likewise prepared in the word-level loss and sentence-level loss model. The strategy additionally tackled the biased issue of those sorts of losses. A stateful human-centered visual captioning system that can extract facial and visual features from the visual contents and pass through a two-layer LSTM with an attention mechanism to produce efficient video caption in which the model was able to extract human actions and objects [22].

An interesting method was proposed using a refocused video encoder [23]. The method with refocused RNN with spatial-visual features was introduced for better video captioning. The refocused video encoder looked up two times with predicted key features to avoid unwanted temporal information. The spatial information from the different regions was passed through a decoder with detailed features. A semantic enhanced encoder-decoder network method with LSTM was used by researchers in [24]. The method presented extracts motion features, appearance features, and global features for video description generations. The method also introduced reinforcement learning, which can ensure a better quality of sentence generation in previous research.

The structure presented in [25], incorporates an action module and motion module for sports video captioning, which is ready to perceive human-related activities with objects during sports. The structure also introduced a pose attribute detection module and a description generation module, the entire system, followed by an attention mechanism. The content branch and the semantic branch encoder were introduced by Liu et al. [26]. The method illustrated a very promising result in terms of two different datasets. The RNN based decoder with soft attention produces the semantic sentence for visual inputs. A model that can produce better captioning due to maintaining the sentences' semantic attention is presented by Xiao et al. [27]. The model introduced two independent attention mechanisms to point the soft attention through visual features and attributes another attention focused on visual signals and text documentation. The framework proposed in [28] introduced an extended version of the encoder-decoder system to overcome the manual selection process. A modal attention network can dynamically point out the important features from the video frames and generate a sentence. Chen et al. [29] proposed a video to text summary method, which also generated the textual caption of videos, and then they again summarize the generated sentences to produce a short description of visual content. Those methods are the footprint of various visual summarization methods.

A joint model to generate video descriptions and produce abstractive summarization from those visual portrayals that can deliver visual captioning and summarization is depicted in [31]. It is one of the cutting edge and productive habits of textual summarization from visual content. Authors in [32] presented a semantic and temporal attention-based technique for a video description in textual formation. The model can expressly join with the high-level visual idea to create temporal attention and included an encoder to extricate the highlights from a visual substance and decoder for the sentence generation. A scene-edge GRU strategy for video descriptions was presented by Hao et al. [33]. The technique presented another encoding framework that can experience the skeleton of a video outline, and the GRU can be ready to search for suspending video frames, which can be prepared with no prior annotation information. The method also outperformed in terms of S2VT models. "Spatio-temporal ranked attention network" for the video description generation in textual format has been performed by Cherian et al. [34]. The two-way strategy was utilized such as spatio-temporal and temporal-spatio. The model is comprised of LSTM based temporal ranking capacity that can powerfully catch activities from the visual frames. The setting outperformed in various datasets and broke the previous top-tier benchmark records.

It can be observed that most of the related models incorporate single layered bi-LSTM, LSTM, or GRU strategies to generate video captions. However, the novelty of this research work is that it proposes a two-layer attention-based bi-directional LSTM as an encoder and a dedicated single layer forward pass LSTM as a decoder for generating video description semantically. While different CNN models of other studies mine the video structures' features to suckle into the encoder and decoder model for the enriched video description generation, our video depiction model, in contrast with the others, is fit for extricating different high-level features. This study also includes an efficient word embedding with a soft attention mechanism in the encoder section to generate qualitative results. We have positioned towards better extricating visual data from the frames, which is revealed in the outcomes. As far as multifaceted computational nature, the proposed model is very reasonable for predicting inconspicuous information.

3. Proposed Methodology

The proposed framework of our video captioning is demonstrated in Figure 1. Our framework includes input video, feature extraction module, encoder module, decoder module, and a sentence generation portion. The method is sequential, and the outcome of the model comprises sentences. This section has discussed the inclusive framework of our model for text generation from visual frames.

3.1. Feature Extraction

To appreciate the visual substance in a video, an elevated level component preparing stage is imperative for separating film attributes. These characteristics are discovered using CNN on various gages that are steady to helper varieties. The previous studies [42], [57] discussed different pre-trained neural networks utilized to catch the highlights through the video frames.

The video is the input of our proposed model, and the video is differentiated into a sequence of visual frames. Every frame is fed into CNN for extracting the features. We have implemented 2D CNN for 2D feature extraction. The 2D CNNs recognize the segmentation map for a solo frame using 2D convolutional kernels. Predictions for segmentation maps for a complete volume are made by one frame at a time. To create predictions, the 2D convolutional kernels can use context throughout the frame's height as well as width (256×256). We have used the pre-trained models VGG-16 [42], InceptionV3 [47], and Xception [48] for mining the 2D features from each visual frame. All those models pre-trained on ImageNet [57] dataset contain approximately 1.2 million images of over 1000 classes. The VGG-16 [42] model consists of 16 layers; the first 13 layers are convolutional layers, and then the three dense layers are for classifying images. The InceptionV3

model consists of 48 layers, and the extended Inception or Xception contains 71 layers. For transfer learning, the last fully connected layer is removed from the model. The 2D CNN models are chosen for reducing the training time for the large MSVD [41] and MST-VTT [35] dataset. Inspired by [11], for the extraction of action features, we have used an Inflated 3D network (I3D) [50] pre-trained in kinetics dataset. The 3D CNNs predict the segmentation maps for the adjacent frames using 3D convolutional kernels. The prediction of the 3D convolutional kernels can use the 2D convolutional kernels incorporated with an additional dimension ($256 \times 256 \times 8$). Another factor of 3D CNN is to consider the training pace. It consumes more time and requires expensive equipment. Therefore, we have used UCF-101 [49], a small dataset for motion feature extraction, and used Faster R-CNN [51] for facial recognition in some cases. At that time, the features are concatenated and fed into an attention-based bi-LSTM encoder.

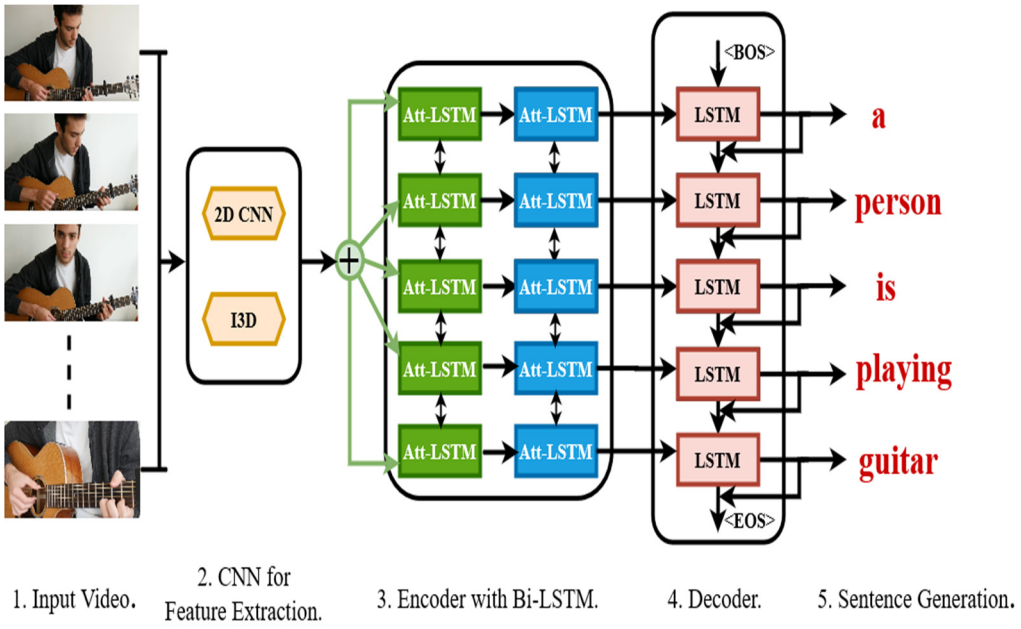

Figure 1. A comprehensive framework of our proposed video description generation model. Here, (**1**) the input of the model is video. (**2**) 2D and I3D CNN are used for feature extraction from the input visual frames. (**3**) The model contains attention-based bi-directional LSTM as an encoder. (**4**) A forward pass LSTM is used for sequential sentence generation. (**5**) Lastly, the sentence generation module is used for captioning the video.

3.2. Encoder-Decoder with Bi-LSTM and Sequential LSTM

LSTM [52] is a distinct version from RNN. LSTMs are expressly intended to keep away from the long-term dependency issue. LSTM consists of several gates, such as the input gate, forget gate, output gate, and a constant memory cell. The LSTM based sequential model is followed by those equations where i_t, f_t, *and* o_t represent the input gate, forget gate, and output gate, respectively, in Equations (1)–(3).

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{3}$$

$$g_t = tanh\left(W_g\left[h_{t-1}, x_t\right] + b_g\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot tanh\left(c_t\right) \tag{6}$$

If the input sequence $(x_1, \ldots, x_t)$, is given to the LSTM, it computes the hidden control unit $(h_1, \ldots, h_t)$, the cell memory sequence $(c_1, \ldots, c_t)$, and b denotes the bias vector. Here, $\odot$ represents the element-wise multiplication, and *tanh* is the non-linear hyperbolic function.

However, the sequential model of LSTM only focuses on present sequences in the temporal order. In our proposed model, we have used bi-LSTM [53] as an encoder, which not only considers the present state but also considers the future state. It improves the standard LSTM network by including another layer. The two layers are opposite in direction and can predict information both from the previous and the forthcoming by applying forward and backward pass. Here, Figure 2 illustrates the traditional LSTM [52] and bi-LSTM [53], and Equation (7) is the equation of bi-LSTM.

$$h_i = \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i}\right] \tag{7}$$

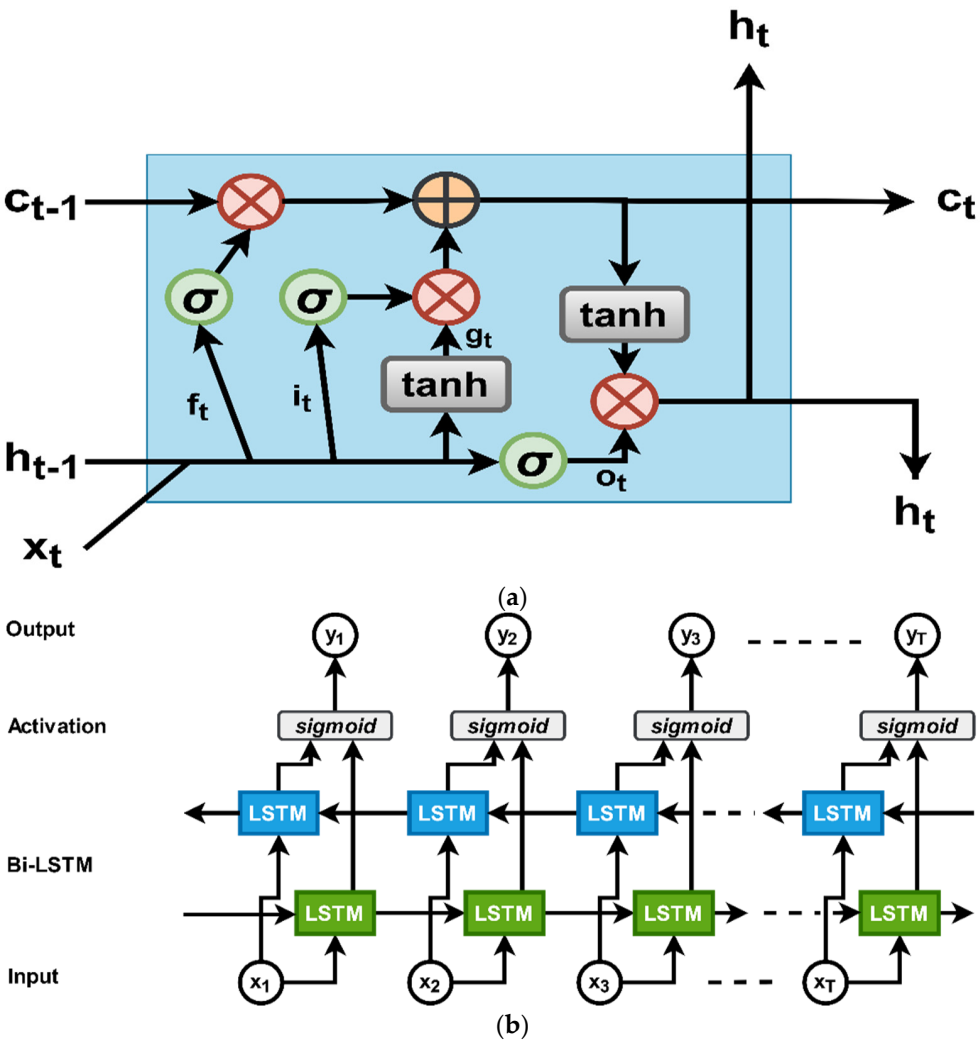

Figure 2. An illustration of LSTM and a bi-directional LSTM model. In our proposed model, we have used bi-LSTM as an encoder and sequential LSTM as a decoder. (**a**) The illustration of an LSTM diagram. Here, the forget gate, input gate, and output gate denote as f_t, i_t, *and* o_t, respectively. (**b**) The illustration of a bi-LSTM model. Two layers, such as the forward pass and backward pass LSTM, followed by the sigmoid function.

Here, $\oplus$ represents the element-wise summation to chain the onward and backward pass LSTM.

The decoder consists of the sequential LSTM [52], which takes input from the encoder. The output y_t and hidden state h_t update based on its earlier position y_{t-1} and h_{t-1} based on time step t. If the encoder embedding represents V as a context vector, then the decoder's equation as Equation (8).

$$\begin{bmatrix} y_t \\ h_t \end{bmatrix} = Dec\ (h_{t-1},\ y_{t-1},\ V) \tag{8}$$

We propose an attention-based bi-LSTM [53] as an encoder in our framework. In the next subsection, we have discussed the attention mechanism of our encoder system.

3.3. Attention Mechanism

Nowadays, attention-based neural networks have shown tremendous achievement in query answering, machine paraphrases, speech acknowledgment, and image/video description generation. Various attention mechanisms have been proposed previously; among them, we have considered the soft attention model inspired by [25] to obtain a dynamic weighted sum of the extracted features from the video frames. Instead of utilizing a frame as an input to the LSTM, we used weighted image characteristics that compensated for attention in soft attention. By multiplying the related segmentation map with a low weight, soft attention discredits unimportant parts of the frame. Therefore, high attention zones retain their original worth, whereas low attention areas approach zero. As a result, the LSTM can generate more accurate predictions.

If we consider that the feature vector F consists of $(F_1,\ \ldots,\ F_n)$ features, we can denote F_i as $F_i\ \in (F_1,\ \ldots,\ F_n)$ and the importance score $s_i^t = \left(s_1^t,\ \ldots,\ s_n^t\right)^T$.

$$s_i^t = W_s\ tanh\left(W_f F_i + W_h h_{t-1} + b_s\right) \tag{9}$$

$$\alpha_i^t = \frac{exp\left(s_i^t\right)}{\sum_{i=1}^{n} exp\left(s_i^t\right)} \tag{10}$$

$$F_t' = \sum_{i=1}^{n} \alpha_i^t F_i \tag{11}$$

Here, $W_s,\ W_f,\ and\ W_h$ are the weight vectors for training and the bias vector b_s. However, α_i^t is the attention at time t. The previous state of the hidden stage (h_{t-1}) and the i-th features of the video frames proceed the standardized weight score s_i^t. Then, the features of the visual frames feed into the bi-LSTM encoder at time t. Finally, the weighted sum of the frame features is calculated through Equation (11). Figure 3 illustrates the bi-LSTM [53] with an attention mechanism for generating meaningful sentences from the visual frame.

3.4. Sentence Generation

The main purpose of the proposed model is to produce sentences from visual input. Previous video description generation approaches generally share a regular visual model and language model, which may prompt extreme information loss. Moreover, most of the time, they neglect the temporal structure of sentences. For that reason, the result of the generated sentence may not be accurate due to duplicating the input. So, we generated sentences from a sequential LSTM model [52] by binding with visual representation and we generated the word recurrently at every time. So, for generating the sentence, we produced the words recurrently for the i-th sentence formulated as Equation (12).

$$Prob(d_t^n | s_{1:i-1},\ d_{t-1}^n,\ F_t; \theta) \tag{12}$$

where F_t is the feature vector, d_{t-1} denotes the last words in the i-th sentence, $s_{1:i-1}$ denotes prior sentences, and θ denotes all the parameters for generating the sentence. The

cost function of the generating sentence is a negative logarithm and is formulated as Equation (13).

$$Loss_{sen} = -\sum_{t=1}^{N} \log(Prob(d_t^n | s_{1:n-1},\ d_{t-1}^n,\ F_t; \theta)) \tag{13}$$

Here, N means all the words in a sentence. By reducing the "$Loss_{sen}$", the circumstantial connection amongst the words in the sentence can be generated clear and even. Inspired by [16], from Equation (13) for our model, we can obtain optimal θ by the Equation (14).

$$\theta^* = argmax \sum_{t=1}^{N} log(Prob(d_t^n | s_{1:n-1},\ d_{t-1}^n,\ F_t; \theta)) \tag{14}$$

Here, the θ^* updates θ by the optimizer in the whole training process. Backpropagation is used for the loss, and the individual LSTM part figures out how to determine a hidden state h_t from the input arrangement. At that point, we execute the $Softmax$ capacity to acquire the possibility of apportionment over the words in the entire vocabulary.

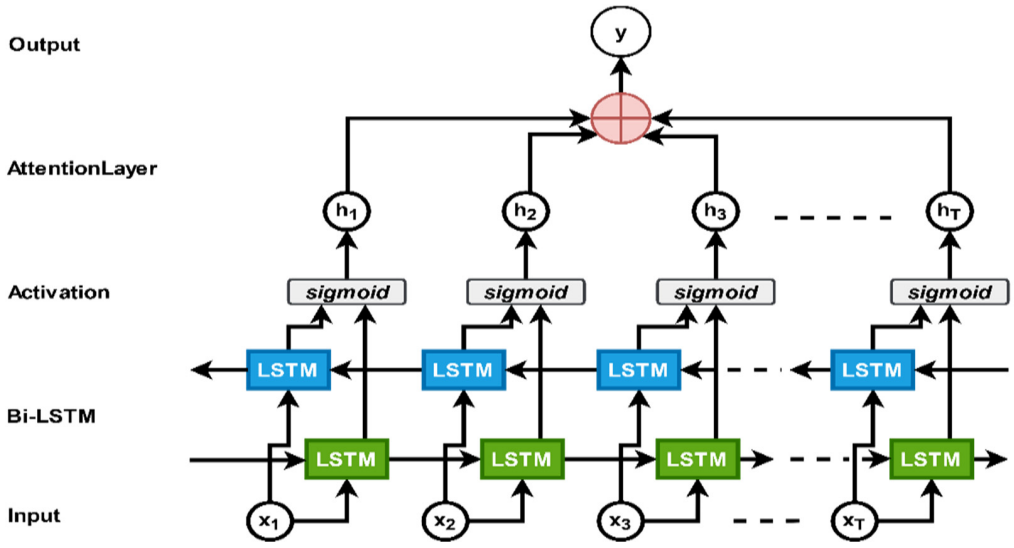

Figure 3. Bi-LSTM with an attention mechanism. For our proposed model, we used the attention mechanism with bi-LSTM as an encoder.

For sentence generation, we used the starting token as <start> to start the sentence and the ending token <end> to stop the sentence, as illustrated in Figure 1. In the next section, we have discussed our model's experimental procedure and analyzed the result produced by our model.

4. Experiment

This section describes the various datasets, evaluation metrics, implementation details, experimental results, and previous methods of our proposed method in detail.

4.1. Datasets

The proposed model has been evaluated on the most popular two datasets, Microsoft Research Video Description (MSVD) [41] and Microsoft Research Video to Text (MSR-VTT) [35].

(i) MSVD [41]: The dataset encloses 1970 YouTube video cuts. The dataset holds approximately 80 k sentences with a vocabulary size of 13,010. The absolute term of this dataset is about 5.3 h. The dataset was split into 1200 recordings for training clips, 100 recordings for validation clips, and the rest for testing.

(ii) MSR-VTT [35]: The dataset is one of the large video datasets, comprises 10 k video cuts alongside an all-out season of 41.2 h. It has 20 unique classes similar to sports, music, gaming, and so forth—the typical length of those clips is around 20 s with a colossal vocabulary of 29,316. The dataset has 6513 training clips, 2990 validation clips, and 497 testing clips.

4.2. Evaluation Metrics

Such as conventional machine interpretation, the produced expressive sentences for the correlative video can be estimated by looking at many reference sentences [46]. As of late, some regular measurements in machine interpretation likewise are utilized for assessing visual subtitling, i.e., Bi-Lingual Evaluation Understudy (BLEU) [37], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [38], Consensus-Based Image Description Evaluation (CIDEr) [39], and Recall Oriented Understudy of Gisting Evaluation (ROUGE) [40].

(i) BLEU [37]: It is one of the most notable measurements for assessing machine-produced sentences. It can be ready to figure out the correspondence between machine-created sentences and ground truth sentences. This measurement can give the best outcome regarding short sentences. The metrics BLEU (N = 1, 2, 3, 4) typically evaluate N-gram matches' precision in a sentence.
(ii) METEOR [38]: It is one of the most utilized measurements for valuation machine-produced sentences. It has unigram-based accuracy and review rates. It has the highlights of equivalent coordinating, stemming, and definite word coordinating. The metrics are more robust than other metrics in terms of the human verdict.
(iii) CIDEr [39]: This measurement is primarily utilized for image subtitling. Video is only a picture containing different ceaseless pictures. It can be ready for an agreement between machine-created sentences and human-explained sentences.
(iv) ROUGE [40]: ROUGE is another measurement for valuation. It has highlights of n-gram events. It has four distinct kinds of measurements, where ROUGE-L is generally utilized for visual inscription assessment.

4.3. Implementation Details

(i) Hardware setup: The experiments have been performed on our local personal computer, which contains an AMD Ryzen 3600x, 6-Core processor (3.80 GHz–4.4 GHz) with 16 GB of DDR4 RAM (3200 MHz). For faster computation, a single Nvidia RTX 3070 GPU (5888 CUDA-Cores) with 8 GB of video memory was also used.
(ii) Preprocessing: For video preprocessing and 2D global temporal feature extraction, the model was tested similarly by dividing 25 frames for every video. A VGG-16 model [42] pre-trained on the ImageNet dataset [57] was used for separating visual appearance highlights. We resized the frames to 224 $\times$ 224 resolution. The VGG-16 model selected an arrangement of 4096-dimensional element vectors delivered by the completely associated with the fully connected fc7 layer. For the InceptionV3 [47] and Xception [48], the pre-trained model, we resized the frames to 299 $\times$ 299 resolution. Moreover, to extract the motion, we used an Inflated 3D network (I3D) pre-trained in kinetics [50] dataset. Then, the classifier layer was tweaked from all the models. The model was very efficient for action identification. We used faster R-CNN [51] for local features such as face detection. In addition, for the sentence portion, we have used word tokenizer NLTK [54] based tool and set the vocabulary of every dataset in the lower-case and removed the punctuations from the sentence. The total vocabulary size of the two datasets is approximately 42 k.
(iii) Experimental Setup: We have proposed an attention-based bidirectional-LSTM model [53] for video captioning. The model can predict both the past and future words and combine them to generate a meaningful sentence. Both forward and in reverse data can be handled to catch relevant data. The attention-based two-layered LSTM [53] has 512 hidden nodes in each layer. The visual substance is separated from the various

proposed CNN models [42,47,48] and the primary layer, along with an attention mechanism to feed the decoder. Moreover, the decoder with sequential LSTM [42] and the attention layer also consists of 512 hidden nodes in each layer. The regularization dropout layer is used in our proposed model. We set the dropout 0.5 at both the encoder and decoder levels. We used the ADAM optimizer to generate the loss function. We defined the starting learning rate 10^{-5} to neglect the gradient burst. In addition, we included a beam search algorithm with size 5 to produce the textual description. Then, we trained our model with a batch size of 64.

In the training stage, to manage sentences with fluctuating spans, we included the <start> and <end> tags to begin and end the sentence. Inspired by [15], we trained our model over 500 epochs and monitor the valuation metrics. If the valuation metrics did not progress in the validation set, we stopped the training in 25 epochs. Lastly, we used the Tensorflow [55] and Keras [56] libraries to implement our model.

5. Experimental Results and Comparison

For our Att-BiL-SL model, we considered the most popular freely available MSVD [41] and MSR-VTT [35] datasets and evaluated the generated captions with automatic machine translator evaluation metrics. BLEU [37], METEOR [38], ROUGE-L [40], and CIDEr [39] are very communal assessment measurements on image and video depictions. The foremost three were primarily planned to assess machine interpretation at the soonest, and CIDEr was proposed to assess the image portrayal with adequate reference sentences. The METEOR could catch the semantic viewpoint since it recognizes all potential matches by extricating precise matches, stem matches, reword matches, and equivalent matches by utilizing the WordNet database and computing sentence level likeness scores per matching loads. To quantitatively assess the presentation of our methodology, we received all the metrics in light of their powerful execution. Here, Table 1 reports the performance measurements of our proposed model.

Table 1. The performance evaluation of our model in MSVD and MSR-VTT datasets. Here, the V, IV3, X, and I3D denote the VGG-16 net, InceptionV3 net, Xception net, and Inflated 3D network, respectively. Text with bold marks denote the best results in evaluation metrics (percentile score).

Model	MSVD				MSR-VTT			
	B-4	**M**	**C**	**R-L**	**B-4**	**M**	**C**	**R-L**
Att-BiL-SL (V + I3D)	48.3	32.5	85.3	68.8	39.0	27.8	46.2	57.8
Att-BiL-SL (IV3 + I3D)	49.1	33.2	86.4	70.3	40.8	28.3	48.0	59.8
Att-BiL-SL (X + I3D)	**51.2**	**35.7**	**86.9**	**72.1**	**41.2**	**28.7**	**49.3**	**60.4**

Table 1 shows the various scores that we evaluated by applying various evaluation metrics. The evaluation metrics are BLUE@4 (B-4), METEOR (M), CIDEr, and ROGUE-L (R-L). The proposed model considered the two most popular datasets, such as MSVD and MSR-VTT, that are publicly available. According to the MSVD dataset, the model with Xception + Inflated 3D scored **51.2%** in B-4, **35.7%** in M, **86.9%** in C, and **72.1%** in R-L, which is the best possible result of our entire proposed model. In the MSR-VTT dataset, our model with Xception + Inflated 3D scored **41.2%** in B-4, **28.7%** in M, **49.3%** in C, and **60.4%** in R-L. We found the best results from the Xception + Inflated 3D attention-based Bi-LSTM and Sequential LSTM CNN model in both datasets. The pictorial view of the outcomes of our proposed method is illustrated in Figures 4 and 5.

Our proposed model shows promising results in all automatic machine translator evaluation metrics. The model trained in three different CNN models for feature extraction and all of them performed very significantly. All the CNN models passed through the same attention-based Bi-LSTM and Sequential LSTM (Att-BiL-SL) encoder-decoder model. Our Att-BiL-SL with VGG net [42] beat many previous baseline methods in two popular datasets. The InceptionV3 net [47] also outperformed in most cases, but the Xception net

model [48] beat most of the previous video captioning models. Here, Figure 4 illustrates the qualitative sentence generation of our Att-BiL-SL method. We compared the qualitative results with the human-annotated ground truth sentences. Although our model generated some relevant but incorrect sentences in some cases, the method has shown the incorrect sentence generations. Figure 5 illustrates the relevant but incorrect and incorrect sentence generations of our method compared with ground truth sentences.

Att-BiL-SL (V + I3D): A person is slicing a carrot.
Att-BiL-SL (IV3+ I3D): Person is cutting a carrot.
Att-BiL-SL (X + I3D): A person is peeling and slicing a carrot.
Ground Truth: A person is peeling and cutting carrot.

Att-BiL-SL (V + I3D): A man is riding motorcycle.
Att-BiL-SL (IV3+ I3D): A man is racing bike.
Att-BiL-SL (X + I3D): Bikers are racing bike on the road.
Ground Truth: Some bikers are racing bike on the road.

Att-BiL-SL (V + I3D): Men are running in the field.
Att-BiL-SL (IV3+ I3D): Many people are running.
Att-BiL-SL (X + I3D): A group of people is running.
Ground Truth: A group of people is running.

Att-BiL-SL (V + I3D): Two persons are fighting.
Att-BiL-SL (IV3+ I3D): Two men are fighting.
Att-BiL-SL (X + I3D): Two men are fighting in a stage and people are watching.
Ground Truth: Two men are boxing in the stage and wearing shorts.

Att-BiL-SL (V + I3D): A woman is showing her makeup.
Att-BiL-SL (IV3+ I3D): A girl is showing some makeup tips.
Att-BiL-SL (X + I3D): A girl is applying makeup on her face.
Ground Truth: A girl is putting makeup on her face.

Figure 4. The sentence generation of our proposed model from several testing videos. Here, V, IV3, X, and I3D denote the VGG-16, InceptionV3, Xception, and Inflated 3D network individually. We compared the generated sentences with the human-annotated ground truth (GT) sentences.

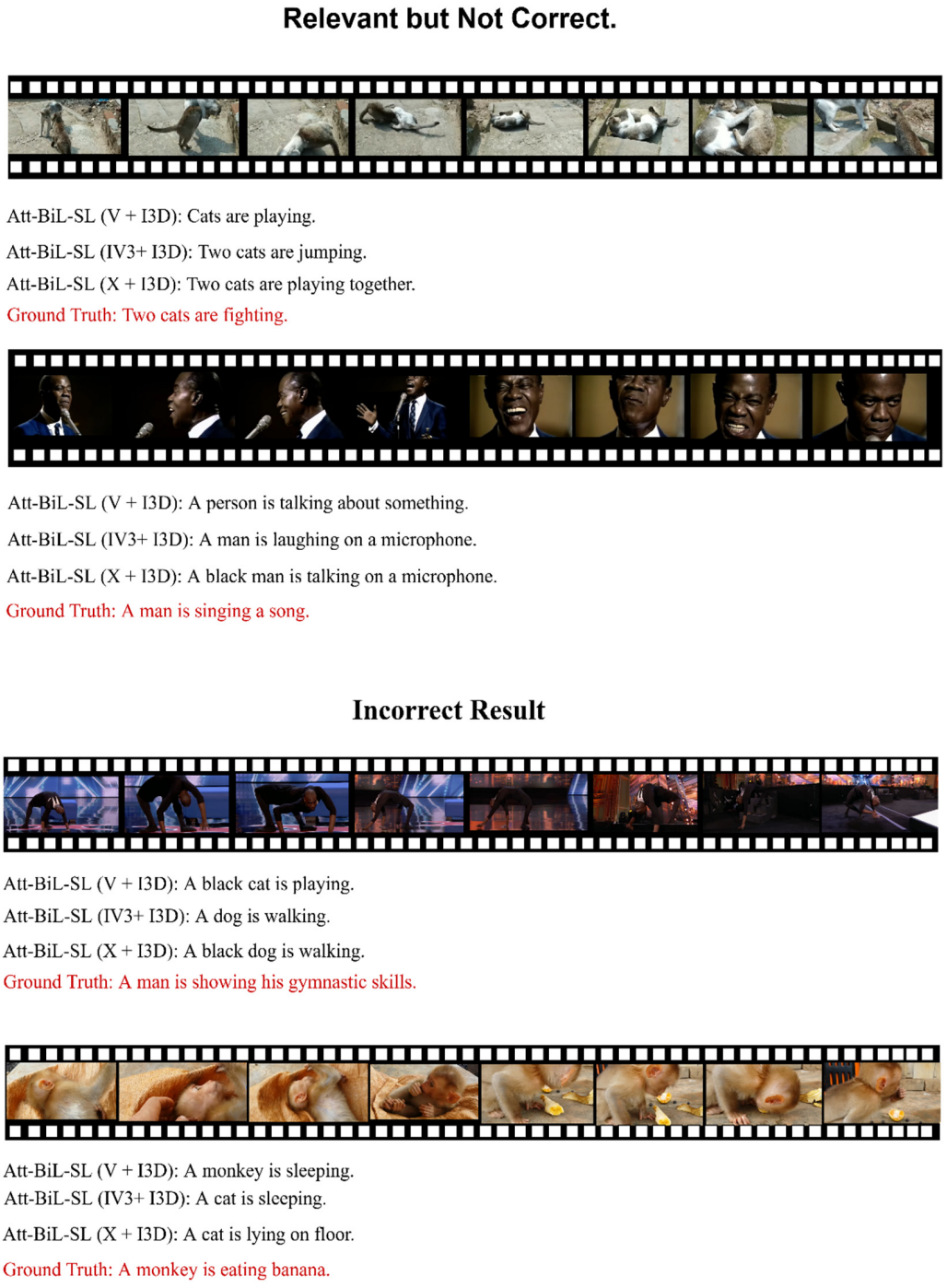

Figure 5. The example of some relevant video captioning but incorrect sentence generation and some incorrect sentence generation of our model. The generated sentences are also compared with ground truth sentences.

5.1. Quantitative Performance Comparison with Baseline Methods

This subsection discusses our method's performance and relates the result with previous baseline methods (compared with the best score). Tables 2 and 3 report the judgment between our method and previous baseline methods. We have chosen these baseline models because all of them use the same dataset and evaluation metrics for generating video captions. Here B-4, M, C, and R-L mean BLEU@4 [37], METEOR [38], CIDEr [39], and ROGUE-L [40] metrics individually.

Table 2. Evaluation results of our method and previous methods for the video description generation on MSVD dataset. Here, "-" indicates the metric is not considered. All the numerical terms are in percentile format. Texts with bold marks are the best scores.

Methods	B-4	M	C	R-L
HATT [9]	**52.9**	33.8	73.8	-
ICA-LSTM [10]	51.3	35.1	82.9	-
SF-SSAG-LSTM [12]	51.2	35.4	74.9	-
LSTM-GAN [13]	42.9	30.4	-	-
DenseLSTM [17]	50.4	32.9	72.6	-
CAM-RNN [20]	42.4	33.4	54.3	69.4
VRE [23]	51.7	34.3	86.7	71.9
MR-HRNN [25]	50.5	32.7	69.6	-
STA-FG-RC [32]	52.7	34.5	-	-
Ours [Att-BiL-SL (V + I3D)]	48.3	32.5	85.3	68.8
Ours [Att-BiL-SL (IV3 + I3D)]	49.1	33.2	86.4	70.3
Ours [Att-BiL-SL (X + I3D)]	51.2	**35.7**	**86.9**	**72.1**

Table 3. Evaluation results of our method and previous methods for the video description generation on the MSR-VTT dataset. Here, "-" indicates the metric is not considered. All the numerical terms are in percentile format. Texts with bold marks are the best scores.

Methods	B-4	M	C	R-L
HATT [9]	41.2	28.5	44.7	60.7
ICA-LSTM [10]	41.2	27.7	43.9	-
SF-SSAG-LSTM [12]	40.8	28.7	46.8	61.5
LSTM-GAN [13]	36.0	26.1	-	-
DenseLSTM [17]	38.1	26.6	42.8	-
CAM-RNN [20]	37.7	27.9	38.8	58.8
VRE [23]	**43.2**	28.0	48.3	**62.0**
MR-HRNN [25]	36.2	25.6	33.5	-
STA-FG-RC [32]	40.8	27.4	-	-
Ours [Att-BiL-SL (V + I3D)]	39.0	27.8	46.2	57.8
Ours [Att-BiL-SL (IV3 + I3D)]	40.8	28.3	48.0	59.8
Ours [Att-BiL-SL (X + I3D)]	41.2	**28.7**	**49.3**	60.4

(i) HATT [9]: The method introduced a two-level LSTM branch as a hierarchical order and dealt with low and high-level attention. The method semantically enriched for its fused multiple modalities for caption generation, but the period of feature extraction was very extraordinary. The method also considered the two public datasets to measure the performance. However, the model scored 52.9% in B-4 metrics. Our model, Att-BiL-SL (compared with the best score), attained a better result than HATT, with 1.9% and 13.1% increases on M and C, respectively, on the MSVD dataset. In terms of the MSR-VTT dataset, our model increased the performance by 0.2% and 13.1% in M and C metrics, respectively.

(ii) ICA-LSTM [10]: The model consisted of multiple LSTM [52] for video captioning and used dual-stage loss functionality. The method was also trained in various CNN models. Our model again accomplished a better outcome than ICA-LSTM, with 0.6% and 4% upsurges on M and C in the MSVD dataset. However, our model increased by 1% and 4% on M and C in the MSR-VTT dataset.

(iii) SF-SSAG-LSTM [12]: The method introduced augmented audio features and semantic filtration. The model generated a single-line video caption. The performance of our method upsurged 0.3% and 12% on M and C metrics in the MSVD dataset and 0.4% and 2.5% on M and C in the MSR-VTT dataset.

(iv) LSTM-GAN [13]: The model introduced a generative adversarial network and attention mechanism for video captioning. Compared to this model, our method performed better than LSTM-GAN, with performance gains of 8.3% and 5.3% on B-4 and M met-

rics in the MSVD dataset. However, the performance increased by 5.2% and 2.6% in the MSR-VTT dataset.

(v) DenseLSTM [17]: Attention-based dense LSTM was considered for video captioning. The model had a backward cell connected with a forward cell. The exhibition of this model did not beat our proposed model. Our model increased by 0.8%, 2.8%, and 14.3% in B-4, M, and R-L measurements on the MSVD dataset. Likewise, it increased 3.1%, 2.1%, and 6.5% on the MSR-VTT dataset.

(vi) CAM-RNN [20]: CAM encoded the visual and literary highlights, and RNN kept up the decoder state to create video captioning. Our model, Att-BiL-SL (contrasted and the best score), achieved a better outcome than CAM-RNN, with 8.8%, 2.3%, 32.6%, and 2.7% increments in B-4, M, C, and R-L, separately (MSVD). Our model built the exhibition with 6.5%, 0.8%, 10.5%, and 1.6% on B-4, M, C, and R-L measurements (MSR-VTT), separately.

(vii) VRE [23]: The proposed method with refocused RNN with spatial-visual features was introduced for better video captioning. The method outperformed on B-4 and R-L metrics with 43.2% and 62.0% in the MSR-VTT dataset. Contrasted with this model, our strategy performed better than VRE, with execution increases of 1.4%, 0.2%, and 0.2% on M, C, and R-L measurements in the MSVD dataset. Nonetheless, the presentation incremented by 0.7%, and 1%, on M and C in the MSR-VTT dataset.

(viii) MR-HRNN [25]: The proposed structure ready to perceive human-related activities with objects during sports. The structure also introduced a pose attribute detection module and a description generation module. In the MSVD dataset, our model expanded the exhibition by 0.7%, 3%, and 17.3% on B-4, M, and C measurements. In the MSR-VTT dataset, the exhibition expanded by 5%, 3.1%, and 15.8% on B-4, M, and C, individually.

(ix) STA-FG-RC [32]: The method with semantic, temporal attention for a video description in textual formation. The model can expressly join with the high-level visual idea to create temporal attention. Our method performed better than STA-FG-RC, with performance gains of 1.2% on M metrics in the MSVD dataset. However, the performance increased by 0.4% and 1.3% in the MSR-VTT dataset.

Our model, Att-BiL-SL (X + I3D), scored 51.2%, 35.7%, 86.9%, and 72.1% on B-4, M, C, and R-L, respectively, in the MSVD dataset. Moreover, the model scored 41.2%, 28.7%, 49.3%, and 60.4% in the MSR-VTT dataset.

6. Quantitative Analysis and Discussion

Throughout recent times, numerous research and techniques for generating video captions have been conducted and proposed. It can be said by analyzing previous research and their frameworks that some of these methods still have a gap in algorithmic level and deployment. In Section 5.1, the improvements of our model were shown and those were compared to ground truth human annotation sentences in different evaluation metrics. Unfortunately, there is no assurance that higher metrics scores will result in better-generated captions. Furthermore, the comparison with existing baseline studies in Tables 2 and 3 do not directly reflect the perfections of the created video captions of our model. Hence, we illustrated our qualitative sentence generation outcomes as shown in Figure 4. We can easily observe that our proposed approach can create more accurate and detailed captions incorporated with human annotation sentences. Most of the time, the sentences generated by the machine are not well furnished. Compared to human annotation sentences, there are many gaps between the machine-generated sentences. Sometimes, the generated sentences are not able to be understood as shown in Figure 5. For example, the ground truth sentence is "A man is showing his gymnastic skills," but the machine-generated sentence is "A black dog is walking," which may confuse a user. If this feature was included in robotics, there would be a high chance of failure because human annotation sentences are very concise in terms of machine language. However, due to continuous improvement, it would have been predicted that the machine sentences must be improved in the near future.

Our model did not consider any audio features of the visual scene except video frames. However, the most important thing is that a video generally contains audio features. The audio might be useful in gathering backend evidence for every occurrence from the visual frames. Although audio feature can be helpful for gathering semantic information, it can also predict the next word generation for the machine. Furthermore, it may boost the performance of the method. Furthermore, due to time reliability and resource constraints, we have only investigated two datasets for our proposed model. Taking into account a variety of huge datasets with a significant number of human annotation sentences, it may help our model reach the highest benchmark score in evaluation metrics.

7. Conclusions

In this paper, we proposed a novel attention-based bi-LSTM and sequential LSTM (Att-BiL-SL) encoder-decoder model for video description generation in textual formation. We introduced a two-layer attentional bi-directional LSTM as an encoder and sequential LSTM as a decoder for sentence generation. From the performance analysis, it can be seen that the model of this study can deal with the global and local features of the visual frames. The model has been trained in different CNN models, and model evaluation has been performed considering standard performance matrices. It has been found that this model shows extraordinary results on two of the most popular public datasets. According to the results, our method has shown outstanding results in various automatic machine evaluation metrics compared to other existing methods. The model can also handle the high-level features for flawless sentence generation from the video. Finally, the suggested model is adaptable to different video modeling approaches.

This literary change lessens the size of video information and empowers us to navigate and explore data quickly. In the future, we can integrate the audio feature to enrich the generated sentences more semantically. Moreover, we may extend our approach to generate video from textual descriptions.

Author Contributions: Conceptualization, S.A. and M.I.H.; Data curation, S.A., M.M.N.S. and M.M.J.; Formal analysis, M.I.H., M.M.N.S., M.M.U.H., S.B.S., J.H., B.S.S. and F.R.; Funding acquisition, M.I.H., M.M.N.S., M.M.J., M.M.U.H., S.B.S., J.H., B.S.S., F.R. and H.M.S.; Investigation, S.B.S.; Methodology, S.A. and A.F.M.S.S.; Project administration, S.A. and A.F.M.S.S.; Resources, M.M.J., J.H. and B.S.S.; Software, M.I.H., M.M.N.S., S.B.S., B.S.S. and F.R.; Supervision, A.F.M.S.S.; Validation, M.I.H., M.M.U.H., and J.H.; Visualization, M.M.U.H.; Writing—original draft, S.A.; Writing—review and editing, S.A., A.F.M.S.S., M.M.J., M.M.U.H. and H.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research funded by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Cisco. *The Zettabyte Era: Trends and Analysis*; Cisco: San Jose, CA, USA, 2017. Available online: https://en.wikipedia.org/wiki/Zettabyte_Era (accessed on 3 April 2021).
2. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2020**, *130*, 370–375. [CrossRef]
3. Sridevi, M.; Kharde, M. Video Summarization Using Highlight Detection and Pairwise Deep Ranking Model. *Procedia Comput. Sci.* **2020**, *167*, 1839–1848. [CrossRef]
4. Chu, Y.-W.; Lin, K.-Y.; Hsu, C.-C.; Ku, L.-W. Multi-Step Joint-Modality Attention Network for Scene-Aware Dialogue System. 2020. Available online: http://arxiv.org/abs/2001.06206 (accessed on 16 September 2020).
5. Huang, J.H.; Worring, M. Query-controllable video summarization. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 242–250. [CrossRef]

6. Li, Z.; Li, Z.; Zhang, J.; Feng, Y.; Zhou, J. Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; IEEE: Piscataway, NJ, USA, 2021; Volume 29, pp. 2476–2483. [CrossRef]
7. Pan, G.; Zheng, Y.; Zhang, R.; Han, Z.; Sun, D.; Qu, X. A bottom-up summarization algorithm for videos in the wild. *EURASIP J. Adv. Signal Process.* **2019**, *2019*, 15. [CrossRef]
8. Nian, F.; Li, T.; Wang, Y.; Wu, X.; Ni, B.; Xu, C. Learning explicit video attributes from mid-level representation for video captioning. *Comput. Vis. Image Underst.* **2017**, *163*, 126–138. [CrossRef]
9. Wu, C.; Wei, Y.; Chu, X.; Weichen, S.; Su, F.; Wang, L. Hierarchical attention-based multimodal fusion for video captioning. *Neurocomputing* **2018**, *315*, 362–370. [CrossRef]
10. Xiao, H.; Xu, J.; Shi, J. Exploring diverse and fine-grained caption for video by incorporating convolutional architecture into LSTM-based model. *Pattern Recognit. Lett.* **2020**, *129*, 173–180. [CrossRef]
11. Jin, T.; Li, Y.; Zhang, Z. Recurrent convolutional video captioning with global and local attention. *Neurocomputing* **2019**, *370*, 118–127. [CrossRef]
12. Xu, Y.; Yang, J.; Mao, K. Semantic-filtered Soft-Split-Aware video captioning with audio-augmented feature. *Neurocomputing* **2019**, *357*, 24–35. [CrossRef]
13. Yang, Y.; Zhou, J.; Ai, J.; Bin, Y.; Hanjalic, A.; Shen, H.T.; Ji, Y. Video Captioning by Adversarial LSTM. *IEEE Trans. Image Process.* **2018**, *27*, 5600–5611. [CrossRef]
14. Song, J.; Guo, Y.; Gao, L.; Li, X.; Hanjalic, A.; Shen, H.T. From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3047–3058. [CrossRef]
15. Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; Shen, H.T. Video Captioning with Attention-Based LSTM and Semantic Consistency. *IEEE Trans. Multimedia* **2017**, *19*, 2045–2055. [CrossRef]
16. Bin, Y.; Yang, Y.; Shen, F.; Xu, X.; Shen, H.T. Bidirectional long-short term memory for video description. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherland, 15–19 October 2016; pp. 436–440. [CrossRef]
17. Zhu, Y.; Jiang, S. Attention-based densely connected LSTM for video captioning. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 802–810. [CrossRef]
18. Francis, D.; Huet, B. L-STAP: Learned spatio-temporal adaptive pooling for video captioning. In Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery-AI4TV'19, Nice, France, 21 October 2019; pp. 33–41. [CrossRef]
19. Xu, N.; Liu, A.-A.; Wong, Y.; Zhang, Y.; Nie, W.; Su, Y.; Kankanhalli, M. Dual-Stream Recurrent Neural Network for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2482–2493. [CrossRef]
20. 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]
21. Xiao, H.; Shi, J. Video Captioning with Adaptive Attention and Mixed Loss Optimization. *IEEE Access* **2019**, *7*, 135757–135769. [CrossRef]
22. Saleem, S.; Dilawari, A.; Khan, U.G.; Iqbal, R.; Wan, S.; Umer, T. Stateful human-centered visual captioning system to aid video surveillance. *Comput. Electr. Eng.* **2019**, *78*, 108–119. [CrossRef]
23. Shi, X.; Cai, J.; Joty, S.; Gu, J. Watch it twice: Video captioning with a refocused video encoder. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 818–826. [CrossRef]
24. Gui, Y.; Guo, D.; Zhao, Y. Semantic enhanced encoder-decoder network (SEN) for video captioning. In Proceedings of the 2nd Workshop on Multimedia for Accessible Human Computer Interfaces-MAHCI'19, Nice, France, 25 October 2019; pp. 25–32. [CrossRef]
25. Qi, M.; Wang, Y.; Li, A.; Luo, J. Sports video captioning by attentive motion representation based hierarchical recurrent neural networks. In Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports, Seoul, Korea, 26 October 2018; pp. 77–85. [CrossRef]
26. Liu, S.; Ren, Z.; Yuan, J. SibNet: Sibling convolutional encoder for video captioning. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 22–26 October 2018; Volume 2, pp. 1425–1434. [CrossRef]
27. Xiao, H.; Shi, J. Video captioning using hierarchical multi-attention model. In Proceedings of the 2nd International Conference on Advances in Image Processing-ICAIP '18, Chengdu, China, 16–18 June 2018; pp. 96–101. [CrossRef]
28. Phan, S.; Miyao, Y.; Satoh, S. MANet: A modal attention network for describing videos. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1889–1894. [CrossRef]
29. Chen, B.-C.; Chen, Y.-Y.; Chen, F. Video to text summary: Joint video summarization and captioning with recurrent neural networks. In Proceedings of the 2017 British Machine Vision Conference, London, UK, 4–7 September 2017; pp. 1–14. [CrossRef]
30. Sah, S.; Kulhare, S.; Gray, A.; Venugopalan, S.; Prud'Hommeaux, E.; Ptucha, R. Semantic text summarization of long videos. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 989–997. [CrossRef]
31. Dilawari, A.; Khan, M.U.G. ASoVS: Abstractive Summarization of Video Sequences. *IEEE Access* **2019**, *7*, 29253–29263. [CrossRef]
32. Gao, L.; Wang, X.; Song, J.; Liu, Y. Fused GRU with semantic-temporal attention for video captioning. *Neurocomputing* **2020**, *395*, 222–228. [CrossRef]

33. Hao, X.; Zhou, F.; Li, X. Scene-Edge GRU for Video Caption. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 1290–1295. [CrossRef]
34. Cherian, A.; Wang, J.; Hori, C.; Marks, T.K. Spatio-temporal ranked-attention networks for video captioning. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1606–1615. [CrossRef]
35. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296.
36. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 664–676. [CrossRef]
37. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
38. Lavie, A.; Agarwal, A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 228–231. Available online: http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75 (accessed on 8 October 2020).
39. Vedantam, R.; Zitnick, C.L.; 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. [CrossRef]
40. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 25–26 July 2004.
41. Chen, D.; Dolan, W. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 190–200.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
43. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
44. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollar, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
45. Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; Saenko, K. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2712–2719. [CrossRef]
46. Park, J.; Song, C.; Han, J.-H. A study of evaluation metrics and datasets for video captioning. In Proceedings of the 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 24–26 November 2017; pp. 172–175. [CrossRef]
47. 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, 27–30 June 2016; pp. 2818–2826.
48. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
49. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv* **2012**, arXiv:1212.0402.
50. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733. [CrossRef]
51. Jiang, H.; Learned-Miller, E. Face Detection with the Faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 650–657. [CrossRef]
52. Hochreiter, S.; Schmidhuber, J.J.U. Long Short-Term Memory. 1997. Available online: http://www7.informatik.tu-muenchen.de/~{}hochreithttp://www.idsia.ch/~{}juergen (accessed on 22 January 2021).
53. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 2, pp. 207–212.
54. Natural Language Toolkit—NLTK 3.5 Documentation. Available online: https://www.nltk.org/ (accessed on 1 April 2021).
55. TensorFlow. Available online: https://www.tensorflow.org/ (accessed on 1 April 2021).
56. Keras Applications. Available online: https://keras.io/api/applications/ (accessed on 1 April 2021).
57. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]

58. Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R.; Saenko, K.; Guadarrama, S. Generating natural-language video descriptions using text-mined knowledge. In Proceedings of the Workshop on Vision and Natural Language Processing, Atlanta, GA, USA, 14 June 2013; pp. 10–19. Available online: http://www.aclweb.org/anthology/W13-1302 (accessed on 22 January 2021).
59. Guo, J.; Liu, H.; Li, X.; Xu, D.; Zhang, Y. An Attention Enhanced Spatial–Temporal Graph Convolutional LSTM Network for Action Recognition in Karate. *Appl. Sci.* **2021**, *11*, 8641. [CrossRef]
60. Peng, L.; Zhu, Q.; Lv, S.-X.; Wang, L. Effective long short-term memory with fruit fly optimization algorithm for time series forecasting. *Soft Comput.* **2020**, *24*, 15059–15079. [CrossRef]
61. Peng, L.; Wang, L.; Xia, D.; Gao, Q. Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. *Energy* **2021**, *238*, 121756. [CrossRef]
62. Qin, Y.; Xiang, S.; Chai, Y.; Chen, H. Macroscopic–Microscopic Attention in LSTM Networks Based on Fusion Features for Gear Remaining Life Prediction. *IEEE Trans. Ind. Electron.* **2019**, *67*, 10865–10875. [CrossRef]