A Multimodal Framework for Video Caption Generation

RESHMI S. BHOOSHAN1 AND SURESH K.2, (Senior Member, IEEE)
1College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala 695016, India
2Government Engineering College, APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala 695016, India

Corresponding author: Reshmi S. Bhooshan (reshmibhooshan@cet.ac.in)

ABSTRACT Video captioning is a highly challenging computer vision task that automatically describes the video clips using natural language sentences with a clear understanding of the embedded semantics. In this work, a video caption generation framework consisting of discrete wavelet convolutional neural architecture along with multimodal feature attention is proposed. Here global, contextual and temporal features in the video frames are taken into account and separate attention networks are integrated into the visual attention predictor network to capture multiple attentions from these features. These attended features with textual attention are employed in the visual-to-text translator for caption generation. The experiments are conducted on two benchmark video captioning datasets - MSVD and MSR-VTT. The results prove an improved performance of the method with a CIDEr score of 91.7 and 52.2, for the aforementioned datasets, respectively.

INDEX TERMS Video captioning, discrete wavelet convolutional model, multimodal feature extraction, visual attention predictor.

I. INTRODUCTION
Video caption generation aims to automatically generate meaningful natural language descriptions about the video. For this, a clear understanding of the semantic details as well as the contextual visual relationship between different objects present in the video is needed. Many algorithms have been developed by researchers in this area of computer vision, for generating descriptions closer to the human perception level. Video caption generation is important in a variety of real-world applications such as content-based video retrieval, video comprehension generation, automatic assistance devices for the visually impaired, subtitles creation in videos, intelligent driving assistance systems, video surveillance and so on [1], [2], [3], [4], [5], [6], [7].

Most of the deep learning frameworks employed for caption generation use an encoder-decoder structure. The encoder utilizes a Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) for extracting the visual and semantic details in the video. It generates a feature vector representation corresponding to the visual content in the video, which is then given to a decoder structure having sequential models such as RNN, Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) that does the visual to natural language translation [8], [9], [10]. In an early attempt made by Venugopal et al., a novel video-to-text generation methodology is presented, which extends image captioning methods by incorporating a 2D-CNN network along with mean pooling and RNN decoder structure [11]. But this method fails to use the temporal details present in the video for caption generation. Subsequently, an S2VT model is proposed in [12] that uses a stacked LSTM network to learn the temporal information in a sequence of frames and then produce a sequence of words. Later, attention mechanisms are included in the spatial as well as temporal domains to achieve better performance [13], [14]. Video descriptions can also be generated by employing attention in the decoder section as well as using multimodal fusion mechanisms of visual, text and audio features [8], [15].

The caption generation of video clips is more challenging compared to image captioning because of the involvement of shots, scenes, activities, random motion of objects of
The results are compared with the existing state-of-the-art using two benchmark datasets - MSVD and MSR-VTT and LSTM. Due to the presence of self-attention layers, the decoder network can exploit spectral information in the input sequences. This network architecture utilizing Discrete Wavelet Transform (DWT) based CNN for extracting more finer visual details from the video frames, which enables better video caption generation. The new architecture can exploit spectral information in the video frames along with the spatial, semantic and temporal details for caption generation. The proposed model includes a Global Feature Extractor (GFE) incorporating DWT based Convolutional Neural Network (2D-WCNN) for the extraction of global features from the video frames, a Contextual Object Relationship Extractor (CORE) for finding out the contextual relationship between different objects in the frames and a Temporal Feature Extractor (TFE) that consists of 3DCNN model for getting the dynamic features from the video. A visual attention predictor network is also incorporated that extracts the attention from these features and finally visual-to-text translation is done using a caption decoder network as in transformer [16]. This greatly solves the long-range dependency issues in the sequential models with self-attention mechanism and allows parallel computing. Due to the presence of self-attention layers, the decoder network can enhance the quality of visual-to-text translation by considering the word-to-word, object-to-object as well as object-to-word interactions in the input sequences. This network utilizes global dependencies between input and output for providing improved performance compared to RNN and LSTM.

The main contributions of the proposed framework are:

- A 2D-WCNN structure using two-level DWT decomposition and CNN layers is employed for extracting the global features in the video frames. The utilization of DWT helps to include the fine-grained spatial, spectral and semantic details in the frames.
- A Contextual Object Relationship Extractor (CORE) makes use of the feature maps obtained from the 2D-WCNN for predicting region proposals and computes the contextual relationship between the different frames in the video.
- A multimodal visual feature attention network that concurrently computes global, contextual and temporal feature attention, capable of increasing the efficiency and prediction accuracy of the entire methodology.

The effectiveness of the proposed method is evaluated using two benchmark datasets - MSVD and MSR-VTT and the results are compared with the existing state-of-the-art methods using the evaluation metrics BLEU, METEOR and CIDEr.

The paper is organized as follows: Section II gives a brief review of the research works existing in this area. The details regarding the proposed architecture and experimental results are described in Section III and IV, respectively and Section V concludes the paper.

II. LITERATURE SURVEY

In early approaches, caption generation of videos is done using classical template-based techniques that employed the SVO-triplets - Subject (S), Verb (V) and Object (O) [17]. These triplets are found out individually and they are combined to form a sentence. Many encoder-decoder architectures have been proposed that use 2DCNN/3DCNN structures as the encoder for generating feature representations and sequential models like RNN, LSTM and GRU as the decoder for language translation [18], [19]. A two-step captioning approach that learns the correspondence between semantic representation labels and verbalization before translating it to natural language is introduced in [20]. An unsupervised Multirate Visual Recurrent Model (MVRM) is presented by [21] that is capable of handling motion speed variations in video frames with a bidirectional reconstruction technique. Dual Memory Recurrent Model (DMRM) utilizing global and temporal details along with semantic supervision for accurate detection of region-of-interest is proposed in [22]. Captions are also generated with latent topic guidance [23], Time Boundary-aware LSTM cell [24], Boosted and Parallel Long Short-Term Memory Networks (BP-LSTM) [25] and Object Relational Graph with Teacher-Recommended Learning (ORG-TRL) system [26]. Another technique utilizes two steps - video Part-of-Speech (POS) tagging and visual cue translation [27]. This can be accomplished using a mixture model for converting visual features to lexical words and sentence templates comprising of POS tags.

A few algorithms are developed by considering both the spatial as well as temporal features simultaneously along with attention mechanisms. A multimodal stochastic recurrent neural network (MS-RNN) that makes use of latent stochastic variables is presented by [28] for video captioning. Hierarchical encoder structures are also proposed by [10] and [24] that give more attention to the temporal details of the video. Descriptions are made by employing attention in the decoder section [8], [15] as well as multimodal fusion mechanisms with aural features in the video [29]. A multimodal temporal attention mechanism incorporating image, motion and audio features is given in [30]. This architecture is developed by assuming that different modalities carry different task-relevant information at different time instances. In another work, captions are produced using a co-attention model based recurrent neural network (CAM-RNN) consisting of a visual attention module, a text attention module and a balancing gate [31]. This algorithm is capable to perform adaptive detection of the most relevant regions in the image and thus
concentrates on the relevant words or phrases in the generated sentence. Recently, transformer-based decoder models are proposed for the generation of video descriptions. Masked transformers can be used for generating end-to-end video captions that employs a masking network to produce differentiable masks from event proposals and thereby maintaining consistency between the proposal and captioning during training [32]. A sparse boundary-aware transformer (SBAT) aligned with a cross-modal encoding scheme can be used to enhance the multimodal interaction thereby providing better captions as mentioned by [33]. Coherent paragraph generation can be made possible using Memory-Augmented Recurrent Transformer (MART) having a memory module to acquire a highly summarized memory state from the video segments and the sentence history [34].

In all these methods, textual descriptions are created by considering only the spatial and temporal features in the video. Enhanced video captions can be achieved by extracting spectral information using DWT and contextual information through visual relationship detection between objects. In order to obtain a good degree of video understanding, 2D multiresolution discrete wavelet convolutional model is used in this work for global and contextual visual feature extraction, which in turn are then fed to the VAP along with the temporal features for identifying the attentive regions in the video. These multimodal features are combined in VTT for producing semantically improved natural language captions. The following section describes the proposed model that integrates the aforementioned techniques.

III. PROPOSED METHOD

Fig. 1 illustrates the overall framework of the proposed methodology. In this, video captions are generated using a sequence of visual feature representations obtained from VFE and VAP. The caption generation is accomplished using VTT consisting of a multimodal attention network along with a softmax layer that describes the video with a sequence of encoded words, \( S_v = [W_1, W_2, \ldots, W_l] \), with \( W_i \in \mathbb{R}^{N_v} \), where \( l \) is the length of the generated caption and \( N_v \) is the vocabulary size. A detailed description of different elements of the proposed model is given in the following subsections.

A. VISUAL FEATURE EXTRACTOR

The VFE comprises of three submodules: GFE, CORE and TFE for the generation of multimodal features.

1) GFE

It employs a 2D-WCNN network having a modified ResNet-50 structure [35] with two-level DWT decomposition to provide a better time-frequency representation of the frames. In DWT decomposition, each frame in the video is decomposed into four subbands, which highlight frequency details in the image. Hence the utilization of the DWT pre-processing stage together with the convolutional neural network helps to extract some of the distinctive spectral features that are more predominant in the subband levels of the frames in addition to the spatial, semantic as well as channel details. The detailed structure of 2D-WCNN network is shown in Fig. 2. Each of the input frames, resized to
224 × 224, is subjected to two-level multi-resolution decomposition using DWT and is further fed to the CNN structure comprising of five levels - CONV1 to CONV5. In level-1 wavelet decomposition, the low pass and high pass filtering of the input frame produces an approximation subband \((ll_1)\) and three detailed subbands \((lh_1, hl_1 \text{ and } hlh_1)\). The \(ll_1\) subband is further filtered out into four subbands - \(ll_2, lh_2, hl_2\) and \(hh_2\) in level-2 decomposition. Then these subbands of \(R, G\) and \(B\) components obtained from both the levels are stacked together and each of these is concatenated with the max pooled output of \(CONV_1\) and \(CONV_2\) layers, respectively as shown in Fig. 2. All four subbands need to be included to extract the features from them in the convolutional layers because each subband carries distinguishable features, which are very essential to have a good visual representation of the frame. The configuration of levels from \(CONV_3\) to \(CONV_5\) is the same as that of ResNet-50 network with three-layer bottleneck blocks along with residue connections. The first convolutional layers of the levels \(CONV_3\) to \(CONV_5\) are having stride 2. The details regarding the filters used in various convolutional layers along with their output sizes are summarized in Table 1. Batch normalization together with the ReLU activation function is used in all the layers. Padding is also employed in every layer. The extracted global feature maps from the \(CONV_5\) layer are given to the CORE for finding out the relationship between the various objects in the frames. These feature maps are also given to the VAP for computing global attention.

2) CORE
Better captions can be generated only by considering the contextual relationship between the objects in the video. The CORE employs a modified configuration of Faster RCNN structure [36]. It utilizes the feature maps produced by the \(CONV_5\) layer of the 2D-WCNN network in the GFE to predict the region proposals for detecting the object relationships as shown in Fig. 3. The object regions are identified using the 2D-WCNN network and Region Proposal Network (RPN), similarly as that of Faster RCNN, along with classifier and regression layers for creating bounding boxes. The identified objects are then paired and different subimages are created with each having the identified object pairs. For uniformity, these subimages are resized to \(32 × 32\) and are given to the CNN layers having two sets of 64 filters, each with receptive field \(3 × 3\) as shown in Fig. 3. The obtained feature maps highlight the spatial relationship between the object pairs. The spatial relation feature maps of each object pair is then stacked together and are given to \(1 × 1 \times 64\) convolution layer to form the contextual spatial relation feature map. These features of the CORE are passed through a fully connected layer to produce a 2048-dimensional feature vector, \(V_{ci}\).

3) TFE
The temporal features are extracted using 3DCNN (C3D network).

The three feature vectors - \(V_{gi}\), \(V_i\) and \(V_t\) representing global features, local features and motion features in the video, respectively, are then provided to the VAP module.

### B. VISUAL ATTENTION PREDICTOR

Content rich video caption generation necessitates a clear understanding of semantics in the video. To accomplish this, the VAP network is incorporated in the model that utilizes *Scaled Dot-product Attention* to compute the global, local and temporal attention features. It consists of a multi-head attention mechanism having \(H\) parallel attention layers or heads, each computing *Scaled Dot-product Attention* on an input having a set of queries \((Q)\), keys \((K)\) and values \((V)\), each of dimension \(\mathbb{R}^d\). In the case of global attention network, all the \(Q, K\) and \(V\) values are set to be equal with \(V_{gi}\) as shown in Fig. 1. Thus the output of the global attention network is given by the expression,

\[
F^g_{att} = \left(h_1 \oplus h_2 \oplus \ldots \oplus h_H\right)W^o
\]

\[
h_i = G_{att}(QW^Q_i, KW^K_i, VW^V_i)
\]

\[
G_{att} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_i}}\right)V
\]

where \(\oplus\) denotes concatenation, \(G_{att}\) represents global *Scaled Dot-product Attention* with independent head projection matrices, \(W^Q_i, W^K_i\) and \(W^V_i\), for \(i = 1, 2, \ldots, H\) in \(\mathbb{R}^d\). \(W^o \in \mathbb{R}^{d \times d}\) is the output projection matrix that combines the output from the various heads, each having a dimension of \(\frac{d}{H}\). Similarly, the outputs for the CORE and temporal attention networks, \(F^c_{att}\) and \(F^t_{att}\) will be computed as in (1) through (3) with the \(Q, K\) and \(V\) inputs set to \(V_{ci}\) and \(V_{ti}\), respectively.

In the VAP module, \(N\) identical attention subnetworks are stacked together separately each for computing the global, contextual and temporal attention features. The attended feature output from the \((i - 1)^{th}\) stage, \(F^{i-1}_{att}\) is used to produce the attended features of the next stage, \(F^{i}_{att}\), in a recursive manner.

### TABLE 1. Details of various convolutional layers in 2D-WCNN model. The input to the network is of size \(224 × 224 \times 3\). The residual blocks are mentioned in square brackets.

| Layer Name | Kernel size/No. of filters/stride | Output size |
|------------|-----------------------------------|-------------|
| \(L_1\)   | \(3 × 3/64, stride\)             | 112 × 112 × 64 |
| \(L_2\)   | \(3 × 3/128, stride\)            | 56 × 56 × 128 |
| \(CONV_1\) | \(3 × 3/64, stride\)            | 224 × 224 × 64 |
| \(CONV_2\) | \(3 × 3/64, stride\)            | 224 × 224 × 64 |
| \(CONV_3\) | \(3 × 3/64, stride\)            | 112 × 112 × 128 |
| \(CONV_4\) | \(3 × 3/64, stride\)            | 112 × 112 × 128 |
| \(CONV_5\) | \(3 × 3/64, stride\)            | 112 × 112 × 128 |
The features, $F_{at}$, are given to the normalization (Norm) and feedforward (FF) networks having two fully-connected layers, ReLU activation function and dropout layers. This introduces non-linearity in the network. In this work, the drop out ratio is set as 0.1. Residual connection and layer normalization are included in all the network layers. Thus $V_{gA}$ is computed as,

$$V_{gA} = \text{Norm}[\text{FF}(\text{Norm}[F_{g}^{g} + V_{g}]) + (\text{Norm}[F_{at}^{g} + V_{g}])]$$

(4)

Similarly, the contextual and temporal attention features are obtained as,

$$V_{cA} = \text{Norm}[\text{FF}(\text{Norm}[F_{at}^{c} + V_{c}]) + (\text{Norm}[F_{at}^{c} + V_{c}])]$$

(5)

$$V_{tA} = \text{Norm}[\text{FF}(\text{Norm}[F_{at}^{t} + V_{t}]) + (\text{Norm}[F_{at}^{t} + V_{t}])]$$

(6)

The attended output features so obtained from the global, local and temporal attention networks are multiplied together and are given to the visual-to-text translator for further processing.

**C. VISUAL-TO-TEXT TRANSLATOR**

The attended visual representations from the VAP of three attention networks, together with the attended ground truth caption word embeddings, $X_{cap}$, are fed to the multi-head attention networks of VTT section of the architecture as shown in Fig. 1. The VTT consists of one masked attention network computing the self-attention within the word embeddings. It uses a mask matrix for improving the self-attention learning process in the caption word embedding during training and each word learns or attends from the words in the previous positions of the output sequence. This self-attention layer of word embeddings is followed by multi-head attention, $M_{att}^{D}$, which computes the guided attention on the word embeddings in accordance with the attended visual representations. The $M_{att}^{D}$ consists of four VTT sub-layers stacked together to produce the attended visual-to-text features, $VT_{att}$ as given below,

$$VT_{att} = \text{Norm}[\text{FF}(\text{Norm}(F_{Datt} + X_{cap}))$$

$$+ \text{Norm}(F_{Datt} + X_{cap})]$$

(7)

The $VT_{att}$ features are fed to a linear network and finally, the prediction of words is performed by the softmax layer. During the training phase, cross-entropy loss $L_{CE}$ from all time steps is used, which is expressed as,

$$L_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_{\theta}(W_{t} | W_{1:t-1}, V_{gi}, V_{ci}, V_{gi}))$$

(8)

where $W_{1:t-1}$ represents the ground truth sequence at timestep, $t$ and $\theta$ denotes the parameters.

The model so designed has to undergone exhaustive evaluation to reveal its effectiveness in video captioning, as discussed below.

**IV. EXPERIMENTS AND RESULTS**

Both qualitative and quantitative analysis of the proposed framework has been carried out with different datasets and performance evaluation metrics. Results of this analysis and a comparative study with the state-of-the-art video captioning techniques are presented in this section.

**A. DATASETS USED**

Experiments are conducted on two benchmark datasets for video captioning: Microsoft Research Video Description Corpus dataset (MSVD) [37] and MSR-Video to Text dataset (MSR-VTT) [38].

1) **MSVD dataset**: It consists of 1,970 YouTube video clips having an average of 40 manually annotated captions per clip. For a fair comparison, we have used the split-up as proposed in [12], that consists of 1,200 videos for training, 100 videos for validation and 670 videos for testing.

2) **MSR-VTT dataset**: It is the largest video captioning dataset having 10K video clips, each annotated with 20 sentences. The standard split-up as mentioned in [38] is adopted for this dataset - 6,513 videos for training, 497 for validation and 2,990 videos for testing.

**B. PERFORMANCE EVALUATION METRICS**

The performance evaluation of the methodology is done using the evaluation metrics - BiLingual Evaluation Understudy (BLEU@4) [39], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [40] and Consensus-based Image Description Evaluation (CIDEr) [41]. These will be
denoted as B@4, MT and CD, respectively, in this work. The B@4 metric is a commonly used metric for evaluating machine translation that measures the 4-gram based accuracy. MT metric measures the harmonic mean of unigram precision and recall between the candidate and the reference sentences. It actually computes the word correlation between the two sentences. The CD metric evaluates the consensus in the generated sentence as assessed by humans. Hence these three metrics can effectively calculate the consistency between occurrences of words in the generated caption and the ground-truth descriptions.

C. IMPLEMENTATION DETAILS
Since the video datasets include videos with different frame rates varying from 6 to 60, input is resampled and made to have uniform frame rate for smooth working of the algorithm for the datasets under consideration. Hence in our method, the video clips are resampled at 10fps and 30 uniformly spaced frames are chosen from each video clip, keeping in mind that adjacent frames in the short clips included in the datasets differ very little in terms of the information content.

2D-WCNN model in the GFE module is pre-trained using ImageNet dataset [42] and C3D model in the TFE module is pre-trained using Sports-1M dataset [43]. For temporal feature extraction, we have considered non-overlapping sequence of 16 frames, same as the default settings. All the visual features are given to individual fully-connected layers with 512 units, to match with the feature dimensionality of the attention networks in the model. In VAP, four stages of attention networks are used for the extraction of attended features.

The pre-processing of all the textual descriptions are done by tokenizing with NLTK toolkit that splits the sentences into words, convert all the words to lowercase and remove punctuations. All those words having an occurrence rate less than 3 are removed. Each word in the caption is represented as a word vector using the 300 dimensional GloVe word embeddings [44] pre-trained on a large-scale corpus. For dimensionality matching, the GloVe embeddings are given to LSTM network with 512 hidden units. The maximum sentence length is limited to 20. In the visual-to-text translator network, four caption decoder stages with model dimensionality matching, the GloVe embeddings are given in Table 3.

D. SELECTION OF APPROPRIATE MOTHER WAVELET
To choose the appropriate mother wavelet, the performance of the proposed method is analyzed with two levels of wavelet decomposition on ten different mother wavelet families - Daubechies wavelets (dbN) [45], biorthogonal Wavelets (biorNr.Nd) [46], Coiflets (coifN) [47] and Symlets (symN) [46], where N represents the number of vanishing moments in the reconstruction and decomposition filters, respectively. The detailed performance results of our method for MSVD and MSR-VTT datasets are given in Table 2. A model that consists of two-level DWT decomposition based CNN along with a single attention network in VAP module is used in the experiment.

For the MSVD dataset bior2.4 secures the highest MT score of about 33.21 but bior1.5 scores better values of B@4 and CD compared to the other wavelets. For MSR-VTT, bior1.5 achieves better B@4, MT and CD score of about 40.54, 26.33 and 49.35, respectively. Hence for the 2D-WCNN network, we have chosen the bior1.5 wavelet for both MSVD and MSR-VTT datasets.

E. SELECTION OF DWT DECOMPOSITION STAGES
A detailed study regarding the results obtained for various DWT decomposition levels are carried out and the results are given in Table 3.
The method with two-level decomposition yields better results than with 1-level DWT decomposition scoring an improvement in the B@4, MT, CD values of about 0.77%, 1.39% and 1.32%, respectively, for the MSVD dataset and about 1.09%, 1.16% and 0.85%, respectively, for the MSR-VTT dataset. The method yields only slight improvements in the performance metric scores with the inclusion of three-level decomposition. Hence considering the computational complexity, a method with two-level decomposition is preferred in the proposed work.

F. PERFORMANCE EVALUATION

Both quantitative and qualitative analysis of the methodology are carried out using the evaluation metrics and is compared with the state-of-the-art methods as detailed below.

1) QUANTITATIVE RESULTS

Table 4 shows the performance results of our method on MSVD dataset along with the comparison on the state-of-the-art methods: HRNE [10], h-RNN [25], LSTM-TSA [48], M3 [49], PickNet [50], RecNet [51], DS-RNN [52], MS-RNN [29], GRU-EVE [53], GFN-POS [54], TRGCN [55], MARN [61], STAT [14] and SBAT [33] in video captioning. From Table 4, it can be noted that our algorithm outperforms the existing methods with an improved MT and CD score of 36.5% and 91.7%, respectively. It also secures a B@4 of 53.6%. This proves the ability of our method in highlighting the finer details in the input video clip. Table 5 shows the quantitative comparison results of the proposed method on MSR-VTT dataset with the existing state-of-the-art methods. These includes the methods ranked in top-3 positions of the Leaderboard of MSR-VTT Challenge 2017 - VideoLAB [56], Aalto [57] and v2t-Navigator [58] along with the methods, MTVC [59], PickNet [50], TVT [60], DS-RNN [52], GRU-EVE [53], GFN-POS [54], TRGCN [55], MARN [61], STAT [14] and SBAT [33] in video captioning. For this dataset also, it achieves an impressive B@4, MT and CD scores of about 44.9%, 29.8% and 52.2%, respectively, which indicates the better performance of our method compared to the existing methods. The inclusion of DWT in the architecture helps to extract the fine visual details present in the video clips more efficiently compared to the other methods. The method extracts three different features from the video for multimodal video representation. Then attention is captured from these three different modalities simultaneously and is combined to acquire all the attentive regions in the video that highlights the underlying video semantics. The textual attention is also interleaved with the aforementioned attention helps to generate captions that are at par with human-generated ones.

2) QUALITATIVE RESULTS

Fig. 4 illustrates the qualitative comparison of the captions generated by STAT [14] and the proposed method for sample videos from both the datasets. From the generated captions, it is evident that the proposed method understands the visual concepts in the video in a superior manner and generates captions reflecting the underlying semantics such as ‘sits on a sofa’, ‘drinks from a bottle’, ‘ride a motorbike’, ‘down the road’, ‘blue and white paper’, ‘eyes’ and ‘brush’, thus conveying more details of the video close to the human-generated ground truth captions.

G. ABLATION STUDY

Experimental studies were conducted to validate the enhanced performance of the method with the inclusion of global, contextual and temporal features in the video with and without discrete wavelet decomposition. An ablation study is also conducted to analyze the effectiveness of the inclusion of multiple attention stages in the VAP and VTT networks. Table 6 shows the results of B@4 and CD scores.

### Table 3. Performance results of the proposed method for different number of decomposition levels.

| Decomposition levels | MSVD | MSR-VTT |
|----------------------|------|---------|
|                     | B@4 | MT    | CD    | B@4 | MT    | CD    |
| 1-level              | 52.87 | 35.14 | 90.39 | 43.83 | 28.68 | 51.37 |
| 2-level              | 53.64 | 36.53 | 91.71 | 44.92 | 29.84 | 52.22 |
| 3-level              | 53.69 | 36.90 | 91.89 | 40.03 | 29.89 | 52.51 |

### Table 4. Performance comparison of our method with other state-of-the-art methods on MSVD dataset. All the values are reported as % and HIGH is good in all columns. (-) indicates that the metric is not reported.

| Method       | B@4 | MT   | CD  |
|--------------|-----|------|-----|
| HRNE [10]    | 43.8 | 33.1 | -   |
| h-RNN [25]   | 49.9 | 32.6 | 65.8|
| LSTM-TSA [48]| 52.8 | 33.5 | 74.0|
| M3 [49]      | 52.8 | 33.3 | -   |
| PickNet [50] | 52.3 | 33.3 | 76.5|
| RecNet [51]  | 52.3 | 34.1 | 80.3|
| DS-RNN [52]  | 53.0 | 34.7 | 79.4|
| MS-RNN [29]  | 53.3 | 33.8 | 74.8|
| GRU-EVE [53] | 47.9 | 35.0 | 78.1|
| GFN-POS [54] | 53.9 | 34.9 | 91.0|
| TRGCN [55]   | 52.6 | 36.3 | 89.6|
| STAT [14]    | 52.0 | 33.3 | 73.8|
| SBAT [33]    | 53.1 | 35.3 | 89.5|
| Ours         | 53.6 | 36.5 | 91.7|

### Table 5. Performance comparison of our method with other state-of-the-art methods on MSR-VTT dataset. All the values are reported as % and HIGH is good in all columns. (-) indicates that the metric is not reported.

| Method       | B@4 | MT   | CD  |
|--------------|-----|------|-----|
| VideoLAB [56]| 39.1 | 27.7 | 44.1|
| Aalto [57]   | 39.8 | 26.9 | 45.7|
| v2t-Navigator[58]| 40.8 | 28.2 | 44.8|
| MTVC [59]    | 40.8 | 28.8 | 47.1|
| PickNet [50] | 41.3 | 27.7 | 44.1|
| TVT [60]     | 40.1 | 27.9 | 47.7|
| DS-RNN [52]  | 42.3 | 29.4 | 46.1|
| MS-RNN [29]  | 39.8 | 26.1 | 40.9|
| GRU-EVE [53] | 38.3 | 28.4 | 48.1|
| GFN-POS [54] | 41.7 | 27.8 | 48.5|
| TRGCN [55]   | 44.6 | 29.5 | 51.4|
| MARN [61]    | 40.4 | 28.1 | 47.1|
| STAT [14]    | 39.3 | 27.1 | 43.9|
| SBAT [33]    | 42.9 | 28.9 | 51.6|
| ORG-TRL [27] | 43.6 | 28.8 | 50.9|
| Ours         | 44.9 | 29.8 | 52.2|
for various network configurations on MSVD and MSR-VTT datasets. Initially, the experiments are conducted in two phases. In the first phase, we used a single attention network with four sublayers to handle the concatenated visual input features - global, contextual and temporal. For this, experiments are carried out with multiple configurations as in Table 6. The configuration with Glob+Temp* achieves the maximum value of CD of about 89.6% and 50.4% for MSVD and MSR-VTT datasets, respectively, for self critical loss. In the second phase of the experimental study, we used three separate attention networks each having four attention blocks, for computing the attention of global, local and temporal features. Here also three different configurations of features are considered and obtained higher CD values of 91.7% and 52.2% for MSVD and MSR-VTT datasets, respectively, for self critical loss. It scores an improvement of about 1.9% and 1.1%, respectively, in CD value for MSVD and MSR-VTT with the inclusion of Self critical training strategy. As mentioned in Table 6, a configuration without WCNN that consists of ResNet-50 for capturing the global features of the frames, Faster RCNN network for extracting contextual information and C3D model for getting temporal details in the video are also done. It achieved a B@4 and CD value of about 89.2% and 50.6%. This enhancement is due to the extraction of spectral information along with the spatial, temporal and semantic details in the input video.

Also, experimental study has been conducted for finding out the optimal number of attention blocks with the introduction of self critical loss. The results are highlighted in Fig. 5. The system performance seems improved with the number of attention blocks in the VAP-caption decoder stages of the transformer network. But it gets saturated after a particular number of attention blocks. Thus, for the proposed model, the optimal number of attention blocks is found to be 4 with B@4 and CD values of about 53.6% and 91.7% for MSVD dataset and 44.9% and 52.2% for MSR-VTT dataset, respectively.

The quality of the captions generated by the method is illustrated with few sample video clips from both the datasets as shown in Fig. 6. Here the baseline method is the network without WCNN as mentioned above.

**H. LIMITATIONS OF THE PROPOSED WORK**

Even though the proposed method gives better performance in the reported evaluation metrics, it still has some limitations.
The method fails to generate correct contextual descriptions of few video clips because of wrong visual content interpretations. In the sample frames of the first video clip in Fig. 7, objects with reflections are visible. In this video clip, we can observe a baby in a red dress looking at himself in the mirror and kissing. Our method identifies this as “two babies in red dress are playing”, producing false result. The method also fails to extract the correct visual interpretations or semantics from those videos, having high motion complex events, similar to the one shown in the frames of the second sample video in Fig. 7, where a motorcyclist is met with an accident by losing his control over the bike and finally...
falls in the water. This activity is identified wrongly as the motorcyclist is “flying a bike” by the method. This is because the attention-based multimodal features may interact with each other degrading the performance of the method with false interpretations of the underlying semantics in the video.

Another limitation of the proposed model is the increase in time cost with the addition of a discrete wavelet pre-processing stage and VAP network that computes the global, local and temporal attention features separately.

V. CONCLUSION

In this work, a deep neural network architecture is introduced for video caption generation by exploiting multimodal feature representation in the video. In this method, the inclusion of two-level discrete wavelet decomposition in 2D visual feature representation helps to extract additional information contained in spatial, temporal and spectral domains in the video. The adoption of three separate attention networks in the visual attention predictor is responsible for extracting more attentive features, leading to more semantic captions in the visual-to-text translator. The performance evaluation of the method is carried out using two benchmark datasets and compared with existing state-of-the-art methods in video captioning.

The results obtained highlights the efficiency of the method in generating meaningful video captions. This method can be further improved by exploiting the audio features also to generate more meaningful captions. The method can be extended to generate textual descriptions of lengthy videos such as the surveillance video system to highlight the abnormal events contained in it.

REFERENCES

[1] W. Li, Z. Qu, H. Song, P. Wang, and B. Xue, “The traffic scene understanding and prediction based on image captioning,” IEEE Access, vol. 9, pp. 1420–1427, 2021.

[2] S. Amirian, A. Farahani, H. R. Arabnia, K. Rasheed, and T. R. Taha, “The use of video captioning for fostering physical activity,” in Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), Dec. 2020, pp. 611–614.

[3] N. Xu, A.-A. Liu, W.-Z. Nie, and Y.-T. Su, “Attention-in-attention networks for surveillance video understanding in Internet of Things,” IEEE Internet Things J., vol. 5, no. 5, pp. 3419–3429, Oct. 2018.

[4] C. Gurrin, “Content-based video retrieval,” in Encyclopedia of Database Systems. Boston, MA, USA: Springer, 2009, pp. 466–473, doi: 10.1007/978-0-387-39940-9_1027.

[5] S. Ding, S. Qu, Y. Xi, and S. Wan, “A long video caption generation algorithm for big video data retrieval,” Future Gener. Comput. Syst., vol. 93, pp. 583–595, Apr. 2019.

[6] S. Fujita, T. Hirao, H. Kamigaito, M. Okamura, and M. Nagata, “SODA: Story oriented dense video captioning evaluation framework,” in Proc. 16th Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K. Berlin, Germany: Springer-Verlag, 2020, pp. 517–531, doi: 10.1007/978-3-030-58589-6_31.

[7] A. Ramani, A. Rao, V. Vidya, and V. B. Prasad, “Automatic subtitle generation for videos,” in Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Mar. 2020, pp. 132–135.

[8] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1112–1131, May 2020.

[9] L. Gao, X. Wang, J. Song, and Y. Liu, “Fused GRU with semantic-temporal attention for video captioning,” Neurocomputing, vol. 395, pp. 222–228, Jun. 2020.

[10] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, “Hierarchical recurrent neural encoder for video representation with application to captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1029–1038.

[11] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” in Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics. Hum. Lang. Technol., 2015, pp. 1494–1504.

[12] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence—Video to text,” in Proc. IEEE Int. Conf. Vis. (ICCV), Dec. 2015, pp. 4534–4542.

[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Describing videos by exploiting temporal structure,” in Proc. Int. Conf. Comput. Vis., Dec. 2015, pp. 4507–4515.

[14] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, “STAT: Spatial–temporal attention mechanism for video captioning,” IEEE Trans. Multimedia, vol. 22, no. 1, pp. 229–241, Feb. 2019.

[15] X. Shi, J. Cai, J. Gu, and S. Joty, “Video captioning with boundary-aware hierarchical language decoding and joint video prediction,” Neurocomputing, vol. 417, pp. 347–356, Dec. 2020.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. 31st Int. Conf. Mach. Learn. (ICML), 2017, pp. 6000–6010.

[17] A. Kojima, T. Tamura, and K. Fukushima, “Natural language description of human activities from video images based on concept hierarchy of actions,” Int. J. Comput. Vis., vol. 50, no. 2, pp. 171–184, Nov. 2002.

[18] S. Mukherjee, S. Ghosh, S. Ghosh, P. Kumar, and P. P. Roy, “Predicting video-frames using encoder-ConvLSTM combination,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 2027–2031.

[19] S. Liu, Z. Ren, and J. Yuan, “SibNet: Sibling convolutional encoder for video captioning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 3259–3272, Mar. 2020.

[20] M. Rohrbach, W. Qu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, “Translating video content to natural language descriptions,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 4349–4356.

[21] L. Zhu, Z. Xu, and Y. Yang, “Bidirectional multistage reconstruction for temporal modeling in videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1339–1348.

[22] Z. Yang, Y. Han, and Z. Wang, “Capturing the temporal regions-of-interest for video captioning,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 146–153.

[23] S. Chen, J. Jin, J. Chen, and A. G. Hauptmann, “Generating video descriptions with latent topic guidance,” IEEE Trans. Multimedia, vol. 21, no. 9, pp. 2407–2418, Sep. 2019.

[24] L. Baraldi, C. Grana, and R. Cucchiara, “Hierarchical boundary-aware neural encoder for video captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3185–3194.

[25] M. Nabi and A. Behrad, “Video captioning using boosted and parallel long short-term memory networks,” Comput. Vis. Image Understand., vol. 190, Jan. 2020, Art. no. 102840.

[26] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, “Object relational graph with teacher-recommended learning for video captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13275–13285.

[27] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, “Joint syntax representation learning and visual cue translation for video captioning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8917–8926.
L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense
J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, “From deter-
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet:
A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT
texts,” in Proc. 10th Asian Conf. Mach. Learn. (ACML), 2017, pp. 1273–1283.
R. Shetty and J. Laaksonen, “Frame- and segment-level features and
captioning with POS sequence guidance based on gated fusion
R. Pasunuru and M. Bansal, “Multi-task video captioning with video
M. Chen, Y. Li, Z. Zhang, and S. Huang, “TVT: Two-view transformer
W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, “Memory-
respective. She is currently pursuing the Ph.D. degree with APJ Abdul Kalam Technological
University, Calicut, Kerala, in 2000, and the Ph.D. degree in signal processing from the Indian
Institute of Science, Bangalore, in 1994, the M.E. degree in instrumentation engineering from the
College of Engineering Thiruvananthapuram, University of Kerala, India, in 2001 and 2005,
respectively. She joined as a Faculty Member at the Department of Technical Education, Government
Engineering College, in 2008. Currently, she is an Assistant Professor with the Department of Electronics and Communication Engineering, Government Engineering College, Barton Hill, Thiruvananthapuram, Kerala, India. Her research interests include signal processing, image and video analytics, machine learning, computer vision, and related areas.

RESHEM S. BHOOSSHAN received the B.Tech. and M.Tech. degrees in applied electronics and instrumentation engineering from the College of Engineering Thiruvananthapuram, University of Kerala, India, in 2001 and 2005, respectively. She is currently pursuing the Ph.D. degree with APJ Abdul Kalam Technological University, Kerala, India. She joined as a Faculty Member at the Department of Technical Education, Government Engineering College, in 2008. Currently, she is an Assistant Professor with the Department of Electronics and Communication Engineering, Government Engineering College, Barton Hill, Thiruvananthapuram, Kerala, India. Her research interests include signal processing, image and video analytics, machine learning, computer vision, and related areas.

SURESH K. (Senior Member, IEEE) received the B.Tech. degree in electronics and communication from the College of Engineering Thiruvananthapuram, University of Kerala, in 1994, the M.E. degree in signal processing from the Indian Institute of Science, Bangalore, in 2000, and the Ph.D. degree in signal processing, in 2010. He joined as a Faculty Member at the Department of Technical Education, Government Engineering College, in 1995. Currently, he is a Professor with Government Engineering College, Wayanad, Kerala, India. His research interests include signal processing, audio signal processing, image/video analytics, machine learning, and related areas.