

Dense-Captioning Events in Videos

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, Juan Carlos Niebles
Stanford University
ICCV 2017

Presenter: Yan-Bo Lin
11-14-2021

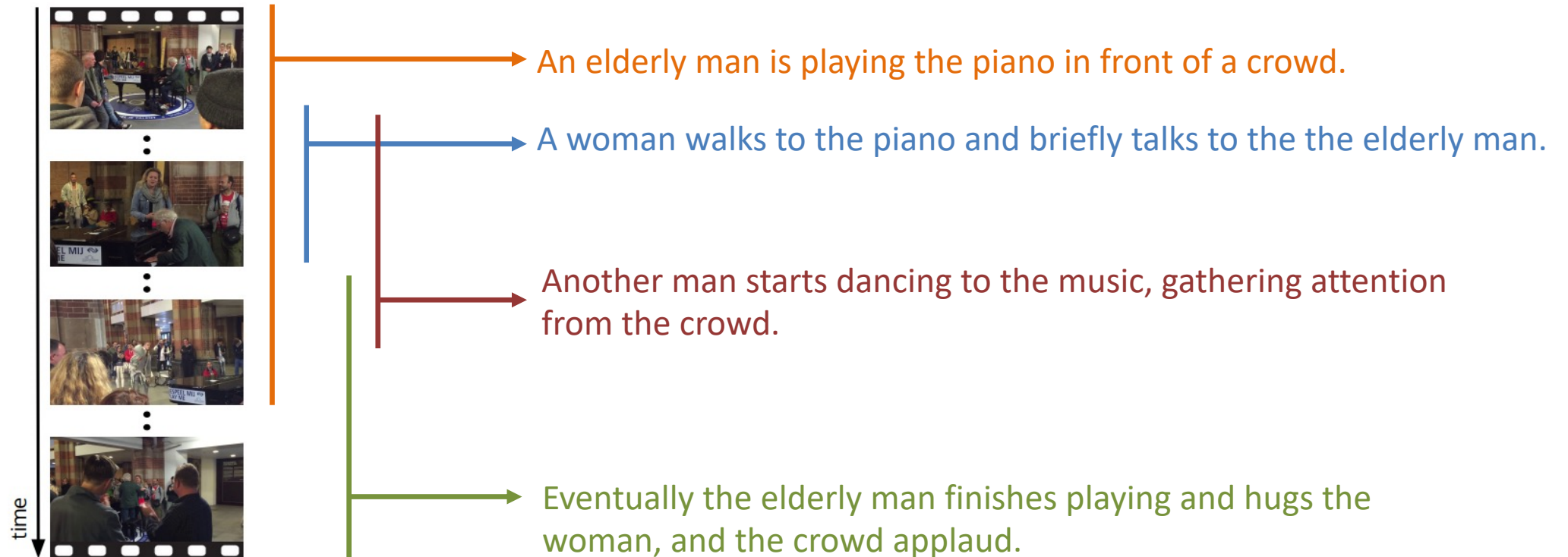
Overview

- Introduction
- Motivation
- Related works
 - Event proposal module
- Proposed framework
- Dataset
- Results
- Conclusion
- Discussion

Introduction

- What is **Dense-Captioning Events in Videos?**

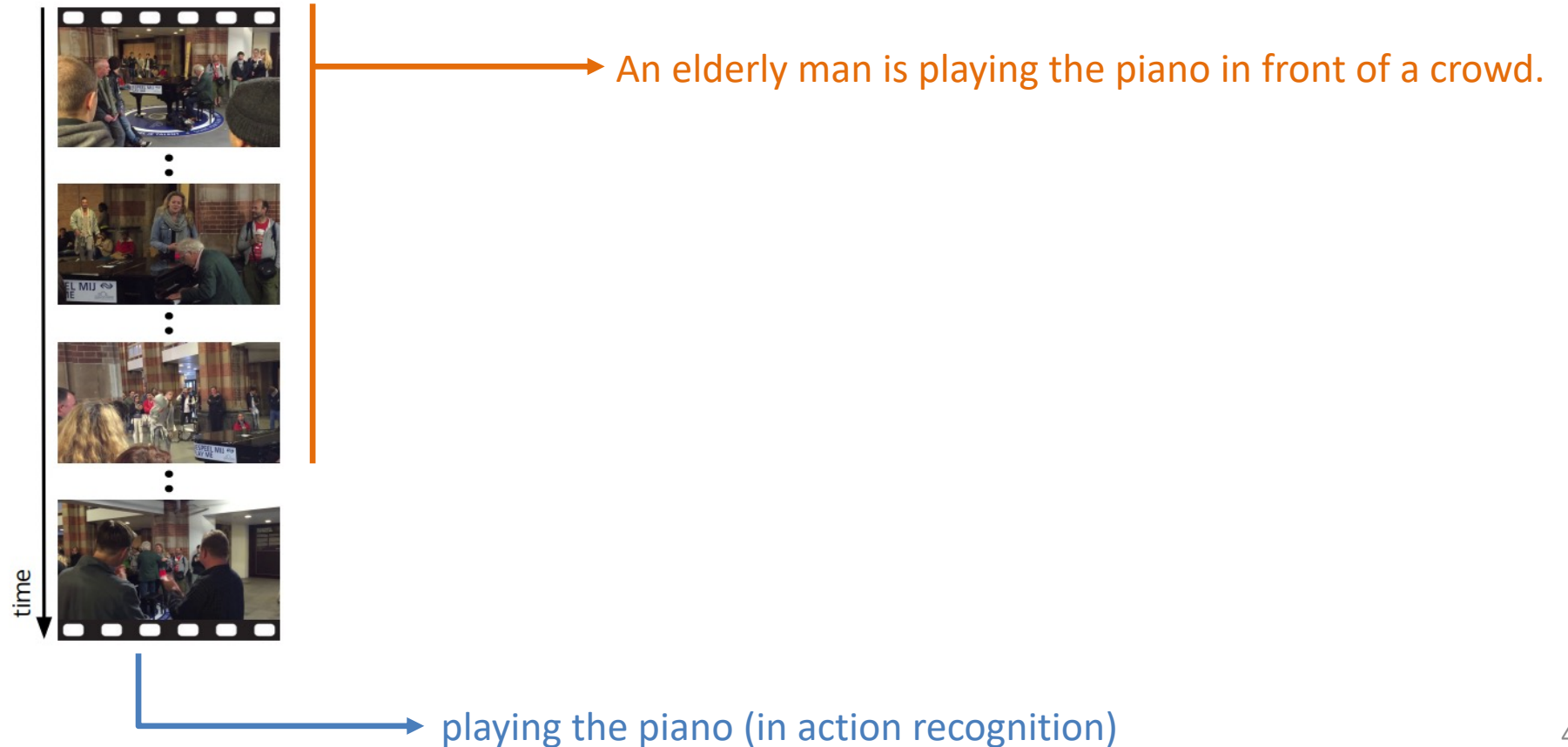
- Input: a video.
- Output: multiple captions for clips in a video.



Motivation

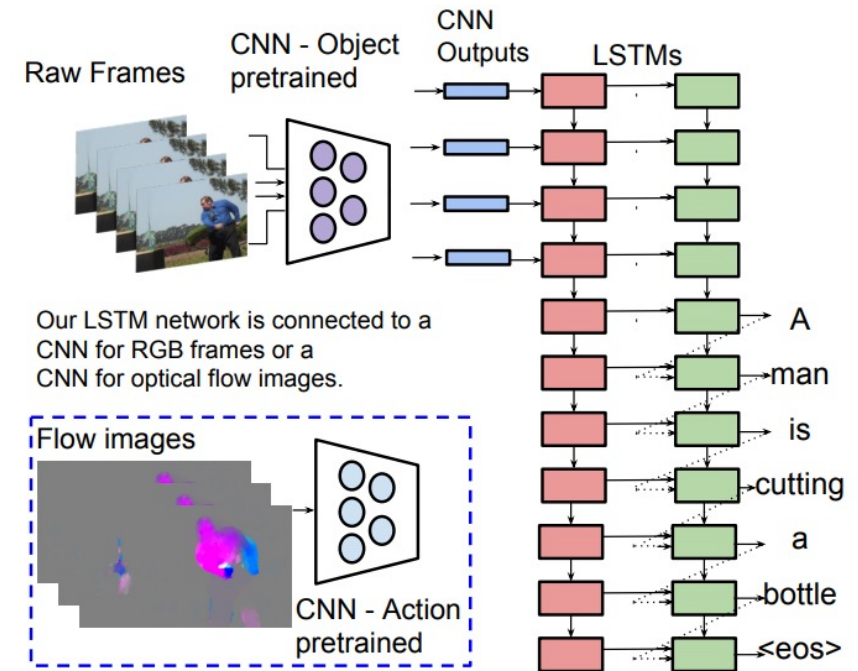
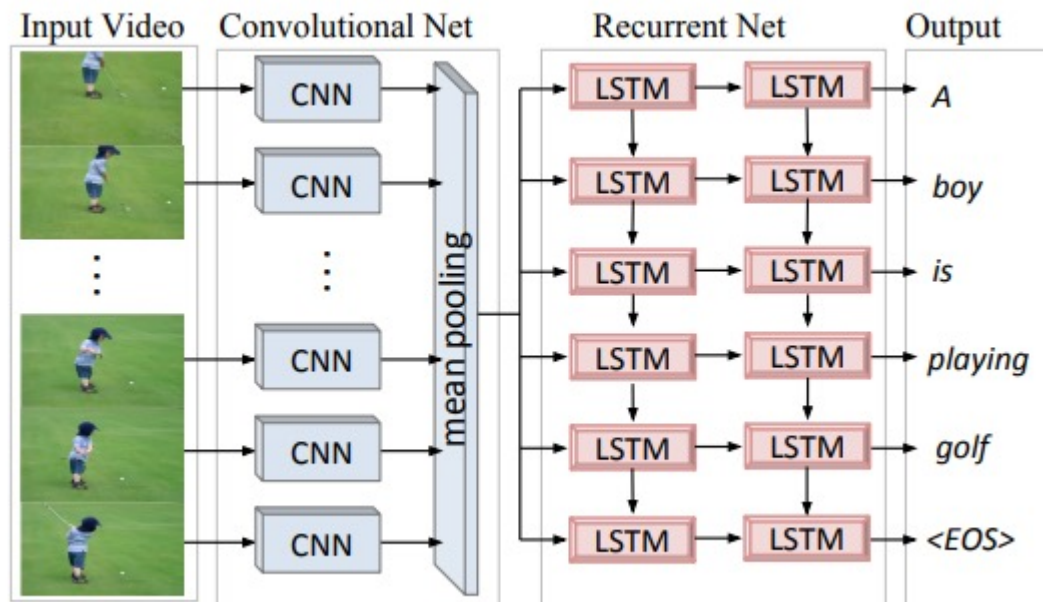
- Why Dense-Captioning Events in Videos:

- Dense caption events requires models to understand details (e.g., scenes, action, and characters...etc).



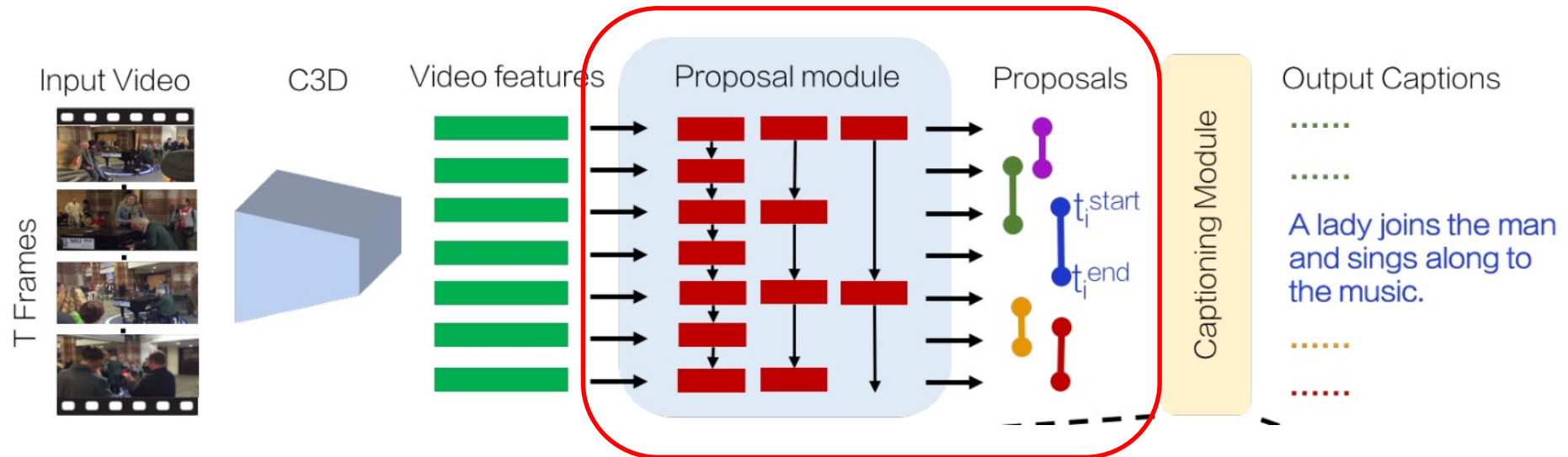
Motivation

- Limitation of traditional works on **Dense-Captioning Events in Videos**:
 - RNN-based methods only works well on short clips.
 - Long video inputs will lead vanishing gradients.



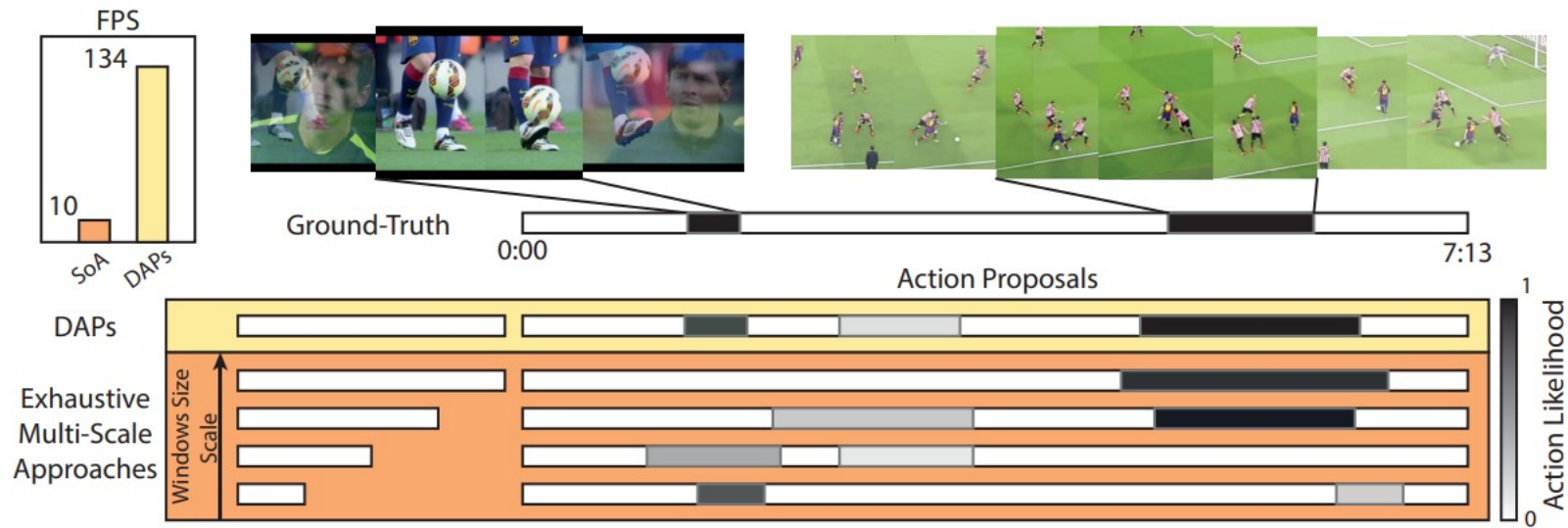
Motivation

- Leverage the action proposal module to detect events in a long video.
 - Alleviate gradient vanish issue in a clip.



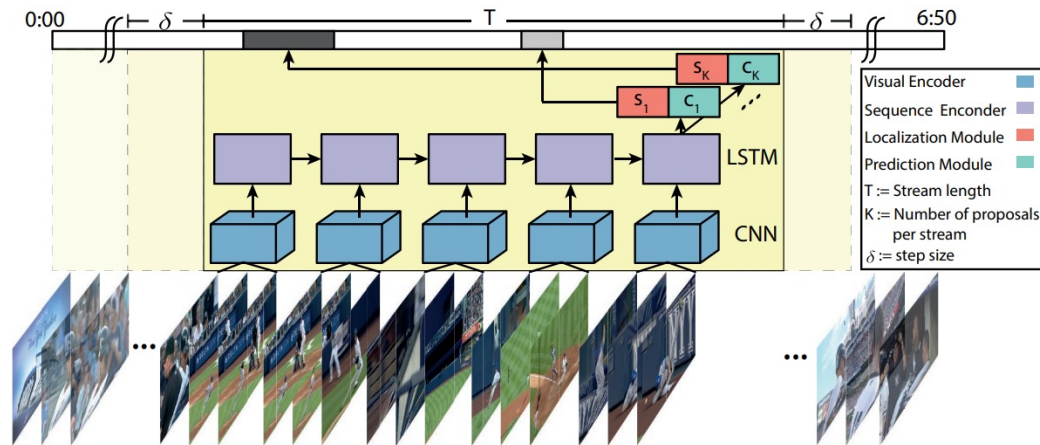
Related works

- DAPs: Deep Action Proposals for Action Understanding
 - Input: a video.
 - Output: temporal boundary and action events.



Related works

- DAPs: Deep Action Proposals for Action Understanding
 - CNN+LSTM module: predict K proposals with *confidences*.
 - $\mathcal{L}_{\text{match}}$ is L2 loss to penalizes matched segments that are distant from action annotations.
 - $\mathcal{L}_{\text{conf}}$ aims to optimize confidence value of matched proposal should be higher than others.

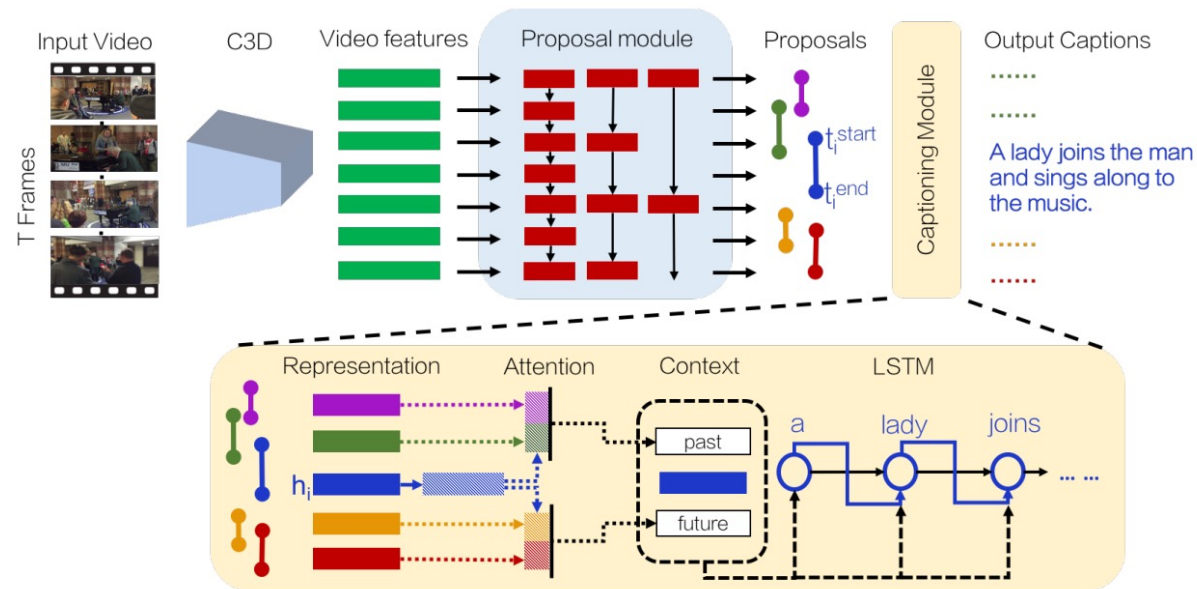


$$(\mathbf{x}^*, \theta^*) = \underset{\mathbf{x}, \theta}{\operatorname{argmin}} \quad \alpha \mathcal{L}_{\text{match}}(\mathbf{x}, S(\theta), A) + \mathcal{L}_{\text{conf}}(\mathbf{x}, C(\theta)) \quad \text{s.t.} \quad x_{ij} \in \{0, 1\}, \quad \sum_i x_{ij} = 1$$

Proposed Method

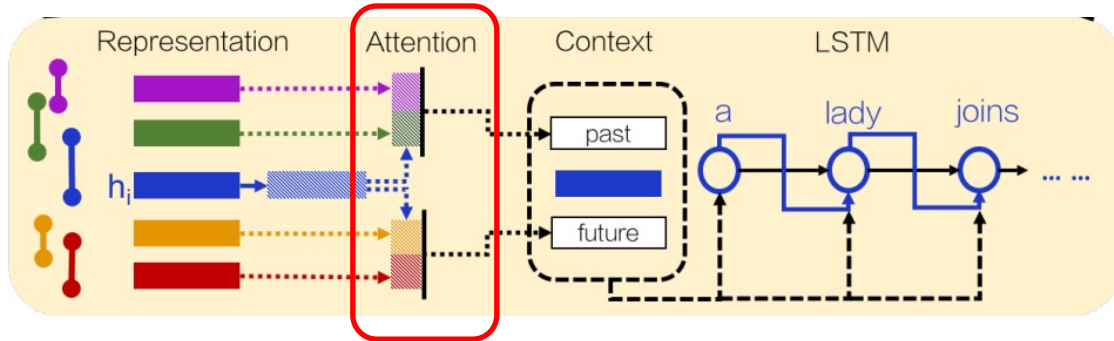
- Framework overview:

- Output/Ground truth: sentence $s_i = \{t^{\text{start}}, t^{\text{end}}, \{v_j\}\}$.
- Proposal module: find temporal proposals of interest in a video.
 - Each proposal consists of an **unique start and end time and a hidden representation**.
- Captioning module: describe the predicted proposals (e.g., higher than threshold) with captions.



Proposed Method

- Captioning module: given a predicted clips, generate captions for each clip.
 - Context module exploits neighboring events information since most events in a video are correlated.



$$h_i^{\text{past}} = \frac{1}{Z^{\text{past}}} \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} < t_i^{\text{end}}] w_j h_j$$

$$h_i^{\text{future}} = \frac{1}{Z^{\text{future}}} \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} \geq t_i^{\text{end}}] w_j h_j$$

At time i , average hidden states before and after i respectively

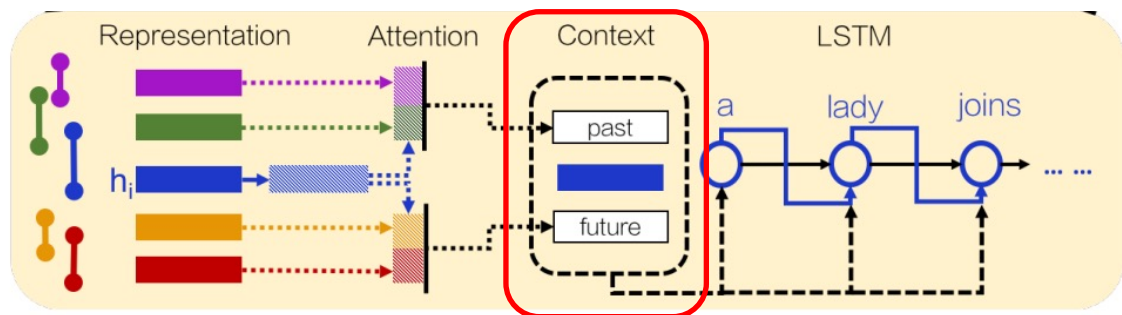
w_j is the relevant score between event i and j :

$$a_i = w_a h_i + b_a \quad \leftarrow \text{mapping } h_i$$

$$w_j = a_i h_j \quad \leftarrow \text{cosine similarity}$$

Proposed Method

- Captioning module: given a predicted clips, generate captions for each clip.
 - Context module exploits neighboring events information since most events in a video are correlated.



simple concatenate

$$h_i^{\text{past}} = \frac{1}{Z^{\text{past}}} \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} < t_i^{\text{end}}] w_j h_j$$

$$h_i^{\text{future}} = \frac{1}{Z^{\text{future}}} \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} \geq t_i^{\text{end}}] w_j h_j$$

At time i , average hidden states before and after i respectively

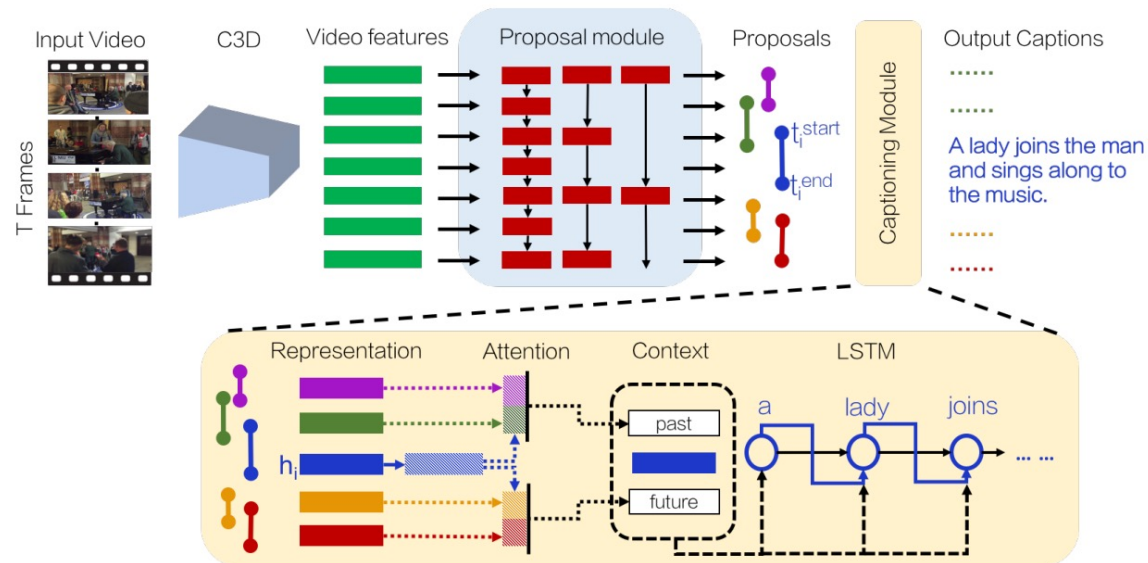
w_j is the relevant score between event i and j :

$$a_i = w_a h_i + b_a \quad \leftarrow \text{mapping } h_i$$

$$w_j = a_i h_j \quad \leftarrow \text{cosine similarity}$$

Proposed Method

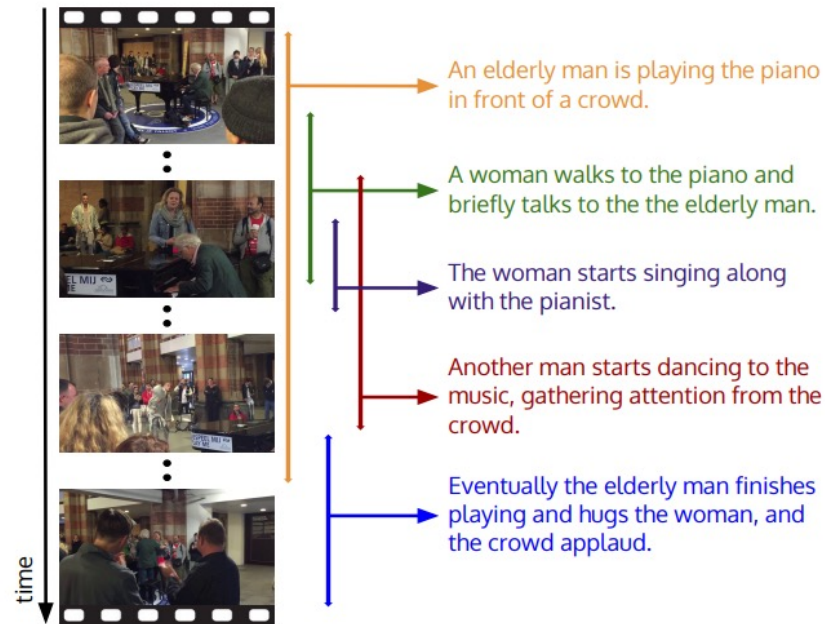
- Training loss function: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cap}} + \lambda_2 \mathcal{L}_{\text{prop}}$
 - Caption loss (\mathcal{L}_{cap}): cross-entropy loss across all words in every sentence.
 - Note that only **accurate predicted proposals (e.g., high IoU between GT)** can pass language model.
 - Proposal loss ($\mathcal{L}_{\text{prop}}$): a weighted cross-entropy loss between predicted confidences for varying proposal length.
 - Weighted cross-entropy loss can alleviate the impact from low-confident proposals.



Experiment

- ActivityNet Captions Dataset:

- ActivityNet Captions contains 20k videos taken from ActivityNet, which contains long videos.
- Each video is annotated with a series of temporally localized descriptions.



Experiment

- Evaluation:
 - Dense-captioning events.
 - Event localization.
 - Video and paragraph retrieval.

Experiment

- Dense-captioning events:

- Metrics: Bleu, METEOR and CIDEr.

- Baseline:

- LSTM-YT pools together video features to describe videos.

- S2VT encodes a video using a RNN.

- H-RNN: two-level RNN. One aims to predict sentence. The other one aims to generate hidden state for next sentence generation.

		with GT proposals					with learnt proposals						
		B@1	B@2	B@3	B@4	M	C	B@1	B@2	B@3	B@4	M	C
same model	→ LSTM-YT [49]	18.22	7.43	3.24	1.24	6.56	14.86	-	-	-	-	-	-
	→ S2VT [50]	20.35	8.99	4.60	2.62	7.85	20.97	-	-	-	-	-	-
	→ H-RNN [64]	19.46	8.78	4.34	2.53	8.02	20.18	-	-	-	-	-	-
	→ no context (ours)	20.35	8.99	4.60	2.62	7.85	20.97	12.23	3.48	2.10	0.88	3.76	12.34
	→ online-attn (ours)	21.92	9.88	5.21	3.06	8.50	22.19	15.20	5.43	2.52	1.34	4.18	14.20
	→ online (ours)	22.10	10.02	5.66	3.10	8.88	22.94	17.10	7.34	3.23	1.89	4.38	15.30
	→ full-attn (ours)	26.34	13.12	6.78	3.87	9.36	24.24	15.43	5.63	2.74	1.72	4.42	15.29
	→ full (ours)	26.45	13.48	7.12	3.98	9.46	24.56	17.95	7.69	3.86	2.20	4.82	17.29

Experiment

- Dense-captioning events:

- Metrics: Bleu, METEOR and CIDEr.

- Baseline:

- LSTM-YT pools together video features to describe videos.

- S2VT encodes a video using a RNN.

- H-RNN: two-level RNN. One aims to predict sentence. The other one aims to generate hidden state for next sentence generation.

	with GT proposals						with learnt proposals					
	B@1	B@2	B@3	B@4	M	C	B@1	B@2	B@3	B@4	M	C
LSTM-YT [49]	18.22	7.43	3.24	1.24	6.56	14.86	-	-	-	-	-	-
S2VT [50]	20.35	8.99	4.60	2.62	7.85	20.97	-	-	-	-	-	-
H-RNN [64]	19.46	8.78	4.34	2.53	8.02	20.18	-	-	-	-	-	-
no context (ours)	20.35	8.99	4.60	2.62	7.85	20.97	12.23	3.48	2.10	0.88	3.76	12.34
online—attn (ours)	21.92	9.88	5.21	3.06	8.50	22.19	15.20	5.43	2.52	1.34	4.18	14.20
online (ours)	22.10	10.02	5.66	3.10	8.88	22.94	17.10	7.34	3.23	1.89	4.38	15.30
full—attn (ours)	26.34	13.12	6.78	3.87	9.36	24.24	15.43	5.63	2.74	1.72	4.42	15.29
full (ours)	26.45	13.48	7.12	3.98	9.46	24.56	17.95	7.69	3.86	2.20	4.82	17.29

Online can only take previous hidden state

mean pooling



Experiment

- Ablation study: sentence order
 - Understand the improvement from past and future context.
 - The results are only for the first three sentences.

	B@1	B@2	B@3	B@4	M	C
no context						
1 st sen.	23.60	12.19	7.11	4.51	9.34	31.56
2 nd sen.	19.74	8.17	3.76	1.87	7.79	19.37
3 rd sen.	18.89	7.51	3.43	1.87	7.31	19.36
online						
1 st sen.	24.93	12.38	7.45	4.77	8.10	30.92
2 nd sen.	19.96	8.66	4.01	1.93	7.88	19.17
3 rd sen.	19.22	7.72	3.56	1.89	7.41	19.36
full						
1 st sen.	26.33	13.98	8.45	5.52	10.03	29.92
2 nd sen.	21.46	9.06	4.40	2.33	8.28	20.17
3 rd sen.	19.82	7.93	3.63	1.83	7.81	20.01

Experiment

- Qualitative result on Dense-captioning events:

	Ground Truth	No Context	Full Context
	Women are dancing to Arabian music and wearing Arabian skirts on a stage holding cloths and a fan.	The women continue to dance around one another and end by holding a pose and looking away.	A woman is performing a belly dancing routine in a large gymnasium while other people watch on.
	Woman is in a room in front of a mirror doing the belly dance.	A woman is seen speaking to the camera while holding up a piece of paper.	She then shows how to do it with her hair down and begins talking to the camera.
	Names of the performers are on screen.	The credits of the video are shown.	The credits of the clip are shown.

(a) Adding context can generate consistent captions.

	Ground Truth	Online Context	Full Context
	A cesar salad is ready and is served in a bowl.	The person puts a lemon over a large plate and mixes together with a.	A woman is in a kitchen talking about how to make a cake.
	Croutons are in a bowl and chopped ingredients are separated.	The person then puts a potato and in it and puts it back	A person is seen cutting up a pumpkin and laying them up in a sink.
	The man mix all the ingredients in a bowl to make the dressing, put plastic wrap as a lid.	The person then puts a lemon over it and puts dressing in it.	The person then cuts up some more ingredients into a bowl and mixes them together in the end.
	Man cuts the lettuce and in a pan put oil, with garlic and stir fry the croutons.	The person then puts a lemon over it and puts an <unk> it in.	The person then cuts up the fruit and puts them into a bowl.
	The man puts the dressing on the luttuces and adds the croutons in the bowl and mixes them all together.	The person then puts a potato in it and puts it back.	The ingredients are mixed into a bowl one at a time.

(b) Comparing *online* versus *full* model.

	Ground Truth	No Context	Full Context
	A male gymnast is on a mat in front of judges preparing to begin his routine.	A gymnast is seen standing ready and holding onto a set of uneven bars and begins performing.	He mounts the beam then does several flips and tricks.
	The boy then jumps on the beam grabbing the bars and doing several spins across the balance beam.	He does a gymnastics routine on the balance beam.	He does a gymnastics routine on the balance beam.
	He then moves into a hand stand and jumps off the bar into the floor.	He dismounts and lands on the mat.	He does a gymnastics routine on the balance beam.

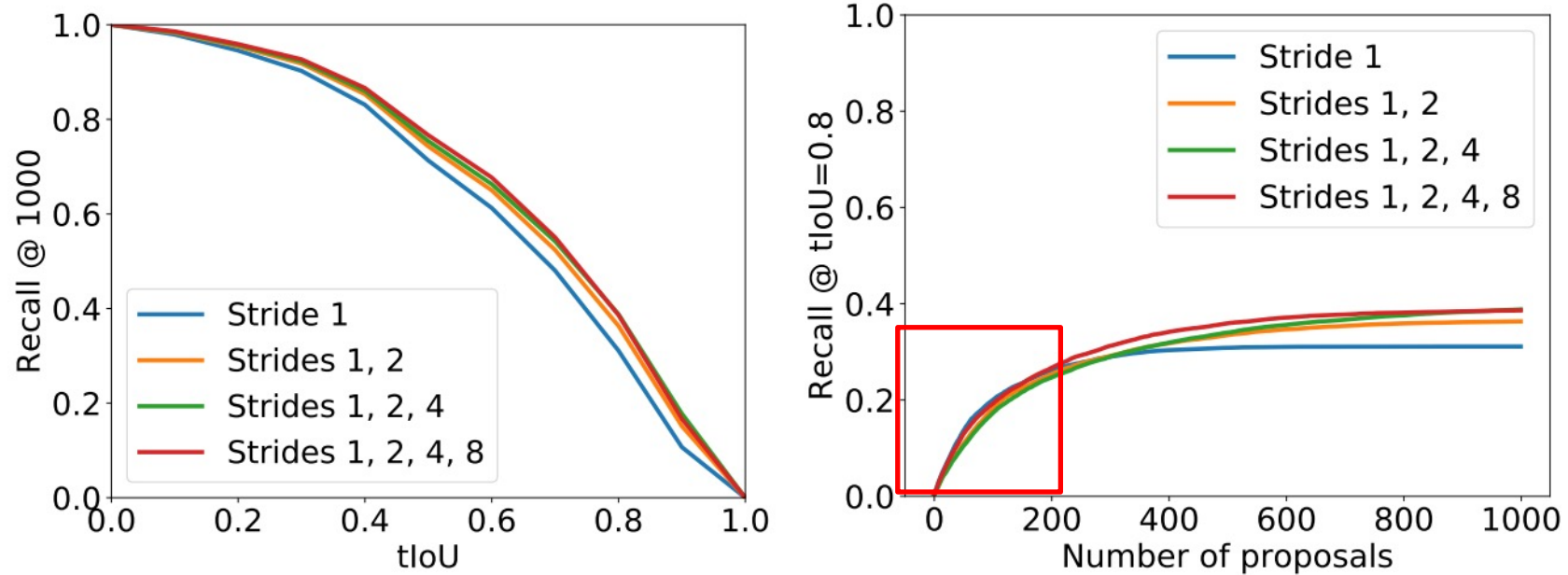
(c) Context might add more noise to rare events.

Full model can find that the vegetables are later mixed in the bow.

Full model may fail to distinguish events in a high overlap video.

Experiment

- Event localization:
 - How well models can predict the temporal location of events
 - Test with strides of 1, 2, 4 and 8. Each stride can be computed in parallel.



When there are a few proposals, the model with stride 1 performs better than any of the multi-stride versions.

Experiment

- Video and paragraph retrieval:
 - Given a set of sentences which describe different parts of a video, retrieval corresponding video, and vice versa.
 - Note that the proposed model is accessible to GT proposals and use captioning module to encode representations.

	Video retrieval				Paragraph retrieval			
	R@1	R@5	R@50	Med. rank	R@1	R@5	R@50	Med. rank
LSTM-YT [49]	0.00	0.04	0.24	102	0.00	0.07	0.38	98
no context [50]	0.05	0.14	0.32	78	0.07	0.18	0.45	56
online (ours)	0.10	0.32	0.60	36	0.17	0.34	0.70	33
full (ours)	0.14	0.32	0.65	34	0.18	0.36	0.74	32

Conclusion

- This paper incorporate event proposal module to find proposals of interest that can generate more detail captions.
- Context module somehow learns long-term information.
- Proposed ActivityNet Caption is a good benchmark including clip description for a long video.

Discussion

- Can attention pooling for hidden states learn long-term information?
- How do these two modules benefit each one?