University of Nottingham
UK | CHINA | MALAYSIA

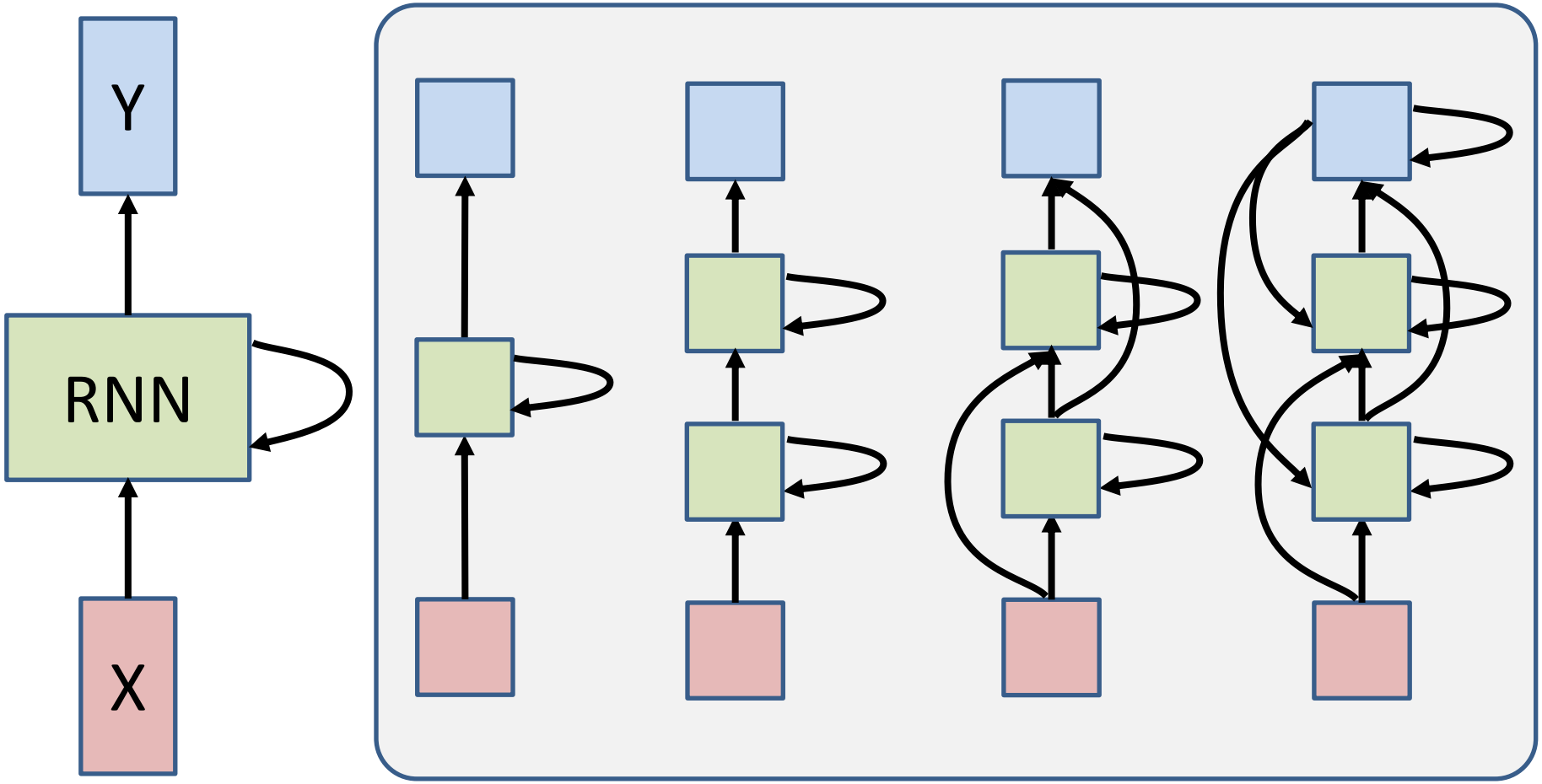
COMP3055 Machine Learning

Topic 14 – Selected Topics on Deep Learning - LSTM

Dr. Zheng LU

2018 Autumn

RNN



RNN can be designed very sophisticatedly with different layers different ways of recurrency

RNN

- Excellent models for problems more than one-to-one
  - Time series prediction and classification
  - Sequence prediction and classification
  - Simplify some problems that are difficult for multi-layer perceptron.

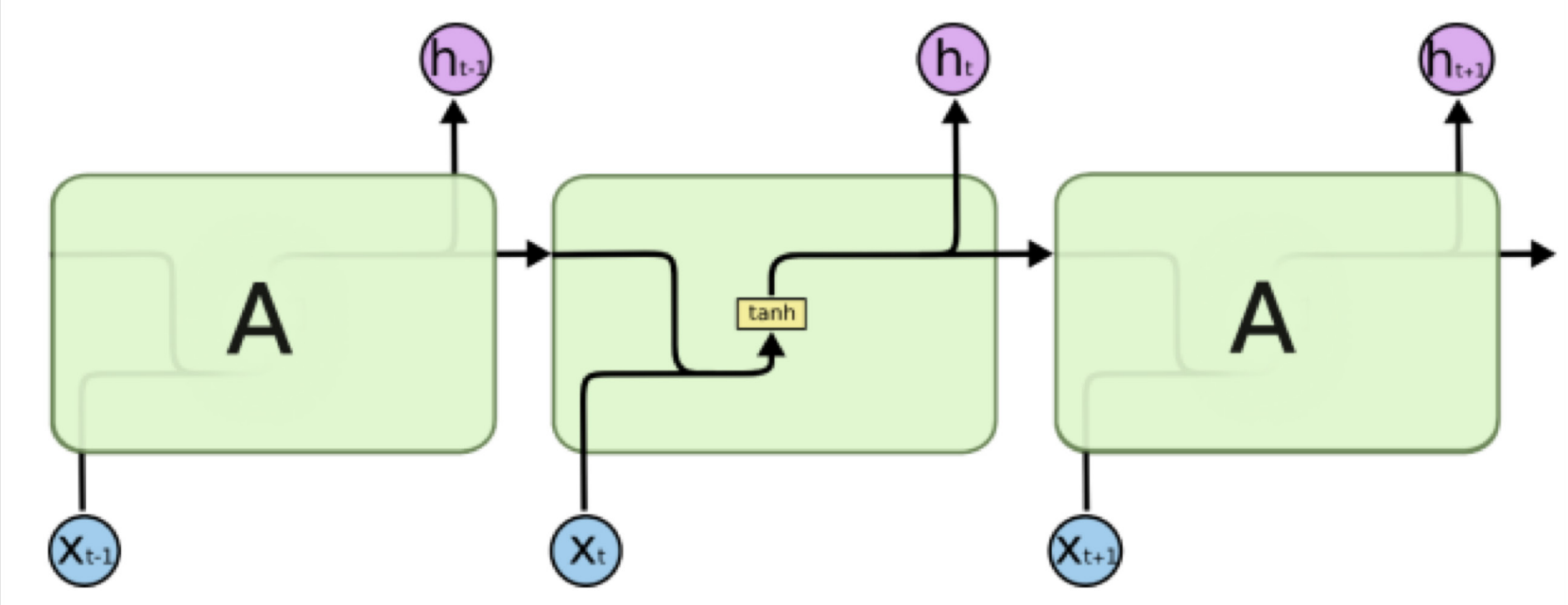
RNN

- In theory RNN retains information from the infinite past.
  - All past hidden state has influence to the future state.
- In practice RNN has little response to the early states.
  - Little memory over what seen before.
  - The hidden outputs blowup or shrink to zeros.
  - The “memory” also depends on activation functions.
  - ReLU and Sigmoid do not work well. Tanh is OK but still not “memorize” for too long.
- Vanishing gradient problem
  - Deeper layers do not have meaningful weights.

LSTM

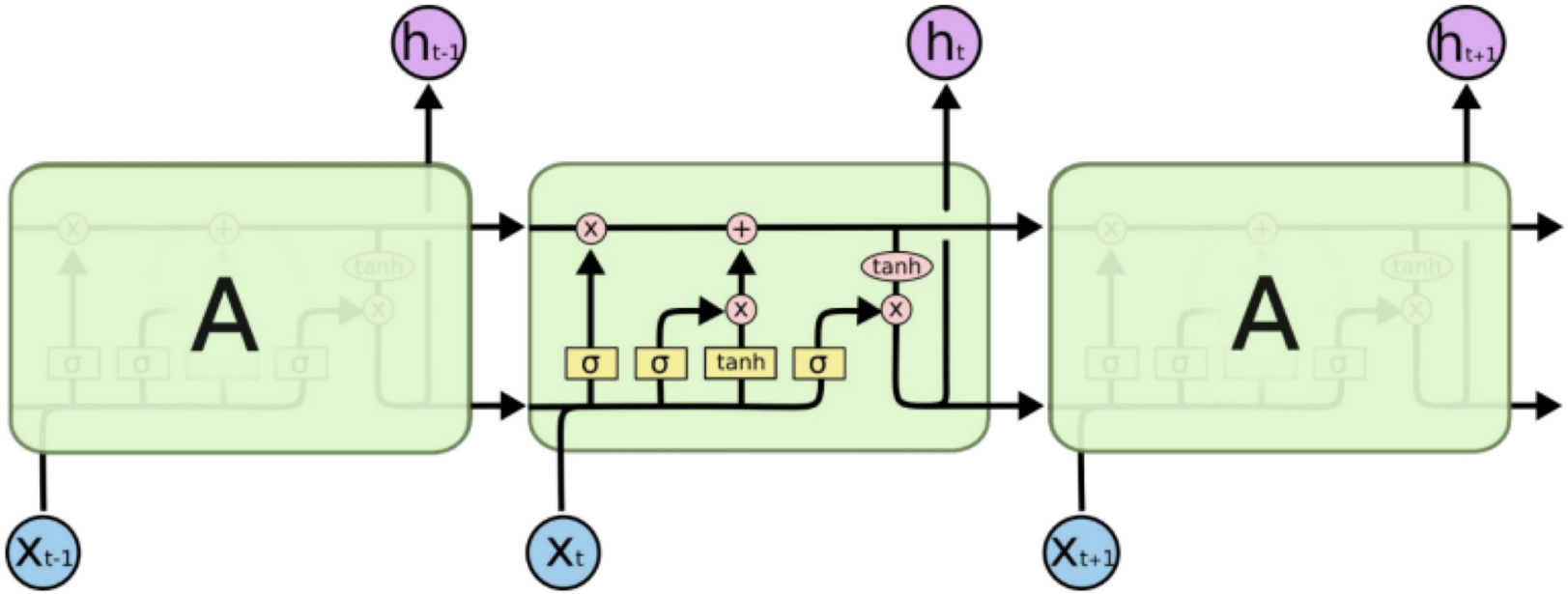
- Long Short-Term Memory
  - Explicitly latch the memory to prevent decay or blowup.
  - Following notes adapted from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

RNN VS LSTM

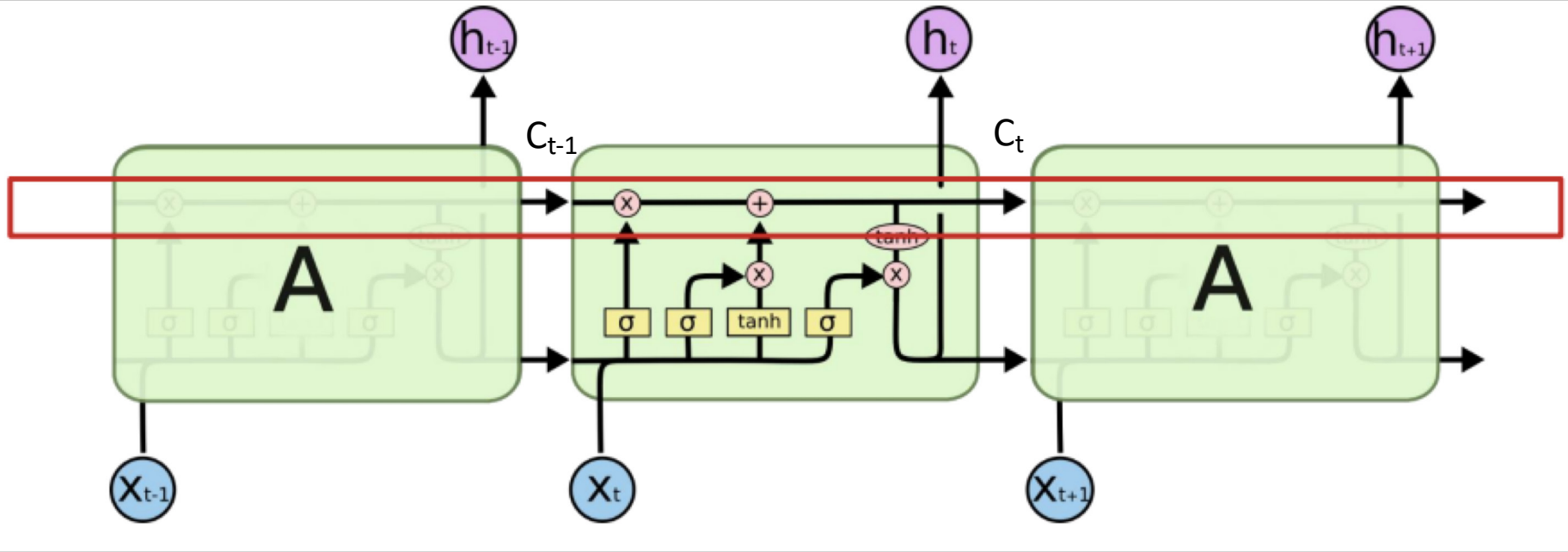


- Recurrent neurons receive past recurrent outputs and current input as inputs.
- Processed through a tanh() activation function
- Current recurrent output passed to next higher layer and next time step.

LSTM



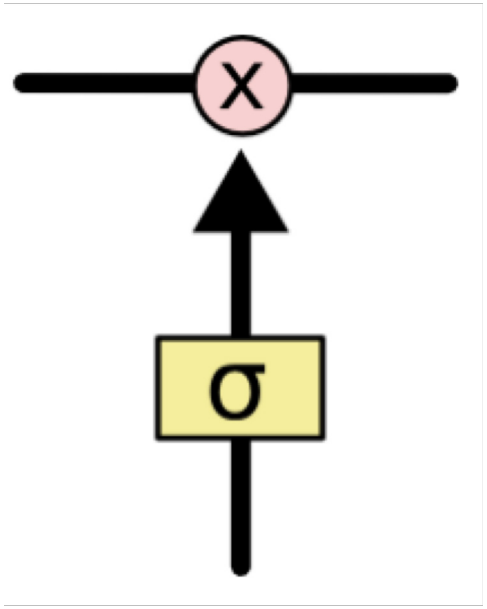
LSTM



Constant Error Carousel

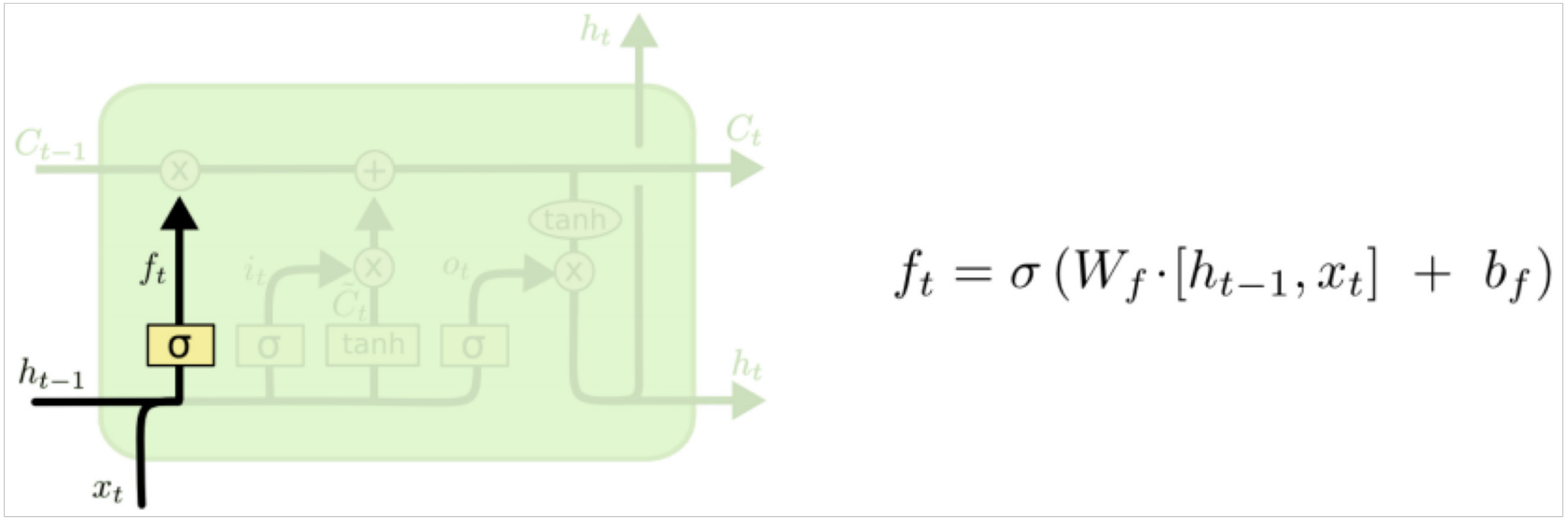
- Key of LSTM: a remembered cell state
- C_t is the linear history carried by the constant error carousel.
- Carries information through and only effected by a gate
  - Addition of history (gated).

LSTM - Gate



- A simple sigmoid function for output in range (0, 1).
- ⊗: element-wise multiplication.

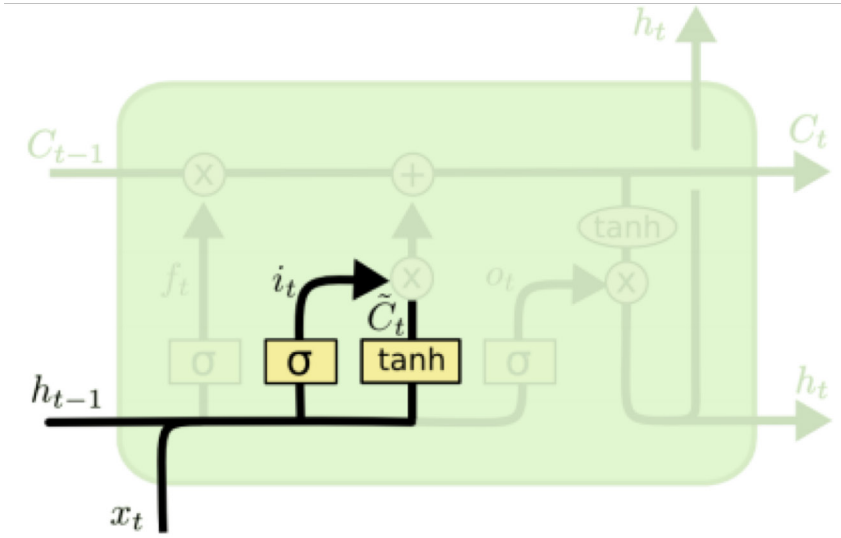
LSTM – Forget Gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

- The first gate determines whether to carry over the history or forget it
  - Called “forget” gate.
  - Actually, determine how much history to carry over.
  - The memory C and hidden state h are distinguished.

LSTM – Input Gate



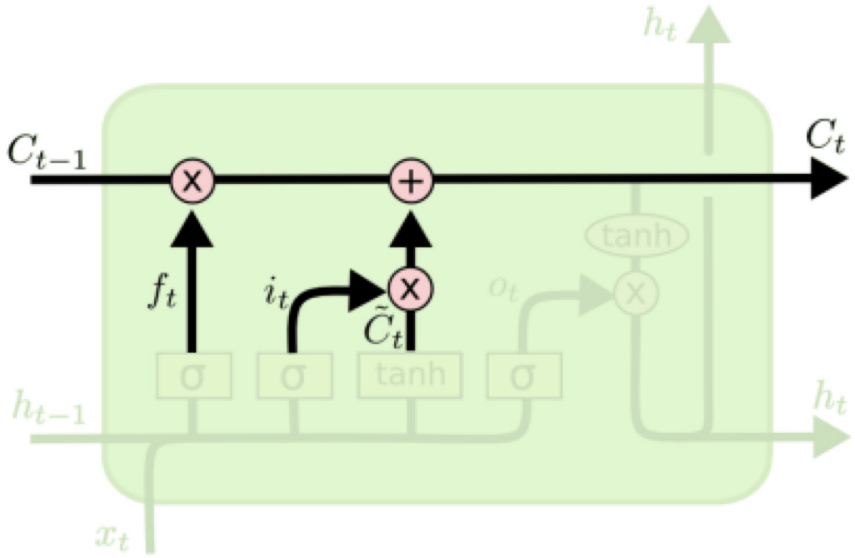
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

The second gate has two parts

- A *tanh* unit determines if there is something new or interesting in the input.
- A gate decides if it is worth remembering.

LSTM – Memory Cell Update

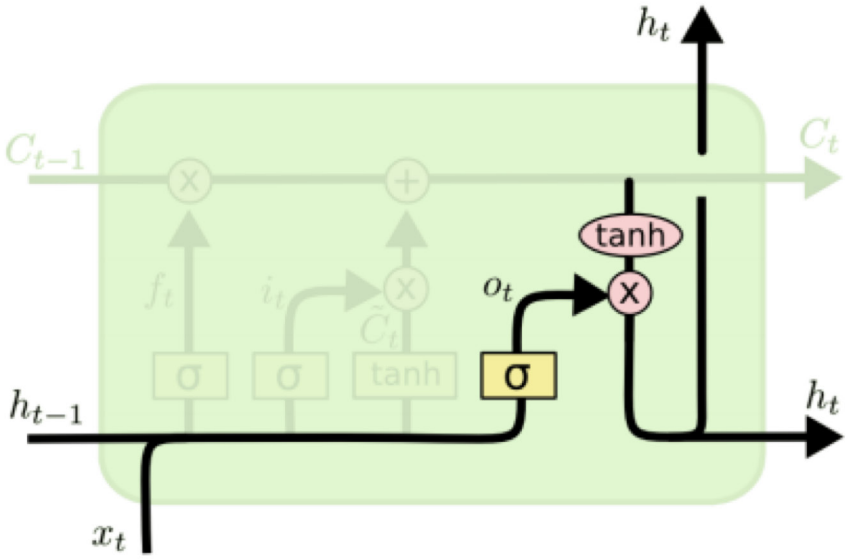


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Add the output of input gate to the current memory cell

- After the forget gate.
- $\bigoplus$: Element-wise addition.

LSTM – Output and Output Gate



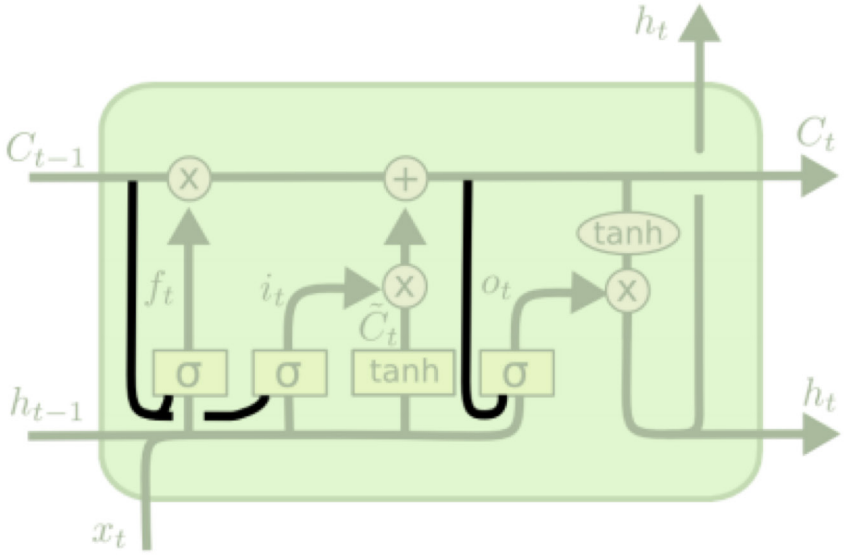
$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

The output of the memory cell

- Similar to input gate.
- A *tanh* unit over the memory to output in range (0, 1).
- Note the memory is carried through without *tanh*.
- A gate to decide if it is worth outputting.

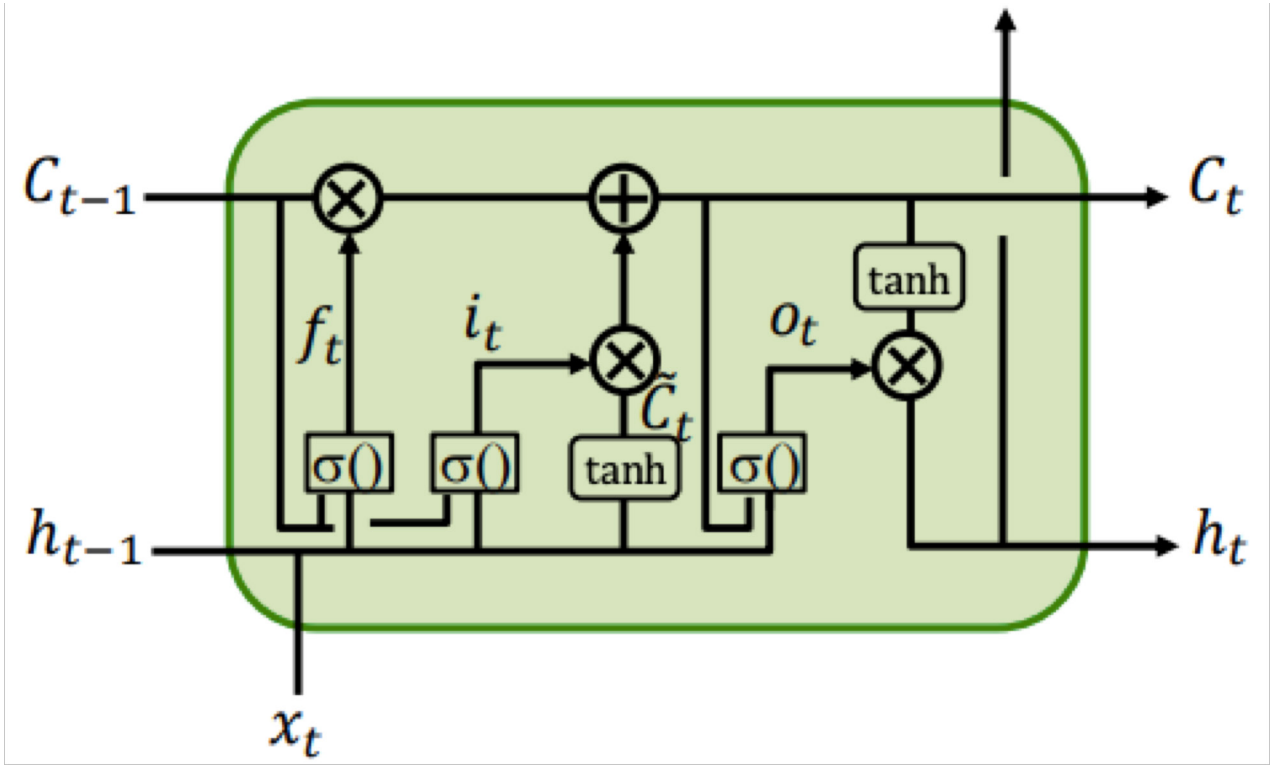
LSTM – the "Peephole" Connection



$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$

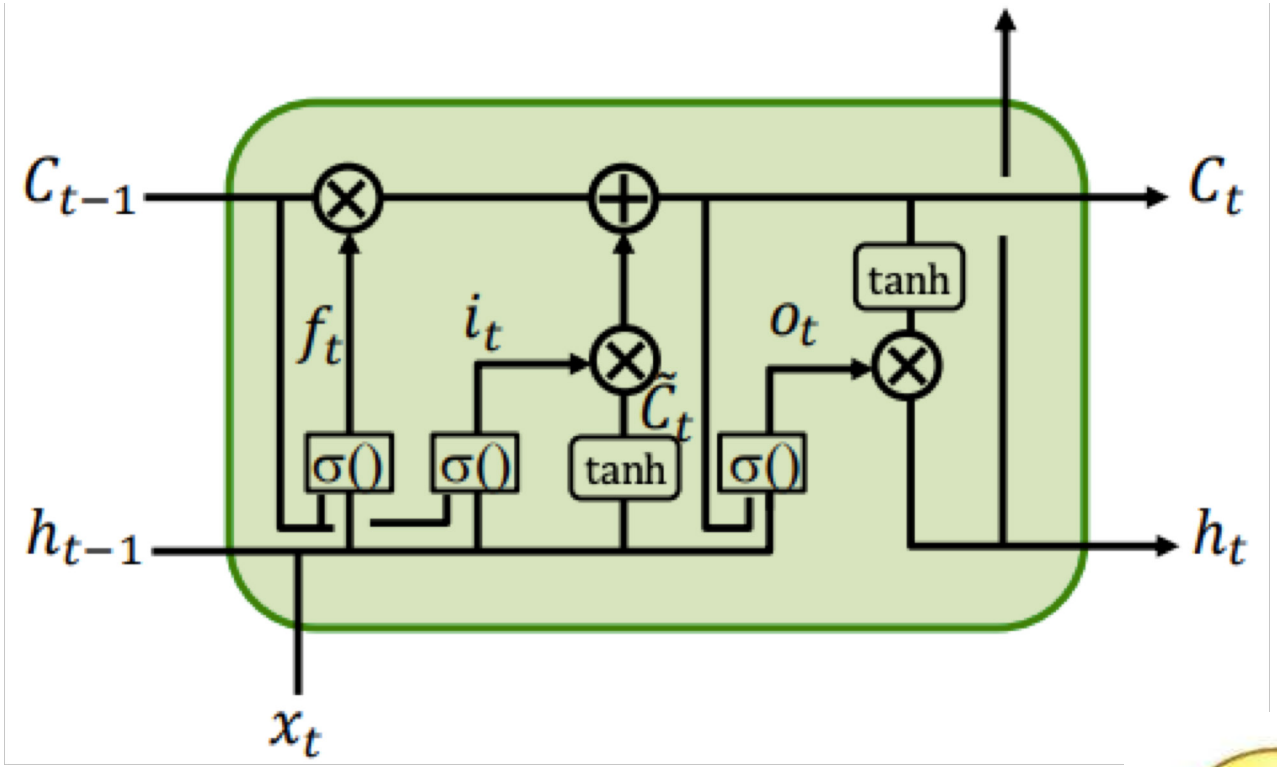
Let the memory cell directly influence the gates!

The Complete LSTM Unit



Input, output, forget gates with peephole connection

Back propagation for Training All the Weights



Lots of equations… You do not want to know.

Applications of LSTM

- Nowadays, considered as the default models for sequence labeling tasks.
- Does not suffer from Vanishing Gradient problem.
- Very powerful, especially in deeper networks.
- Very useful when you have a lot of data.

Applications of LSTM

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

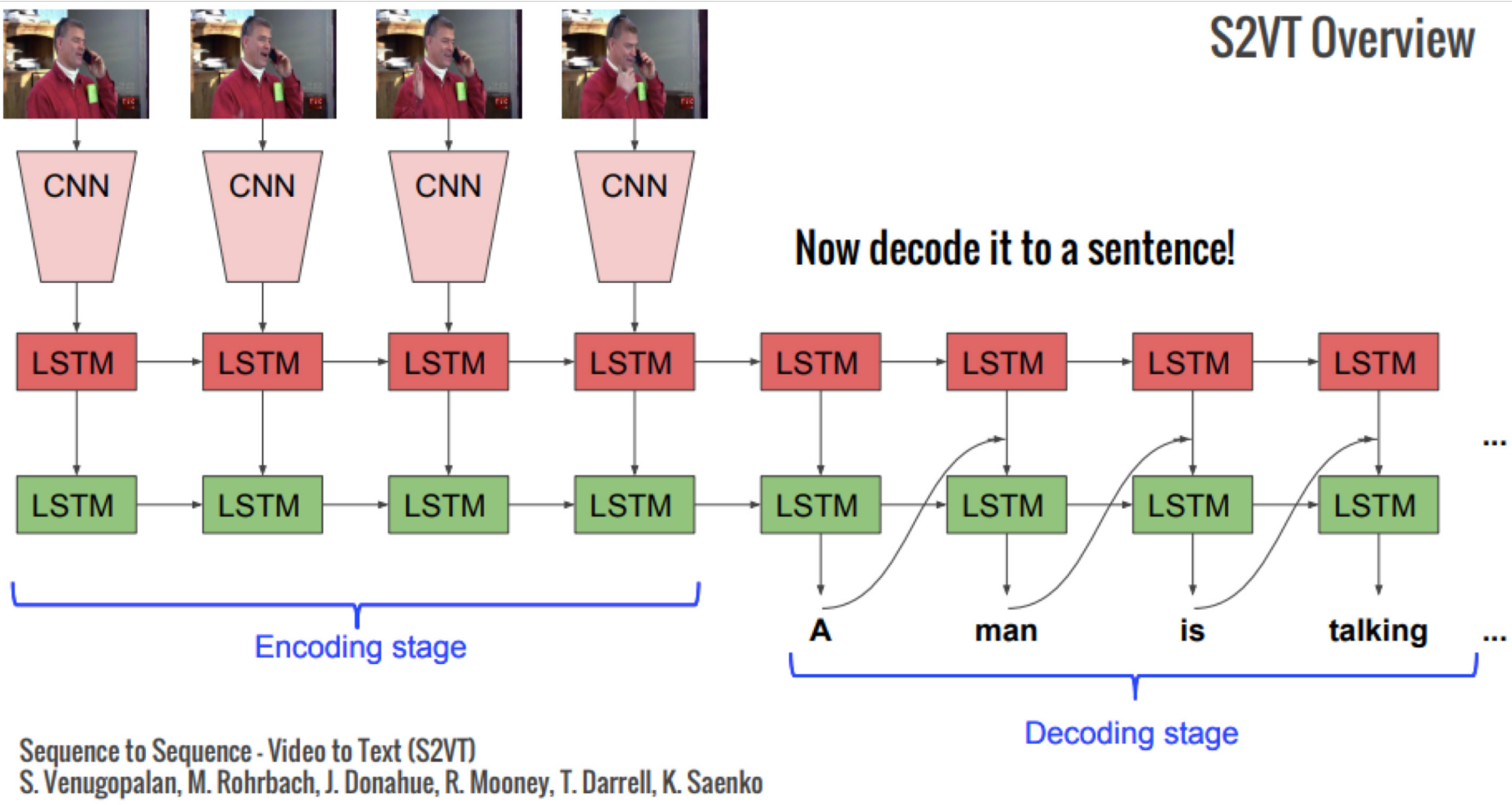
Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Sequence to Sequence Learning by Sutskever et al. 2014

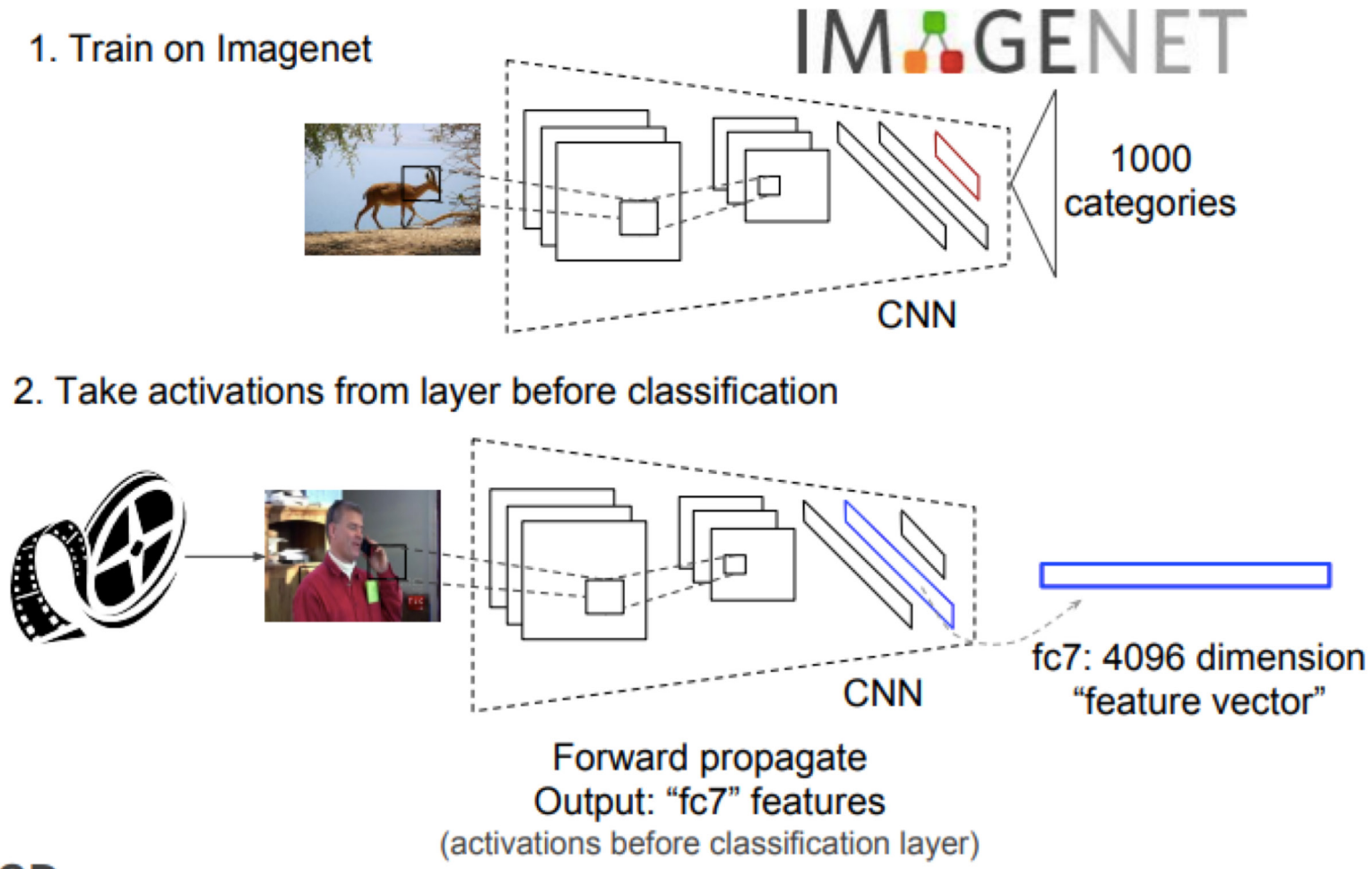
Applications of LSTM

- Sequence to sequence: video to text

Video to Text



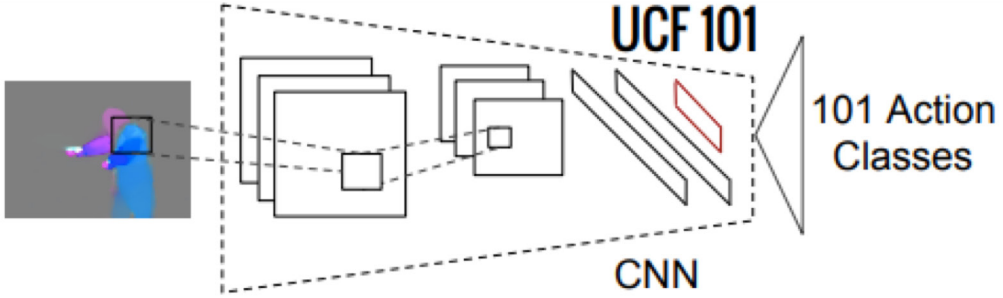
Video to Text



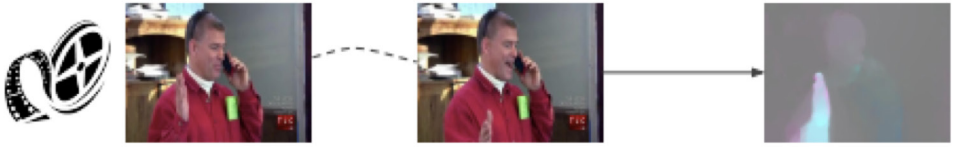
Frames: RGB

Video to Text

1. Train CNN on Activity classes

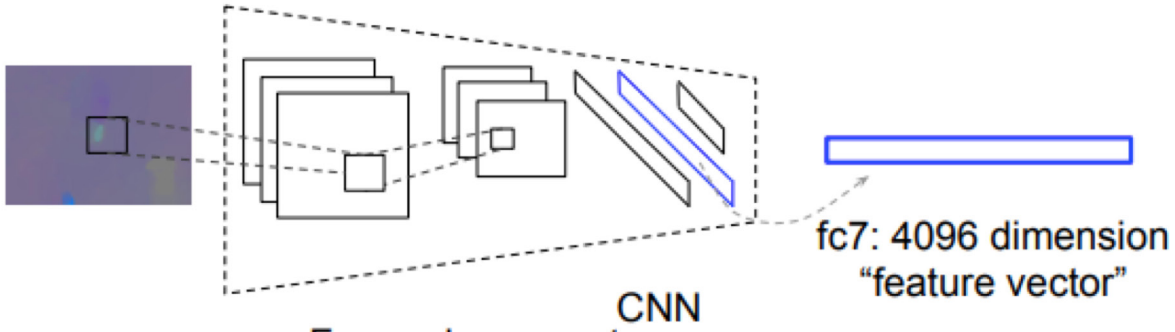


2. Use optical flow to extract flow images.



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

3. Take activations from layer before classification

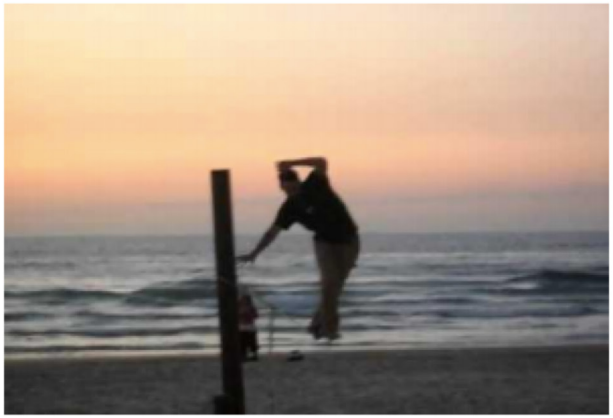


Frames: Flow

Video to Text

Dataset: Youtube

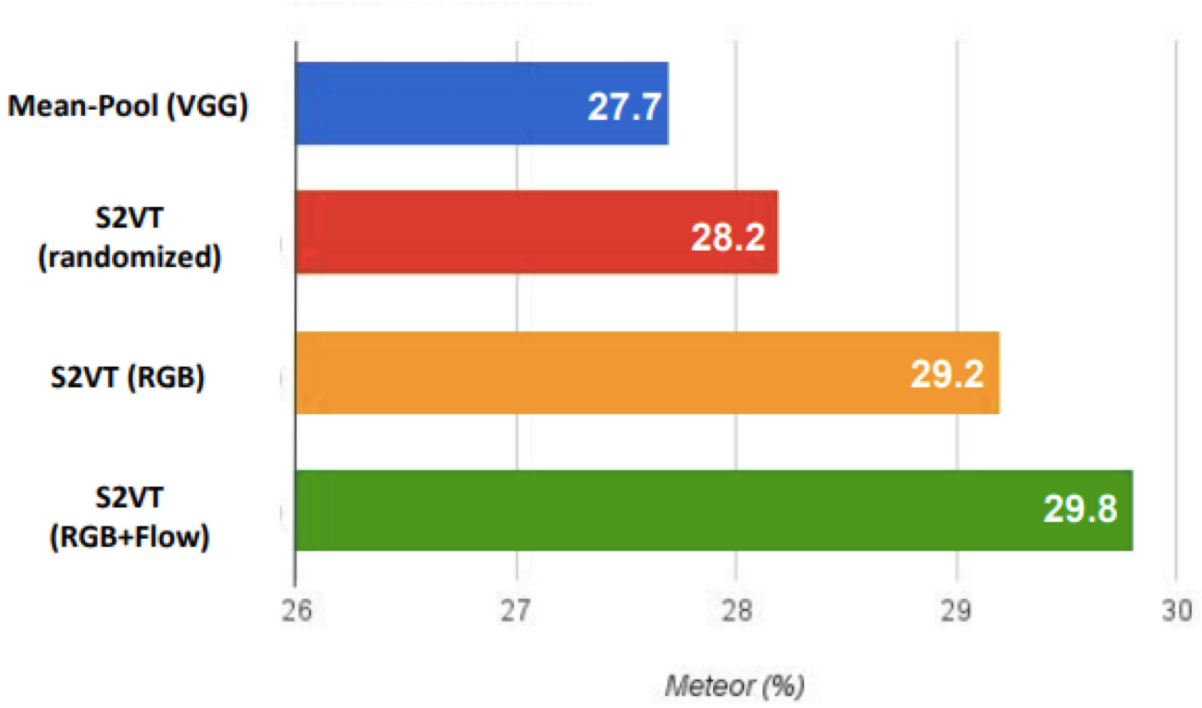
- ~2000 clips
- Avg. length: 11s per clip
- **~40 sentence per clip**
- ~81,000 sentences



- 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.

Video to Text

Results (Youtube)



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

Video to Text

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.

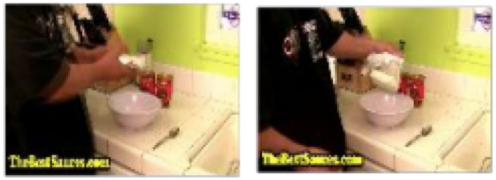
S2VT: A man is shooting a gun at a target.

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.

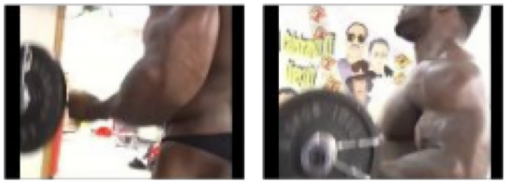


S2VT: A man is spreading butter on a tortilla.

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.

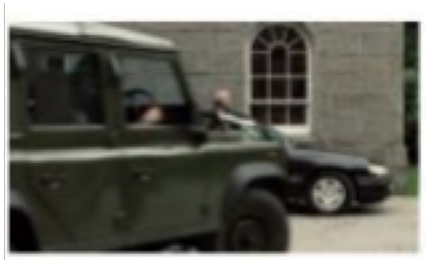
S2VT: A black clip to walking through a path.

Video to Text

Evaluation on movie corpus

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

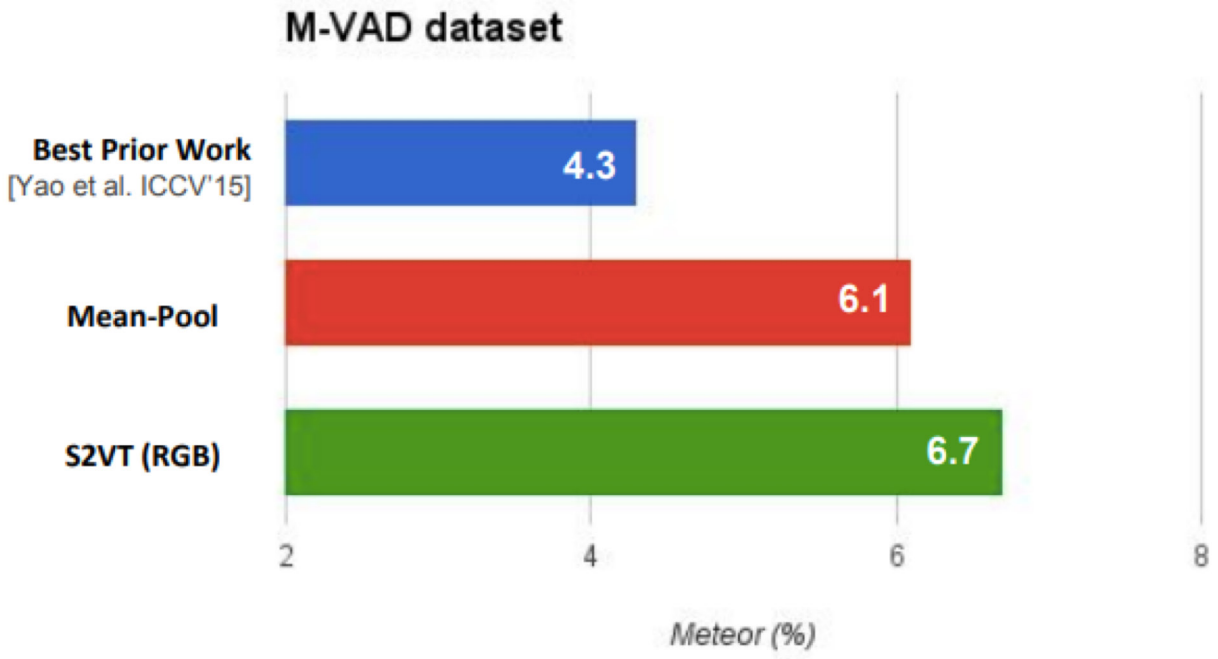


The Land Rover pulls away.

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

Video to Text

Results (M-VAD Movie Corpus)



Video to Text

https://youtu.be/pER0mjzSYaM