Attention - Tutorial

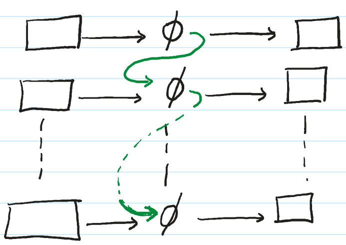
Monday, April 21, 2025 3:57 AM

Problem with RNN

formula $h_t = \phi(W_h h_{t-1} + W_x x_t + b)$

some Activation function

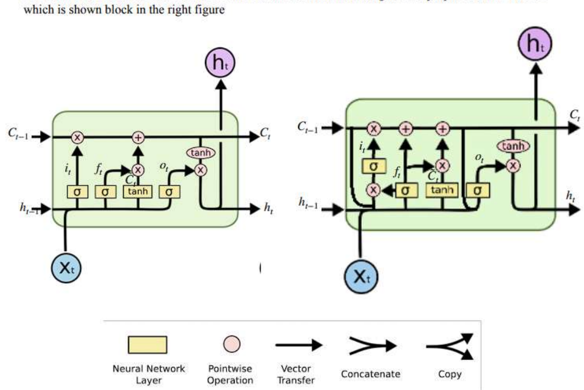
At time t, take input x_t & previous hidden state h_{t-1} to produce h_t



During Backpropogation you keep multiplying numbers that < 1, gradient $\longrightarrow 0$

('Vanishing Gradient Problem')

Recurrent Neural Network
A standard LSTM as discussed in the lectures is shown below in left figure. We proposed a new LSTM which is shown block in the right figure



Write the equations gate ($\tilde{C}_t$), input gate (i_t), output gate (o_t), forget gate (f_t), cell update (C_t) and state update (h_t)

LSTMS

At t, given

1. $x_t \in \mathbb{R}^D$
2. $h_{t-1} \in \mathbb{R}^h$

Compute

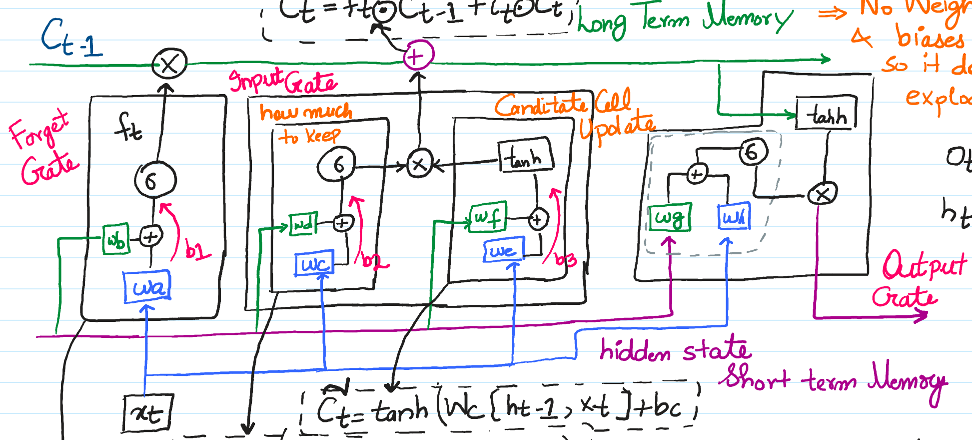
1. Forget Gate :- $f_t \in (0,1)^h$
2. Input Gate :- $i_t \in (0,1)^h$
3. Canditate Gate :- $\tilde{C}_t \in (-1,+1)^h$
4. Output Gate :- $O_t \in (0,1)^h$
5. Hidden State h_t

$\boxed{C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t}$ Cell state Long Term Memory $\Rightarrow$ No Weights & biases here

C_{t-1}

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Cell state
Long Term Memory ⇒ No Weights & biases here so it doesn't explode vanish



$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

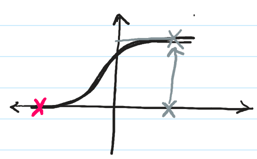
$$h_t = O_t \odot \tanh(C_t)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

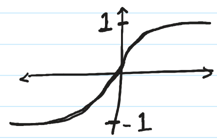
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

$$\sigma = \frac{e^x}{e^x + 1}$$



$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

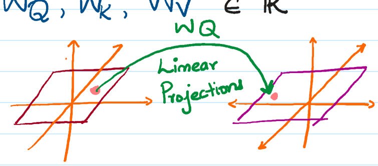


We cannot parallelize an LSTM
Very long Range dependencies fade

Attention

① $X \in \mathbb{R}^{n \times d}$ → Model dimension

② $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$



③ $Q = W_Q X \quad K = X W_K \quad V = X W_V$

$Q, K, V \in \mathbb{R}^{n \times d_k}$

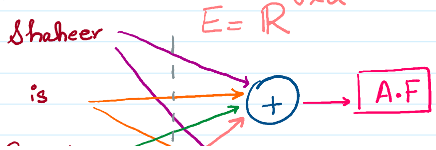
Matrix Multiplication Rules

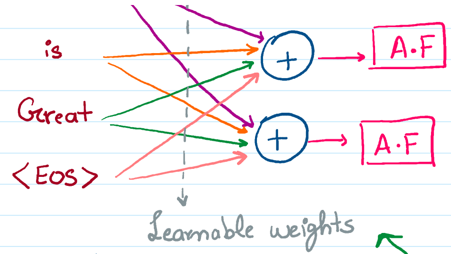
because $n \times d \quad d \times d_k$ → $n \times d_k$

Word Embedding

ONE HOT Encoding

$E = \mathbb{R}^{V \times d}$ ← size of Vocab





Inputs :-

$$\underbrace{\begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\end{bmatrix}}_{\text{Use for Training to learn}}$$

Positional Embedding

One Way is Sinusoidal

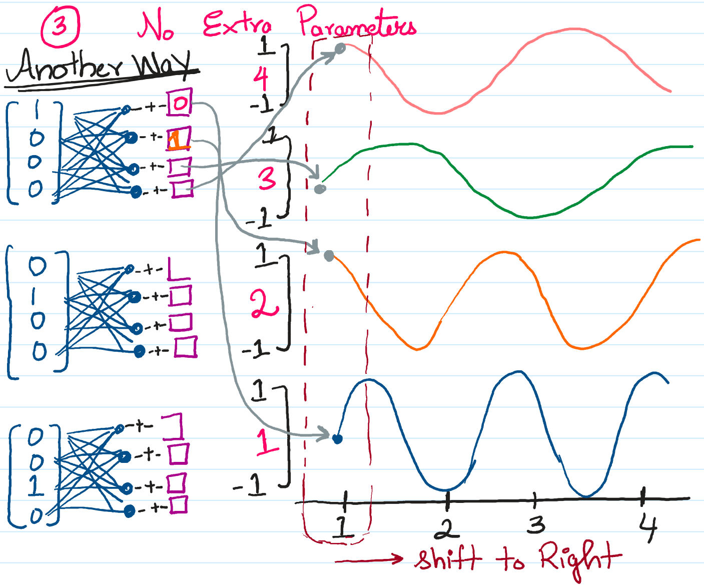
For Positions $t = 0, 1, 2 \dots n-1$
& dimension $d = 0, 1, 2 \dots d_{model} - 1$

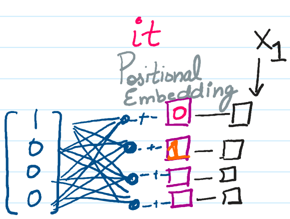
$$PE_{t,i} = \begin{cases} \sin\left(\dfrac{t}{10000^{2i/d_{model}}}\right); & i \text{ even} \\ \cos\left(\dfrac{t}{10000^{2(i-1)/d_{model}}}\right); & i \text{ odd} \end{cases}$$

$$PE \in \mathbb{R}^{n \times d}$$

1. Unique Patterns
2. Smooth functions
3. No Extra Parameters

Another Way

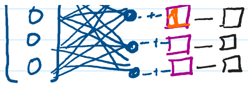




$\boxed{Q_1 = W_{Q_1} X_1}$ $K_2 = X_1 W_{K_1}$ $\boxed{V_2 = X_1 W_{V_1}}$

Step 1

Cosine Sim (Q_1, K_1)



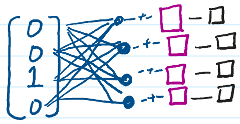
Step1

Cosine Sim (Q_1, K_1)

Cosine Sim. (Q_1, K_2)

Step 2

Cookies X_2

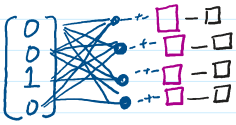


$Q_2 = W_{Q_2} X_2$ $\boxed{K_2 = X_2 W_{K_2}}$ $\boxed{V_2 = X_2 W_{V_2}}$

Cosine Sim. (Q_1, K_3)

step 3

Oven X_2



$Q_3 = W_{Q_3} X_3$ $\boxed{K_3 = X_3 W_{K_3}}$ $\boxed{V_3 = X_3 W_{V_3}}$

Apply Softmax

1.00 0.00 0.00

Someone took some cookies out of the oven & $\boxed{\text{it}}$ smelt so Yum

$\boxed{\text{Self attention}}$

'it' & 'cookies' should have a higher cosine similarity

$$S = QK^T$$

$$\hat{S} = \frac{S}{\sqrt{d_K}}$$

$$A_{ij} = \frac{e^{\hat{S}_{ij}}}{\sum \exp(S_{ij})}$$

$$Z = AV$$

$$Z_i = \sum_j A_{ij} V$$

Diffusion Models

① Forward Process

Variance

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t ; \sqrt{\alpha_t} x_{t-1}, \beta_t I)$$

where $\alpha_t = 1 - \beta_t$

Base Case :-

$$q(x_1 | x_0) = \mathcal{N}(x_1 ; \sqrt{\alpha_1} x_0, \beta_1 I)$$

Where does $q(x_t | x_0)$

$$q(x_{t-1} | x_0) = \mathcal{N}(x_{t-1} ; \sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1}) I)$$

$t-1$

$$q(x_{t-1} \mid x_0) = \mathcal{N}\left(x_{t-1} ; \sqrt{\bar{\alpha}_{t-1}}\, x_0 , (1-\bar{\alpha}_{t-1}) I\right)$$

where
$$\bar{\alpha}_{t-1} = \prod_{s=1}^{t-1} \alpha_s$$
$$= \alpha_1 \alpha_2 \cdots \alpha_{t-1}$$