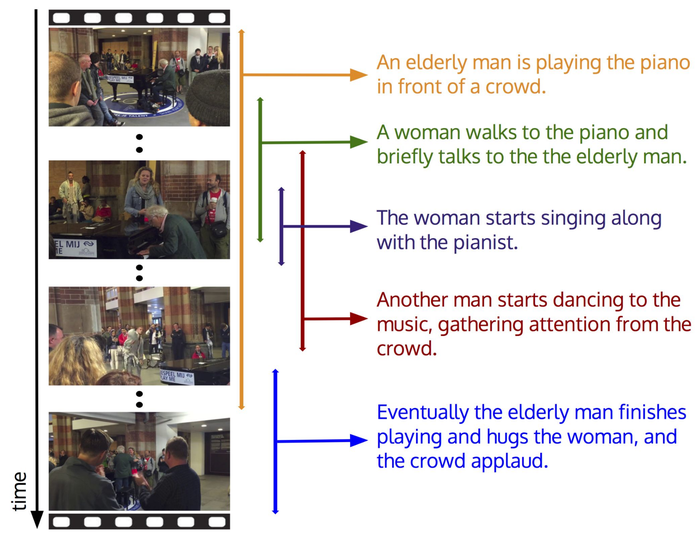
From Caption To Description: Video Dense Captioning with Knowledge Distillation

Wang-Chen Wu, Yizhou Zhao, Yudong Liu

Carnegie Mellon University

Introduction

- Dense Video Captioning

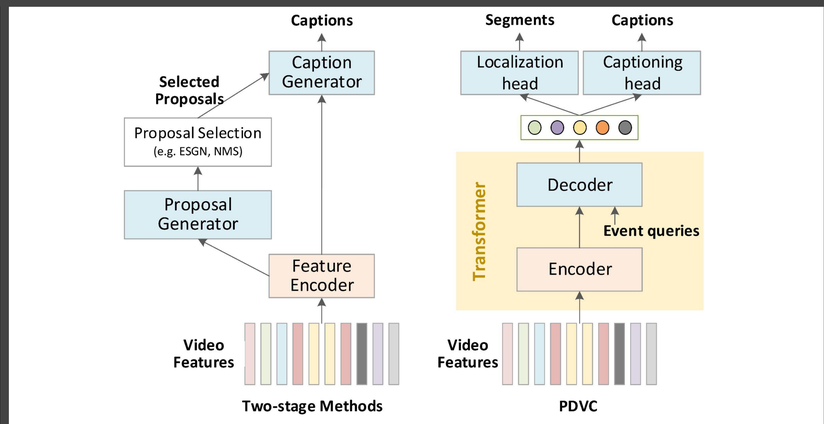


- Motivation
  - From our analysis, major cause for the failure case might be due to the generation procedure of the LSTM model from detected entity tokens of each video event segment

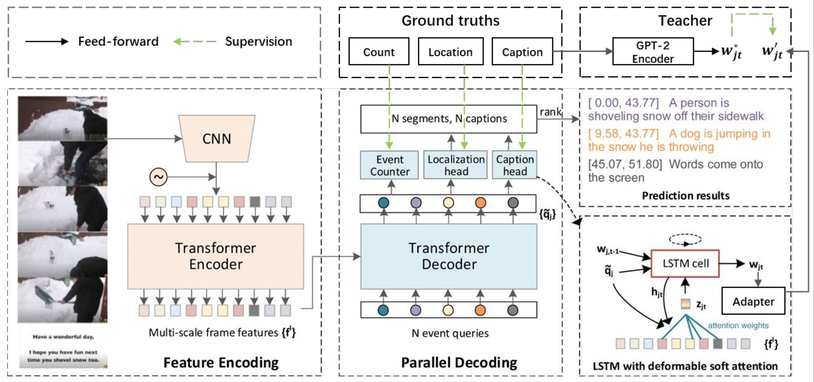
    "add some water to a bowl and add some water to a bowl, and place the salad on the salad"
  - Existing research direction only focus on sentence-level captioning, which might not be specific enough
- Goals
  - Improve sentence-level captioning via knowledge distillation with the help of strong pretrained LMs
  - Generate paragraph-level captioning and provide more details to serve as dense video description

Baseline models

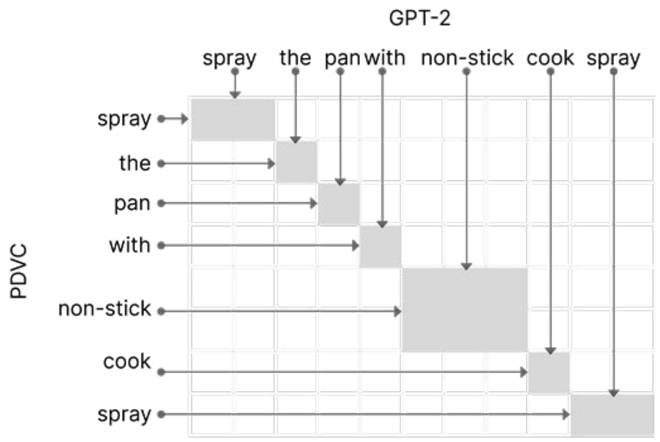
- End-to-end dense video captioning with parallel decoding (PDVC)
  - Formulate the dense caption generation as a set prediction task:
    - Deformable Transformer with an encoder-decoder structure is adopted to capture the inter-frame, inter-event, and event-frame interactions by attention mechanism and produce a set of event query features
    - Two parallel prediction heads predict the boundaries and captions for each event query simultaneously
    - An event counter predicts the event number N_{set} from a global view and selects top N_{set} events



Proposed Method



- **Improve Sentence-level Captioning**
  - Directly apply or append Generative Pretrained Transformer Model (GPT-2) to facilitate sentence-level captions generation of each event - does not work well
  - Knowledge Distillation with GPT-2 as the teacher
    - Feed GPT-2 with ground truth captions and pass its final hidden state to PDVC as a supervision signal
    - Append an adapter to PDVC to imitate the hidden state distribution of GPT-2, drop on inference
    - The adaptor learns to translate video captions to more general sentences and more specific descriptions
    - Word-wise distillation with a masked distillation loss



- **Towards Video Paragraph Captioning**
  - Remove Localization as simple baseline
  - Progressive generation of long captions
    - Coarse-to-fine generation from the initial short text
  - Paragraph caption generation with an adaptor
    - With knowledge from the teacher, the adaptor can generate more complex paragraph-level captions

Evaluation

- **Datasets**
  - YouCook2
    - 2k untrimmed videos
    - 89 cooking recipes
    - 22 video clips for each on average
  - ActivityNet Captions
    - 20k YouTube untrimmed videos
    - 100k caption annotations
    - 120s long with 3.7 events on average

Distillation Loss	β_{dist}	Predicted proposals B4	M	C	SODA_c
Baseline	0.0	1.82	7.48	28.16	5.47
Cos Similarity					

Table 1: Dense captioning on the ActivityNet Captions validation set. B4 / M / C is short for BLEU4 / METEOR / CIDEr. The same below.

Distillation Loss	β_{dist}	Predicted proposals B4	M	C	SODA_c
Baseline	0.0	**0.92**	4.54	**22.96**	4.20
L2 Distance	10.0	0.76	4.49	21.87	4.28
	20.0	0.68	4.45	20.59	**4.31**
	40.0	0.76	4.41	21.93	3.99
Cos Similarity	10.0	0.85	4.56	21.75	4.09
	20.0	0.91	**4.59**	22.33	4.19
	40.0	0.85	4.35	22.03	4.03

Table 2: Dense captioning on YouCook2.

Ongoing Progress

- ActivityNet evaluation
- Error analysis on the results
- Progressive generation of long text
- Human evaluation on paragraph captioning