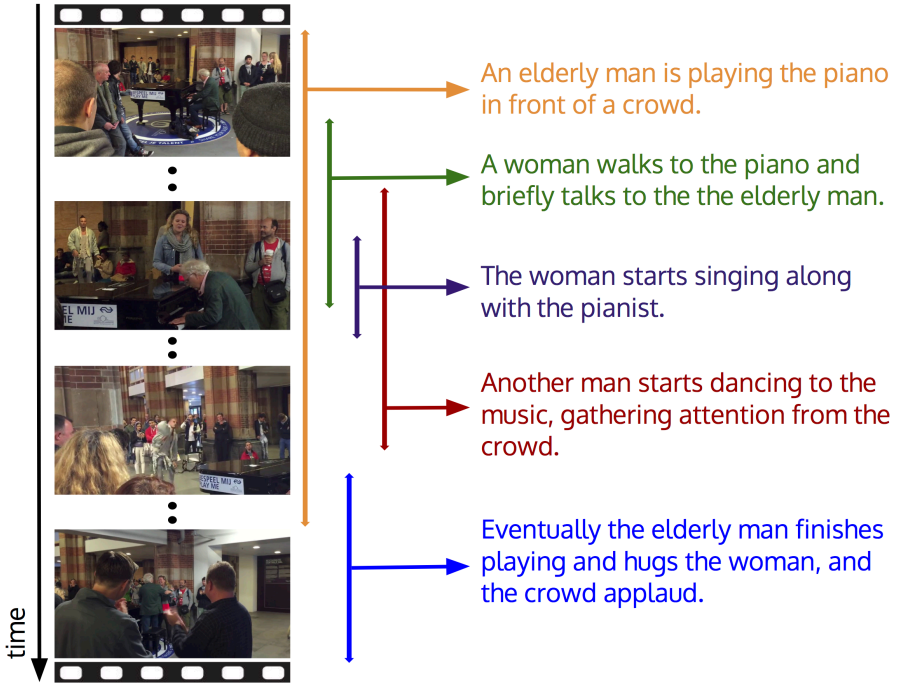
Dense-captioning Events in Videos

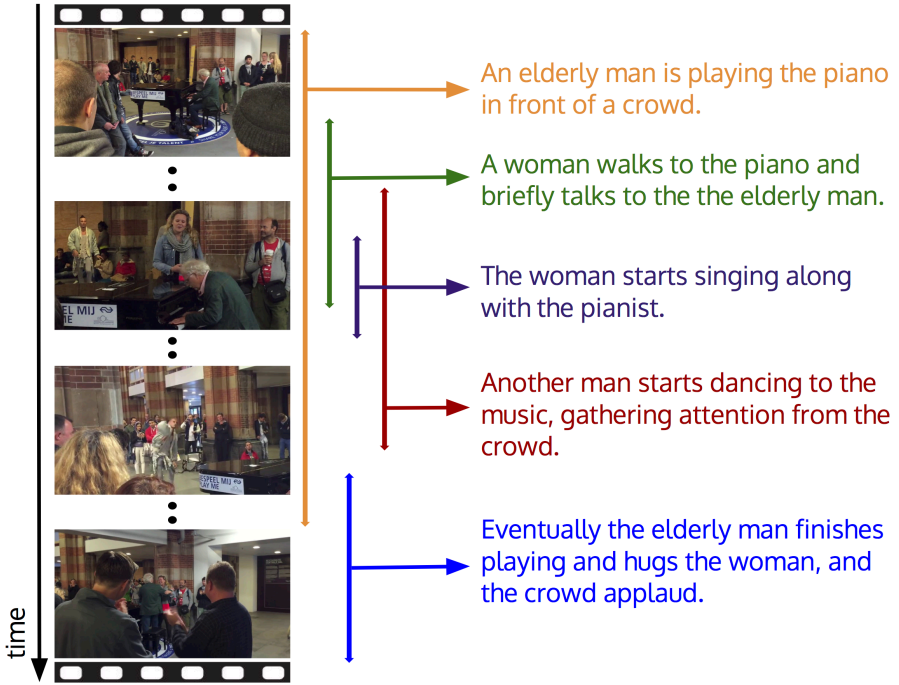
Parisa Asgharzadeh & Candice Tian

Introduction



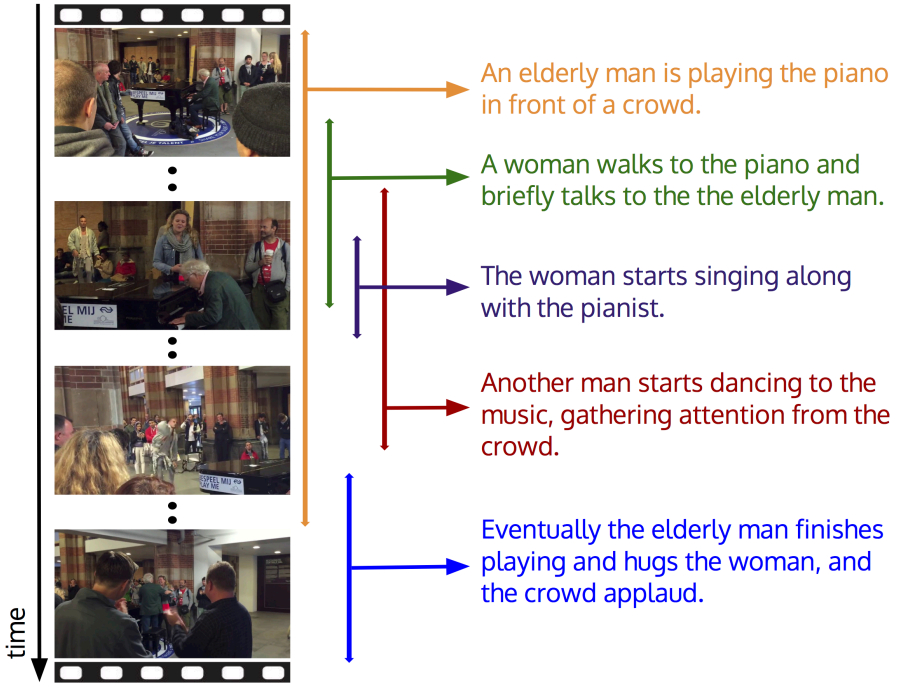
Introduction

- Previous work to describe videos first started with labelling them with a predefined category. playing piano or dancing



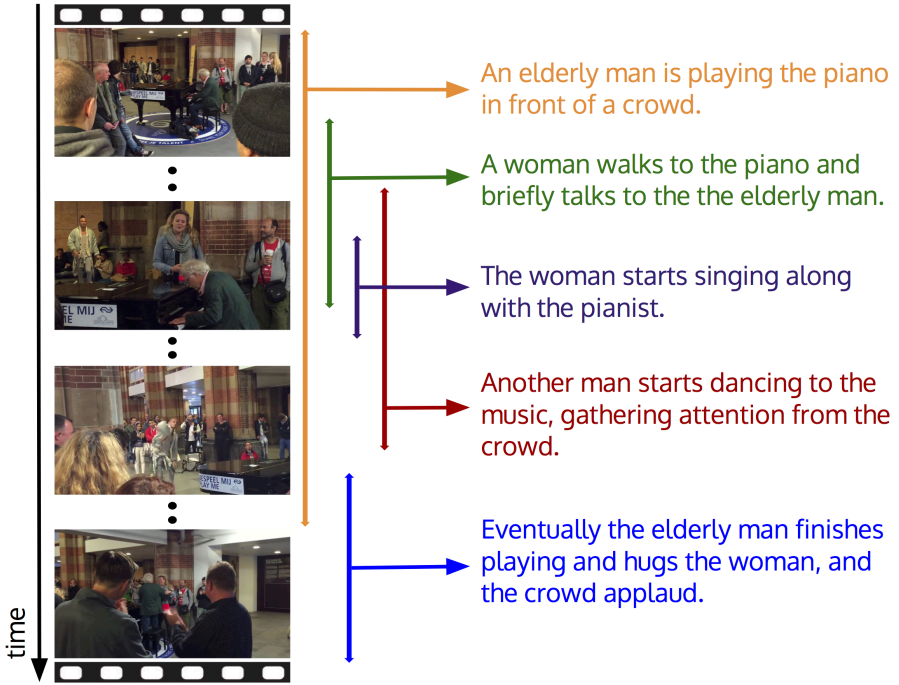
Introduction

- Previous work to describe videos first started with labelling them with a predefined category. playing piano or dancing
- What's missing?



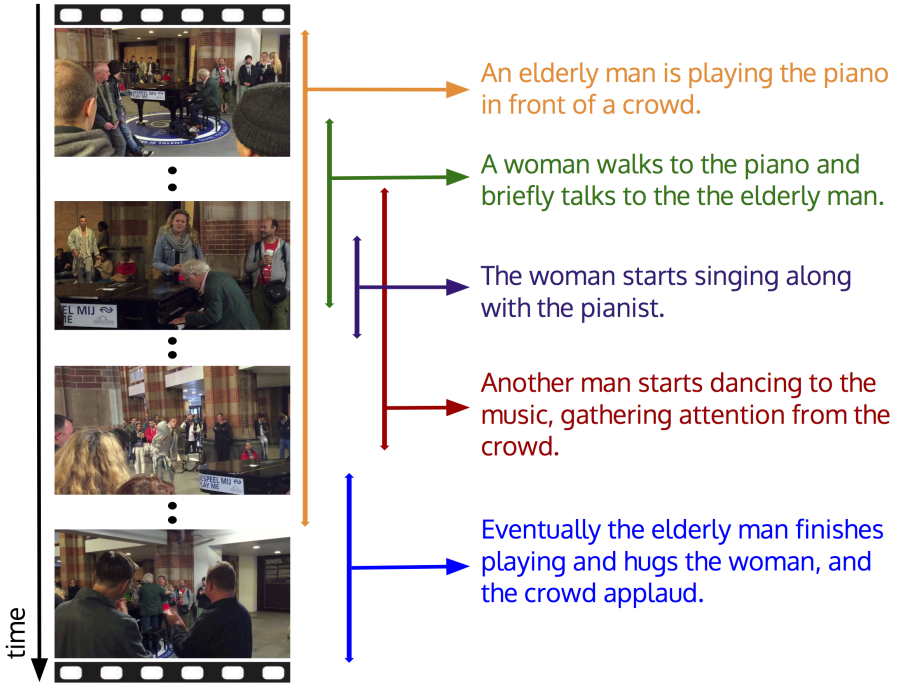
Introduction

- Previous work to describe videos first started with labelling them with a predefined category. playing piano or dancing
- What's missing?
- Detail



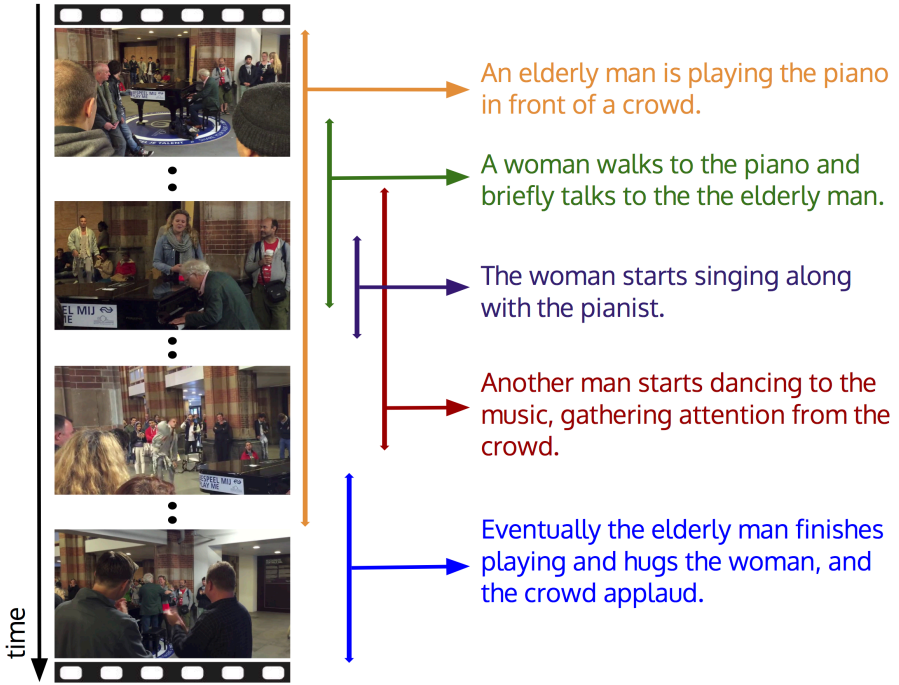
Introduction

- Previous work to describe videos first started with labelling them with a predefined category. playing piano or dancing
- What's missing?
- Detail
- Any solutions?!

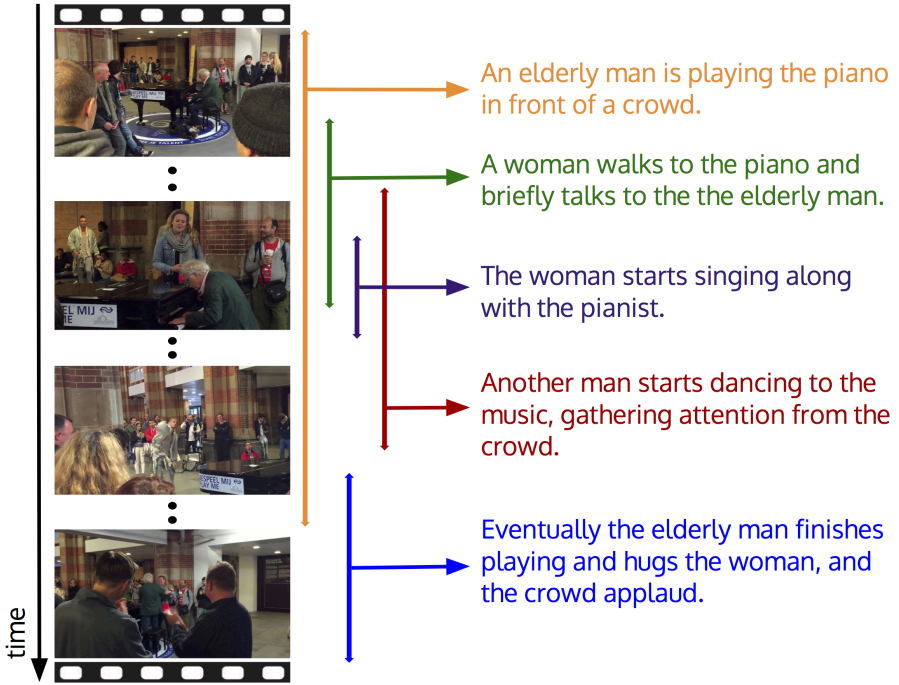


Introduction

- Previous work to describe videos first started with labelling them with a predefined category. playing piano or dancing
- What's missing?
- Detail
- Any solutions?!
- Explaining video semantics using sentence descriptions

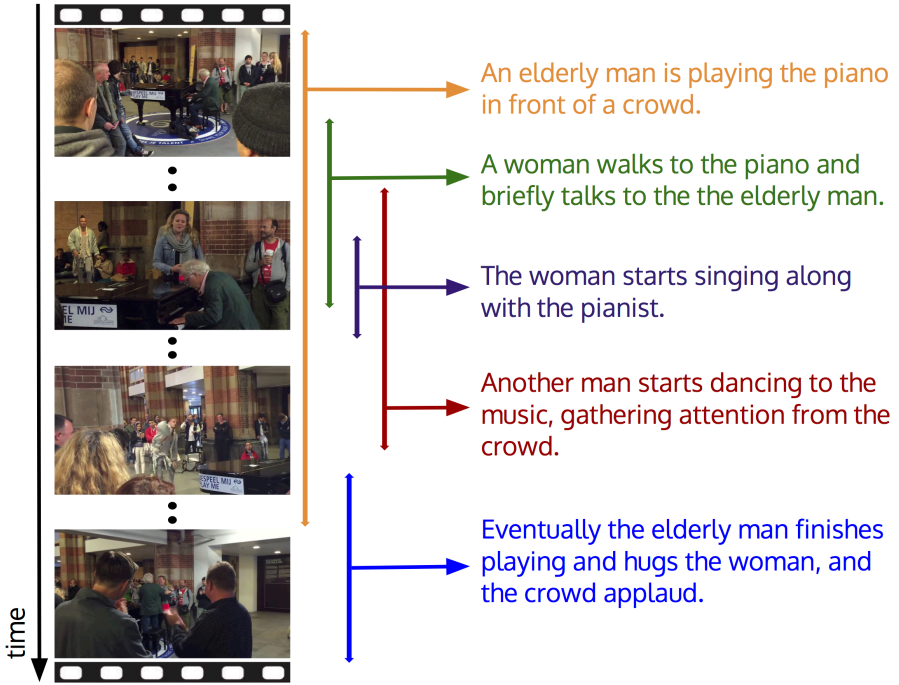


Introduction-cont’d



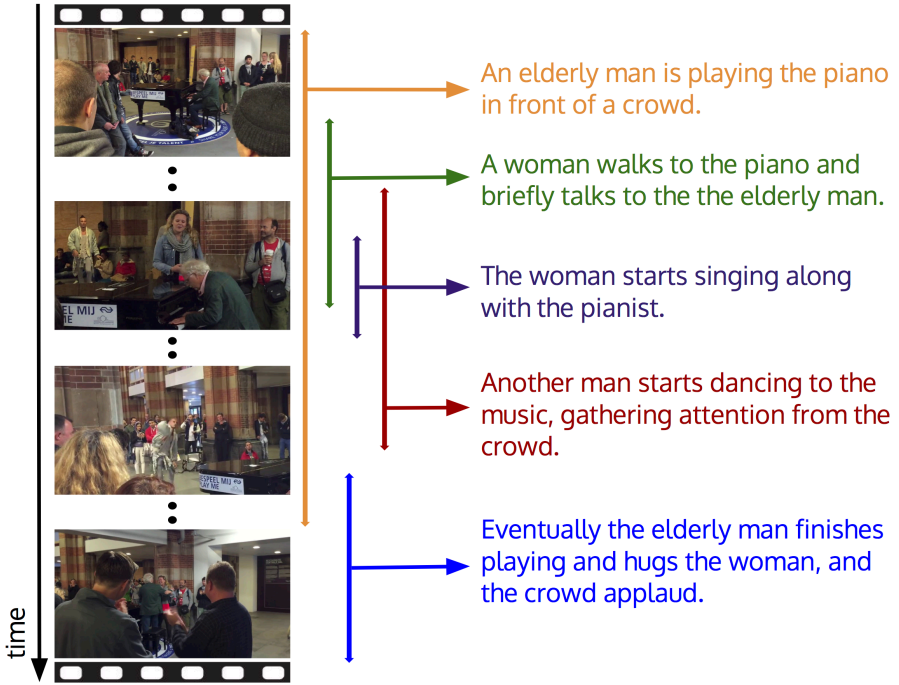
Introduction-cont'd

- So what's the problem?



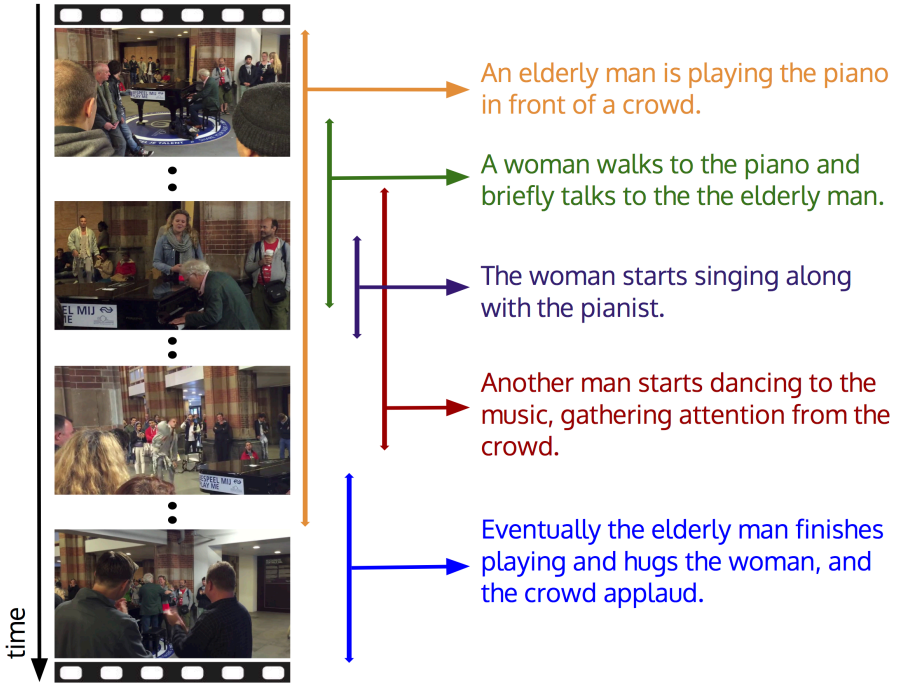
Introduction-cont'd

- So what's the problem?
- It fails to recognize and articulate all the other events in the video



Introduction-cont'd

- So what's the problem?
- It fails to recognize and articulate all the other events in the video
- Dense-captioning events



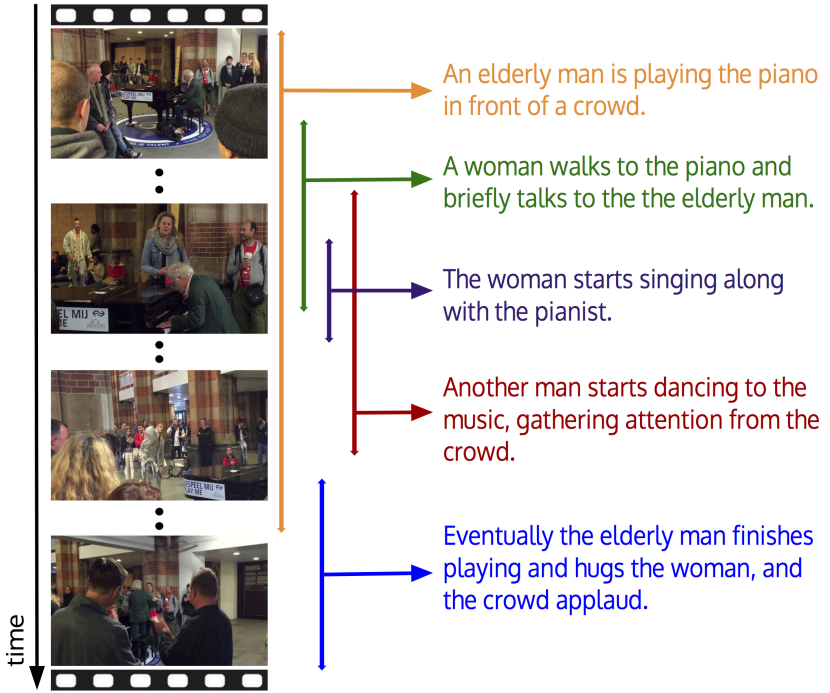
Dense-captioning events vs. dense-image captioning

Dense-captioning events vs. dense-image captioning

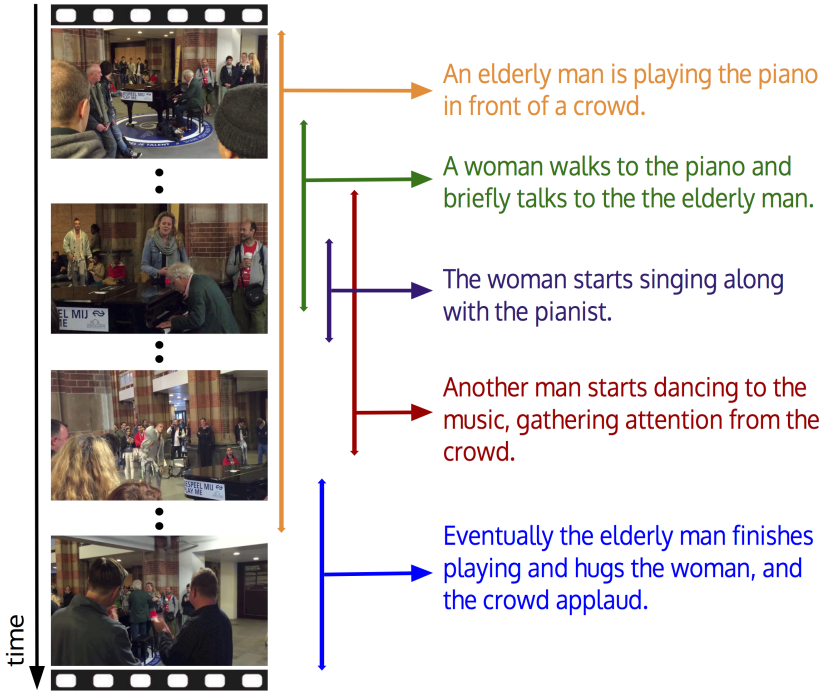
- Event localization in time vs. Localization of regions in space

Dense-captioning events vs. dense-image captioning

- Event localization in time vs. Localization of regions in space
- Events range across multiple time scales and can even overlap
    - Requires encoding short as well as long sequences of video frames to propose events
    - Previous works used mean-pooling or a recurrent neural network (RNN)
    - Vanishing gradients in long video sequences
    - Generating action proposals to multi-scale detection of events.


time
An elderly man is playing the piano in front of a crowd.
A woman walks to the piano and briefly talks to the the elderly man.
The woman starts singing along with the pianist.
Another man starts dancing to the music, gathering attention from the crowd.
Eventually the elderly man finishes playing and hugs the woman, and the crowd applaud.

- The events in a given video are usually related to one another.
  - Use context from surrounding events to caption each event.
  - One solution : describe videos with multiple sentences
  - Problem: generates sentences for sequentially occurring events and highly correlated to the objects in the video
  - Doesn't not generalize to "open" domain videos
  - Solution: using context will solve the problem



Related Work: Action Recognition

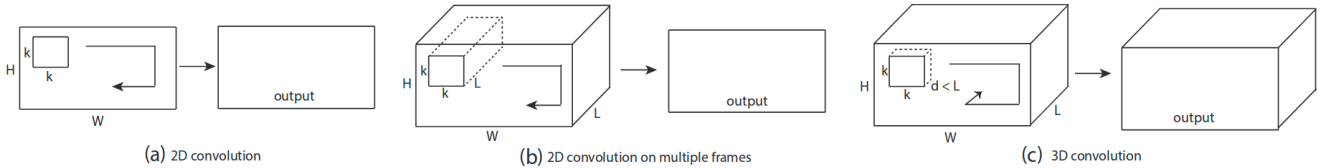


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

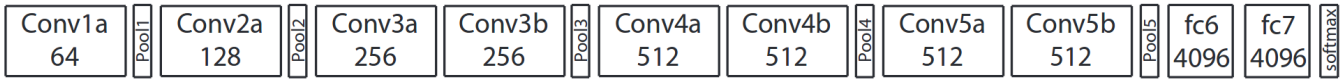


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

Credit: *Learning Spatiotemporal Features with 3D Convolutional Networks*, Tran et al

Related Work: Action Recognition

3D Convolution for extracting features from subsequent frames:

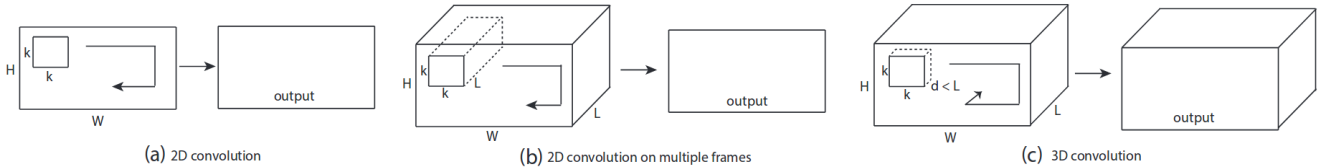


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

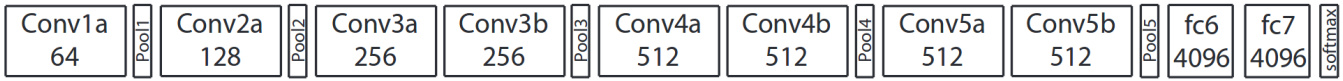


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

Credit: *Learning Spatiotemporal Features with 3D Convolutional Networks*, Tran et al

Related Work: Action Recognition

3D Convolution for extracting features from subsequent frames:

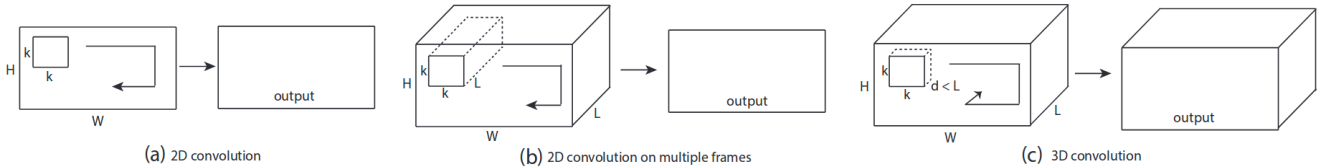


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

CNN built using 3D convolution for action recognition:

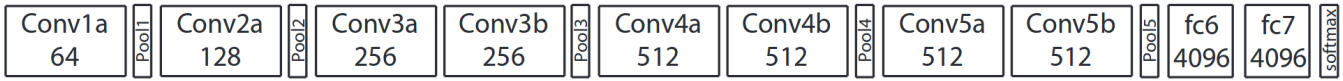


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

Credit: *Learning Spatiotemporal Features with 3D Convolutional Networks*, Tran et al

Related Work: Action Recognition

Method	Accuracy (%)
Imagenet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [18]	65.4
Spatial stream network [36]	72.6
LRCN [6]	71.1
LSTM composite model [39]	75.8
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	**85.2**
iDT w/ Fisher vector [31]	87.9
Temporal stream network [36]	83.7
Two-stream networks [36]	88.0
LRCN [6]	82.9
LSTM composite model [39]	84.3
Conv. pooling on long clips [29]	88.2
LSTM on long clips [29]	88.6
Multi-skip feature stacking [25]	89.1
C3D (3 nets) + iDT + linear SVM	**90.4**

Table 3. **Action recognition results on UCF101**. C3D compared with baselines and current state-of-the-art methods. Top: simple features with linear SVM; Middle: methods taking only RGB frames as inputs; Bottom: methods using multiple feature combinations.

Credit: *Learning Spatiotemporal Features with 3D Convolutional Networks*, Tran et al

Related Work: Temporal Action Proposal

- Problem Formulation:

A set of video frames $\Longrightarrow$ Predicted Segments $S = \{s_i\}_{i=1}^{K}$
and their confidence scores $C = \{c_i\}_{i=1}^{K}$
Which will match with the locations of actions $A = \{a_i\}_{i=1}^{M}$ in the sequence.

- Target Loss Function:

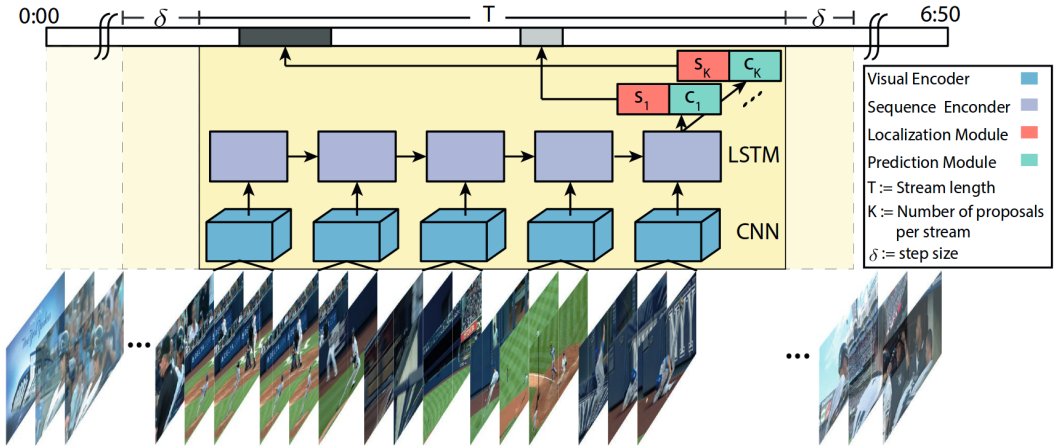
$$(\mathbf{x}^*, \theta^*) = \underset{\mathbf{x},\theta}{\operatorname{argmin}} \; \alpha \mathcal{L}_{\text{match}}(\mathbf{x}, S(\theta), A) + \mathcal{L}_{\text{conf}}(\mathbf{x}, C(\theta))$$

$$\text{s.t.} \quad x_{ij} \in \{0, 1\}, \quad \sum_i x_{ij} = 1$$

Euclidean Distance

Binary Cross-Entropy

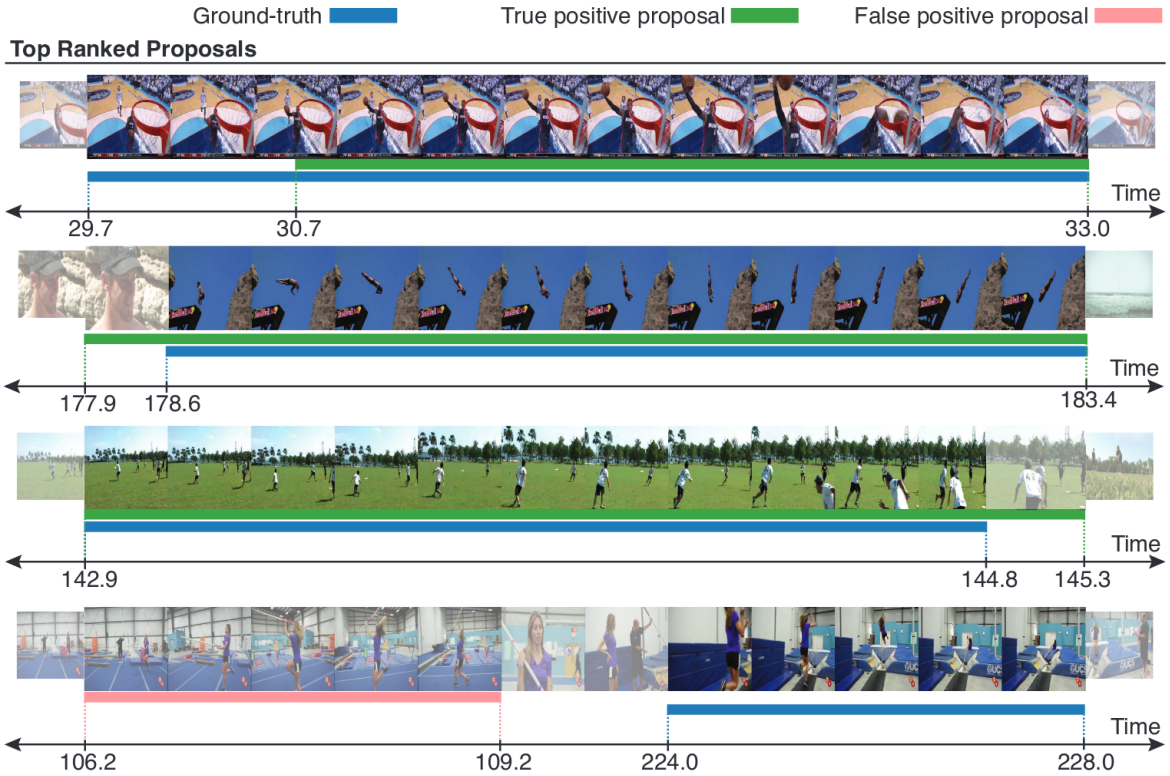
Related Work: Temporal Action Proposal



Localization Module: Predict the location of K proposals inside the stream based on a linear combination of the last state in the sequence encoder. In this way the model can output segments of different lengths in one pass.

Credit: *Daps: Deep action Proposals for action understanding*, Escorcia et al.

Related Work: Temporal Action Proposal



Related Work: Video captioning

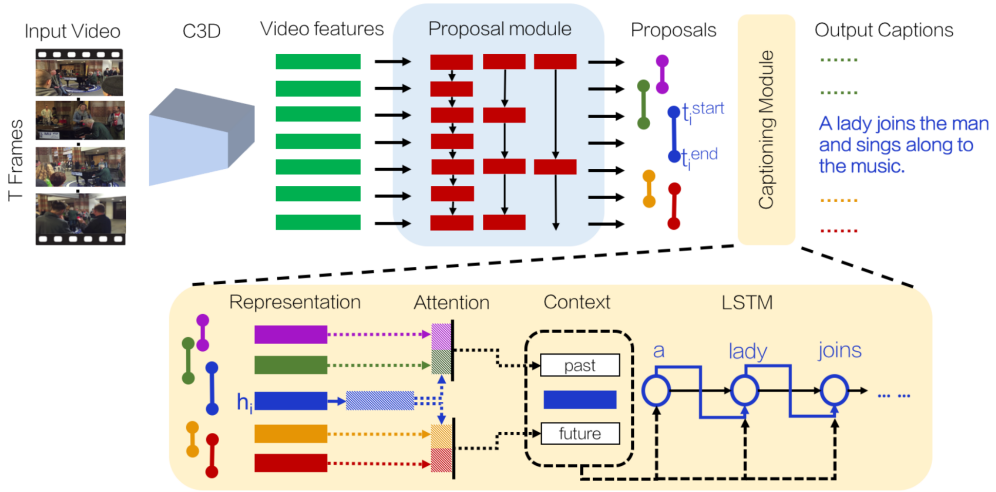
Early Solutions: Mean pooled video frame features and used a pipeline inspired by the success of image captioning

Problems: Only works for short video clips with only one major event

Some solutions:

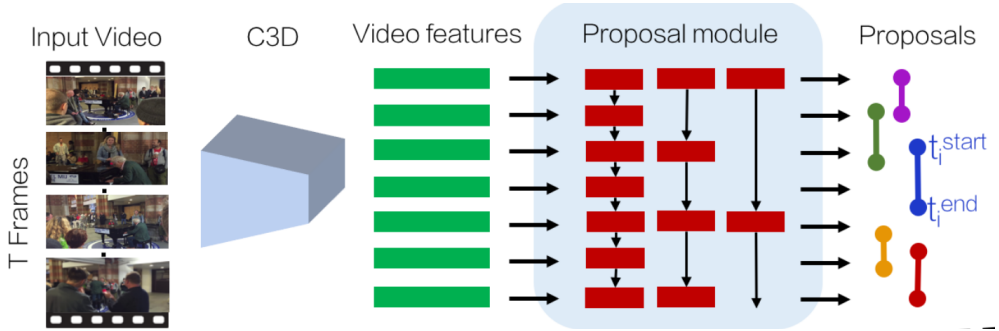
- Hierarchical RNN (Sentences generated are not localized in time. The dataset only contain non-overlapping sequential events)
- Attention mechanism

Model: Overview

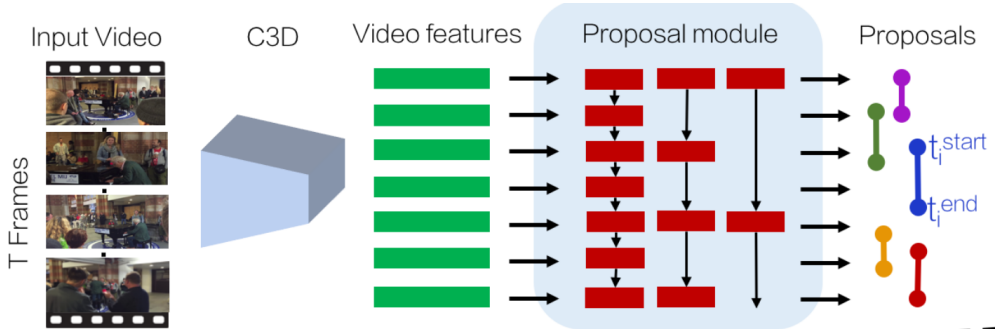


- Input of the model:
  A sequence of video frames $v = \{v_t\}$
  where $t \in 0, ..., T-1$ indexes the frames in temporal order.
- Output of proposal module: $P = \{(t_i^{\text{start}}, t_i^{\text{end}}, \text{score}_i, h_i)\}$
- Output of the model: $s_i = (t^{\text{start}}, t^{\text{end}}, \{v_j\})$

Model: Event Proposal Module

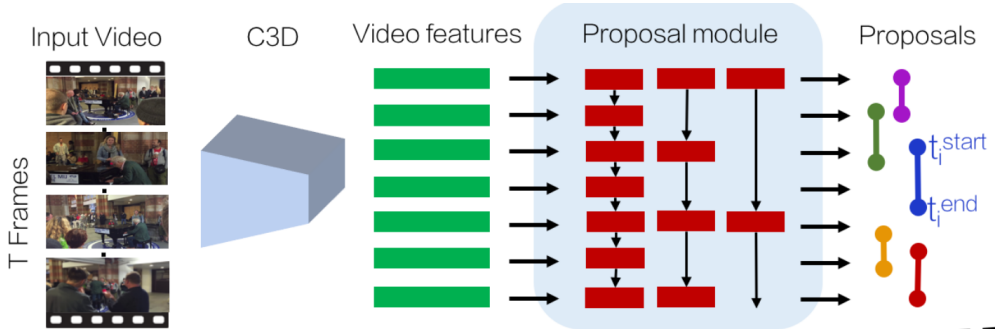


Model: Event Proposal Module



Main difference from traditional DAPs:

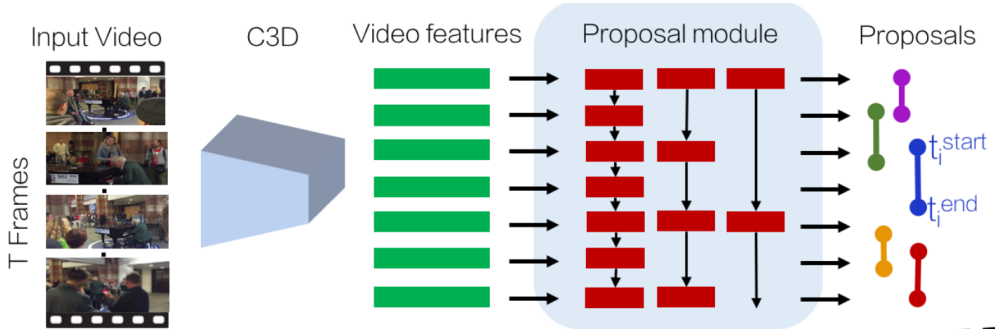
Model: Event Proposal Module



Main difference from traditional DAPs:

- Before feeding into DAPs, sample the video features at different strides (1, 2, 4, and 8 in the paper). The longer strides are able to capture longer events.

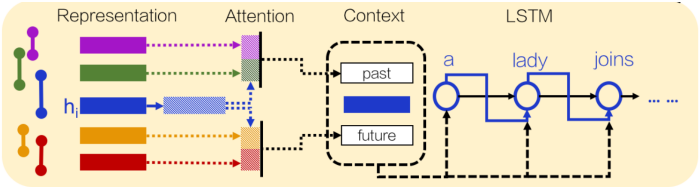
Model: Event Proposal Module



Main difference from traditional DAPs:

- Before feeding into DAPs, sample the video features at different strides (1, 2, 4, and 8 in the paper). The longer strides are able to capture longer events.
- Traditional DAPs uses non-maximum suppression to eliminate overlapping outputs. Here overlapping outputs are kept separately and treated as individual events

Model: Captioning with context



For an event from the proposal module, with hidden representation h_i and start and end times of $[t_i^{\text{start}}, t_i^{\text{end}}]$, we calculate the past and future context representation as follows:

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

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

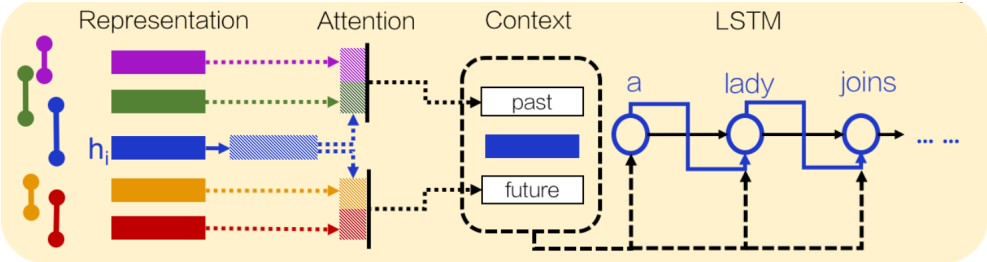
h_j : hidden representation of the other proposed events in the video
w_j : weight determining how relevant event j is to event i.
Z : normalization calculated as $Z^{\text{past}} = \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} < t_i^{\text{end}}]$

Model: Captioning with context

$$a_i = w_a h_i + b_a$$
$$w_j = a_i h_j$$
$$\Longrightarrow \quad w_j = w_a h_i h_j + b_a h_j$$

a_i : attention vector calculated from learnt weights w_a and bias b_a

The concatenation of $(h_i^{\text{past}}, h_i, h_i^{\text{future}})$ is then fed as the input to the captioning LSTM. Each LSTM is initialized to have 2 layers with 512 dimensional hidden representation



Implementation Details

- Loss Function:

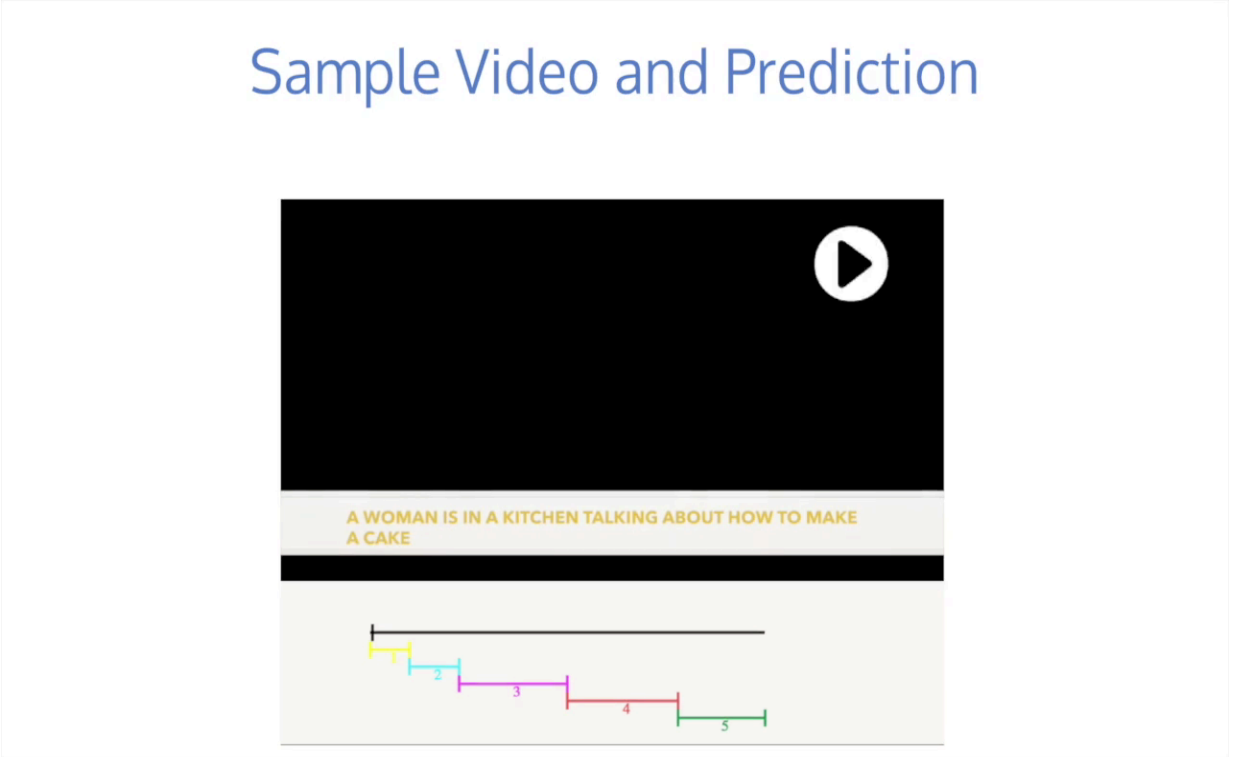
Use two separate losses to train proposal model and captioning model.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cap}} + \lambda_2 \mathcal{L}_{\text{prop}}$$

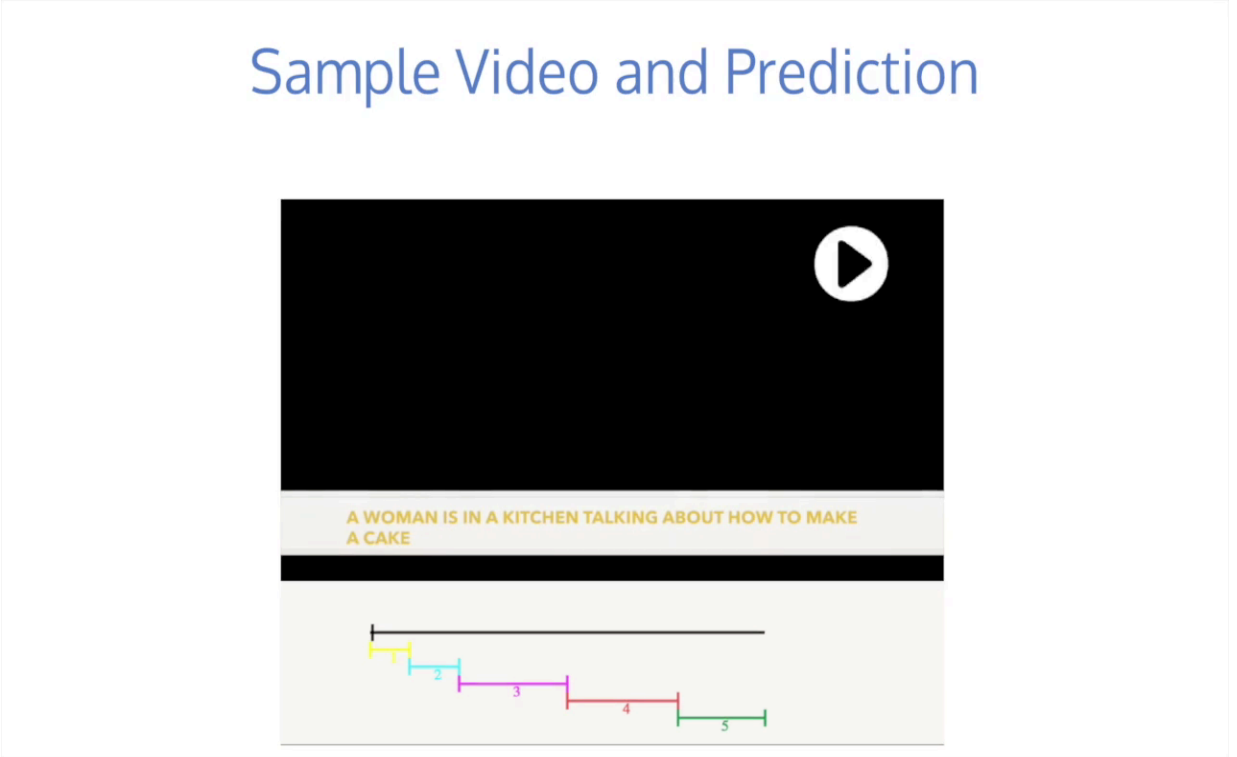
The authors weight the contribution of the captioning loss with $\lambda_1 = 1.0$ and the proposal loss with $\lambda_2 = 0.1$

- The full model is trained by alternating between training the language model and the proposal module every 500 iterations. Training batch size is set to 1. One mini-batch runs in approximately 15.85 ms on a Titan X GPU and the whole model takes two days to converge.

The ActivityNet Captions dataset

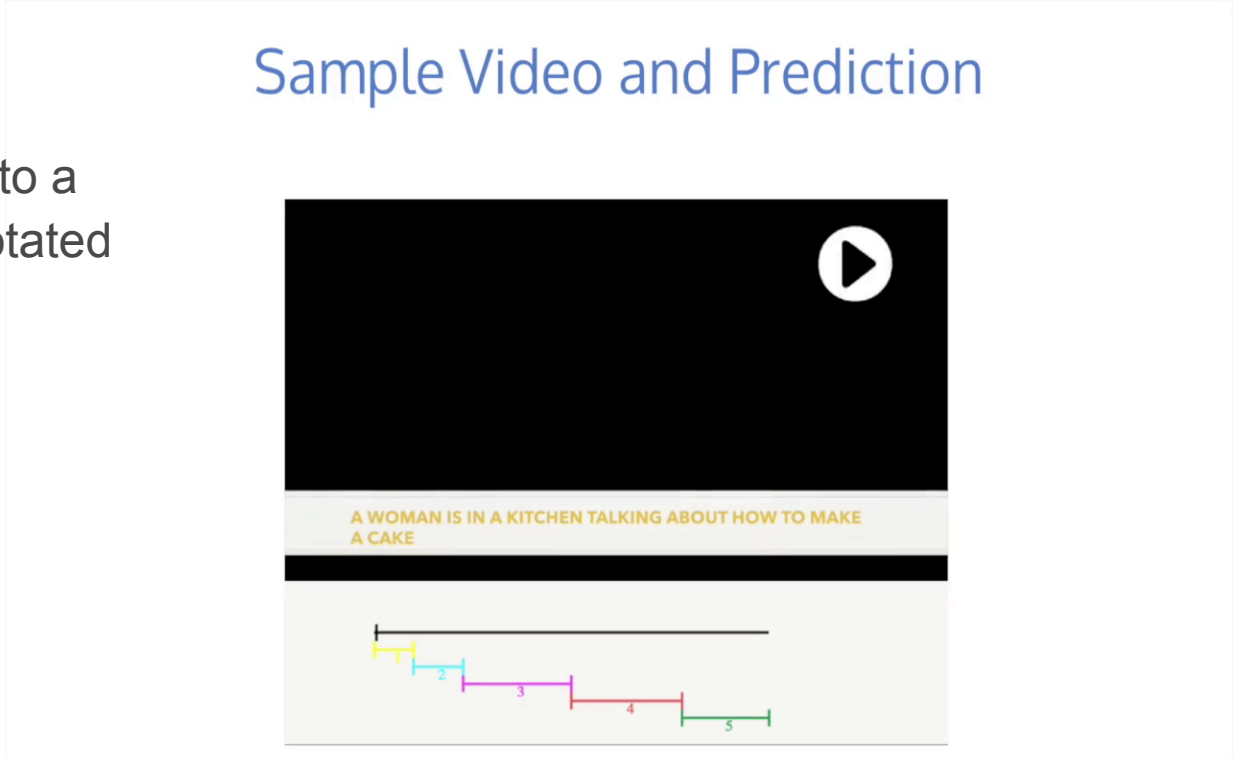


The ActivityNet Captions dataset



The ActivityNet Captions dataset

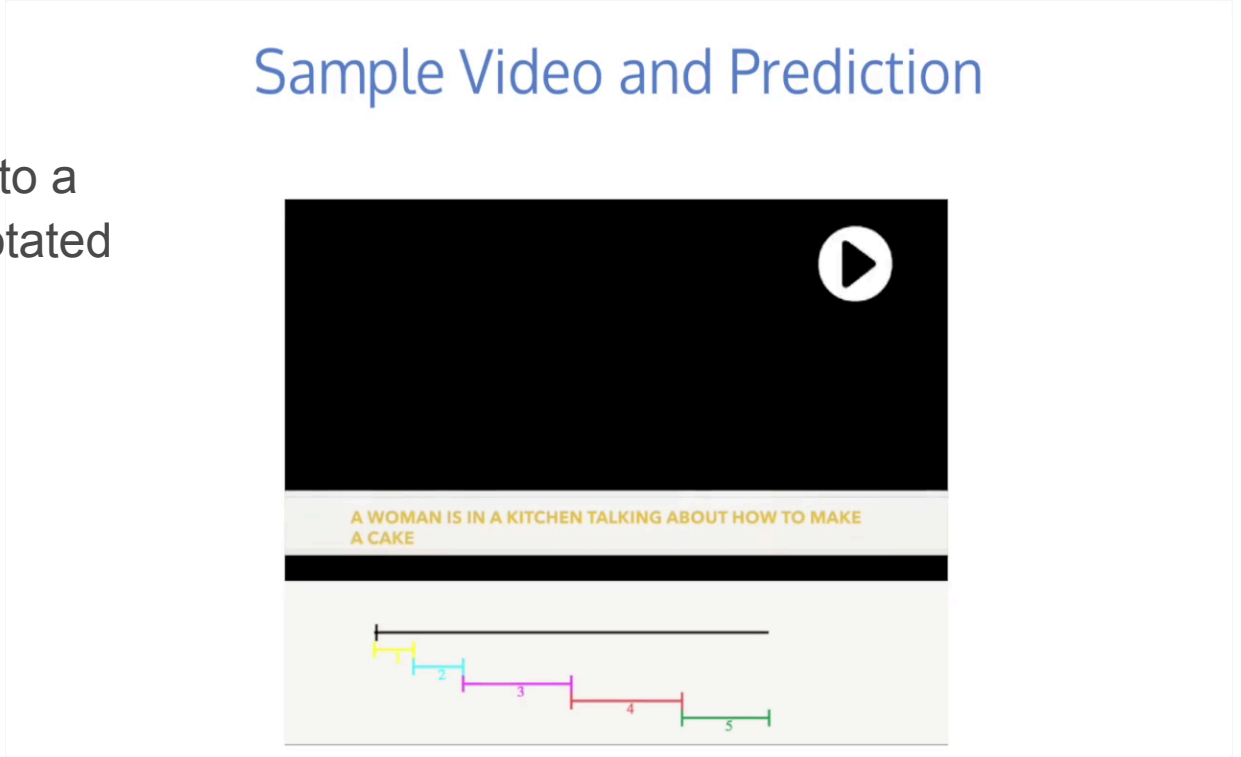
The activitynet captions dataset connects videos to a series of temporally annotated sentence descriptions.



The ActivityNet Captions dataset

The activitynet captions dataset connects videos to a series of temporally annotated sentence descriptions.

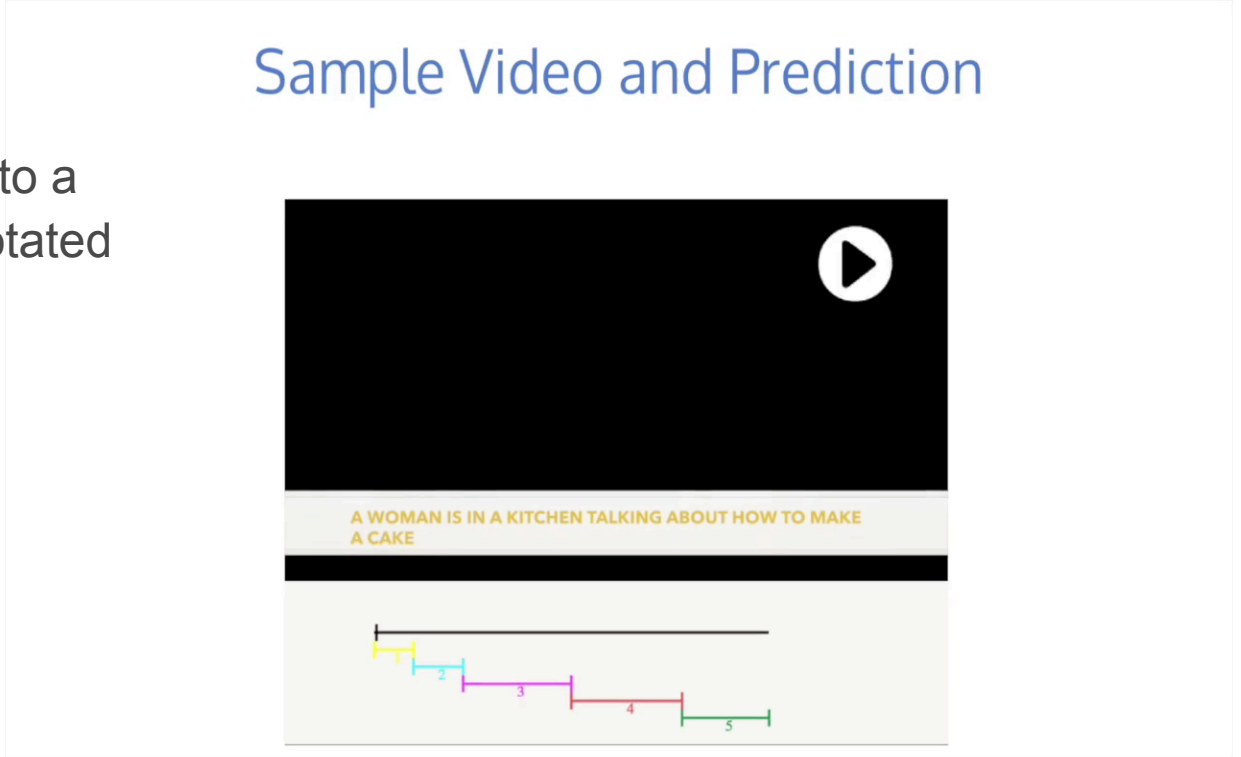
- 20k videos



The ActivityNet Captions dataset

The activitynet captions dataset connects videos to a series of temporally annotated sentence descriptions.

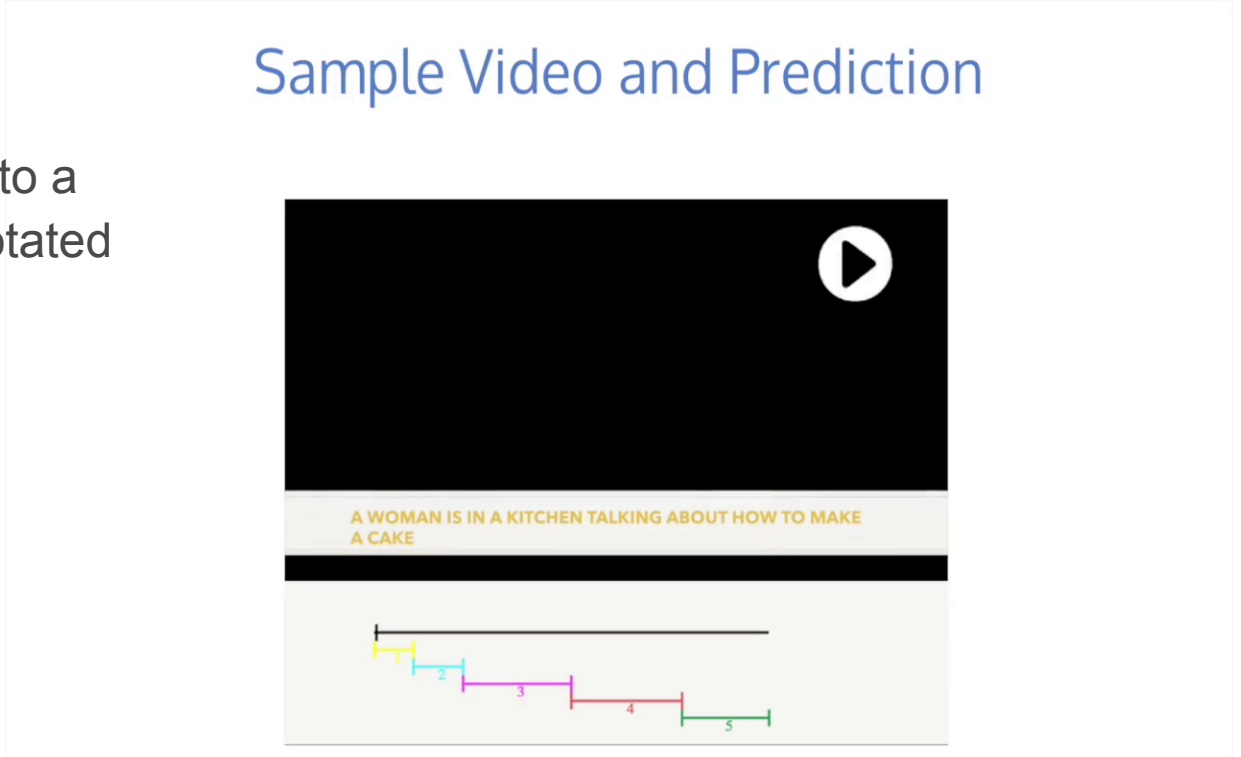
- 20k videos
- Avg. 3.65 temporally localized sentences



The ActivityNet Captions dataset

The activitynet captions dataset connects videos to a series of temporally annotated sentence descriptions.

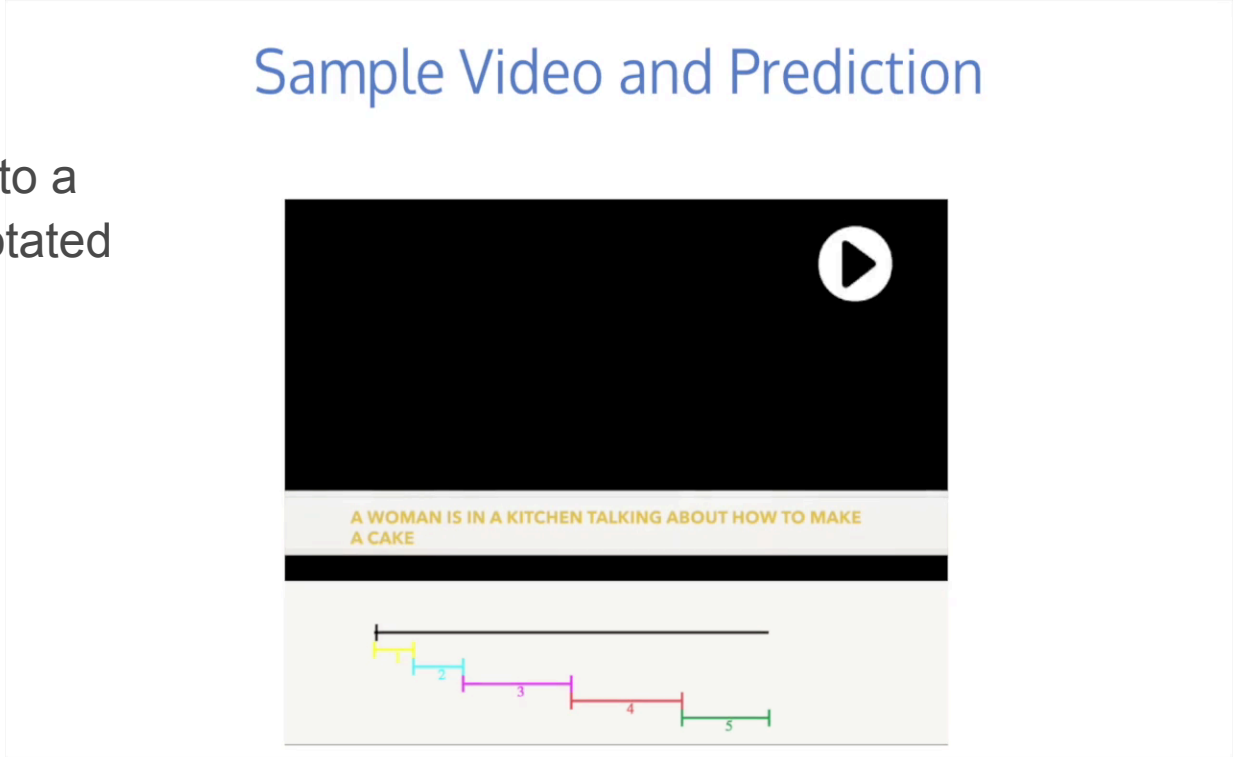
- 20k videos
- Avg. 3.65 temporally localized sentences
- Avg. 13.48 words per sentences



The ActivityNet Captions dataset

The activitynet captions dataset connects videos to a series of temporally annotated sentence descriptions.

- 20k videos
- Avg. 3.65 temporally localized sentences
- Avg. 13.48 words per sentences



Experiments and Results

Experiments and Results

- Evaluation is done by multi-captioning.

Experiments and Results

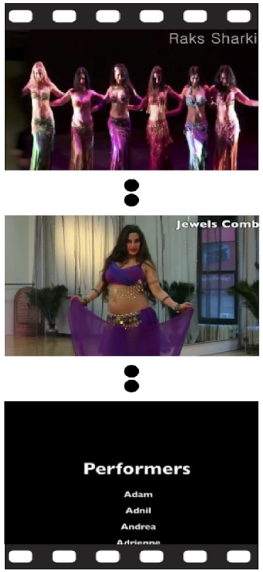
- Evaluation is done by multi-captioning.
- Activitynet captions dataset is used to test model.

Experiments and Results

- Evaluation is done by multi-captioning.
- Activitynet captions dataset is used to test model.
- Baseline results on two additional tasks that are possible:
    - Localization
    - Retrieval

Qualitative Results

Adding context can generate consistent captions.



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 dancing around one another and end by holding a pose and looking away.	A woman is seen 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.

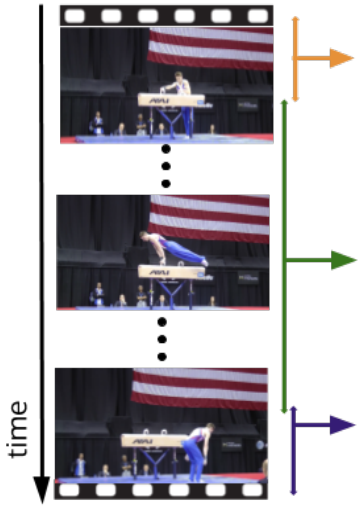
Qualitative Results

Compare online versus full model

GT	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 lettuces 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.

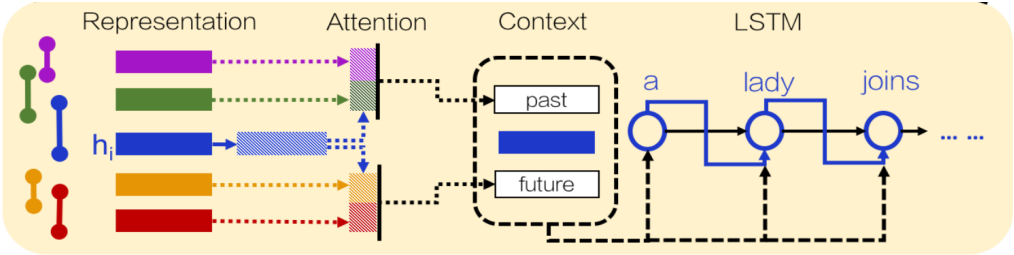
Qualitative Results

Context might add more noise to rare events.



GT	No Context	With 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.

Quantitative Results

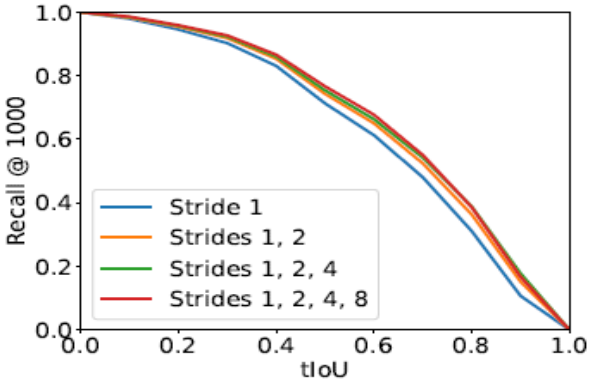


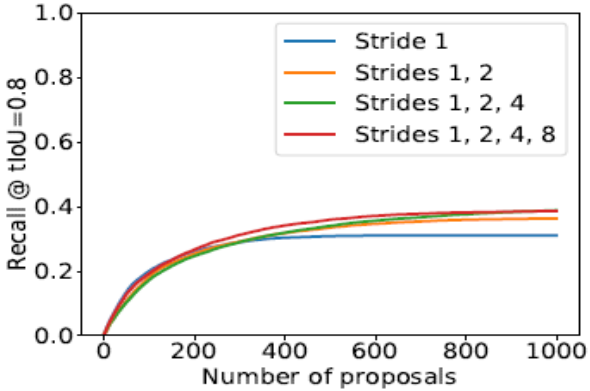
	with GT proposals						with learnt proposals					
	B@1	B@2	B@3	B@4	M	C	B@1	B@2	B@3	B@4	M	C
LSTM-YT	18.22	7.43	3.24	1.24	6.56	14.86	-	-	-	-	-	-
S2VT	20.35	8.99	4.60	2.62	7.85	20.97	-	-	-	-	-	-
H-RNN	19.46	8.78	4.34	2.53	8.02	20.18	-	-	-	-	-	-
no context (ours)	20.35	8.99	4.60	2.62	7.85	20.97	12.23	3.48	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**

Quantitative Results

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

Quantitative Results





	Video retrieval				Paragraph retrieval			
	R@1	R@5	R@50	Med. rank	R@1	R@5	R@50	Med. rank
LSTM-YT	0.00	0.04	0.24	102	0.00	0.07	0.38	98
no context	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

Conclusion

- Dense-captioning events
  - Events can occur within a second or last up to minutes
  - Events in a video are related to one another.

Conclusion

- Dense-captioning events
  - Events can occur within a second or last up to minutes
  - Events in a video are related to one another.
- Proposed model combines a proposal module with a new captioning module
  - The proposal module
  - The captioning module

Conclusion

- Dense-captioning events
  - Events can occur within a second or last up to minutes
  - Events in a video are related to one another.
- Proposed model combines a proposal module with a new captioning module
  - The proposal module
  - The captioning module
- Compare variants of the model and show that context does indeed improve captioning.

Conclusion

- Dense-captioning events
  - Events can occur within a second or last up to minutes
  - Events in a video are related to one another.
- Proposed model combines a proposal module with a new captioning module
  - The proposal module
  - The captioning module
- Compare variants of the model and show that context does indeed improve captioning.
- Release a new dataset for dense-captioning events: activitynet captions.

Pros & Cons and Possible Extensions

Pros:

- Through sampling features at different strides and coming up with context incorporated feature, effectively solved the problem of captioning overlapping events of different lengths in a long video

Cons and Possible Extensions:

- Use more accurate models to extract video features
- Use more attention mechanisms

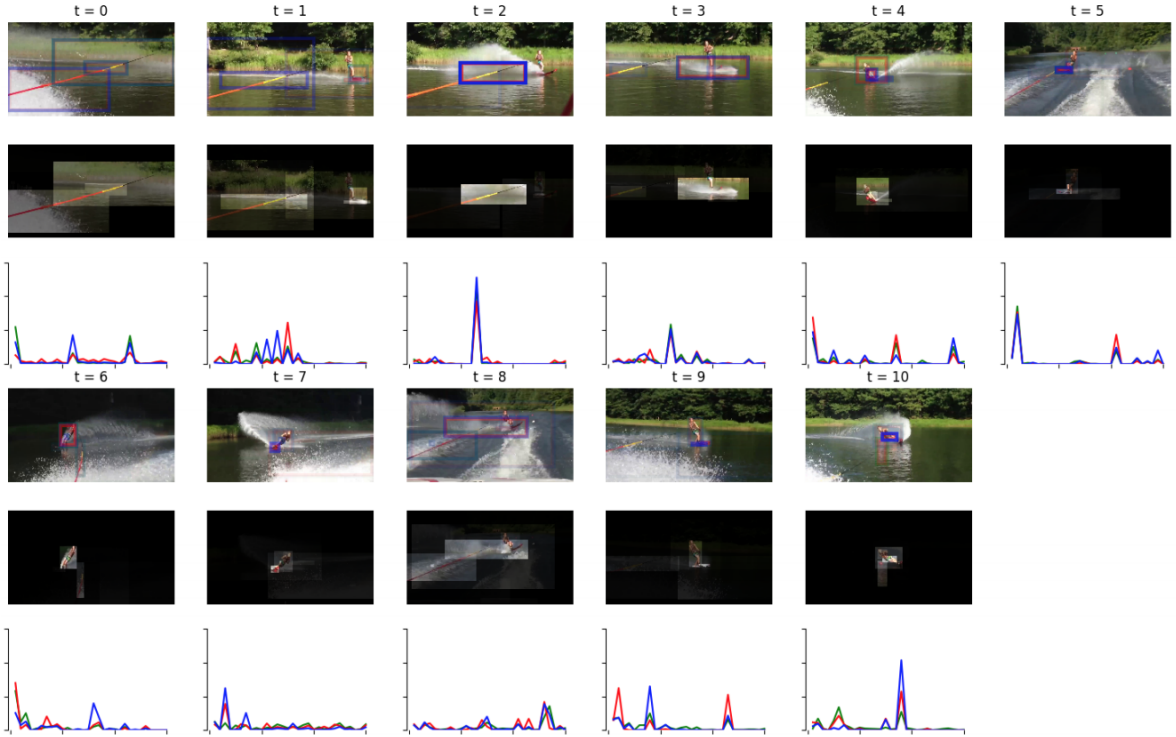


Figure 7. **Water skiing**: Our SINet is able to identify several object relationships and reasons these interactions through time: (1) the rope above the water (2) the wakeboard on the water (3) human riding on the wakeboard (4) rope connecting to the person on the wakeboard. From the distribution of three different attention weights (red, green, blue), we can also see that the proposed attention method not only is able to select objects with different inter-relationships but also can use a common object to discover different relationships around that object when needed. We observed that our method tend to explore the whole scene at the beginning of the video, and focus on new information that is different from the past. For example, while video frame at first few frames are similar, the model focus on different aspect of the visual representation.

Credit: *Attend and Interact: Higher-Order Object Interactions for Video Understanding*, Ma et al

Table 1. Prediction accuracy on the Kinetics validation set. All of our results use only RGB videos sampled at 1 FPS. Maximum number of objects per frame is set to be 30.

Method	Top-1	Top-5
I3D[1] (25 FPS) [6] (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS) [4]	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
Obj feature (mean-pooling)	72.2	90.2
Img + obj feature (mean-pooling)	73.1	91.1
SINet (α-attention)	73.9	91.5
SINet (dot-product attention)	**74.2**	**91.7**

Table 3. METEOR, ROUGE-L, CIDEr-D, and BLEU@N scores on the ActivityNet Captions test and validation set. All methods use ground truth proposal except LSTM-A_3 [55]. Our results with ResNeXt spatial features use videos sampled at maximum 1 FPS only.

Method	B@1	B@2	B@3	B@4	ROUGE-L	METEOR	CIDEr-D
Test set							
LSTM-YT [50] (C3D)	18.22	7.43	3.24	1.24	-	6.56	14.86
S2VT [49] (C3D)	20.35	8.99	4.60	2.62	-	7.85	20.97
H-RNN [56] (C3D)	19.46	8.78	4.34	2.53	-	8.02	20.18
S2VT + full context [21] (C3D)	26.45	13.48	7.21	3.98	-	9.46	24.56
LSTM-A_3 + policy gradient + retrieval [55] (ResNet + P3D ResNet [37])	-	-	-	-	-	12.84	-
Validation set (Avg. 1st and 2nd)							
LSTM-A_3 (ResNet + P3D ResNet) [55]	-	-	-	**3.38**	13.27	7.71	16.08
LSTM-A_3 + policy gradient + retrieval [55] (ResNet + P3D ResNet [37])	-	-	-	3.13	14.29	8.73	14.75
SINet-Caption — img (C3D)	17.18	7.99	3.53	1.47	18.78	8.44	38.22
SINet-Caption — img (ResNeXt)	18.81	9.31	4.27	1.84	20.46	9.56	43.12
SINet-Caption — obj (ResNeXt)	19.07	9.48	4.38	1.92	20.67	9.56	44.02
SINet-Caption — img + obj — no co-attention (ResNeXt)	**19.93**	9.82	**4.52**	2.03	21.08	9.79	44.81
SINet-Caption — img + obj — co-attention (ResNeXt)	19.78	**9.89**	**4.52**	1.98	**21.25**	**9.84**	**44.84**

Credit: *Attend and Interact: Higher-Order Object Interactions for Video Understanding*, Ma et al

References

1. Video paragraph captioning using hierarchical recurrent neural networks. Yu et al.
2. Coherent multi-sentence video description with variable level of detail. Rohrbach et al
3. *Learning Spatiotemporal Features with 3D Convolutional Networks*, Tran et al
4. *Daps: Deep action Proposals for action understanding*, Escorcia et al
5. *Attend and Interact: Higher-Order Object Interactions for Video Understanding*, Ma et al

Questions?!!

Thanks For your Attention!