

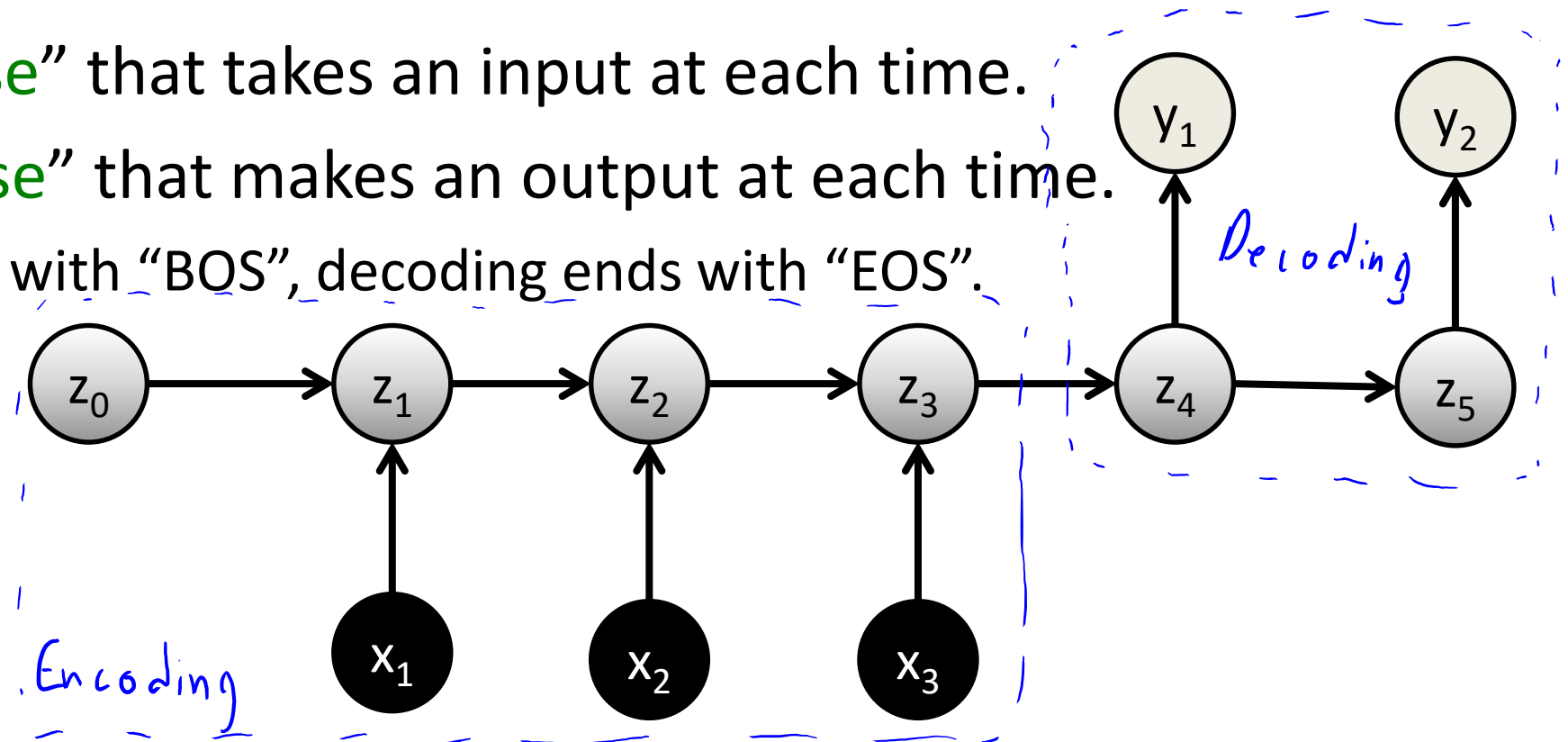
CPSC 540: Machine Learning

Long Short Term Memory

Winter 2020

Previously: Sequence-to-Sequence

- Sequence-to-sequence:
 - Recurrent neural network for sequences of different lengths.
- “Encoding phase” that takes an input at each time.
- “Decoding phase” that makes an output at each time.
 - Encoding ends with “BOS”, decoding ends with “EOS”.



Variations on Recurrent Neural Networks

- **Bi-directional RNNs**: feedforward from past and future.
- **Recursive neural networks**: consider sequences through non-chain data.
- **Graphical models** to explicitly encourage output dependencies:
 - <https://arxiv.org/abs/1711.04956>

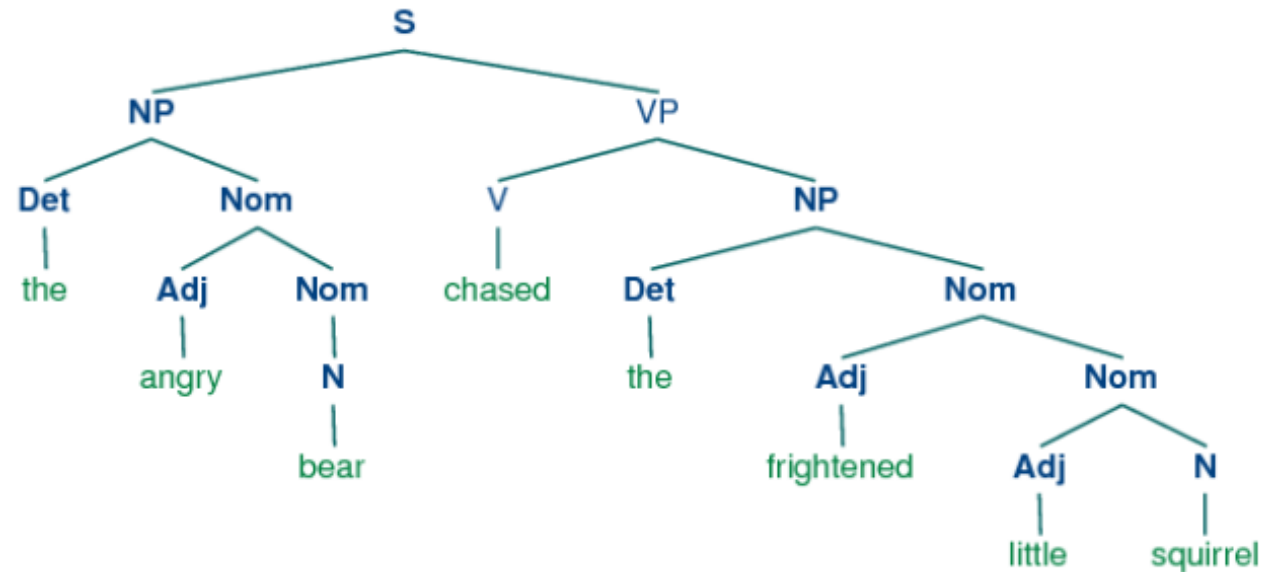
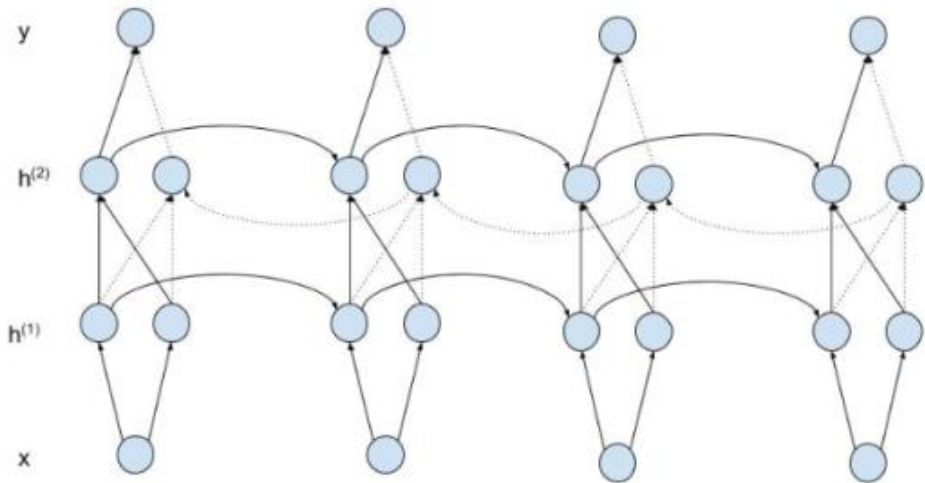


Figure 2: A deep bi-directional RNN with 2 stacked layers

Long Short Term Memory (LSTM)

- Long short term memory (LSTM) models are special case of RNNs:
 - Designed so that model can “remember things for a long time”.
- LSTMs have been the analogy of convolutions for RNNs:
 - “The trick that makes them work in applications.”
- LSTMs are getting impressive performance in various settings:
 - Cursive handwriting recognition.
 - <https://www.youtube.com/watch?v=mLxsbWAYlpw>
 - Speech recognition.
 - Machine translation.
 - Image and video captioning.

LSTMs for Image Captioning

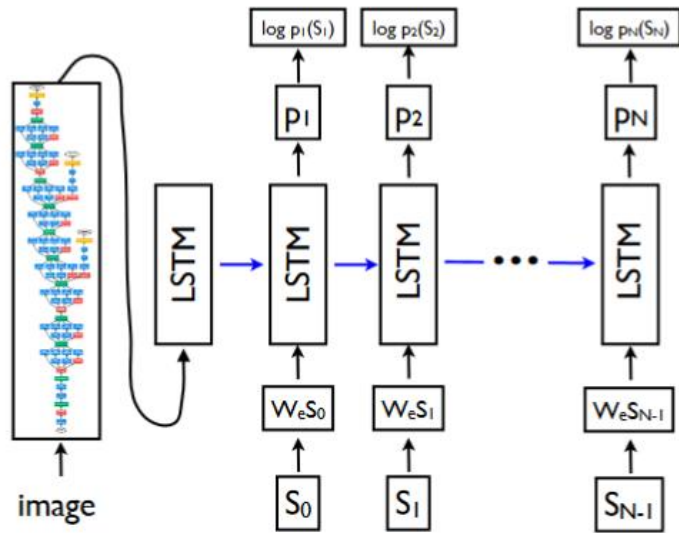
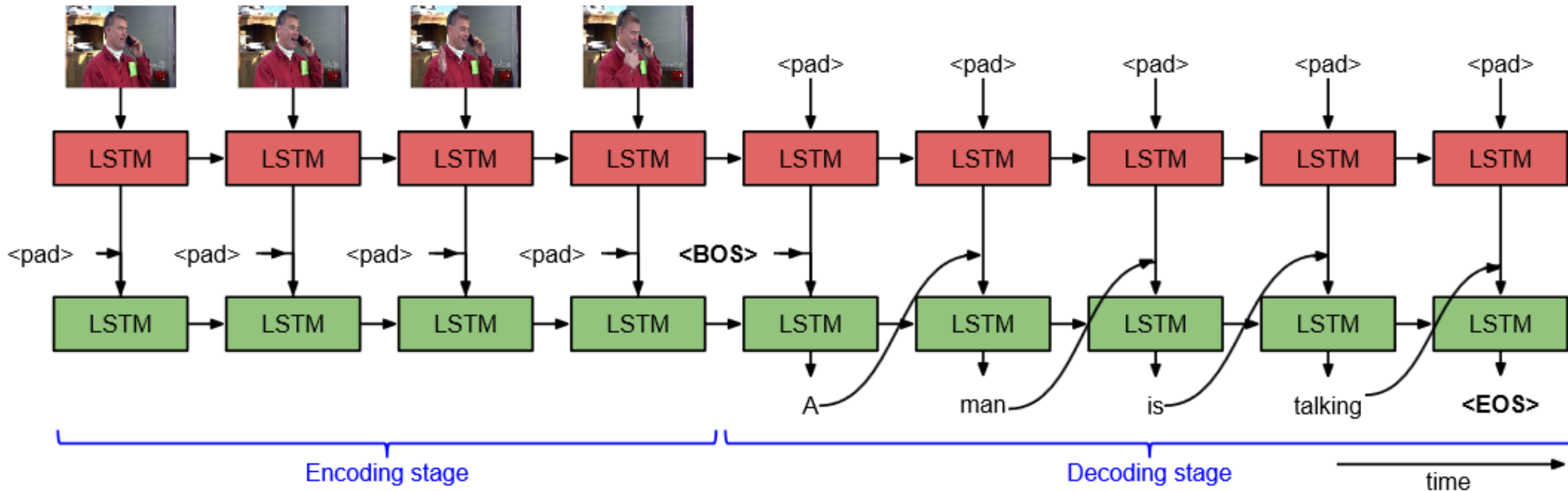


Figure 3. LSTM model combined with a CNN image embedder (as defined in [12]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.



Figure 5. A selection of evaluation results, grouped by human rating.

LSTMs for Video Captioning



LSTMs for Video Captioning

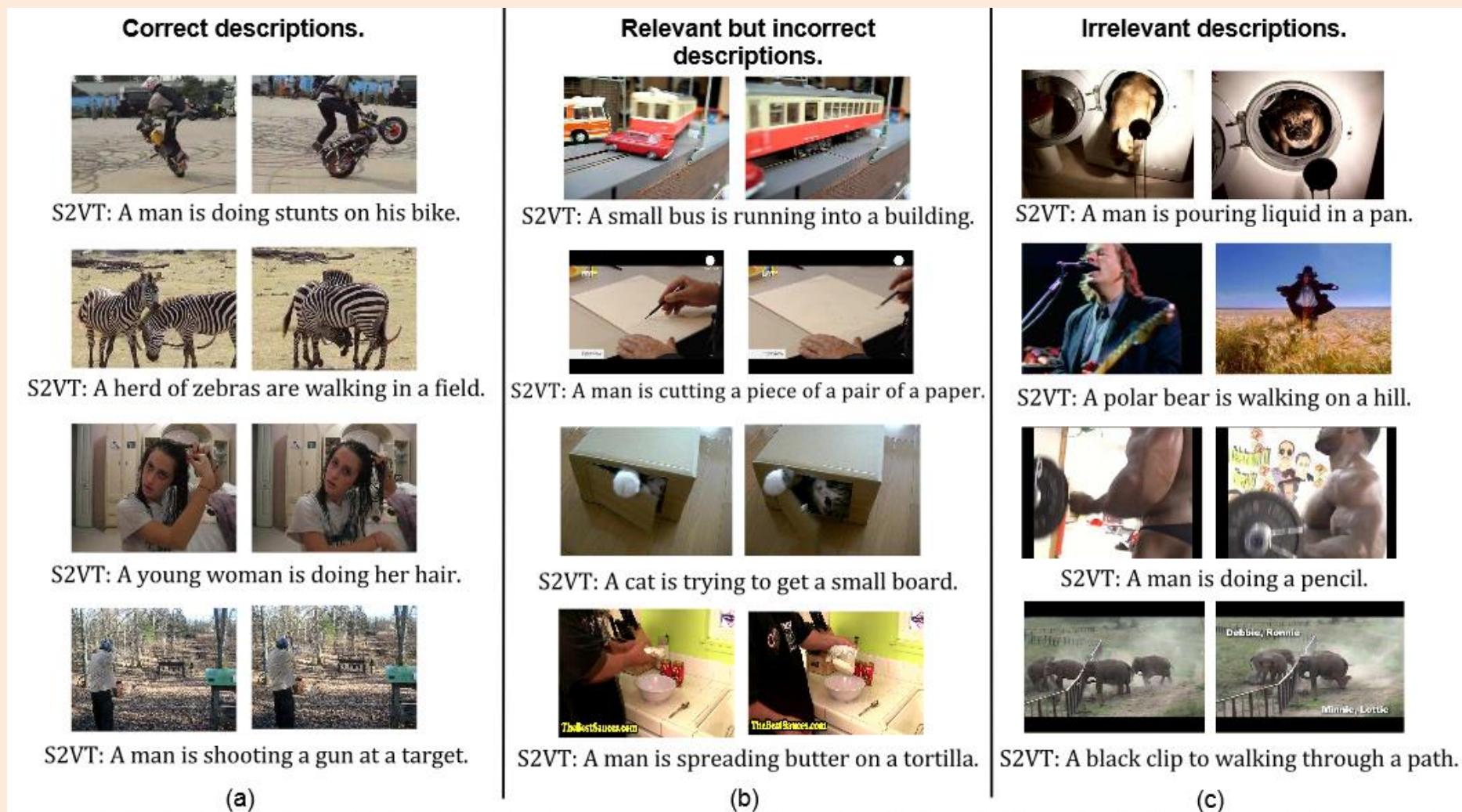
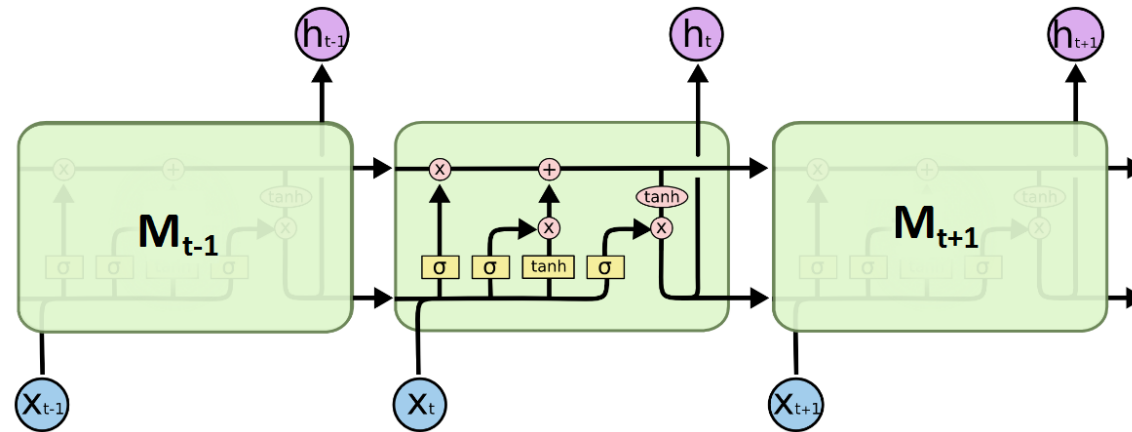


Figure 3. Qualitative results on MSVD YouTube dataset from our S2VT model (RGB on VGG net). (a) Correct descriptions involving different objects and actions for several videos. (b) Relevant but incorrect descriptions. (c) Descriptions that are irrelevant to the event in the video.

Long Short Term Memory

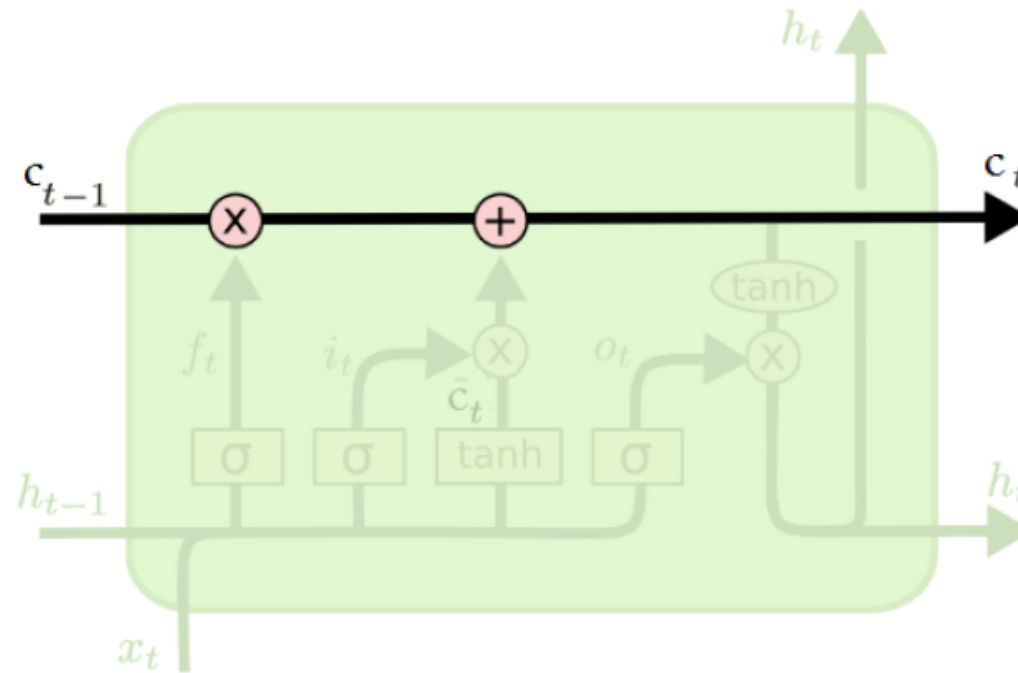
- In addition to usual hidden values 'z', **LSTMs** have **memory cells** 'c':
 - Purpose of memory cells is to **remember things for a long time**.



- “Read/write/forget”:
 - Information **gets into the cell** when its **input gate** is on.
 - Information is **read from the cell** when the **output gate** is on.
 - Information is **thrown away** when the **forget gate** is off.
- “**Gate functions**”: approximate binary operations (like “write or not”).
 - Replace operation by a **sigmoid functions** to make it **continuous/differentiable**.

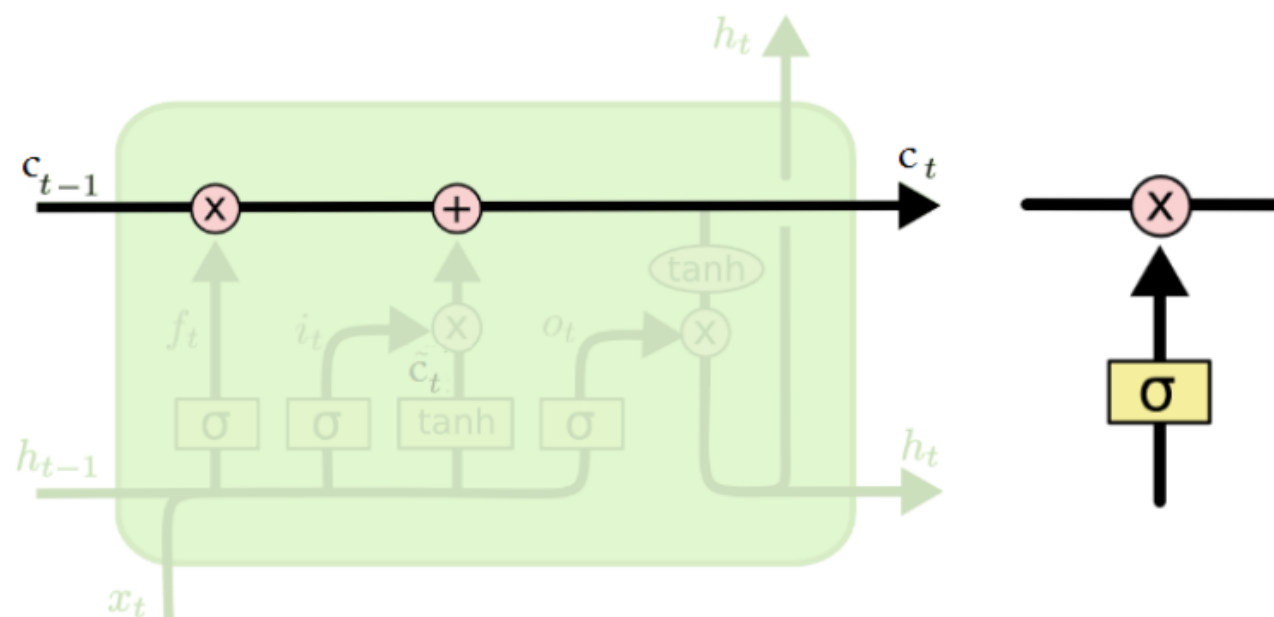
The Core Idea Behind LSTMs: Cell State (Memory Cell)

- Information can flow along the **memory cell unchanged**.
- Information can be **removed** or **written** to the **memory cell**, regulated by gates.



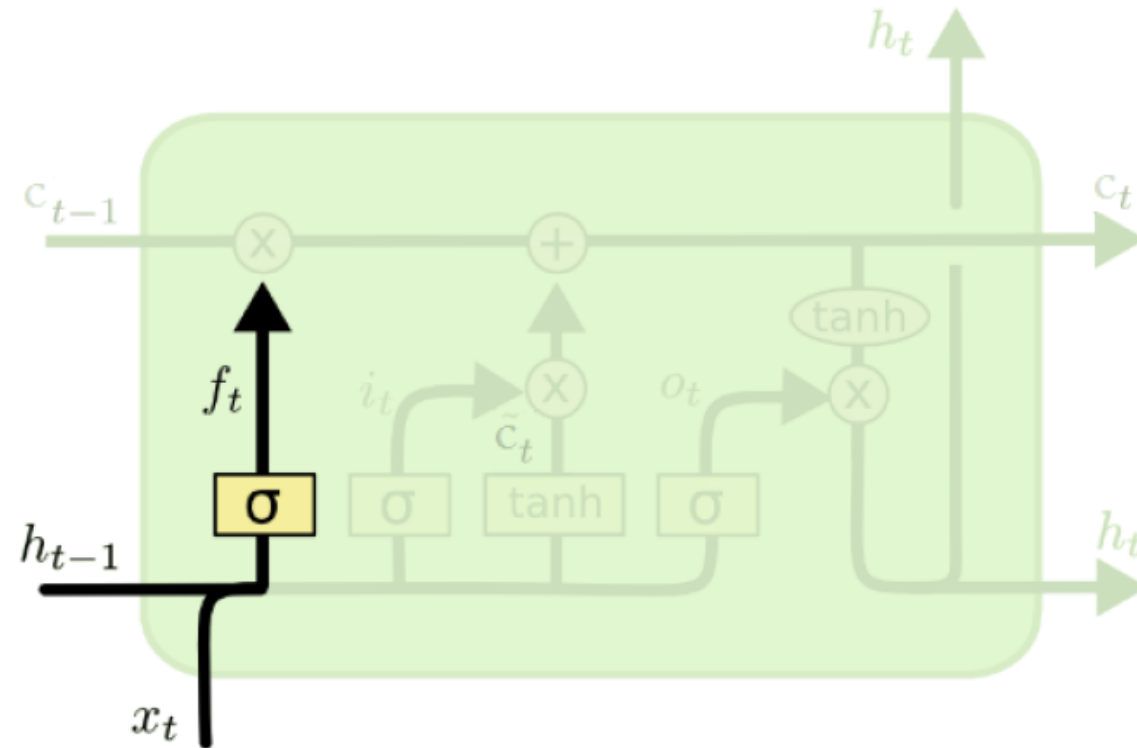
Gates

- **Gates** are a way to optionally let information through.
 - A **sigmoid layer** outputs number between 0 and 1, **deciding** how much of each component should be let through.
 - A pointwise multiplication operation applies the decision.



Forget Gate

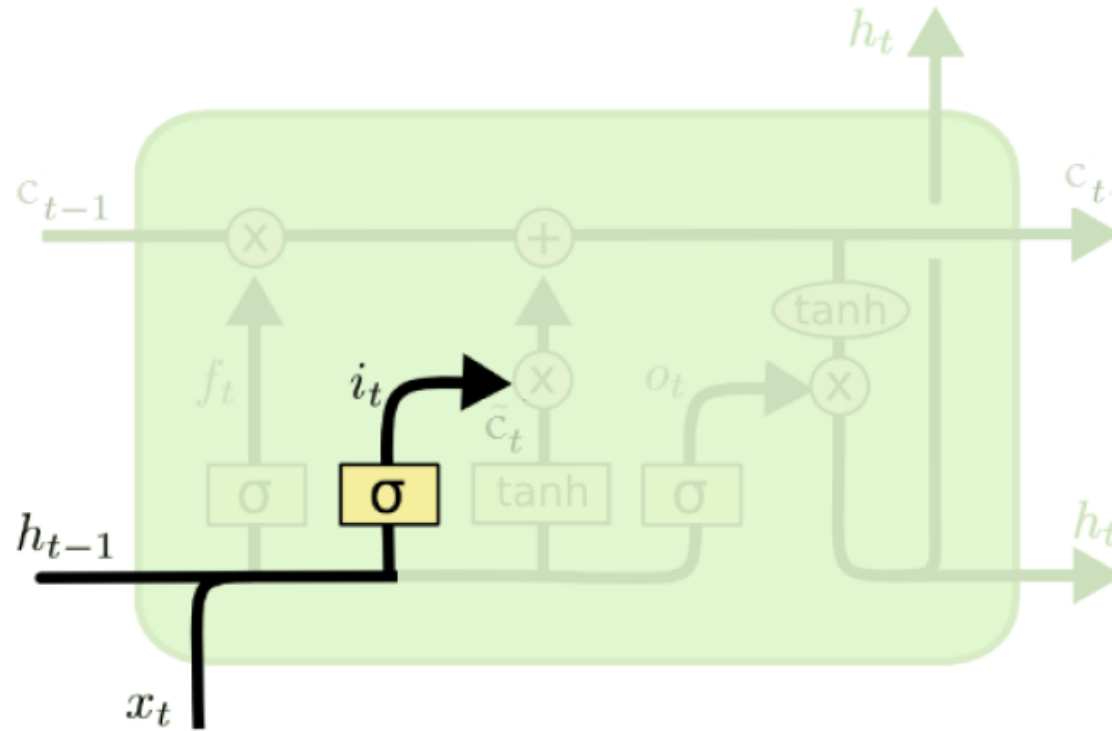
- A **sigmoid** layer, **forget gate**, **decides** which values of the **memory cell** to **reset**.



$$\mathbf{f}_t = \sigma(W_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f)$$

Input Gate

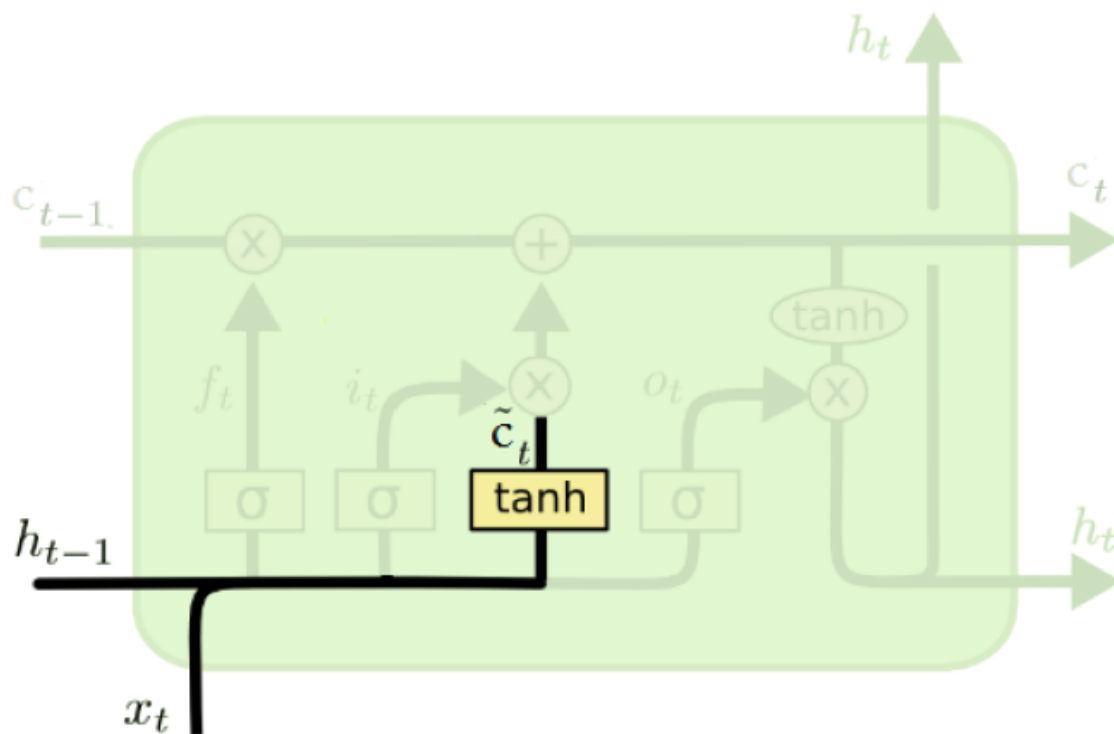
- A **sigmoid** layer, **input gate**, **decides** which values of the **memory cell** to **write** to.



$$\mathbf{i}_t = \sigma(W_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i)$$

Vector of New Candidate Values

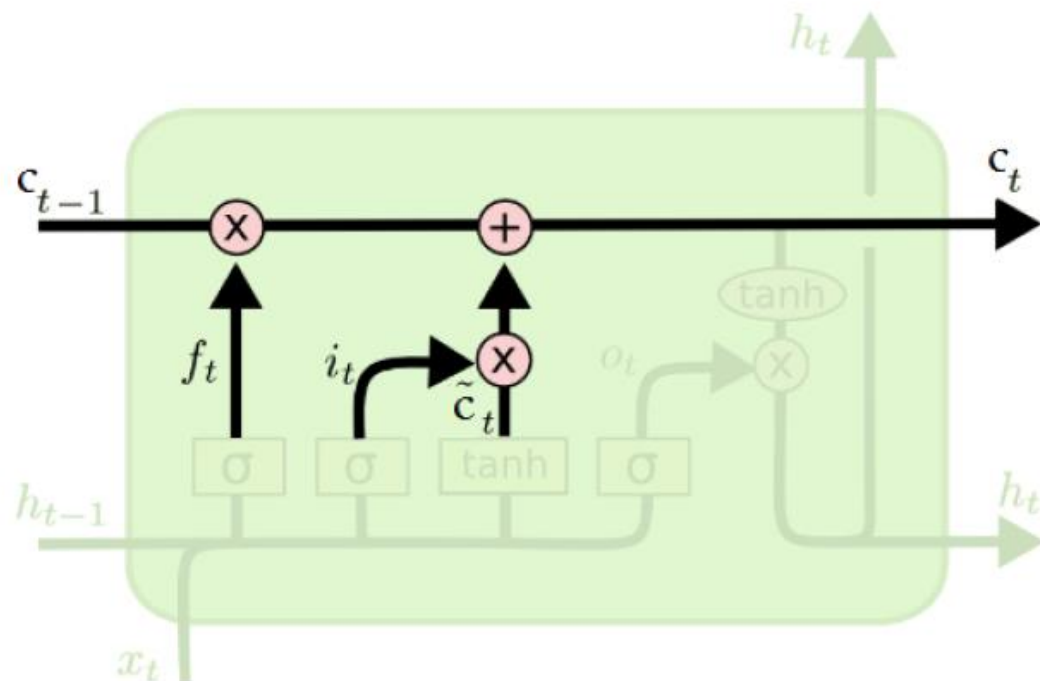
- A **Tanh** layer creates a **vector of new candidate values** \tilde{c}_t to **write** to the **memory cell**.



$$\tilde{c}_t = \text{Tanh}(W_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c)$$

Memory Cell Update

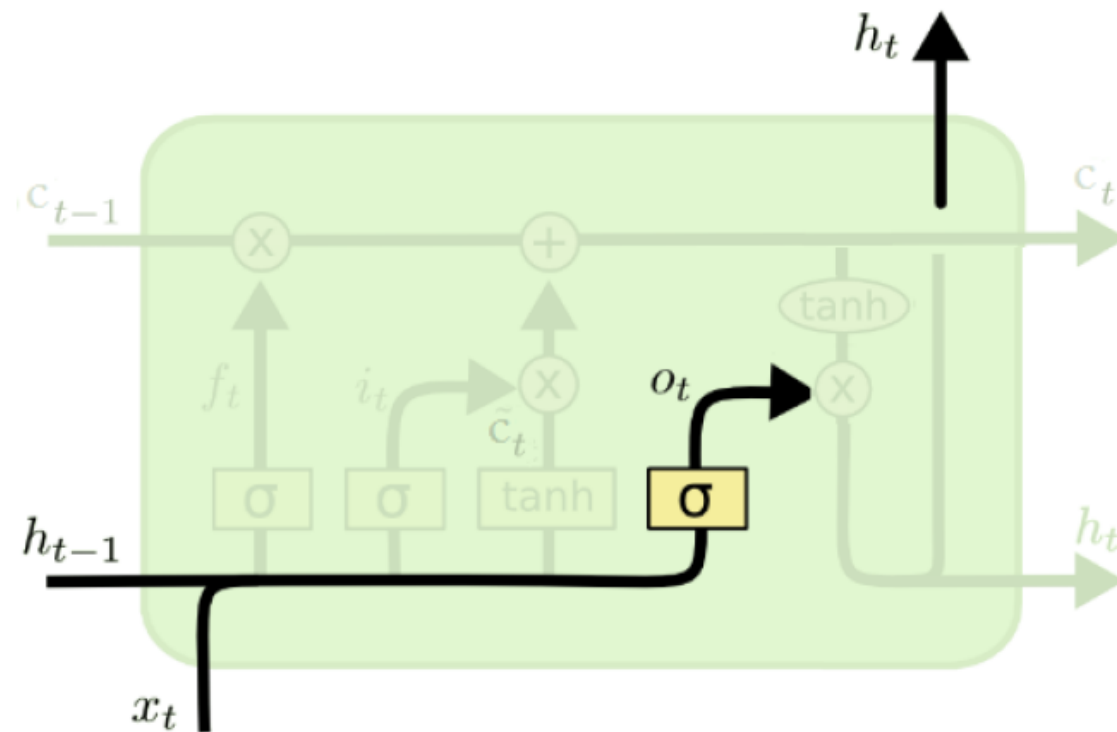
- The previous steps decided which values of the **memory cell** to **reset** and **overwrite**.
- Now the LSTM **applies the decisions** to the **memory cell**.



$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

Output Gate

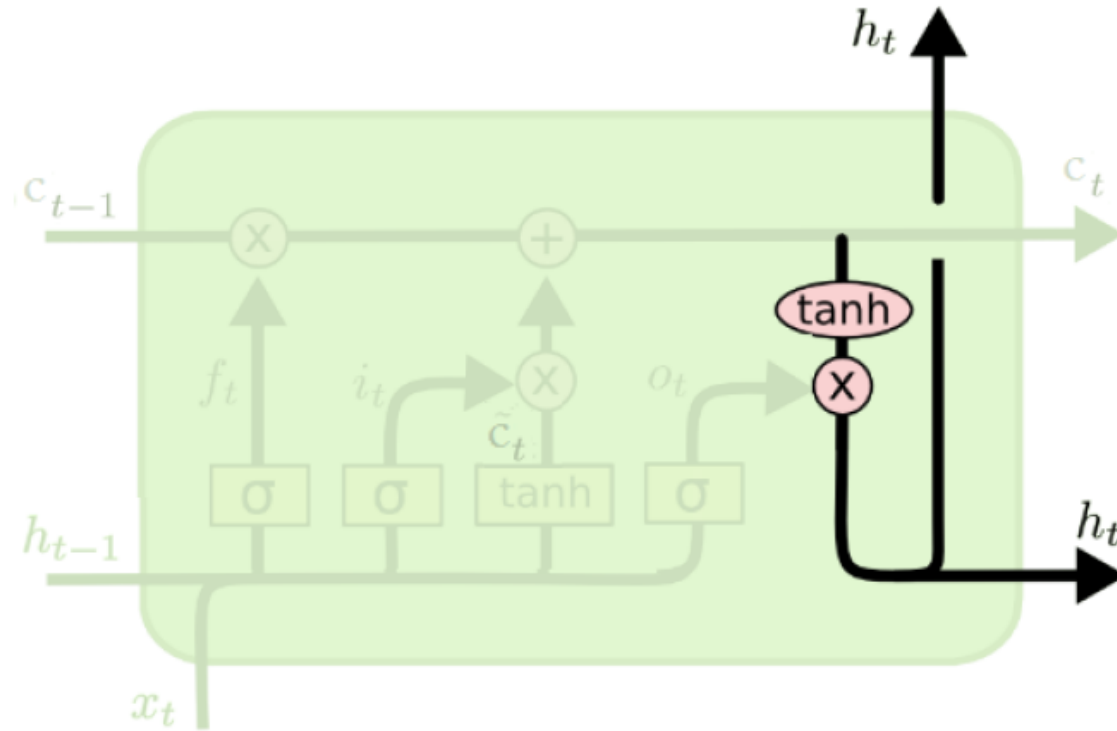
- A **sigmoid** layer, **output gate**, **decides** which values of the **memory cell** to **output**.



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

Output Update

- The **memory cell** goes through **Tanh** and is multiplied by the **output gate**.



$$h_t = o_t * \text{Tanh}(c_t)$$

LSTM Structure

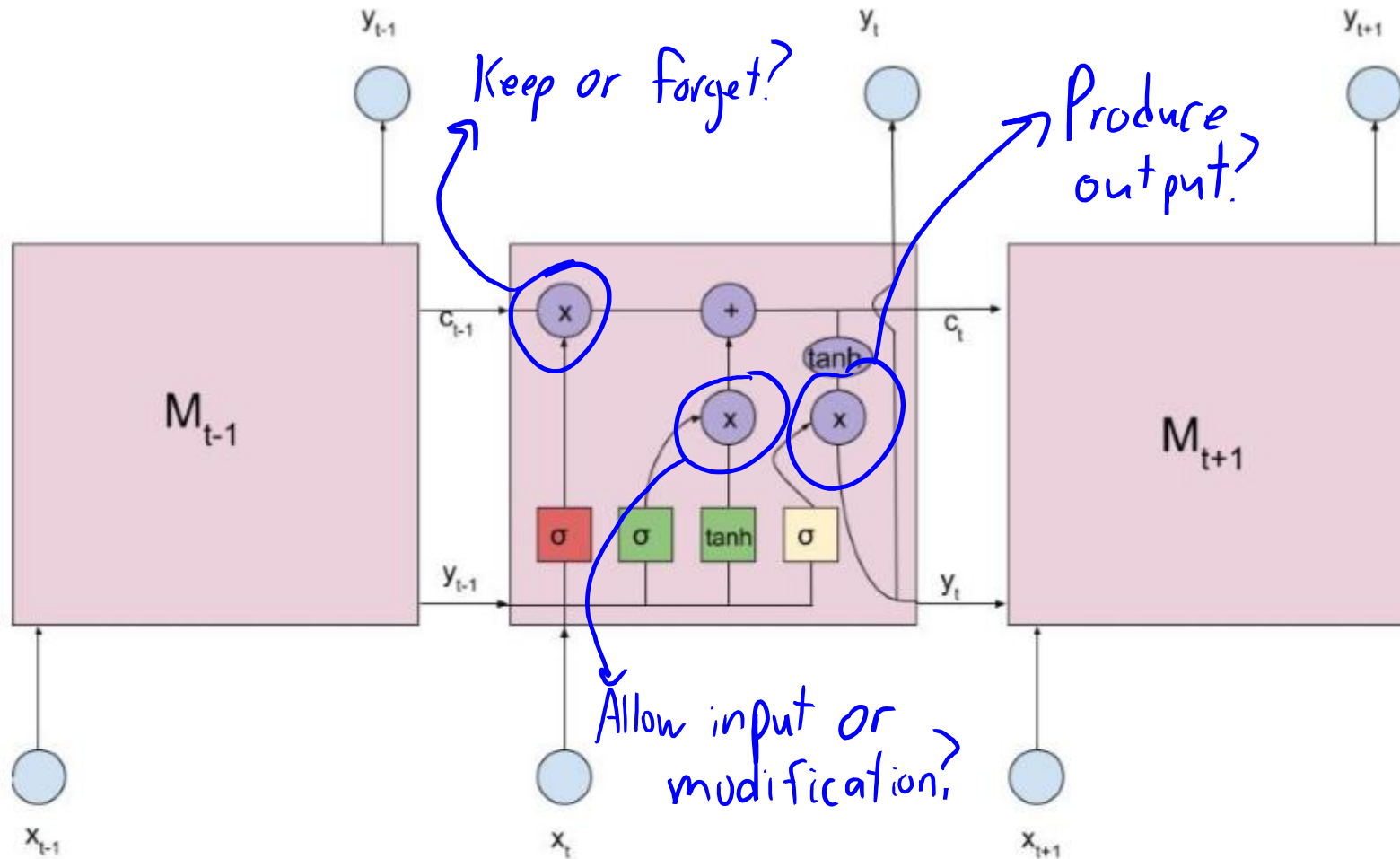


Figure 6: A close look at LSTM structure

Vanilla RNN vs. LSTM

Vanilla Recurrent Neural Network (RNN) has a recurrence of the form

$$h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

Previous layer, same time.
Same layer, previous time.

memory vector c_t^l . At each time step the LSTM can choose to read from, write to, or reset the cell using explicit gating mechanisms. The precise form of the update is as follows:

$$\begin{matrix} \text{Input} & \rightarrow & i \\ \text{Forget} & \rightarrow & f \\ \text{Output} & \rightarrow & o \\ \text{Candidate} & \rightarrow & g \end{matrix} \quad \begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

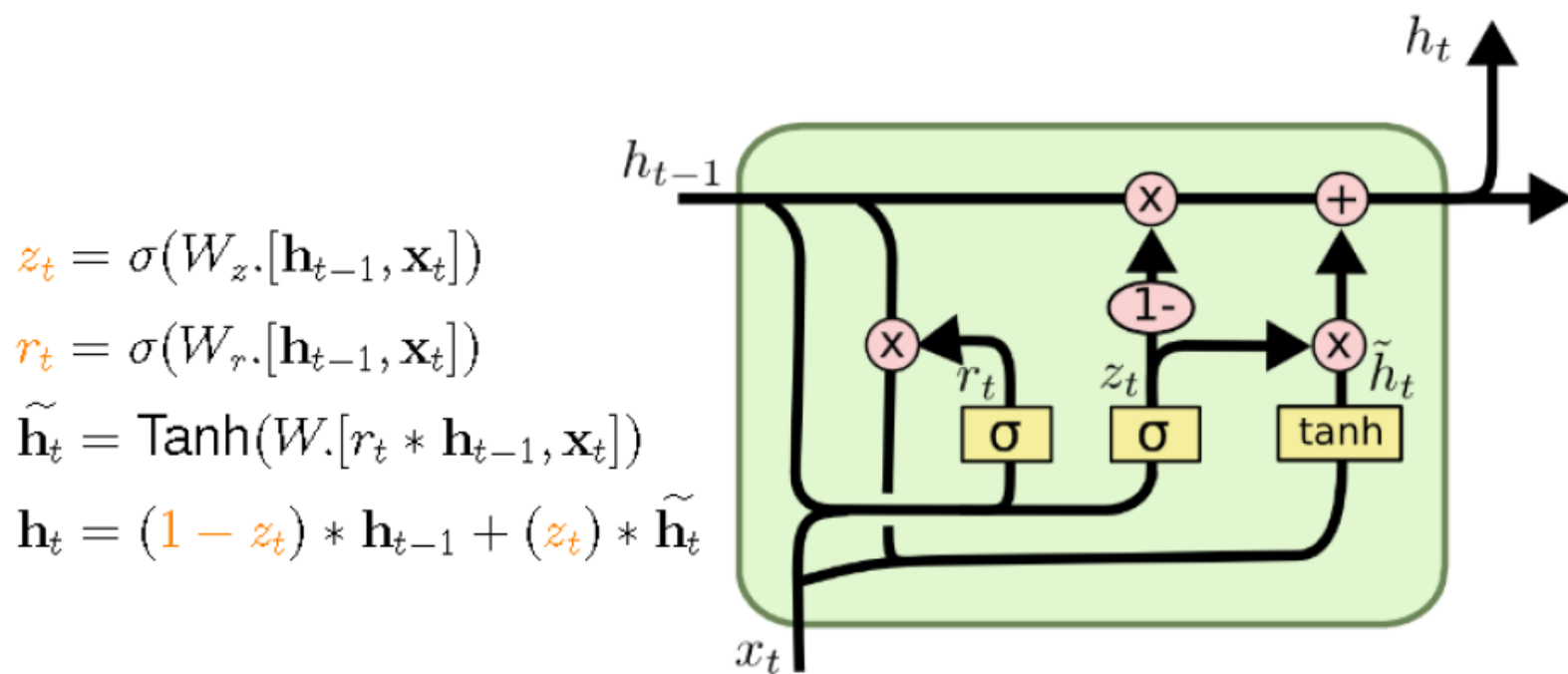
Here, the sigmoid function sigm and \tanh are applied element-wise, and W^l is a $[4n \times 2n]$ matrix.

Cell $\rightarrow c_t^l = f \odot c_{t-1}^l + i \odot g$ Forget times old memory. Input times candidate
Output $\rightarrow h_t^l = o \odot \tanh(c_t^l)$ Output times current memory

- Notice that if “f=1” and “i=0”, then **memory is unchanged**.
 - Memory might only change for specific inputs.
- More recent: **gated recurrent unit (GRU)**:
 - Similar performance but a bit simpler.

Variants on LSTM

- Gated Recurrent Unit (GRU) [Cho et al., 2014]:
 - Combine the **forget** and **input** gates into a single **update** gate.
 - **Merge the memory cell and the hidden state.**
 - ...



Residual Connections

- As in ResNets, modern RNNs are including **residual connections**:

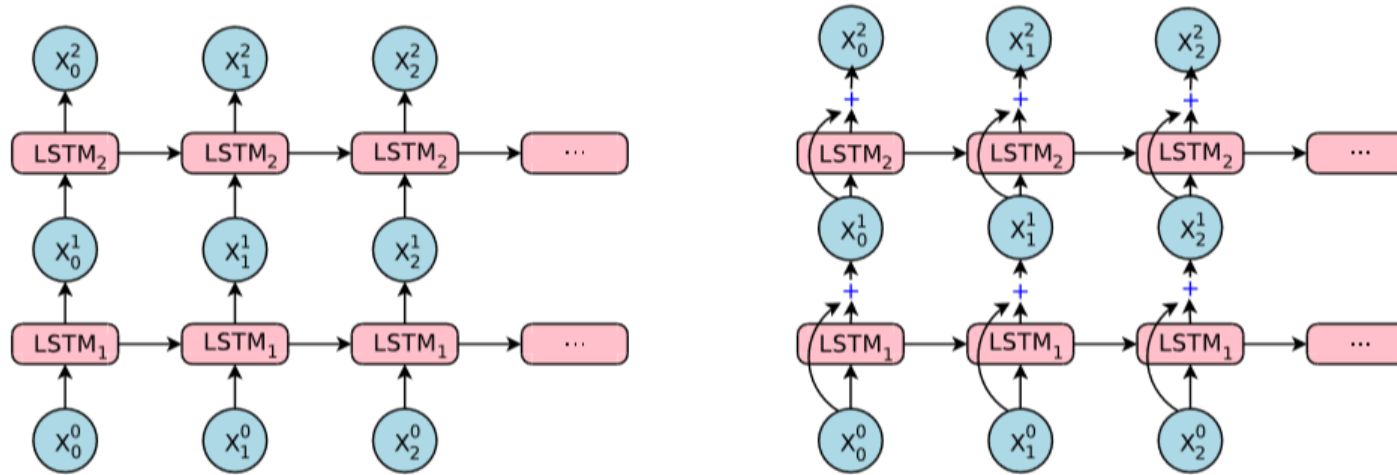
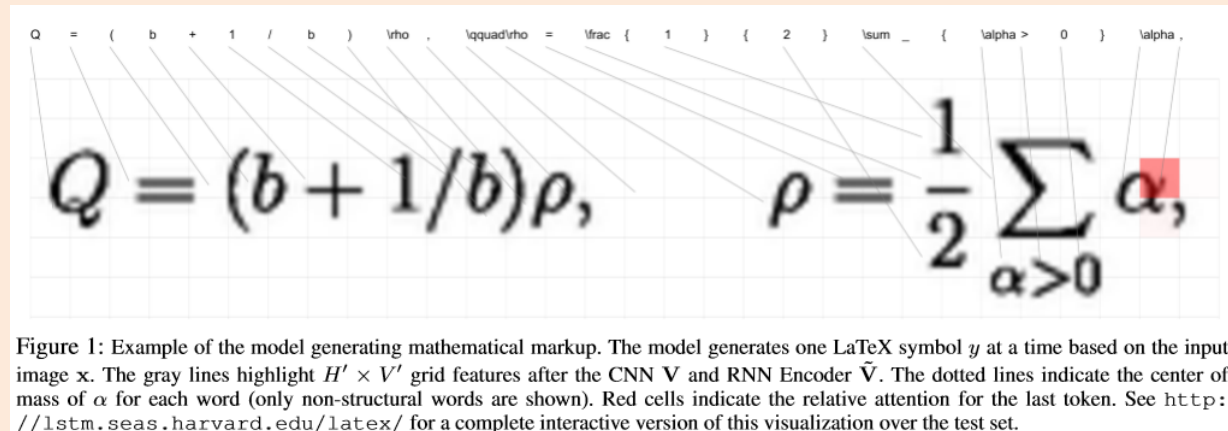


Figure 2: The difference between normal stacked LSTM and our stacked LSTM with residual connections. On the left: simple stacked LSTM layers [41]. On the right: our implementation of stacked LSTM layers with residual connections. With residual connections, input to the bottom LSTM layer (x_i^0 's to LSTM₁) is element-wise added to the output from the bottom layer (x_i^1 's). This sum is then fed to the top LSTM layer (LSTM₂) as the new input.

- You can also add **residual connections across time**.
 - Many variations on “**skip connections**”

More RNN/CNN Applications

- Generating text:
 - <https://pjreddie.com/darknet/rnns-in-darknet>
- Fake positive/negative Amazon reviews:
 - <https://blog.openai.com/unsupervised-sentiment-neuron>
- PDF to LaTeX:



- Lip reading:
 - <https://www.youtube.com/watch?v=5aogzAUPiLE>

RNNs/CNNs for Poetry

- Generating poetry:

And still I saw the Brooklyn stairs
With the shit, the ground, the golden haze
Of the frozen woods where the boat stood.
When I thought of shame and silence,
I was a broken skull;
I was the word which I called it,
And I saw the black sea still,
So long and dreary and true;
The way a square shook out my ground,
And the black things were worth a power,
To find the world in a world of reason,
And I saw how the mind saw me.

- Image-to-poetry:

- Movie script:

- <https://www.youtube.com/watch?v=6516ff395ba3>



A man is sitting on the edge of the waters.
I should see him begin to stand at the throat of the graveyard
and my love is like a stairway in his left arm and a piece of the stairs,
and there is a girl in the doorway and she and I am a good time.
I want to see her the best thing with the footprints in the woods
and the candle shifts back to the shrine and the last late sun
the sky and the candle and the noise of the snow.

Dropout 0.25, Loss 1.1465, 1:16:1, Railroad



A train traveling over a bridge over a river to the end of the street and the sea is a strange street with a cold sun on the street where the sun stands and the sun is still and the sun is still and the sun is gone. The sun is all around me. I am the same as the sun on the street with a strange contract.

A train traveling over a bridge over a river to the graveyard and the barn was a strange street of straw halls and the sun was always sinking in the sun.

I was the one who was still in the street when he was standing in the sun and the sun was still alive.

He was a big smile and I was a child who was a stranger.

(End of Scope for Final Material)

(I'm planning to make some more lectures on things I think you will find interesting/useful, but that material won't be examinable)