IEEE Signal Processing Society | IEEE Transactions on Image Processing

AAP-MIT: Attentive Atrous Pyramid Network and Memory Incorporated Transformer for Multi-Sentence Video Description

Journal:	*Transactions on Image Processing*
Manuscript ID	TIP-26109-2021.R2
Manuscript Type:	Regular Paper
Date Submitted by the Author:	27-Mar-2022
Complete List of Authors:	Prudviraj, Jeripothula; Indian Institute of Technology Hyderabad, computer science and engineering Indrakaran Reddy, Malipatel; Indian Institute of Technology Hyderabad, Computer Science and Engineering Vishnu, C ; Indian Institute of Technology Hyderabad, Computer Science and Engineering Chalavadi, Krishna; IIT Hyderabad, Computer Science and Engg
Subject Category Please select at least one subject category that best reflects the scope of your manuscript:	Image and Video Analysis, Synthesis and Retrieval, Image & Video Sensing, Modeling, and Representation
EDICS:	33. ARS-IIU Image and Video Interpretation and Understanding < Image and Video Analysis, Synthesis and Retrieval, 4. SMR-Rep Image and Video Representation < Image & Video Sensing, Modeling, and Representation, 13. TEC-MLI Machine Learning for Image Processing < Image & Video Processing Techniques, 18. COM-MMC Image and Video Multimedia Communications < Image & Video Communications

Reviewer #1

Comment: The idea of temporal atrous convolution is not new, which can be regarded as a variant of 3D convolution; so using both atrous convolution and optical flow seems redundant

Response: As mentioned in Section 4.2, similar to [14, 32], we first pre-process each video at 2FPS and then extract optical flow features using "global pool layer" of BNInception [10] which is further pre-trained on ActivityNet [3][27]. Usually, these features learn the pattern of apparent motion of various objects between neighbourhood frames. On the other hand, the temporal atrous convolutions convolve such features temporally based on given temporal window. For instance, if we have optical flow features of1, of2, of3, …, of9, the temporal atrous convolution performs temporal convolution among of1, of5, and of9 when filter is 3*1 and atorus rate is 4. From this, we can infer that the optical flow features help to learn relative motion of objects, whereas, the temporal atorus convolutions facilitate the temporal dynamics of the visual scene. In our work, we assume the usage of both atrous convolution and optical flow are not redundant. Moreover, we believe that the atorus convolutions on optical flow features learn differentiations among subtle interactions like walking, jogging, and running, e.t.c.,

Comment: The motivation of using temporal correlation attention on long-range and short-range temporal features is not clear. It would be nice to provide more details.

Response: In order to generate fine-grained captions, we need to learn both short-range and long-range temporal features. For example, to generate a caption "young man holds a violin", the short-range information may be sufficient. but, we need to learn long-range information to generate a caption "the man then begins playing the instrument while moving his hands up and down,"

Comment: The performance improvements seem marginal on ActivityNet Captions dataset (Table 1).

Response: yes, we achieved a marginal improvement on ActivityNet Captions dataset (Table 1). However, this trend can also be observed from the previous state-of-the-art approaches. And, it is hard to achieve a high performance gain in captioning task due to the complexity of the task (no two persons describe the visual content in the same word-to-word manner).

Reviewer #2

Comment: From decoder perspective, I feel MIT and MART are almost the same. It is not very clear for how the MIT improves the MART in decoder part. 1. Give more detailed explanation about difference between the MIT and decoder in MART.

Response: The fundamental difference between MART [14] and the proposed MIT lies in the construction of memory block. MART[14] constructs a two slot memory block using MultiHeadAt (input), tanh (slot 1), and sigmoid (slot 2). On the other hand, we

update the memory block by combining multiple slots effectively (Eq 3) to model complex relations at higher capacity.

Comment: Miss some explanations of the models in table 1. For example, MIT uses the same encoder as MART?

Response: MART uses the simple appearance and flow information as encoder for captioning. Our results in Table 1 with method name MIT demonstrate the results of the same appearance and flow information with MIT decoder.

Comment: In table 3, my observation is that the AAP-MIT is more sensitive to hyper-parameter changes

Response: Yes the performance of the proposed approach varies with the change of hyper parameters. However, the performance change is in between +1 or -1 (Table 3).
Reviewer #4

Comment: The differences between MART and the memory augmented transformer in the proposed MIT are minor.

Response: Yes, we only modify the structure of the memory block which is the core building block of MART.

Comment: The Figure 3 is not that clear since the blocks of different components share similar colors.

Response: Please accept our apologies. We will try to change background colors of different component in the final version of manuscript, if allowed.

Comment: Some important related works about dense video captioning are missing

Response: We apologize for missed references. We will try to include the suggested references in the final draft.

Comment: Multi-scale temporal features can be learnt by stacking several layers of transformer followed by a meshed-memory decoder as in [R1]. What are the differences between this design and the proposed AAP-MIT network?

Response: Yes, the stacked transformer encoder and meshed-memory decoder may achieve comparable results. The stack of transformer layers may resemble COOT architecture [7]. Usually, it can be possible to learn the local features using self-attention and global features using cross-modal attention. However, it may be difficult to model the temporal dependencies among Fi, Fi+r, Fi+2r, Fi+3r… e.t.c. This can be easily modelled using temporal dilated convolutions. For example, if the dilation rate is 6 and convolutional kernel is 3*1, we convolve input features Fi, Fi+6, Fi+12, temporally.
In comparison with meshed-memory decoder (M2T), the M2T constructs the memory-augmented attention by simply extending the self-attention with additional “slots” , However, we construct the multiple memory slots instead of single slot (Eq 3).

Comment: What are the differences between the AAP network and TPN [R2]?

Response: We can see TPN [R2] as a variant of our AAP. However, there are some architectural differences in between TPN and AAP. TPN utilizes the output features of res2, res3, res4, res5 of ResNet, where they are spatially and temporally downsampled to learn features. In contrast, we extract features from linear layer of pre-trained network and maintain the same feature dimension throughout vb0/ob0 to vb3/ob3. Then, we learn the temporal dynamics using set of atrous convolutions at various rates. If we have 10 features, the vb1/ob1 learns the temporal features as output = conv (Fi, Fi+2, Fi+4) and vb3/ob3 as output = conv (Fi, Fi+6, Fi+12). This type of notion is highly different from TPN.

View Reviews

Paper ID	2133
Paper Title	AAP-MIT: Attentive Atrous Pyramid Network and Memory Incorporated Transformer for Multi-Sentence Video Description
Track Name	Main Track

Reviewer #1

Questions

1. [Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences).
This paper proposes an Attentive Atrous Pyramid network and Memory Incorporated Transformer (AAP-MIT) for multi-sentence video description. A temporal pyramid network and a temporal correlation attention module are used to extract rich temporal features from video sequences, and a memory incorporated transformer is introduced to generate highly descriptive natural language sentences.

2. [Relevance] Is this paper relevant to an audience to ACM Multimedia? Please check https://2021.acmmm.org/call-for-paper.html.
Relevant to researchers in subareas only

3. [Significance] Are the results significant?
Significant

4. [Novelty] Are the problems or approaches or applications/systems novel?
Novel

5. [Evaluation] Is the idea proposed in this paper well supported by theoretical analysis or experimental results?
Sufficient

6. [Paper Strengths] Please discuss. Justifying your comments with the appropriate level of details about the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, applications, etc.). For instance, a theoretical paper may need no experiments, while a paper with a new approach or application may require comparisons to existing methods.
1. The idea of incorporating transformer with memory blocks is novel.
2. Extensive experiments are done on several benchmark datasets to verify the effectiveness of the proposed method.
3. The presentation of the paper is good, and it is easy to follow.

7. [Paper Weaknesses] Please discuss. Justifying your comments with the appropriate level of details about the weaknesses of the paper (i.e. lack of novelty – given references to prior work, lack of novelty, technical errors, or/and insufficient evaluation, etc.). Note: If you think there is an error in the paper, please explain why it is an error. It is not appropriate to ask for comparisons with unpublished papers and papers published after the ACM Multimedia deadline. In all cases, please be polite and constructive.

1. The idea of temporal atrous convolution is not new, which can be regarded as a variant of 3D convolution; so using both atrous convolution and optical flow seems redundant.

2. The motivation of using temporal correlation attention on long-range and short-range temporal features is not clear. It would be nice to provide more details.

3. The performance improvements seems marginal on ActivityNet Captions dataset (Table 1).

8. [Preliminary Rating] Please rate the paper according to one of the following choices.
Borderline Accept

10. [Confidence]
Confident

14. [Final Recommendation] Please provide your final recommendation according to the author rebuttal, the discussions with other reviewers and area chairs.
Poster

15. [Final Justification] Please provide your justification of your final recommendation.
The authors addressed my concerns in the rebuttal.

Reviewer #2

Questions

1. [Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences).
This paper improves previous memory augmented recurrent transformer (MART) based video description with Attentive Atrous Pyramid Network (AAP) and memory incorporated transformer (MIT). Based on the visual and optical flow features, AAP captures the spatiotemporal dynamics of the video data. Similar to MART, MIT is used to capture the long-range dependency among video segments and text. Experiments on ActivityNet Captions and YouCookII show the improvement of AAP-MIT over other methods.

2. [Relevance] Is this paper relevant to an audience to ACM Multimedia? Please check https://2021.acmmm.org/call-for-paper.html.
Likely to be of interest to a large proportion of the community

3. [Significance] Are the results significant?
Significant

4. [Novelty] Are the problems or approaches or applications/systems novel?
Somewhat novel or somewhat incremental

5. [Evaluation] Is the idea proposed in this paper well supported by theoretical analysis or experimental results?
Sufficient

6. [Paper Strengths] Please discuss. Justifying your comments with the appropriate level of details about the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, applications, etc.). For instance, a theoretical paper may need no experiments, while a paper with a new approach or application may require comparisons to existing methods.

1. The proposed AAP is well designed and motivated. Experiment results show the effectiveness of the AAP encoder.
2. This paper gives comprehensive experiments. The results show the substantial improvement of the proposed whole system.

7. [Paper Weaknesses] Please discuss. Justifying your comments with the appropriate level of details about the weaknesses of the paper (i.e. lack of novelty – given references to prior work, lack of novelty, technical errors, or/and insufficient evaluation, etc.). Note: If you think there is an error in the paper, please explain why it is an error. It is not appropriate to ask for comparisons with unpublished papers and papers published after the ACM Multimedia deadline. In all cases, please be polite and constructive.

1. From decoder perspective, I feel MIT and MART are almost the same. It is not very clear for how the MIT improves the MART in decoder part.
2. Miss some explanations of the models in table 1. For example, MIT uses the same encoder as MART?

8. [Preliminary Rating] Please rate the paper according to one of the following choices.

Poster

9. [Rebuttal Requests] Please pose questions you want to be answered in the rebuttal. Please do NOT ask the author(s) to include any new results (e.g., experiments and theorems) in the rebuttal.

1. Give more detailed explanation about difference between the MIT and decoder in MART.
2. In table 3, my observation is that the AAP-MIT is more sensitive to hyper-parameter changes.

10. [Confidence]

Confident

14. [Final Recommendation] Please provide your final recommendation according to the author rebuttal, the discussions with other reviewers and area chairs.

Borderline

15. [Final Justification] Please provide your justification of your final recommendation.

Based on the answers from the authors, I don't think there is significant difference between MIT and the decoder in MART. So I think it is somewhat novel for novelty.

Reviewer #4

Questions

1. [Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences).

The paper presents a novel framework to generate multi-sentence descriptions for video, which consists of a pyramid network to model multi-scale temporal contextual information followed by a temporal correlation attention to learn correlations among these multi-scale features as well as a memory incorporated transformer to learn long-range dependencies among video segments and corresponding descriptions. Experiments and analysis are conducted on ActivityNet Captions and YouCookII datasets to validate the superiority of the proposed model.

2. [Relevance] Is this paper relevant to an audience to ACM Multimedia? Please check https://2021.acmmm.org/call-for-paper.html.

Relevant to researchers in subareas only

3. [Significance] Are the results significant?

Highly significant

4. [Novelty] Are the problems or approaches or applications/systems novel?

Somewhat novel or somewhat incremental

5. [Evaluation] Is the idea proposed in this paper well supported by theoretical analysis or experimental results?

Sufficient

6. [Paper Strengths] Please discuss. Justifying your comments with the appropriate level of details about the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, applications, etc.). For instance, a theoretical paper may need no experiments, while a paper with a new approach or application may require comparisons to existing methods.

1) State-of-the-art results are achieved.

2) Extensive experiments are included to demonstrate the effect of different components in the proposed AAP-MIT with various hyperparameters.

3) The paper is well organized and easy to follow.

7. [Paper Weaknesses] Please discuss. Justifying your comments with the appropriate level of details about the weaknesses of the paper (i.e. lack of novelty – given references to prior work, lack of novelty, technical errors, or/and insufficient evaluation, etc.). Note: If you think there is an error in the paper, please explain why it is an error. It is not appropriate to ask for comparisons with unpublished papers and papers published after the ACM Multimedia deadline. In all cases, please be polite and constructive.

1) The differences between MART and the memory augmented transformer in the proposed MIT are minor.

2) The Figure 3 is not that clear since the blocks of different components share similar colors.

3) Some important related works about dense video captioning are missing:

- Streamlined dense video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- Jointly localizing and describing events for dense video captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

8. [Preliminary Rating] Please rate the paper according to one of the following choices.

Borderline Accept

9. [Rebuttal Requests] Please pose questions you want to be answered in the rebuttal. Please do NOT ask the author(s) to include any new results (e.g., experiments and theorems) in the rebuttal.

1) Multi-scale temporal features can be learnt by stacking several layers of transformer followed by a meshed-memory decoder as in [R1]. What are the differences between this design and the proposed AAP-MIT network?

2) What are the differences between the AAP network and TPN [R2]?

[R1] M2: Meshed-Memory Transformer for Image Captioning.

[R2] Temporal Pyramid Network for Action Recognition.

10. [Confidence]

Confident

14. [Final Recommendation] Please provide your final recommendation according to the author rebuttal, the discussions with other reviewers and area chairs.

Weak Reject

15. [Final Justification] Please provide your justification of your final recommendation.

After reading the authors' feedback and other reviewers' comments, I vote for "weak reject" for this paper since my main concern about the limited technical contribution has not been well addressed. From my view, this work is a simple combination of existing techniques (e.g., MART and TPN). Moreover, the reason why this architecture can outperform stacks of transformer layers in learning temporal dynamics has not been well clarified by the authors in the feedback.

Response to the Reviewers Comments

We thank the reviewers for their time and effort in providing the constructive and insightful comments. We have carefully considered the reviewer's comments and incorporated suggested changes to improve the quality of the manuscript.

We have highlighted the changes in the manuscript (track changes) in Blue color. Below table provides point-by-point response to the reviewers comments and suggestions.

Reviewer 1	
Comment	***Response***
1) The implementation of the proposed network is unclear. The authors should details these contents. Is the training of the proposed framework end-to-end?	Yes, our AAP-MIT captioning network is an end-to-end trainable network. However, the input appearance features are extracted from the 'Flatten-673' layer of ResNet-200 and optical flow features are extracted from the 'global pool' layer of BNInception. Both these networks are pre-trained on ActivityNet for action recognition task. As suggested by the reviewer, we have revised the implementation details. **[highlighted in manuscript].**
2) More example of negative results should be reported. Also, the disadvantage of the proposed method should be analyzed.	Now, we have included more negative examples in Figure 6 and analyzed disadvantages of the proposed method in the revised manuscript. **[highlighted in manuscript].**
3) How did the memory bank affect the performance? The ablation study about the memory bank should be conducted.	As per reviewer's suggestion, we have provided ablation study with respect to the memory bank (Table V and Section

	IV.C.2.b). **[highlighted in manuscript].**
4) Some works also explore the attention mechanism in video analysis, such as Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2021 Host-Parasite: Graph LSTM-In-LSTM for Group Activity Recognition. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 32(2): 663-674, 2021 Coherence Constrained Graph LSTM for Group Activity Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019 The authors should introduce these works in the revision.	We have revised our related work section by including the suggested research works. **[highlighted in manuscript].**

Reviewer 2	
Comment	***Response***
1) The related work is still relative short, and could cover more related topics, such as feature pyramid networks, and memory networks in other tasks.	As suggested by the reviewer, we revised the related work section and incorporated the research works suggested by the reviewer. . **[highlighted in manuscript].**
2) The temporal correlation attention is designed for encoding the long-range and short-range temporal features. However, the ablative results for the effect of long-range and short-range temporal features are not shown.	Now we have revised our ablation study with the results of long-range and short-range temporal features (Table IV and Section IV.C.2.a).. **[highlighted in manuscript].**

Response to the Reviewers Comments

We thank the reviewers for their time and effort in providing the constructive and insightful comments. We have carefully considered the reviewer's comments and incorporated suggested changes to improve the quality of the manuscript.

We have highlighted the changes in the manuscript (track changes) in Blue color. Below table provides point-by-point response to the reviewers comments and suggestions.

Reviewer 1	
Comment	***Response***
All issues have been addressed	We thank the reviewer for accepting our manuscript.

Reviewer 2	
Comment	***Response***
1) The justification for the differences with previous methods (explained in the attached rebuttal) should be integrated into the paper as well.	As suggested by the reviewer, we incorporated the differences with previous methods in the revised manuscript. **[highlighted in manuscript].**

AAP-MIT: Attentive Atrous Pyramid Network and Memory Incorporated Transformer for Multi-Sentence Video Description

Jeripothula Prudviraj*, Malipatel Indrakaran Reddy, Chalavadi Vishnu, and C. Krishna Mohan, *Senior Member, IEEE.*

***Abstract*—Generating multi-sentence descriptions for video is considered to be the most complex task in computer vision and natural language understanding due to the intricate nature of video-text data. With the recent advances in deep learning approaches, the multi-sentence video description has achieved an impressive progress. However, learning rich temporal context representation of visual sequences and modelling long-term dependencies of natural language descriptions is still a challenging problem. Towards this goal, we propose an Attentive Atrous Pyramid network and Memory Incorporated Transformer (AAP-MIT) for multi-sentence video description. The proposed AAP-MIT incorporates the effective representation of visual scene by distilling the most informative and discriminative spatio-temporal features of video data at multiple granularities and further generates the highly summarized descriptions. Profoundly, we construct AAP-MIT with three major components: i) a temporal pyramid network, which builds the temporal feature hierarchy at multiple scales by convolving the local features at temporal space, ii) a temporal correlation attention to learn the relations among various temporal video segments, and iii) the memory incorporated transformer, which augments the new memory block in language transformer to generate highly descriptive natural language sentences. Finally, the extensive experiments on ActivityNet Captions and YouCookII datasets demonstrate the substantial superiority of AAP-MIT over the existing approaches.**

***Index Terms*—Multi-sentence video description, dense video captioning, atrous pyramid network, temporal correlation attention, transformers.**

I. Introduction

In recent years, the amount of multimedia data such as videos is growing tremendously due to the extensive use of sensors. With this rapid growth of data, it is essential to understand the multimedia content and describe it in natural language to benefit data management, retrieval, and enhance the content search in streaming platforms. The task of multi-sentence video description (dense video captioning), which describes the content of video in a series of semantically meaningful sentences has received an enormous interest in the intersection of multimedia and computer vision due to its wide range of applications such as storytelling, human-robot interaction, support of disabled, and video indexing. In addition, the dense video captioning is widely adopted in various domains like surveillance video analysis, personal lifelog video creations, and query based action/ emotion recognition. Leaving aside the tremendous applications, the dense video captioning poses two major challenges: i) The obscure nature of multiple events in a video and their spatio-temporal structures make dense video captioning an arduous problem. ii) The generation of coherent, precise, and semantically meaningful descriptions is an extremely complex task.

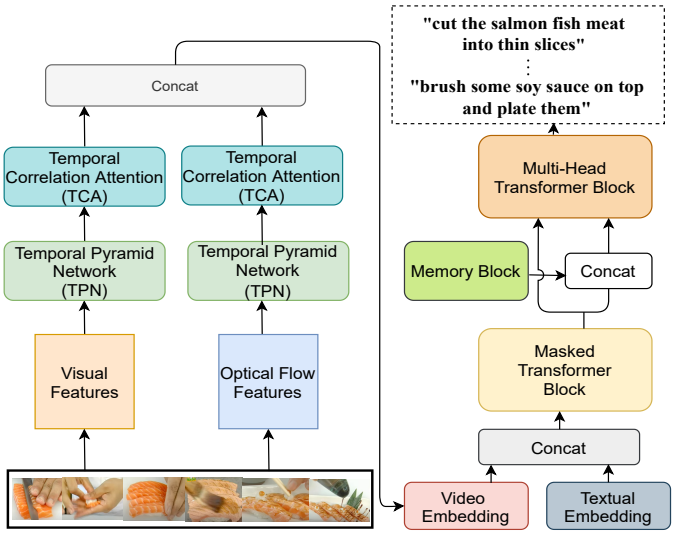


Fig. 1. Overview of the proposed AAP-MIT for Multi-Sentence Video Description

To address above challenges, many works [1], [2], [3], [4], [5], [6] have been proposed since early 2000s [7]. Inspired by machine translation tasks, most of the existing works follow encoder-decoder architecture, where the encoder network embeds the entire video into compact visual representation and the decoder network generates words and sentences in a sequence by utilizing the encoded video features. Due to the rapid development of deep learning approaches, recent works are utilizing convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers in encoder-decoder framework and reporting the state-of-the-art performance. Specifically, Krishna *et al.* [2] utilized C3D-LSTM based encoder-decoder architecture to extract visual information of various temporal event segments of an video and further generate coherent paragraph description. Whereas, Park *et al.* [8] exploited R3D-LSTM network with adversarial learning to generate list of descriptions for video. Further, the combination

*- Corresponding author
The authors are with the Visual Learning and Intelligence Group,, IIT Hyderabad, Hyderabad, India (e-mail: cs17resch01005@iith.ac.in), Page: https://sites.google.com/view/theswath/home

of graph convolutional networks and LSTMs are investigated in [9].

Recently, transformers [10] are showing superior performance over conventional RNNs like LSTM [11] and GRU [12]. Inspired by this, Zhou *et al.* [13] explored temporal convolutional network and masked transformer for dense video captioning. Further, Lei *et al.* [14] utilized two stream (visual and flow) video features with memory augmented transformer to more coherent and less repetitive sentences for multi-sentence video description.

Although the existing works are showing impressive performance over multi-sentence video description, they fail to address the following challenges: i) Some of the methods of dense video captioning task represent the video sequence as a collection of frame-level features by ignoring the fact that the videos contain more sophisticated information such as appearance & motion information, rich video semantics, fine-grained information, and temporal dynamics. ii) Since videos contain complex temporal cues, the content of video representation is often too difficult to be described. Hence, there is a necessity to capture the long-range spatio-temporal context information in order to build informative and discriminative feature hierarchy of video at multiple granularities. iii) Even though the attention mechanisms are widely explored in dense video captioning, learning temporal correlations among different action segments is still a challenging issue. iv) The problem of context fragmentation is associated with transformers, i.e., they operate on fixed-length segments without any additional information flow across the segments.

To address aforementioned issues of multi-sentence video description, we propose attentive atrous pyramid network and memory incorporated transformer (AAP-MIT). The overview of the proposed AAP-MIT is shown in Figure 1. Mainly, we construct the AAP-MIT with three major components, namely, temporal pyramid network (TPN), temporal correlation attention (TCA), and memory incorporated transformer (MIT). On extracting the visual and optical flow features from an input video, the temporal pyramid network employs parallel atrous convolutions over temporal space at multiple rates in order to build the temporal feature hierarchy and capture both short-range & long-range spatio-temporal contextual information of the video. Further, we incorporate a novel attention mechanism, i.e., temporal correlation attention to learn correlation among temporal feature hierarchy. Finally, the memory incorporated transformer accumulates the past history of sentences and video segments through an external memory block in order to generate more precise, diverse, and meaningful descriptions. The main contributions of the our work are summarized as follows:

- We introduce temporal pyramid network to effectively build the saptio-temporal feature hierarchy of an input video.
- A temporal correlation attention mechanism is proposed to align various temporal level features and learn the relations among various temporal cues.
- We design and build the new memory block in transformer architecture to capture long range dependencies over video segments and sentences.
- The efficacy of the proposed attentive atrous pyramid network and memory incorporated transformer (AAP-MIT) is verified quantitatively & qualitatively on two challenging datasets, ActivityNet and YouCookII.

II. Related work

In this section, we first review prominent works of multi-sentence video description and then the significant works related to feature pyramid networks, spatio-temporal coherence, and memory networks are presented as our multi-sentence video description work includes temporal pyramid network (TPN), temporal correlation attention (TCA), memory incorporated transformer (MIT) modules.

A. Multi-sentence video description

In the intersection of computer vision and natural language processing, the multi-sentence video description has received great interest in recent years. Generally, the multi-sentence video description approaches are divided into two categories, RNN-based approaches [15], [16], [2], [17], [18], [19], [20] and transformer-based approaches [13], [14], [21]. The RNN-based approaches leverage either long short term memory networks (LSTMs) or gated recurrent units (GRUs) to generate multi-sentence video descriptions. Whereas, the transformer based approaches utilize the variants of vanilla transformer architecture [10]. Recently, transformers are exhibiting superior performance over conventional RNN based methods on sequence learning task due to the inherent self-attention mechanism.

For the task of dense captioning, the concept of transformer is first explored in [13], where they simply replace LSTM decoder with transformer architecture. Another line of work, Lei *et al.* [14] presented a systematic study of transformer for multi-sentence video description. Mainly, they investigated vanilla transformer [10], transformer-XL [22], and memory augmented transformer to generate coherent paragraph descriptions. Further, Ging *et al* [21] explored cooperative hierarchical transformer (COOT) with MART to generate list of semantic descriptions. Similar to [14], [21], we propose transformer based approach, AAP-MIT for multi-sentence video description, where the attentive atrous pyramid network (AAP) learns the compact spatio-temporal representation of video and the memory incorporated transformer generates highly summarized descriptions with the help of augmented memory block.

B. Feature pyramid networks

Learning short-range and long-range temporal action instance is one of the key factors of action understanding. For example, it is hard to tell an action instance belongs to walking, jogging or running based on its visual appearance alone. However, it is more challenging to capture the subtle visual tempos of action due to their inter-class and intra-class variance across different videos. Recently, many works [23], [24], [25] explored this issue at input-level by leveraging the feature pyramid networks. Huang *et al.* [25] introduced an

attentive temporal pyramid network (ATP-Net) for dynamic scene classification by extracting and accumulating the most discriminative and informative features to construct an efficient representations of dynamic scenes. To incorporate multi-scale modelling for activity detection, Zhang *et al.* [24] proposed dynamic temporal pyramid network (DTPN). The DTPN exploits the temporal context of activities by fusing multi-scale feature maps to learn both local and global temporal contexts. A relation-aware pyramid network is presented in [26], for accurate temporal action proposals that exploits bi-directional long-range relations between local features to distill contextual features. Parsa *et al.* explored spatio-temporal pyramid graph convolutions to learn human actions and associated interactions with objects for human action recognition and postural assessment.

Recently, Dai *et al.* [27] introduced pyramid dilated attention network (PDAN) to model short-term and long-term temporal relations simultaneously by drawing local segments and high temporal receptive fields for action detection task. Yang *et al.* [23] proposed a generic temporal pyramid network (TPN) at the feature-level to model dynamics and temporal scale of an action instance for action recognition. Furthermore, temporal pyramid recurrent neural network is proposed in [28], to learn long-term and multi-scale dependencies in sequential data for various tasks such as masked addition problem, pixel-by- pixel image classification, signal recognition, and speaker identification.

C. spatio-temporal coherence

Recognizing spatial-temporal coherence is considered as one of the prominent characteristics of video understanding plays a vital role in video content analysis. To this end, Shu *et al.* [29] proposed a novel skeleton-joint co-attention recurrent neural networks (SC-RNN) to dynamically learn skeleton-joint co-attention feature map in spatio-temporal space and refine the observed motion information. Tang *et al.* [30] presented a coherence constrained graph LSTM with spatio-temporal context coherence (STCC) and a global context coherence (GCC) for group activity recognition where it captures relevant motions of whole activity while suppressing the some irrelevant motions. Recently, a novel graph LSTM-in-LSTM is introduced in [31] for activity recognition to exploit relationship between the group-level activity and person-level actions at spatio-temporal space.

D. Memory networks

Memory networks are referred to the neural networks which contains an external memory block where the information can be written and read by purposes. In recent years, many works explored memory networks in the domain of computer vision. For instance, Oh *et al.* [32] introduced space-time memory networks for video object segmentation where the memory network stores the current and past frames with object masks along with all the space-time pixel locations. For the task of visual tracking, Yang *et al.* [33] proposed a dynamic memory network in order to adapt the template with the target's appearance variations during tracking.

Memory networks also explored for the task of captioning. By updating existing memory networks, Park *et al.* [34] introduced context sequence memory network towards personalized image captioning. Recently, Cornia *et al.* [35] exploited memory blocks with transformer based architecture for image captioning task to learn a multi-level representation of the visual relationships. Pei *et al.* [36] introduced memory-attended recurrent network (MARN) for video captioning, where the memory structure explores the correspondence between a word and its various similar visual contexts of full-spectrum across videos in training data. To unify textual memory, visual memory and an attribute memory in a hierarchical way, a novel hierarchical memory model is proposed in [37], for video captioning.

III. Method

In this work, we introduce a novel encoder-decoder framework for multi-sentence video description. Given an input video I, with multiple temporally ordered video segments $\{I_1, I_2, \ldots, I_T\}$, our task is to generate sequence of natural language captions $\{s_1, s_2, \ldots, s_T\}$ to describe the content of the video. Here, s_t describes the content of video segment I_t. In the following sub-sections, we first revisit the transformer based dense video captioning approach. Then, we present our attentive atrous pyramid network and memory incorporated transformer (AAP-MIT) approach.

A. Recap of vanilla transformer for dense video captioning

Zhou *et al* [13] first explored the concept of transformer for dense video captioning, which originally introduced in [10] for machine translation task. The vanilla transformer based video paragraph captioning model is shown in Figure 2. The core building block of this vanilla transformer architecture is *scaled dot-product attention*. Given a query matrix (Q), key matrix (K), and value matrix (V), the attention output is computed as

$$A(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}, dim = 1\right)V, \quad (1)$$

where $softmax(:, dim = 1)$ represents that the $softmax$ is performed at the second dimension of the input. Further, we can obtain the *multi-head attention* [10] by combining such h parallel *scaled dot-product attention* blocks. This *multi-head attention* module ($MultiHeadAtt(Q, K, V)$) is quite generic and can be adoptable for several tasks. Specifically, it can be utilized as self/cross-modal attention [10] or memory aggregation [14]. The only difference between the self and cross-modal attention is that the query, key, and value matrix are same for self-attention but the query matrix will be different from the key and value matrix in cross-modal attention. Other important building block of vanilla transformer architecture is feed-forward layer, where it takes the input from multi-head attention and processes through linear projections with ReLU activation. Typically, the vanilla transformer architecture uses such building blocks in both encoder and decoder layers to draw relationships among input features.

As in [10], the baseline dense video captioning framework [13] incorporates L encoder layers and L decoder layers. At

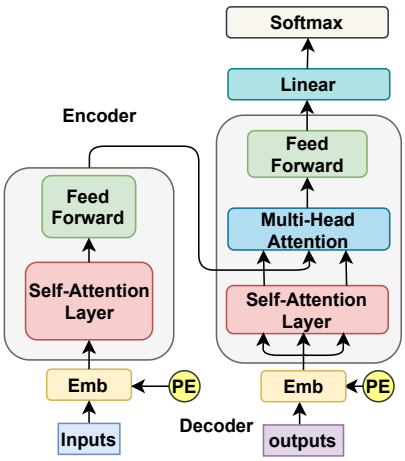


Fig. 2. Baseline transformer model for multi-sentence video description [13]. PE represents positional encoding

each l^{th} layer of encoder, the multi-head attention block takes hidden state from last encoder layer (H^{l-1}) as the input and performs self-attention [10]. Further, these attention outputs are inputted to feed-forward layers at each encoder block. At the l^{th} layer of the decoder, the model first employs the *masked multi-head attention* in order to restrict the model from seeing future words and then encodes the hidden state of last decoder layers. The multi-head attention uses masked outputs as query matrix and hidden states of the l^{th} encoder layer (H^{l}) as key and value matrices. Further a feed-forward layer is used to encode the sentences and accumulate the encoder information. We refer interested readers to [13] for more details.

B. Proposed AAP-MIT

The framework of the proposed AAP-MIT is shown in Figure 3, where we follow the shared encoder-decoder architecture as in [14] . As shown in Figure 3, we first extract visual and optical flow features from an input video, then we feed extracted features to temporal pyramid network (Section III-B1) to capture the effective representation of visual scene and model the temporal dynamics of video data. Further, we learn the temporal correlations of both visual and flow features using novel attention mechanism (Section III-B2). The memory incorporated transformer (Section III-B4) takes these attentive features and generates coherent descriptions of an input video.

1) Temporal pyramid network: In order to capture the most informative and discriminative spatio-temporal features, we should model the temporal dynamics of video information. Usually, there are two ways to explore the temporal cues of video data: i) By leveraging average pooling strategy over temporal space, which simply averages the all features along temporal dimension. This mechanism inevitably destroys the temporal sequential information and introduces unnecessary noise to the input features. The other way is to model the video features as temporal ordered sequences using LSTMs [11] or GRUs [12]. Despite the advantages of these models, the LSTMs and GRUs are complex networks and computational cost is extremely high when compared to CNN based networks. To this end, we propose a temporal pyramid network which exploits atrous convolutions, a special type of convolutional operation to model the temporal dynamics. Mainly, we employ several atrous convolutions on input features parallelly at multiple dilation rates to obtain a temporally convolved features. The construction of temporal pyramid network (TPN) is illustrated in Figure 3 (bottom-left). The TPN can incorporate effective feature representation from the neighbouring frame features and capture long and short temporal cues.

Given an input video, we first extract visual features and optical flow features from a pre-trained two stream network [38]. Thus, we obtain a set of visual features $v_f \in \mathbb{R}^{T\times d_v}$ and optical flow features $o_f \in \mathbb{R}^{T\times d_o}$. On obtaining visual and flow features, we employ temporal pyramid network on each individual feature set as in Figures 1 & 3 to achieve pyramid of spatio-temporal features. Since the construction of temporal pyramid hierarchy is same for both visual feature set and optical flow feature set, we present all the details with respect to visual stream as they can be extended as it is to the optical flow stream.

In our feature pyramid, we first incorporate the original visual features (v_f) as block-zero features ($vb_0 \in \mathbb{R}^{T\times d_{vb}}$, where $d_{vb} = d_v$) by accounting that v_f contains rich local features. Further, we devise $vb_1 \in \mathbb{R}^{T\times d_{vb1}}$, $vb_2 \in \mathbb{R}^{T\times d_{vb2}}$, and $vb_3 \in \mathbb{R}^{T\times d_{vb3}}$ by learning short-range and long-range temporal contextual information. Specifically, we learn both short-range and long-range temporal cues by probing multiple atrous convolutions at temporal dimension with various rates. Temporal atrous convolutions are special type of convolutions that increases the size of temporal receptive field with the increase of dilation rate.

Given a set of input features $\mathcal{E} = \{e_1, e_2, \ldots, e_T\}$, the output $\mathcal{E}^{(r)}$ of atrous convolution with dilation rate r and convolutional kernel W can be defined as

$$e_t^{(r)} = \sum_{i=1}^{w} e_{[t+r.i]} \times W_i^{(r)}, e_t^{(r)} \in \mathbb{R}^d,$$

where $W^{(r)}$ represents atrous convolution with dilation rate r and the $\mathcal{E}^{(r)}$ is the set of $e_t^{(r)}$'s output. For temporal atrous convolution, we employ $f \times 1$ filter to convolve input features at temporal dimension with dilation rate r. And, we learn the long-range temporal features with the increase of dilation rate or filter size. With this notion, we employ three atrous convolutions with dilation rates r_1, r_2, and r_3 in order to construct atrous visual block features vb_1, vb_2, and vb_3, respectively, where $r_1 < r_2 < r_3$. Intuitively, the vb_1 captures the neighbourhood information, vb_2 incorporates the short-range temporal features, and vb_3 acquires the long-range information as r determines the temporal stride with which we sample. For instance, consider the visual example provided in Figure 3, here, the vb_o may capture the representation of "salmon fish", vb_1 may learn the action of "cut", and vb_2 & vb_3 may incorporate the relation between "salmon" and "sushi".

In temporal pyramid network, we first employ the temporal correspondence across temporal dimension d using atrous convolutions, where it learns the rich temporal features. Further,

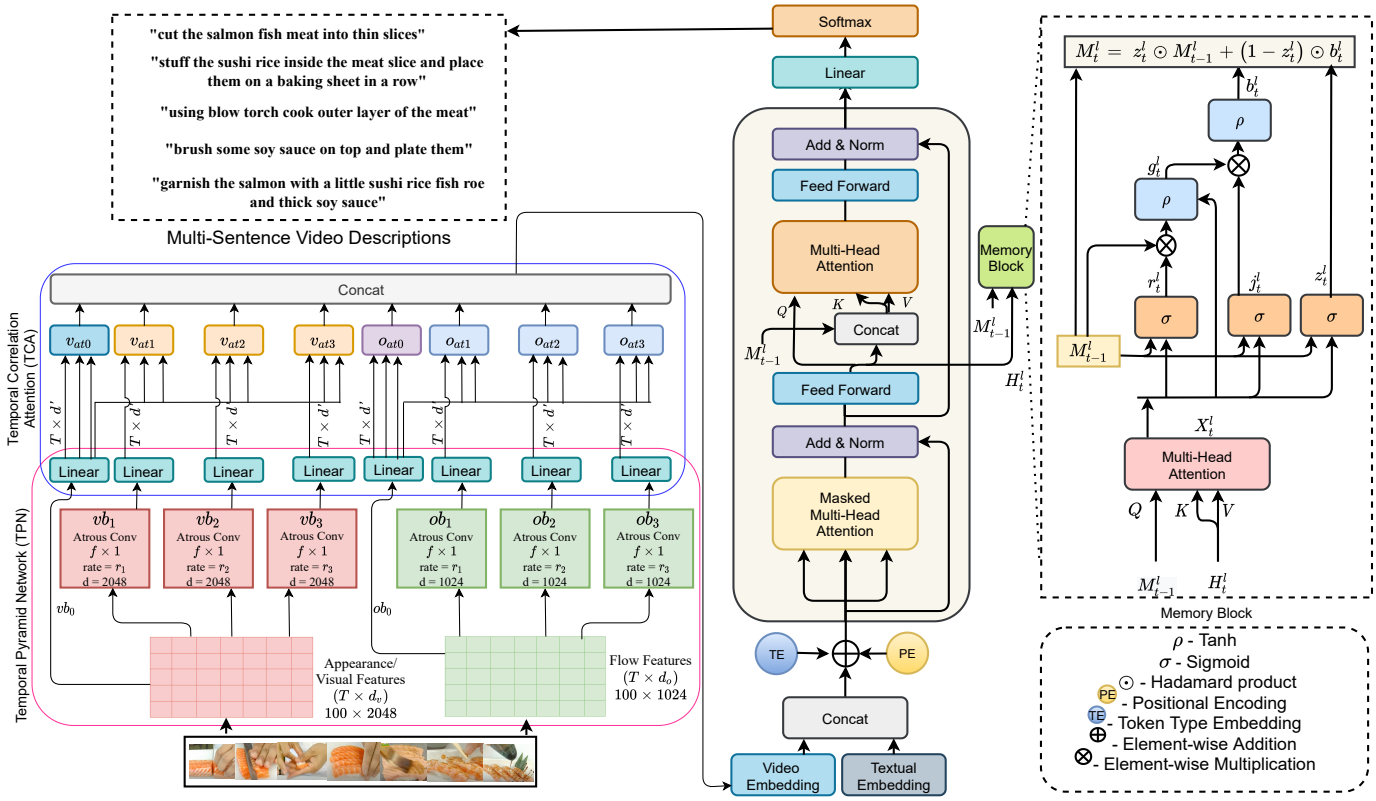


Fig. 3. Framework of the proposed AAP-MIT for multi-sentece video description task. Left: Attentive atrous pyramid which is incorporated with two major components, temporal pyramid network and temporal correlation attention. Right: Proposed memory incorporated transformer.

the attained features are projected onto linear layers to learn the compact spatio-temporal representation. The linear layers squish the output of each visual block feature $vb_j \in R^{T \times d_{vb}}$ to $vb_j \in R^{T \times d'_{vb}}$, where $d'_{vb} < d_{vb}$. The output features obtained from linear layers are now spatially compact and temporally aligned. Similarly, we employ temporal pyramid network on optical flow features in order to generate atrous optical flow block features, ob_0, ob_1, ob_2, and ob_3 . All visual and optical flow features, i.e., $\{vb_0, vb_1, vb_2, vb_3, ob_0, ob_1, ob_2, ob_3\}$ are further fed to the temporal correlation attention to learn inter-dependencies among the features.

2) Temporal correlation attention: Recently, many attention mechanisms [39], [40], [10], [41] have been proposed to generate the attention aware features and learn correlations among the features. In this work, we introduce correlation attention mechanism, similar to [10] in order to learn temporal correlated features. The proposed attention mechanism is

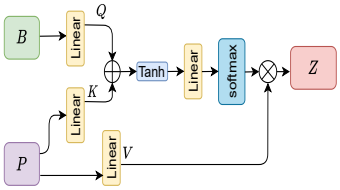


Fig. 4. Temporal correlation attention

shown in Figure 4.

Given two feature maps $B \in \mathbb{R}^{T \times d_b}$ and $P \in \mathbb{R}^{T \times d_p}$, we generate query matrix $Q \in \mathbb{R}^{T \times d_{at}}$ by projecting feature map B to linear layer. And, the key matrix $K \in \mathbb{R}^{T \times d_{at}}$ & value matrix $V \in \mathbb{R}^{T \times d_{at}}$ are constructed using feature map P, where d_{at} is the attention dimension. On extracting query, key, and value matrix, we first perform element wise sum and then feed it into to *Tanh*. The output of *Tanh* is projected to linear layers and applied $softmax$ to generate attention weights. Further, these attention weights are multiplied with value matrix to produce correlated attentive features. This attention mechanism can be presented mathematically as

$$A(Q, K, V) = [softmax(W_3 tanh(W_1 Q + W_2 K))] \odot [W_4 V], \quad (2)$$

where W_1, W_2, W_3, and W_4 are learnable parameters. This attention mechanism can also act as temporal self-attention when feature maps B and P are same. In other words, the query, key, and value matrix are generated from a single feature map.

In our work, we incorporate temporal self-attention on vb_0 & ob_0 and temporal correlation attention on rest of the output features of TPN. For temporal correlation attention, we generate query matrix from atrous features and key & value matrix from original features, i.e., vb_0 for visual stream and ob_0 for optical flow stream. For example, the query matrix for v_{at1}, v_{at2}, & v_{at3} will be vb_1, vb_2, & vb_3, respectively. And, the key matrix & value matrix will be vb_0 for all v_{at1}, v_{at2}, and v_{at3}. All attentive features, i.e., $\{v_{at0}, v_{at1}, v_{at2}, v_{at3}, o_{at0}, o_{at1}, o_{at2}, o_{at3}\}$ are further concatenated and represented as video embedding for our transformer based decoder.

3) Video-text embedding: On extracting attentive visual (v_{at0}, v_{at1}, v_{at2}, v_{at3}) and optical flow features (o_{at0}, o_{at1}, o_{at2}, o_{at3}) from temporal correlation attention (TCA), we concatenate them to construct video embedding $H_I^0 \in \mathbb{R}^{T_I \times d_I}$. Then, we produce textual embedding $H_s^0 \in \mathbb{R}^{T_s \times d_s}$ for input descriptions, where T_I and T_s denote the length of the video and text features, respectively. As shown in Figure 3, these two embeddings are further concatenated as input to the transformer layers, i.e., $H^0 = [H_I^0; H_s^0] \in \mathbb{R}^{T_c \times d}$, where $T_c = T_I + T_s$ and [;] denotes the concatenation. Finally, the combined video-text embedding is coupled with the trainable token type embedding (TE) to note whether the token is from video or text as in [14].

4) Memory incorporated transformer: In our work, we adopt the shared encoder-decoder transformer architecture as in [14], which is further incorporated with novel memory block to learn long-range dependencies among video segments and their corresponding descriptions. The overview of memory block is shown in Figure 3 (right). At l^{th} layer, while decoding the t^{th} video segment at time step t, we use multi-head attention to accumulate the information from both intermediate hidden states $H_t^l \in \mathbb{R}^{T_c \times d}$ and previous memory state $M_{t-1}^l \in \mathbb{R}^{T_m \times d}$ (T_m denotes the memory state length). As shown in Figure 3, we use hidden state information as query matrix ($Q = H_t^l$) and the concatenated memory and hidden state is used as key matrix & value matrix, i.e., $K, V = [M_{t-1}^l; H_t^l] \in \mathbb{R}^{(T_m+T_c) \times d}$. In memory block, the memory state M_{t-1}^l is updated to M_t^l using H_t^l and M_{t-1}^l as

$$
\begin{aligned}
X_t^l &= MultiHeadAtt(M_{t-1}^l, H_t^l, H_t^l), \\
r_t^l &= sigmoid(W_{mr}^l M_{t-1}^l + W_{xr}^l X_t^l + b_r^l), \\
j_t^l &= sigmoid(W_{mj}^l M_{t-1}^l + W_{xj}^l X_t^l + b_j^l), \\
z_t^l &= sigmoid(W_{mz}^l M_{t-1}^l + W_{xz}^l X_t^l + b_z^l), \\
g_t^l &= tanh(W_{mg}^l (r_t^l \odot M_{t-1}^l) + W_{xi}^l X_t^l + b_g^l), \\
b_t^l &= tanh(j_t^l \odot g_t^l), \\
M_t^l &= z_t^l \odot M_{t-1}^l + (1 - z_t^l) \odot b_t^l, \qquad (3)
\end{aligned}
$$

where $\odot$ represents the Hadamard product, W_{**}^l and b_*^l are trainable weight matrices and biases. The r_t^l, j_t^l, and g_t^l are internal cell states. And, z_t^l & b_t^l control the information of memory state. This updating strategy of our memory block is similar to LSTM [11] and GRU [12]. Mainly, our memory block supports multiple memory slots instead of single slot as in LSTM or GRU. This type of memory construction may model complex relations at higher capacity by memorizing the history of previous video segments and generated captions.

In this section, we presented AAP-MIT for multi sentence video description, which combines all three components i.e., temporal pyramid network, temporal correlation attention, and memory incorporated transformer into a unified network in order to generate highly summarized descriptions. In the next section, we present the experimental analysis of our AAP-MIT by quantitatively and qualitatively.

IV. Experiments

This section presents the data pre-processing, implementation details, and experimental results of the proposed approach.

A. Datasets and evaluation Metrics

a) Datasets:: In this work, we verify the effectiveness of the proposed approach on two challenging multi-sentence video description datasets, ActivityNet Captions [2] and YouCookII [42]. Both these datasets are largest datasets of video description task, and contain multiple temporal event segments with corresponding descriptions.

ActivityNet Captions provides $10,009$ videos in train set and $4,917$ videos in validation set. Each video in train set contains a single reference paragraph, while the validation set is provided with two reference paragraphs for each video. For fair comparison with the state-of-the-art models [14], [2], [21], we use widely accepted split provided in [20]. Particularly, the validation set is split into ae-validation and ae-test, where, ae-validation includes 2,460 videos and ae-test contains 2,457 videos. This type of setup makes that the videos of test set will not be seen in validation set.

YouCookII contains 1,333 videos in train set and 457 videos in validation set. These videos are collected from YouTube and cover 89 varieties of recipes. And, each video in YouCookII has single reference paragraph.

b) Evaluation Metrics:: In this work, we evaluate the generated descriptions by following the same evaluation process as in [14], [8], [6]. Specifically, we report results of the proposed approach using standard evaluation metrics like BLEU-4 [43], METEOR [44], ROUGE-L [45], and CIDEr-D [46]. In brief, all these metrics evaluate the coherence between the $N-gram$ occurrences in reference paragraph and generated paragraph. In addition, we evaluate the redundancy among multi-sentence descriptions using $R@4$ as in [14], [8], where it measures the $N-gram$ repetition in the descriptions, here $N = 4$.

B. Data Preprocessing and implementation details

a) Data Preprocessing: As in [13], [14], we represent the videos using both appearance and optical flow features, which are extracted at 2 FPS. Particularly, the appearance features are extracted from *'Flatten-673'* layer of ResNet-200 [47] and the optical flow features are extracted from *'global pool'* layer of BNInception [48]. Both these networks are pre-trained on ActivityNet [49] for action recognition task, developed by [38]. Thus, we achieve 2048D feature vector for appearance features and 1024D feature vector for flow features. In our experiments, we drop the sequences which are longer than 100 for video and 20 for text as in [14]. And, the maximum number of video segments set to 6 for ActivityNet Captions and 12 for YouCookII. Further, the vocabulary is built based on the words that repeat at least 5 times for ActivityNet Captions and 3 for YouCookII. Thus, we achieve a vocabulary of size 3,544 for ActivityNet Captions and 992 for YouCookII.

b) Implementation details: The proposed AAP-MIT is implemented in PyTorch [50] framework. For each video, we use 100 features of appearance with 2048 dimension and flow with 1024 dimension to represent video information. Thus, we achieve visual features $v_f = 100 \times 2048$ and optical flow features $o_f = 100 \times 1024$. On extracting visual and flow features, we employ temporal pyramid network and temporal correlation attention on each individual stream, separately.

In temporal pyramid network (TPN), we set atrous convolutional kernel to 5×1, dilation rates to $2, 4, \&, 6$, and dimension of atrous convolutional feature map to 2048 for visual & 1024 for optical flow. Further, the linear layer dimension of TPN is set to 512.

In temporal correlation attention, we set attention dimension to 512 for visual/appearance features and 256 for optical flow features. Further, we concatenate attentive visual and flow features and obtain feature representation of 100×3072 for each video. Then, we employ a dropout with drop probability of 0.2 on concatenated attentive visual and optical flow features.

In transformer architecture, we set the hidden size to 768, the number of transformer layers to 2, and the attention heads to 12. We follow the fixed scheme as in [10] for positional encoding. For memory block, the length of recurrent memory state is set to 1.

We optimize the proposed model by leveraging the strategy followed in [51]. Further, we use Adam [52] optimizer with a learning rate of $1e^{-4}$ and weight decay of 0.01. We train the proposed model to at most 50 epochs on both the datasets with an early-stop using CIDEr-D. Similar to [14], we use greedy decoding instead of beam search in caption generation. Further, we train the proposed AAP-MIT framework end-to-end. In addition, we present more detailed analysis of the proposed approach with various parameters in section IV-C1.

C. Quantitative and qualitative results

In this paper, we introduce attentive atrous pyramid network and memory incorporated transformer (AAP-MIT) for multi-sentence video description. Specifically, we learn the effective representation of visual scene and model their relations using attentive atrous pyramid network, which is the combination of two modules, temporal pyramid network and temporal correlation attention. In addition, we introduce a novel memory block in baseline transformer architecture as in [14] to learn long-range dependencies among video segments and generated captions. Further, we evaluate the proposed model on two challenging datasets, ActivityNet Captions and YouCookII using standard evaluation metrics like BLEU_n [43], METEOR [44], ROUGE-L [45], CIDEr-D [46], and R@4 [14].

Table I presents the performance of the proposed AAP-MIT on ActivityNet Captions along with the state-of-the-art approaches. Particularly, we compare the performance of the proposed model with the LSTM based models and transformer based approaches. From Table I, we can observer that the proposed approach is outperforming state-of-the-art approaches in all metrics. Specifically, the AAP-MIT is outperforming all other approaches on BLEU_n, METEOR, RougeL, and R@4 but showing a performance deflation on CIDEr-D. However, our memory incorporated transformer is exhibiting superior performance over MART [14] on CIDEr-D when combined with COOT features [21].

Further, the Table II demonstrates the performance of the proposed AAP-MIT on YouCookII dataset with state-of-the-art transformer based approaches. From Table II, we can observe that the proposed model is exhibiting similar performance as in Table I. Mainly, the AAP-MIT is outperforming the all existing approaches on BLEU_n, METEOR, RougeL, & R@4. And, the memory incorporated transformer with COOT is reporting superior performance over all other methods on CIDEr-D. In this paper, we compare our approach with LSTM based models like HSE [53], MFT [6], AdvInf [8], & GVD [20] and transformer based architectures like MART [14] and COOT [21]. In particular, our proposed AAP-MIT is closely relevant to the recent transformer based framework MART [14]. Hence, we conduct a systematic analysis with MART in the next section along with the ablation study.

1) Comparison with MART: MART is a memory augmented transformer which memorizes the sentence history and video segments in order to generate multi-sentence descriptions. Similar to MART [14], we incorporate new memory block along with the effective video encoding mechanism. In specific, the MART [14] approach simply uses the concatenated appearance and optical flow feature by ignoring that the video representation contains rich temporal dynamics and holds concealed relationships. To deal with such characteristics, we propose an attentive atrous pyramid network which is a combination of temporal pyramid network (TPN) (Section III-B1) and temporal correlation attention (TCA) (Section III-B2). Mainly, the TPN learns the local, short-range, and long-range contextual information by probing atrous convolutions at multiple rates. And, the TCA learns the correlation among the output features of the TPN. Further, we augment a new memory block different from MART [14], which incorporates multiple memory slot in order to learn complex and long-range feature representations. From Table I and Table II, we can observe that the proposed AAP-MIT is outperforming MART [14] in all metrics. In addition, we can observe that the COOT with MIT, i.e., $COOT_{video+clip}$ + MIT (Ours) is giving better CIDEr-D score than $COOT_{video+clip}$ + MART [21].

2) Ablation study: We conducted several experiments before finalizing the best possible model. Particularly, the considerable hyperparameters of the model are convolution filter size, number of atrous convolutions used for the construction of TPN, dilation rate, atrous convolution dimension, intermediate linear layer dimension, attention size, video embedding, and text embedding dimension. In addition, we also investigated the transformer attention [10] instead of temporal correlation attention (TCA) and observed very low performance. Please refer Table III for more quantitative results with different hyperparameters on ActivityNet captions dataset. Moreover, we also observed that the effect of memory length and number of memory layers used is not promising. In summary, we achieved best possible results on ActivityNet captions and YouCookII when TPN+TCA is associated with MIT.

TABLE I
COMPARISON OF THE PROPOSED AAP-MIT WITH EXISTING LSTM AND TRANSFORMER BASED APPROACHES ON *ActivityNet Captions* USING ALL EVALUATION METRICS. HERE, AAP-VTRANSFORMER: ATTENTIVE ATROUS PYRAMID NETWORK WITH VANILLA TRANSFORMER, MIT: MEMORY INCORPORATED TRANSFORMER. *-REPRODUCED RESULTS

Method	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr-D	R@4 ↓	RougeL
LSTM based methods								
MFT [6]	-	-	-	10.29	14.73	19.12	17.71	-
HSE [53]	-	-	-	9.84	13.78	18.78	13.22	-
DVcap [2]	26.45	13.48	7.12	3.98	9.46	24.56	-	-
GPaS [9]	-	-	-	1.53	11.04	28.20	-	-
LSTM based methods with detection features								
GVD [20]	-	-	-	11.04	15.71	21.95	8.76	-
AdvInf [8]	-	-	-	10.04	16.60	20.97	5.76	-
Transformer based methods								
VTransformer* [14], [10]	44.51	25.42	15.03	9.21	15.43	21.30	7.49	26.98
Transformer-XL [14], [22]	-	-	-	10.25	14.91	21.71	8.79	-
Transformer-XLRG [14]	-	-	-	10.07	14.58	20.34	9.37	-
MART [14]	-	-	-	9.78	15.57	22.16	5.44	-
$COOT_{video+clip}$ + MART [21]	-	-	17.43	10.85	15.99	28.19	5.35	31.45
AAP-VTransformer (Ours)	48.40	28.70	17.53	10.70	16.26	25.12	7.18	30.42
MIT (Ours)	47.83	28.41	18.29	11.25	17.13	26.23	6.42	31.87
$COOT_{video+clip}$ + MIT (Ours)	-	-	-	11.34	16.41	**30.87**	6.99	31.83
AAP-MIT (Ours)	**49.89**	**29.78**	**18.94**	**12.44**	**17.55**	28.32	**5.29**	**32.91**

TABLE II
COMPARISON OF THE PROPOSED AAP-MIT WITH EXISTING TRANSFORMER BASED APPROACHES ON *YouCookII* DATASET USING ALL STANDARD METRICS. *-REPRODUCED RESULTS

Method	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr-D	R@4 ↓	RougeL
VTransformer* [14], [10]	41.64	22.60	12.15	6.76	15.85	29.17	7.83	29.60
Transformer-XL [14], [22]	-	-	-	6.56	14.76	26.35	6.30	
Transformer-XLRG [14]	-	-	-	6.63	14.74	25.93	6.03	
MART [14]	-	-	-	8.00	15.9	35.74	4.39	
$COOT_{video+clip}$ + MART [21]	-	-	15.75	9.44	18.17	46.06	6.30	34.32
AAP-VTransformer (ours)	42.34	24.33	14.46	8.62	16.97	35.12	3.42	32.79
MIT (ours)	41.89	23.99	15.37	9.26	17.22	39.77	3.74	34.40
$COOT_{video+clip}$ + MIT (Ours)	-	-	-	9.45	18.00	**49.18**	8.55	34.27
AAP-MIT (ours)	**43.98**	**25.73**	**16.76**	**9.82**	**18.23**	43.87	**3.50**	**37.32**

TABLE III
THE PROPOSED ATTENTIVE ATROUS PYRAMID AND MEMORY INCORPORATED TRANSFORMER WITH VARIOUS HYPERPARAMETERS ON ACTIVITYNET CAPTIONS, WHERE TA: TRANSFORMER ATTENTION [10], TCA: TEMPORAL CORRELATION ATTENTION (OURS), VT: VANILLA TRANSFORMER [10], MART: MART [14], MIT: MEMORY INCORPORATED TRANSFORMER (OURS), B@4: BLEU-4, RL: ROUGEL, MT: METEOR, CR: CIDER-D

Visual features	Optical flow features	# Atrous Covolutions (visual, flow)	Size of Conv filter	Dilation rates	Attention mechanism	Size of attentive features	Size of Video embedding	Decoder	B@4	RL	MT	Cr
✓	✓	3 + 3	5×1	2, 4, 6	TA	4×512, 4×256	768	VT	9.09	27.71	15.33	22.68
✓	✓	3 + 3	5×1	2, 4, 6	TA	4×512, 4×256	768	VT	10.70	30.42	16.26	25.12
✓	✓	3 + 3	5×1	2, 4, 6	TCA	4×512, 4×256	768	MART	11.57	31.78	16.72	25.14
✓	✓	6 + 6	5×1	2,4, 6, 8, 10, 12	TCA	7×512, 7×256	2×768	MART	11.42	31.76	16.99	25.40
✓	✓	6 + 6	5×1	4, 6, 8, 10, 12, 14	TCA	7×512, 7×256	2×768	MART	11.45	31.64	16.87	25.42
✓	✓	6 + 6	5×1	4, 6, 8, 10, 12, 14	TCA	7×512, 7×256	2×768	MIT	11.20	31.64	16.84	25.82
✓	✓	3 + 3	3×1	2, 4, 6	TCA	4×512, 4×256	768	MIT	12.01	32.31	16.93	26.12
✓	✓	3 + 3	3×1	4, 6, 8	TCA	4×512, 4×256	768	MIT	11.63	32.00	16.76	25.75
✓	✓	3 + 3	5×1	4, 6, 8	TCA	4×512, 4×256	768	MIT	11.62	31.95	16.83	26.04

TABLE IV
EFFECT OF LONG-RANGE AND SHORT-RANGE TEMPORAL FEATURES ON ACTIVITYNET CAPTIONS, WHERE vb_* AND ob_* ARE VISUAL BLOCKS AND OPTICAL FLOW BLOCKS, RESPECTIVELY.

	List of visual and optical flow features	B@4	RL	MT	Cr
Short-range (SR)	vb_0, ob_0	11.25	31.87	17.13	26.23
Short-range (SR)	vb_0, vb_1, ob_0, ob_1	11.92	32.21	17.28	27.11
Long-range (LR)	vb_2, vb_3, ob_2, ob_3	11.89	32.23	17.45	27.98
SR+LR	vb_0, vb_1, vb_2, vb_3, ob_0, ob_1, ob_2, ob_3	**12.44**	**32.91**	**17.55**	**28.32**

TABLE V
THE ABLATION STUDY OF THE MEMORY MODEL ON ACTIVITYNET CAPTIONS

	hidden layers	mem. len.	B@4	RL	MT	Cr
hidden layers	1	1	12.30	32.80	16.20	27.12
	5	1	**13.20**	**33.10**	16.35	27.54
mem. len. AAP+MIT	2	2	11.90	32.21	15.98	26.43
	2	5	11.75	32.02	15.75	26.21
AAP+MIT	2	1	12.44	32.91	**17.55**	**28.32**

a) Effect of short and long range temporal features: Incorporation of short-range and long-range temporal features are crucial in video understanding [23], [24], [25]. So, we provide an ablative study on effect of long-range and short-range temporal features in IV. From the Table, we can infer that the combined information of short-range and long-range features boosts the performance of the model than individual feature model. Specifically, the performance of the proposed AAP-MIT with short-range and long-range features boosts the performance of Rouge-L from 31.87 to 32.91 and Cider-D from 26.23 to 28.32.

b) Effect of memory model: In our AAP-MIT, the key parameters with respect to memory model are: number of hidden layers and memory length. Table V shows the memory model ablation analysis. From the Table, we can observe that the models with small memory length have better overall performance than the high memory length. From the experiments, we finalize the model with 2 hidden layers and memory length 1 as it is showing a considerable balance between performance and computation.

c) Model architecture comparison with admissible works: Stacking several layers of transformer [10] followed by a meshed-memory decoder [35] may learn multi-scale temporal features. This notion learns the local features using self-attention mechanism and global features using cross-modal attention. Further, the M^2T [35] network can be leveraged for decoder. Although this approach may achieve comparable results, it may be difficult to model the temporal dependencies among input features $F_i, F_{i+r}, F_{i+2r}, \dots F_{i+nr}$. Moreover, the temporal dependencies can be easily modelled using temporal dilated convolutions as in our AAP-MIT. For example, if the dilation rate is 6 and convolutional kernel is 3×1, our AAP-MIT convolves input features in the form of F_i, F_{i+6}, F_{i+12}, temporally to achieve efficient and robust multi-scale temporal features. In comparison with meshed-memory decoder (M^2T) [35], the M^2T constructs the memory augmented attention by simply extending the self-attention with additional "slots". However, our proposed AAP-MIT constructs the multiple memory slots instead of single slot (Equation 3) to learn more complex relations.

We can see the temporal pyramid network (TPN) [23] as a variant of our AAP network. However, there are lot of architectural differences in between TPN and our AAP. TPN utilizes the output features of res_2, res_3, res_4, and res_5 of ResNet, where they are spatially and temporally downsampled to learn features. In contrast, we extract features from linear layer of pre-trained network and maintain the same feature dimension throughout $vb_0/ob_0$to vb_3/ob_3. Then, we learn the temporal dynamics using set of atrous convolutions at various rates. For instance, if we have 10 features, the vb_1/ob_1 learns the temporal features as $output = conv(F_i, F_{i+2}, F_{i+4})$ and vb_3/ob_3 as $output = conv(F_i, F_{i+6}, F_{i+12})$. This type of notion is highly different from model construction of temporal pyramid network [23].

D. Qualitative analysis

The Figure 5 demonstrates the qualitative results of the example images, where we show one example video with generated and ground truth descriptions for ActivityNet captions (Fig 5 (a)) and YouCookII (5 (b)). From the Figure, we

Ours : a man is seen speaking to the camera while holding up a violin, the man then begins playing the instrument while moving his hands up and down, the man continues playing the instrument while looking back to the camera.

GT : A young man holds a violin, then he plays violin, The violinist press the strings with four fingers and move the bow, Then, the man stops a little bit and then continues playing, the young man stops playing and talks.

(a)

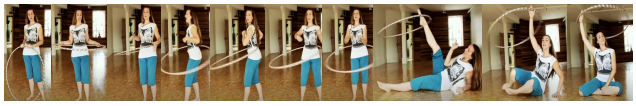
Ours : add chickpeas to a food processor, add salt lemon juice garlic paste salt and oil to the food processor, add salt pepper and lemon juice, grind the ingredients

GT : add chick peas to a food processor, squeeze the garlic paste into the food processor, add salt pepper and oil to a food processor, blend the ingredients in the food processor

(b)

Fig. 5. Qualitative examples on ActivityNet Captions (a) and YouCooKII (b).

can observe that the proposed AAP-MIT is able to generate summarized and semantically meaningful descriptions. Specifically, the generated words from Figure 5 (a) such as *"man"*, *"violin"*, *"instrument"*, *"speaking"*, *"holding"*, *"playing"*, *"begins"*, and *"up and down"* illustrate that the proposed model is able to learn the actions and relations effectively along with the context information. And, the generated words like *"add"*, *"chickpeas"*, *"food processor"*, and *"grind "* from Figure 5 (b) show that the model is able to understand the visual scene and learn the scene context.



Ours : a woman is seen standing in front of the camera holding a hula hoop and begins spinning it around, she continues spinning around the hula hoop while the camera captures her movements, she continues spinning the hula hoop around her body and ends by holding it up.

GT : A woman is standing in a room holding a hula hoop, She begins to use the hula hoop, She lays down on the ground and uses the hula hoop on her foot.

(a)

Ours : heat some oil in a pan and fry the tofu in it, place the patty on the pan , place the bread on the pan,flip the sandwich over, place the cheese on the bread, place the sandwich on the pan

GT : melt butter in a pan, place the bread in the pan, flip the slices of bread over, spread mustard on the bread, place cheese on the bread, place the bread on top of the bread,

(b)

Fig. 6. Failure cases on ActivityNet Captions (a) and YouCooKII (b).

The failure cases of the proposed AAP-MIT are shown in Figure 6. Even though our proposed model failed to match ground truth caption, it is able to detect prominent objects like *"hula hoop"* & *"bread"* and action like *"standing"* & *"flip"*. And, leveraging the previous and future segment information

for video encoding may lead to better understanding of action instances and subtle interactions, which can be explored as the part of future work.

V. Conclusion

In this paper, we present attentive atrous pyramid network and memory incorporated transformer (AAP-MIT) for multi-sentence video description task. The AAP-MIT has three major components: temporal pyramid network (TPN), temporal correlation attention (TCA), and memory incorporated transformer (MIT). The TPN and TCA encode the effective spatio-temporal representation of video sequences and learn the concealed relations of different video segments. And, the MIC memorizes the caption history and video segments information to generate highly descriptive sentences. In a nutshell, our approach provides an effective video-text representation for multi-sentence video description task by combining TPN, TCA, and MIC into a uniform network. The experimental results on two challenging datasets show that the proposed AAP-MIT has superior performance over existing multi-sentence video description approaches.

References

[1] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Video storytelling: Textual summaries for events," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 554–565, 2019.

[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[3] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning by attentive motion representation based hierarchical recurrent neural networks," in *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports*, 2018, pp. 77–85.

[4] J. Zhang and Y. Peng, "Video captioning with object-aware spatio-temporal correlation and aggregation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6209–6222, 2020.

[5] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 426–12 433.

[6] Y. Xiong, B. Dai, and D. Lin, "Move forward and tell: A progressive generator of video descriptions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 468–483.

[7] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[8] J. S. Park, M. Rohrbach, T. Darrell, and A. Rohrbach, "Adversarial inference for multi-sentence video description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6598–6608.

[9] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, "Dense video captioning using graph-based sentence summarization," *IEEE Transactions on Multimedia*, vol. 23, pp. 1799–1810, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[13] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[14] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, and M. Bansal, "Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[16] Z. Wang, Y. Luo, Y. Li, Z. Huang, and H. Yin, "Look deeper see richer: Depth-aware image paragraph captioning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 672–680.

[17] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan, and D. Cai, "Query-biased self-attentive network for query-focused video summarization," *IEEE Transactions on Image Processing*, vol. 29, pp. 5889–5899, 2020.

[18] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[19] Y. Yuan, L. Ma, J. Wang, and W. Zhu, "Controllable video captioning with an exemplar sentence," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1085–1093.

[20] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, "Grounded video description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6578–6587.

[21] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "Coot: Cooperative hierarchical transformer for video-text representation learning," *arXiv preprint arXiv:2011.00597*, 2020.

[22] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[23] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 591–600.

[24] D. Zhang, X. Dai, and Y.-F. Wang, "Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 712–728.

[25] Y. Huang, X. Cao, X. Zhen, and J. Han, "Attentive temporal pyramid network for dynamic scene classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8497–8504.

[26] J. Gao, Z. Shi, G. Wang, J. Li, Y. Yuan, S. Ge, and X. Zhou, "Accurate temporal action proposal generation with relation-aware pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 810–10 817.

[27] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond, "Pdan: Pyramid dilated attention network for action detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2970–2979.

[28] Q. Ma, Z. Lin, E. Chen, and G. Cottrell, "Temporal pyramid recurrent neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5061–5068.

[29] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[30] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph lstm for group activity recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[31] X. Shu, L. Zhang, Y. Sun, and J. Tang, "Host–parasite: graph lstm-in-lstm for group activity recognition," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 663–674, 2020.

[32] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Space-time memory networks for video object segmentation with user guidance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[33] T. Yang and A. B. Chan, "Visual tracking via dynamic memory networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 360–374, 2019.

[34] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 999–1012, 2018.

[35] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[36] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8347–8356.
[37] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "Hierarchical memory modelling for video captioning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 63–71.
[38] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang, "Cuhk & ethz & siat submission to activitynet challenge 2016," *arXiv preprint arXiv:1608.00797*, 2016.
[39] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.
[40] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *arXiv preprint arXiv:1406.6247*, 2014.
[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
[42] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
[44] S. Banerjee and A. Lavie, "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*, 2005, pp. 65–72.
[45] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[46] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
[49] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.
[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[53] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 374–390.

Jeripothula Prudviraj received the B.Tech degree from Sreenidhi Institute of Science and Technology, Hyderabad, India, in 2013, the M.Tech degree from MANIT Bhopal, India, in 2015. He is currently working toward the Ph.D. degree in the department of computer science, IIT Hyderabad. His research interests include video content analysis, deep learning, and computer vision.

Malipatel Indrakaran Reddy received a B.E degree in computer science from MVSR Engineering College, Hyderabad, India in 2017. He is currently pursuing an M.Tech degree in computer science at the Indian Institute of Technology, Hyderabad, India.His research interest lies in video captioning, visual question answering, video action recognition

Chalavadi Vishnu received a B.Tech degree in Computer Science Engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2016. Received an M.Tech degree from the Indian Institute of Technology Hyderabad (IITH), India in Computer Science Engineering, in 2018. He is currently pursuing Ph.D. at IIT Hyderabad in the Department of Computer Science Engineering. His Research interests include learning graph representations on video activities, deep learning for drones, and autonomous vehicles.

Chalavadi Krishna Mohan received the Bachelor of Science Education (B.Sc.Ed.) degree from Regional Institute of Education, Mysore, India, in 1988, the M.C.A. degree from S. J. College of Engineering, Mysore, India in 1991, the M.Tech. degree in system analysis and computer applications from National Institute of Technology Karnataka, Surathkal, India, in 2000, and the Ph.D. degree in computer science and engineering from IIT Madras, India, in 2007. He is currently a Professor with the Department of Computer Science and Engineering, and the Dean of Public and Corporate Relations at the Indian Institute of Technology Hyderabad (IIT Hyderabad), India. He is a senior member of IEEE, member of ACM, and life member of ISTE. His research interests include video content analysis, pattern recognition, and neural networks.