# Indian Institute of Technology Kharagpur
## Class Test III 2020-21

**Date of Examination:** 15 Mar. 2021 **Duration:** 30 minutes
**Subject No.:** CS60010 **Subject:** Deep Learning
**Department/Center/School:** CSE **Credits:** 3 **Full marks:** 20

**Instructions**

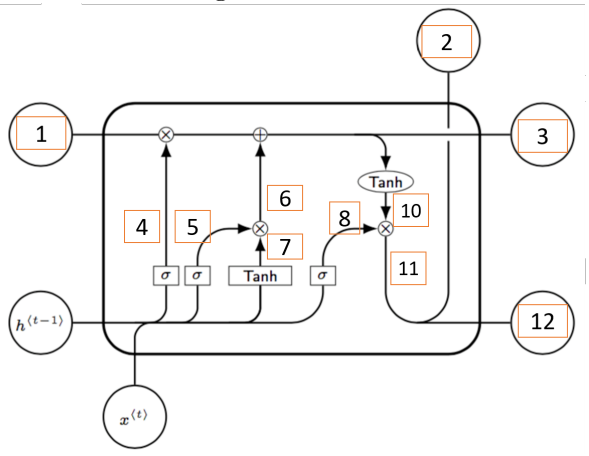
1. Please write your name, roll number, subject name and code, date and time of examination on the answer script before attempting any solution.
2. **Organize your work**, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.
3. **Mysterious or unsupported answers will not receive full marks**. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.
4. **Upload a single pdf file. Please write your name, roll number, and CS60010 CT 3, 15/3/21 on the top of the file.** You must show all your work to get credit.
5. You must answer all questions on your own. You may refer to notes and the internet but you must not take help from anyone including your classmates.

---

1. Consider a classic LSTM cell that implements the following equations:

$$
\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\
\tilde{C}_t &= \tanh\left(W_C[h_{t-1}, x_t] + b_C\right) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
h_t &= o_t \odot \tanh\left(C_t\right)
\end{aligned}
$$

A schematic diagram of a LSTM cell is given below.



(a) (3 points) Correspond the variables below to one of the positions 1 to 12 as printed in the diagram above.

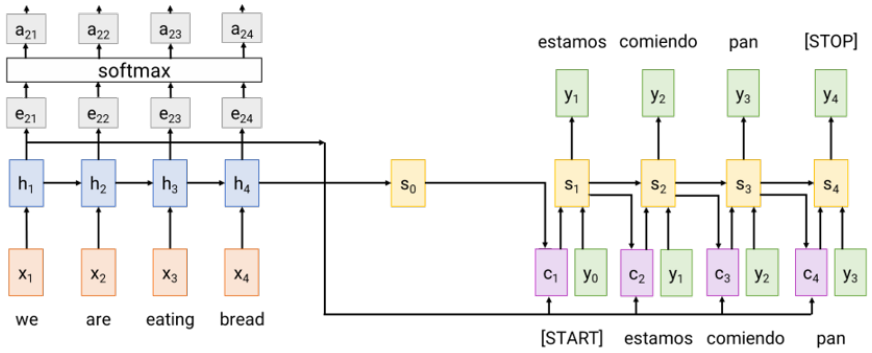
i. $f_t$
ii. $i_t$
iii. $o_t$
iv. $C_t$
v. $C_{t-1}$
vi. $\tilde{C}_t$

(b) (2 points) Suppose $C_t$ has dimension 3. Given the following values of $C_t$ and computed values of $i_t, o_t$ and $\tilde{C}_t$:

$$\begin{aligned} C_{t-1} &= \{0.2, 0.4, 0.8\} \\ f_t &= \{1, 0.5, 0\} \\ i_t &= \{1, 0, 0.5\} \\ o_t &= \{1, 0, 0\} \\ \tilde{C}_t &= \{0.5, 0, 0.1\} \end{aligned}$$

i. Compute the value of $C_t$.
ii. Show the computed value of $h_t$

2. Consider a simple encoder decoder attention model as given in the following figure:




Assume that the alignment scores $e_{ij}$ are computed based on dot-product similarity. Suppose we wish to compute $C_k$ which is the context vector for the $k$th decoder state.

(a) (1 point) Write the expression for $e_{kj}$

(b) (1 point) Compare this model to an encoder decoder model where $C = h_n$ ($h_n$ is the final hidden state of the encoder)is given as an input to each decoder cell $s_i$. How many additional number of parameters are required by the attention model considered?

3. (3 points) Consider a LSTM cell as given above. Suppose $x_t$ has dimension 2 and $c_t$ has dimension 3

(a) What is the dimension of $W_f$ ?

(b) What is the dimension of $b_f$ ?

(c) What are the total number of parameters (weight values) in this LSTM cell?

4. Consider a LSTM with input and output at every time step. Suppose that the loss function in the network is computed only at the final time step $T$ and is denoted by $L = L_T$. Let output at $T$ be given by $V_T = W_V \cdot h_T + b_V$ and $\hat{y}^T = softmax(V_T)$ Assume cross-entropy loss.

   (a) (5 points) Express $\frac{\partial L}{\partial c_t}$ in terms of $\frac{\partial L}{\partial c_{t+1}}$, $\frac{\partial L}{\partial h_t}$ and other non-recursive terms.
   Hint: Consider the backpropagation paths $L \to h_T \to c_T$ and $L \to h_{t+1} \to c_{t+1} \to c_t$

   (b) (2 points) Explain briefly how long term dependency is addressed in LSTM by considering the expression of $\frac{\partial L}{\partial c_1}$ and considering that $f_t = 1$.

5. (3 points) Consider an image of shape $200 \times 200 \times 3$. Suppose depthwise separable convolution is applied on this image where the spatial size of the filter is $4 \times 4$. The stride, padding and the number of output channels are 1, 1 and 10 respectively. Note that for depthwise separable convolution you have to take both depthwise convolution and pointwise convolution into consideration. Assume that padding is applied in 'depthwise convolution' step only. What is the total number of computation here (considering multiplication operations only)? What are the output sizes after each of the depthwise convolution and pointwise convolution operation?