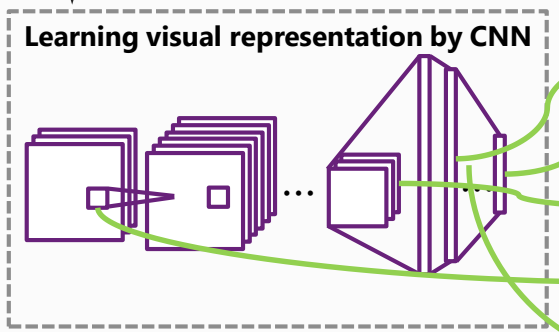
Vision and Language: Some Recent Progresses

Tao Mei

Senior Researcher/Research Manager
Microsoft Research Asia
http://research.microsoft.com/en-us/people/tmei/

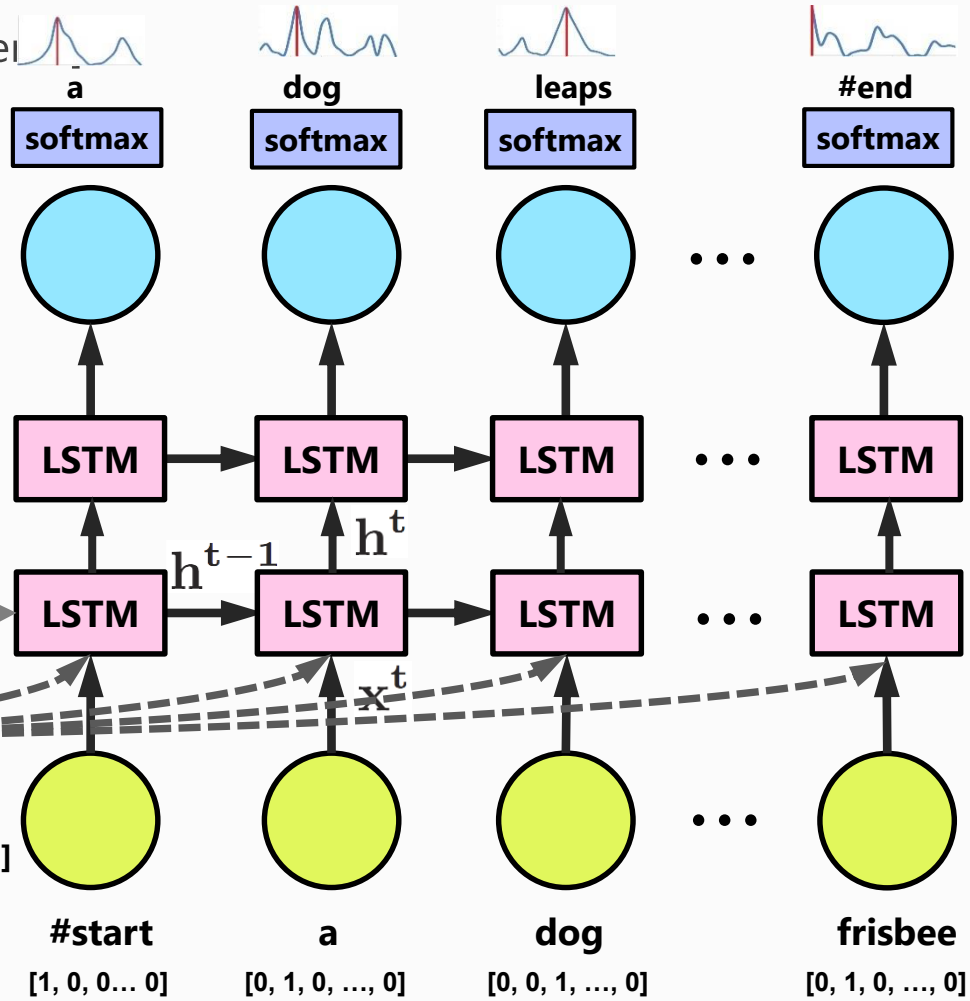
Image/video captioning

- Sequence learning-based approach
  [Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester



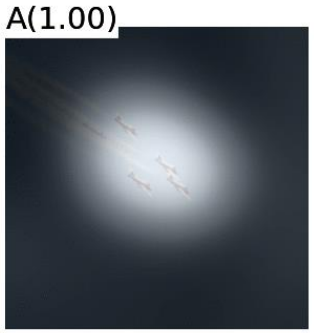
CNN Rep.
Attributes
Attention
Region
LSTM-Encoder

[Vinyals, CVPR15; Chen, CVPR15; Mao, ICLR15]
[Wu, CVPR16; Pan, 2016]
[Xu, ICML15; You, CVPR16]
[Karpathy & Fei-Fei, CVPR15]
[Sutskever, NIPS14]



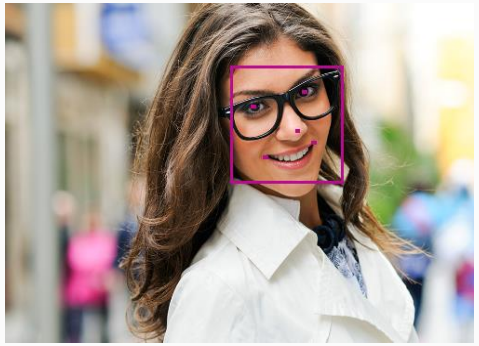
* Note that this figure only shows prediction process.

Image Captioning with X

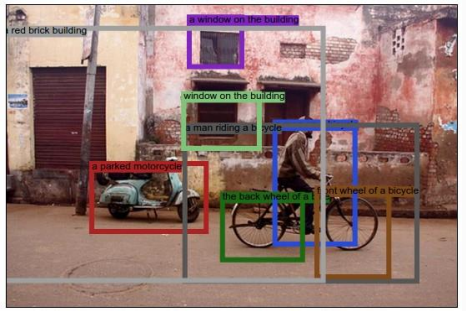


X = visual attention
[Xu, ICML'15]

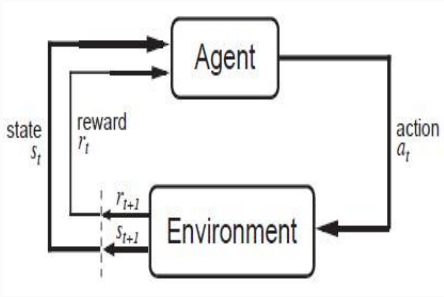
X = visual attributes
[You, CVPR'16; Wu, CVPR'16; Yao, arxiv'16; Pan, CVPR'17]



X = entity recognition
[Tran, CVPR'16]



X = dense caption
[Johnson, CVPR'16]

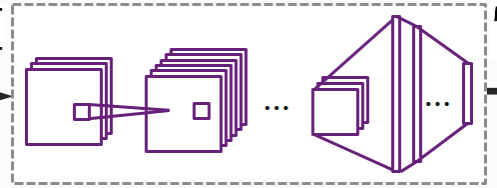


X = reinforcement learning
[Rennie, CVPR'17]

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

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

$\mathbf{I}$



$\mathbf{T}_v$

Visual representation by DCNN

$\mathbf{A}$

[dog: 0.95]
[frisbee: 0.83]
[outdoor: 0.82]
[grass: 0.81]
[leap: 0.45]

Visual Attributes by MIL

Attributes

$\mathbf{T}_a$

LSTM

LSTM

LSTM

LSTM

$\boldsymbol{w_1}$ $\boldsymbol{w_2}$ $\boldsymbol{w_N}$

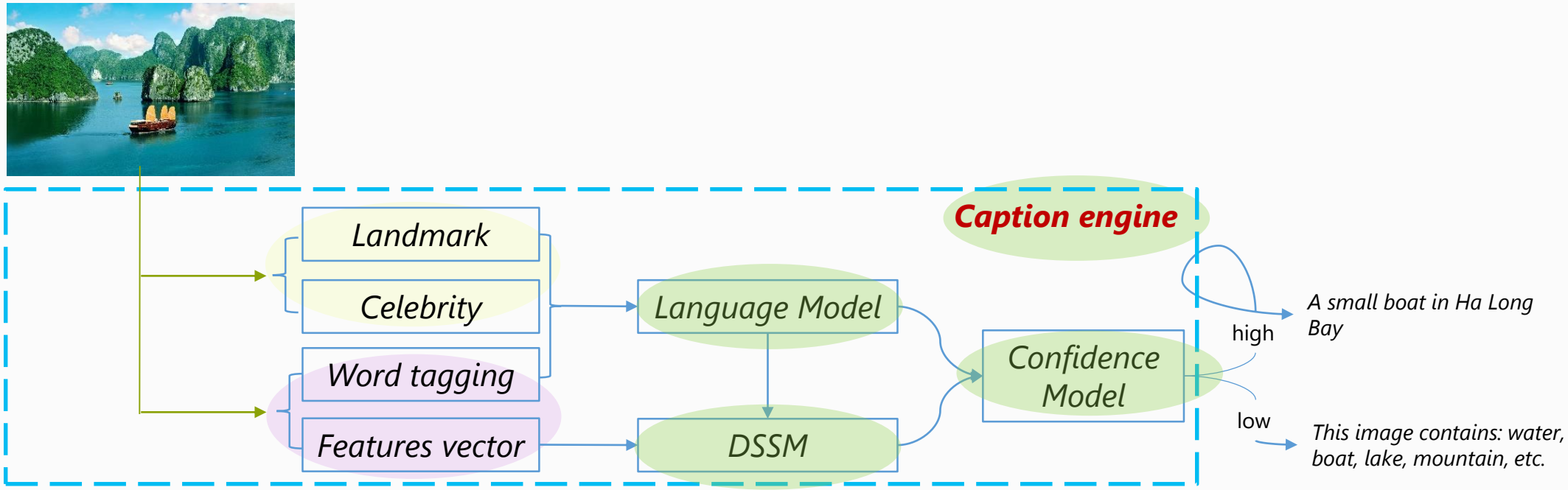
$\mathbf{h}^t$

$\mathbf{x}^t$

$\mathbf{T}_s$ $\mathbf{T}_s$ $\mathbf{T}_s$

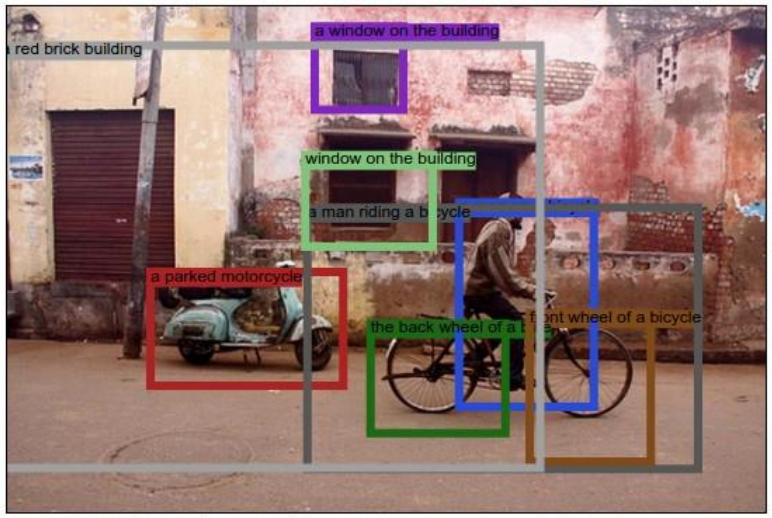
$\boldsymbol{w_0}$ $\boldsymbol{w_1}$ $\boldsymbol{w_{N-1}}$

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

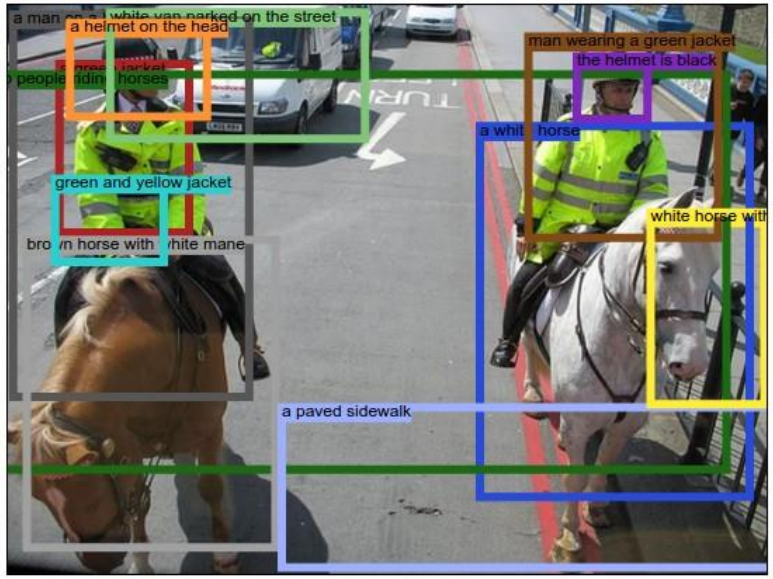


- Entity recognition: extreme classification w/ large set of celebrities (precision 99% coverage ~60%) [Guo, 2016]
- Language model: maximum entropy [Fang, CVPR15]
- Word tagging & feature: ResNet [He, CVPR16]
- Deep Structured Semantic Model [He, CIKM13]

Dense Image Captioning [Johnson & Karpathy, CVPR16]



a parked motorcycle. a man on a bicycle. a man riding a bicycle. the back wheel of a bike. front wheel of a bicycle. a window on the building. a red brick building. window on the building.



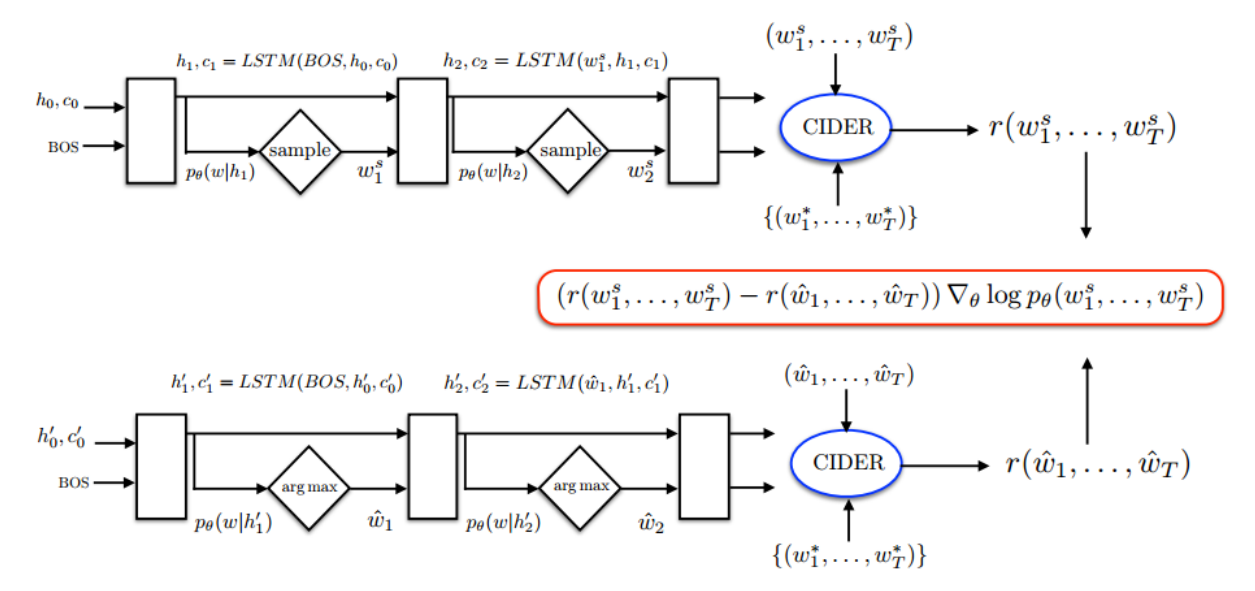
a green jacket. a white horse. a man on a horse. two people riding horses. man wearing a green jacket. the helmet is black. brown horse with white mane. white van parked on the street. a paved sidewalk. green and yellow jacket. a helmet on the head. white horse with white face.

two men playing tennis. man holding a tennis racket. tennis racket in mans hand. man with short hair. tennis racket in mans hand. man wearing a white shirt. a man with short hair. tennis racket in mans hand. a red and black bag. a tennis racket. a white tennis net. a black fence. tennis racket in mans hand. the man is wearing glasses.

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

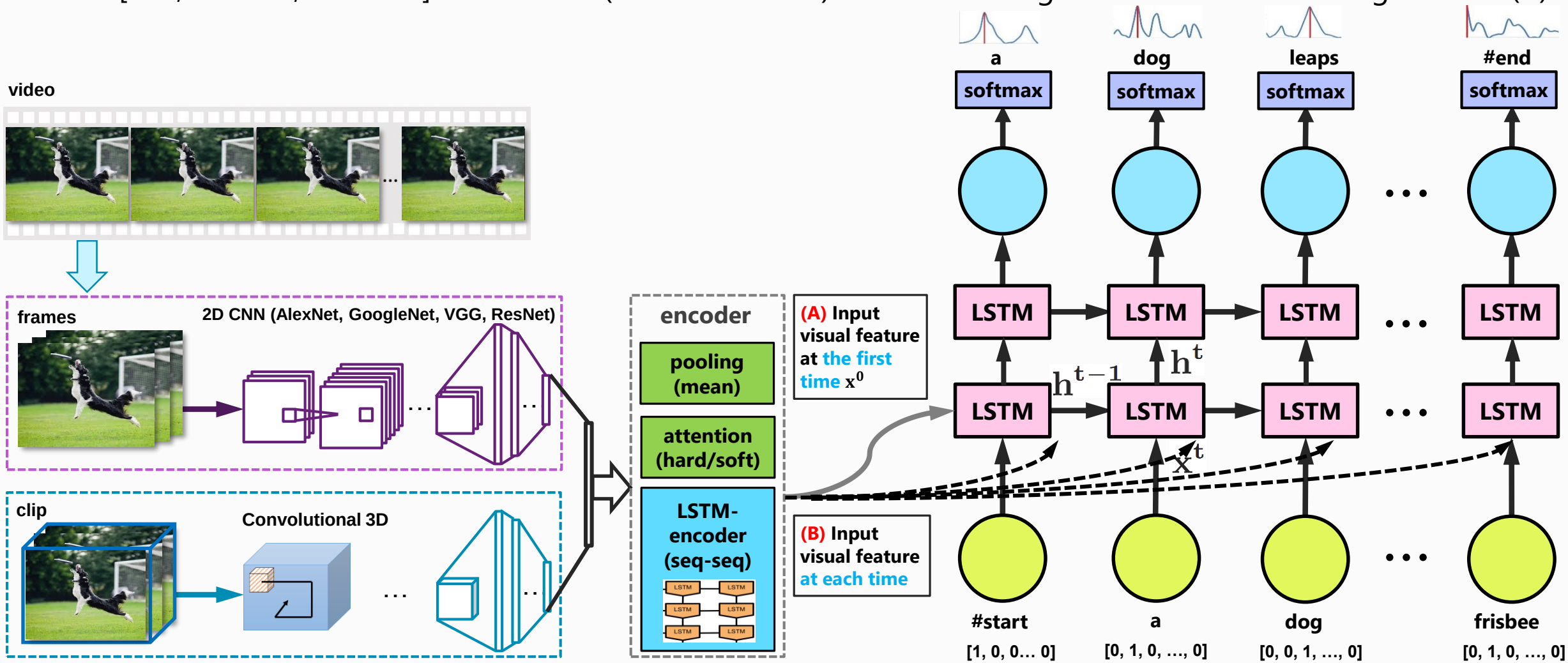
Image Captioning with Reinforcement Learning

- Self-critical sequence training [Rennie, arXiv'16]



- Policy gradient optimization of SPIDEr [Liu, arXiv'16]

- UC Berkeley [Donahue, CVPR'15]: CRF + LSTM encoder-decoder + LSTM (A/B)
- UdeM [Yao, ICCV'15]: (GoogleNet + 3D CNN) + Soft-Attention + LSTM (B)
- UT Austin [Venugopalan, ICCV'15]: (VGG + Optical Flow) + LSTM Encoder-Decoder + LSTM (A)
- UT Austin [Venugopalan, NAACL-HLT'15]: AlexNet + Mean Pooling + LSTM (B)
- MSRA [Pan, LSTM-E, CVPR'16]: (VGG + 3D CNN) + Mean Pooling + Relevance Embedding + LSTM (A)



Visual Question Answering

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

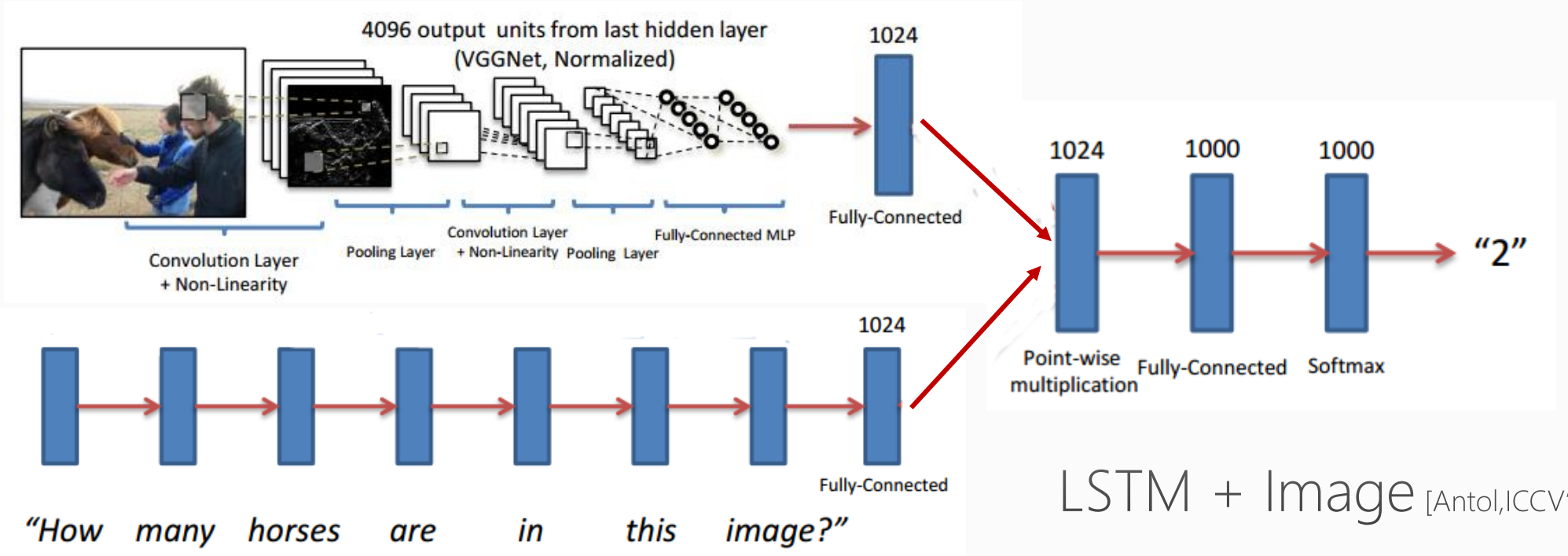
Question:
What game is being played?

Visual Question Answering (VQA)

Answer:
baseball

VQA paradigm and challenges

Model	Acc (%)
LSTM+I	53.7



LSTM + Image [Antol,ICCV'2015]

- 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

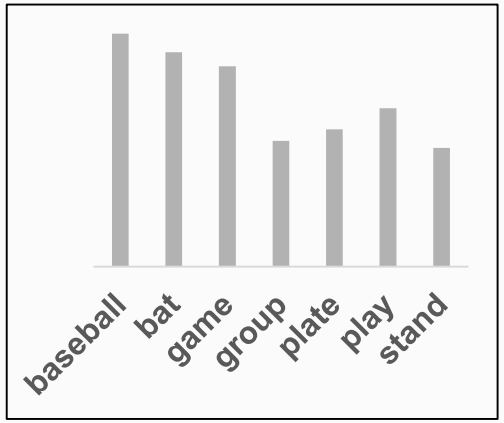
VQA with "X"

Q: what game is being played?

Q: what game is being played?

X = visual attention
[Yang, CVPR'2016; Shih, CVPR'2016]

Q: what game is being played?

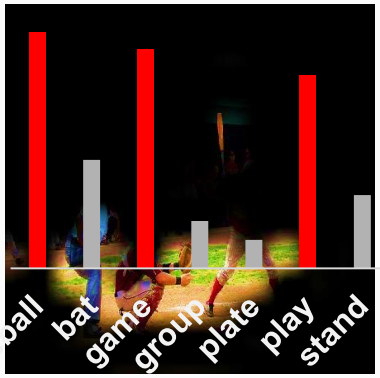


X = visual attributes
[Wu, CVPR'2016]

Q: what game is being played?

X = visual-question co-attention
[Lu, NIPS'2016]

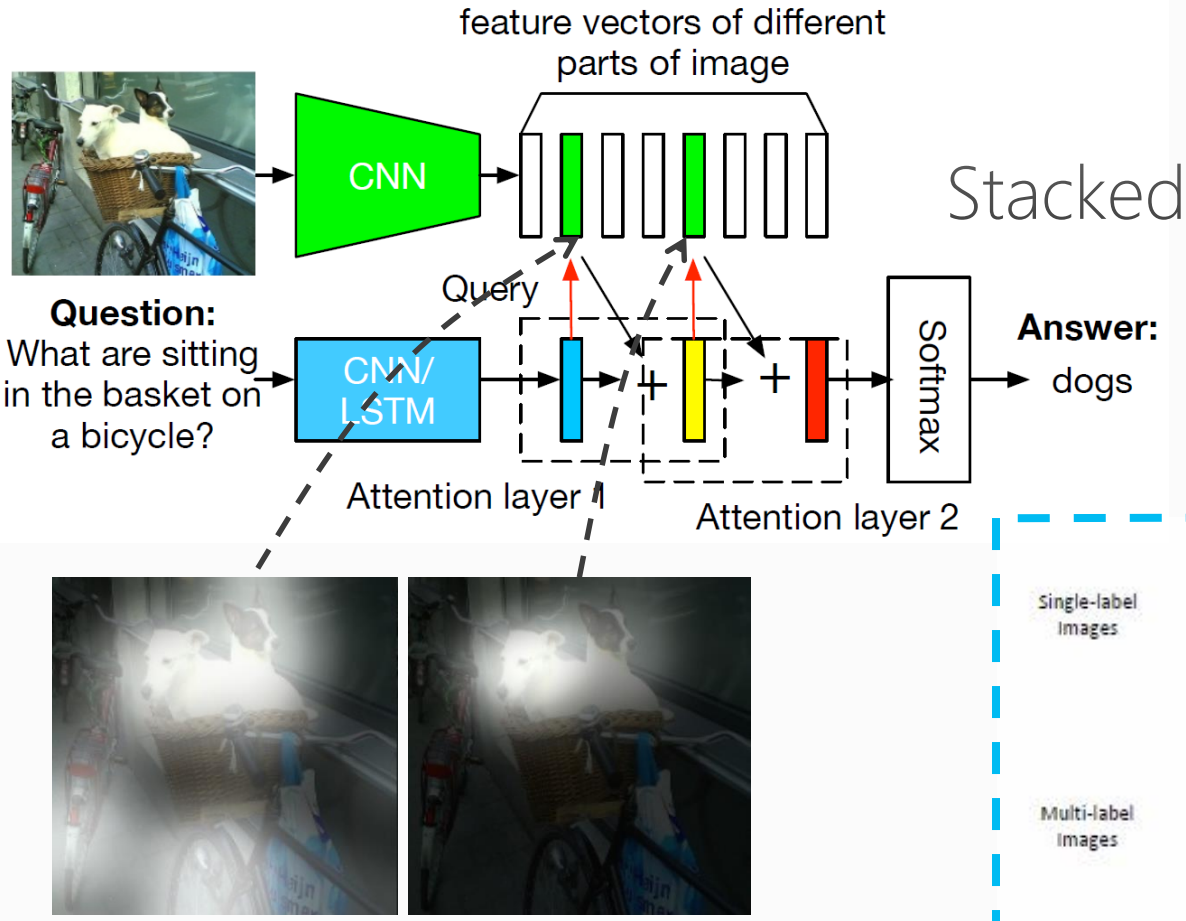
Q: what game is being played?



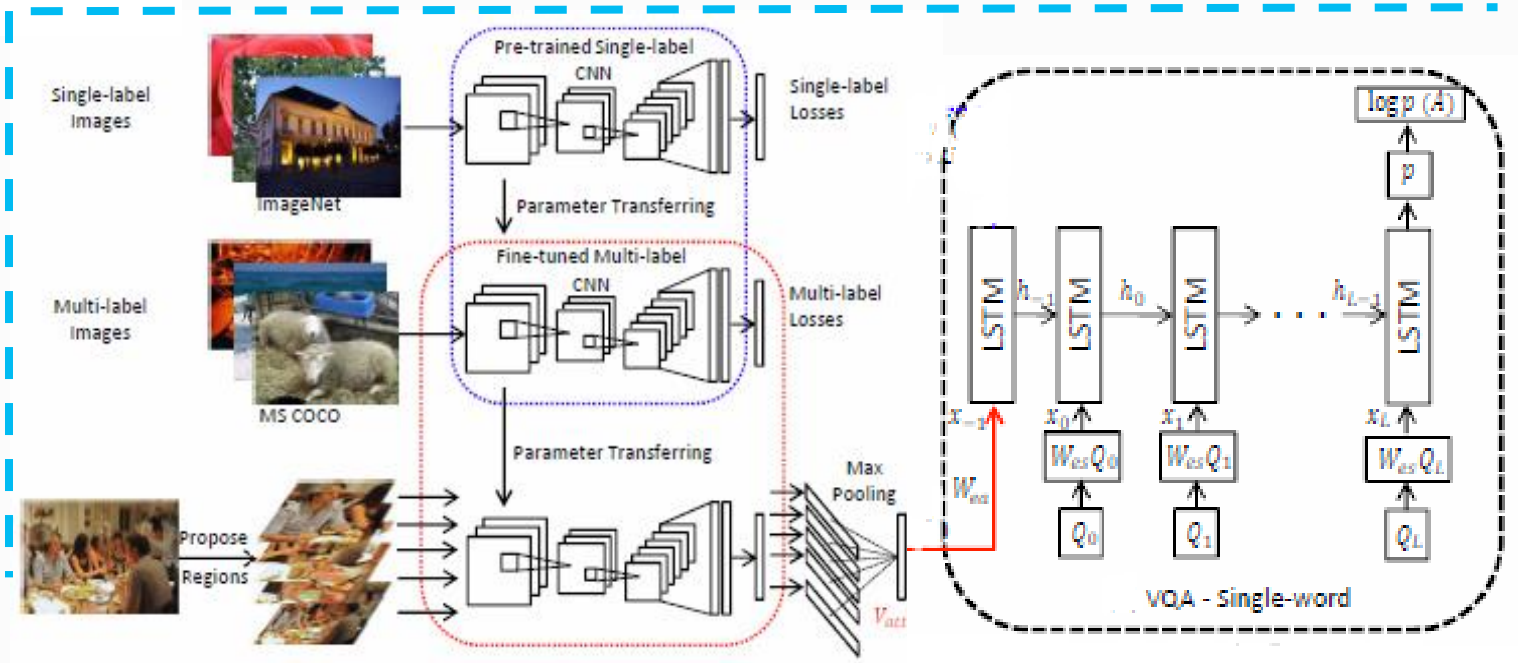
X = multi-level attention
[Yu, CVPR'2017]

Visual attention and attributes

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



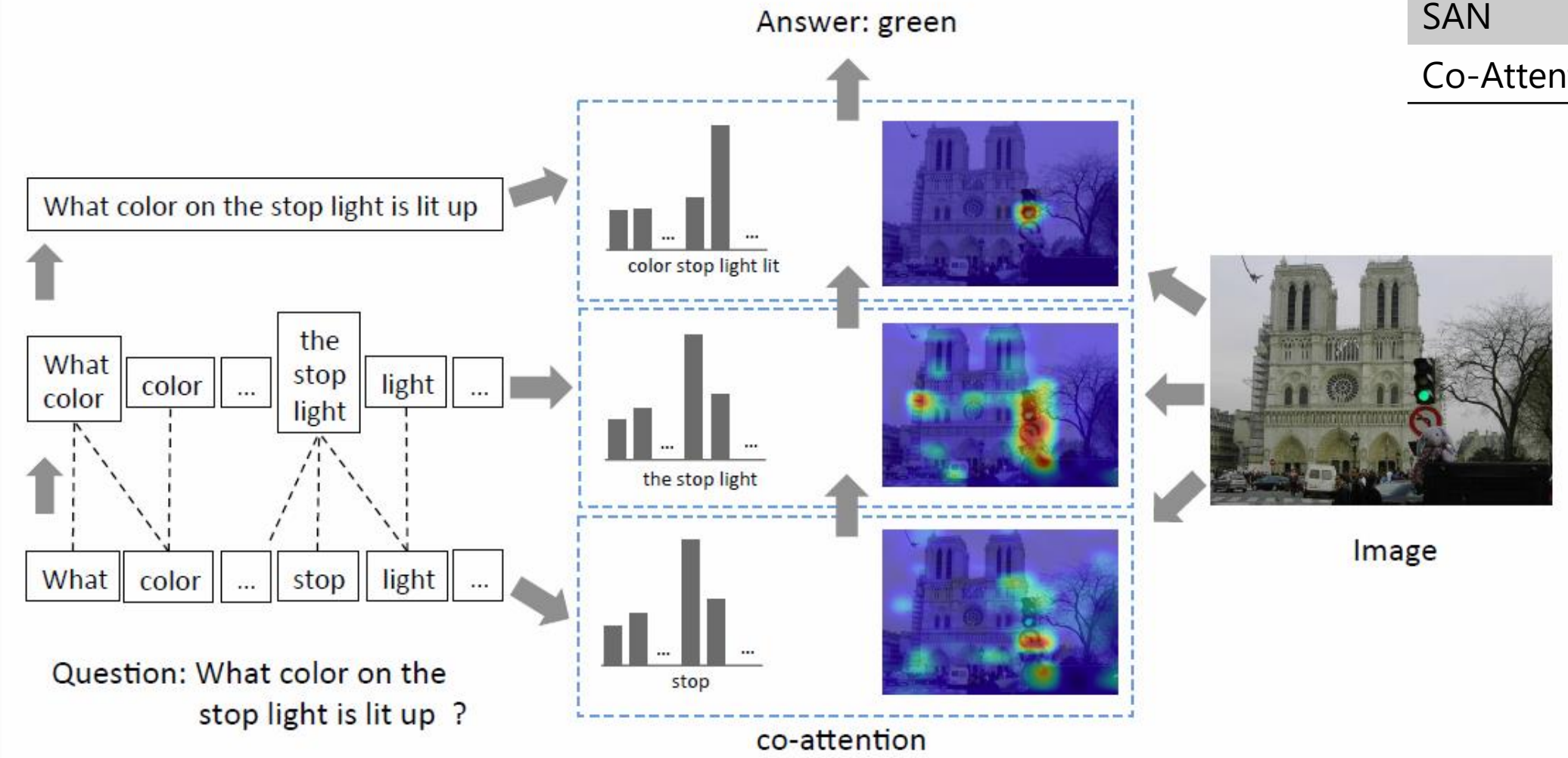
Stacked Attention Networks [Yang,CVPR'2016]



Att-KB LSTM [Wu,CVPR'2016]

Visual-question co-attention

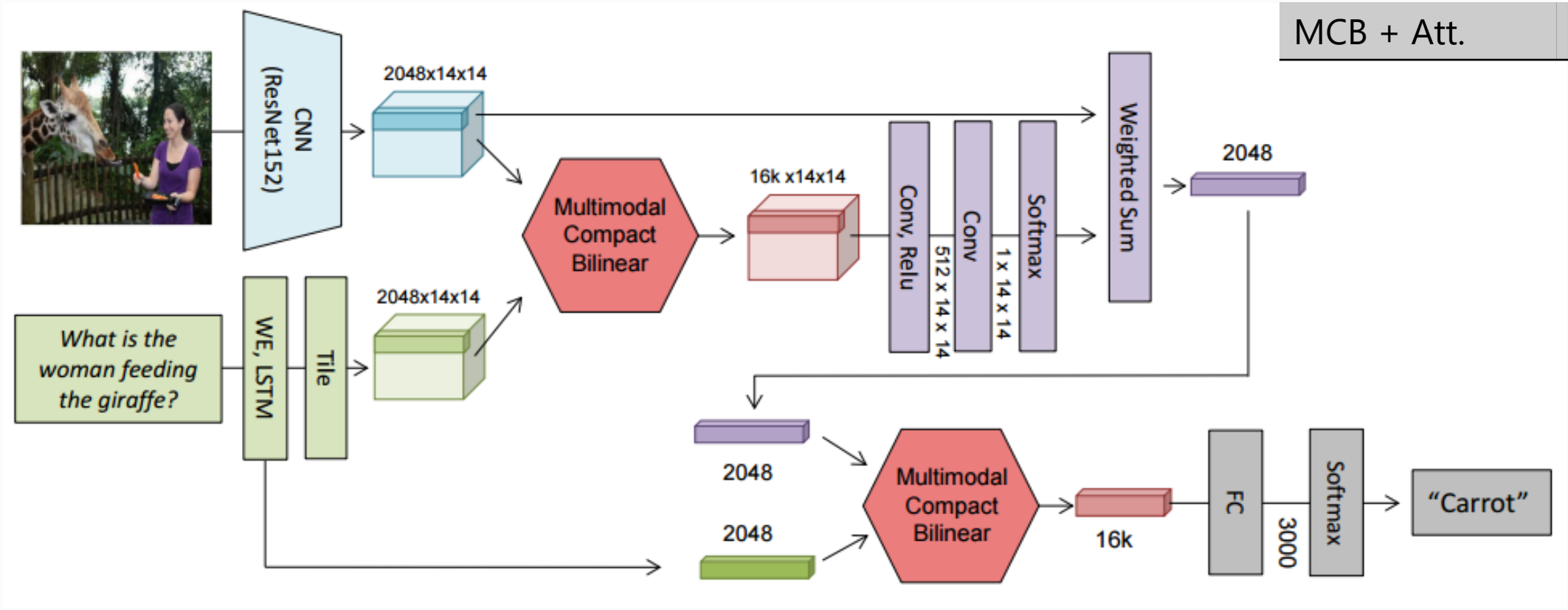
Model	Acc (%)
LSTM+I	53.7
Att-KB+LSTM	57.5
SAN	58.7
Co-Attention	61.8



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

Multi-modality Bilinear Fusion

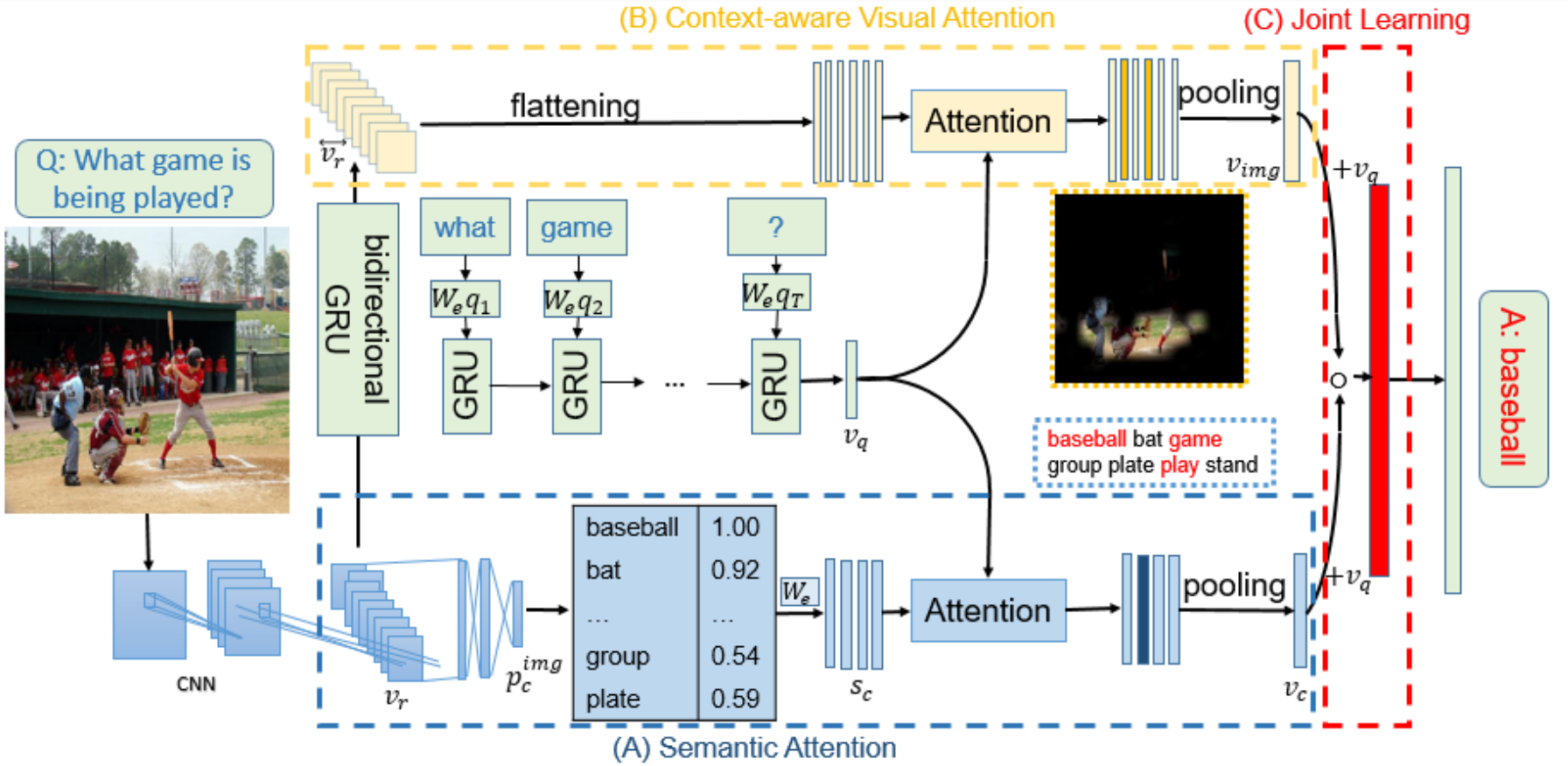
Model	Acc (%)
LSTM+I	53.7
Att-KB+LSTM	57.5
SAN	58.7
Co-Attention	61.8
MCB + Att.	64.2



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

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

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**



Multi-level Attention [Yu, Fu, and Mei, CVPR'2017]

Thanks!

tmei@microsoft.com