

Video and Language

Bridging Video and Language with Deep Learning

Tao Mei

Senior Researcher, Microsoft Research Asia

<http://research.microsoft.com/en-us/people/tmei/>

ECCV-MM Invited Tutorial, 3:30-4:30PM, Oct 15, 2016

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)

Computer vision: 50 years of progress

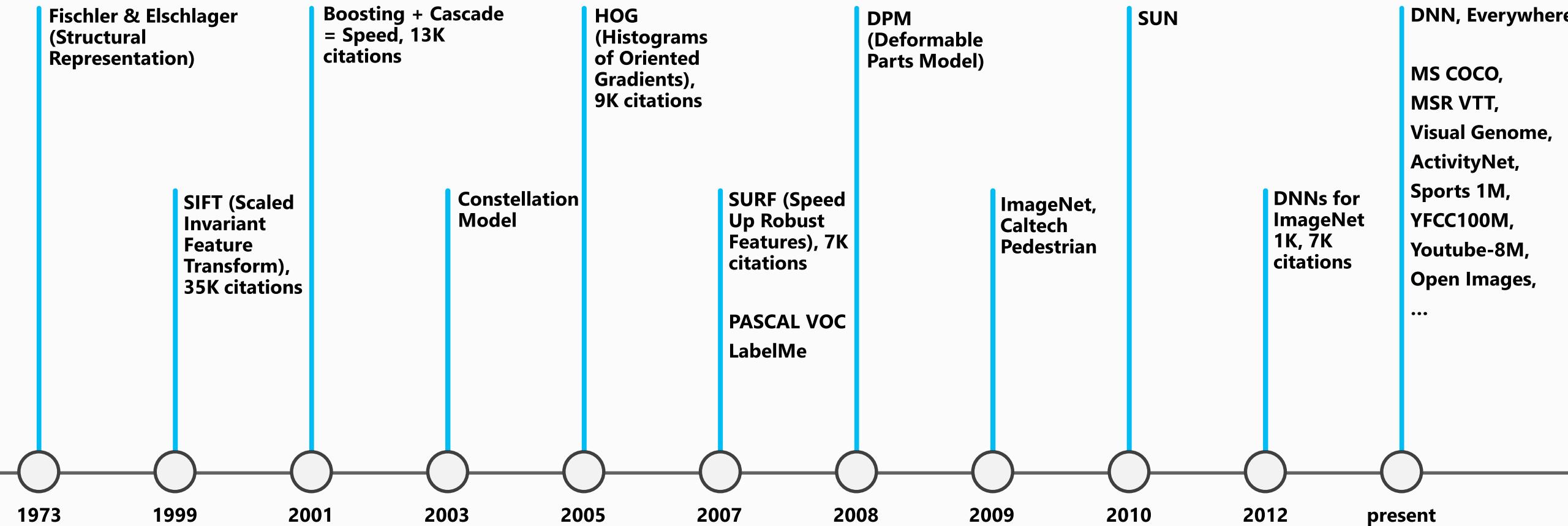
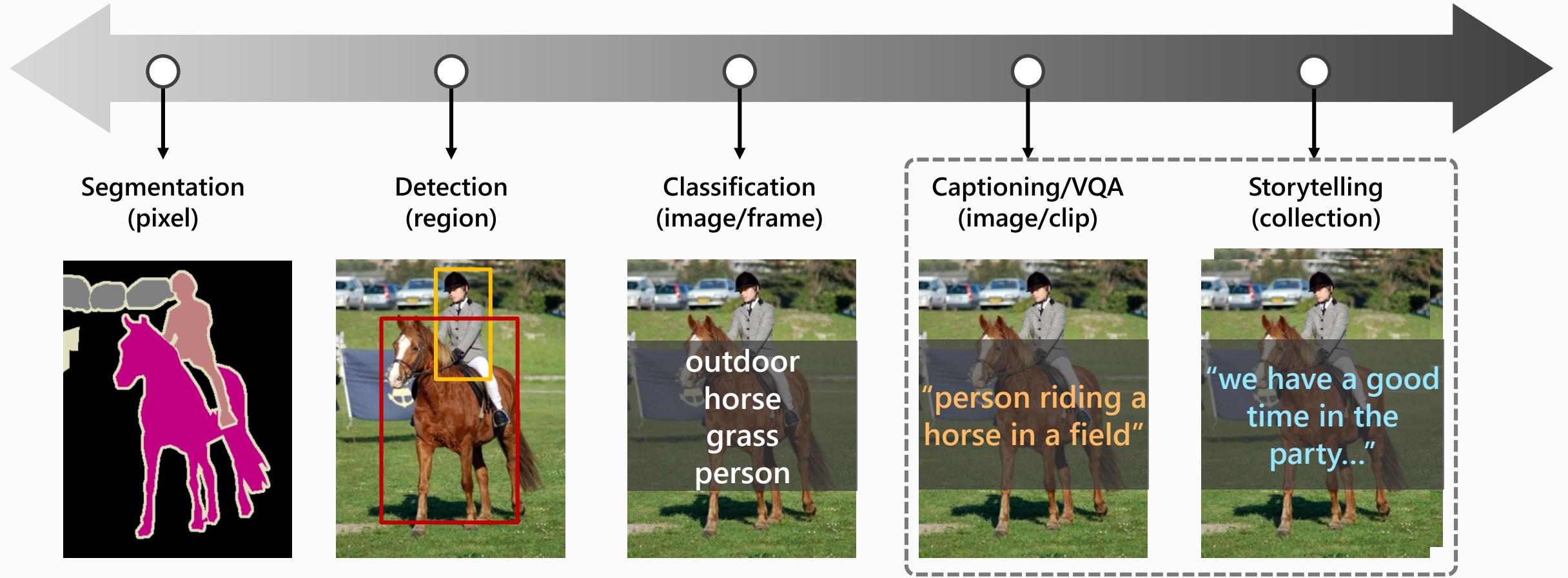


Image and video understanding: core problems



Vision and Language

Deep learning to

“describe what a 3-year-old child sees”

- Image/video recognition: classification, detection, segmentation



“describe what a 5-year-old child sees”

- Vision to language
 - Image captioning
 - Video captioning & commenting
- Visual question-answering



Image Captioning



"I think it's a boat is docked in front of a building."
<https://www.captionbot.ai/> [Microsoft CaptionBot]



"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with *Forbidden City* in the background." [Xiaodong He, 2016]

Video Captioning



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

Video Commenting

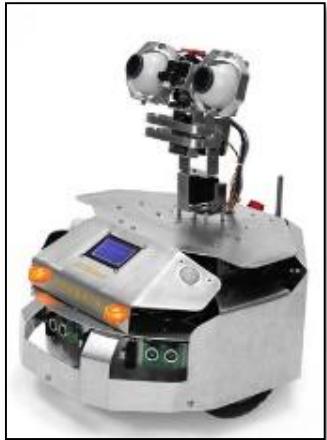


"I love baseball"
"That's how to play baseball"
"That's an amazing play"
[Li, Yao, Mei, MM'16]



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

Vision to Language



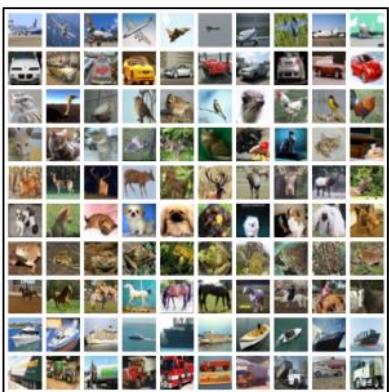
robotic vision



assist for blinded



incident report for surveillance



multimedia search

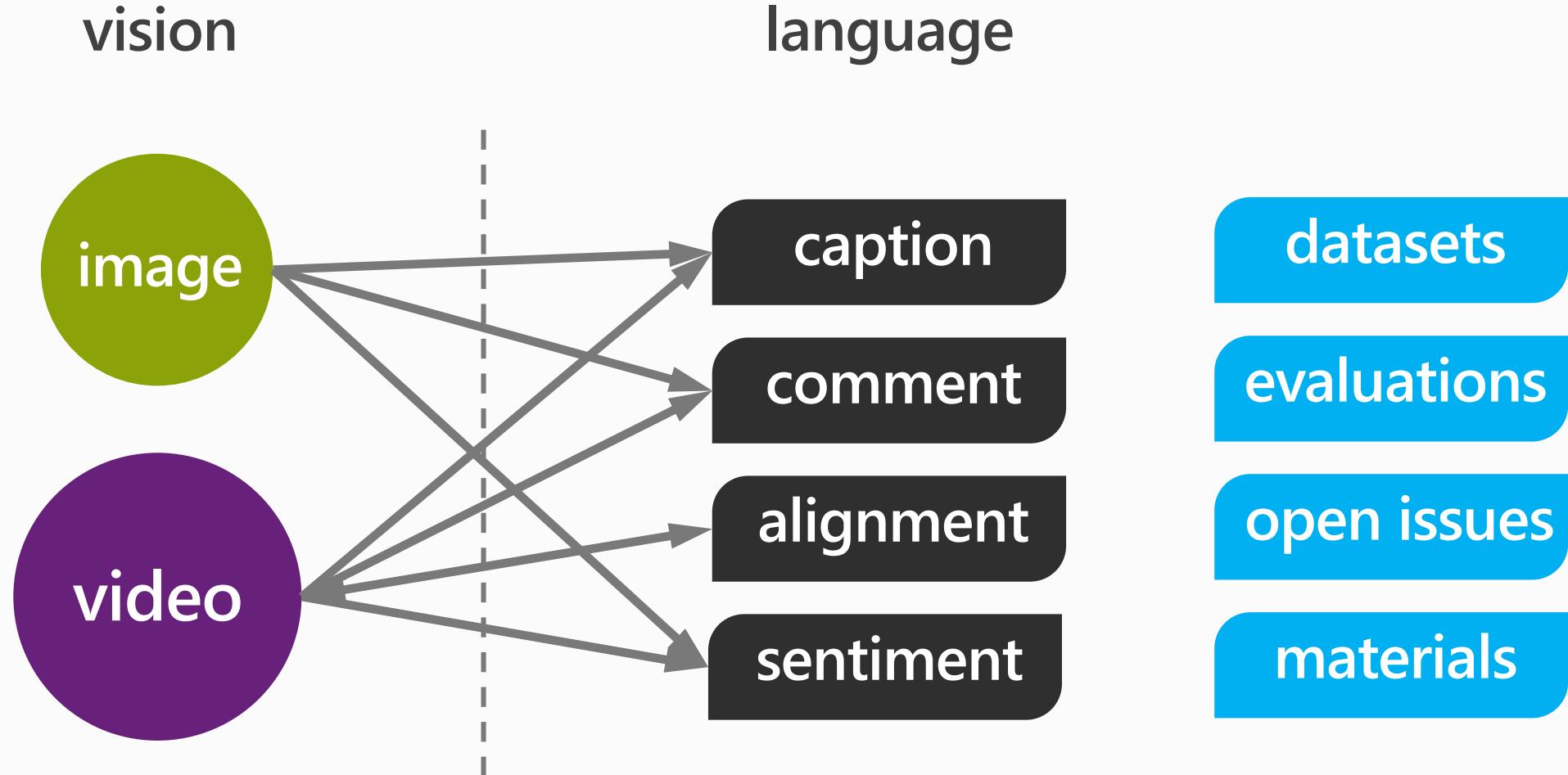


movie description for blinded



seeing chat bot

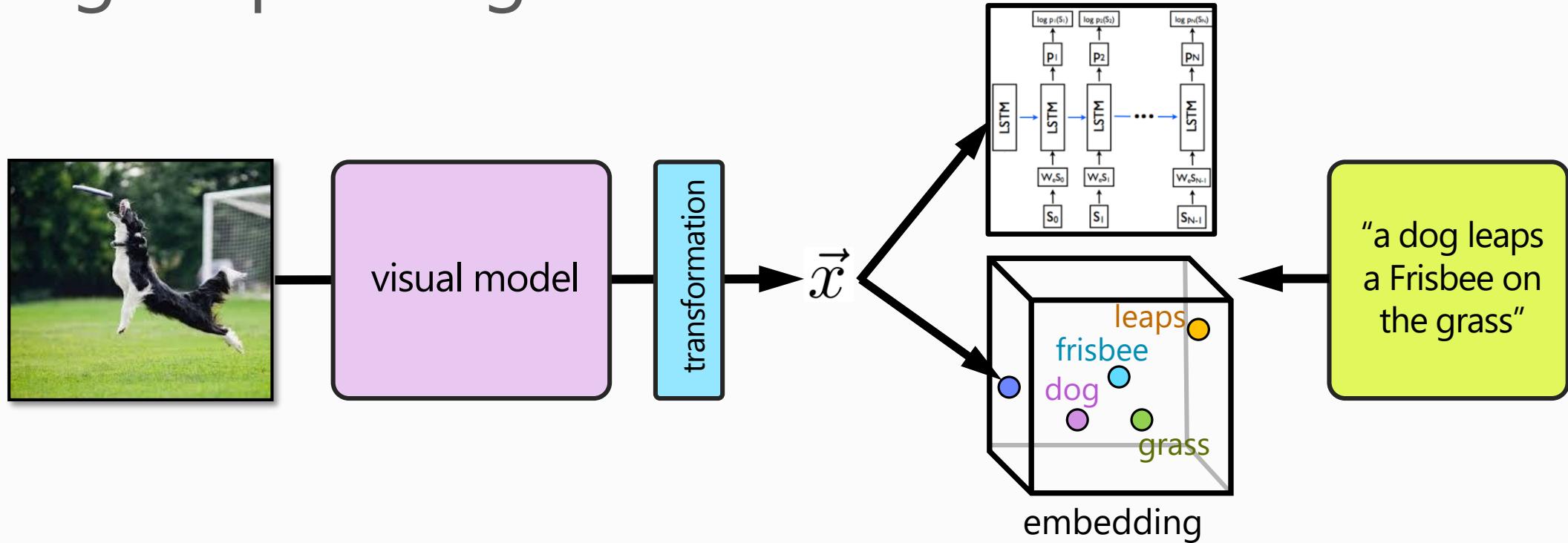
This tutorial will talk about



Outline

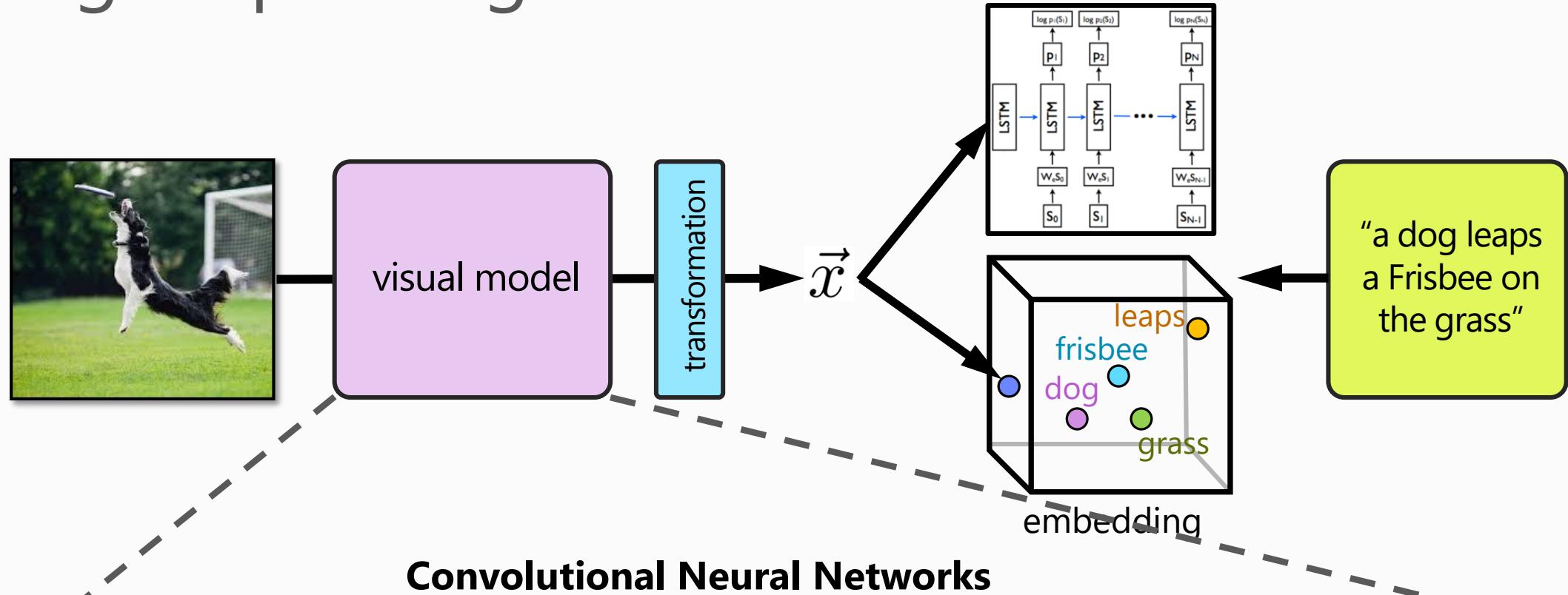
- Image and video captioning
 - caption = object localization/recognition + object relationship + language
 - nouns (objects, people, scenes)
 - adjectives (attributes)
 - verbs (actions)
 - prepositions (relationships)
- Video commenting
- Video and language alignment
- Datasets and evaluations
- Open issues
- Learning materials

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

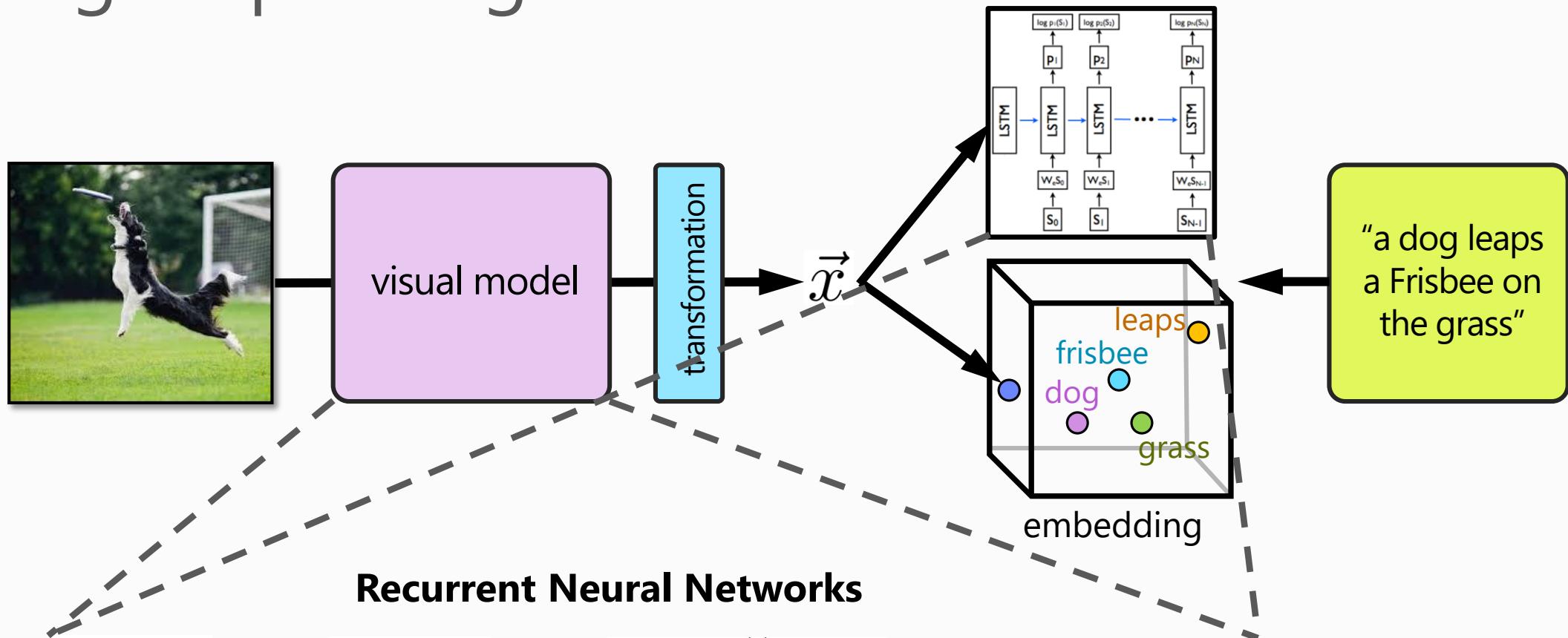
Image captioning: basic idea



Convolutional Neural Networks



Image captioning: basic idea



Recurrent Neural Networks

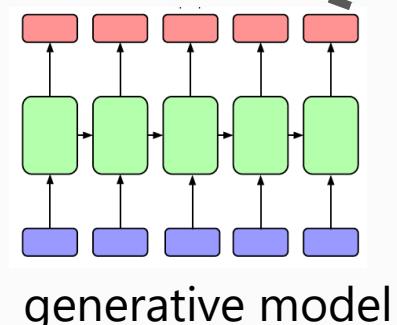
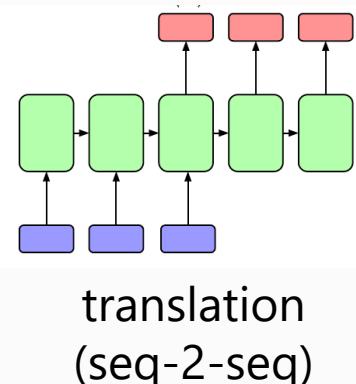
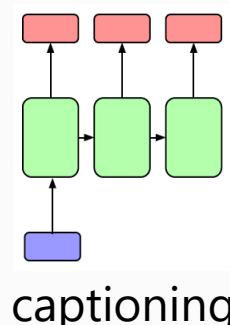
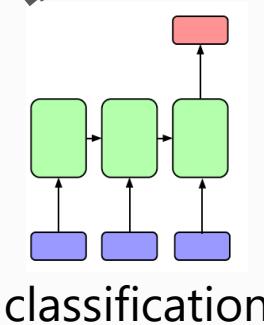


Image captioning

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

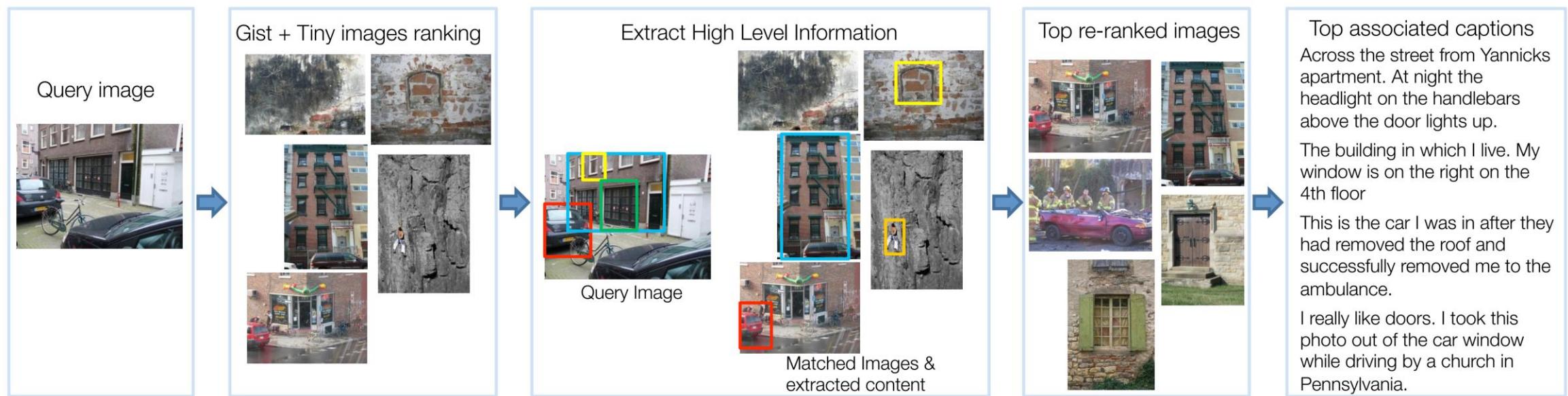


Image captioning

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

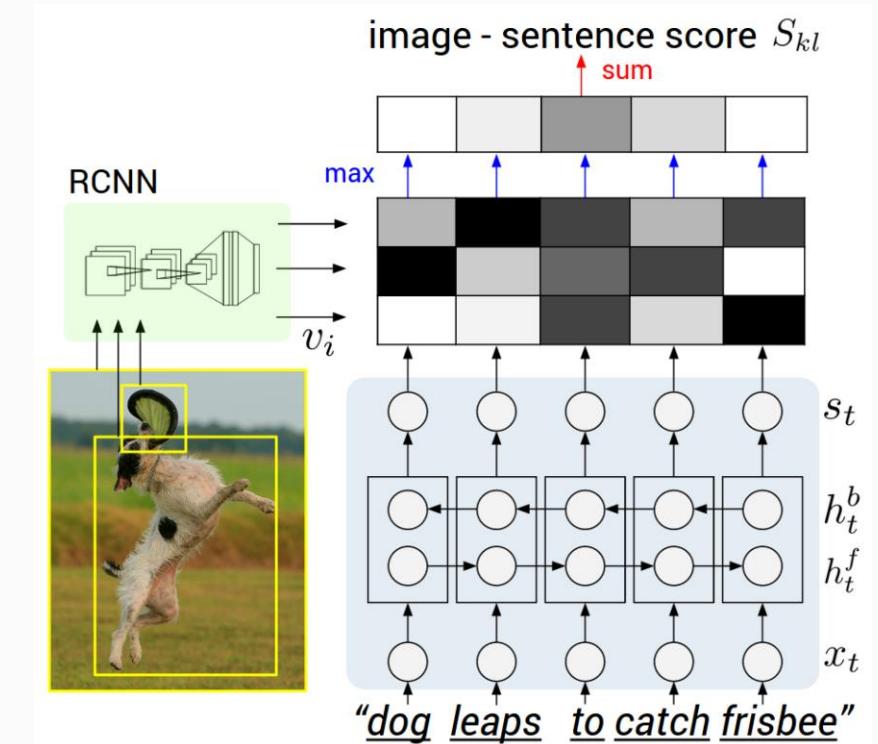
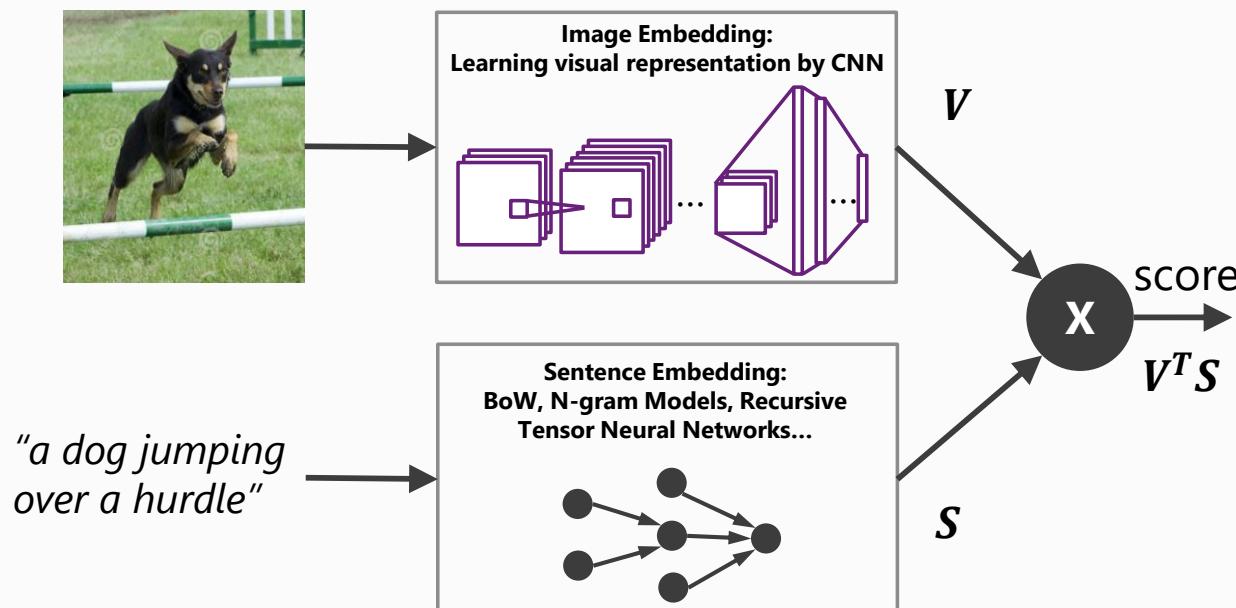


Image captioning

- 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.

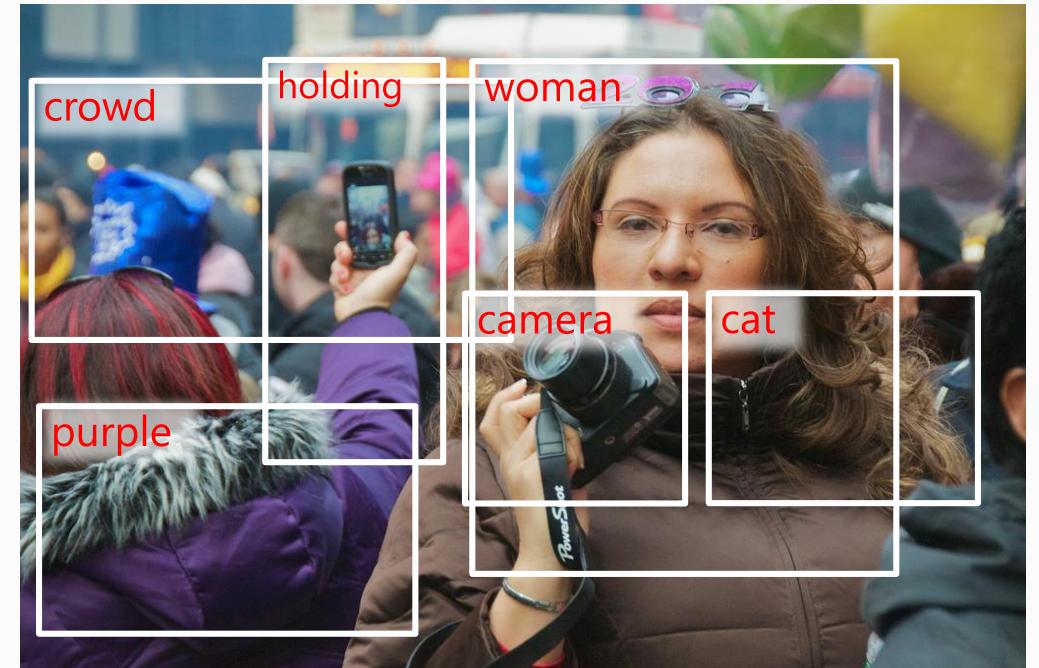
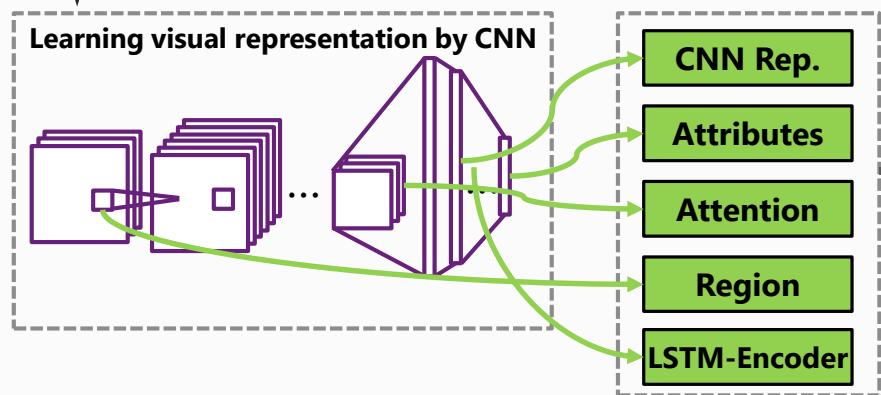


Image captioning

- Sequence learning-based approach

[Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester]

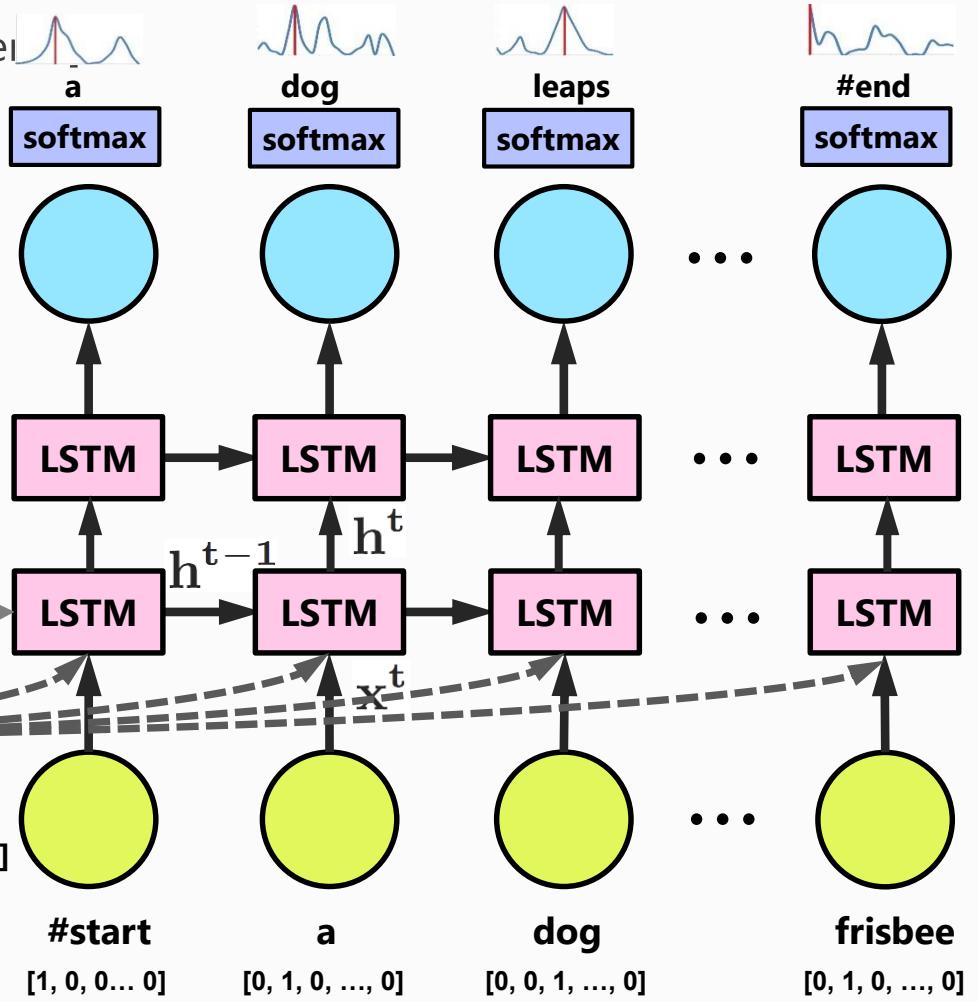


[Vinyals, CVPR15; Chen, CVPR15; Mao, ICLR15]

[Wu, CVPR16; Pan, 2016]

[Xu, ICML15; You, CVPR16]

[Karpathy & Fei-Fei, C]



* Note that this figure only shows prediction process.

Image Captioning with X

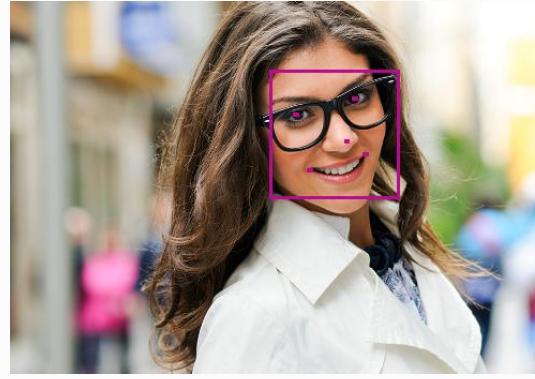
A(1.00)



X = visual attention
[Xu, ICML'15]



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



X = entity recognition
[Tran, CVPR'16]



X = dense caption
[Johnson, CVPR'16]

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.

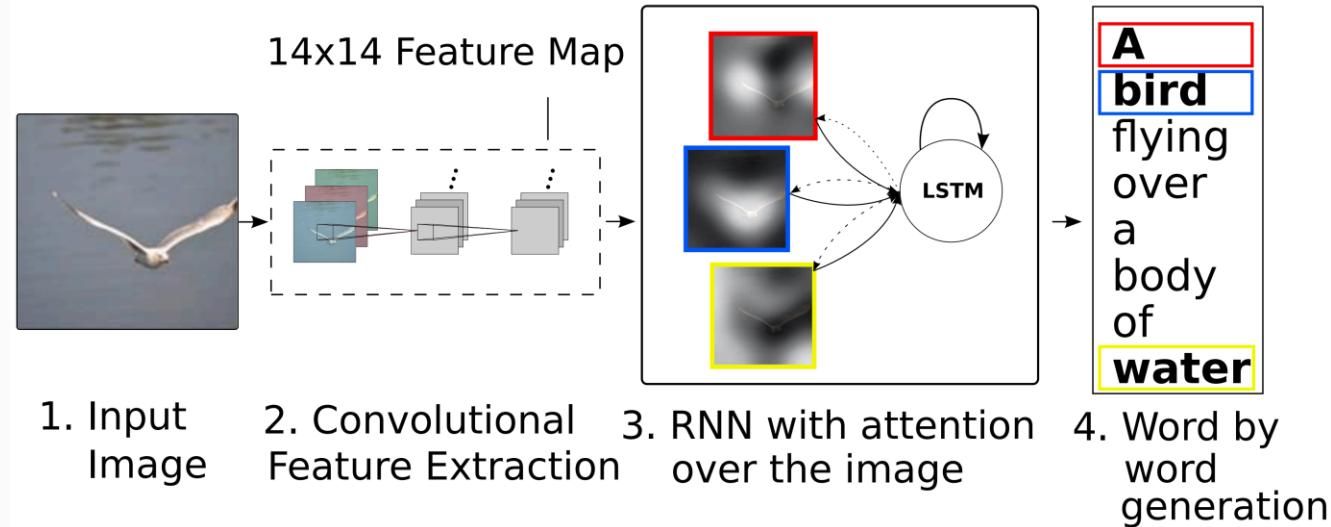
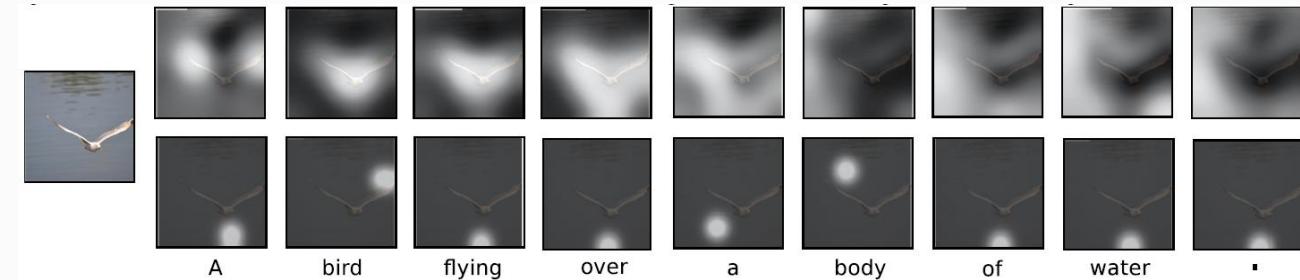


Image Captioning with Visual Attributes

- 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
guitar

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

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

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

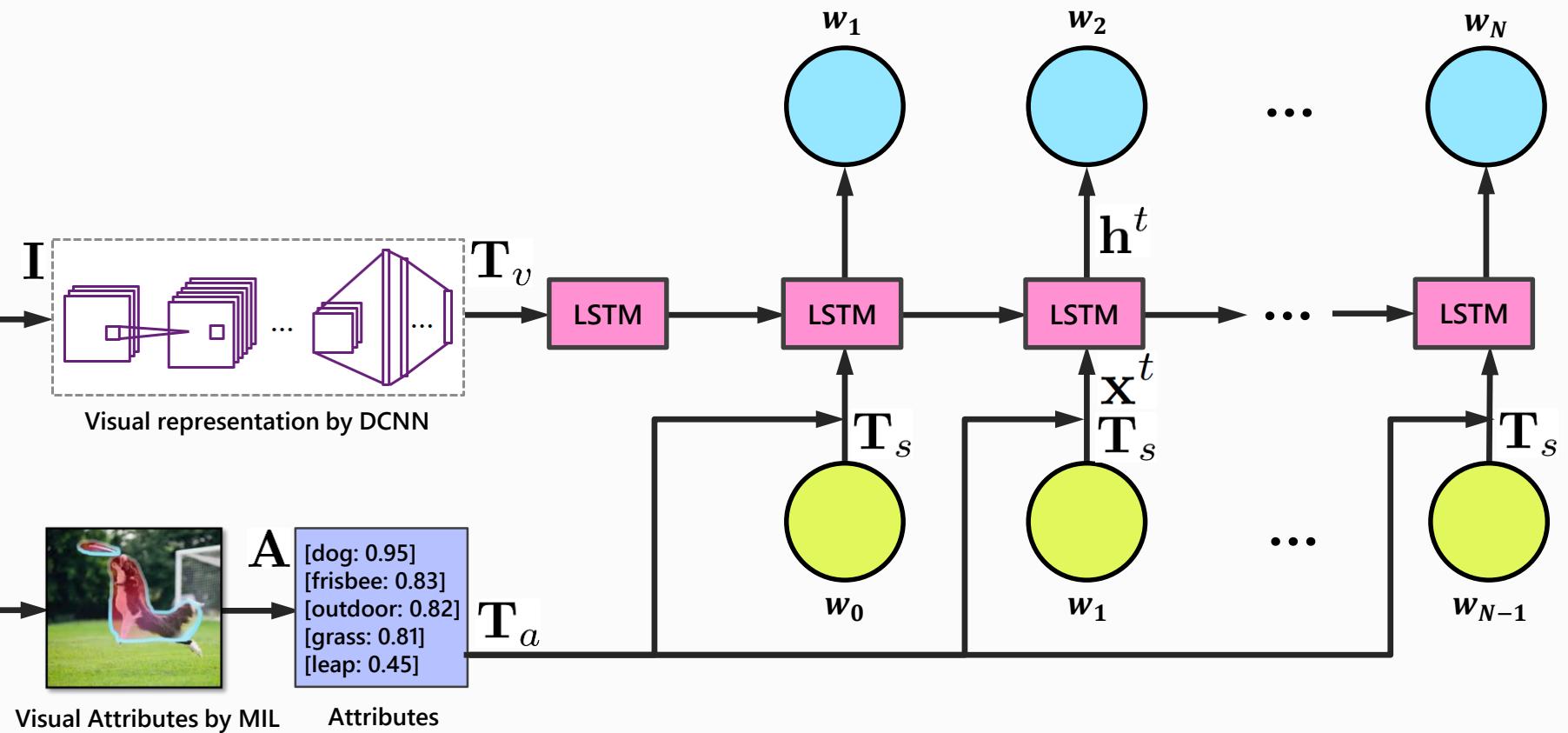
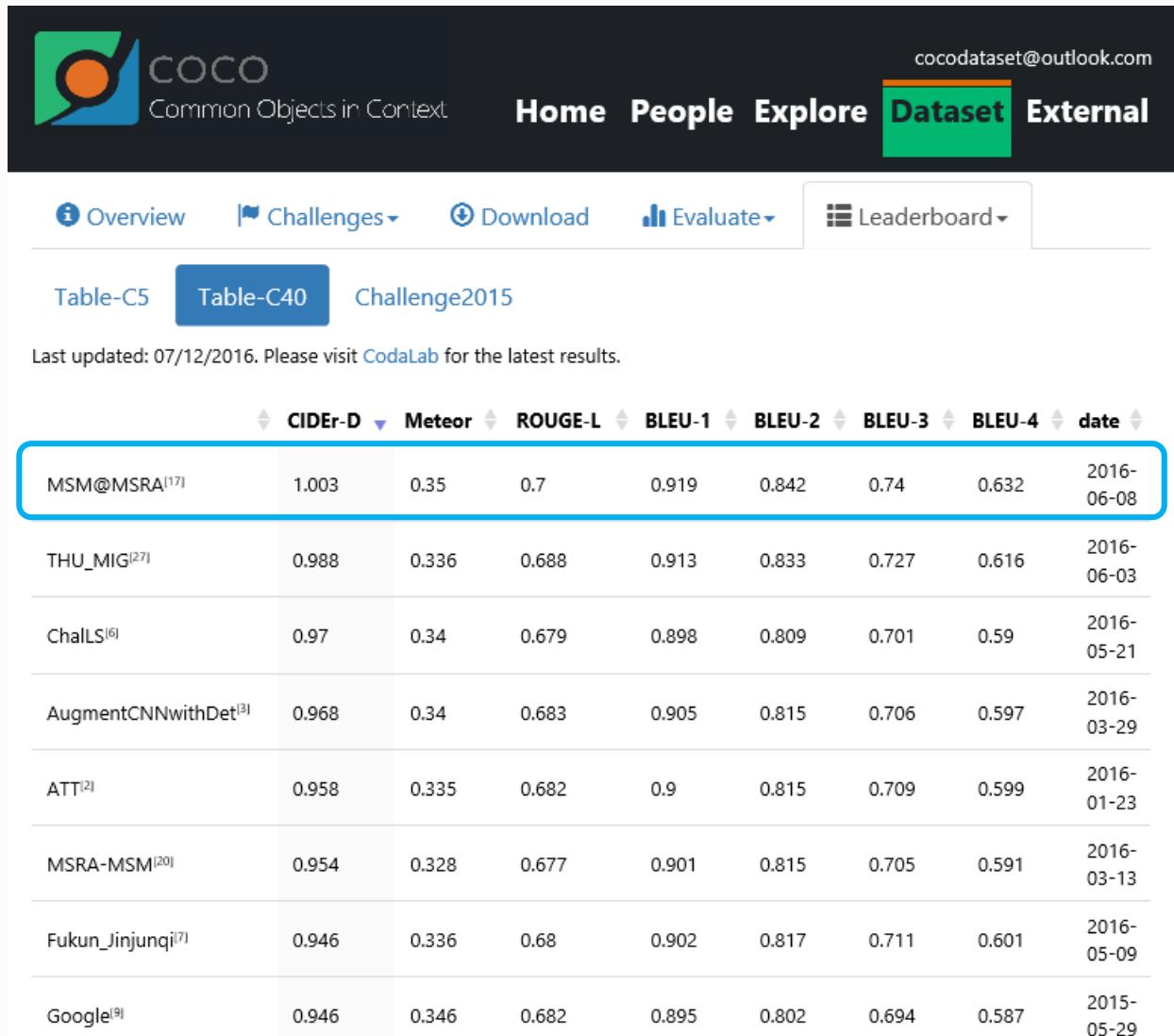


Image captioning

- 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)



The screenshot shows the COCO dataset homepage with the navigation bar: Home, People, Explore, Dataset (highlighted in green), and External. Below the navigation is a menu with Overview, Challenges, Download, Evaluate, and Leaderboard. The Leaderboard section is selected, showing results for Table-C40. The table includes columns for CIDEr-D, Meteor, ROUGE-L, BLEU-1, BLEU-2, BLEU-3, BLEU-4, and date. The top row, 'MSM@MSRA^[17]', is highlighted with a blue border and shows the highest scores across all metrics.

	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



Attributes

[boat: 1]
[water: 0.92]
[river: 0.645]
[small: 0.606]
[dog: 0.555]
[body: 0.527]
[floating: 0.484]

Generated Sentences

LSTM: a group of people on a boat in the water.

CaptionBot: I think it's a man with a small boat in a body of water.

A-LSTM: a man and a dog on a boat in the water.

Ground Truth

- ① an image of a man in a boat with a dog
- ② a person on a rowboat with a dalmatian dog on the boat
- ③ old woman rowing a boat with a dog



Attributes

[bananas: 1]
[market: 0.995]
[outdoor: 0.617]
[bunch: 0.553]
[table: 0.51]
[flowers: 0.454]
[people: 0.431]
[yellow: 0.377]

Generated Sentences

LSTM: a group of people standing around a market.

CaptionBot: I think it's a bunch of yellow flowers.

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

Ground Truth

- ① bunches of bananas for sale at an outdoor market
- ② a person at a table filled with bananas
- ③ there are many bananas layer across this table at a farmers market



Attributes

[flying: 0.877]
[plane: 0.598]
[airplane: 0.528]
[lake: 0.495]
[water: 0.462]
[sky: 0.443]
[red: 0.426]
[small: 0.365]

Generated Sentences

LSTM: a group of people flying kites in the sky.

CaptionBot: I think it's a plane is flying over the water.

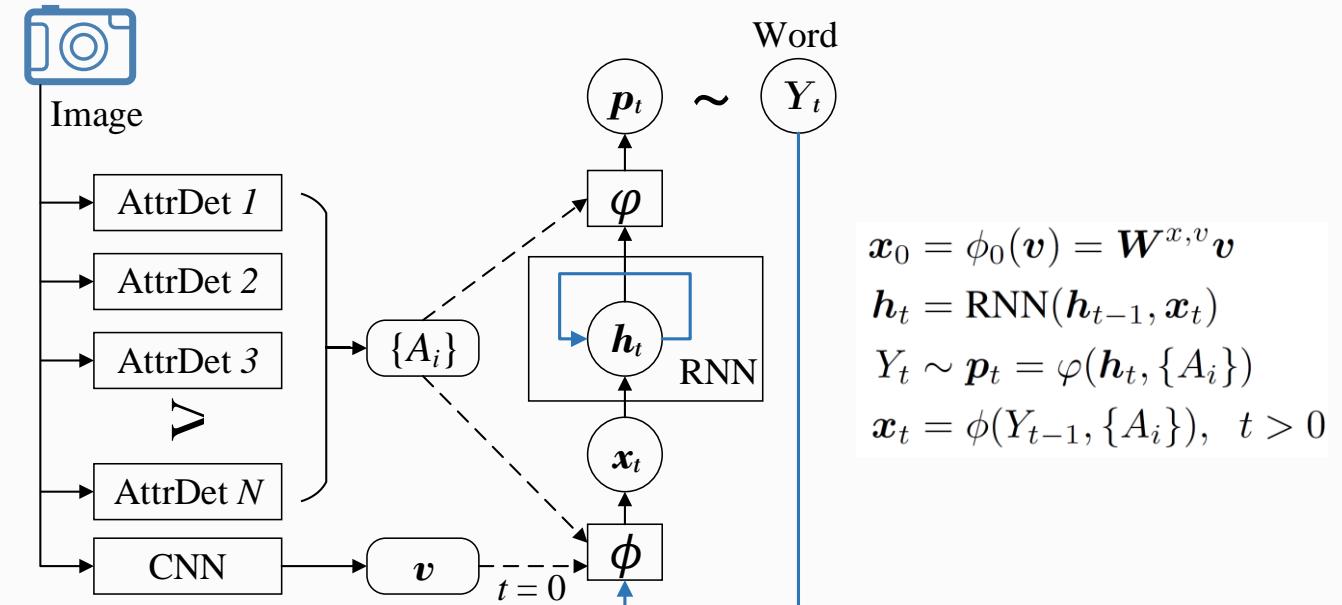
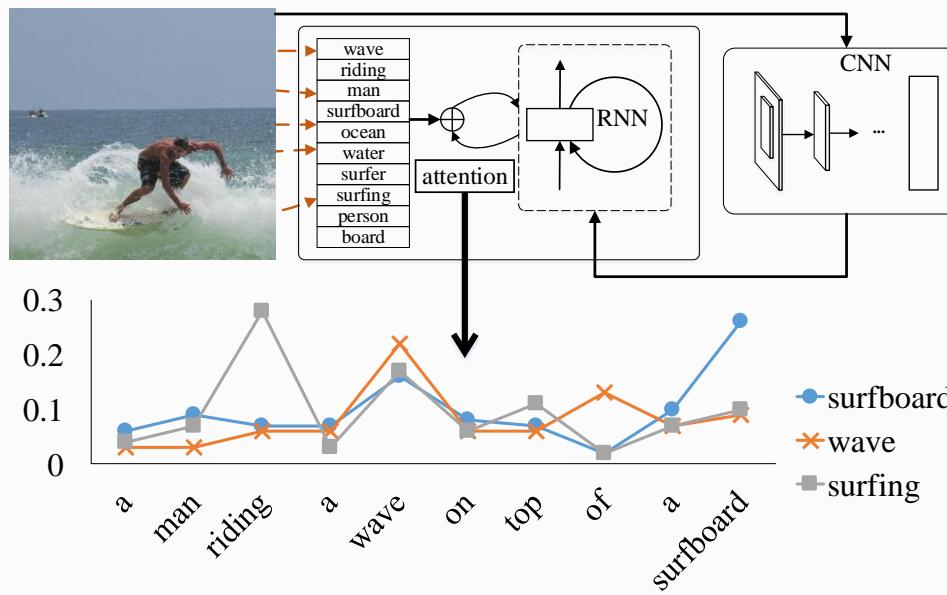
A-LSTM: a red and white plane flying over a body of water.

Ground Truth

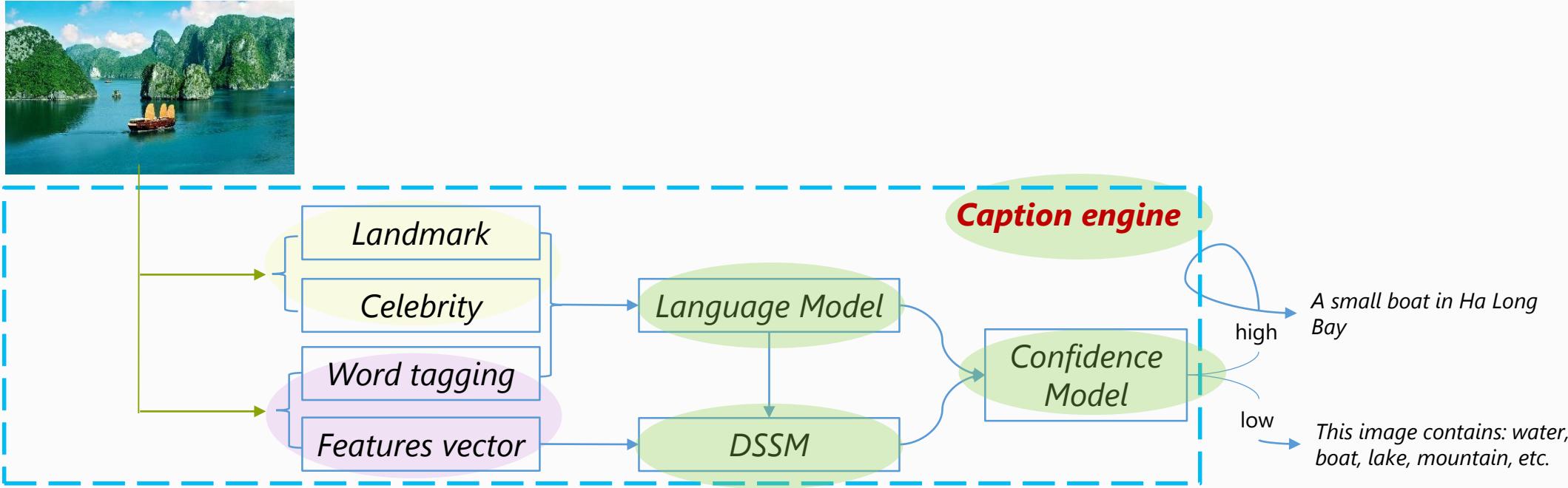
- ① a plane with water skies for landing gear coming in for a landing at a lake
- ② a plane flying through a sky above a lake
- ③ a red and white plane is flying over some water

Image Captioning with Semantic Attention (Attributes)

- Instead of using the same set of attributes at every step, select attributes at each step. [You, CVPR'16]



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]

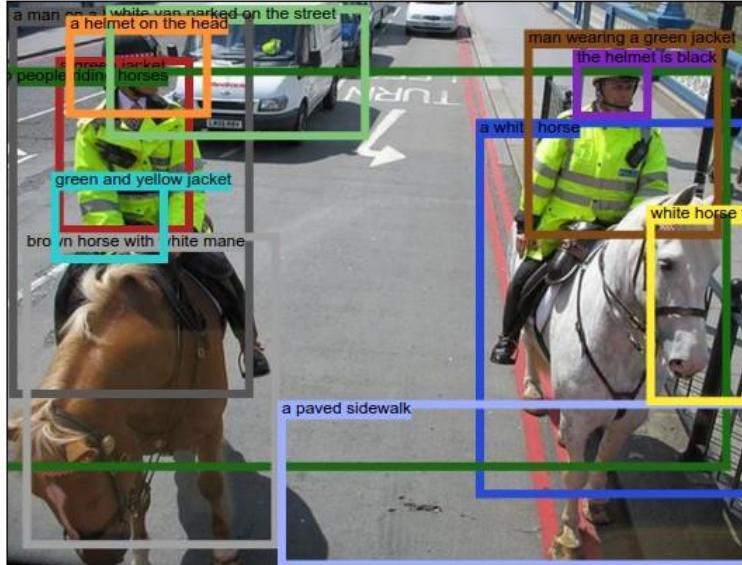


"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with **Forbidden City** in the background." [Xiaodong He, 2016]

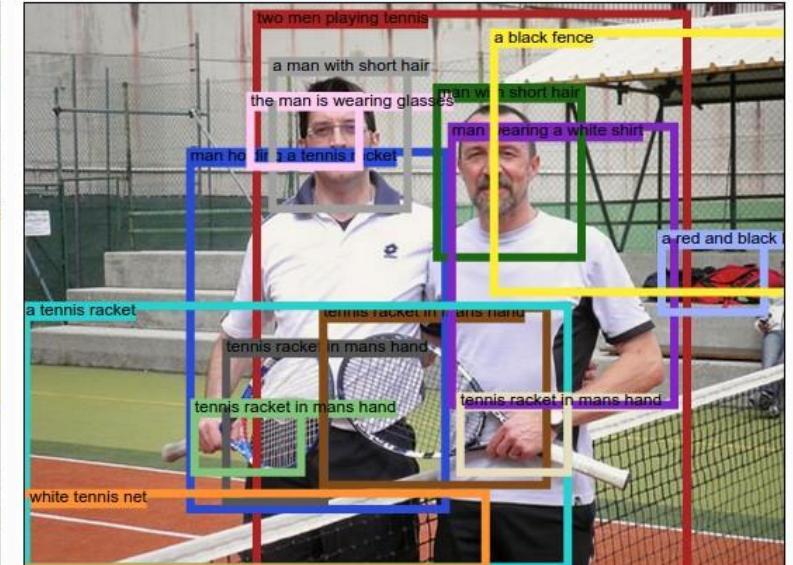
Dense Image Captioning [Johnson & Karpathy, CVPR16]



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.

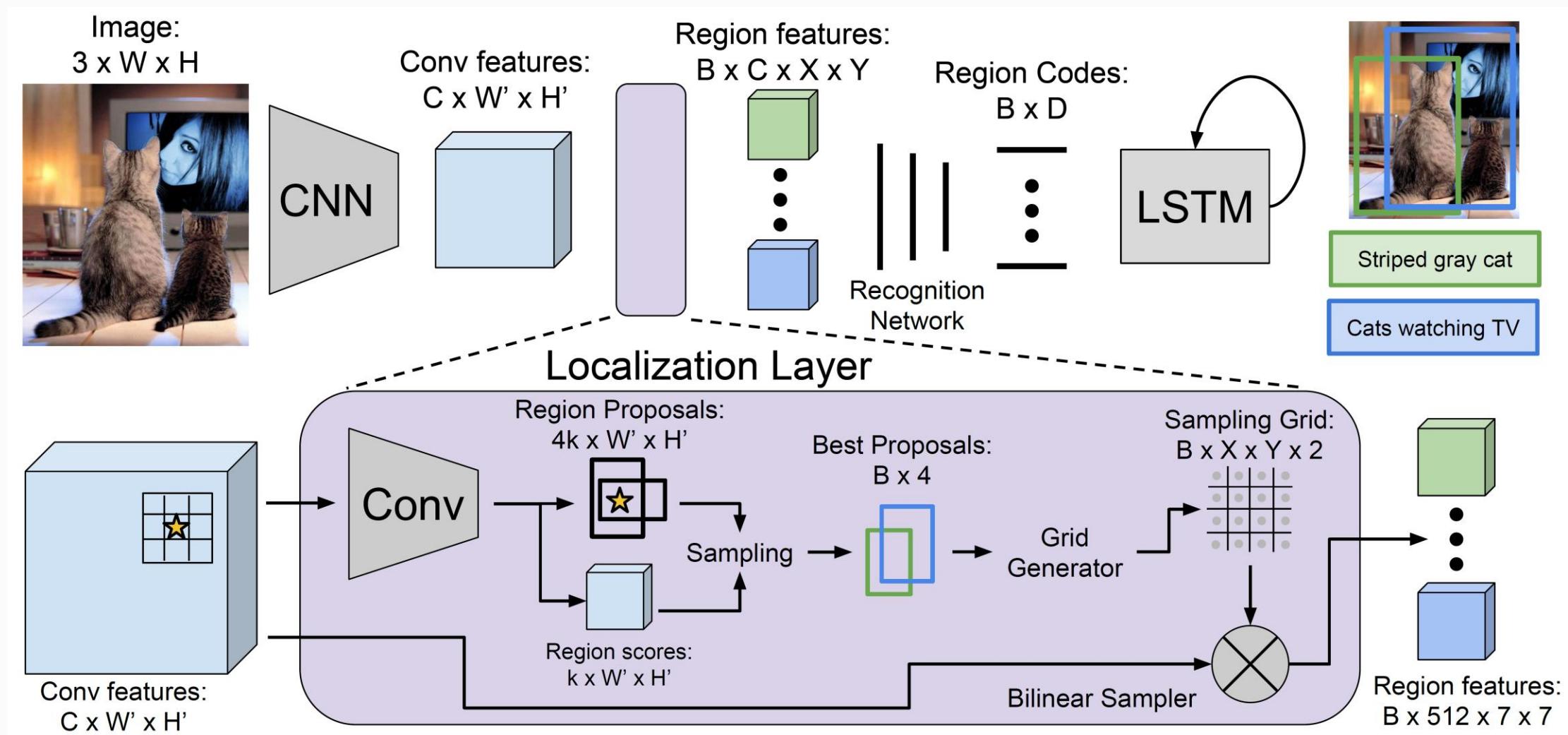


a green jacket. a white horse. a man on a horse. two people riding horses. man wearing a green jacket. the helmet is black. brown horse with white mane. white van parked on the street. a paved sidewalk. green and yellow jacket. a helmet on the head. white horse with white face.



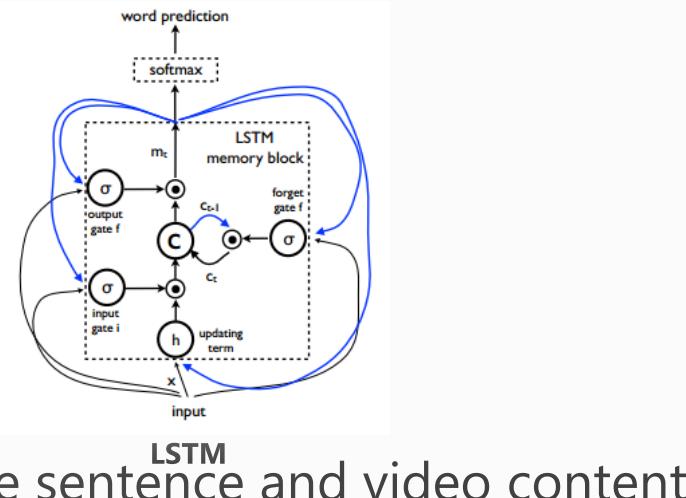
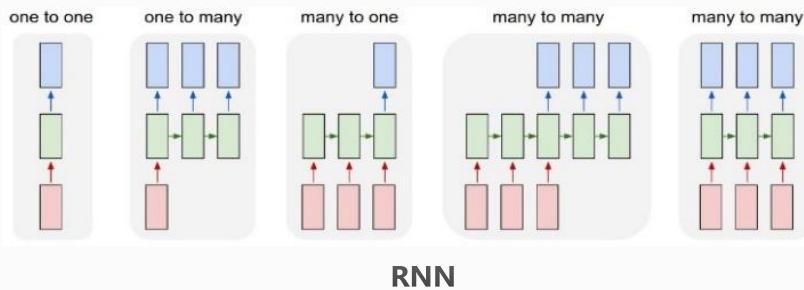
two men playing tennis. man holding a tennis racket. tennis racket in mans hand. man with short hair. tennis racket in mans hand. man wearing a white shirt. a man with short hair. tennis racket in mans hand. a red and black bag. a tennis racket. a white tennis net. a black fence. tennis racket in mans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

Dense Image Captioning [Johnson & Karpathy, CVPR16]

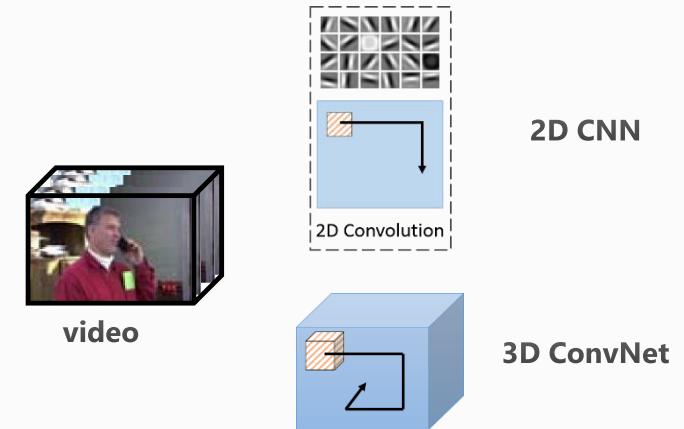


Challenges for video captioning

- 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



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



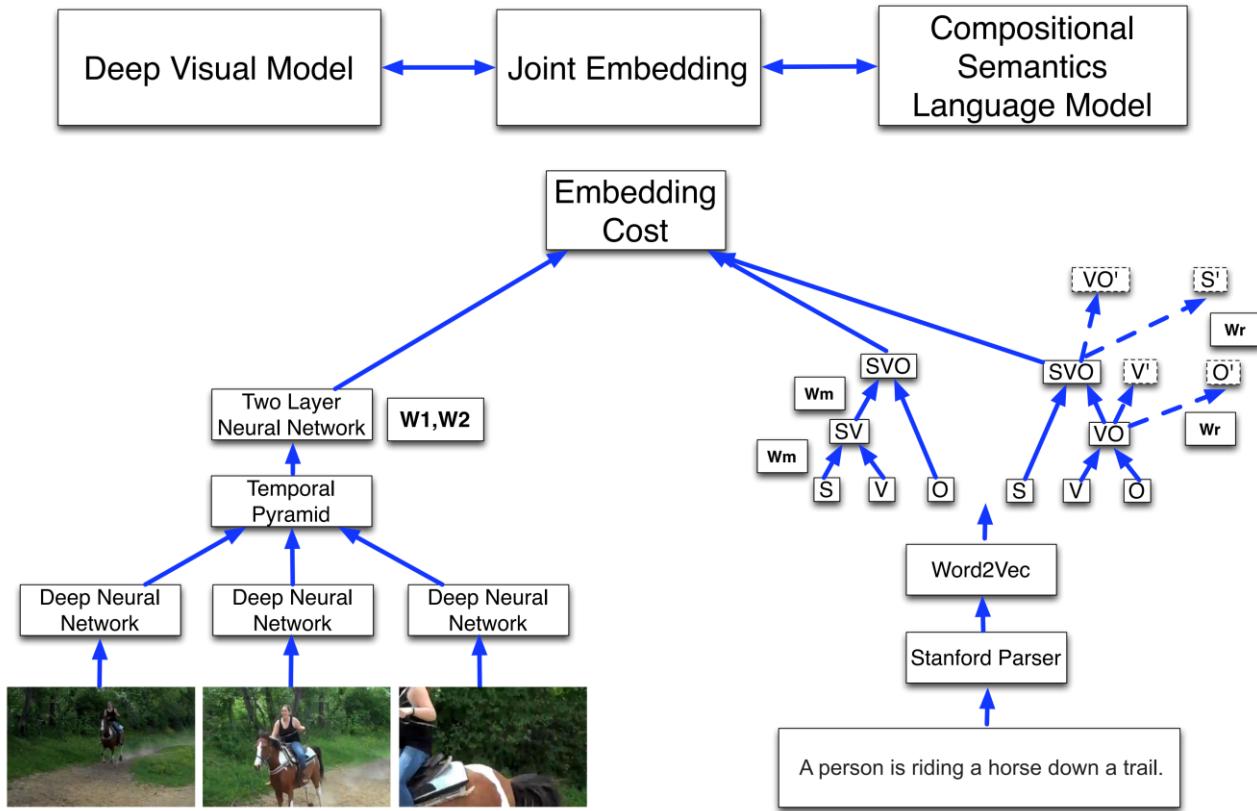
a man wearing a helmet is **racing**



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

Video captioning

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

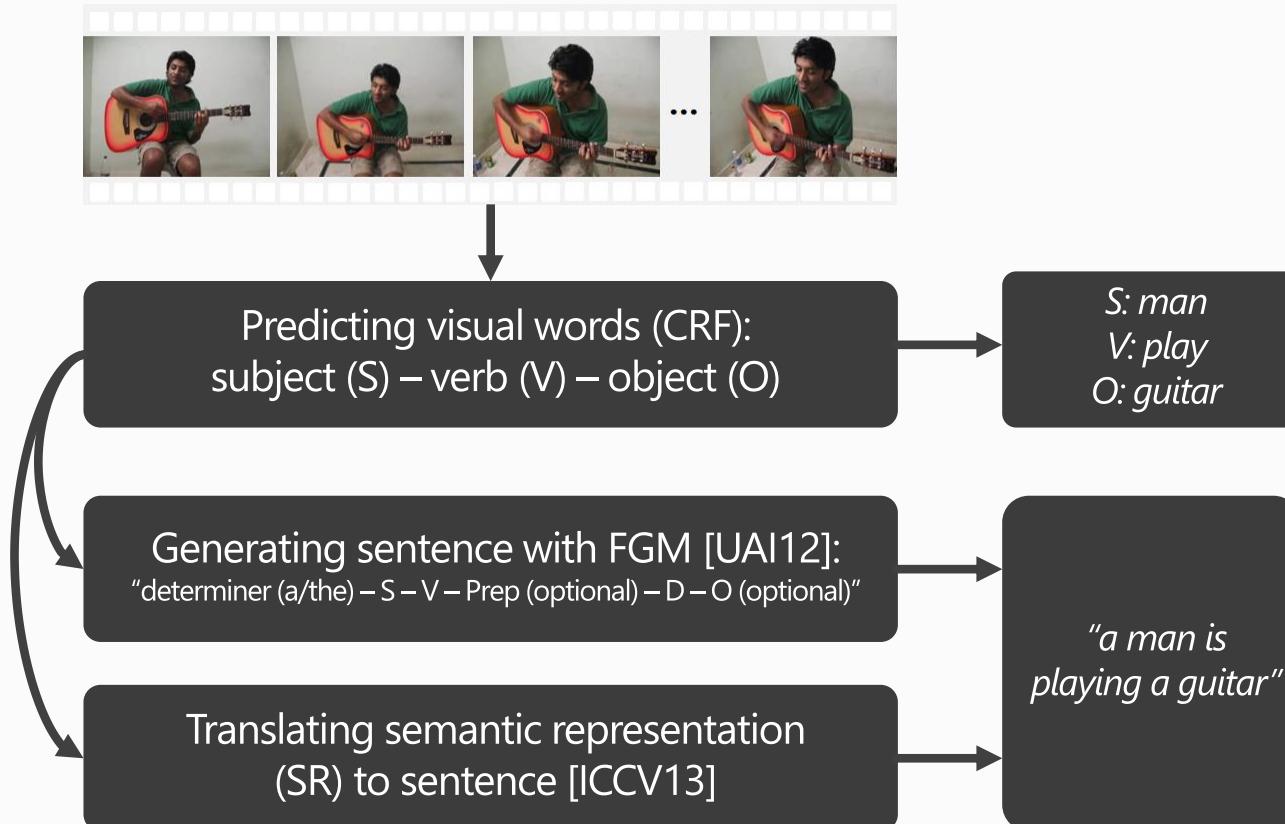


- Deep visual model to learn video representation
- Compositional language model to capture semantic compatibility among concepts
- Joint embedding model to minimize distance of the above two models in video-text space [Xu, AAAI15]

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

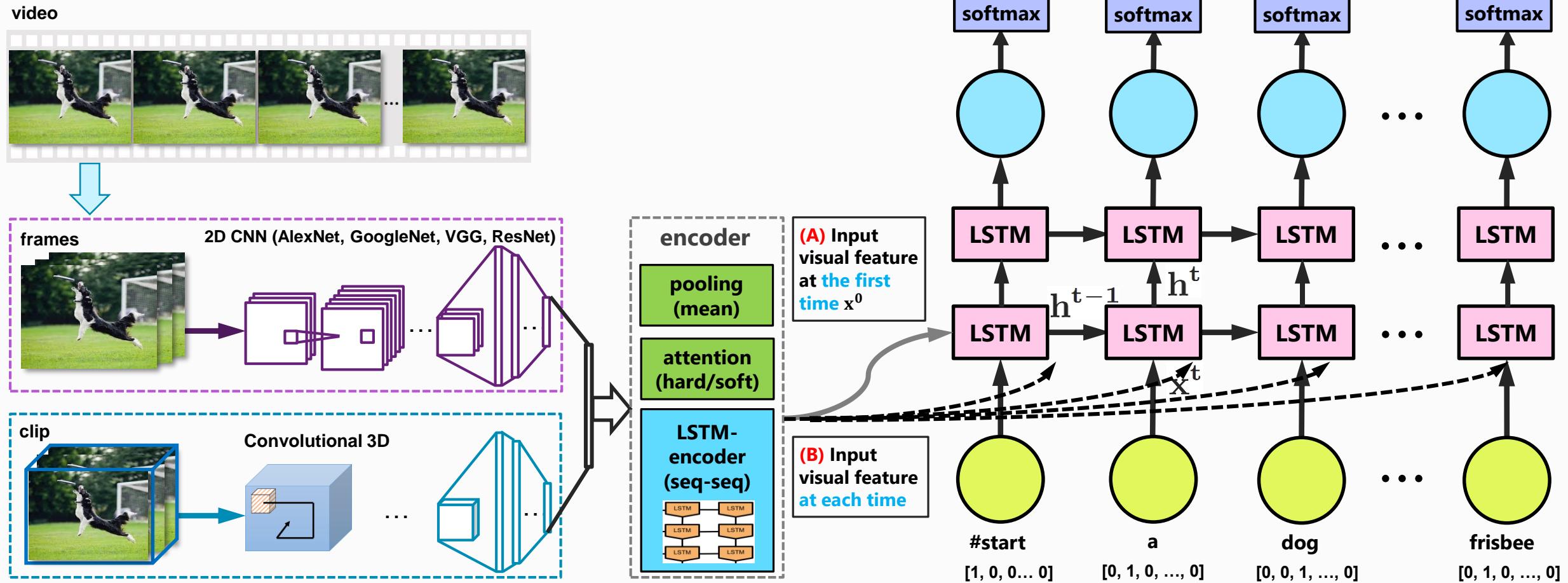
Video captioning

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



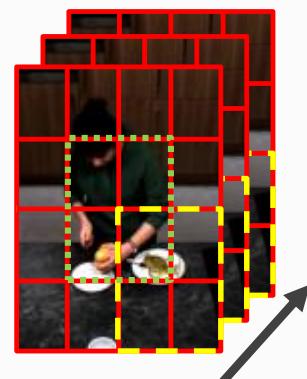
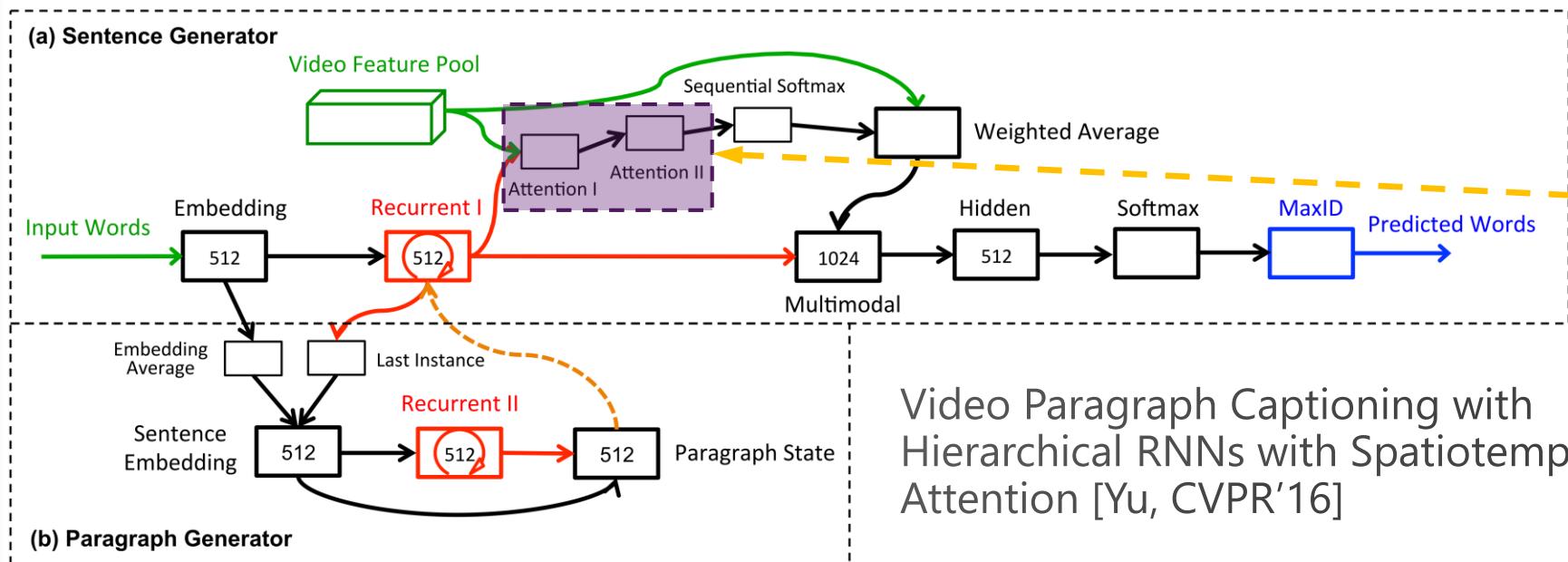
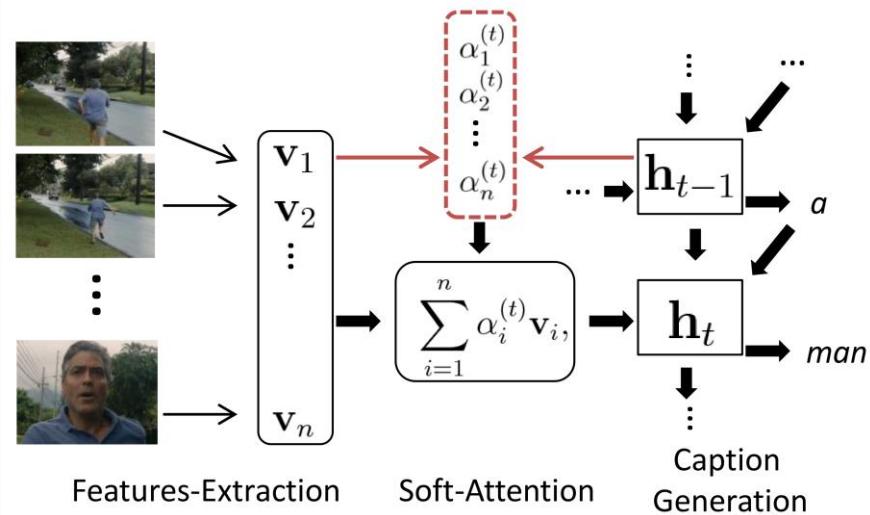
Barbu, et al. "Video In Sentences Out", UAI 2012.
<https://www.youtube.com/watch?v=tu3jMxCJPMw>

- 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)



Video Captioning with Attention

Encoder-decoder LSTM Networks with Temporal Attention [Yao, CVPR'15]



Video Captioning with Semantics

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



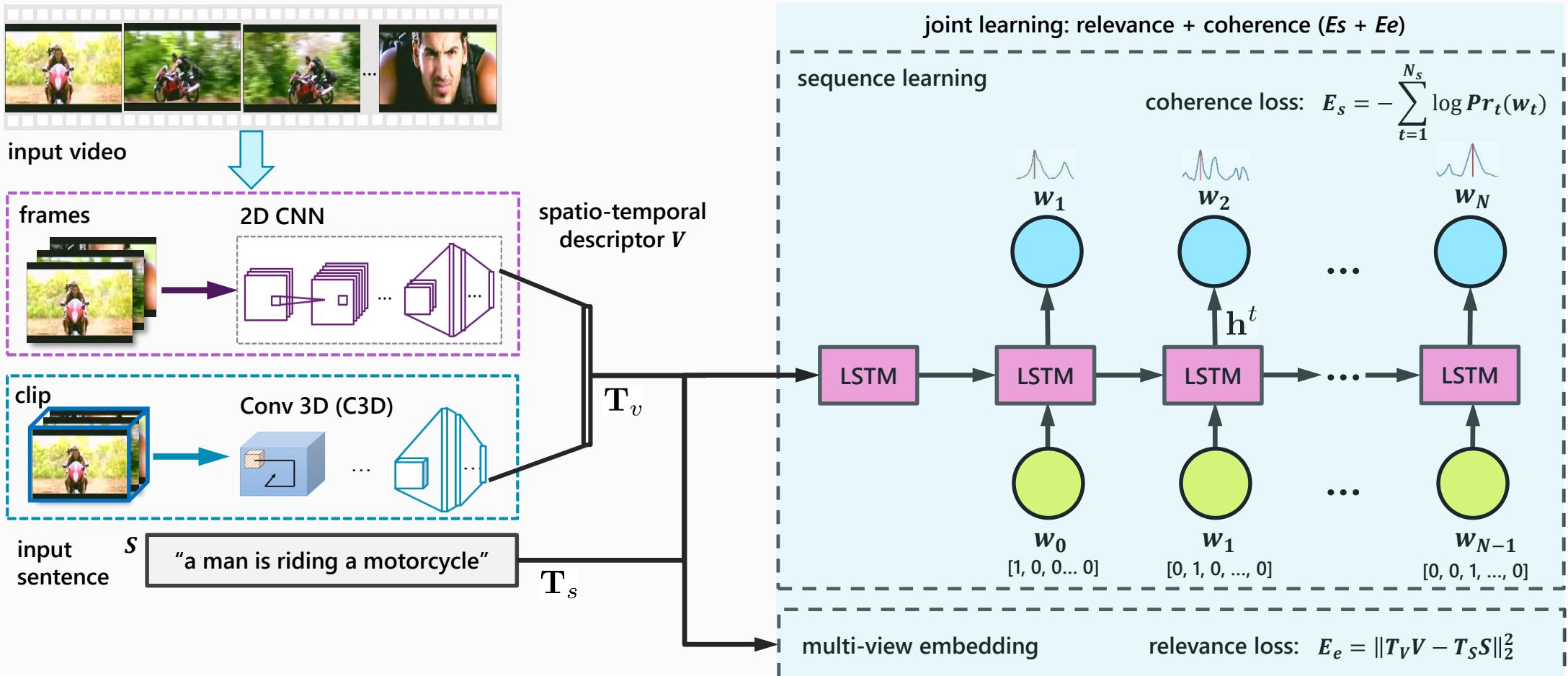
LSTM: a man is playing a **guitar**
LSTM-E: a man is playing a **piano**



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

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



$$E(\mathcal{V}, \mathcal{S}) = (1 - \lambda) \times \|\mathbf{T}_v \mathbf{v} - \mathbf{T}_s \mathbf{s}\|_2^2 - \lambda \times \sum_{t=0}^{N_s} \log \Pr(\mathbf{w}_t | \mathbf{v}, \mathbf{w}_0, \dots, \mathbf{w}_{t-1}; \theta; \mathbf{T}_v; \mathbf{T}_s)$$

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

Performance of video captioning [Sept 2016]

The accuracy of S-V-O triplet prediction.

Model	Team	Subject%	Verb%	Object%
FGM	UT Austin, COLING (2014/08)	76.42	21.34	12.39
CRF	SUNY-Buffalo, AAAI (2015/01)	77.16	22.54	9.25
CCA	Stanford, CVPR (2010/06)	77.16	21.04	10.99
JEM	SUNY-Buffalo, AAAI (2015/01)	78.25	24.45	11.95
LSTM	UC Berkeley, NAACL (2014/12)	71.19	19.40	9.70
LSTM-E	MSRA, arxiv (2015/05)	80.45	29.85	13.88

The performance of sentence generation.

Model	Team	METEOR%	BLEU@4%
LSTM	UC Berkeley, NAACL (2014/12)	26.9	31.2
SA	UdeM, arxiv (2015/02)	29.6	42.2
S2VT	UC Berkeley, arxiv (2015/05)	29.8	--
LSTM-E	MSR Asia, CVPR 2016	31.0	45.3
H-RNN	Baidu, CVPR 2016	32.6	49.9
HRNE	UTS, CVPR 2016	33.1	43.8
GRU-RCN	UdeM, ICLR 2016	31.6	43.3

Microsoft Research Video to Language Grand Challenge



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.

Dataset	Organizer	Context	Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hr)	Baselines
YouCook	SUNY Buffalo	Cooking	Labeled	88	-	2,668	42,457	2,711	2.3	MP-LSTM (VGG, AlexNet)
TACos	MP Institute	cooking	Labeled	123	7,206	18,227	-	-	-	MP-LSTM (C3D + VGG)
TACos M-L	MP Institute	cooking	Labeled	185	14,105	52,593	-	-	-	SA-LSTM (VGG, AlexNet)
M-VAD	UdeM	movie	DVS	92	48,986	55,905	519,933	18,269	84.6	SA-LSTM (C3D + VGG)
MPII	MP Institute	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6	LSTM-E
MSVD	MSR	multi-category	AMT workers	-	1,970	70,028	607,339	13,010	5.3	
MSR-VTT (10K)	MSRA	20 categories	AMT workers	5,942	10,000	200,000	1,535,917	28,528	38.7	
MSR-VTT (20K)	MSRA	20 categories	AMT workers	14,768	20,000	400,000	4,284,032	49,436	87.8	

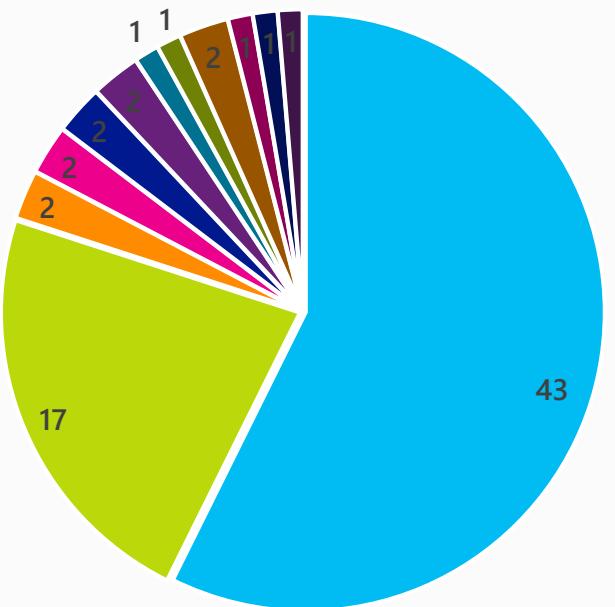
Microsoft Video to Language Challenge

77 teams registered challenge

22 teams submitted results

Awards will be announced at ACMMM

- China
- US
- Finland
- Japan
- Taiwan
- Korea
- Portugal
- Israel
- Australia
- Greece
- Canada
- India

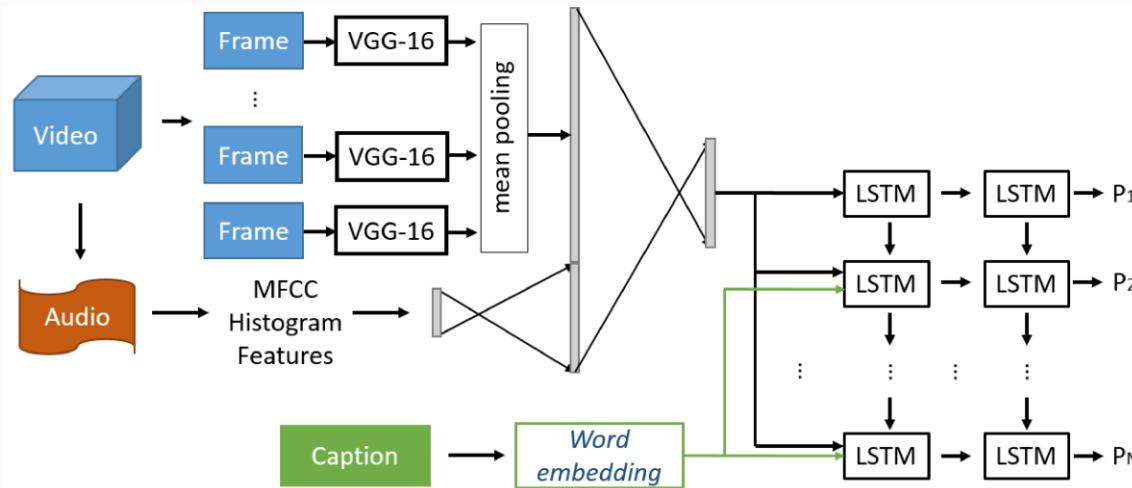


M1	M2	Rank	Team	Organization	BLEU@4	Meteor	CIDEr-D	ROUGE-L
		1	v2t_navigator	RUC & CMU	0.408	0.282	0.448	0.609
		2	Aalto	Aalto University	0.398	0.269	0.457	0.598
		3	VideoLAB	UML & Berkeley & UT-Austin	0.391	0.277	0.441	0.606
		4	ruc-uva	RUC & UVA & Zhejiang University	0.387	0.269	0.459	0.587
		5	Fudan-ILC	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

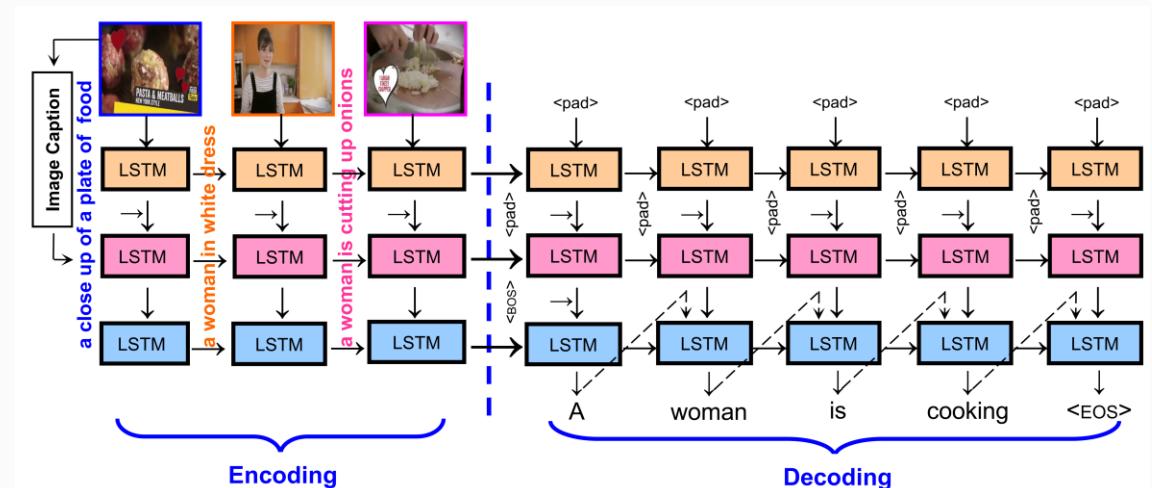
M1	M2	Rank	Team	Organization	C1	C2	C3
		1	Aalto	Aalto University	3.263	3.104	3.244
		2	v2t_navigator	RUC & CMU	3.261	3.091	3.154
		3	VideoLAB	UML & Berkeley & UT-Austin	3.237	3.109	3.143
		4	Fudan-ILC	Fudan & ILC	3.185	2.999	2.979
		5	ruc-uva	RUC & UVA & Zhejiang University	3.225	2.997	2.933
		6	Umich-COG	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

Summary from 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.

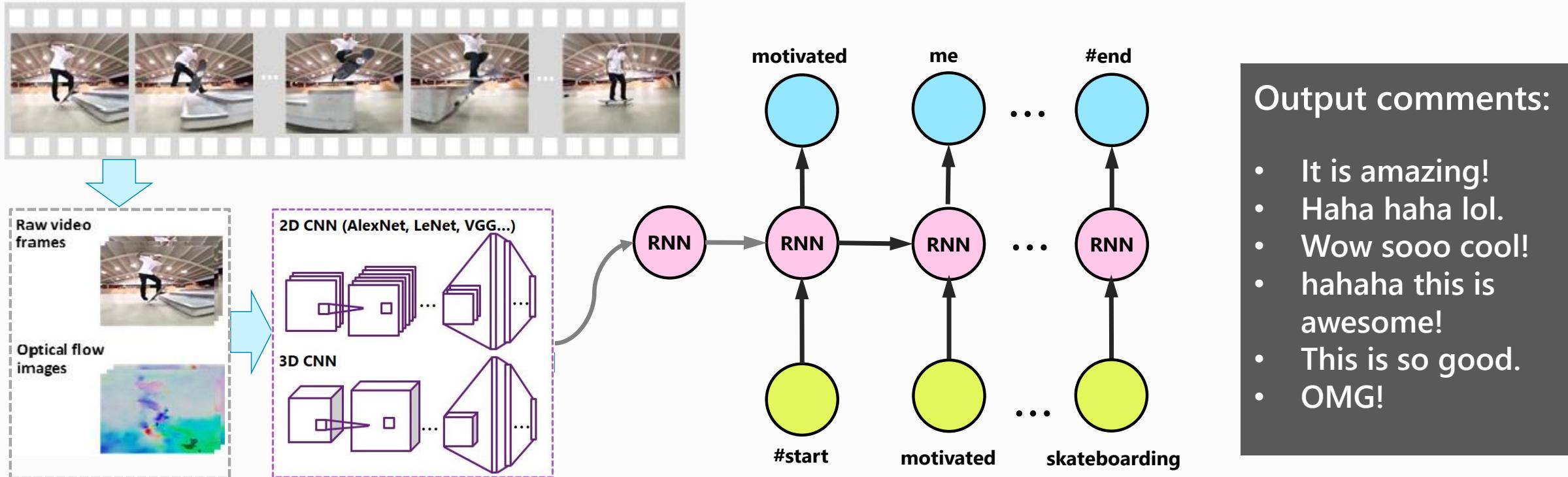
Descriptors	BLEU@4	METEOR	CIDEr	ROUGE-L
categories	0.298	0.228	0.236	0.548
audio	0.301	0.222	0.184	0.544
C3D	0.374	0.264	0.389	0.594
ResNet	0.389	0.269	0.400	0.605
+C3D	0.385	0.267	0.411	0.601
+categories	0.381	0.270	0.418	0.597
+audio	0.395	0.277	0.442	0.610
committee	0.407	0.286	0.465	0.610

- 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]

Outline

- Image and video captioning
- Video commenting
- Video sentiment analysis
- Video and language alignment
- Datasets and evaluations
- Open issues
- Learning materials

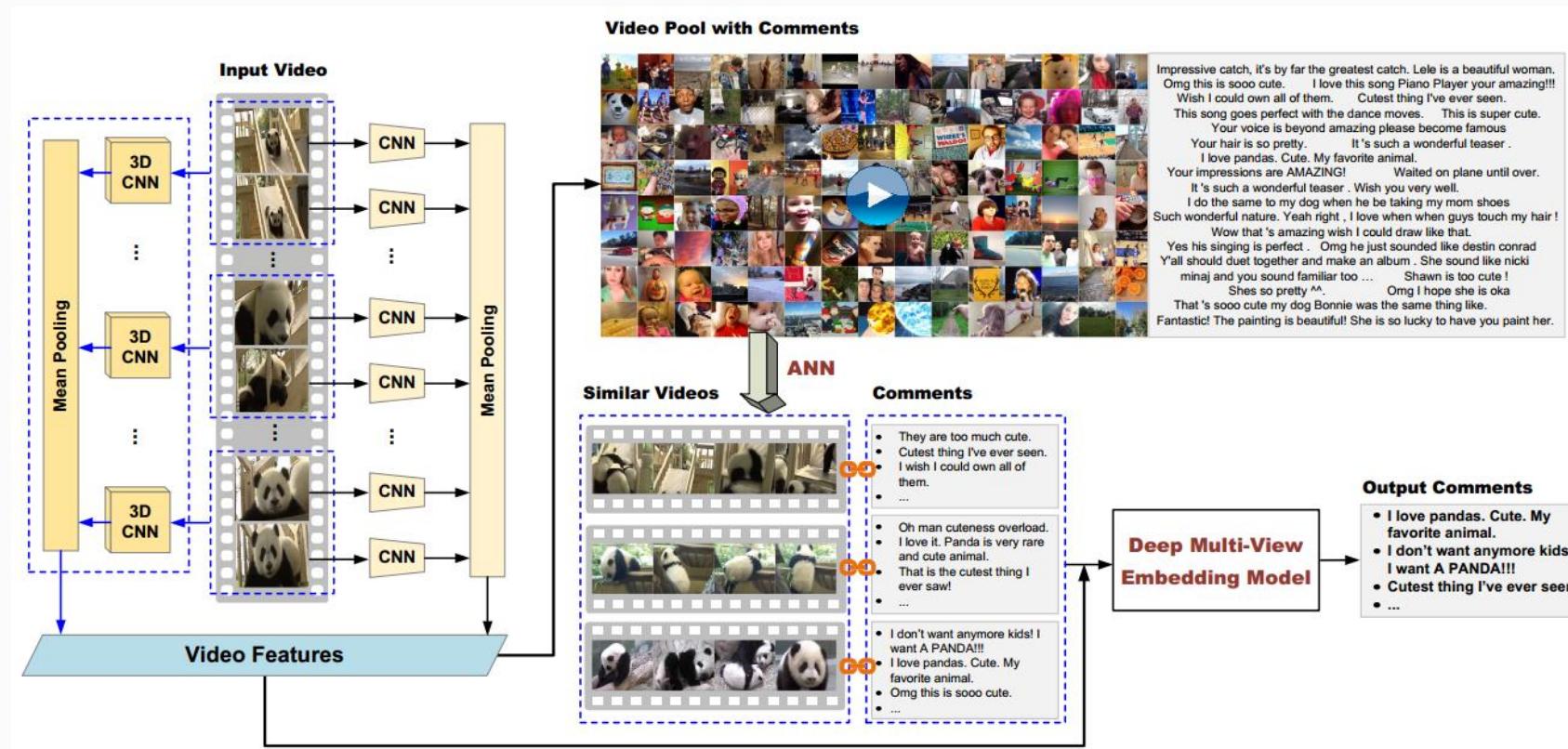
Video commenting [Li, MM'16]



- General-purpose phrases 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

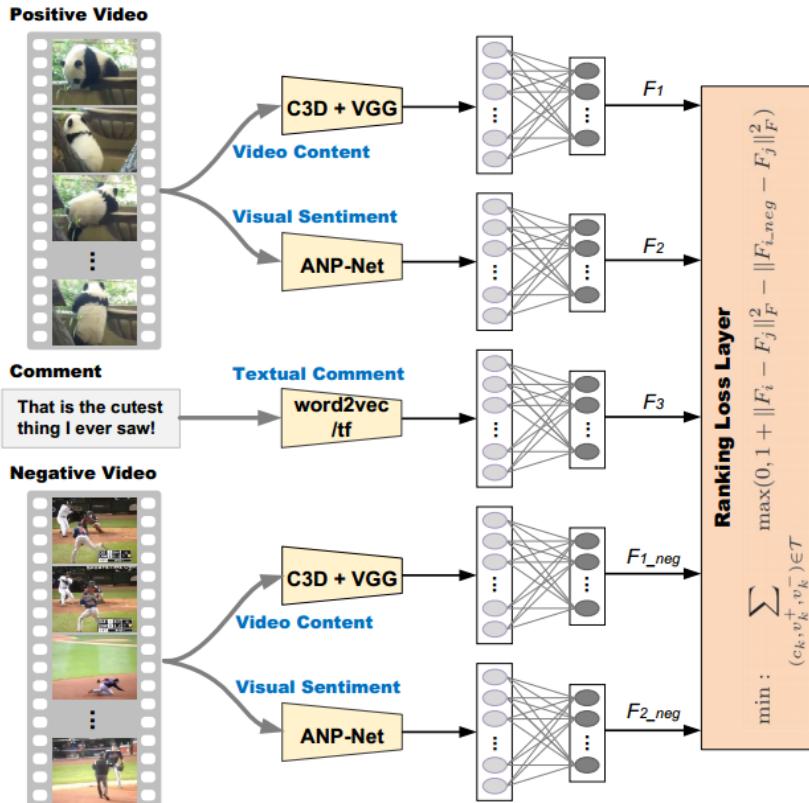
Video commenting

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



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

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

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)

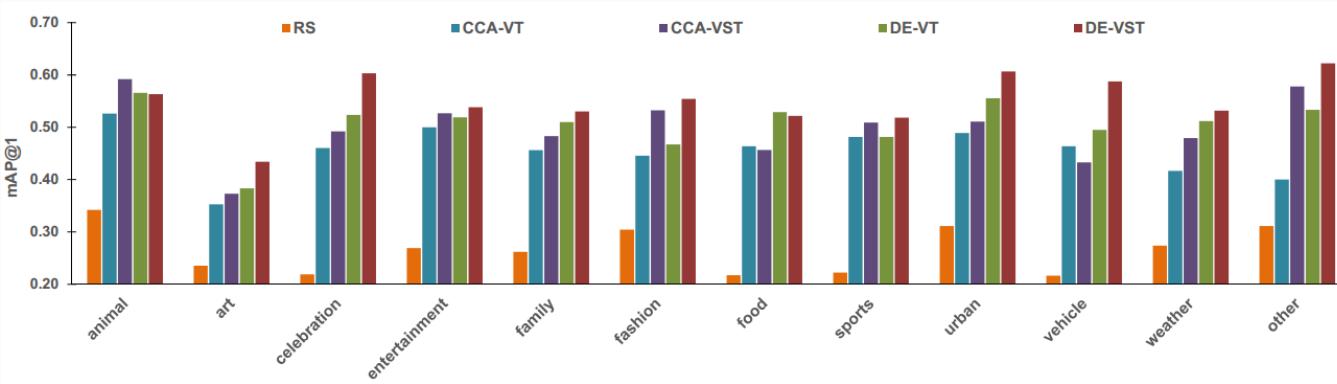


"Haha so cute
and funny at the
same time"
"Glad she is
better. So cute"



"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.

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



- * 今天吃得好淑女 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的小嘴嘴 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



- * 今天吃得好淑女
Eating like a lady with great manner
- * 吃得吧唧吧唧
Eating very enjoyable
- * 每天都在变更漂亮
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
- * 闪眼, 真美
Catches the eyes, so pretty
- * 美美的
Beautiful



- * 冉苒妈24小时陪孩子
Ran's mom stays with her for 24h
- * 看着冉苒每天都在成长进步
Watching冉苒 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
- * 太喜欢冉苒了
Like冉苒 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的小嘴嘴
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
48

Results: auto-commenting



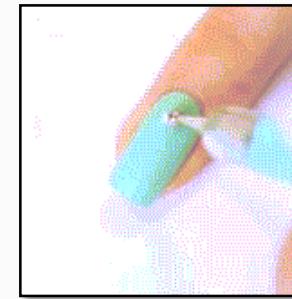
* The eyebrow is pretty 0.5613
* Beautiful 0.5388
* Still looks so pretty 0.5314
* Candy to the eyes 0.5285
* Very beautiful 0.5189



* Such a beautiful daughter 0.4469
* What a cute and beautiful baby 0.4335
* It's too pretty 0.4274
* Such a beautiful baby 0.4237
* Baby is the most beautiful gift of the whole world 0.4181



* What kind of dog is this? very cute 0.4884
* Is this a dog? 0.4714
* It looks exactly like my dog. Even the way they look at you is alike 0.4588
* Your dog is so cute, beautiful lady 0.4573
* Cute puppy 0.4571



* Beautiful manicure takes you into spring 0.4156
* Bohemian manicure 0.4014
* Will do this manicure next time 0.3654
* Beautiful manicure 0.3626
* How do you call those tools used for manicure? 0.3572



* The last one was very harsh 0.3413
* It is red 0.3136
* The last one hurts hatched more 0.2976
* It is all red after been slapped 0.2818
* The last hit hurt me more 0.2813



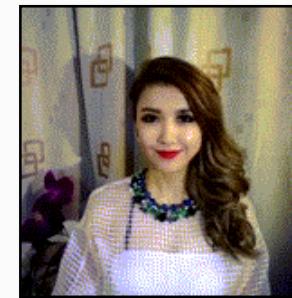
* Behave so much better than my Samoyed 0.6156
* This is Samoyed, right? 0.5723
* So cute that I miss my own Samoyed 0.5272
* The puppy Samoyed is the cutest 0.4863
* I want a Samoyed indeed 0.4768



* Little cutie 0.4643
* The hat is so cute 0.4201
* The eyes are so beautiful. It's too cute and I love it so much 0.4102
* Baby looks so handsome with the hat on. So cute 0.3950
* Such a cute little baby 0.3927



* Mr. Guitar is enjoying it too much 0.4779
* Sounds wonderful, hope that I can hear the whole version of each song 0.4715
* I am moved by the guitar player 0.4507
* Want to hear the final version 0.4373
* Sounds fantastic when put together 0.4341



* It's pretty and I love ancient cloth too 0.4610
* Beautiful Goddess 0.4395
* Super beautiful 0.4253
* it is beautiful 0.4145
* Beautiful 0.4142



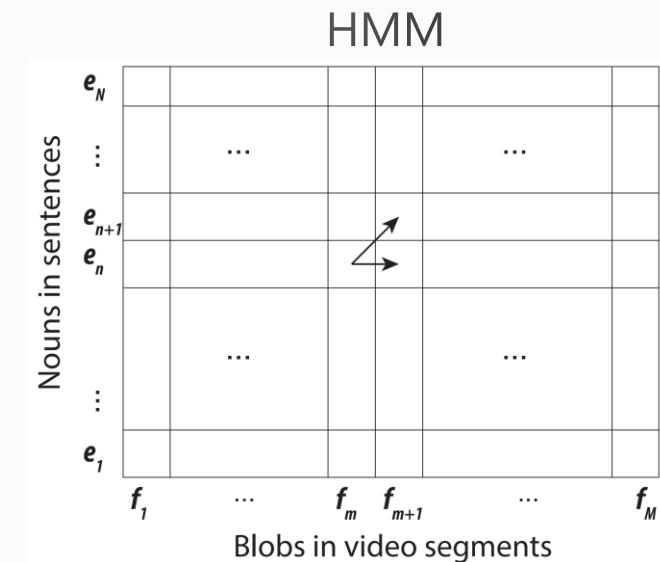
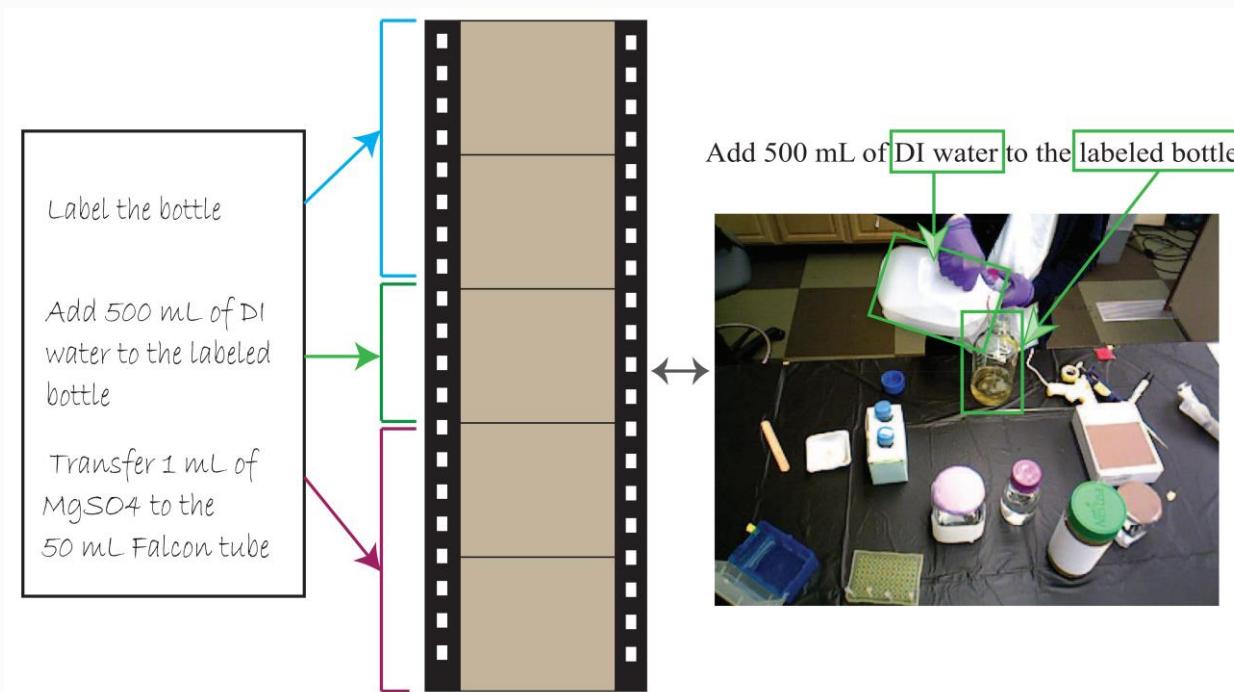
* Such a cute kitty 0.6174
* What kind of cat is this? Too cute 0.6095
* It looks too comfortable and makes me want to be a cat too 0.5817
* Is it Garfield? 0.5575
* What cat is this? So cute 0.5537

Outline

- Image and video captioning
- Video commenting
- **Video and language alignment**
- Datasets and evaluations
- Open issues
- Learning materials

Alignment of Video and Language

- Alignment of language instructions with video segments [Naim, AAAI'14]
 - Aligning nouns to video blobs
 - Model: HMM + IBM 1 [Brown, CL'93]



Probability of generating a set of blobs \mathbf{f} from a set of nouns \mathbf{e} :

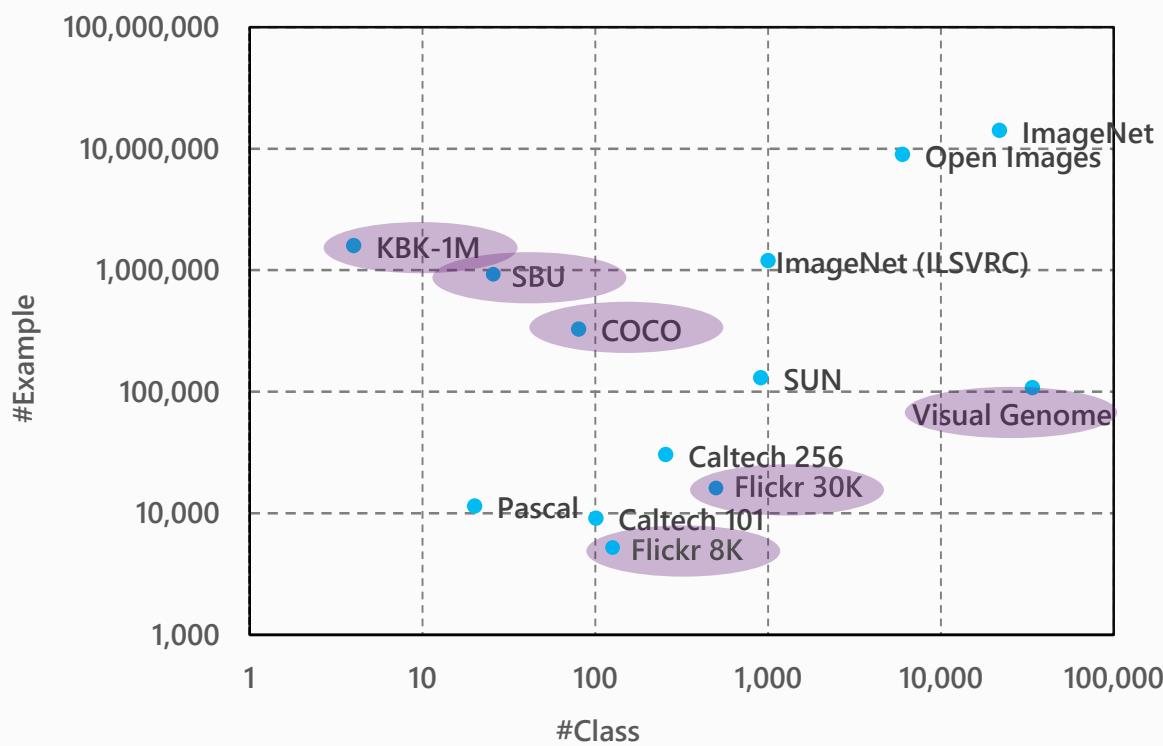
$$P(\mathbf{f}^{(m)} | \mathbf{e}^{(n)}) = \frac{\epsilon}{(I)^J} \prod_{j=1}^J \sum_{i=1}^I p(f_j^{(m)} | e_i^{(n)})$$

Outline

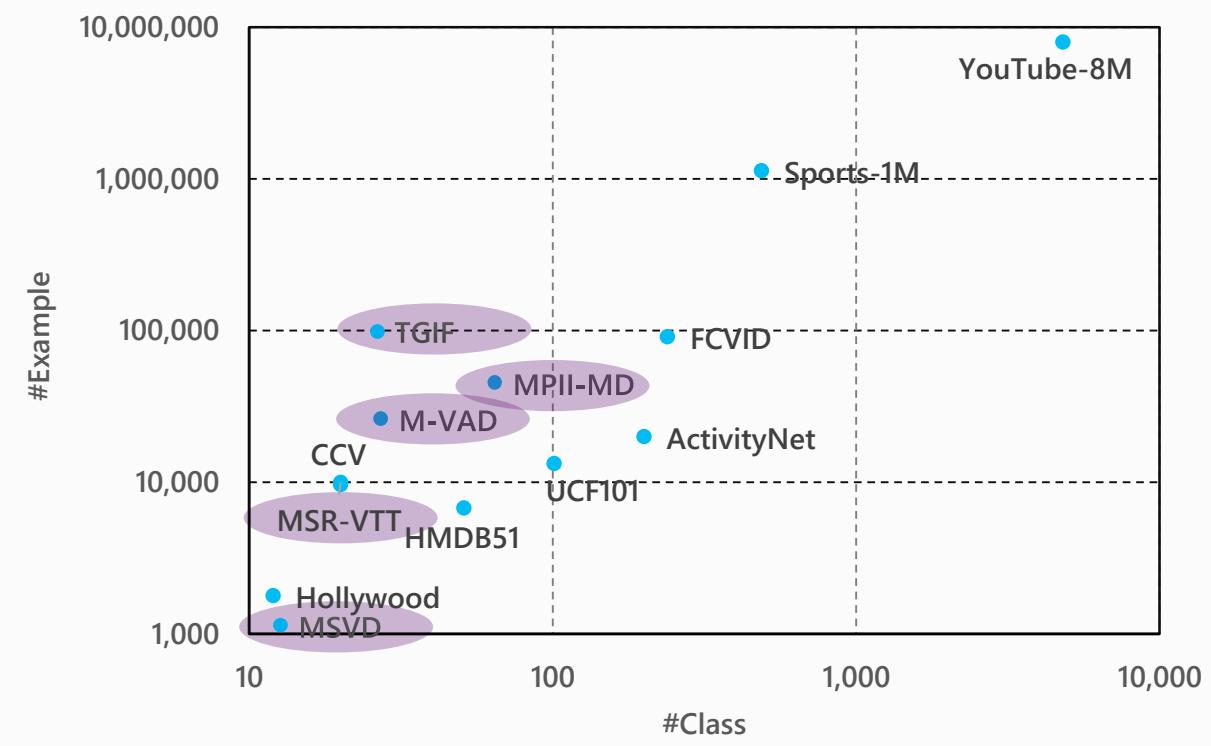
- Image and video captioning
- Video commenting
- Video and language alignment
- **Datasets and evaluations**
- Open issues
- Learning materials

Datasets for Captioning

image



video

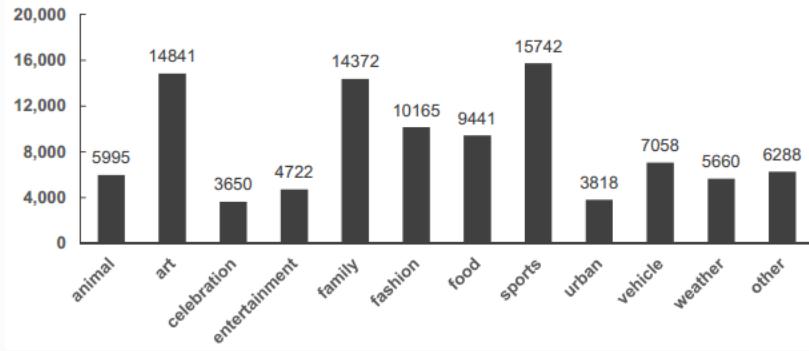


Dataset for captioning.

Note: The class information is unknown for Flickr 8K/30K, SBU, and MSVD, MPII-MD, M-VAD, TGIF.

MSR Video Commenting Dataset [Li, MM'16]

- 101,752 videos from Vine
- 5-15 sec per each video
- ~100 comments per each video
- 12 Categories



- That's so cute where he's waving the flag
- Poor Baby but it was so funny
- he's so cute



- Haha so cute and funny at the same time
- Glad she is better. So cute
- Soo awesome and cute



- I love baseball
- That's how to play baseball
- That an amazing play!



- Such outstanding piano pieces and you play them sublimely :)
- Amazing. I was listening to this while studying!
- Keep it up that's wonderful!

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]
- Subjective metrics – human evaluations
 - Coherence, Relevance, Helpful for Blind [[MSR Video to Language](#)]

Open issues for vision to language

- Rule-based vs. Model-based vs. Data-driven approaches
 - More accurate object/action detection/recognition from videos
- Leveraging more powerful language models
 - Attention model
 - Bi-directional RNN
- Diversity/Natural
 - Sentiment analysis (e.g., adjective-noun pair)
 - Attributes of object (e.g., human body parsing, age)
 - Entity recognition (e.g., celebrity naming, face recognition)
- Multimodal data analysis (e.g., script, speech, audio, comments)
- Visual relationship modeling [Lu, ECCV'16]
- Complex and long videos
 - Data collection from weakly supervised Web data

Reference

- [Captioning] Y. Pan, T. Mei, T. Yao, et al. "Jointly Modeling Embedding and Translation to Bridge Video and Language," CVPR, 2016.
- [Captioning] J. Krishnamurthy, et al. "Generating Natural Language Video Descriptions using Text-mined Knowledge," AAAI, 2013.
- [Captioning] Karpathy, et al. "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2014.
- [Captioning] Vinyals, et al. "Show and Tell: A Neural Image Caption Generator", 2014.
- [Captioning] Kiros, et al. "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", 2014.
- [Captioning] Mao, et al. "Explain Images with Multimodal Recurrent Neural Networks", 2014.
- [Captioning] Donohue, et al. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description", 2014.
- [Captioning] Xu, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", 2015.
- [Commenting] Y. Li, T. Yao, T. Mei, et al. "Share-and-Chat: Achieving Human-Level Video Commenting by Search and Multi-View Embedding," ACM MM, 2016.
- [Sentiment] J. Wang, et al. "Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks," IJCAI, 2016.
- [Alignment] I. Naim, et al. "Unsupervised Alignment of Natural Language Instructions with Video Segments," AAAI, 2014.
- [Alignment] H. Yu, et al. "Grounded Language Learning from Video Described with Sentences," ACL, 2013.
- [Dataset] J. Xu, T. Mei, et al. "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," CVPR, 2016.
- [Dataset] Y. Li, et al. "TGIF: A New Dataset and Benchmark on Animated GIF Description," CVPR, 2016.
-

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
 - LRCN for action 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>

Acknowledgement

- **Ting Yao, Jianlong Fu, Yong Rui, Xiaodong He**
Microsoft Research
- **Yingwei Pan, Jun Xu, Houqiang Li**
University of Science and Technology of China
- **Jiebo Luo**
University of Rochester, USA

Thanks!

tmei@microsoft.com

