Assignment - 2

1) Perceptron calculates: $y = wx + b$
which is a straight line
and XOR has output 0, 1.
Hence, No straight line can separate 0 and 1.

Multilayer has hidden layer and activation functions. As these functions are non linear, multiple layers allow non linear decision making.

2) $y = f_1(f_2 x) = (f_1 f_2)\, x$ → One linear transformation.

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a_n} \times \frac{\partial a_n}{\partial a_{n-1}} \times \cdots \frac{\partial a_1}{\partial w}$$

During backpropogation, as multiple gradient terms get multiplied continuously, it leads to exponential decay. Eg: Sigmoid outputs value b/w 0 and 1 So, it leads to vanishing gradient problem.

ReLU has no saturation, it gradient is equal to 1

3) Positional Embedding is necessary to keep the order of tokens correct which will make semantic meaning of sentence.

Sinusoidal PE has a fixed formula

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

Absolute PE has fixed embeddings

DATE PAGE

4) Query = what I am looking for
Key = what I offer
Value = what is actual content

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

→ Token is most similar to itself

5.) Each head gives attention to different relations.

Multi head gives more broader and richer representation.

d_{model} = Embedding size
h = number of heads
$d_{head} = \frac{d_{model}}{h}$

6) Greedy decoding chooses locally best token.

→ Beam search keeps top-K partial sequences
→ It explores multiple futures.

Eg: let A → probability = 0.7
B → probability = 0.3

A → End = 0.1 , Greedy = 0.07
B → End = 0.9 , Beam search = 0.27

Greedy fails, Beam search succeeds

Coding

DATE PAGE

1)

(a) $d\_head = \frac{d\_model}{h}$

$= \frac{768}{12} = \underline{64}$

(b) Total parameters:

Parameters $= 768 \times 768$

$= \underline{589,824}$

For Q, K, V $= 3 \times 589,824$

$= \underline{1,769,472}$

2) attention scores $= [2.0, 1.0, 0.0]$

$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Total $= e^2 + e^1 + e^0 = 11.107$

Normalise: $\left[\frac{e^2}{11.107}, \frac{e^1}{11.107}, \frac{e^0}{11.107}\right]$

$= [0.665, 0.245, 0.09]$