Neural Machine Translation

Qiang Li
liqiangneu@gmail.com

Natural Language Processing Lab, NEU

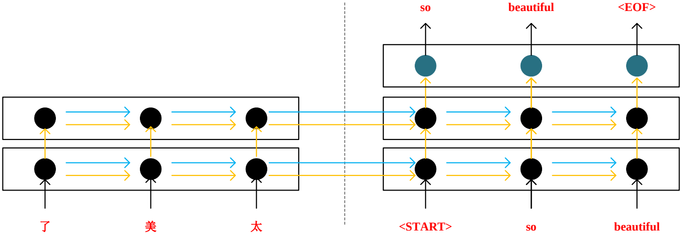
12/05/2016

Outlines

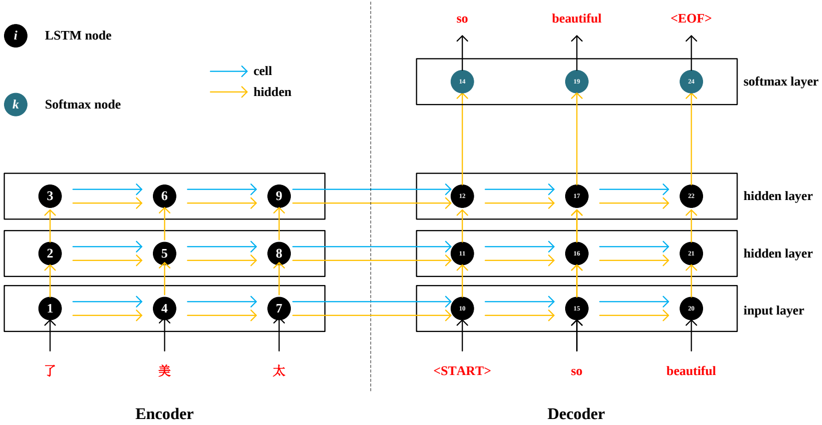
Neural Machine Translation

RNN Encoder-Decoder

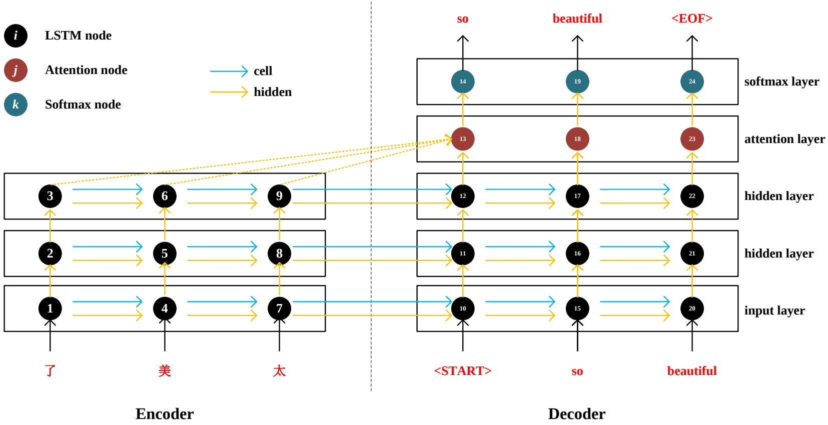
Using a multilayered LSTM to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decoder the target sequence from the vector (Sutskever et al., 2014)



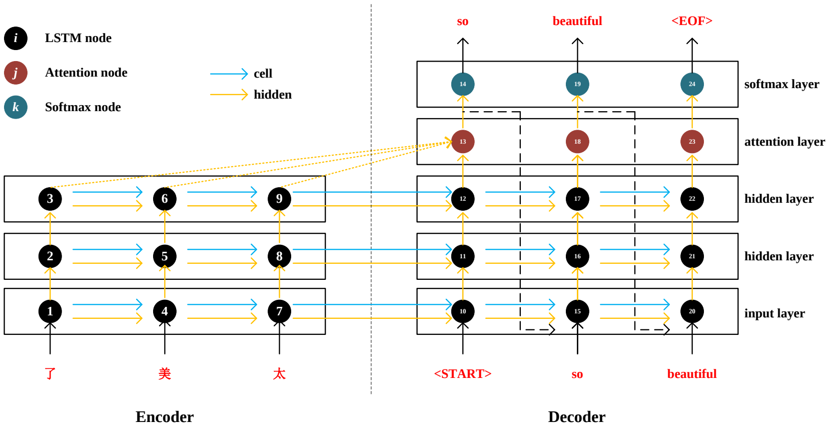
Basic Framework of Encoder-Decoder



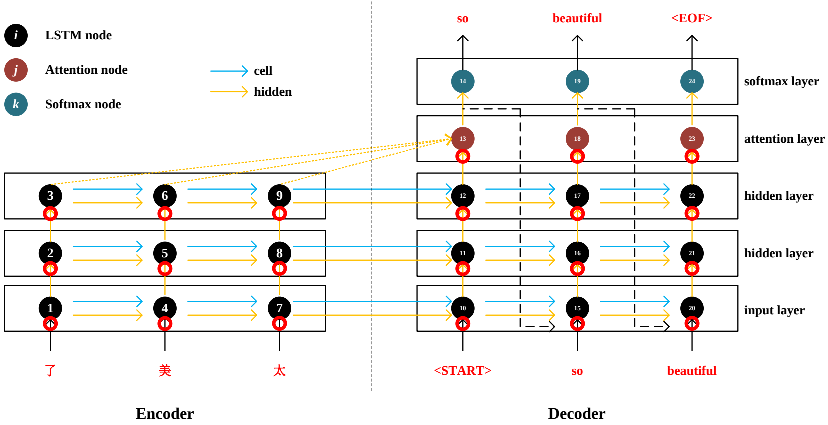
Attention model



Feed Input of Attention Model



Dropout


LSTM node
Attention node
Softmax node
cell
hidden
so
beautiful
<EOF>
softmax layer
attention layer
hidden layer
hidden layer
input layer
了
美
太
<START>
so
beautiful
Encoder
Decoder

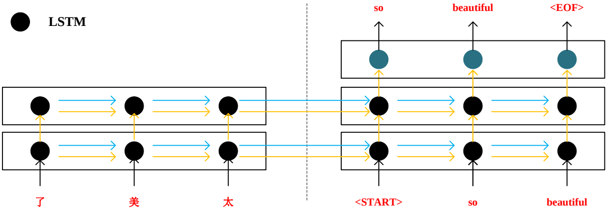
Related Papers

1. Architecture
   - Sutskever et al. 2014. Sequence to sequence learning with neural network.
   - Cho et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation.

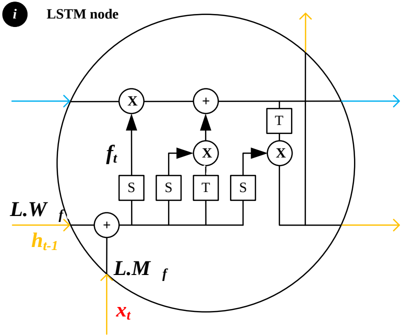
Long Short-Term Memory (LSTM)

Long Short-Term Memory

LSTM is a recurrent neural network (RNN) architecture. Unlike traditional RNNs, an LSTM network is well-suited to learn from experience to classify, process and predict time series when there are very long time lags of unknown size between important events. (from WIKIPEDIA)

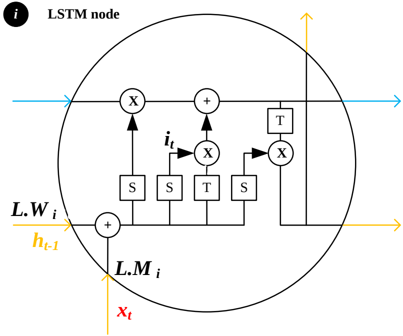


Long Short-Term Memory (LSTM)



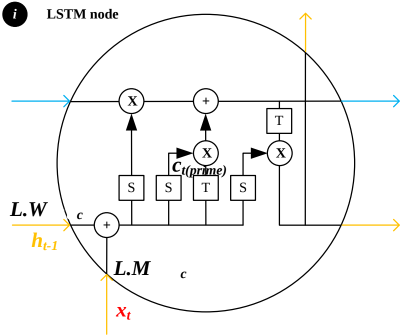
$$f_t = \text{sigmoid}(L.M_f\, x_t + L.W_f\, h_{t-1} + L.b_f)$$

Long Short-Term Memory (LSTM)



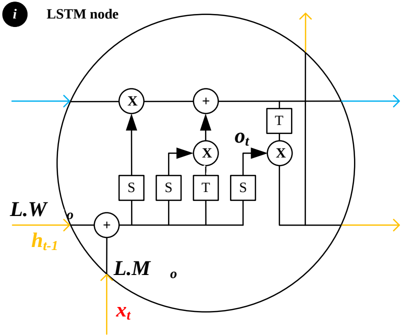
$$i_t = \text{sigmoid}(L.M_i\, x_t + L.W_i h_{t-1} + L.b_i)$$

Long Short-Term Memory (LSTM)



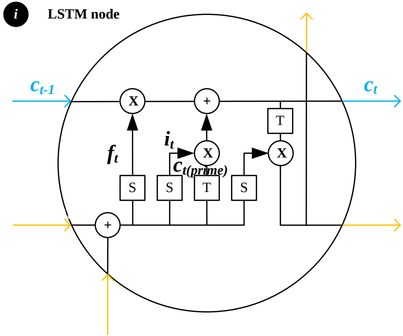
$$c_{t(prime)} = \tanh(L.M_c\, x_t + L.W_c h_{t-1} + L.b_c)$$

Long Short-Term Memory (LSTM)



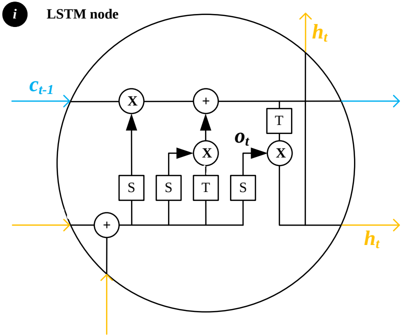
$$o_t = \text{sigmoid}(L.M_o x_t + L.W_o h_{t-1} + L.b_o)$$

Long Short-Term Memory (LSTM)



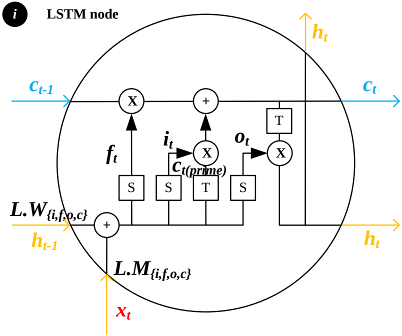
$$c_t = f_t \odot c_{t\text{-}1} + i_t \odot c_{t(prime)}$$

Long Short-Term Memory (LSTM)



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

Long Short-Term Memory (LSTM)



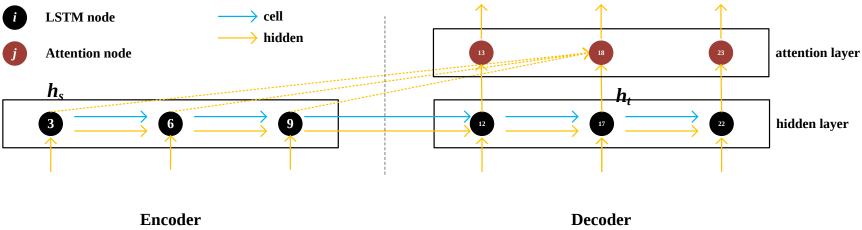
$$f_t = \text{sigmoid}(L.M_f\, x_t + L.W_f\, h_{t-1} + L.b_f)$$
$$i_t = \text{sigmoid}(L.M_i\, x_t + L.W_i h_{t-1} + L.b_i)$$
$$c_{t(prime)} = \tanh(L.M_c\, x_t + L.W_c h_{t-1} + L.b_c)$$
$$o_t = \text{sigmoid}(L.M_o x_t + L.W_o h_{t-1} + L.b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_{t(prime)}$$
$$h_t = o_t \odot \tanh(c_t)$$

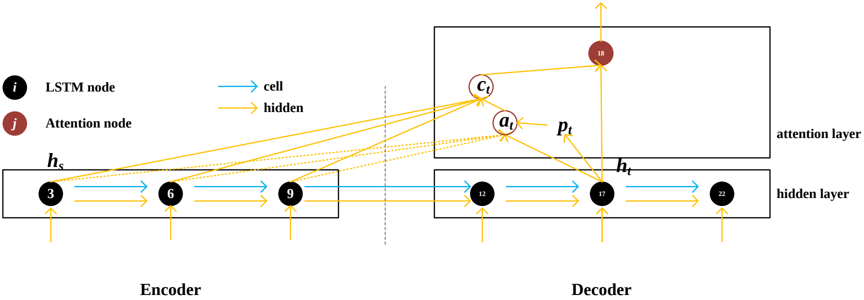
Related Papers

1. LSTM
   - Hochreiter and Schmidhuber. 1997. Long short-term memory.
   - Hochreiter and Schmidhuber. 1997. LSTM can solve hard long time lag problems.
   - Hochreiter et al. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
2. Dropout
   - Zaremba et al. 2014 Recurrent neural network regularization.

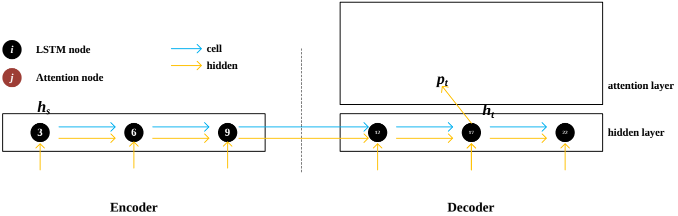
Local Attention



Local Attention



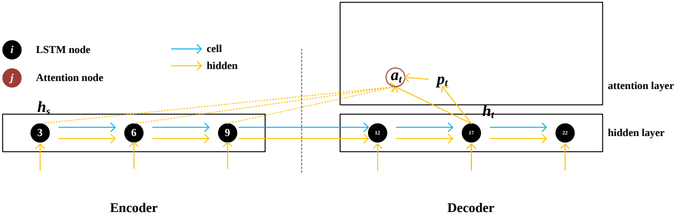
Local Attention



$$p_t = S \cdot \mathrm{sigmoid}(v_p^\top \tanh(W_p h_t))$$

- S is the source sentence length
- v_p and W_p are learned parameters

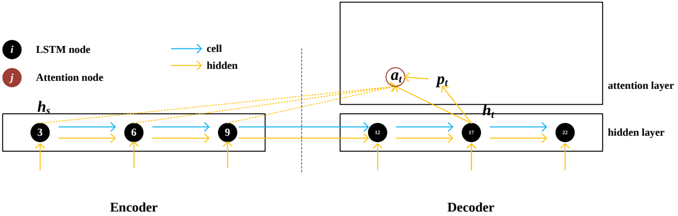
Local Attention



$$a_t(s) = \text{align}(h_t, h_s)\exp(-\frac{(s-p_t)^2}{2\sigma^2})$$

- σ is set to be $D/2$
- s is the source index for that hidden state

Local Attention

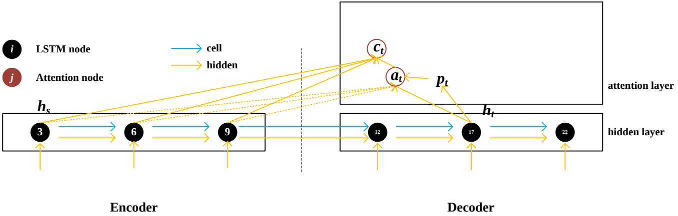


$$\mathrm{align}(h_t, h_s) = \frac{\exp(\mathrm{score}(h_t, h_s))}{\sum_{s'} \exp(\mathrm{score}(h_t, h_s'))}$$

$$\mathrm{score}(h_t, h_s) = h_t^\top W_a h_s$$

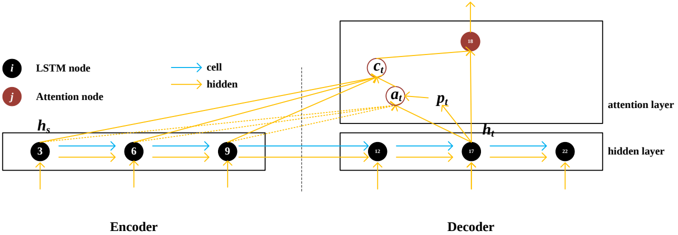
- W_a ia a learnable parameter

Local Attention



- Once all of the alignments are calculated, c_t is created by taking a weighted sum of all source hidden states multiplied by their alignment weight

Local Attention

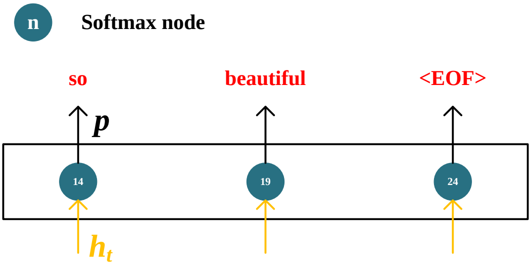


$$\tilde{h}_t = \tanh(W_{c_1} h_t + W_{c_2} c_t + b_c)$$

Related Papers

1. Attention & Feed Input
   - Bahdanau et al. 2014. Neural Machine Translation by Jointly Learning to Align and Translate.
   - Luong et al. 2015. Effective approaches to attention-based neural machine translation.

Probability Distribution with Softmax



$$p(y_t|y_{<t}, x) = \text{softmax}(W_s\tilde{h}_t + b_s)$$

Experiments on 1.8M Chinese-English News

Systems	Settings	Perplexity Train / Valid	Dev MT06	Test MT04	Test MT05	Test MT08
NiuTrans. SMT	$number, $date, $time					
	baseline	- / -	32.1	36.7	31.3	26.0
	+ online_LM	- / -	33.9 + 1.8	38.2 + 1.5	32.8 + 1.5	27.9 + 1.9
Open-source NMT	1 layer, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune	- / -	29.6	33.9	28.6	23.1
	finetune	- / -	34.3	-	-	-
	+ unk	- / -	34.9	41.1	34.3	27.1
	ensemble (4)	- / -	37.3 + 5.2	43.5 + 6.8	35.9 + 4.6	29.3 + 3.3
NiuTrans. NMT	4 layers, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m15)	9.09 / 12.15	30.4	37.5	29.6	23.5
	finetune (m14)	7.34 / 9.72	36.9 + 4.8	42.9 + 6.2	35.5 + 4.2	29.6 + 3.6
	2 layers, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m15)	7.40 / 9.41	38.3	44.2	37.2	30.2
	finetune (m15, 1.2-2, 0)	6.70 / 9.42	38.7	44.7	37.2	30.7
	+ unk	- / -	39.1 + 7.0	45.1 + 8.4	37.7 + 6.4	31.3 + 5.3
	finetune (m15, 1-2, 0.65)	6.70 / 9.42	39.1	45.8	37.9	30.8
	+ unk	- / -	**39.6**	**46.1**	**38.5**	**31.4**
	1 layer, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m15)	8.01 / 10.58	35.9	40.8	34.5	28.8
	finetune (m15)	7.56 / 10.60	36.2	41.2	34.5	28.7
	+ unk	- / -	36.7 + 4.6	41.4 + 4.7	35.0 + 3.7	29.2 + 3.2

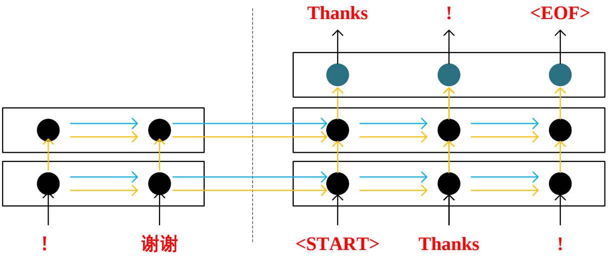
Experiments on 20M Chinese-English Oral

EXP	Beam	Length normalization	Penalty beta	File size of 1best	BLEU of test3
nn-11 old-att j+1	20	0.0	0.0	487k	52.45
		0.2	0.2	501k	53.49
		0.3	0.3	508k	54.51
		0.4	0.4	515k	54.57
		0.45	0.45	520k	54.62
		0.5	0.5	522k	54.45
		0.6	0.6	533k	53.17
		0.65	0.2	513k	54.38
	12	0.4	0.4	514k	54.47
	8			514k	54.42
	4			515k	54.22
	2			512k	54.41
	1			508k	52.42
	8	**0.45**	**0.45**	**519k**	**54.56**
		0.5	**0.5**	**521k**	**54.55**
		0.6	0.6	531k	53.55

EXP	Beam	Length normalization	Penalty beta	File size of 1best	BLEU of test3
nn-11 **new-att**	**8**	0.00	0.00	489k	52.44
		0.20	0.20	502k	53.62
		0.25	0.25	505k	54.03
		0.30	0.30	508k	54.42
		0.35	**0.35**	**511k**	**54.58**
		0.40	0.40	514k	54.41
		0.45	0.45	519k	54.47
		0.50	0.50	522k	54.31
		0.55	0.55	528k	54.00
		0.60	0.60	532k	53.62

Examples of translations produced by NMT

1. 在美国上小学要上几年?
   - SMT: In the last few years in elementary school to the United States?
   - NMT: How many years does it take to go to elementary school in America?
2. 想看看我们的新款衬衫吗?
   - SMT: Want to see our new shirts?
   - NMT: Would you like to see our new shirts?
3. 餐费有包含在内吗?
   - SMT: There are meals included?
   - NMT: Is the meal included?
4. 恐怕我们不能保证27号之后有房间给您了。
   - SMT: I'm afraid we can't guarantee 27 room after you.
   - NMT: I'm afraid we can't guarantee a room for you after 27.


Thanks
!
<EOF>
!
谢谢
<START>
Thanks
!