An Efficient Keyframes Selection Based Framework for Video Captioning

Alok Singh[1], **Loitongbam Sanayai Meetei**[1], **Salam Michael Singh**[1], **Thoudam Doren Singh**[1], and **Sivaji Bandyopadhyay**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
{alok.rawat478,loisanayai,salammichaelcse,thoudam.doren,sivaji.cse.ju}@gmail.com

Abstract

Describing a video is a challenging yet attractive task since it falls into the intersection of computer vision and natural language generation. The attention-based models have reported the best performance. However, all these models follow similar procedures, such as segmenting videos into chunks of frames or sampling frames at equal intervals for visual encoding. The process of segmenting video into chunks or sampling frames at equal intervals causes encoding of redundant visual information and requires additional computational cost since a video consists of a sequence of similar frames and suffers from inescapable noise such as uneven illumination, occlusion and motion effects. In this paper, a boundary-based keyframes selection approach for video description is proposed that allow the system to select a compact subset of keyframes to encode the visual information and generate a description for a video without much degradation. The proposed approach uses $3 \sim 4$ frames per video and yields competitive performance over two benchmark datasets MSVD and MSR-VTT (in both English and Hindi).

1 Introduction

In recent years, we witnessed the exponential growth in multimedia data (especially video) over the Internet (Singh et al., 2019). This large volume of data creates a need for automatic video understanding systems that can describe the video's content, event and action with a short textual description. There are many applications of automatic video description generation such as efficient content indexing and searching, storytelling, the amalgamation of speech with the video description can also help visually impaired people and if the video description approaches are successful in generating a short textual description of the real-world scenes, then the robots can converse with humans effectively (Singh et al., 2020a; Aafaq et al., 2019). The task of generating image and video descriptions are very closely related. But the presence of both temporal and spatial information, which varies with the time in a video, makes the task of video description generation more challenging than image description. So for generating an informative and visually related video description, the efficient encoding of both spatial and temporal features of the video is the basic step in any video description framework.

Being an interdisciplinary problem of both computer vision and natural language processing, researchers form both domain have proposed a numerous approach for describing a video precisely, but still, much work is needed to be done. A video consists of a sequence of similar frames, but various editing effects are included in the video due to the recent advancements in technologies and these editing effects affect the process of selecting informative frames from the video. Existing approaches such as (Singh et al., 2021b; Nabati and Behrad, 2020b; Venugopalan et al., 2014; Gao et al., 2020) encode the visual features of the video either by segmenting the video in the interval of some arbitrary value k (most of time $k = 16$) or by selecting first n frames. Meanwhile, the process of encoding visual features by equal interval sampling does not guarantee that all the selected frames are informative because, in a video it is possible that the selected frames are suffering from different types of noise such as uneven illumination, motion blur, occlusion and object zoom-in/out effects (Chen et al., 2018). In this paper, we address the issue of selecting informative frames by using color information based shot boundary detection followed by keyframe selection from each shot. A shot in a video is a set of continuous similar frames captured uninterruptedly and when the content of these frames get changed, it creates

Proceedings of the 18th International Conference on Natural Language Processing, pages 240–250
Silchar, India. December 16 - 19, 2021.

two types of boundaries (transitions) in the video namely - abrupt and gradual transition. The novel contribution of the proposed work are:

i. We propose a plug-and-play keyframe selection module based on visual color information of the input frames by employing a long video temporal segmentation algorithm. This module is designed by considering the three basic requirements of any video understanding model: flexibility, efficiency and effectiveness.

ii. In the proposed framework, a temporal soft attention mechanism is employed that will focus more on the responsible keyframes from the set of selected keyframes for an input video at every time step.

iii. We perform a detailed qualitative and quantitative analysis on the results generated by the framework for MSVD, English MSR-VTT[1] and Hindi MSR-VTT[2] datasets.

The organization of the remaining part of the paper is as follows. Section 2 report a review of related work on video description. Section 3 discussed the proposed approach. A detail experimental studies is reported in Section 4 followed by conclusion in Section 5.

2 Related Work

Earlier, the process of bridging the gap between visual content and natural language was considered a challenging task. However, with the success of deep learning approaches in recent years, the gap has been reduced. Till now, the approaches proposed for video description can be categorised into three phases: classical method based phase, statistical method based phase and deep learning-based phase (Aafaq et al., 2019). Further, the related work in this section is divided into three subsections based on the type of approach is employed: Sequence-to-Sequence based approaches (S2S), attention-based approaches and boundary-based approaches.

[1] http://ms-multimedia-challenge.com/2017/challenge.

[2] https://github.com/alokssingh/MSR-VTT-Hndi-captionig

2.1 S2S video description approaches

In the early stage of the video description task, most of the approaches proposed for video description are motivated by image description approaches (Singh et al., 2021b). The pioneering work in video description is based on the prediction of SVO (Subject, Verb and Object) and fill them into a predefined templates (Aafaq et al., 2019; Singh et al., 2020a). Recently, the encoder-decoder based framework gains more popularity. Venugopalan et al. (2014) proposed a sequential Convolutional Neural Network (CNN) and Long Short Term Memory based model (CNN-LSTM) for video description. In this framework, Venugopalan et al. (2014) extracted frame-level features for each sampled frame (1 in every ten frames) using a pre-trained model and then passed all the extracted features through a mean pooling layer to get a single vector representation for the whole video. Finally, a description for an input video is generated by employing a two stacked LSTM. Although the proposed approach outperforms the previous SVO based baseline models, the model has few drawbacks such as, it does not preserve the temporal relationship among the frames and represent the whole video with a single features vector which reduces the task of video description to image description due to which lots of vital visual information get lost. To address the issues of previous model Venugopalan et al. (2014) proposed a end-to-end sequential model (Venugopalan et al., 2015) which consists of two LSTM layer. The first LSTM layer encodes the extracted visual features and the second LSTM layer receives the null padded input word concatenated with hidden representation from the first layer and generates an output description. Using a multi-stage refining algorithm (Nabati and Behrad, 2020a) proposed video description framework with content-oriented beam search. This approach involves three stages, namely feature extraction, content-oriented beam search and sentence refining. Wang et al. (2020) proposed a sequential model for encoding spatio-temporal visual representation. Unlike other sequential frameworks in this model, the sequential frame is encoded at every time step and generates the most related word at each step. In this approach, a *"Real-Time Encoder"* is introduced that uses history information of previous time steps to extract informative spatio-temporal visual representation. Recently, the work on de-

scribing a visual entities into multiple languages gained more popularity with the Hindi image captioning (Singh et al., 2021c,a), multi-modal machine translation (Meetei et al., 2019; Singh et al., 2021d) and the release of novel Video to Text (VATEX) (Wang et al., 2019) multilingual dataset (including Chinese and English) for video description. Furthermore, Singh et al. (2020b) proposed a pLSTM framework in the VATEX video captioning challenge. In this framework, two parallel LSTM are employed, which receives the input in different manners. The pLSTM framework was unable to outperform the baseline VATEX model (Wang et al., 2019) in the VATEX dataset.

2.2 Attention based approaches

On observing the effectiveness of soft attention (Xu et al., 2015) and bottom-up, top-down attention (Anderson et al., 2018) in generating visually related words at every time step in image captioning, some approaches based on attention are also proposed in the video description. Yao et al. (2015) proposed an approach that utilizes both temporal and spatial structure of the video for extracting visual features. They employed a temporal attention mechanism for selecting a relevant segment from the video. This approach only considers the first 240 frames of the video, which is the shortcoming of the proposed approach. A hierarchical Recurrent Neural Network h-RNN is proposed by Yu et al. (2016), it exploited the temporal and spatial attention for extracting visual features using Gated Recurrent Unit (GRU). Few other attention-based video captioning frameworks are proposed in (Li et al., 2018; Xiao et al., 2020). Apart from temporal attention, semantic attention is also used for generating temporally and semantically correct video descriptions. Gao et al. (2020) and Xu et al. (2019) proposed a method for video description by exploiting the combination of both semantic and temporal attention. Recently, Singh et al. (2021b) proposed hybrid attention mechanize for Hindi video captioning by utilizing the concept of visual sentinel gate (Lu et al., 2017) proposed for image captioning. The approach proposed in Singh et al. (2021b) differs from Lu et al. (2017) in terms of the implementation of the attention block.

2.3 Boundary aware approaches

An open domain video contains many editing effects, which generates a large number of shots in a video. A video consists of a large number of redundant frames and to minimise the redundancy and improve the computation time, various boundary aware approaches are proposed. (Baraldi et al., 2017) proposed a novel LSTM cell for detecting the temporal boundaries in a video and generates a visual feature vector for the whole video. (Shin et al., 2016) proposed SBD based method for the generation of the multiple sentence video description. In this method, the video is divided into shots by employing sliding windows of different lengths. Based on the assumption that selection of informative frames can improve generated description and reduce computational time (Chen et al., 2018) proposed a plug-and-play PickNet model for selecting relevant frames using reinforcement learning, then finally descriptions are generated for each video. (Sah et al., 2020) proposed a video description approach for a video surveillance system. In this approach for the multi-stream hierarchical video description model, a recurrent layer with a soft attention mechanism is employed with dynamical detected abrupt transitions. Real-time analysis is performed in support of the statement that a video description model could be useful for a video surveillance system. Few other recently proposed boundary-based video description approaches are (Shi et al., 2020; Jin et al., 2020).

3 Proposed Approach

The proposed approach consists of two modules: Boundary detection and Keyframe selection module (Sec 3.1) and Description generation module (Sec 3.2).

3.1 Boundary detection and keyframe selection phase

The main objective of boundary detection is to spot the position at which the content of the video gets changed. In this paper, we are focusing on spotting these abrupt transitions. A color histogram-based approach proposed in Mas and Fernandez (2003) is adopted to detect the temporal discontinuity in a video. The color histogram-based approach is computationally efficient and prevalent in various computer vision-related tasks. In boundary detection and keyframe selection algorithm initially, the color histogram of each frame is computed and then the histogram difference (Δ_i) is computed between the histogram of consecutive frames using Equation 1 where M is the number of bins and h_i is the color histogram of i^{th} frame in a video se-

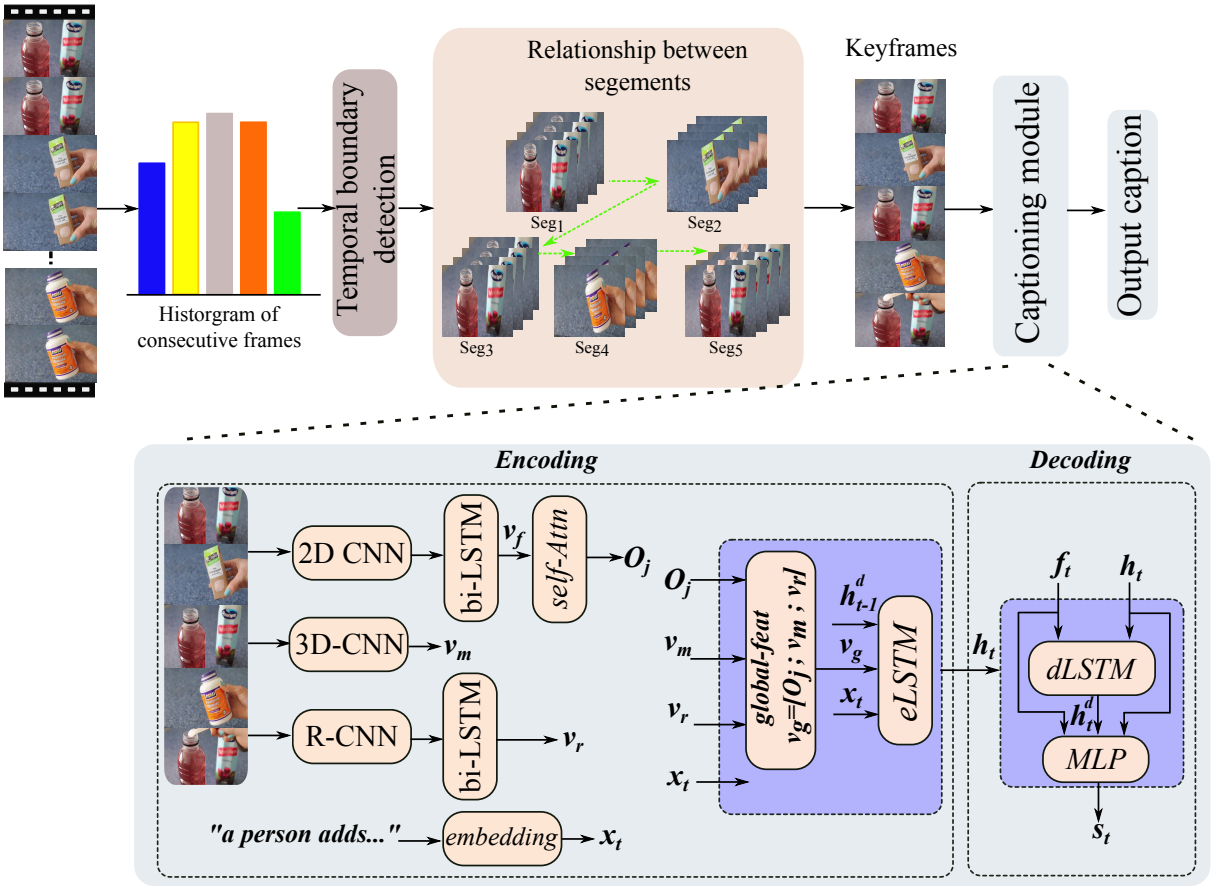

Figure 1: Pictorial representation of whole boundary based video description framework

quence.

$$\Delta_i = \left(\sum_{j=1}^{M} (h_i(j) - h_{i-1}(j))^2 \right)^{\frac{1}{2}} \quad (1)$$

After the computation of histogram difference, to declare temporal boundary at a particular location an adaptive threshold γ ($\gamma = mean(\Delta) + k \times stdev(\Delta)$) employed in Singh et al. (2019) is used, here the value of constant k is set to 5.2 after fine tuning. The mathematical expression for the declaration of temporal boundary is shown in Equation 2, where B_i record the boundary locations.

$$B_i = \begin{cases} i, & (\Delta_i \geq \gamma) \&\& (\Delta_i > \Delta_{i-1}) \\ & \&\& (\Delta_i > \Delta_{i+1}) \\ continue, & Otherwise \end{cases} \quad (2)$$

Keyframe selection: After detecting the temporal boundaries, a video is divided into different segments containing similar frames within it. A simple and computationally efficient approach for video description is the utilization of information present in keyframes of the video rather than using several redundant frames. In the proposed approach, a keyframe is selected from each segment which we get after temporal segmentation. The frame which is selected as a representative frame has a minimum distance to the other frames present in the same shot (segment). This approach is also adopted by Li et al. (2017) for video summarization. Mathematically, it can be described as follow:

$$\underset{i \in [1, n_f]}{min} \left(\sum_{t=1, t \neq i}^{n_f} Euclidean(\tilde{h}_i - \tilde{h}_t) \right) \quad (3)$$

Where n_f is the number of frames in a shot, $\tilde{h}_i$ is color histogram of the selected frame and $\tilde{h}_t$ represent the histogram of other frames within the shot. In this way, a keyframe of each shot is selected based on the visual similarities within the shot.

3.2 Description generation phase

After selecting keyframes for an input video of N frames, we extracted three types of features that are visual appearance features (v_f) which are ex-

Algorithm 1 Temporal segmentation of video with key frame selection

Input: Video, V
Output: Boundaries, $Keyframes$

procedure shot_detection(V)
  $F \leftarrow cv2.VideoCapture(V)$
  $hist_0 \leftarrow cv2.calcHist(vid.read(F(0)), ch, m, h_s, r)$
  for $i = 1$ `to` $length(F)$ **do**
    $hist_i \leftarrow cv2.calcHist(vid.read(F(i)), ch, m, h_s, r)$
    $\Delta_i \leftarrow \left(\sqrt{\sum_{j=1}^{M}(hist_i(j) - hist_{i-1}(j))^2}\right)$
    $hist_{i-1} \leftarrow hist_i$
  for $i = 1$ `to` $length(\Delta) - 1$ **do**
    if $\Delta_i \geq \gamma \&\& (\Delta_i > \Delta_{i-1}) \&\& (\Delta_i > \Delta_{i+1})$ **then**
      $B_i \leftarrow record\ i^{th}$
    else
      $continue$

Function $Keyframe\_sel(frames, B)$
for $k = 1$ to S **do** ▷ i, j ∈ [1,nf], i≠ j
  for $i = 1$ to n_f **do**
    for $j = 1$ to n_f **do**
      $diff_{(i,j)} \leftarrow record\ total\ dissimilarity\ difference$
  $Keyfrm[k] \leftarrow min(diff_{(i,:)})$
$return\ Keyfrm$

tracted using 2D CNN (He et al., 2016a), motion features (v_m) using 3D CNN and object features (v_r) extracted using R-CNN (Ren et al., 2015). Then, the appearance and object features (v_f and v_r) are post-processed using Bi-LSTMs.

3.2.1 Context rich encoding with self attention

Since a video has multiple actions and events, so some the events in earlier frames are responsible for the occurrence of other related events in forthcoming frames. Considering this fact, the post-processed appearance features are passed through a self-attention layer to get more context rich encoded visual features. Self-attention allows the model to look at the visual features of other selected keyframes for better visual encoding. So, initially using the visual features v ($v = vf$) the value of key $K(v)$, value $V(v)$, and query $Q(v)$ are computed using Equation 4 where W_k, W_q and W_v are the weight metrics to be trained.

$$K(v) = W_k v \quad V(v) = W_v v \\ and \quad Q(v) = W_q v \tag{4}$$

Then, to compute context-rich self-attention feature maps ($O_{j...S}$) dot-product attention is applied as follow:

$$\mathbf{O}_j = W_g \left(\sum_{i=1}^{S} \alpha_{i,j} V(v_i) \right) \\ where, \quad \alpha_{i,j} = softmax(\frac{K(v)^T Q(v)}{d_k}) \tag{5}$$

In the Algorithm 1, ch = channels, m = mask, h_s = histSize and r = ranges

In the above equations, S is total number of shots (segments), $v_f \in \mathbb{R}^{S \times l}$, $W_{k,v,q,g} \in \mathbb{R}^{l \times \tilde{l}}$ and the dimension of $K(v)$, $V(v)$ and $Q(v)$ is set to 64 and $d_k = 8$ following the work of Vaswani et al. (2017) for effectiveness of Self attention mechanism. For the encoding of words in the reference caption, the dense embedded representation which is obtained from a word embedding layer is passed to an encoder LSTM (eLSTM). The eLSTM takes the word embedding of input word (x) at current time step, global visual features (v_g) and decoder LSTM's hidden state of last time step as shown Equation 6.

$$h_t = eLSTM(x_{t-1}, v_g, h^d_{t-1}) \tag{6}$$

3.2.2 Decoder

After getting encoded contextually rich representation of input word (h_t) and visual appearance features (O_j) they are passed to the decoder along with motion features (v_m) and object features (v_r). Before passing the self attentive appearance features (O_j) and object features (v_r) to decoder LSTM (dLSTM) they are passed through an attention layer ($Attn(V_x, h)$) as shown in Equation 7 where V ($V = O_i\ or\ v_r$) is encoded features and W_* ($* = h, v$) are trainable weights and b_t is bias.

$$Attn(V_x, h) = \phi(V_j, \alpha_i) \quad where,\ \phi = \sum_{i=1}^{k} \alpha_i V_{i,j} \\ and, \quad \alpha_i = softmax(W_a tanh(W_v V + W_h h_{t-1} + b_t)) \tag{7}$$

After getting attentive appearance features and object features from the attention layer they are concatenated with motion feature (v_m) and passed to decoder LSTM (dLSTM) as shown in Equation 8 where [;] denotes concatenation and h^d_t is used in Equation 6.

$$f_t = [Attn(O_j, h_t); Attn(v_r, h_t); v_m] \\ h^d_t, c^d_t = dLSTM([f_t; h_t]) \tag{8}$$

Further, the word probability s_t at every time step is decoded as follow:

$$s_t = softmax(MLP([f_t; h^d_t; h_t])) \tag{9}$$

The cost function used for maximizing the likelihood of the correct word and minimizing the loss of the model is given by Equation 10.

$$Loss = -\sum_{t=0}^{T} \log Pr(s_t | s_{t-1}, \ldots s_0; F) \tag{10}$$

Table 1: Results of proposed approach on MSVD dataset and its comparison with other approaches.

Methods	BLEU-4	METEOR	CIDEr	ROUGE
Mean pooling				
-$AlexNet$ (Venugopalan et al., 2014)	31.20	26.90	-	-
-$AlexNet$ $(COCO)$ (Venugopalan et al., 2014)	33.30	29.10	-	-
Attention				
- SA (Yao et al., 2015)	40.28	29.00	-	-
-MMN (Li et al., 2018)	48.00	31.60	68.80	-
-$BP-LSTM$ (Nabati and Behrad, 2020b)	42.90	32.00	62.20	68.30
Boundary + Attention				
-$Boundary-aware$ (Baraldi et al., 2017)	42.50	32.40	63.50	-
-$PickNet$ (Chen et al., 2018)	46.10	33.10	69.20	69.20
-MHB (Sah et al., 2020)	43.00	**33.20**	71.10	68.70
Proposed (v_f)	45.55	30.37	68.73	66.44
Proposed (v_f+v_m)	48.66	29.90	68.33	65.97
Proposed (v_f+ v_m+ v_r)	**50.75**	32.50	**71.13**	**70.44**

4 Experimental result and discussion

4.1 Datasets

To manifest the effectiveness of the proposed approach, three benchmark datasets are employed that are: *Microsoft research video description corpus (MSVD)* (Chen and Dolan, 2011), *English Microsoft research video to text (MSR-VTT)* (Xu et al., 2016) and *Hind Microsoft research video to text (hi-MSR-VTT)* (Singh et al., 2021b). The hi-MSR-VTT dataset is recently released dataset for motivating the research on generating video descriptions in the native language. The MSVD dataset include $1,970$ videos with on average 40 descriptions for each video while the en-MSR-VTT and hi-MSR-VTT dataset include $10K$ videos with corresponding 20 descriptions. Table 2 reports the detailed statistics of all the datasets.

Table 2: Detail statistics of all the datasets

Datasets	#Training videos	#Val videos	#Test videos
MSVD	1200	100	670
MSR-VTT	6513	497	2990
hi-MSR-VTT	6513	497	2990

4.2 Metrics

For the validation of the generated descriptions, we employs Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Metric for Evaluation of Translation with Explicit Ordering (METEOR[3]) (Banerjee and Lavie, 2005), Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) and Recall Oriented Understudy of Gisting Evaluation (ROUGE-L) (Lin, 2004). For generating the scores for above discussed automatic evaluation metrics Microsoft COCO[4] toolkit is employed.

4.3 Parameter setting and model implementation

As discussed in section 3.2 for experimentation we employ $ResNet152$ (He et al., 2016b) as 2D CNN model for extracting appearance features of keyframes and C3D model (Karpathy et al., 2014; Tran et al., 2015) as 3D CNN for extracting the motion features. For extracting the region features Faster-RCNN (Ren et al., 2015) trained by (Anderson et al., 2018) is employed, this model extract 36 region features for each keyframes. The model is trained with $ADAM$ optimizer with learning rate 1e-4 and the learning rate is divided by 10 at every 10^{th} epoch. The number of LSTM hidden units is set to 512 and during training, the model having the best $METEOR$ score is saved. To avoid overfitting, a dropout of 0.3 is employed. In the proposed work, we tried to search optimal parameters that work comparatively better than other baseline models in all the datasets, which will minimize the time and effort required to search the best parame-

[3]The METEOR score for Hindi text is generated using: https://github.com/anoopkunchukuttan/meteor_indic

[4]https://github.com/tylin/coco-caption

Table 3: Results of proposed approach on en-MSR-VTT dataset and its comparison with other approaches.

Methods	BLEU-4	METEOR	CIDEr	ROUGE
Mean/Max pooling				
-$LSTM-GAN$ (Yang et al., 2018)	36.00	26.10	-	-
Attention				
-M^3 (Wang et al., 2018)	38.13	26.58	-	-
-MMN (Li et al., 2018)	37.50	26.40	-	-
-$ReBiLSTM$ (Bin et al., 2018)	33.90	26.20	-	-
-$BP-LSTM$ (Nabati and Behrad, 2020b)	36.60	27.00	40.50	58.70
-$MCTA$ (Wei et al., 2020)	38.50	26.90	**43.70**	-
Boundary + Attention				
-$Boundary-aware$ (Baraldi et al., 2017)	36.80	26.70	41.20	58.50
-$PickNet$ (Chen et al., 2018)	**38.90**	**27.20**	42.10	**59.50**
Proposed (v_f)	35.42	25.21	35.36	57.83
Proposed (v_f+v_m)	35.95	25.39	35.66	57.38
Proposed (v_f+ v_m+ v_r)	37.18	26.17	40.90	59.41

ters according to the dataset. All the parameter settings are the same throughout the experimentation for all the datasets. A beam search approach with beam size 7 is employed during testing to generate the final description.

4.4 Results and discussion

Comparison with existing methods: To analyse the performance proposed keyframe based video captioning approach we compare proposed approach with existing methods. For the better understanding and fair comparison all the existing methods are categorised into three type of captioning approaches that are *mean/max pooling*, *attention* and *boundary+attention*. The approaches such as *AlexNet*, *LSTM-GAN* and *pLSTM* are mean/max pooling based approaches, *MMN*, *BP-LSTM*, M^3, *ReBiLSTM* and *MCTA* are attention based while the *PickNet*, *Boundary-aware* and *MHB* are boundary based approaches which employ attention as well.

Table 1, 3 and 4 report quantitative results on MSVD, en-MSR-VTT and hi-MSR-VTT datasets. Our proposed approach outperforms other existing methods on the MSVD and the hi-MSR-VTT dataset, on 3 out of 4 metrics by a reasonable margin. While on the en-MSR-VTT dataset, our model reports comparable scores, although the *PickNet* model reports high scores, but in terms of the average number of frames used to achieve competitive performance, the proposed approach outperforms *PickNet* model. Our model uses $3 \sim 4$ frames per video whereas the *PickNet* model employ $6 \sim 8$ frames per video.

Ablation study: The proposed approach consist of two stage: boundary detection and description generation phase. To evaluate the effectiveness of all the employed visual features the proposed model is experimented with different variations such as with only appearance features, with appearance and motion features and with all three appearance (v_f), motion (v_m) and region features v_r. Table 1, 3 and 4 reports the score of proposed model with all the variation. The effectiveness of proposed method increases when all three features are employed which can be clearly seen in table 1, 3 and 4. In order to validate that whether the proposed model generates more fluent and adequate description along with high automatic scores, we perform a qualitative analysis. Figure 2 shows the description generated by the proposed model along with the output generated by $BP-LSTMs$ and ground truth (GT). On observing the output generated by the proposed model for the videos shown in Figure 2, it is clear that the keyframes based approach can generate better description than $BP-LSTM$, which employ n frames for visual encoding.

4.4.1 Analysis of picked keyframes

We also analysed the efficiency of the boundary based keyframe selection algorithm for selecting the most representative frame from multiple segments of the video. Figure 3 shows the distribution of keyframes selection for both the datasets. From Figure 3 it is observed that for the majority of videos, less than 8 frames are picked as a keyframe

Table 4: Results of proposed approach on hi-MSR-VTT dataset and its comparison with other approaches.

Methods	BLEU-4	METEOR	CIDEr	ROUGE
Mean/Max pooling				
-$pLSTM$ (Singh et al., 2020b)	26.10	33.00	28.50	51.20
Attention				
-$VA + SA$(Singh et al., 2021b)	36.20	39.30	**36.90**	59.80
-RNM (Tan et al., 2020)	38.80	39.10	36.00	60.70
Proposed (v_f)	34.02	38.40	30.76	58.09
Proposed (v_f+v_m)	36.11	39.95	31.12	58.95
Proposed (v_f+ v_m+ v_r)	**41.01**	**44.10**	32.85	**60.80**

Figure 2: Sample videos selected from each dataset with their ground truth (GT) and generated output

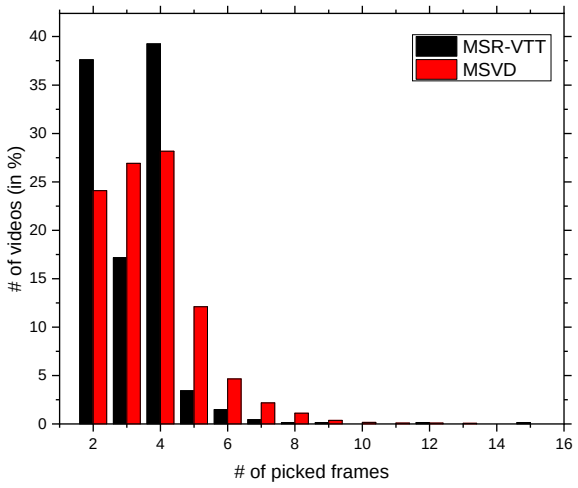

Figure 3: Statistic of picked keyframes for both the datasets

which is due to shorter video length. A video can have a single shot or multiple shots. For a single-shot video, 4 keyframes are selected at the interval of 16 and for a multi-shot video, the keyframe is selected using an approach discussed in section 3.1. From Figure 3, it is clearly observed that around 39% and 28% of videos in MSR-VTT and MSVD respectively, are single-shot videos. The average number of keyframes selected per video is $3 \sim 4$ for both MSVD and MSR-VTT dataset, which helps in avoiding unnecessary visual encoding of redundant frames and signify the efficiency of the proposed approach. Sample examples of picked keyframes are included in supplementary file.

5 Conclusion

In this paper, we employ a boundary-aware keyframe selection framework that acts as a plug-and-play module for downstream video-related tasks, such as video description and video classification. The objective of the boundary aware keyframe selection framework is to select a compact subset of keyframes for input video, which minimises the unnecessary processing of visually similar frames and ensures no degradation in the quality generated description. In the proposed approach, $3 \sim 4$ frames are selected for an input video, which is more efficient than the existing $PickNet$ model, which picks $6 \sim 8$ frames for each video. The experimental results show that the keyframes-based approach can outperform existing methods by picking keyframes and extracting different visual features such as appearance, motion and region features.

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2017. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1666.

Yi Bin, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. 2018. Describing video with attention-based bidirectional lstm. *IEEE transactions on cybernetics*, 49(7):2631–2641.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2018. Less is more: Picking informative frames for video captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 358–373.

Lianli Gao, Xuanhan Wang, Jingkuan Song, and Yang Liu. 2020. Fused gru with semantic-temporal attention for video captioning. *Neurocomputing*, 395:222–228.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. 2020. Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888*.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Jiatong Li, Ting Yao, Qiang Ling, and Tao Mei. 2017. Detecting shot boundary with sparse coding for video summarization. *Neurocomputing*, 266:66–78.

Wei Li, Dashan Guo, and Xiangzhong Fang. 2018. Multimodal architecture for video captioning with memory networks and an attention mechanism. *Pattern Recognition Letters*, 105:23–29.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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*, pages 375–383.

Jordi Mas and Gabriel Fernandez. 2003. Video shot boundary detection based on color histogram. In *TRECVID*.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Masoomeh Nabati and Alireza Behrad. 2020a. Multi-sentence video captioning using content-oriented beam searching and multi-stage refining algorithm. *Information Processing & Management*, 57(6):102302.

Masoomeh Nabati and Alireza Behrad. 2020b. Video captioning using boosted and parallel long short-term memory networks. *Computer Vision and Image Understanding*, 190:102840.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Shagan Sah, Thang Nguyen, and Ray Ptucha. 2020. Understanding temporal structure for video captioning. *Pattern Analysis and Applications*, 23(1):147–159.

Xiangxi Shi, Jianfei Cai, Jiuxiang Gu, and Shafiq Joty. 2020. Video captioning with boundary-aware hierarchical language decoding and joint video prediction. *Neurocomputing*, 417:347–356.

Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. 2016. Beyond caption to narrative: Video captioning with multiple sentences. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3364–3368. IEEE.

Alok Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021a. 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*, pages 65–73. Springer.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2020a. A comprehensive review on recent methods and challenges of video description. *arXiv preprint arXiv:2011.14752*.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2020b. Nits-vc system for vatex video captioning challenge 2020. *arXiv preprint arXiv:2006.04058*.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021b. Attention based video captioning framework for hindi. *Multimedia Systems*, pages 1–13.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021c. An encoder-decoder based framework for hindi image caption generation. *Multimedia Tools and Applications*, pages 1–20.

Alok Singh, Dalton Meitei Thounaojam, and Saptarshi Chakraborty. 2019. A novel automatic shot boundary detection algorithm: robust to illumination and motion effect. *Signal, Image and Video Processing*, pages 1–9.

Salam Michael Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021d. 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)*, pages 2–11, Online (Virtual Mode). INCOMA Ltd.

Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. 2020. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Huiyun Wang, Chongyang Gao, and Yahong Han. 2020. Sequence in sequence for video captioning. *Pattern Recognition Letters*, 130:327–334.

Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. 2018. M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7512–7520.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Ran Wei, Li Mi, Yaosi Hu, and Zhenzhong Chen. 2020. Exploiting the local temporal information for video captioning. *Journal of Visual Communication and Image Representation*, 67:102751.

Huanhou Xiao, Junwei Xu, and Jinglun Shi. 2020. Exploring diverse and fine-grained caption for video by incorporating convolutional architecture into lstm-based model. *Pattern Recognition Letters*, 129:173–180.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Yuecong Xu, Jianfei Yang, and Kezhi Mao. 2019. Semantic-filtered soft-split-aware video captioning with audio-augmented feature. *Neurocomputing*, 357:24–35.

Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. 2018. Video captioning by adversarial lstm. *IEEE Transactions on Image Processing*, 27(11):5600–5611.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.