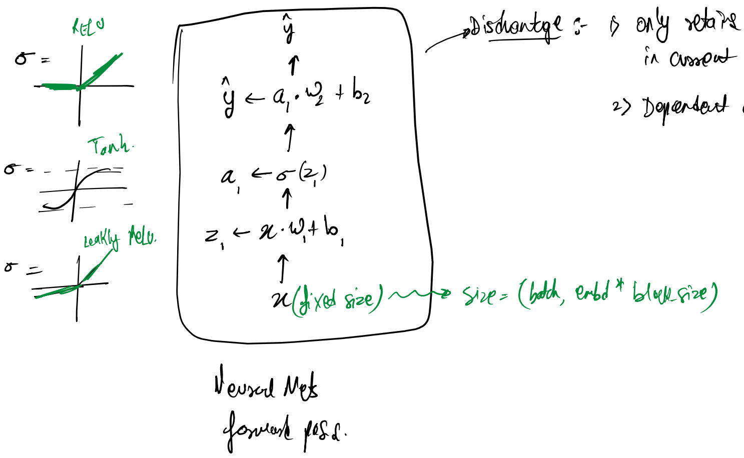
1) RNN Language Model

10 December 2025 01:34 PM

❐ **Architecture of a traditional RNN** — Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. They are typically as follows:

Advantages	Drawbacks
• Possibility of processing input of any length • Model size not increasing with size of input • Computation takes into account historical information • Weights are shared across time	• Computation being slow • Difficulty of accessing information from a long time ago • Cannot consider any future input for the current state

MULTI-LAYER PERCEPTRON

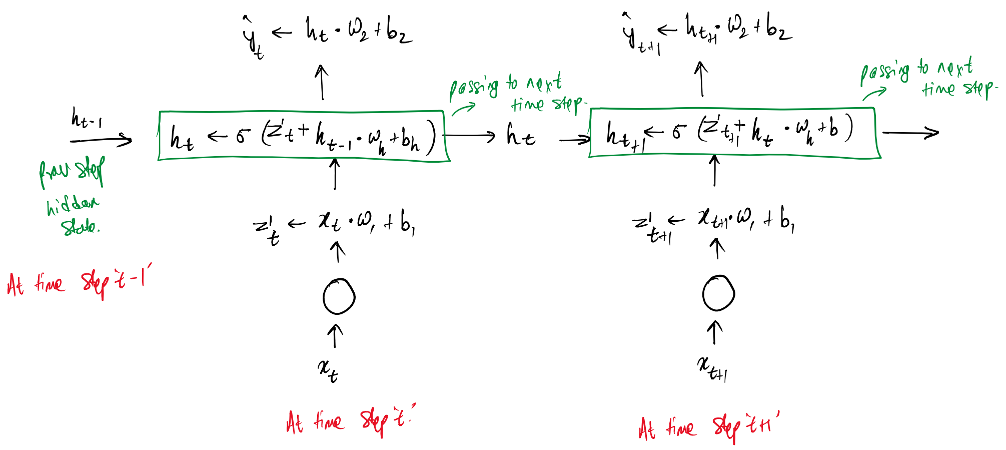


Disadvantage :- 1) only retains the information in current block size.
2) Dependent on input size.

#Solution :- Recurrent Neural Network (RNN).

#RNN. → Transfer everything you learned to next hidden states time step

RNN- Architecture



Basically we have a funcn, performing

$y_t \leftarrow f$ (current_input, prev_hidden_state)

$\Rightarrow \quad \hat{y}_t \leftarrow f(x_t, h_{t-1})$

$\hat{y}_t \leftarrow (\sigma(x_t \cdot \omega_1 + h_{t-1} \cdot \omega_h + b)) \cdot \omega_2 + b_2$ #forward pass for RNN.

$$\boxed{f(x_t, h_{t-1}) = (\sigma(x_t \cdot \omega_1 + h_{t-1} \cdot \omega_h + b)) \cdot \omega_2 + b_2}$$

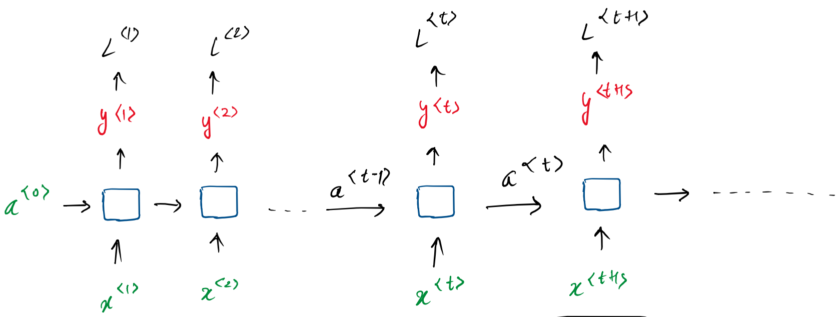
$$\boxed{f(x_t, h_{t-1}) = \left(\sigma\left(x_t \cdot \omega_1 + h_{t-1} \cdot \omega_h + b\right)\right) \cdot \omega_2 + b_2}$$

Activation function

information obtained from past hidden state (past time step)

Controls how much the past hidden state is important.

Generally



where,

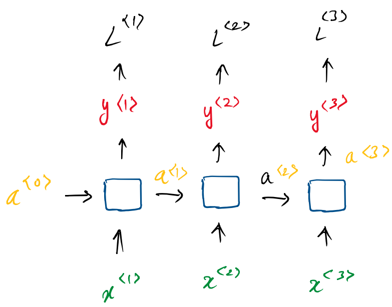
$$a^{\langle t \rangle} = g_1\left(x^{\langle t \rangle} \cdot \omega_x + a^{\langle t-1 \rangle} \cdot \omega_a + b\right)$$

$$y^{\langle t \rangle} = g_2\left(a^{\langle t \rangle} \cdot \omega_y + b_y\right)$$

$$L(\hat{y}, y) = \sum_{i=0}^{t} L^{\langle i \rangle} \; ; \quad L^{\langle t \rangle} = L(y^{\langle t \rangle}, \hat{y})$$

Training RNNs (Backpropagation)

Assume this Architecture.



$$L = L^{\langle 1 \rangle} + L^{\langle 2 \rangle} + L^{\langle 3 \rangle}$$

$$a^{\langle 1 \rangle} = x^{\langle 1 \rangle} \cdot \omega_x + a^{\langle 0 \rangle} \cdot \omega_h + b$$

$$a^{\langle 2 \rangle} = x^{\langle 2 \rangle} \cdot \omega_x + a^{\langle 1 \rangle} \cdot \omega_h + b$$

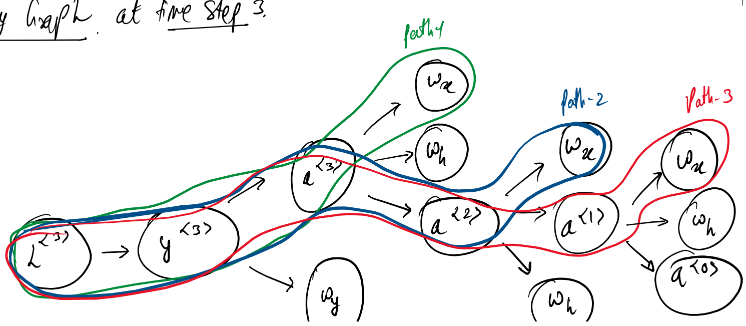
$$a^{\langle 3 \rangle} = x^{\langle 3 \rangle} \cdot \omega_x + a^{\langle 2 \rangle} \cdot \omega_h + b$$

$$y^{\langle 1 \rangle} = a^{\langle 1 \rangle} \cdot \omega_y + b_y$$

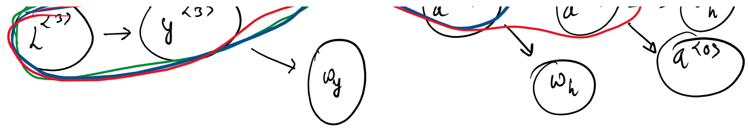
$$y^{\langle 2 \rangle} = a^{\langle 2 \rangle} \cdot \omega_y + b_y$$

$$y^{\langle 3 \rangle} = a^{\langle 3 \rangle} \cdot \omega_y + b_y$$

Dependency Graph at time step 3.



→ This tells us how $a^{\langle 3 \rangle} \rightarrow a^{\langle 2 \rangle} \rightarrow a^{\langle 1 \rangle}$ at time step = 3



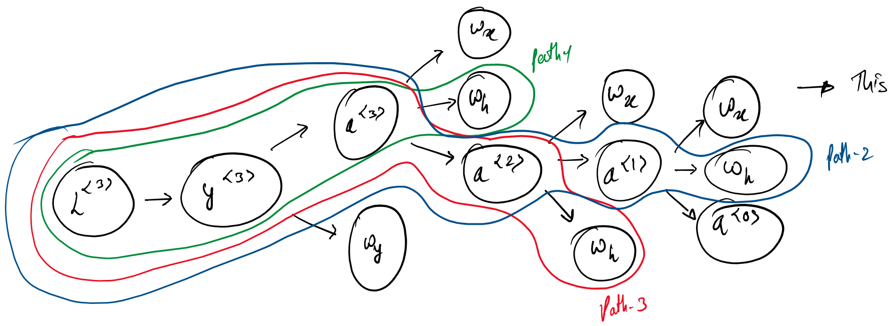
Gradients:-

$$\frac{\partial L^{<3>}}{\partial w_x} = \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial w_x}\right)}_{\text{path-1}} + \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial a^{<2>}} \times \frac{\partial a^{<2>}}{\partial w_x}\right)}_{\text{path-2}} + \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial a^{<2>}} \times \frac{\partial a^{<2>}}{\partial a^{<1>}} \times \frac{\partial a^{<1>}}{\partial w_x}\right)}_{\text{path-3}}$$

$$\boxed{\frac{\partial L^{<T>}}{\partial w_h} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial y^{<t>}} \times \frac{\partial y^{<t>}}{\partial a^{<t>}} \times \frac{\partial a^{<3>}}{\partial w_h}}$$

Backpropagation Through Time (BPTT)

Dependency Graph at time step 3



→ This tells us how $a^{<3>} \to a^{<2>} \to a^{<1>}$

at time step = 3

Gradients:-

$$\frac{\partial L^{<3>}}{\partial w} = \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial w_h}\right)}_{\text{path-1}} + \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial a^{<2>}} \times \frac{\partial a^{<2>}}{\partial w_h}\right)}_{\text{path-3}} + \underbrace{\left(\frac{\partial L^{<3>}}{\partial y^{<3>}} \times \frac{\partial y^{<3>}}{\partial a^{<3>}} \times \frac{\partial a^{<3>}}{\partial a^{<2>}} \times \frac{\partial a^{<2>}}{\partial a^{<1>}} \times \frac{\partial a^{<1>}}{\partial w_h}\right)}_{\text{path-2}}$$

$$\boxed{\frac{\partial L^{<T>}}{\partial w_h} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial y^{<t>}} \times \frac{\partial y^{<t>}}{\partial a^{<t>}} \times \frac{\partial a^{<3>}}{\partial w_h}}$$

Backpropagation Through Time (BPTT)

→ Again Same General formula.

Generalized BPTT

$$\boxed{\frac{\partial L^{<T>}}{\partial w_h} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial y^{<t>}} \times \frac{\partial y^{<t>}}{\partial a^{<t>}} \times \frac{\partial a^{<3>}}{\partial w_h}}$$