Dense-Captioning Events in Videos

Silvio Giancola

Thursday, August 3rd 2017

Teaser

Previous work to describe videos first started with labeling them with a predefined category.

[Krishna et al., 2017] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-Captioning Events in Videos. *arXiv preprint arXiv:1705.00754.*

Outline

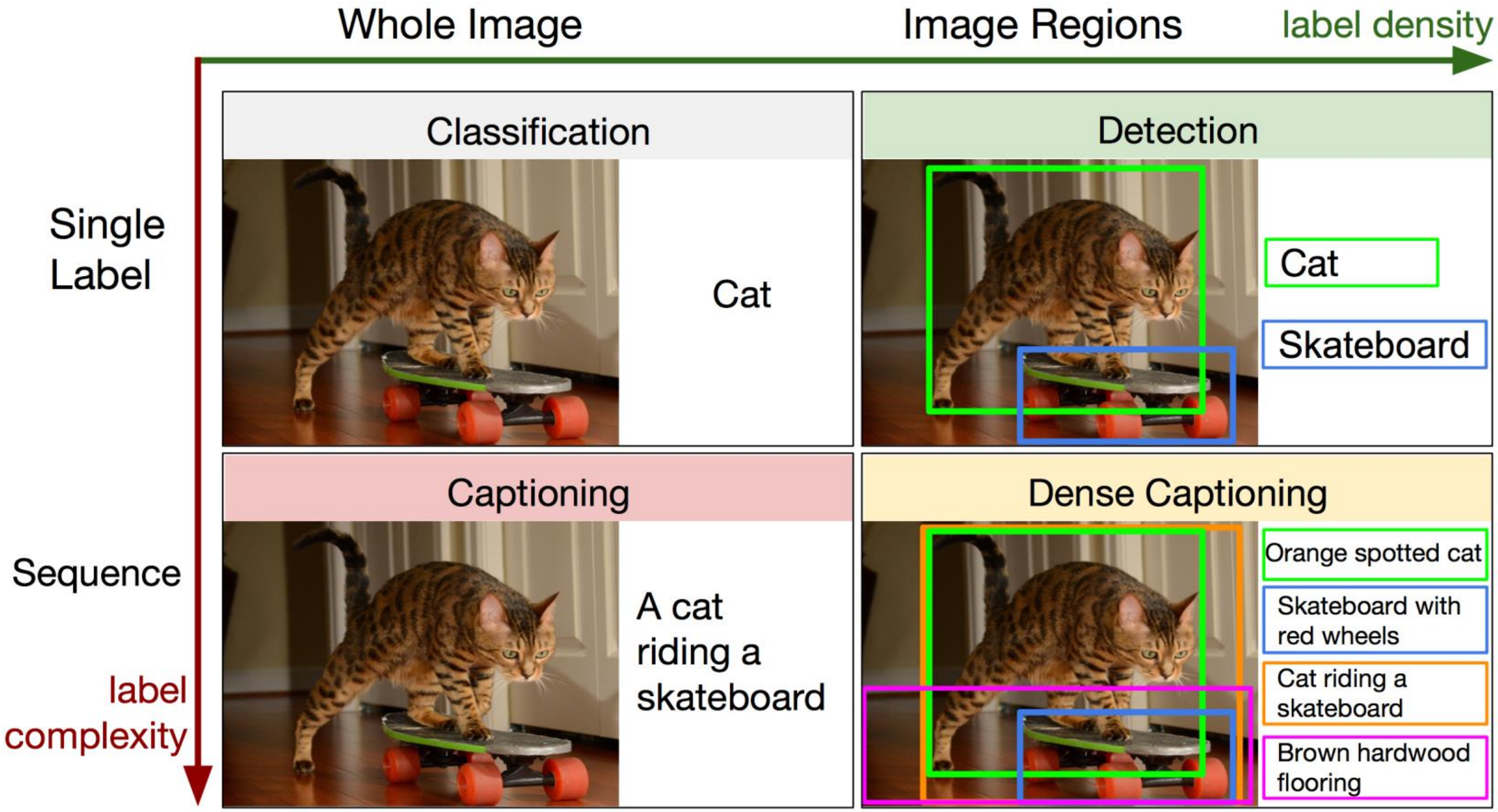
- **Definitions – Dense Captioning Events in Videos**
- Purposes and Applications of Dense Captioning Videos
- Related Works on Video Analysis
- Methodology and Neural Networks Architectures
- Previous and Novel Datasets
- Metrics and Results

Definitions – Dense Captioning

- **Image Recognition**
  - Classification
  - Detection
  - Captioning
  - **Dense Captioning**

- Localizes and describes **regions in space**

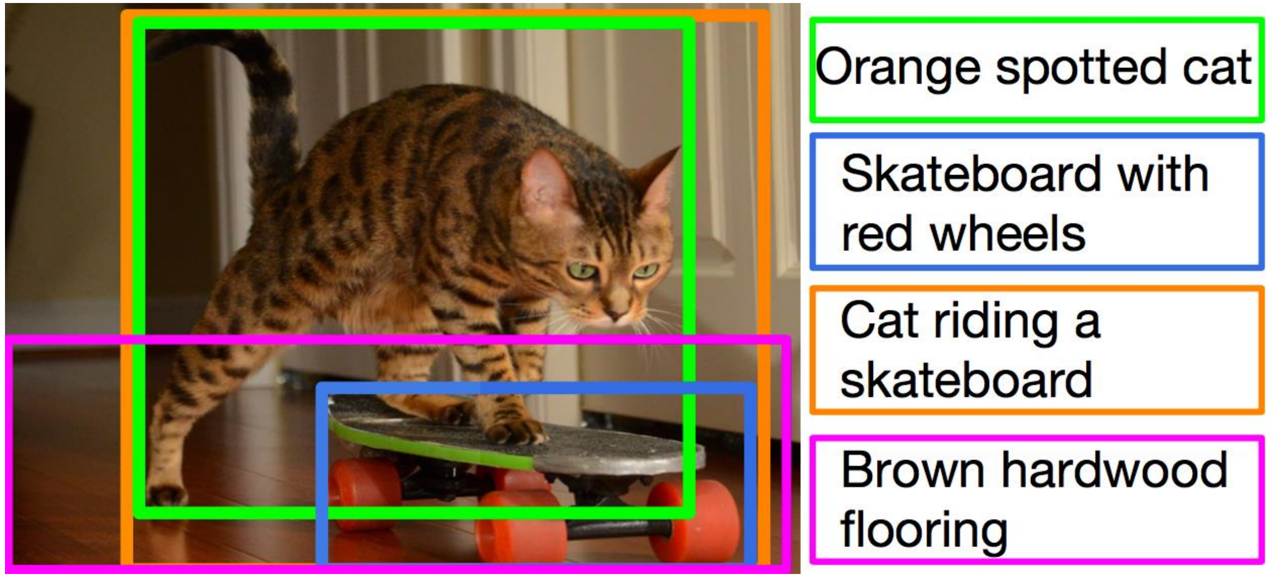
[Johnson et al. 2016]



[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

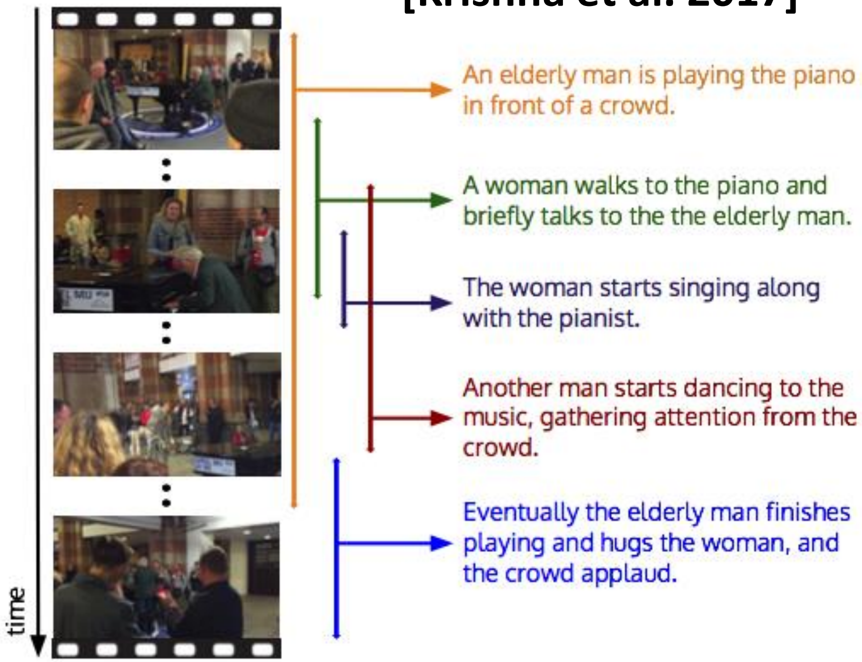
Definitions – Dense Captioning in Video

[Johnson et al. 2016]



Dense captioning *image:* localizes and describes **regions in space**

[Krishna et al. 2017]



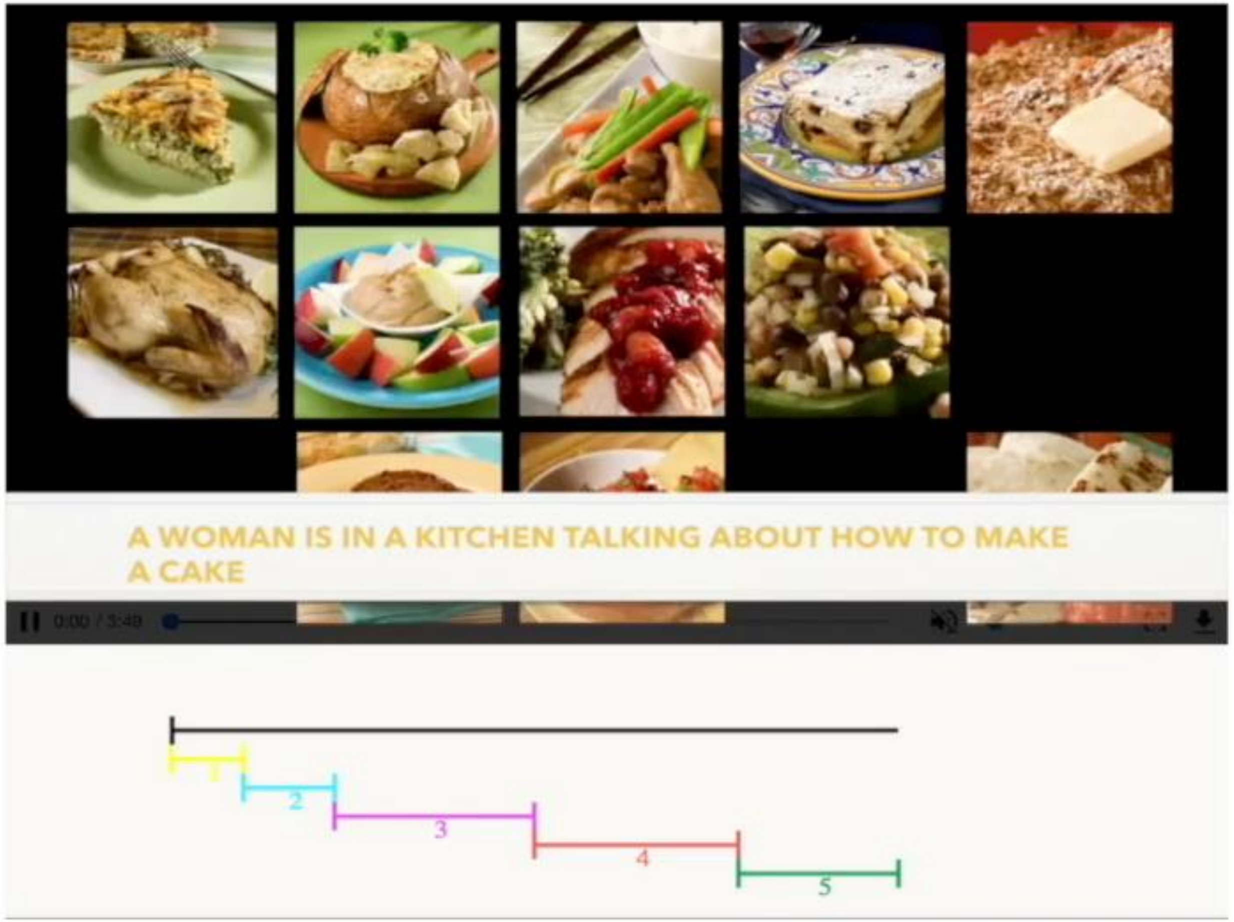
Dense captioning ***events:*** localizes and describes **events in time**

[Krishna et al., 2017] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-Captioning Events in Videos. *arXiv preprint arXiv:1705.00754.*

Definitions – Dense Captioning in Videos

- Events defined by:
  - Time boundaries (from t_{start} to t_{end})
  - Natural Language description

- Images ⟷ Videos
- Region ⟷ Events
- Space ⟷ Time



[Krishna et al. 2017]

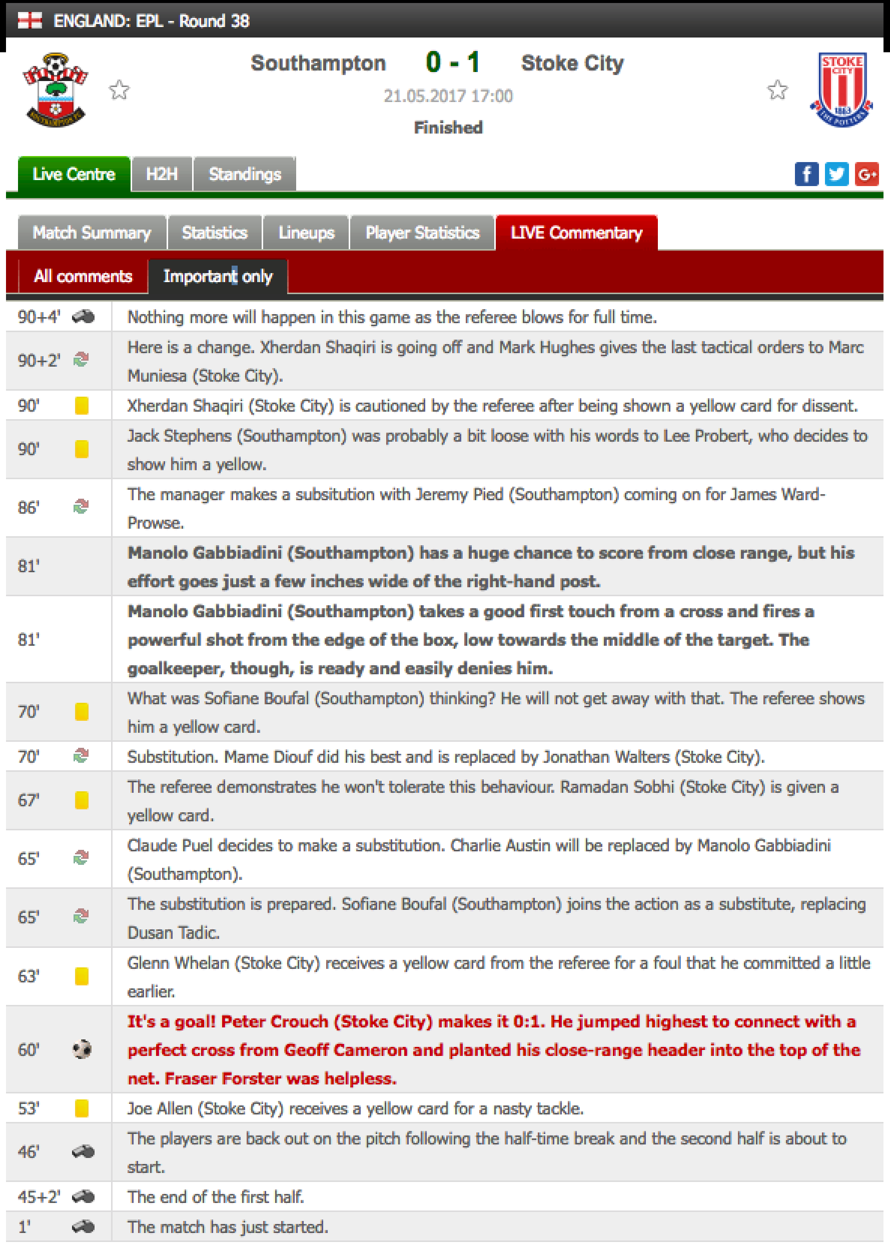
[Krishna et al., 2017] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-Captioning Events in Videos. *arXiv preprint arXiv:1705.00754.*

Outline

- Definitions – Dense Captioning Events in Videos
- **Purposes and Applications of Dense Captioning Videos**
- Related Works on Video Analysis
- Methodology and Neural Networks Architectures
- Previous and Novel Datasets
- Metrics and Results

Applications

- **Automatic Commentary** in sports (Soccer, Basketball, Football,…)
  - Localize actions in the video
  - Generate Natural Language description of the game actions
  - Summarize the game



ENGLAND: EPL - Round 38

Southampton 0 - 1 Stoke City

21.05.2017 17:00

Finished

Live Centre | H2H | Standings

Match Summary | Statistics | Lineups | Player Statistics | LIVE Commentary

All comments | Important only

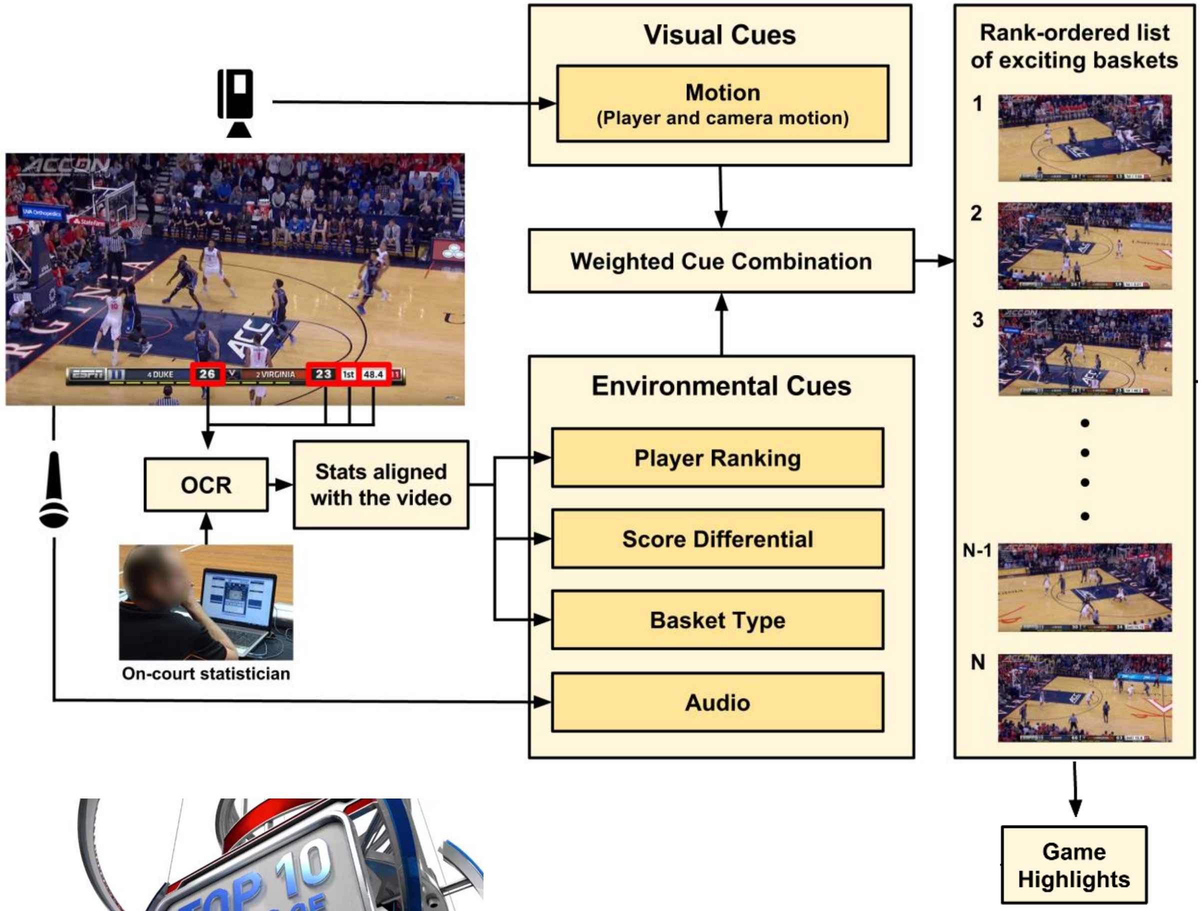
90+4'	Nothing more will happen in this game as the referee blows for full time.
90+2'	Here is a change. Xherdan Shaqiri is going off and Mark Hughes gives the last tactical orders to Marc Muniesa (Stoke City).
90'	Xherdan Shaqiri (Stoke City) is cautioned by the referee after being shown a yellow card for dissent.
90'	Jack Stephens (Southampton) was probably a bit loose with his words to Lee Probert, who decides to show him a yellow.
86'	The manager makes a subsitution with Jeremy Pied (Southampton) coming on for James Ward-Prowse.
81'	**Manolo Gabbiadini (Southampton) has a huge chance to score from close range, but his effort goes just a few inches wide of the right-hand post.**
81'	**Manolo Gabbiadini (Southampton) takes a good first touch from a cross and fires a powerful shot from the edge of the box, low towards the middle of the target. The goalkeeper, though, is ready and easily denies him.**
70'	What was Sofiane Boufal (Southampton) thinking? He will not get away with that. The referee shows him a yellow card.
70'	Substitution. Mame Diouf did his best and is replaced by Jonathan Walters (Stoke City).
67'	The referee demonstrates he won't tolerate this behaviour. Ramadan Sobhi (Stoke City) is given a yellow card.
65'	Claude Puel decides to make a substitution. Charlie Austin will be replaced by Manolo Gabbiadini (Southampton).
65'	The substitution is prepared. Sofiane Boufal (Southampton) joins the action as a substitute, replacing Dusan Tadic.
63'	Glenn Whelan (Stoke City) receives a yellow card from the referee for a foul that he committed a little earlier.
60'	**It's a goal! Peter Crouch (Stoke City) makes it 0:1. He jumped highest to connect with a perfect cross from Geoff Cameron and planted his close-range header into the top of the net. Fraser Forster was helpless.**
53'	Joe Allen (Stoke City) receives a yellow card for a nasty tackle.
46'	The players are back out on the pitch following the half-time break and the second half is about to start.
45+2'	The end of the first half.
1'	The match has just started.

[Bettadapura et al., 2016]

Applications

- **Automatic Highlighting** in sports
    - Retrieve events from contextual information (audio, stats)
    - Rank them base on ESPN data
    - Provides automatic game highlights

    - They use on-court game description to rank their



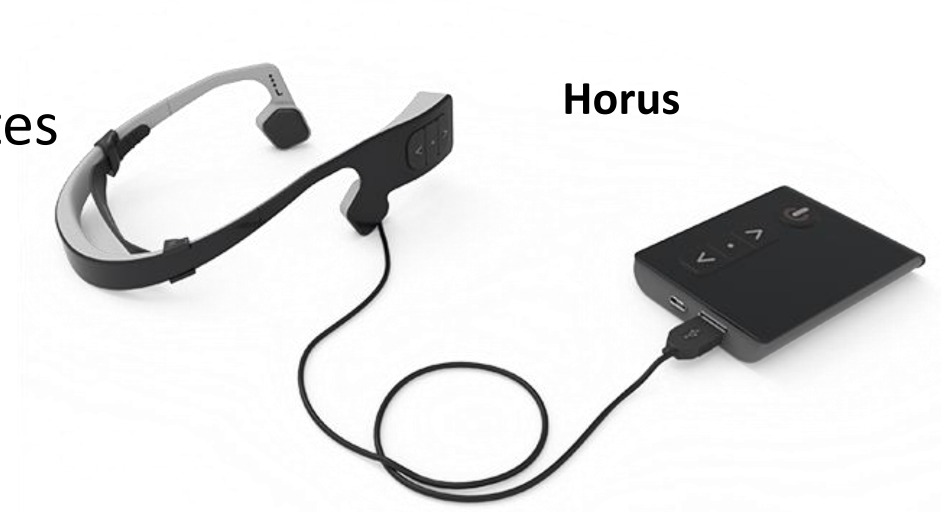
[Bettadapura et al., 2016] Bettadapura, V., Pantofaru, C., & Essa, I. (2016, October). Leveraging Contextual Cues for Generating Basketball Highlights. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 908-917). ACM.

Applications

- **Audio-Description** for visual impaired people
  - Industrial Solutions:
    - HW: Horus, OrCam MyEye, Google Glass…
    - SW: TapTapSee with Google TalkBack
  - Read any printed text from any surface
  - Recognize faces, identify products and bank notes

Google Glass

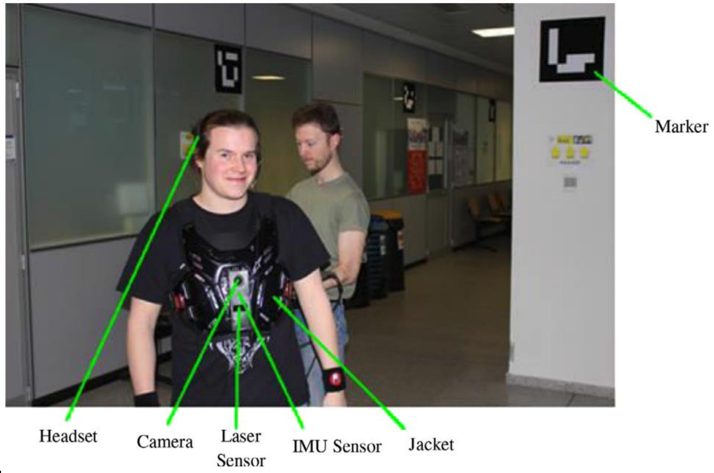
Horus

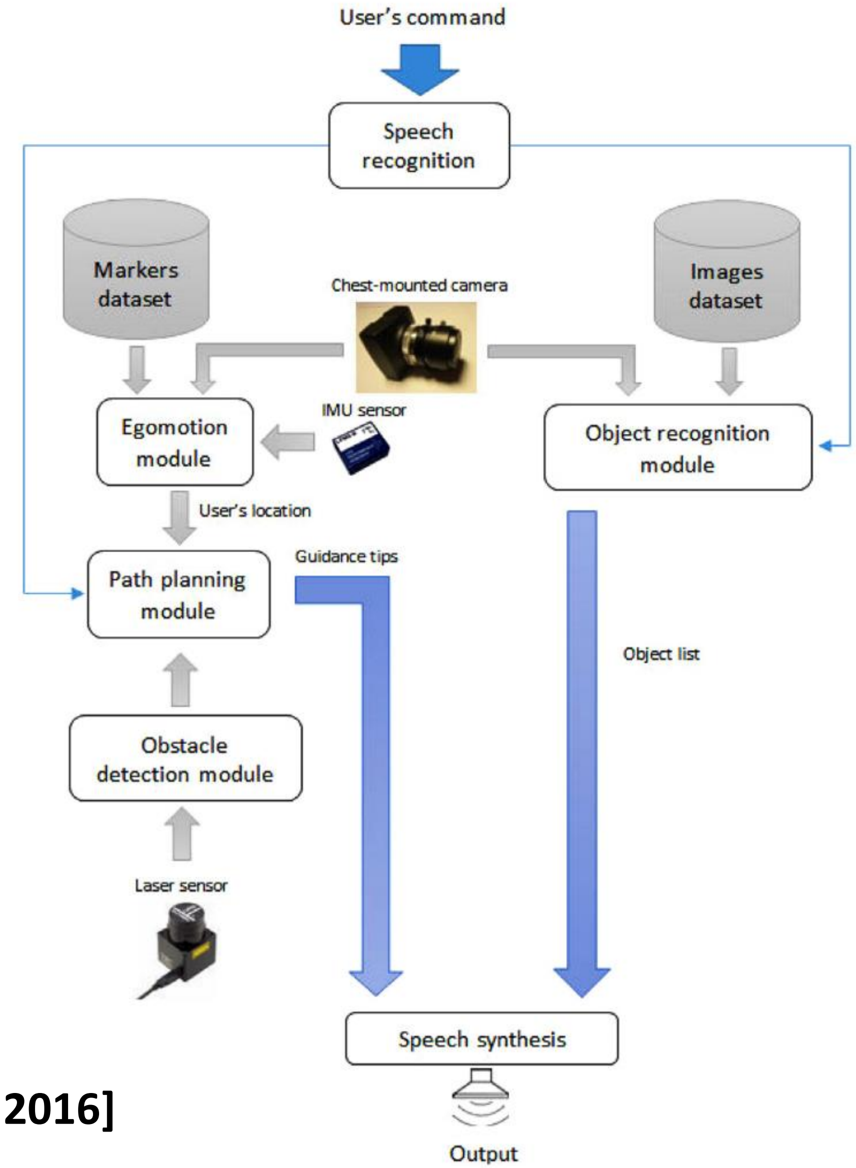


OrCam MyEye

Applications

- **Audio-Description** for visual impaired people
  - R&D projects details (2016)
  - Based on *images* to perform *object* recognition
  - Does not handle interactions, events nor activities





[Mekhalfi et al., 2016]

[Mekhalfi et al., 2016] Mekhalfi, M. L., Melgani, F., Zeggada, A., De Natale, F. G., Salem, M. A. M., & Khamis, A. (2016). Recovering the sight to blind people in indoor environments with smart technologies. *Expert Systems with Applications*, *46*, 129-138.

Outline

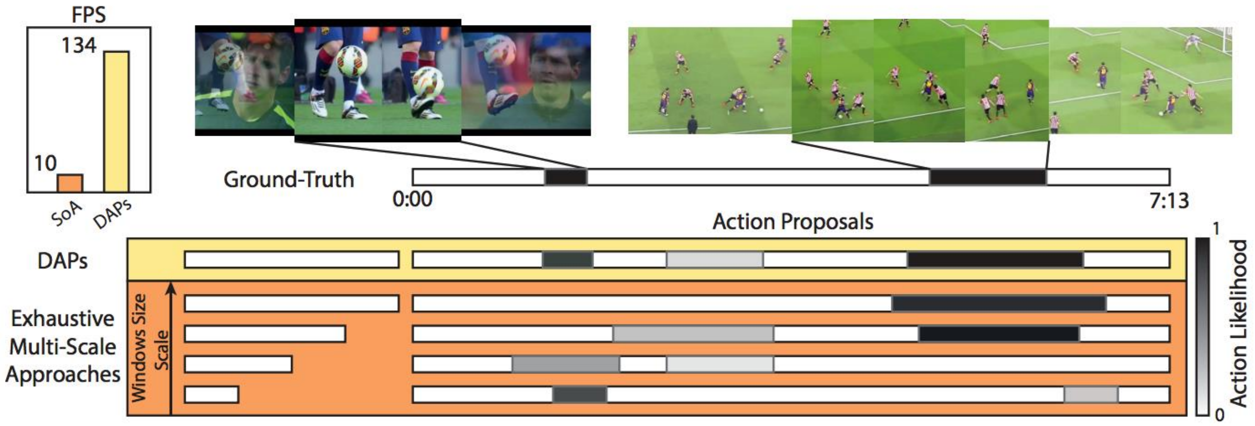
- Definitions – Dense Captioning Events in Videos
- Purposes and Applications of Dense Captioning Videos
- **Related Works on Video Analysis**
- Methodology and Neural Networks Architectures
- Previous and Novel Datasets
- Metrics and Results

Related Works

- **Activity Recognition**
  - Early work on Hidden Markov models [Yamato et al. 1992]
  - Discriminative SVM models
    - Key poses and action grammar [Vahdat et al. 2011], [Ni et al. 2014], [Pirsiavash et al. 2014]
    - Tracking of hand-crafted features [Rohrbach et al. 2012] or object-centric features [Ni et al. 2014]
  - Improved Dense Trajectories [Wang et al. 2014]
  - Deep learning features [Karaman et al. 2014]

➢Describe video with **Activity Label** instead of with **Natural Language**

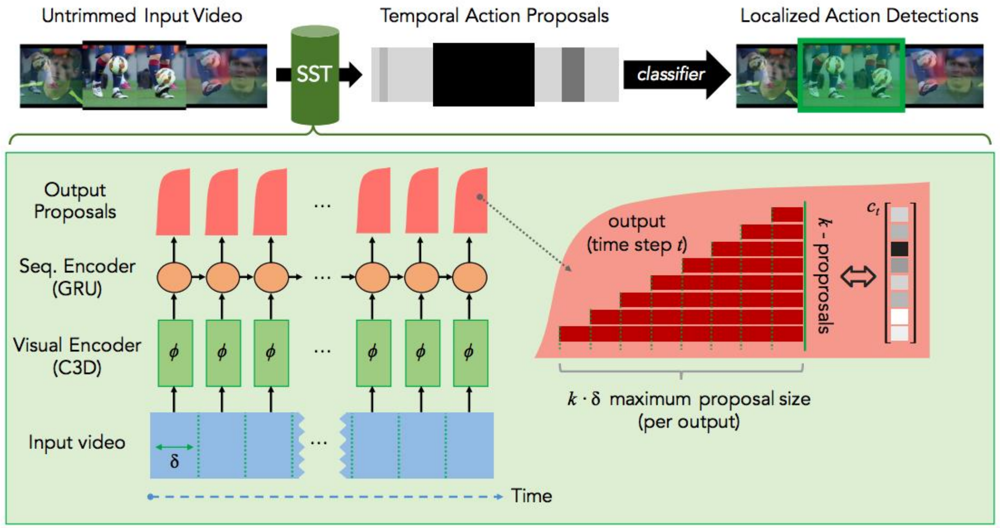
Related Works



[Escorcia et al., 2016]

- **Temporal Action Proposal**
  - Windowing approach [Duchenne et al .2009]
  - Dictionary Learning [Heilbron et al. 2016]
  - RNN Architecture [Escorcia et al. 2016]
  - Single-Stream Approach [Buch et al. 2017]

➢Provides **time boundaries** for activities to describe



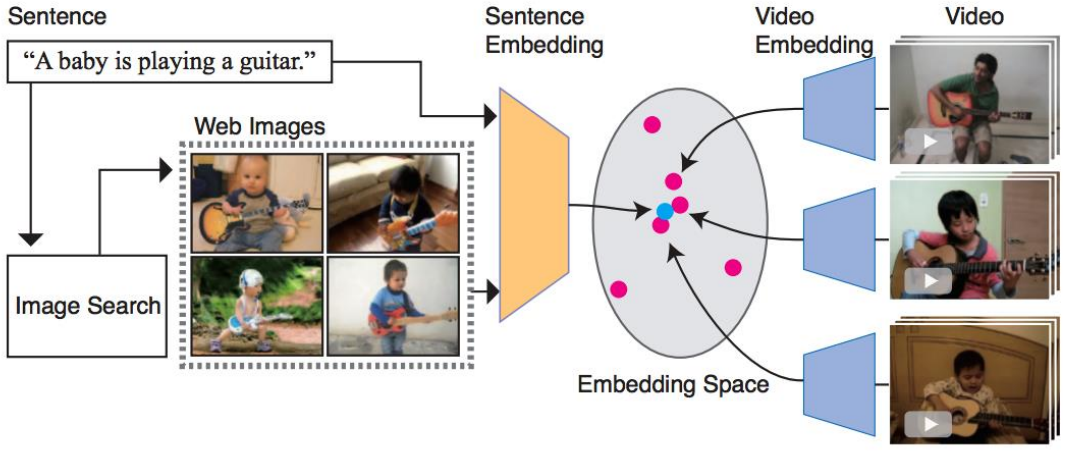
[Buch et al., 2017]

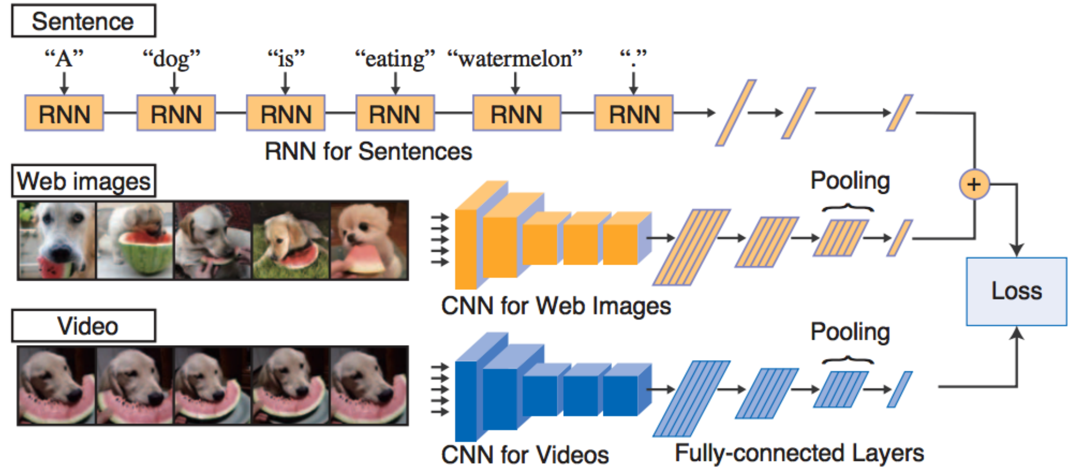
Related Works

- **Video Retrieval with Natural Language**
  - Unified Framework [Xu et al. 2015]
  - Embedding space between language and videos [Otani et al. 2016]

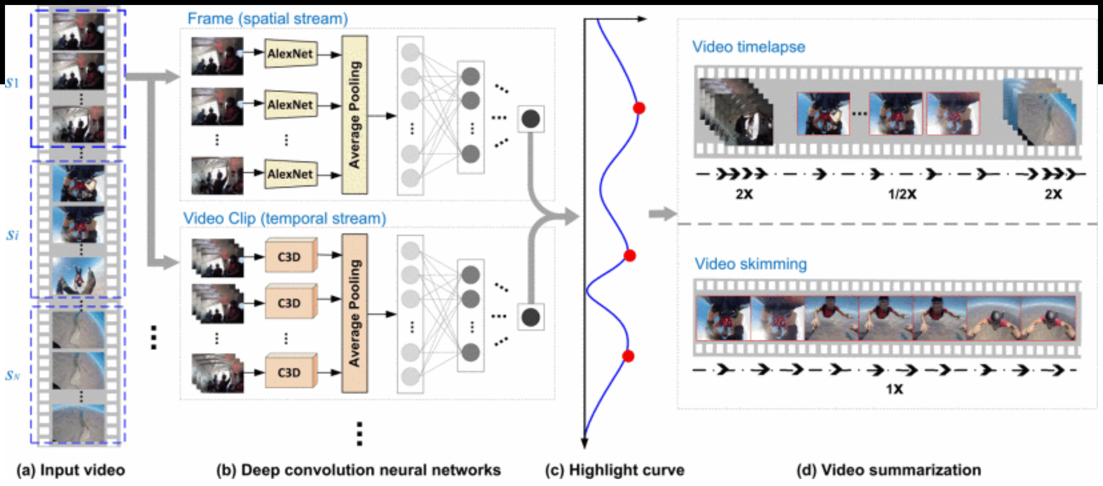
➢ **Inverse problem** of Dense Captioning
  - ➢ Dense Captioning: From Video to Text
  - ➢ Video Retrieval: From Text to Video

[Xu et al. 2015]





Related Works



[Yao et al. 2016]

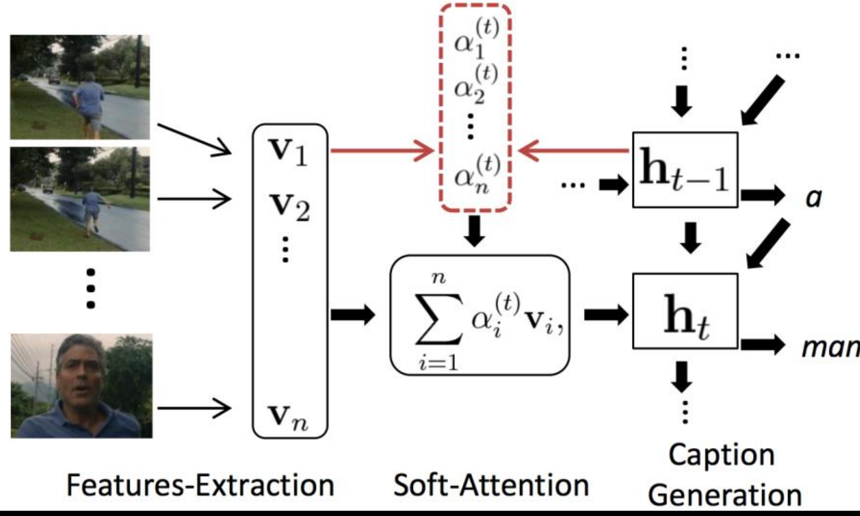
- **Video Summarization**
  - Traditional Image Processing:
    - Low level features such as color [Zhang et al. 1997] and motion [Wolf et al. 1996]
    - Stitching from video to single image [Goldman et al. 2006]
  - Congregate segments of videos that include interesting visual information
    - Detection of uncommon behavior [Boiman et al. 2007]
    - Submodular Mixtures of Objectives [Gygli et al. 2015],
    - DNN to classify between highlights and background [Yang et al. 2015]
    - Finding maximum in interest curves in time [Yao et al. 2016],
  - Additional text inputs from user studies to guide the selection process using *title* [Song et al. 2015], *query* [Liu et al. 2015] or *description* [Yeung et al. 2014]

[Yao et al. 2015]

Related Works

- **Video Captioning**
  - Single-event captioning based on DNN [Venugopalan et al. 2014]
  - Recurrent encoder [Donahue et al. 2015], [Venugopalan et al. 2015], [Xu et al. 2015]
  - Attention mechanism [Yao et al. 2015].

[Yao et al. 2015]

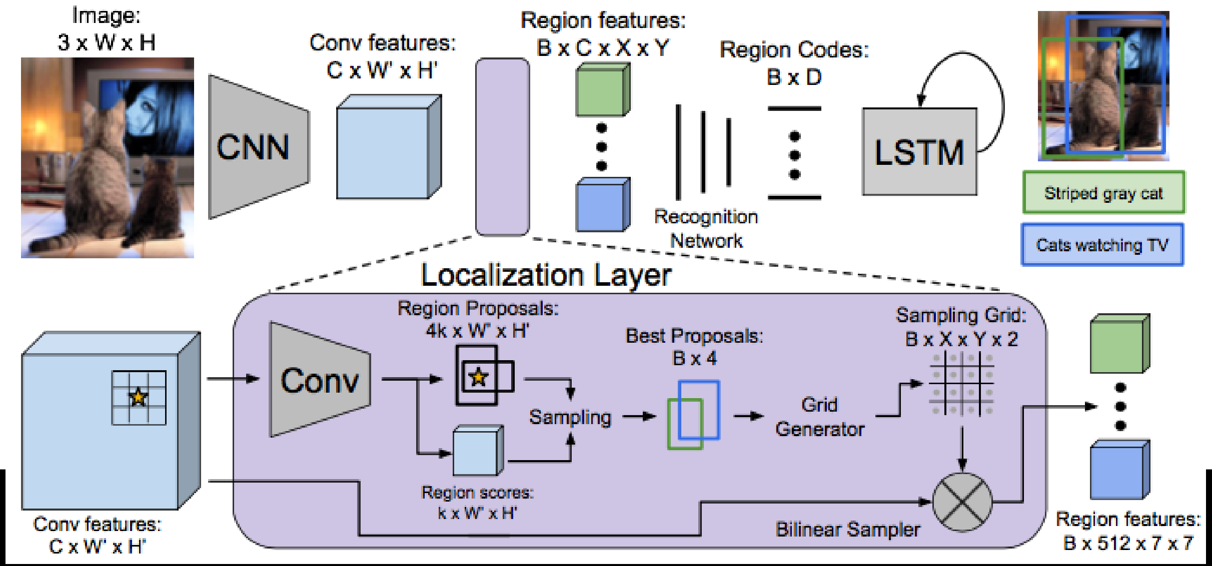


➢**Whole video** labelling

- Make use of paragraph to describe with more details [Mikolov et al. 2010]
  - Limited to cooking video

Related Works

- **Dense Image Captioning**
  - Spatial context to improve captioning [Yang et al. 2016], [Xu et al. 2015]
  - Spatial attention to improve human tracking [Alahi et al. 2016]
  - Localized descriptions for an image [Johnson et al. 2016]



[Johnson et al. 2016]

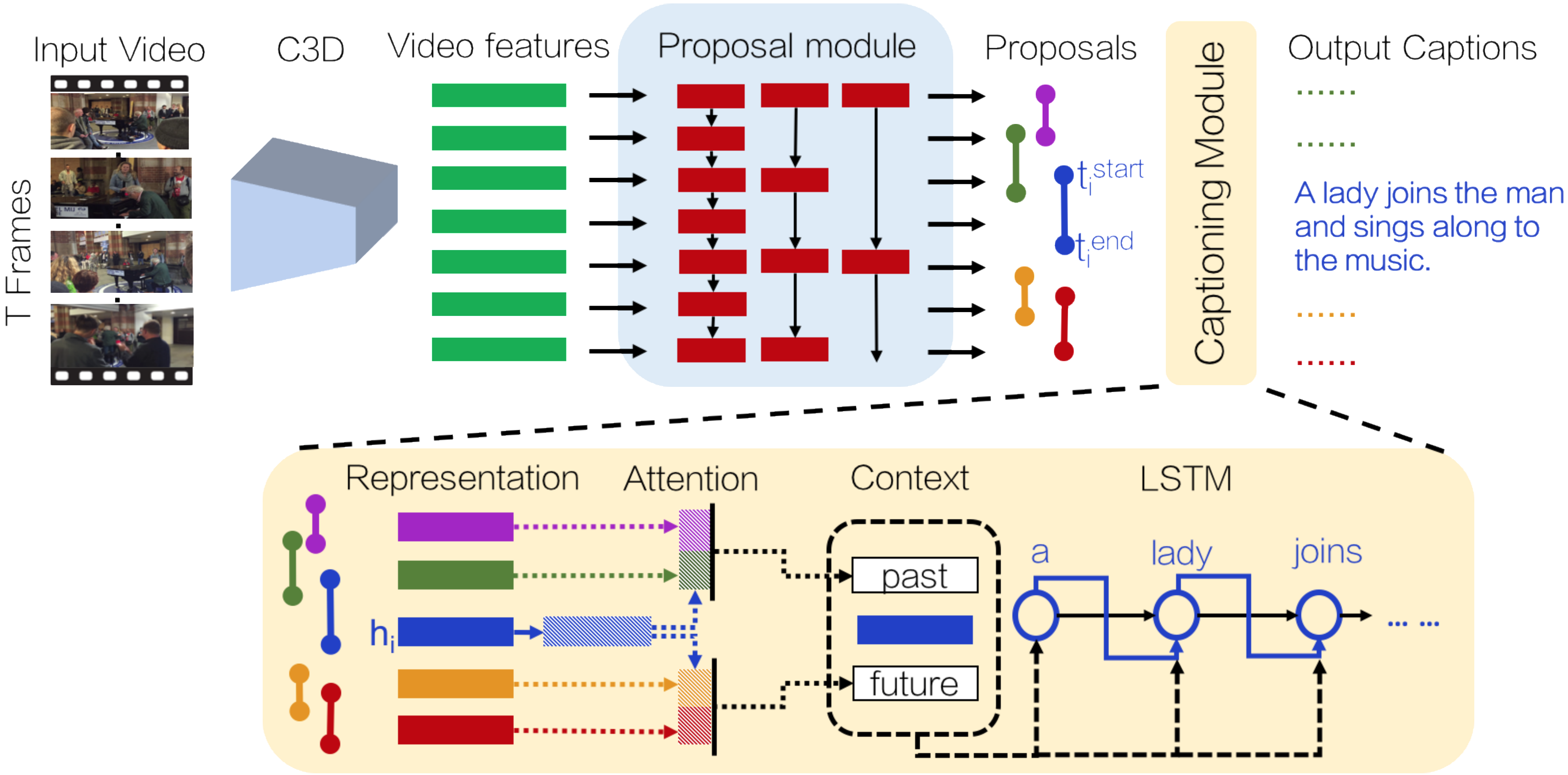
Outline

- Definitions – Dense Captioning Events in Videos
- Purposes and Applications of Dense Captioning Videos
- Related Works on Video Analysis
- **Methodology and Neural Networks Architectures**
- Previous and Novel Datasets
- Metrics and Results

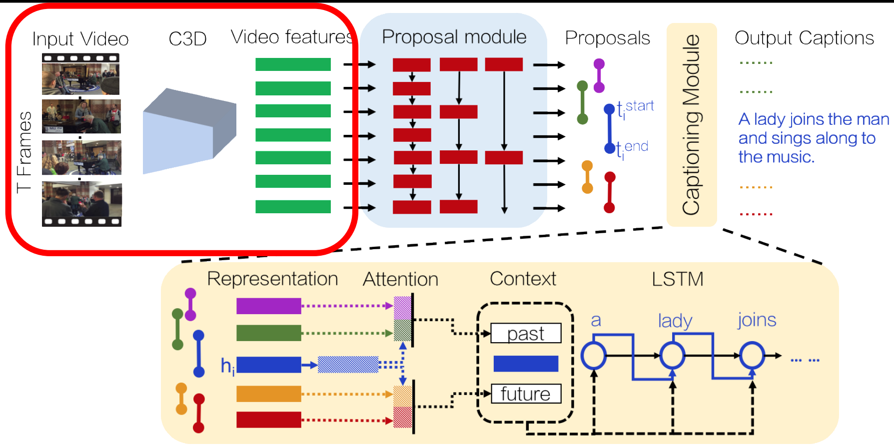
Methodology

- Novel ideas brought by [Krishna et al., 2017]:
  - There are **numerous events** in a video that are **correlated** between each others
  - **Context** around the event helps in the description of an event
  - Events can occurs within **a second** or last **up to minutes**
- Contributions:
  - New variant of an existing **proposal model** that handle multiple time-scales
  - Identification and description of all events with Natural Language **in a single pass.**
  - Novel **captioning module** that uses **contextual information** from past (and eventually future) events
  - **Activity-Net Captions**, a large-scale benchmark for dense-captioning events

Methodology – DNN proposed

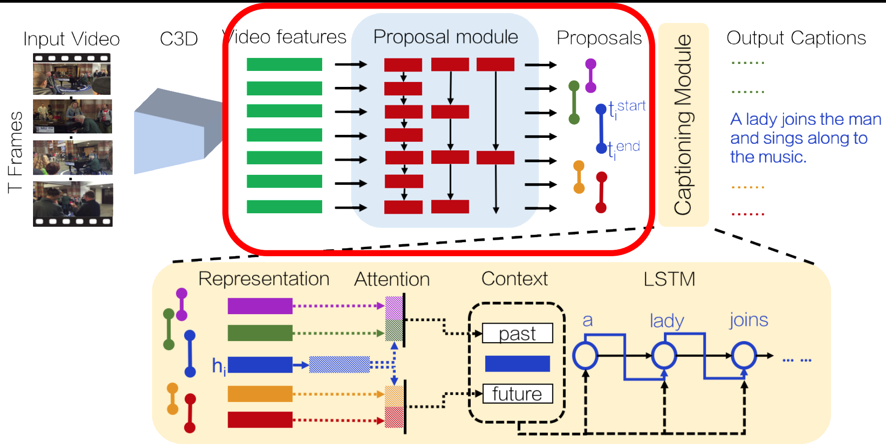


Methodology – DNN proposed



- **Video Features Extraction**
  - **3D CNN** features [Ji et al., 2013]
  - $\{f_t = F(v_t\colon v_{t+\delta})\}$ with $\delta =$ **16 frames**
  - Already trained, **no fine-tuning**
  - Embed 16 frames in a **500-length feature vector**
  - Output $N = \boldsymbol{T/\delta}$ **features**, T being the length of the video

[Ji et al., 2013] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 221-231.

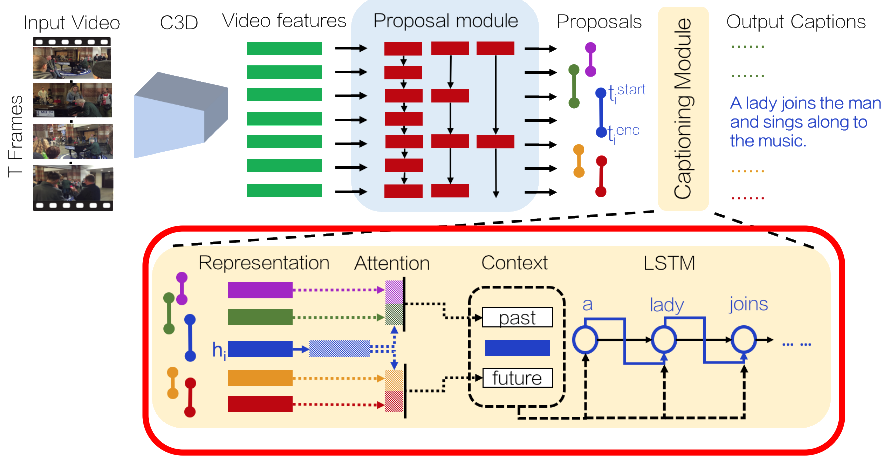


Methodology – DNN proposed

- **Event proposal module**
  - Based on **DAPs** [Escorcia et al., 2016]
  - Sample the video features at **different stride** (1, 2, 4, 8)
  - They do not modify the training, they only use the model for **inference**
  - Can generate **overlapped proposals** for events

[Escorcia et al., 2016] Escorcia, V., Heilbron, F. C., Niebles, J. C., & Ghanem, B. (2016, October). DAPs: Deep action proposals for action understanding. In *European Conference on Computer Vision* (pp. 768-784). Springer International Publishing.

Methodology – DNN proposed



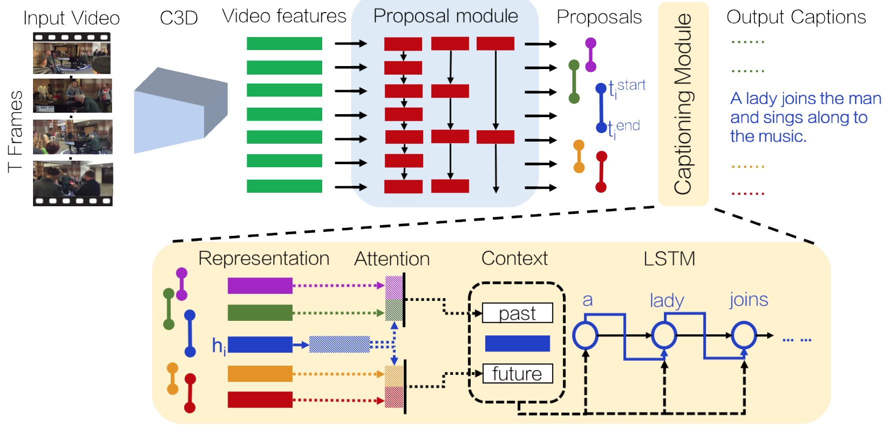
- **Captioning module with context**
  - Language LSTM that uses contextual information for a given event representation h_i

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

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

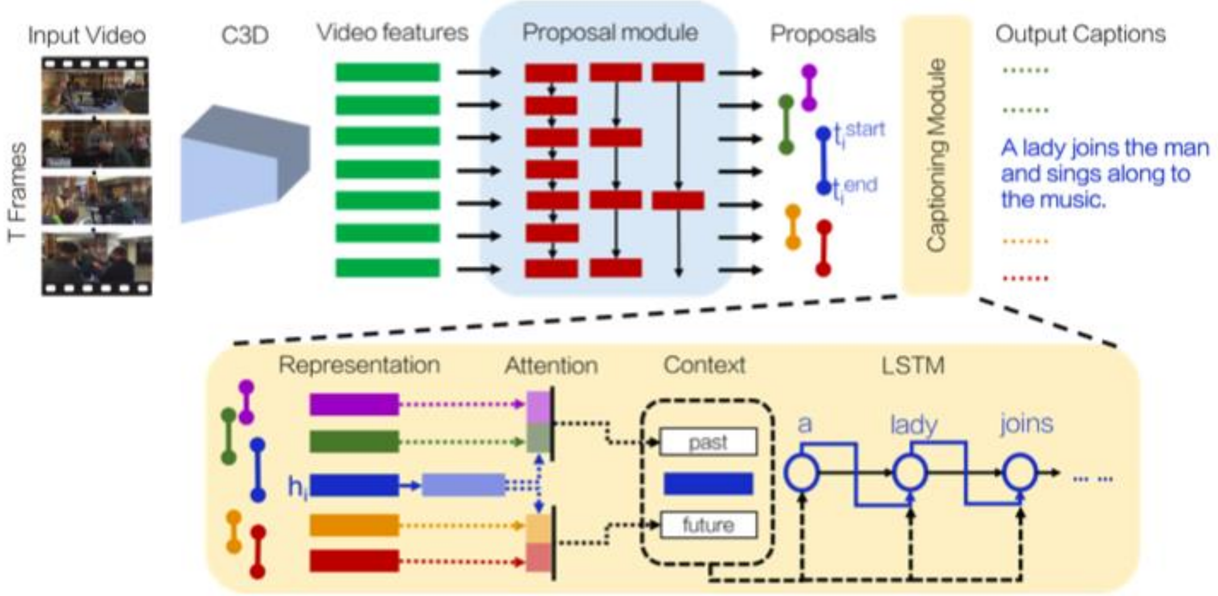
  - The Z are normalization values : $Z^{\text{past}} = \sum_{j \neq i} \mathbb{1}[t_j^{\text{end}} < t_i^{\text{end}}]$
  - An attention in learned: $a_i = w_a h_i + b_a$
  - The weights are estimated in function of a learned attention $w_j = a_i h_j$

Methodology – DNN proposed

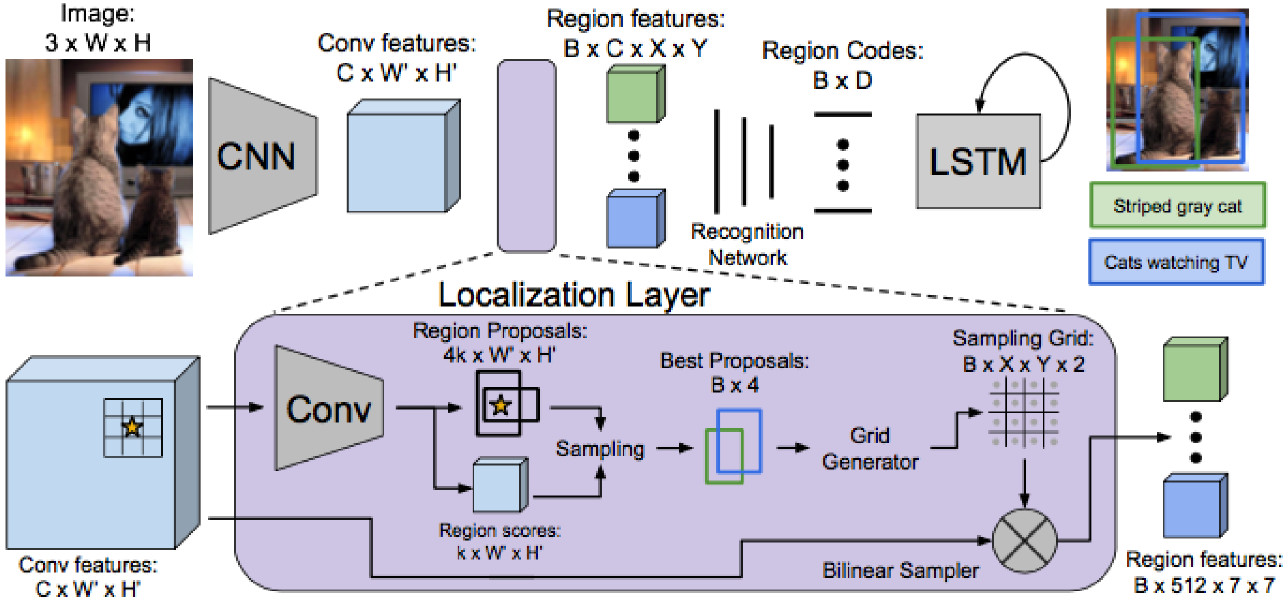


- **Implementation details**
  - The **3DCNN** module is **not fine-tuned**.
  - Minimization of a **combined loss** function $L = L_{caption} + 0.1L_{proposal}$
  - Weight initialized using a Gaussian with $std = 0.01$
  - Alternation of the training between captioning and proposal every 500 iterations
  - **SGD** with $LR_{language} = 0.01, LR_{proposal} = 0.001$ and $momentum = 0.9$
  - Implemented in **PyTorch** on a **Titan X GPU**
  - $15.84\ ms$ per batch ($size = 1$), convergence after **2 days**.

Analogy with Dense Captioning in Images



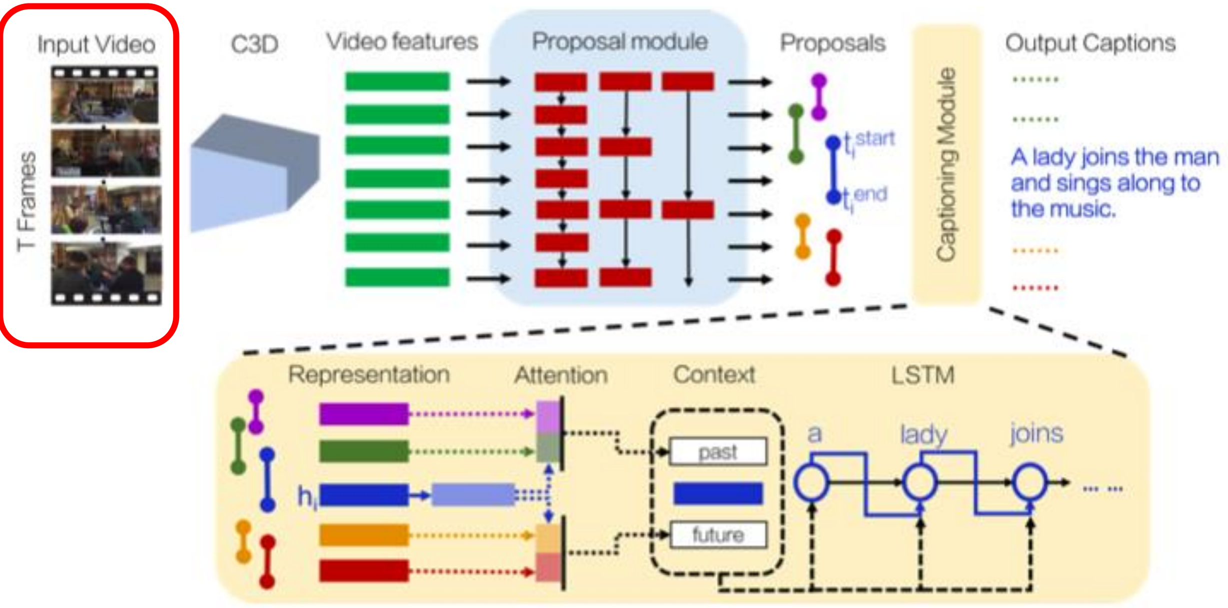
[Krishna et al. 2017]



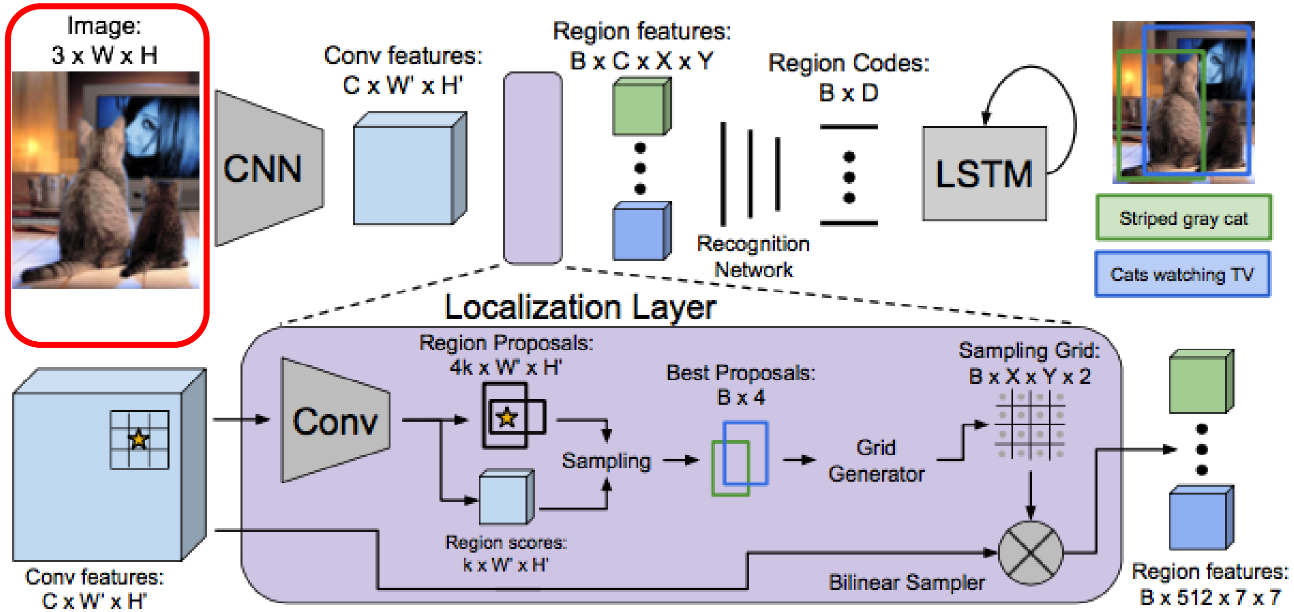
[Johnson et al. 2016]

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]



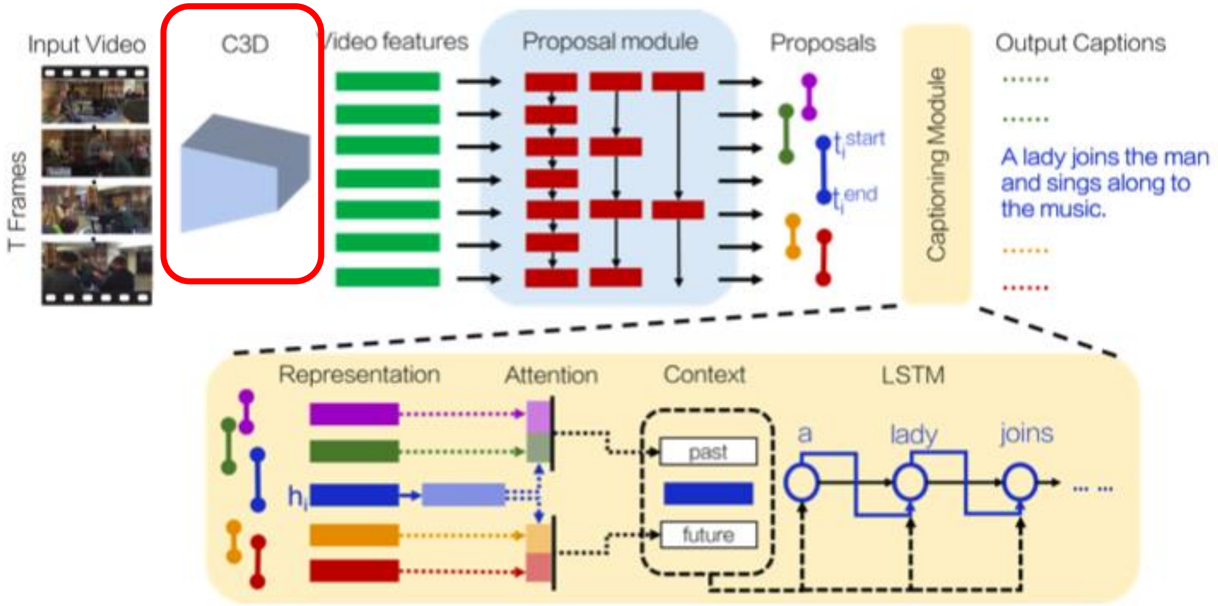
[Johnson et al. 2016]

- Video: Stack of frames
- 4 Dimensions: $T \times 3 \times W \times H$

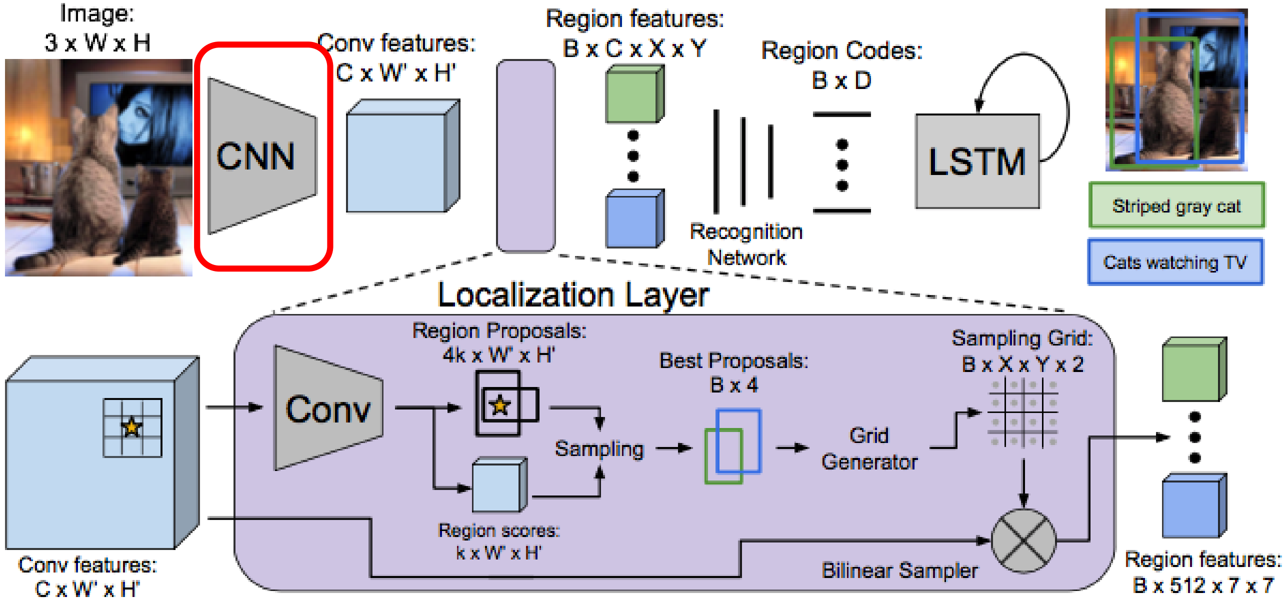
- Image: Single RGB frame
- 3 Dimensions: $3 \times W \times H$

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]

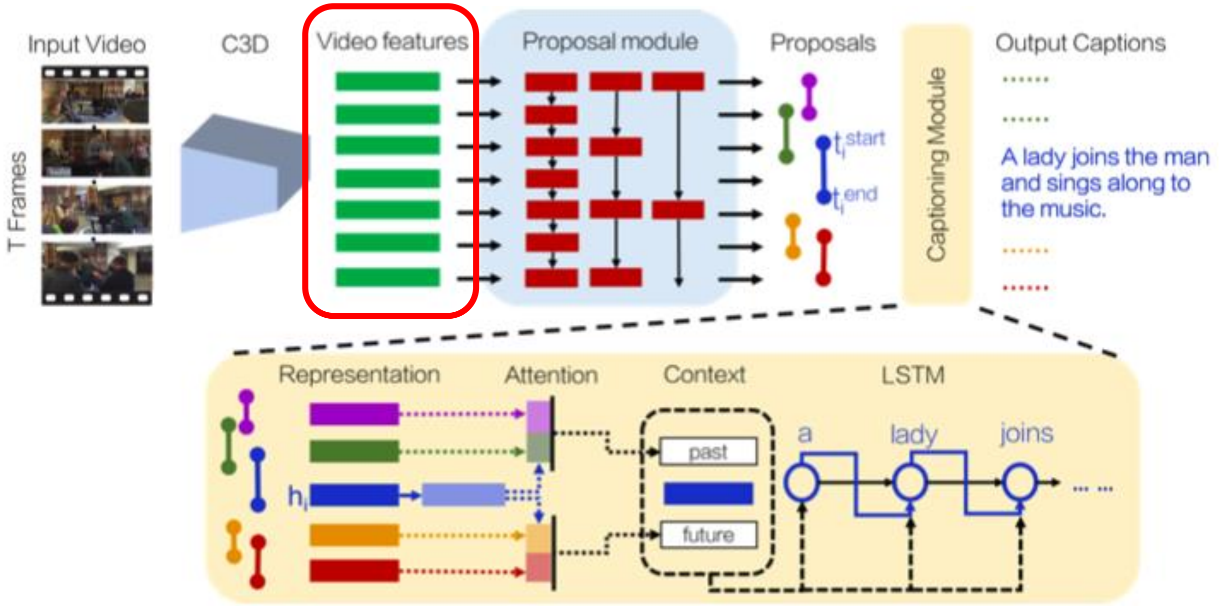


[Johnson et al. 2016]

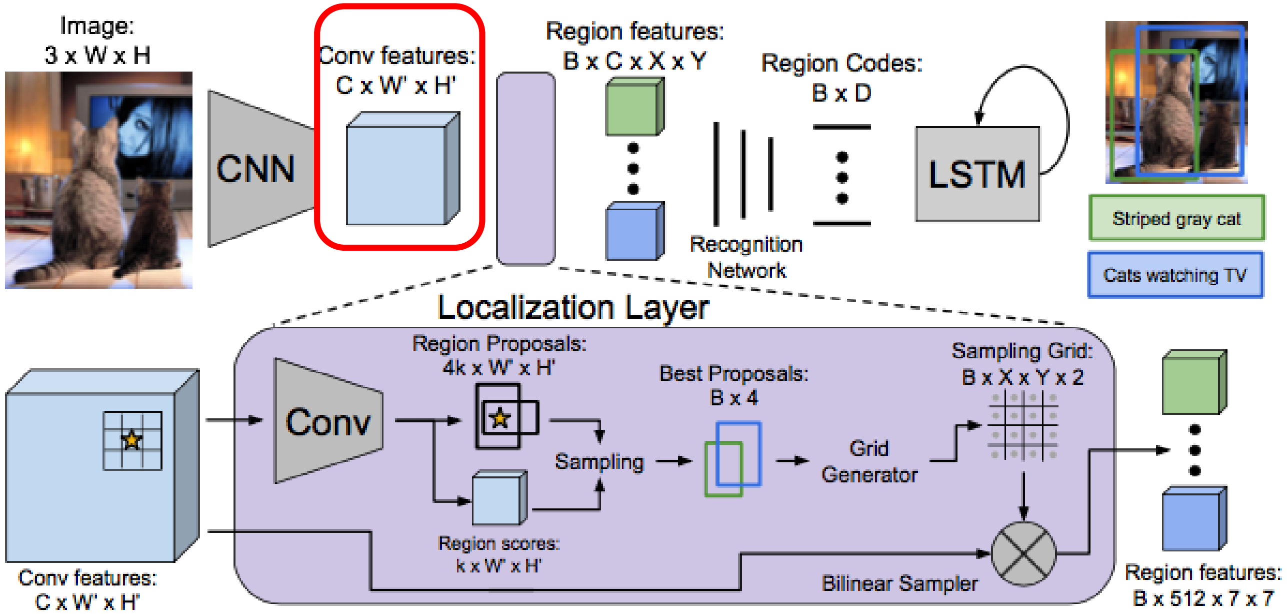
- Embedding: 3D CNN
- Embedding: 2D CNN

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]



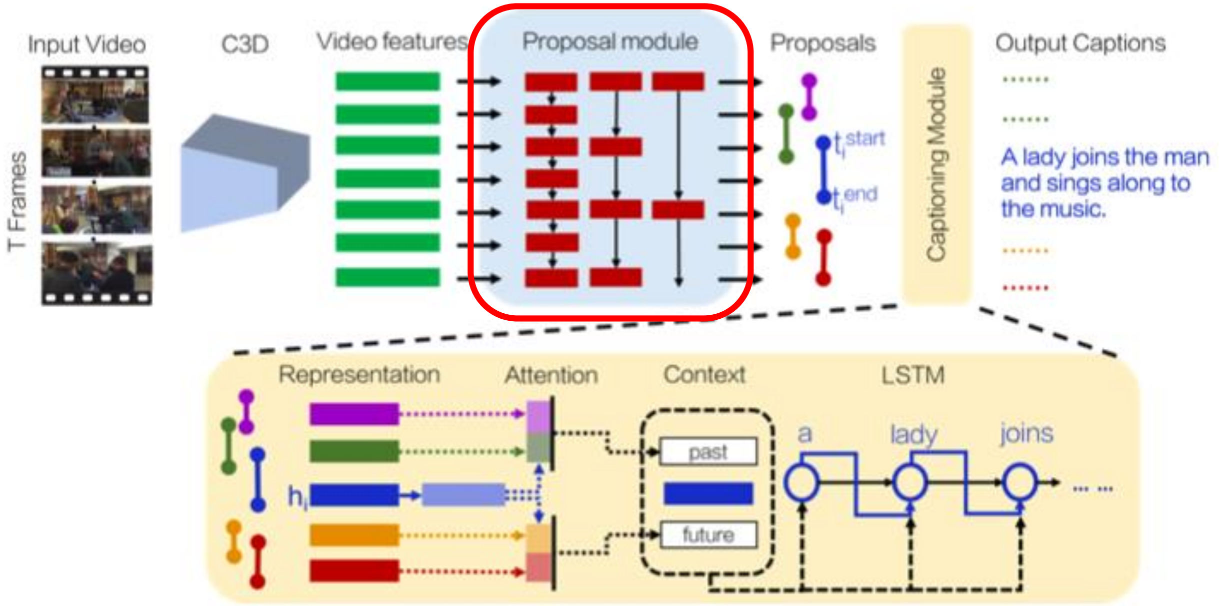
[Johnson et al. 2016]

- Features : $500 \times \frac{T}{\delta}$

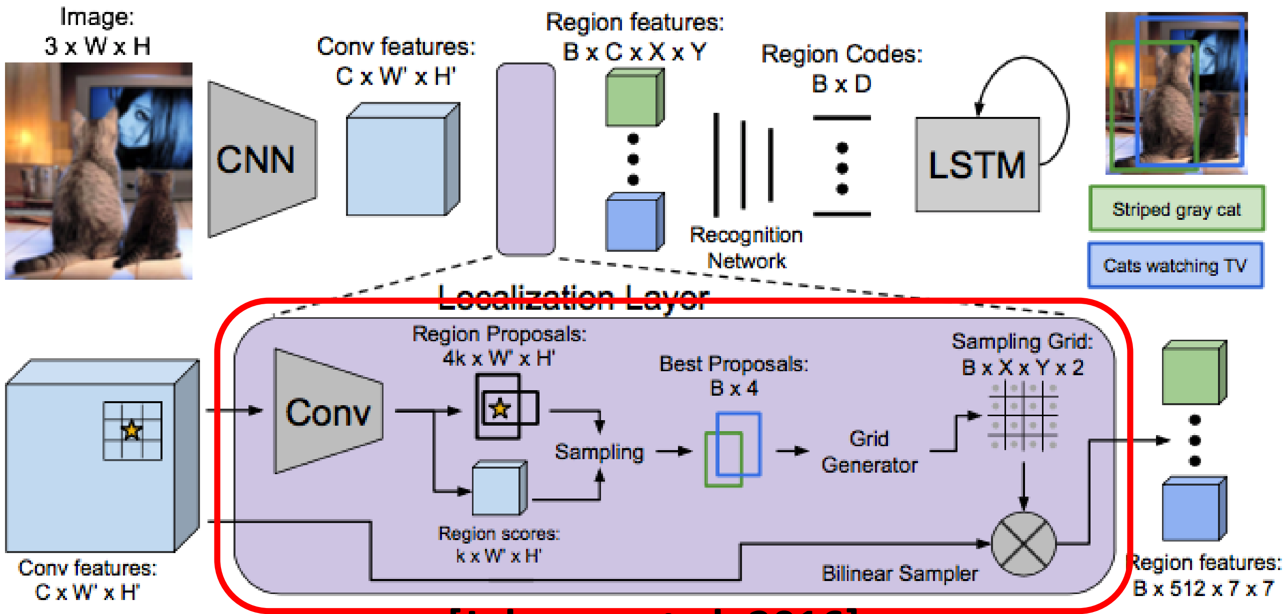
- Features: $C \times W' \times H'$

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]

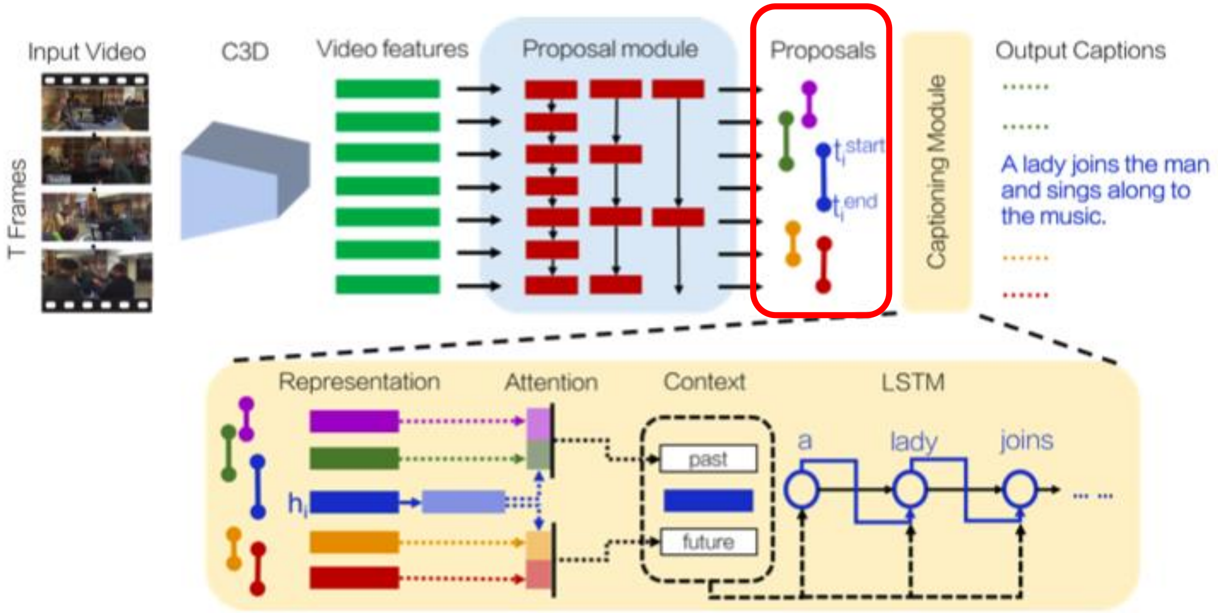


[Johnson et al. 2016]

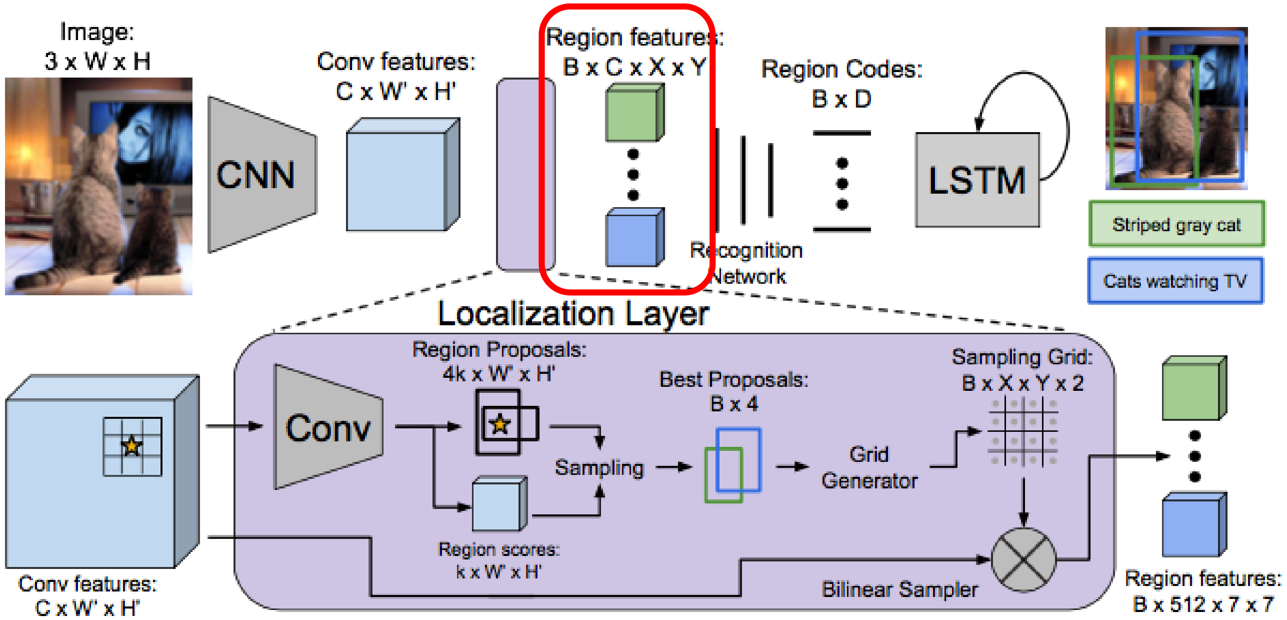
- Temporal proposal method
- Spatial proposal method

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]

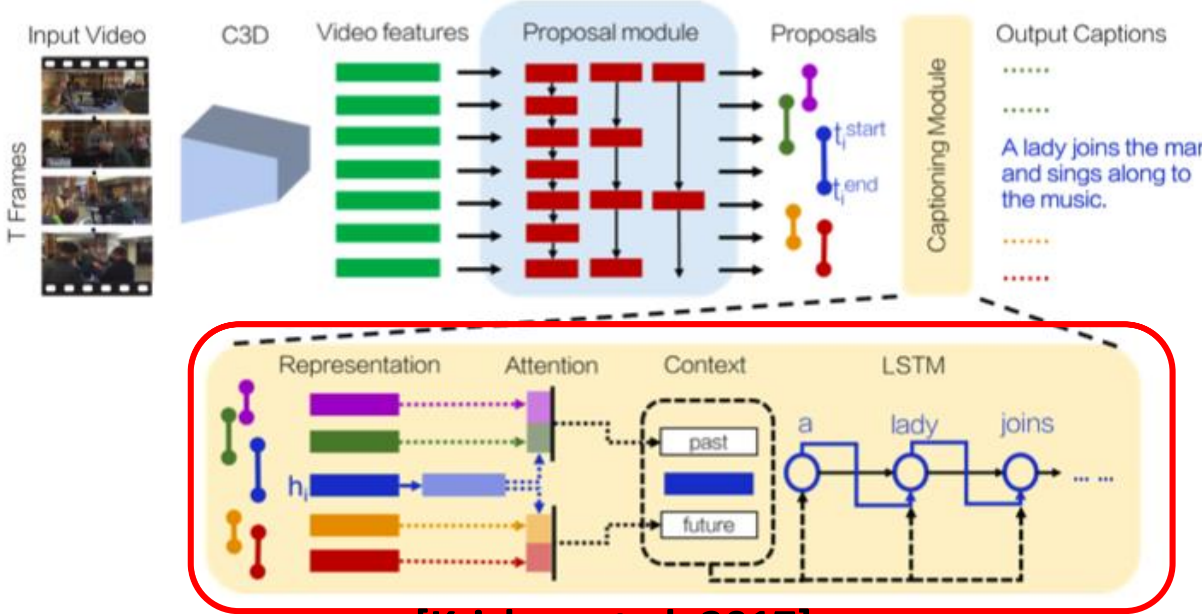


[Johnson et al. 2016]

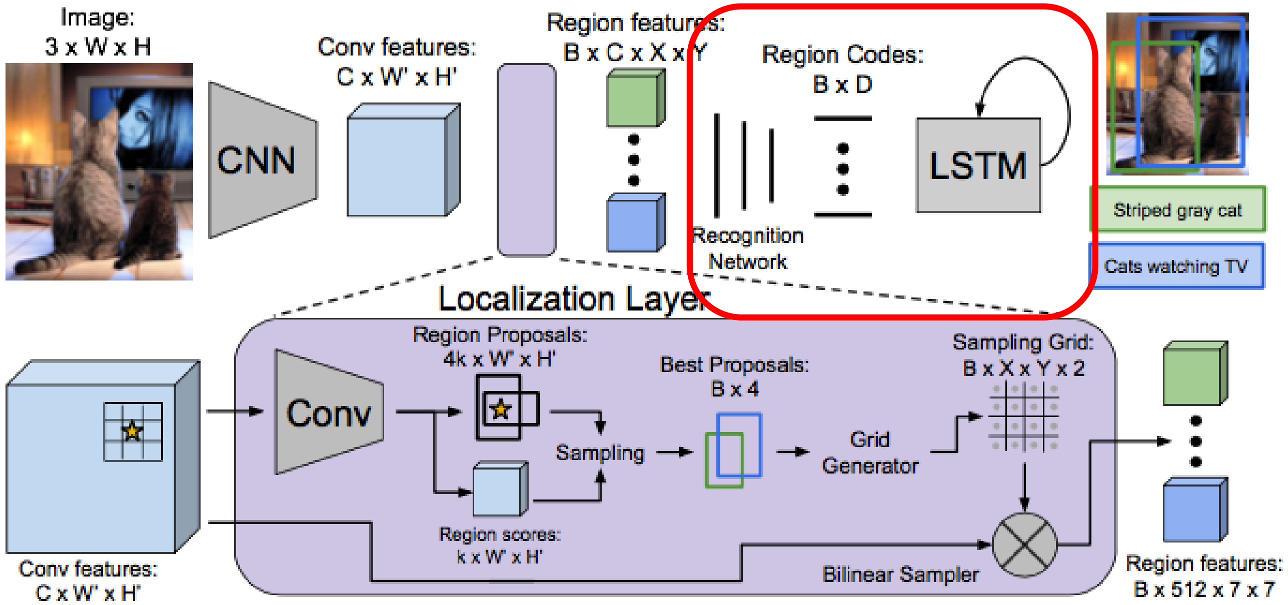
- Temporal proposal
- Spatial proposals

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]



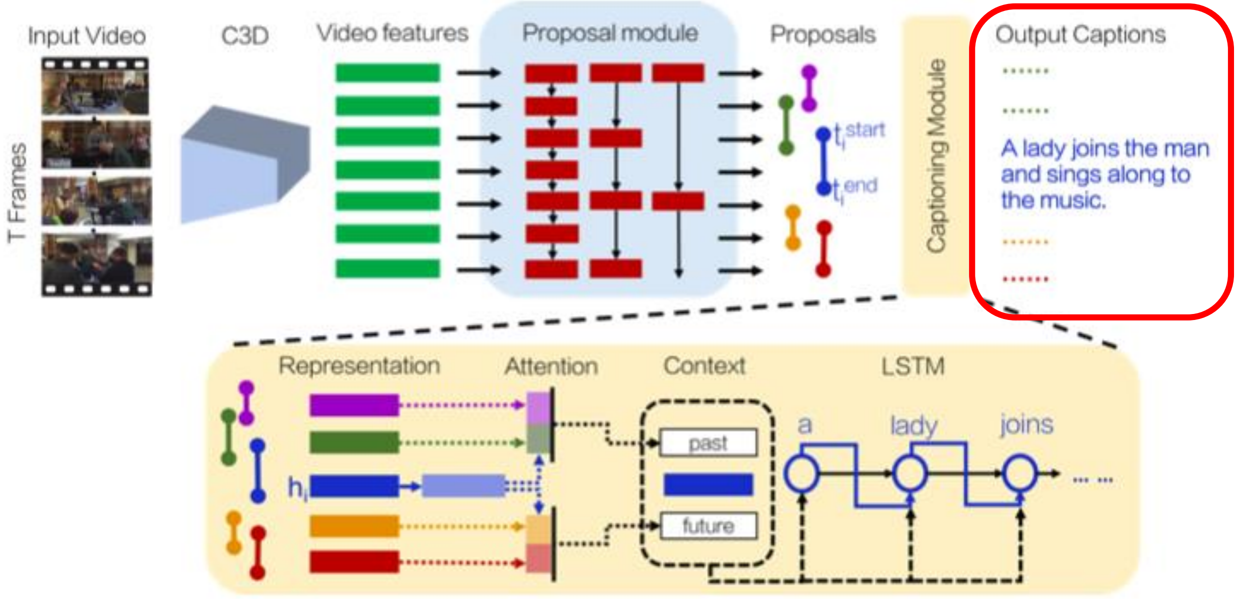
[Johnson et al. 2016]

- Language LSTM module + Attention

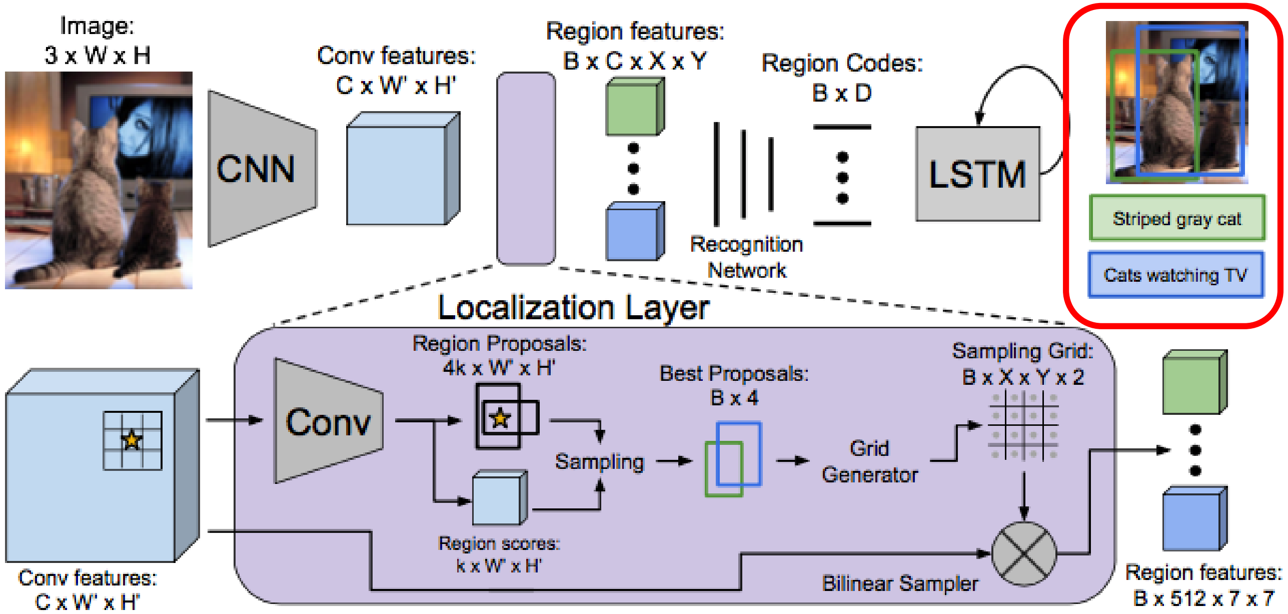
- Language LSTM module

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Analogy with Dense Captioning in Images



[Krishna et al. 2017]



[Johnson et al. 2016]

- Output: Captions
- Output: Bounding boxes

[Johnson et al., 2016] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (oral).*

Outline

- Definitions – Dense Captioning Events in Videos
- Purposes and Applications of Dense Captioning Videos
- Related Works on Video Analysis
- Methodology and Neural Networks Architectures
- **Previous and Novel Datasets**
- Metrics and Results

Dataset – ActivityNet Captions

- **ActivityNet Captions**
  - 20k videos
  - 3.65 ± 1.79 sentences per video,
  - 13.48 ± 6.33 words per sentences
  - 94.6% content coverage
  - Human and non-human activities
  - Focuses on Verbs and Actions
  - Annotated by AMT


```
{
version: "VERSION 1.0",
results: {
  "v_5n7NCViB5TU": [
      {
      sentence: "One player moves all around the net holding the ball", # String description of an event.
      timestamp: [1.23,4.53] # The start and end times of the event (in seconds).
      },
      {
      sentence: "A small group of men are seen running around a basketball court playing a game".
      timestamp: [5.24, 18.23]
      }
  ]
}
external_data: {
  used: true, # Boolean flag. True indicates the use of external data.
  details: "First fully-connected layer from VGG-16 pre-trained on ILSVRC-2012 training set", # This string
 details what kind of external data you used and how you used it.
}
}
```

Dataset – Comparison

Dataset	Domain	# videos	Avg. length	# sentences	Des.	Loc. Des.	paragraphs	overlapping
UCF101 [45]	sports	13k	7s	-	-	-	-	-
Sports 1M [21]	sports	1.1M	300s	-	-	-	-	-
Thumos 15 [15]	sports	21k	4s	-	-	-	-	-
HMDB 51 [25]	movie	7k	3s	-	-	-	-	-
Hollywood 2 [28]	movie	4k	20s	-	-	-	-	-
MPII cooking [40]	cooking	44	**600s**	-	-	-	-	-
ActivityNet [4]	human	20k	**180s**	-	-	-	-	-
MPII MD [39]	movie	68k	4s	68,375	✓	-	-	-
M-VAD [47]	movie	49k	6s	55,904	✓	-	-	-
MSR-VTT [55]	**open**	**10k**	20s	**200,000**	✓	-	-	-
MSVD [6]	**human**	**2k**	10s	**70,028**	✓	-	-	-
YouCook [7]	cooking	88	-	2,688	✓	-	-	-
Charades [43]	**human**	**10k**	30s	16,129	✓	-	-	-
KITTI [12]	driving	21	30s	520	✓	✓	-	-
TACoS [36]	cooking	127	**360s**	11,796	✓	✓	-	-
TACos multi-level [37]	cooking	127	**360s**	52,593	✓	✓	✓	-
ActivityNet Captions (ours)	**open**	**20k**	**180s**	**100k**	✓	✓	✓	✓

Outline

- Definitions – Dense Captioning Events in Videos
- Purposes and Applications of Dense Captioning Videos
- Related Works on Video Analysis
- Methodology and Neural Networks Architectures
- Previous and Novel Datasets
- **Metrics and Results**

Metrics

- **Definition : $\boldsymbol{n}$-gram**
  - Contiguous sequence of n word from a given sequence of text
  - **Example**: “The cat is on the mat” -> sentence of length $k = 6$ words
    - **Uni**-gram: {The, cat, is, on, the, mat} -> **6** elements of **length 1** word
    - **Bi**-gram: {The cat, cat is, is on, on the, the mat} -> **5** elements of **length 2** words
    - **Tri**-gram: {The cat is, cat is on, is on the, on the mat} -> **4** elements of **length 3** words
    - **Quadri**-gram: {The cat is on, cat is on the, is on the mat} -> **3** elements of **length 4** words
    - …
    - n-gram: -> $(\boldsymbol{k}-\boldsymbol{n})$ **elements of length $\boldsymbol{n}$**

Metrics

- **Definition : $\boldsymbol{n}$-gram**
  - $h_k(s)$: number of time a n-gram $\omega_k$occurs in a sentence s

  - Given a set of $\boldsymbol{m}$ **descriptions** $S_i = \{s_{i1}, \dots, s_{im}\}$ for an image i
  - Given **a candidate description** c_i for an image i

  - We define:
    - $h_k(s_{ij})$: number of time an n-gram ω_k occurs in the reference sentence s_{ij}
    - $h_k(c_i)$: number of time an n-gram $\omega_k$occurs in the candidate sentence c_i

Metrics

- **BLEU** (**B**i**L**ingual **E**valuation **U**nderstudy)
  - **Machine Translation** Metric

$$P_n(c_i, S_i) = \frac{\sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_k h_k(c_i)}$$

$$b(C, S) = \begin{cases} 1 & \text{if } l_C > l_S \\ e^{1 - l_S / l_C} & \text{if } l_C \le l_S \end{cases}$$

$$BLEU_N(c_i, S_i) = b(c_i, S_i) \exp\left(\sum_{n=1}^{N} w_n \log P_n(c_i, S_i)\right)$$

k: index for the set of possible n-gram of length n

l_C: total length of candidate sentence C

l_S: closest length of references sentences S

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). ACL.

Metrics

- **ROUGE** (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation)
  - **Text Summarization** Metric

  - Simple n-gram recall-based method

$$ROUGE_N(c_i, S_i) = \frac{\sum_j \sum_k \min(h_k(c_i), h_k(s_{ij}))}{\sum_j \sum_k h_k(s_{ij})}$$

[Lin, 2004] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (Vol. 8).

Metrics

- **ROUGE** (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation)
  - **Text Summarization** Metric
  - Based on the Longest Common Subsequences (LCS)
    - set of word shared by two sentences which occur in the same order.

$$ROUGE_L(c_i, S_i) = \frac{(1+\beta^2)R_l P_l}{R_l + \beta^2 P_l}$$

$$P_l = \max_j \frac{l(c_i, s_{ij})}{|c_i|} \qquad R_l = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|}$$

[Lin, 2004] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (Vol. 8).

Metrics

- **ROUGE** (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation)
  - **Text Summarization** Metric

  - Based on Skip bi-grams:
    - pairs of ordered words in a sentence

$$ROUGE_S(c_i, S_i) = \frac{(1+\beta^2)R_s P_s}{R_s + \beta^2 P_s}$$

$$P_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(c_i)} \qquad R_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(s_{ij})}$$

[Lin, 2004] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (Vol. 8).

Metrics

- **METEOR** (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering)
  - **Machine Translation** Metric

  - Attempt to improve BLEU:
    - The Lack of Recall
    - Use of Higher Order N-grams
    - Lack of Explicit Word-matching Between Translation and Reference
    - Use of Geometric Averaging of N-grams

[Banerjee et al., 2005] Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*

Metrics

- **METEOR** (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering)
  - ch: Chunks
  - m: matching
  - $\alpha, \gamma, \theta, \delta$: hyper parameters
    - tuned for a given language

Language	α	β	γ	δ
English	0.85	0.20	0.60	0.75
Czech	0.95	0.20	0.60	0.80
German	0.95	1.00	0.55	0.55
Spanish	0.65	1.30	0.50	0.80
French	0.90	1.40	0.60	0.65
Universal	0.70	1.40	0.30	0.70

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m}$$

$$Pen = \gamma \left(\frac{ch}{m}\right)^{\theta}$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)}$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})}$$

$$METEOR = (1 - Pen) F_{mean}$$

[Banerjee et al., 2005] Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*

Metrics

- **CIDEr** (**C**onsensus-based **I**mage **D**escription **E**valuation)

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right)$$

$$\mathrm{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}$$

- $|I|$: number of images
- $\boldsymbol{g^n}(s_{ij})$: vector formed by $g_k(s_{ij})$

$$\mathrm{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \mathrm{CIDEr}_n(c_i, S_i)$$

[Vedantam et al., 2015] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575).

Metrics

- **Python**: How to use those metrics?

```
from pycocoevalcap.bleu.bleu import Bleu
from pycocoevalcap.meteor.meteor import Meteor
from pycocoevalcap.rouge.rouge import Rouge
from pycocoevalcap.cider.cider import Cider
    …
cur_res = self.tokenizer.tokenize(res[vid_id])
cur_gts = self.tokenizer.tokenize(gts[vid_id])
score, scores = scorer.compute_score(cur_gts, cur_res)
```

[Chen et al., 2015] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.0032*).

Results

- **Dense Captioning Events**
  - How well can we detect multiple events and describe them?
- **Event Localization**
  - How well can we localize an event with this Dense Captioning system?
- **Video Retrieval**
  - How well can we recover the correct set of sentence given a video?
  - How well can we recover the correct video given a set of sentence?

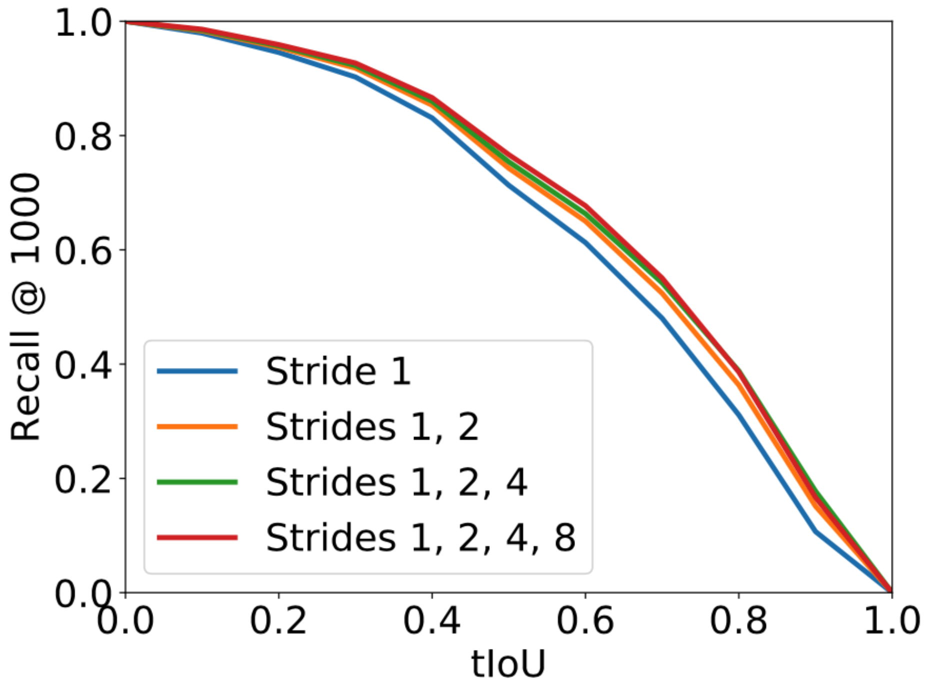
Results - Dense Captioning Events

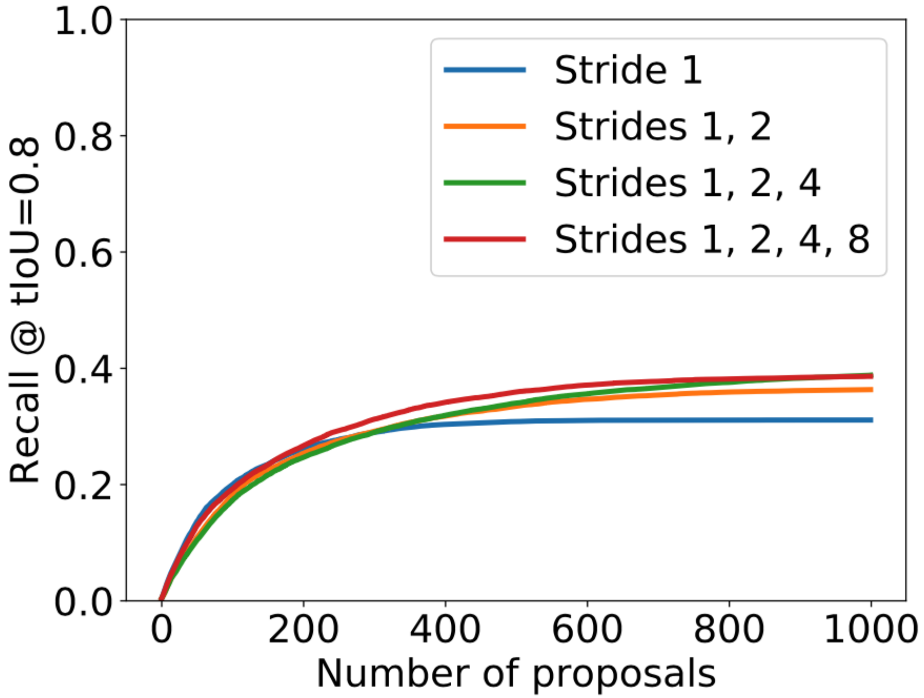
- BLEU(B@k), METEOR (M) and CIDEr (C) captioning scores for the task of **dense-captioning events** respect to state-of-the-art techniques

	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**

Results

- Event Localization





Results

- Video and Paragraph retrieval.
  - R@k measures the recall at varying thresholds k
  - Med. rank measures the median rank the retrieval.

	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

- Introduced the task of dense-captioning events
- Challenges were:
  - Events can occur within a second or last up to minutes
  - Events in a video are related to one other
- Contributions:
  - New variant of an existing proposal module at different time scale in a single pass
  - Captioning module attends over neighboring events
  - Release of a new dataset for dense-captioning events: ActivityNet Captions

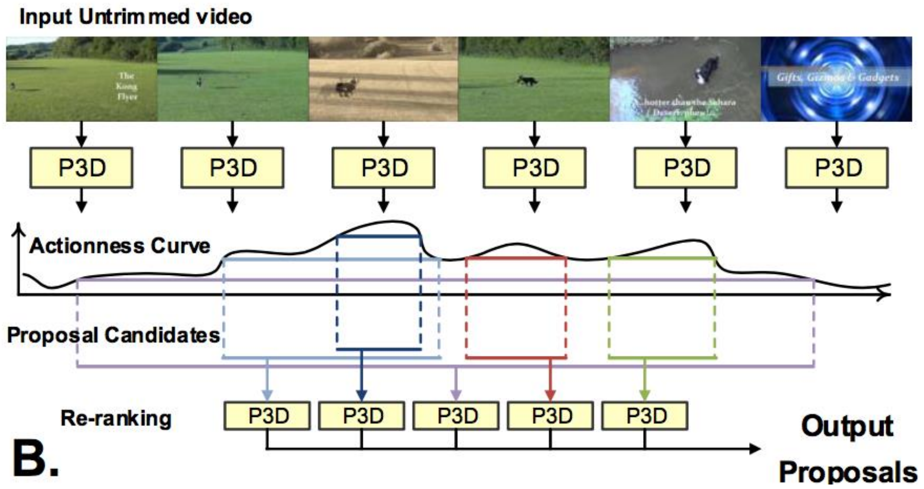
What's next?

- Activity Net Workshop results in CVPR'17

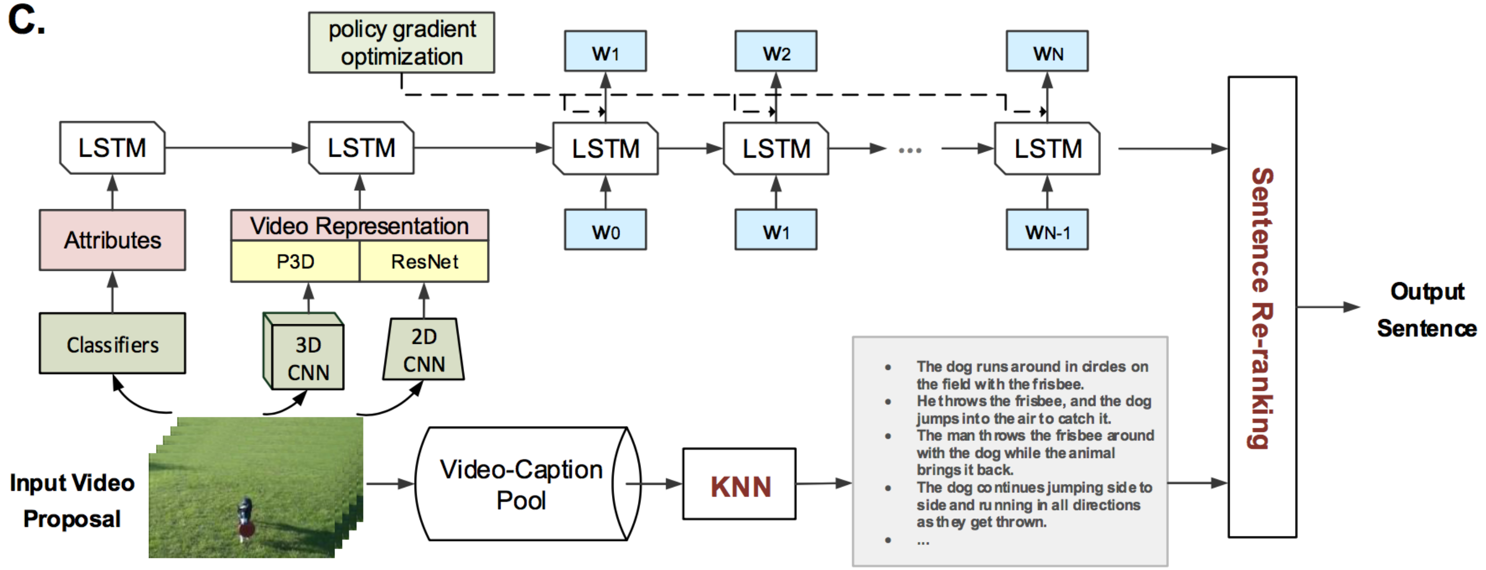
Ranking	Username	Organization	Upload time	Avg. Meteor
1	Ting Yao	Multimedia Search and Mining Group, MSRA	2017-07-17 10:54:38	12.8404
2	Cong Guo	University of Science and Technology of China	2017-07-08 10:49:59	9.8714
3	Qin Jin	RUC-CMU	2017-07-08 03:31:13	9.6154
4	Wonder Woman	Marvel	2017-06-28 04:22:37	1.72147
5	Shizhe Chen	Renmin University of China	2017-07-07 17:30:39	0

What’s next?

- Ting Yao - Multimedia Search and Mining Group, MSRA
  - **Avg. METEOR=12.8404**



Proposal technique based on P3D



Dense Captioning based on P3D, ResNet, Attribute Classifiers and Video-Caption Pooling

Dense-Captioning Events in Videos

Silvio Giancola

Thursday, August 3rd 2017