Captioning of Image Conceptually Using BI-LSTM Technique

Thaseena Sulthana, Kanimozhi Soundararajan, T. Mala, K. Narmatha, G. Meena

Captioning of Image conceptually using BI-LSTM technique

Thaseena Sulthana, kanimozhi.S, Mala. T, Narmatha. K, Meena. G

College of Engineering, Anna University, Chennai, India

E- mail: kanimozhiist21@gmail.com

Abstract:

Due to the fact of increase in amount of video data each day, the need for auto generation of captioning them clearly is inevitable. Video captioning makes the video more accessible in numerous ways. It allows the deaf and hard of hearing individuals to watch videos, helps people to focus on and remember the information more easily, and lets people watch it in sound- sensitive environments. Video captioning refers to the task of generating a natural language sentence that explains the content of the input video clips. The events are temporally localized in the video with independent start and end times. At the same time, some events that might also occur concurrently and overlap in time. Classifying the events into present, past and future as well as separating them based on their start and end times will help in identifying the order of events. Hence the proposed work develops a captioning system that clearly explains each visual feature that is present in the image conceptually. The Blended-LSTM (Bl- LSTM) model with the help of Xception based Convolution Neural Network (CNN) with Fusion Visual Captioning (FVC) system achieves it with the BLEU score of 75.9 %.

Keywords: Blended-LSTM, Fusion Visual Captioning, CNN, Sports video

Introduction:

Video captioning is multimedia analysis which is used to generate a natural language sentences by understanding the given video by considering the events [8] taking place in the video. It creates a great impact in computer vision. Automatic video caption generation [9] includes the understanding of many background concepts. It also includes the detection of every occurrence in the video such as objects, actions taking place in the video, scenes taking places, person to person relations in the context of the video, person to object relations in the video and the temporal order of the events. Video captioning also requires translation [7] of the extracted visual information and grammatically correct natural language description.

Conceptual translation plays major role in captioning where most of the misconception occurs. To overcome that pipeline based methods comes into existence as in [12], where visual features are extracted using machine learning algorithm. Extracted features are indexed with words in the vocabulary for

framing as sentences using some neural network techniques [6]. So framing of words has to be consummate conceptually, which happens only when there is extraction of required features. Feature extraction is effective when there is suitable region [1] inside the image is selected. Hence Region of Interest (ROI) is determined based on edge detection, contour, histogram, etc.

Conceptual captioning can be done only when there is expected regions features are extracted in a frame. Thus the focus of the proposed work is to select keyframes meaningful which leads to proper feature extraction as well as frame sentence conceptually.

Related works:

Recently the auto captioning of images based on the visual information available within it has become major research focus. For producing captions, screening of features is to be done from the initial stage (Keyframe generation) itself. Visual features are more in videos which may be missed if we consider the entire part. So there is a need to convert video into frames [10] for further processing. Selecting a frame from the entire set of frames that represent more visual features is again a huge task. One method for detecting the Keyframe is based on adaptive threshold [4] method. In this threshold value is calculated for a specific region, so that threshold values vary for different regions in a single image. Another method [2] based on HSV histogram, which transfer high-dimensional abstract video image into quantifiable low dimensional data. It results with more appropriate keyframes with low redundancy rate.

Extracting features from the selected keyframes is again difficult task as there is a chance of selecting inappropriate regions. Various methods have been proposed for selecting the regions such as iteration algorithm, region growing and edge detection [6] and so on. In many of these works manual declaration of initial seed is needed which again may create low accuracy. To overcome these issue neural based models like encoder- decoder [5], multimodal layered, etc has been proposed. In which major focus is towards location as well as fixed size region, where there is no need for the manual interruption. The main contribution of this work is as follows:

i) Generating Keyframe from the video using Shot based Adaptive Threshold technique (SAT),

ii) Extracting selective features from the selected Keyframe using Xception based on CNN,

iii) Caption is produced with the extracted features using Fusion Visual Captioning (FVC) model.

Fusion captioning model:

The Fusion Visual Captioning (FVC) model consists of important features of CNN and RNN together. Initially the videos are converted into frames for the clear extraction of object features. The frames are generated ate the rate of 20fps for the video of length 3 min. The keyframe is selected using proposed Shot based Adaptive Threshold (SAT) method as shown in table 1.

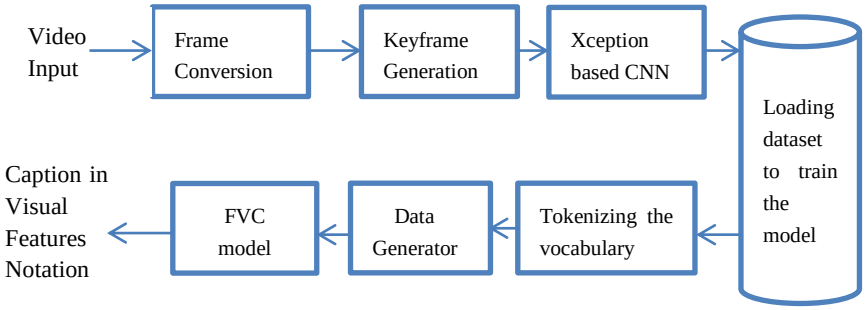

Fig: 1- System design for proposed Bl-LSTM caption model

Next is the visual feature extraction module in which Xception based CNN is employed. CNN is best known visual feature extractor model [14] that fetches the vectors values based on convolution method [15]. Flickr8k_text dataset is used for captioning which consist of five possible labels for each image. In order to integrate the pre-trained Xception based CNN with Flickr dataset last layer is removed to get 2048 feature vector.

In SAT algorithm (table-1) threshold $\emptyset_t$ is calculated with mean of the neighborhood pixels subtracted with the constant $\Delta\emptyset$, whereas constant $\Delta\emptyset$ is calculated by subtracting the weighted sum of the neighborhood pixels. The result of the SAT algorithm is the set of keyframes along with major the features incorporated for further processing.

During training phase, a dictionary is created that maps the images used in the dataset along with its descriptions. A unique index value using tokenizer function [11] is assigned to make matching process easier. Display pattern is created with a starting and ending identifier to each word as (start, object, preposition, action, sport and end). The vector value from the last fully connected layer of Xception based CNN is given to the input layer of the LSTM. To make the training process easier for over 10,000 images the data generator module is used. It creates batches using the input feature vector with its text and corresponding predicted output sequence. Finally, the decoder in FVC model embeds the result from previous module with the LSTM for building the expected captions for each image.

Table-1

Proposed SAT Algorithm
Variables: F_{in} and F_e are the initial and final pixel values of selected region. T is the average pixel value.
Convert pixel values in region P to binary region using the threshold $\emptyset_t$ Assume N is the sum of the number of the non- zero pixels within F_{in} and F_e If T is larger than a predefined threshold, β Then $\emptyset_t = \emptyset_t + \Delta\emptyset$ Repeat the procedure from step 1 Else $\emptyset = \emptyset_t$ End

Experimental Results:

The input to the Bl-LSTM model is given in the form of video due to the fact that future work is focused towards captioning of video sequences. Thus the keyframe is selected in the sense that shows maximum visual representation of entire frame using algorithm in table-1. Highly featured frame is given as input for the Xception based CNN module which extract 2048 vector values for each image with a dense layer. The dimension is reduced to 256 nodes as in [3] which help for embedding the words towards LSTM layer. Totally of about 7822 words are there in the developed vocabulary. The maximum length of description is fixed to 32 as there are five possible captions for each image. The last layer in the FVC model have nodes (256) equal with that of size (7822) of the generated vocabulary.

Fig.2: Examples of final output generated by BL-LSTM caption model

Initially the proposed model predicts the caption as per the visual information presents in the frame as shown in Fig.2. Effectiveness of the system is evaluated based on the BLEU score as in Eq-1, 2 [13] that compare the true captions with the predicted caption and generate the final score.

$$BLEU = min\left(1, \frac{Final_{Length}}{Ref_{Length}}\right)\left(\prod_{i=1}^{n} precision_i\right)^{\frac{i}{n}} \quad \text{Eq-1}$$

$$Precision_i = \frac{\sum_{snt \in cand-corpus} \sum_{i \in snt} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{snt^* \in cand-corpus} \sum_{i^* \in snt^*} m^{i^*}_{cand}} \quad \text{Eq-2}$$

Where n is the size, m^i_{cand} the count of i-gram in candidate, m^i_{ref} is the count of i-gram of reference, w^i_t is the total i-grams in candidate conversion.

In Fig.3 x-axis represents 21 possible captions generated using the proposed Bl-LSTM model and y-axis is the corresponding BLEU score of good vs bad captions of the tested sample images. Good as well as bad captions are differentiated in the graph using two indicators. First with BLEU score, that produces as resultant value of Eq-1. Second, using true vs predicted caption developed by the proposed model. All the generated captions scores are lies between 7.0 – 8.0 BLEU values, which resemble the improvement in the Bl-LSTM system performance.

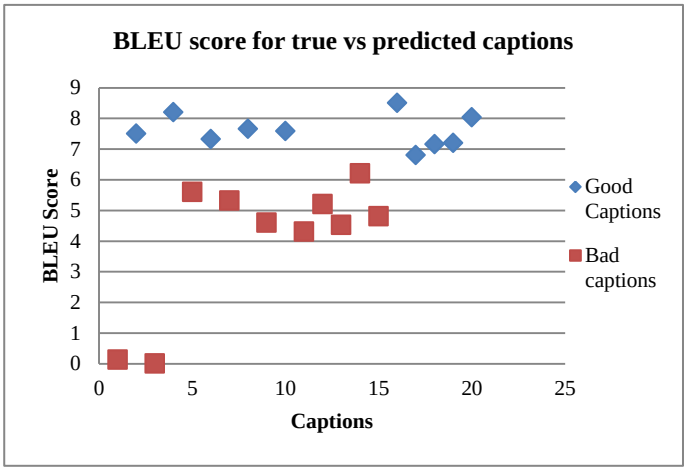

Fig-3: Final Score values for predicted good vs bad captions

The comparative performance result of our Blended-LSTM (Bl-LSTM) model with other state-of-art methods in Flicker30k dataset is given in Table-2. The BLEU value shows the efficiency of proposed Bl-LSTM model in predicting good (true matching with predicted) captions.

Table-2: Performance comparison with state-of-art methods

Methods	**BLEU score**
m-RNN	60.0
R-LSTM	67.7
Bl-LSTM	75.9

Conclusion:

In this paper, the proposed Bl-LSTM aims at captioning the frame with high visual features representation in it. Hence the major focus is divided into two sections. First, entire features of the frame are to be extracted including the foreground and background objects. Second, proper indexing must be done by matching the extracted features with the vocabulary. So fine tuning starts at the keyframe extraction stage itself using the SAT method, which generates the best representation from set of frames for further processing. Xception based CNN model helps in fetching the highlighted information from the key frames. Next the FVC which is a fusion model that combines meaningfully matched captions among the five possible captions to each keyframe. Efficacy of the proposed model is evaluated using the BLEU score and comparative study is done as well with state-of- art methods. Final BLEU score shows that, proposed Bl-LSTM model outperforms other methods by resulting with 75.9 % score value.

Future enhancement is focused towards captioning including visual features of all the events present in entire video. Second, improvement is focused on accuracy for generated captions that should be done, which helps in providing clear description about the video.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 961-971)(2016).
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 961-970)(2015).
3. Escorcia, V., Heilbron, F. C., Niebles, J. C., & Ghanem, B.: Daps: Deep action proposals for action understanding. In Proceedings of the Conference on Computer Vision (pp. 768-784)(2016).
4. Heilbron, F. C., Niebles, J. C., & Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1914-1923)(2016).

5. Lee, S., & Kim, I.: Multimodal feature learning for video captioning. Mathematical Problems in Engineering, (2018).
6. Pirsiavash, H., & Ramanan, D.: Parsing videos of actions with segmental grammars. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 612-619)(2014).
7. Rahman, T., Xu, B., & Sigal, L.: Watch, listen and tell: Multi- modal weakly supervised dense event captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8908-8917)(2019).
8. Xu, H., Li, B., Ramanishka, V., Sigal, L., & Saenko, K.: Joint event detection and description in continuous video streams. In Proceedings of the IEEE winter conference on applications of computer vision (WACV) (pp. 396-405)(2019).
9. Yao, T., Mei, T., & Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 982-990)(2016).
10. Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T.: Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4894-4902)(2017).
11. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78(2014).
12. Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4584-4593)(2016).
13. Zhao, S., Ding, G., Gao, Y., & Han, J. : Approximating discrete probability distribution of image emotions by multi-modal features fusion. Transfer, 1000(1), 4669-4675(2017).
14. Zheng, A., Xu, M., Luo, B., Zhou, Z., & Li, C.: Class: Collaborative low-rank and sparse separation for moving object detection. Cognitive Computation, 9(2), 180-193(2017).
15. Zhong, G., Yan, S., Huang, K., Cai, Y., & Dong, J.: Reducing and stretching deep convolutional activation features for accurate image classification. Cognitive Computation, 10(1), 179-186(2018).