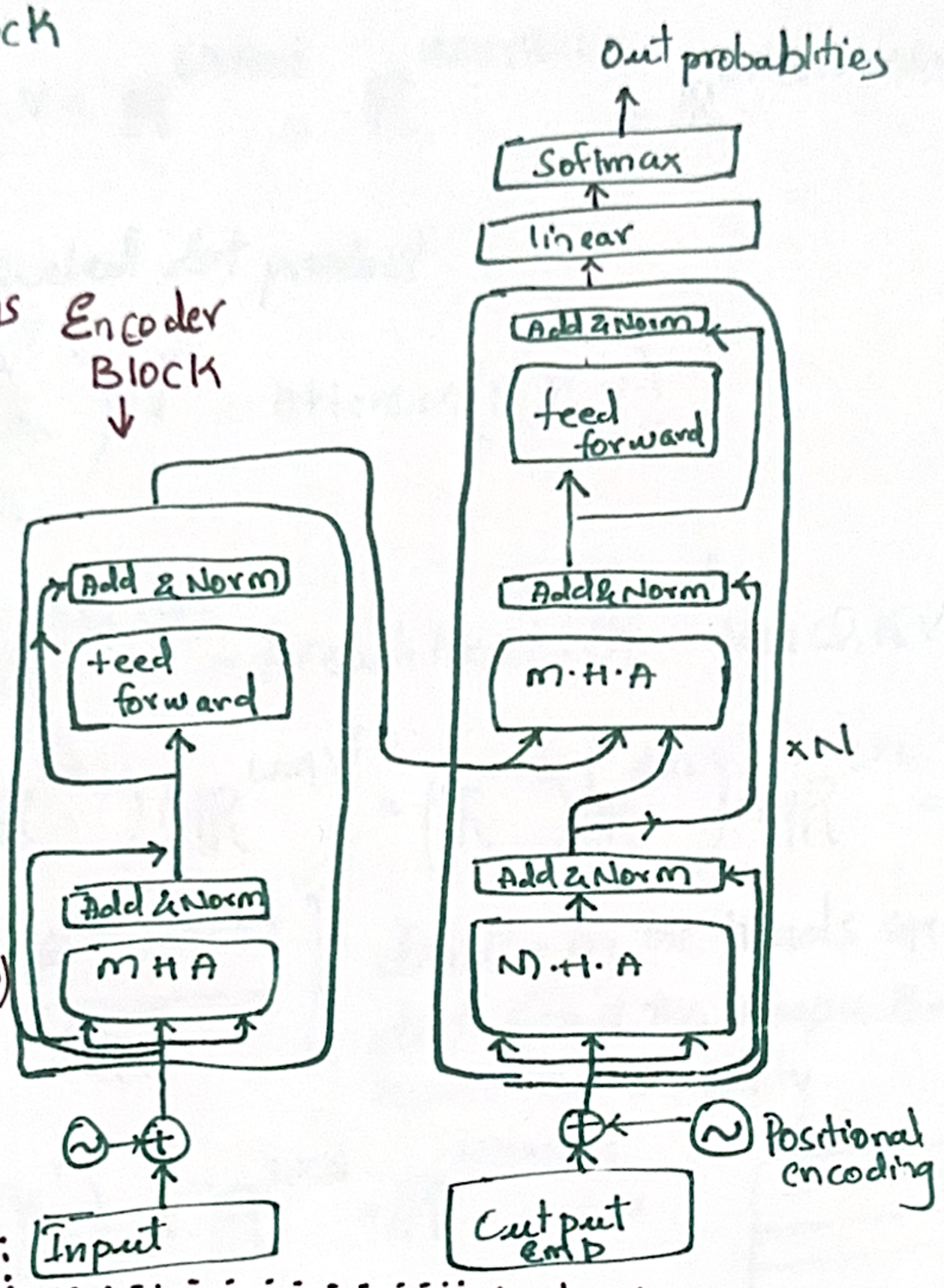
Transformer Block

—

Encoder Block

→ Positional Embeddings
→ Input embeddings
→ M·H·A
→ Projection After M·H·A
→ Add & Norm
(Residual + layer Norm) After Attn
→ feed forward
(two linear layers)
N X →
→ Add & Norm
(Residual + layer Norm After feed forward)

→ Repeat N times
Since Each encoder is N times



Input & positional embeddings.

{ Sequence length = 1024 = d_L
Embd dim = 512 = d_{model}
attention heads = 8 = # heads.
head dim = 512/8 = (64) = $\frac{d_{model}}{heads}$
$d_k = \frac{d_{model}}{heads} = \frac{512}{8}$ } Inputs = $\mathbb{R}^{d_L \times d_{model}}$

Input embedding = $\mathbb{R}^{L \times d_{model}} = \begin{bmatrix} 1024 \times 512 \\ R \end{bmatrix}$
⊕ positional encoding = $\mathbb{R}^{L \times d_{model}} = \begin{bmatrix} 1024 \times 512 \\ \mathbb{R} \end{bmatrix}$

X = Input to multi head attention $\Rightarrow \begin{bmatrix} 1024 \times 512 \\ \mathbb{R} \end{bmatrix}$

Now taking the input & projecting them $X = \mathbb{R}^{L \times d_{model}}$

$Q = X W_Q \quad [W_Q \; W_K \; W_V] =$ learned weights $\in \mathbb{R}^{d_{model} \times d_K}$

$K = X W_K \quad [W_i] = \mathbb{R}^{d_{model} \times d_K}$ • multihead = $d_K = \frac{d_{model}}{no\ of\ head}$

$V = X W_V \quad Q, V, K = \mathbb{R}^{1024 \times \frac{512}{8}} = \left(\mathbb{R}^{1024 \times 64}\right)$ = All heads ⇒ $\mathbb{R}^{1024 \times 64 \times 8}$ ⇒ Reshape $8 \times 1024 \times 64$

$= \mathbb{R}^{(1024 \times 64) * 8}$

$= \mathbb{R}^{1024 \times (64) * 8}$

(X is constant) or same for all K, Q, V

W is are three different weights

Single head in $d_{model} \times d_K$ (M·H·A)

multi head attention

Vanila Attention

$\mathbb{R}^{L \times d_{model}} = K, Q, V$

$\Rightarrow (X = \mathbb{R}^{L \times d_{model}})(W = \mathbb{R}^{d_{model} \times d_K})$

$\Rightarrow Q, V, K = \mathbb{R}^{L \times d_K}$

Now $Q, K, V = \mathbb{R}^{L \times h \times d} = \mathbb{R}^{1024 \times 8 \times 64} = \mathbb{R}^{8 \times (1024 \times 64)}$

Now the scaled dot product

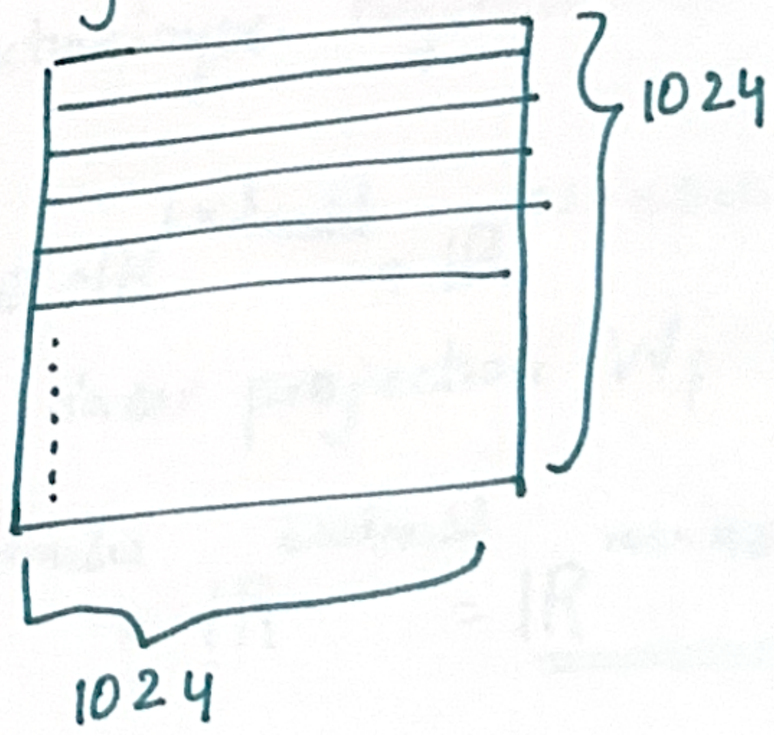
$$\text{Softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) * V = \text{Attention}(Q, K, V)$$

$Q K^T = \mathbb{R}^{h \times (L \times d)}$ = for each head = $\mathbb{R}^{L \times d}$ then $Q, K V$ becomes = $\mathbb{R}^{L \times d}$

$$Q K^T = (\mathbb{R}^{L \times d})(\mathbb{R}^{L \times d})^T = (\mathbb{R}^{L \times d})(\mathbb{R}^{d \times L}) = \mathbb{R}^{L \times L} = \mathbb{R}^{1024 \times 1024}$$

$\frac{Q K^T}{\sqrt{d_k}} = \left[\frac{\mathbb{R}^{1024 \times 1024}}{\sqrt{d_k}}\right]$ dividing By the heads sqrtroot dont affect the shapes But added for numerical stublity.

$$\text{Softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) = \mathbb{R}^{L \times d} = \mathbb{R}^{1024 \times 1024} =$$



Apply Softmax Row By Row

Convert to probablity Rowwise

$$\leftarrow \left[\begin{array}{l} \text{Row-0}\ [\_\ \_\ \_\ \ldots\ \_\ \_\ \_\ ] \\ \qquad\quad 0 \qquad\qquad 1023 \\ \text{Row-1}\ [\_\ \_\ \_\ \ldots\ \_\ \_\ \_\ ] \\ \qquad\quad 0\ 1 \qquad\qquad 1023 \\ \vdots \\ \text{Row-1023}\ [\_\ \_\ \_\ \ldots\ \_\ \_\ \_\ ] \\ \qquad\quad 0 \qquad\qquad 1023 \end{array}\right]$$

$$\text{Attention Score} = \mathbb{R}^{1024 \times 1024} = \text{Softmax}\left[\frac{Q K^T}{\sqrt{d_k}}\right]$$

Now we have attention scores $\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) = \mathbb{R}^{L\times L} = \mathbb{R}^{1024\times 1024}$

Now we do $\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V = \mathbb{R}^{L\times L} * \mathbb{R}^{L\times d_k} = \mathbb{R}^{L\times d_k} = \mathbb{R}^{1024\times 64}$ $d_k = \frac{d_{model}}{heads}$

we Can do the above Computations in parallel Since we have divided Our d_{model} By heads an then we Concatenate these heads.

$$\text{Concatenate heads}\left(\mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}, \mathbb{R}^{1024\times 64}\right)$$

$$\text{Attention}(Q,K,V) \Rightarrow \mathbb{R}^{1024\times(h\cdot d_k)} = \mathbb{R}^{1024\times 8\times 64} = \underline{\mathbb{R}^{1024\times 512}}$$

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \Rightarrow \Uparrow$$

After Concatenating we pass this throug a Projection layer. Projection After MHA

$W_p \in \mathbb{R}^{d_{model}\times d_{model}} = \mathbb{R}^{512\times 512}$

Once we have the multihead attention output $= \mathbb{R}^{L\times d_{model}} = \mathbb{R}^{1024\times 512}$

we pass this Concatenated head through a linear projection $W_p \in \mathbb{R}^{d_{model}\times d_{model}}$

$$(m\cdot HA) * W_p = \mathbb{R}^{L\times d_{model}} \times \mathbb{R}^{d_{model}\times d_{model}} = \mathbb{R}^{L\times d_{model}} = \underline{\mathbb{R}^{1024\times 512}}$$

Add & Norm (Residual Connection

Input of MHA + Output of Projection After M·H·A)

layer normalization dont affcat the Shapes of the output instd it just normalizes the values.

Input of MHA Befor projecting to QKV ↑

(X + Projection After M·H·A)

$(\mathbb{R}^{1024\times 512} + \mathbb{R}^{1024\times 512})$ After adding we do layer normalization.

Shape of output $= \mathbb{R}^{1024\times 512} = \mathbb{R}^{L\times d_{model}}$

Feedforward
(Two linear layers)

In this layer we Consider the output we got after the layernorm

$\mathbb{R}^{1024 \times 512}$

we take this output and pass it through (i) linear projection Expand dimensions
two linear layers (ii) Relu dimension Remain same
(iii) linear projection Project Back to dmodel

first linear projection:

Output from previous layer * weigh of this linear projection

$\mathbb{R}^{L \times d_{model}}$ $\quad$ $\underbrace{W\_first\_linear}_{\mathbb{R}^{d_{model} \times expand}}$ $\quad$ Expand = 2048

$\mathbb{R}^{L \times d_{model}} * \mathbb{R}^{d_{model} \times expand} = \mathbb{R}^{1024 \times 512} * \mathbb{R}^{512 \times 2048} = \mathbb{R}^{1024 \times 2048}$

↳ Projecting Lxdmodel to higher dimension ....

So Now we have projected things from 512 → 2048 dim & then we apply Relu activation.

This keeps the Dimension Same $\Rightarrow \mathbb{R}^{L \times Expand} = \mathbb{R}^{1024 \times 2048}$

After the Relu we apply one more linear transformation and project the ~~weigh~~ things Back to $\mathbb{R}^{L \times d_{model}}$ dimension

$\mathbb{R}^{L \times expand} * \underbrace{\mathbb{R}^{expand \times d_{model}}}_{W\_Second\_linear\_layer} = \mathbb{R}^{L \times d_{model}}$

Projecting Lxexpand Back to $\mathbb{R}^{L \times d_{model}}$

The output of the feed forward layers is $\mathbb{R}^{L \times d_{model}} = \mathbb{R}^{1024 \times 512}$

→ Add and Norm

(Residual Connection + layer Normilization)

The output from the feed forward layer (i.e second projection layer)

$$\mathbb{R}^{1024\times512} = \mathbb{R}^{L\times d_{model}}$$

we add the X the input ~~for~~ to the m·H·A Here

$$X = \mathbb{R}^{L\times d_{model}} = \mathbb{R}^{1024\times512}$$

$$\text{Secondenlinearlayer} = \mathbb{R}^{L\times d_{model}} = \mathbb{R}^{1024\times512}$$

⊕ Residual Connection

$$\Downarrow$$

$$Op = \mathbb{R}^{L\times d_{model}}$$ then we perform layer norm on the (Op) this dont change the shape of the layer. The shape is $\mathbb{R}^{L\times d_{model}} = \mathbb{R}^{1024\times512}$

The above layers and steps are the whole Encoder Block that performs.

Operations.

Input Embeddings $\in \mathbb{R}^{L\times d_{model}}$ ⊕ Positional encodings. $\in \mathbb{R}^{L\times d_{model}}$ → $\mathbb{R}^{L\times d_{model}}$

[Input Embedding ⊕ Positional encoding] $\in \mathbb{R}^{L\times d_{model}}$ × projection (Q, K, V) $W_K, W_Q, W_V \in \mathbb{R}^{d_{model}\times d_k}$ → $Q, K, V \in \mathbb{R}^{L\times d_{model}}$

[Q, K, V] $\in \mathbb{R}^{L\times d_{model}}$ Q, K, V

(M·H·A)

Concatenate heads. $\left[\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)_{\mathbb{R}^{L\times L}} * V\right]_{\mathbb{R}^{L\times d_{model}}}$ → Concatenation of heads $\in \mathbb{R}^{L\times d_{model}}$

$\mathbb{R}^{L\times L}$

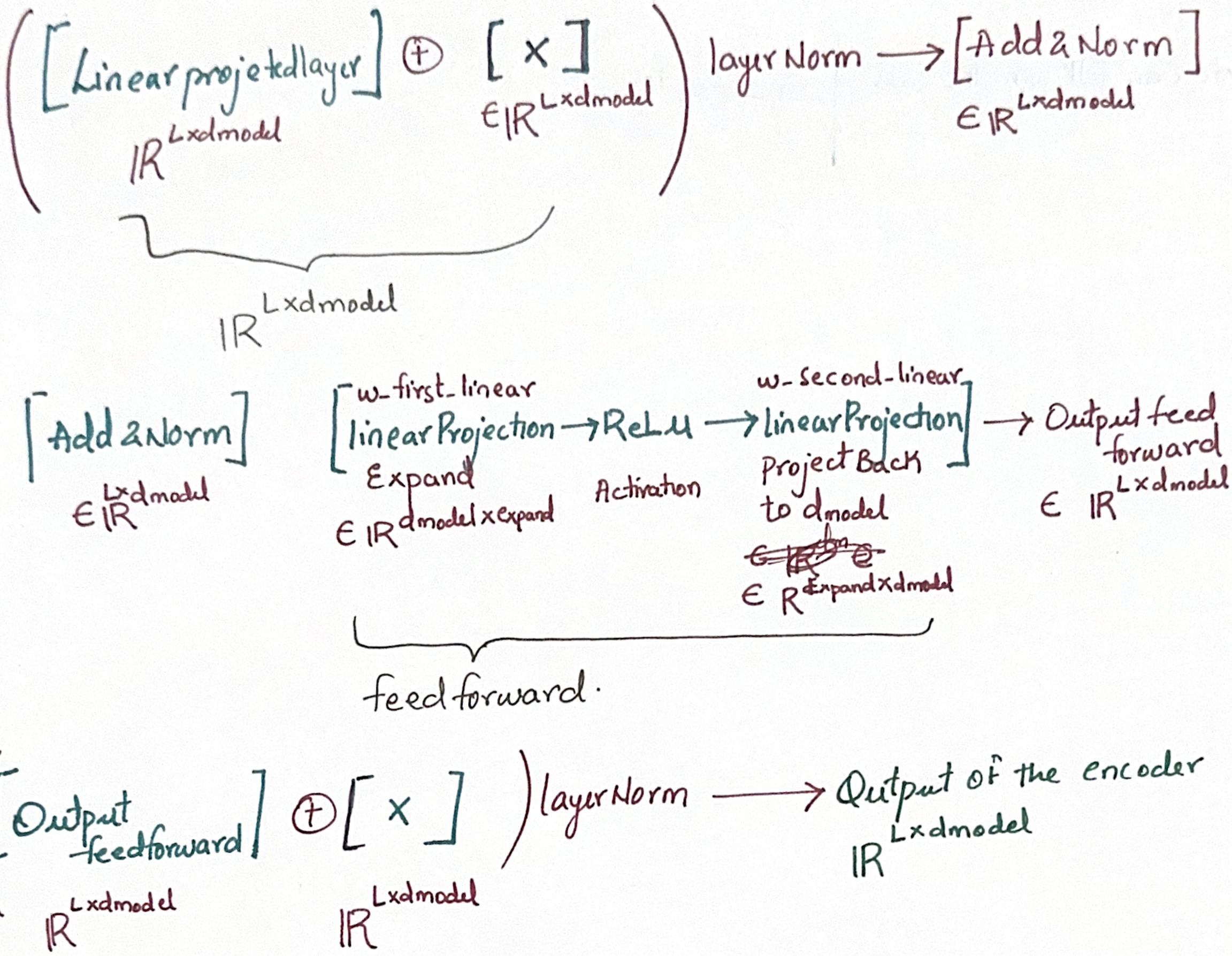
$\mathbb{R}^{L\times d_{model}}$

$\mathbb{R}^{L\times heads\times \frac{d_{model}}{heads}}$

[weights]

Linear projection $W_{PE} \in \mathbb{R}^{d_{model}\times d_{model}}$ → Linear Projected layer $\mathbb{R}^{L\times d_{model}}$

[Concatination of heads.] $\mathbb{R}^{L\times d_{model}}$



The above 7 steps is the Encoder Block form the paper Attention is All you need

Repeat these steps N times & you Get the Encoder Block.

From the paper.