Paper title: An attention-based hybrid deep learning approach for Bengali video captioning

Paper link: https://dl.acm.org/doi/abs/10.1016/j.jksuci.2022.11.015

1. **Summary**
   1.1. **Motivation:**
   This paper aims to establish the best model for generating Bengali captions from Bengali videos which is very prominent in the current working area of research.
   1.2. Contribution:
   The optimal CNN-RNN model combinations with attention mechanisms have been determined in this paper, which has also produced a novel Bengali dataset derived from an English dataset.
   1.3. Methodology:
   This paper conducted its work by using several RNN models like GRU, LSTM, and BiLSTM using encoder-decoder format for taking the video frame features as the input, accomplished its extraction by several CNN models- VGG-19, InceptionV3, and ResNet50V2and provided output through textual description with the addition of attention mechanism and did performance evaluation based on powerful evaluation metrics like- BLEU, ROUGE, and METEOR which strengthen this work potency.
   1.4. Conclusion:
   The best-performing model for Bengali video captioning was found after extensive testing with several noteworthy neural network systems.

2. **Limitations**
   2.1. First Limitation:
   This model does not support longer video clips which is a big issue as videos on different platforms are larger in length as well as it can not generate longer captions and it can not detect more than one action.
   2.2. Second Limitation:
   During dataset translation from Bengali to English, some slang words remain untranslated and for several special characters with Bengali prefixes, and suffixes, the associated valid letters and characters disappeared.

3. **Synthesis:**
   Video captioning is very crucial specifically for the Bengali language because of the emergence of Bengali videos on social platforms, and there has been very limited work done so far in the research zone. Considering these facts, this model may create a huge impact on our research area and open up a vast platform to extensively work on this domain with more precise datasets and more versatile video features like work on multiple actions and length of a video and captions based on longer videos clips.