Dense Video Captioning

Srikumar Brundavanam[1] **Mukul Ganwal**[1] **David Ologan** [1] **Shang Shi**[1]
Carnegie Mellon University
Department of Mechanical Engineering
{vbrundav,mganwal,dologan,shangs}@andrew.cmu.edu

Abstract

In recent years, there has been a large focus on the research of video captioning. Video captioning can be used in many fields such as self-driving cars and video surveillance. However, the current models struggle to generalize to unseen videos. Therefore, the team proposes a project which will employ large language models to help to generalize the model to produce better captions on unseen videos. The datasets the team will use to train and validate our model include ActivityNet and YouCook2 among others. The baseline model the team chose is End-to-end dense video captioning with parallel decoding (PDVC) [1]. On top of PDVC, the team will inject a large language model into different locations and compare the performance against the baseline.

1 Introduction

Video captioning has been a large focus of research in recent years. A good video captioning model has huge implications in a variety of fields. It can aid those with visual impairments, self-driving cars, caption surveillance [2] and even automatically labeling data. Current video captioning models struggle to provide meaningful captions to videos not seen during training. Therefore this project aims to generate detailed captions of videos by making use of large language models. One example of this is to generate semantic features using a large language model which can align with the video features [3] before feeding into the transformer decoder. Large language models will allow us to build upon existing video captioning implementations to make them more generalizable to unseen video data.

2 Dataset Descriptions

This project aims to leverage the following preexisting data sets to test and validate our model. ActivityNet Captions' dataset includes 20,000 footage of people engaging in various activities. Each movie lasts for roughly two minutes and contains four descriptive sentences. For training, validation, and testing, the dataset is divided into three sections. Another dataset called YouCook2 contains 2,000 cooking films that are each about five minutes long and have eight or so description sentences. For training, validation, and testing purposes, this data set is additionally divided into three sections.

3 Project Description

The primary goal of our project is to develop a highly generalized model capable of producing quality captions of a variety of video types. Utilizing PDVC (Parallel Decoding for Video Captioning) as our baseline model, the project aims to deploy large language models into PDVC to improve overall performance. For example, aligning visual feature inputs with semantic features [3] from a large language model before feeding into the encoder. Another direction the team will explore is to improve

Introduction to Deep Learning, 11-785 (IDL 2023), Pittsburgh, PA

the caption head by injecting that with a large language model as well for a more generalized output caption. The team will then evaluate the performance of the new model against PDVC to see the improvements.

Currently, we plan to take advantage of PyTorch libraries to build our model, while using collaborative tools and editors like Git and VSCode to streamline our workflow. AWS/GCP will also be used to train our models on the cloud.

4 Literature Review

SWINBERT: End-to-End Transformers with Sparse Attention for Video Captioning: [4]
This paper proposes a new end-end fully transformer-based architecture called SwinBERT for video captioning. The SWINBERT model is a video-based pure-Transformer architecture designed for caption generation. It consists of two modules: Video Swin Transformer (VidSwin) and Multimodal Transformer Encoder. VidSwin is used to extract spatial-temporal video representations from raw video frames, while the Multimodal Transformer Encoder takes the video representations and outputs a natural language sentence through sequence-to-sequence generation. The VidSwin module tokenizes the grid features of the raw video frames along the channel dimension to generate video tokens. The Multimodal Transformer Encoder module performs seq2seq generation to form a natural language sentence, with textual and visual modalities inputs, and uses a causal self-attention mask to simulate a uni-directional seq2seq generation process. All textual tokens have full attention to the video tokens. This helps with improved training and improves the performance of the longer videos.

End-to-end dense video captioning with parallel decoding: [1]
This paper proposes a dense video captioning framework with parallel decoding (PDVC). Previous methods usually separate the task into event localization and captioning forming a two-stage pipeline. This causes several issues such as the downstream captioning task being highly dependent on the quality of the generated event proposals. PDVC avoids this by feeding the enhanced representations of event queries into the localization and captioning head in parallel. This creates a deep connection between the two tasks allowing them to be mutually optimized. The overall model first adopts a pre-trained video feature extractor and transformer encoder to obtain frame-level features. The features are sent through a transformer decoder to output an event counter which ensures the correct amount of events predicted, localization head and caption head. There exists room for improvement on top of PDVC since only visual features are being fed into the transformer encoder. In addition, a better captioning model can be deployed instead of the current LSTM.

5 Baseline Model

The baseline model that we use for this project is the dense video captioning with parallel decoding (PDCC) mentioned in Wang et al.[1]. The proposed PDVC model decodes frame features, extracted from a Vision Transformer, into an event set with their respective locations and captions. PDVC aims to directly exploit inter-task association at the feature level by applying localization head and captioning head in parallel. Since the quality of dense captioning is dependent on the size of the event set, a newly proposed event counter is added on top of the decoder to accurately predict the number of final events.

Wang et al. evaluate the localization performance, dense captioning performance, and paragraph captioning performance of PDVC with different metrics. For localization performance, the average precision, recall across Intersection over Union (IOU), and F1 scores were considered. Average precision measurement metrics like CIDEr[5], BLEU4 [6], and METEOR [7] were used to evaluate the dense captioning performance of PVDC. Further SODAc [8] was used for the overall performance evaluation of PDVC. Again CIDEr, BLEU4, and METEOR were used for evaluating paragraph captioning performance. These will be the evaluation metrics our team plans to use and improve through our proposed approach.

6 Team Members, Responsibilities

Table 1: Team Member Responsiblities

Name	Description
Srikumar Brundavanam	Responsible for setting up and training the baseline PVDC model
Mukul Gamwal	Responsible for setting up and training the baseline PVDC model
David Ologan	Responsible for research and integration of new large language models
Shang Shi	Responsible for research and integration of new large language models

7 Midterm Milestone

Before the midterm milestone deadline, we hope to train the baseline model on both ActivityNet and YouCook2 datasets and reach comparable accuracy mentioned in Wang et al.[1]. Our team will also research language models to integrate with the above baseline model to further improve accuracy of our captions.

References

[1] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. *CoRR*, abs/2108.07781, 2021.

[2] Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia. Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access*, 8:218386–218400, 2020.

[3] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. *CoRR*, abs/2111.13196, 2021.

[5] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

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

[7] Alon Lavie and Michael Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115, 09 2009.

[8] S.Fujita, T.Hirao, H.Kamigaito, M.Okumura, and M.Nagata. Soda: Story-oriented dense video captioning evaluation framework. *Proc. Eur. Conf. Comput. Vis.*, 2020.