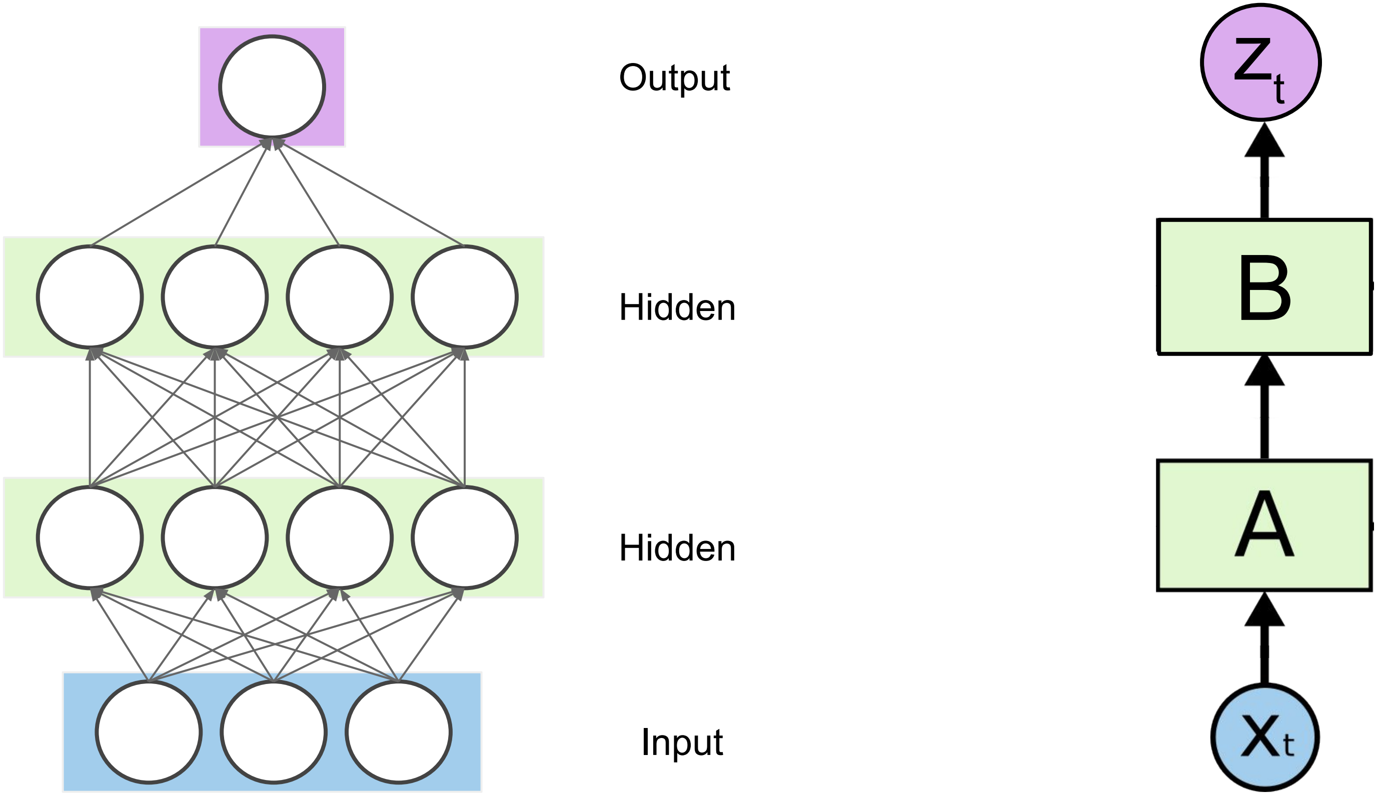
LSTMs Overview

Subhashini Venugopalan

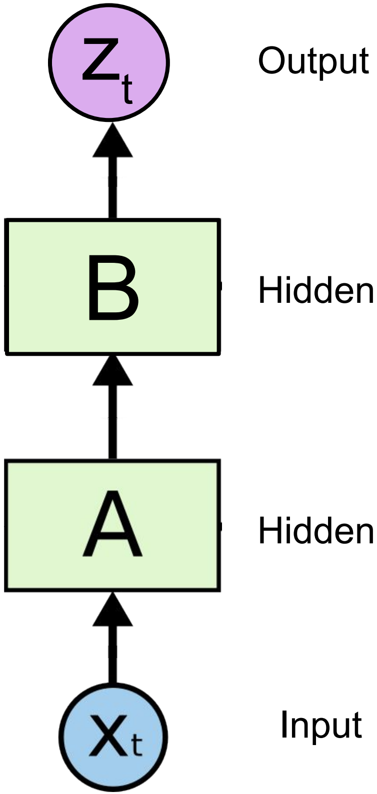
Neural Networks



WHY RNNS/LSTMS?

Can we operate over sequences of inputs?

Limitations of vanilla Neural Networks

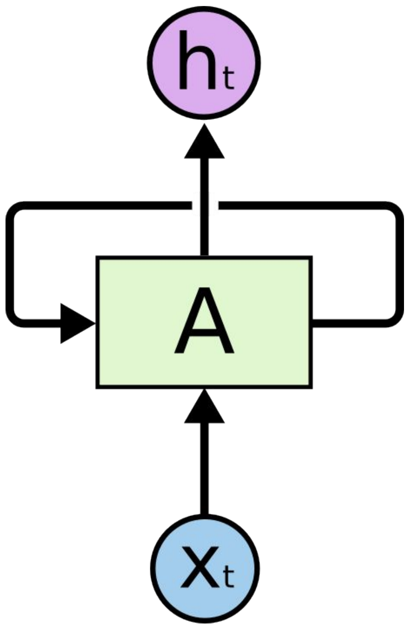


Outputs a fixed size vector.

Performs a fixed number of computations (#layers).

Accepts only fixed size input e.g 224x224 images.

Recurrent Neural Networks

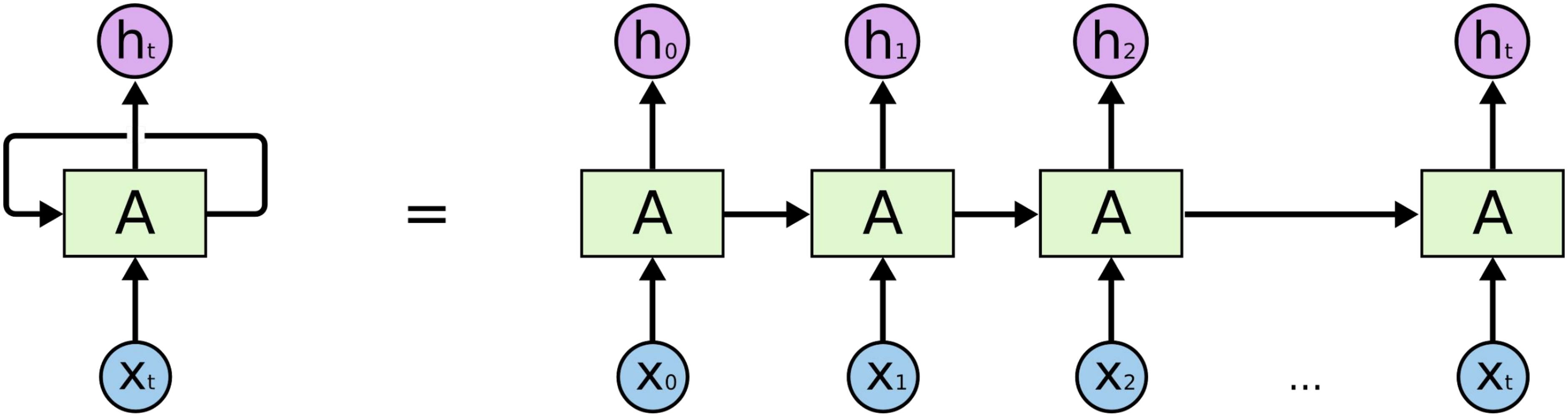


They are networks with loops. [Elman '90]

Image Credit: Chris Olah

Un-Roll The Loop

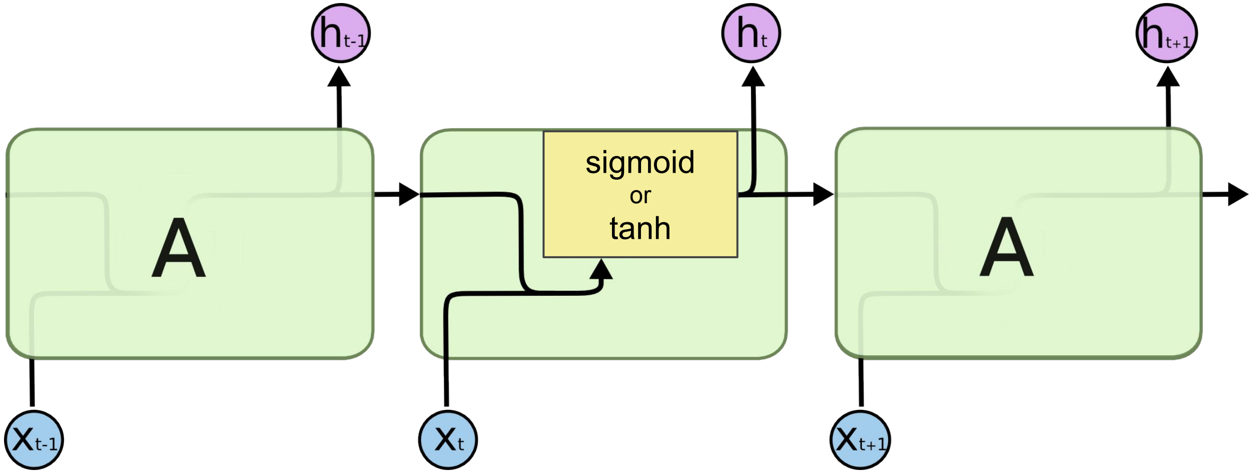
Recurrent Neural Network "unrolled in time"



- Each time step has a layer with the same weights.
- The repeating layer/module is a sigmoid or a tanh.
- Learns to model $(h_t \mid x_1, \ldots, x_{t-1})$

Image Credit: Chris Olah

Simple RNNs



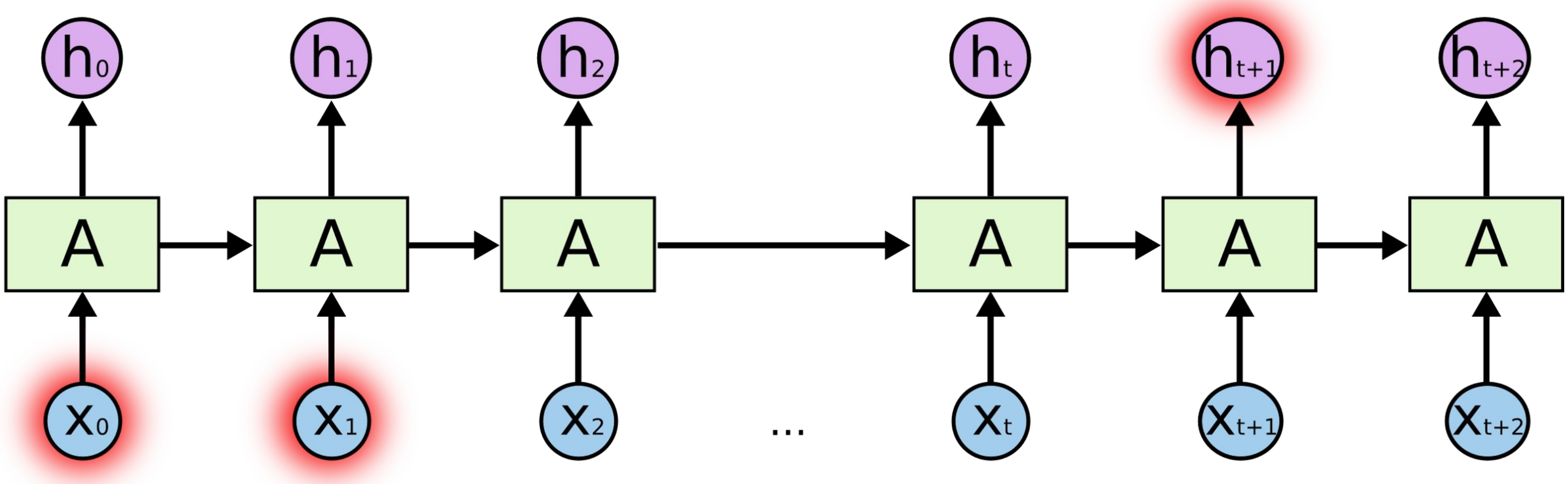
$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

$$z_t = g(W_{hz}h_t + b_z)$$

sigmoid or tanh

Image Credit: Chris Olah

Problems with Simple RNNs

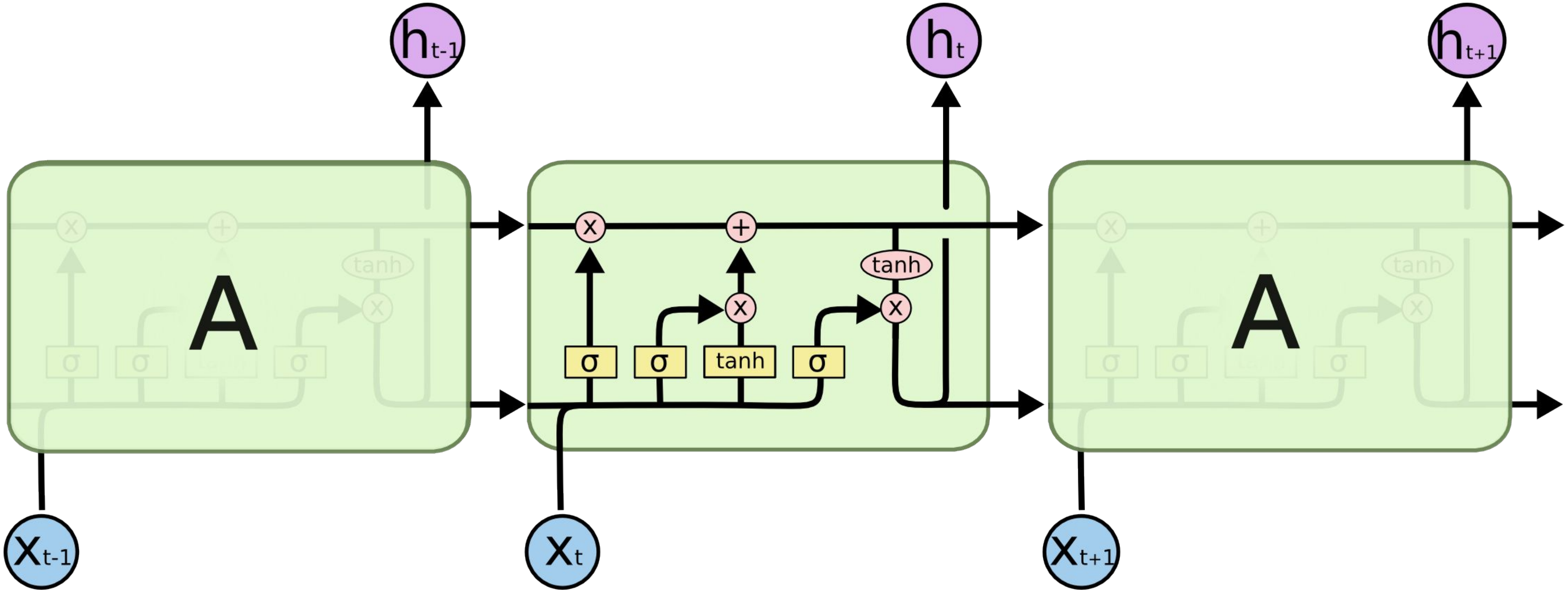


- Can't seem to handle "long-term dependencies" in practice
- Gradients shrink through the many layers (Vanishing Gradients)

[Hochreiter '91]
[Bengio et. al. '94]

Image Credit: Chris Olah

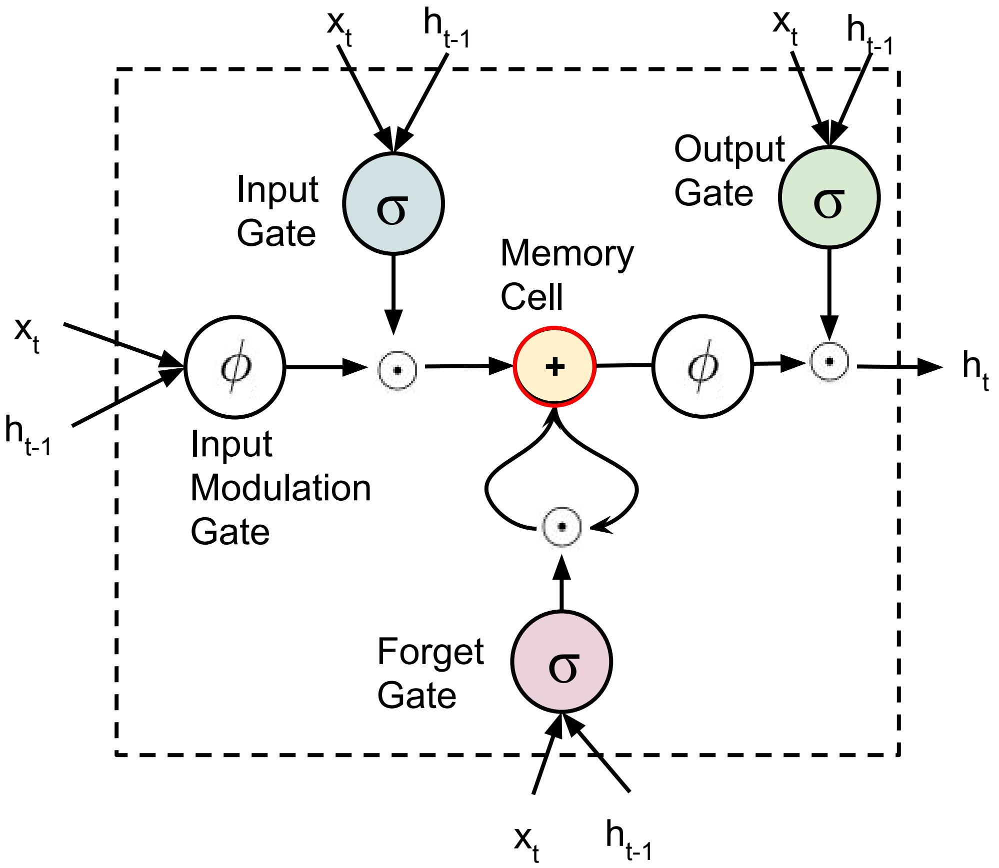
Long Short Term Memory (LSTMs)



[Hochreiter and Schmidhuber '97]

Image Credit: Chris Olah

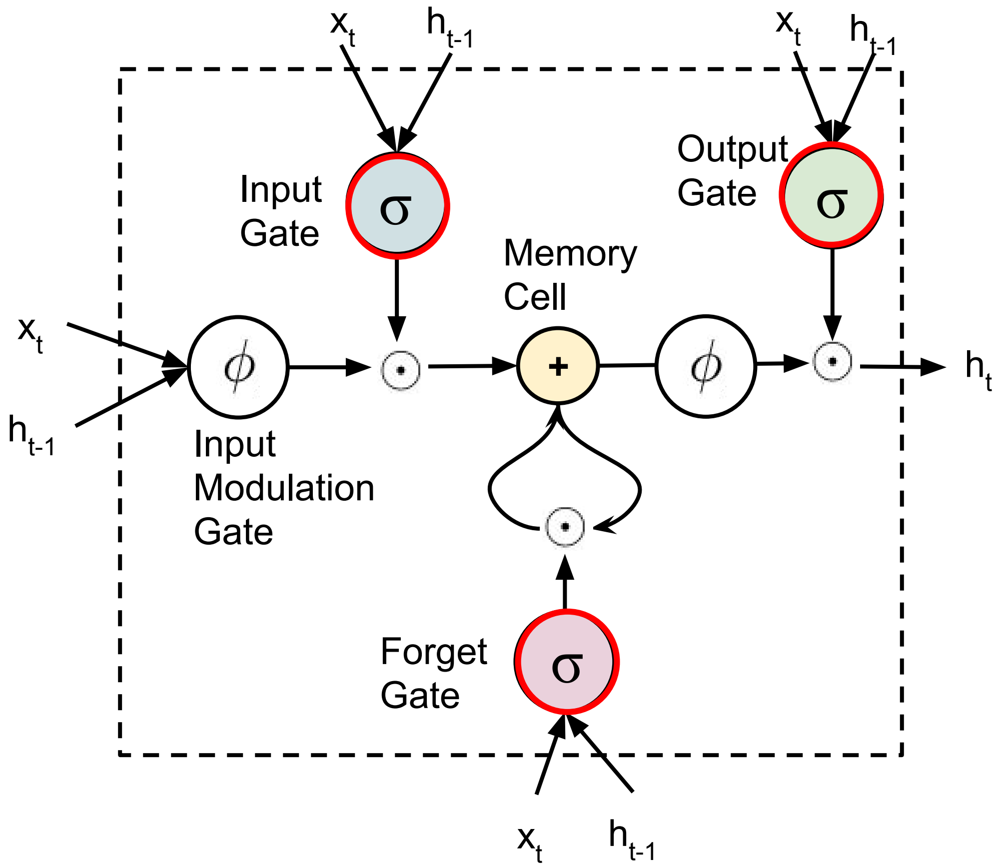
LSTM Unit



Memory Cell:
Core of the LSTM Unit
Encodes all inputs observed

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

LSTM Unit

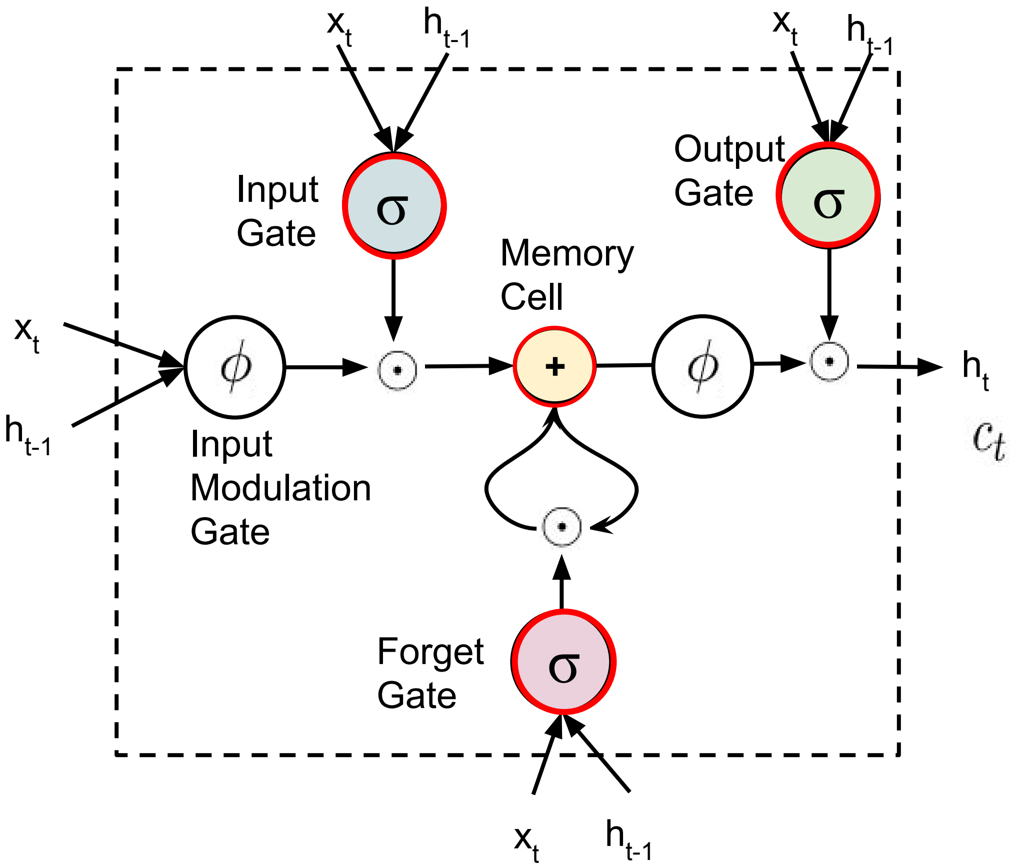


Memory Cell:
Core of the LSTM Unit
Encodes all inputs observed

Gates:
Input, Output and Forget
Sigmoid [0,1]

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

LSTM Unit



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

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

Update the Cell state

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

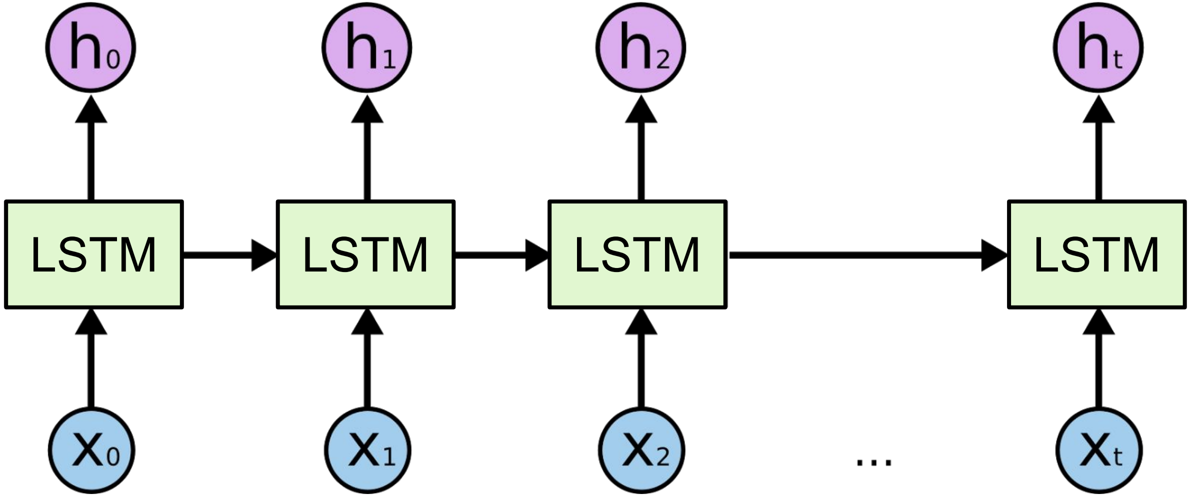
Learns long-term dependencies

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

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

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

Can Model Sequences

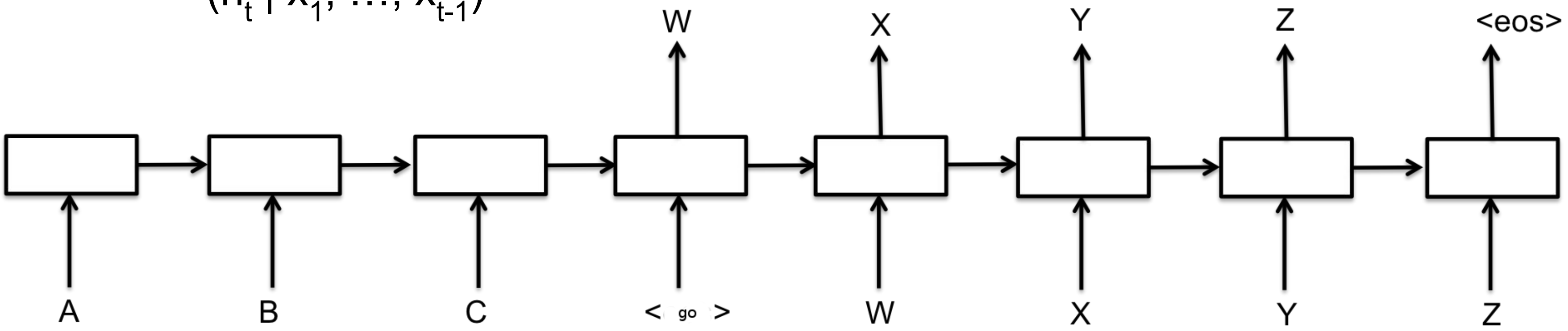


- Can handle longer-term dependencies
- Overcomes Vanishing Gradients problem
- GRUs - Gated Recurrent Units is a much simpler variant which also overcomes these issues.

[Cho et. al. ‘14]

Putting Things Together

Encode a sequence of inputs to a vector.
$(h_t \mid x_1, \ldots, x_{t-1})$



Decode from the vector to a sequence of outputs.
$Pr(x_t \mid x_1, \ldots, x_{t-1})$

Image Credit: Sutskever et. al.

SOLVE A WIDER RANGE OF PROBLEMS

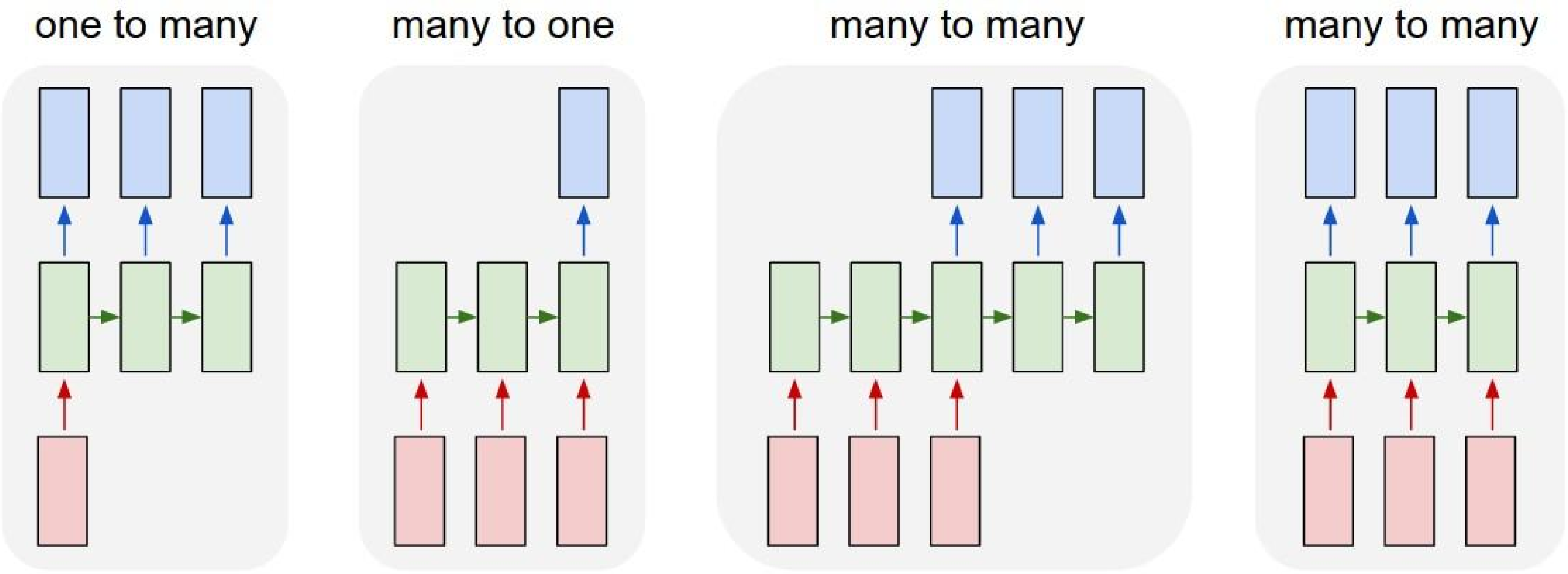


Image Captioning

Vinyals et. al. ‘15,
Donahue et. al. ‘15

Activity Recognition

Donahue et. al. ‘15

Sequence to Sequence

Machine Translation Sutskever et. al. ‘14, Cho et. al. ‘14
Speech Recognition Graves & Jaitly ‘14
Video Description V. et. al. ‘15, Li et. al. ‘15
VQA, POS tagging, ... 3 of 4 papers to be discussed this class

Image Credit: Andrej Karpathy

Resources

- Graves' paper - LSTMs explanation. Generating sequences with recurrent neural networks. Applications to handwriting and speech recognition.
- Chris' Blog - LSTM unit explanation.
- Karpathy's Blog - Applications.
- Tensorflow and Caffe - Code examples.

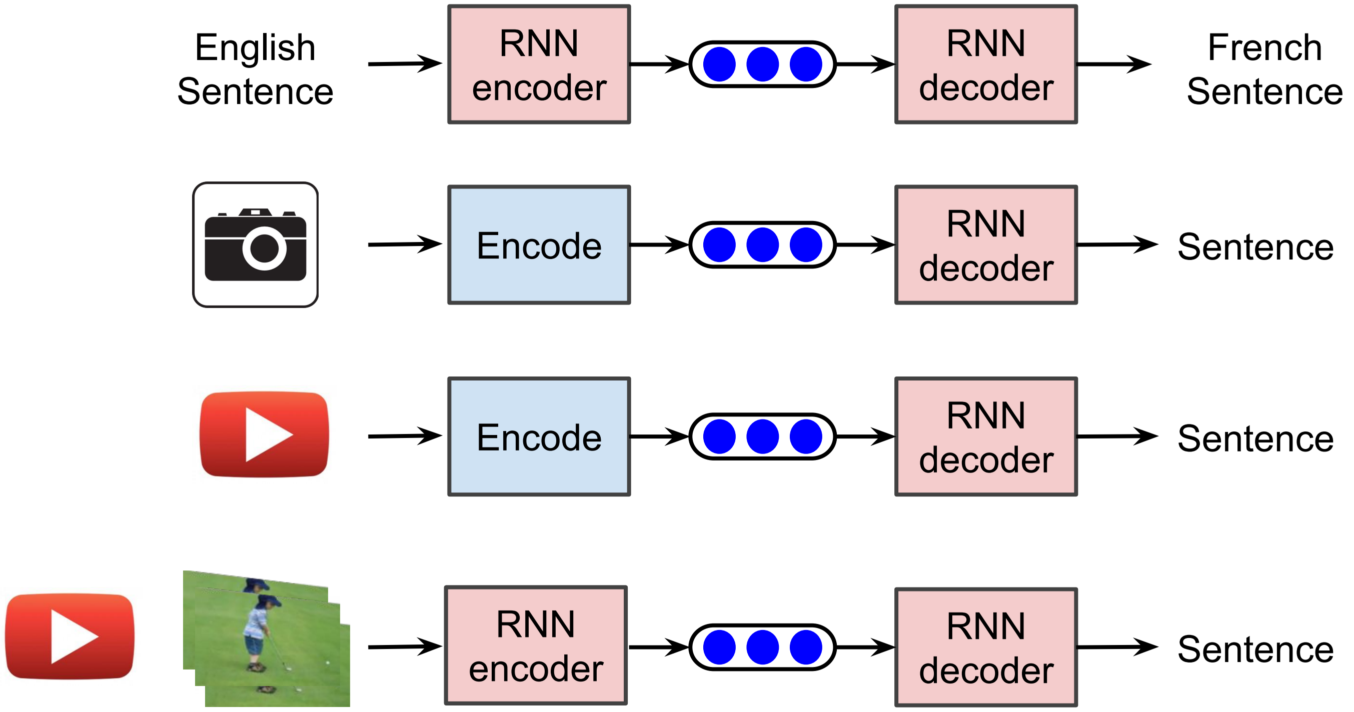
Sequence to Sequence
Video to Text

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue
Raymond Mooney, Trevor Darrell, Kate Saenko

Objective

A monkey is pulling a dog’s tail and is chased by the dog.

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



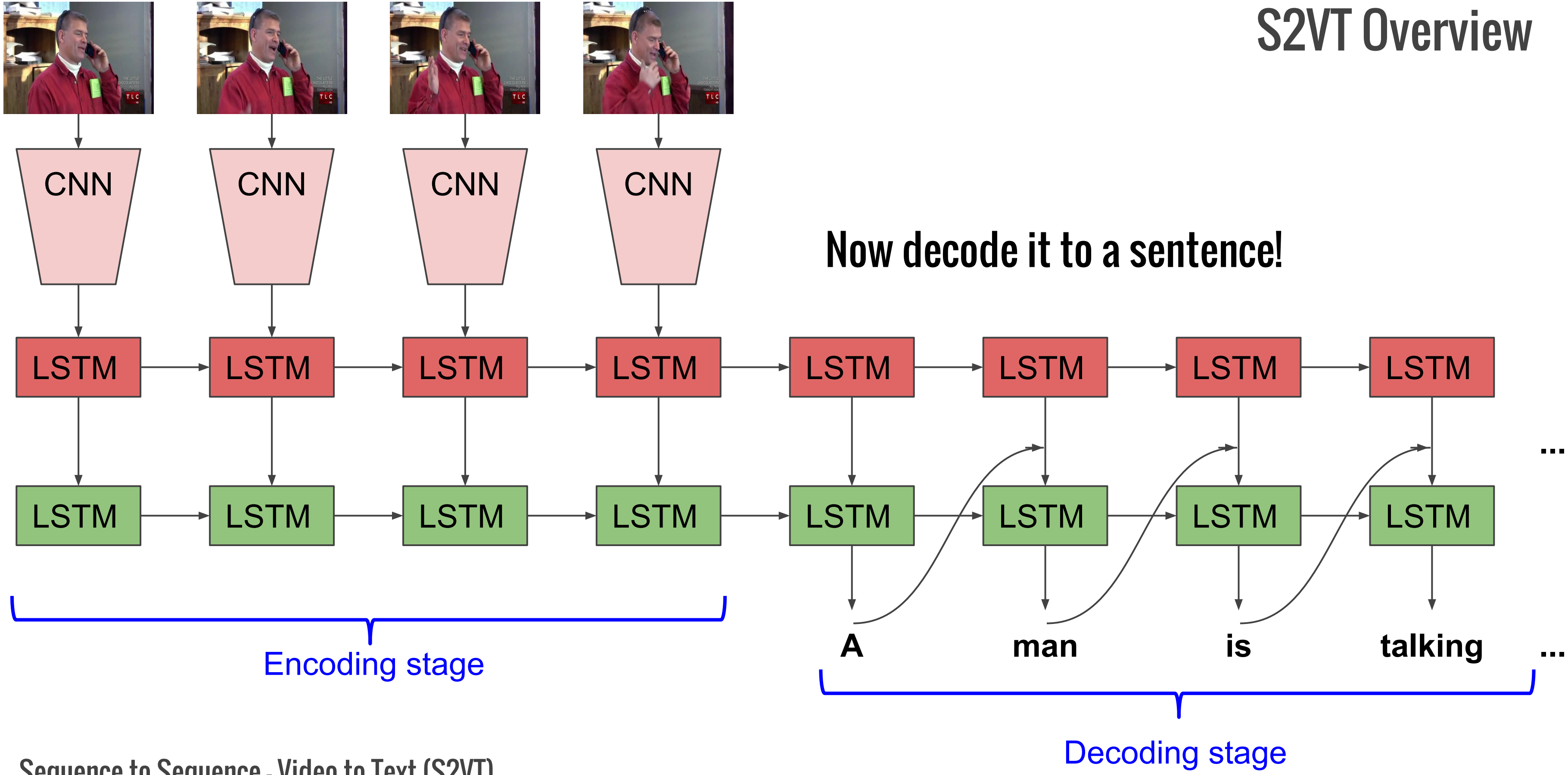
[Sutskever et al. NIPS’14]

[Donahue et al. CVPR’15]
[Vinyals et al. CVPR’15]

[Venugopalan et. al. NAACL’15]

[Venugopalan et. al. ICCV’ 15] (this work)

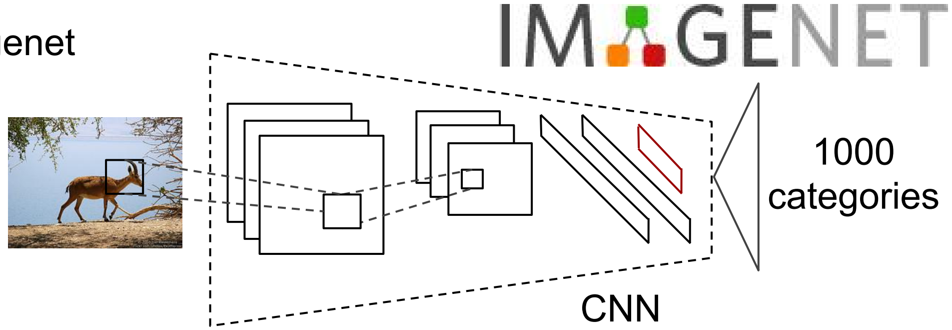
S2VT Overview



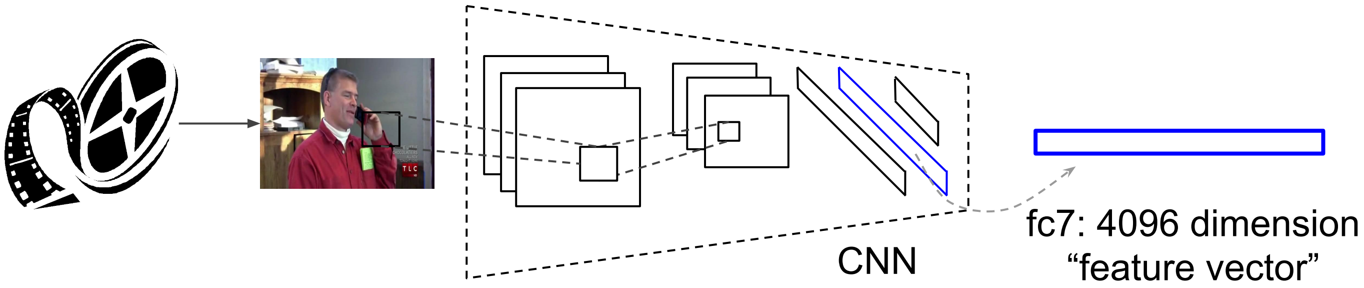
Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko

Frames: RGB

1. Train on Imagenet



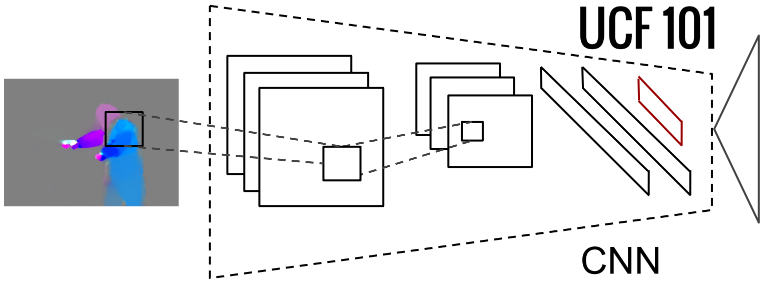
2. Take activations from layer before classification



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

Frames: Flow

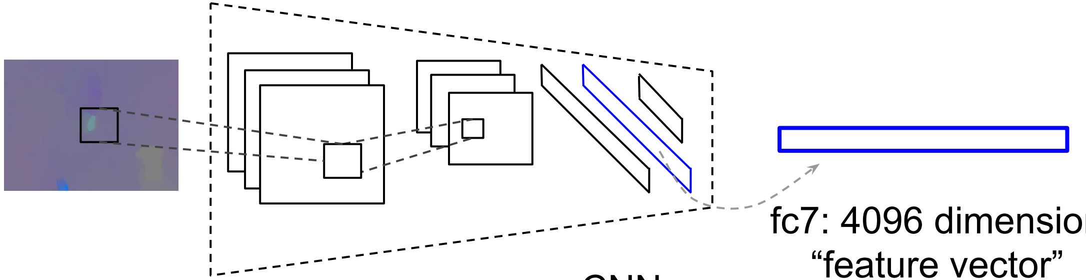
1. Train CNN on Activity classes



2. Use optical flow to extract flow images.

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

3. Take activations from layer before classification

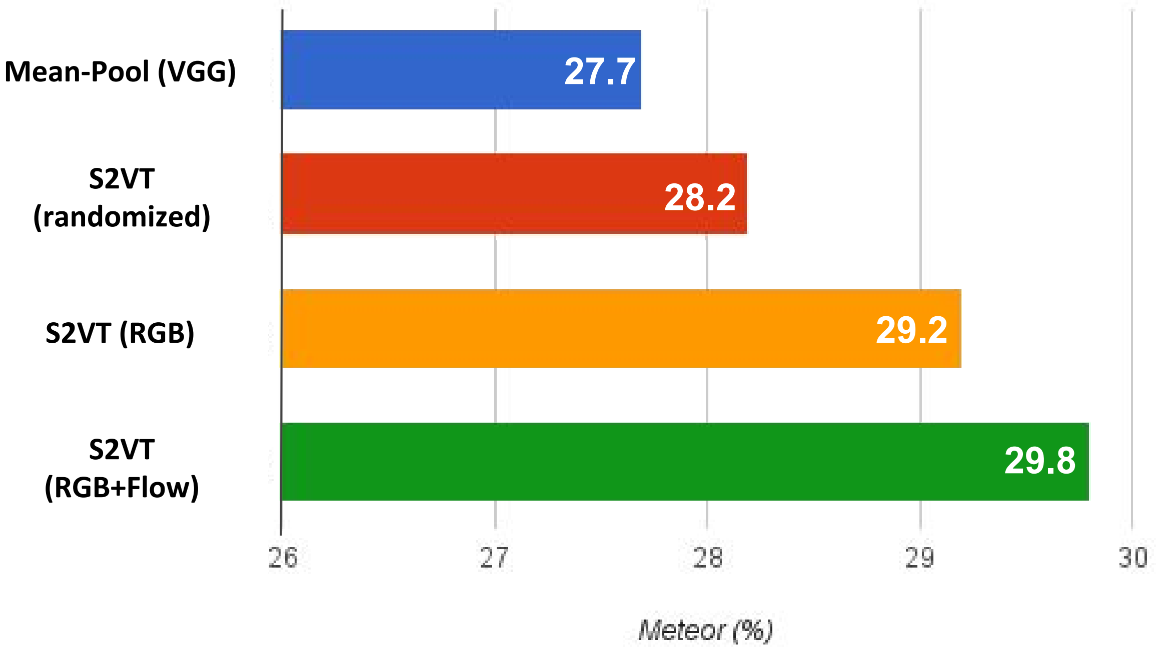


Dataset: Youtube

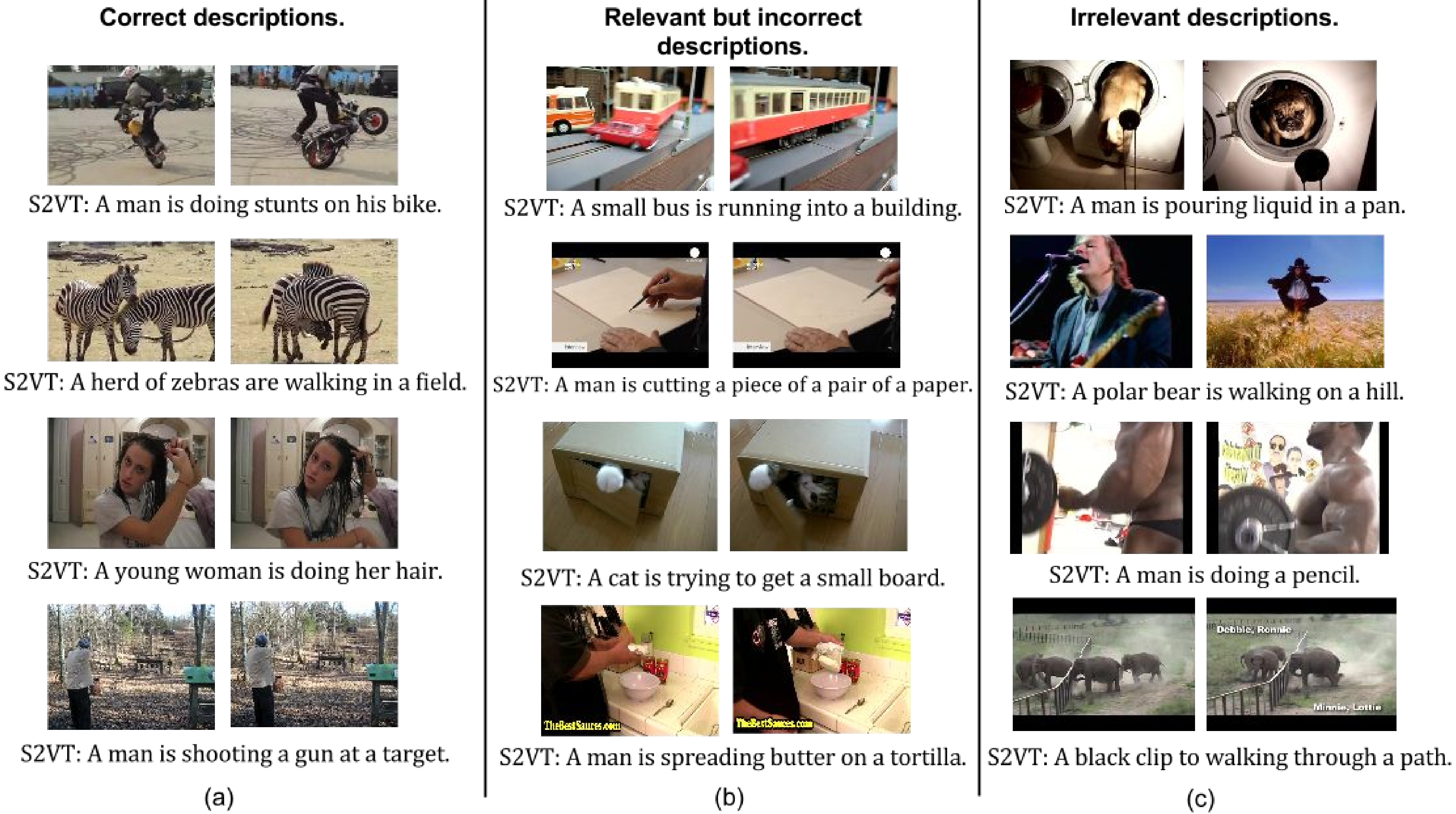
- ~2000 clips
- Avg. length: 11s per clip
- **~40 sentence per clip**
- ~81,000 sentences

- 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.

Results (Youtube)



METEOR: MT metric. Considers alignment, para-phrases and similarity.


Correct descriptions.
S2VT: A man is doing stunts on his bike.
S2VT: A herd of zebras are walking in a field.
S2VT: A young woman is doing her hair.
S2VT: A man is shooting a gun at a target.
(a)
Relevant but incorrect descriptions.
S2VT: A small bus is running into a building.
S2VT: A man is cutting a piece of a pair of a paper.
S2VT: A cat is trying to get a small board.
TheBestSauces.com
S2VT: A man is spreading butter on a tortilla.
(b)
Irrelevant descriptions.
S2VT: A man is pouring liquid in a pan.
S2VT: A polar bear is walking on a hill.
S2VT: A man is doing a pencil.
Debbie, Ronnie
Minnie, Lottie
S2VT: A black clip to walking through a path.
(c)

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: automated speech extraction
- 92 movies
- 46,009 clips
- Avg. length: 6.2s per clip
- **1-2 sentences per clip**
- 56,634 sentences

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.

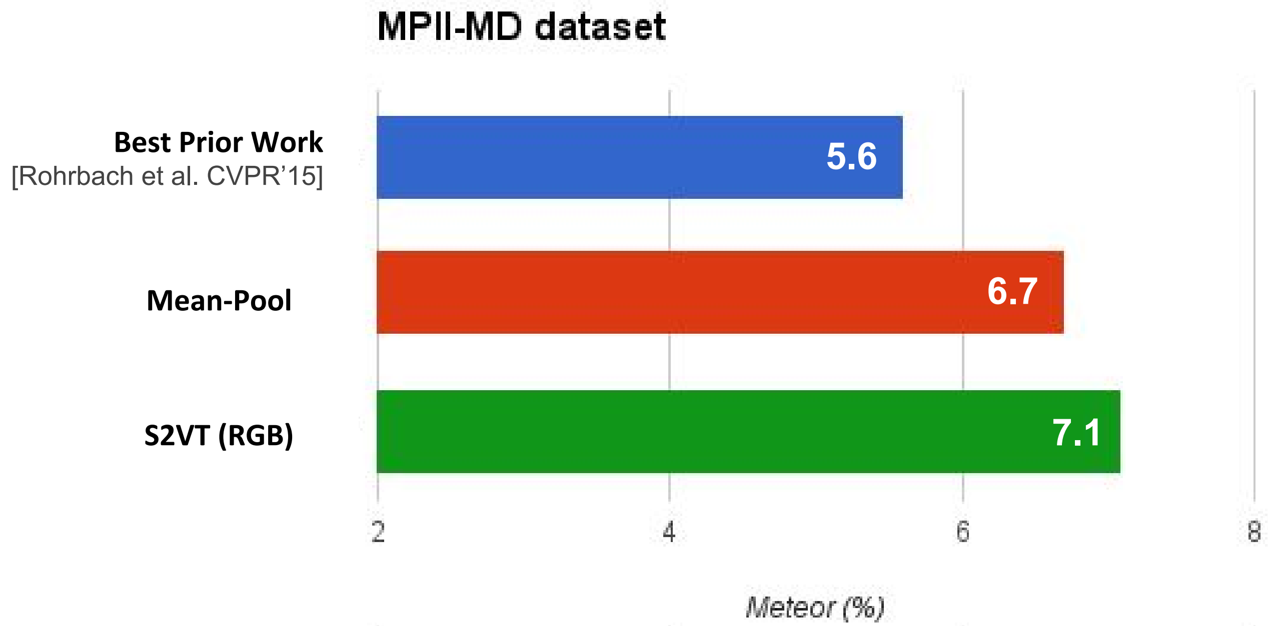
The Land Rover pulls away.

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

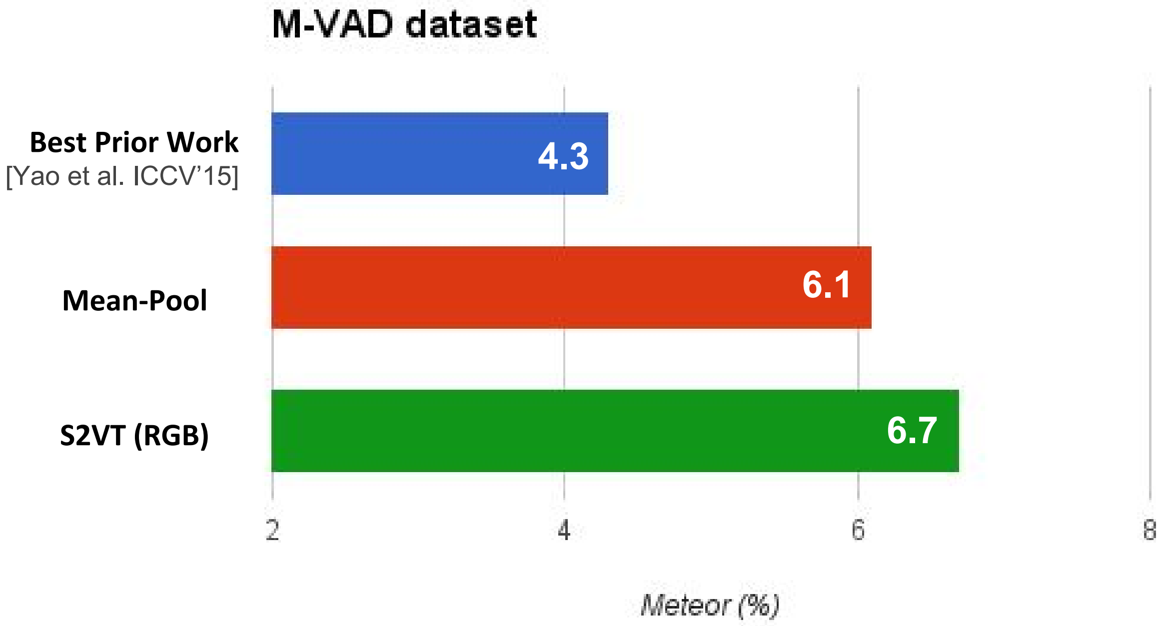
Processed: Looking troubled, someone descends the stairs.

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

Results (MPII-MD Movie Corpus)



Results (M-VAD Movie Corpus)



M-VAD: https://youtu.be/pEROmjzSYaM

Discussion

- What are the advantages/drawbacks of this approach?
  - End-to-end, annotations
- Detaching recognition and generation.
- Why only METEOR (not BLEU or other metrics)?
- Domain adaptation, Re-use RNNs (youtube -> movies, activity recognition)
- Languages other than English.
- Features apart from Optical Flow, RGB; temporal representation.

Edit-Distance	$k = 0$	$k <= 1$	$k <= 2$	$k <= 3$
MSVD	42.9	81.2	93.6	96.6
MPII-MD	28.8	43.5	56.4	83.0
MVAD	15.6	28.7	37.8	45.0

Table 3. Percentage of generated sentences which match a sentence of the training set with an edit (Levenshtein) distance of less than 4. All values reported in percentage (%).

Code and more examples
http://vsubhashini.github.io/s2vt.html

Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko