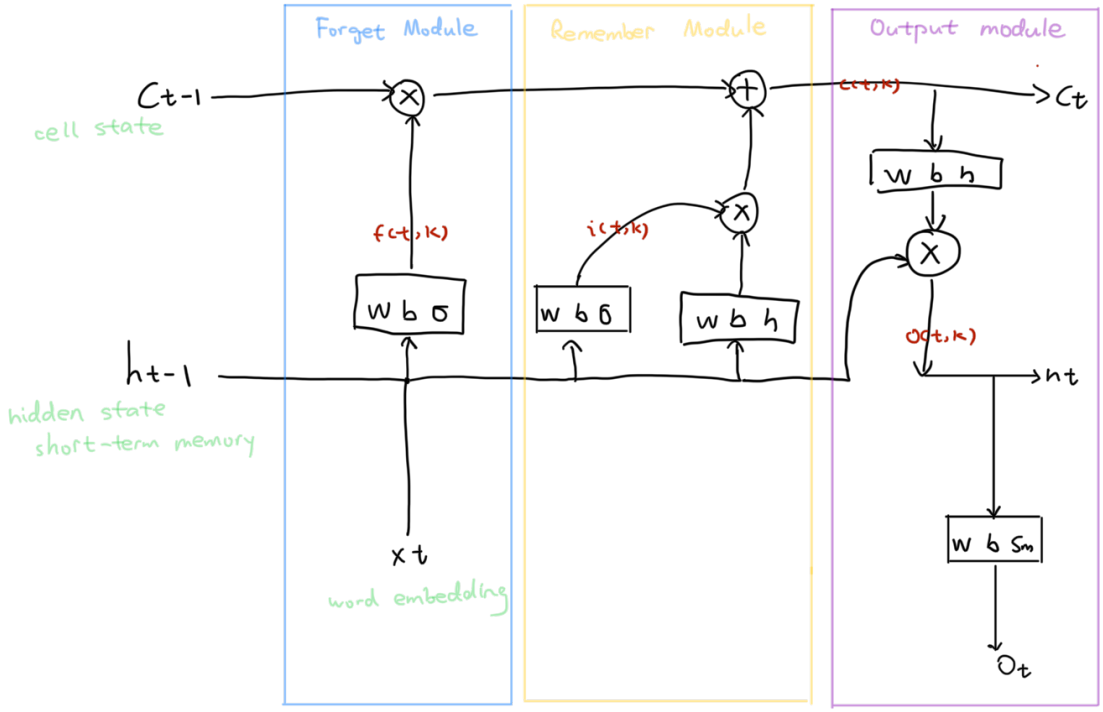
LSTM

- Plain LSTM Cell



- Input

one word at a time in the order

① document delimiter query

② query delimiter document

delimiter: |||

each document query pair is passed as a long sequence

- Architecture

input x(t)

output y(t)

$$y(t) = y'(t,1) \,||\, \cdots \,||\, y'(t,k)$$

output for entire LSTM is the concatenat of outputs from each layer

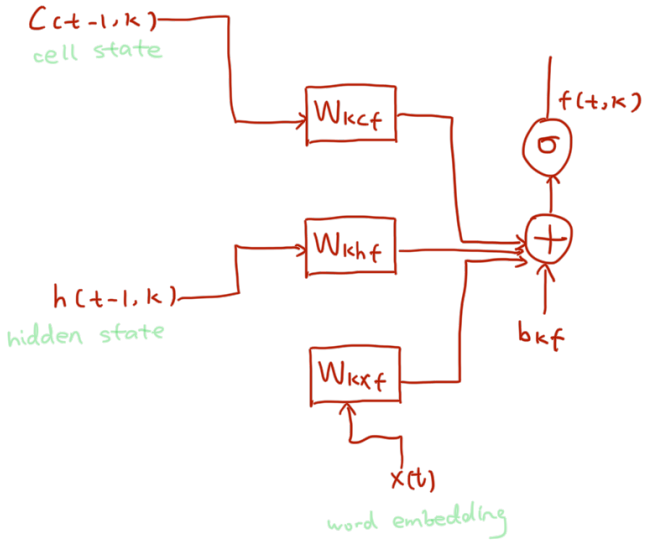
in our two layer deep LSTM Reader, output of the entire LSTM reader is the concatenation of output from e layer

$$x'(t,k) = x(t) \,||\, y'(t,k-1)$$

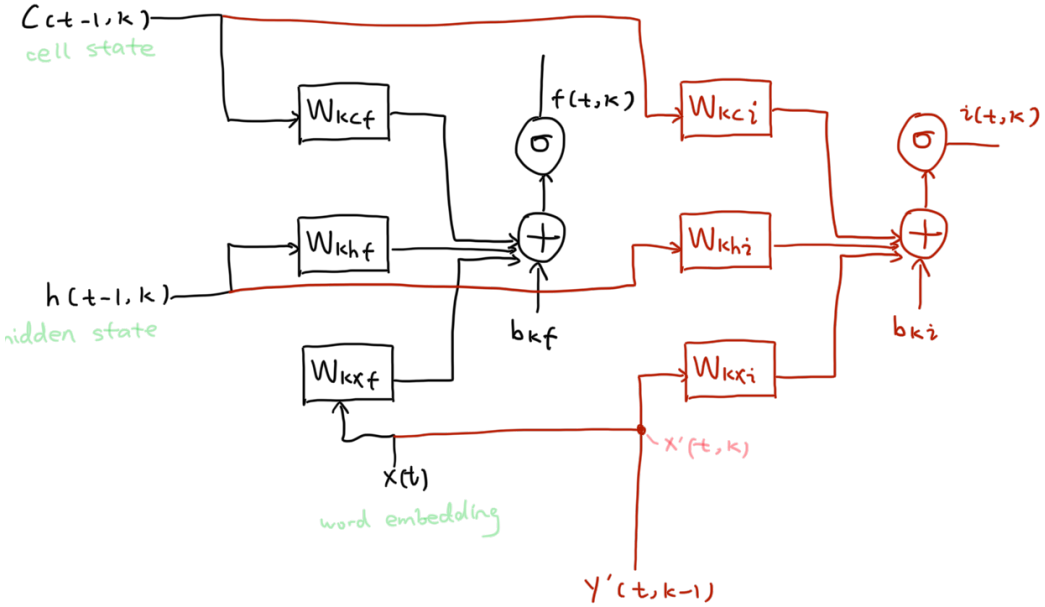
input for each hidden layer of LSTM Reader i the concatenation of original input to the LST Reader (embedded document query pair) and

the output from previous layer

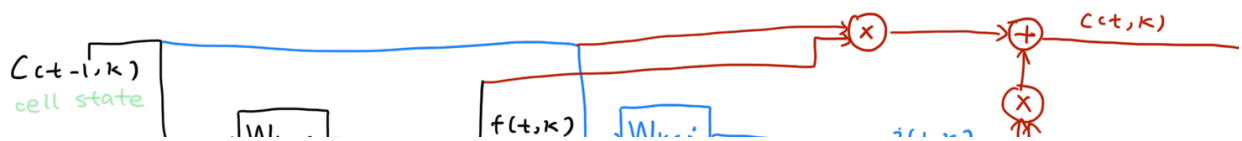
$$f(t,k) = \sigma\left(W_{kxf}\, x(t) + W_{khf}\, h(t-1,k) + W_{kcf}\, c(t-1,k) + b_{kf}\right)$$

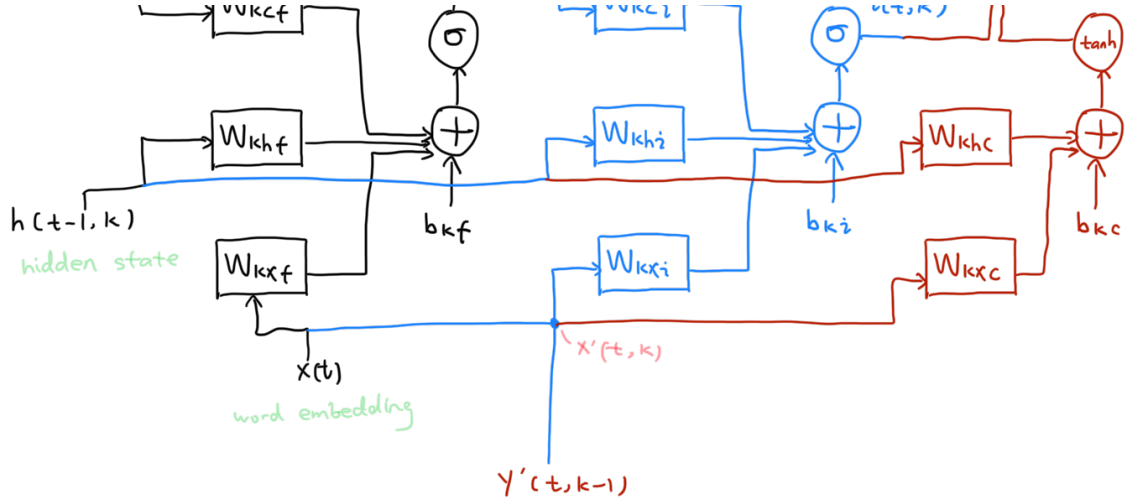


$$i(t,k) = \sigma\left(W_{kxi}\, x'(t,k) + W_{khi}\, h(t-1,k) + W_{kci}\, c(t-1,k) + b_{ki}\right)$$

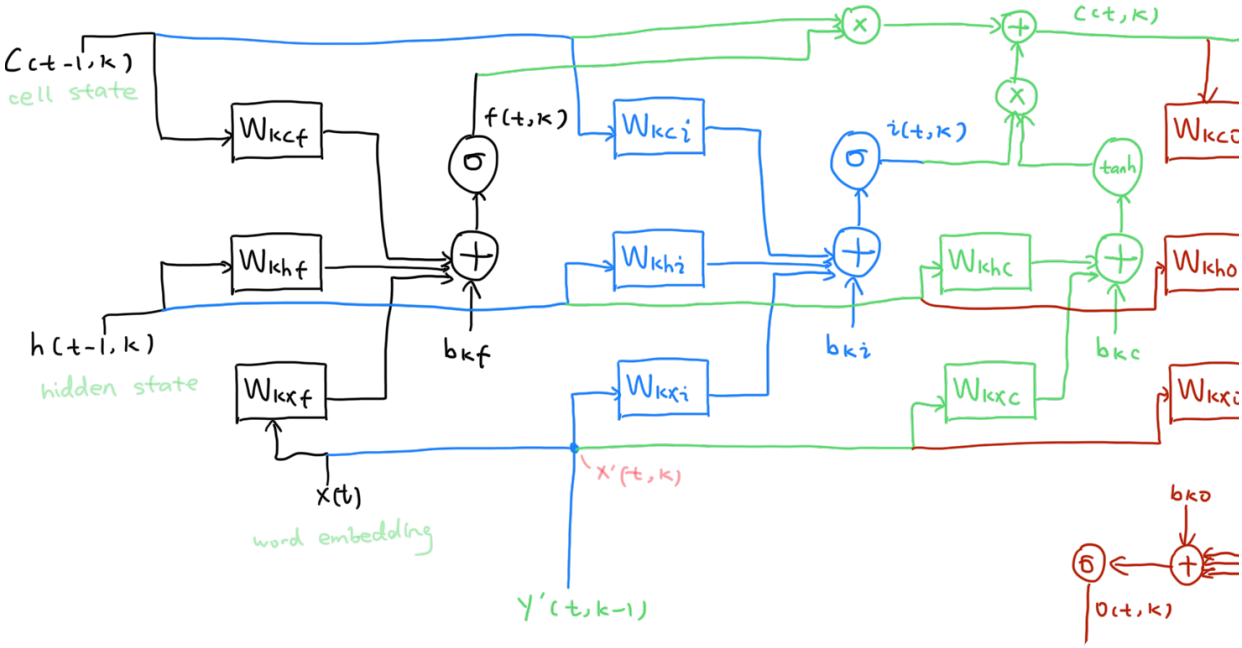


$$c(t,k) = f(t,k)\, c(t-1,k) + i(t,k) \tanh\left(W_{kxc}\, x'(t,k) + W_{khc}\, h(t-1,k) + b\right.$$

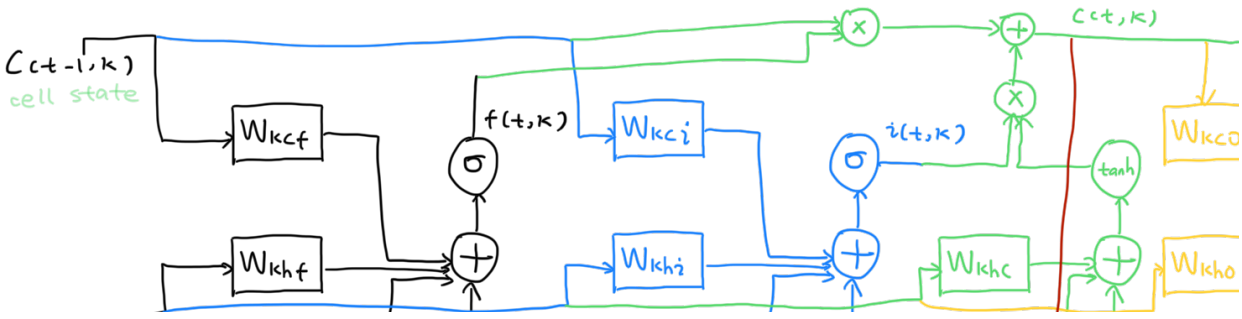


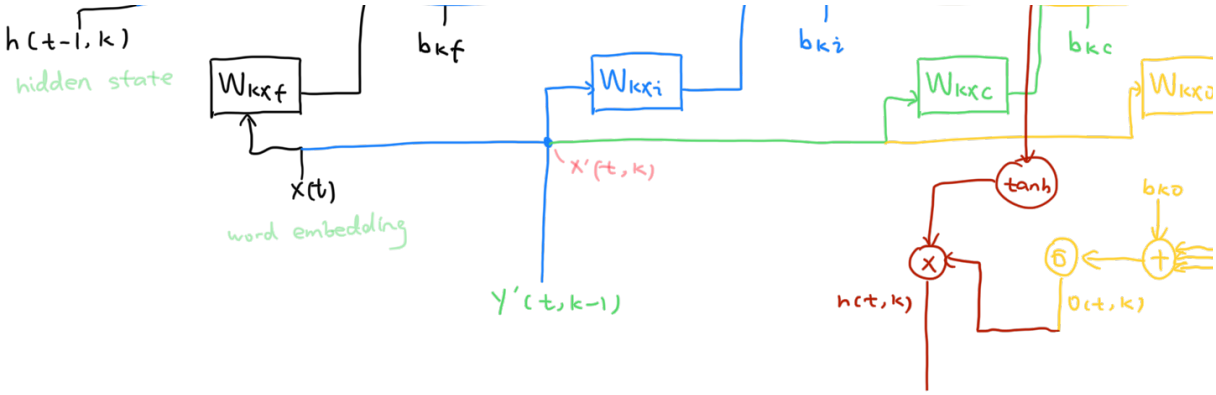


$$O(t,k) = \sigma\left(W_{kxo}\, x'(t,k) + W_{kho}\, h(t-1,k) + W_{kco}\, c(t,k) + b_{ko}\right)$$

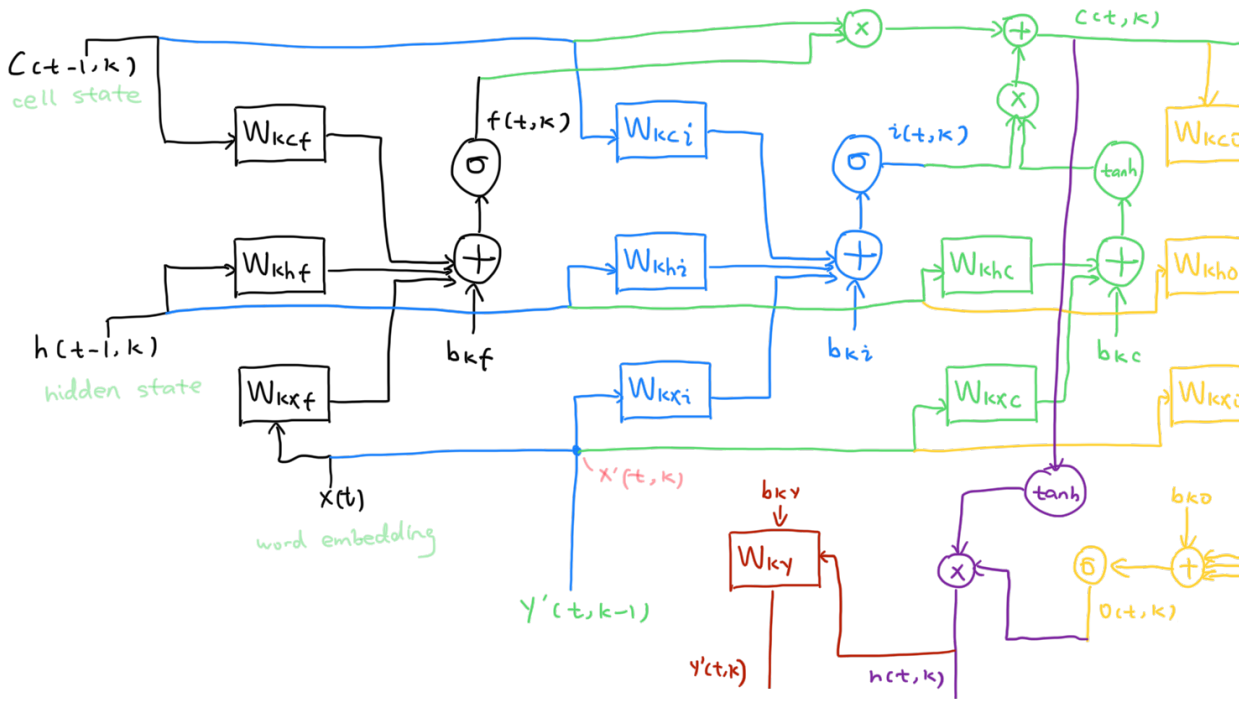


$$h(t,k) = O(t,k) \tanh\left(C(t,k)\right)$$





$$y'(t,k) = W_{ky}\, h(t,k) + b_{ky}$$



$$g^{LSTM}(d,q) = y(|d| + |q|)$$

Recall $y(t) = y'(t,1) \,||\, \cdots \,||\, y'(t,k)$

- Model Hyperparameters

* indicates optimal configuration

Hidden Size	Learning Rate	Batch Size	Dropout	dept
64	1^{-3}	16	0.0	# lay
128	$5x1^{-4}$	32	0.1	
256	1^{-4}		0.2	
	$5x1^{-5}$			

Input Form

cqa (context document, delimiter, question)

qca (question, delimiter, context document)

RMS Drop momentum 0.9 decay 0.95

- Model Summary