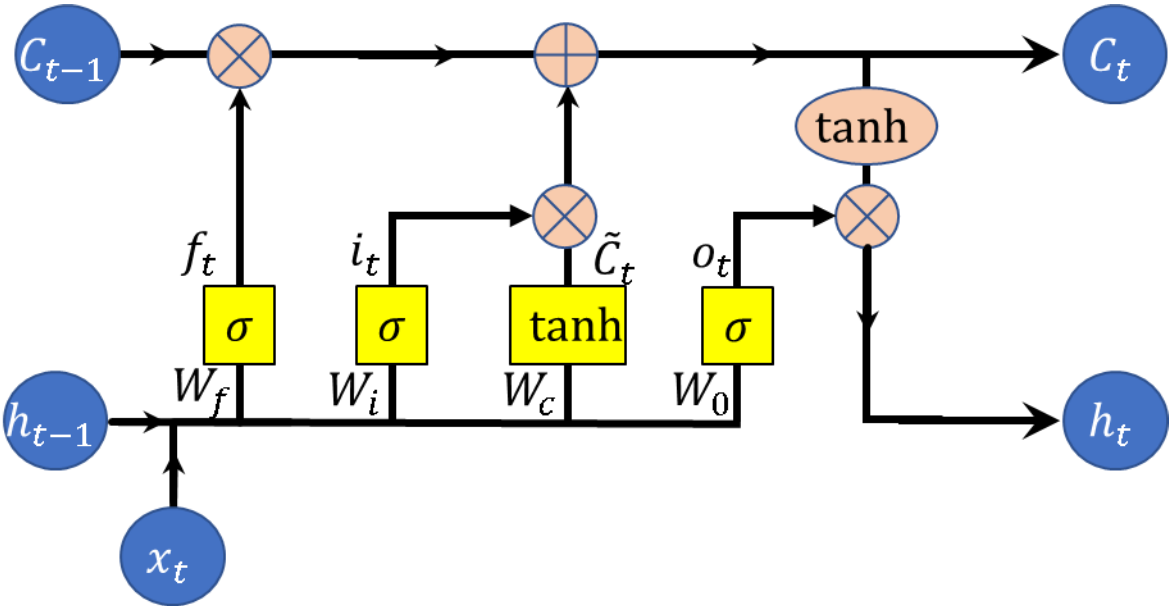
1. Explain the problem of exploding and vanishing gradients.

2. Why the sigmoid activation function is unable to prevent the vanishing gradient problem. Why is Relu activation function helpful towards mitigating the vanishing gradient problem.

3. Why are non zero-centered activation functions a problem in backpropagation?

4. Give the intuition behind input gate (selective read), forget gate (selective forget), and output gate (selective forget) in LSTMs.

5. The typical LSTM network architecture is shown below:



(In concatenating h_{t-1}+ x_t, treat h_{t-1} as the prefix and x_t as the suffix)
Given:

$$\sigma(x) = \frac{1}{1+e^{-x}};\ \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$x_t =$	0.77	$h_{t-1} =$	0.98	$C_{t-1} =$	0.57
	0.46		0.41		0.30
	0.21				

$W_f =$	0.36	0.71	0.24	0.81	0.39
	0.67	0.35	0.05	0.26	0.31

W_i	0.58	0.22	0.61	0.90	0.32
=	0.18	0.44	0.83	0.46	0.33

$W_c =$	0.61	0.16	0.82	0.53	0.21
	0.22	0.50	0.97	0.53	0.34

W_o	0.37	0.28	0.10	0.34	0.42
=	0.53	0.35	0.83	0.36	0.71

Compute $f_t, i_t, \widetilde{C}_t, o_t, C_t, h_t$