
VATEX2020: PLSTM FRAMEWORK FOR VIDEO CAPTIONING

Authors	Alok Singh, Salam Michael Singh, Loitongbam Sanayai Meetei, Ringki Das, Thoudam Doren Singh, Sivaji Bandyopadhyay
Published by	Elsevier B.V
Published in	Procedia Computer Science Journal
Date of Publication	2023
Cited in	1(in papers)
DOI	10.1016/j.procs.2023.01.101

Abstract:

- The paper focuses on video captioning, which involves converting video content into text.
- Video captioning has applications in video sentiment analysis, machine translation, visual question-answering, and humanitarian aid.
- The paper discusses the architecture of the pLSTM framework used for the VATEX-2020 video captioning challenge.
- It employs a sequential method for encoding visual features, using a 3D Convolutional Neural Network (C3D) pretrained on the Sports-1M dataset.
- In the decoding phase, input captions and visual features are combined separately in Long Short-Term Memory networks (LSTM).
- The final output is obtained by performing an element-wise dot product on the outputs of both LSTMs.
- The model achieves BLEU-4 scores of 0.20 and 0.22 on publicly available and private test datasets, respectively.

Keywords:

1. Video captioning
2. 3D CNN (Convolutional Neural Network)
3. Encoder-decoder architecture
4. VATEX-2020
5. LSTM (Long Short-Term Memory)
6. C3D (3D Convolutional Neural Network)
7. BLEU-4 scores: 0.20 (public test data) and 0.22 (private test data)

I. Introduction:

- Rapid growth in multimedia data, particularly images and videos, due to technological advancements and the internet.
- Challenges related to data categorization, indexing, and understanding visual entities.
- Focus on visual recognition and classification in AI; limited work on describing visual entities in natural language.
- Video captioning lies at the intersection of natural language processing (NLP) and computer vision (CV).
- Video captioning aims to create systems that understand and describe the content of videos in natural language.
- Bridging the gap between CV and NLP for a deeper understanding of visual scenes and activities.
- Applications of video captioning: video retrieval, indexing, and video-guided machine translation (MT).
- Accessibility for the visually impaired through video captions.
- Video captioning is motivated by MT, with encoded visual context as the source and video caption as the target.
- Introduction of multiple modalities (sound, vision, text) for learning and understanding.
- Role of automatic visual scene understanding systems in handling large volumes of content.
- Mention of various datasets for video captioning in different languages and domains.
- Focus on multilingual image captioning, including Hindi and Assamese.

- The paper presents a sequential pLSTM approach for English video captioning on the VATEX dataset (2020).
- Two fusion strategies employed simultaneously using parallel LSTMs in the video description framework.
- Detailed qualitative and quantitative analysis of the proposed approach's output captions on the English VATEX video captioning dataset.

Key Findings:

1. A video description framework uses two fusion strategies with parallel LSTMs.
2. The proposed approach is analyzed qualitatively and quantitatively using the English VATEX video captioning dataset from VATEX-2020 challenge.

II. Background Knowledge:

- Video captioning frameworks primarily inspired by image captioning but face unique challenges.
- Video description is more challenging than image captioning due to dynamic scenes and frame redundancy.
- Approaches for video description categorized into classical, statistical, and deep learning-based methods.
- Classical approaches include template-based matching and retrieval-based methods but have limitations in fluency and coverage.
- First successful method for video description used Subject-Verb-Object (SVO) structures but limited to short videos.
- Deep Learning (DL) is increasingly used due to advancements in computer vision.
- Convolutional Neural Networks (CNN) commonly used for visual feature encoding.
- Recurrent Neural Networks (RNN) used for encoding text sequences and decoding in machine translation and video captioning.
- DL models include encoding and sequence decoding stages, with variations in the use of CNN, RNN, and their variants.
- Video captioning approaches categorized based on encoder-decoder architectures, such as CNN-RNN and RNN-RNN.

- Visual feature encoding can be variable-sized or fixed-sized.
- Early DL models proposed for video captioning used LSTM and CRF-based prediction.
- Sequence-to-sequence models, like LSTM-based models, gained popularity in video captioning.
- Models like h-RNN use spatial and temporal attention for encoding informative visual features.
- Challenges in video captioning include illumination and motion effects.
- Multilingual datasets like MSVD and VATEX are available for video captioning.
- Efforts to address multilingual needs, such as translating MSR – VTT dataset into Hindi.
- The paper outlines its organization, with Section 3 discussing the proposed framework, Sections 4 and 5 covering system performance and the conclusion, respectively.

Key Points:

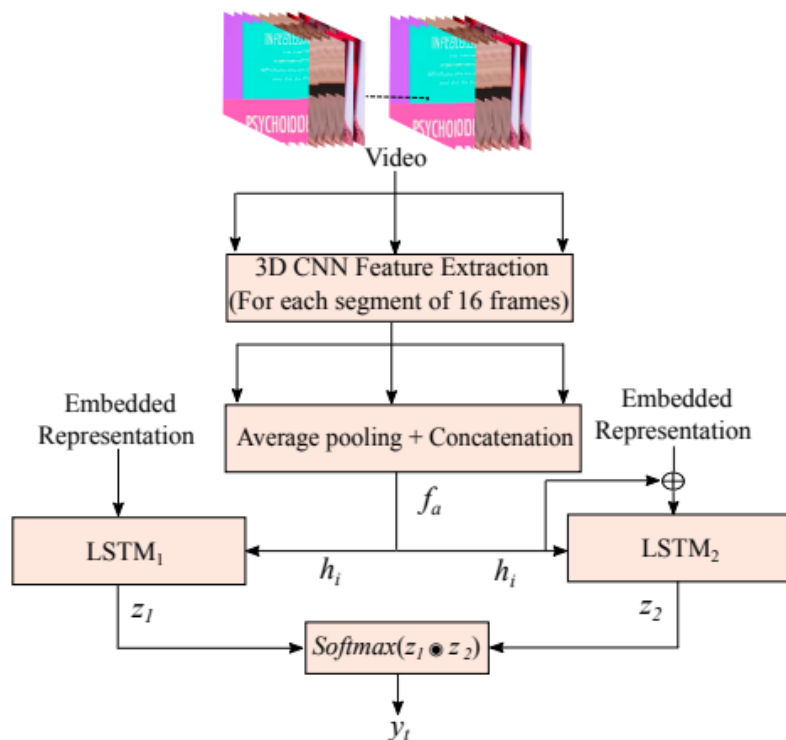
1. Video captioning methods categorized into classical, statistical, and deep learning-based approaches.
2. DL, especially CNN and RNN, is widely used in video captioning for visual feature encoding and sequence decoding.
3. Sequence-to-sequence models, spatial and temporal attention mechanisms are popular in DL-based video captioning.
4. Challenges in video captioning include dynamic scenes, frame redundancy, illumination, and motion effects.
5. Efforts to address multilingual video captioning, including translation of datasets like MSR – VTT into Hindi.

III. Proposed System:

3.1 Visual Feature Encoding:

- The system employs a conventional encoder-decoder technique for visual feature encoding.
- The input video is divided into "n" segments spaced by 16 for creating the visual context vector.

- The visual context vector is represented as " $f = \{s_1, s_2 \dots s_n\}$ " and is obtained using a pretrained C3D network.
- The visual context vector has a dimension of " $f \in R^{n \times m_x}$ " where " $m_x = 4096$."
- To reduce feature dimensionality and prevent redundancy, average pooling with a filter size of 5 is applied.
- A sequence of averaged pooled features is created to preserve temporal relationships, which are then passed to the decoder for caption generation.



3.2. Caption Generator:

- Two independent LSTMs are used as decoders.
- Input captions are processed through an embedding layer to obtain a dense embedded representation.
- The first LSTM is initialized with a visual context vector and takes the dense embedded representation of input words as input.
- The second LSTM takes both the concatenated visual context vector and the dense embedded representation as input.

- An element-wise dot product is performed between the output of the two LSTMs to generate the correct word in a sequence.
- Mathematical representation of the procedure includes equations for embedding input captions, LSTM outputs, and the loss function.
- Equation 1 represents the embedding of input captions (X) with weights (W_e) and biases (b_e).
- Equation 2 and Equation 3 denote the outputs of the first and second LSTMs, respectively.
- The cross-entropy loss function, represented in Equation 5, is used to maximize the likelihood of the correct word at a specific time step "t" while minimizing the loss. " $y_0 \dots T$ " represents the generated words in a sequence, and "F" is encoded visual features.

$$\tilde{y} = W_e X + b_e \quad (1)$$

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

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

$$y_t = softmax(z_1 \odot z_2) \quad (4)$$

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

IV. Ablation Study:

4.1. Dataset:

- The VATEX dataset, released in 2020, is used for evaluating the proposed framework.
- The dataset provides ten captions for each video in both Chinese and English, encouraging multilingual video captioning.
- The dataset includes public and private test sets, with ground truth available for the public test set and withheld for the private test set.

4.2. Evaluation Metrics:

- Multiple evaluation metrics are employed to assess the quality of automatically generated captions.
- The following metrics are used: BLEU (B), CIDEr, ROUGE-L, and METEOR.
- Automatic scores are evaluated by the organizers of the VATEX challenge.

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

4.3. Experimental Setup:

- During training, markers "BOS >" and "EOS >" are appended to each caption to indicate the start and end of the caption generation process.
- The allowed number of caption words is capped at 30.
- A vocabulary of 15,000 words is obtained using the Stanford tokenizer.
- In the testing phase, caption generation begins by monitoring the start marker and the visual context vector.
- The learning rate is 2×10^{-4} , and an ADAM optimizer with cross-entropy loss function is used.
- LSTMs have hidden units set to 512, and a dropout of 0.5 is used to mitigate overfitting.

- The proposed architecture undergoes training for 50 epochs with batch sizes of 64 and 256, and the smaller batch size shows better performance in terms of automatic evaluation scores.

4.4. Results:

- Results on both public and private test sets are presented.
- The proposed model achieves a CIDEr score of 0.24 on the public test set and 0.27 on the private test set.
- Evaluation metrics including BLEU, CIDEr, ROUGE-L, and METEOR are reported.
- Qualitative analysis of the proposed system's output reveals that it generates long, grammatically correct, and natural descriptions.
- Comparisons with other models on the private test dataset show that the proposed method performs competitively.




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

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.

		
<p>Output A man is lifting a heavy weight over his head and then drops it</p> <p>(a)</p>	<p>Output A group of people are skiing down a hill and one of them falls down</p> <p>(b)</p>	<p>Output A woman is demonstrating how to apply mascara to her eye-lashes</p> <p>(c)</p>

V. Conclusion and Future Work:

- In this paper, we introduced a novel framework called aLSTMs, designed to optimize relevance loss and semantic cross-view loss simultaneously.
- The paper describes the pLSTM framework used in the VATEX2020 challenge for video captioning.
- Video captioning is a challenging task due to the complexity and diversity of video content.
- The proposed model addresses these challenges by employing an encoder-decoder architecture for generating English captions.
- Visual features are encoded using a pretrained Convolutional Neural Network (CNN).
- The decoder consists of two parallel Long Short-Term Memory networks (LSTMs) that fuse visual features with word embeddings in two different ways simultaneously.
- The final description is generated by calculating a dot product between the outputs of the two LSTMs.
- The proposed system achieved BLEU-4 scores of 0.20 and 0.22 on public and private video captioning test sets, respectively.
- The system's performance is compared to other models proposed by challenge participants.
- Future work will focus on improving the model's computational efficiency and the accuracy of automatic evaluation scores.

This paper presents a video captioning model that successfully generates descriptions for videos and aims to improve its efficiency and evaluation scores in the future.

VI. Acknowledgements:

- The research work is supported by the Scheme for Promotion of Academic and Research Collaboration (SPARC) with the project code P995.
- This support comes from the Ministry of Human Resource Development (MHRD), Government of India, under the grant number SPARC/2018-2019/119/SL(IN).

References:

Among the total 37 references, here are some commonalities and patterns:

Dataset Creation and Utilization:

- References [1], [3], and [10] discuss the creation and use of large-scale datasets for video-and-language research, such as VATEX and MSR-VTT.
- References [4], [5], [6], [9], and [37] involve generating captions or translations for images and videos using different approaches and languages.
- Reference [8] specifically focuses on Assamese news image caption generation.

Video Captioning Models and Approaches:

- References [11], [15], [23], [24], [25], and [28] discuss various methods and models for video captioning, including recurrent neural networks (RNNs), attention mechanisms, and convolutional architectures.
- References [12] and [22] provide comprehensive reviews of methods and challenges related to video description.

Evaluation Metrics:

- References [4] and [32] discuss evaluation metrics for assessing the quality of generated captions, such as BLEU and CIDEr.
- Reference [34] introduces METEOR as a language-specific translation evaluation metric.

Image Captioning and Multimodal Approaches:

- References [13], [19], [26], [27], and [28] explore deep learning models and approaches for image captioning.
- Reference [16] discusses automated textual descriptions for a wide range of video events, including human actions.
- Reference [29] presents a video-to-text framework using an automatic shot boundary detection algorithm.

Deep Learning and Medical Image Analysis:

- *References [30] and [31] extend beyond video* and image captioning to deep learning in medical image analysis.

---X---X---X---