Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 1229–1237

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

VATEX2020: pLSTM framework for video captioning

Alok Singh[a,b,*], Salam Michael Singh[a,b], Loitongbam Sanayai Meetei[a,b], Ringki Das[a,b], Thoudam Doren Singh[a,b], Sivaji Bandyopadhyay[a,b]

[a] *Department of Computer Science and Engineering, National Institute of Technology Silchar Assam, India*
[b] *Center for Natural Language Processing, National Institute of Technology Silchar Assam, India*

Abstract

Captioning a video involves condensing the video's information into text, which can be useful in video sentiment analysis, video-guided machine translation (VMT), visual question-answering and humanitarian aid. This paper discusses the details of the architecture of the pLSTM framework that is employed for the VATEX-2020 video captioning challenge. In this work, a sequential method is employed wherein to encode visual features a 3D convolutional neural network (C3D) is used. C3D was pretrained using the Sports-1M dataset. In the decoding phase, the input captions and visual features are fused separately in Long Short Term Memory networks (LSTM). The element-wise dot product is performed on the output of both LSTMs to get the final output. On both publicly available and private test data sets, our model achieves BLEU-4 scores of 0.20 and 0.22, respectively.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering

Keywords: Video captioning; 3D CNN; Encoder-decoder; VATEX; LSTM; C3D;

1. Introduction

In the last few decades, artificial intelligence (AI) has focused on developing models that can understand the visual entities around us and draw some reasoning to describe them into a natural language. There is an exponential growth in multimedia data (especially images and videos) all across the globe due to the advances in technologies and the Internet. This massive growth in multimedia data raised several issues such as automatic categorisation of digital data, indexing for fast retrieval and recognising the visual entities. A significant amount of work has been carried out on visual recognition and classification in the last few decades. As a result, a computer can identify the visual entity and classify them into their respective classes. However, research into comprehending and describing visual entities through natural language is still in its infancy. The field of generating a short description (or caption) for a video has attracted many researchers. This falls at the intersection of natural language processing (NLP) and computer

[*] Corresponding author.
E-mail address: alok.rawat478@gmail.com

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering
10.1016/j.procs.2023.01.101

vision (CV). Video captioning aims to design an automatic system that can *'understand'* the relationship among the objects and attributes present in the visual scene and *'communicate'* them in natural language. Video captioning is the stepping stone in bridging the gap between CV and NLP to produce a deeper understanding of visual scenes, activities and tasks of a video in the form of natural language. Appropriate treatment of video captioning challenges can help in improving the performance of various video-related tasks such as video retrieval based on content, video indexing, and the most recent emerging field of video-guided MT [1]. Understanding the video scenes in the form of texts can aid visually impaired ones to understand the video content. The task of video captioning is predominantly motivated by MT, where the output in the target language T is produced by translating the sentence S of the source language. In video captioning, the encoded visual context vector is the source and the caption of the video is the target.

Multiple modalities such as sound, vision and text assist human in learning and understanding the activities happening around them. The tasks involving scene understanding, recognition and interaction in natural language seem accomplishable for human. However, it becomes challenging for human to handle as the volume of the content in these tasks increases. The role of automatic visual scene understanding systems becomes important in such a situation. But, the task of visual scene understanding for these systems is not as easy as for human. For instance, a quick glimpse of a video is sufficient to *'remember'*, *'recall'* and *'describe'* all the minute details present in the video for human. It is challenging to perform all three tasks simultaneously and effectively for the AI models. Various datasets have been proposed in multiple languages and domains to carry out video captioning. $YouCook$ and $TACoS$ are two domain-specific datasets that specifically contain the activity related to cooking. $MSVD$ [2] and $MSR - VTT$ [3] are two recent open-domain datasets for video captioning. The $MSR - VTT$ dataset contain video of 20 different classes [4]. India is a multilingual country. As per the Eighth Schedule of the Indian Constitution, there are 22 official languages. Among all the oldest and widely spoken languages, Hindi is one of them, and it is the most spoken in India and South Asia. Recently, the multilingual image captioning gained more popularity with the introduction of various Hindi image captioning [5, 6, 7], Assamese news image captioning [8], multi-model machine translation [5, 9] approaches and the release of Hindi video captioning dataset by Singh et al. [10].

This work reports details about the sequential pLSTM approach for English video captioning on the VATEX dataset that is released in 2020 for the VATEX challenge. The findings of the proposed work are as follows:

i. A video description framework is proposed in which two different fusion strategies are employed simultaneously using two parallel LSTMs.

ii. A detailed qualitative and quantitative analysis is performed on the output captions produced by the proposed approach on the English VATEX video captioning dataset, which is released in the VATEX-2020 video captioning challenge.

2. Background Knowledge

The majority of video captioning frameworks are inspired by image captioning approaches, which aim to generate a description of static entities in an image. However, generating a video description is more challenging than image captioning due to multiple scenes that change dynamically with time and the redundancy of frames. All the proposed methods for video description are broadly categorised into classical, statistical and recent deep learning based approaches [11, 12]. The template based matching and retrieval based approaches are older classical approaches employed for captioning. The captions generated using template based matching usually lack fluency compared to the retrieval based approach due to their rigid structures. Nevertheless, both approaches are unable to produce adequate captions [13]. The first successful method for video description is proposed using SVO (Subject, Verb, Object) [4]. These classical approaches for captioning have two steps. The first step is identifying attributes and their related activities in the video and then the caption is generated, aiming to map the recognised object and action to SVO and fill them into fixed templates. The effectiveness of these approaches is limited to videos with shorter lengths or videos that have less SVO. Numerous approaches have been proposed in [14, 15, 16, 17] for recognising the objects, events and actions in a video. These classical approaches work on detecting the events and actions which are predefined.

The preprint of the paper is available at https://arxiv.org/pdf/2006.04058.pdf

Thus, the effectiveness of these approaches for open domain video is not reasonable. Moreover, these approaches are unable to describe the video with more events and objects.

To overcome the shortcomings of classical approaches and the performance of Deep Learning (DL) approaches in almost all the sub-fields of CV motivates researchers to use DL techniques to address the challenges of captioning. Since the Convolutional Neural Network (CNN) outperform all the classical techniques of encoding visual features in action recognition and object recognition, that is why CNN is used most widely for encoding the visual features of a video [18, 19]. For encoding the sequence of text, Recurrent Neural Network (RNN) is successfully used in MT. Similarly, RNN is also used in video captioning. The DL models include: the encoding and sequence decoding stage [4]. In the visual encoding stage, either CNN or RNN are employed to encode visual features. Whereas, RNN and various variants of the recurrent network are used for decoding. In this section, the approaches for video captioning are grouped based on different encoder-decoder architectures such as CNN-RNN and RNN-RNN. In the CNN-RNN, CNN is employed for visual encoding and RNN for sequence decoding, while in RNN-RNN based architectures, both the encoding and decoding stages are RNN and its variants. The process of visual features encoding can be categorised into two classes: variable-sized and fixed-sized [4]. The first DL based model for video captioning was proposed by [20]. They proposed three different architectures for video captioning based on LSTM and CRF based prediction. [21] proposed end-to-end trainable network for video description. This model can learn both the semantic and syntactic structure of the language. [20] used domain-specific cooking video for the evaluation of the model whereas [21] used open domain video dataset. Inspired by the success of the sequence-to-sequence model in MT, [22] proposed a model sequential model in which two LSTM layers are employed. A collection of frames is given as an input to the first recurrent layer of the LSTM layer and then using the hidden state representation of the first LSTM with null padded input words the second LSTM layer generates the description. In [22], all the frames in the video are given the same importance during visual encoding, so there is a possibility of more redundant information, which is less likely to generate video caption [23] since the video include lots of redundant frames. To address the issue of giving equal weightage to all the frames, [24, 25] employed a temporal attention-based mechanism so that the informative frame is focused on generating the meaningful caption. [26] proposed an LSTM-LSTM based video description model. LSTM is employed for the encoding and decoding of the visual context vector in the form of text in this model. A hierarchical RNN (h-RNN) based model is proposed by [25] for generating a paragraph. For encoding informative visual features this model employs both spatial and temporal attention. In h-RNN, firstly Gated Recurrent Unit (GRU) is used for visual feature encoding and generating a short sentence. Then by employing another recurrent layer a paragraph is generated. Some other attention based approaches are [27, 28].

To address the issue of giving equal weightage to all the frames, [24, 25] employed a temporal attention-based mechanism so that the informative frame is focused on generating the meaningful caption. [26] proposed an RNN-RNN based video description model in which LSTM is used for extracting the visual context vector and then passes the vector to the subsequent LSTM for decoding the vector in the form of text. A video suffers various challenges such as illumination and motion effect, due to which the performance of video captioning frameworks gets affected to address these issues [29] proposed a video to text approach using a shot boundary approach. Deep learning approaches [30, 31] proposed for image analysis can also be used for effective visual feature encoding. The work cited above is done in specific languages due to the presence of datasets in limited languages (such as German, English, Chinese, etc.). The MSVD dataset [2] and most recent dataset VATEX [1] are two multilingual dataset available for video captioning. Considering the necessity of caption generation in Hindi, [10] translated the available $MSR-VTT$ dataset into Hindi using Google translator and proposed a Hindi video captioning model.

Furthermore, the remaining part of the paper is organised as follows. Section 3 reports the description of the proposed framework. Then Sections 4 and 5 discuss the system's performance and the conclusion, respectively.

3. Proposed System

The architecture of the system and the details of visual feature extraction from the video are covered in this part.

3.1. Visual Feature Encoding

A conventional encoder-decoder based technique is employed for visual encoding. For encoding, the input video is uniformly divided into n segments spaced by 16 to create the visual context vector. For visual context vector $f = \{s_1, s_2 \ldots s_n\}$ a pretrained C3D network is employed. Each video's visual context vector has a dimension of $f \in \mathbb{R}^{n \times m_x}$ ($m_x = 4096$). A high-dimensional visual context vector degrades feature quality since it is prone to carrying redundant information [23]. Average pooling with a filter size of 5 is applied for the dimension reduction of features. To preserve the temporal relationship between the averaged pool features, all features are concatenated in a sequence and then passed to a decoder for captioning generation

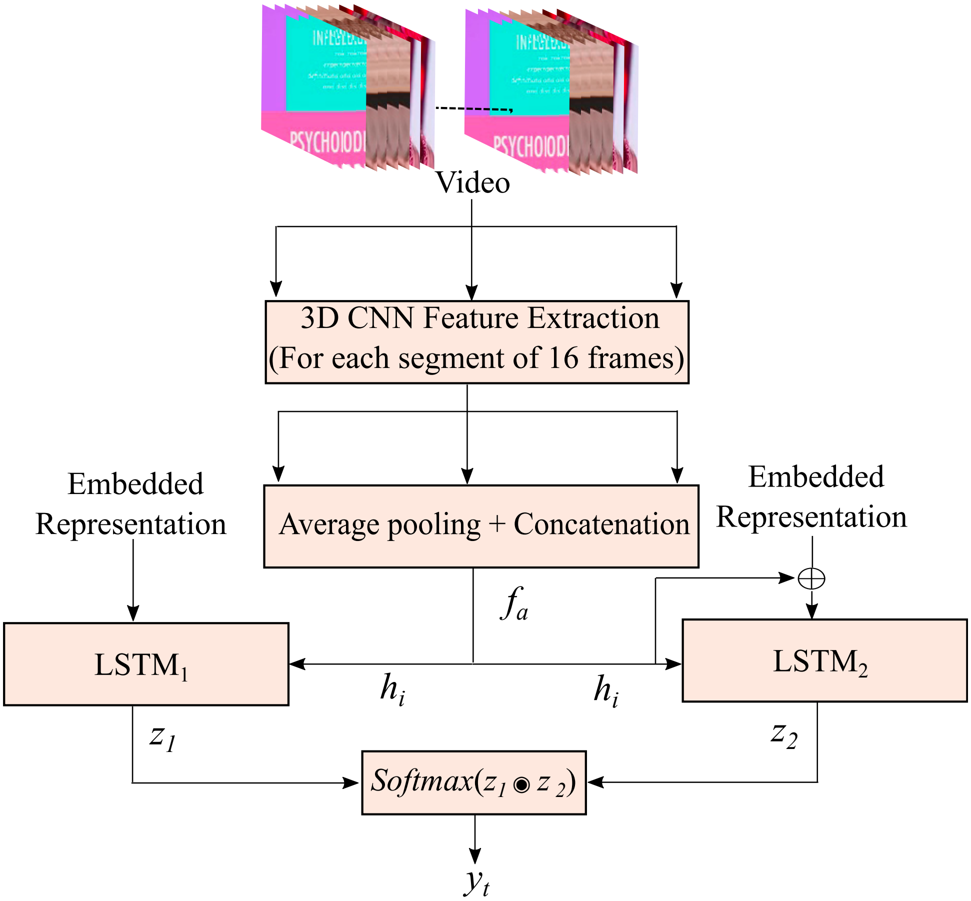


Fig. 1: The pictorial representation of the proposed framework

3.2. Caption Generator

As a decoder, two independent LSTMs are used. To extract a dense embedded representation the words in the input caption are passed through the embedding layer. After getting a dense embedded representation it is passed to LSMTs in the decoder separately. The initial states of the first LSTM are initialised with a visual context vector and take a dense embedded representation of the input word as an input. While the second LSTM takes concatenated visual context vector and dense embedded representation as input as shown in Figure 1. Finally, to get the correct word in a sequence an element-wise dot product is performed between the output z_1 and z_2. Mathematically, the whole procedure can be described as:

$$\tilde{y} = W_e X + b_e \tag{1}$$

$$z_1 = LSTM_1(\tilde{y}, h_i) \tag{2}$$

$$z_2 = LSTM_2([\tilde{y}; f_a]) \tag{3}$$

$$y_t = softmax(z_1 \odot z_2) \tag{4}$$

Embedding of the input captions (X) is represented by Equation 1, where W_e and b_e are weights and biases, respectively. The outputs of LSTM layers are z_1 in Equation 2 and z_2 in Equation 3. The hidden states of $LSTM_1$ and $LSTM_2$ are initialised by averaged pooled features ($h_i = f_a$). $\odot$ is the product of z_1 and z_2, which is then passed to the $softmax$ layer. A cross-entropy loss function is used to maximise the likelihood of the correct word at a particular time step t while minimising the loss, as shown by Equation 5, $y_{0...T}$ represent the generated words in a sequence and F is encoded visual features.

$$Loss = -\sum_{t=0}^{T} \log P(y_t | y_{t-1}, \ldots y_0; F) \tag{5}$$

4. Ablation study

This section presents a detailed study of the dataset used, evaluation measures, experimental setup and the system output is carried out.

4.1. Dataset

The VATEX dataset [1] released in 2020 is employed for the evaluation of the proposed framework. By associating each video with ten captions in both Chinese and English this dataset is intended to encourage multilingual video captioning. Since this dataset was introduced in a challenge so for evaluation purposes two sets were presented: public and private test sets. The ground truth for the public test set was made available publicly but for the private test set, it was held out for evaluation.

4.2. Evaluation Metrics

There have been many different suggestions made for evaluation metrics that can be used to judge the level of quality of captions that are generated automatically. To quantitatively assess the quality of the output generated in this

[;] represent the concatenation

Table 1: Statistics derived from the dataset that was used

Dataset separation	**#Videos**	**#English description**	**#Chinese description**
Training	25,991	259,910	259,910
Validation	3,000	30,000	30,000
Private set	6,287	62,780	62,780
Public set	6,000	60,000	60,000

paper, the following metrics are used: BLUE (B) [32], CIDEr [33], and ROUGE-L, METEOR [34]. All the automatic scores are evaluated by the organisers of the VATEX challenge.

4.3. Experimental Setup

As part of the training process, two unique markers, $BOS >$ and $EOS >$, are appended to each caption to tell the model when the caption generation process should begin and end. We capped the allowed number of caption words at 30, and if a caption was shorter than that, it would be "zero-padded" to reach the target length. In order to retain only the most frequently used words, the Stanford tokenizer [35] is used to tokenize the captions after that a vocabulary of size $15K$ is obtained. Caption generation in the testing phase begins by keeping an eye on the start marker and visual context vector from there, the most likely word is sampled out and added to the pool of words and visual feature vectors that will be used in the next iteration until a unique end marker is produced. The learning rate of 2×10^{-4} and $ADAM$ optimiser with cross-entropy loss function is used. Both LSTMs have their hidden units set to 512, and a dropout of 0.5 is used to mitigate the overfitting problem. The proposed architecture undergoes training for a total of 50 epochs, with batch sizes of 64 and 256. The proposed approach performs better in smaller batch size which is manifested by the automatic evaluation scores.

4.4. Results

Table 2 displays the results on both test sets. On both public and private test sets, the proposed model receives a CIDEr score of 0.24 and 0.27 respectively.

Table 2: Results obtained on both test sets

Evaluation Measure	**Performance on public set**	**Performance on private set**
B-1	0.63	0.65
B-2	0.43	0.45
B-3	0.30	0.32
B-4	0.20	0.22
CIDEr	0.24	0.27
ROUGE-L	0.42	0.43
METEOR	0.18	0.18

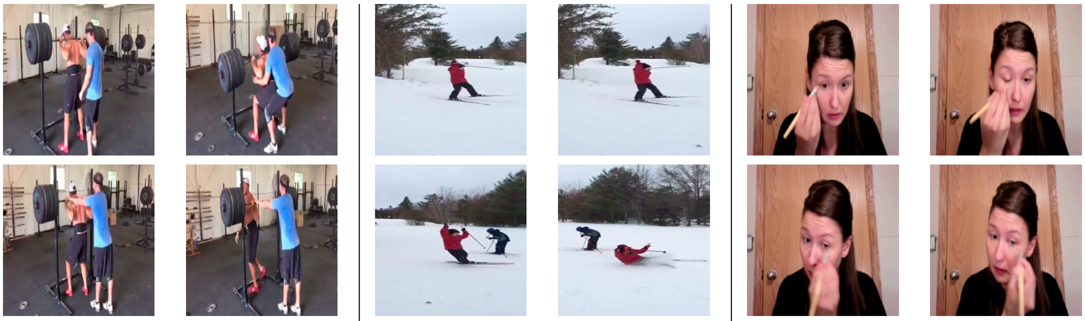
We also sampled videos at random from the test sets and analysed them qualitatively to see how well they matched the generated description. As an example of the proposed system's output captions, the sample output captions are listed in Table 4. Analyzing the caption produced by the proposed system reveals that the model not only generates long, grammatically correct descriptions but also describes the scene in a natural and eloquent manner.

Check out the CodaLab website for the complete set of results.

Table 3: Analyzing the proposed method's performance in relation to other models on a private test dataset

Method	**B-4**	**METEOR**	**CIDEr**	**ROUGE**
Proposed (Our)	0.220	0.430	0.180	0.270
ORG-TRL [36]	0.321	0.489	0.222	0.497
Top-down + X-LAN [37]	0.392	0.527	0.250	0.760
Baseline [1]	0.285	0.470	0.216	0.451

Table 4: Illustration of the proposed system's generated caption output.



Output
A man is lifting a heavy weight over his head and then drops it

(a)

Output
A group of people are skiing down a hill and one of them falls down

(b)

Output
A woman is demonstrating how to apply mascara to her eyelashes

(c)

5. Conclusion and future work

This paper reports the description of our pLSTM framework which is employed in the VATEX2020 challenge. It's difficult to give a complete description of a video because there are so many different things to describe. This paper proposes a model that successfully deals with these issues. To generate the English caption, we employ a video captioning framework based on an encoder-decoder pair. The visual features are encoded with the help of a pretrained Convolutional Neural Network, and the decoder end uses two parallel LSTMs to fuse the visual features with the word embedding in two different ways simultaneously. A dot product is then calculated between the two LSTMs' outputs to arrive at the final description. In terms of public and private video captioning test sets, the proposed system achieved a BLEU-4 of 0.20 and 0.22, respectively. During the challenge, the proposed system is compared to other models proposed by participants.

Our future efforts will centre on enhancing the model's computational efficiency and the accuracy of the automatic evaluation scores.

Acknowledgments

This work is supported by Scheme for Promotion of Academic and Research Collaboration (SPARC) Project Code: P995 of No: SPARC/2018-2019/119/SL(IN) under MHRD, Govt of India.

References

[1] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, W. Y. Wang, Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4581–4591.

[2] D. L. Chen, W. B. Dolan, 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, 2011, pp. 190–200.

[3] J. Xu, T. Mei, T. Yao, Y. Rui, Msr-vtt: A large video description dataset for bridging video and language, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5288–5296.

[4] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, M. Shah, Video description: A survey of methods, datasets, and evaluation metrics, ACM Computing Surveys (CSUR) 52 (6) (2019) 1–37.

[5] L. S. Meetei, T. D. Singh, S. Bandyopadhyay, Wat2019: English-hindi translation on hindi visual genome dataset, in: Proceedings of the 6th WorkObject relational shop on Asian Translation, 2019, pp. 181–188.

[6] A. Singh, L. S. Meetei, T. D. Singh, S. Bandyopadhyay, Generation and evaluation of hindi image captions of visual genome, in: Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India, Springer, 2021, pp. 65–73.
URL `https://doi.org/10.1007/978-981-33-4084-8_7`

[7] A. Singh, T. D. Singh, S. Bandyopadhyay, An encoder-decoder based framework for hindi image caption generation, Multimedia Tools and Applications (2021) 1–20.
URL `https://doi.org/10.1007/s11042-021-11106-5`

[8] R. Das, T. D. Singh, Assamese news image caption generation using attention mechanism, Multimedia Tools and Applications 81 (7) (2022) 10051–10069.

[9] S. M. Singh, L. Sanayai Meetei, T. D. Singh, S. Bandyopadhyay, Multiple captions embellished multilingual multi-modal neural machine translation, in: Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021), INCOMA Ltd., Online (Virtual Mode), 2021, pp. 2–11.
URL `https://aclanthology.org/2021.mmtlrl-1.2`

[10] A. Singh, T. D. Singh, S. Bandyopadhyay, Attention based video captioning framework for hindi, Multimedia Systems 28 (1) (2022) 195–207.
URL `https://doi.org/10.1007/s00530-021-00816-3`

[11] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys (CSUR) 51 (6) (2019) 1–36.

[12] A. Singh, T. D. Singh, S. Bandyopadhyay, A comprehensive review on recent methods and challenges of video description, arXiv preprint arXiv:2011.14752 (2020).

[13] X. He, L. Deng, Deep learning for image-to-text generation: A technical overview, IEEE Signal Processing Magazine 34 (6) (2017) 109–116.

[14] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.

[15] X. Wu, G. Li, Q. Cao, Q. Ji, L. Lin, Interpretable video captioning via trajectory structured localization, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6829–6837.

[16] P. Hanckmann, K. Schutte, G. J. Burghouts, Automated textual descriptions for a wide range of video events with 48 human actions, in: European Conference on Computer Vision, Springer, 2012, pp. 372–380.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[18] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[19] Q. Wang, A. B. Chan, Cnn+ cnn: Convolutional decoders for image captioning, arXiv preprint arXiv:1805.09019 (2018).

[20] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[21] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, arXiv preprint arXiv:1412.4729 (2014).

[22] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence-video to text, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534–4542.

[23] Y. Xu, J. Yang, K. Mao, Semantic-filtered soft-split-aware video captioning with audio-augmented feature, Neurocomputing 357 (2019) 24–35.

[24] Y. Tu, X. Zhang, B. Liu, C. Yan, Video description with spatial-temporal attention, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1014–1022.

[25] H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, Video paragraph captioning using hierarchical recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4584–4593.

[26] N. Srivastava, E. Mansimov, R. Salakhudinov, Unsupervised learning of video representations using lstms, in: International conference on machine learning, 2015, pp. 843–852.

[27] W. Li, D. Guo, X. Fang, Multimodal architecture for video captioning with memory networks and an attention mechanism, Pattern Recognition Letters 105 (2018) 23–29.

[28] H. Xiao, J. Xu, J. Shi, Exploring diverse and fine-grained caption for video by incorporating convolutional architecture into lstm-based model, Pattern Recognition Letters 129 (2020) 173–180.

[29] A. Singh, T. D. Singh, S. Bandyopadhyay, V2t: video to text framework using a novel automatic shot boundary detection algorithm, Multimedia Tools and Applications (2022) 1–21.
URL `https://doi.org/10.1007/s11042-022-12343-y`

[30] P. Rastogi, V. Singh, M. Yadav, Deep learning and big datatechnologies in medical image analysis, in: 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2018, pp. 60–63.

[31] P. Rastogi, K. Khanna, V. Singh, Gland segmentation in colorectal cancer histopathological images using u-net inspired convolutional network, Neural Computing and Applications 34 (7) (2022) 5383–5395.

[32] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[33] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[34] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.
URL http://www.aclweb.org/anthology/P/P14/P14-5010

[36] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, Z.-J. Zha, Object relational graph with teacher-recommended learning for video captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13278–13288.

[37] K. Lin, Z. Gan, L. Wang, Multi-modal feature fusion with feature attention for vatex captioning challenge 2020, arXiv preprint arXiv:2006.03315 (2020).