A Literature Survey on Image and Video Captioning using Deep Learning and Knowledge Graph

19MIS1018 – B DEVI PRASAD

1. ABSTRACT:

Image and Video captioning are considered to be intellectually challenging problems in imaging science. The application domains include automatic captioning generation for images and videos for people who suffer from various visual impairment and languages problems. Caption generating for videos and images requires more visual representation along with the generating process. The existing video captioning methods focus on making an exploration of spatial temporal representations and their relationships to produce inferences. Such methods only exploit the superficial association contained in the video itself without considering the intrinsic visual commonsense knowledge that existed in a video data set, which may hinder their capabilities of knowledge cognitive to reason accurate descriptions. The main task for video or image captioning is to describe visual content in terms of natural language processing. The current systems need somewhere to fill the gap between low-level and high-level features while mapping. Therefore, to tackle this problem, there is a need to describe the latest research and methods to overcome difficulties and to propose effective solutions. Which can be performed using potential of Deep learning methods, Knowledge Graph and Natural Language processing.

Keywords: Image/Video Captioning, Deep Learning, Knowledge Graph, Natural language processing, Caption generating, Visual representation.

2. INTRODUCTION:

The significance and applications of artificial intelligence (AI) in automatically describing visual data in images and videos, particularly focusing on image and video captioning. Image captioning involves generating a natural language description of an image, while video captioning extends this to describe events and actions in a video, considering temporal dynamics. Various deep learning methods, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Knowledge Graphs (KGs), have been employed to improve image and video

captioning. These methods use neural networks to learn features from images and videos, aiding in tasks like object and attribute detection, understanding interactions, and establishing relationships among visual elements.

The text highlights the challenges involved in image and video captioning, emphasizing the need for accurate and grammatically correct descriptions. The applications of these technologies span across surveillance, security, healthcare, and assisting visually impaired individuals. Image captioning is described as a significant field in AI and computer vision, with applications in information retrieval, education, natural language processing, and social media. The evolution of vision tasks, such as action and image classification, recognition of objects and scenes, is noted, emphasizing the progress made in these areas. The text also introduces the concept of generating automatic descriptions for images, essential for effective communication between machines and humans.

Video captioning is presented as an attractive task in AI, computer vision, and knowledge graphs. The goal is to provide human-understandable language descriptions for video content, capturing temporal dynamics and offering more information compared to still images. Applications of video captioning include Video Retrieval Systems, Visual Questioning Answering, assistance for visually impaired individuals, Text-to-speech technology, and more.

3. PRELIMINARIES:

3.1 Algorithms/Model Used for Image and Video Captioning using Deep Learning:

- (CNN+GRU) (Convolutional Neural Network + Gated Recurrent Unit.)
- Recurrent Neural Networks (RNNs)
- Generative Adversarial Networks (GANs)
- Transformer Networks
- A Streamlined approach C3D+GRU+RNN
- LSTM + GAN (Long short term Memory + Gated Recurrent unit)
- Convolutional Neural Network +Transformer
- CNN+RNN(Convolutional Neural Network + Recurrent Neural Network)
- CNN and RNN-LSTM
- Graph Convolutional Network (GCN)
- Visual Commonsense-aware Representation Network

- Visual Semantic Reasoning (VSR).

3.2 Dataset Used:

- MSR-VVT
- MSVD, MSR-VVT, M-VAD and MPII-MD
- MSCOCO
- ActivityNet Captions
- Flickr8K
- MSR-VTT (Microsoft Research Video to Text)
- MSVD (Microsoft Research Video Description)
- Flickr30K
- From YouTube of average duration 9 second created by Amazon Mechanical Turk workers
- ActivityNet dataset

4. LITERATURE REVIEW:

[1] **Multimodal feature fusion based on object relation for video captioning (Zhiwen Yan, Ying Chen, Jinlong Song, Jia Zhu y)**

- Proposed a novel video captioning framework (ORMF) based on the object relation graph and multimodal feature fusion.ORMF uses the similarity and Spatio-temporal relationship of objects in videoAt the same time, ORMF also constructs a multimodal features fusion network to learn the relationship between different modal features. The multimodal feature fusion network is used to fuse the features of different models.
- The data collected The Microsoft video captioning corpus (MSVD) datasetcontains 1970 short videos with an average duration of 10s.

[2] Prompt Learns Prompt: Exploring Knowledge-Aware Generative Prompt Collaboration for Video Captioning (Liqi Yan, Cheng Han, Zenglin Xu, Dongfang Liu and Qifan Wang)

- Proposed a Video-Language Prompt tuning (VL-Prompt) approach for video captioning, which frst effciently pre-train a video-language model to exatract key information with fexibly generated Knowledge-Aware Prompt(KAP). Then,

designed a Video-Language Prompt (VLP) to utilize the knowledge from KAP and fnetune the model to generate full captions.

- Two datasets are used MSR-VTT consists of 10K video clips. Each video clip has 20 ground-truth captions. We use the standard split, which has 6.5K training videos, 497 validation videos and 2.9K testing videos. MSVD is a collection of 2K video clips downloaded from YouTube. Each video clip has roughly 40 ground-truth captions written by humans.

[3] **Visual Commonsense-aware Representation Network for Video Captioning (Pengpeng Zeng, Haonan Zhang, Lianli Gao, Xiangpeng Li, Jin Qian and Heng Tao Shen, Fellow, IEEE)**

- Proposed Visual Commonsense-aware Representation Network (VCRN), for video captioning. Specifically, we construct a Video Dictionary, a plugand-play component, obtained by clustering all video features from the total dataset into multiple clustered centers without additional annotation. Each center implicitly represents a visual commonsense concept in the video domain, which is utilized in our proposed Visual Concept Selection (VCS) to obtain a video related concept feature.
- The dataset used isMSR-VTT consists of 10K video clips. Each video clip has 20 ground-truth captions. We use the standard split, which has 6.5K training videos, 497 validation videos and 2.9K testing videos. b)MSVD is a collection of 2K video clips downloaded from YouTube. Each video clip has roughly 40 ground-truth captions written by humans.

[4] Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods (Mohammad Saif Wajid, Hugo Terashima-Marin, Peyman Najafirad, Mohd Anas Wajid.)

- The developed model has Captioning methods into two groups: the deep learning approaches and knowledge graph based approaches. Each category is based on each research method's fundamental characteristics and differences. Many researchers have employed various scene interpretation techniques, including the encoder-decoder and attention mechanisms.
- The rapid development of this field of study have been the availability of labeled datasets for video descriptions. Except for a few datasets containing many phrases or even paragraphs per video sample, most datasets only assign one caption per video. Here we mentioned benchmark datasets for video captioning

[5] **Exploring deep learning approaches for video captioning: A comprehensive review (Adel Jalal Yousif a, Mohammed H. Al-Jammas)**

- The proposed deep learning video captioning with different aspects including fundamental concepts, benchmark datasets, evaluation metrics, methods, and challenges. this study focused on understanding the key components of video captioning, such as the encoder, decoder, and word embedding techniques.
- The datasets like MSR-VTT, MSVD, MPII-MD, YouCook II, and ActivityNet Captions, which serve as valuable resources for training and evaluating deep learning models.

[6] Cptr: Full Transformer Network For Image Captioning (Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, Jing Liu)

- The developed model Image captioning as a sequenceto-sequence prediction task and propose CPTR, a full Transformer model to replace the conventional "CNN+Transformer" procedure.
- MS COCO dataset which is the most commonly used benchmark for image captioning. To be consistent with previous works, we use the "Karpathy splits" which contains 113,287, 5,000 and 5,000 images for training, validation and test, respectively.

[7] **Video captioning using transformer network (Mubashira I. Nechikkat, Bhagyasree V. Pattilikattil, Soumya Varma)**

- The proposed deep learning algorithms into various existing video captioning architectures. The insights of the detailed study lead us to propose a transformer network for video captioning that is proving promising. This transformer architecture is based on multi-head self-attention and encoder-decoder attention. Transformer is different from the LSTM based models as it processes the input as a whole and parallel. The decoder of this network is auto-regressive because it makes the prediction one at each time step.
- This dataset contains 1970 short videos taken from YouTube of average duration 9 second created by Amazon Mechanical Turk workers. Each video is described with 40 different captions. In this dataset, 1200 videos are used for training, 100 and 670 for validation and testing respectively.

[8] Text with Knowledge Graph Augmented Transformer for Video Captioning (Xin Gu1,Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, Longyin Wen)

- Proposed knowledge graph augmented transformer for video captioning, which aims to integrate external knowledge in knowledge graph and ex ploit the multi-modality information in video to mitigate long-tail words challenge. Externsive experiments con ducted on four challenging datasets demonstrate the effec tiveness of the proposed method
- Datasets YoucookII contains 2, 000 long untrimmed videos from 89 cooking recipes, including 1, 333 videos for training, 457 videos for testing, ActivityNet Captions is a large-scale challenging dataset for video captioning, with 10, 024 videos for training, 4, 926 for validation, and 5, 044 for testing, MSR-VTT includes 10, 000 video clips with 15 seconds on average, MSVD consists of 1970 video clips collected from YouTube and each clip is approximately 10 seconds long.

[9] Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap (Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, And Hamid R. Arabnia)

- The Developed Methodologies that utilize Deep Learning offer great potential for applications that automatically attempt to generate captions or descriptions about images and video frames.
- The application domains include automatic caption (or description) generation for images and videos for people who suffer from various degrees of visual impairment; the automatic creation of metadata for images and videos (indexing) for use by search engines; general-purpose robot vision systems; and many others.
- Datasets are widely used to evaluate and compare image captioning methods: Flickr8K , Flickr9K , Flickr30k , and Microsoft COCO

[10] **Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph (Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo)**

- The Developed entity-aware image captioning method that constructs a multi-modal knowledge graph by exploring external knowledge from the web. The method can simultaneously associate visual cues with named entities and capture the fine-grained relationships.
- The proposed method first constructs a text sub-graph that consists of the named entities and their relationships, and an image sub-graph containing the visual objects in the image, and then con nects the similar named entities and visual objects using a cross-modal entity matching module.

- Two news image captioning datasets, Goodbyes and NYTimes800k. The images, captions and news articles in these datasets are collected from New York Times, and each image is an notated with one ground-truth caption.

[11] Leveraging Weighted Fine-Grained Cross-Graph Attention for Visual and Semantic Enhanced Video Captioning Network (Deepali Verma, Arya Haldar, Tanima Dutta)

- The Developed video captioning approach to generate human-like captions based on visual and semantic information. It uses a graph construction module to generate weighted visual regions and semantic knowledge graphs for a given video and its available captions.
- Microsoft Video description corpus (YouTube2Text) (Guadarrama et al. 2013) is a benchmark VC dataset which contains 1970 short YouTube clips. There are about 8000 clips description pairs and each video has approximate 10-20 descriptions.

[12] **Bidirectional transformer with knowledge graph for video captioning (Maosheng Zhong, Youde Chen)**

- The Developed Model based on transformer architecture have risen to prominence for video captioning. However, most models are only to improve either the encoder or the decoder, because when we improve the encoder and decoder simultaneously, the shortcomings of either side may be amplified. Based on the transformer architecture, we connect a bidirectional decoder and an encoder that integrates fine-grained spatio-temporal features, objects, and relationships between the objects in the video.
- Datasets, MSVD and MSR-VTT, demonstrate the effectiveness of BTKG, which achieves state-of-the-art performance in significant metrics.

[13] Dense Video Captioning using BiLSTM Encoder (Jyoti Madake, Shripad Bhatlawand, Swarali Purandare, Swati Shilaskar, Yash Nikhare)

- The Developed Video captioning has been a widely researched topic integrating visual information and natural language but performing video captioning on long untrimmed videos is still challenging as the video contains multiple events and the model has to describe each event.
- This model is trained and tested on MSVD dataset which has around 2000 videos and their corresponding captions. The proposed framework shows increased accuracy in video captioning in terms of BLEU score 0.78 and METEOR score 0.34.

[14] Image Captioning Using CNN and RNN (S Rohitharun, L Uday Kumar Reddy, S Sujana)

- The Developed Image Caption is a concept of gathering the right description of the given image on the internet use Computer Vision and natural language processing. The following is achieved using the Deep learning techniques called as convolution neural network and recurrent neural network.
- The dataset used for implementation is called as the Flickr8_k Dataset. The model uses the combination of convolution neural network which helps in extraction whereas the recurrent neural network helps in generation of the right text.

[15] **Modeling graph-structured contexts for image captioning (Jiahui Wei , Feicheng Huang a, Huifang Ma b.)**

- The proposed image captioning has been significantly improved recently through deep neural network architectures combining with attention mechanisms and reinforcement learning optimization. Exploring visual relationships and interactions between different objects appearing in the image, however, is far from being investigated.
- he MSCOCO dataset and the Flickr30k dataset validate the superiority of our SGT model, which can realize state-of-the-art results in terms of all the standard evaluation metrics.

5. TOOLS SUPPORTED AND USED:

Jupyter Notebook or Google Colab:

Jupyter Notebook is a project to develop open-source software, open standards, and services for interactive computing across multiple programming languages.

PYTORCH:

PyTorch is a machine learning framework based on the Torch library, used for applications such as computer vision and natural language processing,

6. CONCLUSION:

The evaluation of image/video captioning and dense video captioning methods, categorizing them into deep learning and knowledge graph-based approaches. The

study highlights popular datasets and evaluation metrics, such as MSCOCO, Flickr8K, Flickr30K, MSVD, and MSR-VTT. It emphasizes the effectiveness of knowledge graph-based methods in improving semantic understanding, consistency, and coherence of generated captions. The three enhancements for video/Image captioning, incorporating Graph Convolutional Networks (GCN) for relationship features, a multimodal feature fusion network, and a caption length loss for richer captions. The model exhibits superior performance on MSR-VTT and MSVD datasets, affirming its effectiveness.

7. REFERENCES:

[1] S. Rohitha run, L. Uday Kumar Reddy, S. Sujana, Image captioning using CNN and RNN, in: 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1–8

[2] A. Tripathi, M.K. Gupta, C. Srivastava, P. Dixit, S.K. Pandey, Object detection using YOLO: a survey, in: 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 747–752.

[3] J. Jin, J. Ye, X. Lin, L. He, Pseudo-query generation for semi-supervised visual grounding with knowledge distillation, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.

[4] Ankit Kumar, Rakesh Kumar Yadav, DilipKumar Jang Bahadur Saini, Create and mplement a new method for robust video face recognition using convolutional neural network algorithm, e-Prime - advances in electrical engineering, Electron. Energy Volume 5 (2023)

[5] Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR), 52(6), 1-37.

[6] Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. Cptr: Full transformer network for image captioning. arXiv 2021. arXiv preprint arXiv:2101.10804.

[7] Nechikkat, M. I., Pattilikattil, B. V., Varma, S., & James, A. (2022, October). Video captioning using transformer network. In AIP Conference Proceedings (Vol. 2494, No. 1). AIP Publishing.

[8] Zhao, W., & Wu, X. (2023). Boosting entity-aware image captioning with multi-modal knowledge graph. IEEE Transactions on Multimedia.

[9] Verma, D., Haldar, A., & Dutta, T. (2023, June). Leveraging Weighted Cross-Graph Attention for Visual and Semantic Enhanced Video Captioning Network. In

Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 2, pp. 2465-2473).

[10] Zhong, M., Chen, Y., Zhang, H., Xiong, H., & Wang, Z. (2023). Bidirectional transformer with knowledge graph for video captioning. Multimedia Tools and Applications, 1-20.

[11] S. Rohitharun, L. Uday Kumar Reddy and S. Sujana, "Image Captioning Using CNN and RNN," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-8, doi: 10.1109/ASIANCON55314.2022.9909146.

[12] Li, Z., Wei, J., Huang, F., & Ma, H. (2023). Modeling graph-structured contexts for image captioning. Image and Vision Computing, 129, 104591.

[13] Zeng, P., Zhang, H., Gao, L., Li, X., Qian, J., & Shen, H. T. (2023). Visual commonsense-aware representation network for video captioning. IEEE Transactions on Neural Networks and Learning Systems.

[14] Yan, L., Han, C., Xu, Z., Liu, D., & Wang, Q. (2023, August). Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI). International Joint Conferences on Artificial Intelligence Organization. https://doi. org/10.24963/ijcai (Vol. 180).

[15] Yan, Z., Chen, Y., Song, J., & Zhu, J. (2023). Multimodal feature fusion based on object relation for video captioning. CAAI Transactions on Intelligence Technology, 8(1), 247-259.