

# Visual Data Analysis, Understanding, and Mining

**Tao Mei (梅 涛)**

JD AI Research (京东AI研究院)  
CCDM 2018 Tutorial, Aug 6, 2018

## Outline

### Part I:

- Recent advances in vision and language (15 min)
- Image to language (captioning & poetry) (45 min)
- Break (15 min)

### Part II

- Video to language (captioning & commenting) (45 min)
- Visual question answering (15 min)
- Break (15 min)

### Part III

- Image and video generation (generation & translation) (15 min)
- Datasets and evaluations (15 min)
- Open issues and Q&A (10 min)

# Computer Vision

## Since the beginning of Artificial Intelligence

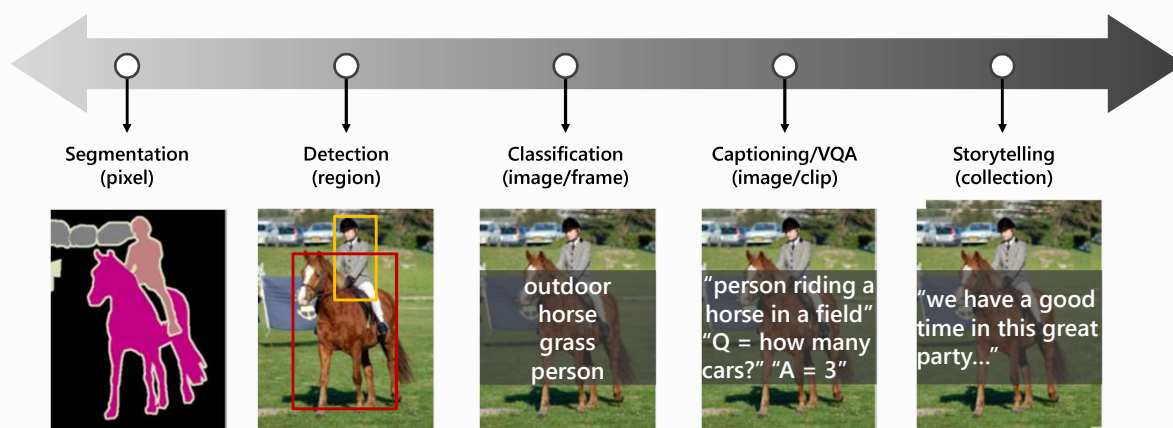


"Connect a television camera to a computer and get the machine to **describe** what it sees."

—Marvin Minsky (1966)

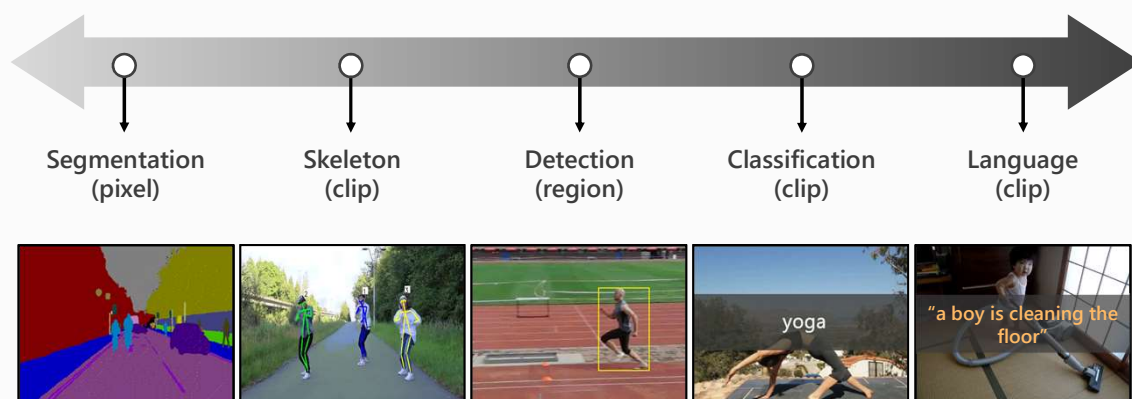
3

## Computer vision: image understanding

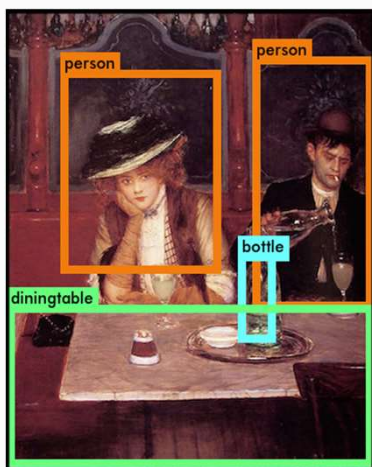


4

# Computer vision: video understanding



## Recent advances in object detection



| Approach                          | Pascal 2007 (mAP) | Speed   |            |
|-----------------------------------|-------------------|---------|------------|
| DPM<br>[Felzenszwalb, CVPR'10]    | 33.7              | .07 FPS | 14 s/img   |
| R-CNN<br>[Girshick, CVPR'14]      | 66.0              | .05 FPS | 20 s/img   |
| Fast R-CNN<br>[Girshick, ICCV'15] | 70.0              | .5 FPS  | 2 s/img    |
| Faster R-CNN<br>[Ren, NIPS'15]    | 73.2              | 7 FPS   | 140 ms/img |
| YOLO<br>[Redmon, CVPR'16]         | 69.0              | 45 FPS  | 22 ms/img  |
| YOLO 9000<br>[Redmon, CVPR'17]    | 76.8              | 67 FPS  | 15 ms/img  |





## Recent advances in image captioning



*"a group of zebras grazing in a field with a rainbow in the sky"*

[Yao, Pan, Li, Mei, ECCV'18]



*"a little girl holding an umbrella in a shopping cart"*

[Yao, Pan, Li, Mei, ECCV'18]

9

## Image to Language



双排扣大衣略带一点中性的帅气，加上legging，曲线美尽显。



黑色风衣是经典必备的基础款，显瘦显气质，绝对不会穿出错，搭配白色衬衫和深色小脚裤，简单的搭配很显气质，知性而利落。

the sea says: my love fear  
my soul can sail with sea  
for thee shall sea thee in the sea

[Liu, Fu & Mei, 2018]

四野里无人徘徊  
而人们不懂  
到时候我自己也涌出悲哀

诗人们弹出的心曲  
就在这苦酸的世界里  
晒太阳的懒猪

[微软小冰]



## Video Captioning



*"a group of people are dancing"*  
[Pan and Mei, CVPR'16'17]



*"I love baseball"*  
*"That's how to play baseball"*  
*"That's an amazing play"*  
[Li, Yao, Mei, MM'16]



*"Not just beautiful"*  
*"You are so beautiful"*  
*"Goddess doesn't need plastic surgery"*  
[Li, Yao, Mei, MM'16]

11

## Vision and Language

*"describe what a 3-year-old child sees"*  
classification, detection, segmentation

*"describe what a 5-year-old child sees"*  
vision to language  
visual question-answering

*"do what a 7-year-old child does?"*  
drawing, generation, design?



# Language to Vision

text to image

"this small bird has *short* beak and *dark stripe* down the top, the wings are a *mix of brown, white and black* and the upper breast is *white* and has *black strips*."

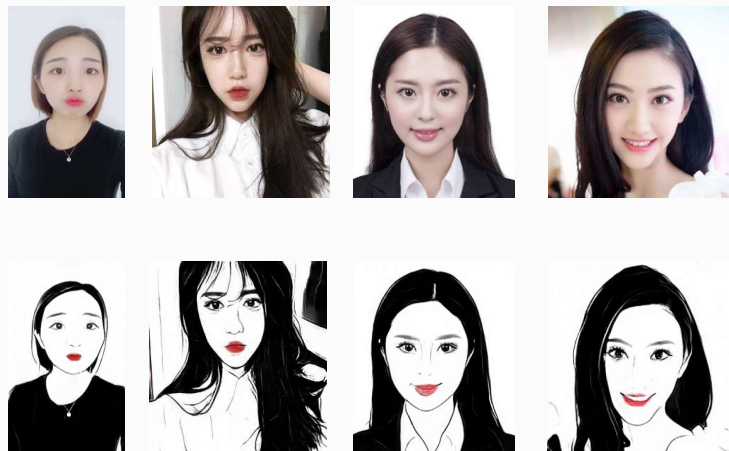
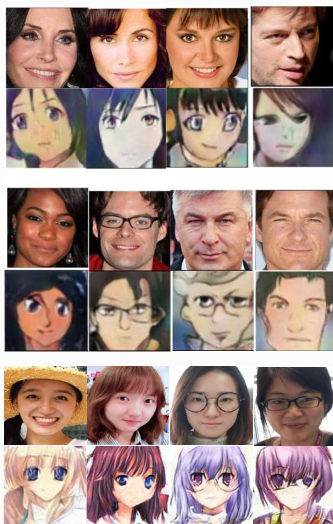


b/w to color



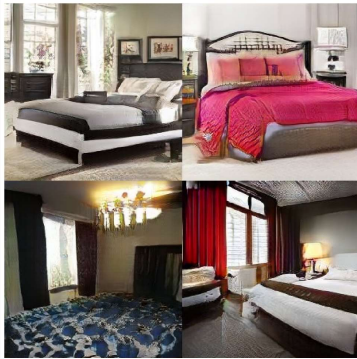
[Ma, Fu, Chen & Mei, CVPR 2018]

# Image to Image



[Ma, Fu, Chen & Mei, CVPR 2018]

## Vision to Design



synthesized bedroom layout  
[Karas, ICLR'18]



sketch to cartoon [CVPR'18]



sketch to design

## Video generation from captions



digit 8 is moving left and right



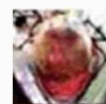
digit 2 is up and down and digit 0  
is left and right



a cook puts noodles into  
some boiling water



a person is cutting beef

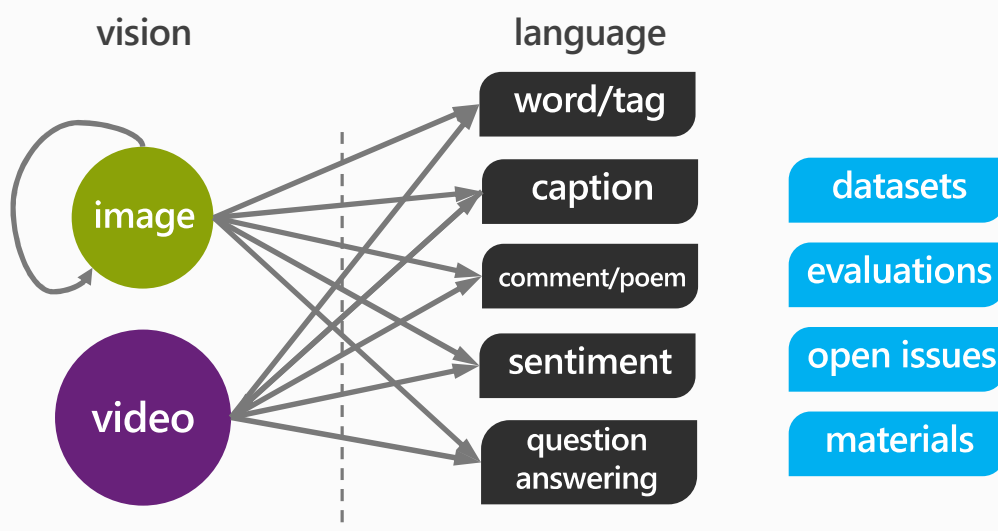


a chef is stirring a soup

Video generation w/ TGANs-C [Pan, Yao, Mei, ACM MM 2017]

16

## This tutorial will talk about



17

## Outline

### Part I:

- Recent advances in vision and language (15 min)
- **Image to language (recognition & captioning & poetry)** (45 min)
- Break (15 min)

### Part II

- Video to language (recognition & captioning & commenting) (45 min)
- Visual question answering (15 min)
- Break (15 min)

### Part III

- Image and video generation (generation & translation) (10 min)
- Datasets and evaluations (10 min)
- Open issues and Q&A (5 min)

18



## Fine-grained image recognition: challenges



CUB-200-2011 [P. Welinder et.al. 2010]



Stanford-Dogs [Fei-fei Li et.al. 2011]

### Challenges:

- ❑ Discriminative region localization
  - Localizing the very marginal visual differences from highly-localized regions
- ❑ Fine-grained feature learning
  - Describing the subtle visual differences by representative visual features

### The state-of-the-art general recognition network (ResNet-152)

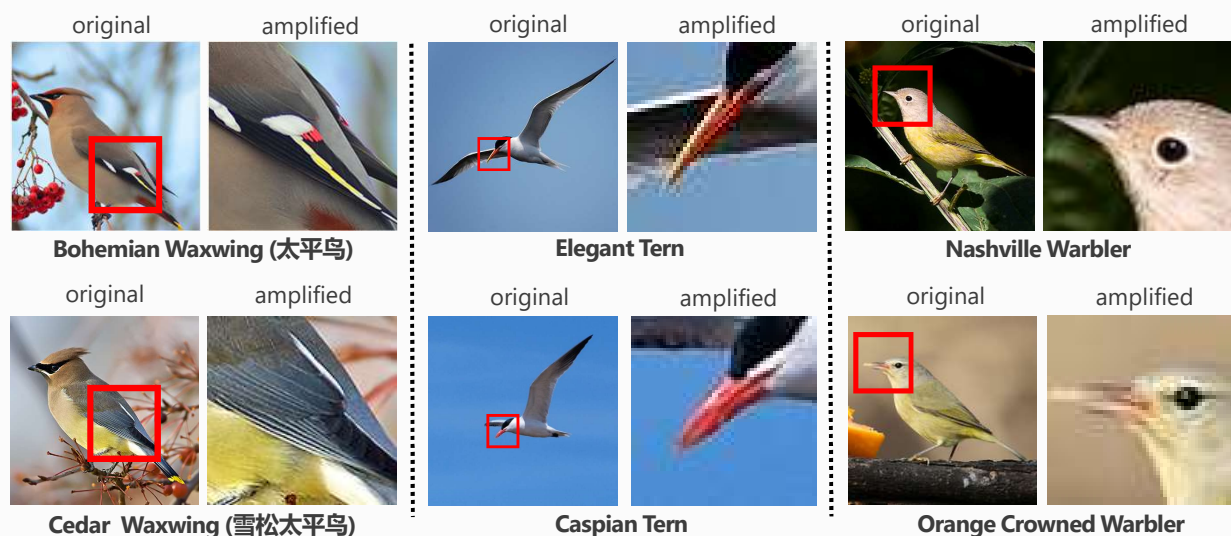
| Dataset  | CUB-Birds<br>(200 categories) | Stanford-Cars<br>(196 categories) | Stanford-Dogs<br>(120 categories) |
|----------|-------------------------------|-----------------------------------|-----------------------------------|
| Accuracy | 77.3%                         | 87.5%                             | 87.3%                             |

### Our fine-grained image recognition network (CVPR & ICCV 2017)

| Dataset  | CUB-Birds<br>(200 categories) | Stanford-Cars<br>(196 categories) | Stanford-Dogs<br>(120 categories) |
|----------|-------------------------------|-----------------------------------|-----------------------------------|
| Accuracy | <b>86.5%</b> <b>↑9.2%</b>     | <b>93.8%</b> <b>↑6.3%</b>         | <b>89.3%</b> <b>↑2.0%</b>         |

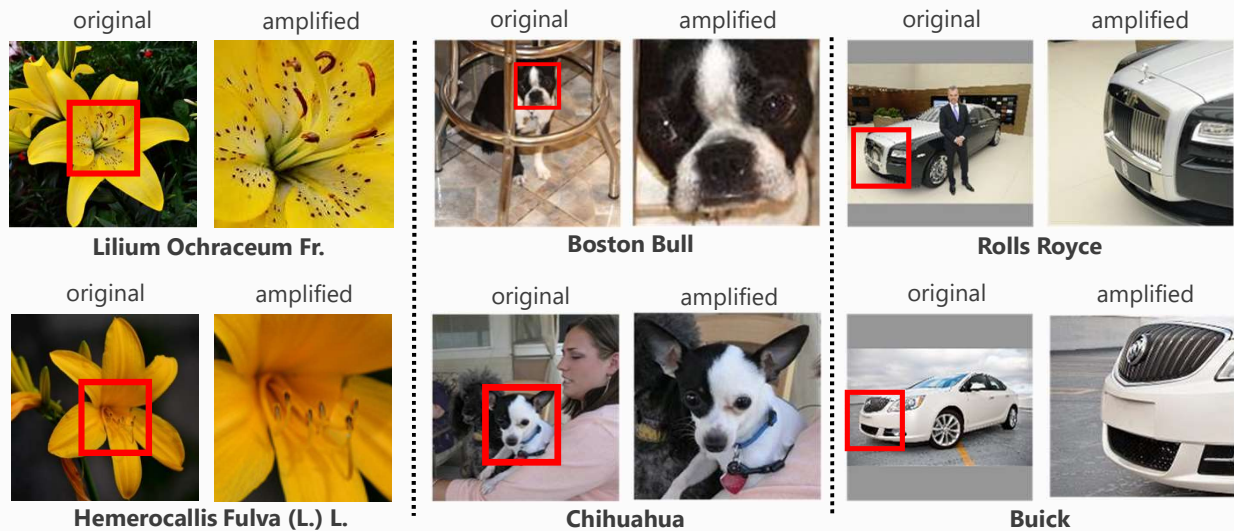
19

## Fine-grained image recognition: challenges



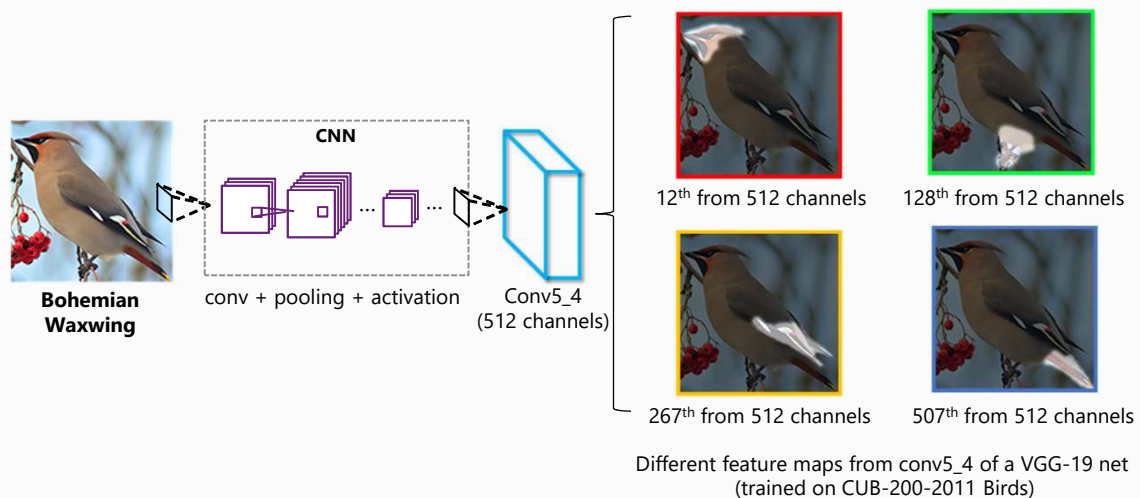
Observation: the subtle visual differences can be clearly represented from amplified parts.

## Fine-grained image recognition: challenges



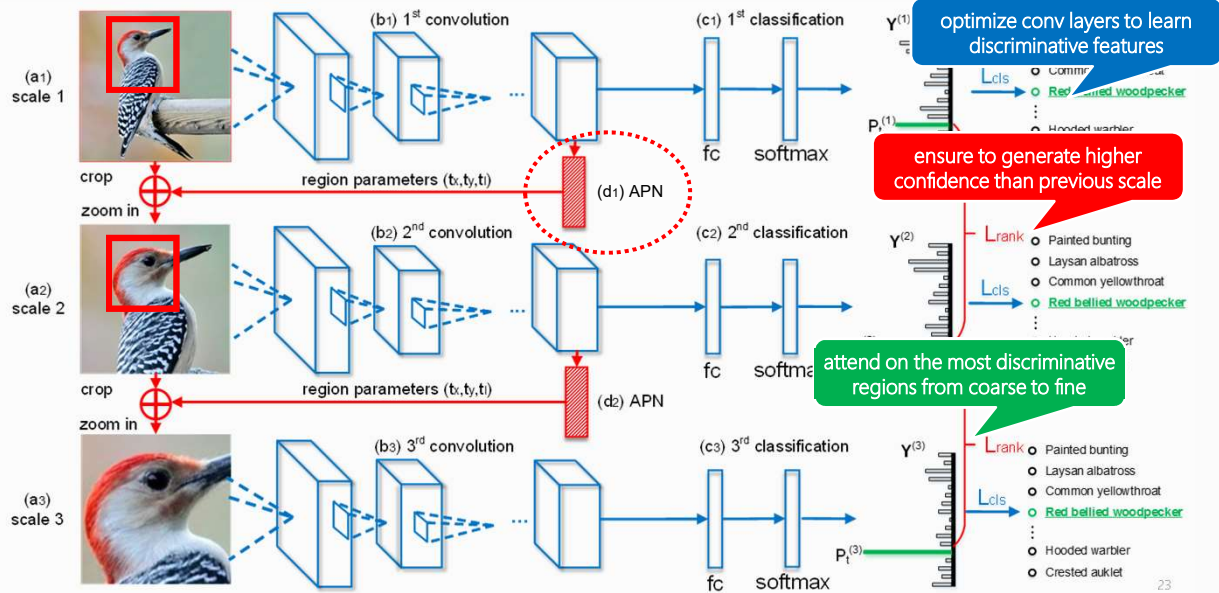
Observation: the subtle visual differences can be clearly represented from amplified parts.

## Fine-grained image recognition



Observation: the subtle visual differences can be clearly represented from amplified parts.

## Look Closer To See Better: Recurrent Attention CNN [Fu, CVRP'17]



## Recurrent Attention CNN: Multi-task Formulation

### Classification Net

$$p(\mathbf{X}) = f(\mathbf{W}_c * \mathbf{X})$$

### Attention Proposal Net

$$[t_x, t_y, t_l] = g(\mathbf{W}_e * \mathbf{X})$$

### Input for the Next Scale

$$\mathbf{X}^{att} = \mathbf{X} \odot \mathbf{M}(t_x, t_y, t_l)$$

$$\mathbf{X}_{(i,j)}^{amp} = \sum_{\alpha, \beta=0}^1 |1 - \alpha - \{i/\lambda\}| |1 - \beta - \{j/\lambda\}| \mathbf{X}_{(m,n)}^{att}$$

where  $\mathbf{M}$  indicates an attention mask.

### Optimization

$$L(\mathbf{X}) = \sum_{s=1}^3 \{L_{cls}(\mathbf{Y}^{(s)}, \mathbf{Y}^*)\} + \sum_{s=1}^2 \{L_{rank}(p_t^{(s)}, p_t^{(s+1)})\}$$

where  $L_{cls}$  indicates classification loss in each scale,  $L_{rank}$  denotes pair-wise ranking loss.

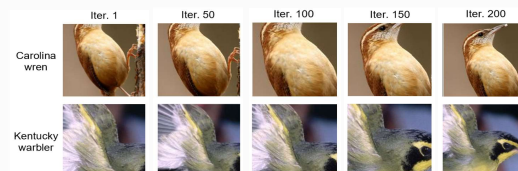


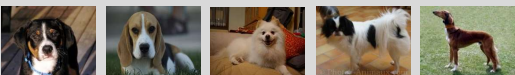



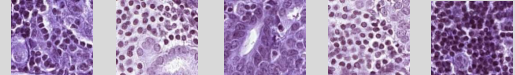


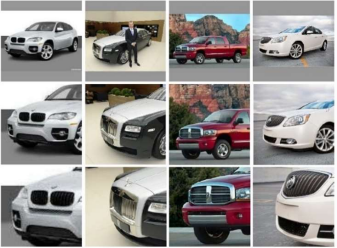


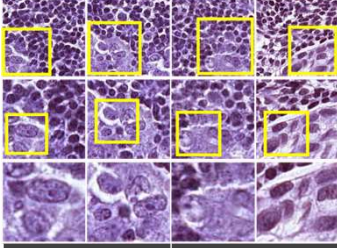


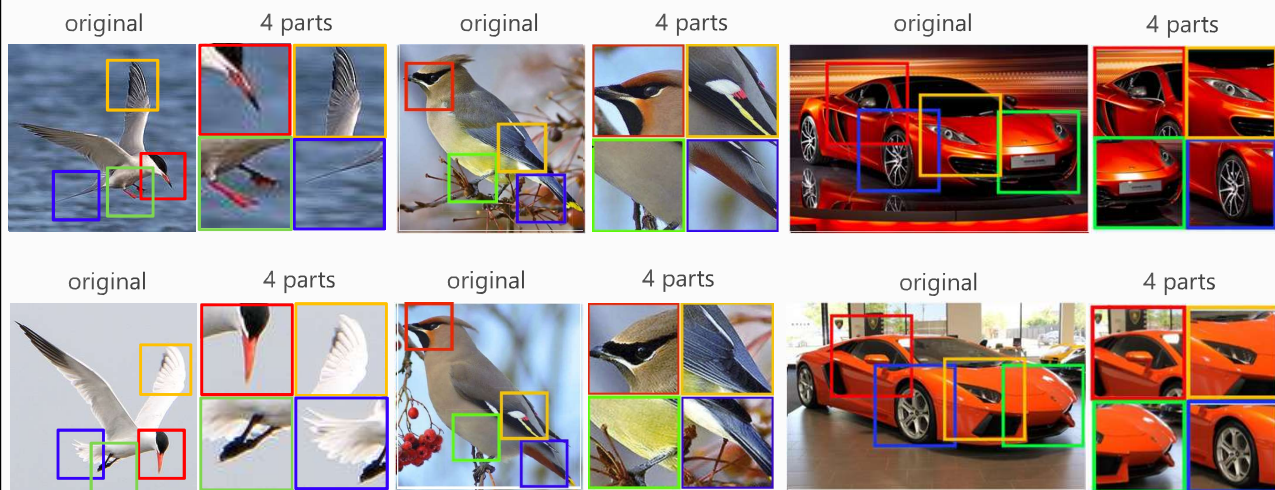
Illustration of learning process for region attention.



| Evaluation       | Domain  | # Category | # Training | # Testing  |
|------------------|---|------------|------------|------------|
| 1. Flower-102    |  | 102        | 2,040      | 6,000      |
| 2. CUB-200-2011  |  | 200        | 5,994      | 5,794      |
| 3. Stanford-Dogs |  | 120        | 12,000     | 8,580      |
| 4. Stanford-Cars |  | 196        | 8,144      | 8,041      |
| 5. FGVC-Aircraft |  | 100        | 6,667      | 3,333      |
| 6. VIREO Food    |  | 172        | 66,144     | 33,072     |
| 7. Camelyon16    |  | binary     | 270 slides | 130 slides |

|   |   |  |
|---|---|--|
|  <div> <div>Bird</div> <div>Accuracy</div> <div>state-of-art 82.6%</div> <div>Ours (CVPR) 85.3%</div> </div>     |  <div> <div>Dog</div> <div>Accuracy</div> <div>state-of-art 84.2%</div> <div>Ours (CVPR) 87.3%</div> </div>  |  <div> <div>Car</div> <div>Accuracy</div> <div>state-of-art 89.1%</div> <div>Ours (CVPR) 92.5%</div> </div> |
|  <div> <div>Aircraft</div> <div>Accuracy</div> <div>state-of-art 84.1%</div> <div>Ours (CVPR) 87.8%</div> </div> |  <div> <div>Food</div> <div>Accuracy</div> <div>state-of-art 82.0%</div> <div>Ours (CVPR) 84.8%</div> </div> |  <div> <div>Tumor</div> <div>AUC</div> <div>state-of-art 97.0%</div> <div>Ours (CVPR) 95.8%</div> </div>    |

## Extension: from single to multi- attention [Zheng, ICCV'17]



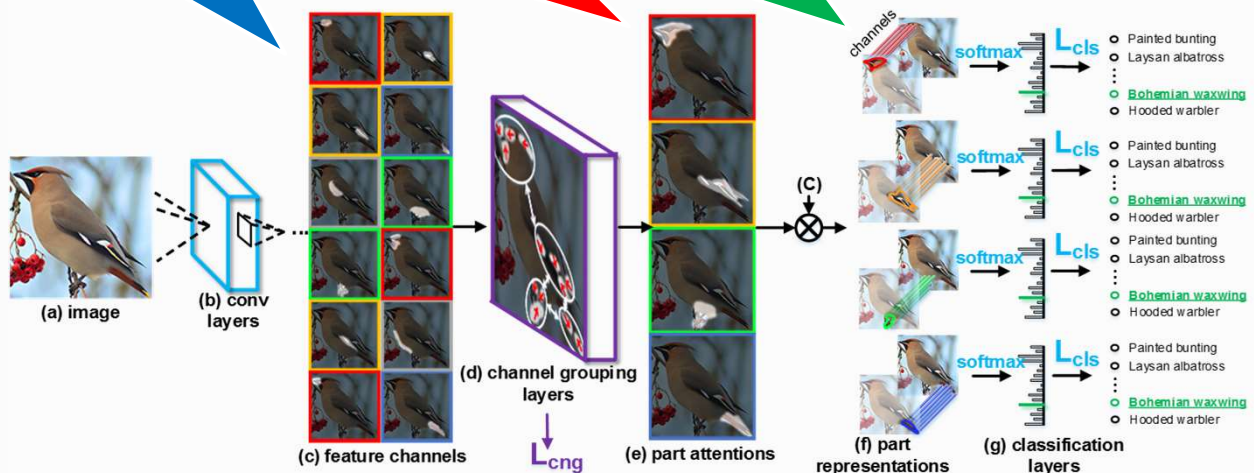
27

## Multi-Attention Recurrent CNN [Zheng, ICCV'17]

different channels exhibit diff. responses in diff. parts.

channel grouping groups small parts into semantic regions

region-based pooling makes image-level label into region-level label





# Loss Function

$$L(X) = \sum_{i=1}^N [L_{cls}(Y^{(i)}, Y^*)] + L_{cng}(M_1, \dots, M_N)$$

where  $N$  is the number of parts,

$$M_i(X) = \text{sigmoid}\left(\sum_{j=1}^c d_j [W * X]_j\right)$$

$L_{cls}$  indicates classification loss for each part,

$L_{cng}$  denotes channel grouping loss.

$d_j$  is the weight vector over different channels,

$[\cdot]_j$  denotes the  $j^{th}$  feature channel.

The specific form of channel grouping:

$$L_{cng}(M_i) = Dis(M_i) + \lambda Div(M_i)$$

$Dis(\cdot)$  is a distance function, which encourages a compact distribution for each  $M_i$ .

$$Dis(M_i) = \sum_{(x,y) \in M_i} m_i(x,y) [\|x - t_x\|^2 + \|y - t_y\|^2]$$

where  $(t_x, t_y)$  is the coordinate from the peak response of  $M_i$ .




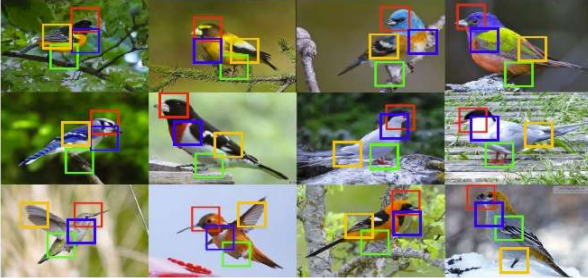
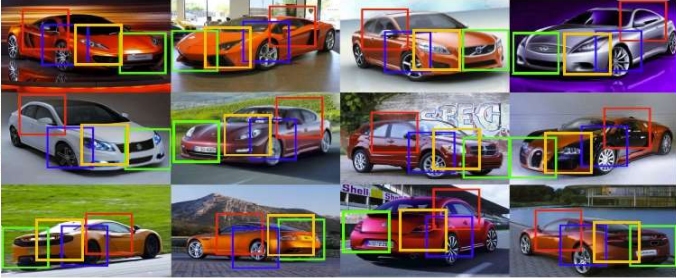

$Div(\cdot)$  is designed to favor a diverse attention distribution from different part attention maps.

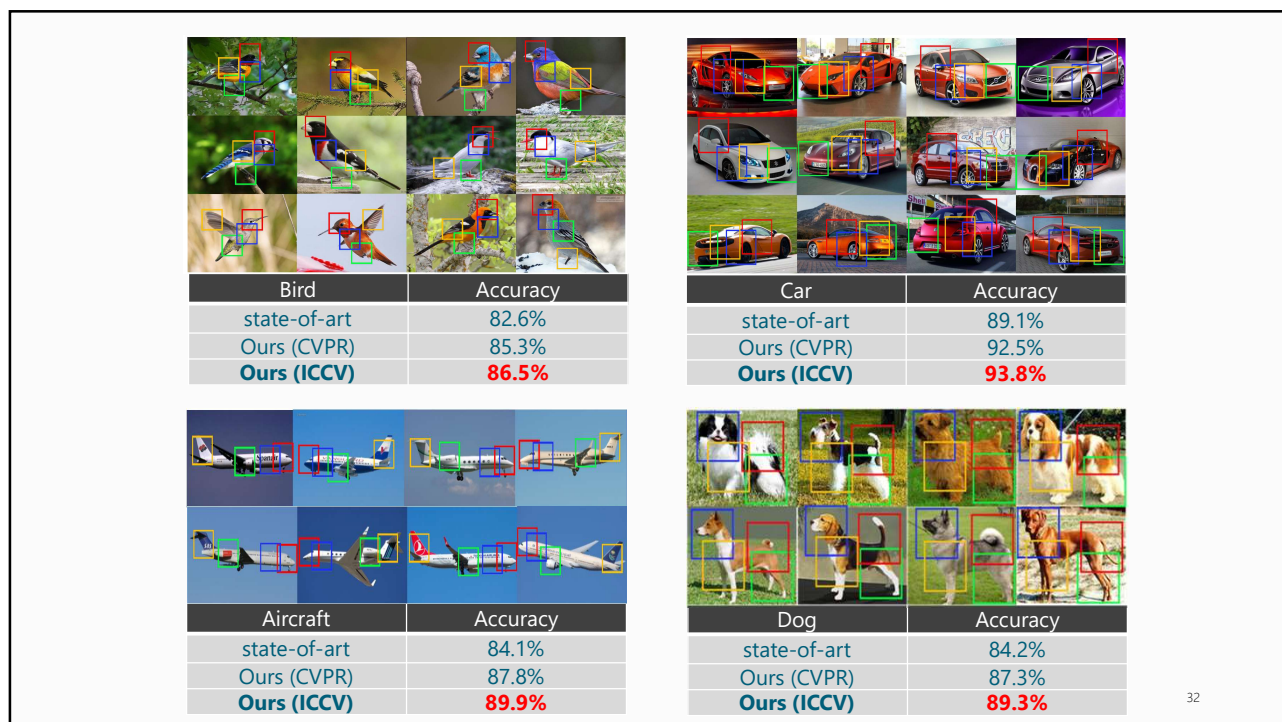
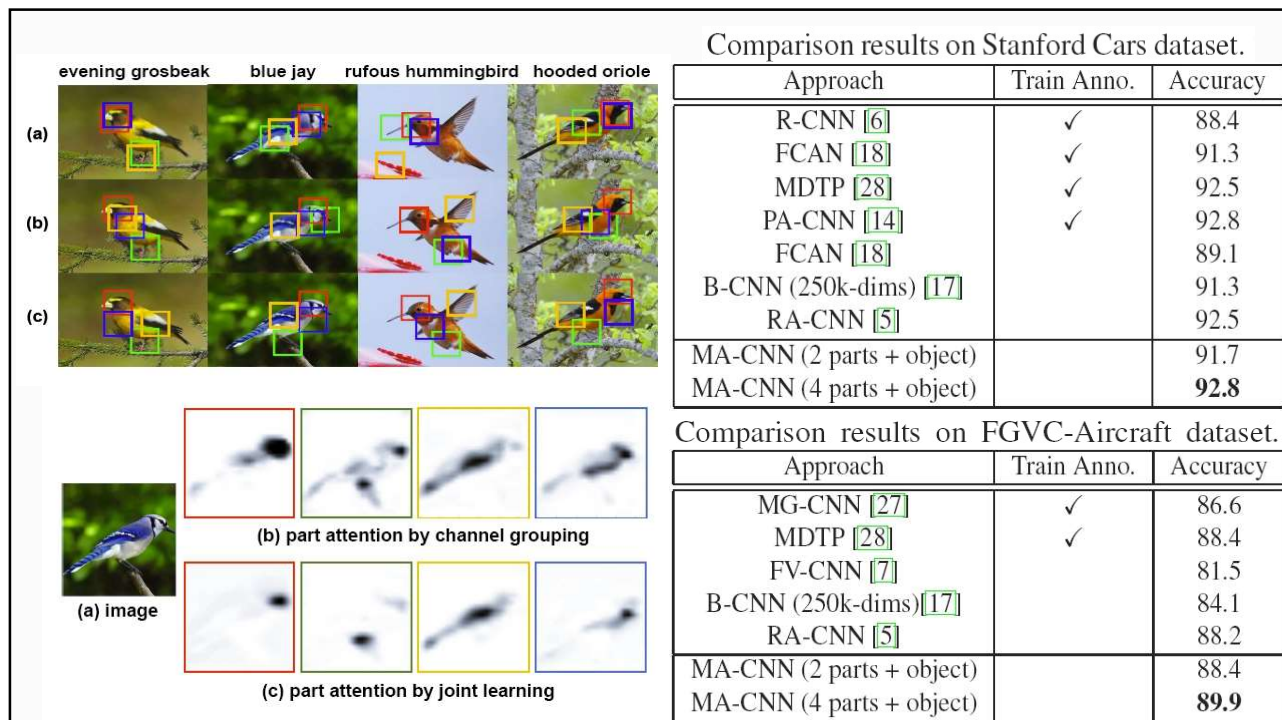
$$Div(M_i) = \sum_{(x,y) \in M_i} m_i(x,y) [\max_{k \neq i} m_k(x,y) - mrg]$$

where  $mrg$  represents a margin.

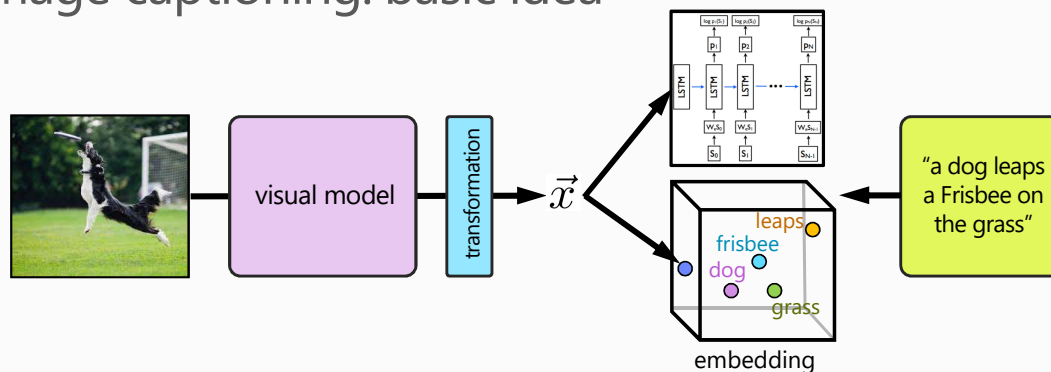


29

| Evaluation  | Domain  | # category | # Training | # Testing |
|---|---|------------|------------|-----------|
| 1. CUB-200-2011   |  | 200        | 5,994      | 5,794     |
| 2. Stanford-Cars  |  | 196        | 8,144      | 8,041     |
| 3. FGVC-Aircraft  |  | 100        | 6,667      | 3,333     |
| <div>   </div> <div> <span>(a) CUB-Birds</span> <span>(b) Stanford-Cars</span> </div> <div>  <span>(c) FGVC-Aircraft</span> </div> |   |            |            |           |



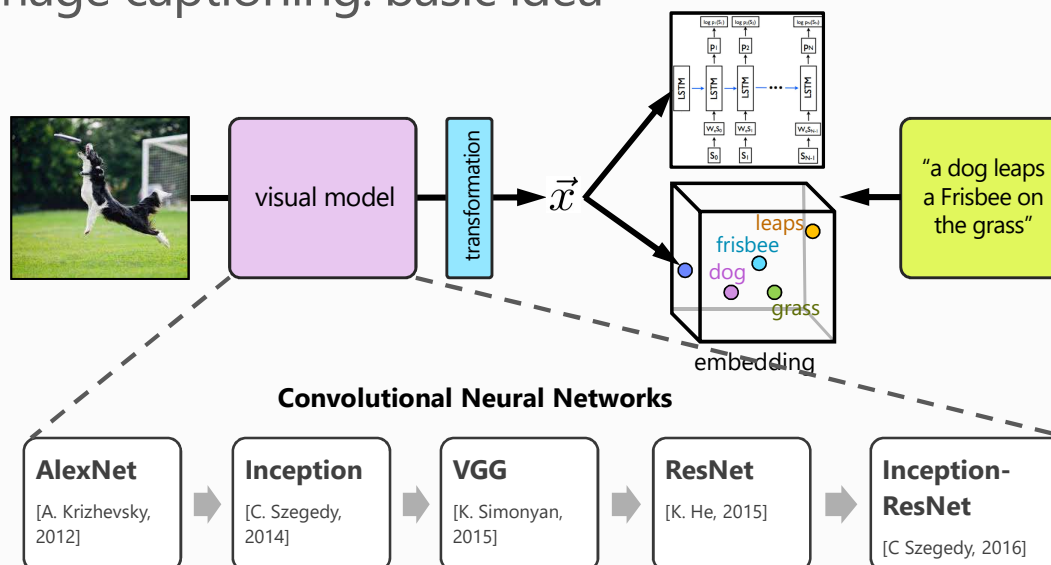
## Image captioning: basic idea



- Transforming an image to a vector in visual space
  - CRF, CNN, Semantic Vector, CNN+Attention
- Transforming description to a vector in semantic space
  - Collection of words (BoW), sequence of words (RNN)
- Creating an embedding space
  - Language template (FGM, ME), RNNs (Encoder-Decoder), LSTM
- Methodologies
  - Search-based
  - Language template-based
  - Sequence learning-based
    - Generation: learning-decoder
    - Translation: encoder-decoder

33

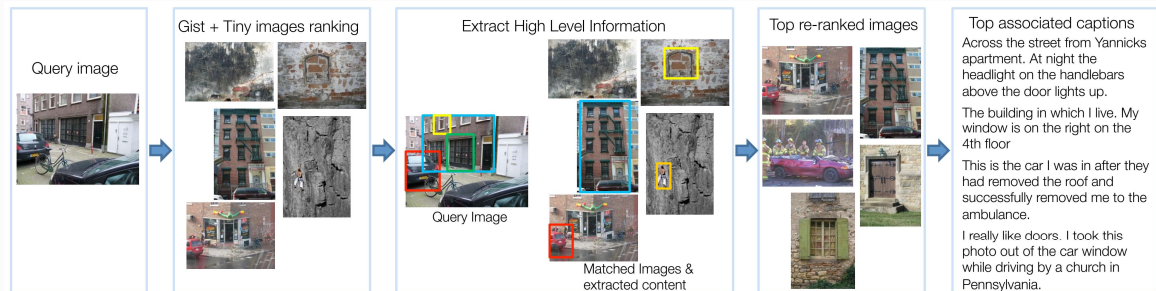
## Image captioning: basic idea



34

# Image captioning

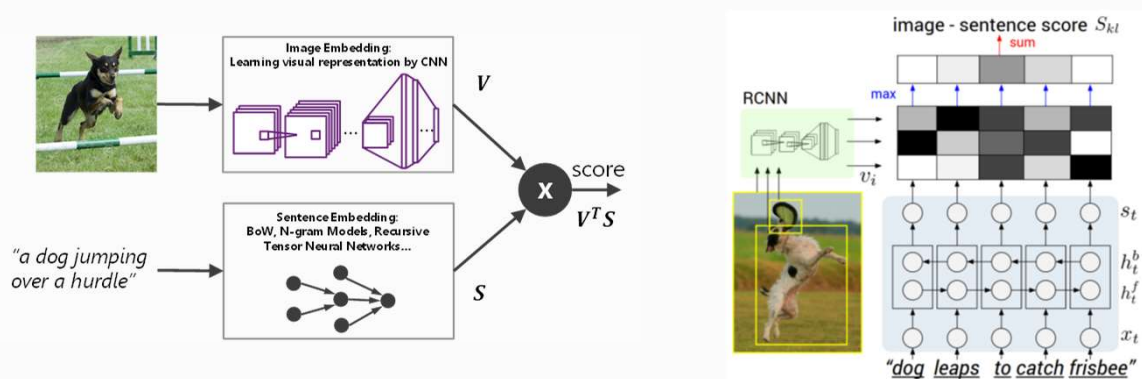
- Search-based approach [Farhadi, ECCV10; Ordonez, NIPS11; Frome, NIPS13; Socher, NIPS14; Karpapthy, CVPR15; Devlin, ACL15]



35

# Image captioning

- Search-based approach [Farhadi, ECCV10; Ordonez, NIPS11; Frome, NIPS13; Socher, NIPS14; Karpapthy, CVPR15; Devlin, ACL15]



36



# Image captioning

- Language template-based approach [Feng, ACL10; Yang, EMNLP11; Kulkarni, PAMI13; **Fang, CVPR15**]

## Image word detection (s-v-o)

Woman, crowd, cat, camera, holding, purple.

## Language generation (maximum entropy)

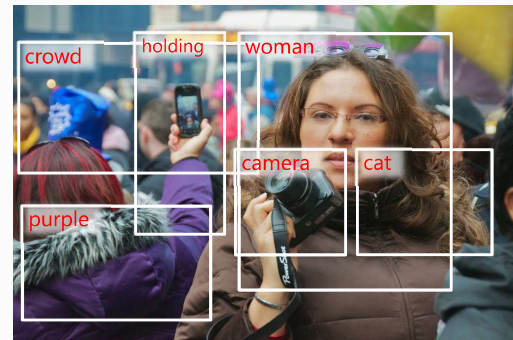
A purple camera with a woman.

A woman holding a camera in a crowd.

A woman holding a cat.

## Semantic re-ranking (deep embedding)

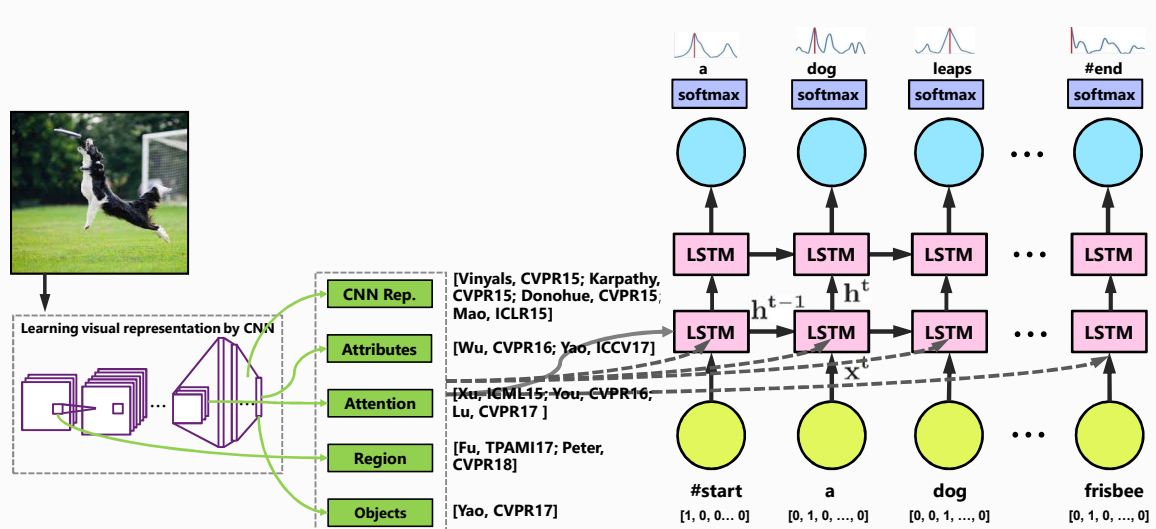
A woman holding a camera in a crowd.



37

# Sequence learning-based approach

[Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester16, UAdelaide16, Virginia Tech17, THU17, MSR17&18]

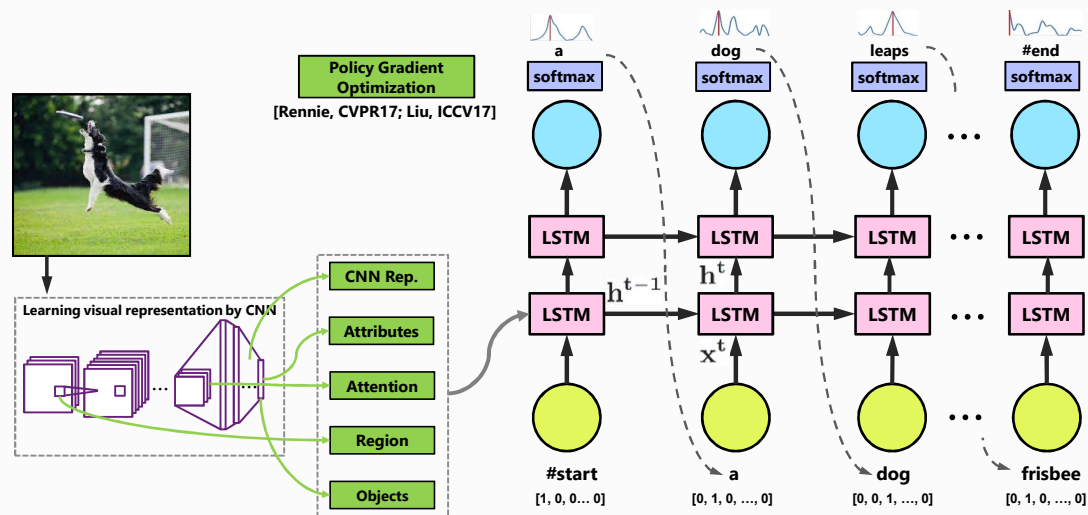


\* Note that this figure only shows prediction process.

38



# Sequence learning-based approach with reinforcement learning [IBM17, U of Oxford & Google17]



\* Note that this figure only shows prediction process.

39

## Image captioning with X

**X = visual attention**  
[Xu, ICML15; Lu, CVPR17]

A(1.00)



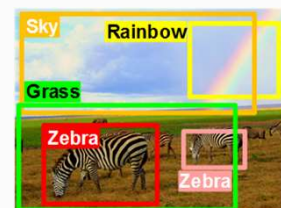
**X = visual attributes**  
[You, CVPR16; Wu, CVPR16; Yao, ICCV17]



**X = object/entity recognition**  
[Yao, CVPR17]



**X = region**  
[Fu, TPAMI17; Peter, CVPR18; Yao, ECCV18]



40

# Image captioning with X

## X = visual attention

[Xu, ICML15; Lu, CVPR17]

A(1.00)



## X = visual attributes

[You, CVPR16; Wu, CVPR16; Yao, ICCV17]



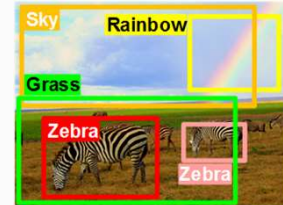
## X = object/entity recognition

[Yao, CVPR17]



## X = region

[Fu, TPAMI17; Peter, CVPR18; Yao ECCV18]



### Detected Objects:

cat, **suitcase**, clothes, bag, luggage, eyes ...

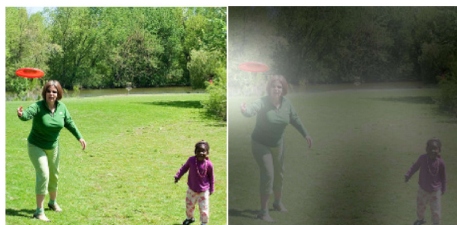
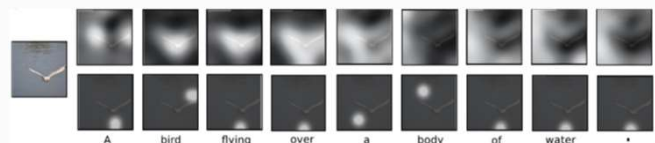
Image captioning: a cat sitting on top of a red chair

Novel object captioning: a cat laying on a **suitcase**

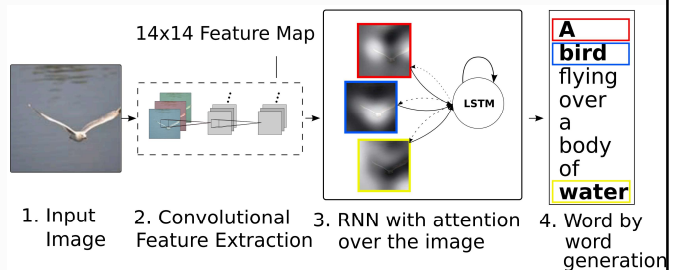
41

# Image Captioning with Visual Attention

- Image captioning with attention mechanism [Xu, ICML'15; Cho, 2015]
- Learning stochastic "hard" vs. deterministic "soft" attention




A woman is throwing a **frisbee** in a park.




## Image Captioning with Visual Attributes

- Visual attributes: a high-level representation w/ concept detector responses
  - Video search with high-level concepts [TRECVID, 2006]
  - Object bank for image classification [Li & Fei-Fei, NIPS'10]
  - High-level concepts for captioning and question-answering [Wu & Shen, CVPR'16]



**Attributes:**   
 [piano: 0.930] [hand: 0.71]  
 [music: 0.672] [keyboard: 0.624]  
**LSTM:** a man is playing a  
**LSTM-E:** a man is playing  
 a **piano**



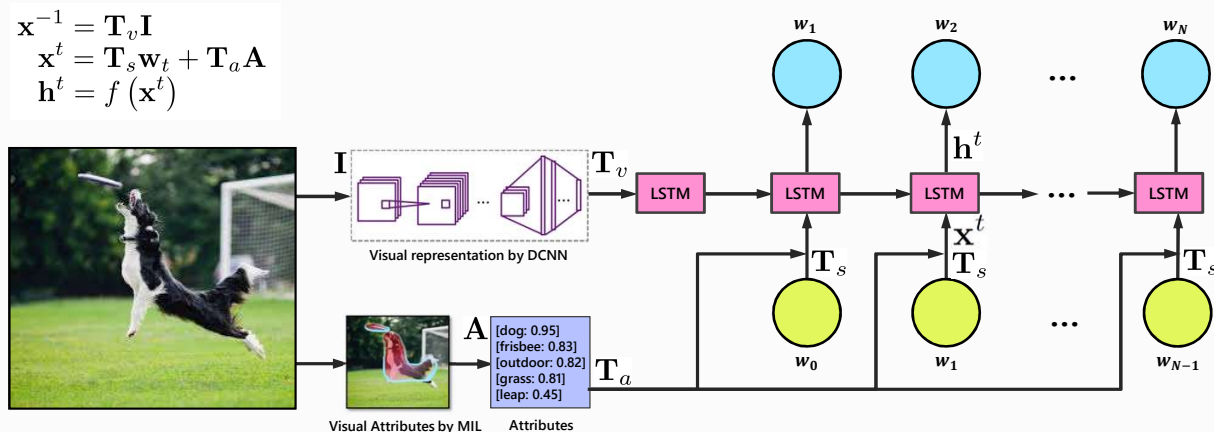
**Attributes:**   
 [bananas: 1] [market: 0.995] [bunch: 0.553] [table: 0.51] [flowers: 0.454]  
 [people: 0.431] [yellow: 0.377]  
**LSTM:** a group of people standing  
 around a market.  
**A-LSTM:** a group of people standing  
 around a bunch of **bananas**.

- Joint learning w/ recognizable attributes: relevance + coherence [Pan, CVPR'16]
  - Image captioning [A-LSTM]: explicitly emphasize attributes together with visual content
  - Video captioning [LSTM-E]: implicitly emphasize video content with "relevance" regularizer

43

## A-LSTM: image captioning w/ attribute-LSTM [Yao & Mei, arxiv16]

$$\begin{aligned} \mathbf{x}^{-1} &= \mathbf{T}_v \mathbf{I} \\ \mathbf{x}^t &= \mathbf{T}_s \mathbf{w}_t + \mathbf{T}_a \mathbf{A} \\ \mathbf{h}^t &= f(\mathbf{x}^t) \end{aligned}$$



44

# Image captioning

- [Leaderboard](#) of MS COCO image captioning
  - Rank 1 in both external and internal ranking lists, in terms of all performance metrics (July 21, 2017)
- COCO dataset
  - 123,287 images (82,783 for training + 40,504 for validation)
  - 5 sentences per image (AMT workers)

**COCO**  
Common Objects in Context

Home People Explore **Dataset** External

cocodataset@outlook.com

Overview Challenges Download Evaluate Leaderboard

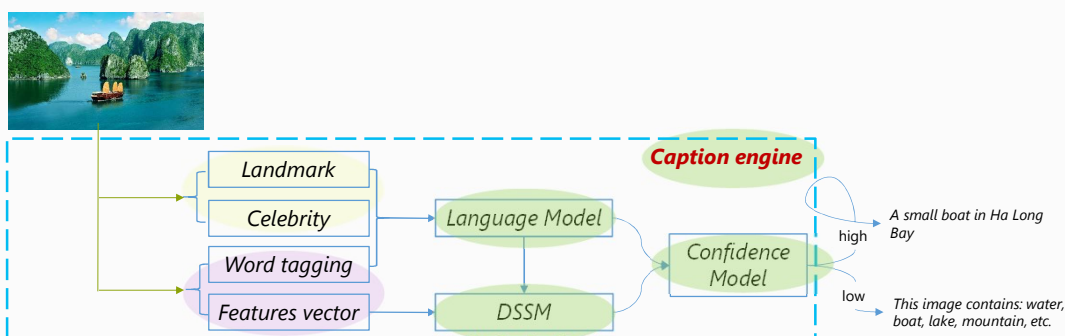
Table-C5 Table-C40 Challenge2015

Last updated: 07/12/2016. Please visit [CodaLab](#) for the latest results.

|                                  | CIDEr-D | Meteor | ROUGE-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | date       |
|----------------------------------|---------|--------|---------|--------|--------|--------|--------|------------|
| MSM@MSRA <sup>(1)</sup>          | 1.003   | 0.35   | 0.7     | 0.919  | 0.842  | 0.74   | 0.632  | 2016-06-08 |
| THU_MIG <sup>(2)</sup>           | 0.988   | 0.336  | 0.688   | 0.913  | 0.833  | 0.727  | 0.616  | 2016-06-03 |
| ChallS <sup>(3)</sup>            | 0.97    | 0.34   | 0.679   | 0.898  | 0.809  | 0.701  | 0.59   | 2016-05-21 |
| AugmentCNNwithDet <sup>(4)</sup> | 0.968   | 0.34   | 0.683   | 0.905  | 0.815  | 0.706  | 0.597  | 2016-03-29 |
| ATT <sup>(5)</sup>               | 0.958   | 0.335  | 0.682   | 0.9    | 0.815  | 0.709  | 0.599  | 2016-01-23 |
| MSRA-MSM <sup>(6)</sup>          | 0.954   | 0.328  | 0.677   | 0.901  | 0.815  | 0.705  | 0.591  | 2016-03-13 |
| Fukun_Jinjunqi <sup>(7)</sup>    | 0.946   | 0.336  | 0.68    | 0.902  | 0.817  | 0.711  | 0.601  | 2016-05-09 |
| Google <sup>(8)</sup>            | 0.946   | 0.346  | 0.682   | 0.895  | 0.802  | 0.694  | 0.587  | 2015-05-29 |

45

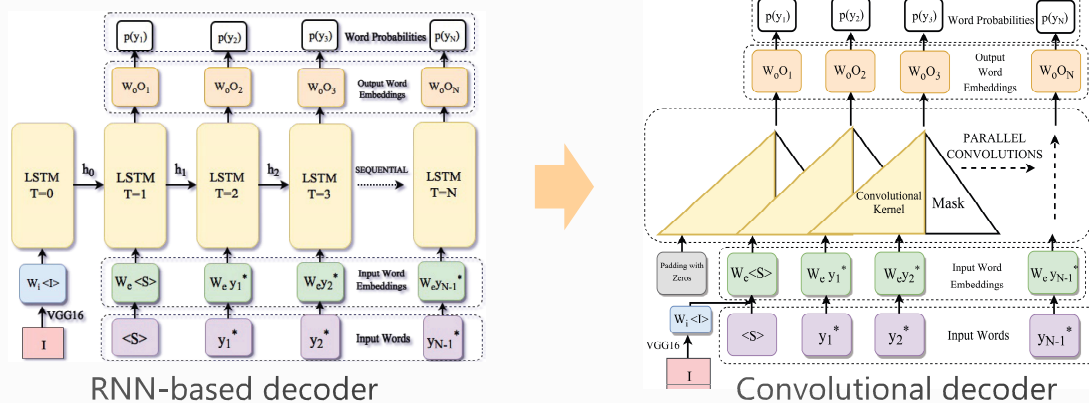
## Rich Image Captioning in the Wild [Tran, CVPR'16]



- Entity recognition: extreme classification w/ large set of celebrities (precision 99% coverage ~60%) [Guo, 2016]
- Language model: maximum entropy [Fang, CVPR15]
- Word tagging & feature: ResNet [He, CVPR16]
- Deep Structured Semantic Model [He, CIKM13]

46

## What's next? RNN -> Convolutions [Aneja, CVPR18]



- Ease the optimization
- Support parallel computation during training

Figure courtesy of [Aneja et al., CVPR18]

47

## Evaluations on COCO

- Performance comparison

| Model    | B@1  | B@2  | B@3  | B@4  | METEOR | ROUGE-L | CIDEr-D |
|----------|------|------|------|------|--------|---------|---------|
| LSTM     | 70.4 | 52.8 | 38.4 | 27.8 | 24.1   | 51.7    | 87.6    |
| CNN+Attn | 70.8 | 53.4 | 38.9 | 28.0 | 24.1   | 51.7    | 87.2    |

- Train time comparison

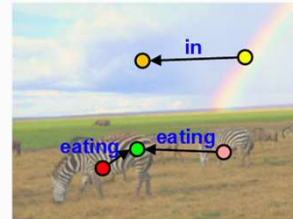
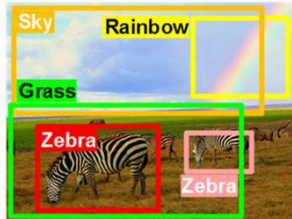
| Model    | #Parameters | Train time per epoch |
|----------|-------------|----------------------|
| LSTM     | 13M         | 1529s                |
| CNN+Attn | 20M         | 1620s                |

48



## What's next? visual relationship

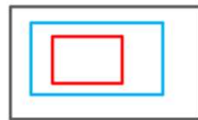
- Semantic relationship (20+ types from Visual Genome)



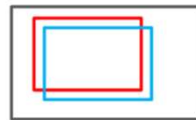
- Spatial relationship (11 types)



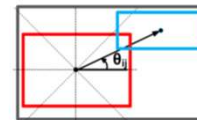
Class 1 (C1): Inside



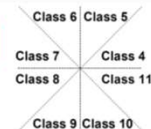
Class 2 (C2): Cover



Class 3 (C3): Overlap



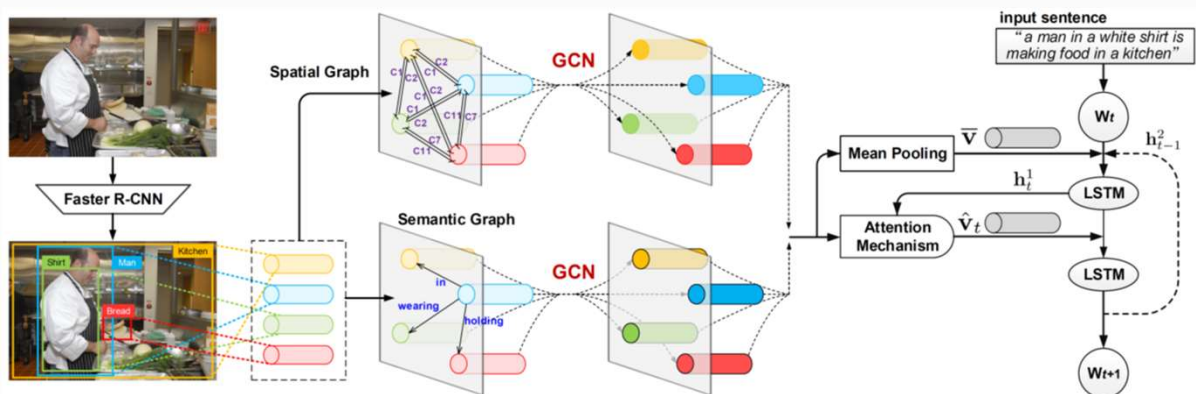
Class 4-11 (C4-11): Index =  $\left\lceil \frac{\theta_{ij}}{45^\circ} \right\rceil + 3$



49

## Image captioning with visual relationship

[GCN (Graph Convolutional Networks)-LSTM, Yao *et al.*, ECCV'18]

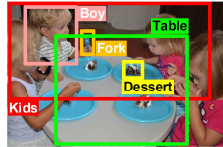
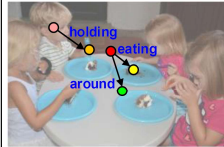
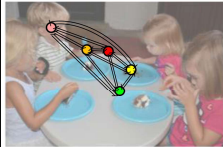
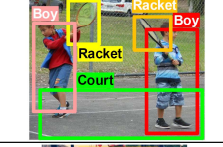

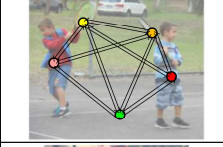
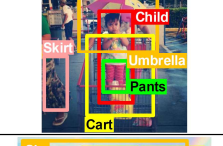

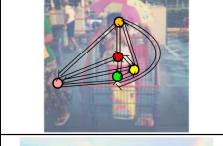
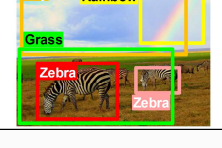

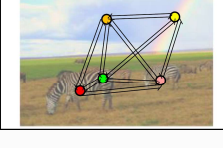


50

# Evaluations on COCO test server

| Model                    | Group                  | B@4         |             | METEOR      |             | ROUGE-L     |             | CIDEr-D      |              |
|--------------------------|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|--------------|
|                          |                        | c5          | c40         | c5          | c40         | c5          | c40         | c5           | c40          |
| <b>GCN-LSTM</b>          | <b>JD AIR, ECCV'18</b> | <b>38.7</b> | <b>69.7</b> | <b>28.5</b> | <b>37.6</b> | <b>58.5</b> | <b>73.4</b> | <b>125.3</b> | <b>126.5</b> |
| <b>Up-Down</b>           | MSR, CVPR'18           | 36.9        | 68.5        | 27.6        | 36.7        | 57.1        | 72.4        | 117.9        | 120.5        |
| <b>LSTM-A</b>            | MSRA, ICCV'17          | 35.6        | 65.2        | 27          | 35.4        | 56.4        | 70.5        | 116          | 118          |
| <b>Watson Multimodal</b> | IBM, CVPR'17           | 34.4        | 63.6        | 26.8        | 35.3        | 55.9        | 70.4        | 112.3        | 114.6        |
| <b>G-RMI</b>             | Google, ICCV'17        | 33.1        | 62.4        | 25.5        | 33.9        | 55.1        | 69.4        | 104.2        | 107.1        |
| <b>MetaMind/VT_GT</b>    | Salesforce, CVPR'17    | 33.6        | 63.7        | 26.4        | 35.9        | 55          | 70.5        | 104.2        | 105.9        |
| <b>DLTC@MSR</b>          | MSR, CVPR'17           | 33.1        | 63.1        | 25.7        | 34.8        | 54.3        | 69.6        | 100.3        | 101.3        |
| <b>reviewnet</b>         | CMU, NIPS'16           | 31.3        | 59.7        | 25.6        | 34.7        | 53.3        | 68.6        | 96.5         | 96.9         |
| <b>ATT</b>               | Rochester, CVPR'16     | 31.6        | 59.9        | 25          | 33.5        | 53.5        | 68.2        | 94.3         | 95.8         |
| <b>Google</b>            | Google, CVPR'15        | 30.9        | 58.7        | 25.4        | 34.6        | 53          | 68.2        | 94.3         | 94.6         |

51

|   |   |   |  |
|---|---|---|--|
|  |  |  | <p><b>GT:</b> a group of children sitting at a table eating pieces of cake</p> <p><b>LSTM:</b> a group of people sitting at a table with a cake</p> <p><b>Up-Down:</b> a group of children sitting at a table with a cake</p> <p><b>GCN-LSTM:</b> a group of children sitting at a table eating a cake</p> |
|  |  |  | <p><b>GT:</b> two young boys are playing with tennis rackets</p> <p><b>LSTM:</b> a young boy playing a game of tennis</p> <p><b>Up-Down:</b> two young boys playing tennis on a tennis court</p> <p><b>GCN-LSTM:</b> two young boys playing with tennis rackets on a court</p>                             |
|  |  |  | <p><b>GT:</b> a baby girl standing in a shopping cart holding an umbrella</p> <p><b>LSTM:</b> a woman walking down a street holding an umbrella</p> <p><b>Up-Down:</b> a little girl holding an umbrella in a street</p> <p><b>GCN-LSTM:</b> a little girl holding an umbrella in a shopping cart</p>      |
|  |  |  | <p><b>GT:</b> a herd of zebras grazing in a field and a rainbow</p> <p><b>LSTM:</b> a group of zebras standing in a field</p> <p><b>Up-Down:</b> a group of zebras and a rainbow in the sky</p> <p><b>GCN-LSTM:</b> a group of zebras grazing in a field with a rainbow in the sky</p>                     |

52

# Dense Image Captioning [Johnson & Karpathy, CVPR16]

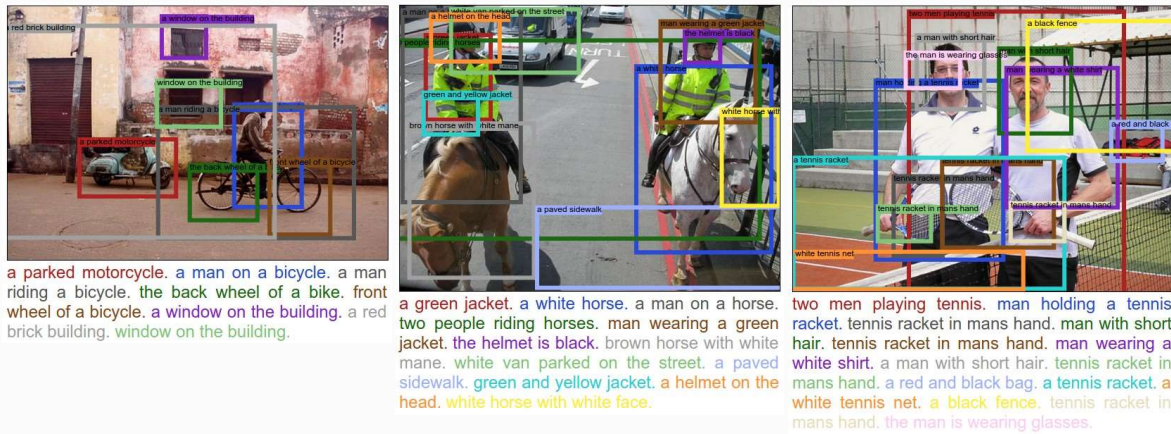
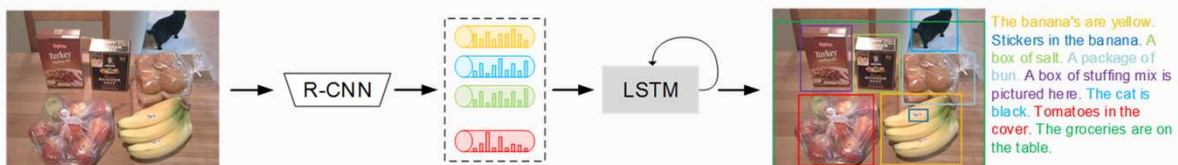


Figure courtesy of [Johnson, Karpathy, and Fei-Fei, CVPR16]

53

## RCNN + LSTM architecture [Johnson, CVPR16; Yang, CVPR17]



54

## Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training [Liu, MM'18]



### Description:

A **falcon** is **eating** during sunset.  
The falcon is **standing** on earth.

### Poem:

Like a falcon by the night  
**Hunting** as the black **knight**  
**Waiting** to take over the **fight**  
With all of it's mind and might



Collect two datasets:

- MultiM-Poem: image and poem pair dataset
- UniM-Poems: large poem corpus



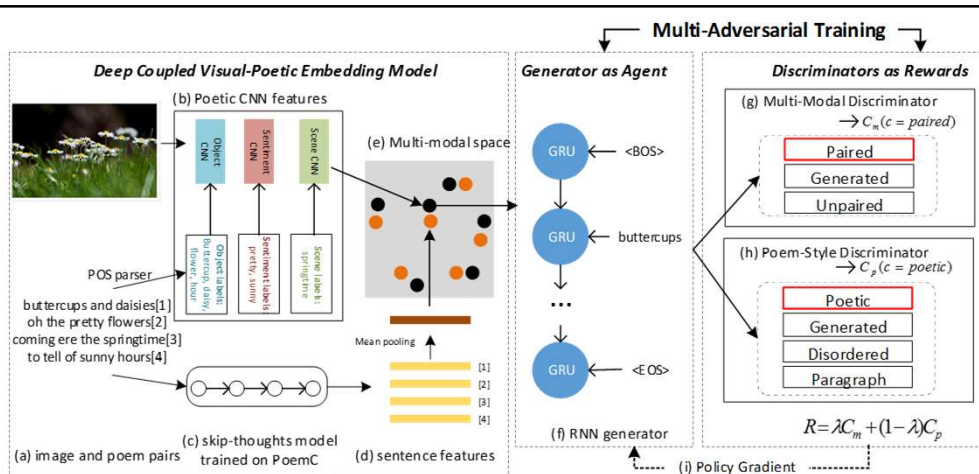
Deep coupled visual-semantic embedding

- Learn poetic clues from images to poems
- Enlarge image and poem pair dataset



CNN-RNN generator with multi-adversarial training (two discriminators) and further optimize it by policy gradient

55



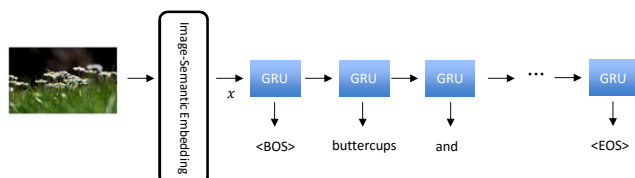
## Framework

- Deep coupled visual-poetic embedding model
- CNN-RNN-based poem generator as agent
- Two discriminators provide rewards to policy gradient

56



## Poem Generator as Agent



- Use image-semantic embedding from trained visual- embedding model
- We use the generated poem to compute rewards and take word selection in each time-step as an action
- The rewards of policy gradient is computed as the sum over all future actions

57

## Discriminators as Rewards



Multi-modal discriminator ( $D_m$ ):

Whether the generated sentences are related to the image

Three classes: paired, unpaired, generated



Poem-style discriminator ( $D_p$ ):

Whether the generated sentences are poems

Four classes: poetic, disordered, paragraphic, generated



Rewards:

Linear combination of probabilities of classifying generated poem as positive classes

$$R(\mathbf{y}|\cdot) = \lambda C_m(c = \textit{paired}|\mathbf{x}, \mathbf{y}) + (1-\lambda) C_p(c = \textit{poetic}|\mathbf{y})$$

58

## Multi-Adversarial Training

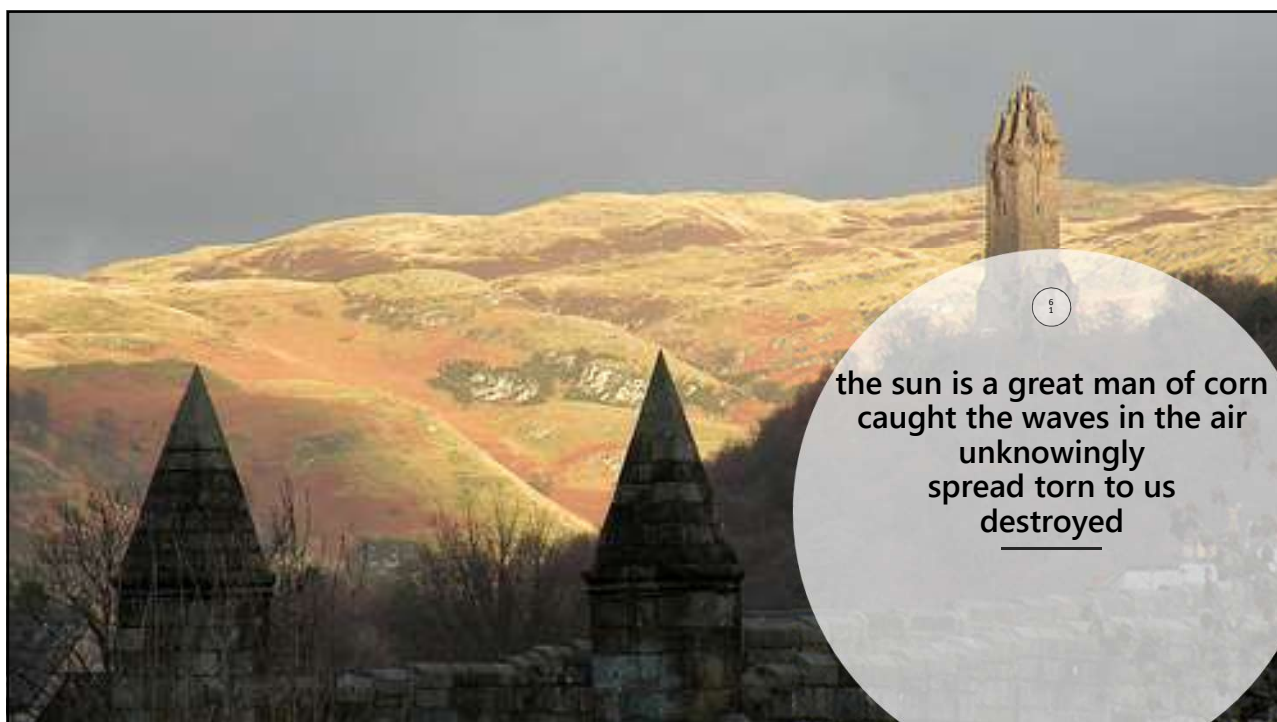
- Before adversarial training, pre-train a generator based on image captioning model
- Generator and discriminators are iteratively updated in an adversarial way
  - Generator: generate poems that have higher rewards for both discriminators
  - Discriminators: distinguish the generated poems from paired and poetic poems

59

in the morning light  
is warm and dark  
it is  
beautiful to be  
a dream



60



## Outline

### Part I:

- Recent advances in vision and language (15 min)
- Image to language (recognition & captioning & poetry) (45 min)
- Break (15 min)

### Part II

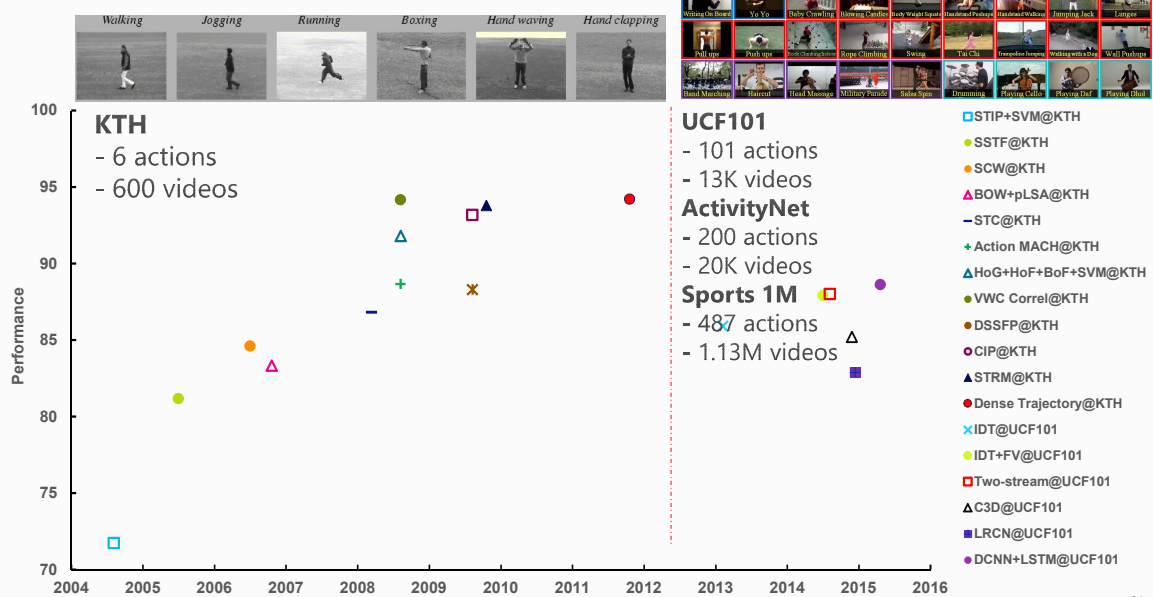
- Video to language (recognition & captioning & commenting) (45 min)
- Visual question answering (15 min)
- Break (15 min)

### Part III

- Image and video generation (generation & translation) (10 min)
- Datasets and evaluations (10 min)
- Open issues and Q&A (5 min)



## Video Classification



64

## Input video



## Output

- **Table-of-Content**  
story -> scene -> shot -> clip -> keyframe
- **Objects**  
man, skateboard
- **Highlights**  
skateboarding
- **Tags (hash code + object + action)**  
man, skateboarding, indoors, ceiling, light
- **Captions**  
"a man is doing a trick on a skateboarding"
- **Comments**  
Motivated me to go beyond my limits in skateboarding!
- **Visual question answering**  
[Q] "what is the man playing with?"  
[A] "skateboard"

Parsing

Representation  
Learning

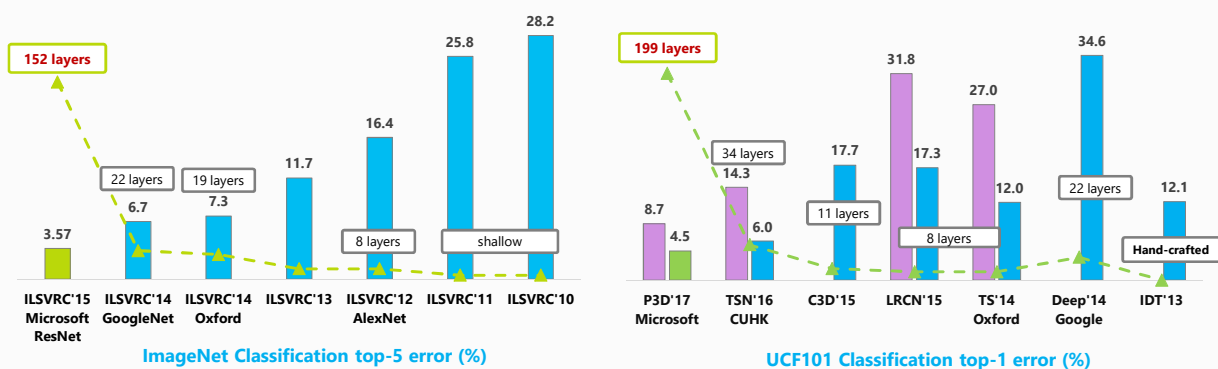
Action Recognition

Captioning

Commenting

65

## Learning video representation is harder than image!



66



## Video representation learning

2011

### Hand-crafted feature

Action recognition by dense trajectories. [Wang, CVPR'11]

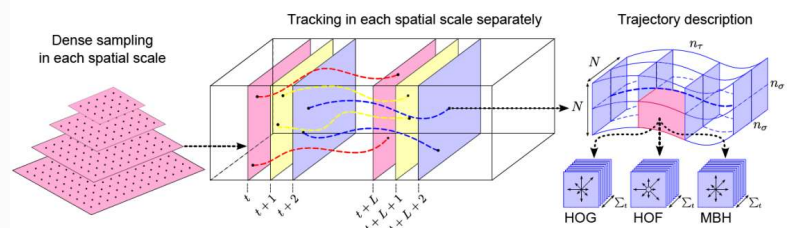
2012

2013

2014

2015

2016



- Suffer from camera motion and illumination change in video
- Not contain high-level semantic information
- High dimensionality
- Too expensive for real-time computation

67

## Video representation learning

2011

### 2D Convolutional Neural Network

Large-scale Video Classification with Convolutional Neural Networks. [Karpthy, CVPR'14]

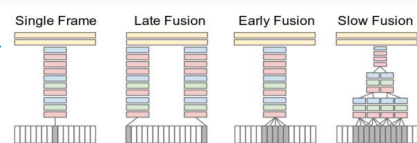
2012

2013

2014

2015

2016



- Treat video as a bag of short, fixed-sized clips
- Extend the connectivity of network in time dim.

Two-Stream Convolutional Networks for Action Recognition in Videos. [Simonyan, NIPS'14]

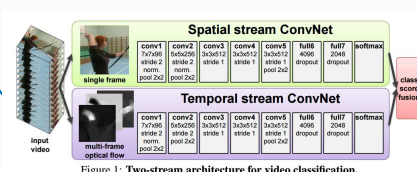
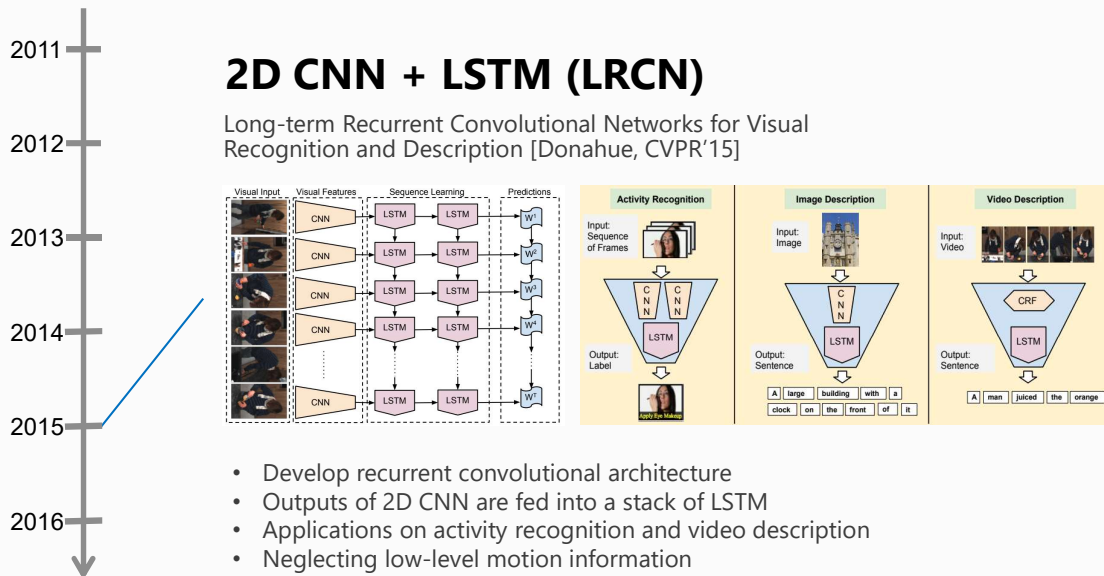


Figure 1: Two-stream architecture for video classification.

- Two-stream: frame + motion (stacked optical flow)
- 2D CNN for frame is pre-trained on ImageNet
- 2D CNN for motion is trained from scratch

68

## Video representation learning



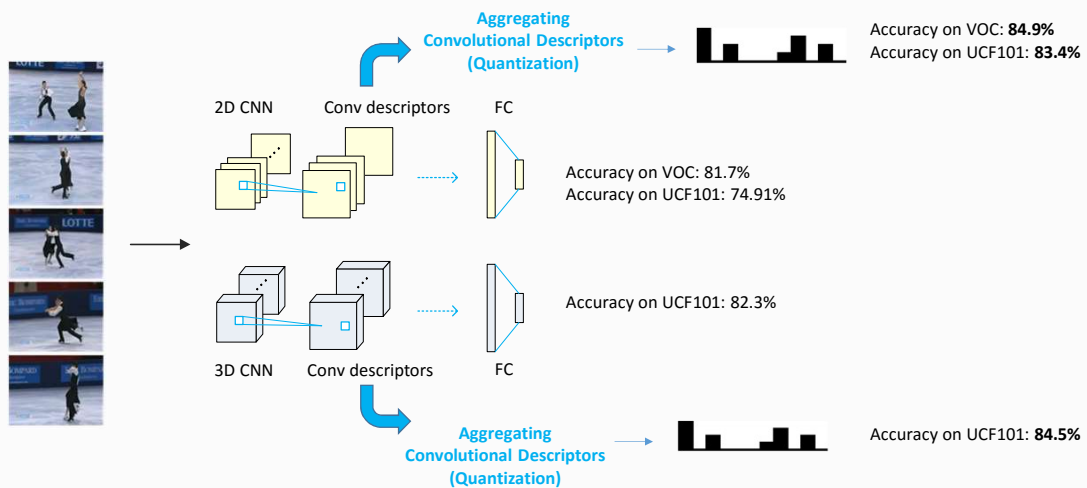
69

## Learning video representations

- Question 1: how to learn good holistic representations from convolution layers?
  - Deep Quantization (DQ) [CVPR'17]
- Question 2: how to learn good low-level representations (appearance + motion) from a video clip?
  - Pseudo 3D Residual Network (P3D) [ICCV'17]

70

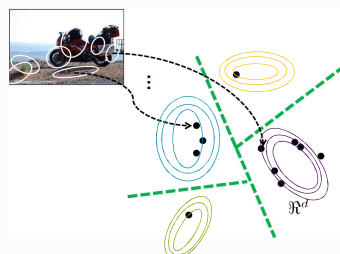
## Deep Representation Quantization



71

## Quantization Mechanisms

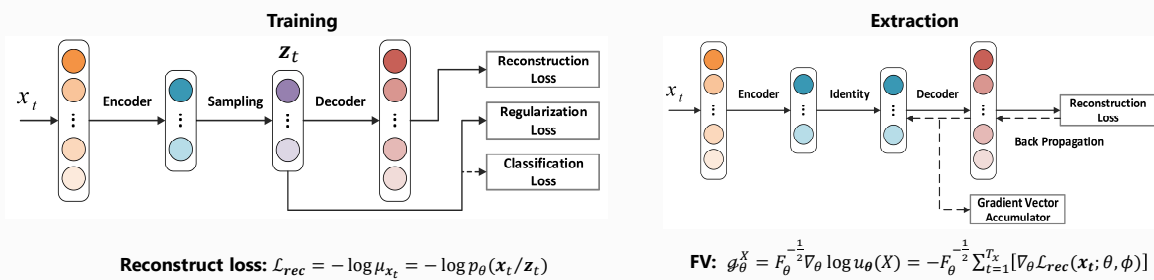
- Bag-of-visual-words (BoW)
- Vector of Locally Aggregated Descriptors (VLAD)
- Fisher Vector (FV)



72

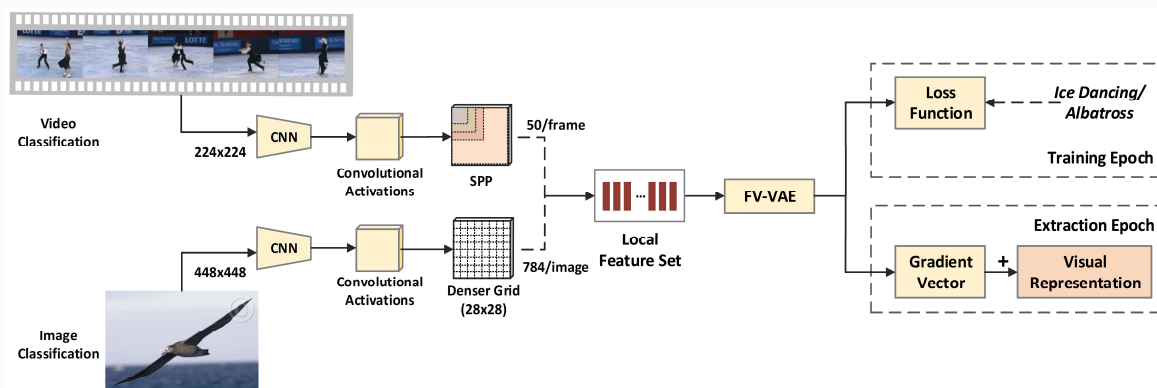
## Fisher Vector Meets Variational Auto-Encoder (FV-VAE)

- Assumption of FV
  - Data is generated from Gaussian Mixture Model, which may not hold in practice
- VAE
  - *Encoder* ( $q_\phi(\mathbf{z}/\mathbf{x})$ ): learn new representations  $\mathbf{z}$  for the given input  $\mathbf{x}$
  - *Decoder* ( $p_\theta(\mathbf{x}/\mathbf{z})$ ): generate FV of new representations  $\mathbf{z}$



73

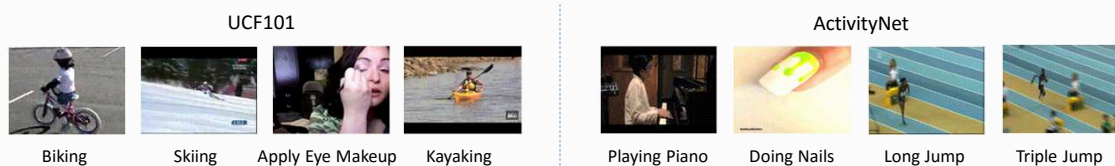
## Visual Representation Learning Framework



74

## Evaluations

- UCF101
  - 101 action categories
  - 13,320 videos (~9,500 training, ~3,700 testing)
  - ~6 sec on average for each video
- ActivityNet
  - 200 action categories, multi-label
  - 19,994 videos (10,024 training, 4,926 validation, 5,044 testing)
  - 1.9 min on average for each video, 648 hours in total



75

## Performance on UCF101

Comparisons of different quantization methods on UCF101 split1

| Feature   | Dimension | Accuracy      |
|---|-----------|---------------|
| Global Activations  | 4,096     | 74.91%        |
| Concatenation   | 25,088    | 75.89%        |
| AVE (ave of conv)   | 512       | 73.25%        |
| FV [F. Perronnin <i>et al.</i> , XRCE, ECCV'10]             | 131,072   | 78.85%        |
| VLAD [H. Jegou <i>et al.</i> , INRIA, CVPR'10]              | 131,072   | 80.67%        |
| Bilinear Pooling [T.-Y. Lin <i>et al.</i> , UMass, ICCV'15] | 262,144   | 81.39%        |
| FV-VAE (ours)   | 131,072   | <b>83.45%</b> |

Comparisons with state-of-the-art methods on UCF101

| Method  | Acc          |
|---|--------------|
| Two-stream ConvNet [A. Zisserman <i>et al.</i> , U of Oxford, NIPS'14]    | 88.1%        |
| C3D (3 nets) [R. Fergus <i>et al.</i> , FAIR & NYU, ICCV'15]              | 85.2%        |
| Factorized ST-ConvNet [L. Sun <i>et al.</i> , HKUST, ICCV'15]             | 88.1%        |
| Two-stream + LSTM [O. Vinyals <i>et al.</i> , Google, CVPR'15]            | 88.6%        |
| Two-stream fusion [A. Zisserman <i>et al.</i> , U of Oxford, CVPR'16]     | 92.5%        |
| Long-term temporal ConvNet [G. Varol <i>et al.</i> , INRIA, TPAMI'17]     | 91.7%        |
| Key-volume mining CNN [W. Zhu <i>et al.</i> , Tsinghua, CVPR'16]          | 93.1%        |
| TSN (3 modalities) [L. Wang <i>et al.</i> , CUHK, ECCV'16]                | 94.2%        |
| IDT [C. Schmid <i>et al.</i> , INRIA, ICCV'13]                            | 85.9%        |
| C3D + IDT [R. Fergus <i>et al.</i> , FAIR & NYU, ICCV'15]                 | 90.4%        |
| TDD + IDT [L. Wang <i>et al.</i> , CUHK, CVPR'15]                         | 91.5%        |
| Long-term temporal ConvNet+IDT [G. Varol <i>et al.</i> , INRIA, TPAMI'17] | 92.7%        |
| FV-VAE-pool5  | 83.9%        |
| FV-VAE-pool5 optical flow   | 89.5%        |
| FV-VAE-res5c  | 86.6%        |
| FV-VAE-(pool5 + pool5 optical flow)                                       | 93.7%        |
| FV-VAE-(res5c + pool5 optical flow)                                       | <b>94.2%</b> |
| FV-VAE-(res5c + pool5 optical flow) + IDT                                 | <b>95.2%</b> |



# ActivityNet Challenge

- FV-VAE: Rank 1 in terms of single representations among all teams
- Our ensemble model: Rank 3 in the ActivityNet Challenge 2016

Comparisons on ActivityNet

| Methods   | Top-1         | Top-3         | MAP           |
|---|---------------|---------------|---------------|
| VGG_19-GA [A. Zisserman <i>et al.</i> , U of Oxford, ICLR'15] | 66.59%        | 82.70%        | 70.22%        |
| ResNet_152-GA [MSRA, CVPR'16]                                 | 71.43%        | 86.45%        | 76.56%        |
| C3D-GA [R. Fergus <i>et al.</i> , FAIR & NYU, ICCV'15]        | 65.80%        | 81.16%        | 67.68%        |
| IDT [C. Schmid <i>et al.</i> , INRIA, ICCV'13]                | 64.70%        | 77.98%        | 68.69%        |
| FV-VAE-pool5  | 72.51%        | 85.68%        | 77.25%        |
| FV-VAE-res5c  | <b>78.55%</b> | <b>91.16%</b> | <b>84.09%</b> |

**ACTIVITYNET**  
Large Scale Activity Recognition Challenge

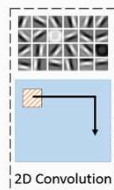
Home People Important Dates Program Guidelines Evaluation Contact Us **CVPR2016**

**Leaderboard - Untrimmed Video Classification**

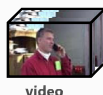
| Ranking | IT | Username               | IT | Organization                                  | IT | Upload time         | IT | mAP     | IT | Top-1   | IT | Top-3   | IT |
|---------|----|------------------------|----|---|----|---------------------|----|---------|----|---------|----|---------|----|
| 1       |    | Limin Wang             |    | CUHK & ETHZ & SIAT                            |    | 2016-06-08 14:10:36 |    | 0.93233 |    | 0.88136 |    | 0.96421 |    |
| 2       |    | Ruxin Wang             |    | QCIS  |    | 2016-06-09 06:47:55 |    | 0.92413 |    | 0.87792 |    | 0.97084 |    |
| 3       |    | Ting Yao               |    | Multimedia Search and Mining Group, MSRA      |    | 2016-06-09 07:26:06 |    | 0.91937 |    | 0.86685 |    | 0.95535 |    |
| 4       |    | Linchao Zhu            |    | UTS   |    | 2016-05-08 12:01:45 |    | 0.87163 |    | 0.849   |    | 0.9504  |    |
| 5       |    | Masatoshi Hidaka       |    | The University of Tokyo                       |    | 2016-06-09 05:28:49 |    | 0.86458 |    | 0.80434 |    | 0.9262  |    |
| 6       |    | Ke Ning                |    | Zhejiang University                           |    | 2016-06-09 05:22:02 |    | 0.84104 |    | 0.8339  |    | 0.93525 |    |
| 7       |    | Cong Guo               |    | University of Science and Technology of China |    | 2016-05-18 17:18:53 |    | 0.84067 |    | 0.79654 |    | 0.91355 |    |
| 8       |    | Yi Zhu                 |    | UC Merced                                     |    | 2016-06-06 22:05:16 |    | 0.831   |    | 0.78444 |    | 0.91072 |    |
| 9       |    | Cesar Roberto de Souza |    | Xerox Research Center Europe                  |    | 2016-06-08 15:58:55 |    | 0.82607 |    | 0.78524 |    | 0.89584 |    |
| 10      |    | Gurkirt Singh Singh    |    | Oxford Brookes                                |    | 2016-06-09 07:28:29 |    | 0.82546 |    | 0.7677  |    | 0.89401 |    |
| 11      |    | Yingwei Pan            |    | USTC  |    | 2016-05-03 01:05:36 |    | 0.8165  |    | 0.75461 |    | 0.88954 |    |
| 12      |    | Xianming Liu           |    | University of Illinois, Urbana-Champaign      |    | 2016-06-08 18:59:46 |    | 0.75765 |    | 0.7406  |    | 0.88035 |    |
| 13      |    | Qun Zhong              |    | SJTU  |    | 2016-05-31 15:47:33 |    | 0.75391 |    | 0.73513 |    | 0.86436 |    |
| 14      |    | Kirill Gavriluk        |    | University of Amsterdam                       |    | 2016-05-01 13:05:08 |    | 0.7414  |    | 0.73921 |    | 0.84561 |    |
| 15      |    | Po-Yao Huang           |    | CMU   |    | 2016-06-08 09:42:04 |    | 0.7358  |    | 0.78082 |    | 0.78082 |    |
| 16      |    | Bharat Singh           |    | University of Maryland, College Park          |    | 2016-06-09 04:09:09 |    | 0.71388 |    | 0.69133 |    | 0.85341 |    |
| 17      |    | Alberto Montes         |    | UPC   |    | 2016-06-08 15:06:29 |    | 0.58741 |    | 0.58757 |    | 0.75548 |    |
| 18      |    | Tackgeun You           |    | POSTECH                                       |    | 2016-06-08 16:01:52 |    | 0.53767 |    | 0.58898 |    | 0.66117 |    |
| 19      |    | mujtaba hasan          |    | IIT-Delhi                                     |    | 2016-06-08 12:14:05 |    | 0.5089  |    | 0.51156 |    | 0.70581 |    |
| 20      |    | Christian Rupprecht    |    | Technische Universität München                |    | 2016-06-07 22:04:24 |    | 0.37167 |    | 0.4635  |    | 0.66298 |    |

## Video representation learning: from 2D CNN to 3D CNN

**ResNet:**  
[MSRA, CVPR'16]



**3D CNN:**  
[FAIR & NYU, ICCV'15]



### Network comparison on Sports-1M

| Network | Depth | Model Size | Video hit@1 |
|---------|-------|------------|-------------|
| ResNet  | 152   | 235 MB     | 64.6%       |
| C3D     | 11    | 321 MB     | 61.1%       |
| C3D     | 100+  | ~3 GB      | --          |

- Training 3D CNN is very computationally **expensive**
- Difficult to train very **deep** 3D CNN
- Fine-tuning** 2D CNN is better than 3D CNN

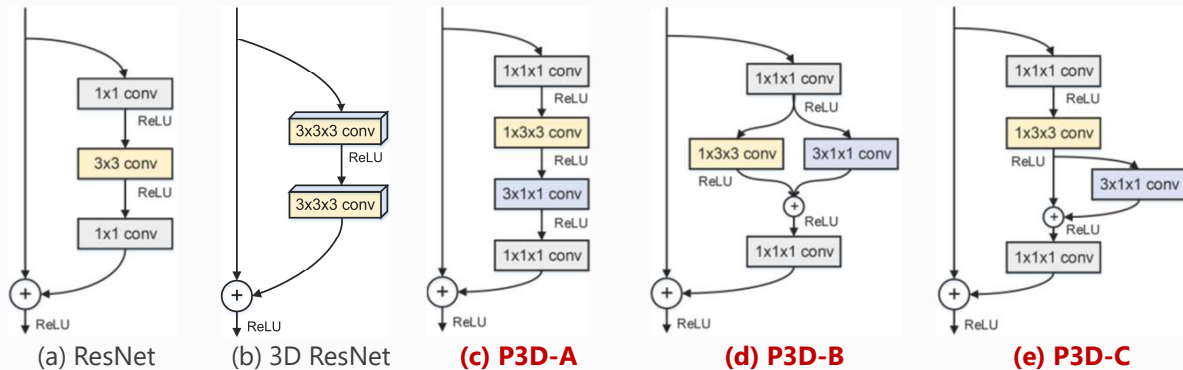
## [Yao &amp; Mei, ICCV'17]



## Pseudo-3D Residual Networks (P3D) [Qiu, Yao, Mei, ICCV'17]



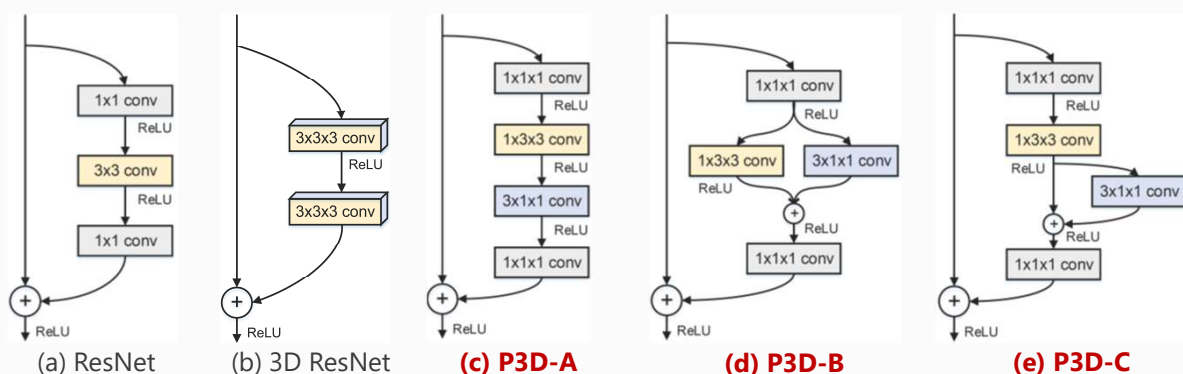
## P3D: architectures



- Mix different P3D blocks to replace Residual Units in a **152-layer ResNet**
- Train on **Sports-1M dataset** (1.13M videos annotated with 487 labels)
- Learn a generic spatiotemporal video representation with **199** layers
- <https://github.com/ZhaofanQiu/pseudo-3d-residual-networks> [ICCV'17]

81

## P3D <https://github.com/ZhaofanQiu/pseudo-3d-residual-networks> [Qiu, Yao, Mei, ICCV'17]

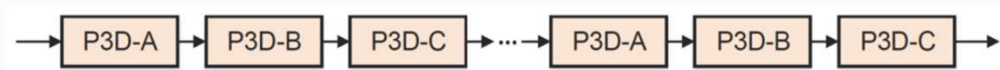


- (a)  $(\mathbf{I} + \mathbf{S}) \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) = \mathbf{x}_{t+1}$   
 (b)  $(\mathbf{I} + \mathbf{TS}) \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{TS}(\mathbf{x}_t) = \mathbf{x}_{t+1}$   
 (c)  $(\mathbf{I} + \mathbf{T} \cdot \mathbf{S}) \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)) = \mathbf{x}_{t+1}$   
 (d)  $(\mathbf{I} + \mathbf{S} + \mathbf{T}) \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) + \mathbf{T}(\mathbf{x}_t) = \mathbf{x}_{t+1}$   
 (e)  $(\mathbf{I} + \mathbf{S} + \mathbf{T} \cdot \mathbf{S}) \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)) = \mathbf{x}_{t+1}$

82

## Pseudo-3D ResNet

- P3D-A, -B, and -C ResNet, by replacing all the Residual Units in a ResNet-50 with one certain kind of P3D block
- P3D: Mix different P3D blocks



Comparisons between P3D ResNet variants

| Method       | Model Size | Speed (GPU time) | Accuracy |
|--------------|------------|------------------|----------|
| ResNet-50    | 92MB       | 15.0 frame/s     | 80.8%    |
| P3D-A ResNet | 98MB       | 9.0 clip/s       | 83.7%    |
| P3D-B ResNet | 98MB       | 8.8 clip/s       | 82.8%    |
| P3D-C ResNet | 98MB       | 8.6 clip/s       | 83.0%    |
| P3D ResNet   | 98MB       | 8.8 clip/s       | 84.2%    |

- Dataset: UCF101
- GPU: One Nvidia K40

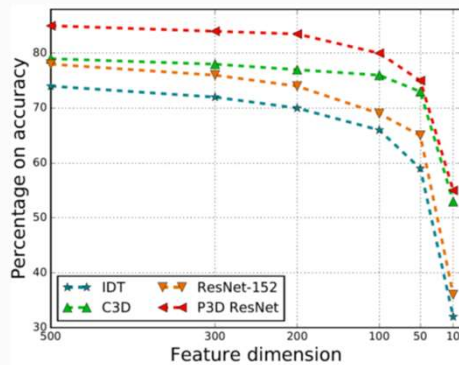
## Spatio-Temporal Representation Learning

- Perform on a **deeper 152-layer ResNet**, and then produce a generic spatio-temporal video representation
- Learn on **Sports-1M datasets**, roughly containing about **1.13 million videos** annotated with 487 sports labels

Comparisons in terms of pre-train data and accuracy on Sports-1M

| Method  | Pre-train Data | Clip Length | Clip hit@1   | Video hit@1  | Video hit@5  |
|---|----------------|-------------|--------------|--------------|--------------|
| Deep Video (Slow Fusion) [L. Fei-fei et al., CVPR'14] | ImageNet1K     | 1           | 41.1%        | 59.3%        | 77.7%        |
| Deep Video (Slow Fusion) [L. Fei-fei et al., CVPR'14] | ImageNet1K     | 10          | 41.9%        | 60.9%        | 80.2%        |
| C3D [R. Fergus et al., FAIR & NYU, ICCV'15]           | -              | 16          | 44.9%        | 60.0%        | 84.4%        |
| C3D [R. Fergus et al., FAIR & NYU, ICCV'15]           | I380K          | 16          | 46.1%        | 61.1%        | 85.2%        |
| ResNet-152 [MSRA, CVPR'16]                            | ImageNet1K     | 1           | 46.5%        | 64.6%        | 86.4%        |
| P3D ResNet  | ImageNet1K     | 16          | <b>47.9%</b> | <b>66.4%</b> | <b>87.4%</b> |

# P3D <https://github.com/ZhaofanQiu/pseudo-3d-residual-networks> [Qiu, Yao, Mei, ICCV'17]



| Networks               | CPU runtime (ms) | GPU runtime (ms) |
|------------------------|------------------|------------------|
| ResNet-152 (16 frames) | 5,600            | 400              |
| P3D-199 (16 frames)    | 1,500            | 150              |

P3D ResNet consistently outperforms others at each dimension (16 frames/clip).

P3D ResNet performs 3 times faster than ResNet on a single clip (16 frames).

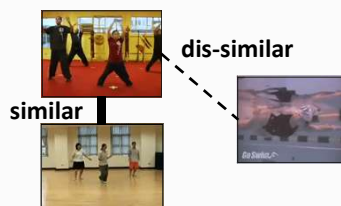
85

- ActivityNet Untrimmed Task
- Action Recognition



Walking the dog

- ASLAN
- Action Similarity Labeling



- YUPENN, Dynamic Scene
- Scene Recognition



beach

| Method                     | Top-1         | Top-3         | MAP           |
|----------------------------|---------------|---------------|---------------|
| IDT [INRIA, ICCV'13]       | 64.70%        | 77.98%        | 68.69%        |
| C3D [FAIR, ICCV'15]        | 65.80%        | 81.16%        | 67.68%        |
| VGG [U of Oxford, ICLR'15] | 66.59%        | 82.70%        | 70.22%        |
| ResNet [MSRA, CVPR'16]     | 71.43%        | 86.45%        | 76.56%        |
| <b>P3D ResNet</b>          | <b>75.12%</b> | <b>87.71%</b> | <b>78.86%</b> |

| Method                    | Accuracy     | AUC          |
|---------------------------|--------------|--------------|
| MIP [Tel Aviv U, ECCV'12] | 65.5%        | 71.9%        |
| IDT+FV [INRIA, ICCV'13]   | 68.7%        | 75.4%        |
| C3D [FAIR, ICCV'15]       | 78.3%        | 86.5%        |
| ResNet [MSRA, CVPR'16]    | 70.4%        | 77.4%        |
| <b>P3D ResNet</b>         | <b>80.8%</b> | <b>87.9%</b> |

| Method                 | Dynamic Scene | YUPENN       |
|------------------------|---------------|--------------|
| [U Penn, CVPR'12]      | 43.1%         | 80.7%        |
| [York U, CAN, CVPR'14] | 77.7%         | 96.2%        |
| C3D [FAIR, ICCV'15]    | 87.7%         | 98.1%        |
| ResNet [MSRA, CVPR'16] | 93.6%         | 99.2%        |
| <b>P3D ResNet</b>      | <b>94.6%</b>  | <b>99.5%</b> |



## Action recognition w/ P3D

- **Dataset:** UCF101  
(101 categories w/ 13,320 videos)
- **Task:** Action Recognition
- **Methods:**
  - ① end-to-end CNN
  - ② CNN-based representation + SVM
  - ③ fused w/ IDT

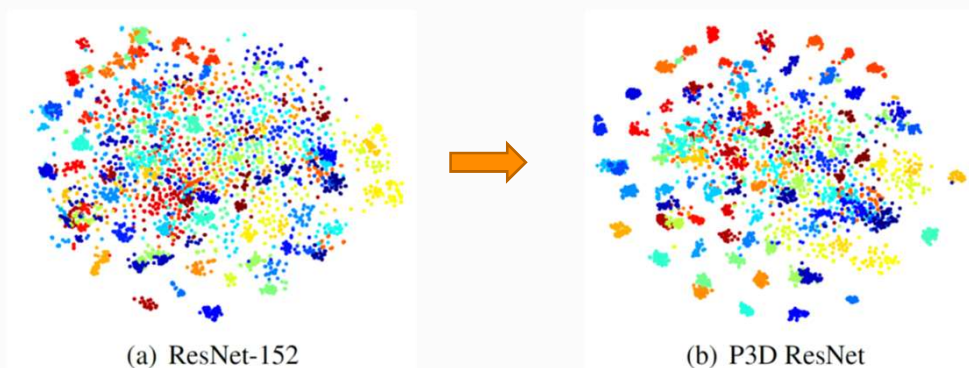


action: taichi

| Method   | Accuracy             |
|--|----------------------|
| <b>End-to-end two-stream CNNs w/ fine-tuning (+optical flow)</b> |                      |
| Two-stream ConvNet [U of Oxford, NIPS'14]                        | 73.0% (88.0%)        |
| Two-stream + LSTM [Google, CVPR'15]                              | 82.6% (88.6%)        |
| Two-stream fusion [U of Oxford, CVPR'16]                         | 82.6% (92.5%)        |
| TSN [CUHK, ECCV'16]  | 85.7% (94.0%)        |
| P3D ResNet (ours)  | <b>89.7% (94.8%)</b> |
| P3D ResNet + Deep Quantization (ours)                            | <b>91.3% (95.5%)</b> |
| <b>CNN-based representation + linear SVM</b>                     |                      |
| C3D [FAIR & NYU, ICCV'15]  | 82.3%                |
| ResNet-152 [MSRA, CVPR'16]                                       | 83.5%                |
| P3D ResNet (ours)  | <b>88.6%</b>         |
| <b>Method fusion w/ IDT</b>                                      |                      |
| IDT [INRIA, ICCV'13]   | 85.9%                |
| C3D + IDT  | 90.4%                |
| P3D ResNet (ours) + IDT  | <b>93.7%</b>         |

## Representation embedding visualization

- Video rep. by P3D ResNet are better semantically separated than those of ResNet-152



# Examples



P3D: Waterskiing;  
Fun sliding down;  
Wakeboarding

ResNet: Canoeing;  
Fun sliding down;  
Rafting

C3D: Wakeboarding;  
River tubing;  
Waterskiing



P3D: Preparing pasta;

Preparing salad;  
Playing blackjack

ResNet: Preparing pasta;  
Making an omelette;  
Preparing salad

C3D: Preparing pasta;  
Playing blackjack;  
Beer pong



P3D: Mopping floor;  
Disc dog;  
Walking the dog

ResNet: Cumbia;  
Washing face;  
Grooming dog

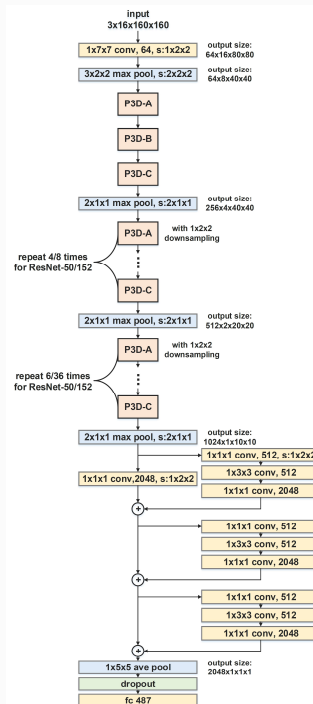
C3D: Cumbia;  
Brushing hair;  
Carving jack-o-lanterns



P3D: Removing curlers;  
Getting a haircut;  
Braiding hair

ResNet: Removing curlers;  
Gargling mouthwash;  
Knitting

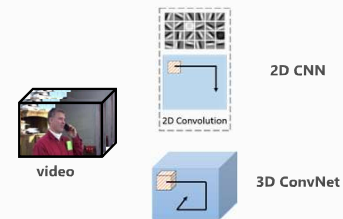
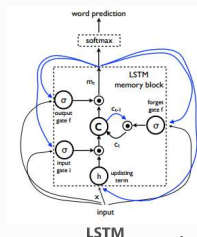
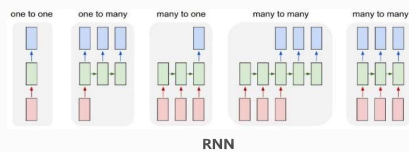
C3D: Getting a haircut;  
Blowing leaves;  
Removing curlers



90

# Challenges for video captioning

- Video captioning is much more complicated
- Learning video representation
  - frame: visual objects (AlexNet, GoogLeNet, VGG)
  - segment: temporal dynamics (3D CNN, optical flow)
  - video: pooling/alignment on frame and/or segment



- Sentence generation
  - multi-layer RNN (LSTM)
  - semantic relationship between entire sentence and video content

91

## What if simply applying image captioning to video?

### Video-to-sentence:



LSTM-E: a man is riding a motorcycle

### Image-to-sentence (keyframe-based): <http://deeplearning.cs.toronto.edu/i2t>



there is a black motorcycle sitting in front of a small amount of cars



someone is holding a hole in the background



a close up of a pair of scissors with his hand



a man wearing a helmet is racing

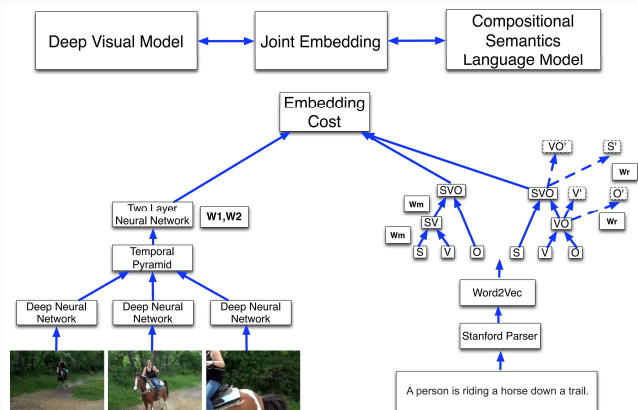


a flock of birds flying over the rock of water on a cliff

92

# Video captioning

- Search (embedding)-based approach [Xu, AAAI15; Yu, ACL13 & AAAI15]



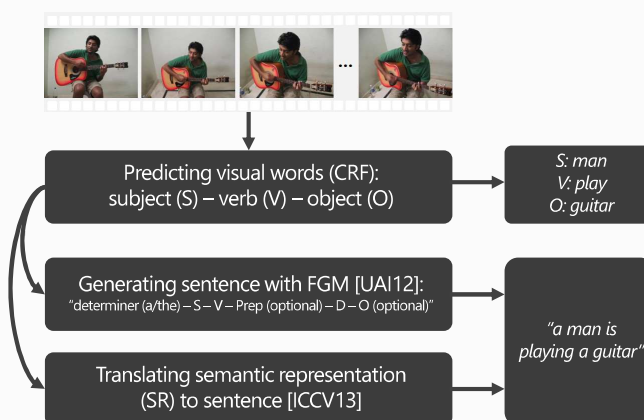
- Deep visual model to learn video representation
- Compositional language model to capture semantic compatibility among concepts
- Joint embedding model to minimize distance of the above two models in video-text space [Xu, AAAI15]

$$J(V, T) = \sum_{i=1}^N (E_{embed}(V, T) + \sum_{p \in \mathbf{NT}} E_{rec}(p|W_m, W_r)) + r$$

93

# Video captioning

- Language model-based approach [Thomason, COLING14; Barbu, UAI12; Rohrbach, ICCV13; Krishnamoorthy, AAAI13]



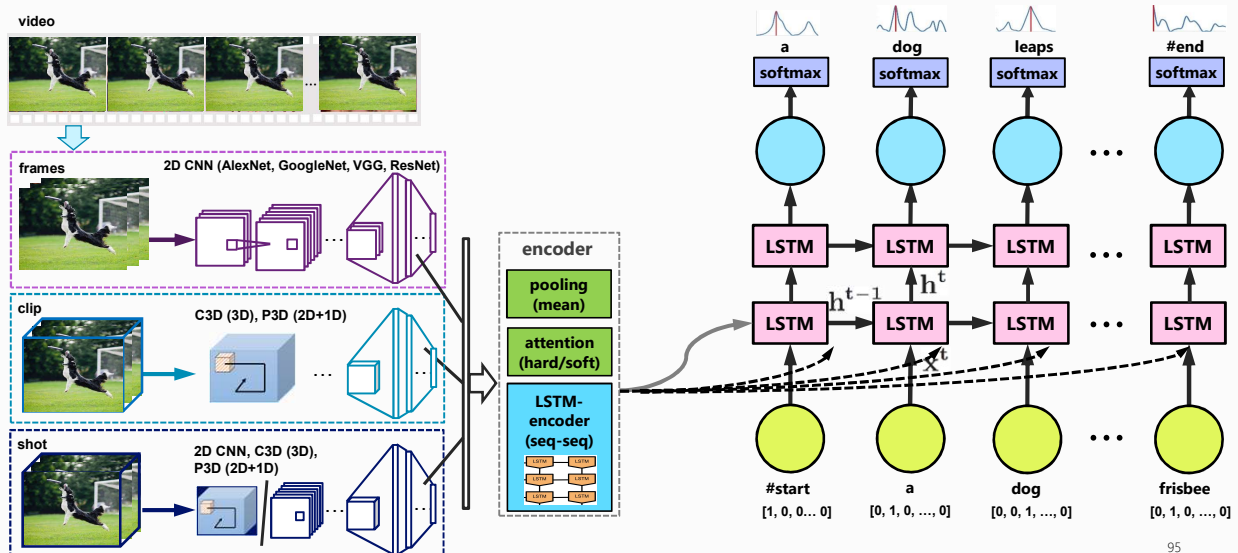
Barbu, et al. "Video In Sentences Out", UAI 2012.  
<https://www.youtube.com/watch?v=tu3jMxCJPMw>

94



# Sequence learning-based approach

[Donahue, CVPR15; Yao, ICCV15; Venugopalan, ICCV15; Yu, CVPR'16; Pan, CVPR16&17; Baraldi, CVPR17]



95

## Video Captioning with X



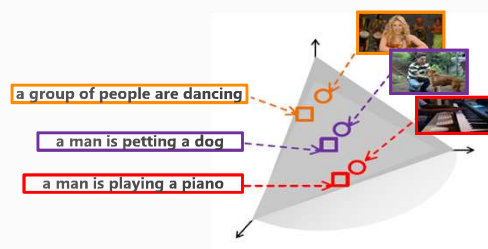
A **man** is **shooting** a **gun**  
**X = temporal attention**  
 [Yao, CVPR'15]



**X = spatiotemporal attention**  
 [Yu, CVPR'16]



**X = visual attributes**  
 [Pan, CVPR'16'17; Yu, CVPR'17]



**X = semantic embedding**  
 [Pan, CVPR'16]

96

# Video Captioning with Semantics

- Key issues in sentence generation
  - relevance**: relationship between sentence (S, V, O) semantics and video content
  - coherence**: sentence grammar



LSTM: a man is playing a **guitar**  
 LSTM-E: a man is playing a **piano**

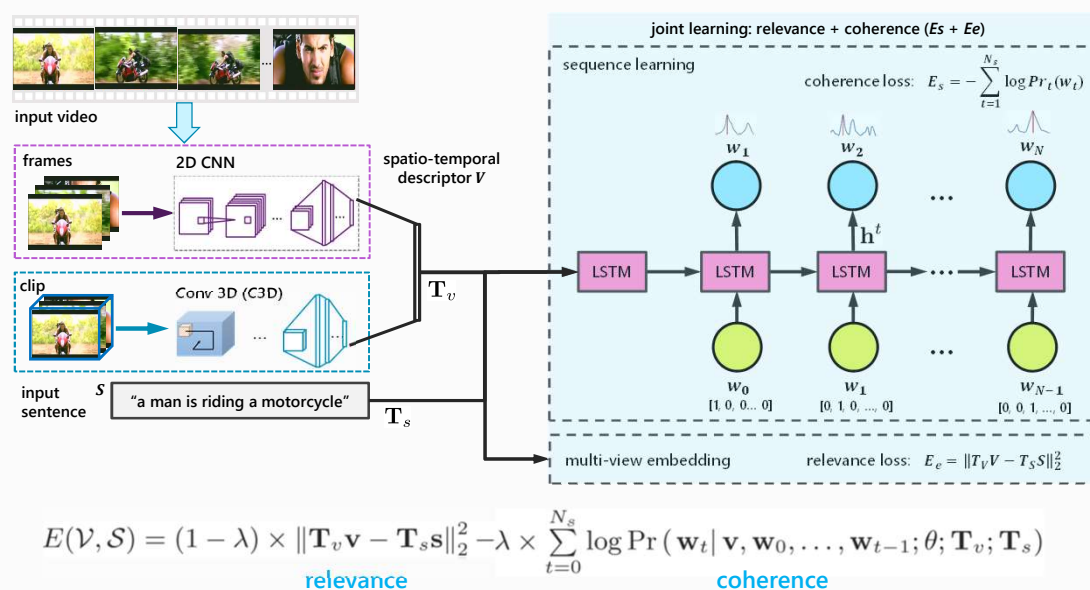


LSTM: **a man** is dancing  
 LSTM-E: **a group of people** are dancing

- Joint learning (LSTM-E): relevance + coherence [Pan, CVPR'16]
  - Explicitly and holistically emphasize video content with "relevance" regularizer

97

## LSTM-E for video captioning [Pan & Mei, CVPR'16]



98

# Evaluations

- Dataset ([MSR Video Description Corpus](#), a.k.a. YouTube2Text)
  - 1,970 Youtube video snippets (1,200 training, 100 validation, 670 testing)
  - 10-25 sec for each clip
  - ~40 human-generated sentences for each clip (by AMT)
  - dictionary: 15,903 -> 7,000; 45 S-groups, 218 V-groups, 241 O-groups
- Training: 12 hrs in one single CPU; testing: ~5 sec per clip



1. a man is petting a dog
2. a man is petting a tied up dog
3. a man pets a dog
4. a man is showing his dog to the camera
5. a boy is trying to see something to a dog



1. a man is playing the guitar
2. a men is playing instrument
3. a man plays a guitar
4. a man is singing and playing guitar
5. the boy played his guitar



1. a kitten is playing with his toy
2. a cat is playing on the floor
3. a kitten plays with a toy
4. a cat is playing
5. a cat tries to get a ball



1. a man is singing on stage
2. a man is singing into a microphone
3. a man sings into a microphone
4. a singer sings
5. the man sang on stage into the microphone

99

| Dataset                    | Organizer    | Context              | Annotation         | #Video        | #Clip         | #Sentence      | #Word            | Vocabulary    | Duration (hr) |
|----------------------------|--------------|----------------------|--------------------|---------------|---------------|----------------|------------------|---------------|---------------|
| YouCook                    | SUNY Buffalo | Cooking              | Labeled            | 88            | -             | 2,668          | 42,457           | 2,711         | 2.3           |
| TACos                      | MP Institute | cooking              | Labeled            | 123           | 7,206         | 18,227         | -                | -             | -             |
| TACos M-L                  | MP Institute | cooking              | Labeled            | 185           | 14,105        | 52,593         | -                | -             | -             |
| M-VAD                      | UdeM         | movie                | DVS                | 92            | 48,986        | 55,905         | 519,933          | 18,269        | 84.6          |
| MPII                       | MP Institute | movie                | DVS+Script         | 94            | 68,337        | 68,375         | 653,467          | 24,549        | 73.6          |
| MSVD                       | MSR          | multi-category       | AMT workers        | -             | 1,970         | 70,028         | 607,339          | 13,010        | 5.3           |
| T-GIF                      | Yahoo & UR   | Tumblr GIF           | AMT workers        | -             | 102,068       | 125,782        | 1,418,555        | 12,228        | 103           |
| <b>ActivityNet Caption</b> | KAUST        | multi-category       | AMT workers        | 20,000        | 20,000        | 100,000        | 1,348,000        | -             | 849           |
| <b>MSR-VTT (10K)</b>       | <b>MSRA</b>  | <b>20 categories</b> | <b>AMT workers</b> | <b>7,180</b>  | <b>10,000</b> | <b>200,000</b> | <b>1,856,523</b> | <b>29,136</b> | <b>41.2</b>   |
| <b>MSR-VTT (20K)</b>       | <b>MSRA</b>  | <b>20 categories</b> | <b>AMT workers</b> | <b>14,768</b> | <b>20,000</b> | <b>400,000</b> | <b>4,284,032</b> | <b>49,436</b> | <b>87.8</b>   |



1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.



1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.



1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

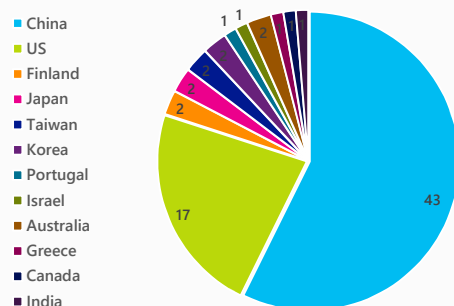


1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

100

## Microsoft Video to Language Challenge 2016

77 teams registered challenge  
22 teams submitted results  
Awards will be announced at ACMMM

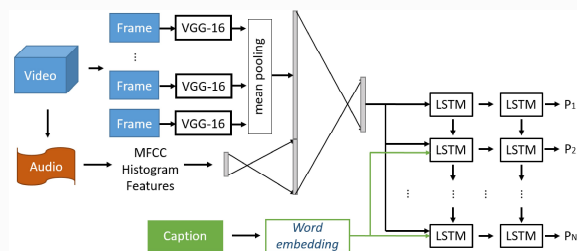


| M1   |               | M2                              |        |        |         |         |
|------|---------------|---------------------------------|--------|--------|---------|---------|
| Rank | Team          | Organization                    | BLEU@4 | Meteor | CIDEr-D | ROUGE-L |
| 1    | v2t_navigator | RUC & CMU                       | 0.408  | 0.282  | 0.448   | 0.609   |
| 2    | Aalto         | Aalto University                | 0.398  | 0.269  | 0.457   | 0.598   |
| 3    | VideoLAB      | UML & Berkeley & UT-Austin      | 0.391  | 0.277  | 0.441   | 0.606   |
| 4    | ruc-uva       | RUC & UVA & Zhejiang University | 0.387  | 0.269  | 0.459   | 0.587   |
| 5    | Fudan-ILC     | Fudan & ILC                     | 0.387  | 0.268  | 0.419   | 0.595   |
| 6    | NUS-TJU       | NUS & TJU                       | 0.371  | 0.267  | 0.410   | 0.590   |
| 7    | Umich-COG     | University of Michigan          | 0.371  | 0.266  | 0.411   | 0.583   |
| 8    | MCG-ICT-CAS   | ICT-CAS                         | 0.367  | 0.264  | 0.404   | 0.590   |
| 9    | DeepBrain     | NLPR_CASIA & IQIYI              | 0.382  | 0.259  | 0.401   | 0.582   |
| 10   | NTU MIRA      | NTU                             | 0.355  | 0.261  | 0.383   | 0.579   |

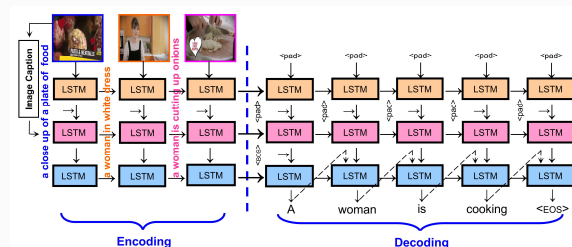
| M1   | M2            |                                 |       |       |       |
|------|---------------|---------------------------------|-------|-------|-------|
| Rank | Team          | Organization                    | C1    | C2    | C3    |
| 1    | Aalto         | Aalto University                | 3.263 | 3.104 | 3.244 |
| 2    | v2t_navigator | RUC & CMU                       | 3.261 | 3.091 | 3.154 |
| 3    | VideoLAB      | UML & Berkeley & UT-Austin      | 3.237 | 3.109 | 3.143 |
| 4    | Fudan-ILC     | Fudan & ILC                     | 3.185 | 2.999 | 2.979 |
| 5    | ruc-uva       | RUC & UVA & Zhejiang University | 3.225 | 2.997 | 2.933 |
| 6    | Umich-COG     | University of Michigan          | 3.247 | 2.865 | 2.929 |
| 7    | NUS-TJU       | NUS & TJU                       | 3.308 | 2.833 | 2.893 |
| 8    | DeepBrain     | NLPR_CASIA & IQIYI              | 3.259 | 2.878 | 2.892 |
| 9    | NLPRMMC       | CASIA & Anhui University        | 3.266 | 2.868 | 2.893 |
| 10   | MCG-ICT-CAS   | ICT                             | 3.339 | 2.800 | 2.867 |

## MSR Video to Language Grand Challenge 2016

- CNN-LSTM [1, 2, 4, 5, 7]



- Sequence-to-Sequence (encoder-decoder) [3, 6, 9, 10]



- Image features
  - VGG-19 [1][2][5][6][9][10]
  - GoogleNet [2][4][5]
  - ResNet [3][5][8]
  - VGG-16 [5][7][8]
  - PlaceNet [5][9]
- Motion features
  - C3D [1][2][3][4][5][9][10]
  - IDT [1][2]
  - Optical flow [8]
- Acoustic features
  - MFCCs [1][3][7]
- Text features
  - ASR [1]
  - Video category [3][4]

## Summary from Video to Language Grand Challenge

Team [6] shows performance improve by ResNet, data augmentation and dense trajectory.

|                        | B@4  | MET. | ROU. | CID. |
|------------------------|------|------|------|------|
| VGG+C3D                | 32.3 | 25.8 | 56.7 | 29.6 |
| VGG+C3D+Aug.           | 33.3 | 26.6 | 57.2 | 32.5 |
| VGG+C3D+Res.           | 34.6 | 26.9 | 58.3 | 37.9 |
| VGG+C3D+Res.+Aug.      | 35.3 | 27.4 | 58.9 | 38.3 |
| VGG+C3D+Res.+Tra.      | 36.5 | 27.1 | 59.2 | 40.3 |
| VGG+C3D+Res.+Aug.+Tra. | 35.6 | 27.0 | 58.9 | 38.1 |

Team [3] shows performance gain by audio and category information.

| Descriptors | BLEU@4 | METEOR | CIDEr | ROUGE-L |
|-------------|--------|--------|-------|---------|
| categories  | 0.298  | 0.228  | 0.236 | 0.548   |
| audio       | 0.301  | 0.222  | 0.184 | 0.544   |
| C3D         | 0.374  | 0.264  | 0.389 | 0.594   |
| ResNet      | 0.389  | 0.269  | 0.400 | 0.605   |
| +C3D        | 0.385  | 0.267  | 0.411 | 0.601   |
| +categories | 0.381  | 0.270  | 0.418 | 0.597   |
| +audio      | 0.395  | 0.277  | 0.442 | 0.610   |
| committee   | 0.407  | 0.286  | 0.465 | 0.610   |

### • Other observations

- Additional training data from MS-COCO [2][7][8]
- Additional data from FCVID [4]
- Additional data from Youtube2Text [9]
- Captioning with tag based sentence reranking [4]
- Data augmentation (sampling from different frames and horizontally flipped frames) [5]
- Use PCA to reduce the dimensionality of low-level feature [8]

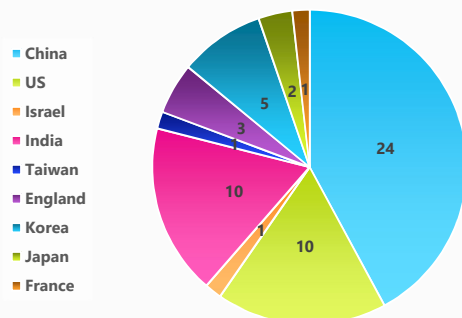
103

## Microsoft Video to Language Challenge 2017

57 teams registered challenge

8 teams submitted results

Awards will be announced at ACMMM'17



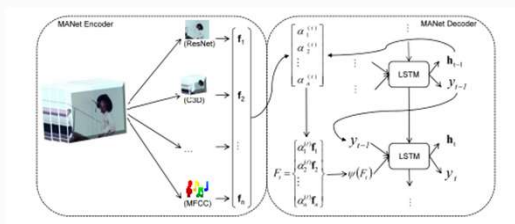
| M1   | M2          |                                   |        |        |         |         |
|------|-------------|-----------------------------------|--------|--------|---------|---------|
| Rank | Team        | Organization                      | BLEU@4 | Meteor | CIDEr-D | ROUGE-L |
| 1    | RUC+CMU_V2T | RUC & CMU                         | 0.390  | 0.255  | 0.315   | 0.542   |
| 2    | TJU_Media   | TJU                               | 0.359  | 0.226  | 0.249   | 0.515   |
| 3    | NII         | National Institute of Informatics | 0.359  | 0.234  | 0.231   | 0.514   |
| 4    | MIC_TJU     | Tongji University                 | 0.351  | 0.226  | 0.236   | 0.509   |
| 5    | Illusion    | IIT Delhi                         | 0.304  | 0.213  | 0.206   | 0.494   |
| 6    | LVIC_AS     | CEA LIST                          | 0.289  | 0.203  | 0.175   | 0.487   |
| 7    | TJU-NUS     | TJU & NUS                         | 0.265  | 0.191  | 0.151   | 0.456   |
| 8    | AFRL        | AFRL                              | 0.240  | 0.186  | 0.160   | 0.427   |

| M1   | M2          |                                   |       |       |       |
|------|-------------|-----------------------------------|-------|-------|-------|
| Rank | Team        | Organization                      | C1    | C2    | C3    |
| 1    | RUC+CMU_V2T | RUC & CMU                         | 4.437 | 3.437 | 3.567 |
| 2    | NII         | National Institute of Informatics | 4.078 | 3.359 | 3.570 |
| 3    | TJU_Media   | TJU                               | 4.032 | 2.962 | 3.048 |
| 4    | MIC_TJU     | Tongji University                 | 3.844 | 2.789 | 2.978 |
| 4    | Illusion    | IIT Delhi                         | 4.042 | 2.583 | 2.921 |
| 6    | TJU-NUS     | TJU & NUS                         | 3.762 | 2.364 | 2.376 |
| 7    | AFRL        | AFRL                              | 3.109 | 2.343 | 2.411 |
| 8    | LVIC_AS     | CEA LIST                          | 3.477 | 2.322 | 2.321 |

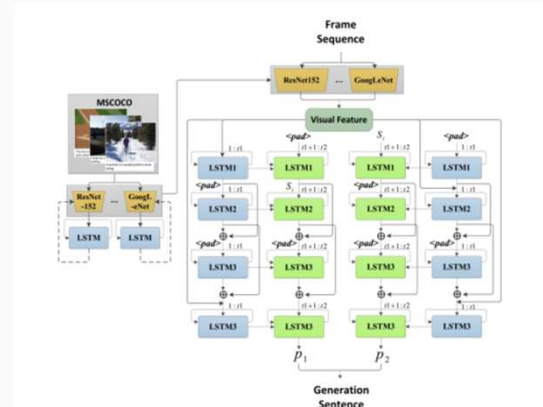


## MSR Video to Language Grand Challenge 2017

- Some other observations
  - Sentence Reranking [1]
  - Additional data from MSCOCO [7]
  - Additional semantic information [7]
  - Video category information [1]
  - Multi-modality fusion [1][2]



[3]



[4]

105

## Video Captioning vs. Dense Video Captioning

|                               |   |  |  |
|-------------------------------|---|--|--|
| <b>Input Video</b>            |   |  |  |
| <b>Video Captioning</b>       | A man is playing frisbee with a dog.  |  |  |
| <b>Dense Video Captioning</b> | <div> <div> <div></div> <div>A man and a dog are outdoors and waiting for their turn to play on a fenced in green field.</div> </div> <div> <div></div> <div>The man and the dog runs onto the field and he throws the frisbee a far distance and the dog runs and fetches it, then returns it back to the man and they repeat the process 6 times.</div> </div> <div> <div></div> <div>When they are done, another man runs to them and hands the man a leash and he leashes his dog.</div> </div> </div> <div> <div> <div></div> <div>The whole time there are people on the sidelines watching them and taking pictures.</div> </div> <div> <div></div> <div>A man and a dog walk onto a field.</div> </div> <div> <div></div> <div>A man throws a frisbee and the dog chases after it.</div> </div> <div> <div></div> <div>The dog brings the frisbee back to the man.</div> </div> </div> <div> <div>Start time</div> <div>End time</div> </div> |  |  |

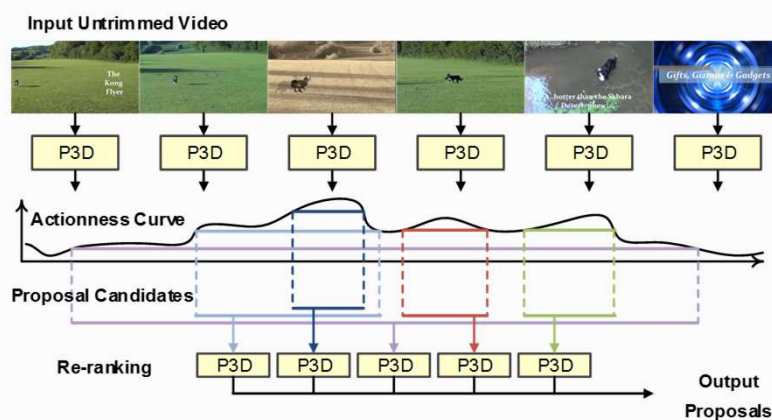
106

## Challenges for dense video captioning [ActivityNet'17]

- Powerful video representation
  - Pseudo-3D ResNet (P3D) [Z. Qiu, T. Yao and T. Mei, ICCV'17]
- Accurate event localization
  - Actionness detection + grouping + re-ranking
- Good video description
  - Captioning with attributes [T. Yao *et al.*, ICCV'17] + retrieval

107

## Event localization: actionness detection + grouping + re-ranking



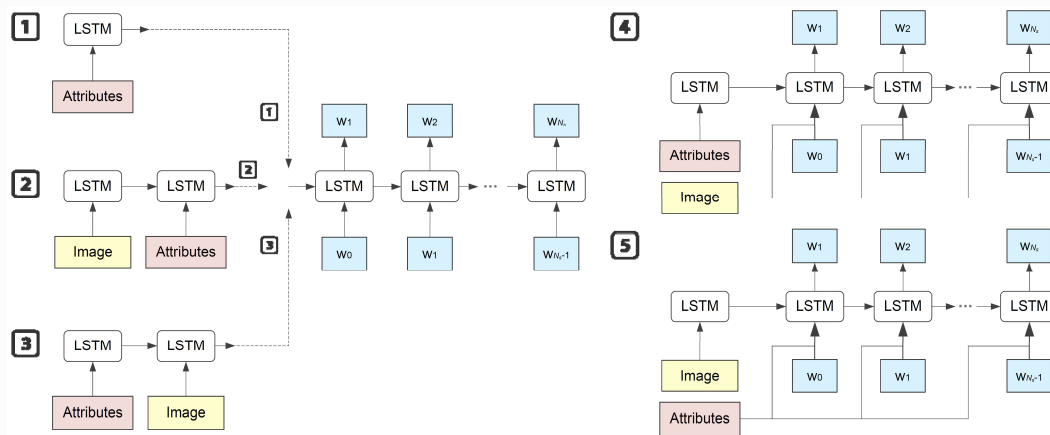
1. actionness = proposal (highlight)
2. Kinetics is the dataset for trimmed video classification in ActivityNet

Performance on validation set in temporal action proposal task

| Network         | Pre-trained | AUC   |
|-----------------|-------------|-------|
| ResNet          | ImageNet    | 59.03 |
| ResNet          | +Kinetics   | 60.13 |
| P3D ResNet      | Sports-1M   | 60.76 |
| P3D ResNet      | +Kinetics   | 61.13 |
| Fusion (4 in 1) | --          | 63.12 |
| Test Server     | --          | 64.18 |

108

## Observation 1: attributes work for captioning



109

## Observation 1: attributes work for captioning

- COCO Image Captioning [Yao, ICCV'17]
  - 82,783 training, 5,000 validation and 5,000 testing
  - 5 sentences per image
  - Image feature: GoogleNet; Attribute detector: Multiple Instance Learning [Fang *et al.*, CVPR15]

| Model               | Group                | B@1         | B@2         | B@3         | B@4         | METEOR      | ROUGE-L     | CIDEr-D      | SPICE       |
|---------------------|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|
| LRCN                | UC Berkeley, CVPR'15 | 69.7        | 51.9        | 38.0        | 27.8        | 22.9        | 50.8        | 83.7         | 15.8        |
| HA                  | UdeM, ICML'15        | 71.8        | 50.4        | 35.7        | 25.0        | 23.0        | -           | -            | -           |
| SA                  |                      | 70.7        | 49.2        | 34.4        | 24.3        | 23.9        | -           | -            | -           |
| ATT                 | Rochester, CVPR'16   | 70.9        | 53.7        | 40.2        | 30.4        | 24.3        | -           | -            | -           |
| SC                  | Michigan, arxiv'17   | 72.0        | 54.6        | 40.4        | 29.8        | 24.5        | -           | 95.9         | -           |
| LSTM-A <sub>1</sub> | Ours                 | 72.9        | 56.2        | 42.4        | 31.9        | 25.1        | 53.4        | 97.5         | 18.1        |
| LSTM-A <sub>2</sub> |                      | 73.3        | 56.5        | 42.7        | 32.3        | 25.3        | 53.9        | 99.1         | 18.3        |
| LSTM-A <sub>3</sub> |                      | <b>73.5</b> | 56.6        | 42.9        | 32.4        | <b>25.5</b> | 53.9        | 99.8         | 18.5        |
| LSTM-A <sub>4</sub> |                      | 72.1        | 55.5        | 41.7        | 31.4        | 24.9        | 53.2        | 95.7         | 17.8        |
| LSTM-A <sub>5</sub> |                      | 73.4        | <b>56.7</b> | <b>43.0</b> | <b>32.6</b> | 25.4        | <b>54.0</b> | <b>100.2</b> | <b>18.6</b> |
| LSTM-A*             | Upper bound          | <b>95.7</b> | <b>82.5</b> | <b>68.5</b> | <b>55.9</b> | <b>34.1</b> | <b>67.3</b> | <b>150.5</b> | <b>26.8</b> |

110

## Observation 2: generation + retrieval

- Video captioning (e.g., YouTube2Text, MSR-VTT)



1. A man is playing the guitar.
  2. A men is playing instrument.
  3. A man plays a guitar.
  4. A man is singing and playing guitar.
  5. The boy played his guitar.
- ....

Descriptive (objective)

- Dense-captioning events (ActivityNet video captions)



"They seem to be having a great time as the camera pulls back to show them all."

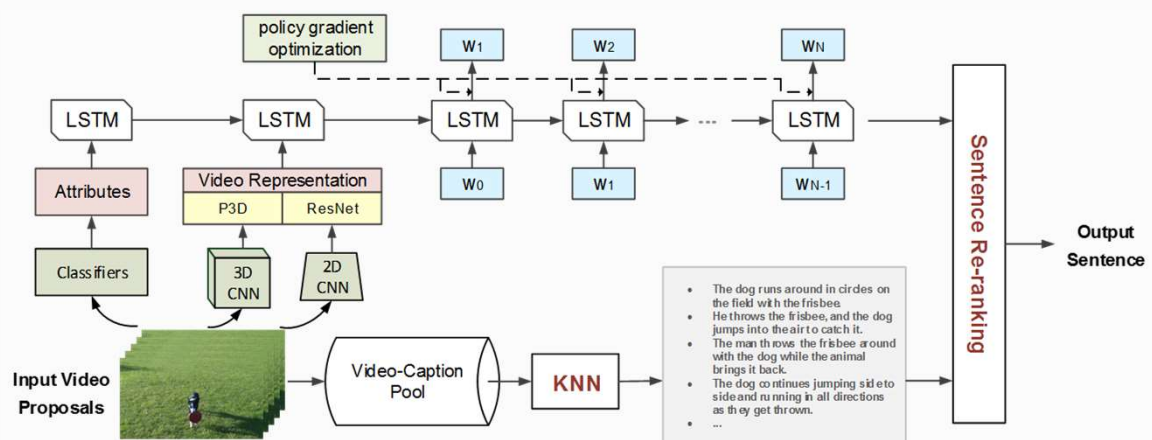
Subjective

Generation + Retrieval

111

## Event Description Generation

- Attribute-augmented captioning + retrieval



112

## Evaluations (Rank 1 @ ActivityNet'17)

- ActivityNet captions: 19,994 Youtube videos (10,024 training, 4,926 validation, 5,044 testing)
- ~3.65 event proposals for each video, one ground-truth sentence for each event proposal
- Word representation: one-hot vector
- Video representation: ResNet (Kinetics) + P3D ResNet (Sports-1M)
- Attributes: 200 categories in untrimmed video classification dataset

| Model   | BLEU@4%     | METEOR%      | ROUGE-L%     | CIDEr-D%     |   | Team                  | METEOR% |
|---|-------------|--------------|--------------|--------------|---|-----------------------|---------|
| <b>LSTM-A<sub>3</sub></b>                               | <b>3.38</b> | 7.71         | 13.27        | <b>16.08</b> | 1 | MSRA                  | 12.84   |
| <b>LSTM-A<sub>3</sub> + policy gradient</b>             | 3.07        | 8.47         | 14.28        | 13.82        | 2 | USTC                  | 9.87    |
| <b>LSTM-A<sub>3</sub> + policy gradient + retrieval</b> | 3.13        | <b>8.73</b>  | <b>14.29</b> | 14.75        | 3 | RUC & CMU             | 9.61    |
| <b>Test Server</b>                                      | -           | <b>12.84</b> | -            | -            | 4 | Stanford U (baseline) | 4.82    |

## Dense video captioning results



1. A man is standing in a room playing a game of a tennis ball and hitting the ball in the court. [0.271, 4.000]



2. Two men are seen standing in a room with a tennis racket and hitting a ball around. [4.271, 25.337]



3. The man is then seen playing a ball in the room and hitting a ball around. [32.090, 34.382]



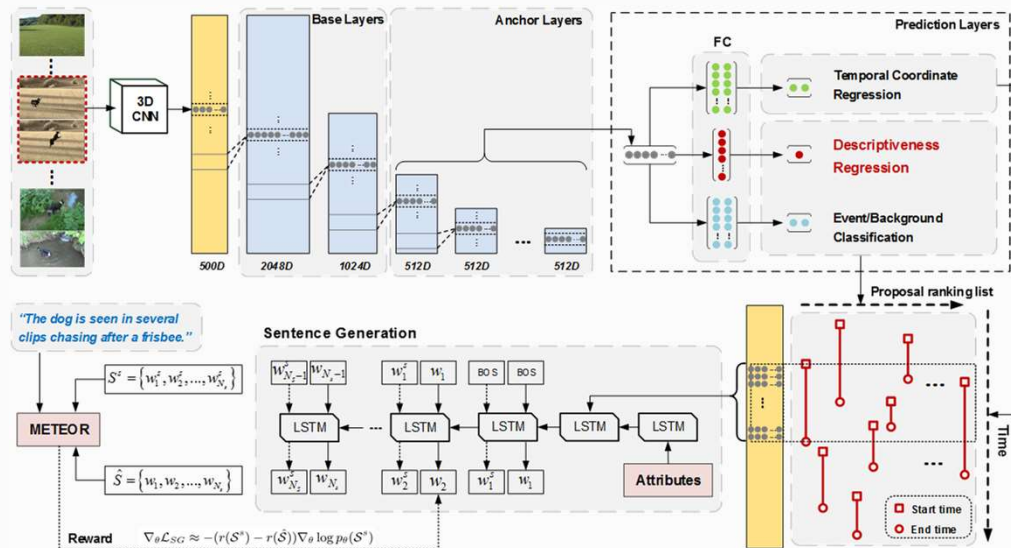
4. Two men are seen holding tennis rackets and hitting a ball back and forth to one another off the wall. [8.457, 35.596]



5. The men continue hitting the ball back and forth to one another while racing around the room to hit the ball. [28.344, 37.936]



# Joint Localization and Generation [Li, Yao, Mei, CVPR'18]



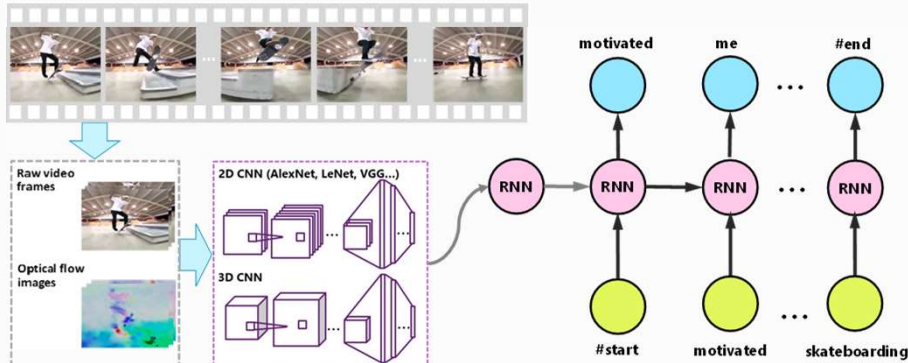
115

## Evaluations on ActivityNet Captions test server

| Team                                  | METEOR%      |
|---------------------------------------|--------------|
| <b>Li, Yao, Mei, CVPR'18</b>          | <b>12.96</b> |
| MSRA, ActivityNet challenge@CVPR17    | 12.84        |
| Tencent AI & SCUT, CVPR'18            | 10.72        |
| Salesforce, CVPR'18                   | 10.12        |
| RUC&CMU, ActivityNet challenge@CVPR17 | 9.61         |
| Boston & Disney Research, arxiv'18    | 8.81         |
| Stanford, ICCV'17                     | 4.82         |

116

# Video commenting [Li, MM'16]

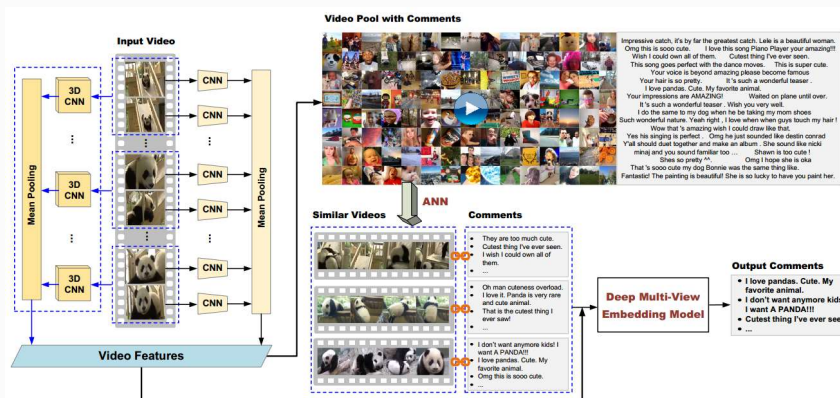


- General-purpose phases often appear  
"It is amazing." "OMG that was awesome!" "That is cool!"
- Comments in the training data are very diverse  
"I love how you ride a skateboard." "After I saw this I wish I could skate board."
- Difficult to establish a mapping from video to comments

117

# Video commenting

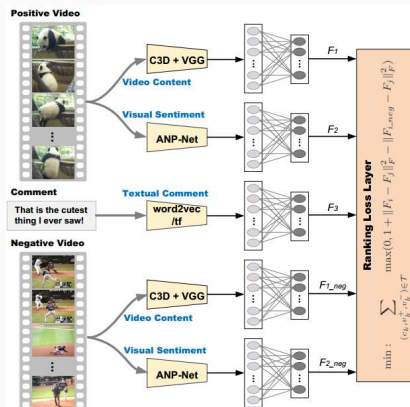
- Video Commenting by Search and Multi-View Embedding [Li, MM'16]
  - Similar video search (VS)
  - Comment dynamic ranking (DR)



118

# Video commenting

- Video Commenting by Search and Multi-View Embedding [Li, MM'16]
  - Similar video search (VS)
  - Dynamic ranking of comments (DR)



- Ranking loss

$$\min : \sum_{(c_k, v_k^+, v_k^-) \in \mathcal{T}} \max(0, 1 + \|F_i - F_j\|_F^2 - \|F_{i\_neg} - F_j\|_F^2)$$

$$s.t. \quad i, j = 1, \dots, 3, \quad i \neq j, \quad i \neq 3.$$

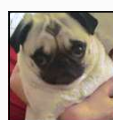
- Prediction

$$r(\hat{v}, \hat{c}) = \|F_1(\hat{v}) - F_3(\hat{c})\|_F^2 + \|F_2(\hat{v}) - F_3(\hat{c})\|_F^2.$$

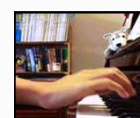
119

# Video commenting

- Dataset
  - 102K videos from vine.com
  - 10.6M comments from 12 categories
  - 5~15 sec for each video clip
- Video representation
  - Video content: C3D, VGG, C3D + VGG
  - Comments: TF, word2vector
  - Visual sentiment: ANP (adj-noun pairs)
- Approaches
  - Random Selection (RS)
  - Two-view CCA (CCA-VT)
  - Three-view CCA (CCA-VST)
  - Deep Two-view Embedding (DE-VT)
  - Deep Three-view Embedding (DE-VST)

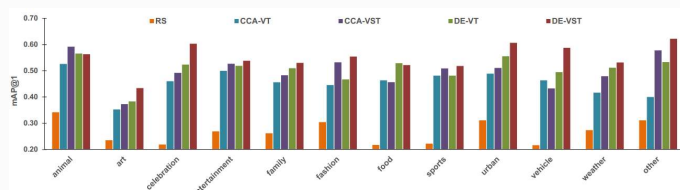


"Haha so cute and funny at the same time"  
"Glad she is better. So cute"



"Such outstanding piano pieces and you play them sublimely :)"  
"Amazing. I was listening to this while studying!"

| Approach | mAP@1 | mAP@2 | mAP@3 | mAP@4 | mAP@5 |
|----------|-------|-------|-------|-------|-------|
| RS       | 0.259 | 0.244 | 0.219 | 0.203 | 0.191 |
| CCA-VT   | 0.458 | 0.421 | 0.399 | 0.389 | 0.382 |
| CCA-VST  | 0.501 | 0.465 | 0.439 | 0.429 | 0.419 |
| DE-VT    | 0.504 | 0.469 | 0.447 | 0.433 | 0.422 |
| DE-VST   | 0.549 | 0.513 | 0.486 | 0.471 | 0.459 |



The mAP@1 performance for all the 12 categories.

120

## Results: auto-commenting

### Test video:



- \* 不止漂亮 0.522  
Not just beautiful
- \* 你好漂亮 0.497589  
You are so beautiful
- \* 好美, 喜欢看自拍视频的 0.4942  
Gorgeous. Love to watch homemade video
- \* 心目中的女神是不整容的 0.4904  
Goddess doesn't need plastic surgery
- \* 美丽! 0.4857  
Beautiful



- \* 今天吃得真淑女 0.4519  
Eating like a lady with great manner
- \* 吃的越来越干净了 0.4238  
Getting better at learning how to eat
- \* 好想亲下momo的小嘴唇 0.3901  
Want to kiss momo's little lips
- \* 吃得吧唧吧唧 0.3600  
Eating very enjoyable
- \* 看看吃饭是一种享受 0.3573  
It is enjoyable just to watch someone eats

### Top-K similar videos:



- \* 很漂亮  
so beautiful
- \* 笑容好美  
beautiful smile
- \* 美美美  
pretty
- \* 哪里出的美女  
where did this beautiful lady come from
- \* 好美啊  
so beautiful



- \* 今天吃得真淑女  
Eating like a lady with great manner
- \* 吃得吧唧吧唧  
Eating very enjoyable
- \* 每天都在变得更漂亮  
Become prettier every single day
- \* 不然不容易消化  
It will be hard to digest
- \* 不要在吃饭的时候教她说话  
Don't teach her talking while eating



- \* 不止漂亮  
Not just beautiful
- \* 好美, 喜欢看自拍视频的  
Gorgeous. Love to watch homemade video
- \* 有点韩国人的感觉  
Looks a bit like Korean
- \* 眨眼, 真美  
Catches the eyes, so pretty
- \* 美美的  
Beautiful



- \* 冉冉妈24小时陪孩子  
Ran's mom stays with her for 24h
- \* 看着冉冉每天都在成长进步  
Watching 冉冉 grow and progress every single day
- \* 小宝宝怕冷也怕热, 穿的少了舒服  
Baby is sensitive to both cold and hot
- \* 下班回去我带  
I will take care of her after work
- \* 太喜欢冉冉了  
Like 冉冉 too much



- \* 你好漂亮  
You are so beautiful
- \* 心目中的女神是不整容的  
Goddess doesn't need plastic surgery
- \* 很好看, 没有大浓妆, 但很抢眼  
Great look, no heavy makeup but it catches the eyes
- \* 女神  
Goddess
- \* 美哒哒  
Beautiful



- \* 吃的真香  
Enjoying the yummy food
- \* 好享受的样子  
It seems so enjoyable
- \* 小吃货  
Little Foodie
- \* 包括米粉么?  
Does it include rice noodles?
- \* 不像混血, 反而像中国BB  
Doesn't look like MIX but a Chinese baby



- \* 五官真好看  
Beautiful facial
- \* 美女耶  
Pretty lady
- \* 你好自恋哦! 美女  
You are such a narcissist
- \* 美女  
Beautiful lady
- \* 大众美女脸  
Generally beautiful face



- \* 好喜欢朵朵  
Liking 朵朵 so much
- \* 吃的真文明  
Eating with such great manner
- \* 朵朵好会吃饭  
朵朵 can eat so well
- \* 干吃面没菜菜啊  
Just noodles?
- \* 用牛肉汤煮的  
Cook it with beef stock



- \* 美丽!  
Beautiful
- \* 美美哒  
Beautiful
- \* 白衬衣美哭了  
The white shirt is so pretty
- \* 太阳女神美美哒  
The Goddess of Sun is beautiful
- \* 美翻了啦  
Outrageously beautiful



- \* 吃的越来越干净了  
Getting better at learning how to eat
- \* 好想亲下momo的小嘴唇  
Want to kiss momo's little lips
- \* 看看吃饭是一种享受  
Enjoyable just to watch someone eats
- \* momo吃的好香啊  
Momo is enjoying her food
- \* 14 months

## Outline

### Part I:

- Recent advances in vision and language (15 min)
- Image to language (recognition & captioning & poetry) (45 min)
- Break (15 min)

### Part II

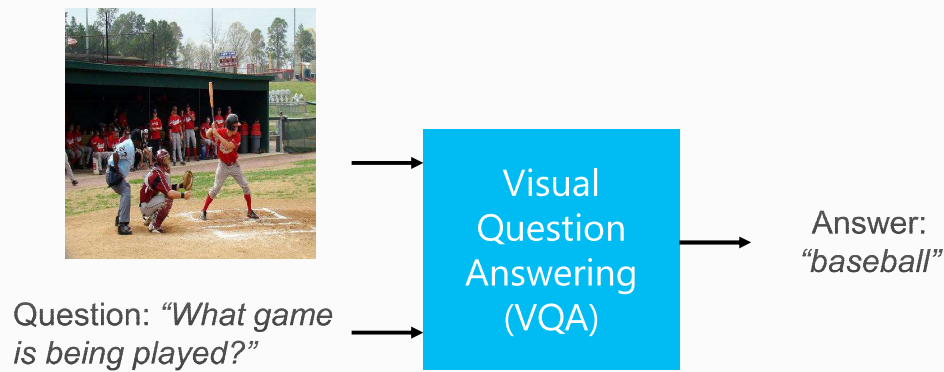
- Video to language (recognition & captioning & commenting) (45 min)
- **Visual question answering** (15 min)
- Break (15 min)

### Part III

- Image and video generation (generation & translation) (10 min)
- Datasets and evaluations (10 min)
- Open issues and Q&A (5 min)

# Visual Question Answering

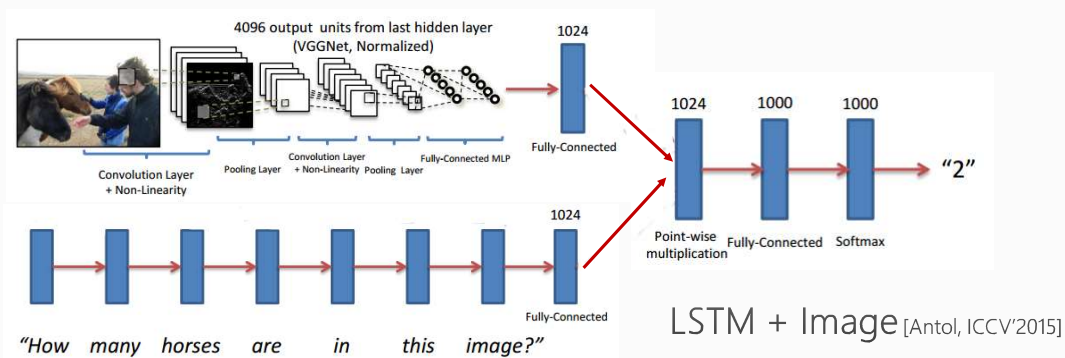
- Answer natural language questions according to the content of a reference image



123

## VQA paradigm and challenges

| Model  | Acc (%) |
|--------|---------|
| LSTM+I | 53.7    |



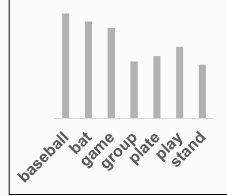

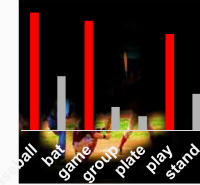



- Image modeling
  - CNN, Semantic Vector, CNN + Attention, Multi-level Attention
- Question modeling
  - Bag-of-Words (BOW), RNN, Sentence-CNN, Textual Attention
- Multimodal feature fusion
  - Element-wise Multiplication, Compact Bilinear Pooling, Low-rank Bilinear Pooling

124

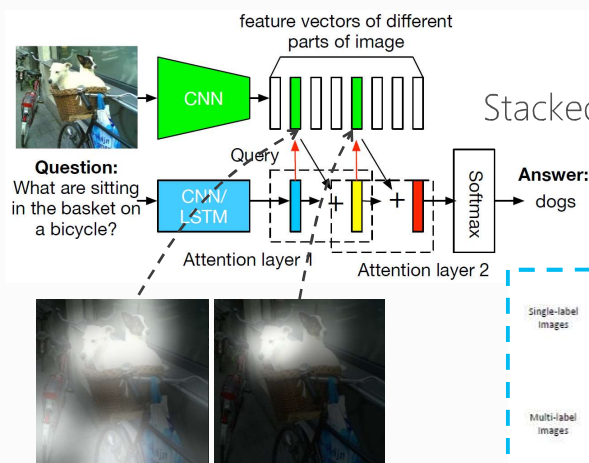


## VQA with "X"

|   |   |   |   |  |   |
|---|---|---|---|--|---|
| Q: what game is being played?   | Q: what game is being played?   | Q: what game is being played?   | Q: <b>what game</b> is being <b>played</b> ?                                      | Q: what game is being played?  | Q: <b>what game</b> is being <b>played</b> ?<br>Candidate Ans: <b>baseball</b>      |
|  |  |  |  |  |  |
| <b>X = visual attention</b><br>[Yang, CVPR'2016;<br>Shih, CVPR'2016]              | <b>X = visual attributes</b><br>[Wu, CVPR'2016]                                   | <b>X = visual-question co-attention</b><br>[Lu, NIPS'2016]                        | <b>X = multi-level attention</b><br>[Yu, CVPR'2017]                               | <b>X = &lt;Q,I,A&gt; triple reasoning</b><br>[Bai, ECCV'2018]                      |   |

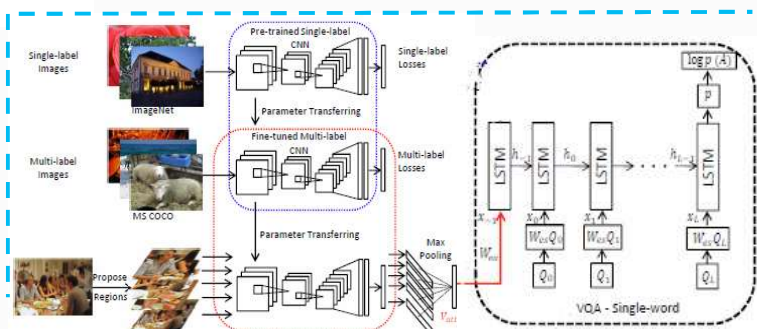
125

## Visual attention and attributes



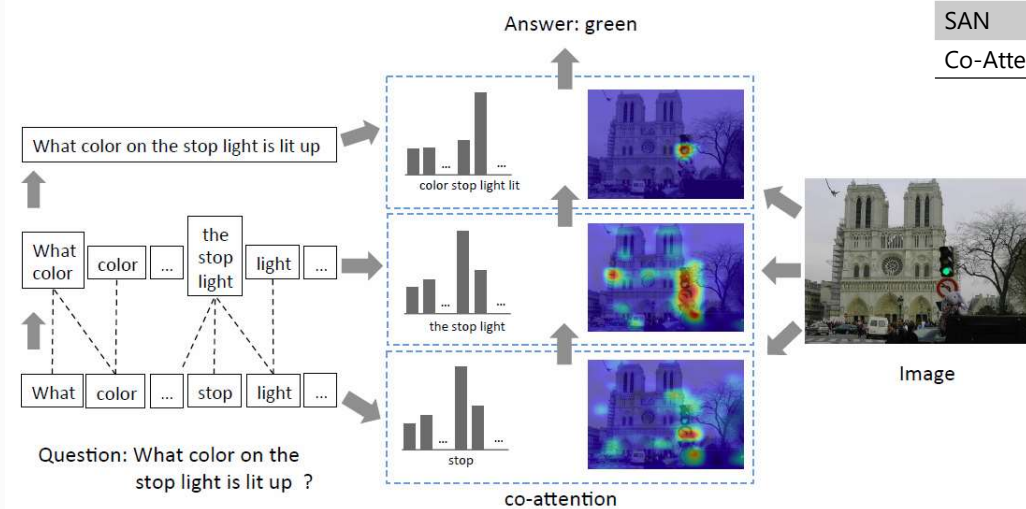
Stacked Attention Networks [Yang, CVPR'2016]

| Model       | Acc (%) |
|-------------|---------|
| LSTM+I      | 53.7    |
| Att-KB+LSTM | 57.5    |
| SAN         | 58.7    |



126

## Visual-question co-attention

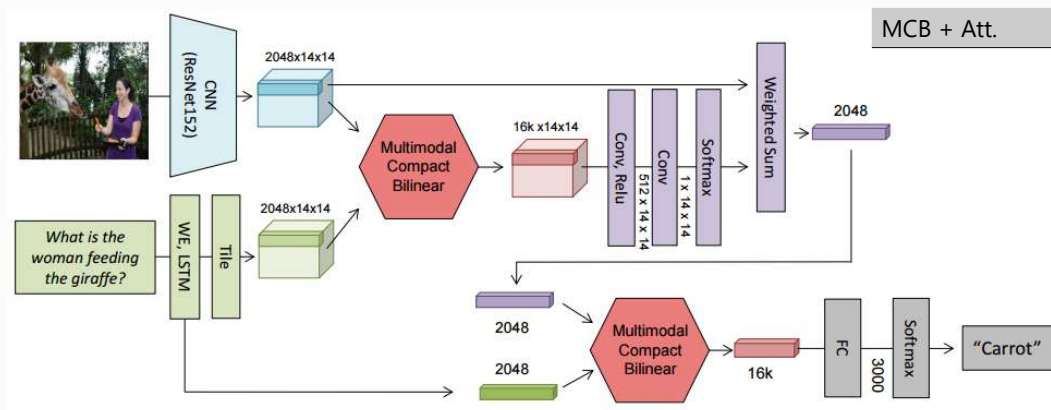


Visual-question Co-Attention [Lu, NIPS'2016]

| Model        | Acc (%) |
|--------------|---------|
| LSTM+I       | 53.7    |
| Att-KB+LSTM  | 57.5    |
| SAN          | 58.7    |
| Co-Attention | 61.8    |

127

## Multi-modality Bilinear Fusion

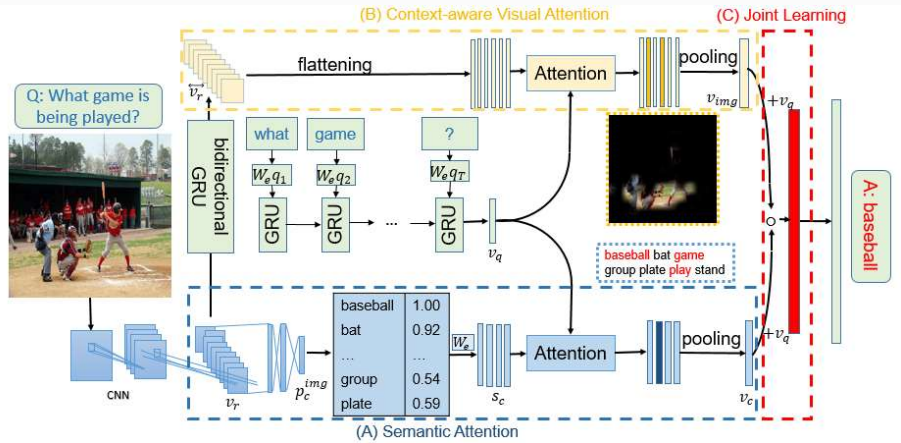


Multimodal Compact Bilinear with Attention [Fukui, EMNLP'2016]

| Model        | Acc (%) |
|--------------|---------|
| LSTM+I       | 53.7    |
| Att-KB+LSTM  | 57.5    |
| SAN          | 58.7    |
| Co-Attention | 61.8    |
| MCB + Att.   | 64.2    |

128

# Multi-level Attention [Yu & Mei, CVPR'17]

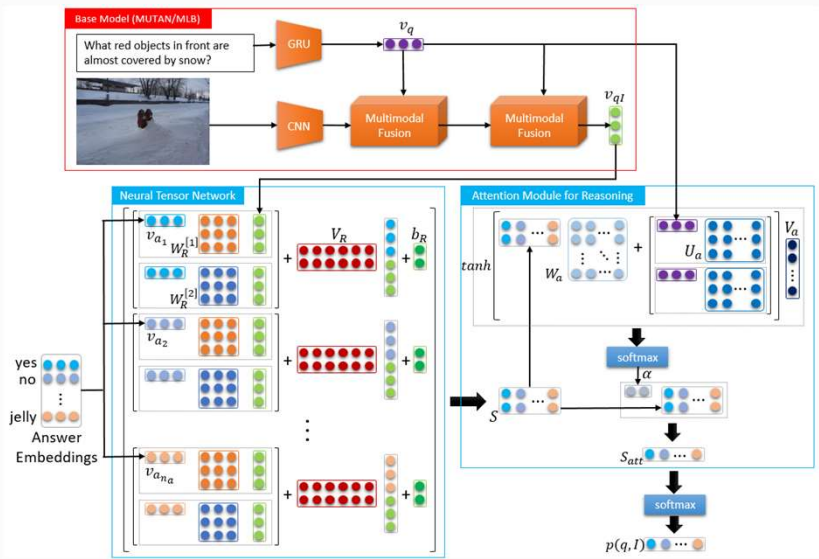


Multi-level Attention [Yu@MSRA, CVPR'2017]

| Model            | Acc (%) |
|------------------|---------|
| LSTM+I           | 53.7    |
| Att-KB+LSTM      | 57.5    |
| SAN              | 58.7    |
| Co-Attention     | 61.8    |
| MCB + Att.       | 64.2    |
| Multi-level Att. | 65.4    |

129

# Triplet Reasoning [Bai & Mei, ECCV'18]



| Model             | Acc (%) |
|-------------------|---------|
| LSTM+I            | 53.7    |
| Att-KB+LSTM       | 57.5    |
| SAN               | 58.7    |
| Co-Attention      | 61.8    |
| MCB + Att.        | 64.2    |
| Multi-level Att.  | 65.4    |
| Triplet Reasoning | 67.6    |

Deep Attention Neural Tensor Network [Bai@JD, ECCV'2018]

130

## Outline

### Part I:

- Recent advances in vision and language (15 min)
- Image to language (recognition & captioning & poetry) (45 min)
- Break (15 min)

### Part II

- Video to language (recognition & captioning & commenting) (45 min)
- Visual question answering (15 min)
- Break (15 min)

### Part III

- Image and video generation (generation & translation) (10 min)
- Datasets and evaluations (10 min)
- Open issues and Q&A (5 min)

131

## Text to Image Synthesis Comparisons with State-of-the-arts (StackGAN, AttnGAN vs. DA-GAN)

This is a small bird. Its **head and back** are a vivid blue. Its **breast** is white, with a brown throat patch, the **eyes** are large relative to its size. The wings have white wing bars.

这是一只小鸟，它有蓝色得头和背部。它得胸部是白色的并带有棕色的纹理。它有相对于它的体型来说大大的眼睛，它的翅膀上有白色的带状纹理。



StackGAN



DA-GAN

the bird has a **yellow** crown and a **black** eyering that **is round**



AttnGAN (Xiaodong He, etc.)

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. Han Zhang, Tao Xu, etc. ICCV 2017.

AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, Tao Xu, Xiaodong He, etc.

This small bird has short beak and **dark stripe** down the top, the **wings** are a **mixture of brown, white and black** and the **upper breast** is **white** and has **black strips**.

这是小鸟有**短短的嘴巴**，身体从上到下覆盖有**深色的条纹**。**翅膀**混合有**白色，黑色和深褐色**。它的**上胸部**是**白色的**并有**黑色条纹**。



StackGAN



DA-GAN(ours)

This bird has a **yellow underbelly** and **chest with black stripes** and **grey, black, and white** feathers on his **head and wings**.

这只鸟有**黄色的腹部和胸部**，上面覆盖有**白色条状纹理**。在它的**头部和翅膀**上有**灰色，黑色和白色**的羽毛。

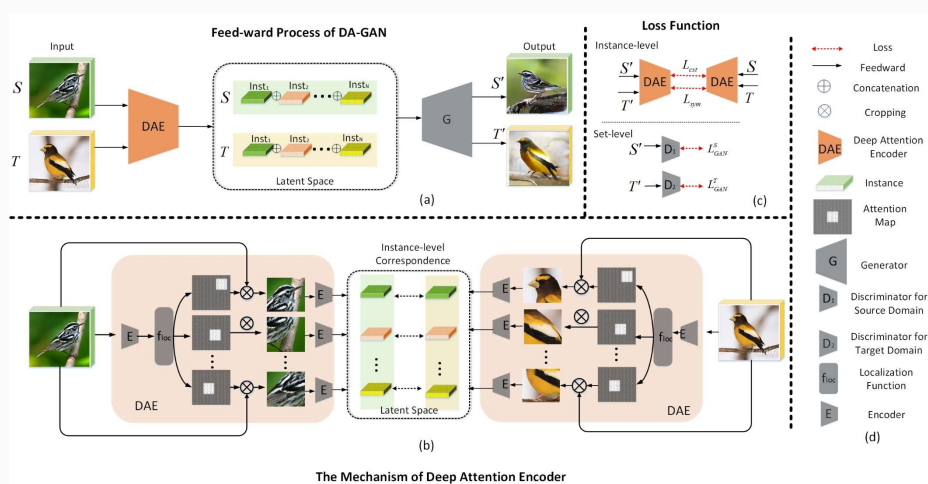


StackGAN



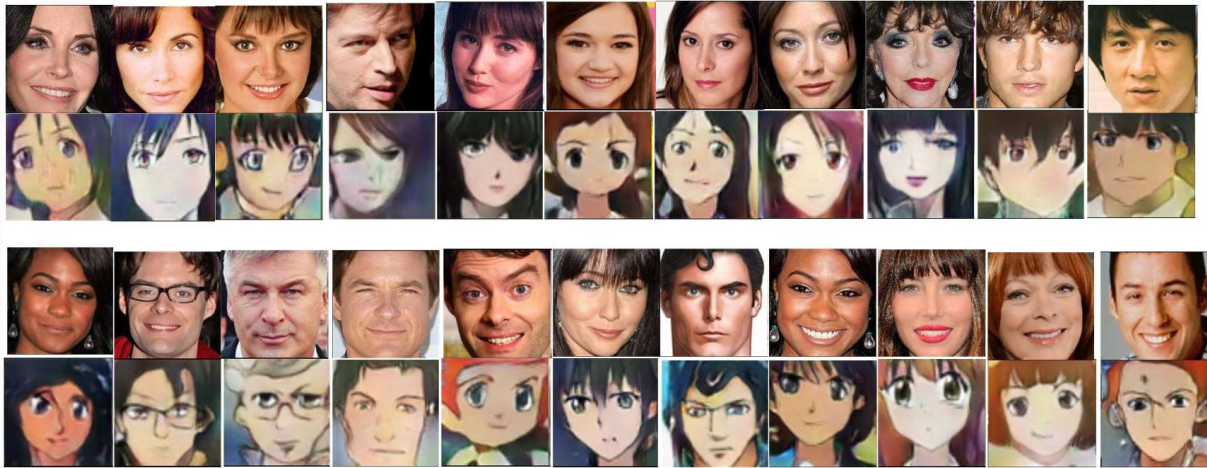
DA-GAN(ours)

## DA-GAN: Deep Attention Generative Adversarial Networks [Ma, CVPR'18]





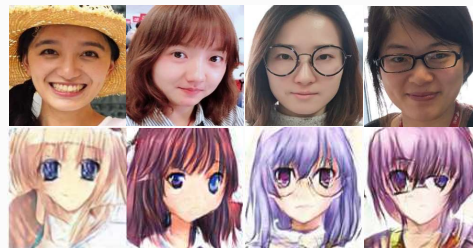
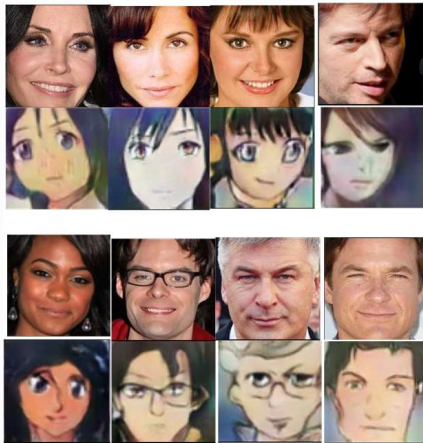
## Human face to Cartoon face Generation Results by DA-GAN



## Skeleton to Cartoon Figure Generation

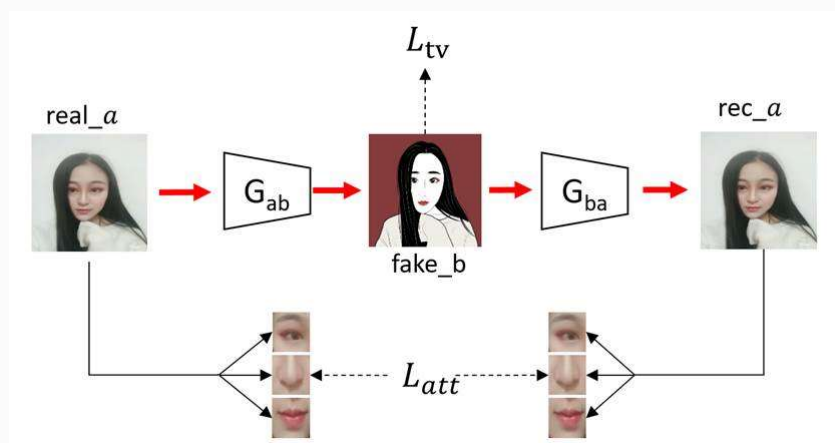


## Cartoon Face Generation - Japanese style



137

## Cartoon Face Generation - Hand-painted style



**Total variation loss**  $L_{tv} = E_{x \sim P_{data}}[|\nabla G_{ab}(x)|_1]$

**Attention loss**  $L_{att} = E_{x \sim P_{data}}[|G_{ba}(G_{ab}(x))_m - x_m|_1], m \in \{eye, nose, mouth\}$

138

## CV for JD 618 Promotion



**Selfie Segmentation**  
mIOU > 94%



**First day launching: 390K UV , 430K PV**  
**Accumulated UV : 1.4M; Accumulated PV : 1.6M**

139

## Language-to-video generation

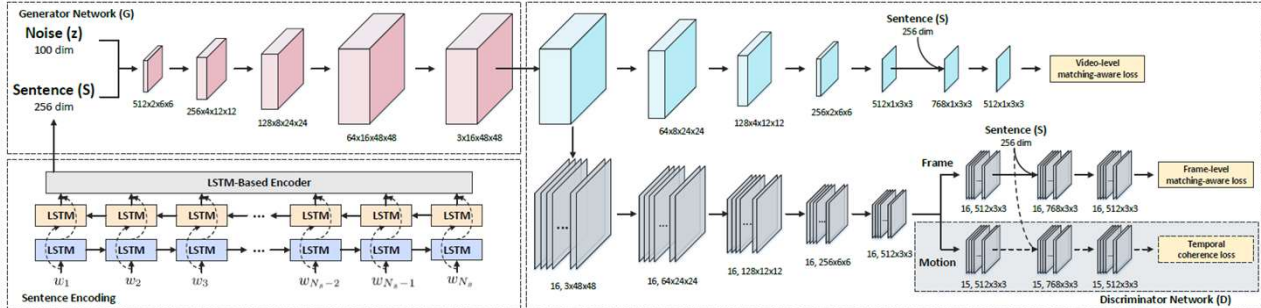
- Key issues in language-to-video
  - *Temporal coherence*: adjacent video frames are smoothly connected over time
  - *Semantic match*: relevance between video and caption
- Joint adversarial learning (TGANs-C): Temporal coherence + Semantic match
  - 2D generator -> 3D generator
  - Basic discriminator-> frame-level/video-level matching-aware discriminator
  - Motion discriminator: distinguish the displacement between consecutive real or generated frames

140



# Temporal GANs conditioning on Captions (TGANs-C)

[Pan, Yao, Mei, ACM MM 2017]

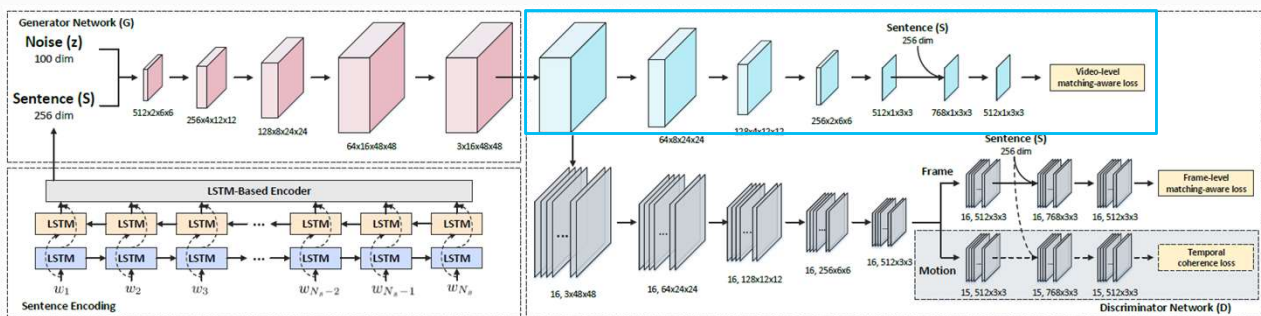


- Video discriminator  $D_0(v, S) (\{\mathbb{R}^{d_v}, \mathbb{R}^{d_s}\} \rightarrow [0, 1])$
- Frame discriminator  $D_1(f^i, S) (\{\mathbb{R}^{d_f}, \mathbb{R}^{d_s}\} \rightarrow [0, 1])$
- Motion discriminator  $D_2(f^i, f^{i-1}) (\{\mathbb{R}^{d_f}, \mathbb{R}^{d_f}\} \rightarrow \mathbb{R}^{d_{c0} \times d_{h0} \times d_{d0}})$

141

# Temporal GANs conditioning on Captions (TGANs-C)

[Pan, Yao, Mei, ACM MM 2017]



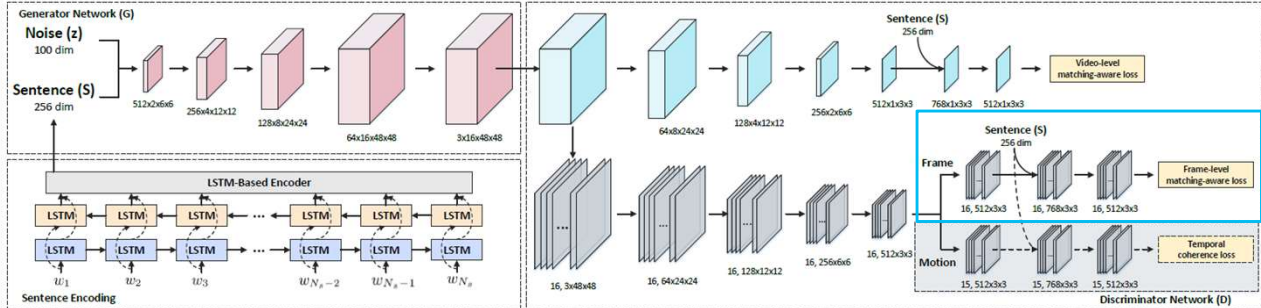
- Video discriminator  $D_0(v, S) (\{\mathbb{R}^{d_v}, \mathbb{R}^{d_s}\} \rightarrow [0, 1])$ 
  - Distinguish real video from synthetic one and align video with the correct caption
  - Video-level matching-aware loss:

$$\mathcal{L}_v = -\frac{1}{3} [\log(D_0(v_{real+}, S)) + \log(1 - D_0(v_{real-}, S)) + \log(1 - D_0(v_{syn+}, S))]$$

142

# Temporal GANs conditioning on Captions (TGANs-C)

[Pan, Yao, Mei, ACM MM 2017]



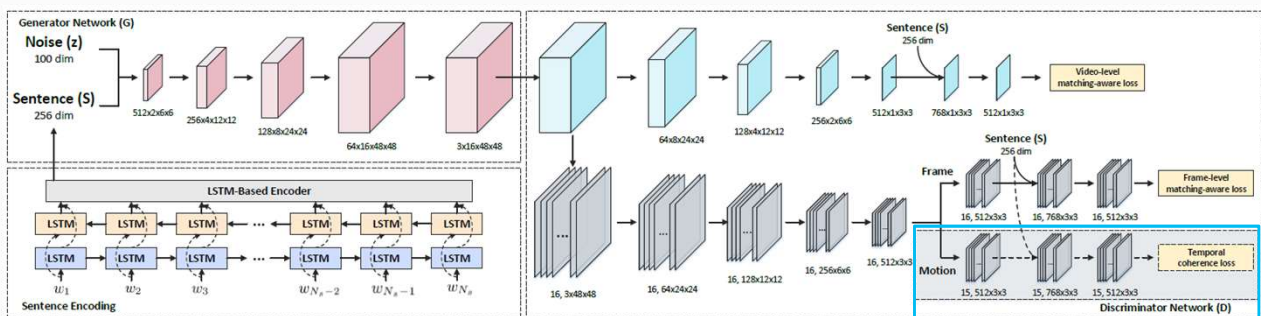
- Frame discriminator  $D_1(f^i, S) (\{\mathbb{R}^{d_f}, \mathbb{R}^{d_s}\} \rightarrow [0, 1])$ 
  - Determine whether each frame is real/fake and semantically matched/mismatched with caption
  - Frame-level matching-aware loss:

$$\mathcal{L}_f = -\frac{1}{3d_l} \left[ \sum_{i=1}^{d_l} \log(D_1(f_{real+}^i, S)) + \sum_{i=1}^{d_l} \log(1 - D_1(f_{real-}^i, S)) \right. \\ \left. + \sum_{i=1}^{d_l} \log(1 - D_1(f_{syn+}^i, S)) \right]$$

143

# Temporal GANs conditioning on Captions (TGANs-C)

[Pan, Yao, Mei, ACM MM 2017]



- Motion discriminator  $D_2(f^i, f^{i-1}) (\{\mathbb{R}^{d_f}, \mathbb{R}^{d_f}\} \rightarrow \mathbb{R}^{d_{c0}} \times d_{h0} \times d_{d0})$

- Exploit temporal coherence between consecutive frames

- Temporal coherence loss:

Scheme 1: Temporal coherence **constraint** loss

$$\mathcal{D}(f^i, f^{i-1}) = \|\mathbf{m}_{f^i} - \mathbf{m}_{f^{i-1}}\|_2^2 = \|\bar{\mathbf{m}}_{f^i}\|_2^2$$

$$\mathcal{L}_t^{(1)} = \frac{1}{d_l - 1} \sum_{i=2}^{d_l} \mathcal{D}(f_{syn+}^i, f_{syn+}^{i-1})$$

Scheme 2: Temporal coherence **adversarial** loss

$$\mathcal{L}_t^{(2)} = -\frac{1}{3(d_l-1)} \left[ \sum_{i=2}^{d_l} \log(\Phi_2(\bar{\mathbf{m}}_{f_{real+}^i}, S)) \right. \\ \left. + \sum_{i=2}^{d_l} \log(1 - \Phi_2(\bar{\mathbf{m}}_{f_{real-}^i}, S)) \right. \\ \left. + \sum_{i=2}^{d_l} \log(1 - \Phi_2(\bar{\mathbf{m}}_{f_{syn+}^i}, S)) \right]$$

144

## Dataset

- Single-Digit Bouncing MNIST GIFs (SBMG)
  - 12,000 GIFs with a single 28\*28 digit moving left-right or up-down
  - Each GIF is 16 frames long
- Two-Digit Bouncing MNIST GIFs (TBMG)
  - 12,000 GIFs with two 28\*28 digits moving left-right or up-down
  - Each GIF is 16 frames long
- Microsoft Research Video Description Corpus (MSVD)
  - Manually filter out 518 cooking videos from 1,970 YouTube video snippets
  - ~40 human-generated sentences for each clip

sentence: "digit 2 is left and right."



sentence: "digit 3 is up and down."



(a) Single-Digit Bouncing MNIST GIFs

sentence: "digit 6 is up and down and digit 3 is left and right."



sentence: "digit 9 is up and down and digit 8 is up and down."



(b) Two-Digit Bouncing MNIST GIFs

sentence: "a woman is slicing a cucumber into pieces."



sentence: "a man is pouring pancake mixture into a frying pan."



(c) Microsoft Research Video Description Corpus

145

## Evaluation metrics for captioning

- Objective metrics
  - Accuracy of S%, V%, O%
  - ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 04]
  - BLEU@4 (BiLingual Evaluation Understudy) [Papineni, ACL'02]  
[modified n-gram precision](#)
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee, ACL05]  
[similar with f-score combining precision and recall with a weight](#)
  - CIDEr (Consensus-based Image Description Evaluation) [Vedantam, 2014; COCO evaluation]
  - SPICE (Semantic Propositional Image Caption Evaluation) [Anderson, ECCV16]  
[how effectively captions recover objects, attributes and their relations over scene graphs](#)
- Subjective metrics – human evaluations
  - Coherence, Relevance, Helpful for Blind

146



## Reference

- T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring Visual Relationship for Image Captioning," ECCV, 2018.
- Y. Li, T. Yao, Y. Pan, et al. "Jointly Localizing and Describing Events for Dense Video Captioning," CVPR, 2018.
- Y. Bai, T. Mei, et al. "Deep Attention Neural Tensor Network for Visual Question Answering," ECCV, 2018.
- T. Yao, Y. Pan, Y. Li and T. Mei, "Boosting Image Captioning with Attributes," ICCV, 2017.
- T. Yao, T. Mei, et al. "Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects," CVPR, 2017.
- T. Yao, T. Mei, et al. "MSR ASIA MSM at ActivityNet Challenge 2017: Trimmed Action Recognition, Temporal Action Proposals and Dense-Captioning Events in Videos," CVPR ActivityNet Challenge Workshop, 2017.
- Y. Pan, T. Mei, T. Yao, et al. "Video Captioning with Transferred Semantic Attributes," CVPR, 2017.
- Y. Pan, T. Mei, T. Yao, et al. "Jointly Modeling Embedding and Translation to Bridge Video and Language," CVPR, 2016.
- J. Xu, T. Mei, et al. "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," CVPR, 2016.
- Karpathy, et al. "Deep Visual-Semantic Alignments for Generating Image Descriptions," CVPR 2014.
- Vinyals, et al. "Show and Tell: A Neural Image Caption Generator", 2014.
- Kiros, et al. "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," 2014.
- Donohue, et al. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," 2014.
- Xu, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," 2015.
- Yao, et al. "Describing Videos by Exploiting Temporal Structure," ICCV, 2015.
- Venugopalan, et al. "Sequence to sequence-video to text," ICCV, 2015.
- Yu, et al. "Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks," CVPR, 2016.
- Krishna, et al. "Dense-Captioning Events in Videos," ICCV, 2017.
- Zhou, Xiong, et al. "End-to-End Dense Video Captioning with Masked Transformer," CVPR, 2018.
- Yang, et al. "Dense Captioning with Joint Inference and Visual Context," CVPR 2017.
- Johnson, et al. "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," CVPR 2016.
- ...

147

## Learning materials

- Source codes for image captioning:
  - <https://github.com/karpathy/neuraltalk>, <https://github.com/karpathy/neuraltalk2>
  - LRCN for image caption: [https://github.com/jeffdonahue/caffe/tree/54fa90fa1b38af14a6fca32ed8aa5ead38752a09/examples/coco\\_caption](https://github.com/jeffdonahue/caffe/tree/54fa90fa1b38af14a6fca32ed8aa5ead38752a09/examples/coco_caption)
  - LRCN for action recognition: [https://github.com/LisaAnne/lisa-caffe-public/tree/lstm\\_video\\_deploy/examples/LRCN\\_activity\\_recognition](https://github.com/LisaAnne/lisa-caffe-public/tree/lstm_video_deploy/examples/LRCN_activity_recognition)
  - Show attend and tell <https://github.com/kelvinxu/arctic-captions>
- Source codes for video captioning:
  - Sequence to Sequence - Video to Text: <https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt>
  - Soft-attention: <https://github.com/yaoli/arctic-capgen-vid>
- Leaderboards:
  - COCO Image Captioning: <http://cocodataset.org/#captions-leaderboard>
  - Dense-Captioning Events in Videos: [http://activity-net.org/challenges/2017/evaluation.html#leaderboard\\_captioning](http://activity-net.org/challenges/2017/evaluation.html#leaderboard_captioning)
  - VQA challenge 2018: <http://visualqa.org/roe.html>

148

京东AI研究院

# 京东葡萄树计划



AI Scholars  
(visiting + projects)

AI Stars  
(internship)

AI Innovation  
(challenges)

[http://neuhub.jd.com/detail\\_page/jdgrapevine.html](http://neuhub.jd.com/detail_page/jdgrapevine.html)

2018 京东人工智能创新峰会  
智汇京东 · 开放共赢



Thanks!  
tmei@jd.com

2018