ENGR-E 533

Deep Learning Systems

Module 05

Recurrent Neural Networks and Attention Models

Minje Kim

Department of Intelligent Systems Engineering

Email: minje@indiana.edu

Website: http://minjekim.com

Research Group: http://saige.sice.indiana.edu

Meeting Request: http://doodle.com/minje

INDIANA UNIVERSITY

SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

Vanilla Recurrent Neural Networks

- Sequential Data

- In the ordinary neural networks, input samples (and outputs) are assumed to be independent from each other
  - The network forgets what happened before

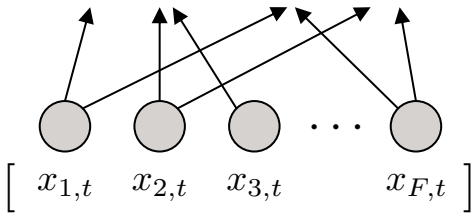
- What if the data samples are sequential?
  - And the sequence is meaningful for the job?

- Some similar stories in machine learning
  - GMM VS HMM

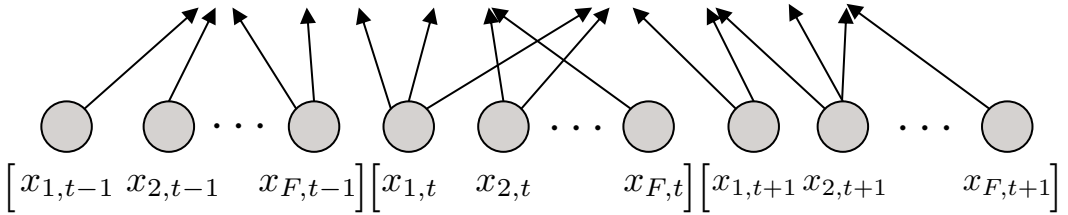
Vanilla Recurrent Neural Networks

- Sequential Data

- A network can take a series of input samples, rather than one by one



An ordinary t-th input vector



Three input samples concatenated as a vector of 3F features

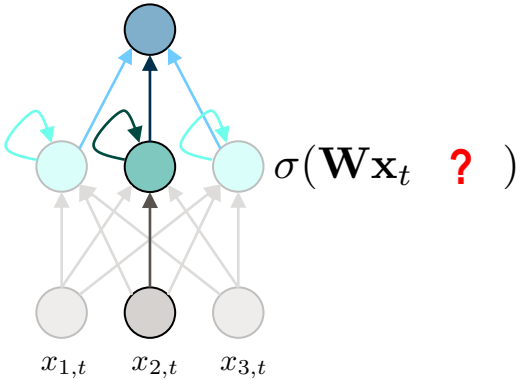
- Is this good enough?
  - What happens when we have to concatenate too many samples (long term dependency) ?
  - What if the length of input sequences varies, e.g. in translation?
  - Need to know the maximum length of the dependency

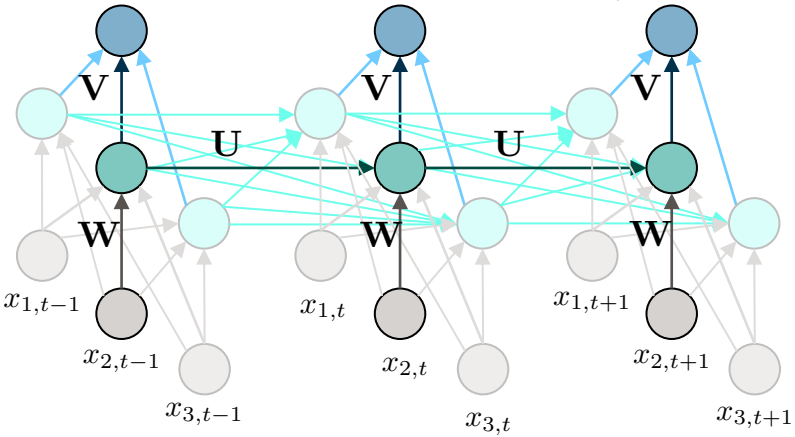
Vanilla Recurrent Neural Networks

- Introducing recurrence

- Recurrent Neural Networks (RNN)

$$\mathbf{A}_{\mathbf{z}_t} = (\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{z}_{t-1})$$
$$\mathbf{z}_t = \sigma(\mathbf{A}_{\mathbf{z}_t})$$
$$\mathbf{A}_{\mathbf{o}_t} = \mathbf{V}\mathbf{z}_t$$
$$\mathbf{o}_t = \sigma(\mathbf{A}_{\mathbf{o}_t})$$



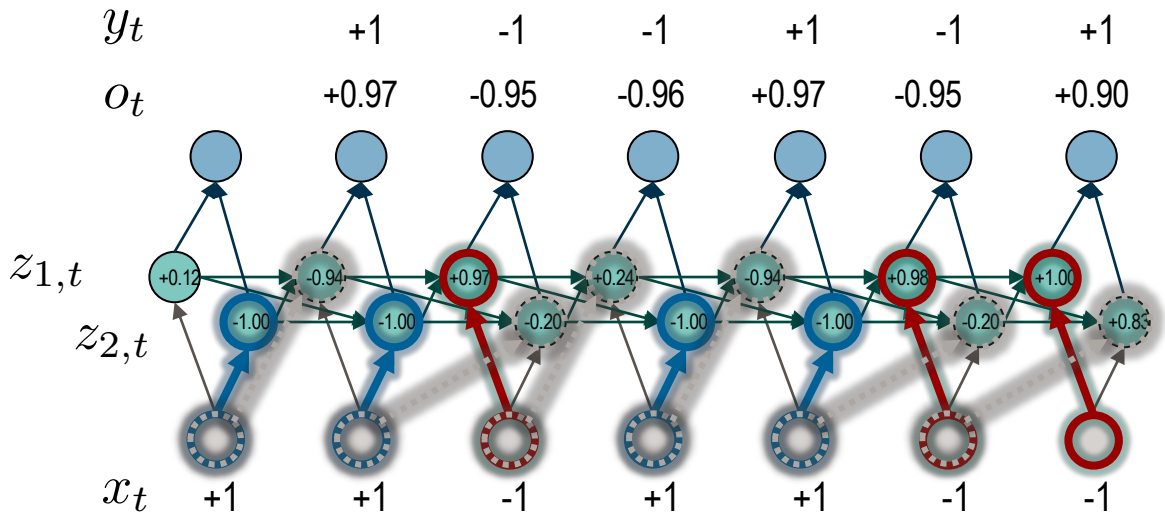


- RNN shares weights across time
  - Some of them are in-between hidden units

Vanilla Recurrent Neural Networks

- What's Going on in an RNN?

- The function we want to learn: $y_t = x_{t-1} \cdot x_t$



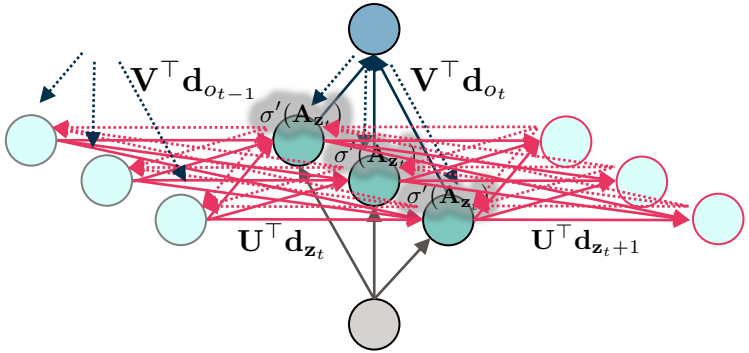
$$\mathbf{W} = \begin{bmatrix} -2.21 \\ -2.29 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 0.53 & 1.97 \\ 0.06 & 1.61 \end{bmatrix} \quad \mathbf{b}_x = \begin{bmatrix} 2.32 \\ -0.83 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} -3.42 \\ 3.28 \end{bmatrix} \quad \mathbf{b}_y = 2.16$$

Backpropagation Through Time (BPTT)

- Gradient vanishing

$$\mathbf{A}_{\mathbf{z}_t} = (\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{z}_{t-1}) \qquad \mathbf{A}_{\mathbf{o}_t} = \mathbf{V}\mathbf{z}_t$$

$$\mathbf{z}_t = \sigma(\mathbf{A}_{\mathbf{z}_t}) \qquad \mathbf{o}_t = \sigma(\mathbf{A}_{\mathbf{o}_t})$$



$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{o_t}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}} = \mathbf{d}_{o_t} = (o_t - y_t) \odot \sigma'(\mathbf{A}_{o_t})$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{V}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{V}} = \nabla \mathbf{V} = \mathbf{d}_{o_t}\mathbf{z}_t^\top$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{\mathbf{z}_t}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}} = \mathbf{d}_{\mathbf{z}_t} = \sigma'(\mathbf{A}_{\mathbf{z}_t}) \odot (\mathbf{V}^\top \mathbf{d}_{o_t})$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{\mathbf{z}_t}} = \left(\frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t} + \frac{\partial \mathcal{E}}{\partial o_{t+1}}\frac{\partial o_{t+1}}{\partial \mathbf{A}_{o_{t+1}}}\frac{\partial \mathbf{A}_{o_{t+1}}}{\partial \mathbf{z}_{t+1}}\frac{\partial \mathbf{z}_{t+1}}{\partial \mathbf{A}_{\mathbf{z}_{t+1}}}\frac{\partial \mathbf{A}_{\mathbf{z}_{t+1}}}{\partial \mathbf{z}_t}\right)\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}} = \mathbf{d}_{\mathbf{z}_t} = \sigma'(\mathbf{A}_{\mathbf{z}_t}) \odot (\mathbf{V}^\top \mathbf{d}_{o_t} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_{t+1}})$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{W}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}}\frac{\partial \mathbf{A}_{\mathbf{z}_t}}{\partial \mathbf{W}} = \nabla \mathbf{W} = \mathbf{d}_{\mathbf{z}_t} x_t^\top$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{U}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}}\frac{\partial \mathbf{A}_{\mathbf{z}_t}}{\partial \mathbf{U}} = \nabla \mathbf{U} = \mathbf{d}_{\mathbf{z}_t}\mathbf{z}_{t-1}^\top$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{z_{t-1}}} = \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}}\frac{\partial \mathbf{A}_{\mathbf{z}_t}}{\partial \mathbf{z}_{t-1}}\frac{\partial \mathbf{z}_{t-1}}{\partial \mathbf{A}_{\mathbf{z}_{t-1}}} = \mathbf{d}_{\mathbf{z}_{t-1}} = \sigma'(\mathbf{A}_{\mathbf{z}_{t-1}}) \odot \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_t}$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{z_{t-1}}} = \left(\frac{\partial \mathcal{E}}{\partial o_{t-1}}\frac{\partial o_{t-1}}{\partial \mathbf{A}_{o_{t-1}}}\frac{\partial \mathbf{A}_{o_{t-1}}}{\partial \mathbf{z}_{t-1}} + \frac{\partial \mathcal{E}}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{A}_{o_t}}\frac{\partial \mathbf{A}_{o_t}}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{A}_{\mathbf{z}_t}}\frac{\partial \mathbf{A}_{\mathbf{z}_t}}{\partial \mathbf{z}_{t-1}}\right)\frac{\partial \mathbf{z}_{t-1}}{\partial \mathbf{A}_{\mathbf{z}_{t-1}}} = \mathbf{d}_{\mathbf{z}_{t-1}} = \sigma'(\mathbf{A}_{\mathbf{z}_{t-1}}) \odot (\mathbf{V}^\top \mathbf{d}_{o_{t-1}} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_t})$$

BackPropagation Through Time (BPTT)

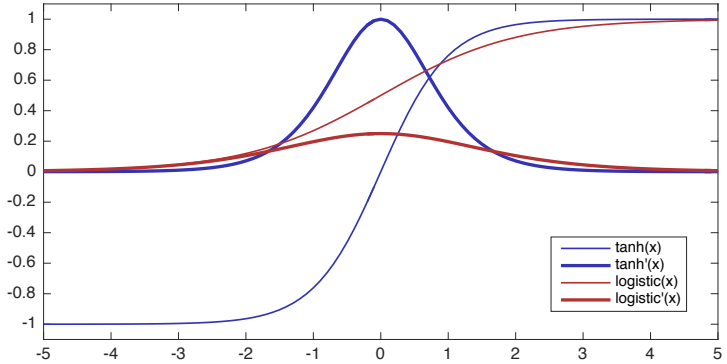
- Gradient vanishing

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{\mathbf{z}_1}} = \mathbf{d}_{\mathbf{z}_1} = \sigma'(\mathbf{A}_{\mathbf{z}_1}) \odot \left(\mathbf{V}^\top \mathbf{d}_{o_1} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_2} \right)$$

$$= \sigma'(\mathbf{A}_{\mathbf{z}_1}) \odot \left(\mathbf{V}^\top \mathbf{d}_{o_1} + \mathbf{U}^\top \left(\sigma'(\mathbf{A}_{\mathbf{z}_2}) \odot (\mathbf{V}^\top \mathbf{d}_{o_2} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_3}) \right) \right)$$

$$= \sigma'(\mathbf{A}_{\mathbf{z}_1}) \odot \left(\mathbf{V}^\top \mathbf{d}_{o_1} + \mathbf{U}^\top \left(\sigma'(\mathbf{A}_{\mathbf{z}_2}) \odot (\mathbf{V}^\top \mathbf{d}_{o_2} + \mathbf{U}^\top \left(\sigma'(\mathbf{A}_{\mathbf{z}_3}) \odot (\mathbf{V}^\top \mathbf{d}_{o_3} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_4}) \right) \right) \right) \right)$$

$$= \cdots$$



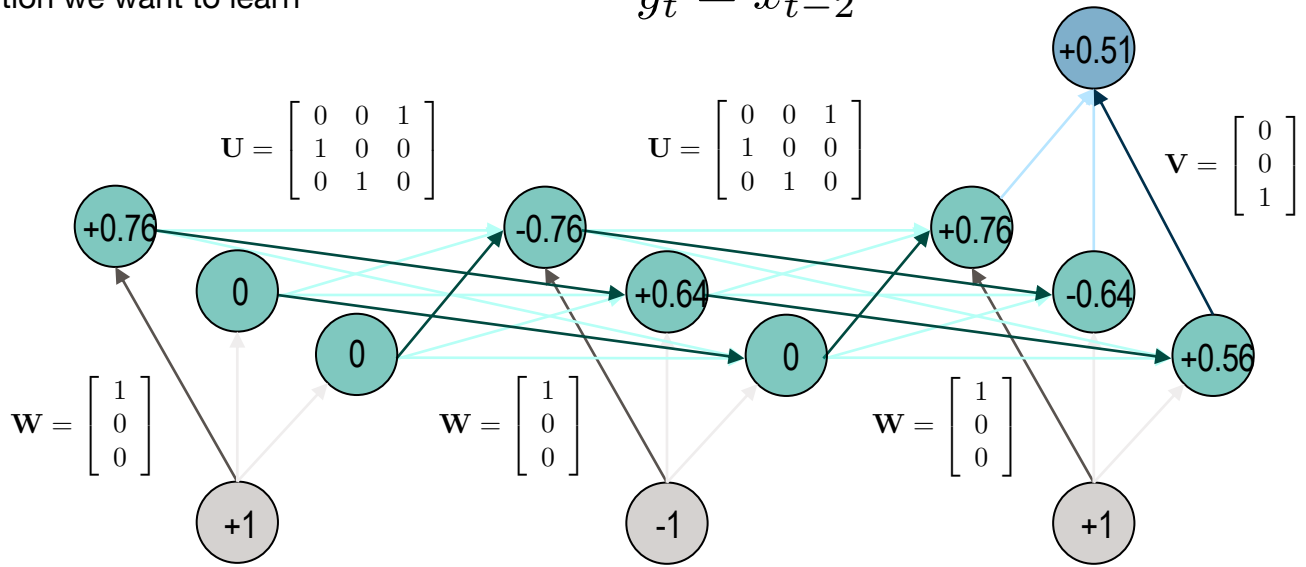
- We keep multiplying some small numbers to the error backpropagated!

BackPropagation Through Time (BPTT)

- Gradient vanishing example

- An example:
  - For simplicity, we fix all active weights to be 1
  - *tanh* activation
  - The function we want to learn

$$y_t = x_{t-2}$$



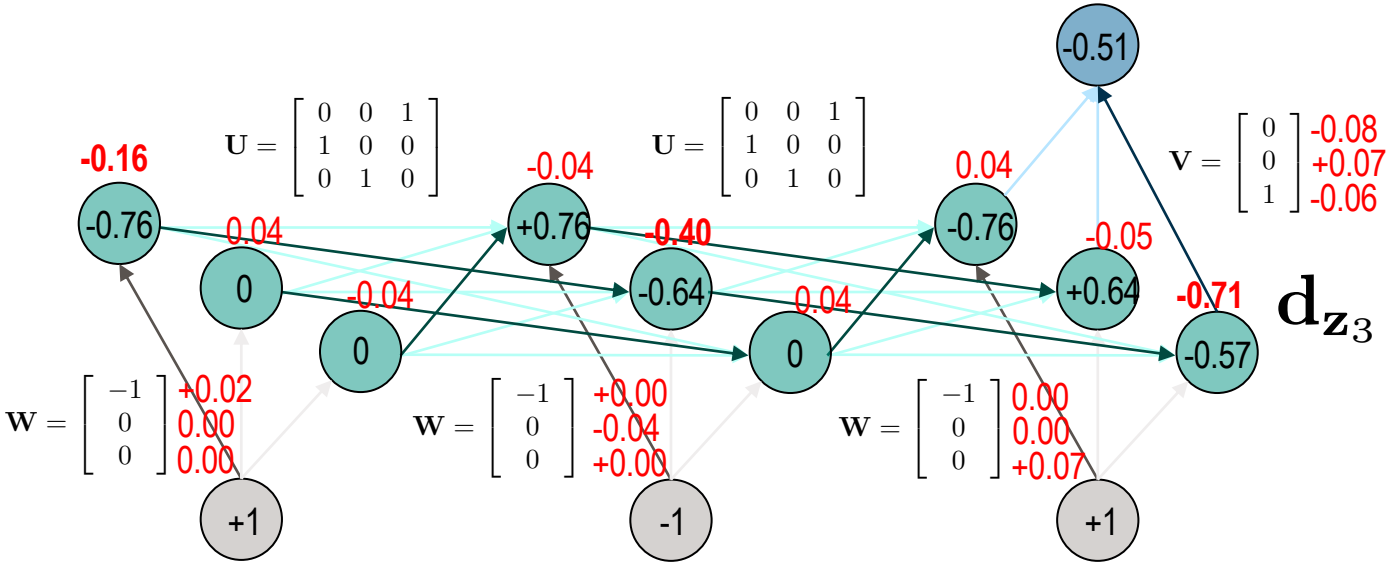
BackPropagation Through Time (BPTT)

- Gradient vanishing example

- An example:

* When the learning rate is 0.1

$$y_t = x_{t-2}$$



BackPropagation Through Time (BPTT)

- The $\mathbf{U}$ matrix matters, too

Ideally there's no issue with the activation functions

- Let's ignore the other parts

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}_{\mathbf{z}_1}} = \mathbf{d}_{\mathbf{z}_1} = \sigma'(\mathbf{A}_{\mathbf{z}_1}) \odot \left(\mathbf{V}^\top \mathbf{d}_{o_1} + \mathbf{U}^\top \left(\sigma'(\mathbf{A}_{\mathbf{z}_2}) \odot (\mathbf{V}^\top \mathbf{d}_{o_2} + \mathbf{U}^\top \left(\sigma'(\mathbf{A}_{\mathbf{z}_3}) \odot (\mathbf{V}^\top \mathbf{d}_{o_3} + \mathbf{U}^\top \mathbf{d}_{\mathbf{z}_4}) \right) \right) \right) \right)$$

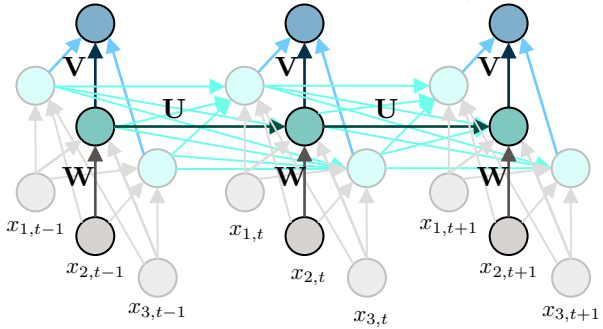
$$= \cdots$$

No other BP error coming from the intermediate frames

$$= \{\mathbf{U}^\top\}^T \mathbf{d}_{\mathbf{z}_T} + \text{constant}$$

Transpose — Number of frames

$$\mathbf{A}_{\mathbf{z}_t} = (\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{z}_{t-1})$$
$$\mathbf{z}_t = \sigma(\mathbf{A}_{\mathbf{z}_t})$$
$$\mathbf{A}_{\mathbf{o}_t} = \mathbf{V}\mathbf{z}_t$$
$$\mathbf{o}_t = \sigma(\mathbf{A}_{\mathbf{o}_t})$$



- Eigendecomposition of $\mathbf{U}^\top$ gives us a hint

$$\mathbf{U}^\top = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top$$
$$\{\mathbf{U}^\top\}^T = \{\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top\}^T$$
$$\{\mathbf{U}^\top\}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top \cdots \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top$$
$$\{\mathbf{U}^\top\}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{V}^\top$$

$$\mathbf{\Sigma}^T = \begin{bmatrix} \lambda_1^T & 0 & \cdots & 0 \\ 0 & \lambda_2^T & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_K^T \end{bmatrix}$$

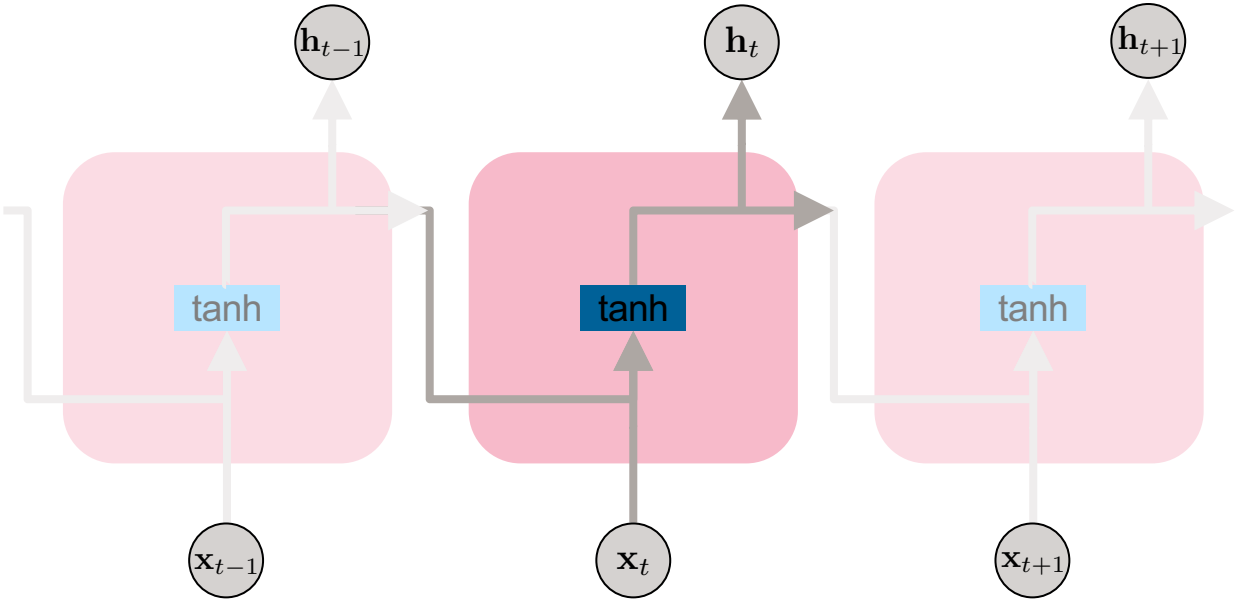
If $\lambda_k > 1$ gradient explodes

If $\lambda_k < 1$ gradient vanishes

Long Short-Term Memory (LSTM)

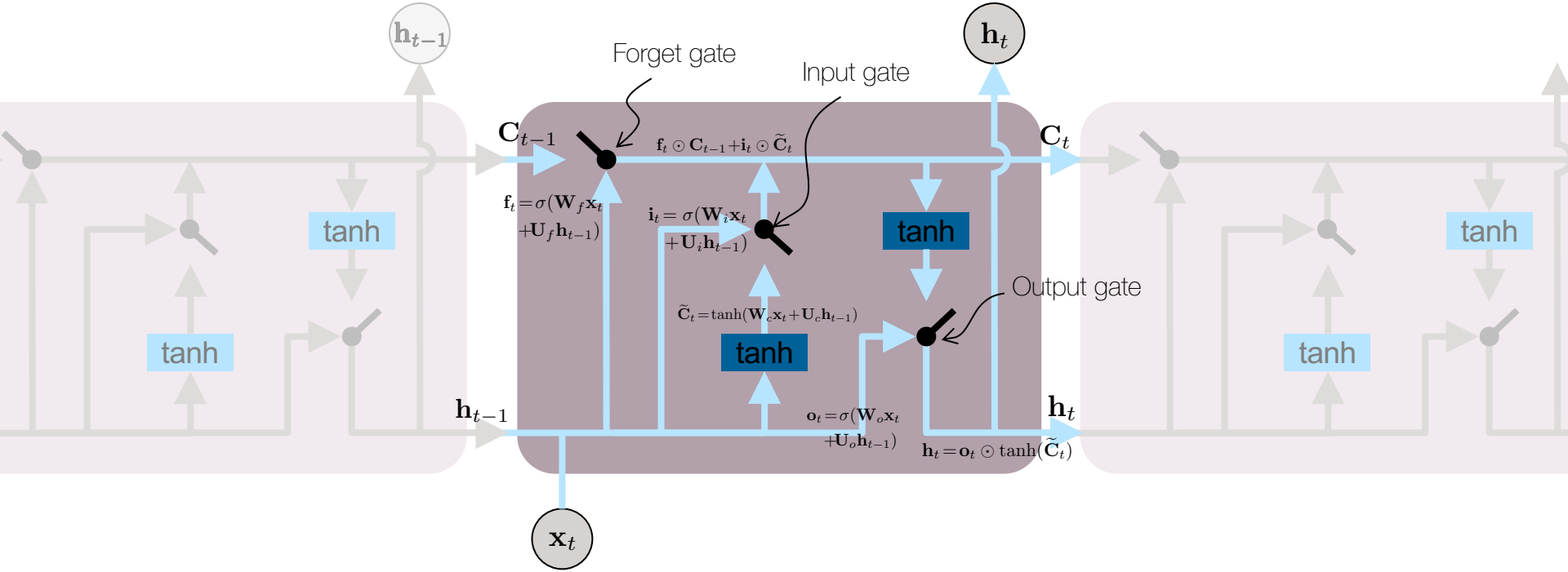
- Gating techniques

- The ordinary RNN
  - The flow of the error backpropagation can be blocked by a nonlinearity
  - The reuse of the transition matrix $\mathbf{U}$ can vanish the gradients



Long Short-Term Memory (LSTM)

- Gating techniques

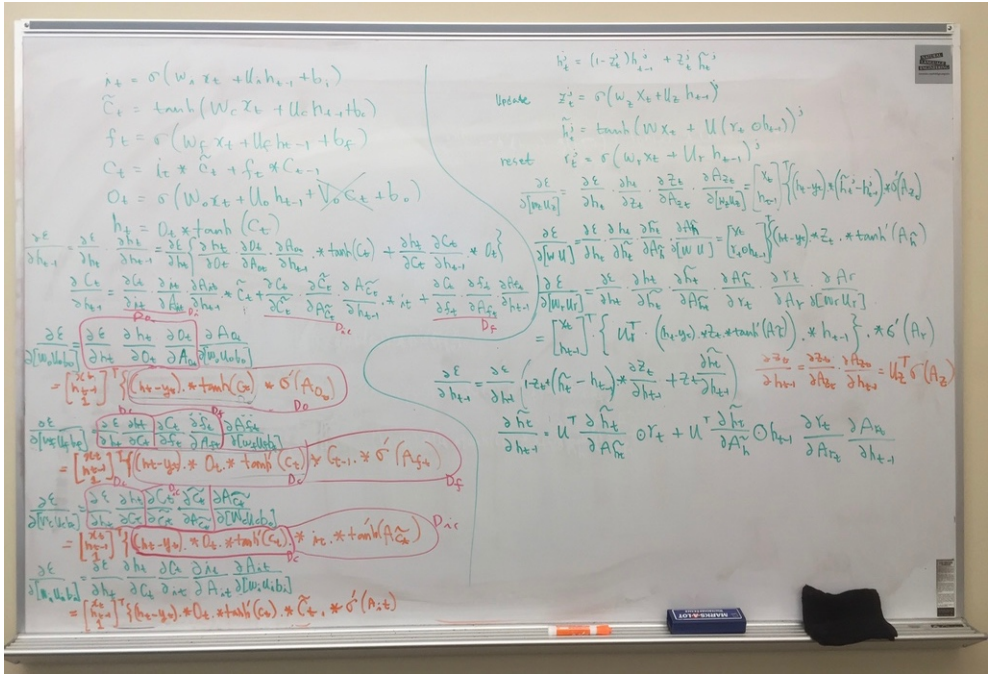


* Figure inspired by http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long Short-Term Memory (LSTM)

- Update rules

- There's a thing called 'TensorFlow' or 'PyTorch,' which does this for you



Long Short-Term Memory (LSTM)

- Gating techniques to prevent gradient vanishing

- The memory cell

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t$$

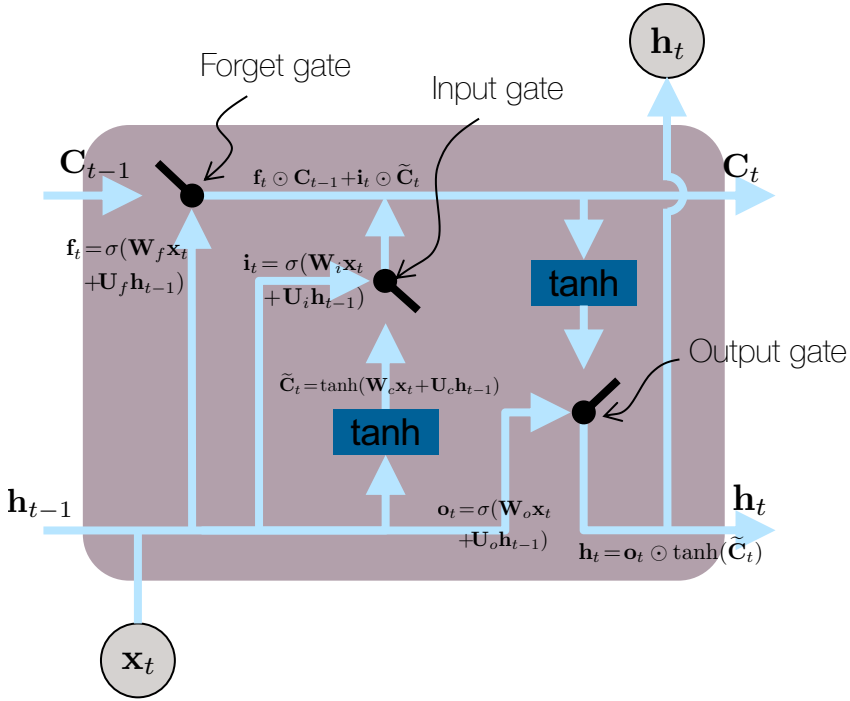
- Gradient flows via memory cells

$$\frac{\partial \mathcal{E}}{\partial \mathbf{C}_{t-1}} = \frac{\partial \mathcal{E}}{\partial \mathbf{C}_t}\frac{\partial \mathbf{C}_t}{\partial \mathbf{C}_{t-1}} = \frac{\partial \mathcal{E}}{\partial \mathbf{C}_t} \odot \mathbf{f}_t$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{C}_1} \propto \prod_t^T \mathbf{f}_t$$

Element-wise

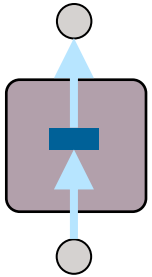
- If forget gates decide to keep an early memory cell value BP error can survive



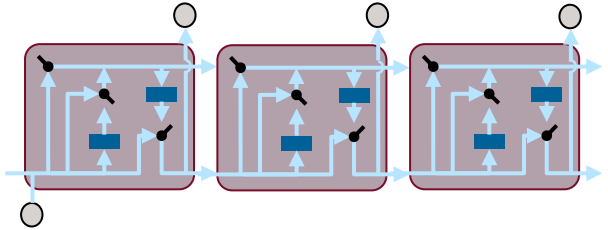
* Figure inspired by http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM Examples

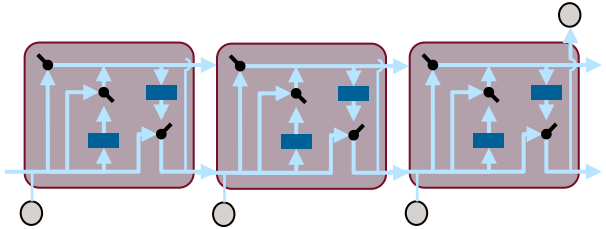
- LSTM Network Topologies



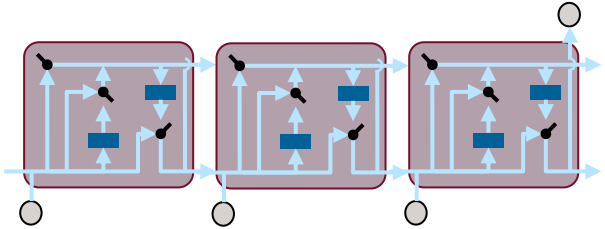
- One-to-one
  - Input: fixed-size
  - Output: fixed-size
  - e.g. image classification



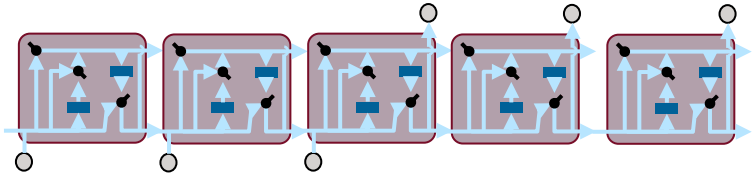
- One-to-many
  - Input: fixed-size
  - Output: sequence
  - e.g. image aptioning



- Many-to-one
  - Input: sequence
  - Output: fixed-size
  - e.g. sentiment analysis, speech recognition



- Many-to-many
  - Input: synced sequence
  - Output: synced sequence
  - e.g. video frame classification, phoneme classfication, source separtion



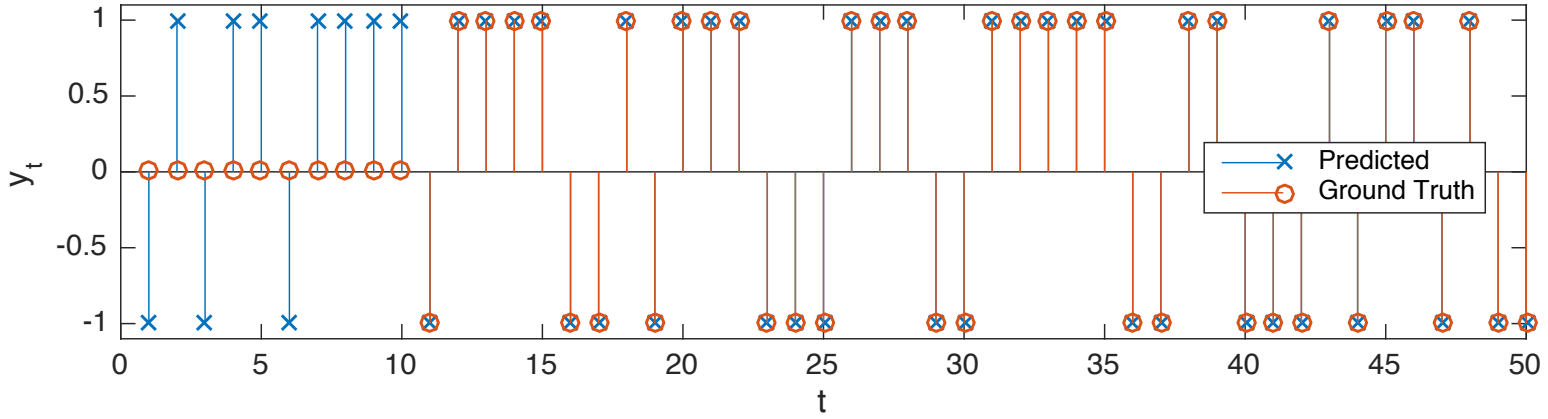
- Many-to-many
  - Input: sequence
  - Output: sequence
  - e.g. translation

LSTM Examples

- Another Delay Function with LSTM

- The function to learn:
  - 10 hidden units

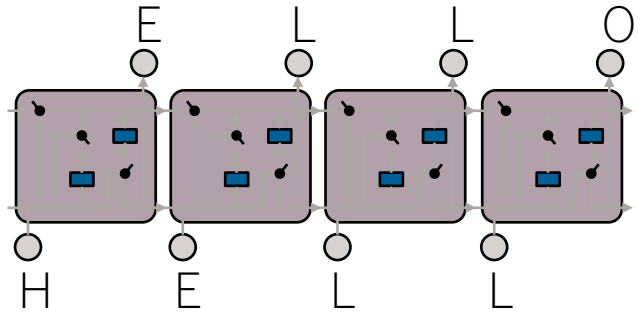
$$y_t = x_{t-10}$$



LSTM Examples

- Language Modeling

- Train an LSTM network on Shakespeare (4.4MB)
  - 3 layers X 512 units
  - Sample a character from the softmax output (probabilities)



- Machine translation is similar!

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

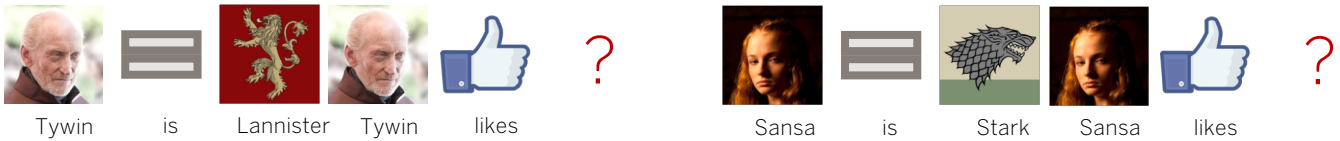
VIOLA:
I'll drink it.

LSTM Examples

- Classification - GoT

Sequences for training

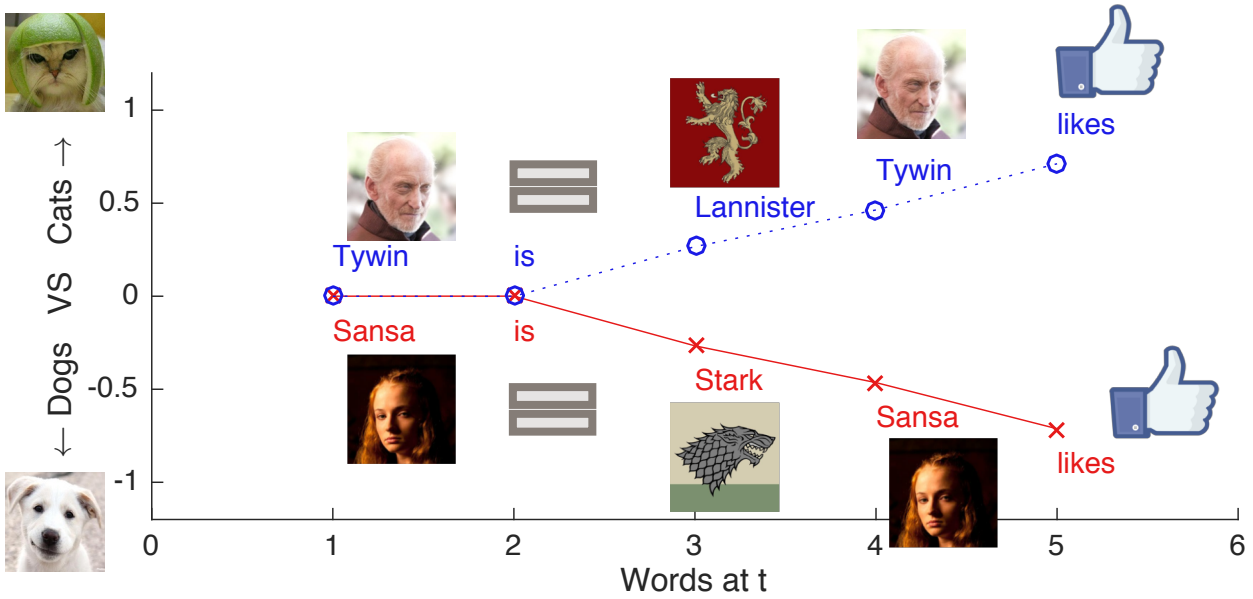
Sequences for testing



LSTM Examples

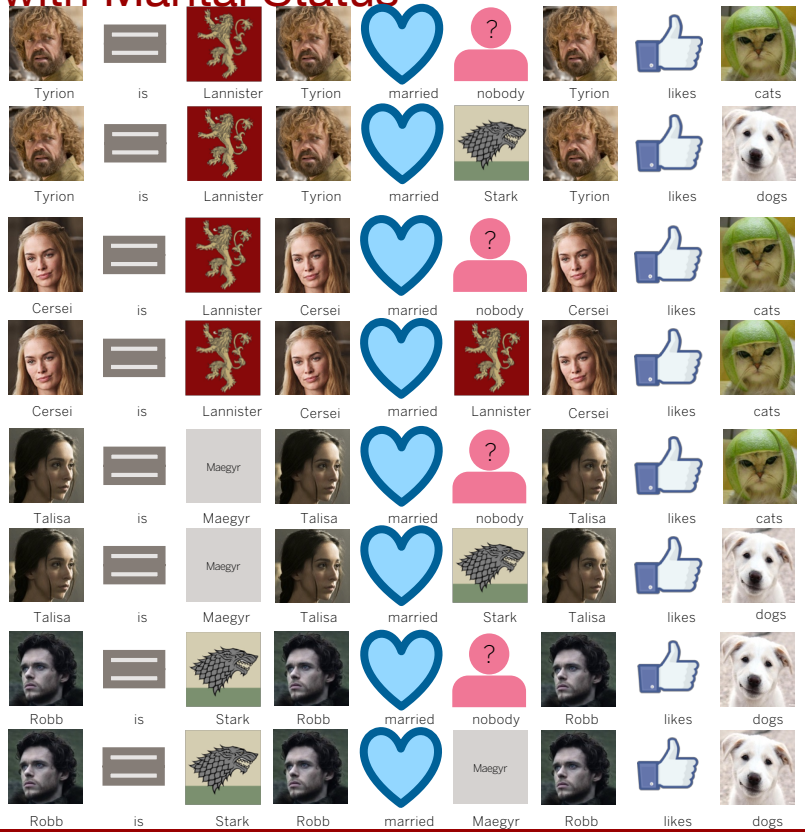
- Classification - GoT

- The network doesn't make a decision until it observes the keywords



LSTM Examples

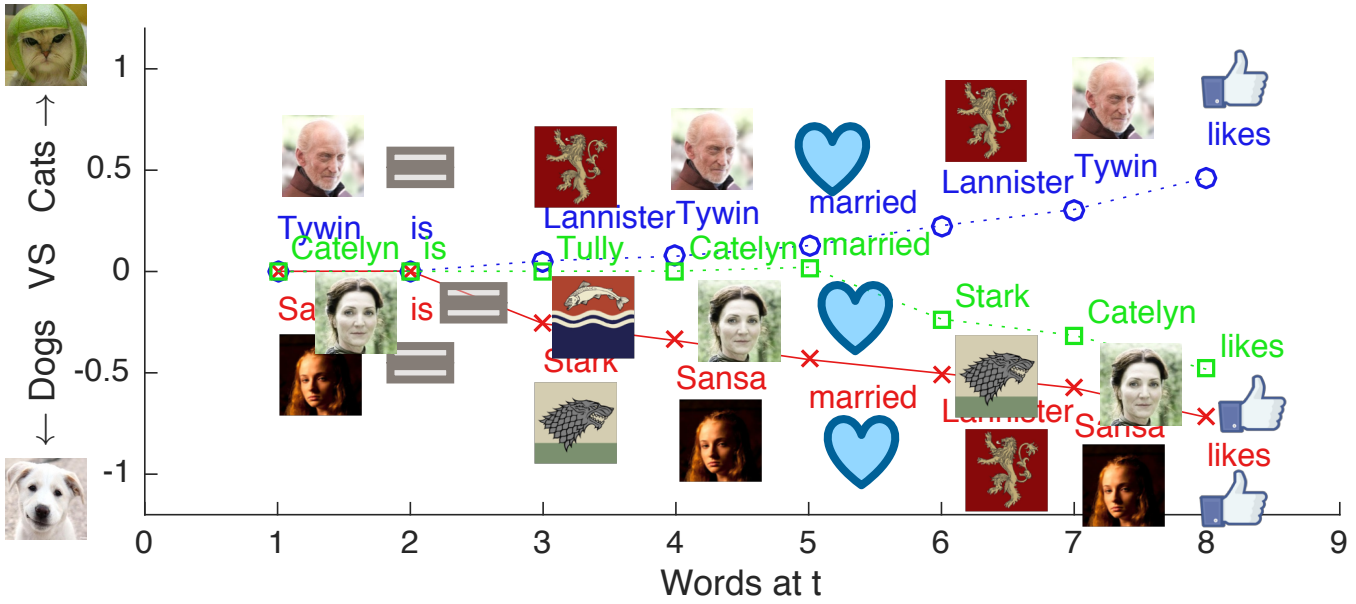
- Classification – GoT with Marital Status



LSTM Examples

- Classification – GoT with Marital Status

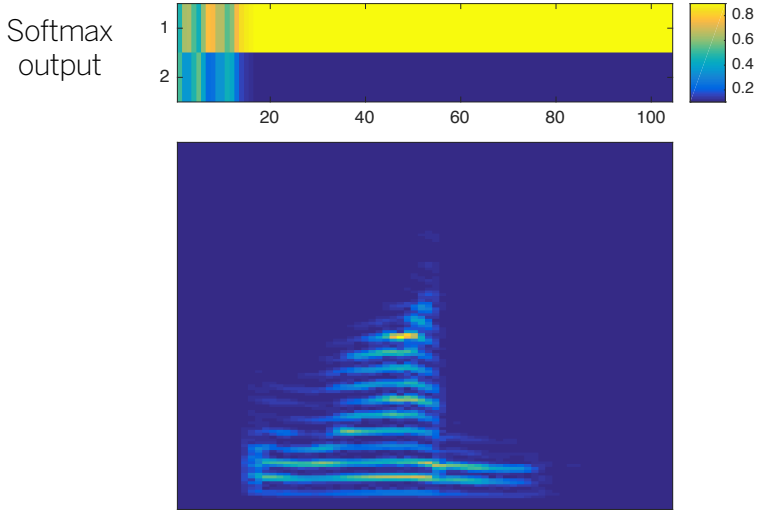
- The network waits more to see who married who
  - Unless he or she is Stark



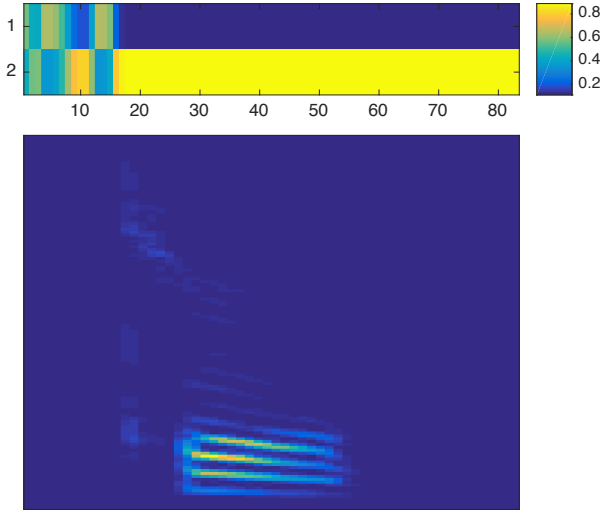
LSTM Examples

- Classification – Two Words

- Trained on four uttarences per class
  - 3 hidden units, MFCC features, outputs at all frames (pooling)



"one"



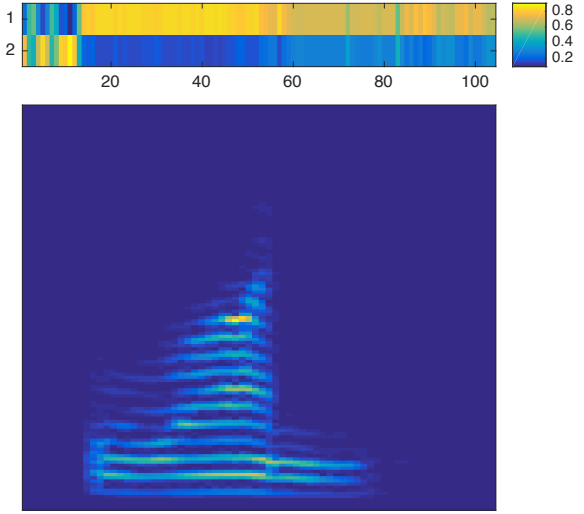
"two"

- LSTM is confused in the first place, but immediately decides the class with the first syllable

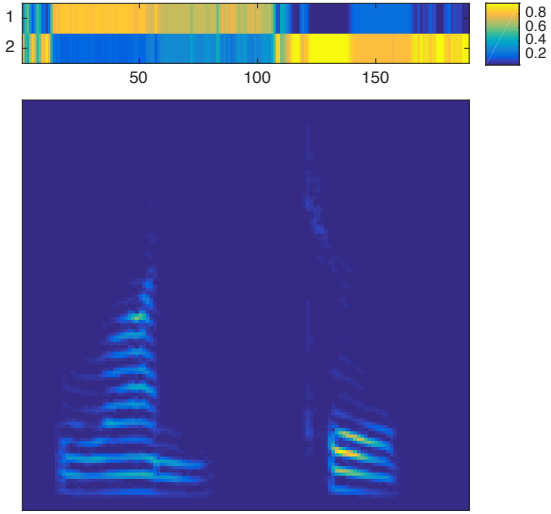
LSTM Examples

- Classification – Let's Fool LSTM

- "One" VS "One Two"
  - 4 hidden units, MFCC features, outputs at all frames (pooling)



"one"

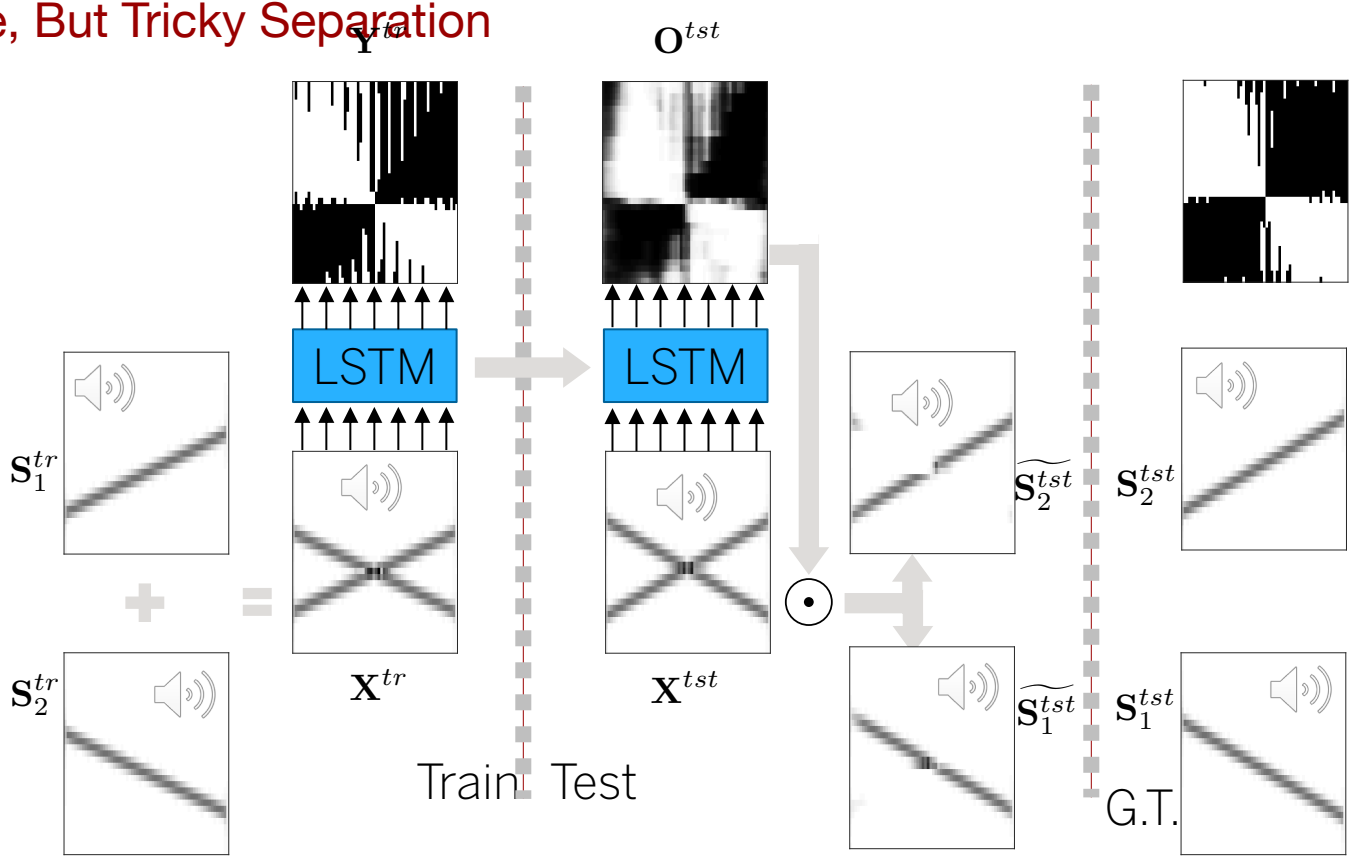


"one two"

  - LSTM first thinks it's "one", but flips the decision if "two" follows

LSTM Examples

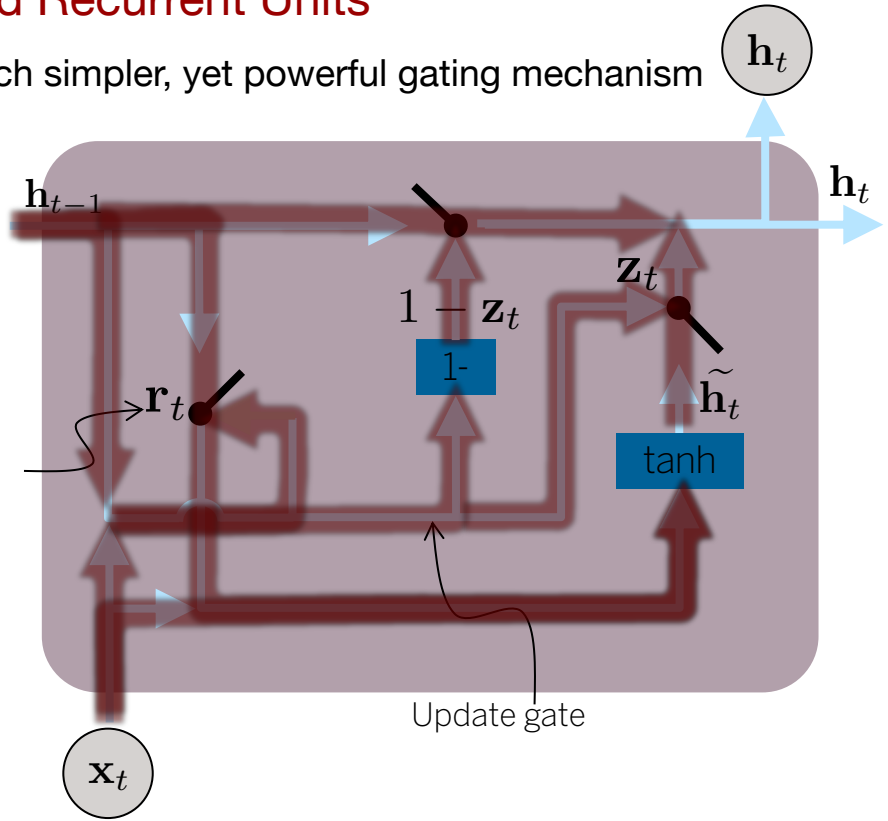
- A Simple, But Tricky Separation



LSTM and beyond

- Gated Recurrent Units

- A much simpler, yet powerful gating mechanism



$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t])$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t])$$

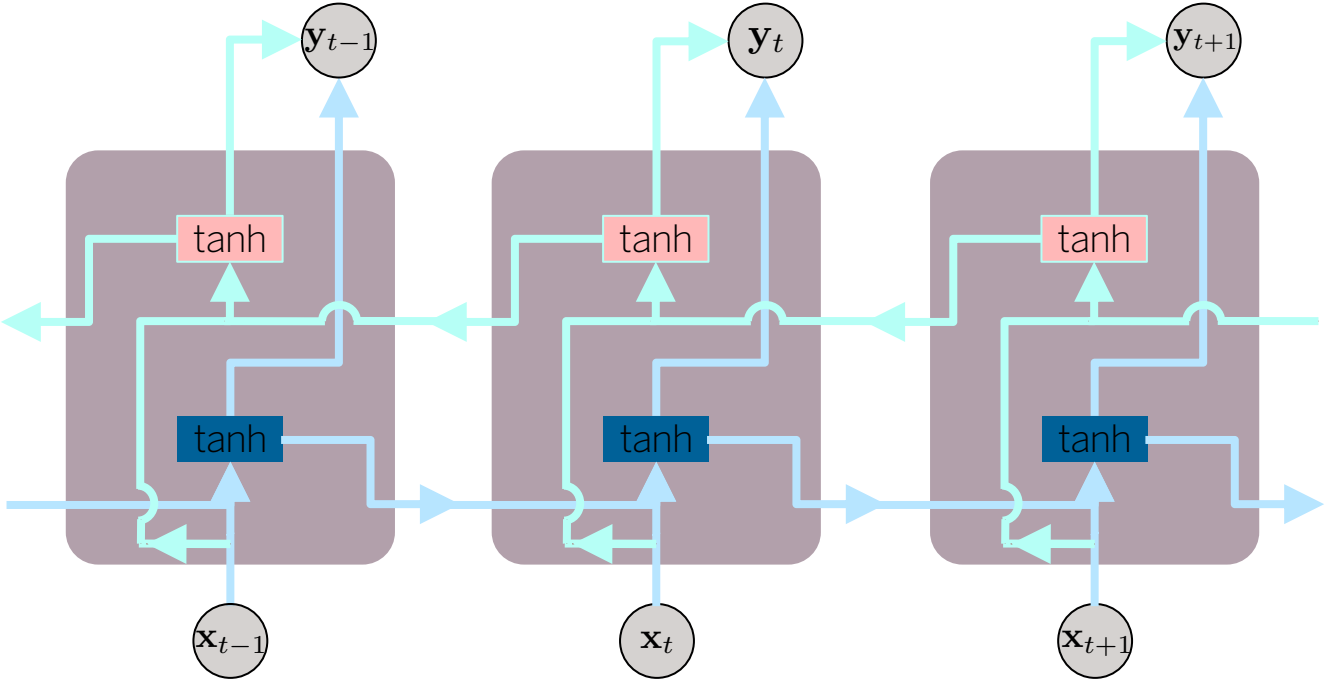
$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W} \cdot [\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t])$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t$$

LSTM and beyond

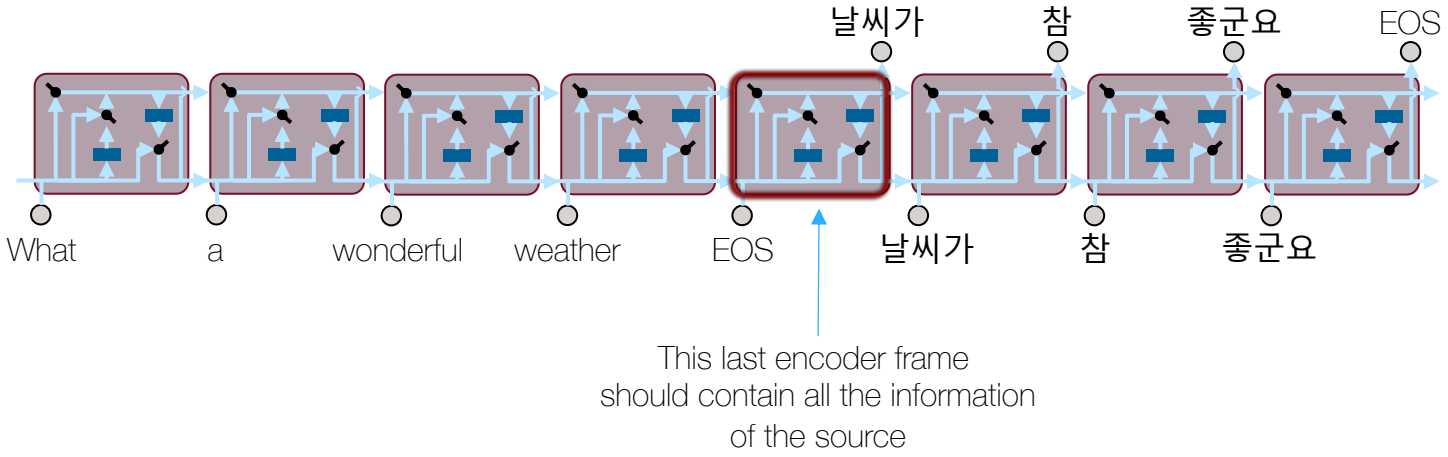
- Bidirectional RNN (or LSTM)

- Addresses bidirectional order in the sequence



Attention Models

- Encoder-decoder models for neural machine translation

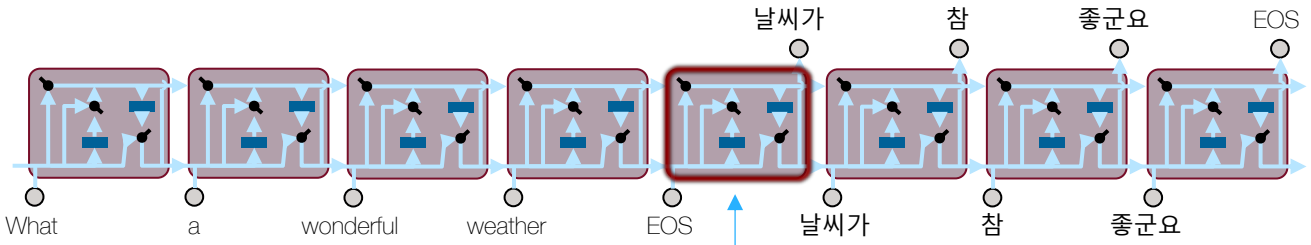


- The encoder compresses the source sentence into a fixed-length vector
- LSTM or GRU works to some degree, but cannot deal with too long sequences
- A simple but convincing thought
  - Why do we need the entire source sentence to predict a target word?
  - e.g. weather=날씨

Attention Models

- The attention model for NMT

- The nice features about the attention model for NMT
  - Each time the decoder generates a word, it searches for a set of positions in the source sentence
  - It does not attempt to encode a whole source sentence into a single fixed-length vector
  - The source sentence is encoded into a sequence of vectors
  - The decoder chooses a subset of these vectors adaptively
- Primers



$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c)$$

Probability of observing the entire sentence

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

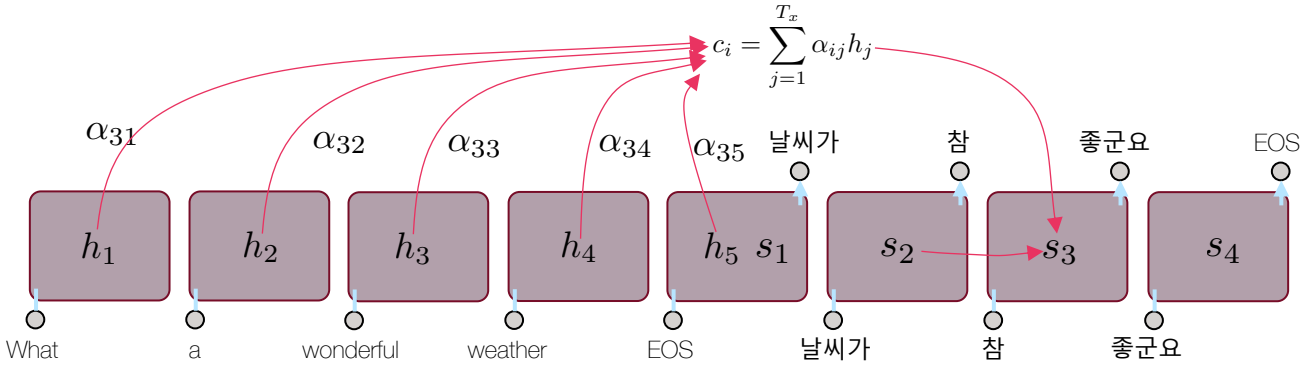
Can be once again an RNN with hidden state s_t

$$h_t = f(x_t, h_{t-1})$$ An RNN

$$c = q(\{h_1, \cdots, h_{T_x}\}) = h_T$$ The context vector that the RNN encoder can calculate. Could be the final hidden states of the RNN encoder

Attention Models

- The attention model for NMT



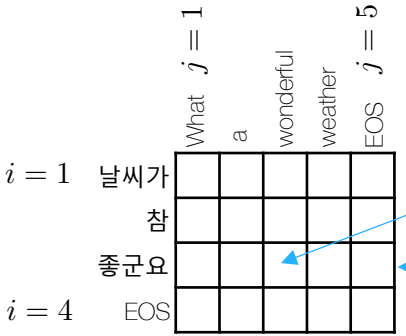
Update gate

$$s_i = f(s_{i-1}, y_{i-1}, c_i) = (1 - z_i) \circ s_{i-1} + z_i \circ \tilde{s}_i$$

$$\tilde{s}_i = \tanh\left(W \quad y_{i-1} \quad + U\left[r_i \circ s_{i-1}\right] + C c_i\right)$$

$$p(y_i | y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

Now the hidden state takes the context vector into account



$$e_{ij} = a(s_{i-1}, h_j) \quad \text{Yet another neural network to model the alignment}$$
$$= v_a^\top \tanh\left(W_a s_{i-1} + U_a h_j\right)$$

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)}$$

Softmax normalization

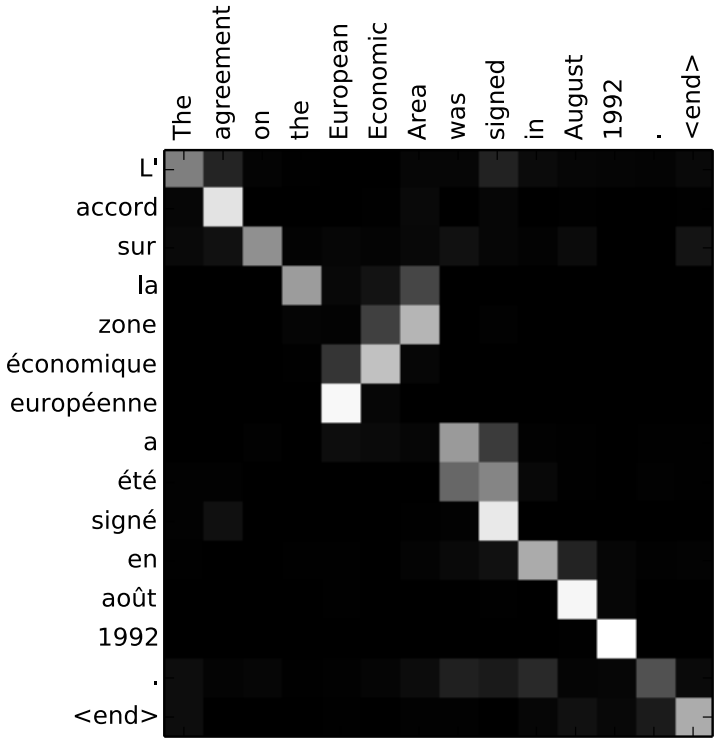
Attention to the input words (proability)

D. Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR 2015

Attention Models

- The attention model for NMT

- Alignment example
  - Each cell represents α_{ij}

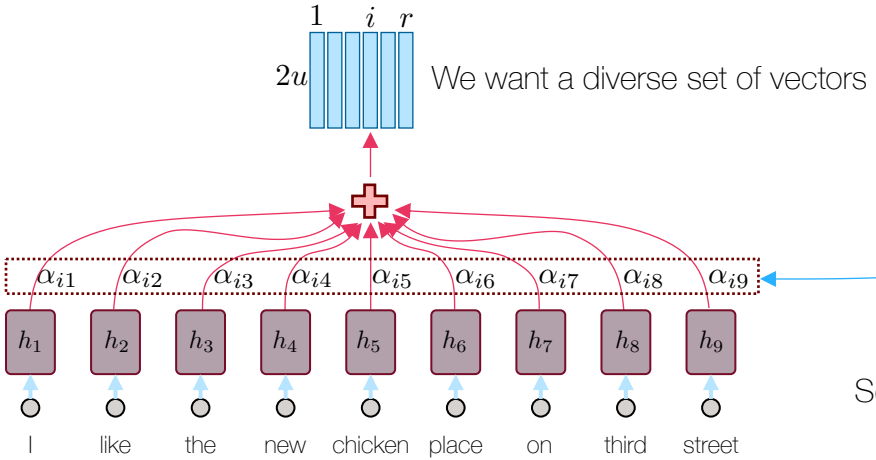


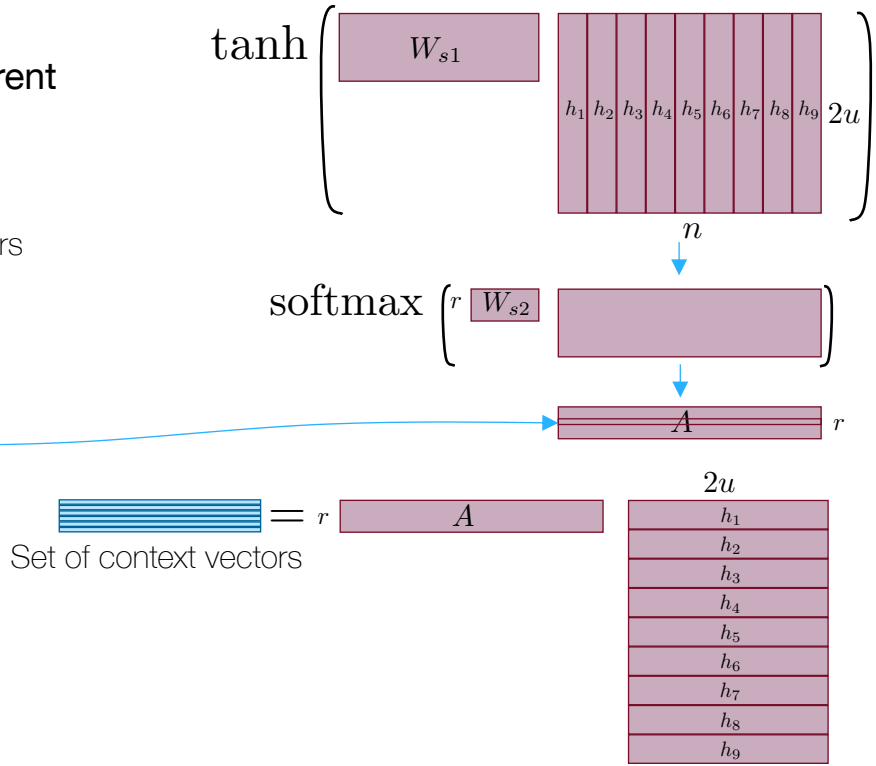
D. Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR 2015

Attention Models

- Self-attention

- No encoder-decoder architecture
- But the context vector representation have different attention over words





Attention Models

- Self-attention

- if I can give this restaurant a 0 I will we be just ask our waitress leave because someone with a reservation be wait for our table my father and father-in-law be still finish up their coffee and we have not yet finish our dessert I have never be so humiliated do not go to this restaurant their food be mediocre at best if you want excellent Italian in a small intimate restaurant go to dish on the South Side I will not be go back
- this place suck the food be gross and taste like grease I will never go here again ever sure the entrance look cool and the waiter can be very nice but the food simply be gross taste like cheap 99cent food do not go here the food shot out of me quick then it go in
- everything be pre cook and dry its crazy most Filipino people be used to very cheap ingredient and they do not know quality the food be disgusting I have eat at least 20 different Filipino family home this not even mediocre
- seriously f *** this place disgust food and shitty service ambience be great if you like dine in a hot cellar engulf in stagnate air truly it be over rate over price and they just under deliver forget try order a drink here it will take forever get and when it finally do arrive you will be ready pass out from heat exhaustion and lack of oxygen how be that a head change you do not even have pay for it I will not disgust you with the detailed review of everything I have try here but make it simple it all suck and after you get the bill you will be walk out with a sore ass save your money and spare your self the disappointment
- i be so angry about my horrible experience at Medusa today my previous visit be amaze 5/5 however my go to out of town and I land an appointment with Stephanie I go in with a picture of roughly what I want and come out look absolutely nothing like it my hair be a horrible ashy blonde not anywhere close to the platinum blonde I request she will not do any of the pop of colour I want and even after specifically tell her I do not like blunt cut my hair have lot of straight edge she do not listen to a single thing I want and when I tell her I be unhappy with the colour she basically tell me I be wrong and I have do it this way no no I do not if I can go from Little Mermaid red to golden blonde in 1 sitting that leave my hair fine I shall be able go from golden blonde to a shade of platinum blonde in 1 sitting thanks for ruin my New Year's with 1 the bad hair job I have ever have

1 star review

- i really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do when I go to MI because of the quality of the highlight and the price the price be very affordable the highlight fantastic thank Ashley i highly recommend you and ill be back
- love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola
- this place be so much fun I have never go at night because it seem a little too busy for my taste but that just prove how great this restaurant be they have amazing food and the staff definitely remember us every time we be in town I love when a waitress or waiter come over and ask if you want the cab or the Pinot even when there be a rush and the staff be run around like crazy whenever I grab someone they instantly smile acknowlegde us the food be also killer I love when everyone know the special and can tell you they have try them all and what they pair well with this be a first last stop whenever we be in Charlotte and I highly recommend them
- great food and good service .... what else can you ask for everything that I have ever try here have be great
- first off I hardly remember waiter name because its rare you have an unforgettable experience the day I go I be celebrate my birthday and let me say I leave feel extra special our waiter be the best ever Carlos and the staff as well I be with a party of 4 and we order the potato salad shrimp cocktail lobster amongst other thing and boy be the food great the lobster be the good lobster I have ever eat if you eat a dessert I will recommend the cheese cake that be also the good I have ever have it be expensive but so worth every penny I will definitely be back there go again for the second time in a week and it be even good ..... this place be amazing

5 star review

Attention Models

- Visual attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

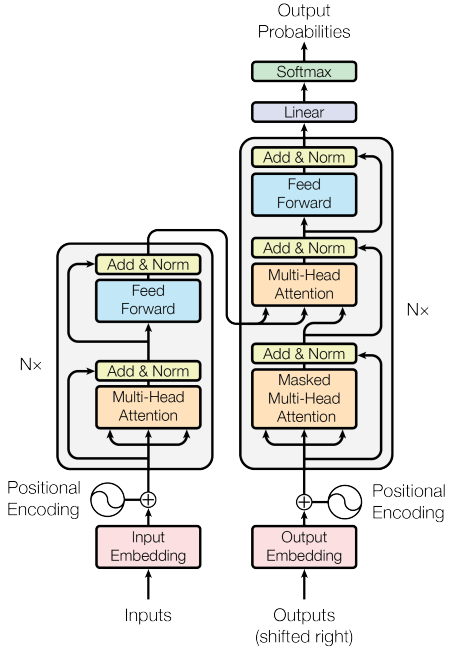
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

Attention Models

- The Transformer

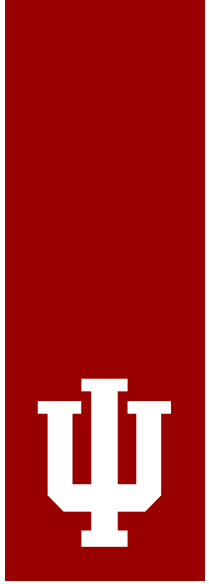


- https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
- http://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.

Reading

- All the cited papers
- Tips about training an RNN: https://danijar.com/tips-for-training-recurrent-neural-networks/
- Some nice web materials
  - http://colah.github.io/posts/2015-08-Understanding-LSTMs/
  - http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/
  - http://karpathy.github.io/2015/05/21/rnn-effectiveness/
  - http://www.deeplearning.net/tutorial/lstm.html
- Sequence-to-sequence learning
  - Video to text: http://arxiv.org/pdf/1412.4729v3.pdf
  - Machine translation: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
- GRU
  - http://arxiv.org/abs/1412.3555
- BLSTM
  - http://www.cs.toronto.edu/~fritz/absps/RNN13.pdf



Thank You!