An attention-based hybrid deep learning approach for bengali video captioning

AUTHORS :Md. Shahir Zaoad, M.M. Rushadul Mannan, Angshu Bikash Mandol, Mostafizur Rahman, Md. Adnanul Islam, Md. Mahbubur Rahman

Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka 1216, Bangladesh

PUBLISHED BY: Elsevier B.V. on behalf of King Saud University

Article history:

Received 2 July 2022

Revised 6 November 2022

Accepted 25 November 2022

Available online 5 December 2022

ABSTRACT

- Video captioning is an automated process to add text descriptions to videos.
- Bengali video captioning is an underexplored field compared to English.
- The research focuses on generating Bengali captions for videos and finding the best model for this task.
- Several sequence-to-sequence models, including LSTM, BiLSTM, and GRU, are used, taking video frame features from CNN models (VGG-19, Inceptionv3, ResNet50v2).
- An Attention mechanism is incorporated, a first in Bengali video captioning.
- A new Bengali video captioning dataset is created from an English dataset using a deep learning translator and manual editing.
- Model performance is evaluated using BLEU, METEOR, and ROUGE metrics.
- An attention-based hybrid model outperforms existing models, setting a new benchmark for Bengali video captioning.

INTRODUCTION

- Video captioning is the process of generating textual descriptions from videos, combining Natural Language Processing (NLP) and Computer Vision (CV) techniques.
- Video captioning involves two primary tasks: video feature extraction and generating textual descriptions.
- It's a challenging task as videos are more dynamic and contain more information than static images.
- The growth of video content on platforms like YouTube and Facebook has increased the need for analyzing and describing video content.
- Automatic caption generation for images and videos is crucial for various purposes, including news, emergencies, and weather forecasts.
- Most video captioning research has been focused on the English language.
- Bengali, despite being one of the most spoken languages globally (7th worldwide with 268 million speakers), has limited research in video captioning.
- Possible reasons for the limited research in Bengali video captioning include the scarcity of suitable Bengali datasets and the complexity of the task.
- Developing video captioning in Bengali can benefit the Bengali-speaking community and various application domains.
- **LSTM and Vanishing Gradient**: Conventional RNNs face vanishing gradient issues and are limited in handling information from many time steps. LSTM (Long Short-Term Memory) addresses these challenges with memory units to manage hidden conditions and long-term dependencies.
- **GRU as an Alternative**: GRU (Gated Recurrent Unit) also uses gates to control information flow, with a simpler architecture that makes it faster to train compared to LSTM.
- **BiLSTM for Enhanced Information**: Bidirectional LSTM (BiLSTM) increases available information to the network, improving context accessibility to the algorithm.

- **Incorporating Attention Mechanism**: This research marks the first-ever attempt in Bengali video captioning to incorporate the attention mechanism, enhancing caption generation.
- **Language Effect Analysis**: The study evaluates the language effect on the performance of different RNN models in Bengali video captioning.
- **Dataset Development**: A novel dataset for Bengali video captioning is developed from the MSVD dataset.
- **Major Contributions Summarized**:
    - Exploration of state-of-the-art CNN and RNN models to identify the best CNN-RNN combination for Bengali video captioning.
    - Introduction of the attention mechanism in Bengali video captioning and a comparative analysis of language effects on model performance.
    - Development of a new dataset for Bengali video captioning, achieving benchmark results in the field.

LITERATURE SURVEY

- **LSTM and Video Captioning**: LSTM-based models have been effectively used in video captioning, addressing the limitations of conventional RNNs by handling long-term dependencies and incorporating memory units.
- **BiLSTM for Bidirectional Modeling**: Bidirectional LSTM (BiLSTM) enhances video data encoding by considering both forward and backward dependencies, offering a more comprehensive understanding of video content.
- **Introduction of Attention Mechanism**: The introduction of an attention mechanism in video captioning architectures allows for the selection of salient features and consideration of correlations between textual and visual content.
- **Visual Attention in LSTM Networks**: Dai et al. proposed an end-to-end two-stream attention-based LSTM network to overcome the issue of ignoring visual attention. It selectively focuses on relevant features in the input.
- **Attention-Based Dual Learning**: Ji et al. introduced an attention-based dual learning (ADL) approach that combines information from input videos and generated captions, using a multi-head attention mechanism.
- **Transformer-Based Models**: Lin et al. proposed an end-to-end transformer-based video captioning model that processes densely sampled video frames, showing the potential benefits of using more densely sampled video frames.
- **Multilingual Video Captioning**: Several studies have explored video captioning in languages other than English, such as Chinese and Hindi, using attention mechanisms, and large datasets.
- **Video Captioning in Bengali**: Despite extensive work in different languages, video captioning in Bengali remains unexplored, primarily due to the lack of large, semantically sound datasets.
- **Bengali Image Captioning**: Some related work in Bengali includes the creation of a Bengali image captioning dataset (BNLIT), transformer-based models, and attention-based encoder-decoder models for image captioning.
- **Transformer-Based Models for Bengali**: Transformer-based models have shown promise in Bengali image captioning, improving performance and training speed.

EXPERIMENTAL SETUP

- **Google Colaboratory Usage**: The majority of experiments are conducted using Google Colaboratory, a Python development environment that provides essential deep learning libraries like NumPy, TensorFlow, and Keras. A local machine is also used for dataset preparation and video feature extraction, using the Spyder IDE.
- **Hyperparameter Selection**: The choice of hyperparameters is crucial for model performance. Various hyperparameters are considered to determine the best combination for each specific model.
- **Dataset Creation**: Scarcity of proper Bengali datasets for video captioning is a challenge. To address this, an English video captioning dataset (MSVD) is used, and a deep learning-based translator is employed to automatically translate the English captions into Bengali. Manual efforts are undertaken to refine the quality of machine-translated Bengali captions.
- **Manual Evaluation of Translations**: Nine participants with proficiency in both Bengali and English languages evaluate the machine-translated captions for consistency with natural Bengali language. This refined dataset is then used for training and evaluation.
- **Training**: Training deep learning models involves mapping between video input and corresponding captions. Different combinations of CNN (VGG19, ResNet50v2, Inception v3) and RNN models (LSTM, GRU, BiLSTM) are trained, along with an attention mechanism. The choice of epochs is crucial to avoid overfitting or underfitting, and early stopping is used to address this concern.
- **Hyperparameter Selection for Training Accuracy**: The best hyperparameter combination for training accuracy is selected for each combination of CNN and RNN models, which are then used for performance evaluation.

RESULT AND CONCLUSION

Quantitative Analysis with Greedy Search:

- Quantitative evaluation is conducted on a testing dataset consisting of 100 video snippets.
- Evaluation is performed for each combination of CNN and RNN models, incorporating an attention mechanism.
- Greedy search, which selects the most likely word at each stage, is used for caption generation, resulting in speed but potentially suboptimal output.
- Performance scores for BLEU, METEOR, and ROUGE metrics are reported for each CNN and RNN combination with and without the attention mechanism.

Quantitative Analysis with Beam Search:

- Beam search is a heuristic search method that considers multiple alternatives at each decoding stage.
- A beam width of 3 ($k = 3$) is used, and the search procedure stops when the maximum sequence length or a likelihood threshold is reached.
- Performance scores (BLEU, METEOR, ROUGE) are reported for beam search for each CNN and RNN combination, with and without the attention mechanism.

Qualitative Analysis:

- Sample outputs of the proposed models are presented to illustrate the quality of generated captions.
- Predicted captions convey the intended meaning of actions in video snippets and closely resemble reference captions.
- The qualitative analysis is performed on a subset of 35 video clips and reference captions, validated using evaluation metrics.

Performance Comparison:

- A performance comparison is made between the proposed models for video captioning and existing models for video and image captioning in Bengali.
- While video captioning in Bengali has seen limited work, several studies have been conducted on Bengali image captioning.
- Videos are considered continuous sequences of images, and the field of video captioning in Bengali is highlighted.

CONCLUSION

- **Rigorous Experimentation**: The research conducted extensive experiments with various neural network architectures to identify the best-performing model for Bengali video captioning.
- **Dataset Development**: A significant contribution of this research is the creation of a syntactically and semantically consistent dataset derived from MSVD, which can be used for future research in this field.
- **State-of-the-Art Models**: State-of-the-art models were trained on the dataset to determine the model with the highest accuracy. The attention mechanism was introduced to achieve benchmark performance.
- **Best Performing Model**: LSTM combined with VGG19 was identified as the best-performing model among generic RNN architectures. The attention-based GRU with VGG19 produced more natural captions, making it the overall best performing model.
- **Comparative Analysis**: A thorough comparative analysis was conducted on the model's performance using two different search techniques and three popular evaluation metrics (BLEU, METEOR, and ROUGE).
- **Future Work**: While this research established a solid Bengali dataset for video captioning, future work should focus on expanding the dataset for more comprehensive training. Addressing specific language features of Bengali remains a challenge, and further work is needed. Additionally, future research can explore models for longer video clips, generating longer captions, and detecting multiple actions in a single video.