Natural Language Video Description using Deep Recurrent Neural Networks

Thesis Proposal
23 Nov. 2015

Subhashini Venugopalan
University of Texas at Austin

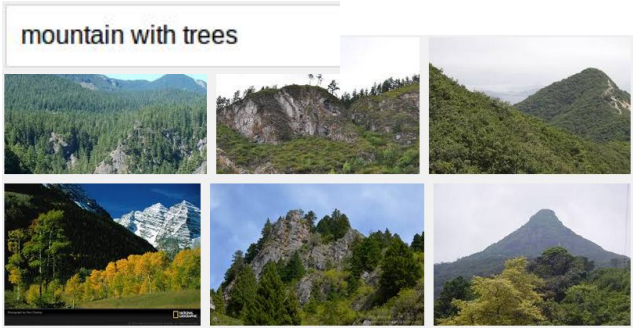
Problem Statement

Generate descriptions for events depicted in video clips

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

Applications

Image and video retrieval by content.



Video description service.

Human Robot Interaction

Video surveillance

Outline

- Related Work
- Completed Research
  - Translating Videos to Text using LSTMs [Venugopalan et. al. NAACL'15]
  - Sequence to Sequence - Video to Text [Venugopalan et. al. ICCV'15]
- Proposed Research
  - [near-term] Linguistic Knowledge, Attend to objects
  - [long-term] Multi-activity videos
  - [bonus] Movie character names
- Conclusion

Related Work

Related Work - 1: Language & Vision

Language: Increasingly focused on **grounding** meaning in perception.
Vision: Exploit linguistic ontologies to **"tell a story"** from images.

[Farhadi et. al. ECCV'10]

(animal, stand, ground)

[Kulkarni et. al. CVPR'11]

There are one cow and one sky.
The golden cow is by the blue sky.

Many early works on Image Description
Farhadi et. al. ECCV'10, Kulkarni et. al. CVPR'11, Mitchell et. al. EACL'12, Kuznetsova et. al. ACL'12 & ACL'13

Identify objects and attributes, and combine with linguistic knowledge to "tell a story".

Dramatic increase in interest the past year. (8 papers in CVPR'15)

[Donahue et. al. CVPR'15]

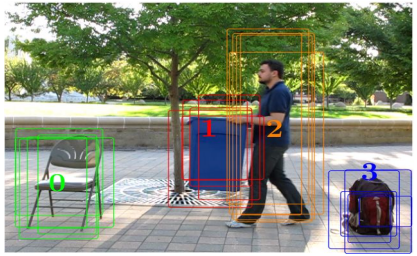
A group of young men playing a game of soccer.

Relatively little on Video Description

Need videos for semantics of wider range of actions.

Related Work - 2: Video Description

[Krishnamurthy, et al. AAAI'13]



[Yu and Siskind, ACL'13]

[Rohrbach et. al. ICCV'13]

- Extract object and action descriptors.
- Learn object, action, scene classifiers.
- Use language to bias visual interpretation.
- Estimate most likely agents and actions.
- Template to generate sentence.

Others: Guadarrama ICCV'13, Thomason COLING'14

Limitations:

- Narrow Domains
- Small Grammars
- Template based sentences
- Several features and classifiers

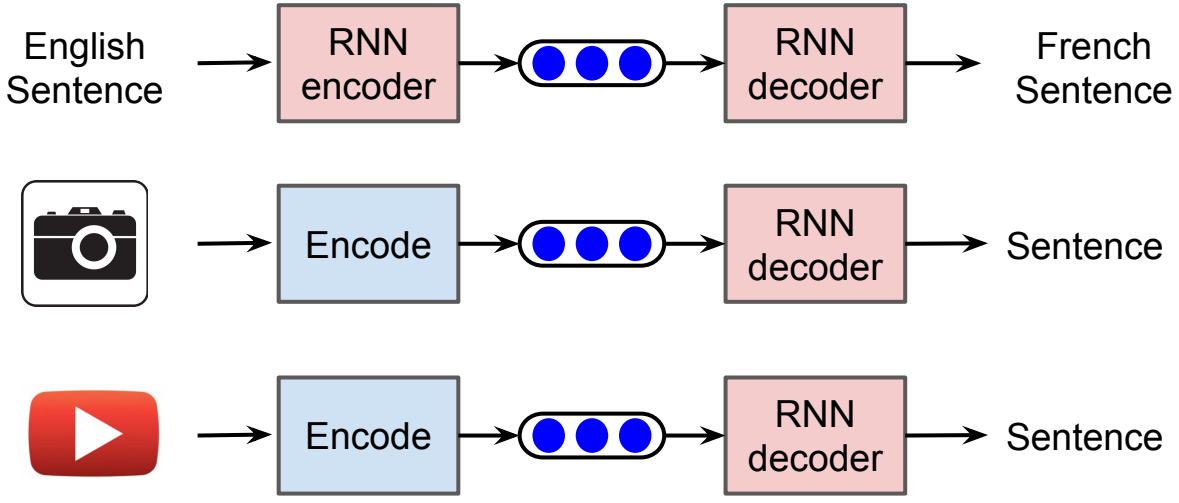
Which objects/actions/scenes should we build classifiers for?

Can we learn directly from video sentence pairs?

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

[Venugopalan et. al. NAACL'15]

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

In this section

- Background - Recurrent Neural Networks
- 2 Deep methods for video description
  - First, learns from image description.
    (ignores temporal frame sequence in videos)
  - Second is temporally sensitive to input.

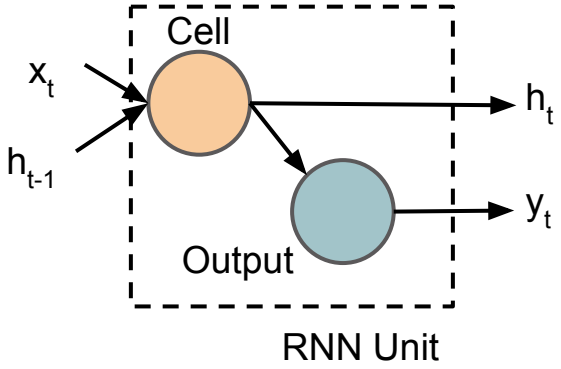
[Background] Recurrent Neural Networks

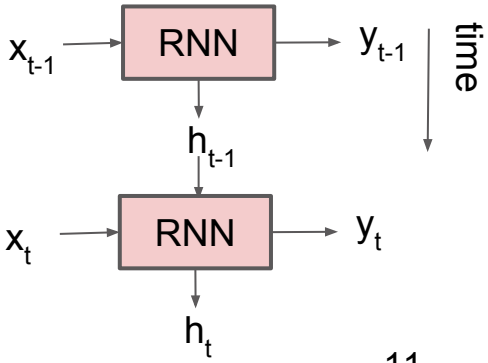
Successful in translation, speech.

RNNs can map an input to an output sequence.

$$\Pr(\text{out } y_t \mid \text{input}, \text{out } y_0 \ldots y_{t-1})$$

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



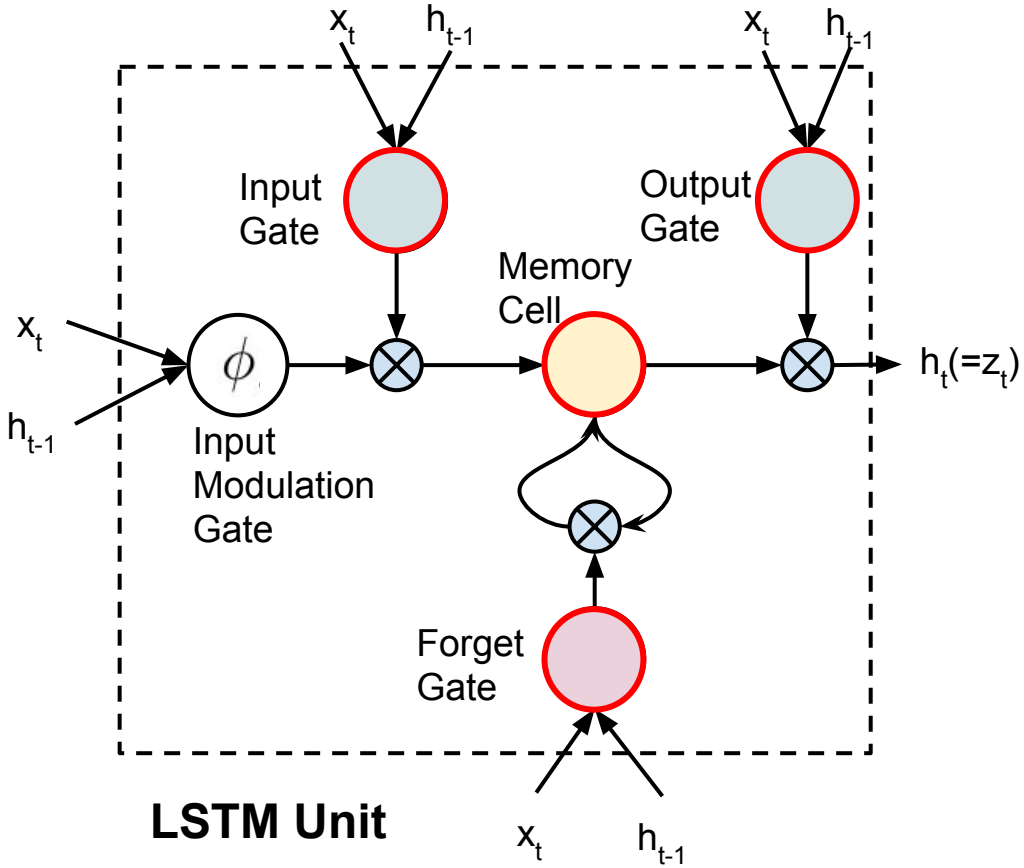


Problems:
1. Hard to capture long term dependencies
2. Vanishing gradients (shrink through many layers)

Solution: Long Short Term Memory (LSTM) unit

[Background] LSTM

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



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

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

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

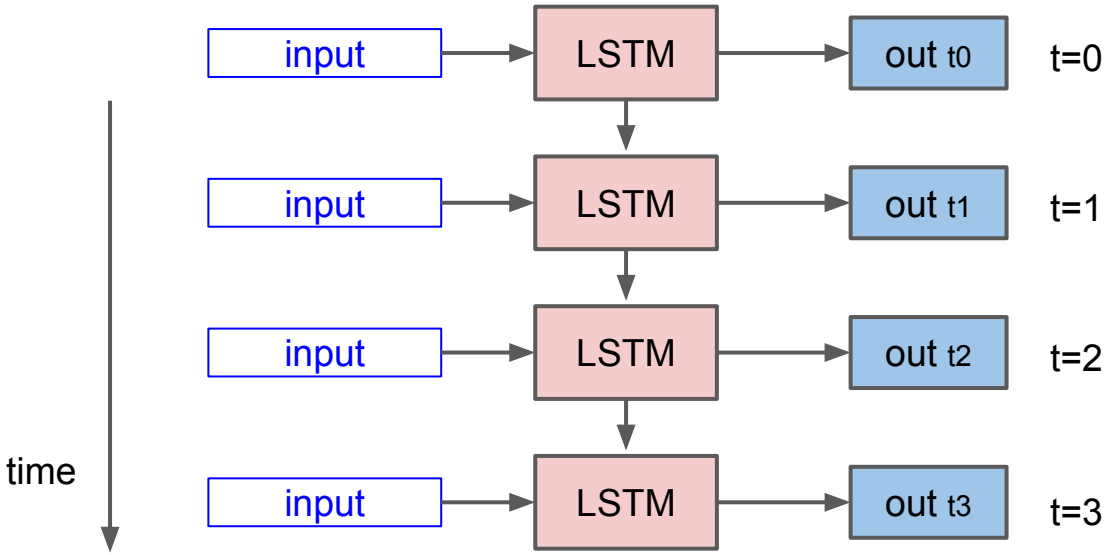
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

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

[Background] LSTM Sequence decoders

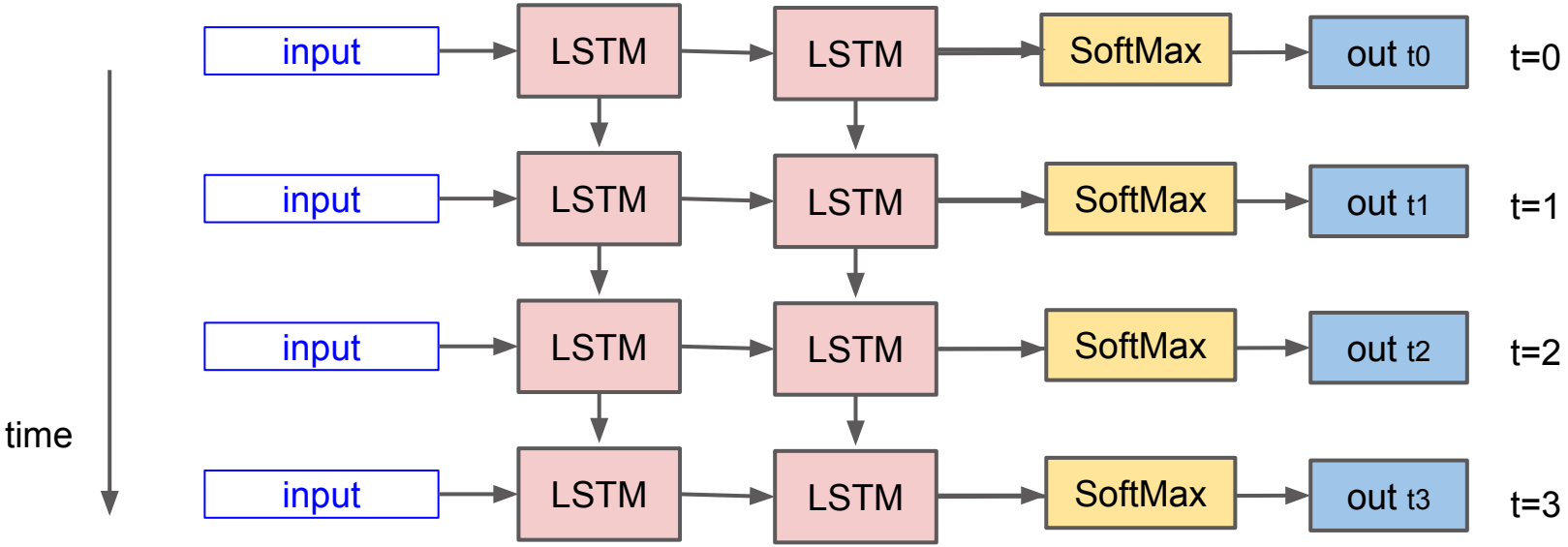
Functions are differentiable.
Full gradient is computed by backpropagating through time.
Weights updated using Stochastic Gradient Descent.

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]

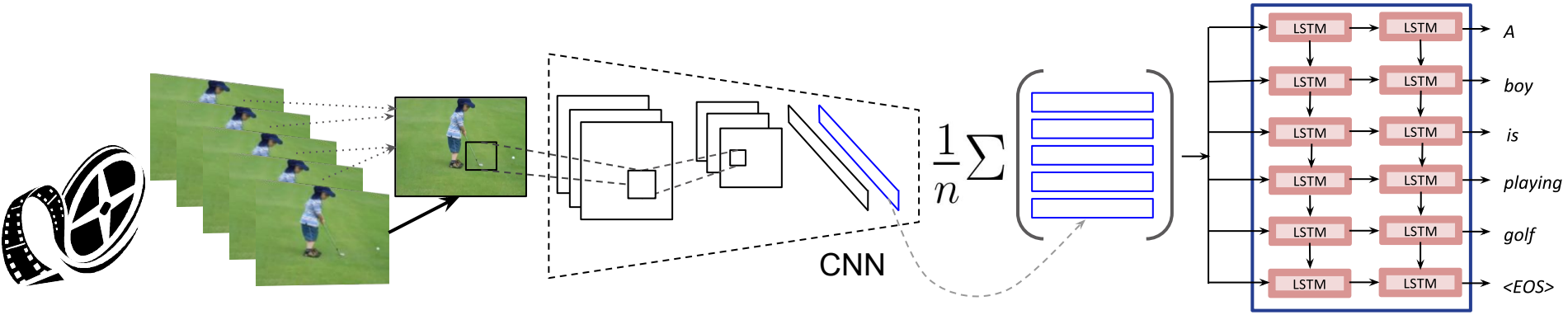


LSTM Sequence decoders

Two LSTM layers - 2nd layer of depth in temporal processing.
Softmax over the vocabulary to predict the output at each time step.

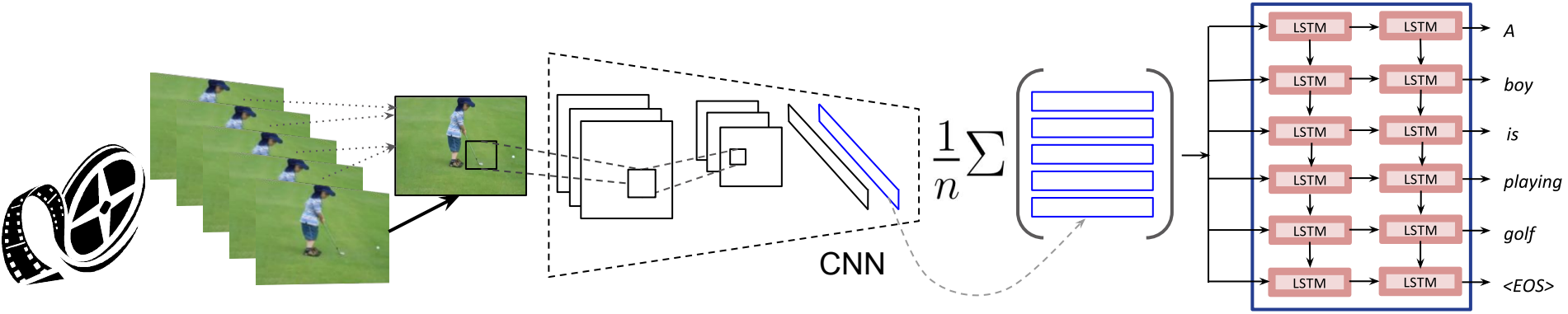


Translating Videos to Natural Language



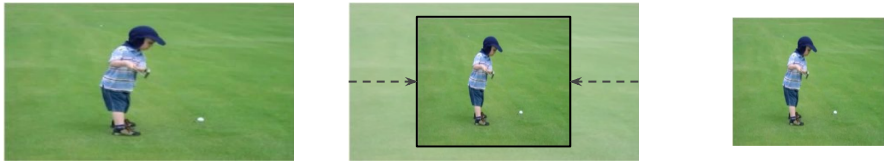
[Venugopalan et. al. NAACL'15]

Test time: Step 1



(a)

Input Video → Sample frames @1/10



Frame → Scale

(b)

[Background] Convolutional Neural Networks (CNNs)

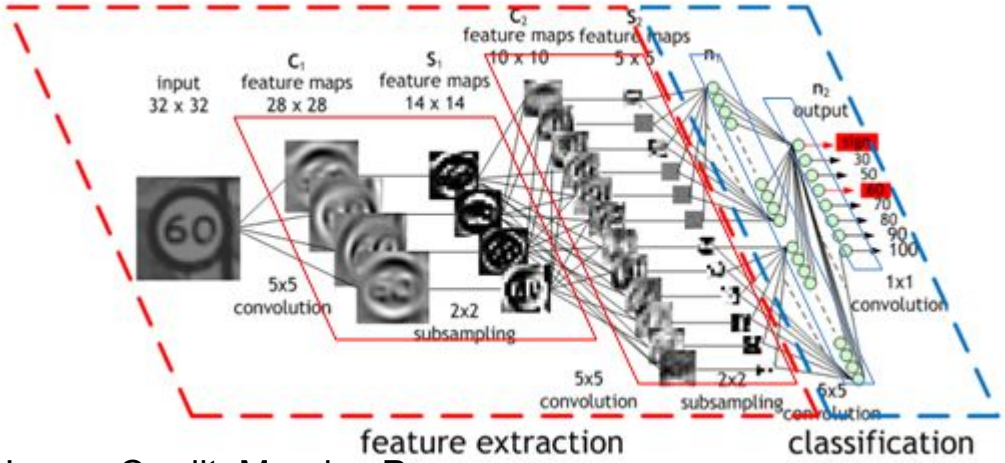


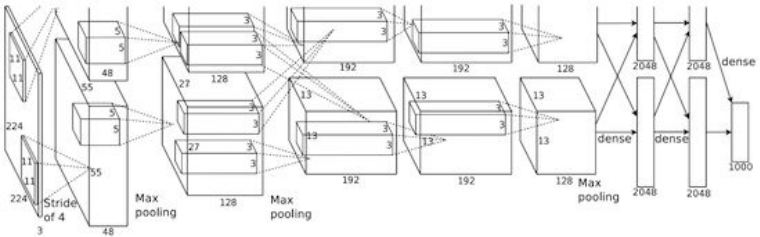
Image Credit: Maurice Peeman

Successful in semantic visual recognition tasks.

Layer - linear filters followed by non linear function. Stack layers.

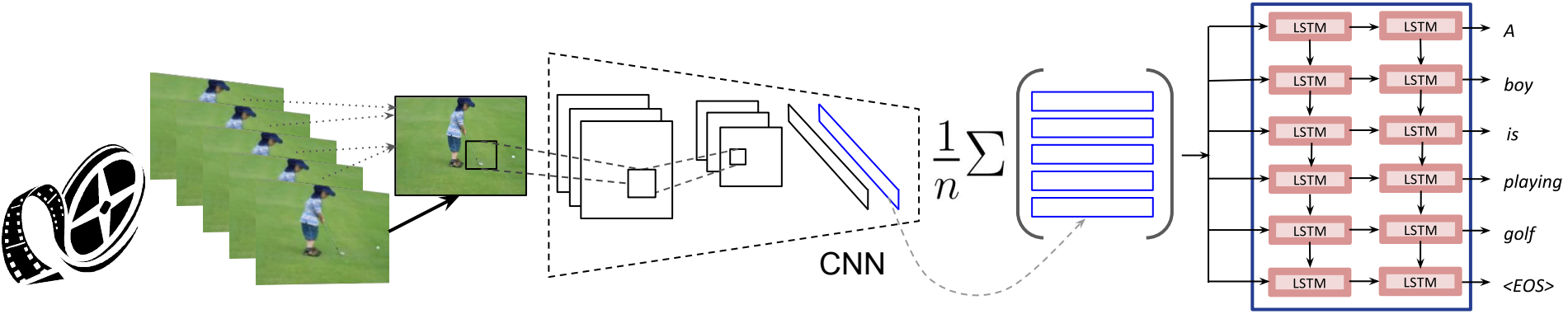
Learn a hierarchy of features of increasing semantic richness.

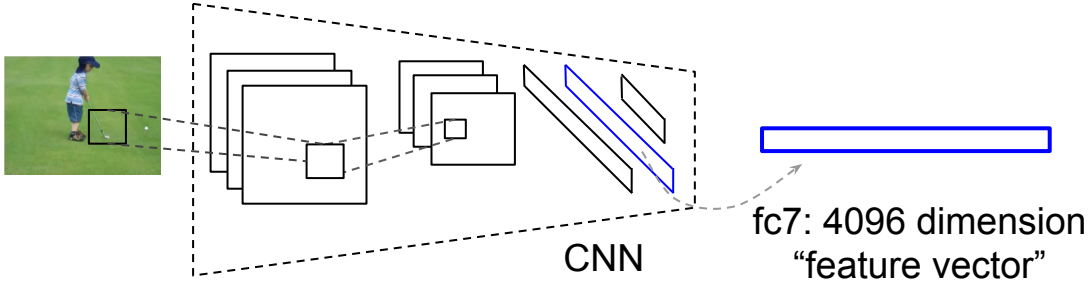
>>



Krizhevsky, Sutskever, Hinton 2012
ImageNet classification breakthrough

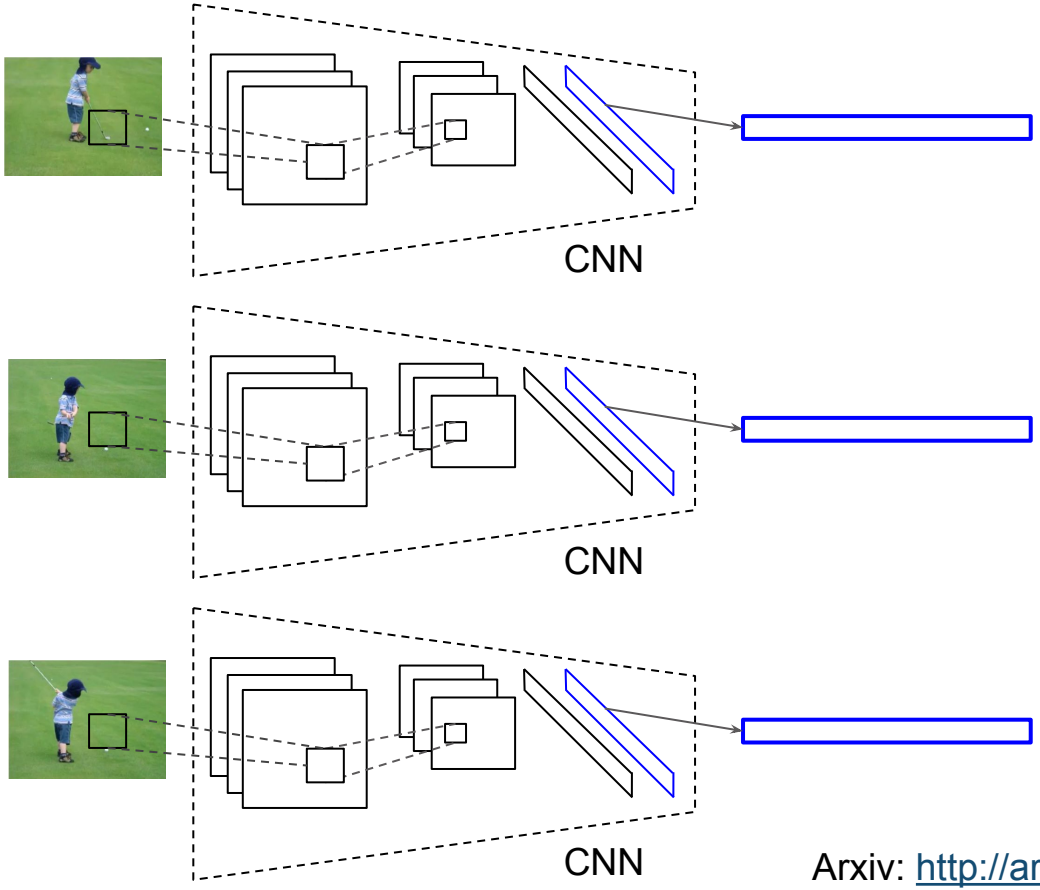
Test time: Step 2 Feature extraction





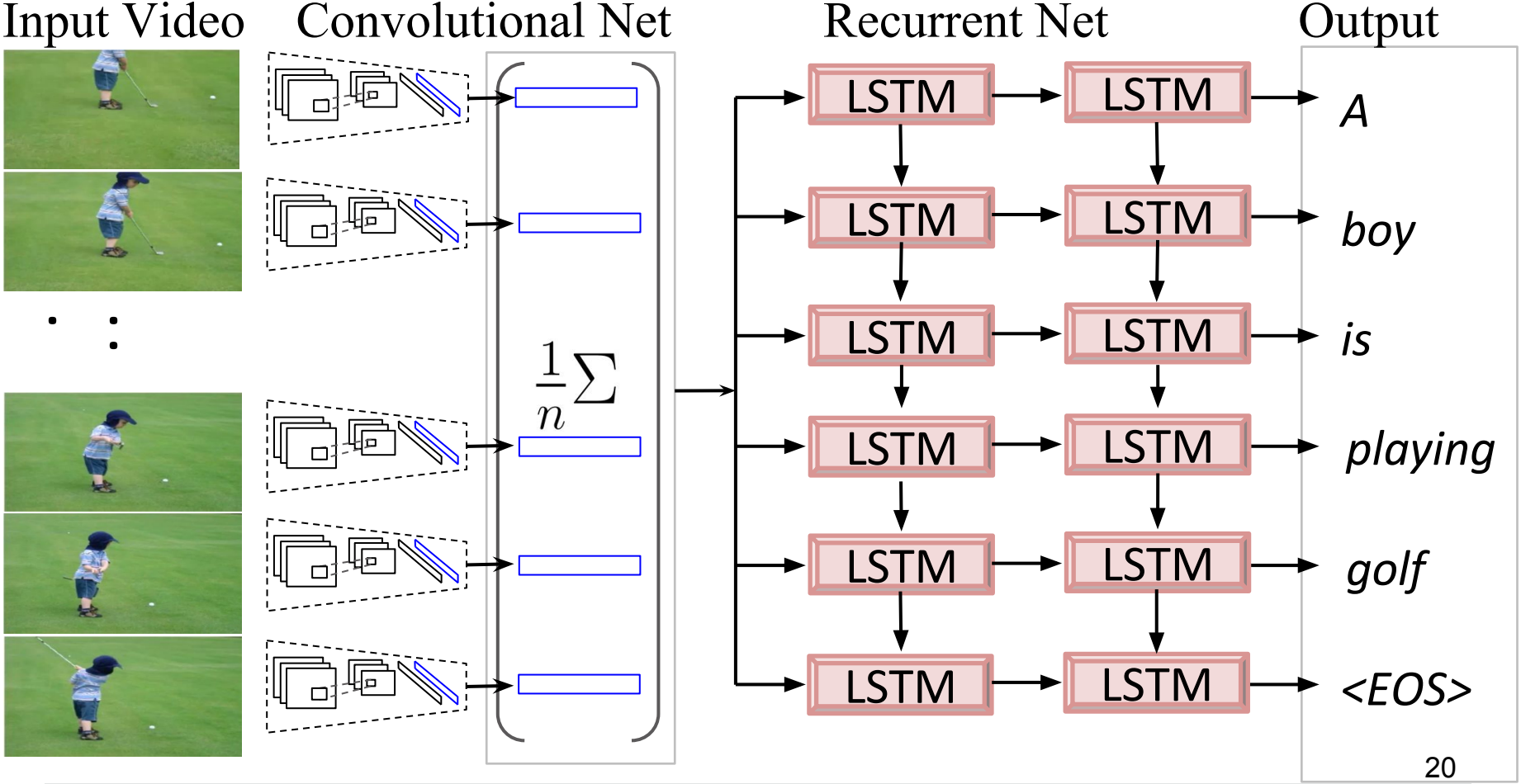
Forward propagate
Output: "fc7" features
(activations before classification layer)

Test time: Step 3 Mean pooling



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

Test time: Step 4 Generation



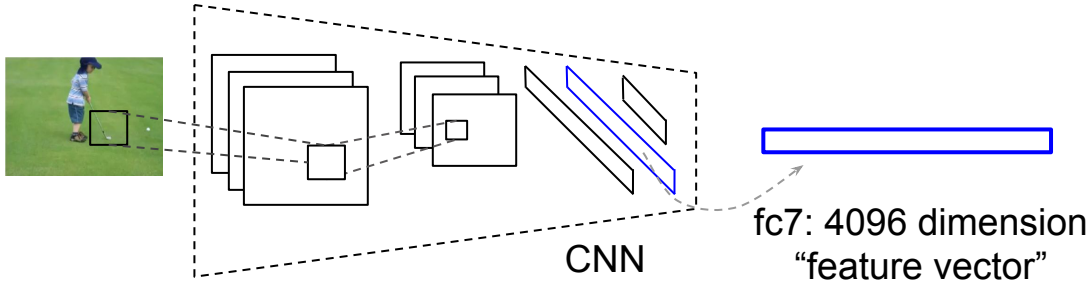
Training

Annotated video data is scarce.

Key Insight:

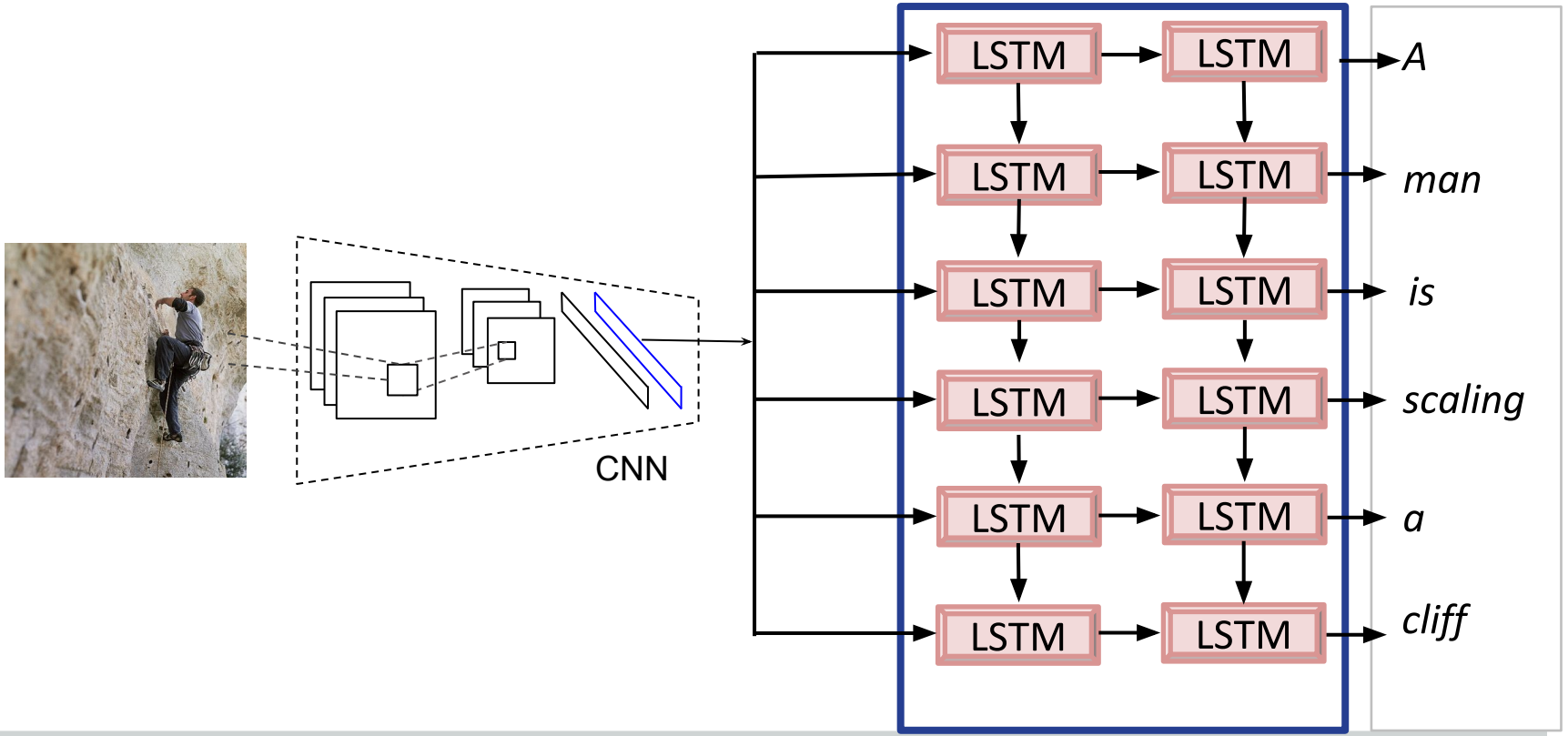
Use supervised pre-training on data-rich auxiliary tasks and transfer.

Step1: CNN pre-training

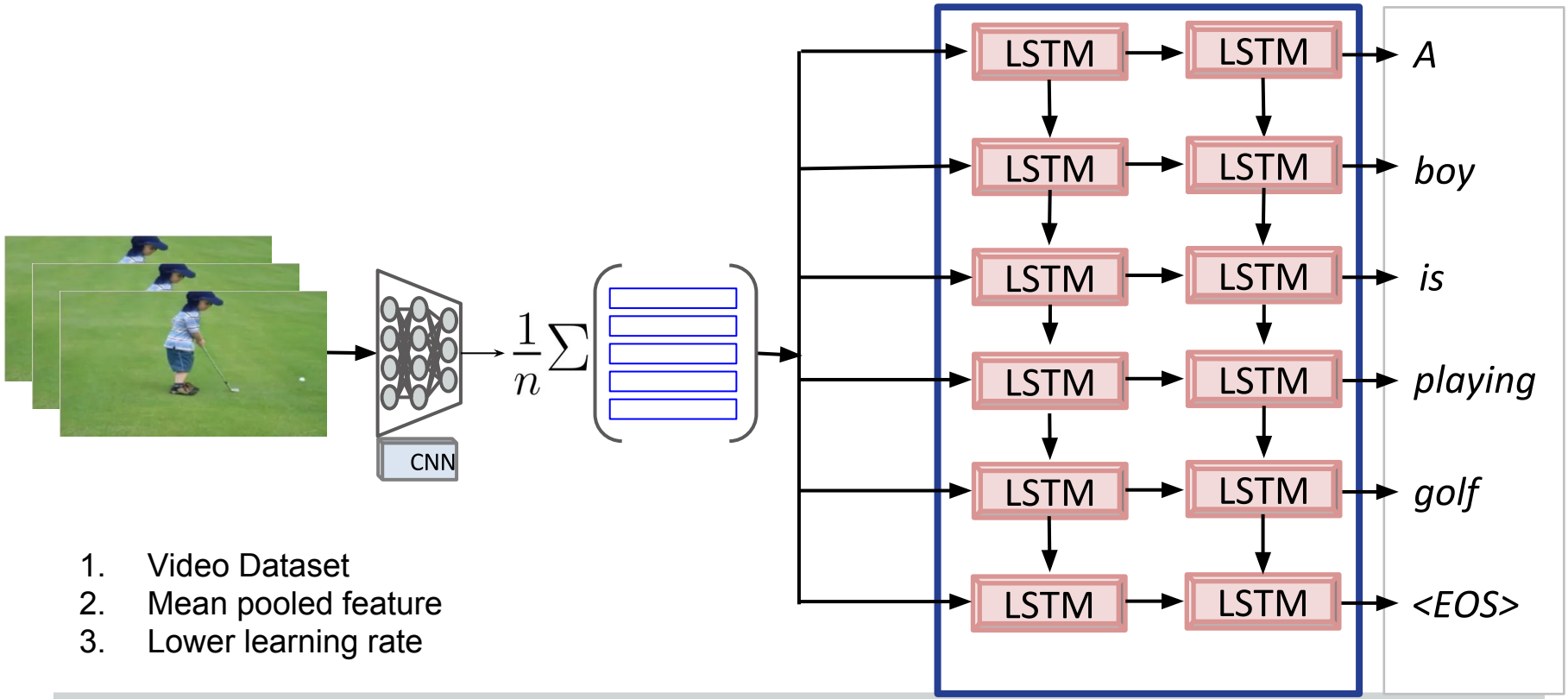


- Based on Alexnet [Krizhevsky et al. NIPS’12]
- 1.2M+ images from ImageNet ILSVRC-12 [Russakovsky et al.]
- Initialize weights of our network.

Step2: Image-Caption training



Step3: Fine-tuning



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

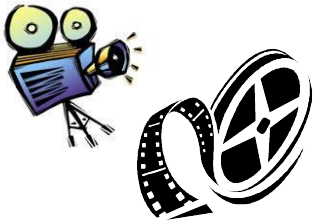
Experiments: Dataset

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

Augment Image datasets



YouTube

Training videos - 1300

Flickr30k - 30,000 images, 150,000 descriptions

MSCOCO - 120,000 images, 600,000 descriptions

Sample video and gold descriptions

- A man appears to be **plowing** a rice field with a plow being pulled by two **oxen**.
- A team of **water buffalo pull** a plow through a rice paddy.
- Domesticated **livestock** are helping a man **plow**.
- A man **leads** a team of oxen down a muddy path.
- Two **oxen walk** through some mud.
- 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.

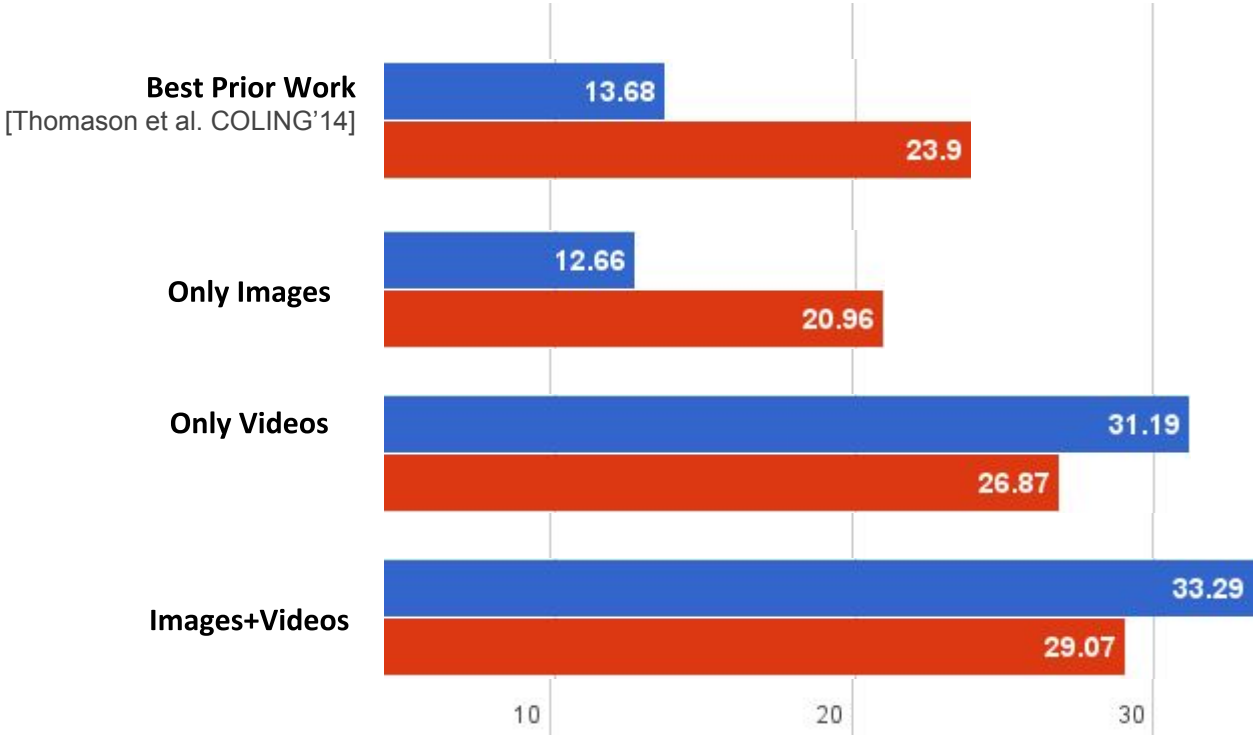
- A man is **walking** on a **rope**.
- A man is **walking** across a **rope**.
- A man is **balancing** on a **rope**.
- A man is **balancing** on a **rope** at the beach.
- A man **walks** on a **tightrope** at the beach.
- A man is **balancing** on a **volleyball net**.
- A man is **walking** on a **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.

Evaluation

- Machine Translation Metrics
  - BLEU
  - METEOR
- Human evaluation

Results - Generation

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



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

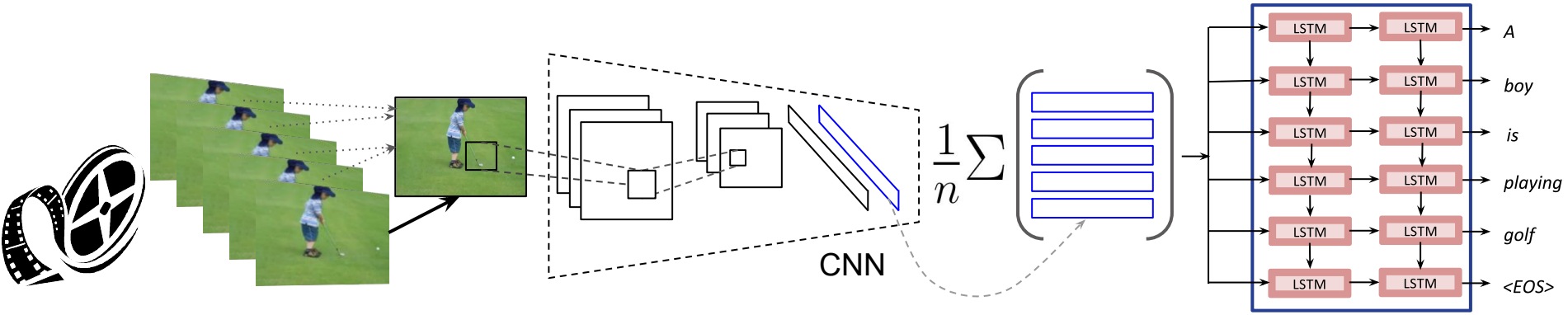
Multiple sentences can have same rating.

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

More Examples

Translating Videos to Natural Language



Does not consider temporal sequence of frames.

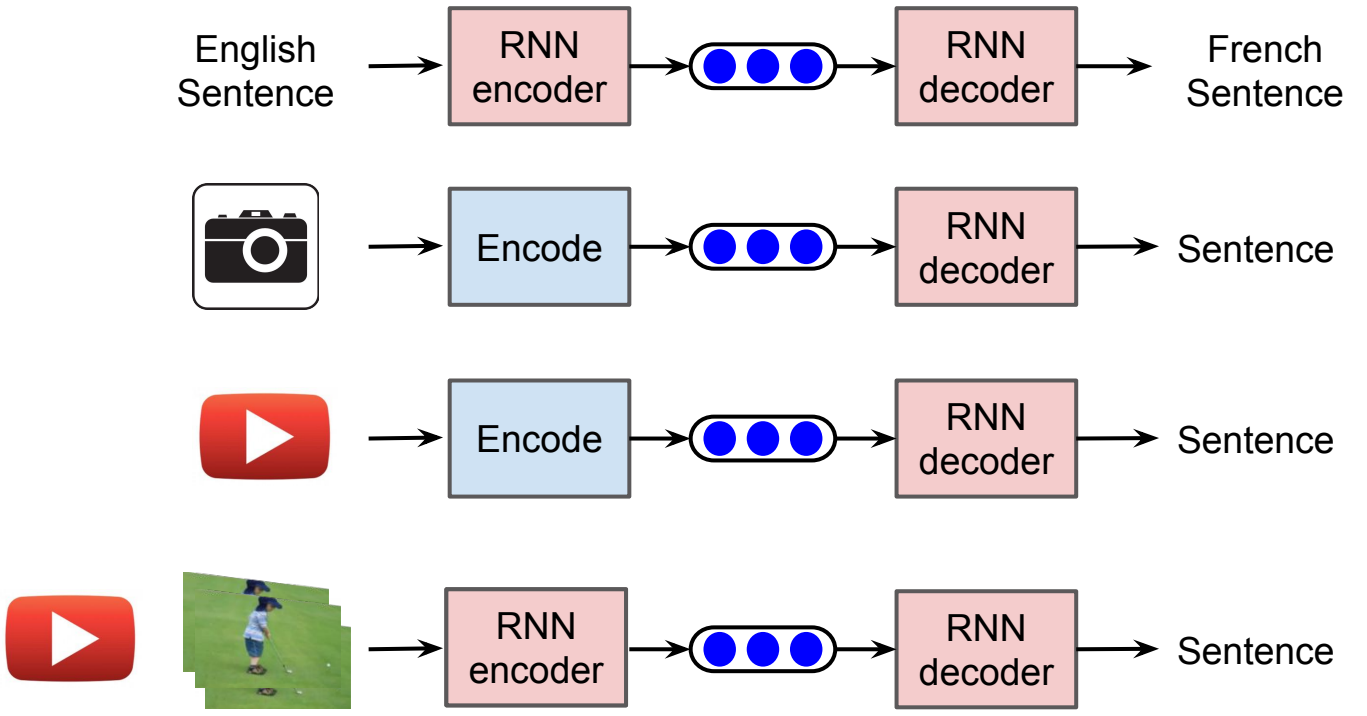
[Venugopalan et. al. NAACL'15]

Can our model be sensitive to temporal structure?

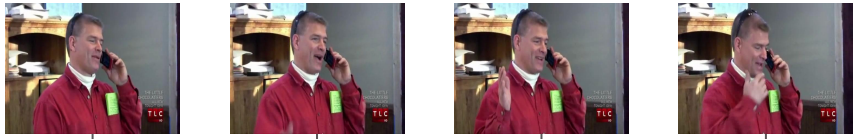
Allowing both input (sequence of frames) and output (sequence of words) of variable length.

[Venugopalan et. al. ICCV'15]

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



S2VT: Sequence to Sequence Video to Text



CNN CNN CNN CNN

LSTM LSTM LSTM LSTM LSTM LSTM LSTM LSTM ...

LSTM LSTM LSTM LSTM LSTM LSTM LSTM LSTM

Encoding stage

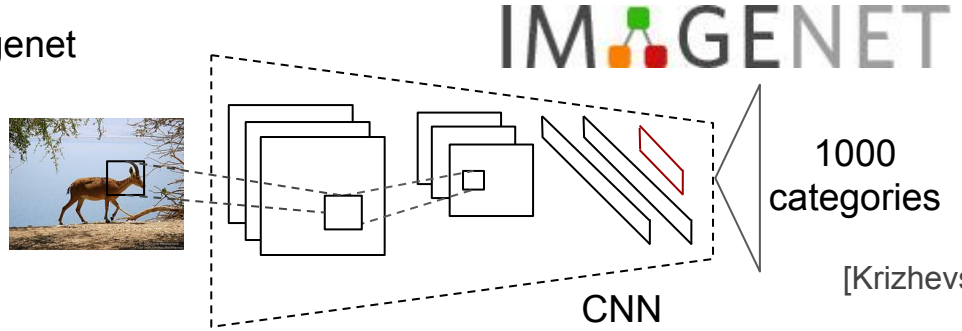
Now decode it to a sentence!

A man is talking ...

Decoding stage

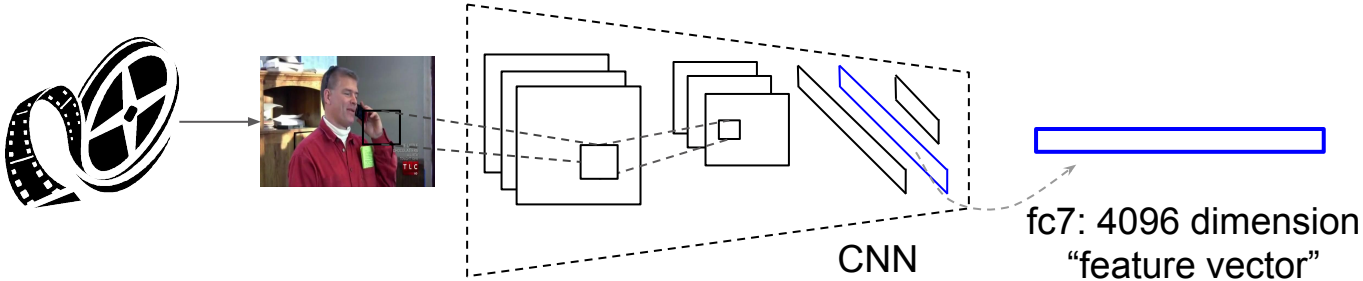
[Venugopalan et. al. ICCV'15]

1. Train on Imagenet



[Krizhevsky et al. NIPS'15]

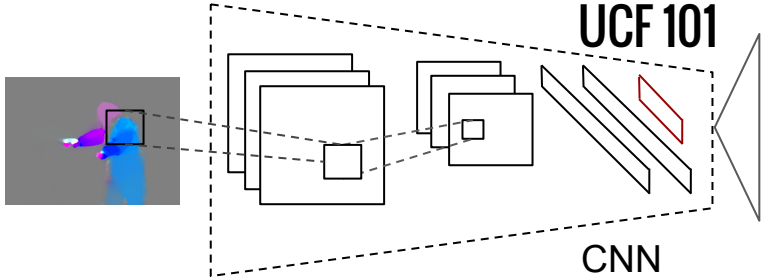
2. Take activations from layer before classification



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

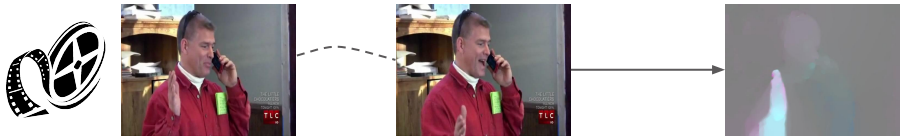
Frames: RGB

1. Train CNN on Activity classes



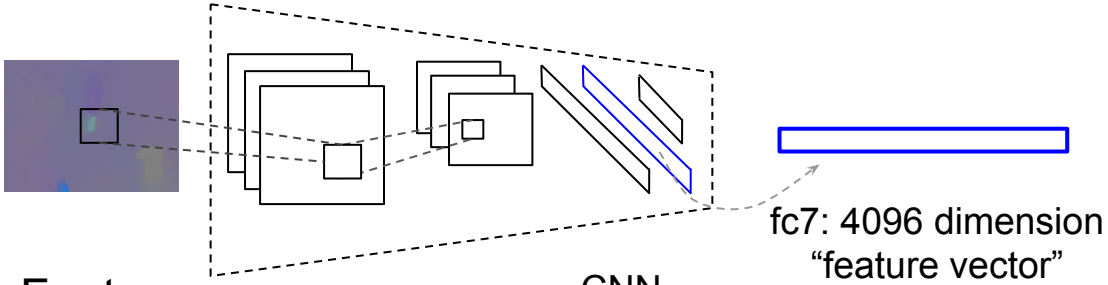
[Donahue et al. CVPR'15]

2. Use optical flow to extract flow images.



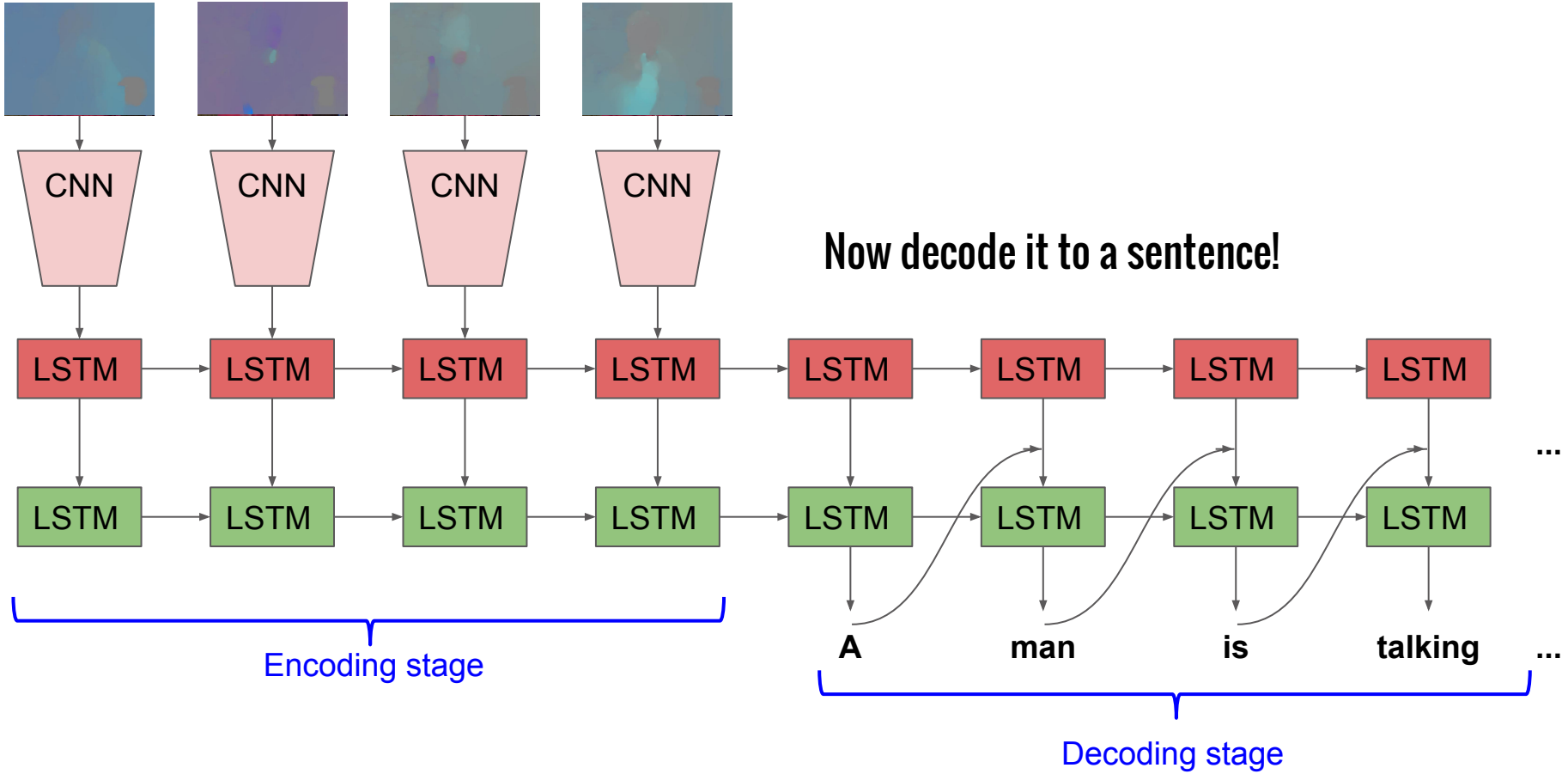
[T. Brox et. al. ECCV '04]

3. Take activations from layer before classification



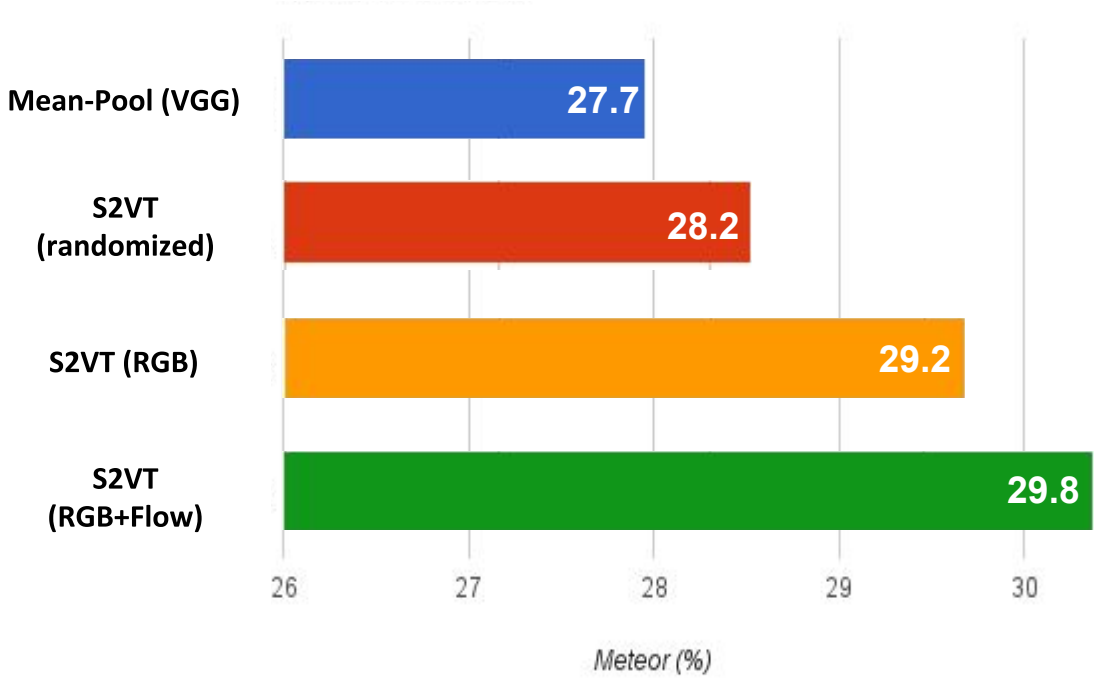
Explicit Activity Recognition Features

Frames: Flow



[Venugopalan et. al. ICCV'15]

Results (Youtube)



Movie Corpus - DVS

CC: Queen: "Which estate?"

DVS: Looking troubled, the Queen descends the stairs.

The Queen rushes into the courtyard. She then puts a head scarf on . . .

. . . and gets into the driver's side of a nearby Land Rover.

The Land Rover pulls away.

Three bodyguards quickly jump into a nearby car and follow her.

Processed:
Looking troubled, someone descends the stairs.

Someone rushes into the courtyard. She then puts a head scarf on ...

Evaluation: Movie Corpora

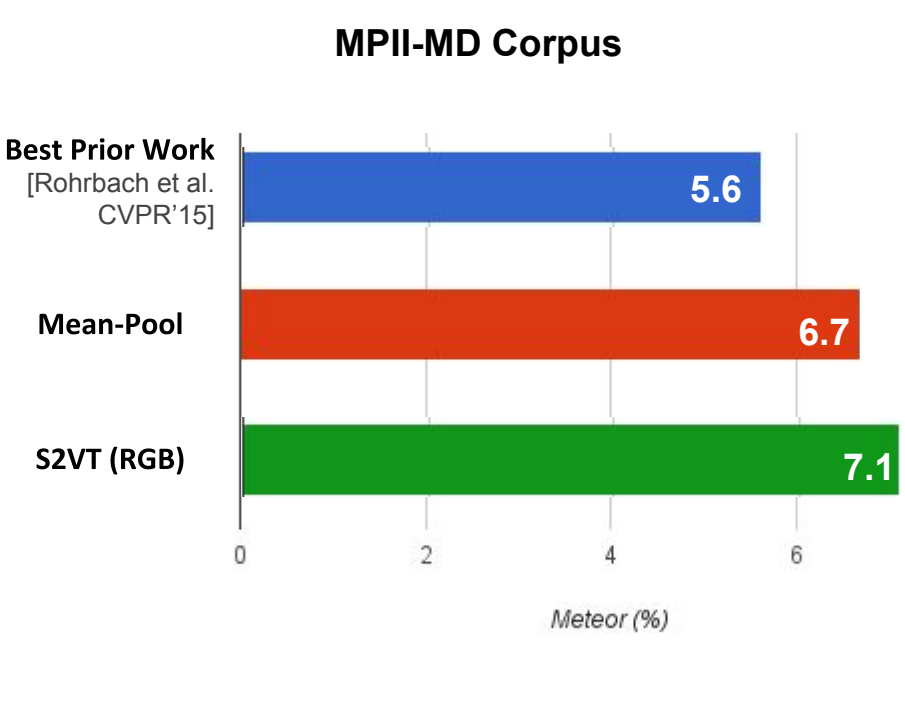
MPII-MD

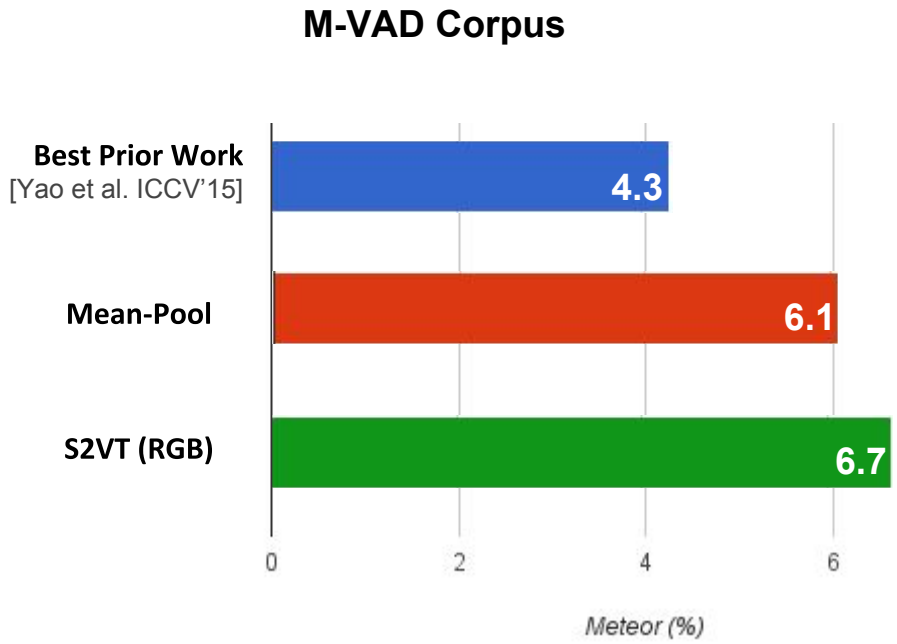
- MPII, Germany
- DVS alignment: semi-automated and crowdsourced
- 94 movies
- 68,000 clips
- Avg. length: 3.9s per clip
- **~1 sentence per clip**
- 68,375 sentences

M-VAD

- Univ. of Montreal
- DVS alignment: semi-automated and crowdsourced
- 92 movies
- 46,009 clips
- Avg. length: 6.2s per clip
- **1-2 sentences per clip**
- 56,634 sentences

Results (M-VAD Movie Corpus)





Examples (M-VAD Movie Corpus)

MPII-MD: https://youtu.be/XTq0huTXj1M
M-VAD: https://youtu.be/pER0mjzSYaM

Summary of completed work

- Two models for video description.
- Transfers from image-captioning task.
- Temporally sensitive.
  - Additionally includes activity features

Limitations

- Rely only on language in the paired text corpora.
  - Trained only on images and videos paired with captions.
  - Limits vocabulary.
- No mechanism to focus on particular objects.
- Can’t handle multiple events in longer videos.
  - Trained on pre-segmented clips.
- Generic DVS descriptions without character names.

Proposed Research

- Integrate external linguistic knowledge
- Attend to specific objects
- Segment multi-activity videos
- Character names for DVS

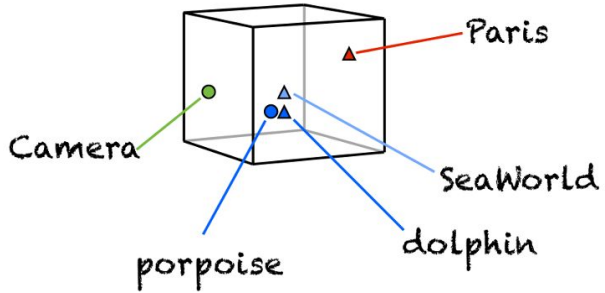
near-term

long-term

bonus

Integrating Linguistic Knowledge

1. Utilize word embeddings trained on external text corpora.



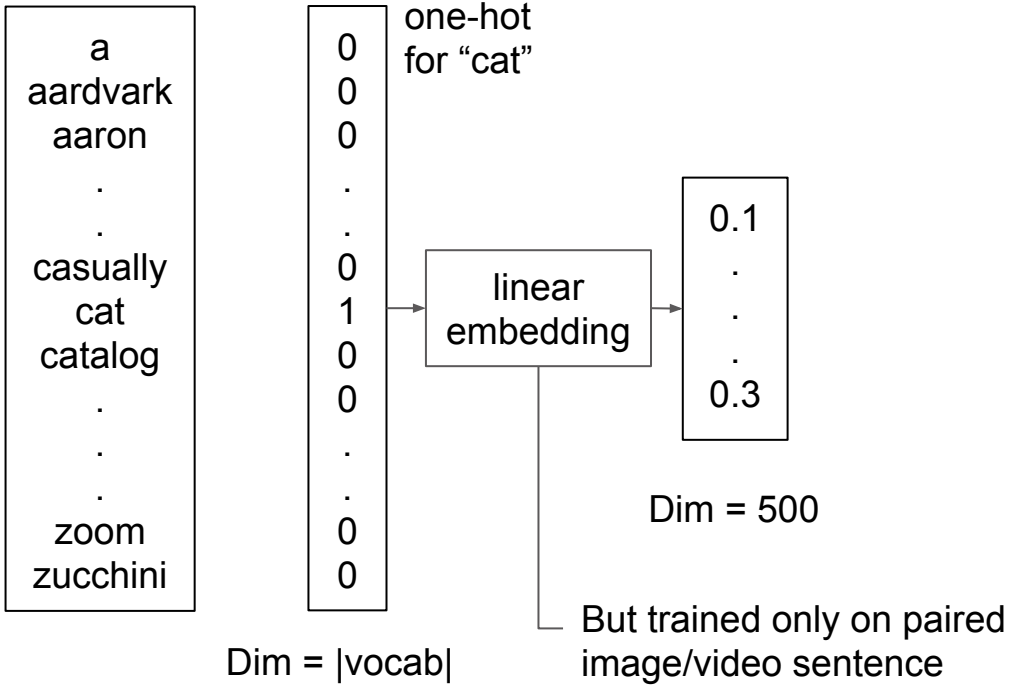
2. Pre-train the model (relevant layers) on text-only corpora.

3. Use an external language model.

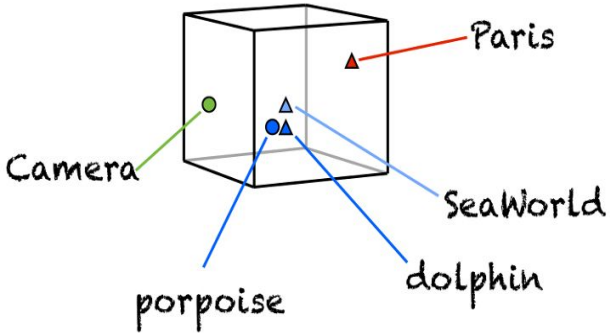
Integrating Linguistic Knowledge-1

Input Representation of Words

Vocabulary



Distributional Vectors



1. Captures semantics
2. Includes more context
3. Larger Vocabulary
4. Lower dimension

e.g. Word2Vec [Mikolov NIPS’13]
Glove [Pennington EMNLP’14]

Integrating Linguistic Knowledge-1

Representation of Words

Vocabulary

a
aardvark
aaron
.
.
casually
cat
catalog
.
.
.
zoom
zucchini

0
0
0
.
.
.
0
1
0
0
0
.
.
.
0
0

one-hot
for “cat”

linear
embedding

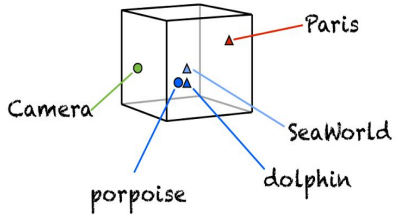
0.1
.
.
.
0.3

Dim = 500

Dim = |vocab|

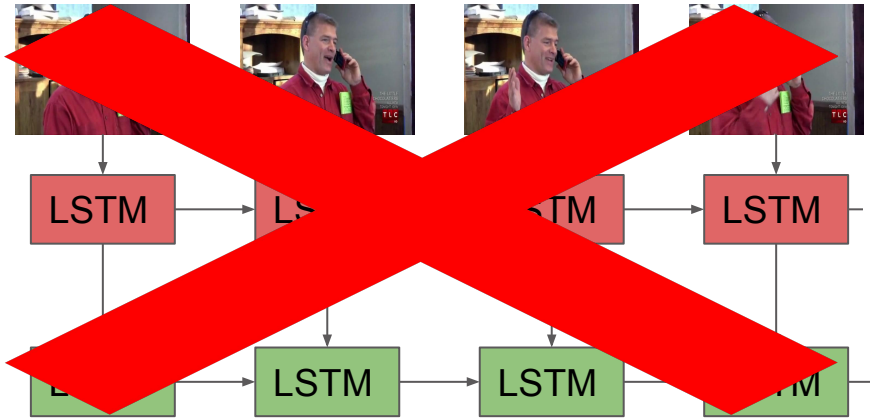
But trained only on paired
image/video sentence

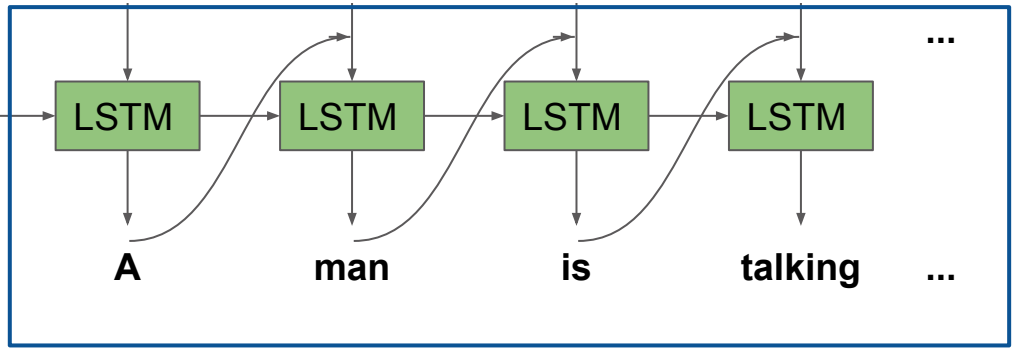
Distributional Vectors



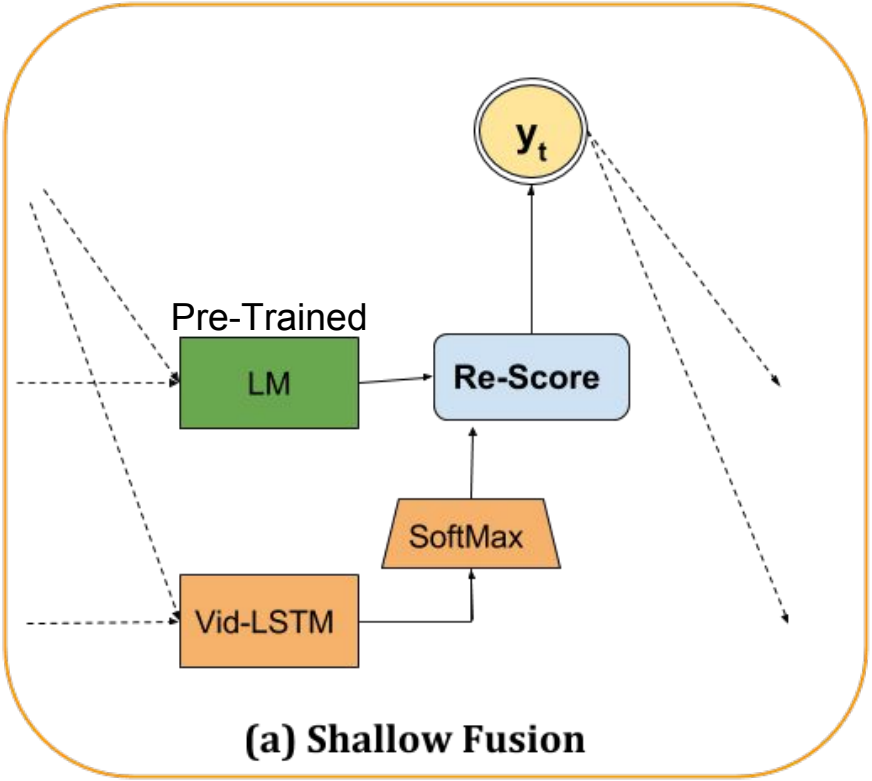
How?
1. Initialize Embedding
2. Learn mapping from distributional vectors
3. Concatenate both vectors

Integrating Linguistic Knowledge-2

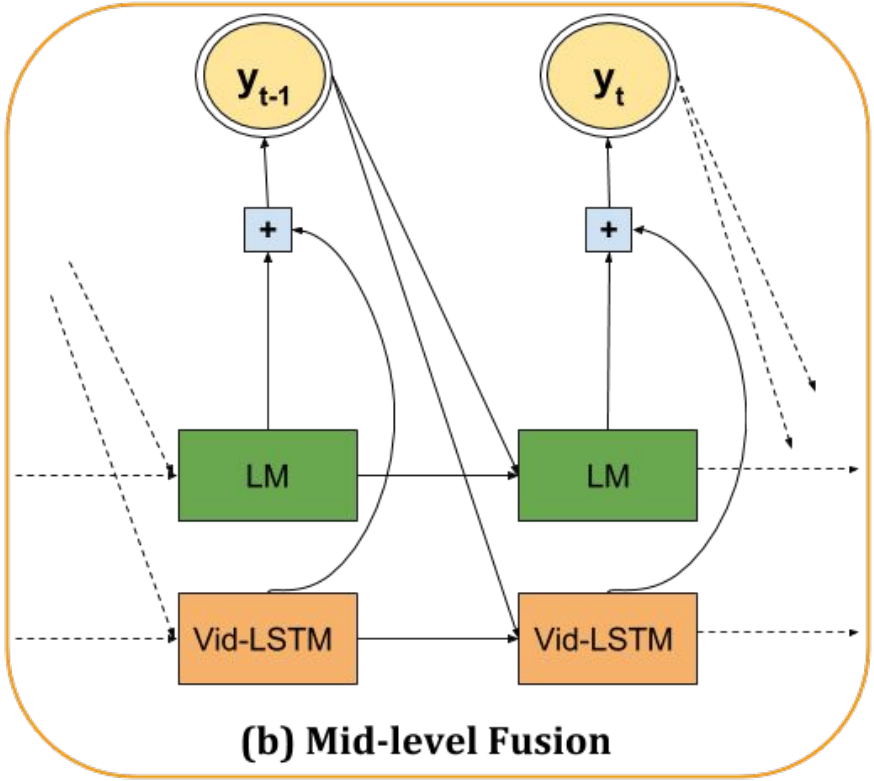




Integrating Linguistic Knowledge-3



(a) Shallow Fusion

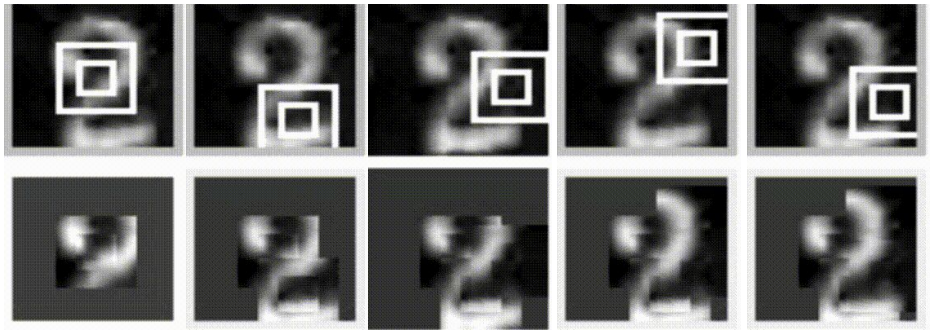


(b) Mid-level Fusion

Attention to focus on objects

"Attention": Sequentially processes regions in a single image.
Objective: Model learns "where to look" next.

Classify house numbers and translated MNIST digits.



[Mnih et al. NIPS'14]

Image Captioning

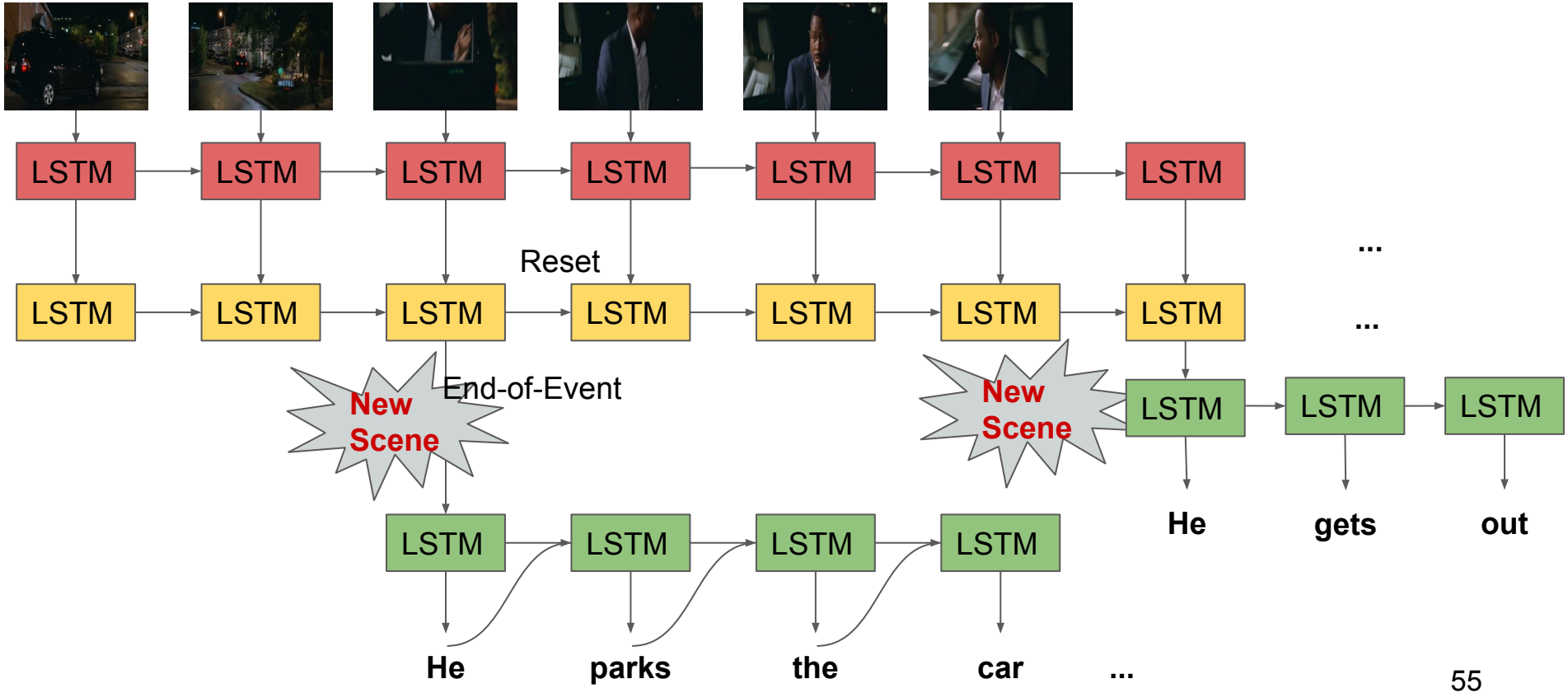
[Xu et al. ICML'15]

Attention

Attend to different regions/objects at each time step based on caption.

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

Longer Videos - 1



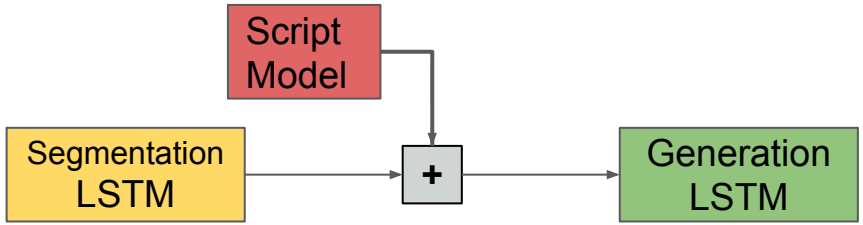
Longer Videos - 2

"scripts" encode stereotypical events

- ordered sequence of sub-events

E.g. open -> pour -> mix is a more likely event sequence than mix->open->pour

- Used to infer next event or missing event in the sequence.
- LSTMs to model scripts. [Pichotta and Mooney AAAI'15]



Movie Character Names [Bonus]

Proper Names replaced by “someone” during DVS training

+ Makes learning problem easier.
- Descriptions don’t associate characters with actions.

“Someone nods his head”

“Someone is driving the car”

“Someone opens the door for someone”

Movie Character Names [Bonus]

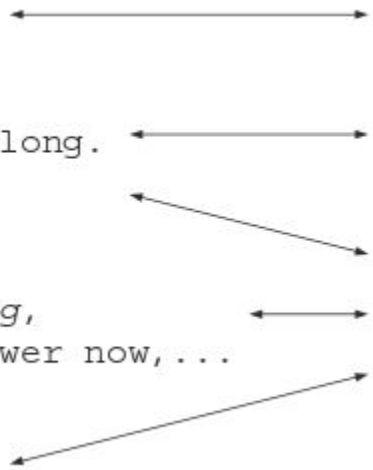
[Everingham et. al BMVC'06, Cour et. al CVPR'09, Cour et. al JMLR'11]

Subtitles

```
00:18:55,453 --> 00:18:56,086
Get out!

00:18:56,093 --> 00:19:00,044
- But, babe, this is where I belong.
- Out! I mean it.

00:19:00,133 --> 00:19:03,808
I've been doing a lot of reading,
and I'm in control of my own power now,...

00:19:03,893 --> 00:19:05,884
..so we're through.
```

Movie Script

```
HARMONY
Get out.

SPIKE
But, baby... This is where I belong.

HARMONY
Out! I mean it. I've done a lot of
reading, and, and I'm in control
of my own power now. So we're
through.
```



Dialogues
Has Timing!
No character names :(

Has Characters!
Has conversation!
No timing information :(

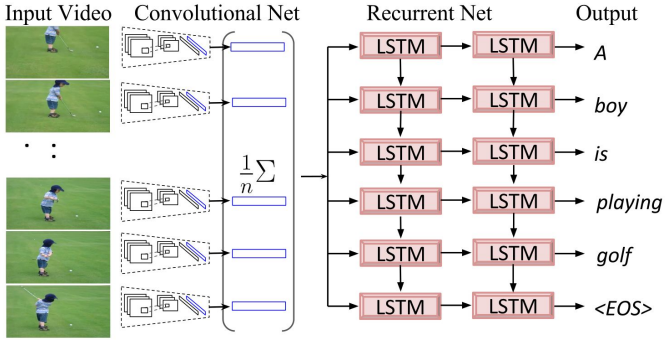
Movie Character Names

- Align Scripts and Subtitles
- Identify times at which characters appear.
- Face Detection + Multiple Instance Learning

Alternate: Learn actor faces. Use acting credits.

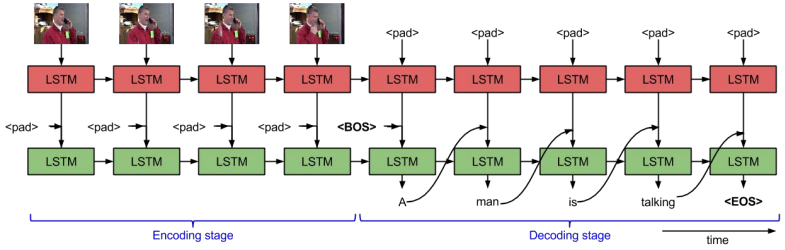
Conclusion

1. Completed Research



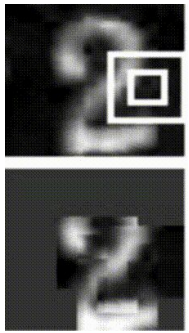
Two fully deep models to generate descriptions for videos.

1. Learns to transfer from paired image-captions to videos.
2. Jointly models a sequence of frames and sequence of words.

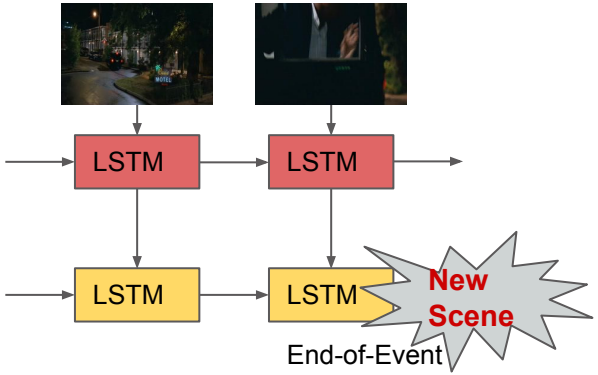


2. Proposed Directions

Include external linguistic knowledge



Attend to objects



Multi-activity videos

Hermione pours it into the pot.

DVS character names

Publications

[1] **S. Venugopalan**, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. ICCV 2015.

[2] **S. Venugopalan**, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. NAACL-HLT 2015.

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, **S. Venugopalan**, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. CVPR 2015.

[4] J. Thomason*, **S. Venugopalan***, S. Guadarrama, K. Saenko, R. Mooney. Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild. COLING 2014.

[5] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, **S. Venugopalan**, T. Darrell, R. Mooney, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. ICCV 2013.

Thank You

Mean-Pool model (data and code): https://gist.github.com/vsubhashini/3761b9ad43f60db9ac3d
S2VT (code): https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt