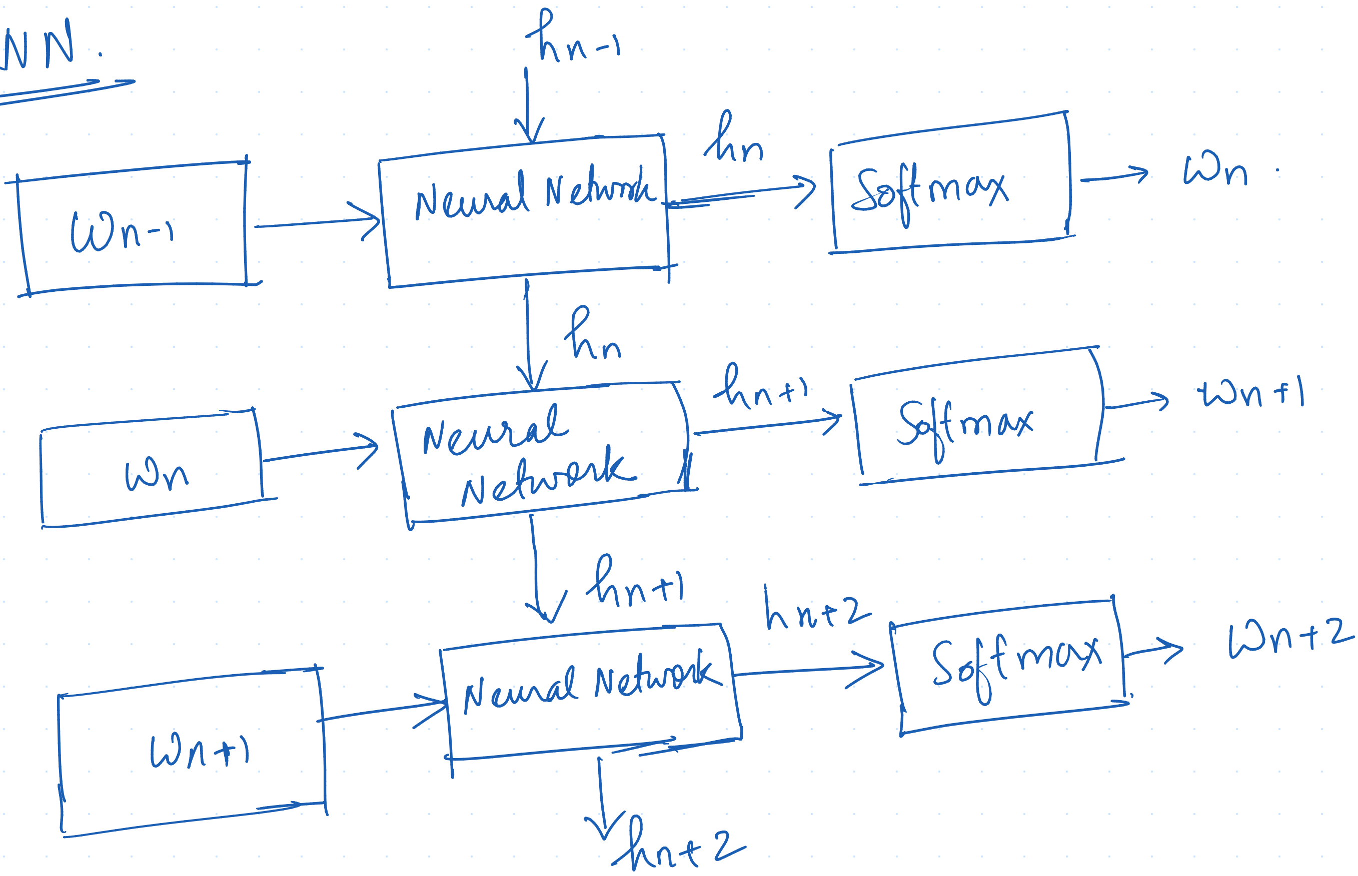
Lecture - 9

15/03/2026.

RNN.

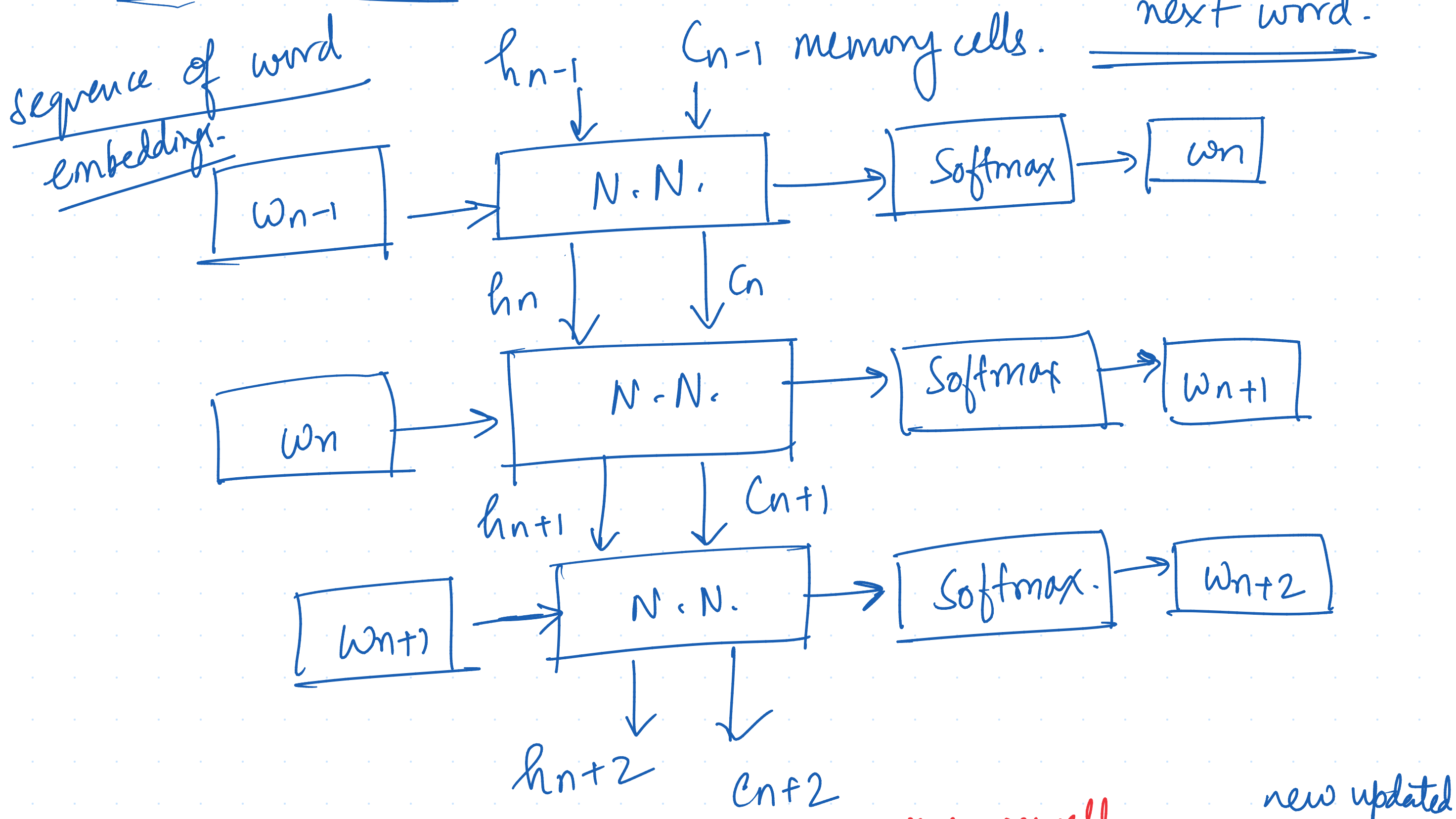


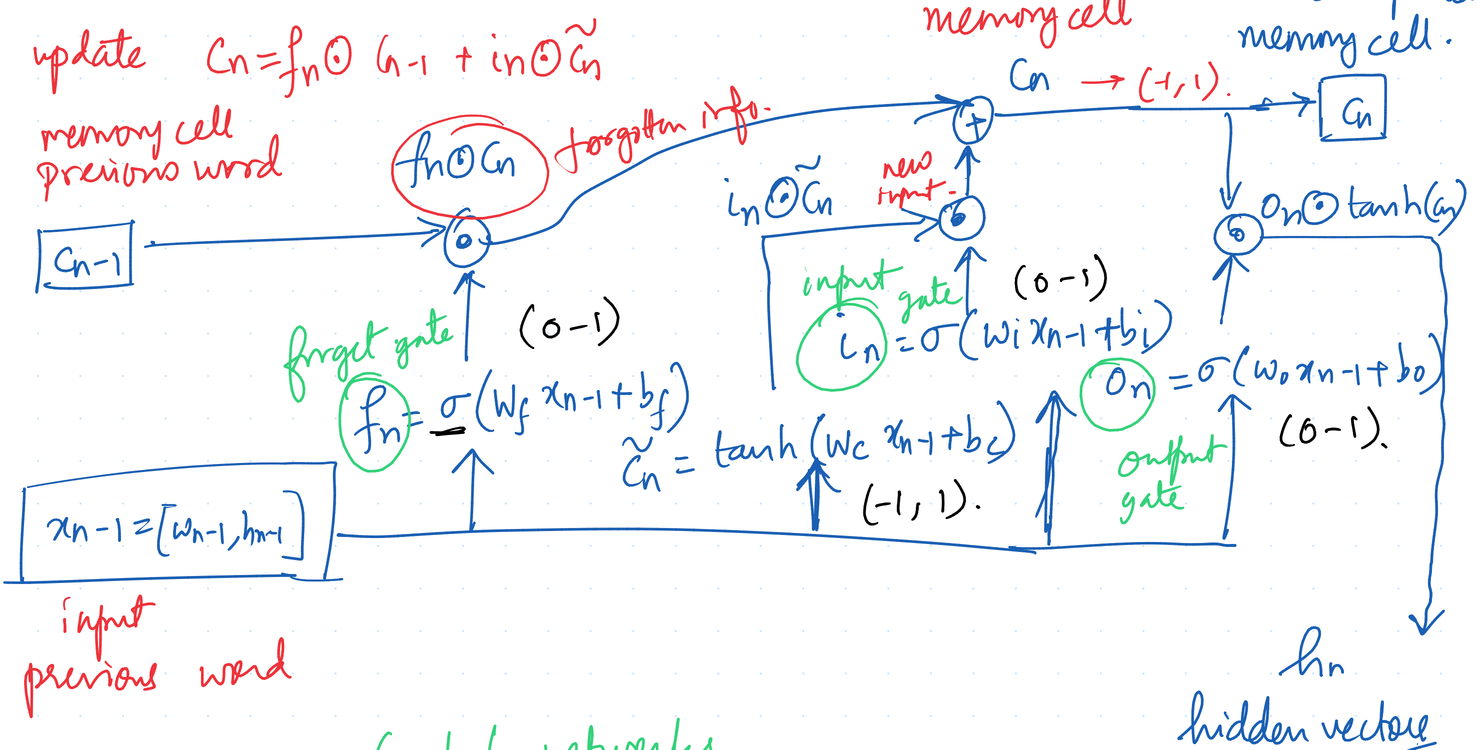
$$h_n = \tanh(W x_{n-1} + b)$$

$$p(w_n \mid w_{n-1}, h_{n-1}) = \text{Softmax}(U h_n + B).$$

$$w_1, w_2, \ldots, \quad w_{n-1}, \quad w_n$$

Long-short-term Memory





Control networks

f_n, i_n, O_n.

controls the amount of information within the models.

$$[x_1, x_2, \dots, x_n]$$

weighted multiplication $\times$

$$[0.5, 0.1, \dots 1 \dots 0.3] \rightarrow$$

higher weightage — lower — full weightage — less weightage

forget gate (0-1).

- If we want to forget some information, we multiply it with a very low value. (0.1, 0.001)
- If we want to retain some information we multiply it with some high value (1, 0.99).
- Input gate controls the amount of new data to the neural network.
- Output gate – controls the degree to which the memory cell goes to the o/p of the hidden vector.

I/p : x_t, h_{t-1}, C_{t-1} (memory cell). (LSTM)

$$g_t = \tanh(W_c x_t + U_c h_{t-1})$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$ → forget gate.

$$i_t = \sigma(W_i x_t + U_i h_{t-1})$$ i/p gate

$$C_t = f_t C_{t-1} + i_t g_t$$

$$O_t = \sigma(W_o x_t + U_o h_{t-1})$$ o/p gate -

$$h_t = O_t \tanh(C_t)$$

GRU

Input : x_t, h_{t-1}. : no memory cell.

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$ → update gate

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$ → reset gate.

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}))$$

$$h_t = (1 - z_t) h_{t-1} + z_t \tilde{h}_t$$

Different types of attention Mechanism :-

Additive attention

$$a(q, k) = w_v^T \tanh(W_q \cdot q + W_k \cdot k).$$

General

$$a(q, k) = q^T W_a k$$

Dot-product attention

$$a(q, k) = q^T k.$$

Scaled-dot product attention

$$a(q, k) = \frac{q^T k}{\sqrt{d_k}}$$ → softmax in transformers. to lie b/w (0-1)

Transformer.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Hard attention : non-differentiable / gradient estimation (RL)

Soft Attention : differentiable / softmax based.