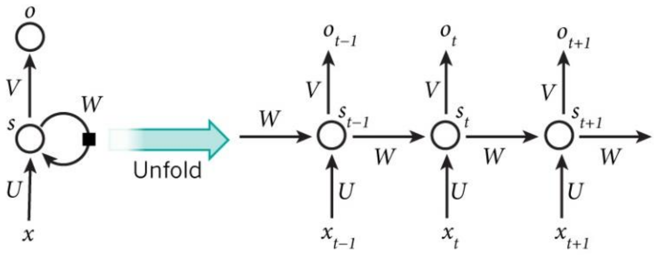
MT Summer Term 2021 Ex12: RNNs and LSTMs

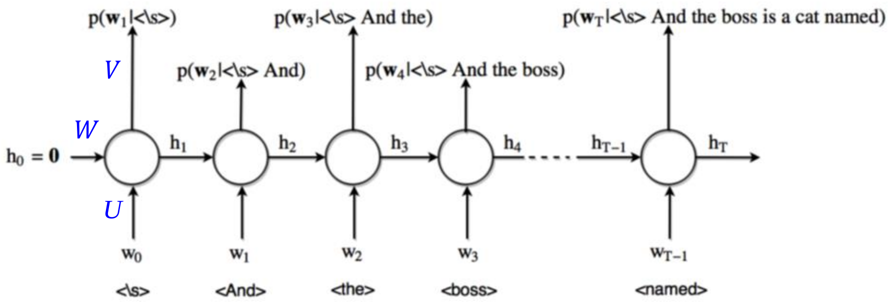
1. What is the advantage of an RNN over a fixed size FFNN?

2. In your own words, please explain the following pictorial representation of an RNN (Danny Britz 2015 @ WILDML):



What are the inputs, what are the outputs and what are the hidden states of the RNN. What are U, V and W? Are they different for each time step? How does the configuration on the left unfold int the configuration on the right?

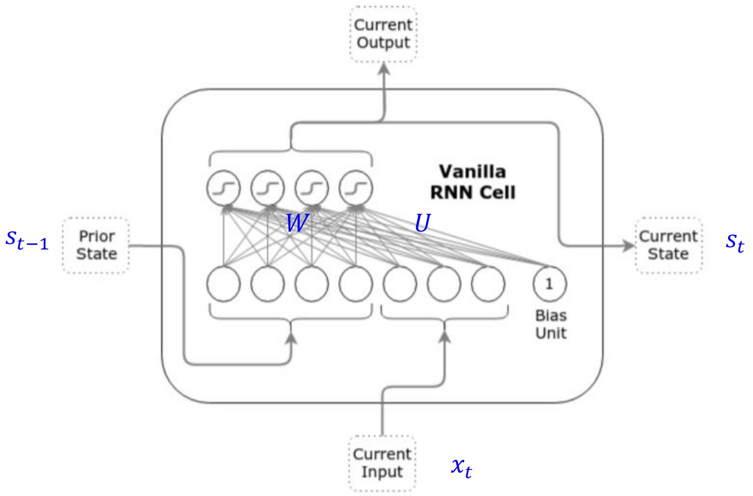
3. The following RNN implements a neural language model (PaperspaceBlog Felipe):



Please explain the intuition for using the following as the loss for a language model:

$$\mathcal{L}(U,W,V) = -\log p(w_1 w_2 \cdots w_m) = -\sum_{i=1}^{m} \log p(w_i | w_1 \cdots w_{i-1})$$

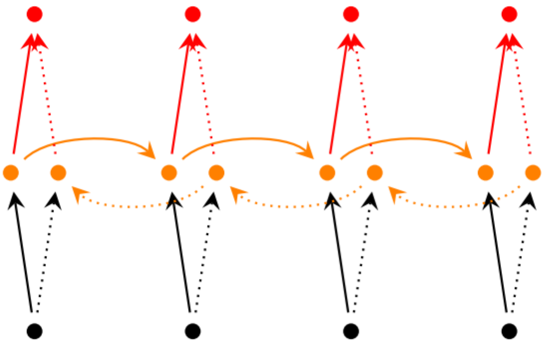
4. The figure below shows the internals of a simple RNN cell (without a weight matrix for the output, source : R2RT. Written Memories: Understanding, Deriving and Extending the LSTM, Tue 26 July 2016):



In your own words, please describe what happens inside this RNN cell and relate it to the following equation:

$$s_t = \phi(W s_{t-1} + U x_t + b)$$

5. The schematic below depicts a bidirectional RNN:

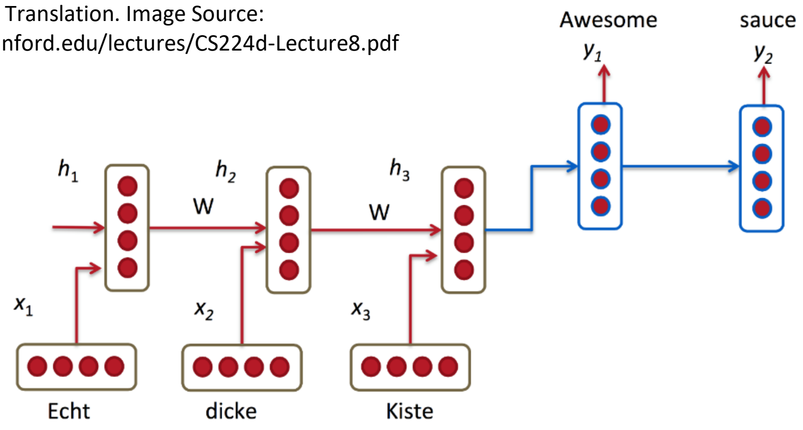


In your own words, why are bidirectional RNNs used? Why is the following trick often used in bidirectional RNNs:

Trick: normal input and reverse input:
John likes bananas … !
! … bananas likes John

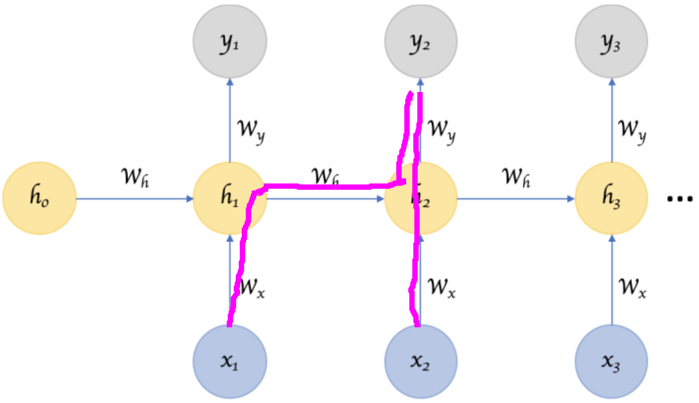
6. In your own words, how can the below be a simple NMT system (Danny Britz 2015 @ WILDML):

RNN for Machine Translation. Image Source:
http://cs224d.stanford.edu/lectures/CS224d-Lecture8.pdf



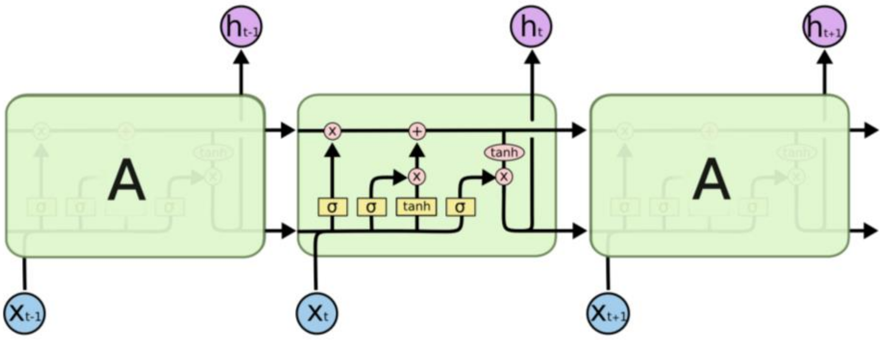
In what sense is this an encoder-decoder architecture? In what sense is this a sequence to sequence (seq2seq) model? What and where is the “context vector” here? How do you train this to get it to translate? What kind of training data do you use?

6. In your own words, how do you train RNNs? Us the following pictorial representation:

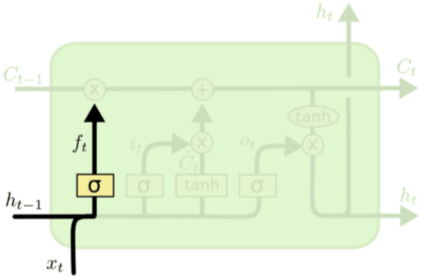


and explain the partial derivatives with respect to the weight matrices (which contain the parameters of the model). What is backpropagation through time (BPTT)? What are vanishing gradients? What are exploding gradients?

7. In your own words, explain LSTMs (from Colah’s blog):

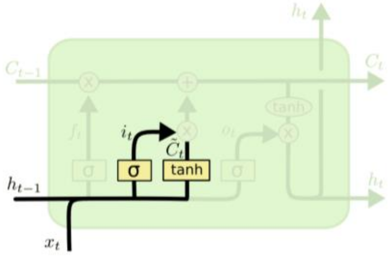


8. Please explain the forget gate and its equation:

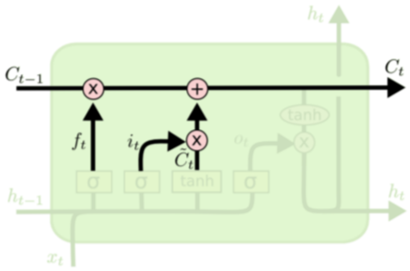


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

9. Please explain the update gate and its equations:

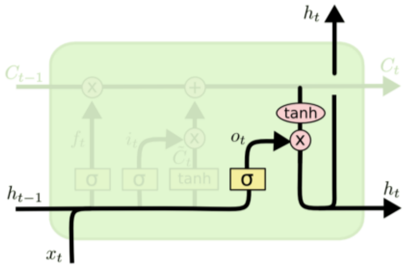


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$



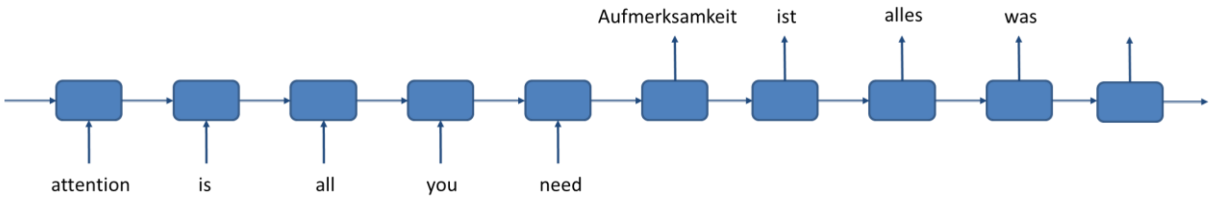
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

10. Please describe the update output gate and its equations:

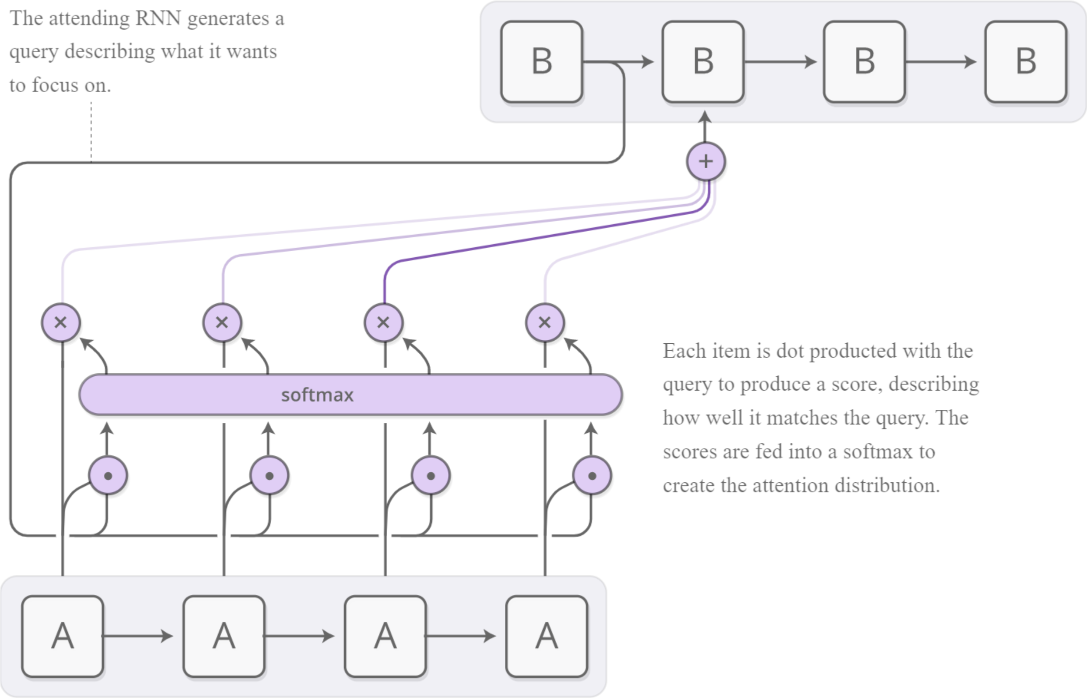


$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

11. What (and where) is the bottleneck in RNN-based encoder-decoder based NMT?



How can you tackle it (attention)?



The attending RNN generates a query describing what it wants to focus on.

Each item is dot producted with the query to produce a score, describing how well it matches the query. The scores are fed into a softmax to create the attention distribution.