國立清華大學
NATIONAL TSING HUA UNIVERSITY

VSLab

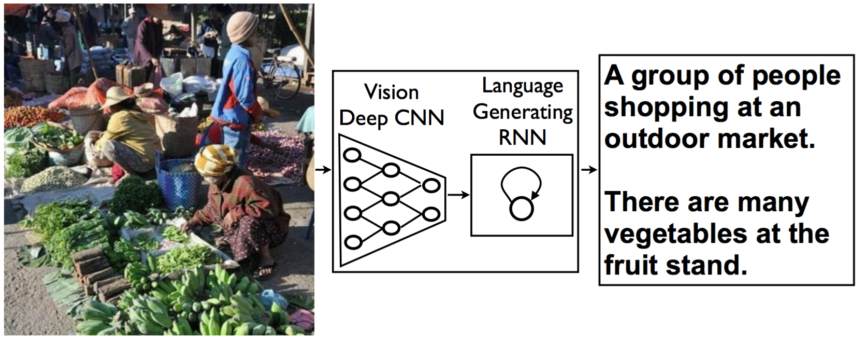
Vision Guided Language Generation

Min Sun

National Tsing Hua University

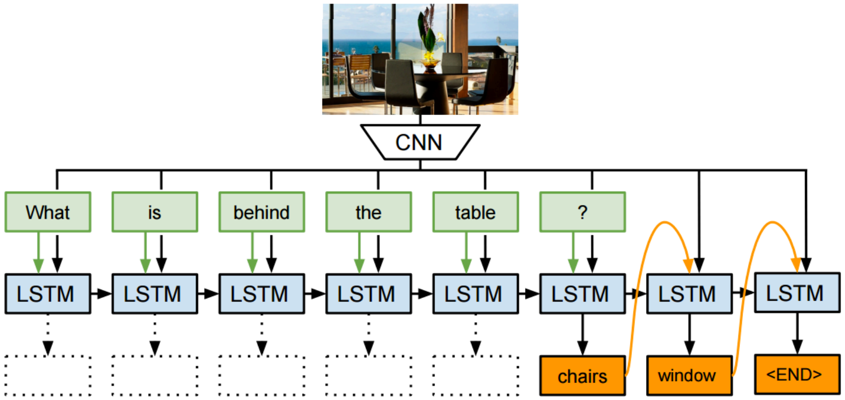
Overview

- Captioning
  - For Image
  - Referring Expression
  - For Video
- Question Answering
  - For Image
  - For Video
- Others
  - Storytelling
  - Visual-aware HCI



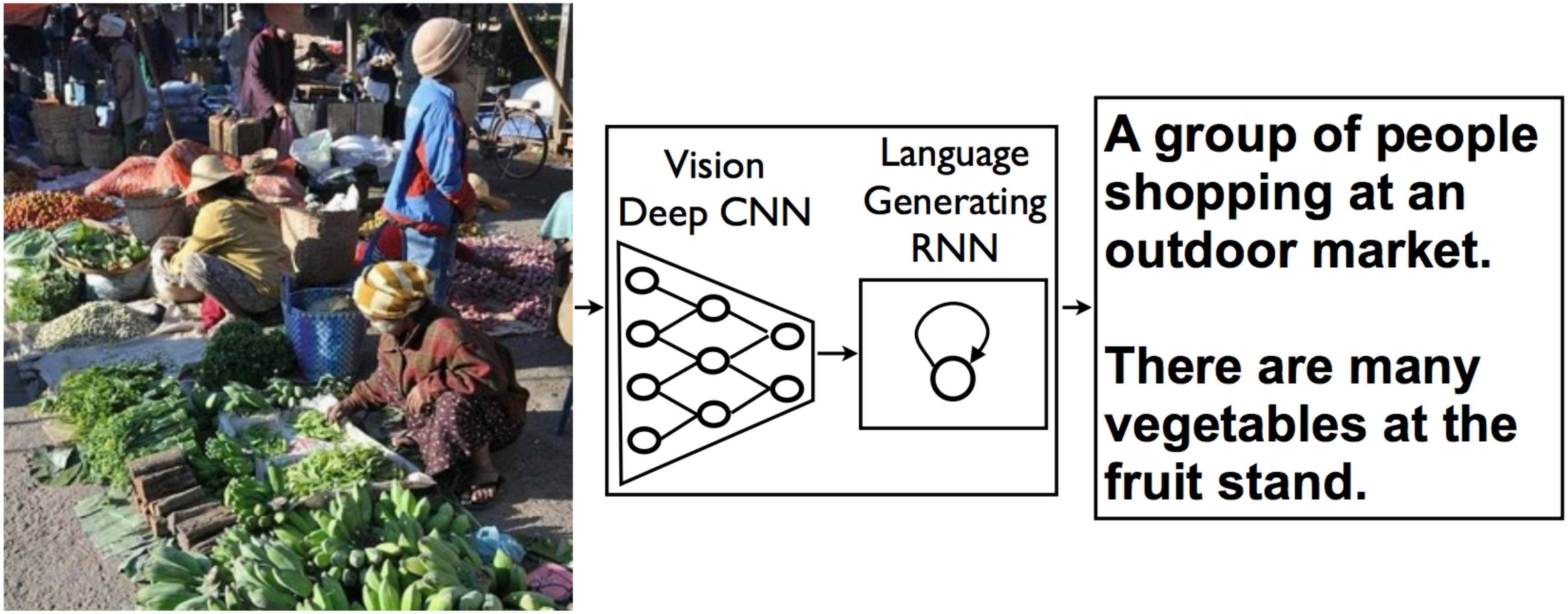
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.



Captioning

- I have a **CNN**, I have a **RNN** -> **Novel Sentences**



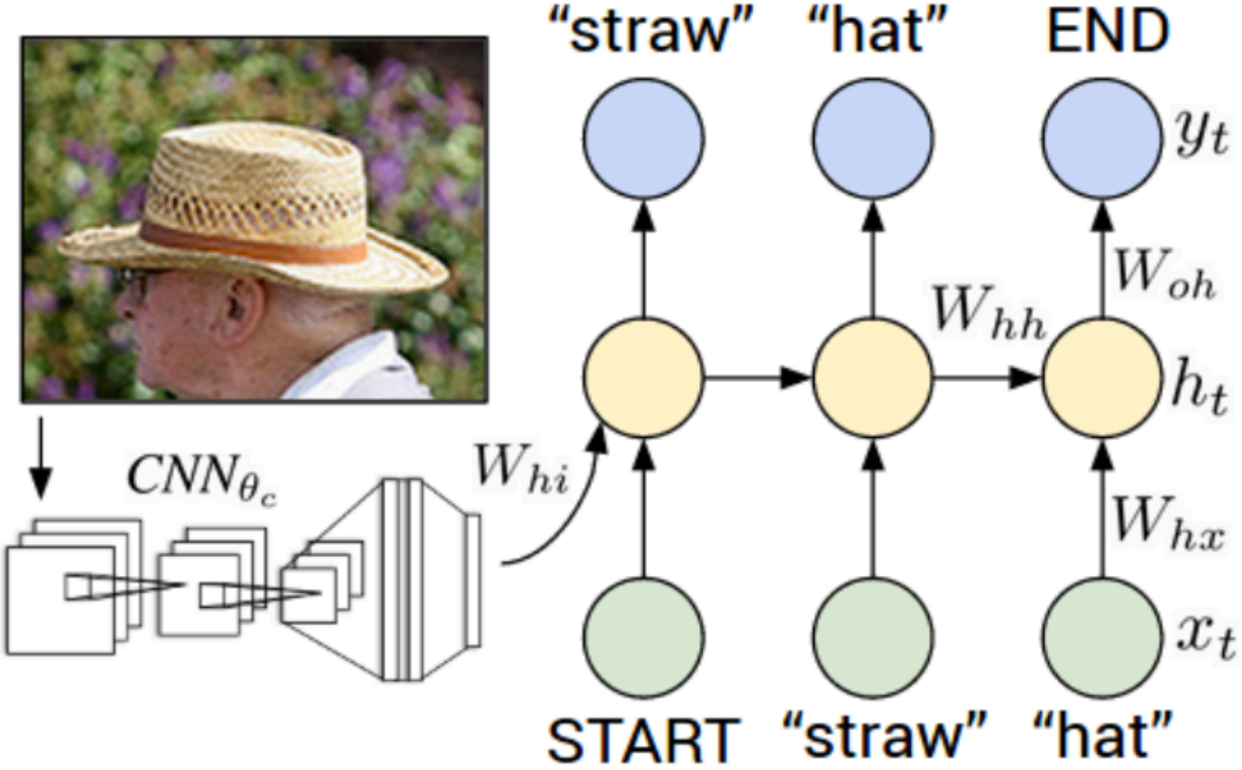
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. CVPR 2015

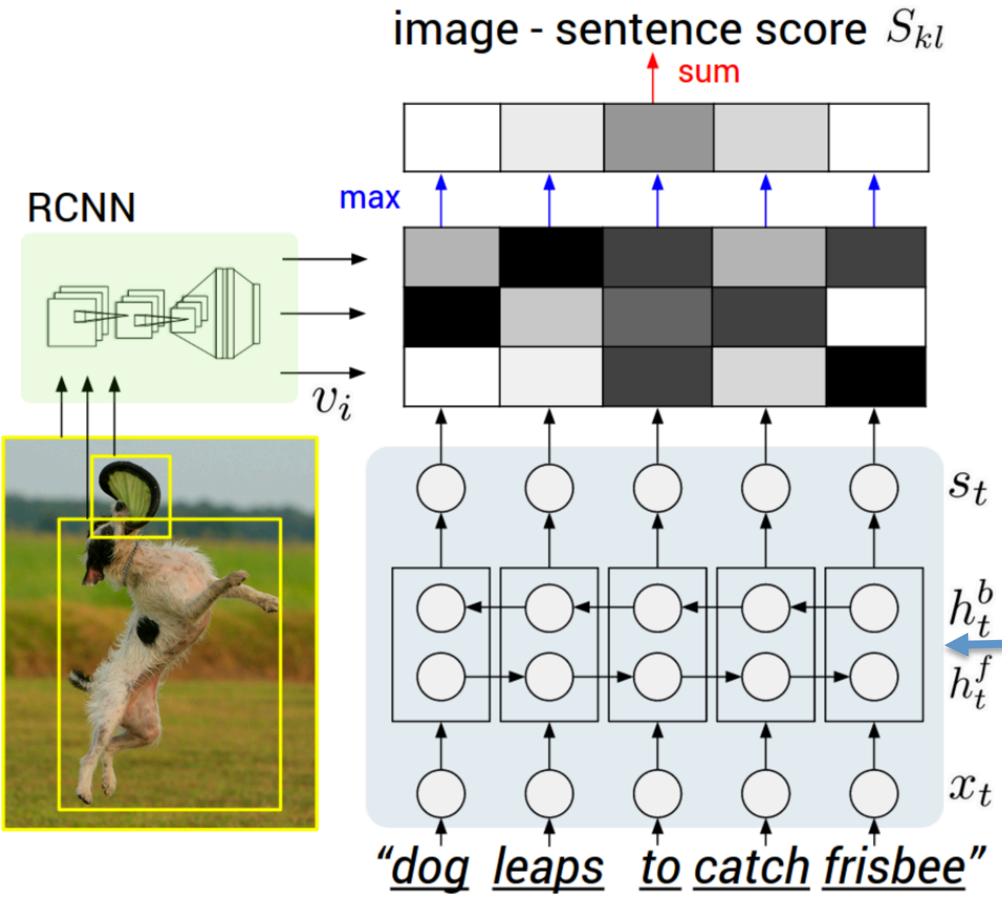
Google

Captioning



Andrej Karpathy, Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. CVPR 2015

Captioning

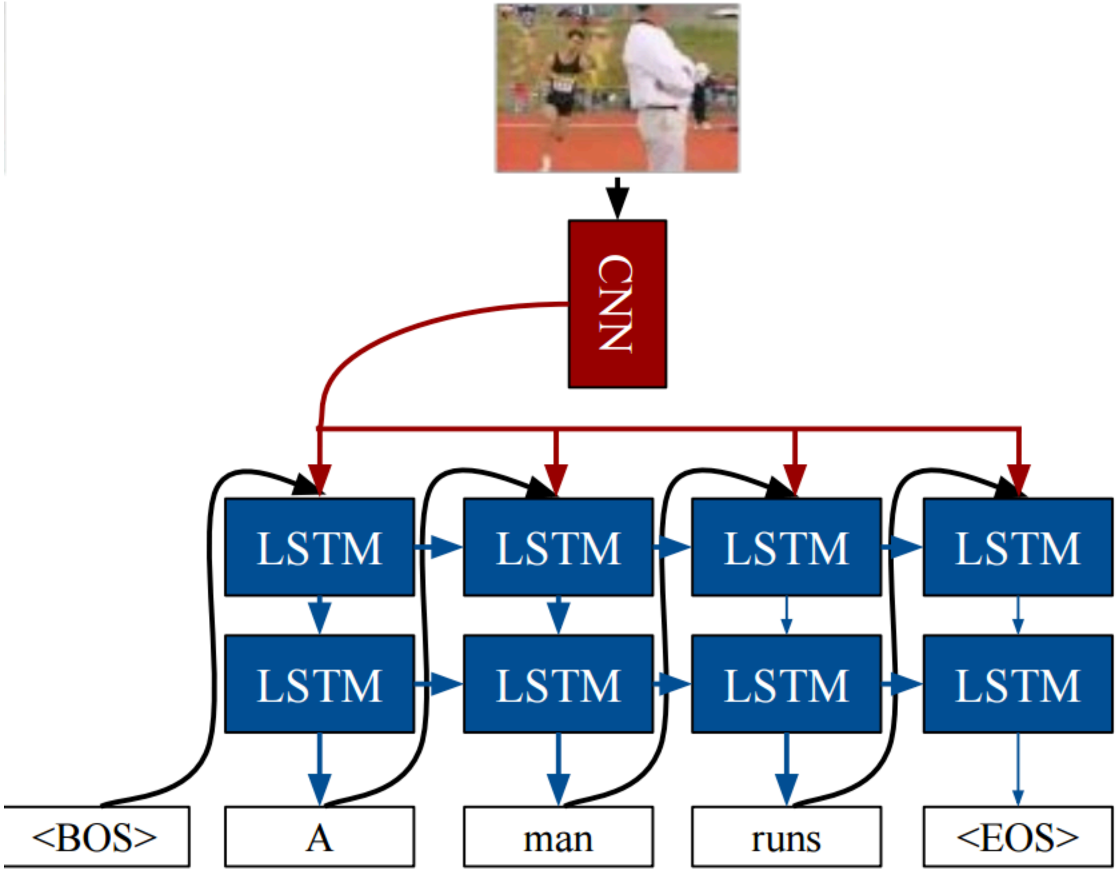


Bi-directional RNN

Andrej Karpathy, Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. CVPR 2015

Captioning

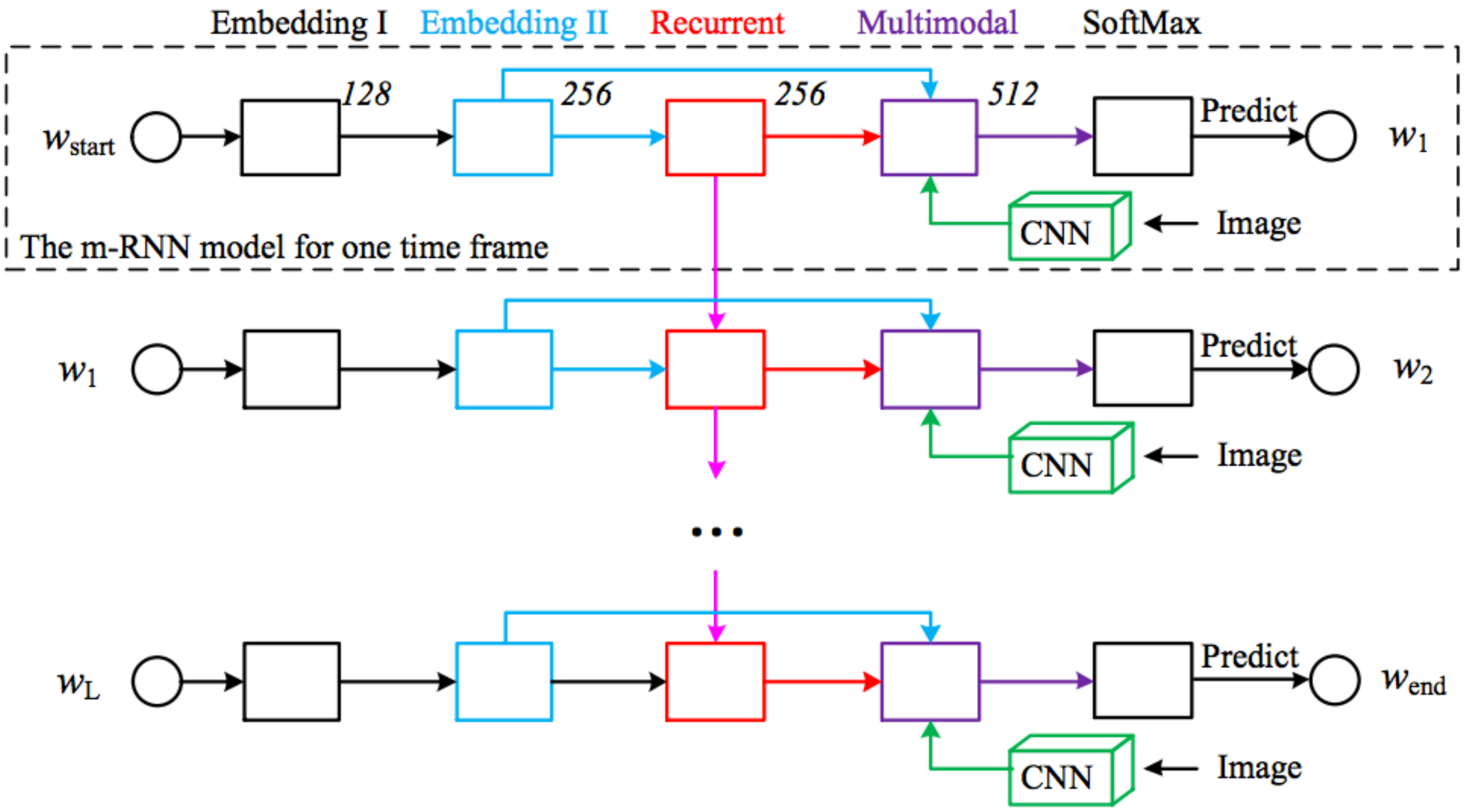
Sequences in the Output



Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

Captioning

Two-layer word embedding system



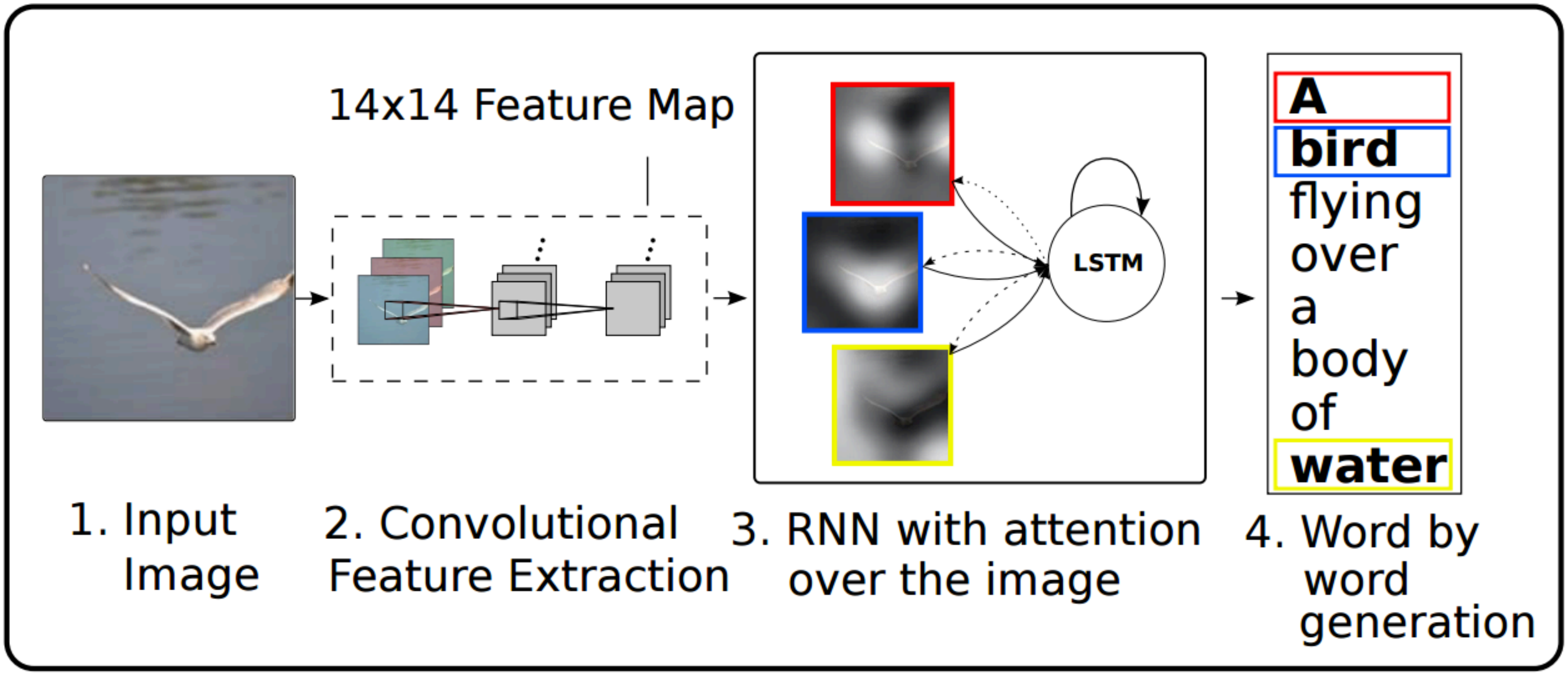
(b). The m-RNN model

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). ICLR 2015

baidu

Captioning - *Attention*

- **Attention** mechanism: per word attention.

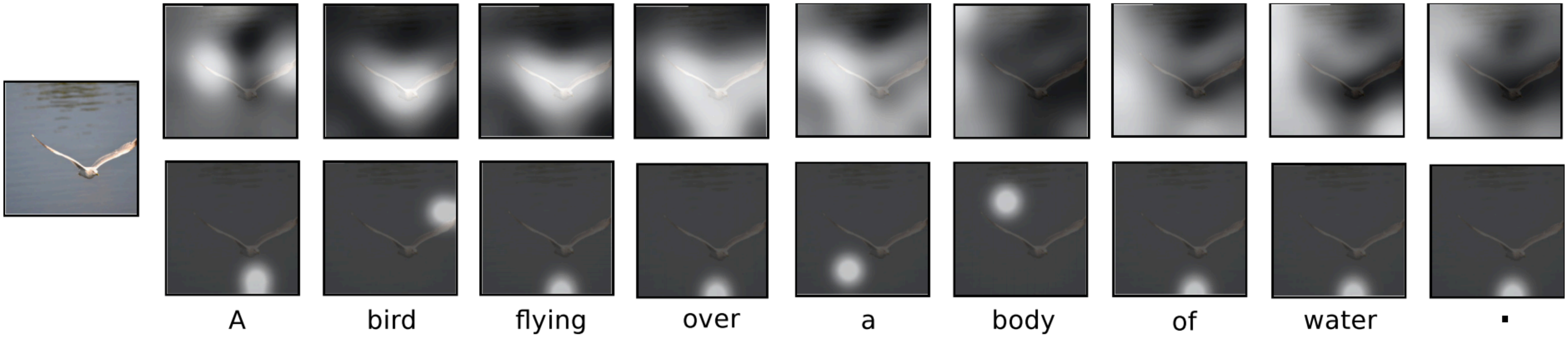

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, **Attend** and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Captioning - *Attention*

- Attention mechanism: per word attention.

Soft-attention, determinisitc, Backpropagation

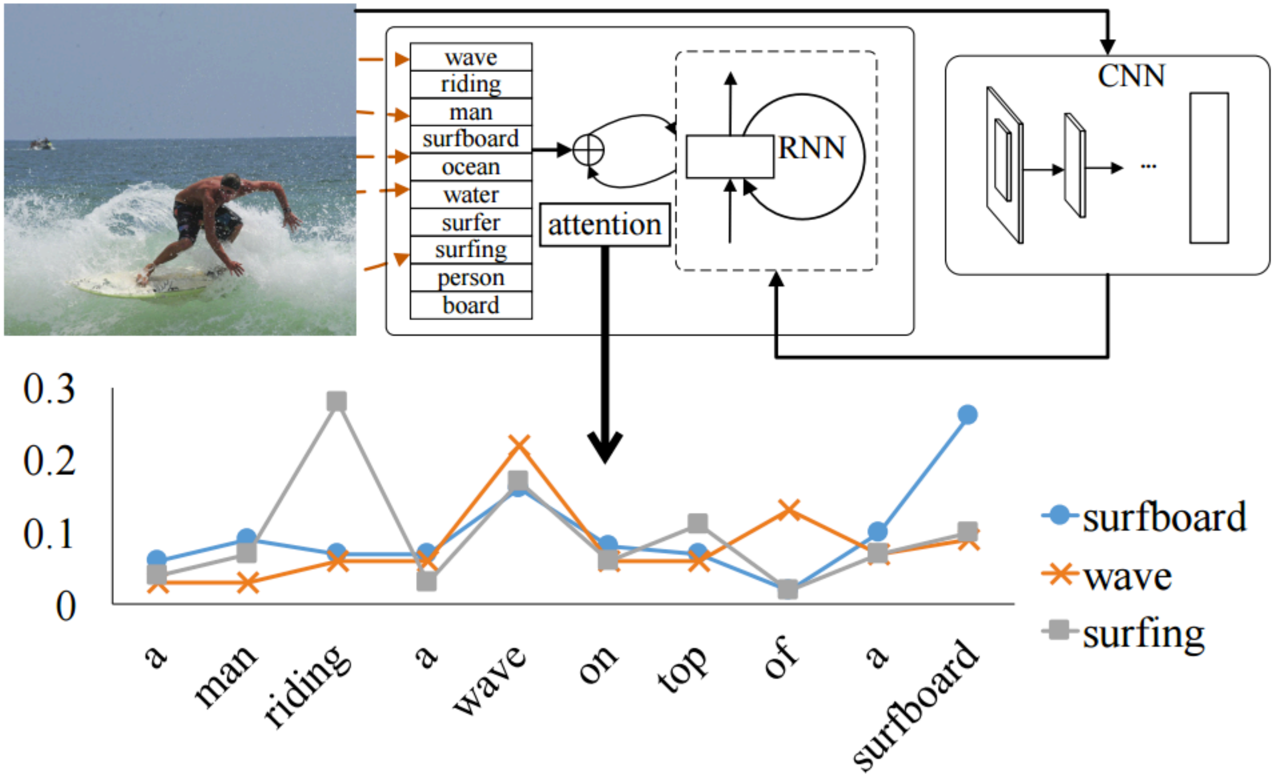


Hard-attention, stochastic, lower-bound or RL

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Captioning - *Attributes*

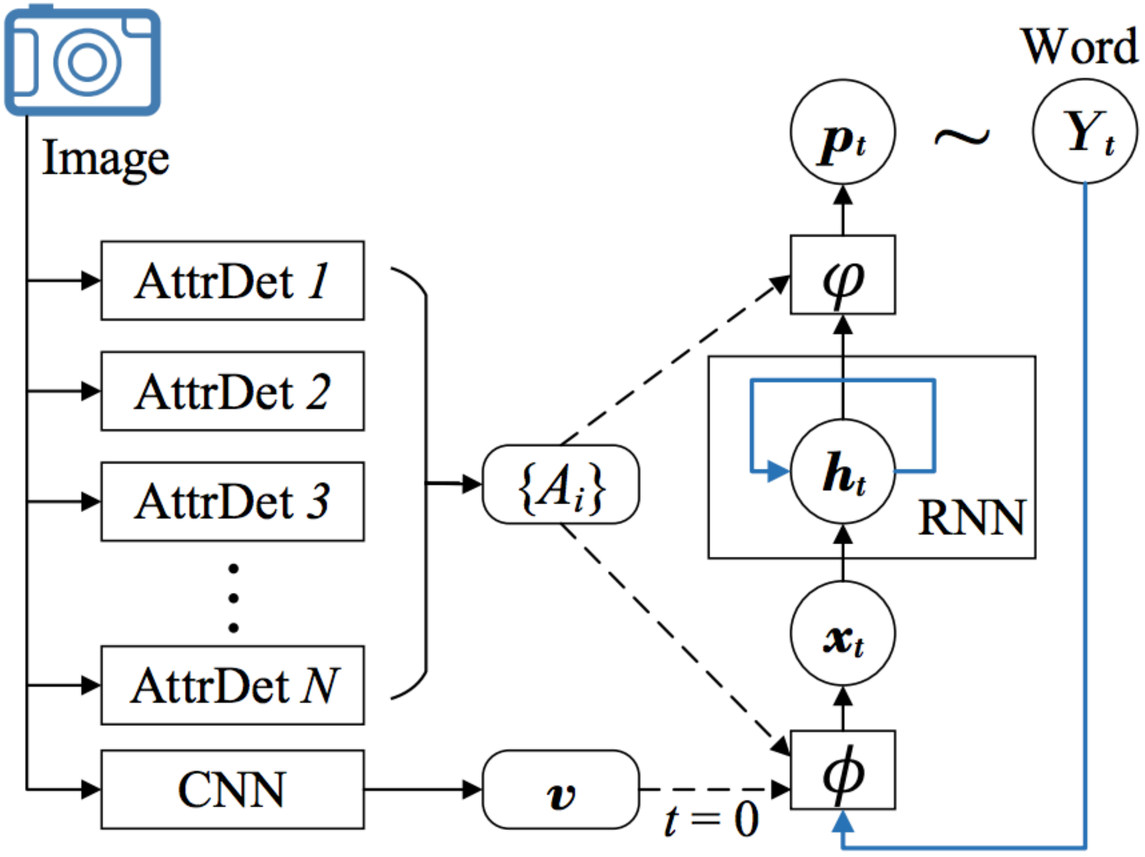
- **Semantic** Attention (e.g., surfboard, etc.)



Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo. Image Captioning With Semantic Attention. CVPR 2016

Captioning - *Attributes*

- **Semantic** Attention (e.g., surfboard, etc.)



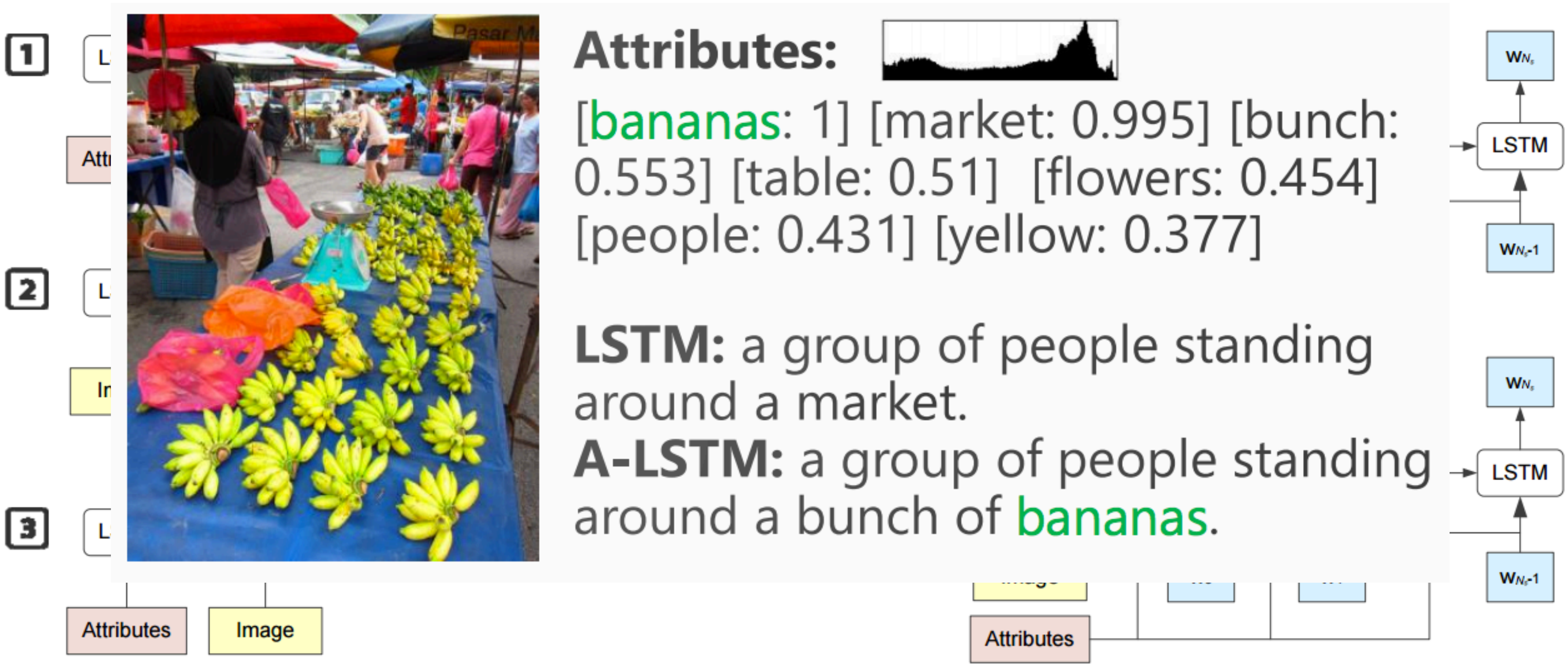
Bottom-up

Top-down

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo. Image Captioning With Semantic Attention. CVPR 2016

Captioning - *Attributes*

- **Great performance on COCO captioning**



Attributes:

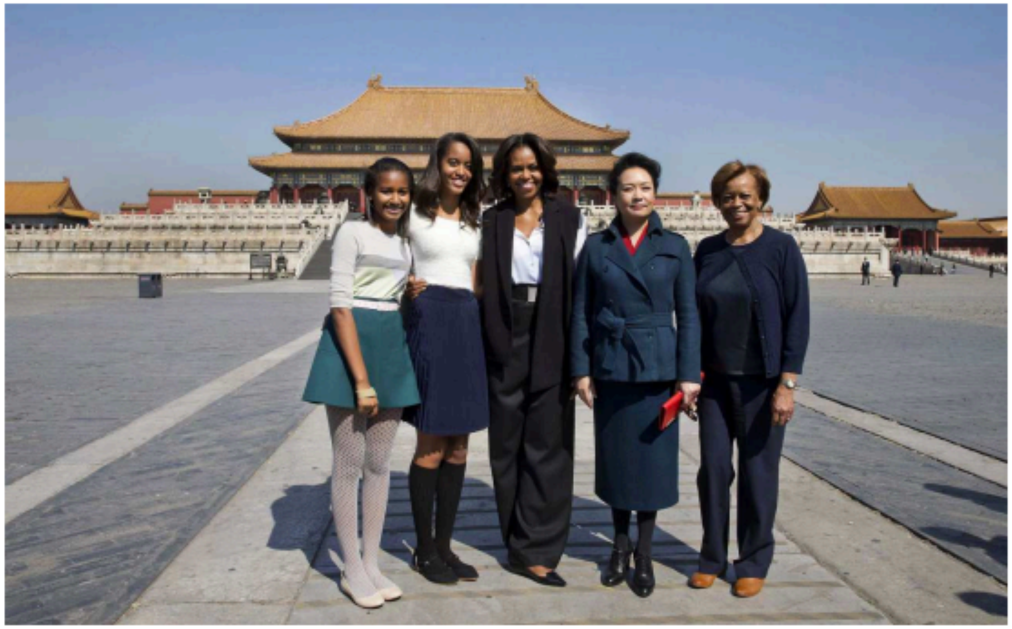
[bananas: 1] [market: 0.995] [bunch: 0.553] [table: 0.51] [flowers: 0.454] [people: 0.431] [yellow: 0.377]

LSTM: a group of people standing around a market.
A-LSTM: a group of people standing around a bunch of bananas.

BOOSTING IMAGE CAPTIONING WITH ATTRIBUTES
Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, Tao Mei. ICLR 2017 under review

Captioning – *Rich Caption*

- **Landmark and Celebrity**

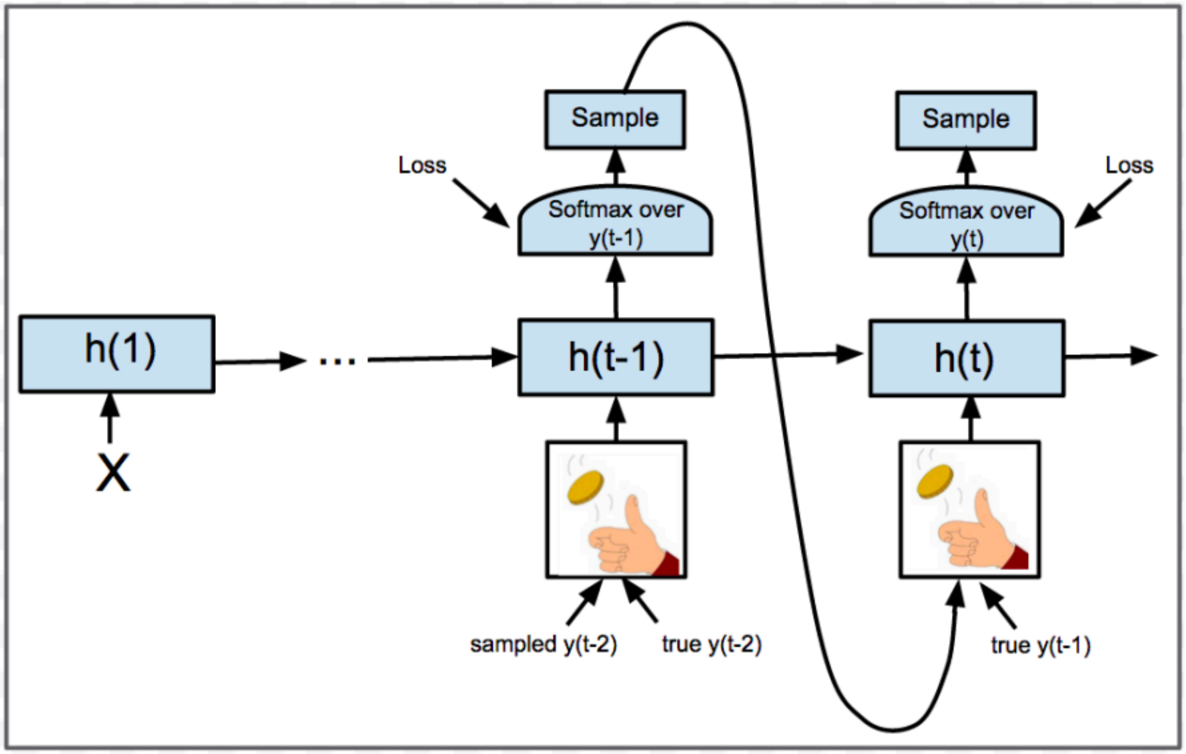


"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background."

Kenneth Tran et al., Rich Image Captioning in the Wild. CVPR 2016

Captioning – *Curriculum Learning*

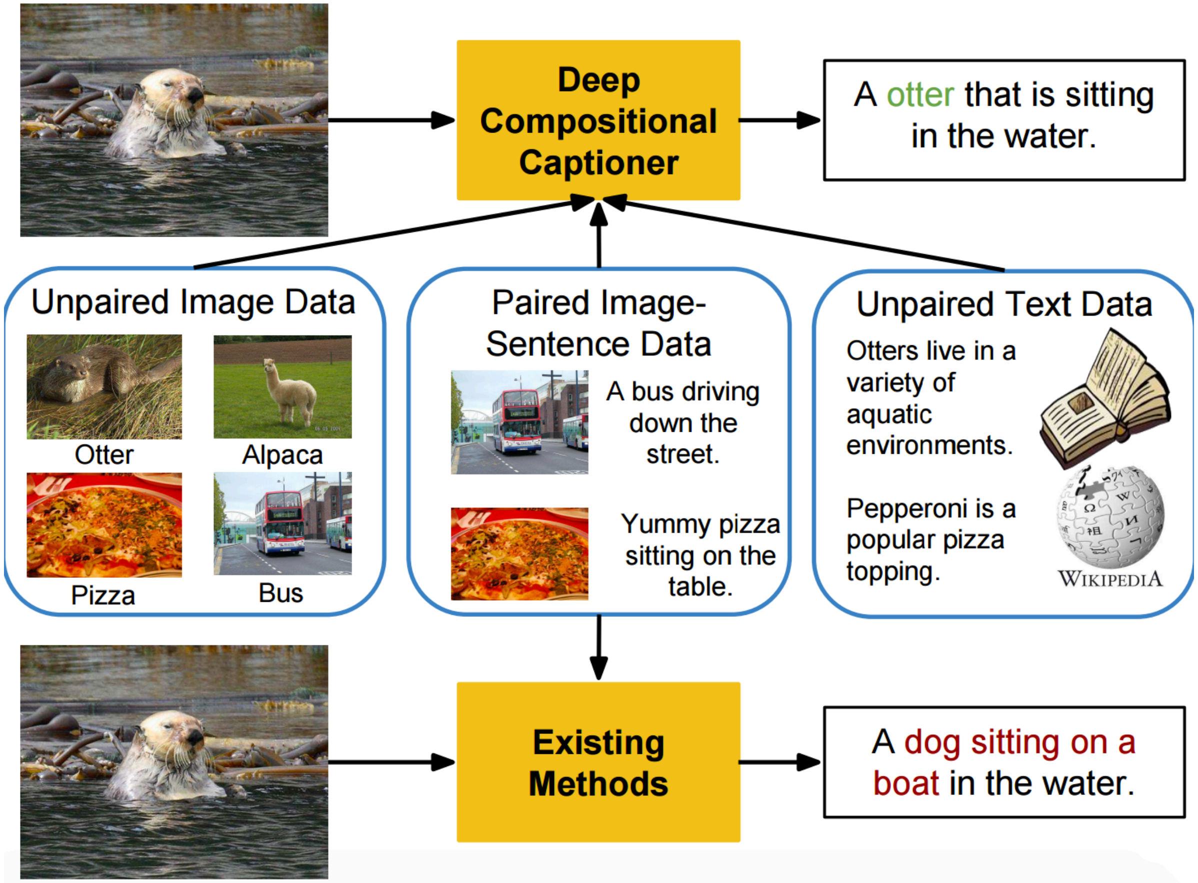
- **gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead**



Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, NIPS 2015

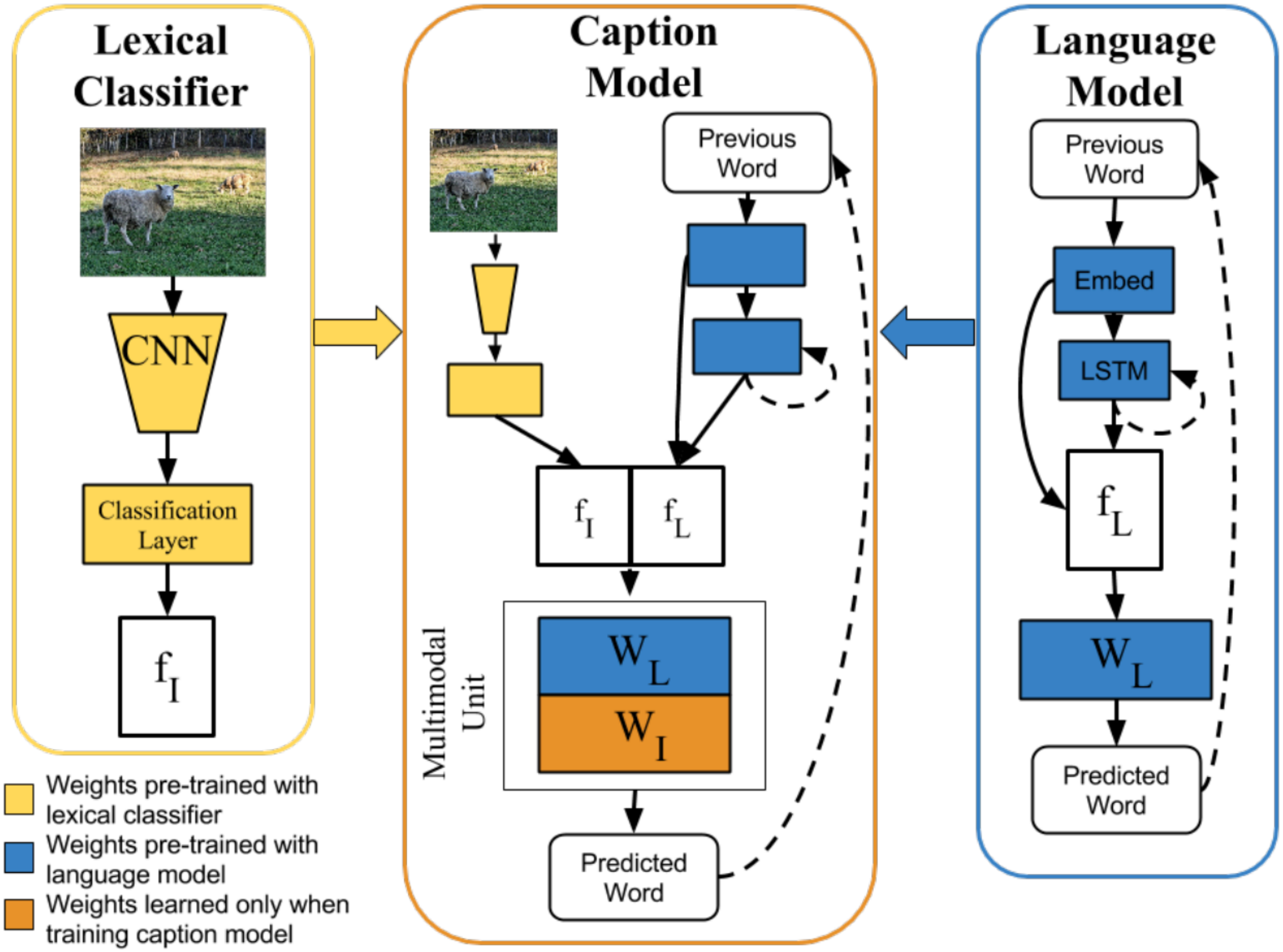
Captioning – *Unpaired Data*

- generating descriptions of novel objects which are not present in paired image-sentence datasets


Deep Compositional Captioner
A otter that is sitting in the water.
Unpaired Image Data
Otter
Alpaca
Pizza
Bus
Paired Image-Sentence Data
A bus driving down the street.
Yummy pizza sitting on the table.
Unpaired Text Data
Otters live in a variety of aquatic environments.
Pepperoni is a popular pizza topping.
WIKIPEDIA
Existing Methods
A dog sitting on a boat in the water.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell. Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data. CVPR 2016

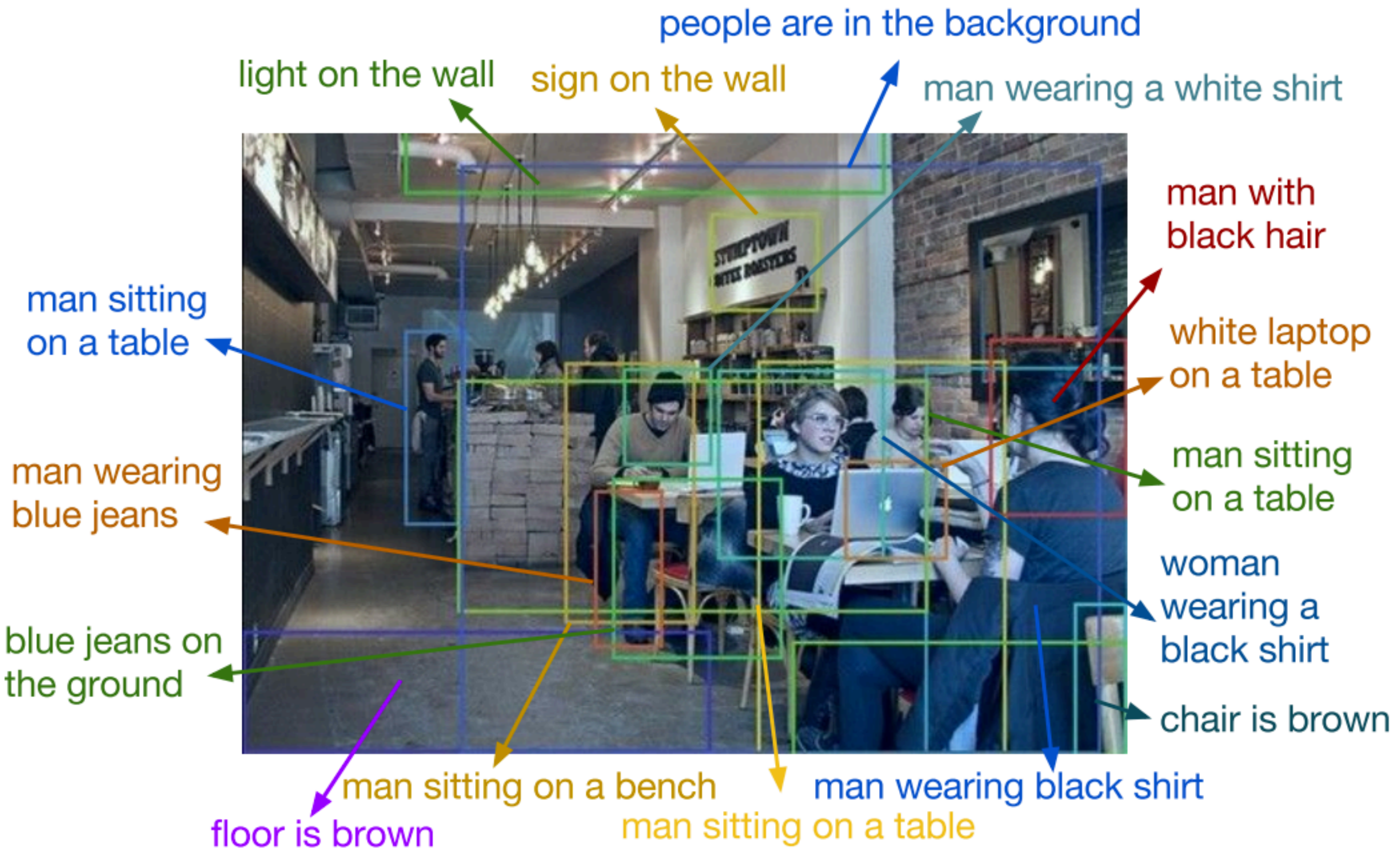
Captioning – *Unpaired Data*



Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell. Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data. CVPR 2016

Captioning – *DenseCap*

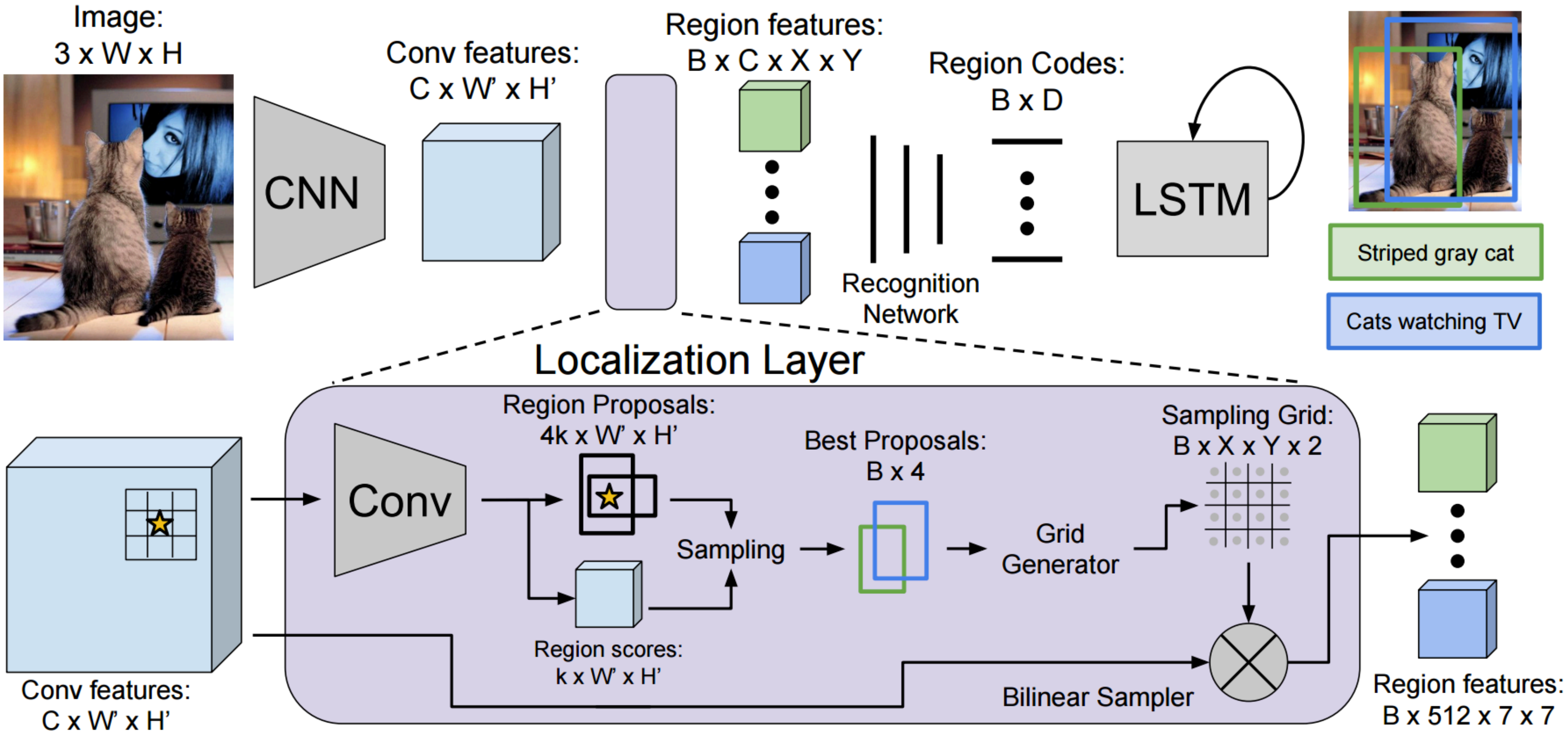
- Both localize and describe salient regions in images in natural language.



Justin Johnson, Andrej Karpathy, Li Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning, CVPR 2016

Captioning – *DenseCap*

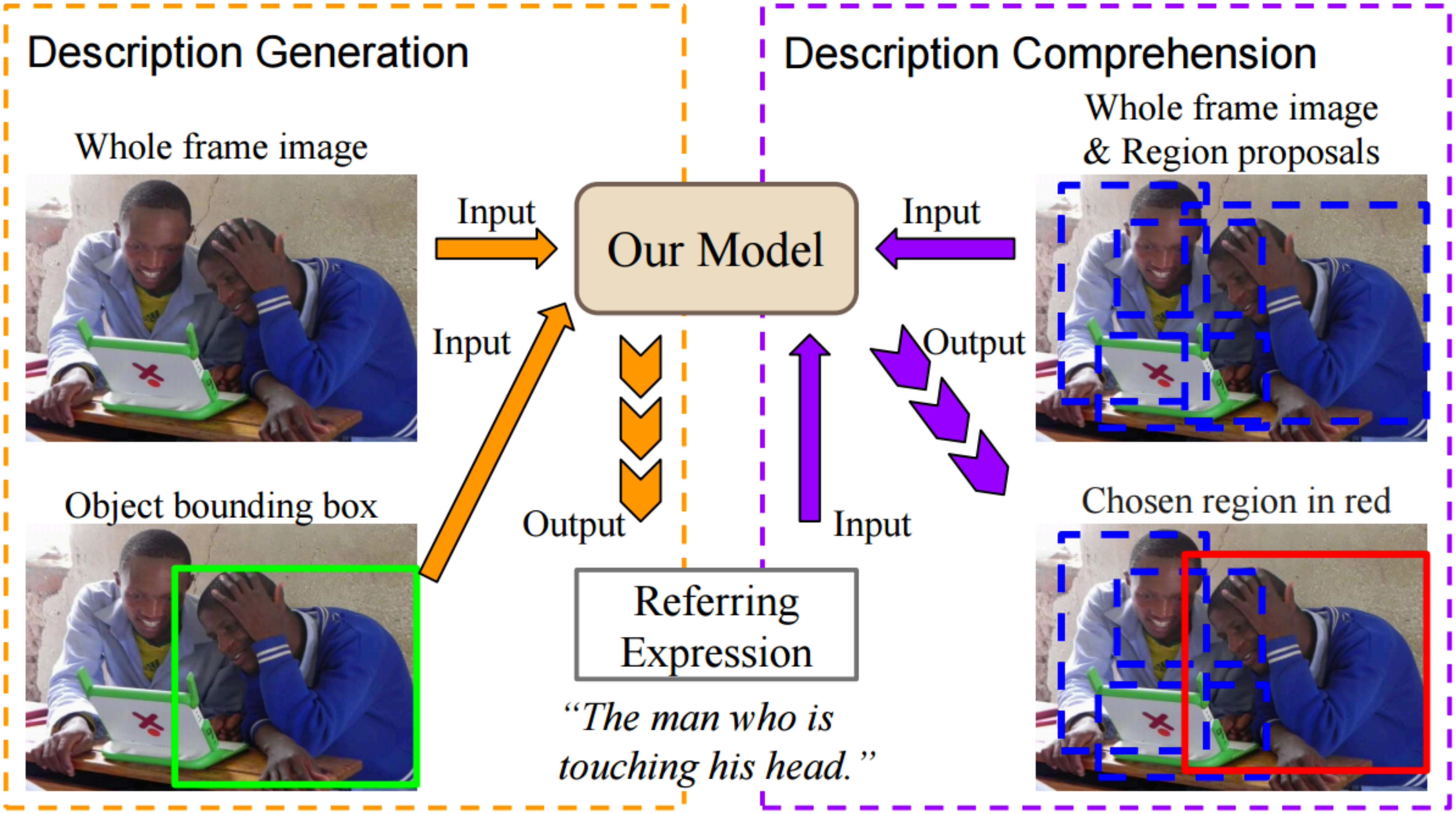
- Caption and localization layers end-to-end trainable



Justin Johnson, Andrej Karpathy, Li Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning, CVPR 2016

Captioning – *Object Descriptions*

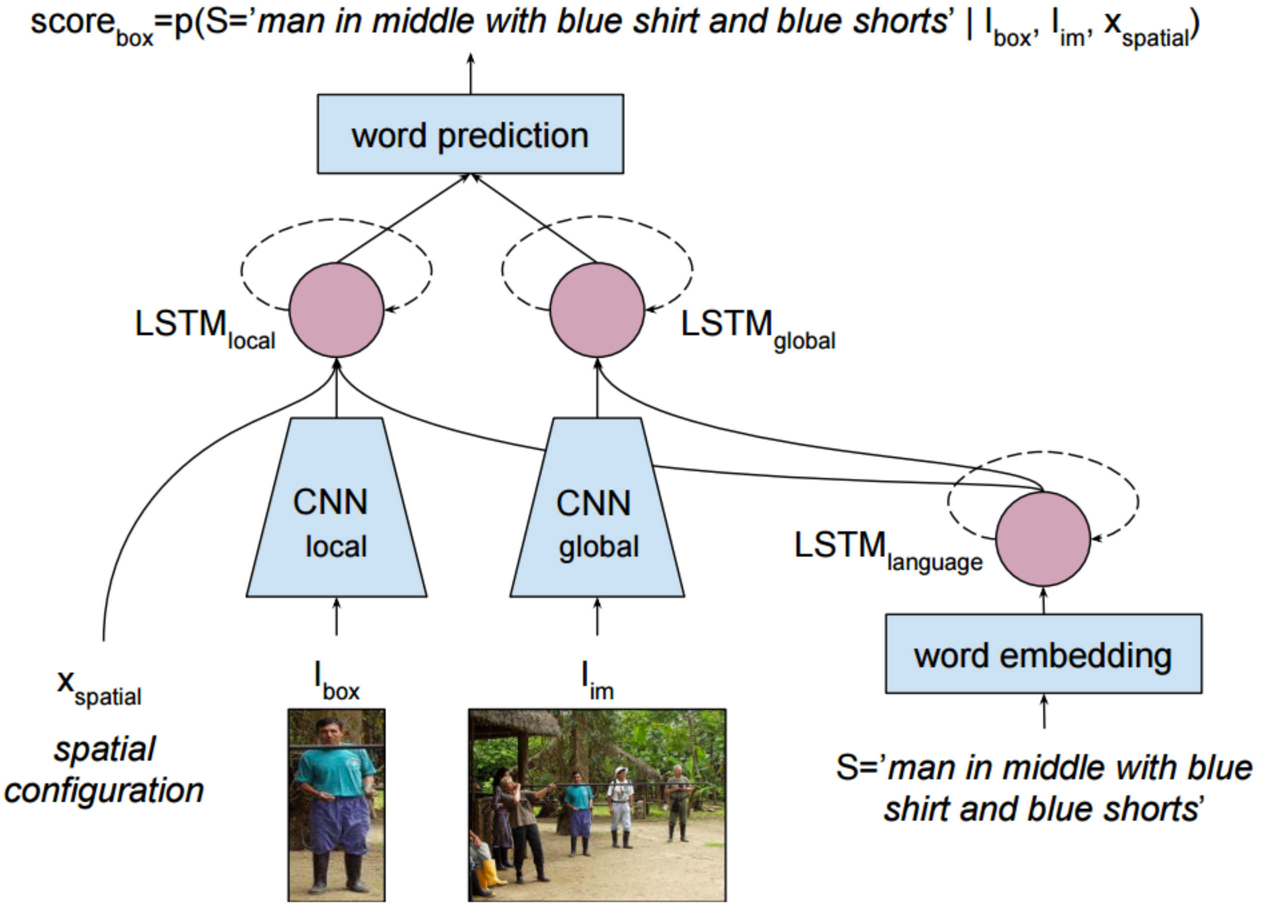
- Generation and Comprehension of Unambiguous OD



Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. CVPR 2016

Captioning – *Object Retrieval*

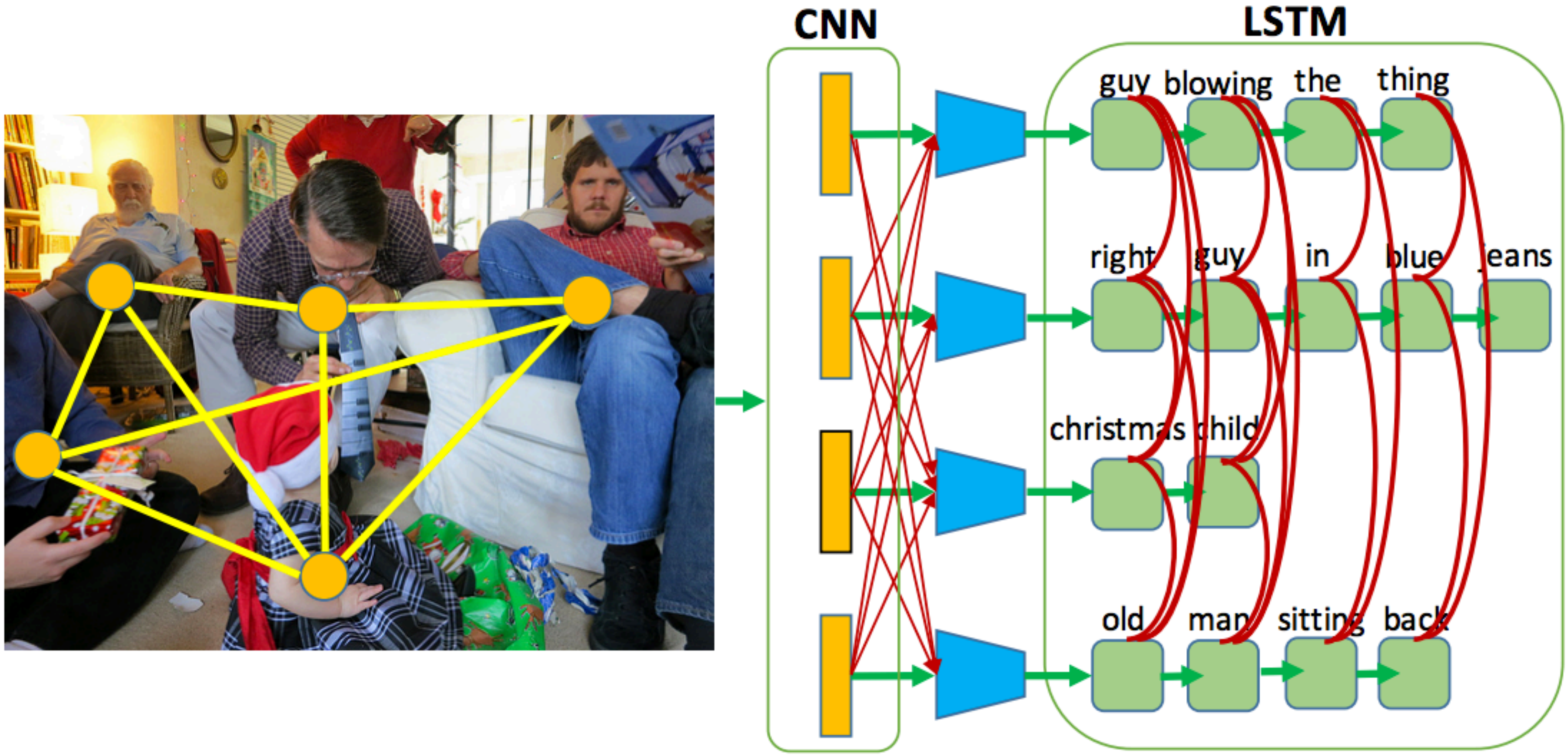
- using Natural Language


score$_{box}$=p(S='*man in middle with blue shirt and blue shorts*' | I$_{box}$, I$_{im}$, x$_{spatial}$)
word prediction
LSTM$_{local}$
LSTM$_{global}$
CNN local
CNN global
LSTM$_{language}$
word embedding
x$_{spatial}$
spatial configuration
I$_{box}$
I$_{im}$
S='*man in middle with blue shirt and blue shorts*'

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, Trevor Darrell. Natural Language Object Retrieval CVPR 2016

Captioning – *Referring Expression*

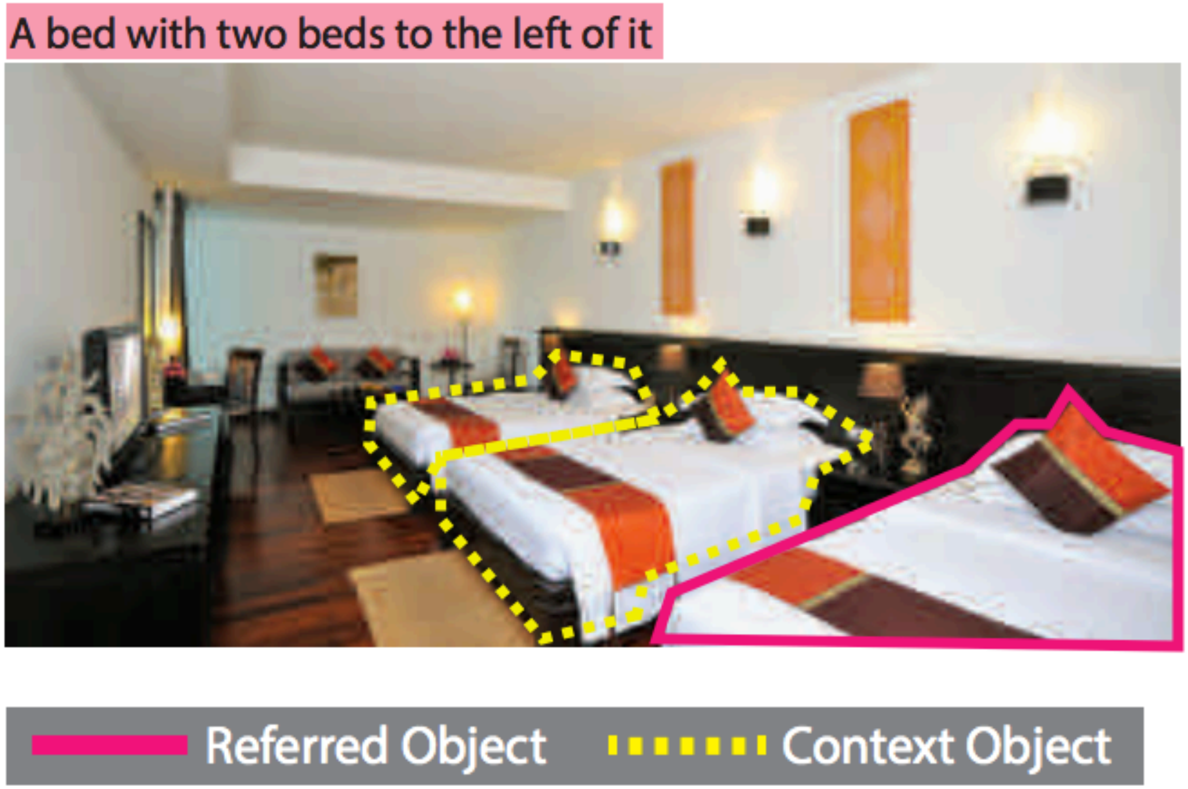
- Joint inference among all objects



Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg, Modeling Context in Referring Expressions, ECCV 2016

Captioning – *Referring Expression*

- Context Between Objects for Referring Expression



Varun K. Nagaraja Vlad I. Morariu Larry S. Davis, Modeling Context Between Objects for Referring Expression Understanding, ECCV 2016

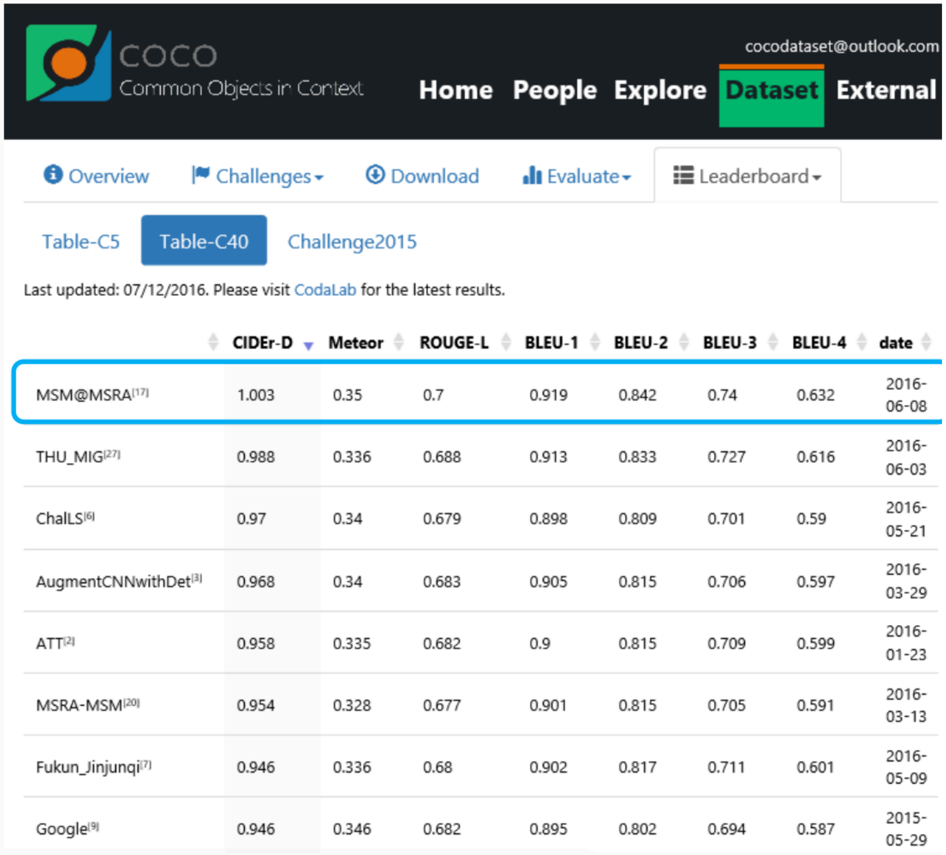
Captioning – Dataset

- Peter Young, Alice La,i Micah Hodosh, Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. (**Flickr30K**) TCAL 2014
- Tsung-Yi Lin et al., **Microsoft COCO**: Common Objects in Context. ECCV 2014
- Ranjay Krishna et al., **Visual Genome**: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Arxiv 2016

Challenge

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)
  - COCO dataset
    - 123,287 images (82,783 for training + 40,504 for validation)
    - 5 sentences per image (AMT workers)

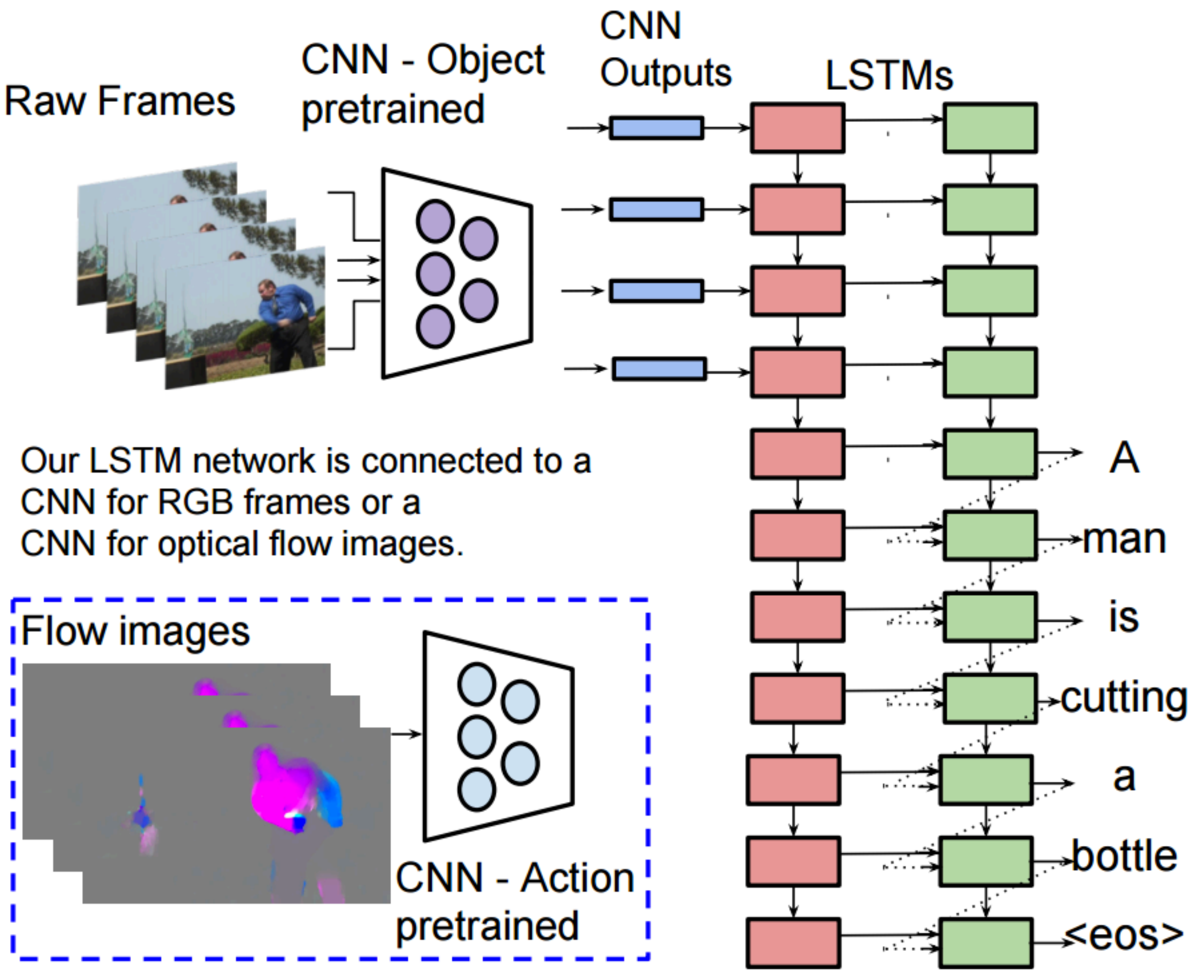


Last updated: 07/12/2016. Please visit CodaLab for the latest results.

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	date
MSM@MSRA[17]	1.003	0.35	0.7	0.919	0.842	0.74	0.632	2016-06-08
THU_MIG[27]	0.988	0.336	0.688	0.913	0.833	0.727	0.616	2016-06-03
ChalLS[6]	0.97	0.34	0.679	0.898	0.809	0.701	0.59	2016-05-21
AugmentCNNwithDet[3]	0.968	0.34	0.683	0.905	0.815	0.706	0.597	2016-03-29
ATT[2]	0.958	0.335	0.682	0.9	0.815	0.709	0.599	2016-01-23
MSRA-MSM[20]	0.954	0.328	0.677	0.901	0.815	0.705	0.591	2016-03-13
Fukun_Jinjunqi[7]	0.946	0.336	0.68	0.902	0.817	0.711	0.601	2016-05-09
Google[9]	0.946	0.346	0.682	0.895	0.802	0.694	0.587	2015-05-29

Video Captioning

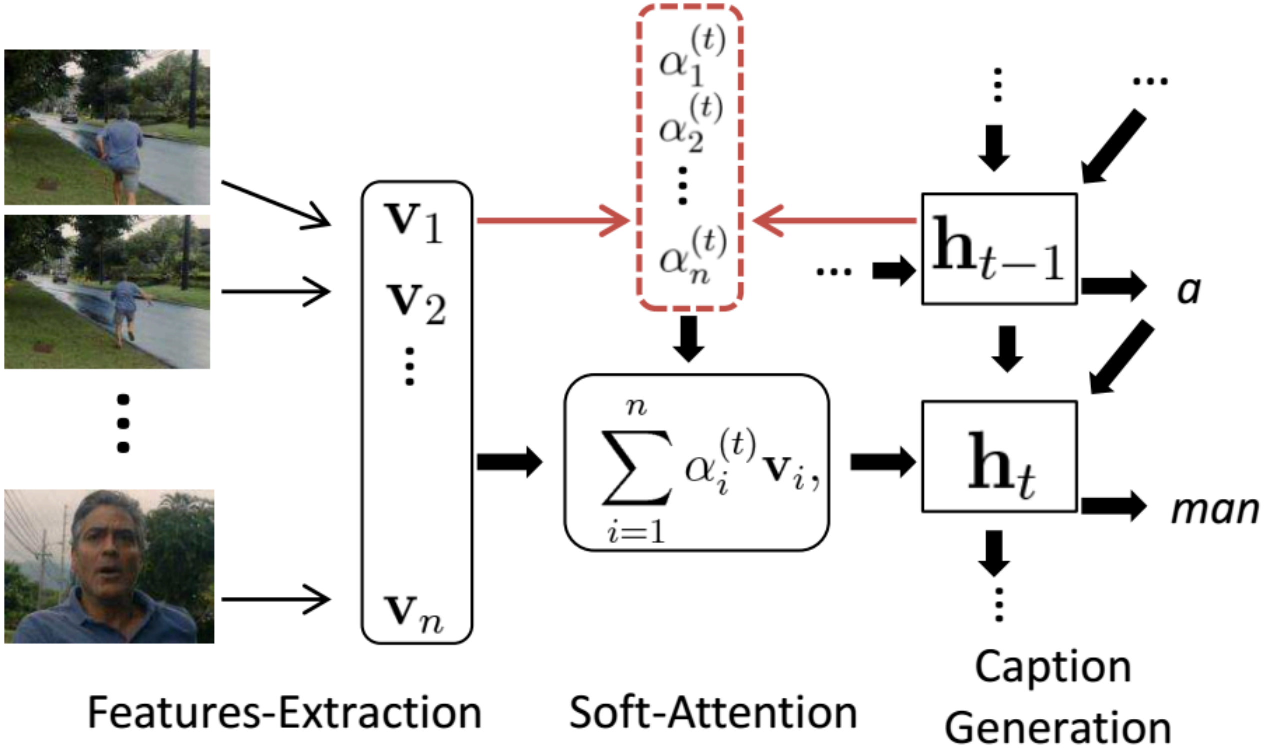
- I have a RNN-Encode, I have a RNN-Decoder: Video-to-Text



Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko. Sequence to Sequence – Video to Text. ICCV'15

Video Captioning-Attention

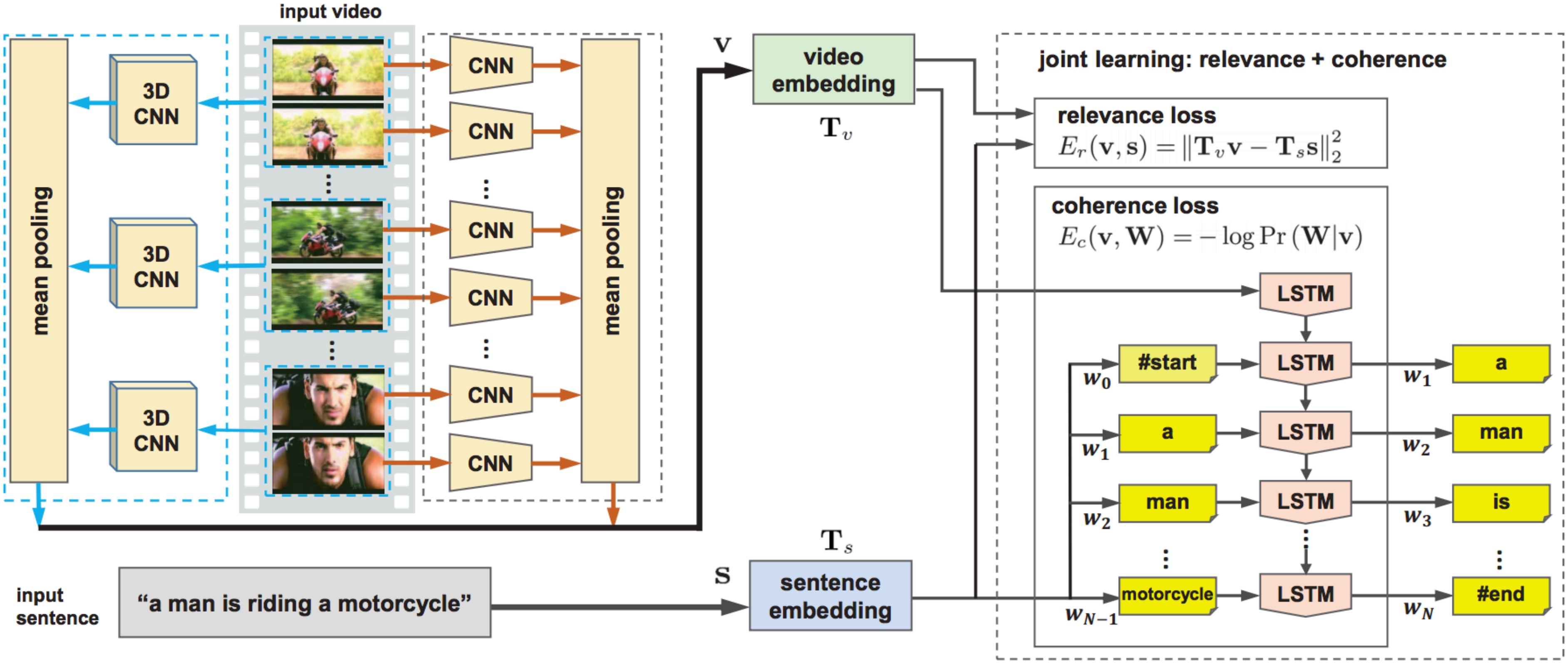
- Frame-level soft-attention



Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville, Describing Videos by Exploiting Temporal Structure. ICCV 2015

Video Captioning-Joint Embedding

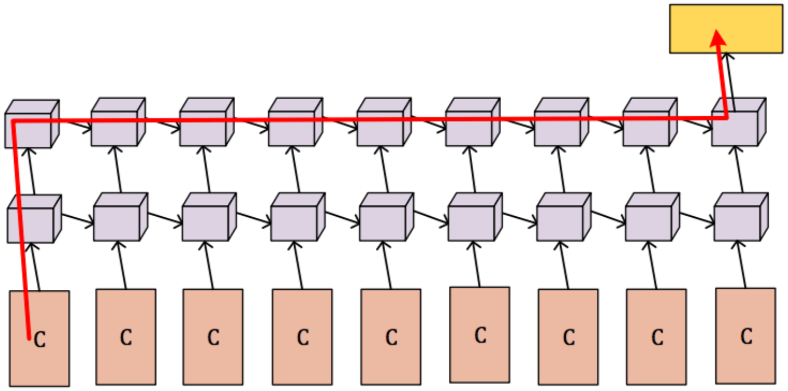
- Additional Relevance Loss



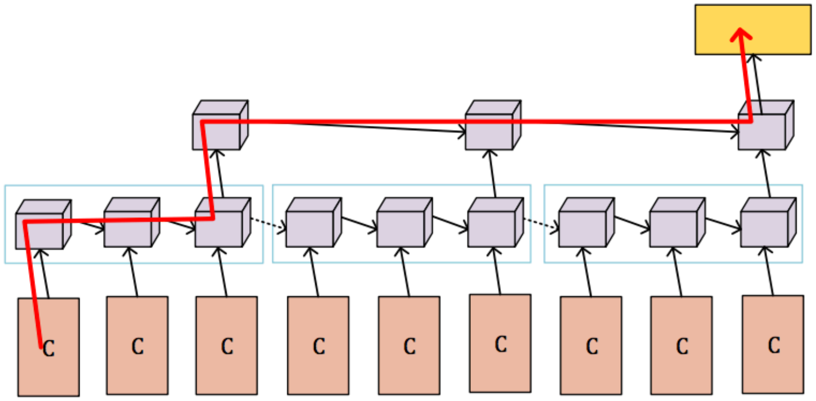
Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language, CVPR 2016

Video Captioning-Hierarchical Encoder

- Exploiting video temporal structure in a longer range



(a) Stacked LSTM video encoder

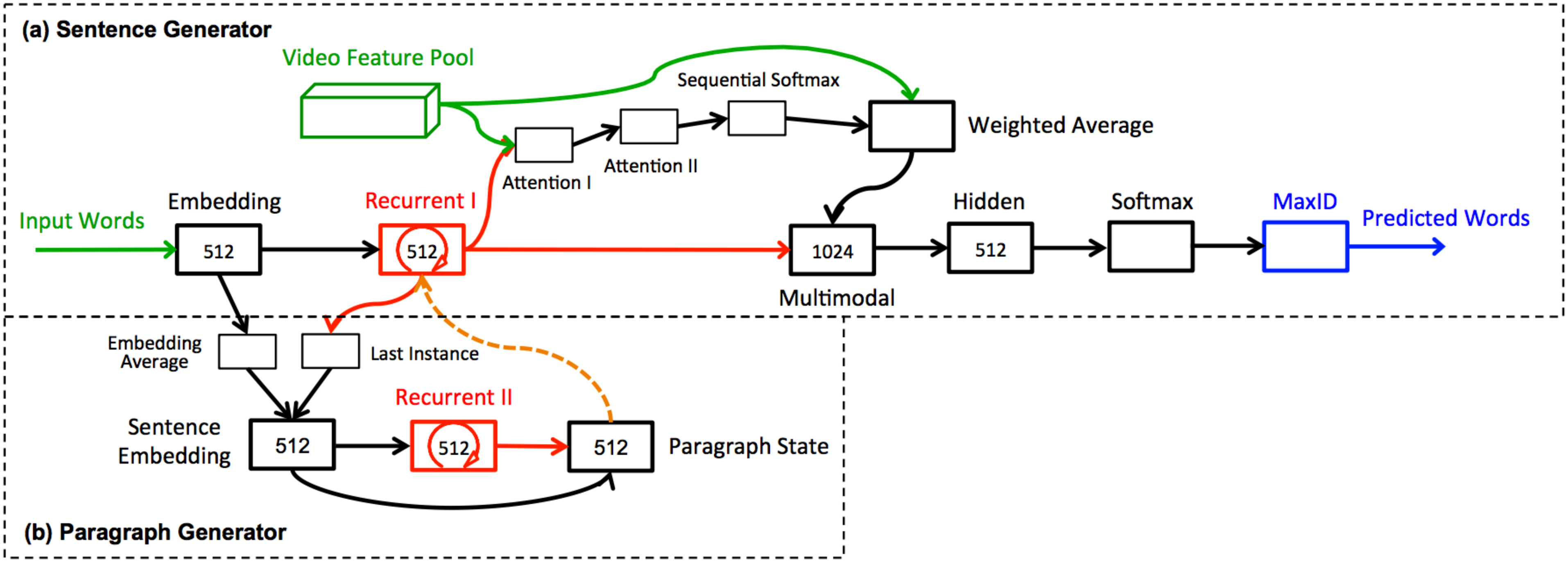


(b) Hierarchical Recurrent Neural Encoder

Pingbo Pan et al., Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. CVPR 2016

Video Captioning-Generate a Paragraph

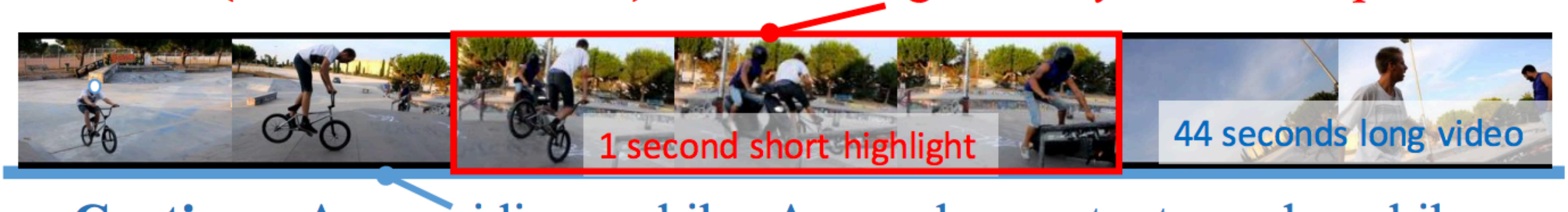
- A sentence generator and a paragraph generator
- Spatial- and Temporal-Attention
- Paragraph state to initialize Recurrent I



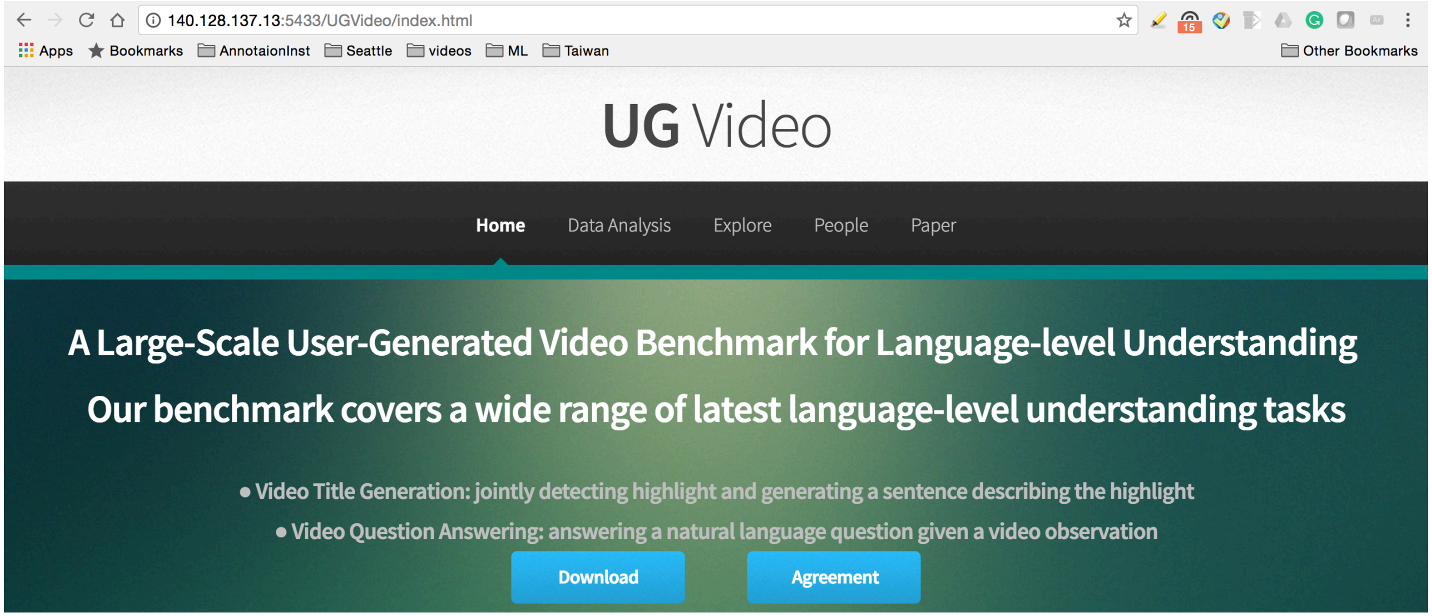
Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, Wei Xu, Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks. CVPR 2016

Video Captioning – Title (Highlight)

Title (most salient event): Bmx rider gets *hit by scooter* at park



Captions: A man riding on bike. A man does a stunt on a bmx bike.



UG Video

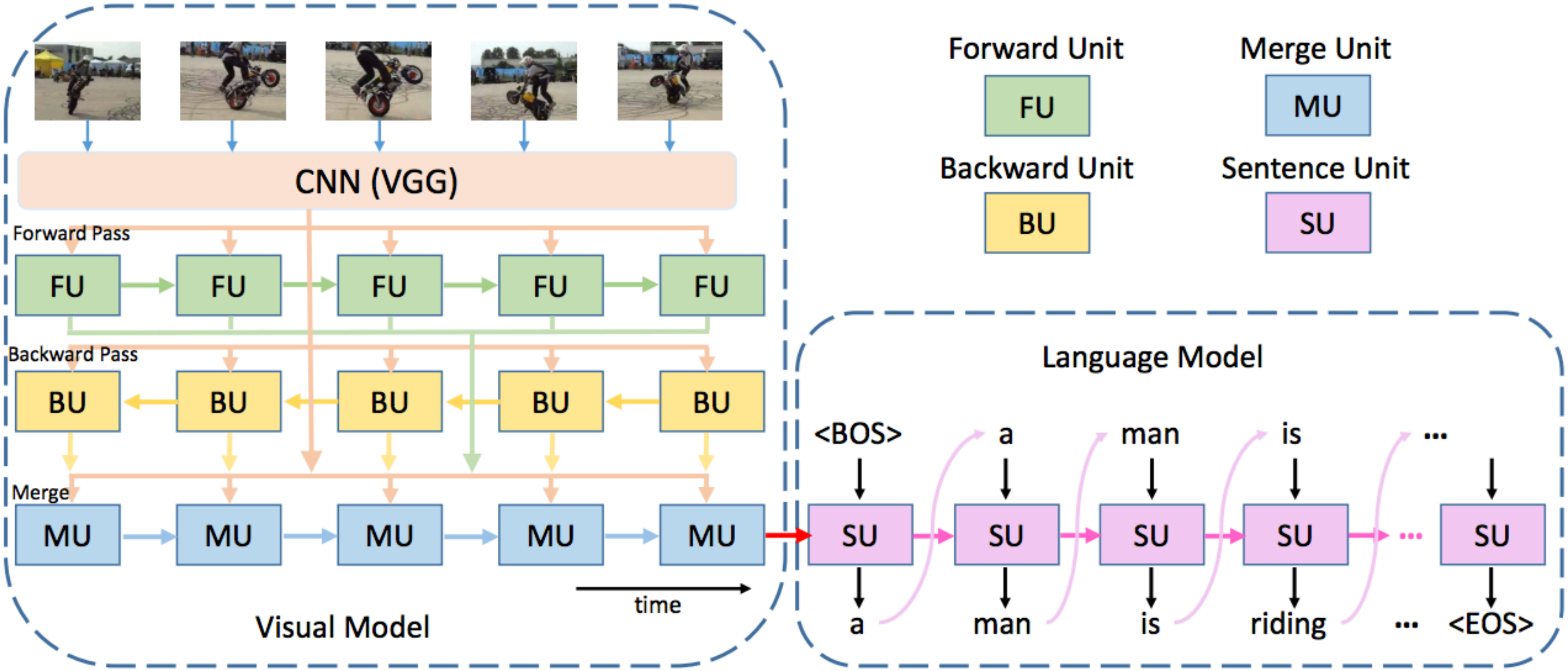
A Large-Scale User-Generated Video Benchmark for Language-level Understanding

Our benchmark covers a wide range of latest language-level understanding tasks

- Video Title Generation: jointly detecting highlight and generating a sentence describing the highlight
- Video Question Answering: answering a natural language question given a video observation

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, Min Sun. Title Generation for User Generated Videos. ECCV 2016

Video Captioning–Bi-direction

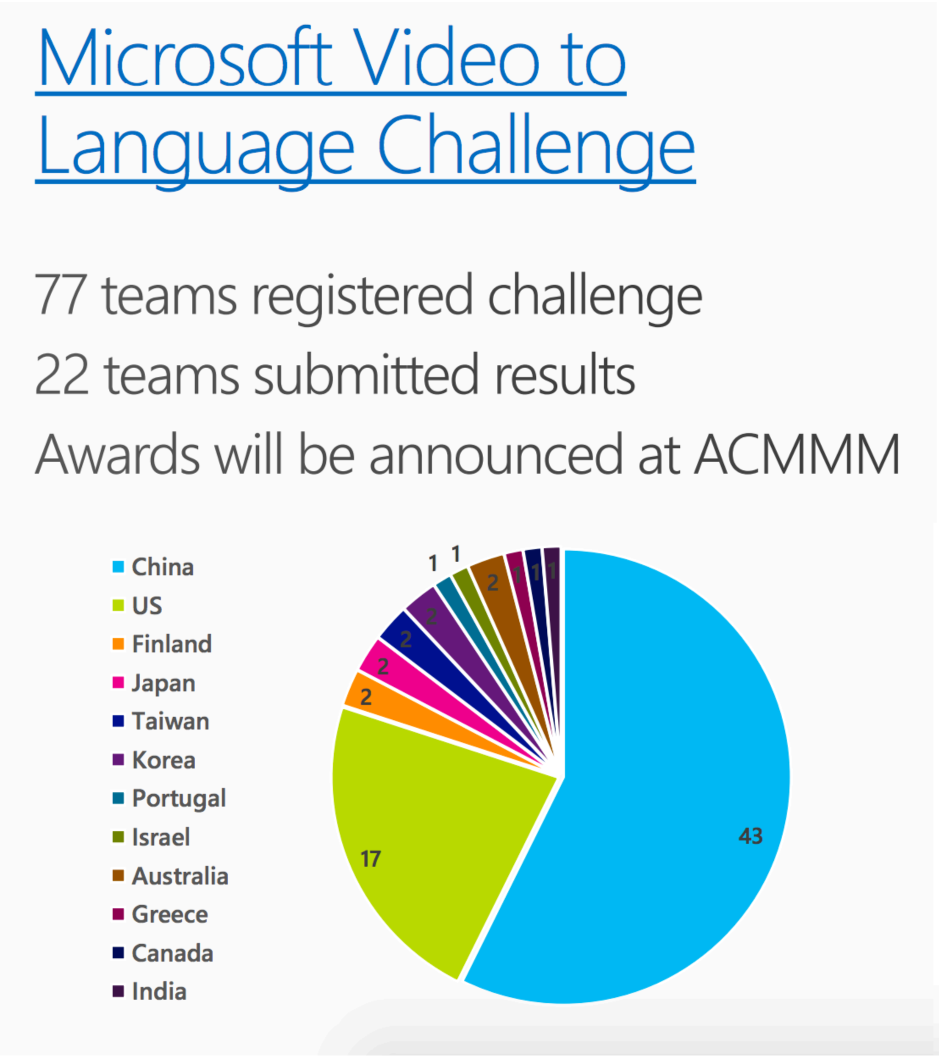


Yi Bin et al., Bidirectional Long-Short Term Memory for Video Description. ACM MM 2016

Video Captioning – Dataset

- Rohrbach et al. MPII Movie Description (MPII-MD). CVPR 2016
- Torabi et al. Montreal Video Annotation Dataset (M-VAD). Arxiv 2016
- Jun Xu , Tao Mei , Ting Yao and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. CVPR 2016
- Zhen et al. Video Title in the Wild (VTW). ECCV 2016

Challenge

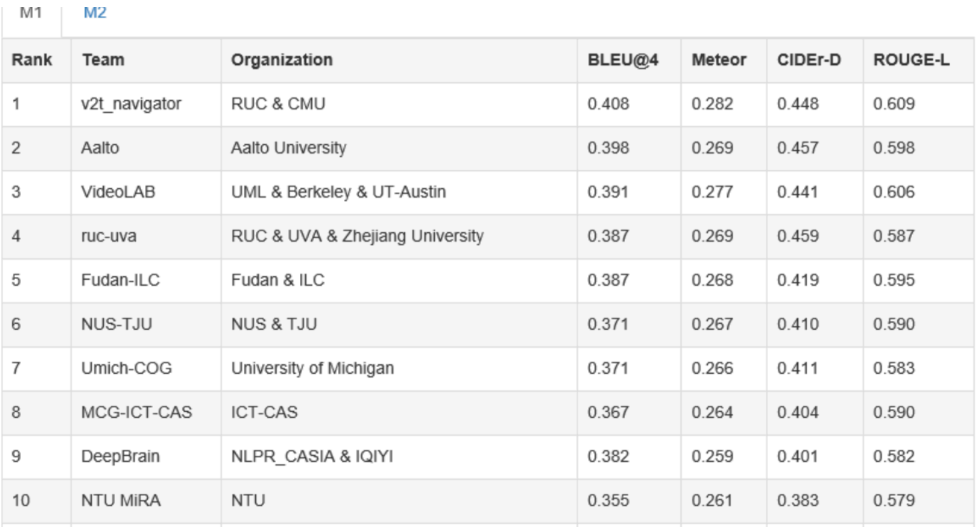


Microsoft Video to Language Challenge

77 teams registered challenge

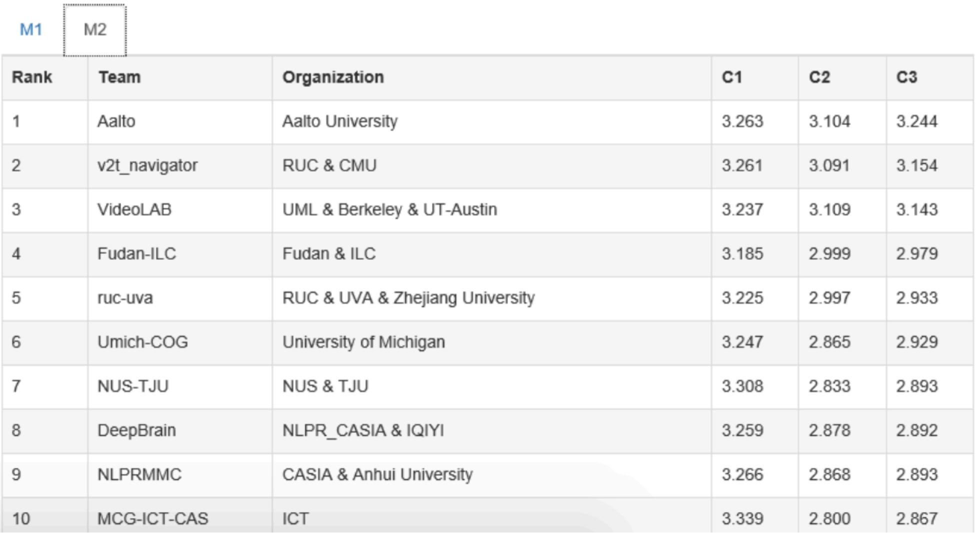
22 teams submitted results

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



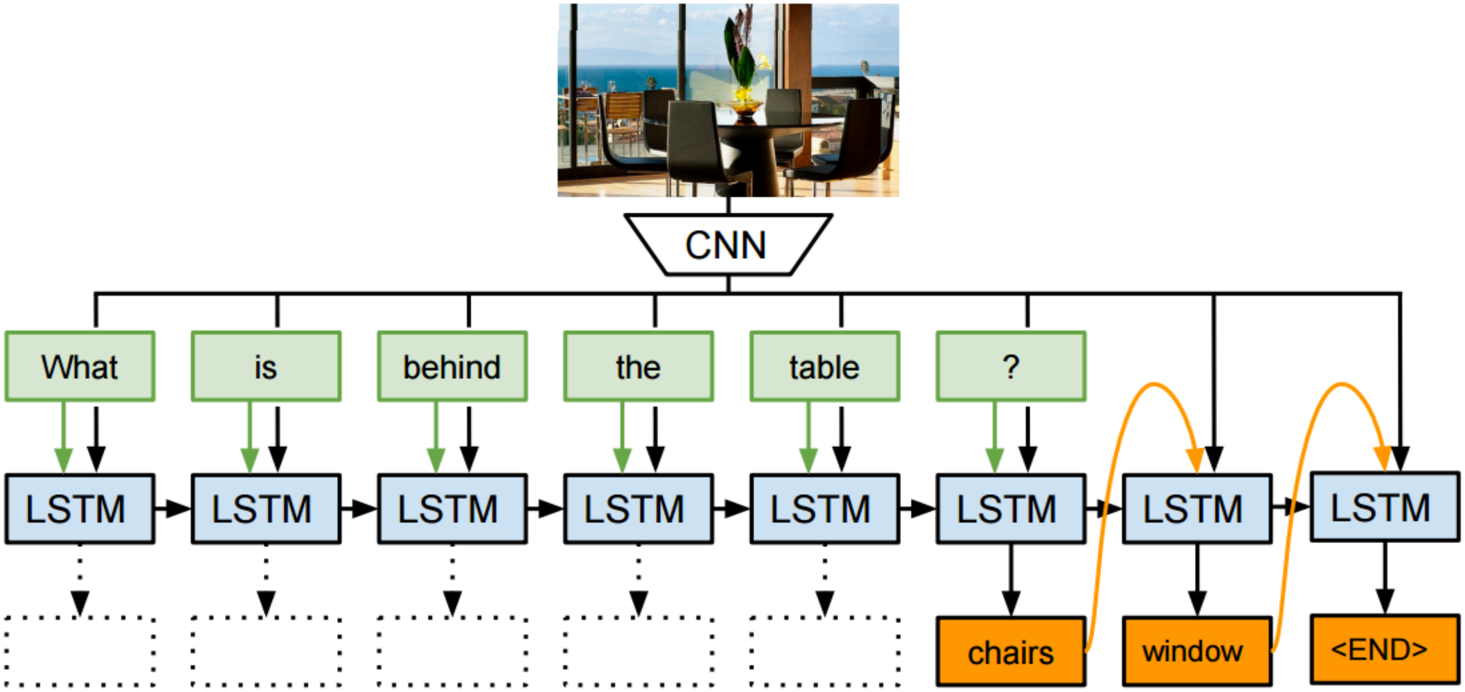
M1 M2

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

Question Answering

Question Answering

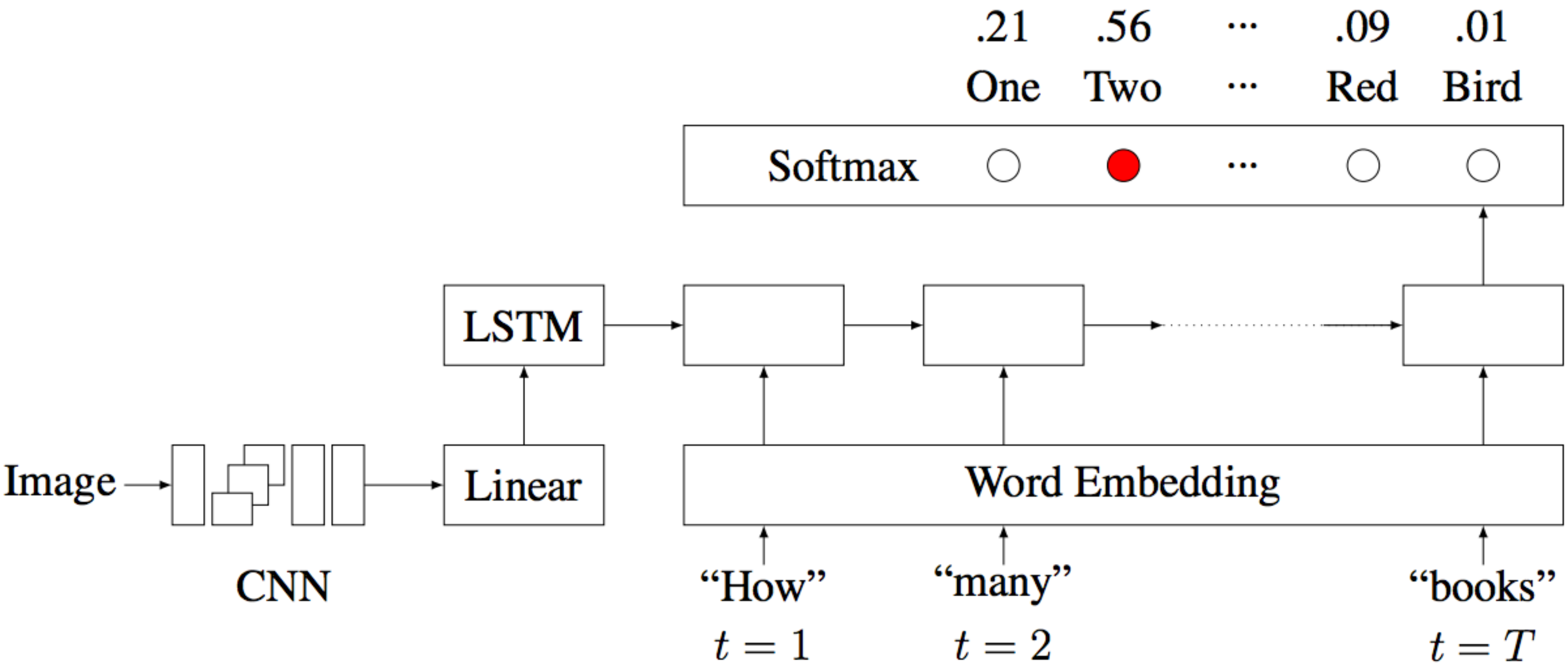
- RNN to encode a question and Image; RNN to decode an answer (multiple words); Single-RNN



Mateusz Malinowski, Marcus Rohrbach, Mario Fritz, Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, ICCV 2015

Question Answering

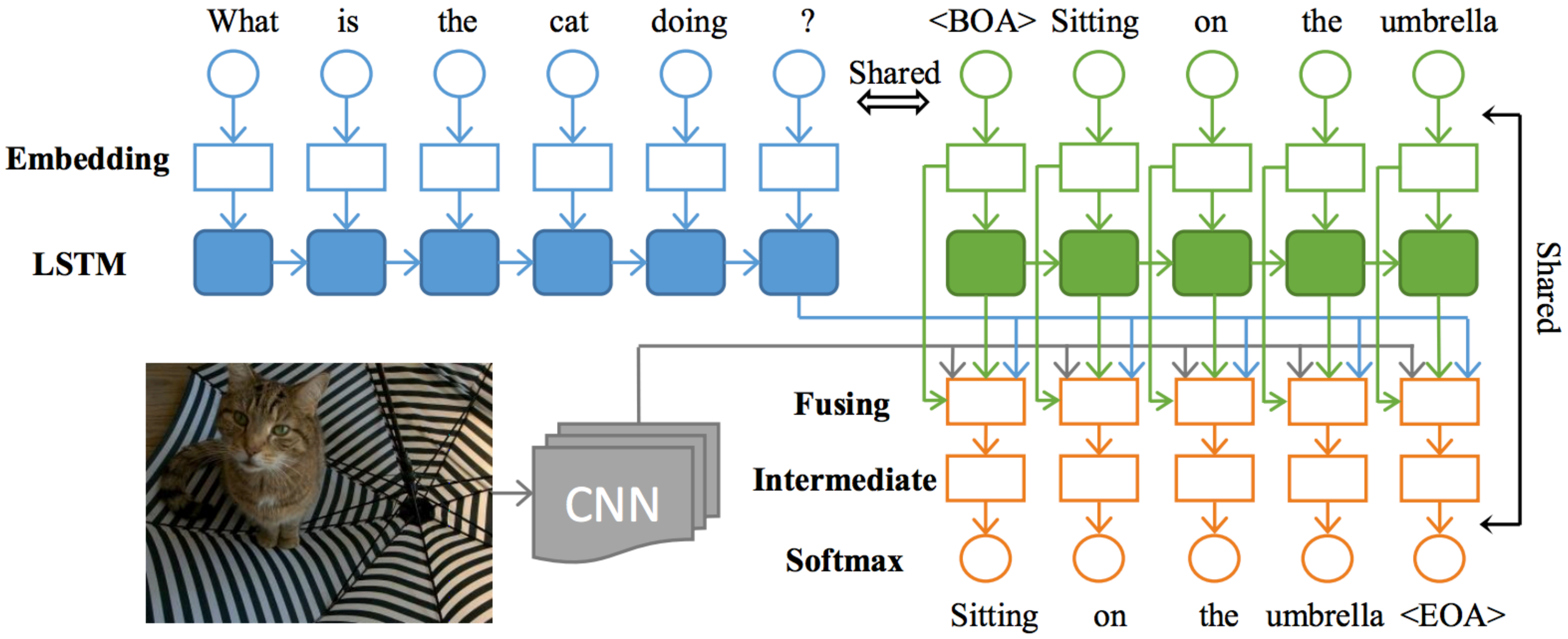
- limited answer space for easy evaluation



Mengye Ren, Ryan Kiros, Richard Zemel, Exploring Models and Data for Image Question Answering, ICML 2015

Question Answering

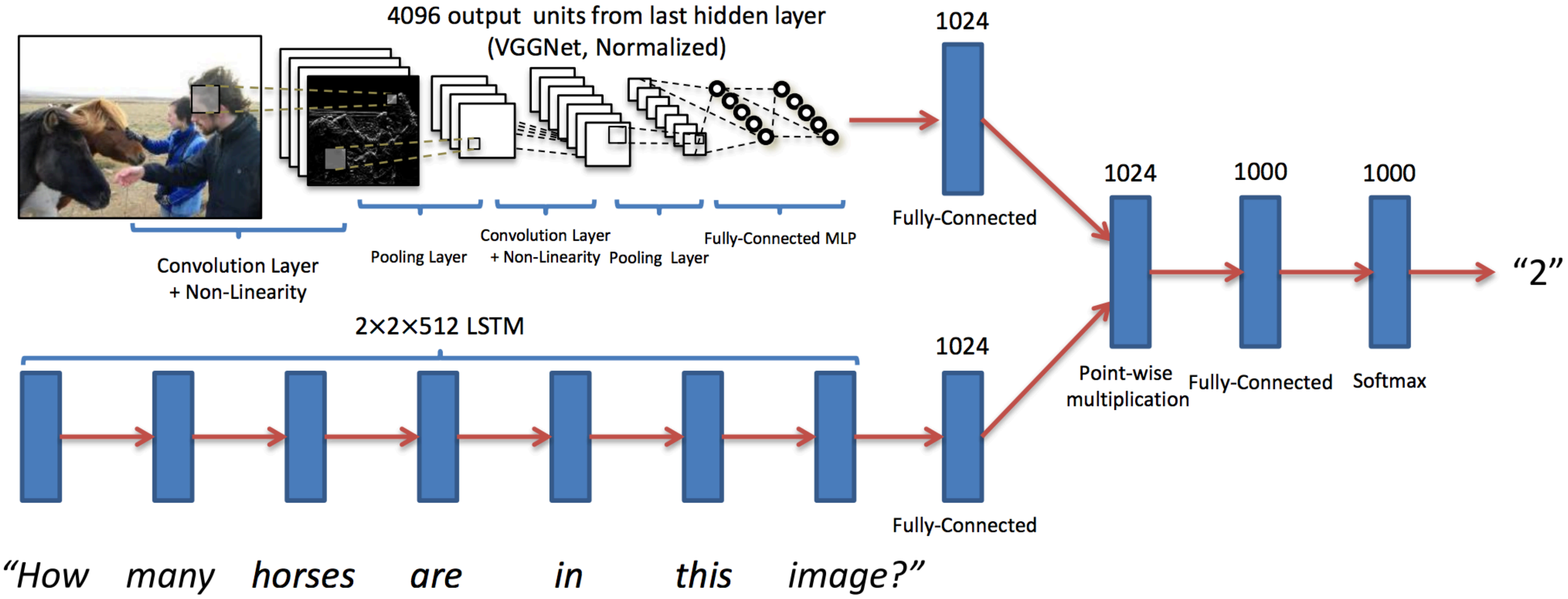
- Separate LSTM-Q and LSTM-A



H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and methods for multilingual image question answering. NIPS 2015

Question Answering

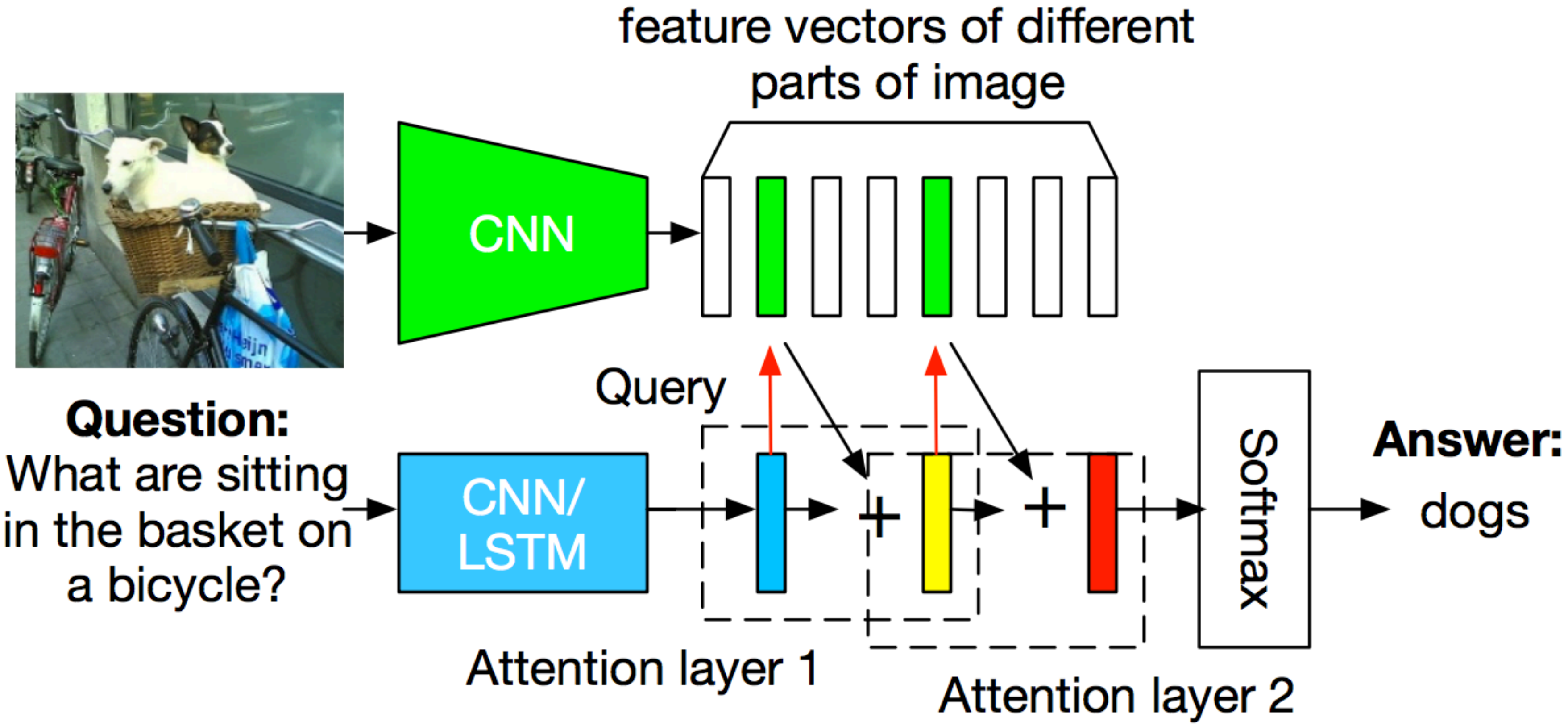
- Point-wise multiplication



S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. ICCV 2015

Question Answering – Attention

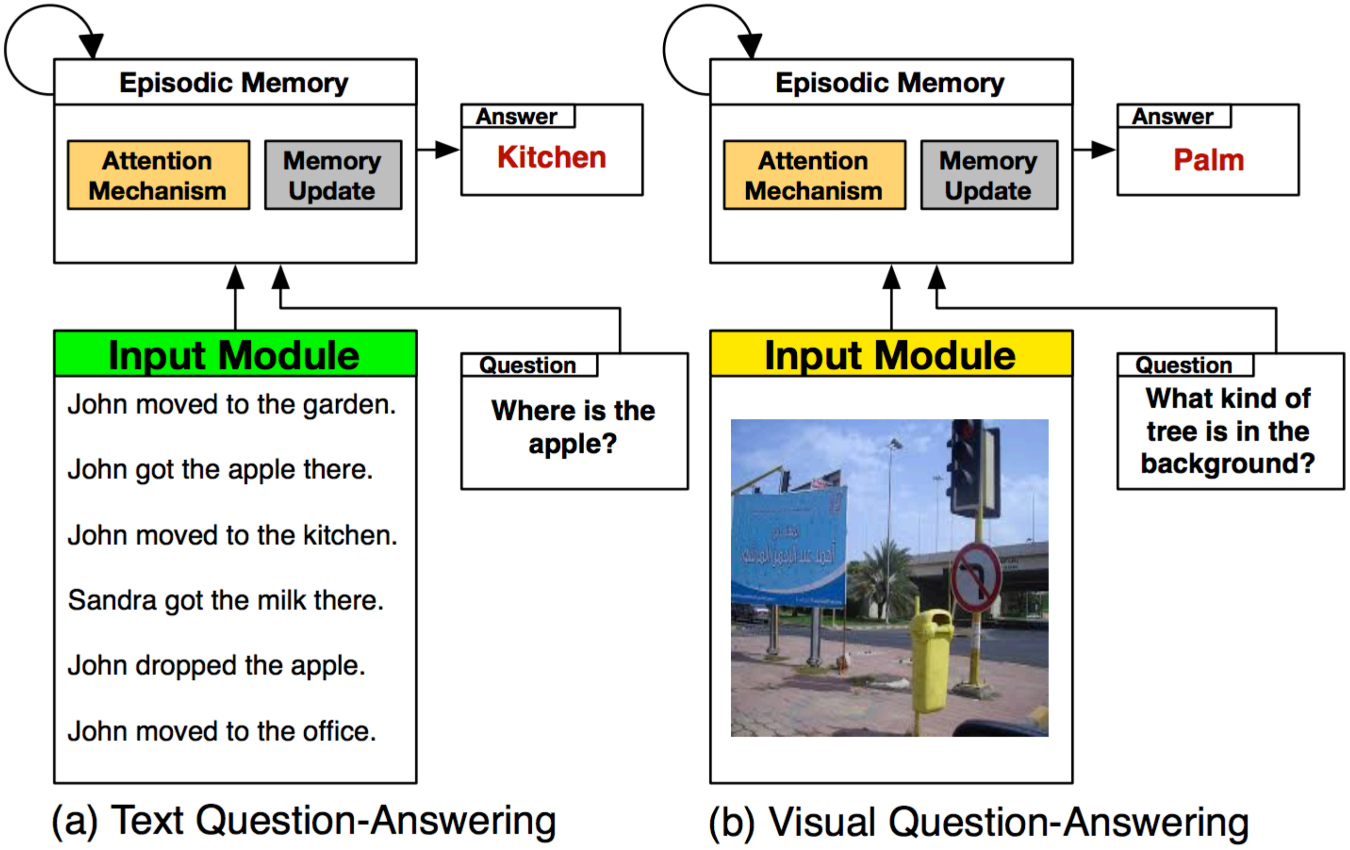
- Stack Attention Network (SAN)


feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, Stacked Attention Networks for Image Question Answering, CVPR 2016

Question Answering – Attention

- Dynamic Memory Network

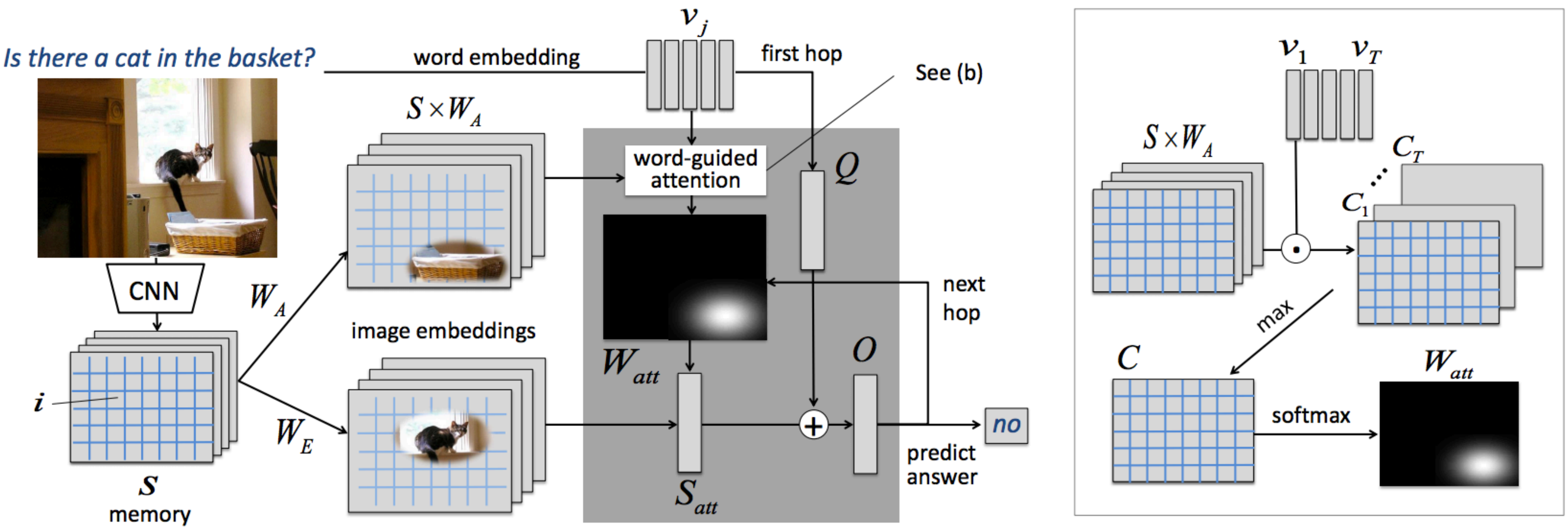


(a) Text Question-Answering (b) Visual Question-Answering

Caiming Xiong, Stephen Merity, Richard Socher, Dynamic Memory Networks for Visual and Textual Question Answering, ICML 2016

Question Answering – Attention

- word to patch at 1st hop; whole Q at 2nd hop.



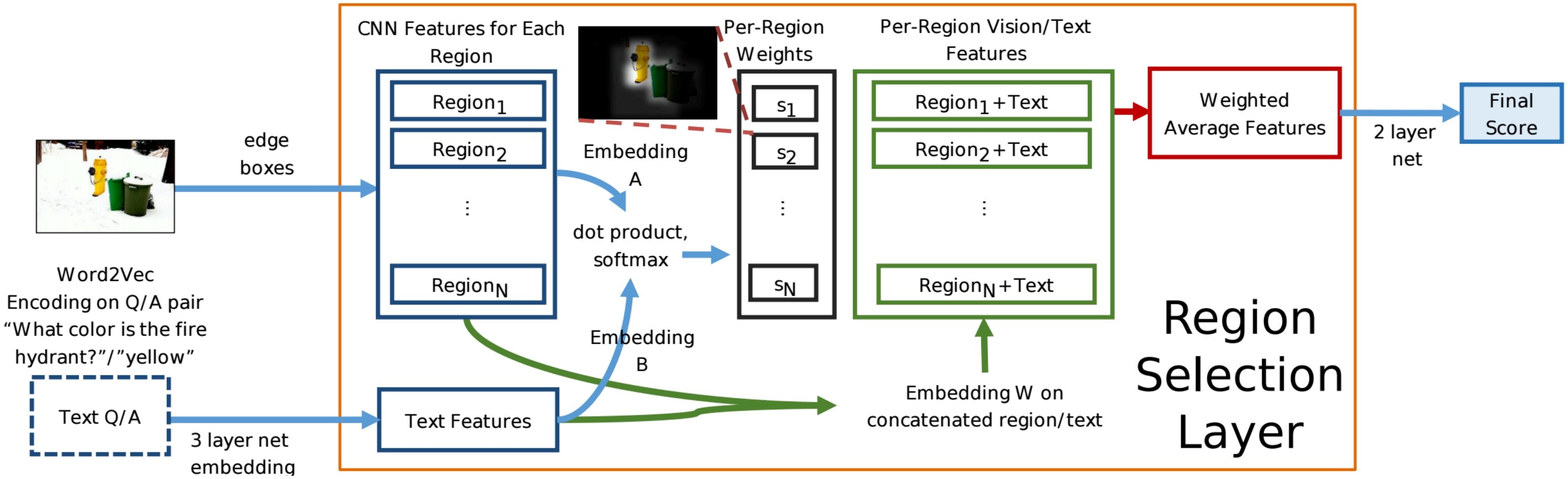
(a) Overview

(b) Word-guided attention

Huijuan Xu, Kate Saenko, Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering, ECCV 2016

Question Answering – Attention

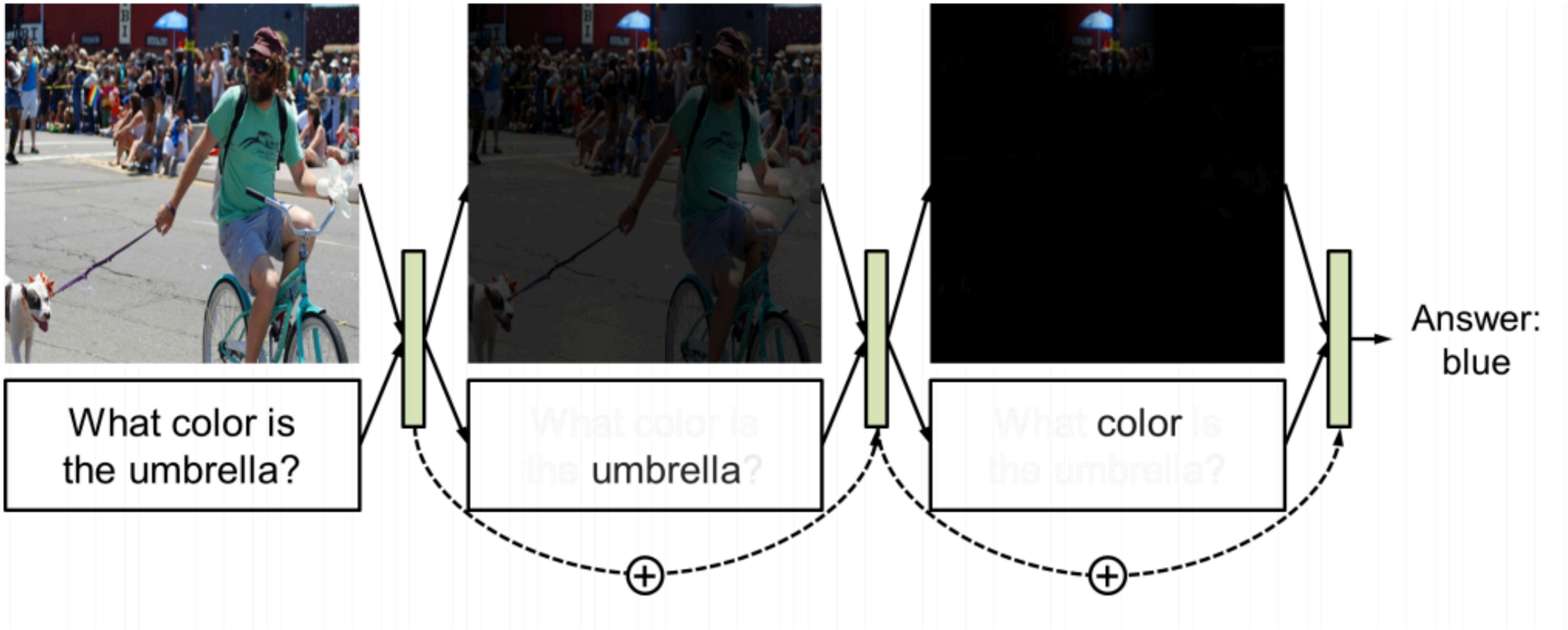
- Averaged representation of word2vec vectors for language



Kevin J. Shih, Saurabh Singh, Derek Hoiem, Where To Look: Focus Regions for Visual Question Answering, CVPR 2016

Question Answering – Attention

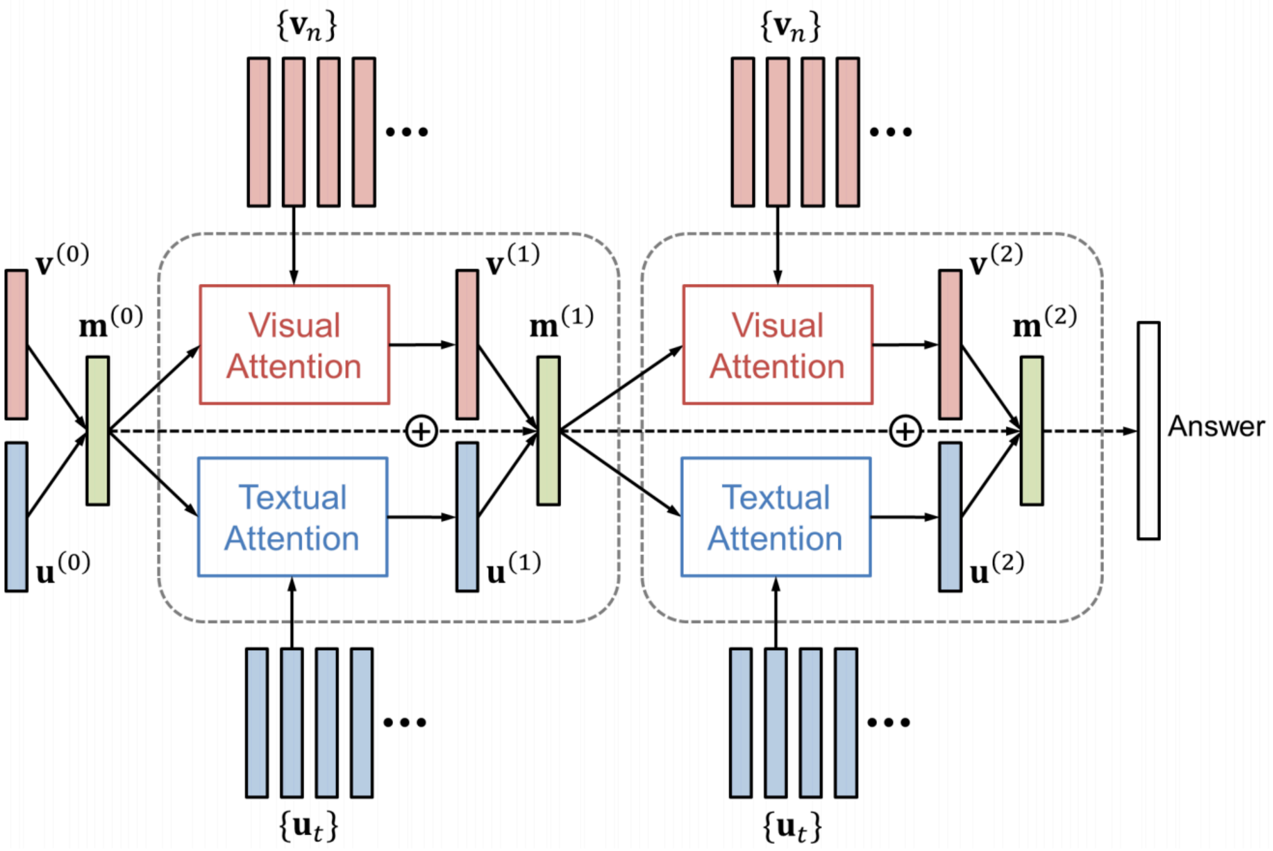
- Dual Attention


What color is the umbrella?
What color is the umbrella?
What color is the umbrella?
Answer: blue

Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. CVPR'16 VQA Challenge Workshop

Question Answering – Attention

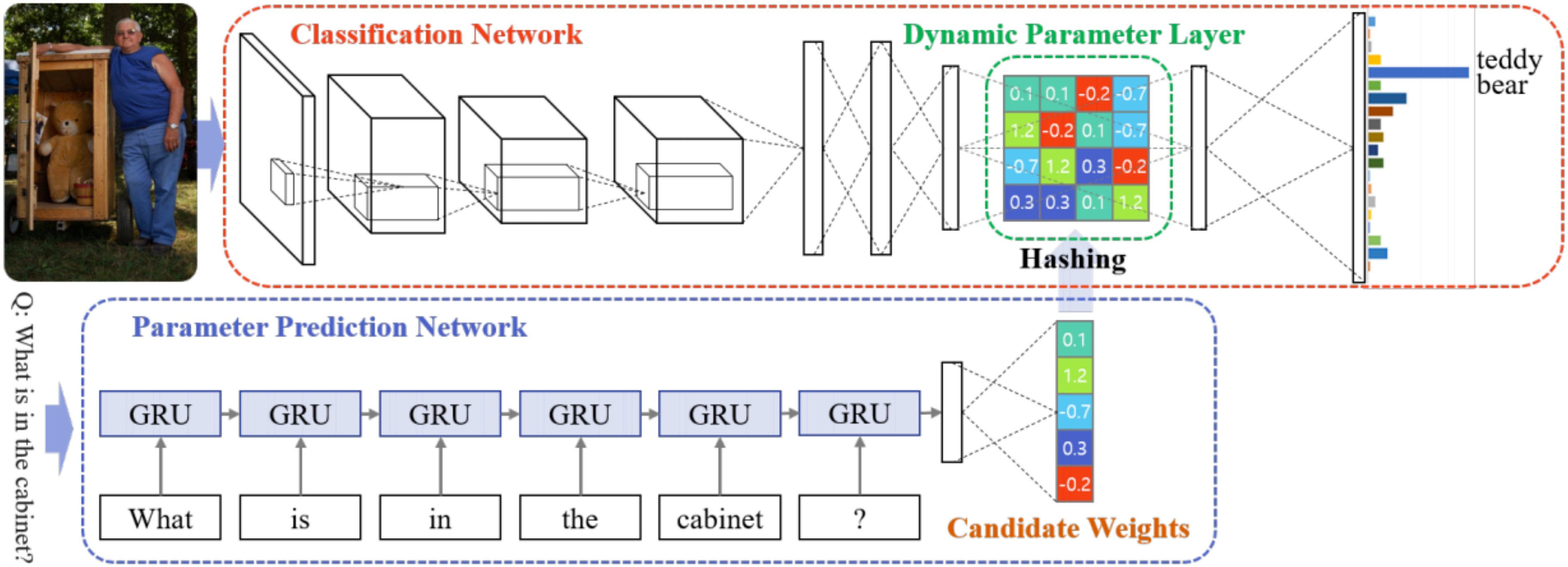
- Dual Attention



Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. CVPR'16 VQA Challenge Workshop

Question Answering – Dynamic Parameter

- Dynamic Parameter Layer (hashing)

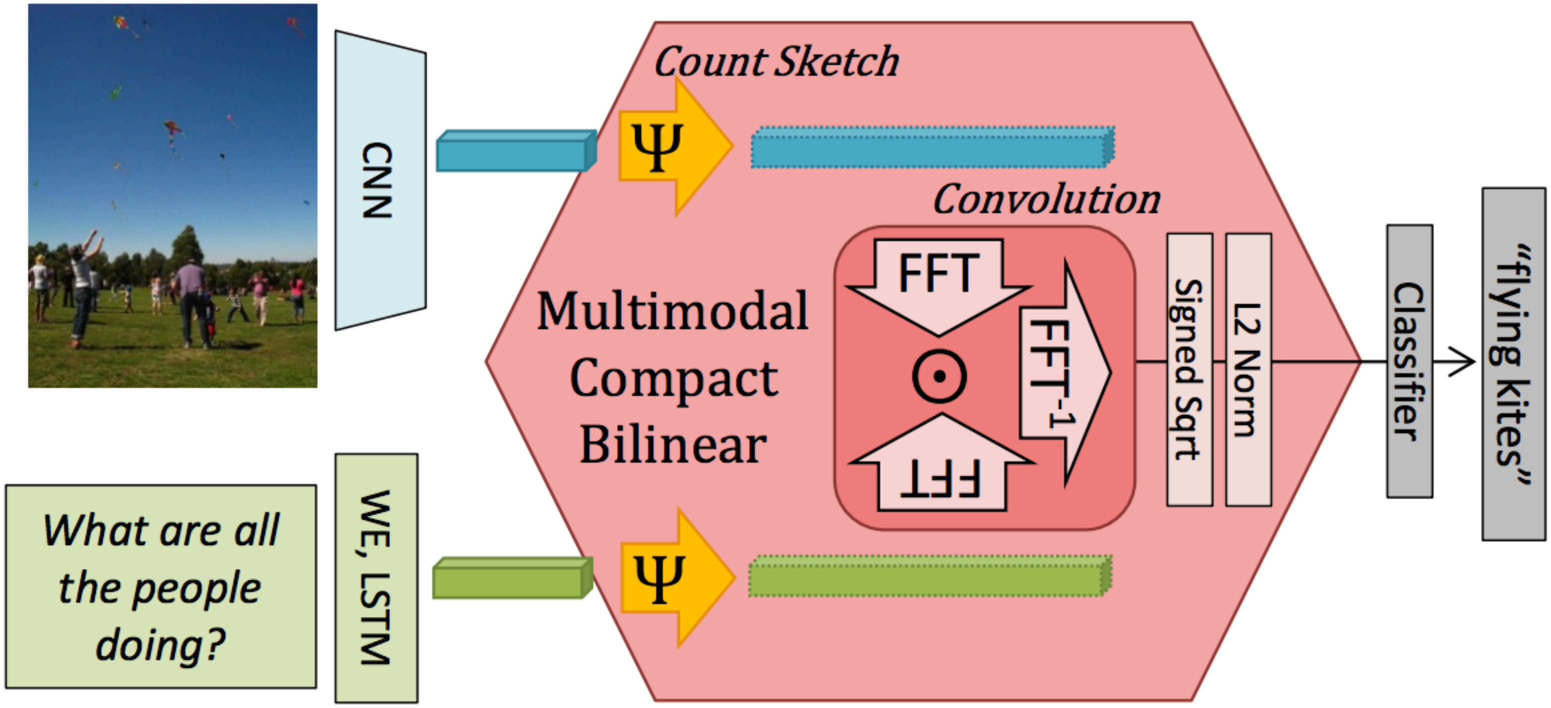


0.1	0.1	-0.2	-0.7
1.2	-0.2	0.1	-0.7
-0.7	1.2	0.3	-0.2
0.3	0.3	0.1	1.2

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction, CVPR 2016

Question Answering – Bilinear Pooling

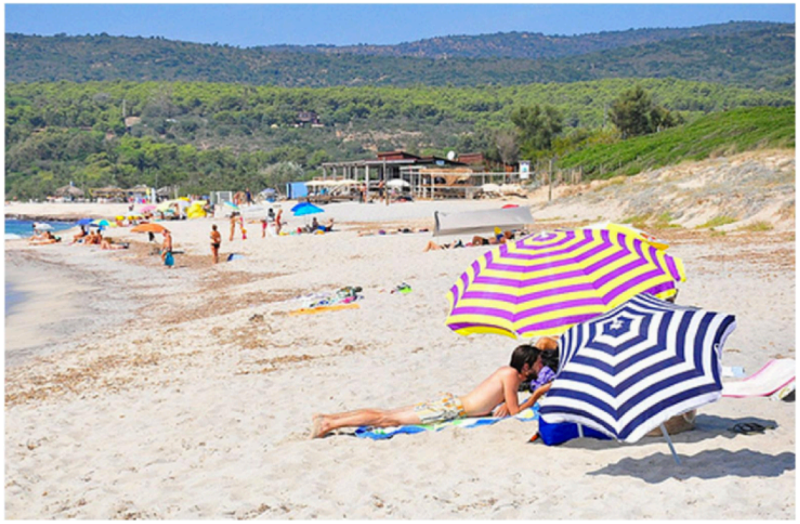
- Outer product of the visual and textual vectors



Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. CVPR'16 VQA Challenge Workshop.

Question Answering – Knowledge

- External Knowledge



Attributes:
umbrella
beach
sunny
day
people
sand
laying
blue
green
mountain

Internal Textual Representation:
A group of people enjoying a sunny day at the beach with umbrellas in the sand.

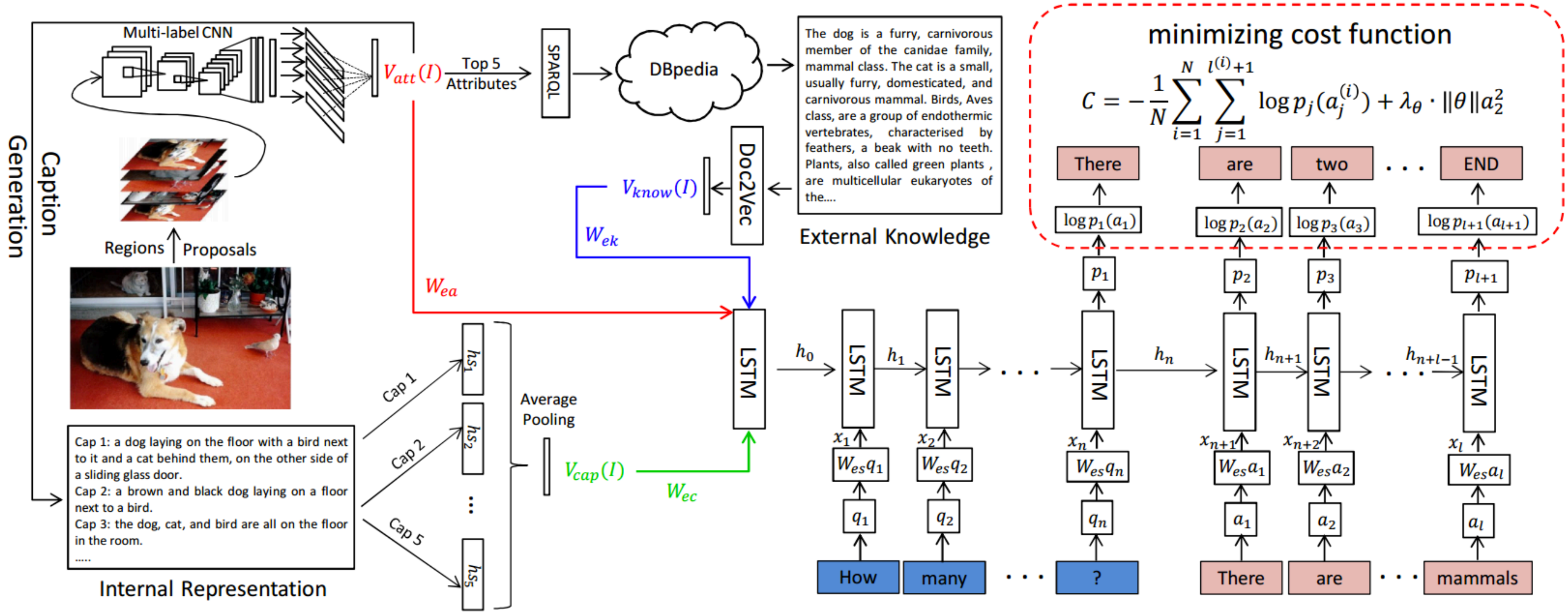
External Knowledge:
An umbrella is a canopy designed to protect against rain or sunlight. Larger umbrellas are often used as points of shade on a sunny beach. A beach is a landform along the coast of an ocean. It usually consists of loose particles, such as sand....

Question Answering:
Q: Why do they have umbrellas? **A :** Shade.

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. CVPR 2016

Question Answering – Knowledge

- External Knowledge



The dog is a furry, carnivorous member of the canidae family, mammal class. The cat is a small, usually furry, domesticated, and carnivorous mammal. Birds, Aves class, are a group of endothermic vertebrates, characterised by feathers, a beak with no teeth. Plants, also called green plants , are multicellular eukaryotes of the....

Cap 1: a dog laying on the floor with a bird next to it and a cat behind them, on the other side of a sliding glass door.
Cap 2: a brown and black dog laying on a floor next to a bird.
Cap 3: the dog, cat, and bird are all on the floor in the room.
.....

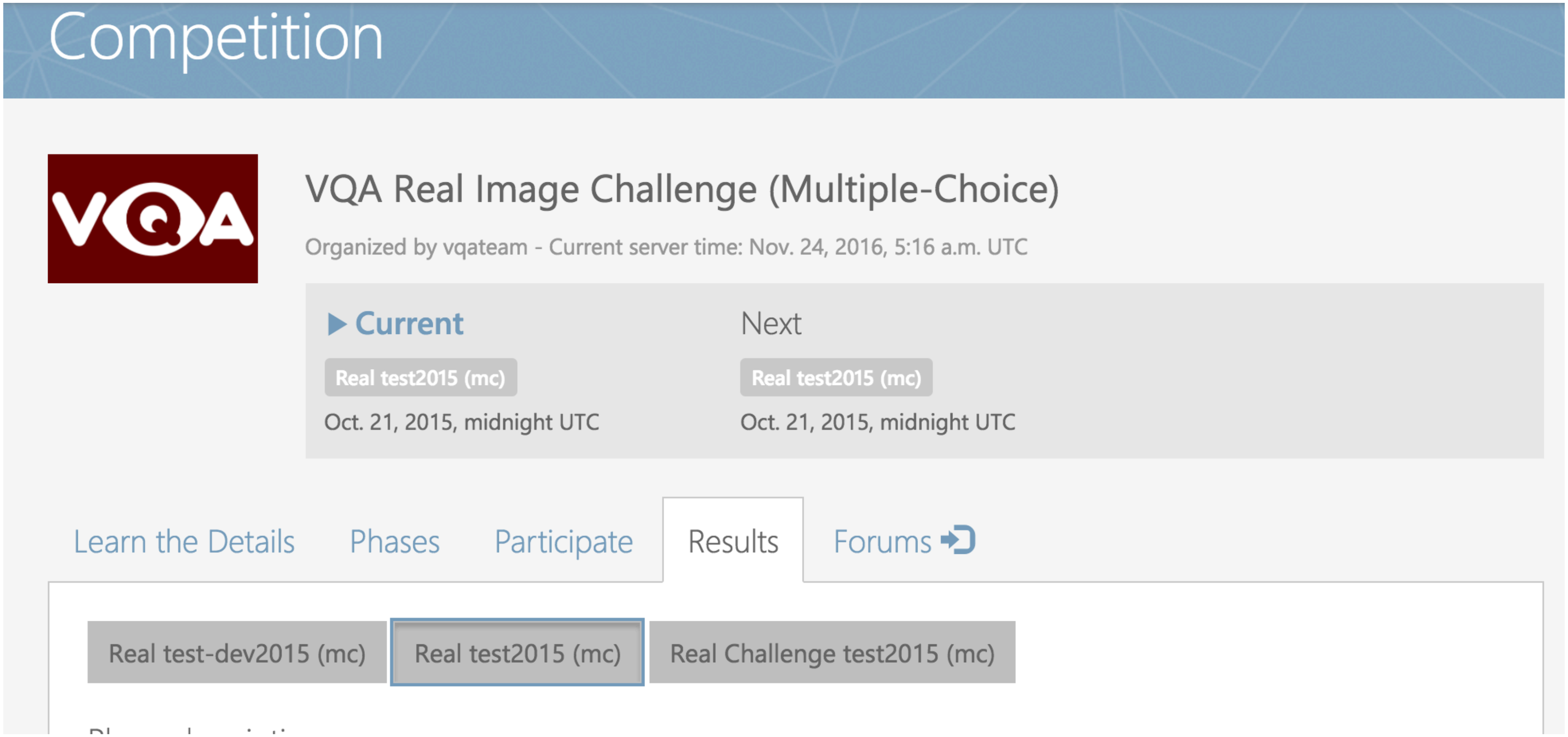
$$C = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l^{(i)}+1}\log p_j(a_j^{(i)}) + \lambda_\theta \cdot \|\theta\|a_2^2$$

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. CVPR 2016

Question Answering – Dataset

- DAQUAR – Malinowski and Fritz. NIPS 2014
- VQA - based on MSCOCO images. ICCV 2015
- COCO-QA - based on MSCOCO images. Ren et al. ICML 2015
- FM-IQA - built from scratch by Baidu - in Chinese, with English translation. Gao et al. NIPS 2015
- Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei, Visual7W: Grounded Question Answering in Images, CVPR 2016.

Challenge



Competition

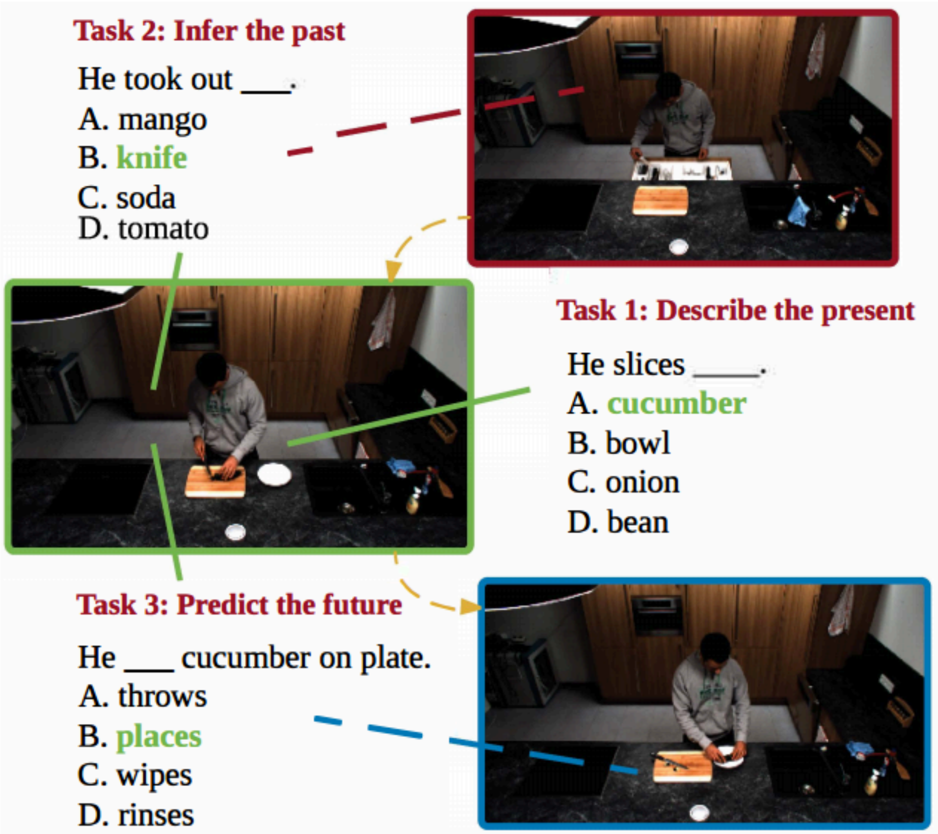
VQA Real Image Challenge (Multiple-Choice)

Organized by vqateam - Current server time: Nov. 24, 2016, 5:16 a.m. UTC

▶ Current	Next
Real test2015 (mc)	Real test2015 (mc)
Oct. 21, 2015, midnight UTC	Oct. 21, 2015, midnight UTC

Video Question Answering

- Learning to rank multiple choices



Task 2: Infer the past

He took out ___.
A. mango
B. knife
C. soda
D. tomato

Task 1: Describe the present

He slices ____.
A. cucumber
B. bowl
C. onion
D. bean

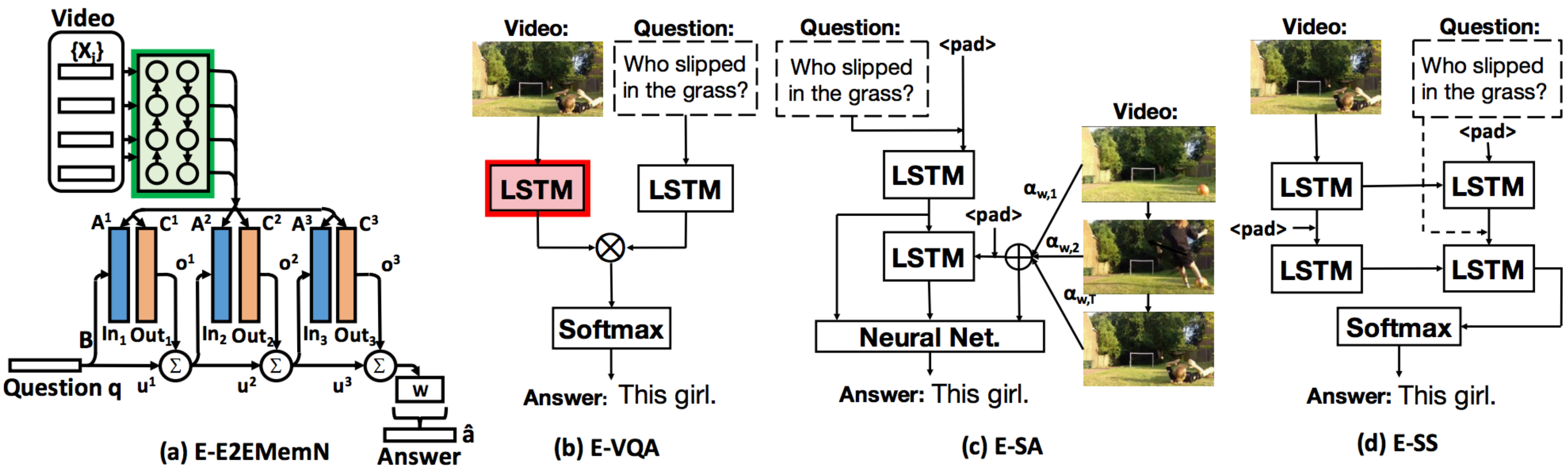
Task 3: Predict the future

He ___ cucumber on plate.
A. throws
B. places
C. wipes
D. rinses

Linchao Zhu, Zhongwen Xu, Yi Yang, Alexander G. Hauptmann. Uncovering Temporal Context for Video Question and Answering. arXiv 2015

Video Question Answering

- Multiple Extensions for Video-QA

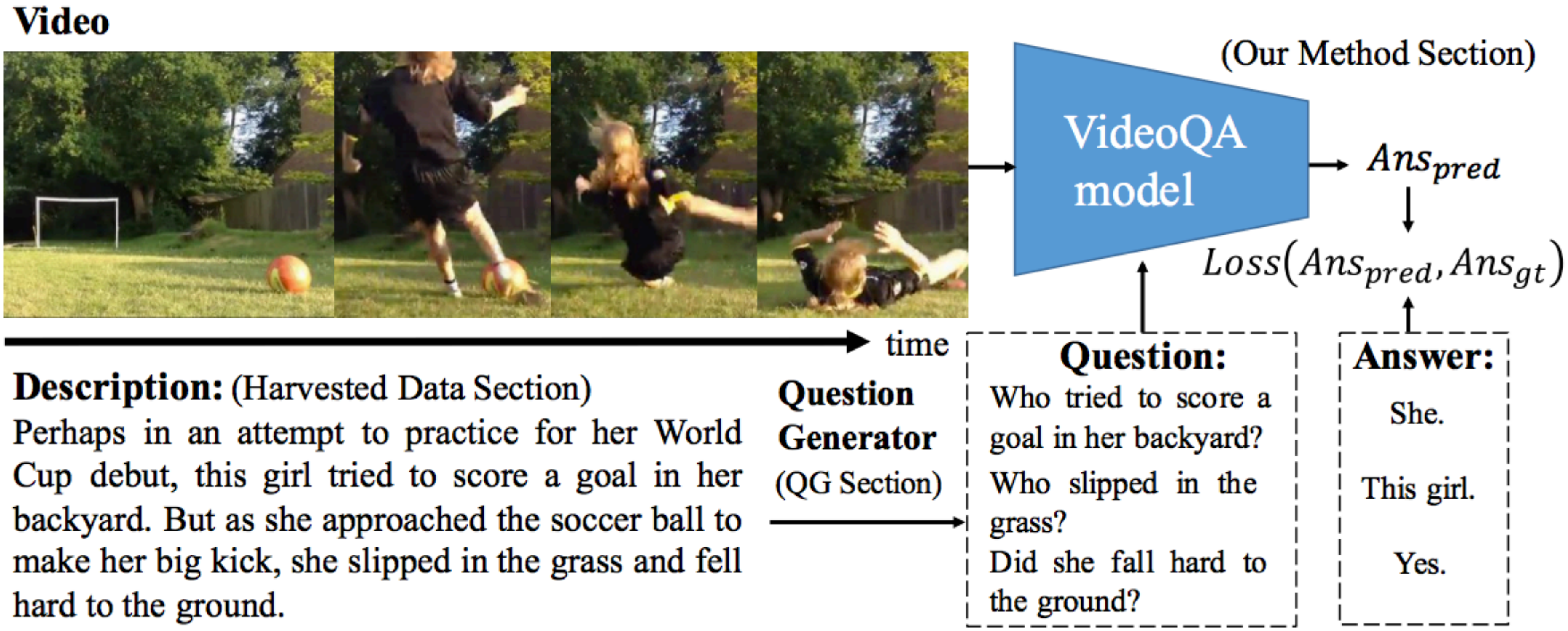


(a) E-E2EMemN (b) E-VQA (c) E-SA (d) E-SS

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. AAAI 2017

Video Question Answering – Dataset

- VTW Video-QA dataset



Description: (Harvested Data Section)
Perhaps in an attempt to practice for her World Cup debut, this girl tried to score a goal in her backyard. But as she approached the soccer ball to make her big kick, she slipped in the grass and fell hard to the ground.

Question:
Who tried to score a goal in her backyard?
Who slipped in the grass?
Did she fall hard to the ground?

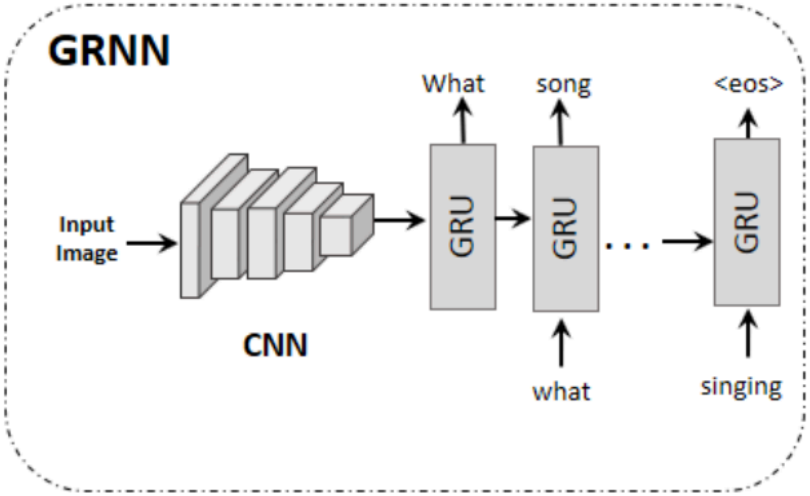
Answer:
She.
This girl.
Yes.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. AAAI 2017

Video Question Generation

- How many horses are in the field? ✗

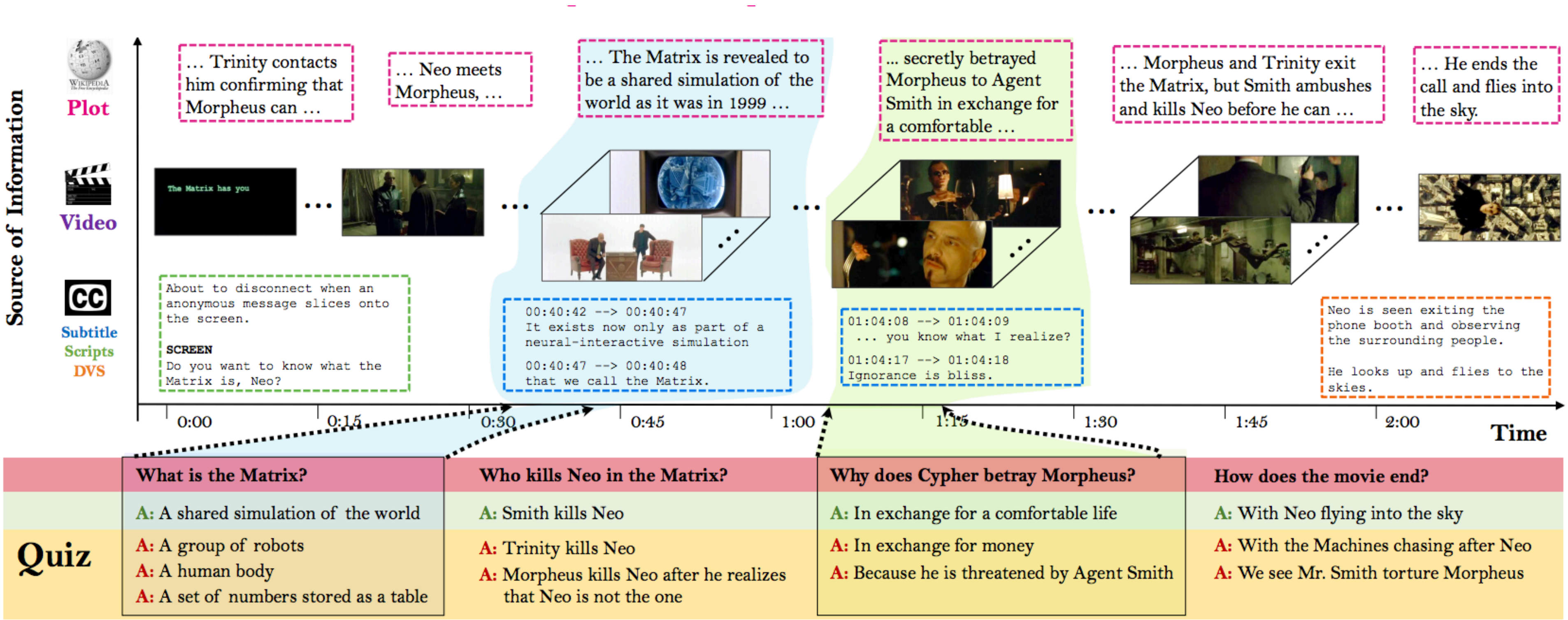
- Who won the race? ✓



Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, Lucy Vanderwende, Generating Natural Questions About an Image, ACL 2016

Video Question Answering – Dataset

- Moive-QA



… Trinity contacts him confirming that Morpheus can …

… Neo meets Morpheus, …

… The Matrix is revealed to be a shared simulation of the world as it was in 1999 …

... secretly betrayed Morpheus to Agent Smith in exchange for a comfortable …

… Morpheus and Trinity exit the Matrix, but Smith ambushes and kills Neo before he can …

… He ends the call and flies into the sky.

About to disconnect when an anonymous message slices onto the screen.

SCREEN
Do you want to know what the Matrix is, Neo?

00:40:42 --> 00:40:47
It exists now only as part of a neural-interactive simulation

00:40:47 --> 00:40:48
that we call the Matrix.

01:04:08 --> 01:04:09
… you know what I realize?

01:04:17 --> 01:04:18
Ignorance is bliss.

Neo is seen exiting the phone booth and observing the surrounding people.

He looks up and flies to the skies.

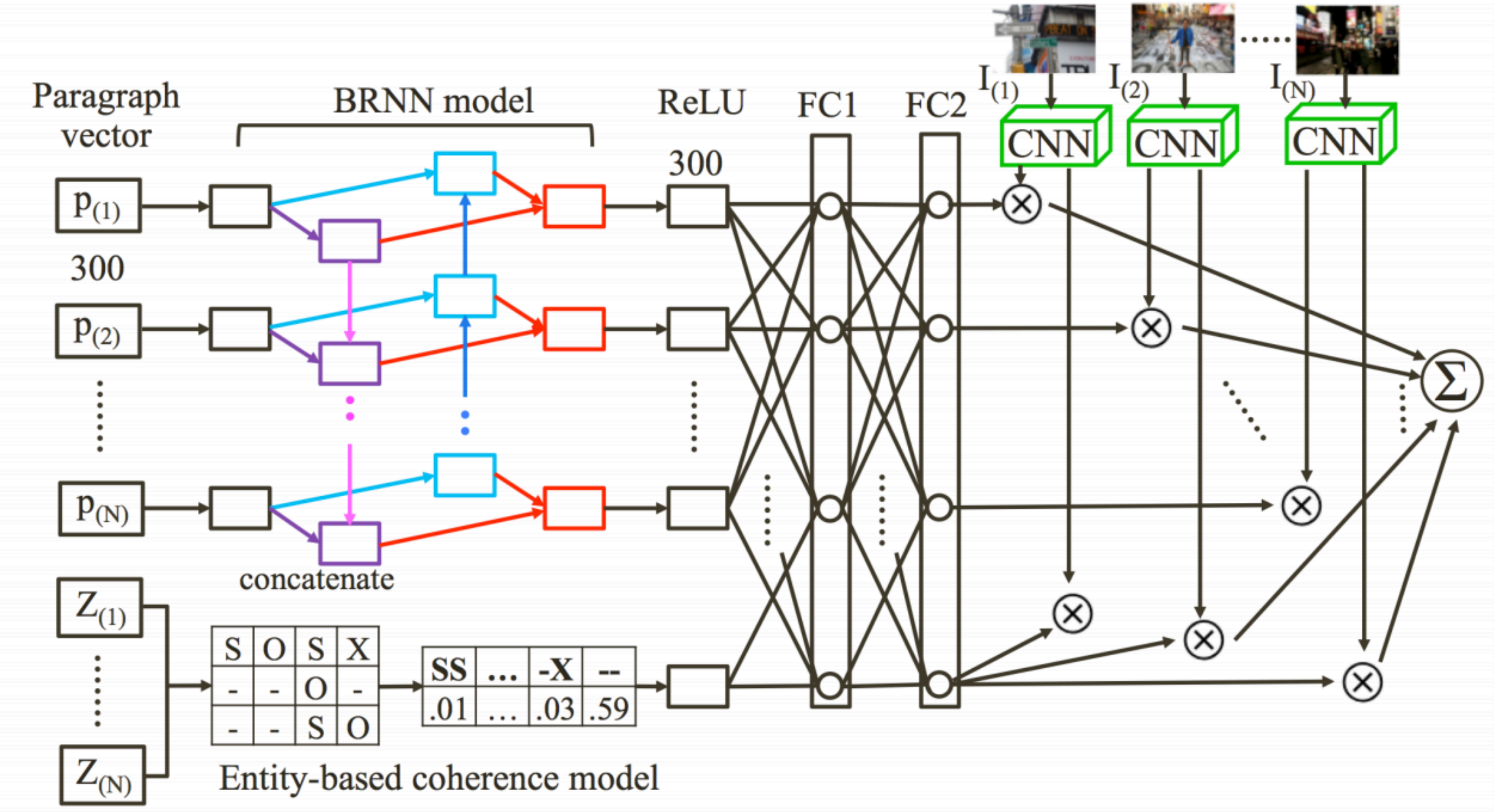
	What is the Matrix?	Who kills Neo in the Matrix?	Why does Cypher betray Morpheus?	How does the movie end?
	A: A shared simulation of the world	A: Smith kills Neo	A: In exchange for a comfortable life	A: With Neo flying into the sky
Quiz	A: A group of robots A: A human body A: A set of numbers stored as a table	A: Trinity kills Neo A: Morpheus kills Neo after he realizes that Neo is not the one	A: In exchange for money A: Because he is threatened by Agent Smith	A: With the Machines chasing after Neo A: We see Mr. Smith torture Morpheus

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, Sanja Fidler, MovieQA: Understanding Stories in Movies through Question-Answering, CVPR 2016.

Others

Storytelling – Retrieval

- **Retrieve** fluent sequential multiple sentences



S	O	S	X
-	-	O	-
-	-	S	O

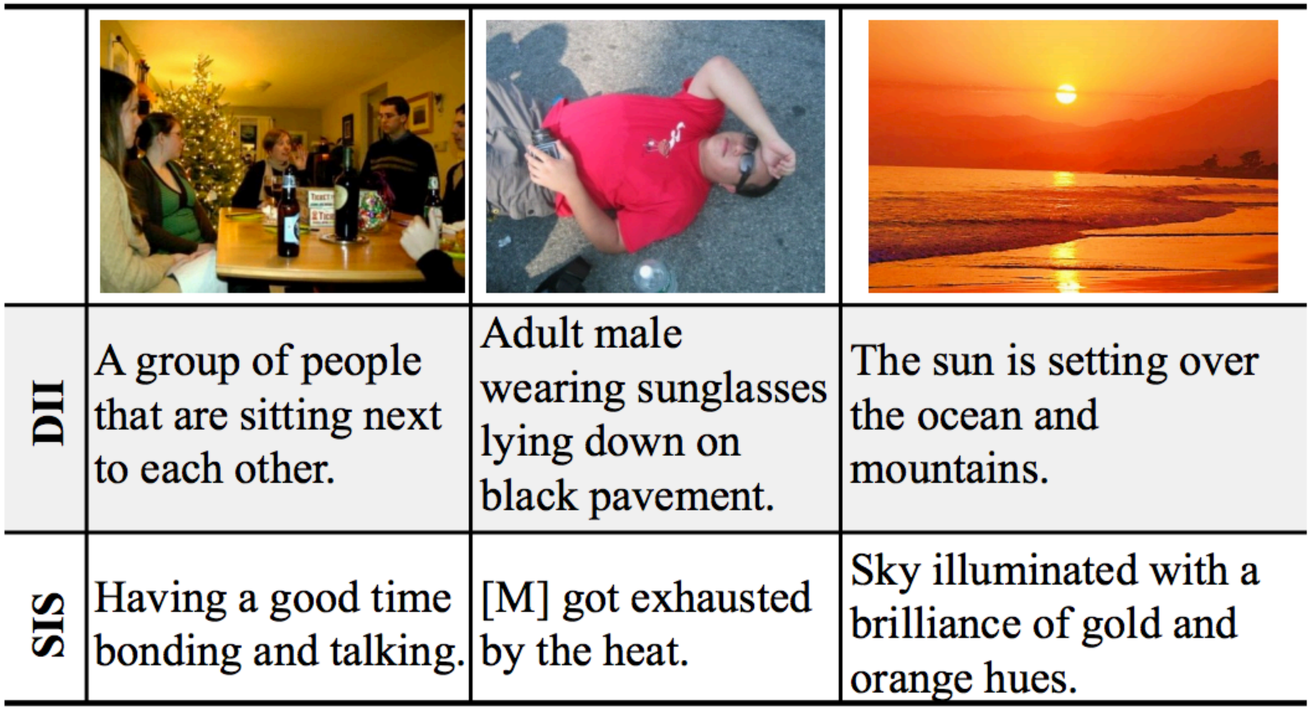
SS	…	-X	--
.01	…	.03	.59

(b) The CRCN architecture

Cesc Chunseong Park, Gunhee Kim. Expressing an Image Stream with a Sequence of Natural Sentences. NIPS 2015

Storytelling – Generate

- **Sequential Images Narrative Dataset (SIND)**



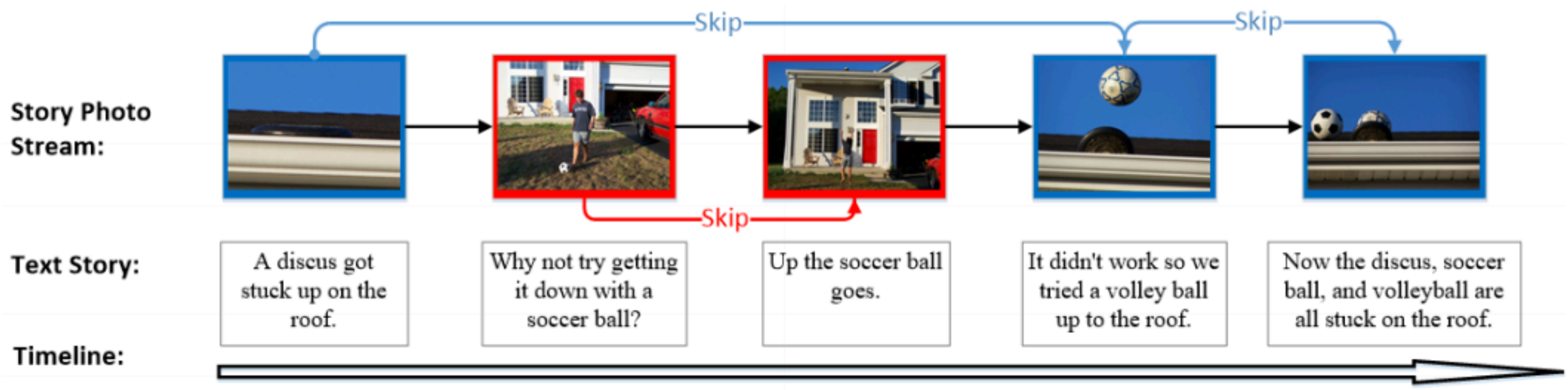
DII	A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.	The sun is setting over the ocean and mountains.
SIS	Having a good time bonding and talking.	[M] got exhausted by the heat.	Sky illuminated with a brilliance of gold and orange hues.

Figure 1: Example language difference between descriptions for images in isolation (DII) vs. stories for images in sequence (SIS).

Ting-Hao (Kenneth) Huang et al., Visual Storytelling, NAACL 2016

Storytelling - Skip

- sGRU



Skip

Skip

Story Photo Stream:

Skip

Text Story:

A discus got stuck up on the roof.

Why not try getting it down with a soccer ball?

Up the soccer ball goes.

It didn't work so we tried a volley ball up to the roof.

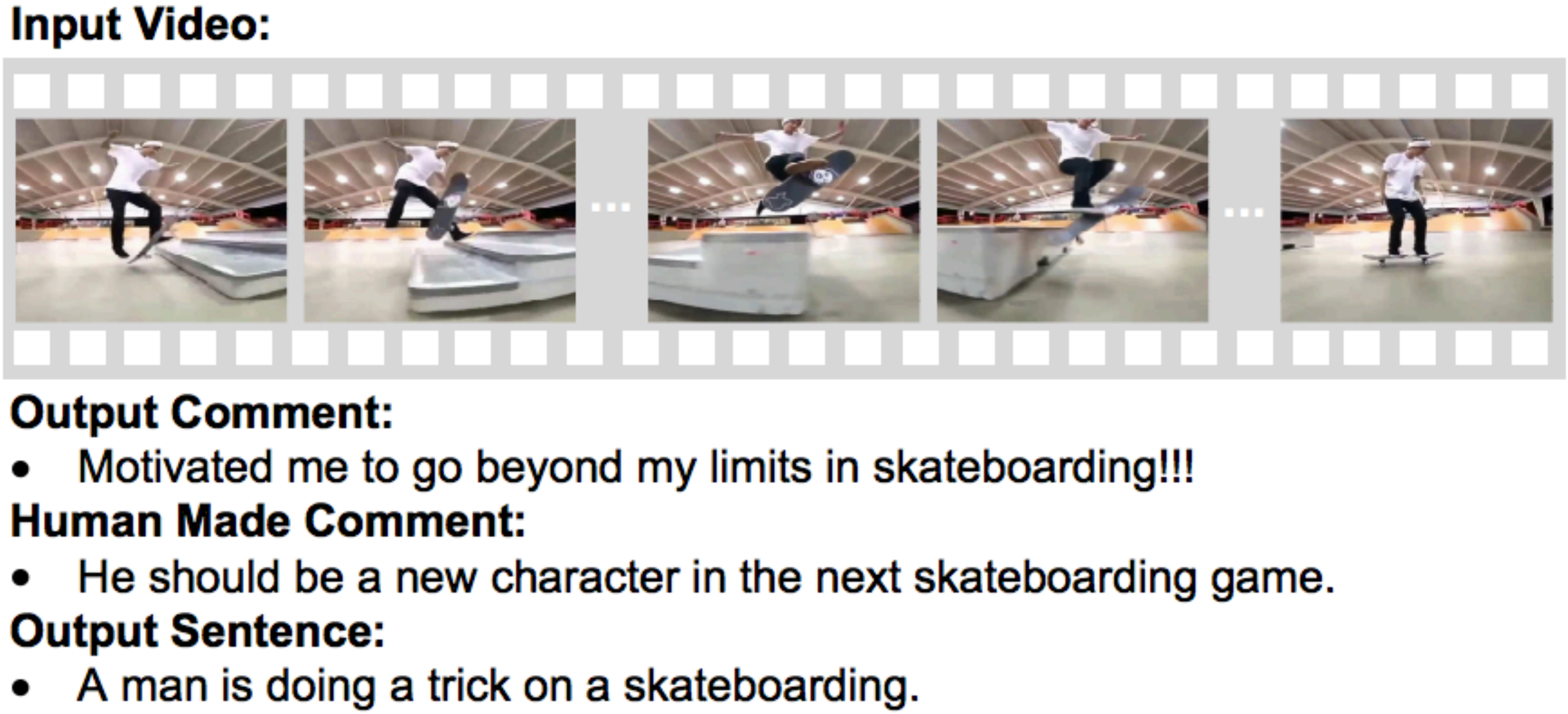
Now the discus, soccer ball, and volleyball are all stuck on the roof.

Timeline:

Yu Liu, Jianlong Fu, Tao Mei, Chang Wen Chen. Storytelling of Photo Stream with Bidirectional Multi-thread Recurrent Neural Network. Arxiv 2016

Video-Commenting

- **Deep Multi-View Embedding Model**



Input Video:

Output Comment:

- Motivated me to go beyond my limits in skateboarding!!!

Human Made Comment:

- He should be a new character in the next skateboarding game.

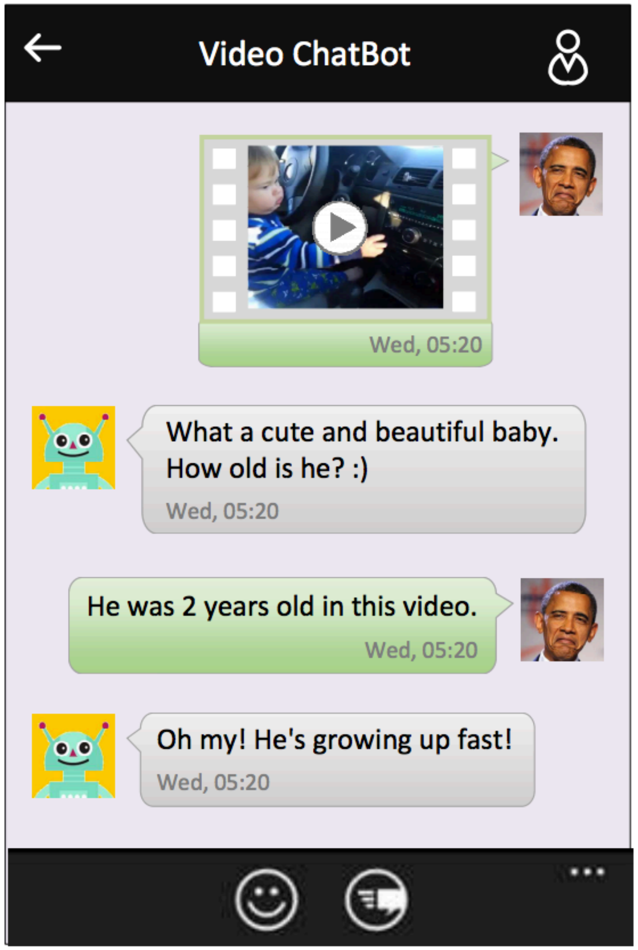
Output Sentence:

- A man is doing a trick on a skateboarding.

Yehao Li, Ting Yao, Tao Mei, Hongyang Chao, Yong Rui, "Share-and-Chat: Achieving Human-Level Video Commenting by Search and Multi-View Embedding," ACM Multimedia (MM), 2016

Video-Chatbot

- human-level emotional comments



Wed, 05:20

What a cute and beautiful baby. How old is he? :)

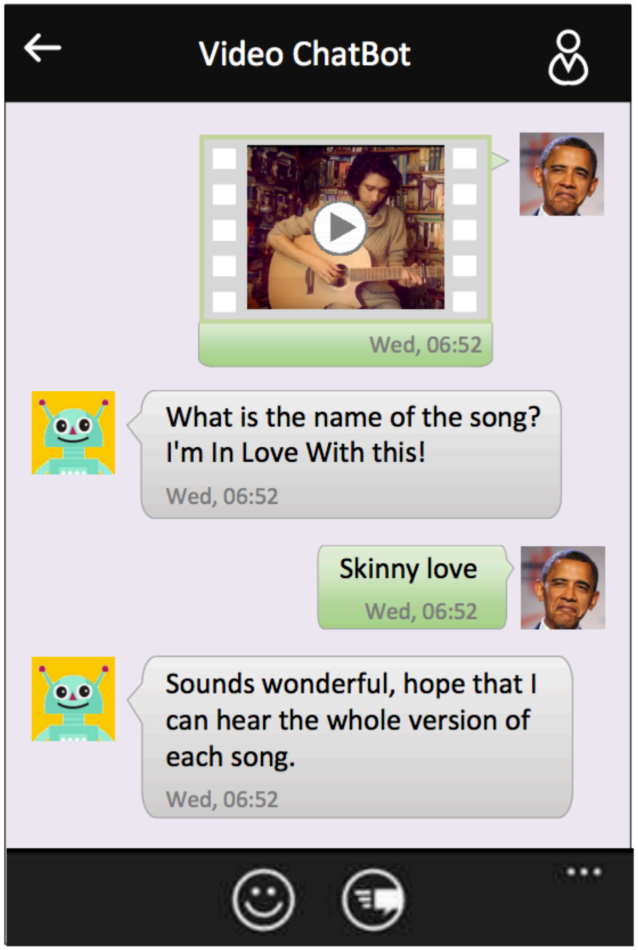
Wed, 05:20

He was 2 years old in this video.

Wed, 05:20

Oh my! He's growing up fast!

Wed, 05:20



Wed, 06:52

What is the name of the song? I'm In Love With this!

Wed, 06:52

Skinny love

Wed, 06:52

Sounds wonderful, hope that I can hear the whole version of each song.

Wed, 06:52

Yehao Li, Ting Yao, Rui Hu, Tao Mei, Yong Rui. Video ChatBot: Triggering Live Social Interactions by Automatic Video Commenting. ACM Multimedia (MM), 2016

Future

- Vision Guided Language-based HCI
  - Chatbot
  - Smart Assistant
- Storytelling
  - Public events (e.g., newspaper)
  - Personal events (e.g., personal blog post)

Thanks!