AI VIDEO CAPTIONING

STUDENTS: HSIAO-TING HUANG, RUBY LIN, PRANAV KAMATH,
BHAVAN RAJAN NEELAKANDAN, SADJYOT HEMANT GANGOLLI

WYZE

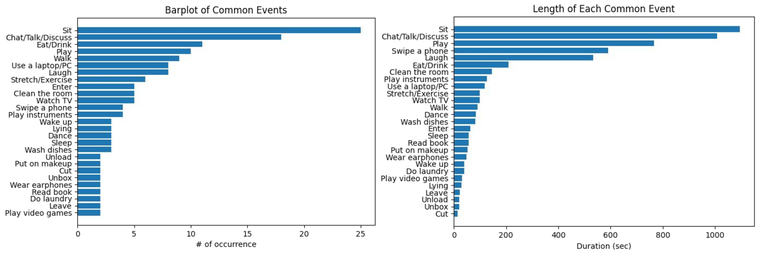
42

Introduction

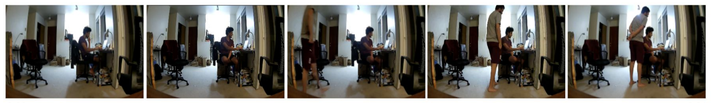
- The team is focusing on exploring technical solutions of video captioning for indoor/outdoor scenarios and super long video (> 1 hour), video content understanding by bridging visual and language information and leveraging video data to generate captions to describe the semantic content in video recordings.
- To generate the captions for the super long video, the team worked to do in-depth research on academia-leading solutions on dense video captioning and improve state-of-the-art method, PDVC model, to generate descriptions on Wyze device captured videos.
- Different from normal video captioning, the task of dense video captioning involves both detecting and describing events in a video; we can describe the super long video well.
- Our model and software developed will be used to create meaningful captions that help users monitor their daily events and identify situations that need their attention.

Dataset

- Our dataset consists of indoor home scenarios:
  - 100 videos within 20~30 sec (25 from Wyze cam + 75 from Internet)
  - 10-min video*1 + 1-hr videos*5 (all from Wyze cam)
  - 27 Most common events
- The total duration of the dataset is 349.25 minutes and the average length is 197.69 seconds. The 27 most frequently occurring events are shown in the barplot. The most common event is “Sit”, which appears in 25 videos. Next comes the event “Chat/Talk/Discuss”, with 18 occurrences. On the rightmost side of the barplot are the least frequent events, appearing in only 2 videos.



DATASET EXAMPLES:



0.00 — 0.21

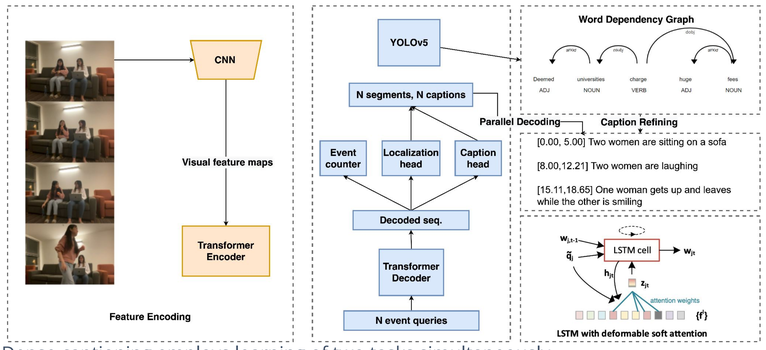
GT: A man is sitting on a chair working on the laptop **GT**: A man enters the room and two people are discussing while looking at the laptop

0.00 — 0.28

GT: A woman is sleeping in the bed **GT**: A woman opens the curtains and makes the bed **GT**: A woman starts stretching on the floor

Enhanced PDVC Model

- Our proposed solution is an end-to-end machine learning model with an Encoder-Decoder type architecture.
- The encoder is a 3D CNN which extracts multi-scale frame features from the pre-trained feature extractor.
- The decoder has three parallel heads to generate captions, predict temporal boundaries for events in a video and count the number of events in the input.
- The generated caption is refined using the YOLO Object Detection Model for obtaining the correct object class in the video

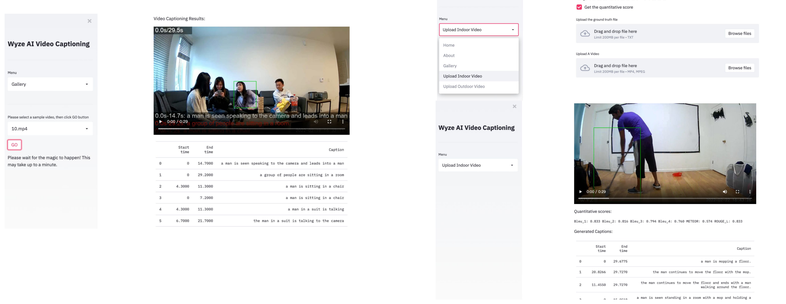


- Dense captioning employs learning of two tasks simultaneously
  - Localization of events
  - Captioning of events
- PDVC uses the same intermediate features so that information is shared between the two tasks.

Software System

- Using Streamlit with Ngrok and Colab to build the website.
- Features:
  - Gallery: User can test some sample videos.
  - Upload indoor/outdoor video:User can upload indoor/outdoor videos and get the quantitative results through their mobile and computer.

Software System:



Results

EVALUATION DATASET:

- A dataset designed to evaluate the performance of the video caption generated on about 50 indoor videos and 10 outdoor videos using the metrics mentioned below.

METRIC NAME	METHODOLOGY
BLEU	N-GRAM PRECISION
ROUGE	N-GRAM RECALL
METEOR	N-GRAM WITH SYNONYM MATCHING
CIDEr	TF-IDF WEIGHTED N-GRAM SIMILARITY

QUANTITATIVE RESULTS: (INDOOR)

MODEL	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
Pre-Trained	0.683	0.606	0.543	0.488	0.375	0.683	2.009
Fine-Tuned	0.785	0.725	0.669	0.620	0.424	0.749	2.854

QUANTITATIVE RESULTS: (OUTDOOR)

MODEL	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
Pre-Trained	0.538	0.357	0.222	0.132	0.244	0.484	0.445
Fine-Tuned	0.776	0.698	0.626	0.570	0.446	0.756	2.521
Fine_Tuned with YOLO	0.897	0.875	0.850	0.825	0.607	0.920	2.956

QUALITATIVE RESULTS:

0.00 — 0.30

GT: A man is standing in a kitchen and washing dishes
Pre-Trained: A man is washing dishes standing in a kitchen
Fine-Tuned: A man is standing in a kitchen with a rag and washing dishes

0.00 — 0.30

GT: A man is walking around the table with a dog in the room
Pre-Trained: A man is walking around the room
Fine-Tuned: A man is walking around the room and a dog in a room

GT: The man continues washing the dishes in the kitchen
Pre-Trained: A man starts doing dishes in a kitchen
Fine-Tuned: The man continues washing dishes in a kitchen

GT: A dog is walking around the table
Pre-Trained: A dog is walking in the room
Fine-Tuned: A dog starts walking on the floor

Future Work and References

- Future Work
  - Enhancement of the PDVC Model Architecture.
  - Creating an enhanced UI capable of user authentication and authorization.
  - Integration of the system with the Wyze Hardware for seamless integration as a single system.
  - Generating better captions for anomaly, vehicle based and outdoor scenarios.
- References
1. Alwassel, Humam, Silvio Giancola, and Bernard Ghanem. "Tsp: Temporally-sensitive pretraining of video encoders for localization tasks." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
2. Wang, Teng, et al. "End-to-End Dense Video Captioning with Parallel Decoding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISORS: HUNG-MIN HSU, JENQ-NENG HWANG, ZHONGWEI CHENG, CHENG-YEN YANG

SPONSORS: WYZE LABS