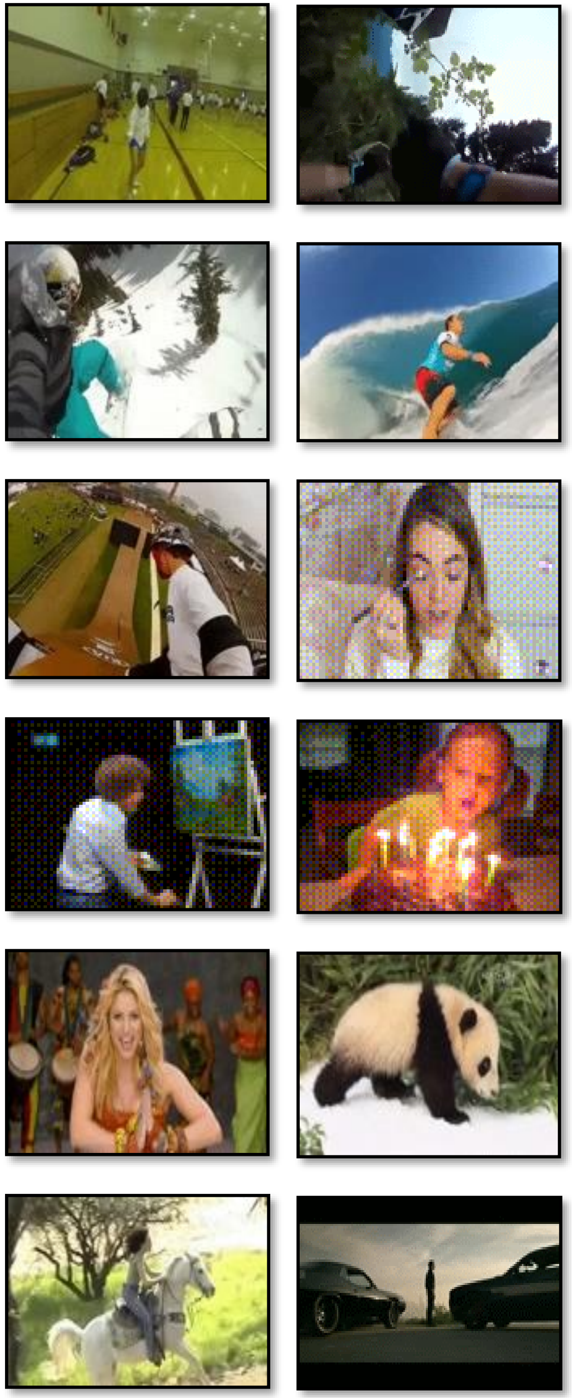
Deep Learning for Intelligent Video Analysis - Part II

Tao Mei, Senior Research Manager
Cha Zhang, Principal Applied Science Manager

Microsoft AI & Research

Microsoft



50 years of progress in visual understanding

feature engineering + model learning

$\mathbf{f} = f(I)$ $\quad y = g(\mathbf{f}, \theta)$

deep l

$y = g(I, \theta)$, wher

before 1990s
- Fischler & Elschlager (Structural Representation), 1973
- Canny edge detection, 1986 (27K)
- Snake, 1988 (18K)
- Eigenfaces, 1991 (16K)

1999
- SIFT (Scaled Invariant Feature Transform) (55K)

2001
- Boosting + Cascade = Speed (30K)

2003
- Constellation Model

2005
- HOG (Histograms of Oriented Gradients), 19K citations

2007
- SURF (Speed Up Robust Features), (18K)
- DPM
- PASCAL VOC
- LabelMe

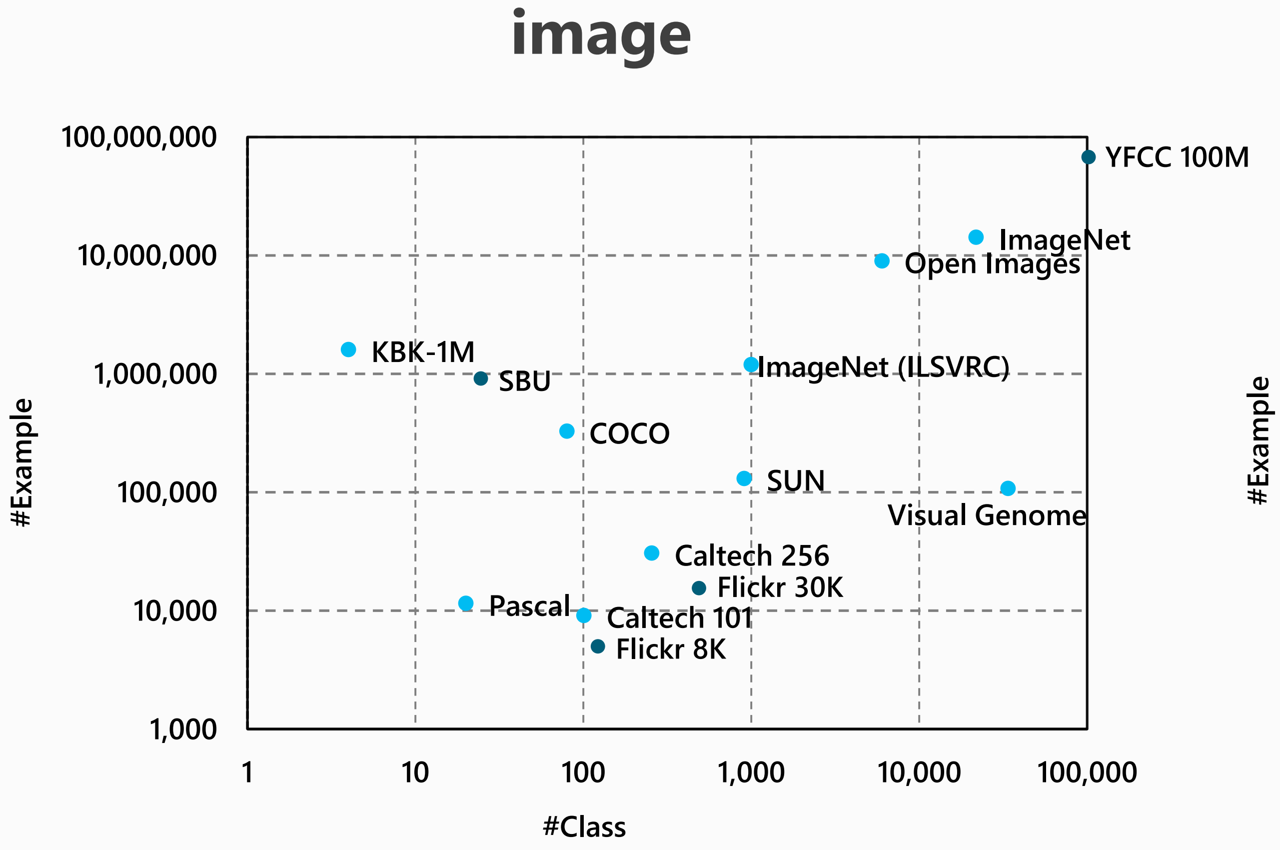
2009
- ImageNet (4K)
- Caltech Pedestrian

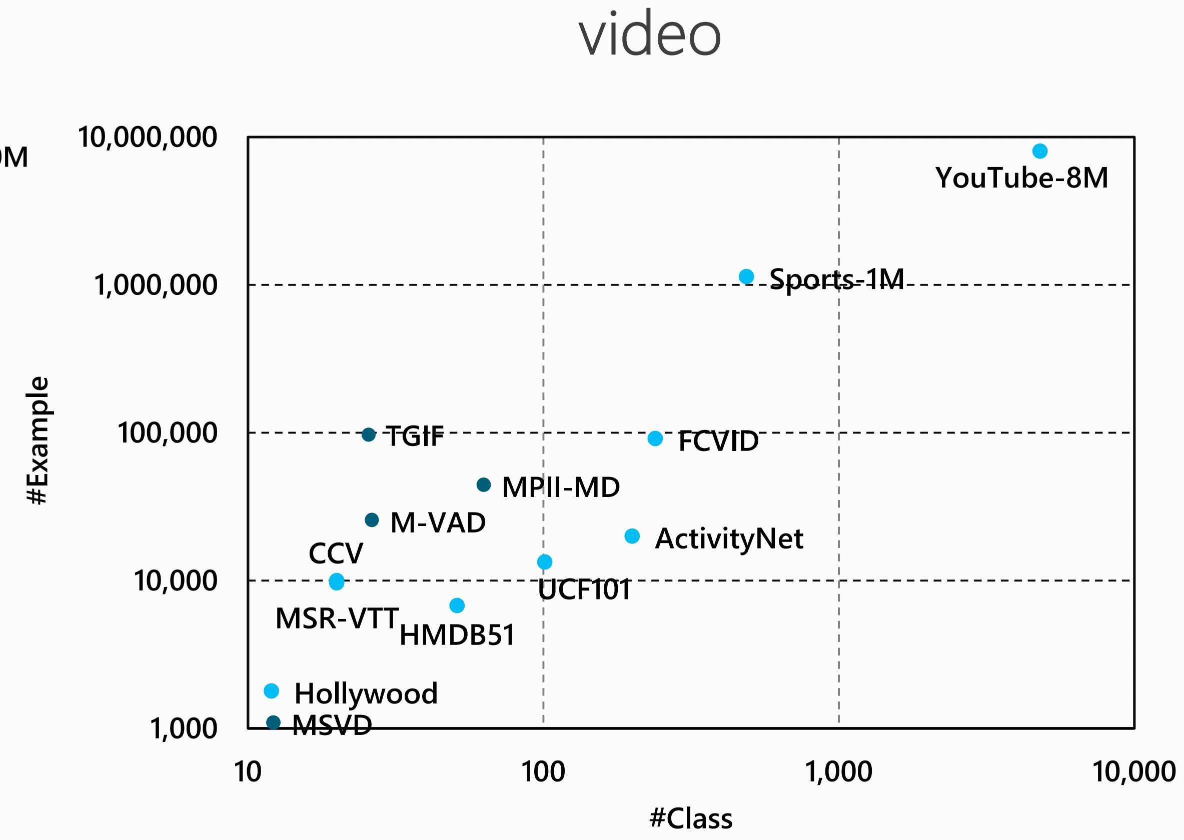
2012~ present
- DNN, Everywhere
- CNNs for ImageNet-1K (AlexNet) (13K)
- MS COCO,
- MSR VTT,
- Visual Genome,
- ActivityNet,
- Sports-1M,
- YFCC100M,
- Youtube-8M,
- Open Images 9M,
- …

2014
- GoogLeNet (3.5K)
- VGG Net (5K)
- Generative Adversarial Networks
- Vision and Language

20

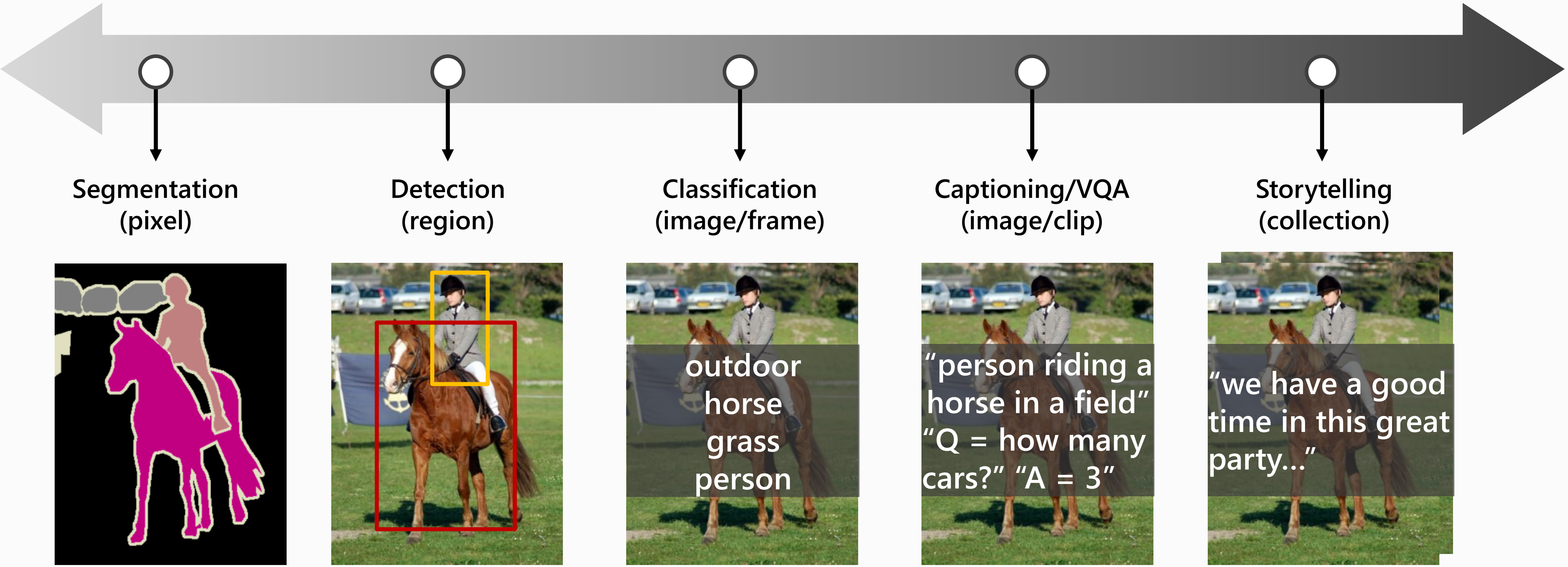
Computer vision: 50 years of progress



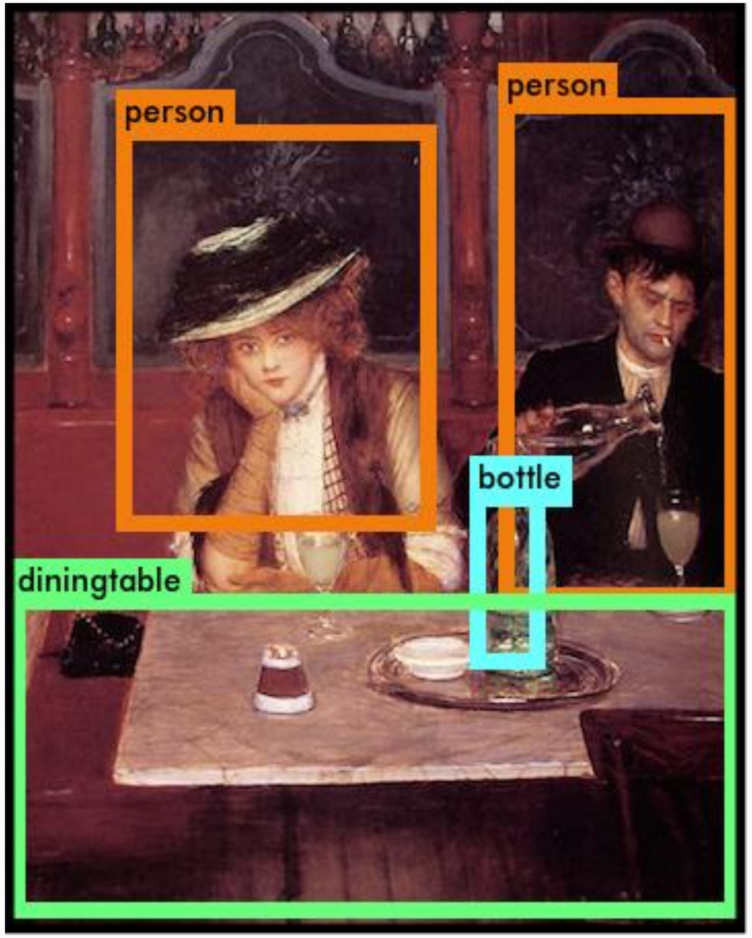


Note: The class information is unknown for Flickr 8K/30K, SBU, and MSVD, MPII-MD, M-VAD, TGIF.

Image understanding: core problems



Deep learning for object detection



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

Deep learning for semantic segmentation

Results on the first 5k images from the COCO test set is available at https://github.com/daijifeng001/TA-FCN

Deep learning for image captioning

"I think it's a boat is docked in front of a building." https://www.captionbot.ai/ *[Microsoft CaptionBot, 2015]*

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

Deep Learning for Image Poem Generation

- 学习了1920年以来的519位中国现代诗人的100K行诗
  Learned 100K lines of poems from 519 Chinese poets since 1920

- 每学习一轮需要0.6分钟，经过10K次100个小时的迭代学习
  0.6 min for each learning iteration, 100 hrs overall training time with 10K iterations

- 精选139首结集出版
  Published with 139 selected poems

Deep learning to

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

- visual recognition: classification, detection, segmentation

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

- vision to language
- image captioning & poeming
- visual question-answering

Some statistics about video

55%

of people watch videos online every day

3.7 Billion

daily views for video at facebook

facebook

500 Million

hours of videos watched daily in Youtube

YouTube

30%

video ad spend increased 30% from 2015 to 2016

2.6 X

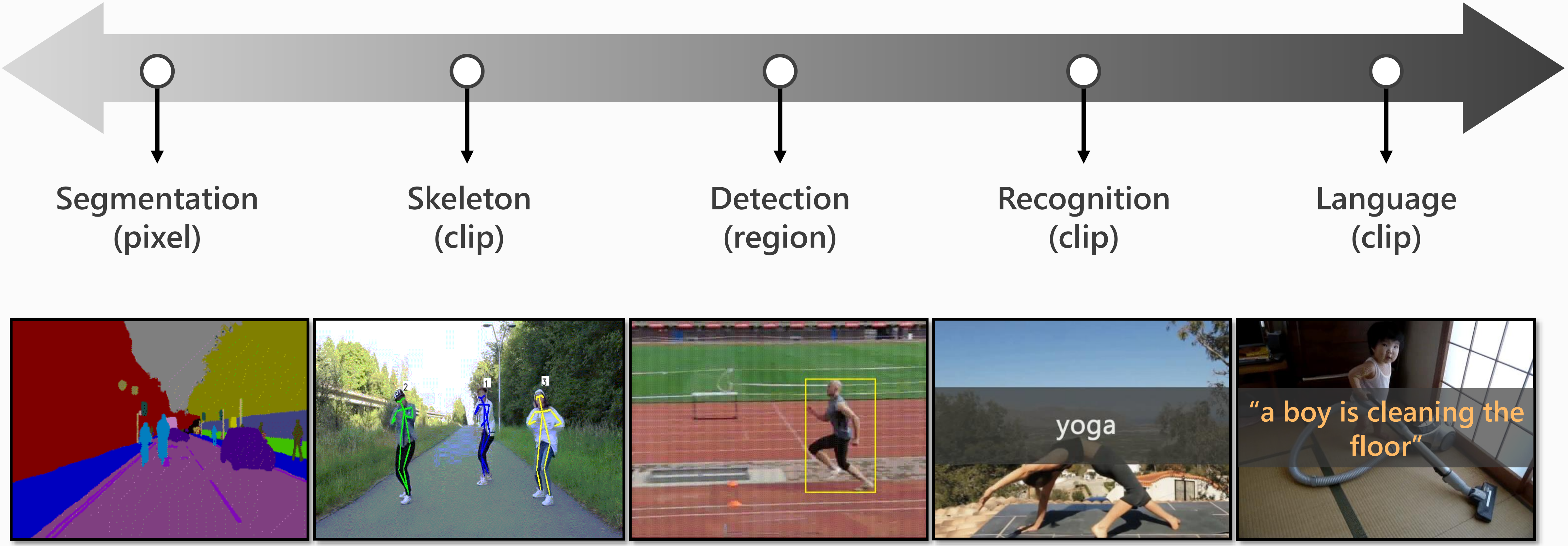
people spend 2.6x more time on pages w/ video than w/o

1200%

video generates 1200% more shares than text and image


AutoRace
Badminton
Baseball
Basketball
Beach Handball
Beach Tennis
Beach
Volleyball
Biking
Blowing
Candles
Bodybuilding
BrushHai
r
Billiards
Bowling
Boxing
Climbing
Cricket
Curling
Diving
Fencin
g
Cooking
Typing
Handstand
Fishing
Football
Golf
Handball
Hockey
HorseRiding
IceHockey
Judo
Knitting
Makeup
NailArtDesign
Kayak
Motorcycling
Rafting
Rowing
Sailing
ShootGun
Skateboarding
Skating
Painting
PullUps
PushUps
Skiing
SkippingRope
Skydiving
Soccer
Softball
Surfing
Swimming
TableTennis
Situp
Treadmill
Decorating
ChristmasTree

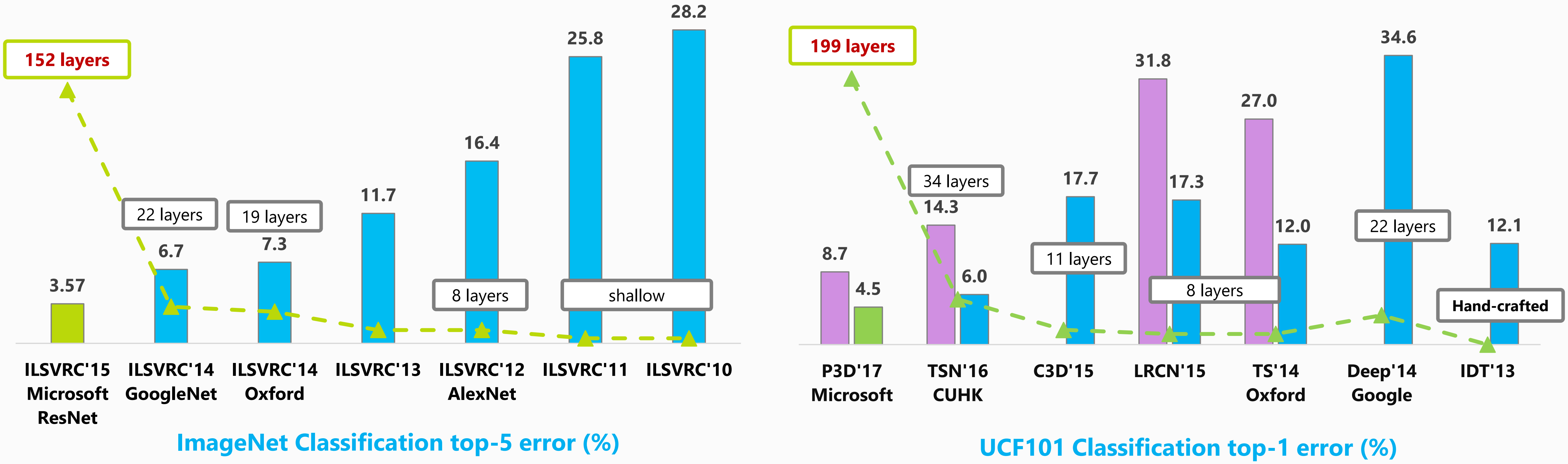
Video understanding: core problems



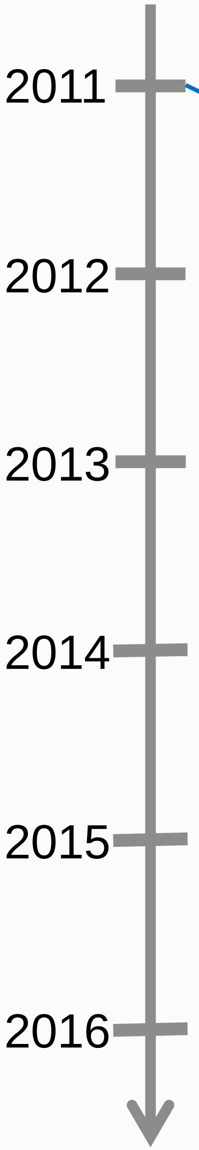
This part

- Video representation learning
- Video classification (a.k.a. action recognition)
- Video captioning
- Semantic video segmentation

Learning video representation is harder than image!

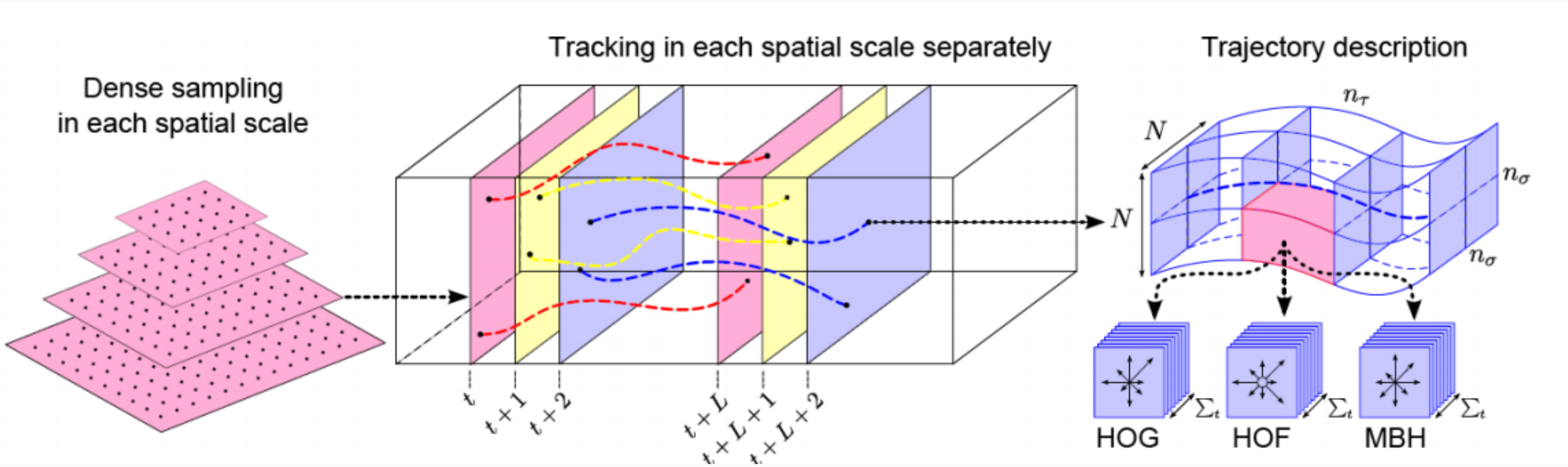


Video representation learning



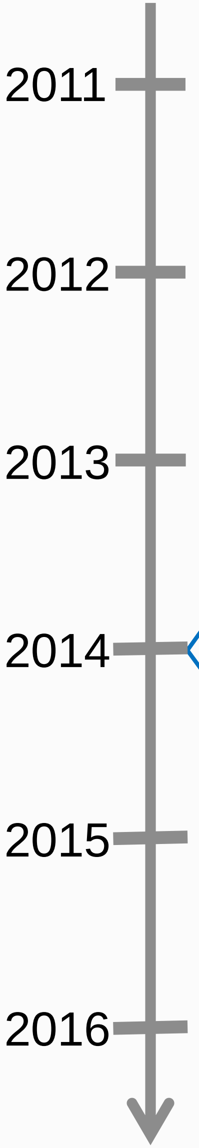
Hand-crafted feature

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



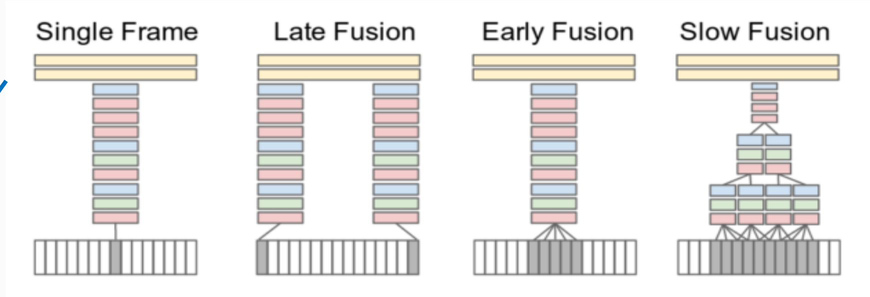
- Suffer from camera motion and illumination change in video
- Not contain high-level semantic information
- High dimensionality
- Too expensive for real-time computation

Video representation learning



2D Convolutional Neural Network

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



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

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

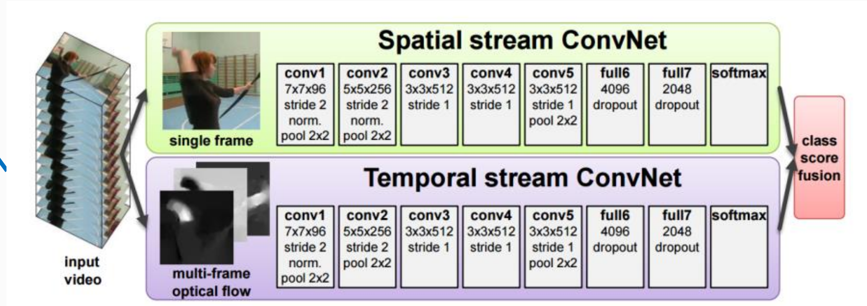
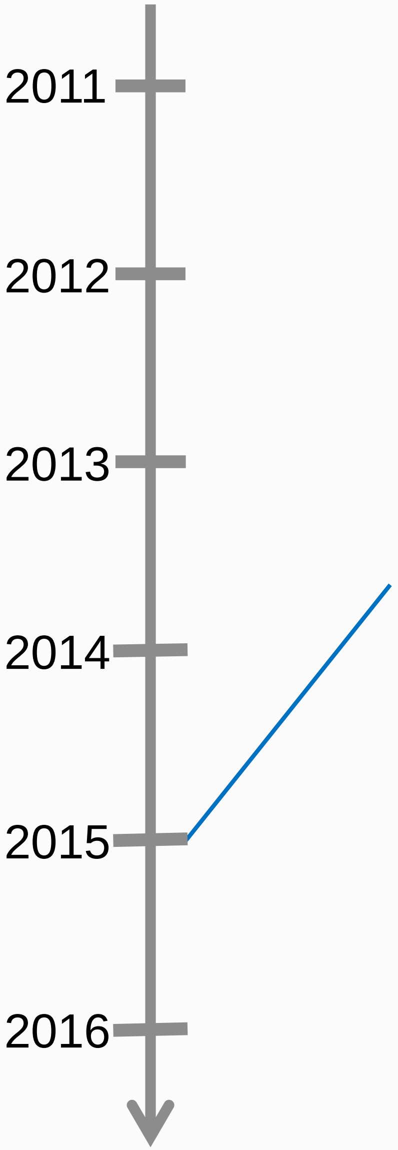


Figure 1: **Two-stream architecture for video classification.**

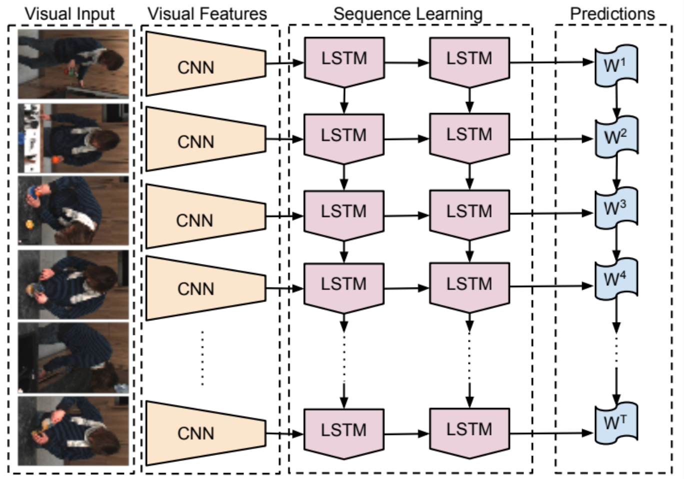
- Two-stream: frame + motion (stacked optical flow)
- 2D CNN for frame is pre-trained on ImageNet
- 2D CNN for motion is trained from scratch

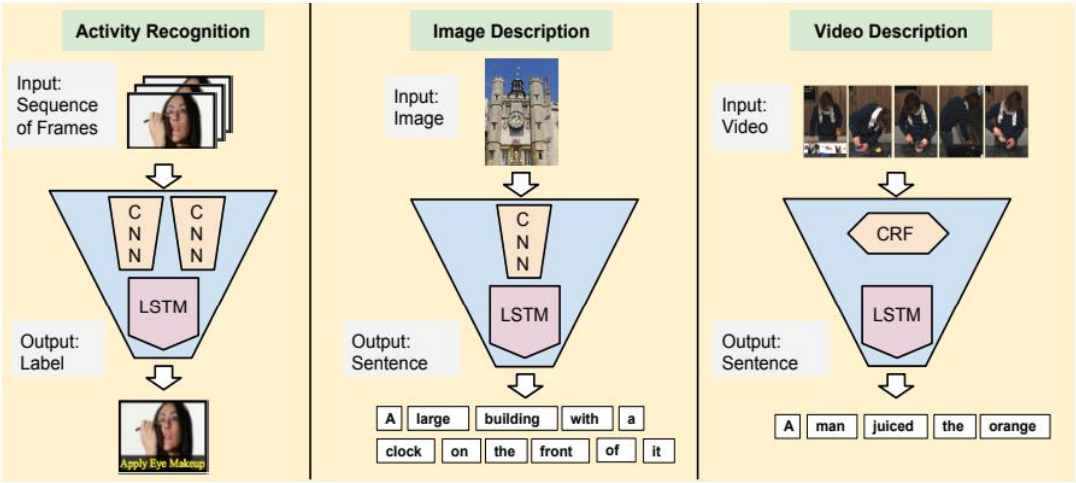
Video representation learning



2D CNN + LSTM (LRCN)

Long-term Recurrent Convolutional Networks for Visual Recognition and Description [Donahue, CVPR'15]

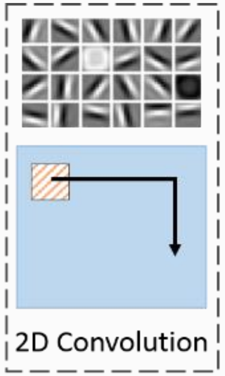




- Develop recurrent convolutional architecture
- Outputs of 2D CNN are fed into a stack of LSTM
- Applications on activity recognition and video description
- Neglecting low-level motion information

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

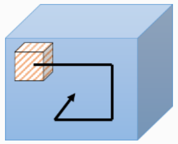
ResNet:
[MSRA, CVPR'16]

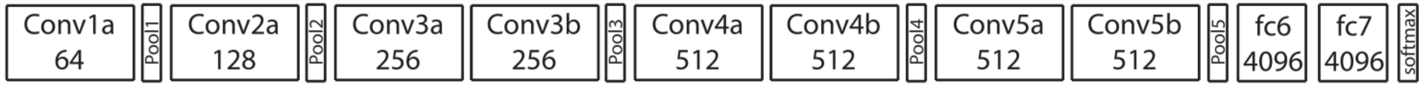


Network comparison on Sports-1M

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

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

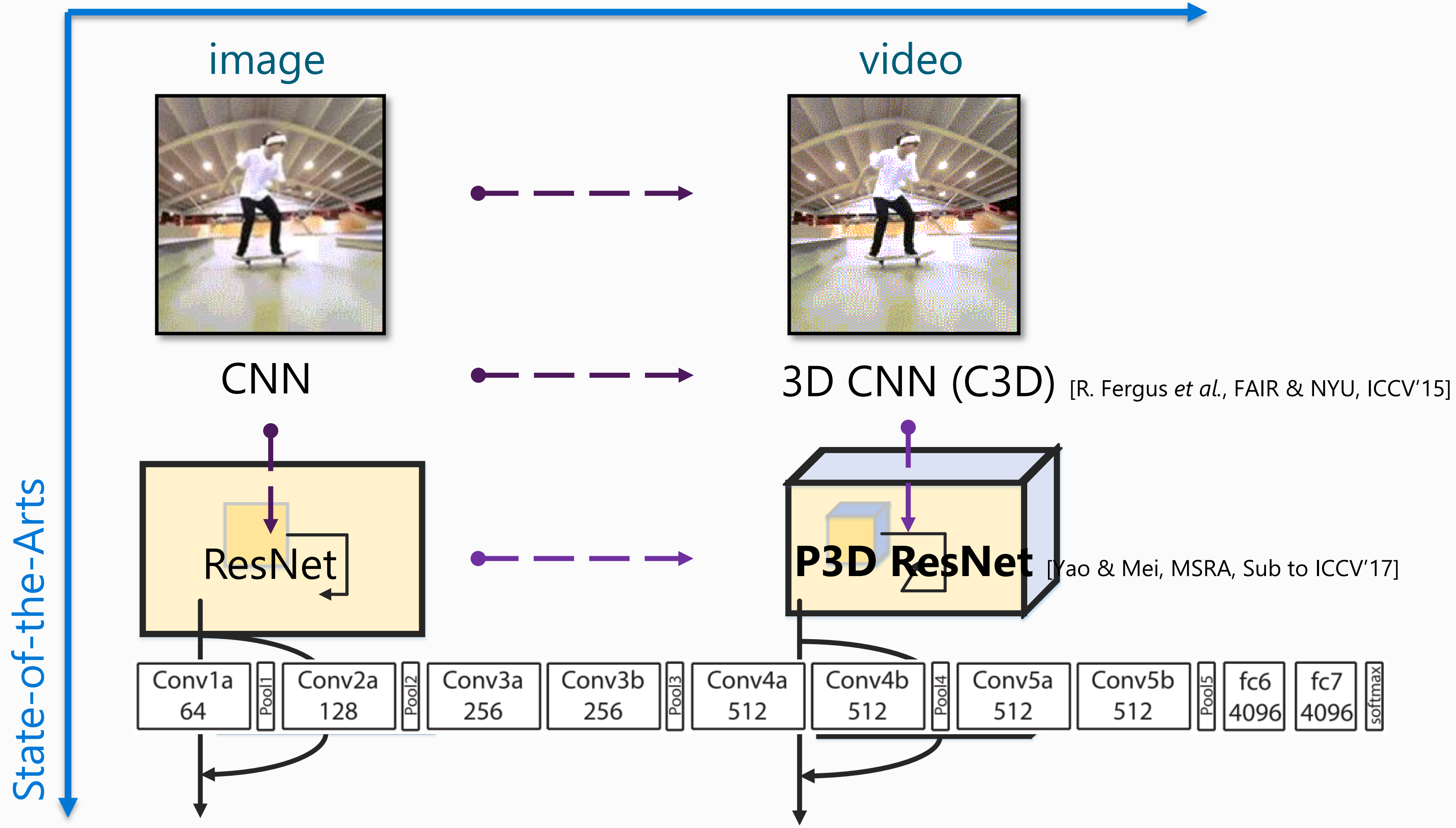


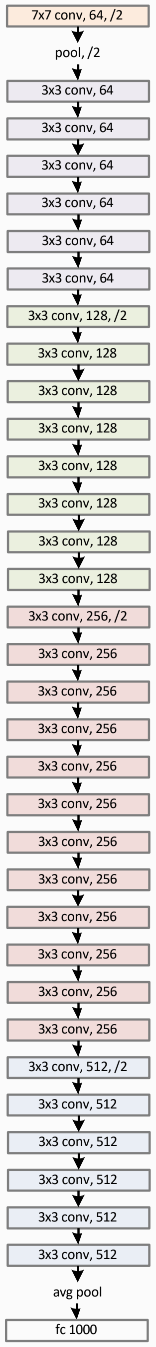


- Training 3D CNN is very computationally **expensive**
- Difficult to train very **deep** 3D CNN
- **Fine-tuning** 2D CNN is better than 3D CNN

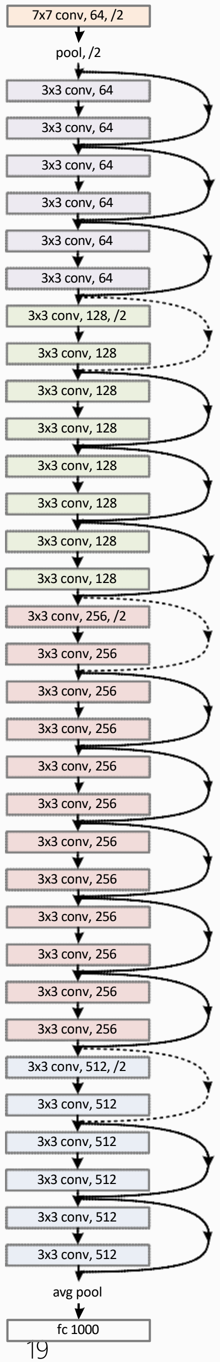
Pseudo-3D Residual Networks (P3D)

[Yao & Mei, ICCV'17]



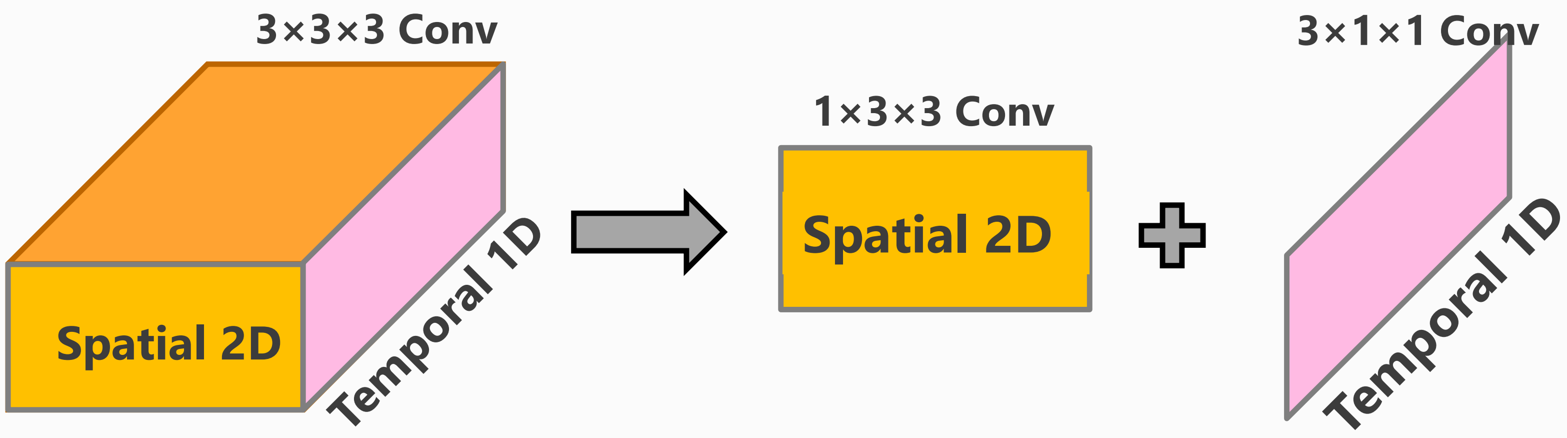


Plain Net

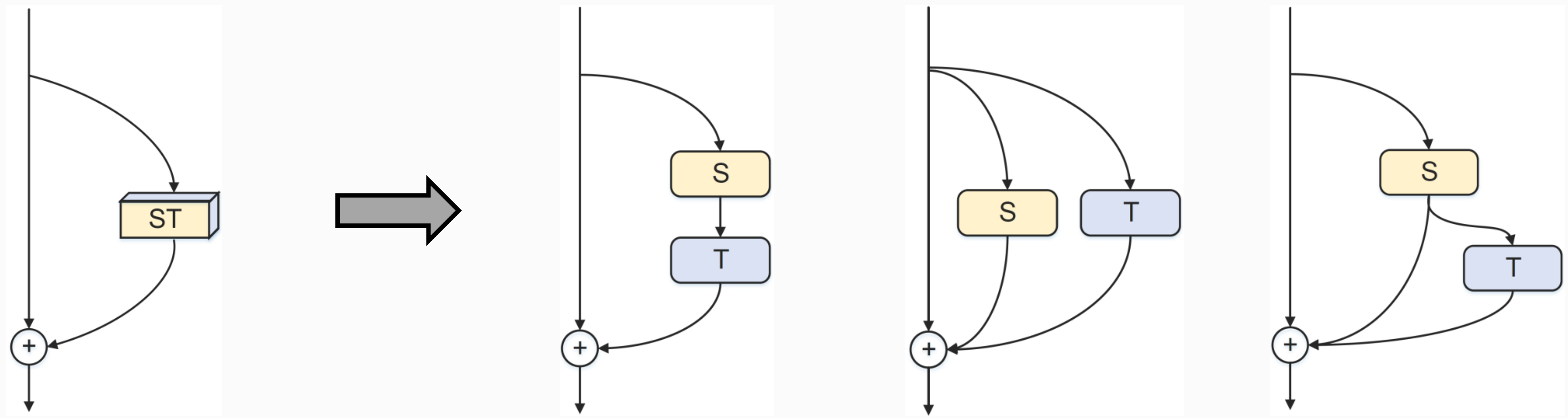


ResNet

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

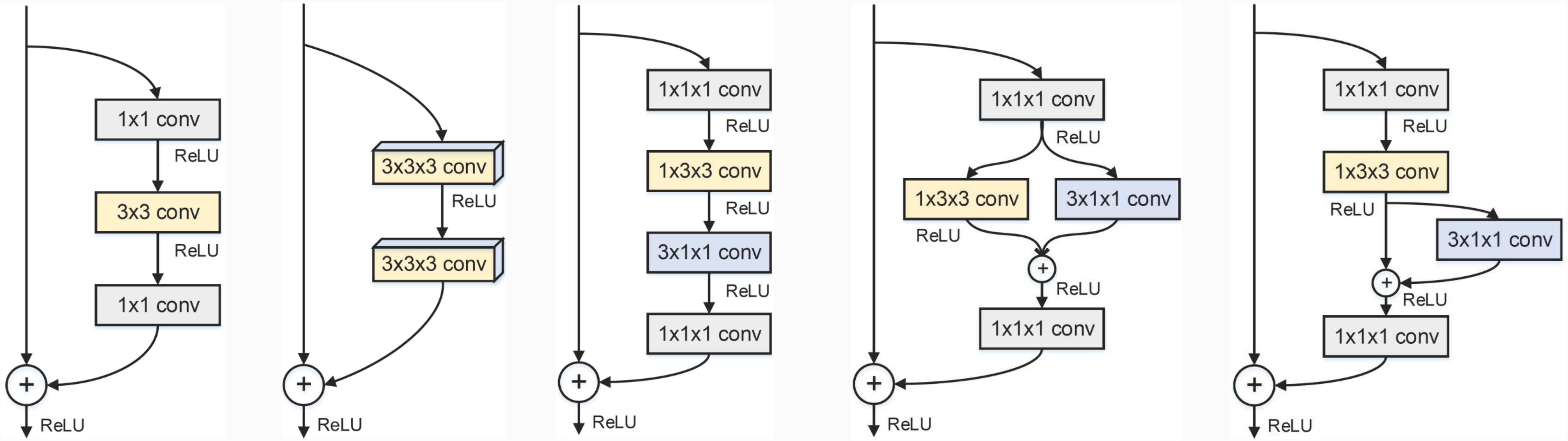


- Reduce model size
- Fully leverage pre-learned 2D CNN from image
- Enhance the structural diversity



P3D-A P3D-B P3D-C

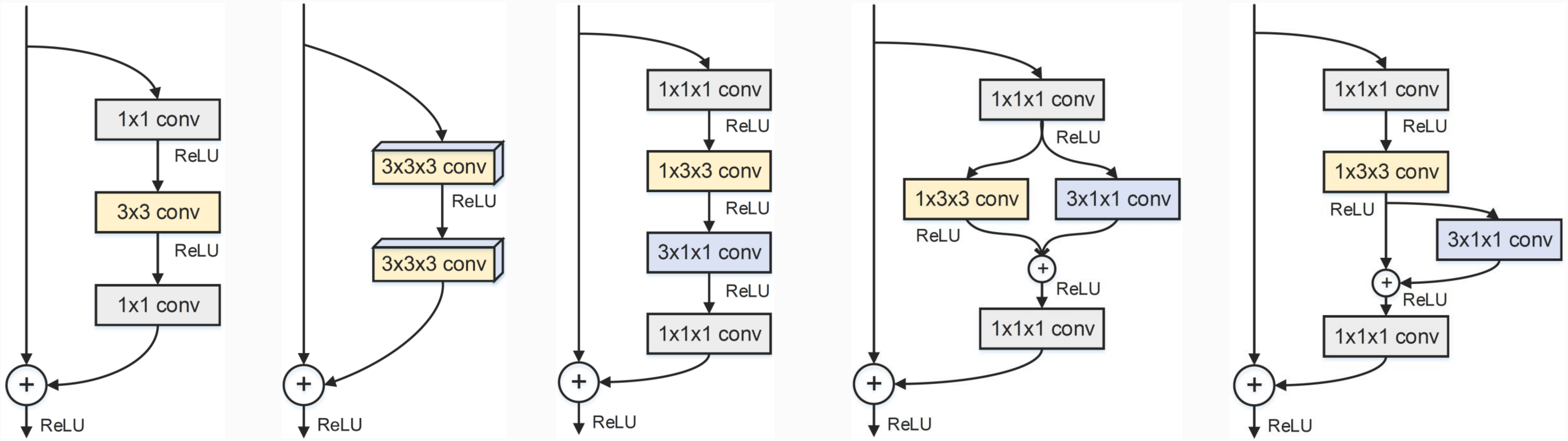
P3D: architectures



(a) ResNet (b) 3D ResNet **(c) P3D-A** **(d) P3D-B** **(e) P3D-C**

- 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]

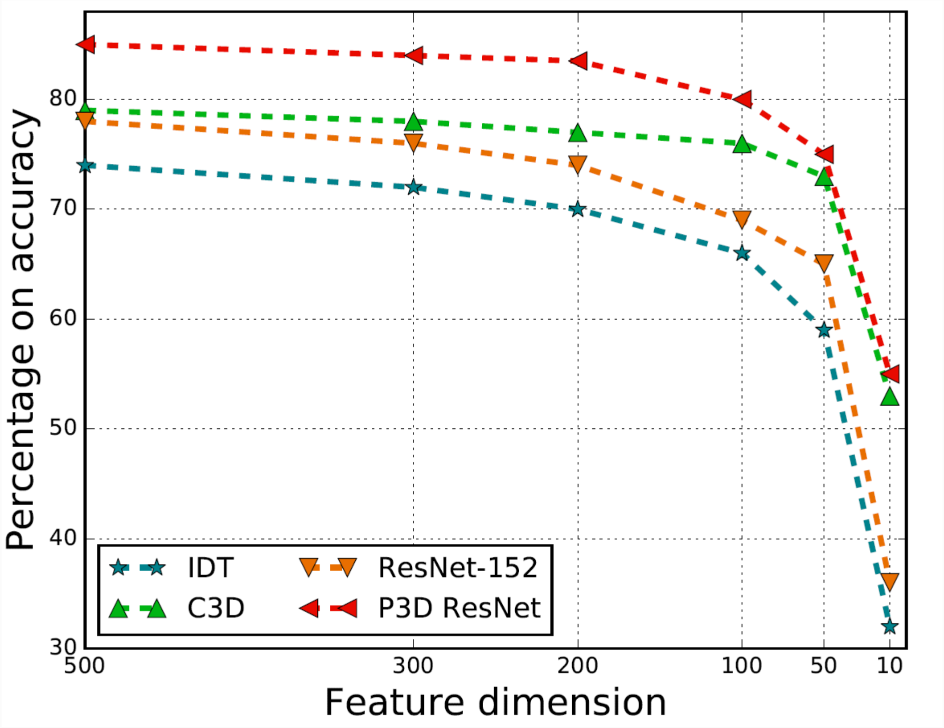
P3D https://github.com/ZhaofanQiu/pseudo-3d-residual-networks [Qiu, Yao, Mei, ICCV'17]



(a) ResNet (b) 3D ResNet **(c) P3D-A** **(d) P3D-B** **(e) P3D-C**

$$
\begin{aligned}
&\text{(a)} \quad (\mathbf{I}+\mathbf{S})\cdot\mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) = \mathbf{x}_{t+1} \\
&\text{(b)} \quad (\mathbf{I}+\mathbf{TS})\cdot\mathbf{x}_t := \mathbf{x}_t + \mathbf{TS}(\mathbf{x}_t) = \mathbf{x}_{t+1} \\
&\text{(c)} \quad (\mathbf{I}+\mathbf{T}\cdot\mathbf{S})\cdot\mathbf{x}_t := \mathbf{x}_t + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)) = \mathbf{x}_{t+1} \\
&\text{(d)} \quad (\mathbf{I}+\mathbf{S}+\mathbf{T})\cdot\mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) + \mathbf{T}(\mathbf{x}_t) = \mathbf{x}_{t+1} \\
&\text{(e)} \quad (\mathbf{I}+\mathbf{S}+\mathbf{T}\cdot\mathbf{S})\cdot\mathbf{x}_t := \mathbf{x}_t + \mathbf{S}(\mathbf{x}_t) + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)) = \mathbf{x}_{t+1}
\end{aligned}
$$

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

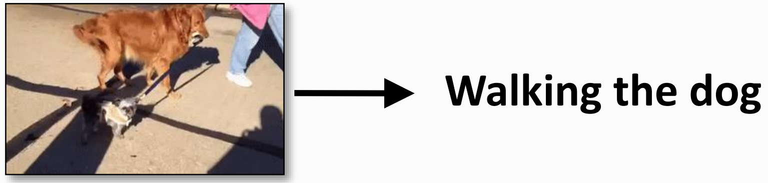


Networks	CPU runtime (ms)	GPU runtime (ms)
ResNet-152 (16 frames)	5,600	400
P3D-199 (16 frames)	1,500	150

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).

- ActivityNet Untrimmed Task
- Action Recognition

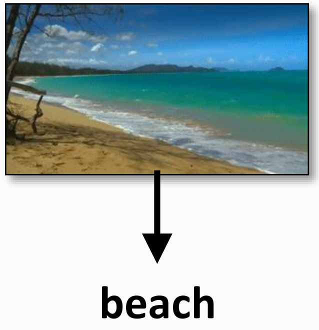


Method	Top-1	Top-3	MAP
IDT [INRIA, ICCV'13]	64.70%	77.98%	68.69%
C3D [FAIR, ICCV'15]	65.80%	81.16%	67.68%
VGG [U of Oxford, ICLR'15]	66.59%	82.70%	70.22%
ResNet [MSRA,CVPR'16]	71.43%	86.45%	76.56%
P3D ResNet	**75.12%**	**87.71%**	**78.86%**

- ASLAN
- Action Similarity Labeling

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%
P3D ResNet	**80.8%**	**87.9%**

- YUPENN, Dynamic Scene
- Scene Recognition



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%
P3D ResNet	**94.6%**	**99.5%**

This part

- Video representation learning
- Action recognition
- Video captioning
- Semantic video segmentation

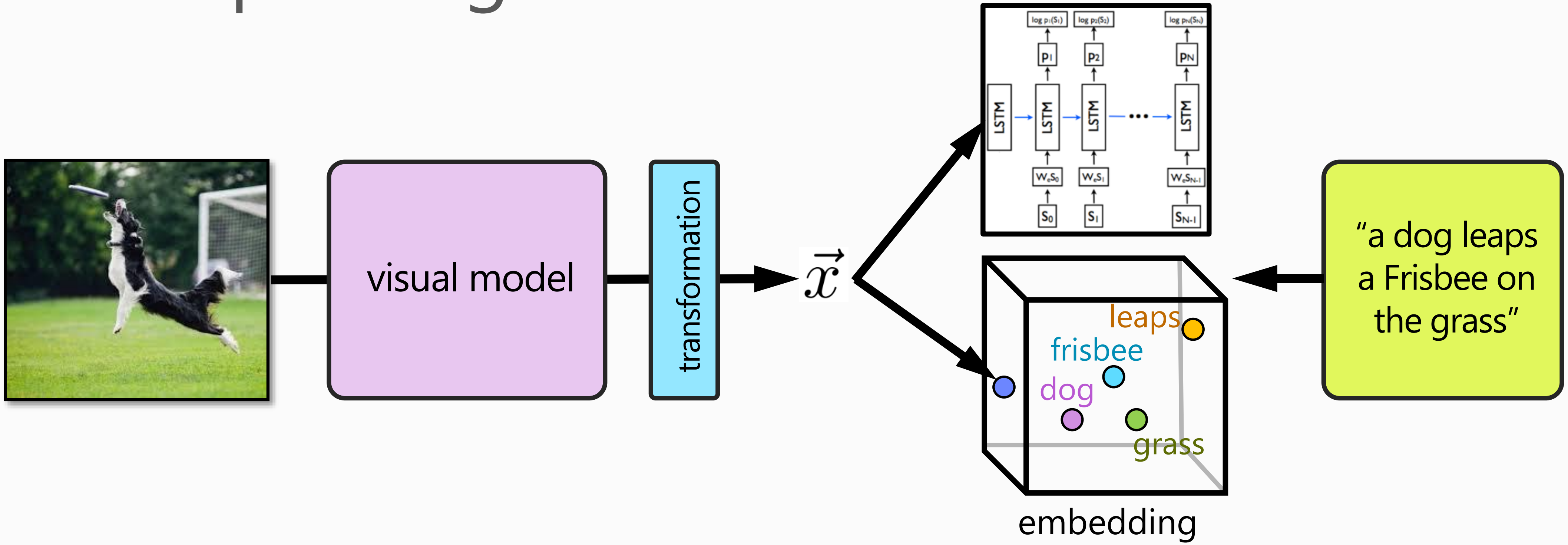
Vision to language: video captioning

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

"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]

Video captioning: basic idea

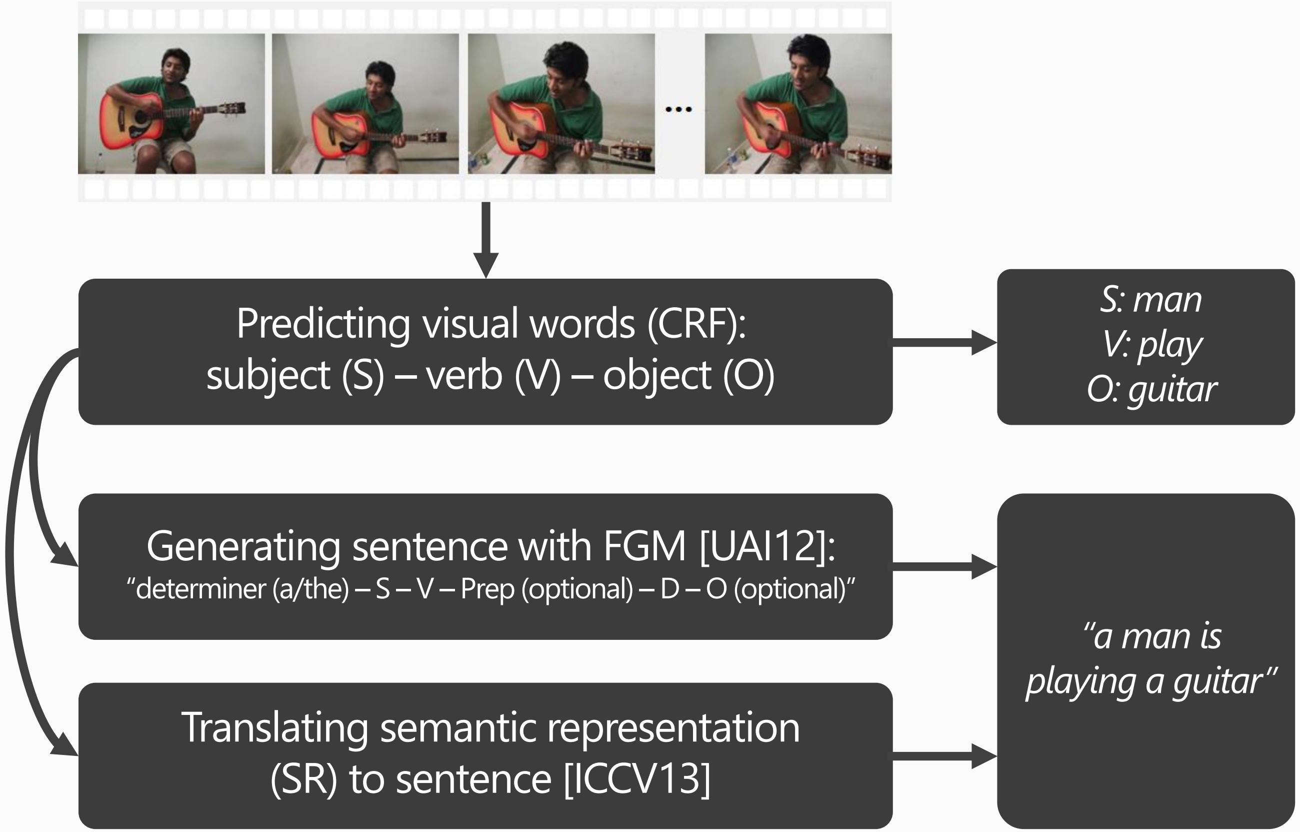


- Transforming an image/clip to a vector in visual space
  - CRF, CNN, Semantic Vector, CNN+Attention
- Transforming description to a vector in semantic space
  - Collection of words (BoW), sequence of words (RNN)
- Creating an embedding space
  - Language template (FGM, ME), RNNs (Encoder-Decoder), LSTM

- Methodologies
  - Search-based
  - Language template-based
  - Sequence learning-based
    - Generation: learning-decoder
    - Translation: encoder-decoder

Video captioning

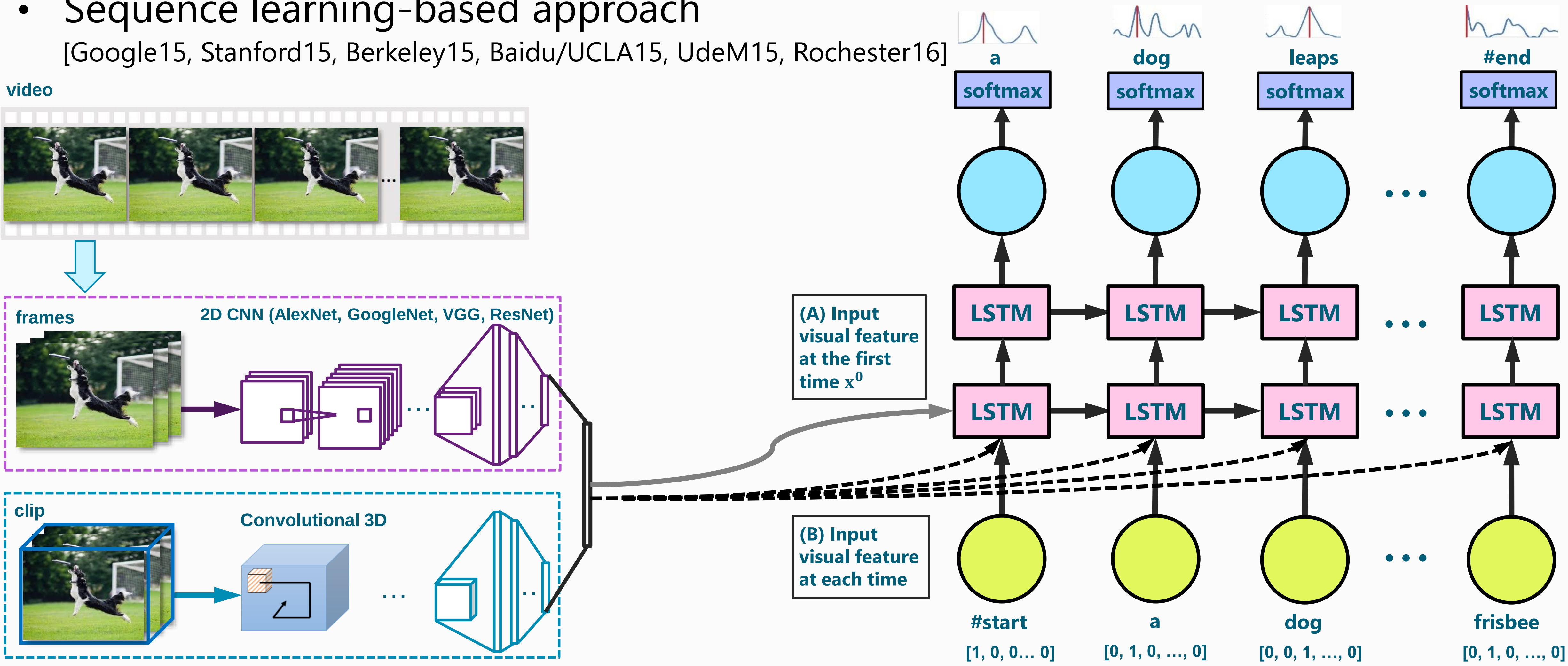
- **Language model-based approach** [Thomason, COLING14; Barbu, UAI12; Rohrbach, ICCV13; Krishnamoorthy, AAAI13]



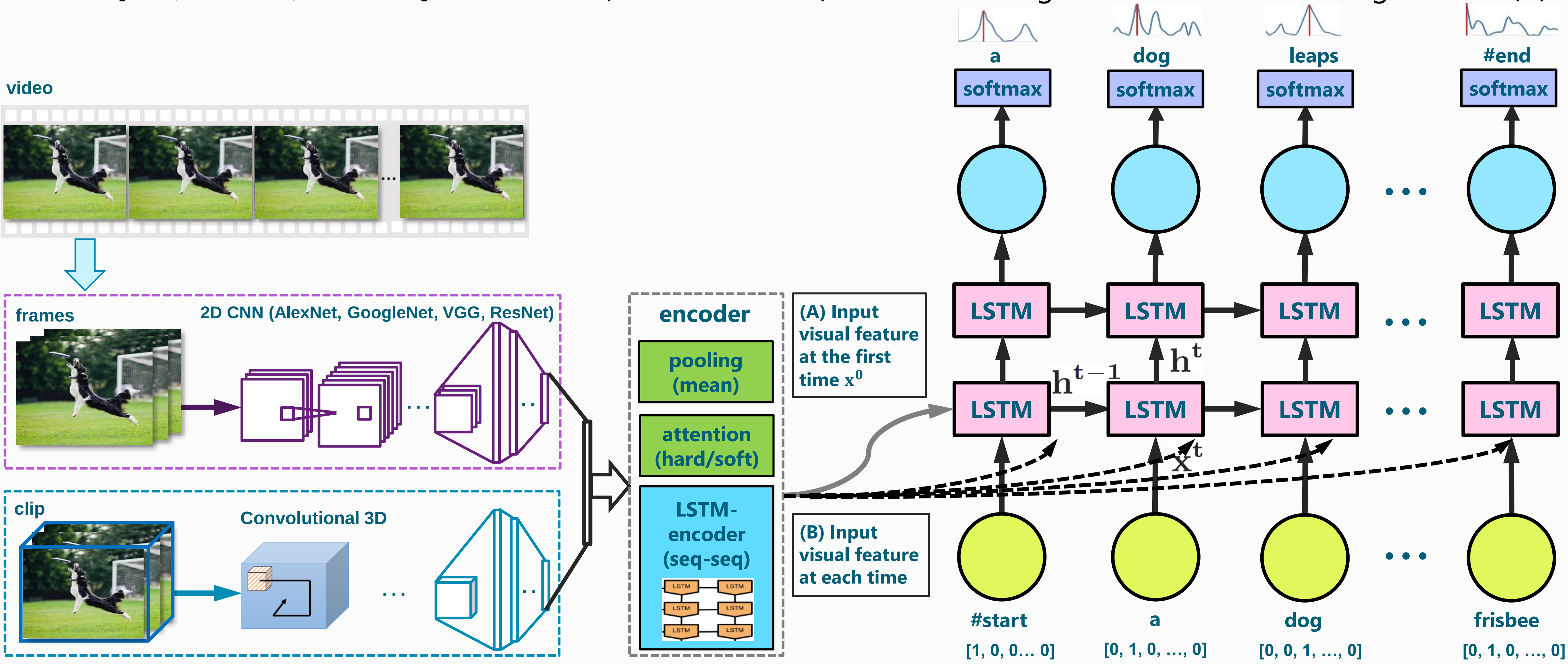
Barbu, et al. "Video In Sentences Out", UAI 2012.
https://www.youtube.com/watch?v=tu3jMxCJPMw

Video captioning

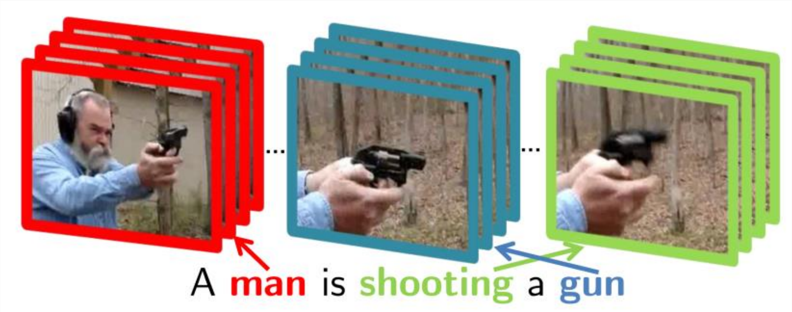
- Sequence learning-based approach
  [Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester16]



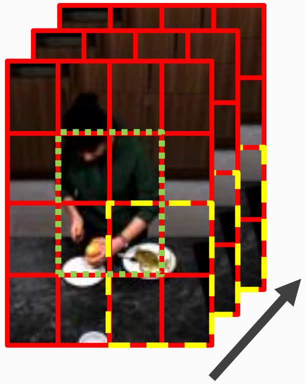
- UC Berkeley [Donahue, CVPR'15]: CRF + LSTM encoder-decoder + LSTM (A/B)
- UdeM [Yao, ICCV'15]: (GoogleNet + 3D CNN) + Soft-Attention + LSTM (B)
- UT Austin [Venugopalan, ICCV'15]: (VGG + Optical Flow) + LSTM Encoder-Decoder + LSTM (A)
- UT Austin [Venugopalan, NAACL-HLT'15]: AlexNet + Mean Pooling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



Video Captioning with **X**

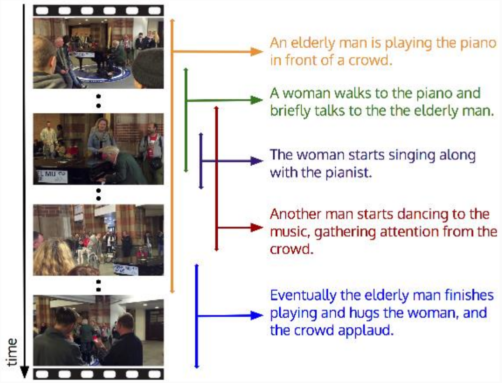


X = temporal attention
[Yao, CVPR'15]



X = spatiotemporal attention
[Yu, CVPR'16]

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



X = dense caption
[Krishna, ArXiv'17; Shen, CVPR'17]

Video Captioning with Semantics

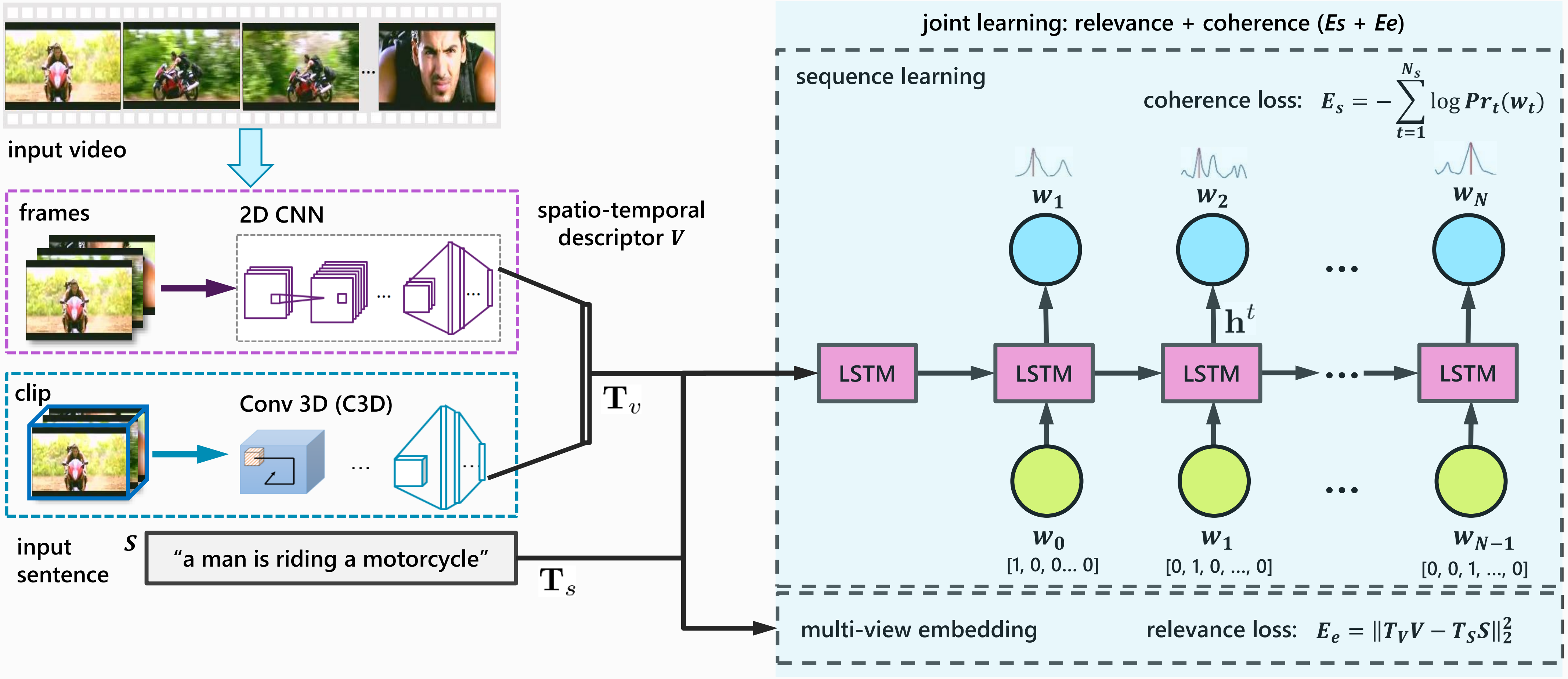
- Key issues in sentence generation
  - *relevance*: relationship between sentence (S, V, O) semantics and content
  - *coherence*: sentence grammar

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

Video captioning w/ LSTM-E [CVPR'16'17]



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

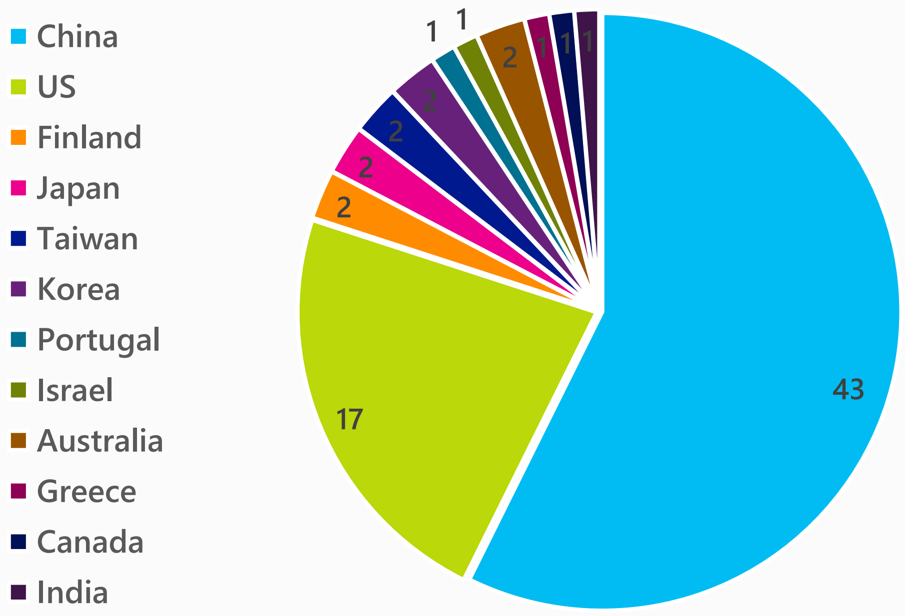
relevance coherence

Microsoft Video to Language Challenge 2016

77 teams registered challenge

22 teams submitted results

Awards will be announced at ACMMM



M1 M2

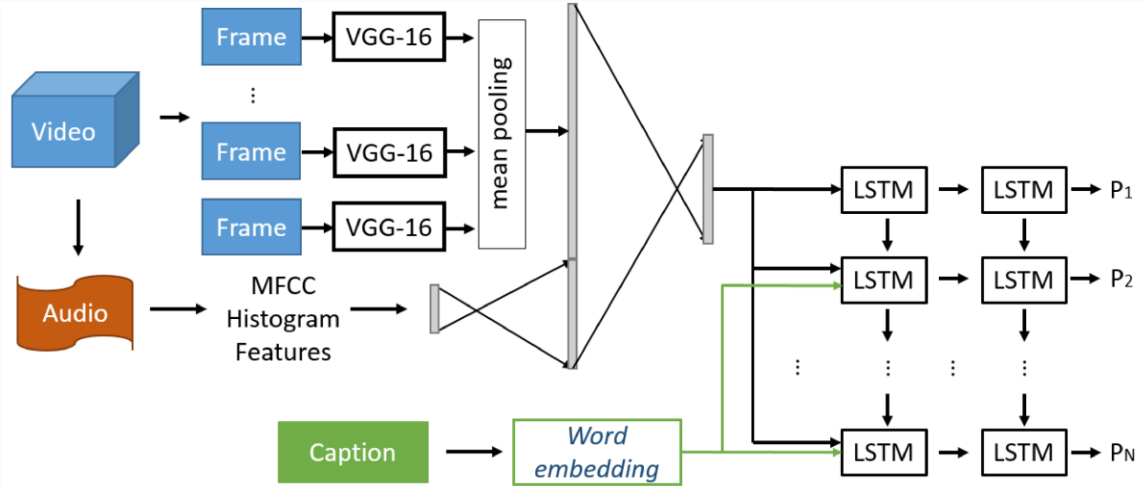
Rank	Team	Organization	BLEU@4	Meteor	CIDEr-D	ROUGE-L
1	v2t_navigator	RUC & CMU	0.408	0.282	0.448	0.609
2	Aalto	Aalto University	0.398	0.269	0.457	0.598
3	VideoLAB	UML & Berkeley & UT-Austin	0.391	0.277	0.441	0.606
4	ruc-uva	RUC & UVA & Zhejiang University	0.387	0.269	0.459	0.587
5	Fudan-ILC	Fudan & ILC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

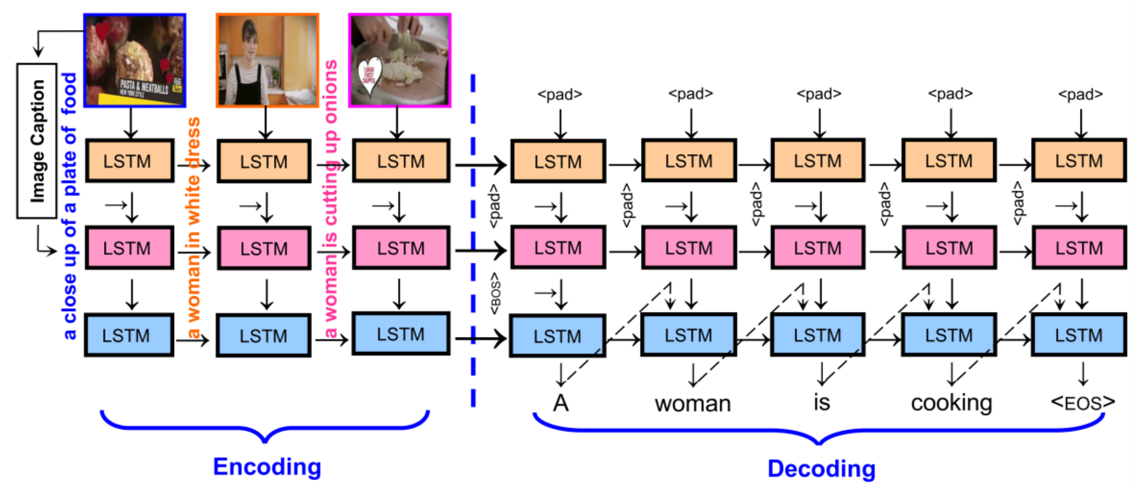
M1 M2

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]





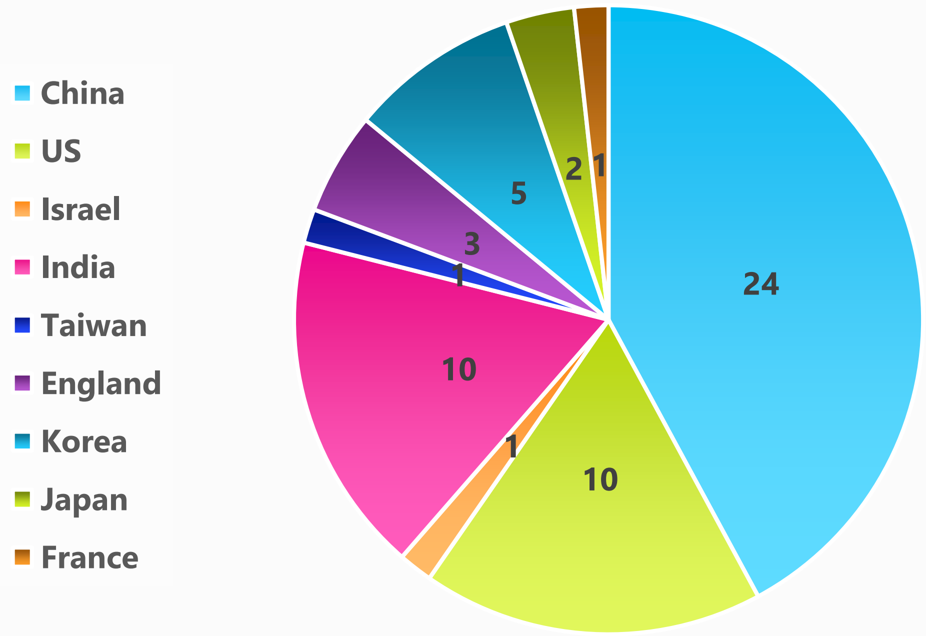
- Image features
  - VGG-19 [1][2][5][6][9][10]
  - GoogleNet [2][4][5]
  - ResNet [3][5][8]
  - VGG-16 [5][7][8]
  - PlaceNet [5][9]
- Motion features
  - C3D [1][2][3][4][5][9][10]
  - IDT [1][2]
  - Optical flow [8]
- Acoustic features
  - MFCCs [1][3][7]
- Text features
  - ASR [1]
- Video category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

Microsoft Video to Language Challenge 2017

57 teams registered challenge

8 teams submitted results

Awards will be announced at ACMMM'17



M1 M2

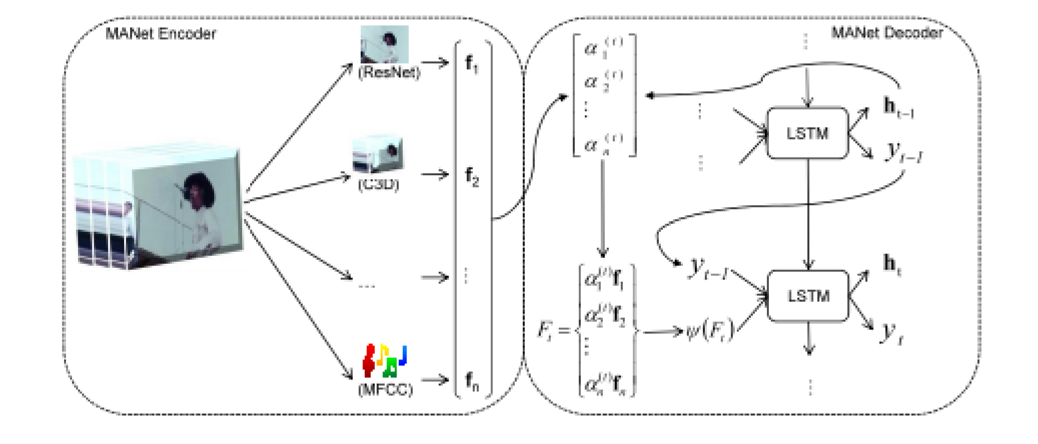
Rank	Team	Organization	BLEU@4	Meteor	CIDEr-D	ROUGE-L
1	RUC+CMU_V2T	RUC & CMU	0.390	0.255	0.315	0.542
2	TJU_Media	TJU	0.359	0.226	0.249	0.515
3	NII	National Institute of Informatics	0.359	0.234	0.231	0.514
4	MIC_TJU	Tongji University	0.351	0.226	0.236	0.509
5	Illusion	IIT Delhi	0.304	0.213	0.206	0.494
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

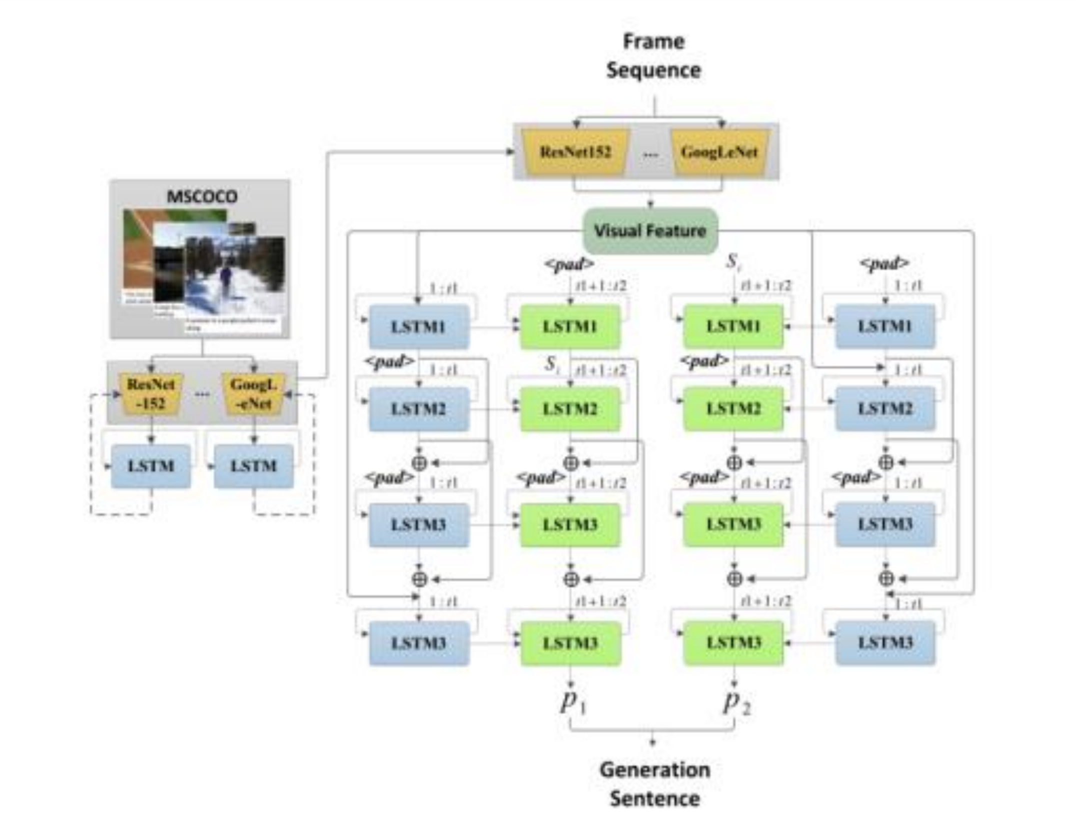
Rank	Team	Organization	C1	C2	C3
1	RUC+CMU_V2T	RUC & CMU	4.437	3.437	3.567
2	NII	National Institute of Informatics	4.078	3.359	3.570
3	TJU_Media	TJU	4.032	2.962	3.048
4	MIC_TJU	Tongji University	3.844	2.789	2.978
4	Illusion	IIT Delhi	4.042	2.583	2.921
6	TJU-NUS	TJU & NUS	3.762	2.364	2.376
7	AFRL	AFRL	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]

Vision to language: auto-commenting [Li, MM'16]

* 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

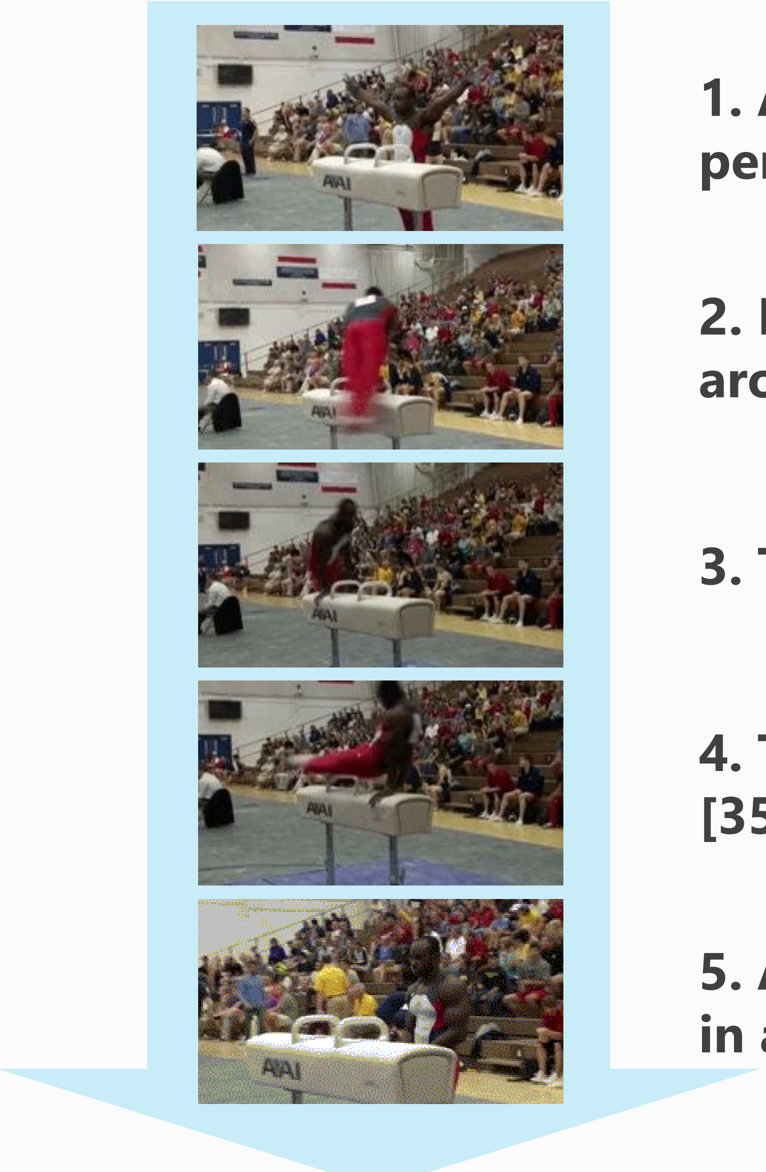
* Little cutie 0.4643
* The hat is so cute 0.4201
* The eyes are so beautiful. It's too cute and I love it so much 0.4102
* Baby looks so handsome with the hat on. So cute 0.3950
* Such a cute little baby 0.3927

* 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

Dense video captioning w/ P3D [Yao & Mei, CVPR'17]



1. An athletic man is seen standing before a beam and begins performing a gymnastics routine. [1.997, 11.198]

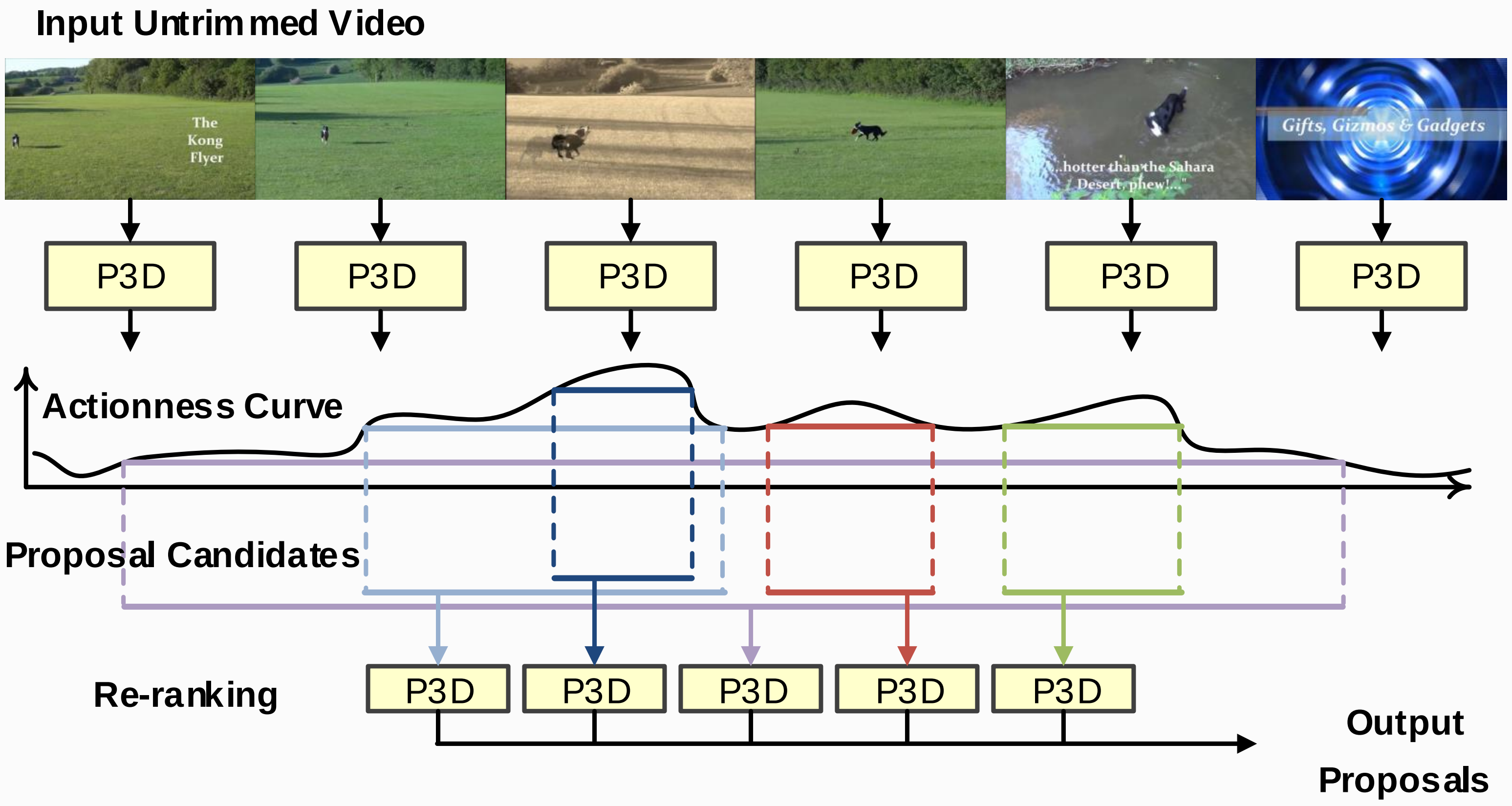
2. He then performs a gymnasts routine while swinging himself all around the bar and ends by jumping down. [19.723, 43.410]

3. The man is then shown flips on the bars. [25.841, 40.337]

4. The man is then shown on the bars and jumping off the bars. [35.758, 46.546]

5. A man is seen standing on a set of bars and performing a routine in a gym on the bars. [0.000, 47.056]

Event localization: actionness detection + grouping + re-ranking

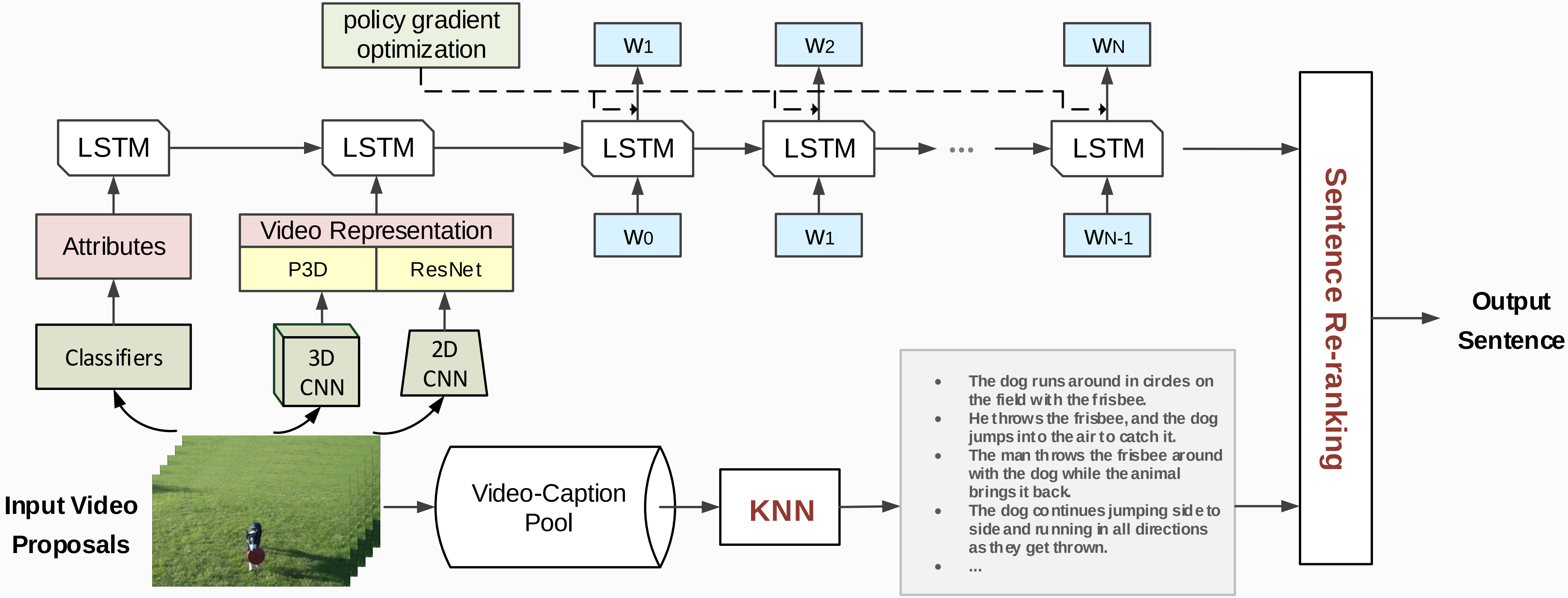


Performance on validation set in temporal action proposal task

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**

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

Dense video captioning w/ P3D [Yao & Mei, CVPR'17]

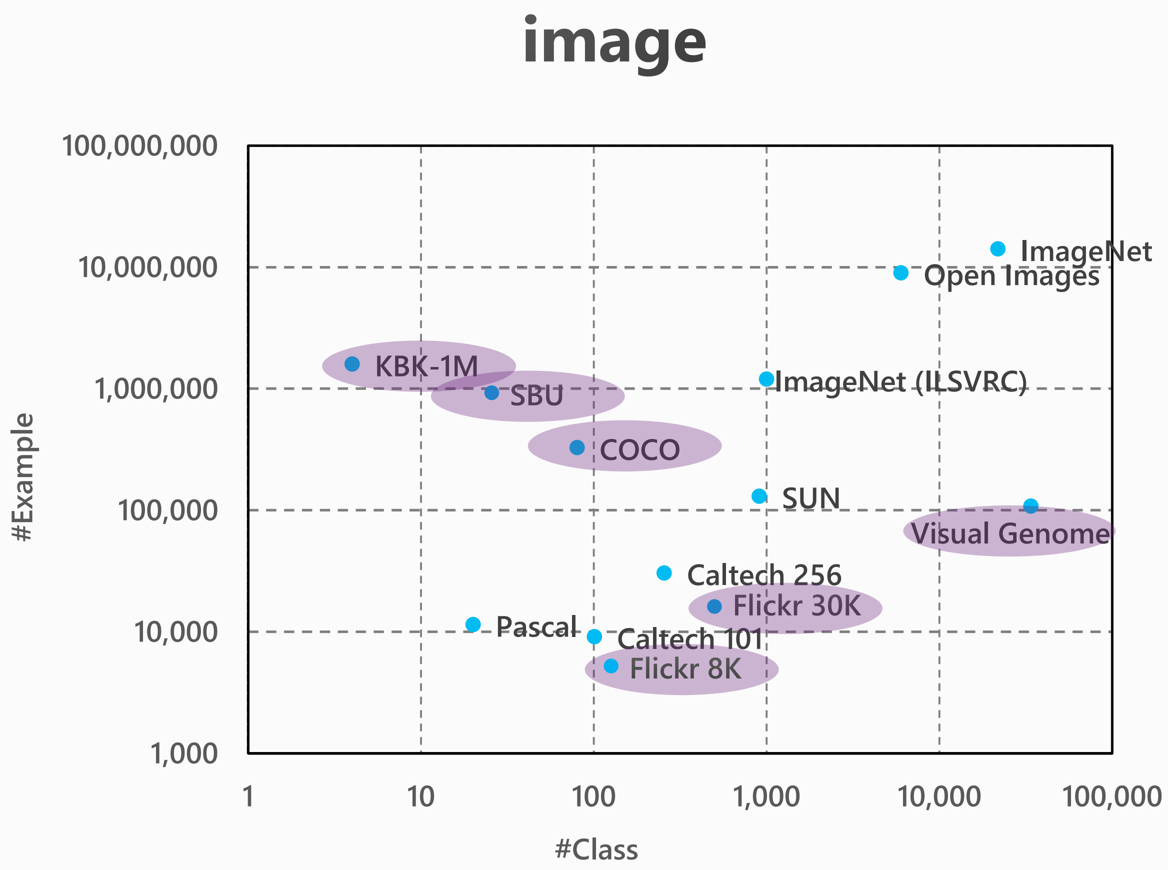


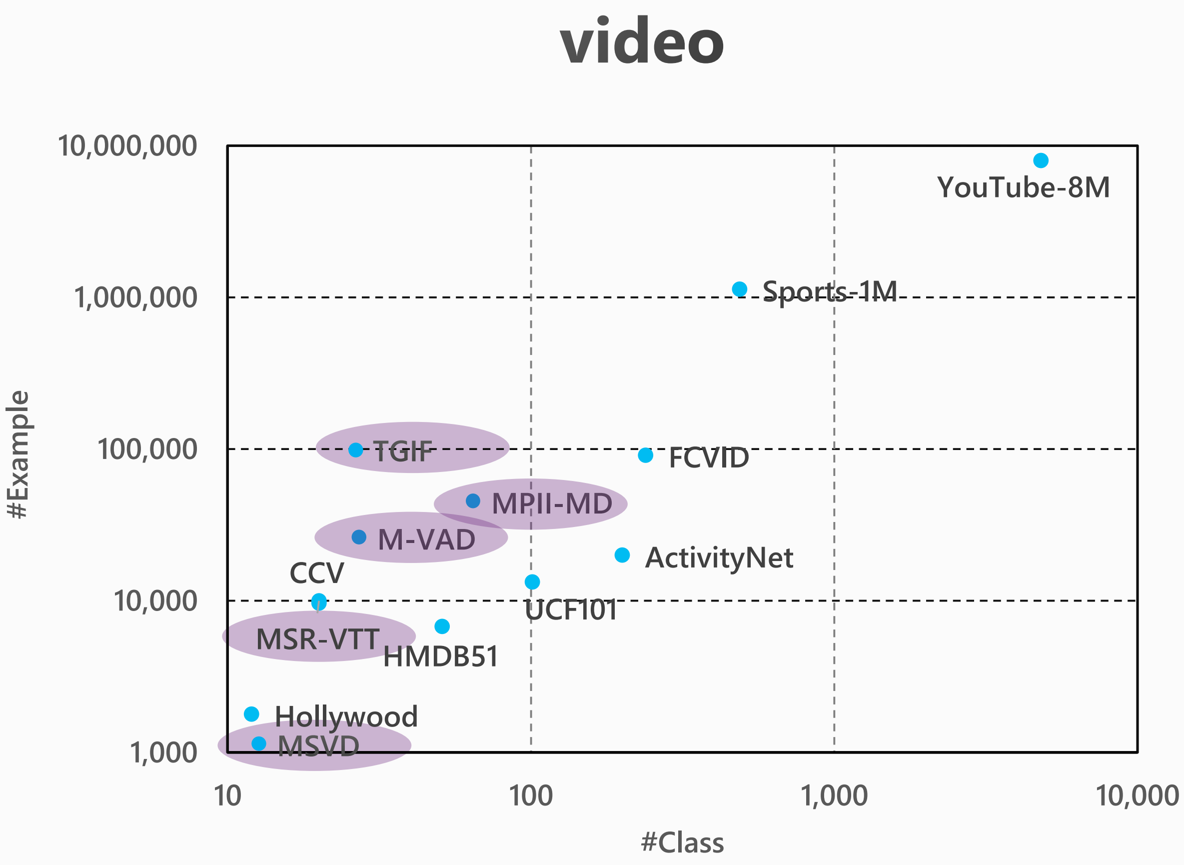
Rank 1 in the ActivityNet challenge 2017

- 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
- Video representation: ResNet (Kinetics) + P3D ResNet
- Attributes: 200 categories in the untrimmed video classification dataset

#	Team	METEOR%
1	**Microsoft Research Asia**	**12.84**
2	University of Science and Technology of China	9.87
3	Renmin University of China & Carnegie Mellon University	9.61
4	Stanford University	4.82

Datasets for Image/Video Captioning





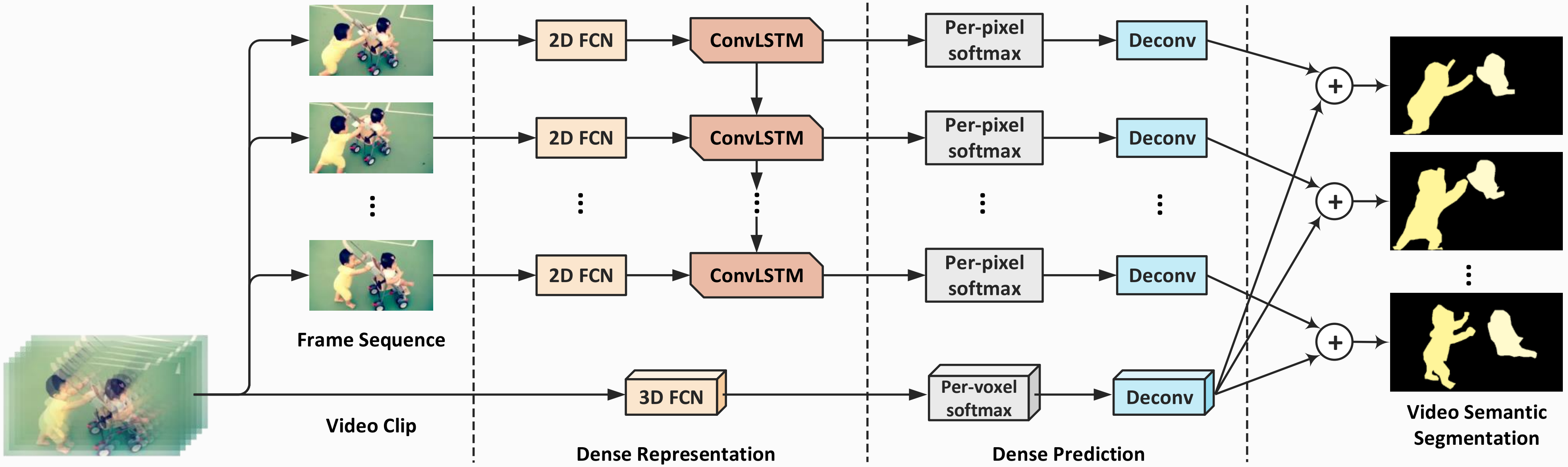
Dataset for captioning.

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

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]

Video segmentation with P3D [Qiu, TMM'17]

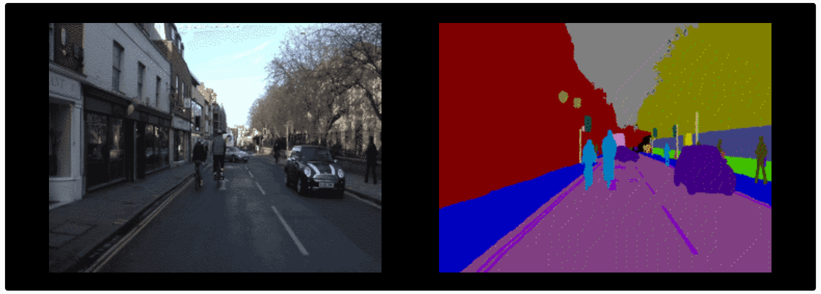


Learning Spatiotemporal Dependency for Semantic Video Segmentation

- Frame sequence: 2D FCN to learn spatial dependency + ConvLSTM to learn sequential information
- Video clip: 3D FCN to learn voxel-level spatio-temporal dependency

Video segmentation with P3D

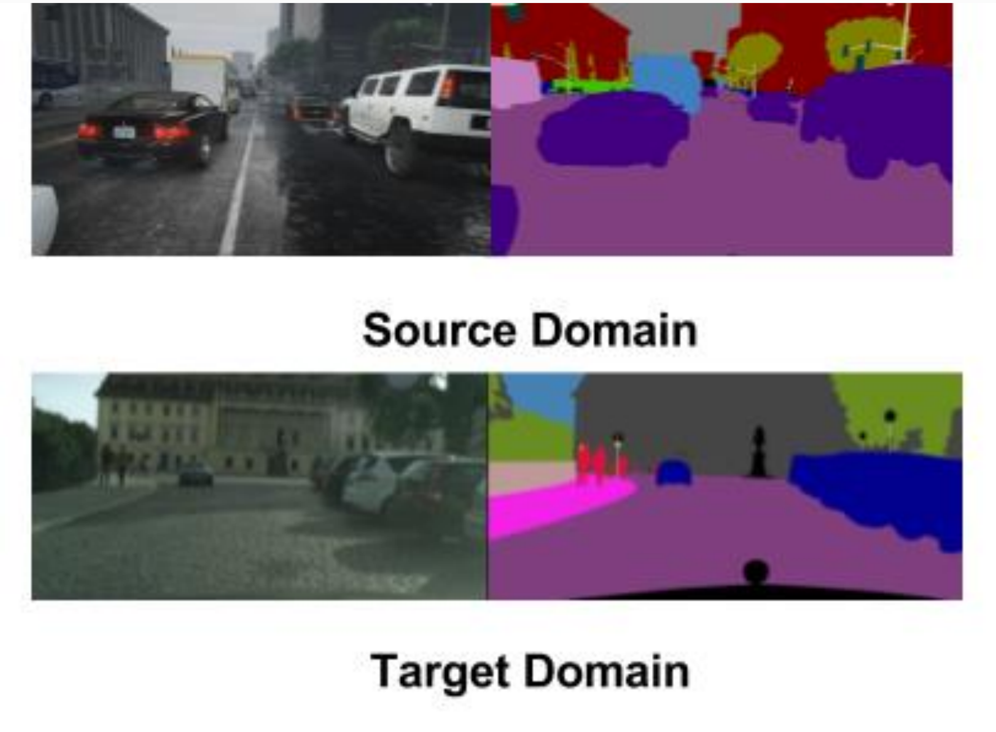
- CamVid dataset
  - 11 class labels
  - 701 labeled frames in 5 videos
- A2D dataset
  - 7 actor + 9 actions
  - 3,782 videos



CamVid	Pix-Acc	mIoU
Active Inference [Liu, CVPR'15]	82.8 %	47.2 %
FSO [Kundu, CVPR'16]	-	66.1 %
Dilation8 [Yu, ICLR'16]	-	65.3 %
2D FCN	91.8 %	66.4 %
2D FCN + LSTM	92.0 %	68.1 %
3D FCN	89.7 %	62.2 %
DST-FCN	92.2 %	68.8 %

A2D	Pix-Acc	mIoU
Trilayer [Xu, CVPR'15]	72.9 %	--
GPM [Xu, CVPR'16]	83.8 %	--
2D FCN	91.6 %	25.1 %
2D FCN + LSTM	92.5 %	29.9 %
3D FCN	91.3 %	28.0 %
DST-FCN	93.0 %	33.4 %

Leaderboard of segmentation challenge at ICCV'17



#	Team Name	Affiliation	Score
1	**RTZH**	**Microsoft Research Asia**	**47.5**
2	_piotr_	University of Oxford, Active Vision Laboratory	44.7
3	whung	University of California, Merced, Vision and Learning Lab	42.4

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

- Codes for P3D
  - https://github.com/ZhaofanQiu/pseudo-3d-residual-networks

- 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

- 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

Thank you! We are hiring!

tmei@microsoft.com