Transfomers (Chapter 12 of the UDL Book)

Natural Language processing

- ↳ Machine translation
- ↳ Parts of speech tagging
- ↳ Generating missing words
- ↳ Autocomplete

] Understanding language → vectors or a sequence of number

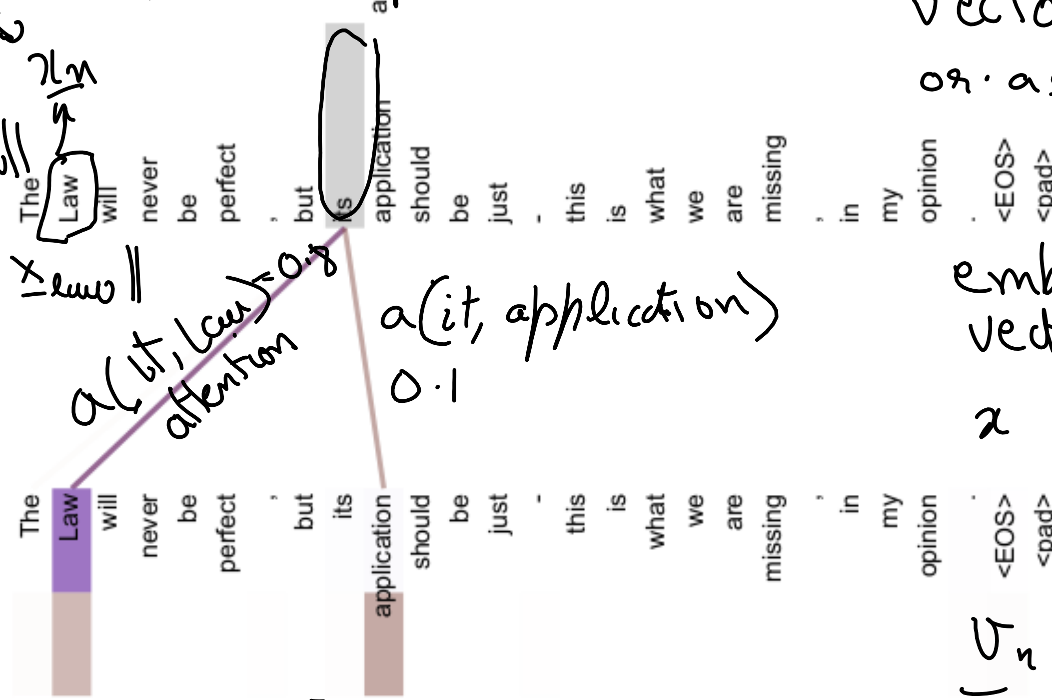
embedding vector

x

Rule → x_r ↔ x_n

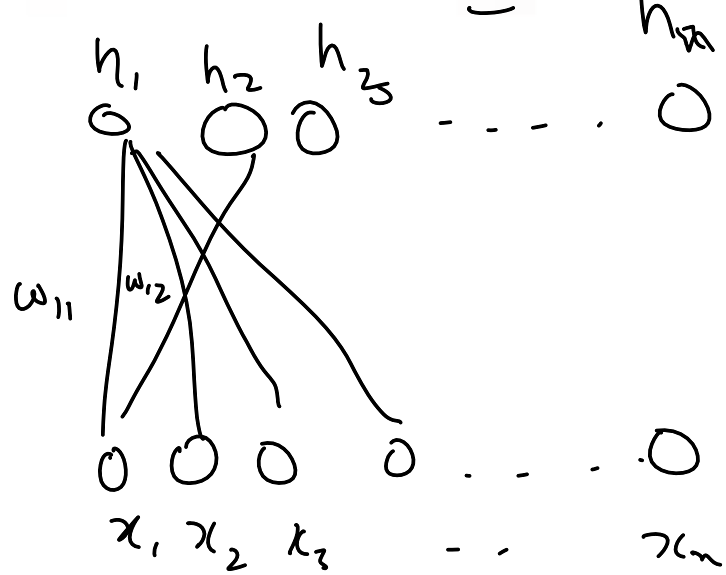
$\|\underline{x}_{rule} - \underline{x}_{law}\|$

$< \|\underline{x}_{chows} - \underline{x}_{law}\|$

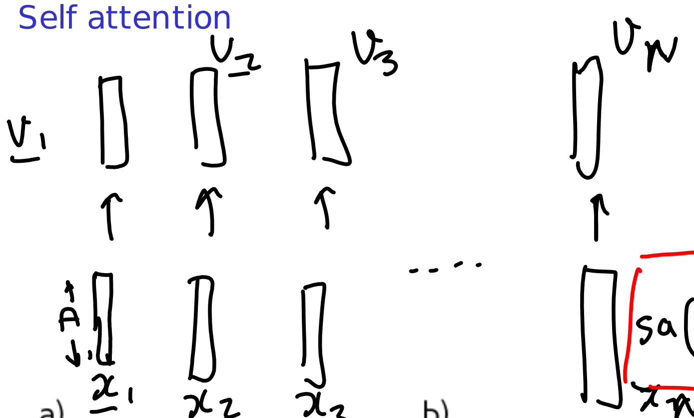


$\underline{v}_1$ $\underline{v}_{23}$ $\underline{v}_n$

$$\underline{h} = a(W\underline{x} + \underline{b})$$



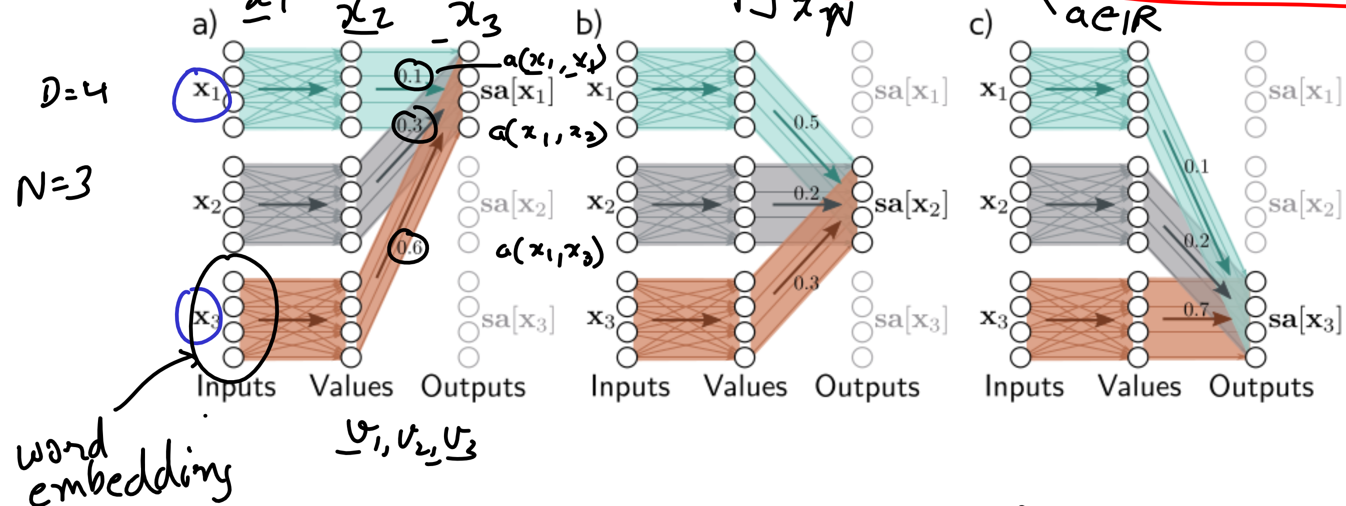
Self attention



$$\underset{D\times 1}{v_n} = \underset{D\times D}{\Omega_v}\, \underset{D\times 1}{x_n} + \underset{D\times 1}{\beta_v}$$

$$sa(x_m) = \sum_{n=1}^{N} a(x_n, x_m)\, v_n$$

$a \in \mathbb{R}$



$$a(x_n, x_m) \in [0,1] \quad \forall\, n \in [1, N] \text{ and } m \in [1, N]$$

$$\sum_{m=1}^{N} a(x_n, x_m) = 1$$

(x_m ↑ input)

Dot product self attention

$a(\underline{x}_n, \underline{x}_m)$

value vector $\quad \underline{v}_n = \underline{\Omega}_v \underline{x}_n + \underline{\beta}_v$ $\quad$ for each word in the sentence $\forall\ n \in [1, N]$

key vector $\quad \underline{k}_n = \underline{\Omega}_k \underline{x}_n + \underline{\beta}_k$

query vector $\quad \underline{q}_n = \underline{\Omega}_q \underline{x}_n + \underline{\beta}_q$

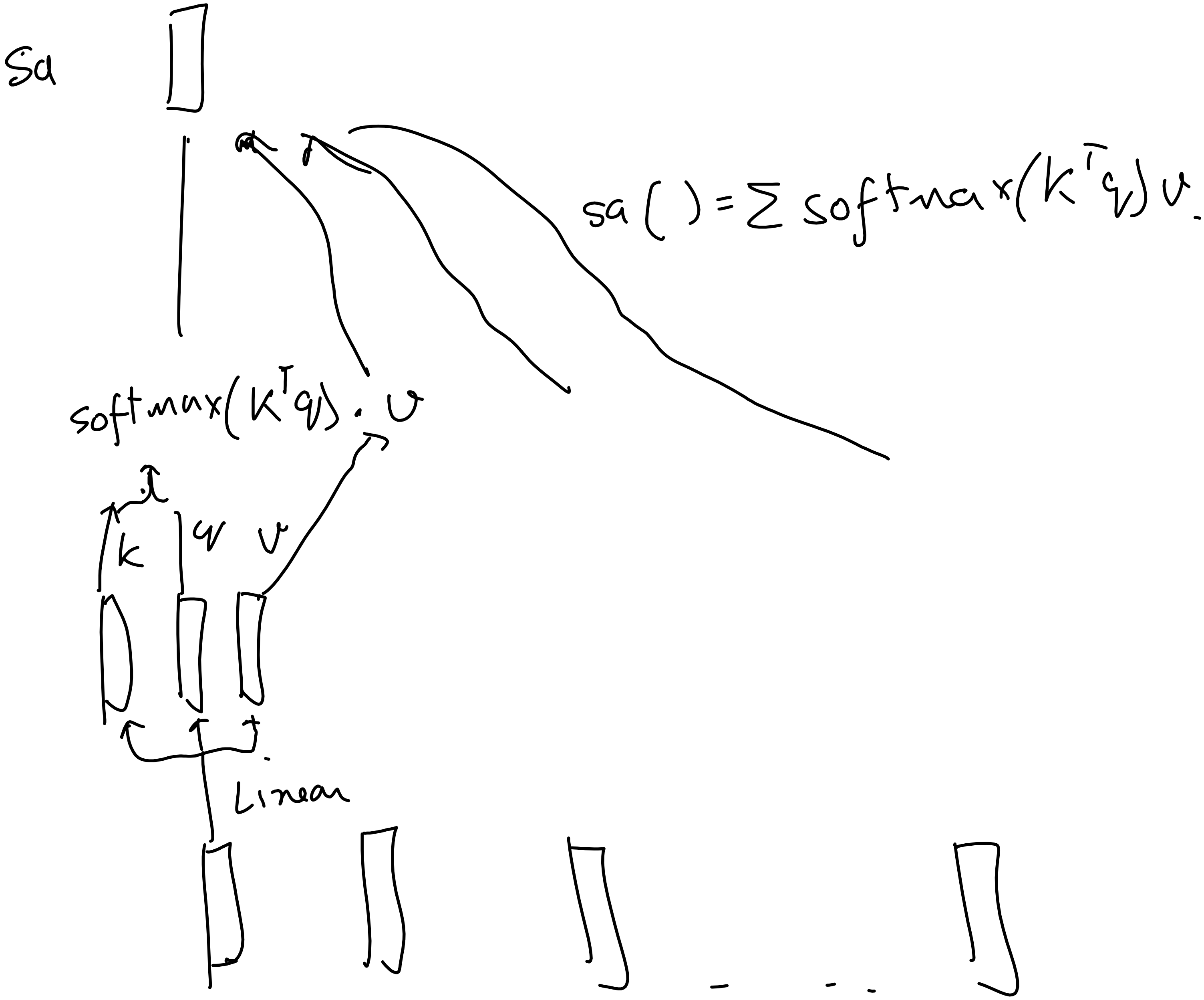
$$a(\underline{x}_n, \underline{x}_m) = \text{softmax}_m \left(\underline{k}_m^T \underline{q}_n \right)$$

($\underline{x}_n$: output, $\underline{x}_m$: input)

$$= \frac{\exp(\underline{k}_m^T \underline{q}_n)}{\sum_{m'=1}^{N} \exp(\underline{k}_{m'}^T \underline{q}_n)}$$

$$sa(\underline{x}_n) = \sum_{m=1}^{N} \underbrace{\text{softmax}_m \left(\underline{k}_m^T \underline{q}_n \right)}_{\in \mathbb{R}} \underline{v}_m$$

$\underline{v}_m = \Omega_v \underline{x}_m + \beta_v$



key
query
value

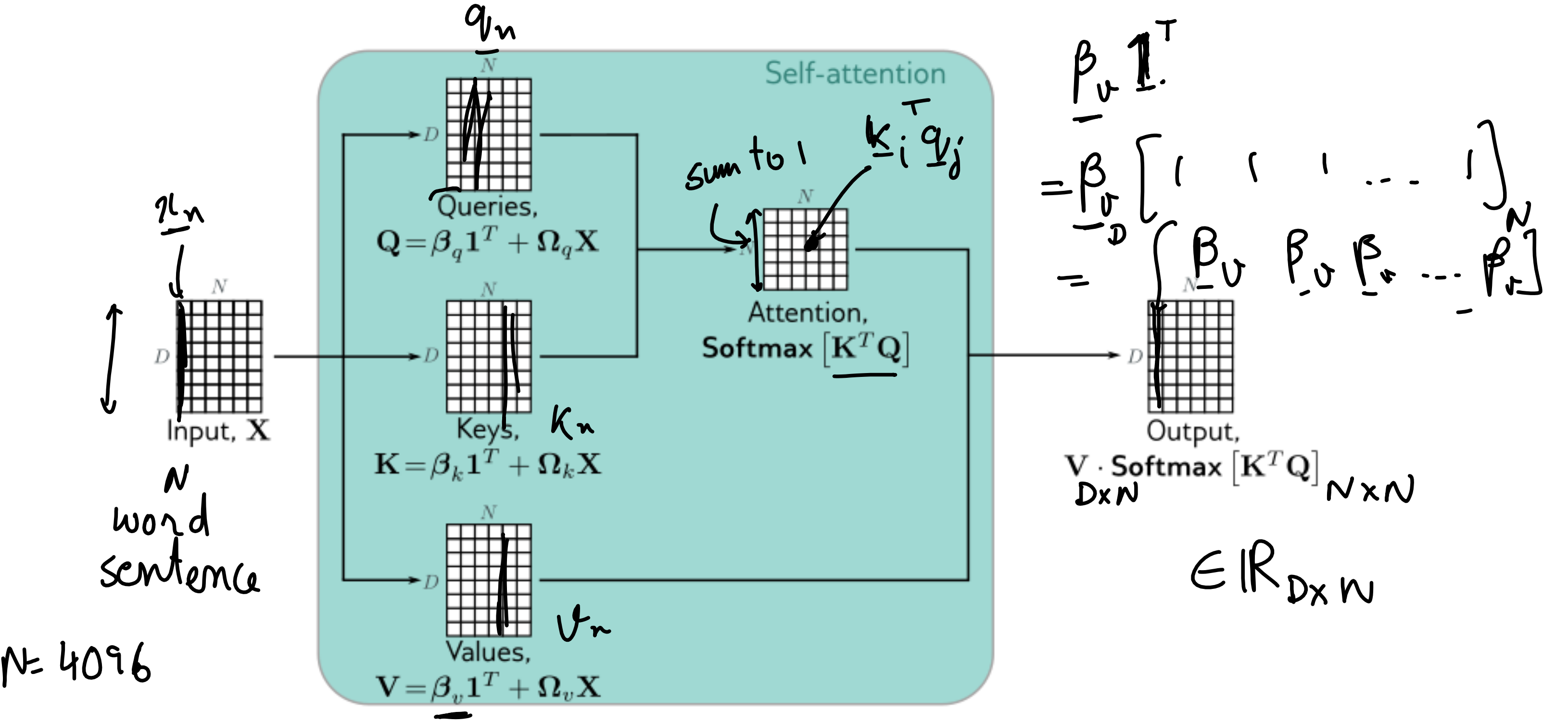
softmax $(K^T q)$

Database

key - value pair

question - answer pairs

query is a new key that you want to search in the database



$$sa(x_n) = \sum_{m=1}^{N} softmax_m(k_m^T q_n) \, v_m$$

$\in \mathbb{R}$

Scaled Dot product self attention

$$Sa[X] = V[X] \, Sofmax(K(X)^T Q(X))$$

$$Var(K(X)_{D \times N}) = 1$$
$$Var(Q(X)_{D \times N}) = 1$$

$$Var(K(X)^T_{N \times D} \, Q(X)_{D \times N}) = D$$

$$Sa[X] = V[X]\,\text{softmax}\left(\frac{K(X)^T Q(X)}{\sqrt{D}}\right)$$

Positional encoding

The sentence The woman ate the raccoon has a quite different meaning to The raccoon ate the woman.

$\underline{x}_1$ for The

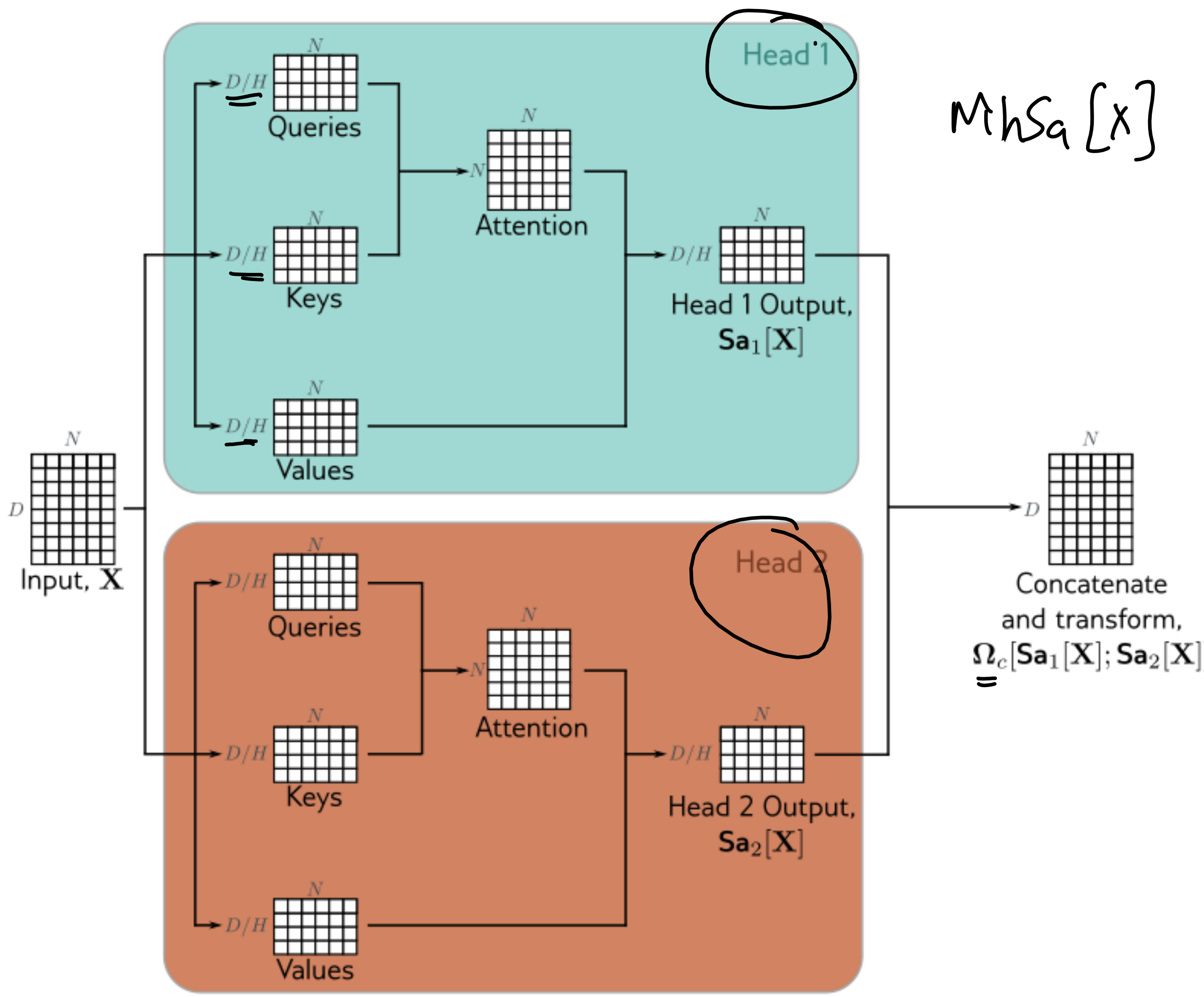
$\underline{x}_2$ for Raccon

$\underline{x}_i = [\ ]$ WE (D) ; $[\ ]$ PE (D')

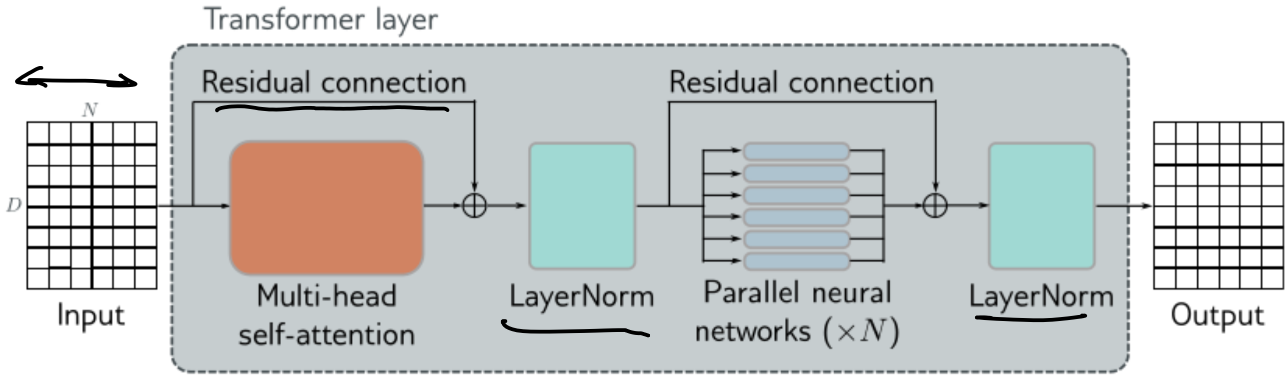
$$PE_{(pos,\,2i)} = \sin\left(pos / 10000^{2i/D'}\right)$$

$$PE_{(pos,\,2i+1)} = \cos\left(pos / 10000^{2i+1/D'}\right)$$

Scaled dot product self-attention


Head 1
Queries
Keys
Values
Attention
Head 1 Output,
$\mathbf{Sa}_1[\mathbf{X}]$
MhSa[X]
Input, $\mathbf{X}$
Head 2
Queries
Keys
Values
Attention
Head 2 Output,
$\mathbf{Sa}_2[\mathbf{X}]$
Concatenate
and transform,
$\boldsymbol{\Omega}_c[\mathbf{Sa}_1[\mathbf{X}]; \mathbf{Sa}_2[\mathbf{X}]]$

Transformer Layer



Layer Norm

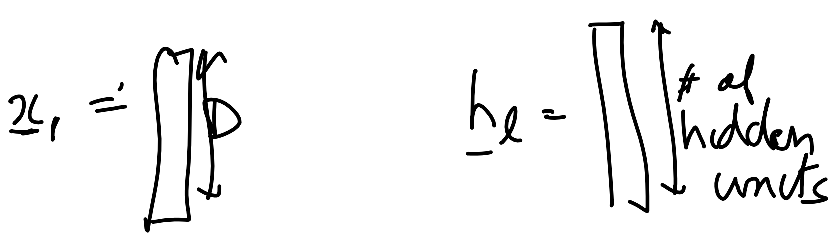
$$\text{Batch Norm}(x_n) = \frac{x_n - \mu}{\sigma}$$

$$\mu = \frac{1}{B}\sum_{b=1} x_b \qquad \sigma^2 = \frac{1}{B}\sum (x_b - \mu)^2 \qquad x_1, x_2, x_3 \dots x_B$$

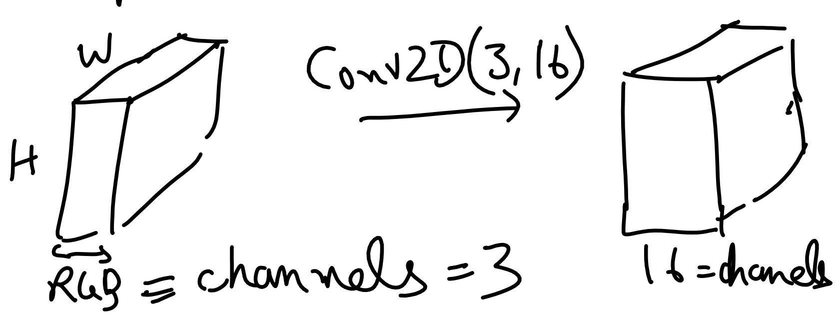
Layer Norm

the mean and variance are computed over the "channel" dimension

a) Linear layer

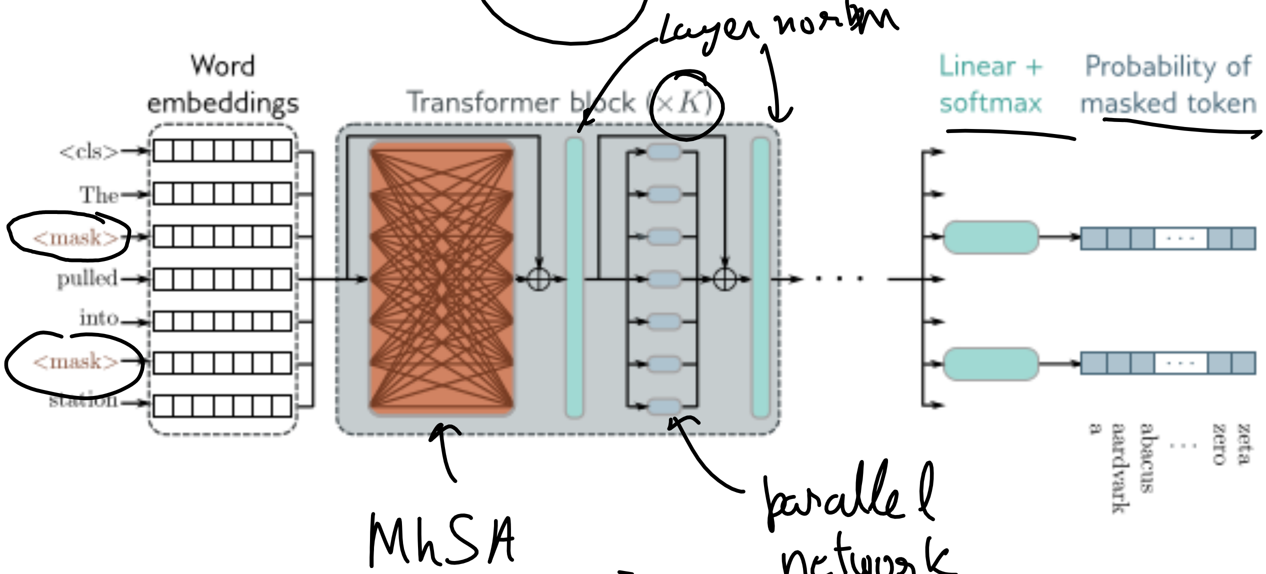


b) Conv layer



c) NLP os Sq

Channels = N = number of word in the sentence



Self-supervised pretraining

GPT = Generative Pretrained Transformers