ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097128

CLIP4VIDEOCAP: RETHINKING CLIP FOR VIDEO CAPTIONING WITH MULTISCALE TEMPORAL FUSION AND COMMONSENSE KNOWLEDGE

Tanvir Mahmud, Feng Liang, Yaling Qing, Diana Marculescu

The University of Texas at Austin, Austin, TX, USA

ABSTRACT

In this paper, we propose CLIP4VideoCap for video captioning based on large-scale pre-trained CLIP image and text encoders together with multi-scale temporal reasoning and commonsense knowledge. In addition to the CLIP-image encoder operating on successive video frames, we introduce a knowledge distillation-based learning scheme that aims to exploit the CLIP-text encoder to generate rich textual knowledge from the image features. For improved temporal reasoning over the video, we propose a multi-scale temporal fusion scheme that accumulates temporal features from different temporal windows. In addition, we integrate various commonsense aspects in the caption generation which greatly enhances the caption quality by extracting the commonsense features from the video in the intermediate phase. Combining these strategies, we achieve state-of-the-art performance on the benchmark MSR-VTT dataset confirming that our framework significantly outperforms existing approaches.

Index Terms— CLIP, captioning, fusion, commonsense

1. INTRODUCTION

Understanding the human perception of the visual world by describing events with language is one of the widely explored research questions in computer vision [1], audio processing [2], and natural language processing [3]. Bridging the gap between visual world and its linguistic interpretation introduces interesting applications such as image and video captioning [4, 5], video retrieval [6], and video question answering [7]. Grounding multi-modal perception is one of the predominant challenges to achieve human-level performance in these applications. In this paper, we are primarily interested in video captioning for certain applications, such as describing movies for visually-impaired people [8] and automated human-robot interaction [9], which require complex temporal reasoning of visual-linguistic knowledge.

To ground the visual context and generate realistic captions, most existing approaches primarily utilize the encoder-decoder framework in both image and video captioning [4, 5, 10]. The encoder module grounds the visual cues and the decoder generates the caption from the grounded knowledge. Recently, CLIP (Contrastive Language-Image Pretraining) has demonstrated its superior performance on various visual-linguistic tasks relying on large-scale contrastive pre-training with image-text pairs [11]. Several approaches use the pre-trained CLIP encoders for captioning [5, 10]. Since generating captions from image or video initially operates on the visual modality and later on the grounded visual knowledge in the existing framework, incorporating both the pre-trained text encoder and image/video encoder for caption generation is a challenging task that requires simultaneous processing of visual-linguistic features. Most existing approaches primarily use a visual encoder, an approach that struggles to utilize the large-scale pre-training knowledge extracted through contrastive image-text pairs [5, 10].

Unlike image captioning that operates on static images [10], video captioning introduces additional temporal context with complex event transitions in sequential video frames [4]. Exploiting the temporal context is particularly important for video captioning, yet it is undoubtedly more challenging than image captioning. Several existing approaches attempted to address this challenge by integrating optical flow with visual features [12], using a simple transformer [5] or Long Short Term Memory module [4] on the sequence of image features. However, it is necessary to generalize the event dynamics across several events in the video, thereby requiring different scales of temporal reasoning for recognizing the events and perceiving the complex sequential event interactions.

To summarize the videos, humans usually extract several commonsense aspects for complex reasoning, such as the intended action of the subject, attributes of the ongoing events, and their subsequent implications on the subject and final objective. Rather than attempting to do caption generation, prior work has found such reasoning on commonsense knowledge to be beneficial [4]. However, the primary objective of earlier work is to analyze the performance of commonsense aspect generation given the video and/or captions, an approach that lacks a robust end-to-end framework to integrate diverse commonsense aspects in the generated captions.

To solve these challenges, in this paper we introduce Clip4VideoCap by properly utilizing large-scale pre-trained CLIP image and text encoders for video captioning and integrating multi-scale temporal reasoning with various commonsense aspects. In summary, our contributions are as follows:

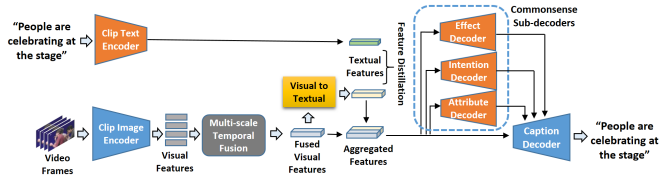

Fig. 1. The overview of our proposed CLIP4VideoCap model. For a given video, we first extract frame features with the pre-trained CLIP image encoder. Frame features are fed into the proposed multi-scale temporal fusion module. The visual-to-textual (*Vis2Text*) module mimics the CLIP text encoder through feature distillation and commonsense sub-decoders integrate diverse commonsense aspects to enrich caption.

1. We introduce a knowledge distillation-based rich textual feature learning scheme by using a pre-trained CLIP text encoder with image-encoder in an end-to-end training.
2. We propose a multi-scale temporal fusion scheme for improved temporal reasoning of event transitions.
3. We introduce diverse commonsense aspects from video into the generated captions for improved reasoning.
4. We achieve state-of-the-art performance through extensive experiments on the MSR-VTT dataset.

2. METHODOLOGY

2.1. CLIP4VideoCap: Overview

Clip4VideoCap exploits large scale pre-trained CLIP image and text encoders in an effective manner along with multi-scale temporal fusion and commonsense reasoning. Initially, the sequence of video frames is processed through the image encoder E_I to generate sequential visual feature representations from each frame. Since video frames are processed with the image encoder that doesn't operate on the temporal domain, it is necessary to use the sequential visual features to extract different scales of temporal interaction between events. We introduce a *multi-scale temporal fusion* module that relies on sequential temporal squeeze-fusion-expansion operations to gather the temporal context from different observation windows and process the sequential information. In parallel to the video feature processing within the CLIP-image encoder, we feed the CLIP-text encoder with the ground truth captions, thereby generating linguistic feature representations of the events. Since CLIP image and text encoders are pre-trained with large scale image-text pairs by contrastive learning, the representative visual-lingusitic features are supposed to have considerable feature correlation. Using these linguistic features, we introduce a knowledge distillation-based supervision on *Vis2Text* module to learn the mapping between the visual and linguistic representations of events. Moreover, various

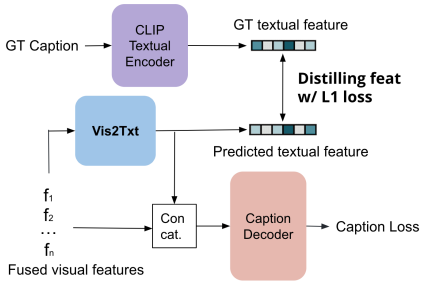

Fig. 2. The illustration of the proposed knowledge distillation based learning with CLIP text encoder.

commonsense aspects of the events are generated utilizing the sub-decoders before attempting the final caption decoding. Finally, the caption decoder generates enriched captions utilizing fused features and commonsense aspects. Both the primary decoder and the commonsense sub-decoders are supervised with ground truth annotations of captions and commonsense aspects. The whole framework is presented in Figure 1.

2.2. Introducing CLIP image/text encoders

Since video captioning is a multi-modal task, we propose to use state-of-the-art multi-modal pre-trained CLIP image and text encoders. CLIP is pre-trained on image-text pairs using a contrastive objective [11] which pulls the image-text features from the same pair together, while pushing unpaired features apart. Benefiting from large publicly available datasets (400M image-text pairs), the pre-trained CLIP models are helpful for many applications [10, 5, 13]. Hence, in the Clip4VideoCap framework, the sequence of video frames $(x_1, \ldots, x_n)$ is fed into the CLIP visual/image encoder (E_I) that generates sequence of visual features $(v_1, \ldots, v_n)$. Later, the sequential visual features are processed with multi-scale temporal fusion module $(\mathcal{F})$ to generate fused features $(f_1, \ldots, f_n)$, given by

$$(v_1, \ldots, v_n) = E_I(x_1, \ldots, x_n) \quad (1)$$

$$(f_1, \ldots, f_n) = \mathcal{F}(v_1, \ldots, v_n) \quad (2)$$

The video captioning is processed as a sequential transformation from the visual to textual representations which makes the proper utilization of both the CLIP image and text encoders complicated. Existing approaches primarily focused on the CLIP image encoder only for captioning [5], an approach which fails to exploit the complete contrastive knowledge of both CLIP encoders in an end-to-end fashion. In contrast to existing approaches, we feed the ground truth caption $C = (y_1, y_2, \ldots, y_n)$ to the CLIP textual encoder (E_T) in parallel to generate the grounded textual features T_{CLIP}. We introduce a visual-to-textual (*Vis2Text*) module to generate the representative textual features T_{vis2text} from the fused visual features (Figure 2). The *Vis2Text* : $\mathcal{V} \rightarrow T_{\text{Vis2Text}}$ module is

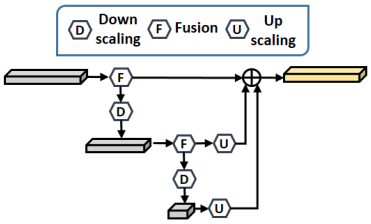

Fig. 3. Illustration of the multi-scale temporal fusion module.

designed such that,

$$vt_i = \sigma_2(F_2(\sigma_1(F_1(f_i)))); \ \forall i \in \{1, 2, \ldots, n\} \quad (3)$$

$$T_{\text{Vis2Text}} = \{vt_1, vt_2, \ldots, vt_n\} \quad (4)$$

where F_1, F_2 represent two fully connected layers and σ_1, σ_2 represent rectified non-linear activation functions, respectively.

The objective of the *Vis2Text* module is to learn the mapping from the visual features to their linguistic representation to generate rich textual features during inference without using the CLIP text encoder. To provide supervision to the *Vis2Text* module and mimic the feature representation of the CLIP textual encoder, we use L1 distillation loss between T_{CLIP} and T_{vis2text} given by:

$$T_{\text{CLIP}} = E_T(y_1, \ldots, y_n) \quad (5)$$

$$\mathcal{L}_{\text{distill}} = \text{minimize} \ ||T_{\text{CLIP}} - T_{\text{Vis2Text}}||_1 \quad (6)$$

2.3. Multi-scale Temporal Fusion

Since the baseline CLIP model is pre-trained with image-caption pairs, incorporating temporal fusion to adapt to the video context is particularly important for better performance. The proposed fusion module (Figure 3) directly operates on the extracted CLIP image features in order to integrate the temporal context into the caption decoder. As the video contains several sequential frames that include complex temporal interactions of the subject and its surroundings, it is necessary to generalize the temporal context from different temporal scales. We gradually downscale the temporal resolution of the sequential visual features to obtain high level temporal representations, perform temporal fusion on each scale to integrate temporal contexts, and perform upscaling of the squeezed representations for subsequent feature aggregation. To exploit the sequential nature of the video features, temporal feature processing modules, such as long short term memory (LSTM), 1D convolutional layers, and multi-headed transformer layers, can be very effective as fusion blocks on each scale. Finally, all the temporal features extracted from various scales are aggregated to generate the fusion feature vector $(f_1, \ldots, f_n)$.

2.4. Integrating Commonsense Knowledge in Captioning

In contrast to image, video events contain richer details, thereby requiring reasoning for improving caption quality. As mentioned in V2C [4], the commonsense knowledge can be described on three dimensions: intention, effect, and attribute. For example, the caption "*There is a man in black cutting the green leaves on the countertop*" have the following commonsense aspects: The action *intended* by the person is "*to cook something*", the *effect* of the action is "*getting clean dishes*", and the *attribute* of the person is "*hungry*".

Different than the V2C [4] scheme that primarily focuses on generating different aspects of commonsense knowledge separately one at a time given the video frames and captions, we introduce all three aspects of commonsense knowledge generation (intention, effect, and attribute) as an intermediate step of the final caption generation as illustrated in Fig 1. Separate sub-decoders are integrated such that they operate on the grounded visual features to generate feature representations for the three commonsense aspects. The purpose of commonsense sub-decoders is to generate different commonsense aspects of visual events, and thus, to enrich the final caption generated from the primary decoder. The objective caption generation loss $\mathcal{L}_{caption}$, and aggregated loss ($\mathcal{L}$) can be defined by,

$$\mathcal{L}_{caption} = \alpha_1 L_{cap} + \alpha_2 L_{int} + \alpha_3 L_{eff} + \alpha_4 L_{att} \quad (7)$$

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{caption} \quad (8)$$

where α denote the loss weight, L_{cap} denotes the output caption generation loss, and L_{int}, L_{eff}, L_{att} denote the intention, effect, and attribute generation loss, respectively.

Methods	CIDEr	METEOR	ROUGE-L
VNS-GRU [14]	52.0	29.5	63.3
MSAN [15]	52.4	29.5	-
topic-guided [16]	51.8	29.6	62.8
AVSSN [17]	46.9	28.8	61.7
SemSyn [18]	50.1	28.8	62.5
Uni-VL [19]	49.9	28.8	61.2
Clip4Caption [5]	57.7	30.7	63.7
Clip4VideoCap (Prop.)	**62.2**	**31.5**	**65.8**

Table 1. Performance comparison of the proposed method with other state-of-the-art approaches on MSR-VTT dataset.

Visual encoder	CIDEr	METEOR	ROUGE-L	BLEU-1
ResNet152	46.8	25.0	61.1	68.8
CLIP ViT-B32	60.1	27.4	**64.4**	73.2
CLIP ViT-B16	**60.4**	**27.8**	63.1	**73.9**

Table 2. Effect of the CLIP visual encoder: CLIP visual encoder performs significantly better than ResNet152.

3. RESULTS

3.1. Experimental Setup

Dataset

We use the most well-known video captioning dataset MSR-VTT [20] that contains around $9,721$ videos each, 10 to 30 seconds long. Video-to-Commonsense(V2C) [4] further complements MSR-VTT with event-level commonsense annota-

distillation loss ratio	CIDEr	METEOR	ROUGE-L	BLEU-1
0 (w/o CLIP txt)	59.9	28.8	63.7	73.3
1	61.5	30.1	64.9	74.5
5	61.9	31.0	65.2	75.0
10	62.2	**31.5**	**65.8**	**75.3**
20	**62.4**	31.2	65.5	75.1

Table 3. Effect of the knowledge distillation scheme with CLIP-text encoder and *Vis2Text* module.

Fusion Module	#Scales	CIDEr	METEOR	ROUGE-L	BLUE-1
LSTM	1	60.1	29.4	64.0	73.9
	3	**62.2**	**31.5**	**65.8**	**75.3**
Transformer	1	59.2	29.7	63.7	73.3
	3	61.6	30.8	64.9	74.4
Conv1D	1	58.4	26.7	63.6	73.1
	3	59.3	27.5	64.2	74.0

Table 4. Effect of the multi-scale hierarchical temporal fusion with ViT-B16 encoder.

tions, *i.e.*, event descriptions with intentions, effects and attributes, where each caption is supported by five commonsense annotations. The dataset is officially split with a training set of $6,819$ videos, and a test set of $2,903$ videos. We report performance as evaluated by automatic scores following the protocols from [4].

3.2. Comparison of the Proposed Method with State-of-the-art Approaches

We compare the performance of the proposed Clip4VideoCap with other state-of-the-art approaches on the MSR-VTT dataset. The results are summarized in Table 1. It is noticeable that the CLIP based approach significantly improves the captioning performance compared to other approaches with more than 10% improvements of the CIDEr score. The large-scale contrastive pretraining of the CLIP encoder greatly contributes such improvement. Moreover, our proposed approach considerably improves the performance over the contemporary CLIP based approach [5] with 4.5% improvement of CIDEr score, 2.1% improvement of the ROUGE-L score, and 0.8% improvement of the METEOR score.

4. ABLATION STUDIES

4.1. Effect of the CLIP visual encoder

We experiment with three visual encoders: ResNet-152, CLIP ViT-B32, and CLIP ViT-B16 (Table 2). CLIP visual encoders achieve 13.6% improvement of CIDEr score, 2.8% improvement of METEOR score, 3.3% improvement of ROUGE-L score, and 5.1% improvement of BLUE-1 score over ResNet thanks to pre-trained multi-modal representation.

4.2. Effect of knowledge distillation of CLIP-text encoder

We study the effect of the *Vis2Text* module on the final captioning performance with different distillation ratio (Table 3).

Sub-decoders	CIDEr	METEOR	ROUGE-L	BLUE-1
w/o	60.1	27.3	64.1	74.4
Intention	61.3	29.8	64.7	75.0
Effect	61.0	29.6	64.8	74.9
Attribute	60.7	29.5	64.4	74.6
Combined	**62.2**	**31.5**	**65.8**	**75.3**

Table 5. Effect of the multi-task commonsense adaptation with ViT-B16 and LSTM encoder.

With increasing distillation loss ratio, the performance considerably improves upon the baseline. The performance gain gradually saturates since the combined weighted loss puts less weight on the final captioning objective.

4.3. Effect of the Multi-scale Temporal Fusion

We investigate the effect of multi-scale features in the proposed hierarchical fusion module (Table 4). We note that the performance improves considerably for all choices of fusion modules with increasing feature scales. The best performance is achieved with multi-scale LSTM based fusion.

4.4. Effect of the Commonsense Adaptation

We study the effect of commonsense knowledge generation on the final captioning performance (Table 5). We note that incorporating any commonsense aspect improves the caption quality over the baseline that uses no commonsense sub-decoder. The combination of all three commonsense aspects provides the highest accuracy gain over using a single one, thereby showing the effectiveness of commonsense knowledge adaptation.

5. CONCLUSION

In this paper, we introduce Clip4VideoCap, a commonsense enriched caption generation framework with multi-scale temporal fusion. The CLIP enriched visual features are found to be very effective when incorporating learning from large-scale multi-modal pre-training. The proposed knowledge distillation based scheme properly exploits the textual knowledge that greatly encourages the model to incorporate the contrastive multi-modal features. The proposed multi-scale temporal fusion scheme exploits short and long range temporal interactions for generalizing different scales of temporal contexts. Moreover, we find that generating commonsense knowledge as an intermediate sub-task greatly enhances the generated caption quality. We carried out extensive experiments on the MSR-VTT dataset and achieved state-of-the-art performance thereby showing the effectiveness of the proposed approach.

Acknowledgements

This research was supported in part by the Office of Naval Research, Minerva Program, and a UT Cockrell School of Engineering Doctoral Fellowship.

6. REFERENCES

[1] Iwan Setyawan and Reginald L Lagendijk, "Human perception of geometric distortions in images," in *Security, Steganography, and Watermarking of Multimedia Contents VI*. SPIE, 2004, vol. 5306, pp. 256–267.

[2] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.

[3] Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino, "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche," *Science advances*, vol. 5, no. 9, pp. eaaw2594, 2019.

[4] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang, "Video2commonsense: Generating commonsense descriptions to enrich video captioning.," *arXiv preprint arXiv:2003.05162*, 2020.

[5] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li, "Clip4caption: Clip for video caption.," in *Proceedings of the 29th ACM International Conference on Multimedia.*, 2021.

[6] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval.," *arXiv preprint arXiv:2104.08860*, 2021.

[7] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun, "Leveraging video descriptions to learn video question answering," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[8] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[9] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli, "Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017," *Advanced Robotics*, vol. 33, no. 15-16, pp. 764–799, 2019.

[10] Ron Mokady, Amir Hertz, and Amit H Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision.," *International Conference on Machine Learning.*, 2021.

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

[13] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021.

[14] Haoran Chen, Jianmin Li, and Xiaolin Hu, "Delving deeper into the decoder for video captioning," *arXiv preprint arXiv:2001.05614*, 2020.

[15] Liang Sun, Bing Li, Chunfeng Yuan, Zhengjun Zha, and Weiming Hu, "Multimodal semantic attention network for video captioning," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1300–1305.

[16] Shizhe Chen, Qin Jin, Jia Chen, and Alexander G Hauptmann, "Generating video descriptions with latent topic guidance," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2407–2418, 2019.

[17] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez, "Attentive visual semantic specialized network for video captioning," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5767–5774.

[18] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez, "Improving video captioning with temporal composition of a visual-syntactic embedding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3039–3049.

[19] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

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