Translating Videos to Natural Language Using Deep Recurrent Neural Networks






Subhashini Venugopalan
UT Austin

Huijuan Xu
UMass. Lowell

Jeff Donahue
UC Berkeley

Marcus Rohrbach
UC Berkeley

Raymond Mooney
UT Austin

Kate Saenko
UMass. Lowell

Subhashini Venugopalan
University of Texas at Austin

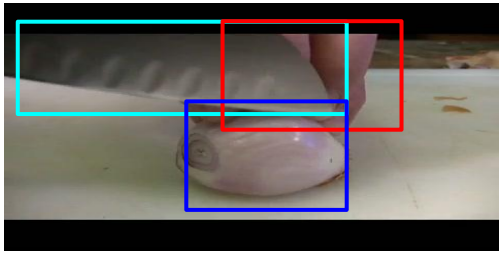
Problem Statement

Generate descriptions for events depicted in video clips

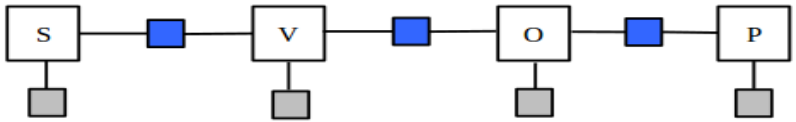
A monkey pulls a dog's tail and is chased by the dog.

Prior Work (Pipelined approach)

[Thomason et al. COLING'14]



Subjects		Verbs		Objects		Scenes	
person	0.95	slice	0.19	egg	0.31	kitchen	0.64
monkey	0.01	chop	0.11	onion	0.21	sky	0.17
animal	0.01	play	0.09	potato	0.20	house	0.07
.		.		.		.	
parrot	0	speak	0	piano	0	snow	0



A person is slicing an onion in the kitchen.

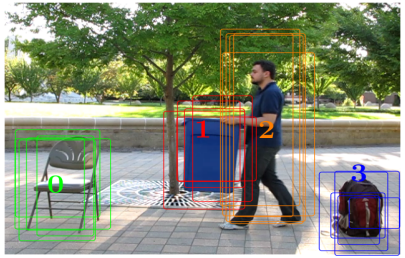
- Detect objects
- Classify actions and scenes

- Visual confidences over entities and actions

- Bias with language statistics
- Factor Graph Model (FGM) estimates most likely entities

- Template based sentence generation.

Prior Work

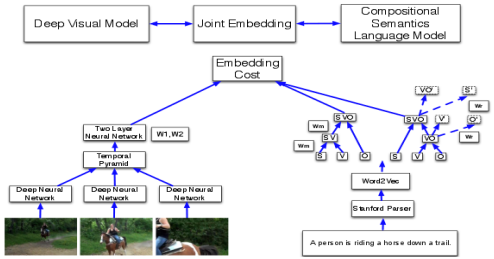


Yu and Siskind, ACL'13

Detect and track objects. Learning HMMs for actions.

Rohrbach et. al. ICCV'13

Cooking videos. CRFs.



Xu et. al. AAAI'15

Embed video and words in same space. Retrieval. CRFs for generation.

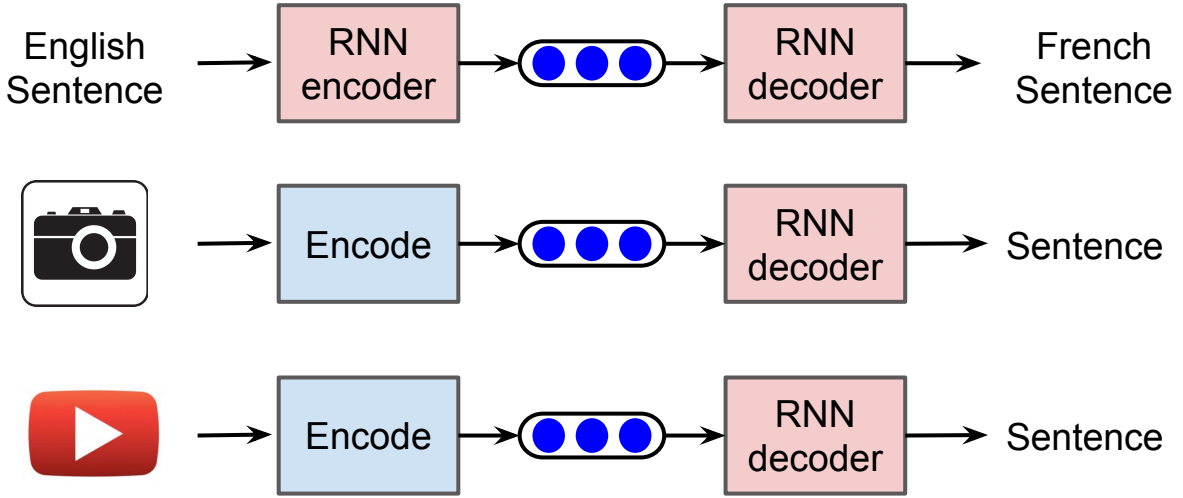
Lots of work on image to text but relatively little on video to text.

Downside: which objects/actions/scenes should I build classifiers for?

Can we learn directly from video sentence pairs?

Without having to explicitly learn object/action/scene classifiers for our dataset.

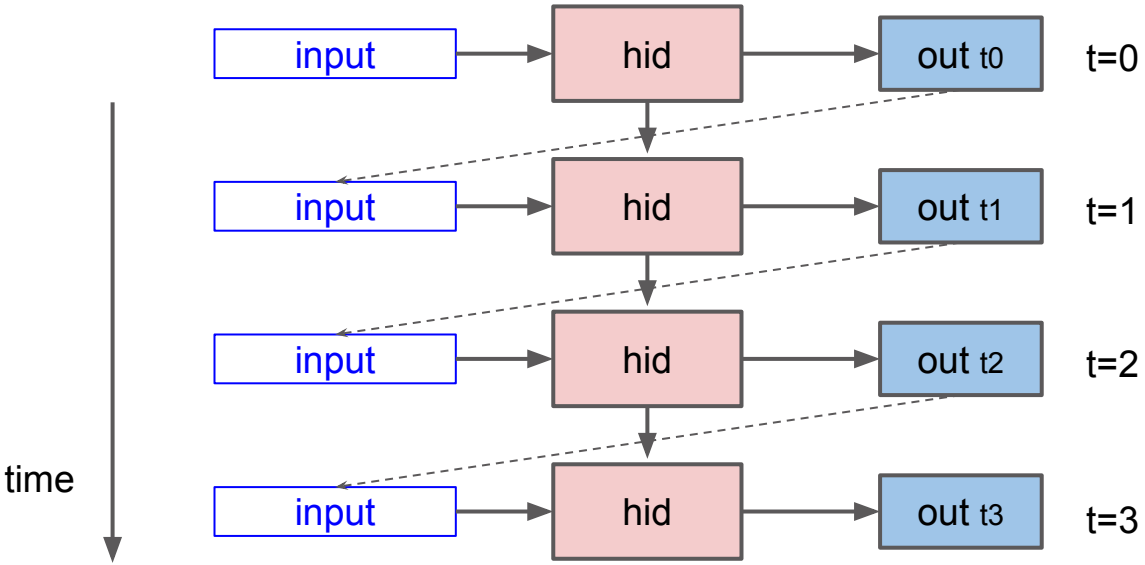
Recurrent Neural Networks (RNNs) can map a vector to a sequence.



Key Insight:

Generate feature representation of the video and "decode" it to a sentence

Recurrent Neural Networks (RNNs)



Insight: Each time step has a layer with the same weights.

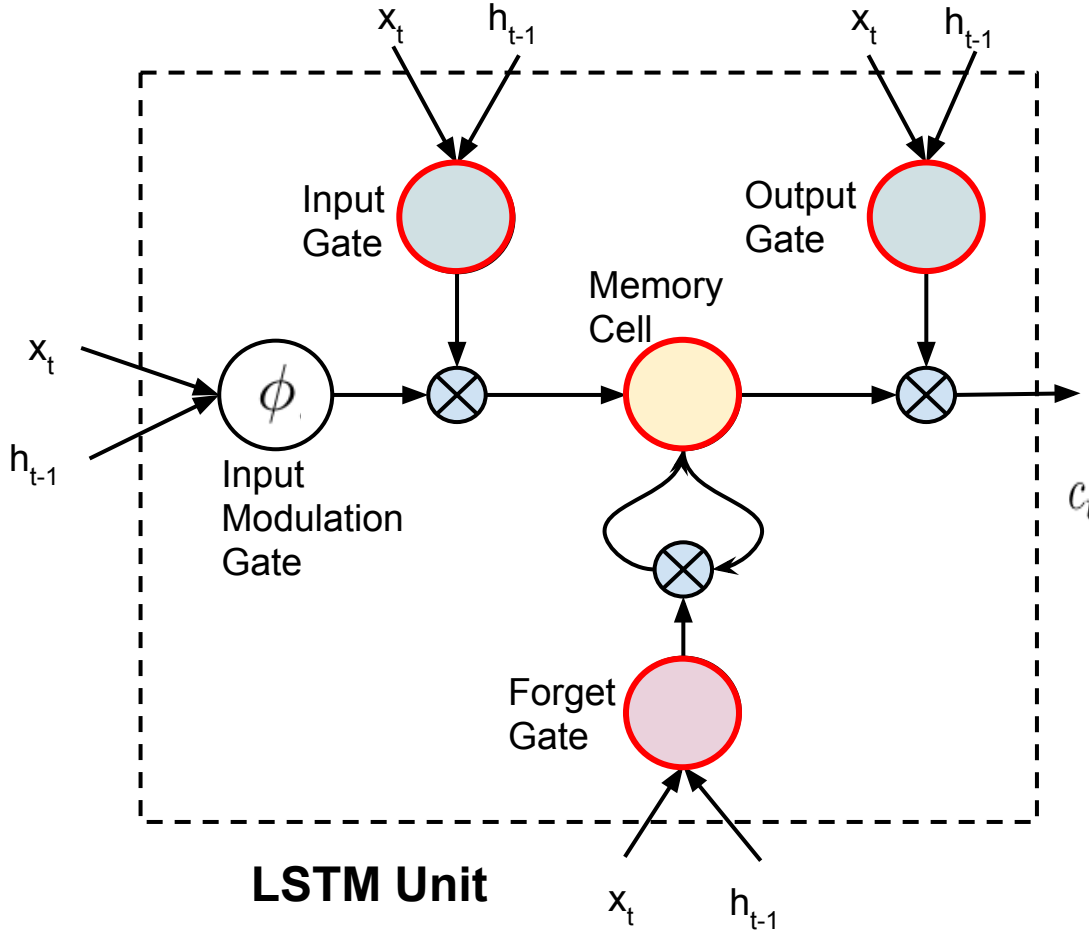
$Pr(\text{out } t_n \mid \text{input}, \text{out } t_0 \ldots t_{n-1})$

Problems -
1. Hard to capture long term dependencies
2. Vanishing gradients (shrink through many layers)

Solution: Long Short Term Memory (LSTM) unit

LSTM

[Hochreiter and Schmidhuber '97]
[Graves '13]



$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1})$$

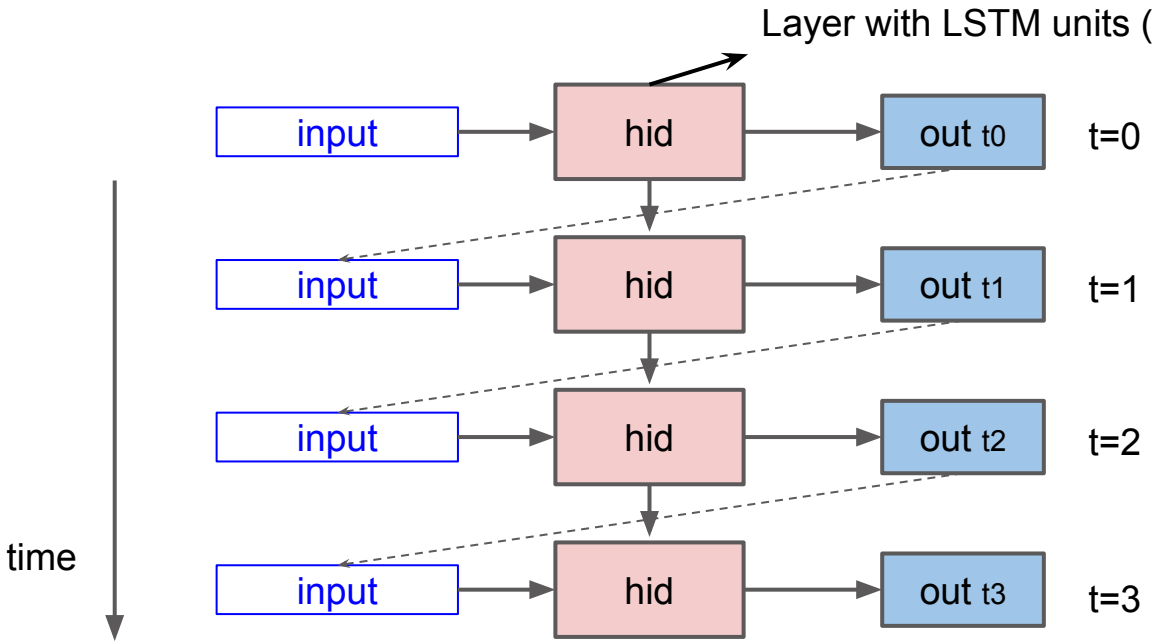
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1})$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

$$h_t = o_t \odot \phi(c_t)$$

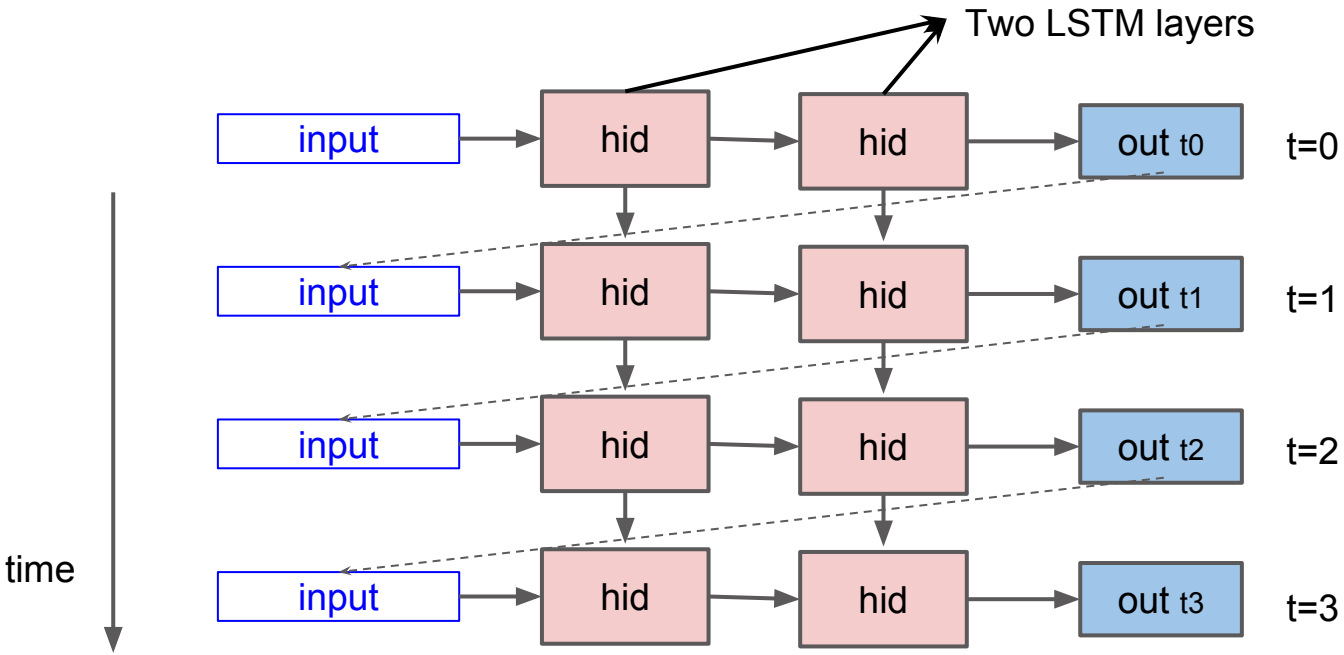
LSTM Sequence decoders



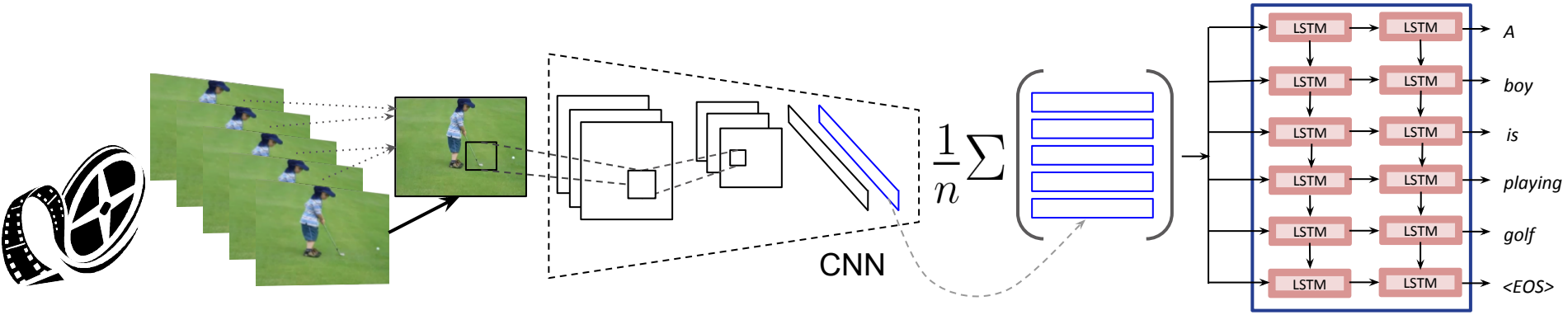
Full gradient is computed by backpropagating through time.

Matches state-of-the-art on:
Speech Recognition
[Graves & Jaitly ICML'14]
Machine Translation (Eng-Fr)
[Sutskever et al. NIPS'14]
Image-Description
[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

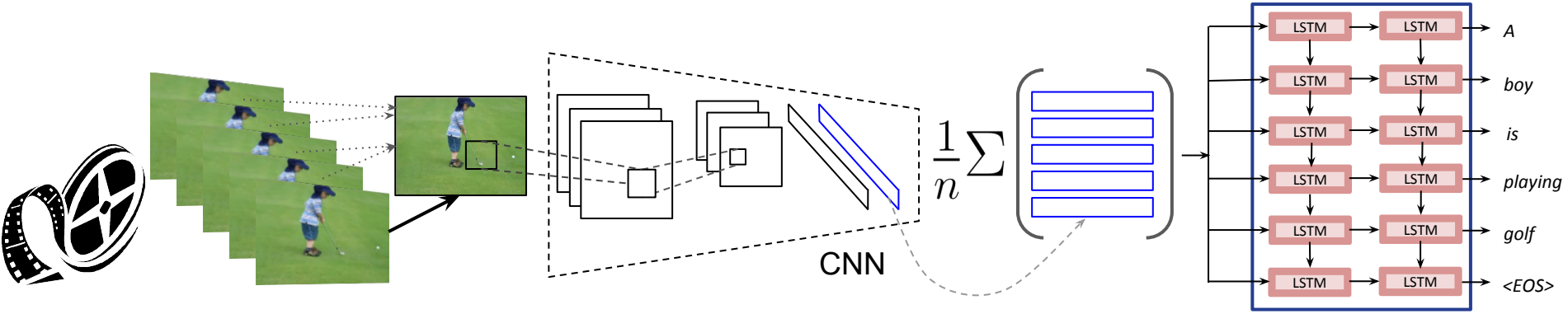
LSTM Sequence decoders



Translating videos to natural language

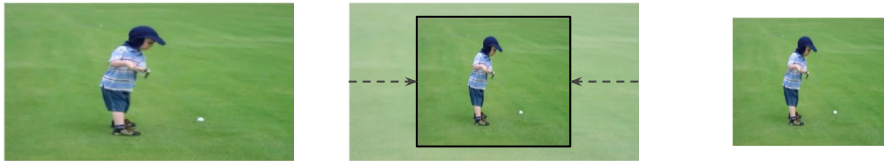


Test time: Step 1



(a)

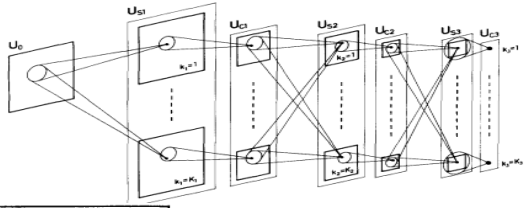
Input Video → Sample frames @1/10



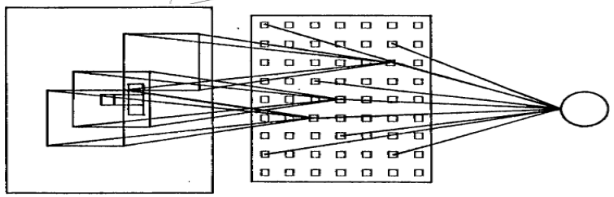
Frame → Scale

(b)

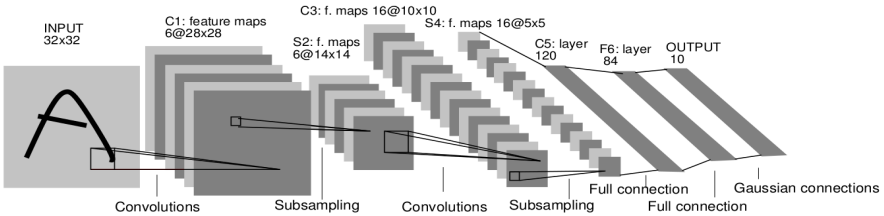
Convolutional Neural Networks (CNNs) for feature learning



Fukushima, 1980
Neocognitron.

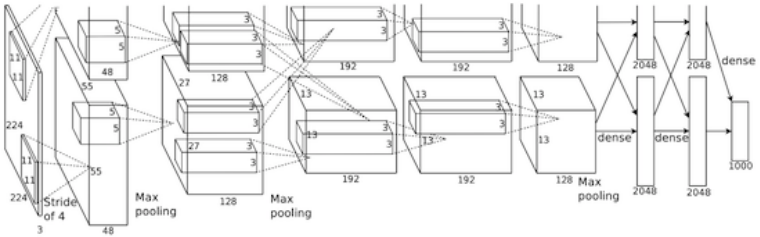


Rumelhart, Hinton, Williams 1986
"T" vs "C"



LeCun et al. 1989-1998
Handwritten digit recognition

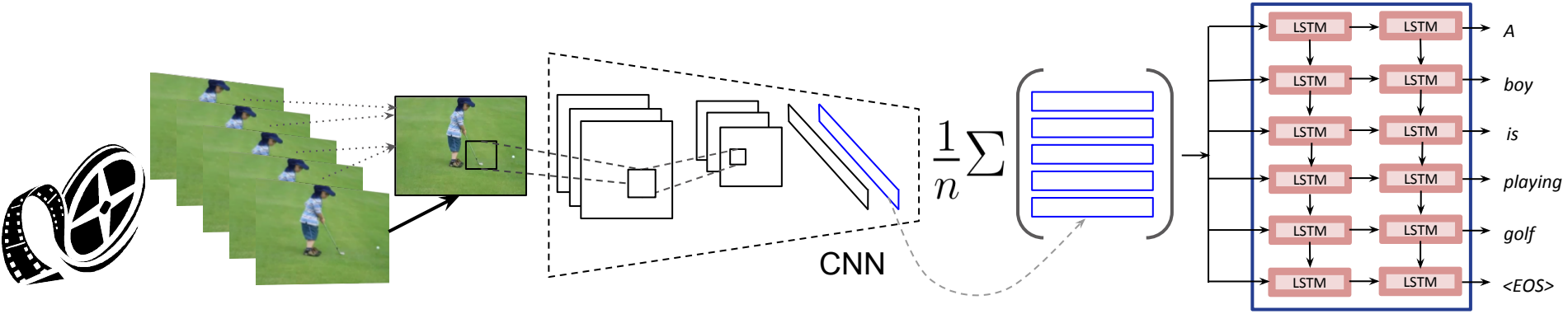
>>

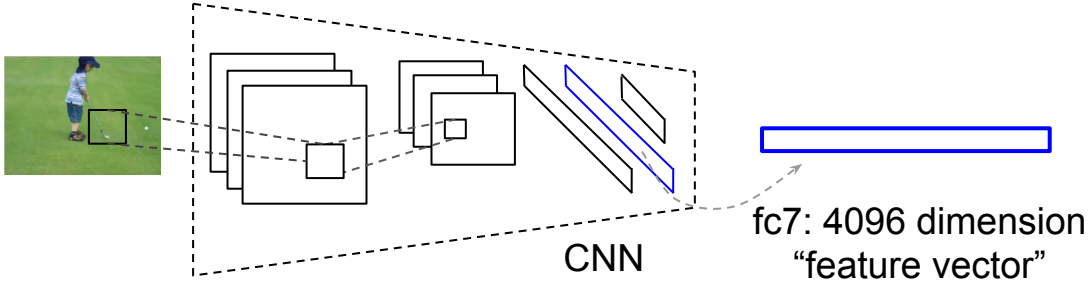


Krizhevsky, Sutskever, Hinton 2012
ImageNet classification breakthrough

Credits: R. Girshick

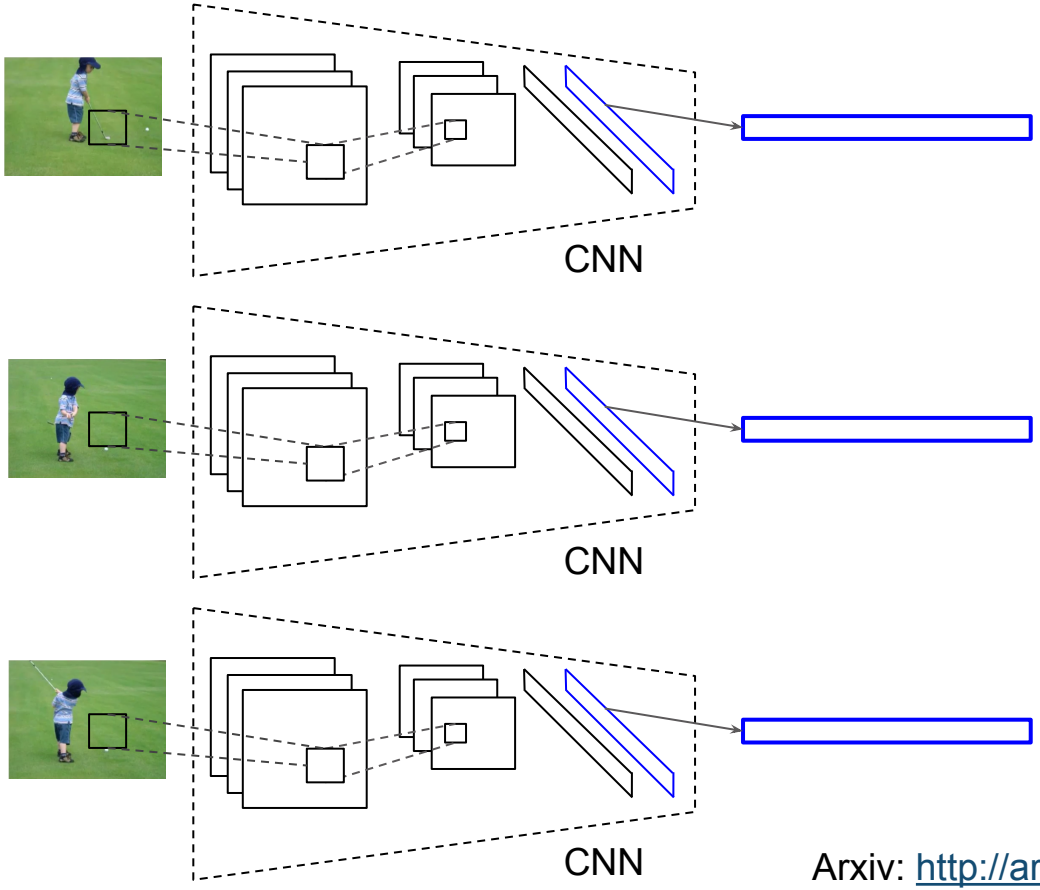
Test time: Step 2 Feature extraction





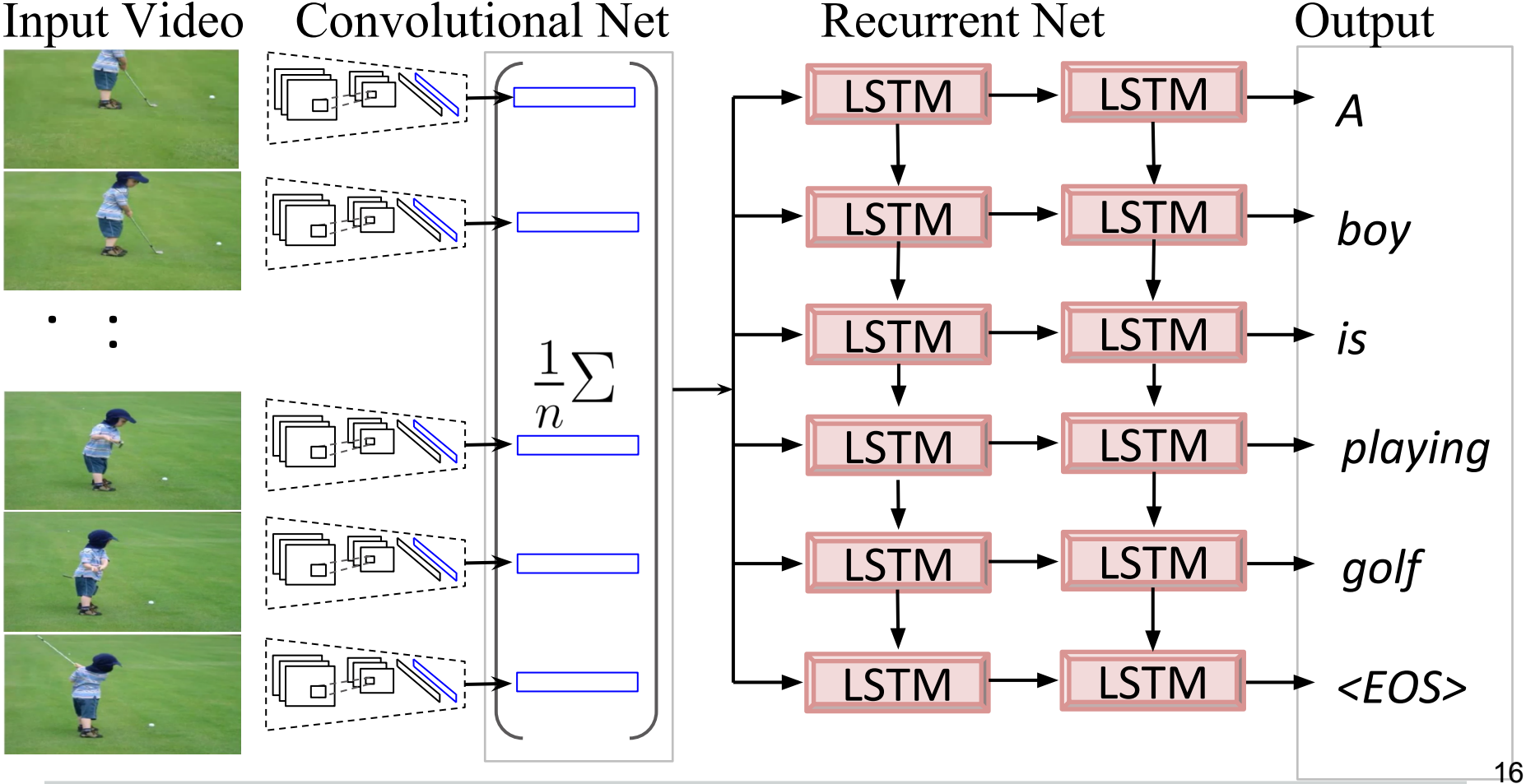
Forward propagate
Output: "fc7" features
(activations before classification layer)

Test time: Step 3 Mean pooling



Arxiv: http://arxiv.org/abs/1505.00487

Test time: Step 4 Generation



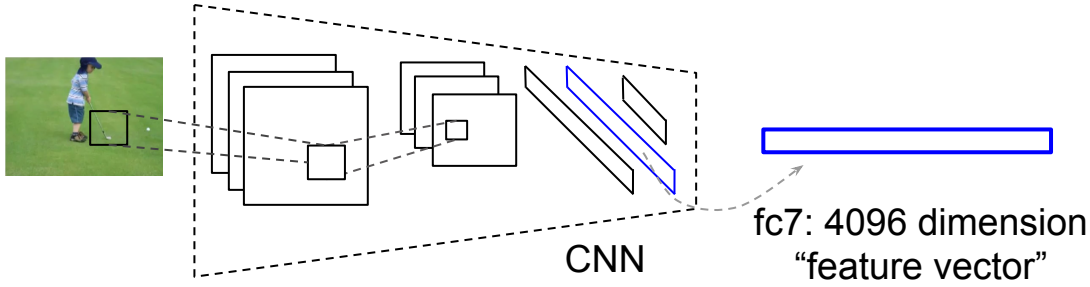
Training

Annotated video data is scarce.

Key Insight:

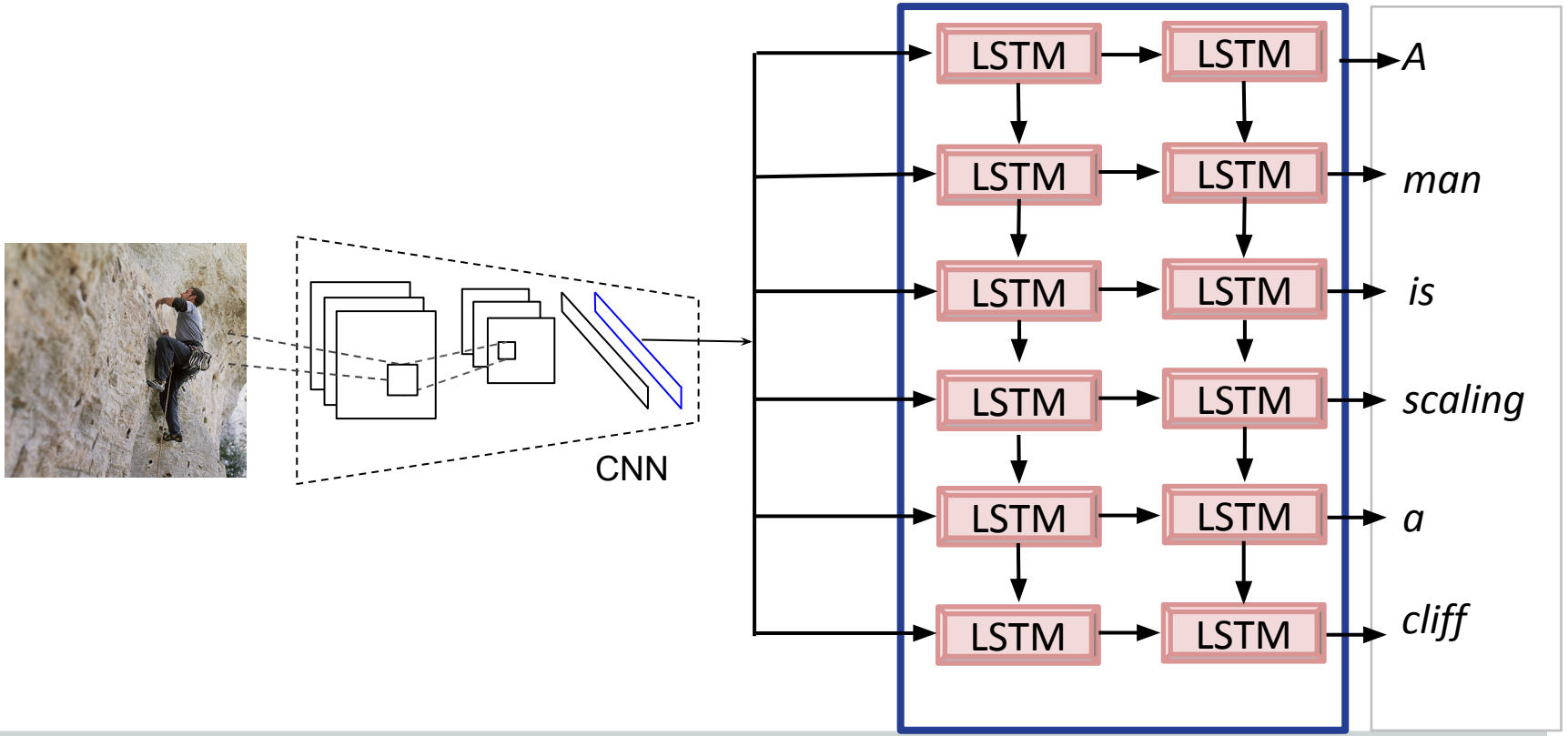
Use supervised pre-training on data-rich auxiliary tasks and transfer.

Step1: CNN pre-training

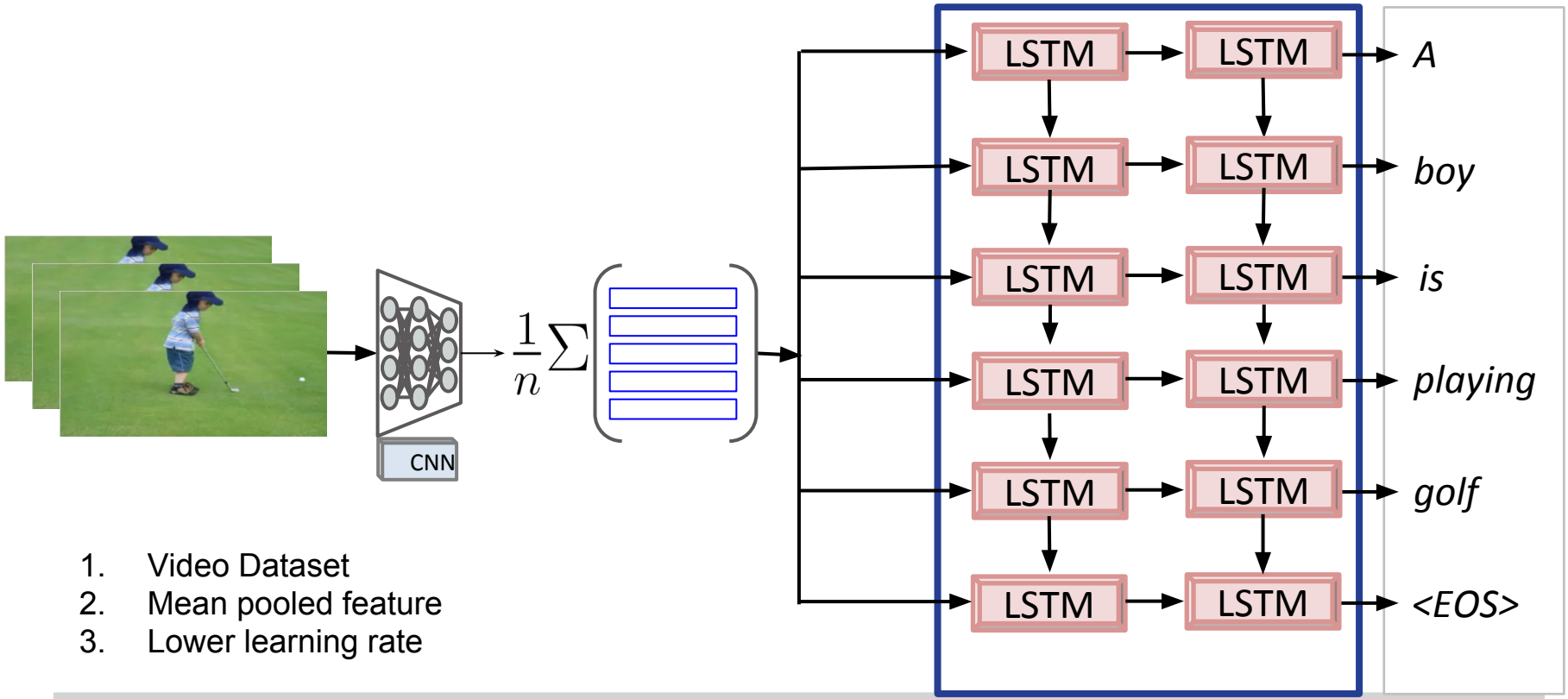


- Caffe Reference Net - variation of Alexnet [Krizhevsky et al. NIPS’12]
- 1.2M+ images from ImageNet ILSVRC-12 [Russakovsky et al.]
- Initialize weights of our network.

Step2: Image-Caption training



Step3: Fine-tuning



1. Video Dataset
2. Mean pooled feature
3. Lower learning rate

Experiments: Dataset

Microsoft Research Video Description dataset [Chen & Dolan, ACL'11]

Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

- 1970 YouTube video snippets
  - 10-30s each
  - typically single activity
  - no dialogues
  - 1200 training, 100 validation, 670 test
- Annotations
  - Descriptions in multiple languages
  - ~40 English descriptions per video
  - descriptions and videos collected on AMT

Sample video and descriptions

- A man appears to be plowing a rice field with a plow being pulled by two oxen.
- A man is plowing a mud field.
- Domesticated livestock are helping a man plow.
- A man leads a team of oxen down a muddy path.
- A man is plowing with some oxen.
- A man is tilling his land with an ox pulled plow.
- Bulls are pulling an object.
- Two oxen are plowing a field.
- The farmer is tilling the soil.
- A man in ploughing the field.

- A man is walking on a rope.
- A man is walking across a rope.
- A man is balancing on a rope.
- A man is balancing on a rope at the beach.
- A man walks on a tightrope at the beach.
- A man is balancing on a volleyball net.
- A man is walking on a rope held by poles
- A man balanced on a wire.
- The man is balancing on the wire.
- A man is walking on a rope.
- A man is standing in the sea shore.

Augment Image datasets

Training videos - 1300

Flickr30k - 30,000 images, 150,000 descriptions

MSCOCO - 120,000 images, 600,000 descriptions

Evaluation

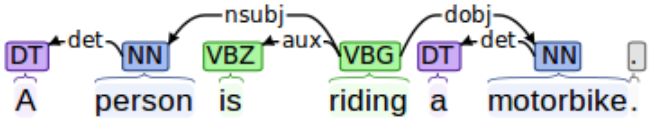
- Subject, Verb, Object accuracy (extracted from generated sentences)
- BLEU
- METEOR
- Human evaluation

Evaluation: Extracting SVO

Extracting Subject-Verb-Object (SVO) from sentences.

Consider the dependency parse of a sentence.

```
det(person-2, A-1)
nsubj(riding-4, person-2)
aux(riding-4, is-3)
root(ROOT-0, riding-4)
det(motorbike-6, a-5)
dobj(riding-4, motorbike-6)
```

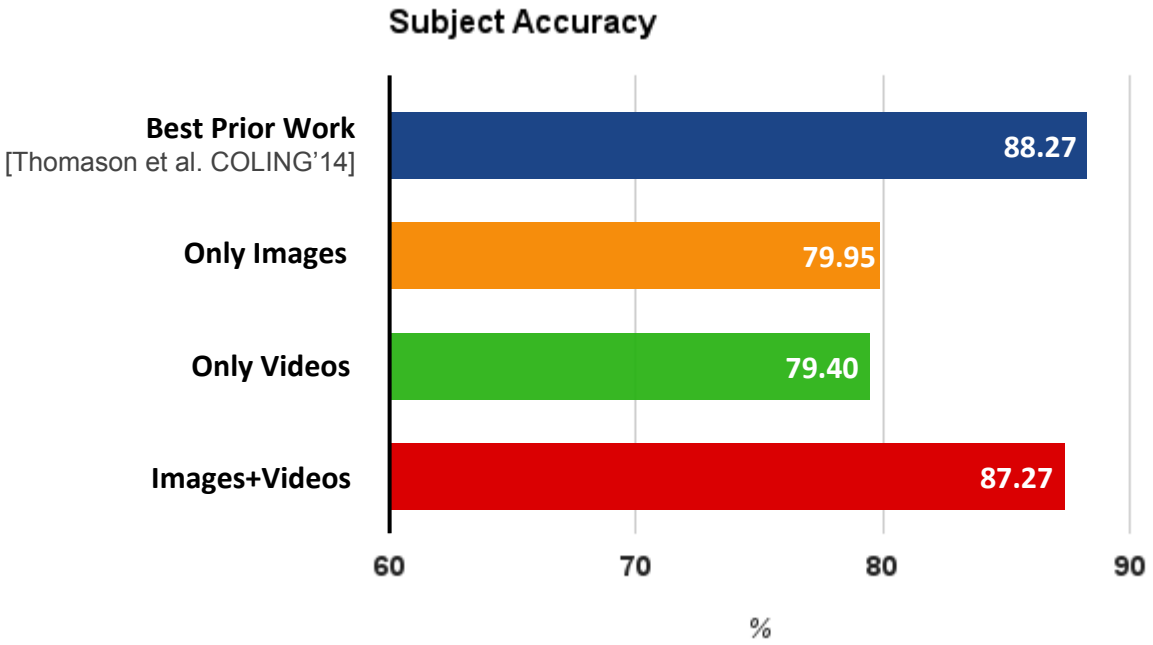


Extract Subject, Verb, Object.

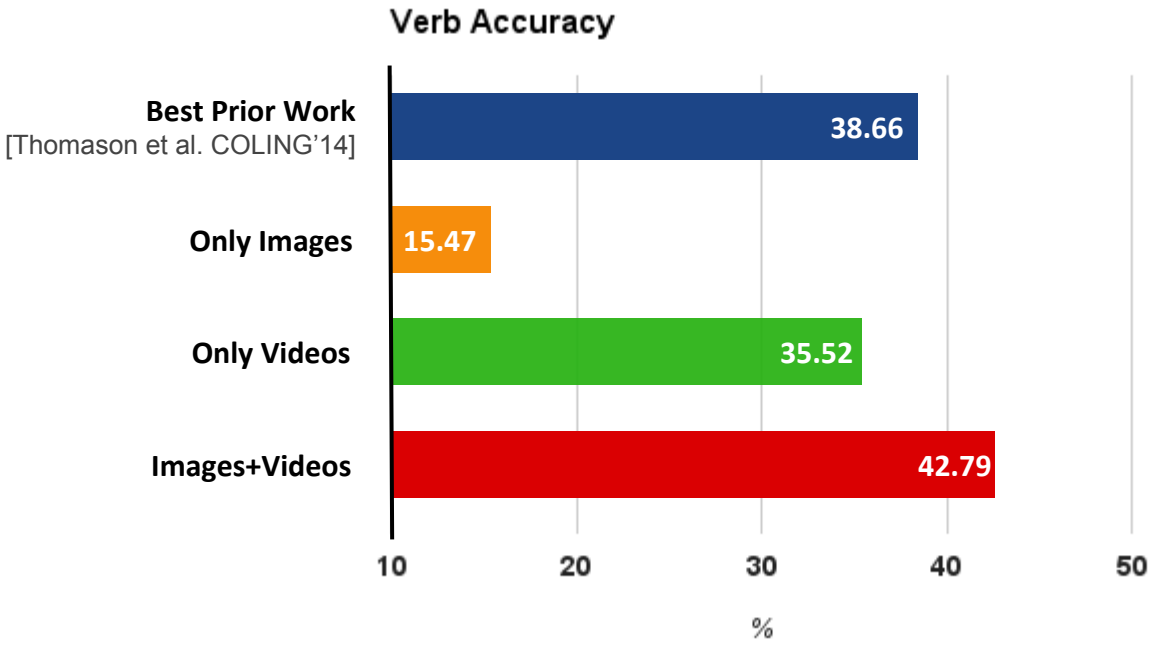
(person, ride, motorbike)

Accuracy - any valid ground truth S, V, O

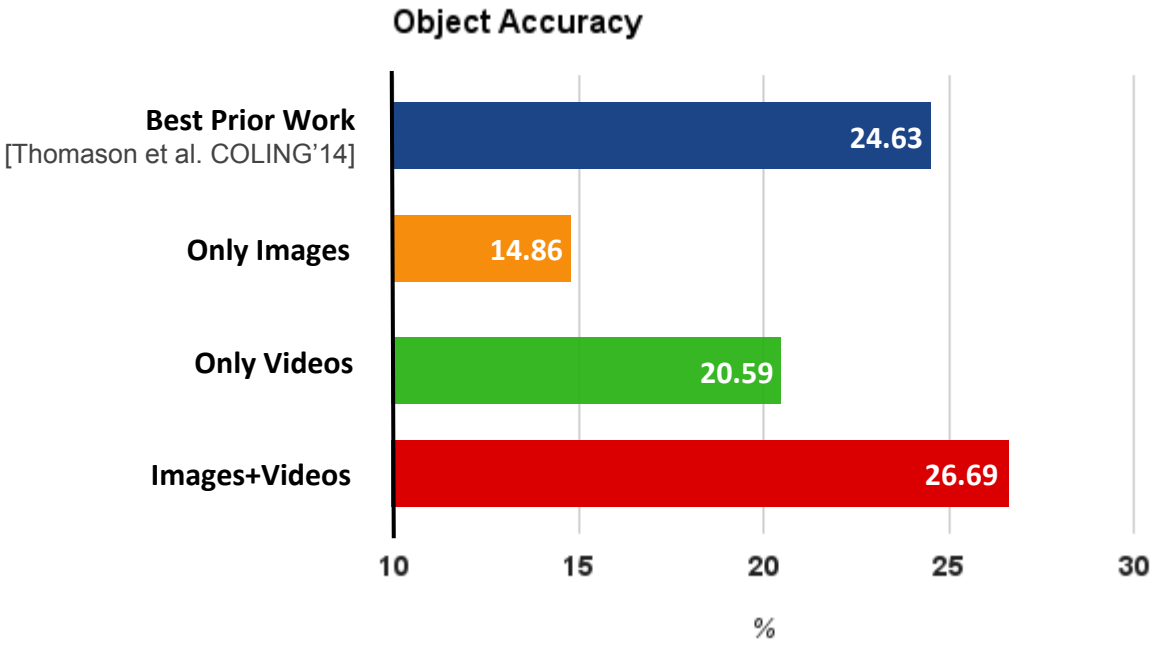
SVO - Subject accuracy



SVO - Verb accuracy



SVO - Object accuracy



Results - Generation

MT metrics (BLEU, METEOR) to compare the system generated sentences against (all) ground truth references.

Model	BLEU	METEOR
Best Prior Work [Thomason et al. COLING'14]	13.68	23.90
Only Images	12.66	20.96
Only Video	31.19	26.87
Images+Video	**33.29**	**29.07**

Human Evaluation

Relevance

Rank sentences based on how accurately they describe the event depicted in the video.

Least relevant — Most Relevant

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

No two sentences can have the same rank.

Grammar

Rate the grammatical correctness of the following **sentences**.

Incorrect — Grammatically correct

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

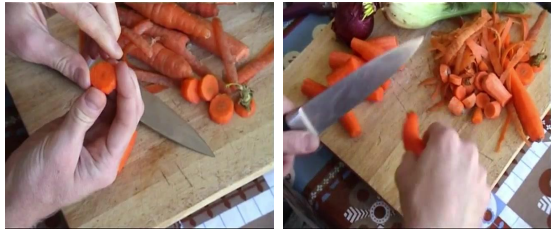
Multiple sentences can have same rating.

Results - Human Evaluation

Model	Relevance	Grammar
Best Prior Work [Thomason et al. COLING'14]	2.26	**3.99**
Only Video	2.74	3.84
Images+Video	**2.93**	3.64
Ground Truth	4.65	4.61

Examples

FGM: A person is dancing with the person on the stage.
YT: A group of men are riding the forest.
I+V: **A group of people are dancing.**
GT: Many men and women are dancing in the street.



FGM: A person is cutting a potato in the kitchen.
YT: A man is slicing a tomato.
I+V: **A man is slicing a carrot.**
GT: A man is slicing carrots.

FGM: A person is walking with a person in the forest.
YT: A monkey is walking.
I+V: **A bear is eating a tree.**
GT: Two bear cubs are digging into dirt and plant matter at the base of a tree.

FGM: A person is riding a horse on the stage.
YT: A group of playing are playing in the ball.
I+V: **A basketball player is playing**.
GT: Dwayne wade does a fancy layup in an allstar game.

Examples: Relevant but not always correct

FGM: A person is cutting the water in a pool.
YT: A man is pouring some sauce.
I+V: **A person is cutting a pizza.**
GT: Someone opens a pizza box containing pepperoni pizza .

FGM: A person is playing the guitar on the stage.
YT: **A person is flying**.
I+V: A man is doing the **air**.
GT: A female gymnast does a flip.

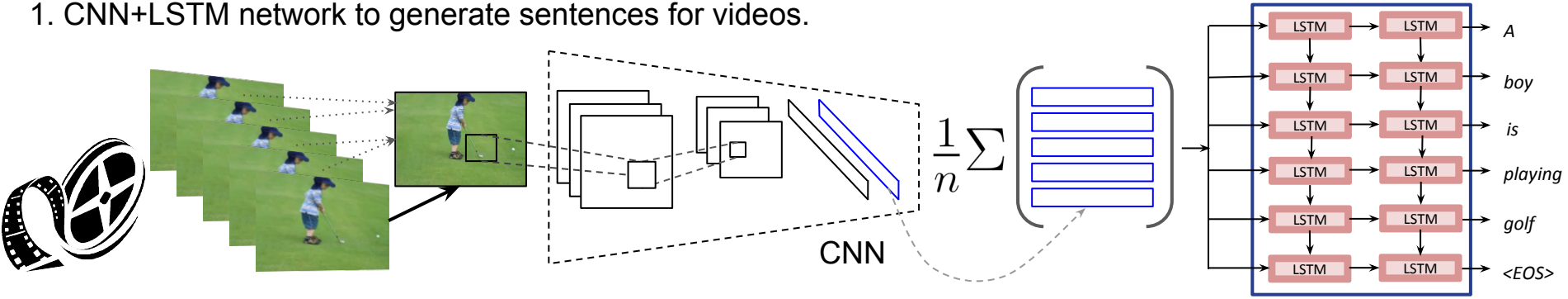
FGM: A person is walking with a person in the kitchen.
YT: A monkey is walking.
I+V: **A elephant is walking.**
GT: A baby elephant is walking and wraps his trunk around a leafy green plant .

FGM: A person is playing a person in the sky.
YT: **A dog is playing in the snow**.
I+V: A **dog** is **walking on** a ball**.**
GT: Two polar bears are wrestling in the snow.

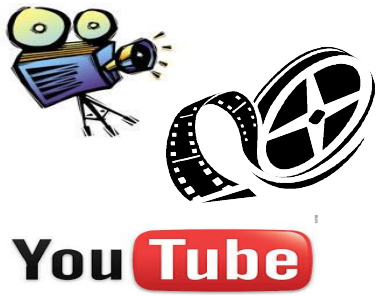
More Examples

Conclusion

1. CNN+LSTM network to generate sentences for videos.



2. Augment training with image-caption datasets.



Future Work: Incorporate temporal sequence information: http://arxiv.org/abs/**1505.00487**

Thank You

Code: https://github.com/vsubhashini/caffe/tree/recurrent/examples/youtube
We use Caffe! http://caffe.berkeleyvision.org
- Clean & fast CNN library in C++ with Python and MATLAB interfaces
- Will soon include LSTMs PR #1873
Future Work: Incorporate temporal sequence information: http://arxiv.org/abs/**1505.00487**