Recognition on videos

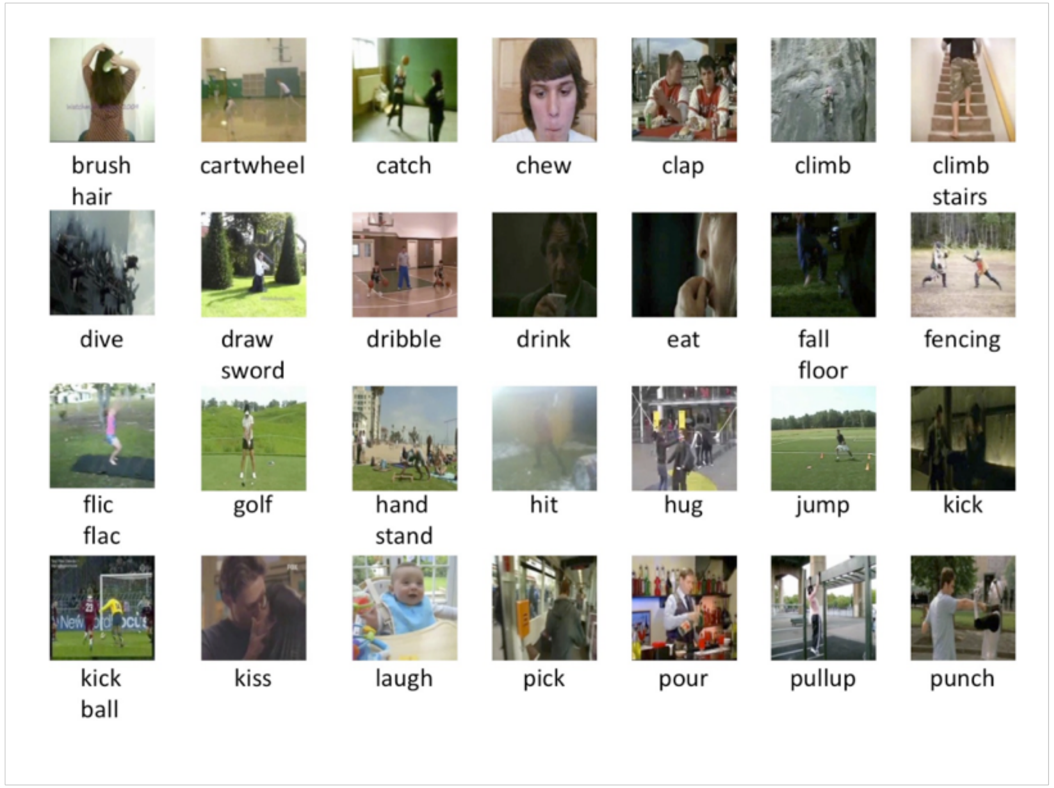
Video classification

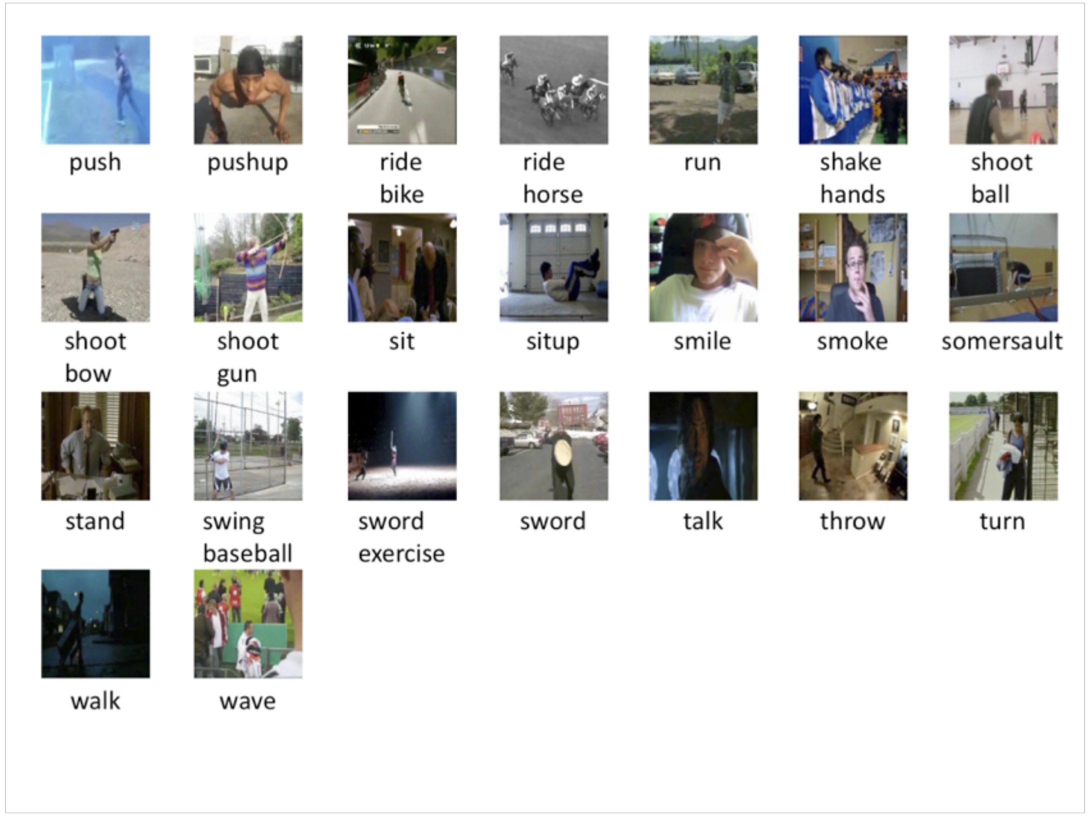
- Significantly harder to collect data
- 1 video = hundreds of frames
- UCF 101 : 13k videos from 101 actions
- HMDB: 7k videos from 51 actions

UCF 101

Recognition of 101 human actions from videos in the wild, Khurram Soomro, Amir Roshan Zamir and Mubarak Shah,
In arXiv preprint arXiv:1212.0402, November, 2012

HMDB 51





H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. ICCV, 2011.

Is a video just a collection of images?

Is a video just a collection of images?

Is a video just a collection of images?

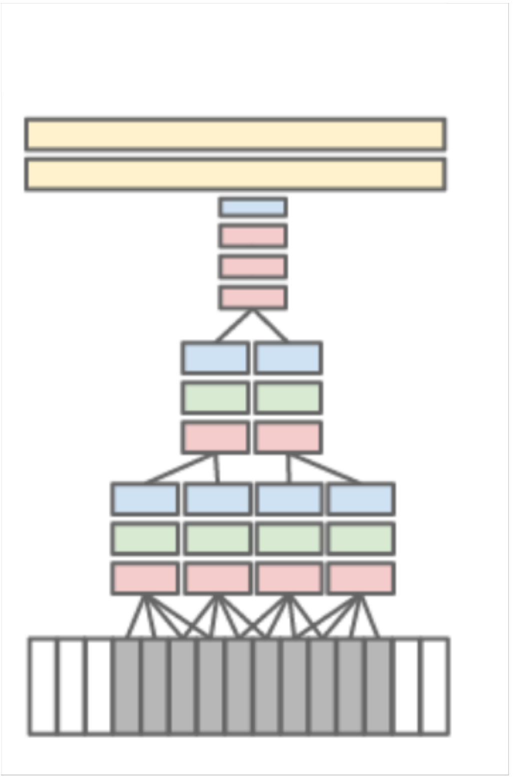
Is a video just a collection of images?

Is a video just a collection of images?

- Mostly a question of classes involved
- Actions involving object interactions can be identified from single image
  - e.g., riding horse, playing tennis, pitching ball
- Actions that occur in particular scenes can be identified from single images
  - e.g., diving, swimming, rock-climbing

Video recognition with convolutional networks

Extending convolutional networks to 3D



Extending convolutional networks to 3D

Method	Accuracy
Best non-deep approach	85.9%
Temporal convolution	65.4%

The challenge of applying convnets to videos

- Each video = multiple images: memory constraints
- Cost of convolution increases with dimension: time constraints

Convnets and Videos

- Can’t use pretraining
- Can’t train with long videos
- 3D convolution too expensive

A step back: frame-based convnets

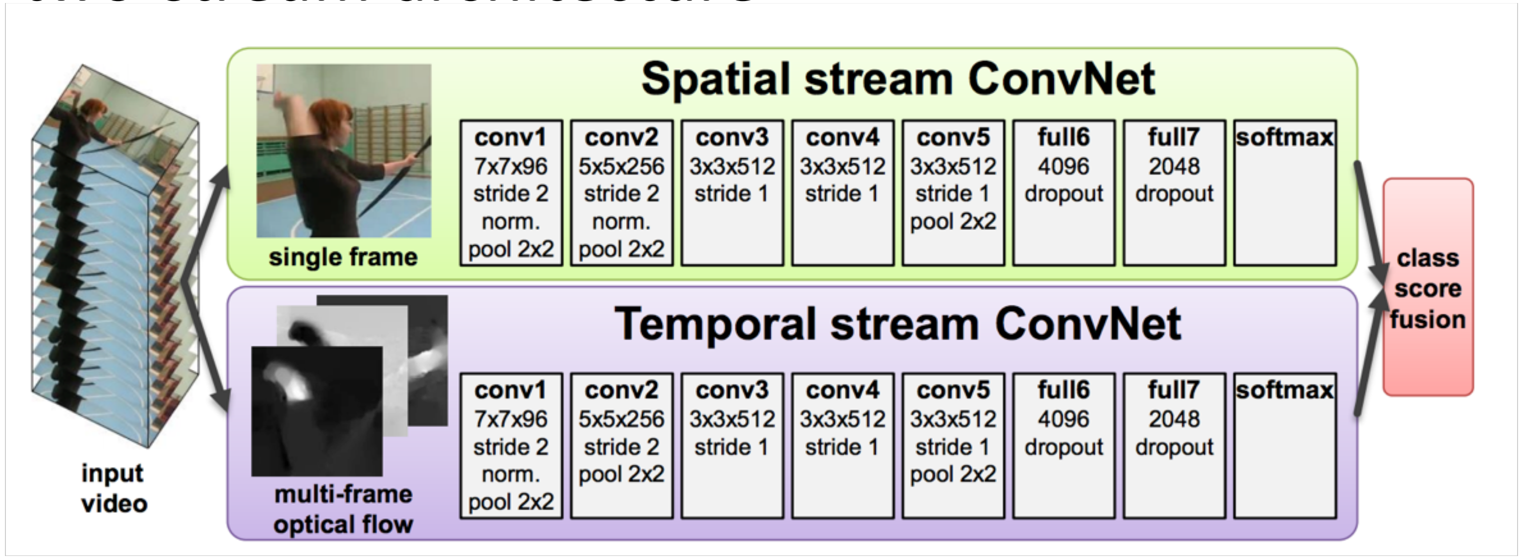
- Idea: train single-frame classifier
  - Can use pretraining
  - Can sample frames from videos
  - Cheap

- *Average* scores over frames

Method	Accuracy (UCF 101)
Best classical approach	85.9%
Temporal convolution	65.4%
Avg-of-frame-scores	72.8%

Bringing back motion

- How do we take motion into account?
- Key idea: use optical flow
- Add flow as additional channel?
  - Can't use pretraining!
- Idea: Use *two-stream* architecture



Bringing back motion

Method	Accuracy (UCF 101)
Best dense trajectories + bag-of-words	85.9%
Temporal convolution	65.4%
Avg-of-frame-scores (color)	72.8%
Avg-of-frame-scores (flow)	81.2%
Avg-of-frame-scores (both)	86.2%

Two-Stream Convolutional Networks for Action Recognition in Videos. Karen Simonyan and Andrew Zisserman. In *NIPS,* 2014.

Beyond frame averaging

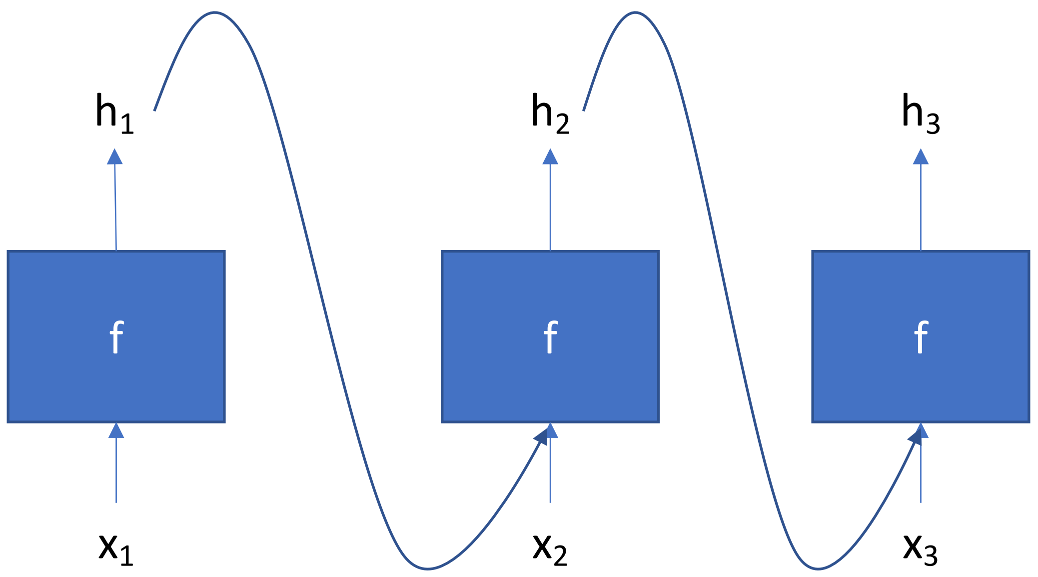
- Do better than averaging frames?
  - without increasing cost!
- Idea: video is *sequence* of frames
- Convnet converts each frame into feature vector
- Sequence of feature vectors → class scores?

Detour: Sequence modeling

Sequence modeling

- Internal state *h*
- Observations *x*
- See observations over time
- Update state

$$h_t = f(x_t, h_{t-1})$$



Recurrent Neural Networks (RNNs)

Problem: vanishing/exploding gradients

$$h_t = f(x_t, h_{t-1})$$

$$\frac{\partial z}{\partial h_{t-1}} = \frac{\partial f(x_t, h_{t-1})}{\partial h_{t-1}} \frac{\partial z}{\partial h_t}$$

“Long Short-term Memory”

- Instead of repeated applications of arbitrary function f
- Have repeated application of something that does not decay or scale up
  - Addition!

(Not) LSTM - 1

$$\cancel{h_t = f(x_t, h_{t-1})}$$

$$h_t = h_{t-1} + x_t$$

$$\frac{\partial z}{\partial h_{t-1}} = \frac{\partial z}{\partial h_t}$$

(Not) LSTM - 2

$$\cancel{h_t = f(x_t, h_{t-1})}$$

$$h_t = f(c_t)$$
$$c_t = c_{t-1} + x_t$$

(Not) LSTM - 3

$$\cancel{h_t = f(x_t, h_{t-1})}$$

$$h_t = f(c_t)$$

$$c_t = c_{t-1} + g_t$$

$$g_t = g(x_t)$$

(Not) LSTM - 3

$$\cancel{h_t = f(x_t, h_{t-1})}$$

$$h_t = f(c_t)$$

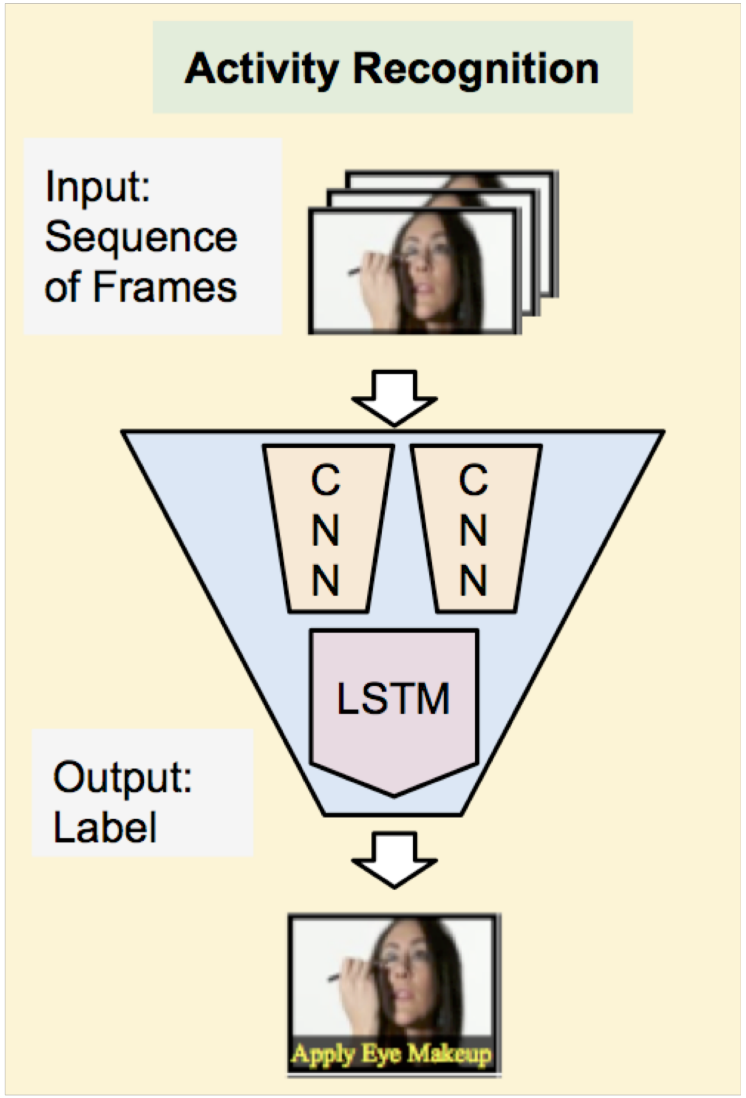
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$f_t = \sigma(W[h_{t-1}, x_t]) \qquad i_t = \sigma(V[h_{t-1}, x_t])$$

Long Short-Term Memory. Sepp Hochreiter and Jurgen Schmidhuber. In *Neural Computation*

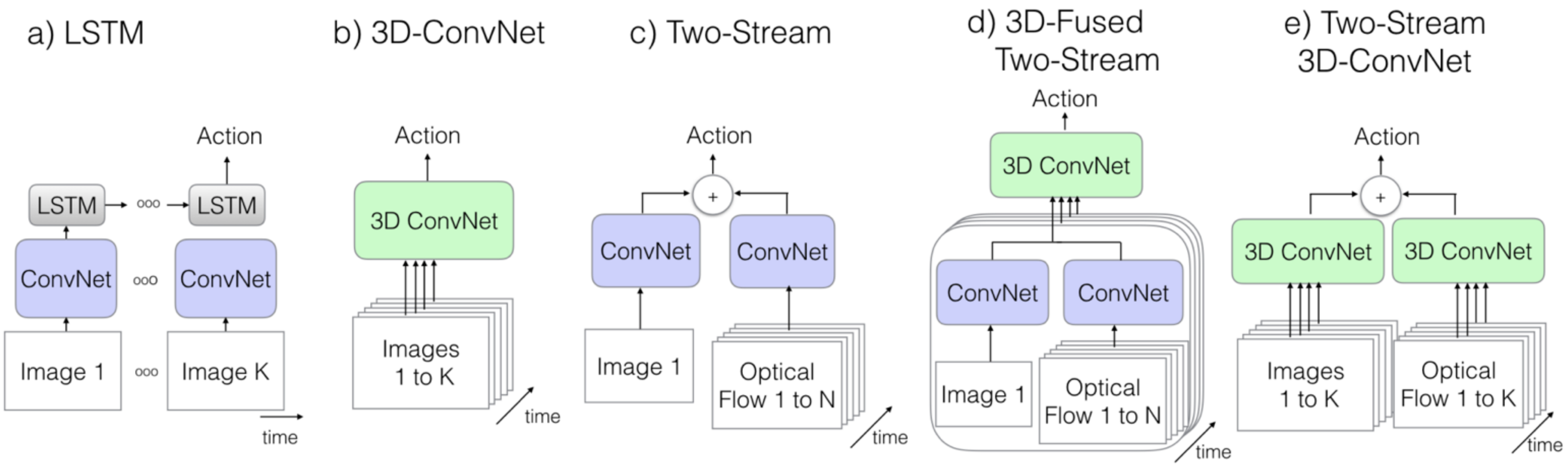
Using LSTMs for frame accumulation

Method	Accuracy (UCF 101)
Simple average	78.9%
LSTM	82.3%



Long-term Recurrent Convolutional Networks. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. In *CVPR,* 2015.

Revisiting video classification architectures

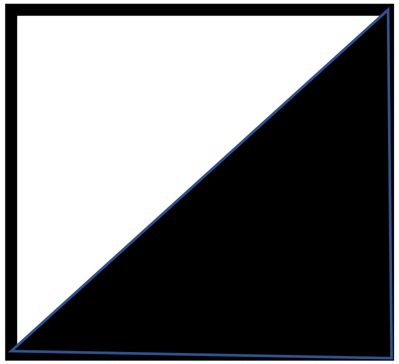


Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. J. Carreira and A. Zisserman. In *CVPR,* 2017.

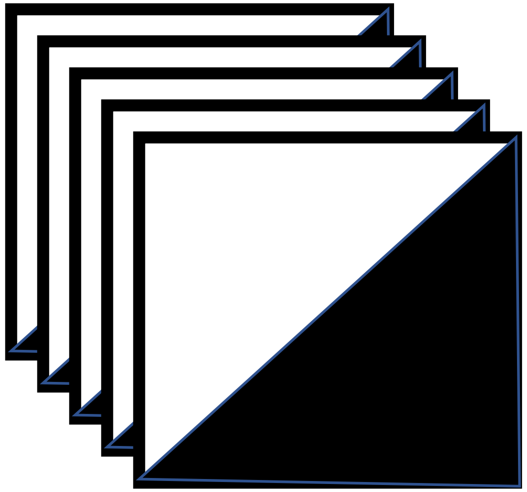
Revisiting video classification architectures

Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

Pre-training 3D convolutional networks



Pre-training 3D convolutional networks



Revisiting video classification architectures

Architecture	UCF-101		
	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–
(b) 3D-ConvNet	51.6	–	–
(c) Two-Stream	83.6	85.6	91.2
(d) 3D-Fused	83.2	85.8	89.3
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**

Towards better video datasets

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

The Kinetics Human Action Video Dataset. Kay, Will et al. Arxiv 2017.

Other video problems

- Action “detection / localization” in time
- Temporally consistent object detection / semantic segmentation
- Tracking

Vision and Language

Image captioning - The task

A group of young men playing soccer.

Image captioning - why?

- Alt-text for visually impaired
- Test for true understanding?

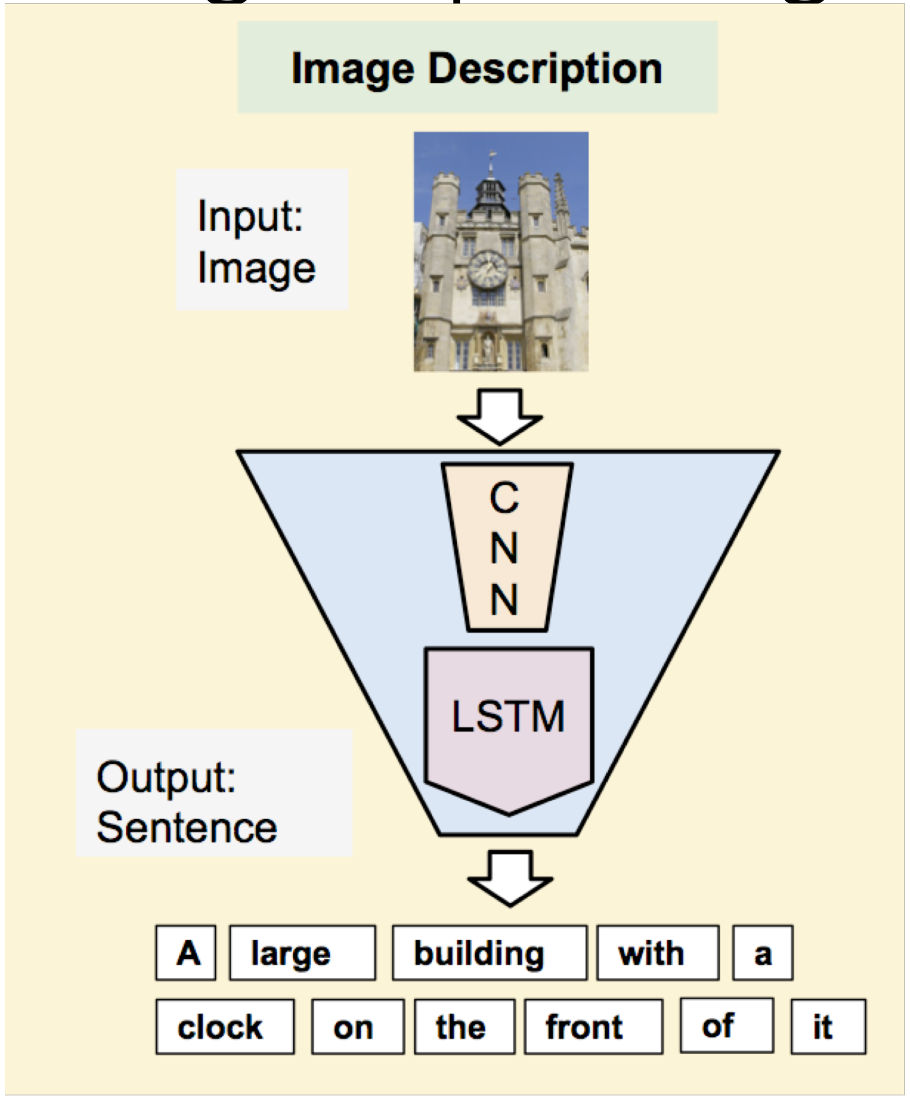
Image captioning - evaluation

- Given computer-generated caption and human caption, compute match
- BLEU from machine translation community
- Computes (modified) n-gram precision

Reference: A group of people playing soccer

Candidate: People playing baseball.
BLEU: 1/3

Image captioning

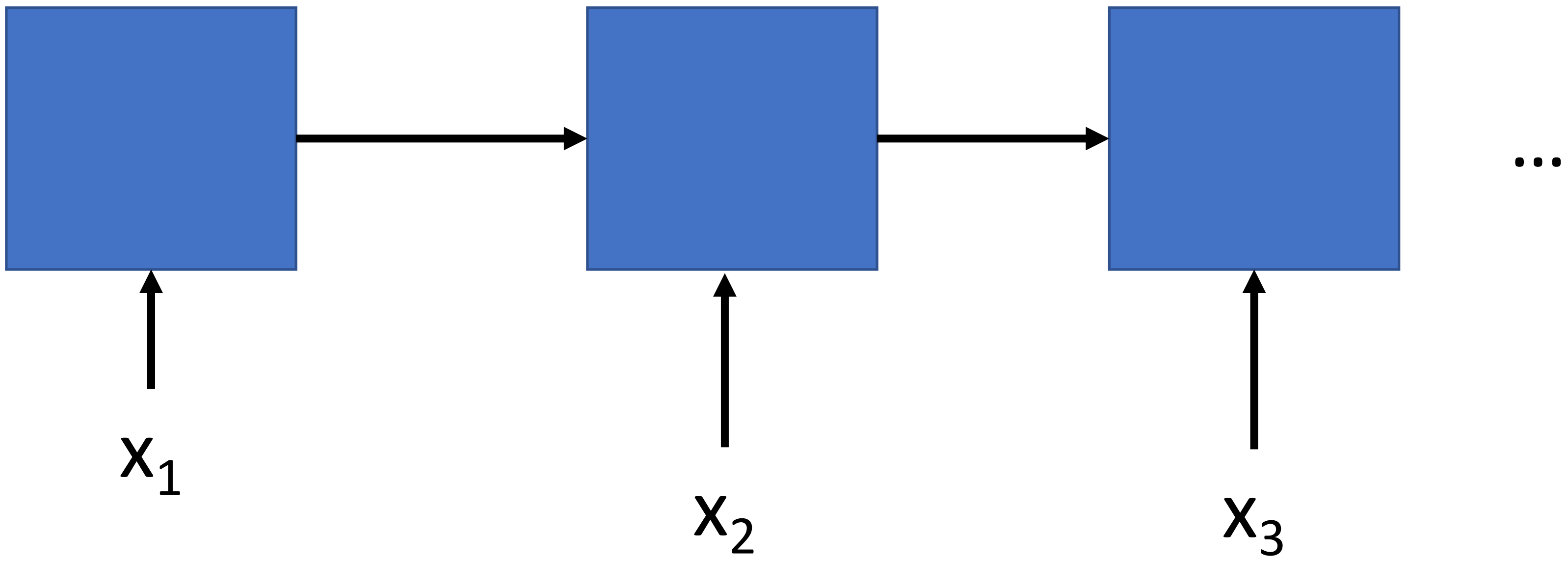


Long-term Recurrent Convolutional Networks. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. In *CVPR,* 2015.

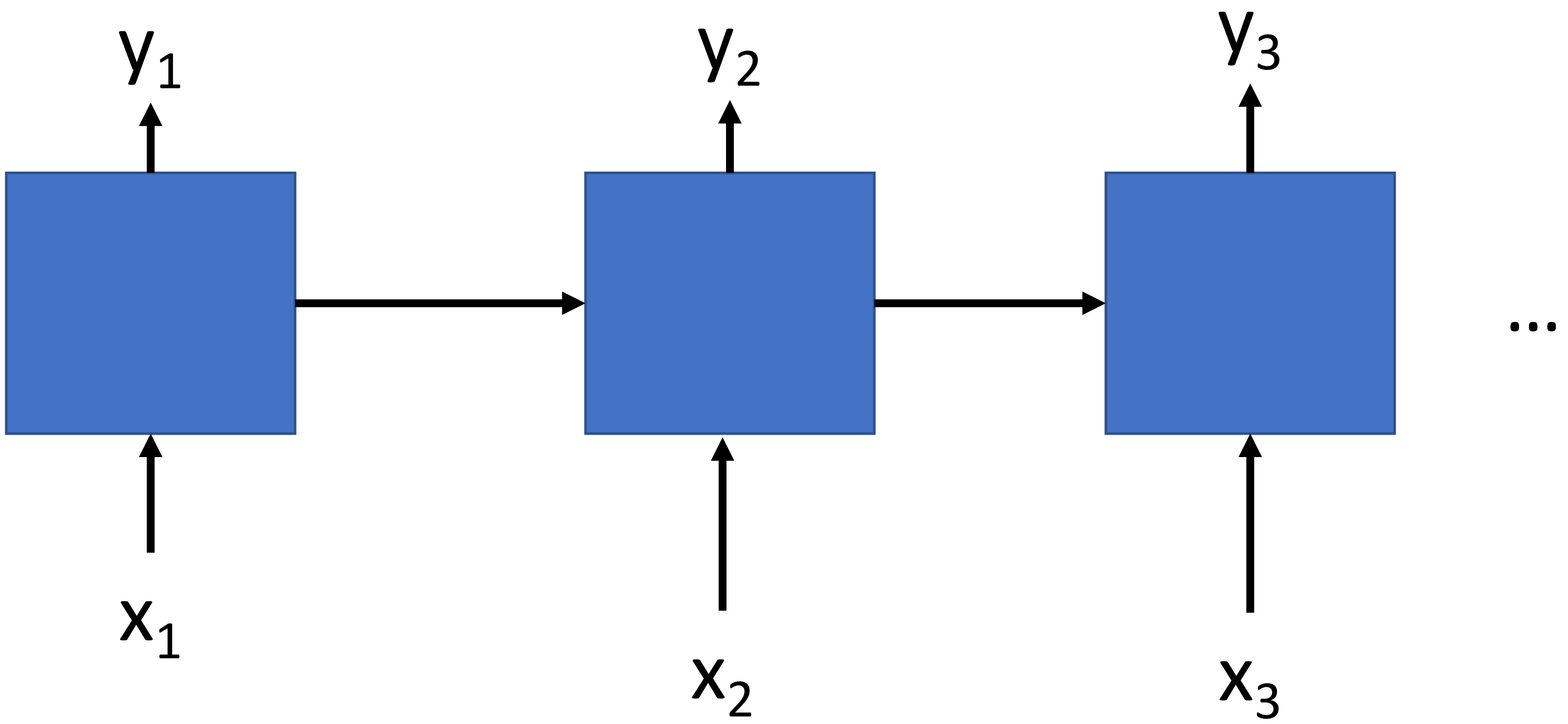
Deep Visual-Semantic Alignments for Generating Image Descriptions. Andrej Karpathy and Li Fei-Fei. In *CVPR,* 2015

Show and tell: A neural image caption generator
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan.
In *CVPR,* 2015.

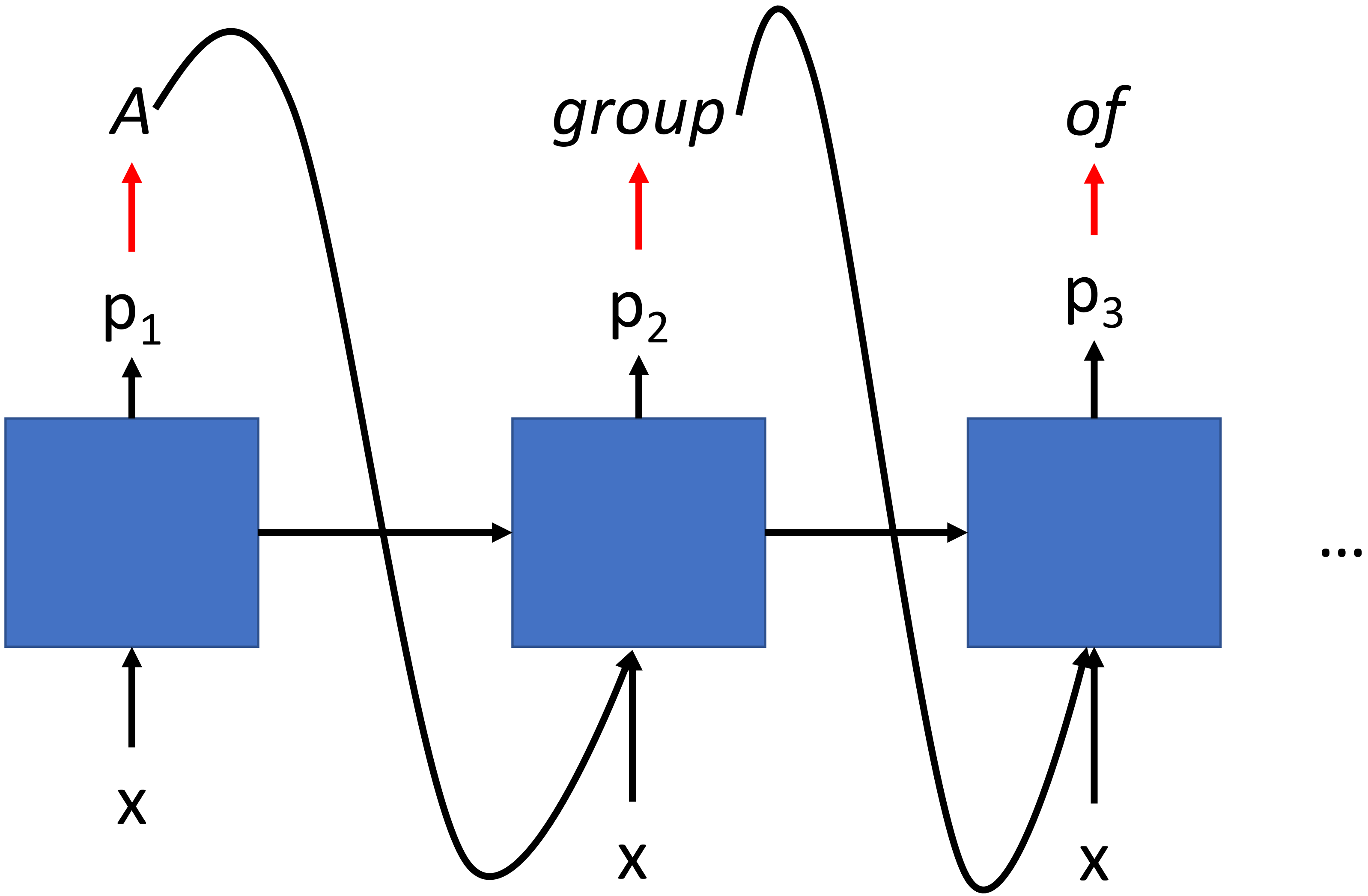
Generating sequences with LSTMs



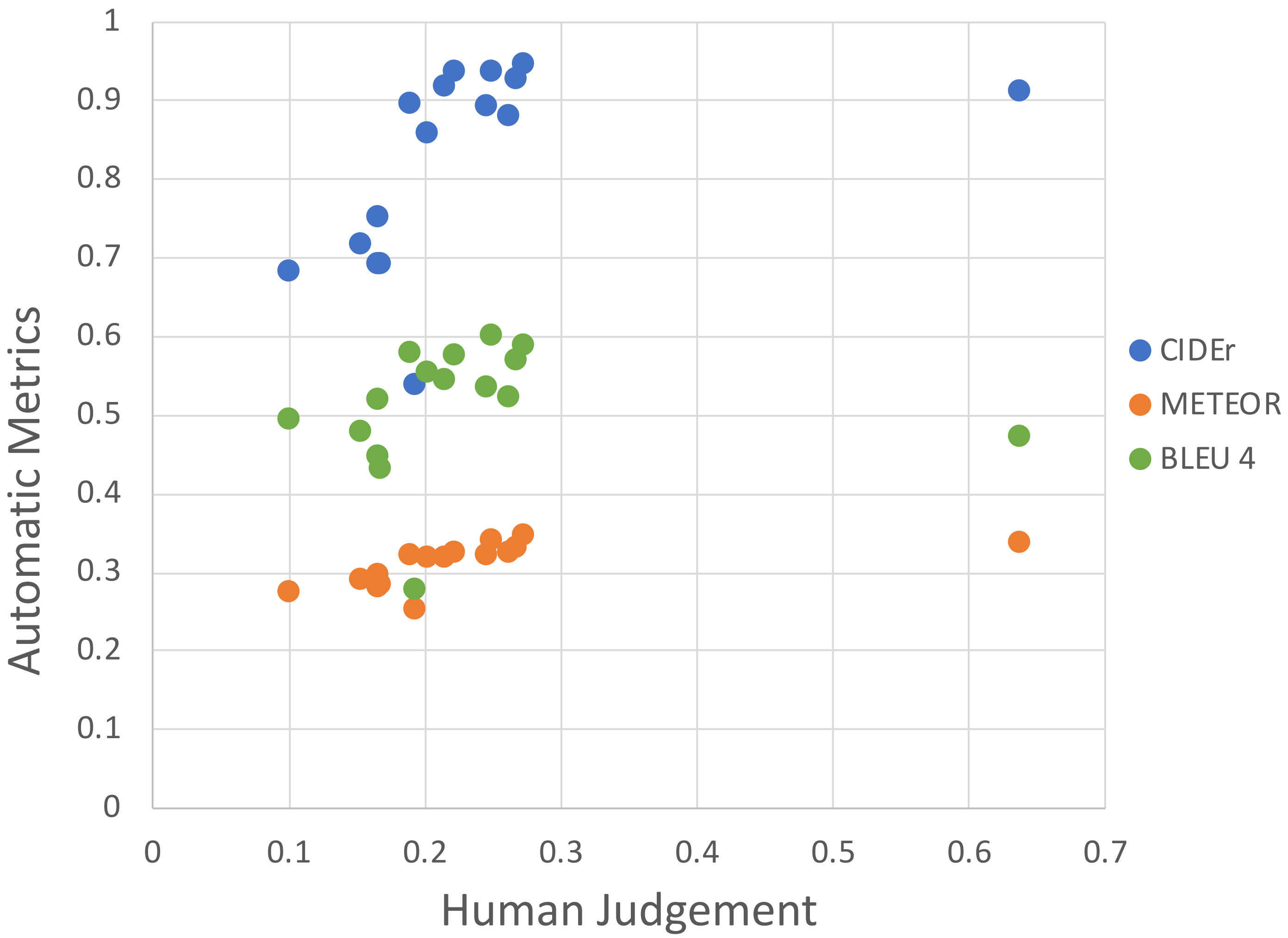
Generating sequences with LSTMs



Generating sequences with LSTMs

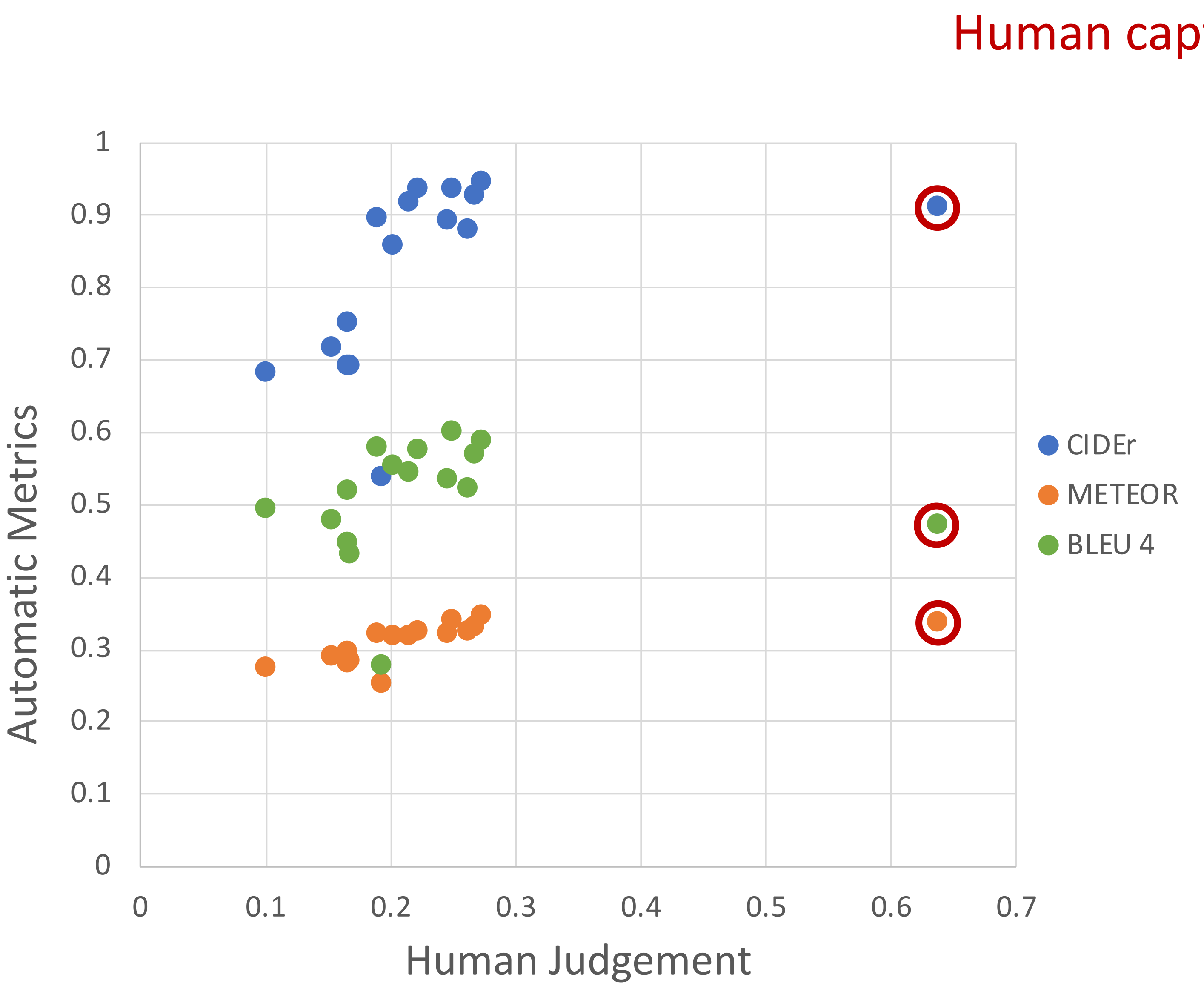


Evaluation Metrics



Slide credit: Larry Zitnick

Evaluation Metrics



Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

“surfboard”

Slide credit: Larry Zitnick

The post-captioning world

- Captioning is hard to evaluate!
  - Frame task so that it is easy to evaluate objectively
- Datasets are biased!
  - Control dataset bias

Stephanie Melnick
@unicornsteph96

Follow

I'm going to crush the rebellion... but first, let me take a selfie. #captionbot

I am not really confident, but I think it's a man taking a selfie in front of a building.

Reasoning

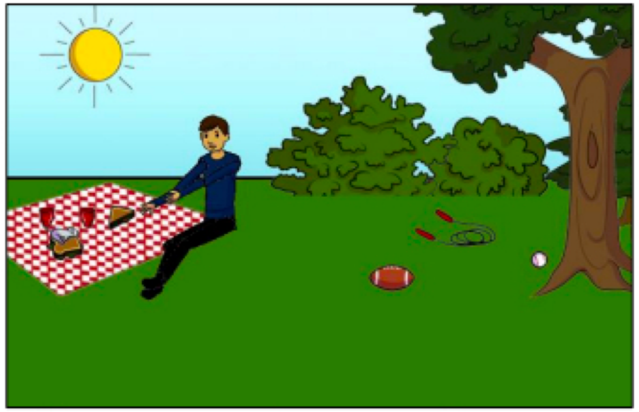
- Want vision systems to reason about what is going on
  - Identify objects and scenes
  - Identify relationships between objects
  - Understand physics of the world
  - Understand social interactions, intent etc.
  - Incorporate knowledge: common sense, pop culture, …

Visual Question Answering

- Direct motivation: assistive technology
- Indirect motivation: sandbox for reasoning

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

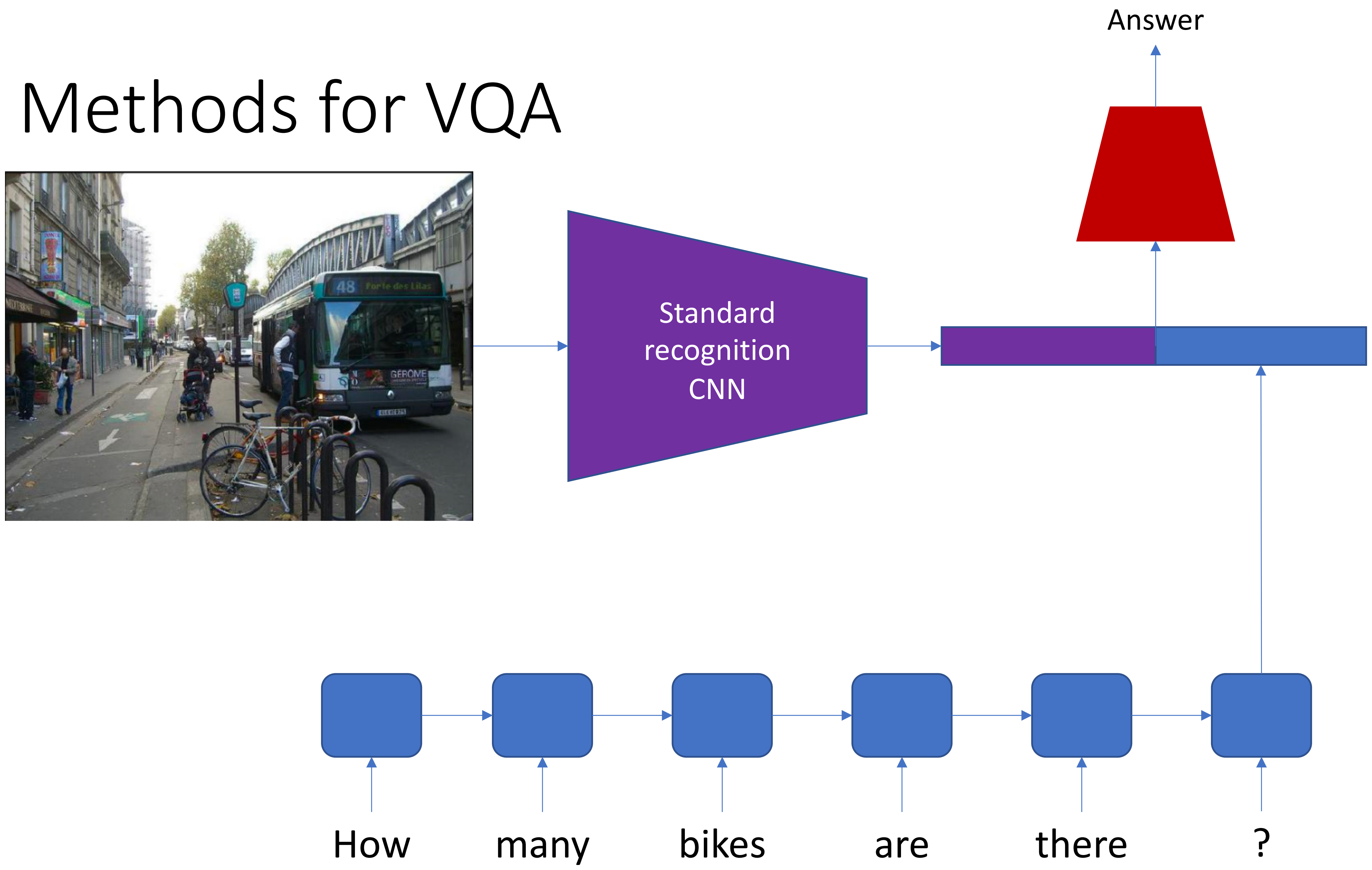
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. In *ICCV,* 2015.

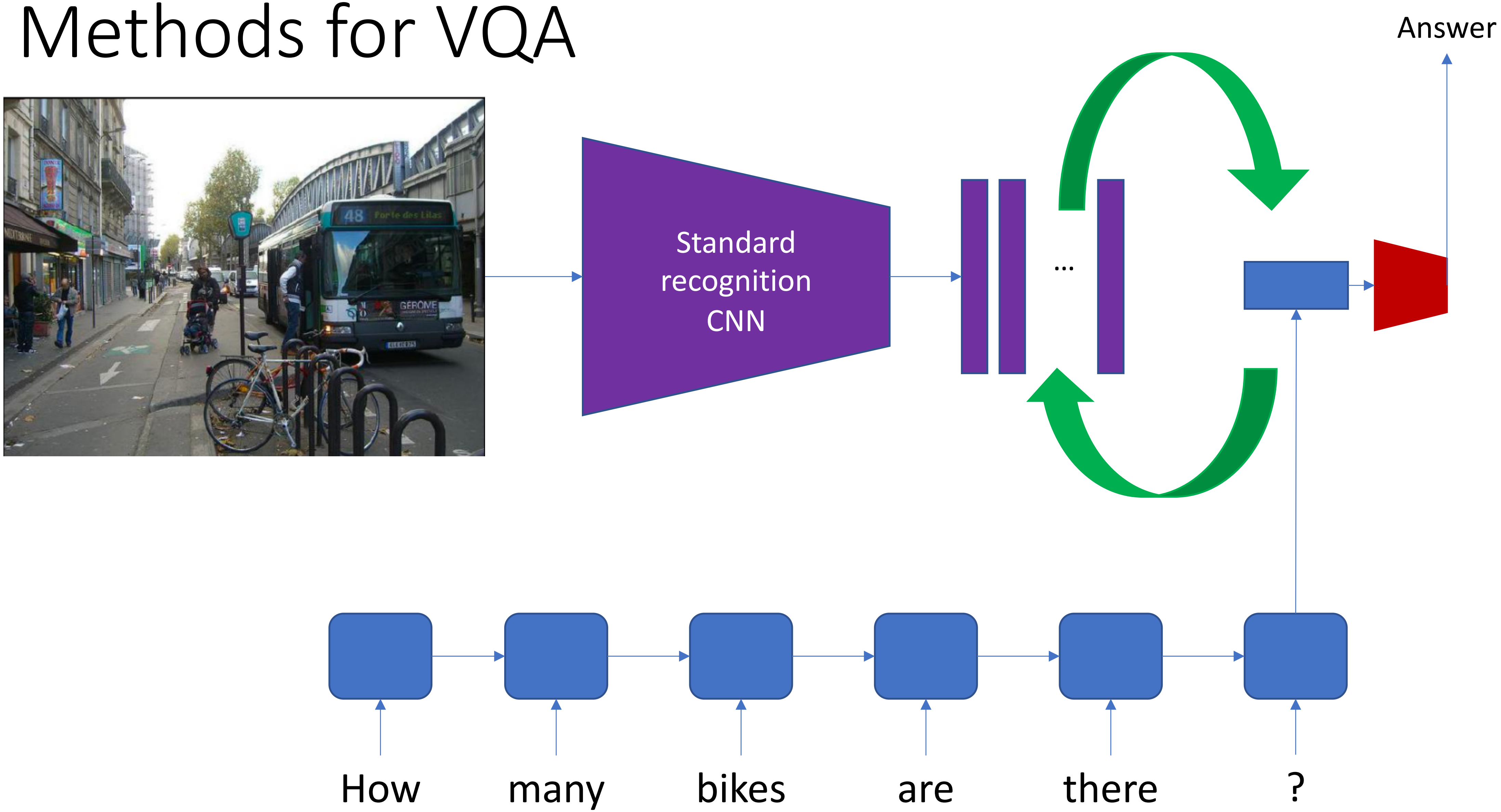
Visual Question Answering

"We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot! Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image."

Methods for VQA

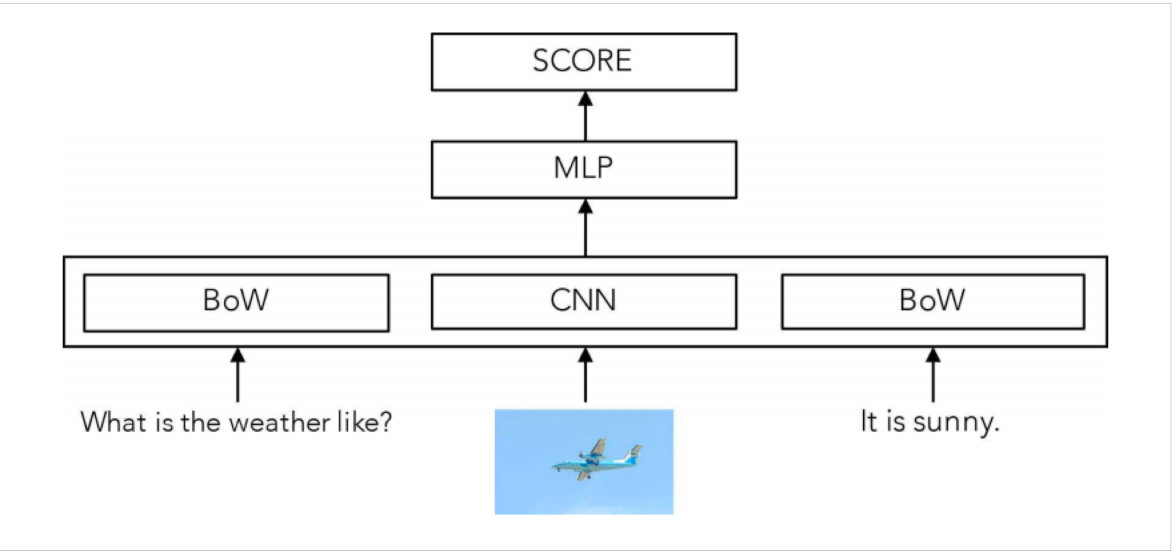


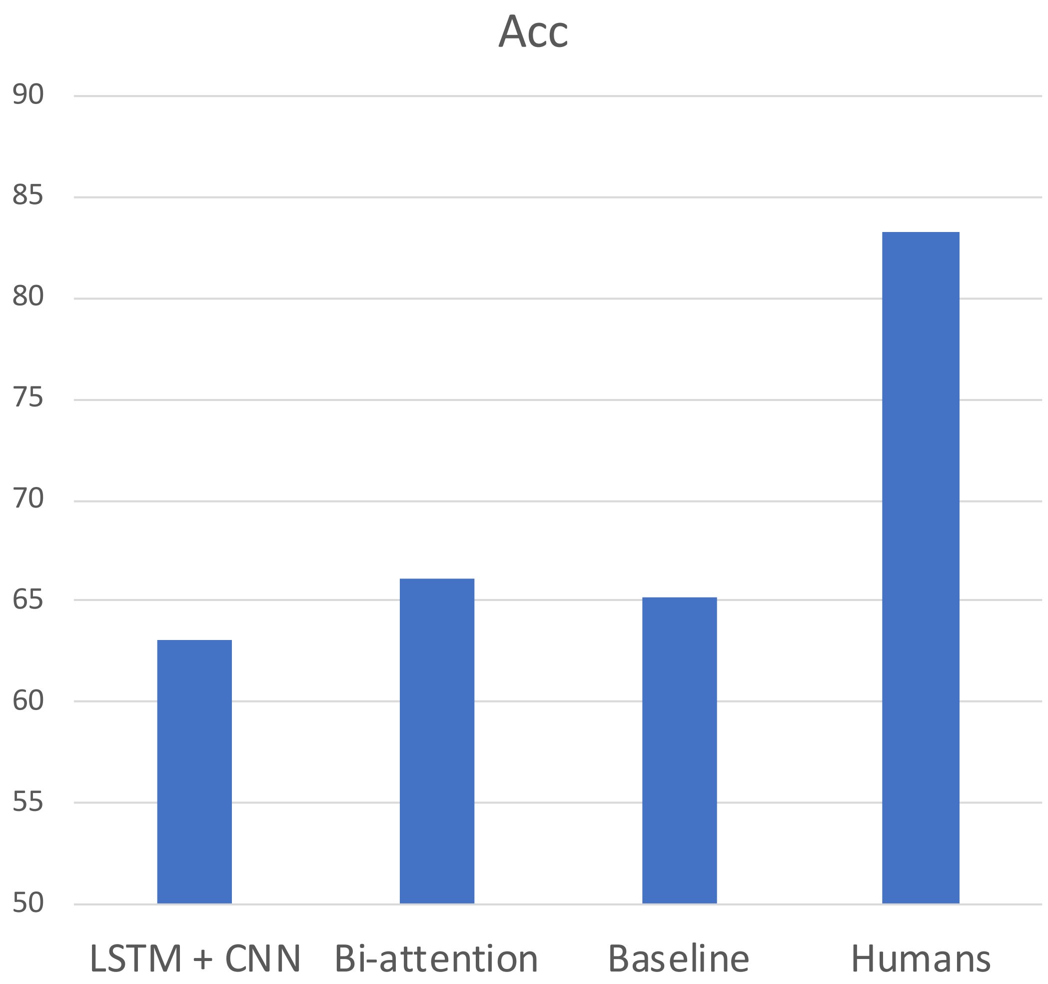
Methods for VQA



Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016

The Unreasonable Effectiveness of Baselines





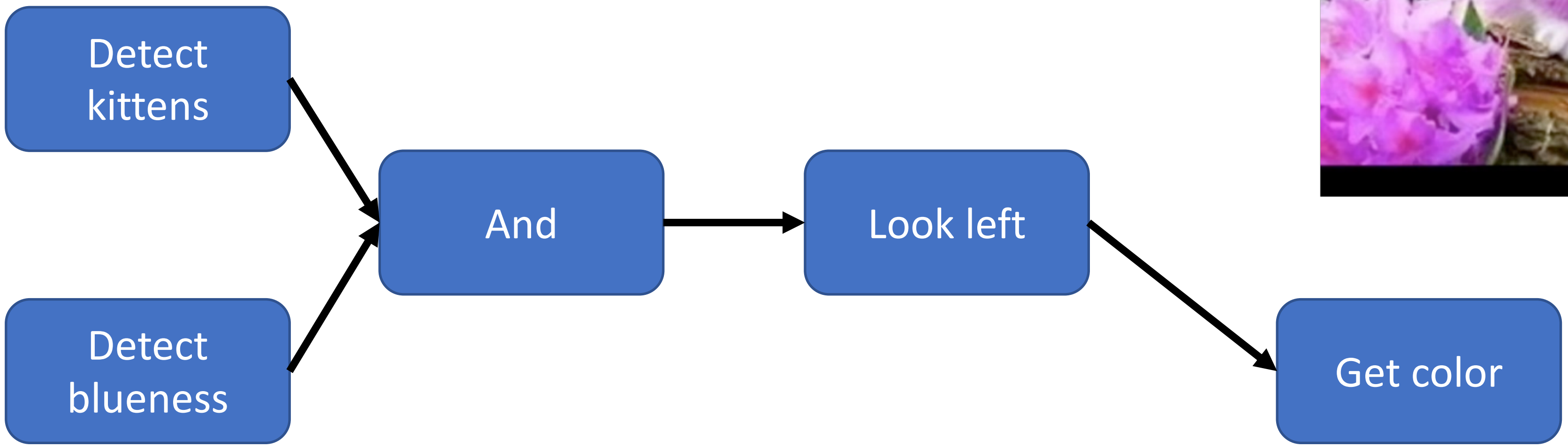
A. Jabri, A. Joulin, L. van der Maaten. Revisiting visual question answering baselines. In *ECCV,* 2016.

Compositional reasoning

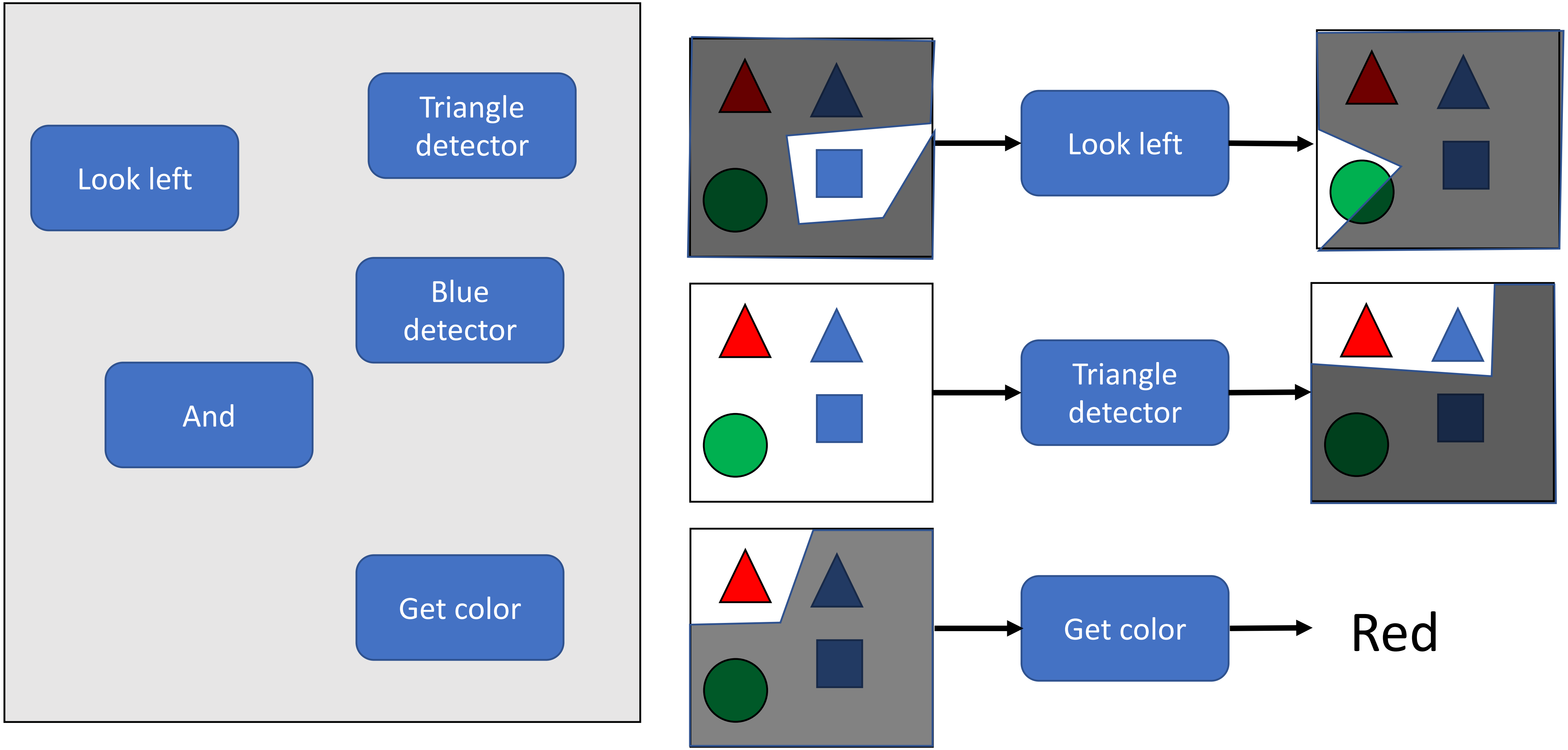
What is the color of the kitten to the left of the blue kitten?

Compositional reasoning

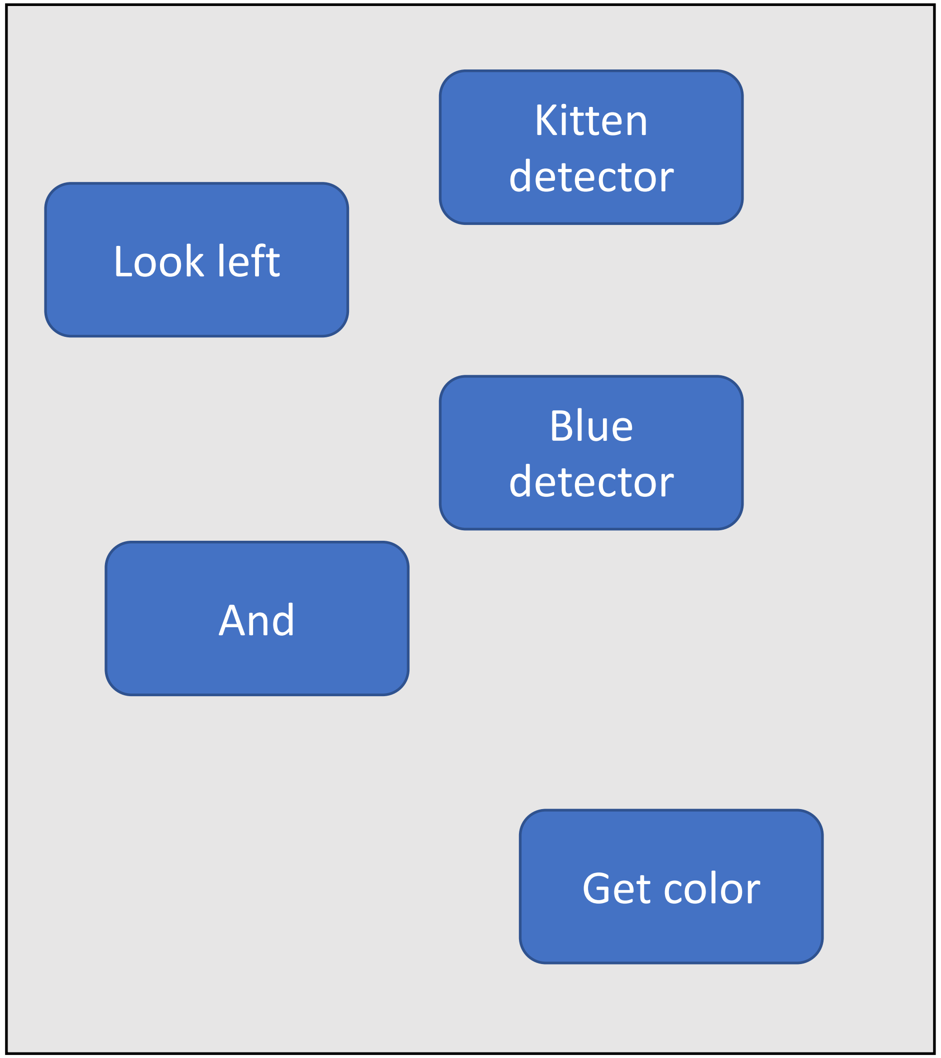
What is the color of the kitten to the left of the blue kitten?



Compositional reasoning



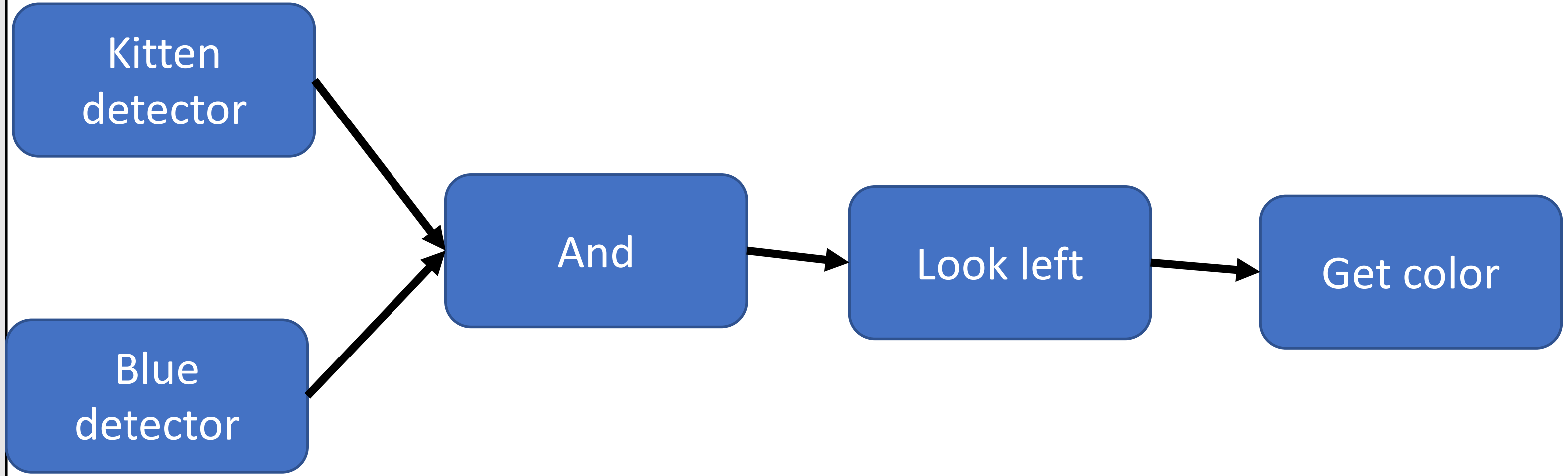
Compositional reasoning



What is the color of the kitten to the left of the blue kitten?

Compositional reasoning

What is the color of the kitten to the left of the blue kitten?



Compositional reasoning

- How do we learn a mapping from language to trees?
  - Problem: semantic parsing
  - Option 1: Syntactic parsing
  - Option 2: Use supervision

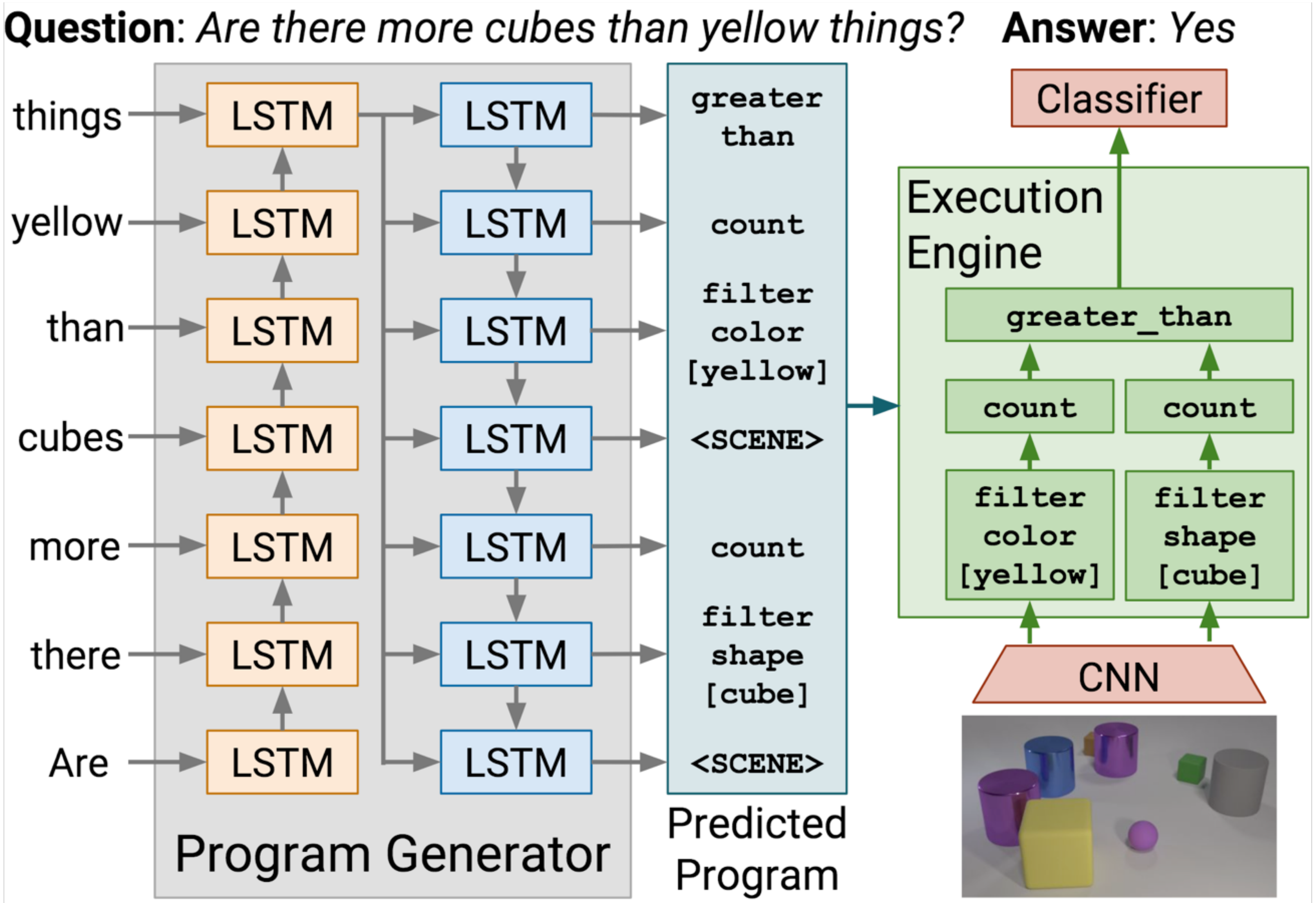
Neural module networks. Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. CVPR 2016

Learning to compose neural networks for question answering. Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. NAACL 2016

Learning to reason: End-to-end module networks for visual question answering. Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko. ICCV 2017

Inferring and Executing Programs for Visual Reasoning
Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. *ICCV*, 2017

Compositional reasoning



The problem with VQA

- Dataset biases allow cheating
  - Only-question Bag-of-Words: 53.7% (vs ~65% for state-of-the-art)
- Require common sense to answer
  - “What is the moustache made of?”
- Hard to diagnose error
  - Is the problem understanding the question?
  - Or understanding the image?

What color are her eyes?
What is the mustache made of?

Clever Hans

Current state of vision and language

- Still an active area of research
- Much better datasets but…
- Non-compositional models still win out
- The search for a better task continues