

Transfer learning

pretrained models
use + specific data
for finetuning

not in RNN



Word Embeddings

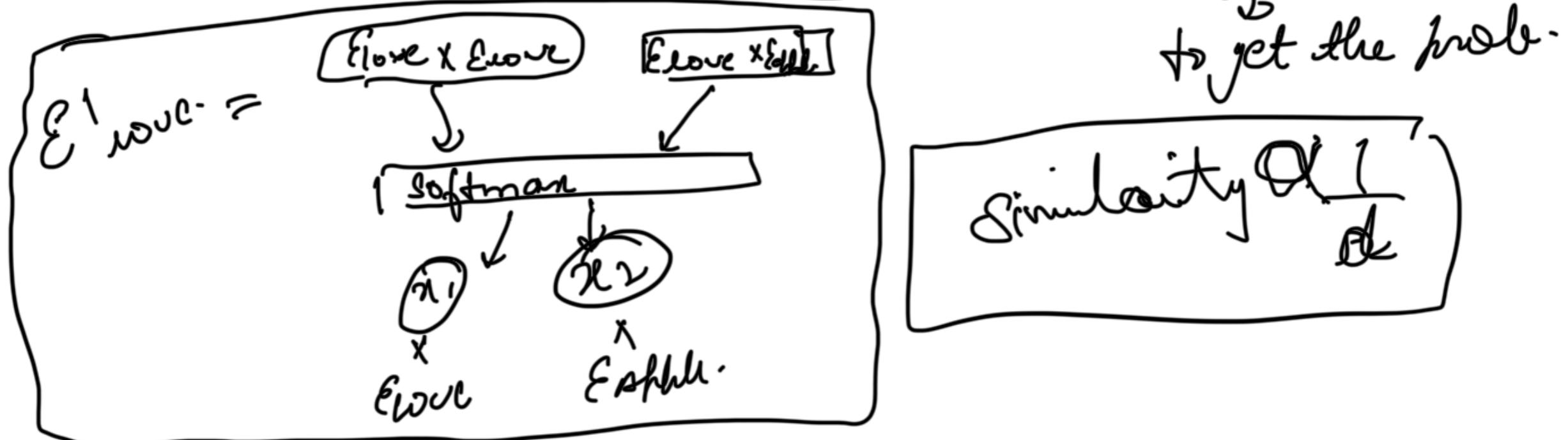
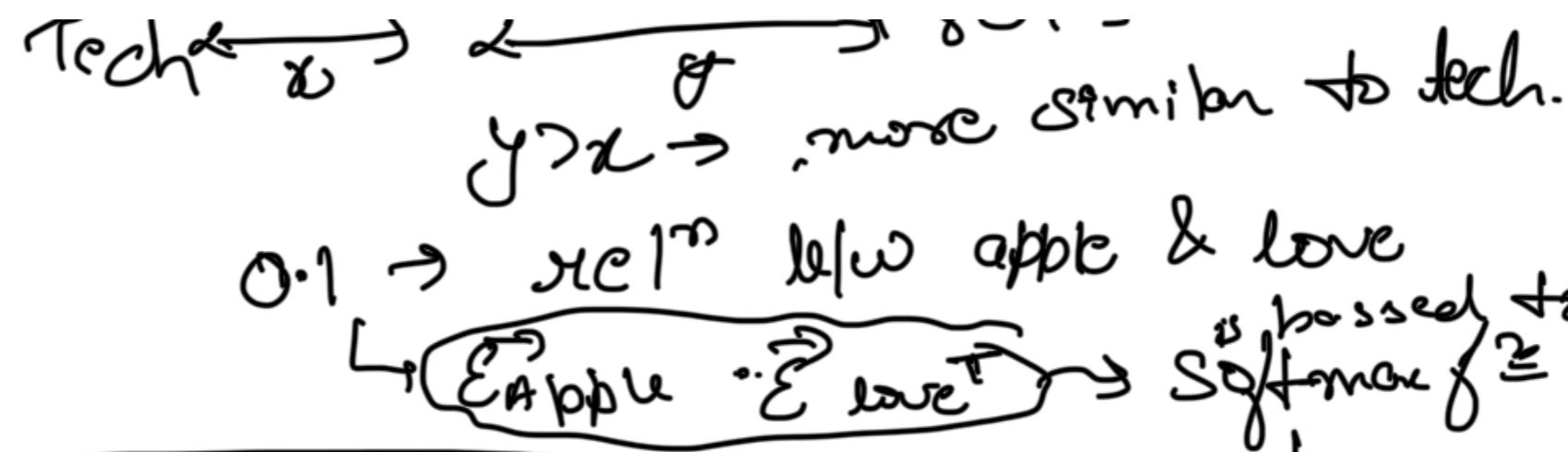
love apple phones

$$\text{Apple} = 0.1 \times E_{\text{love}} + 0.5 \times E_{\text{apple}} + 0.4 \times E_{\text{phones}}$$

Word Emb.
space



should apply



Problems

→ no learnable params.

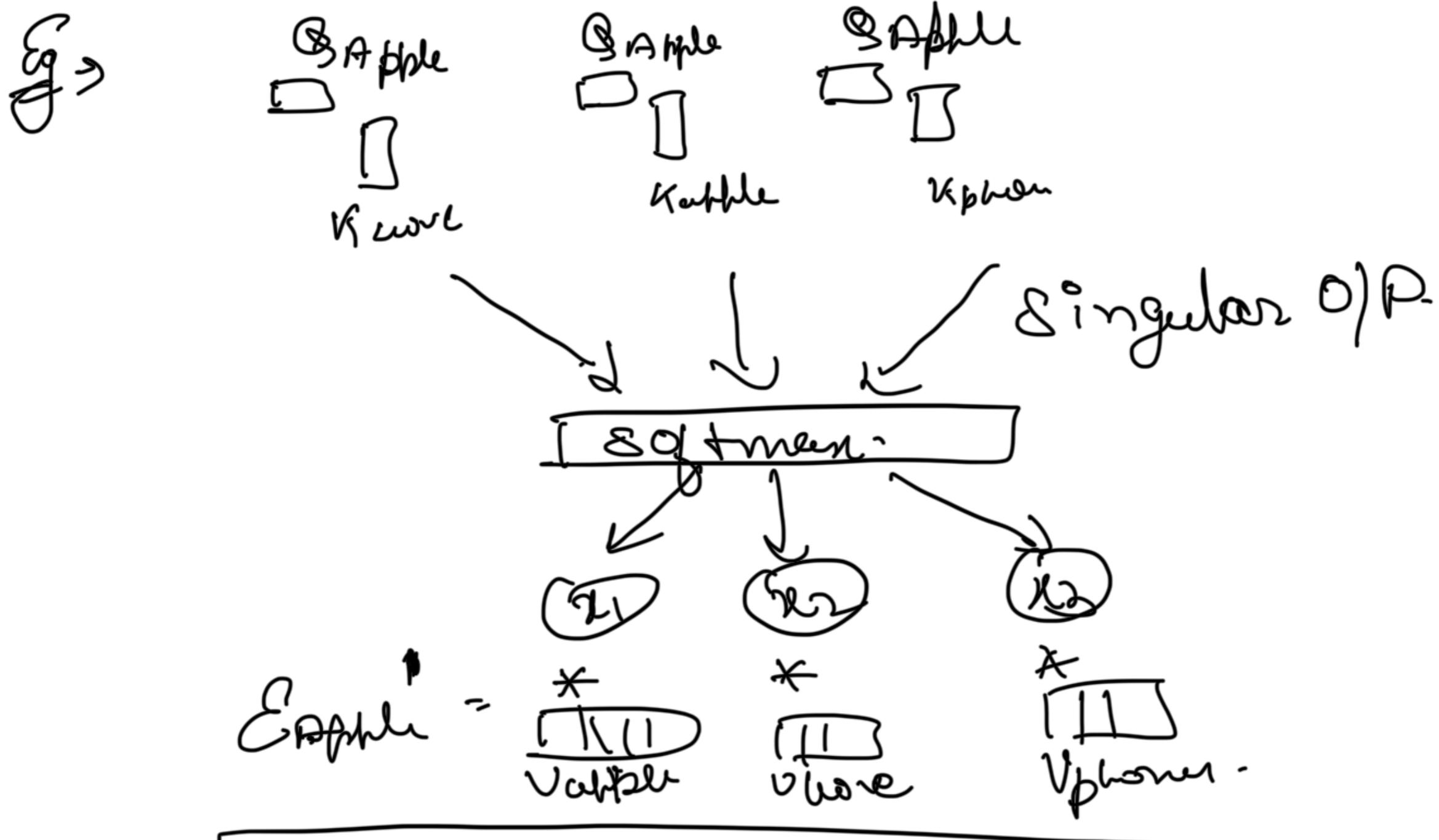
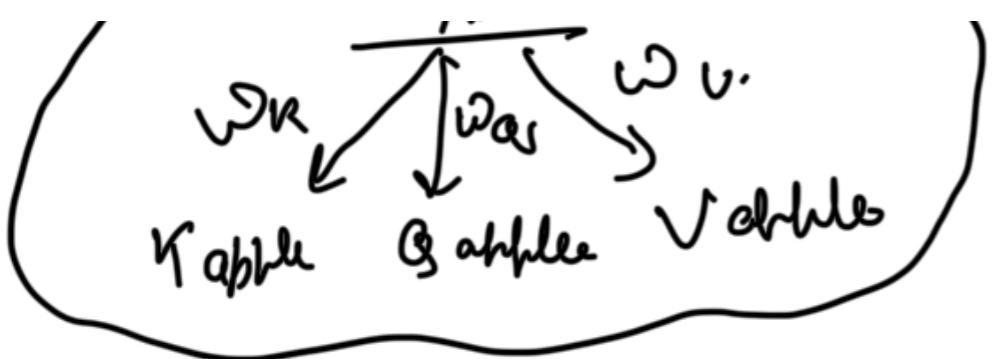
use of same vector

To solve

Example ω_q

love ω_k

apple



$$E_{\text{Affine}} = \alpha_1 \cdot V_{\text{Affine}} + \alpha_2 \cdot V_{\text{Linear}} + \alpha_3 \cdot V_{\text{Bilinear}}$$

no dependence of E of lone δ 's.

So, they eliminate sequential processing
& allows parallelization.

long range dependencies

Software $(Q \cdot R^T) \cdot V$ is done to add learnable parameters in the matrices

* Linear Transformation → changes dimension
→ extracting such features
see - OLSR

Juice

- Affine

very absent

• Watch
time

• Tech in space so
not much relation can be captured.

↳ Change dimension
 $y = w \cdot x$ adjust like
ANN.

	Apple	Orange	Watch
isTech	2	0	1
isWatch	0	0	3
isFruit	2	3	0

$$\begin{bmatrix} 2 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 4 & 3 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 8 \end{bmatrix}$$

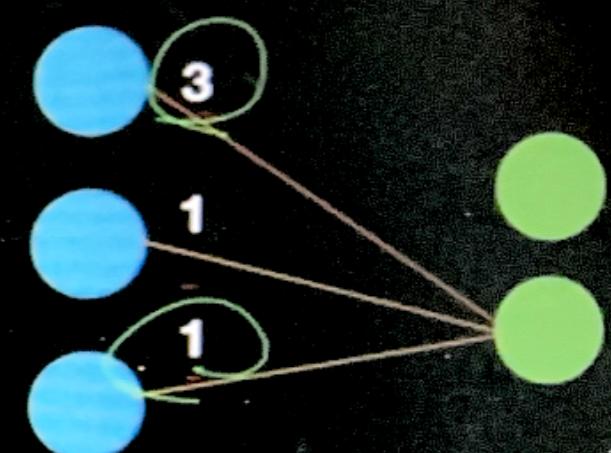
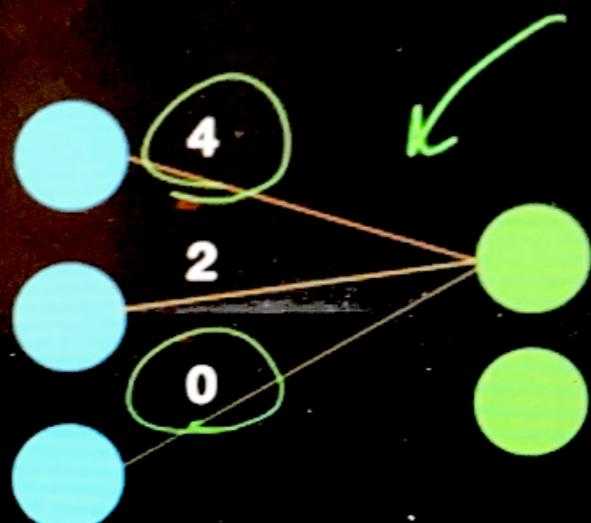
$$2 \times 4 + 0 \times 2 + 2 \times 0$$

$$2 \times 3 + 0 \times 1 + 2 \times 1$$

Apple

$$\begin{bmatrix} 0 & 3 \end{bmatrix}$$

Orange



$$\begin{bmatrix} 1 & 3 & 0 \end{bmatrix}, \begin{bmatrix} 4 & 3 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}$$

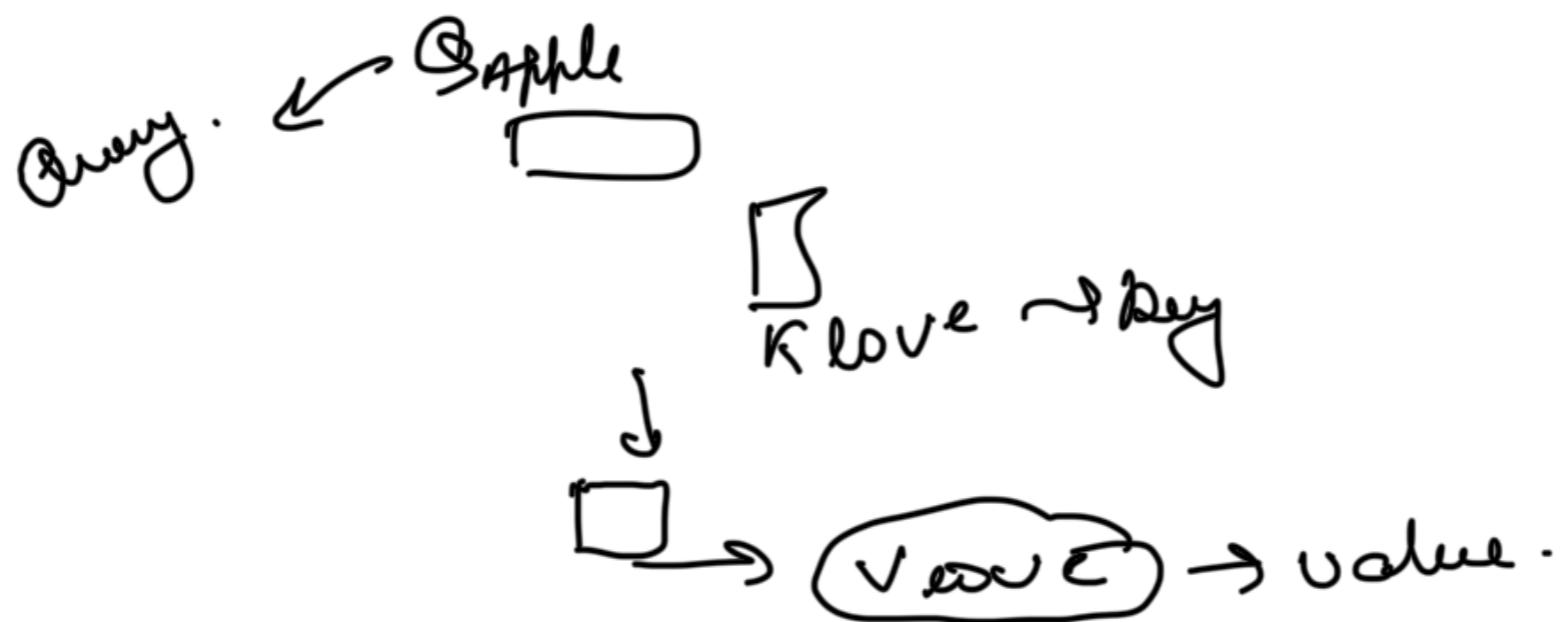
$$\begin{bmatrix} 10 & 6 \end{bmatrix}$$

Watch

*

Query, key, value
Q, K, V.

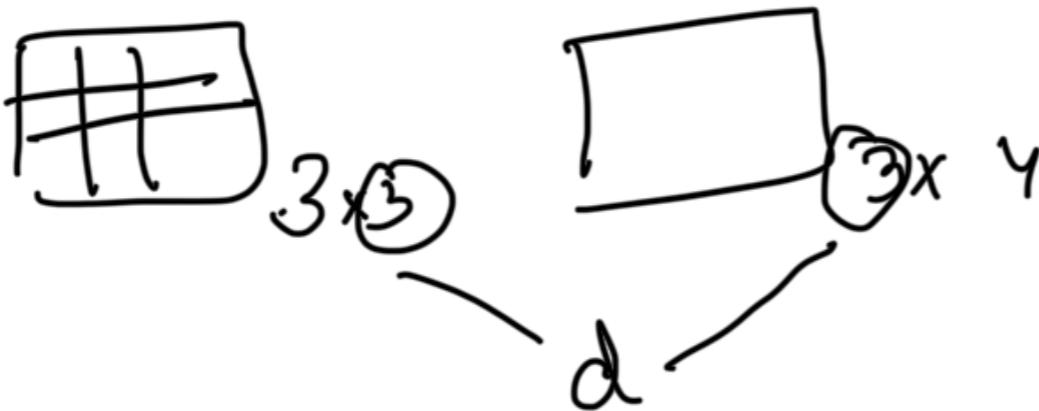
sound of [dog] =
↓ ↓
query key
Bam
value.



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

* Why $\sqrt{d_K}$ → dim. of $\{S, K, V\}$
 ↓
 hyperparameter

variance $(A \cdot B) \propto n$, $n = \text{common dim. of } A \& B$.



High var → Vanishing grad
 pulse

Hinders training

so we use $\sqrt{d_K}$

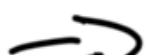
$$C = A \cdot B$$

$$\text{Var}(C) = n \cdot \text{Var}(A) \cdot \text{Var}(B)$$

→ zero mean

→ indep. random var.

→ entries of $A \& B$ have identical values.



$X = Q \cdot K^T \int \text{var}(x) Q dK$ To reduce variance we use \sqrt{dK} . \Rightarrow keep $\text{var}(X)$ same.

$$X \Rightarrow \text{var}(X)$$

$$Q \cdot K \rightarrow Q^2 \text{var}(X)$$



Multi-Head Attention

As complexity of model ① It is tough to capture the meaning of different words & other things

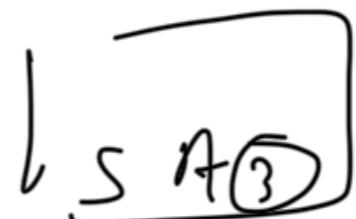
Spatial Rel^{ns}



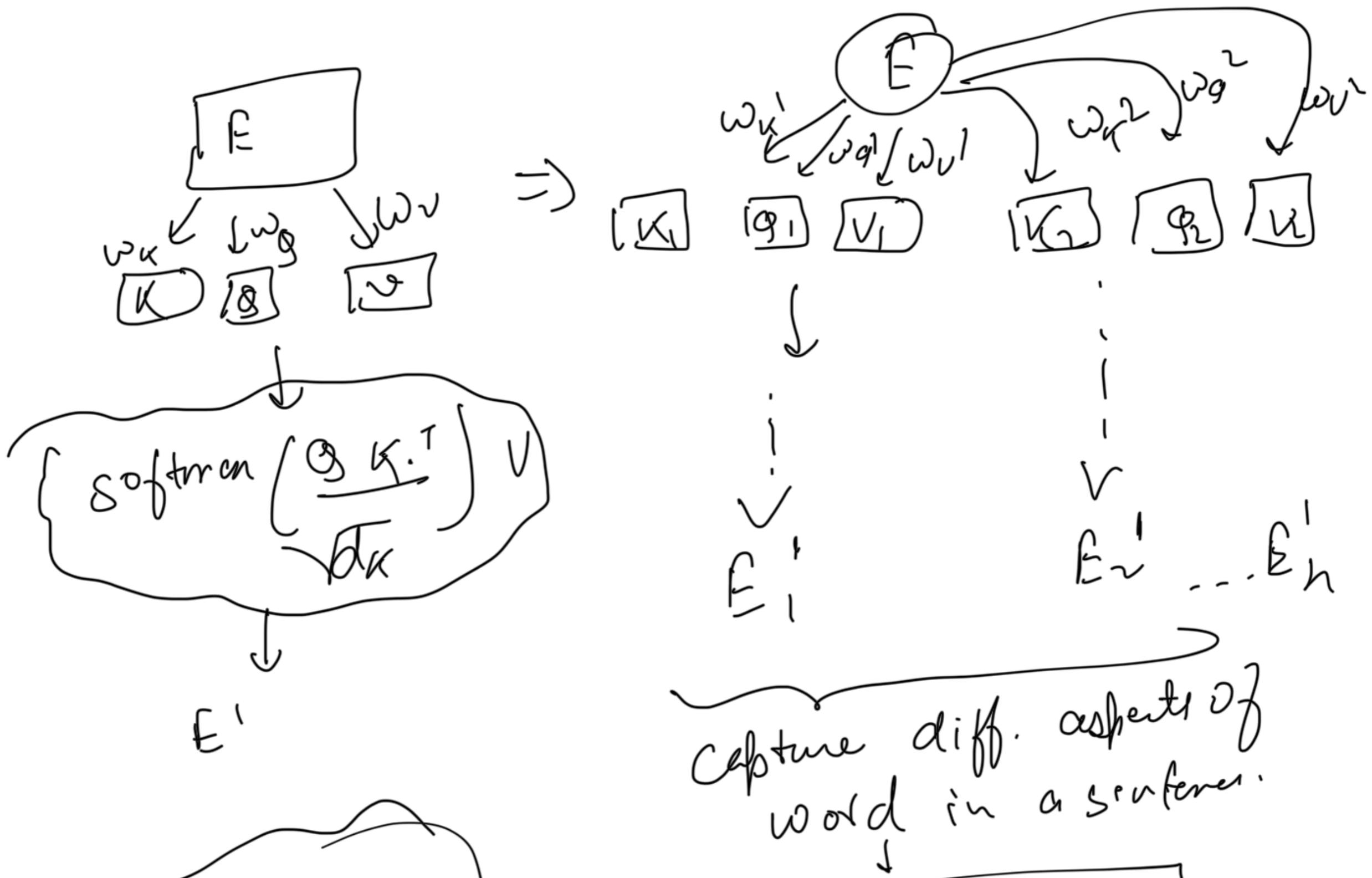
Sub-verb
obj

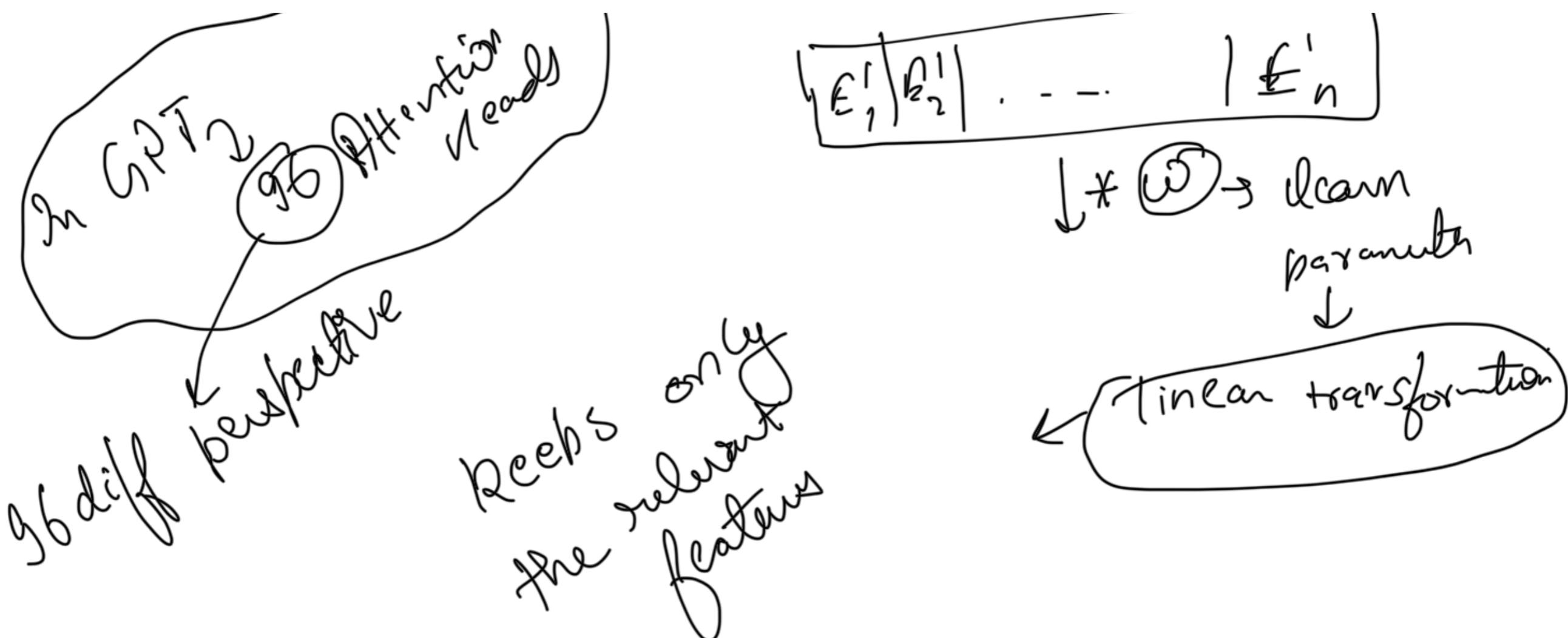


Time \rightarrow



to provide diverse content
just like CNN





* Positional Encoding

To due to lack of sequential understanding.

If individually sinusoids used
then \rightarrow due to periodicity 2 words

may be identical

use $\sin(u) \cos(x)$
together

In Og transform

512 dim \rightarrow 264 pair of cosine

$\boxed{\sin(u) \cos(u) \sin(\frac{u}{2}) \cos(\frac{u}{2}) \dots}$

reduced freq

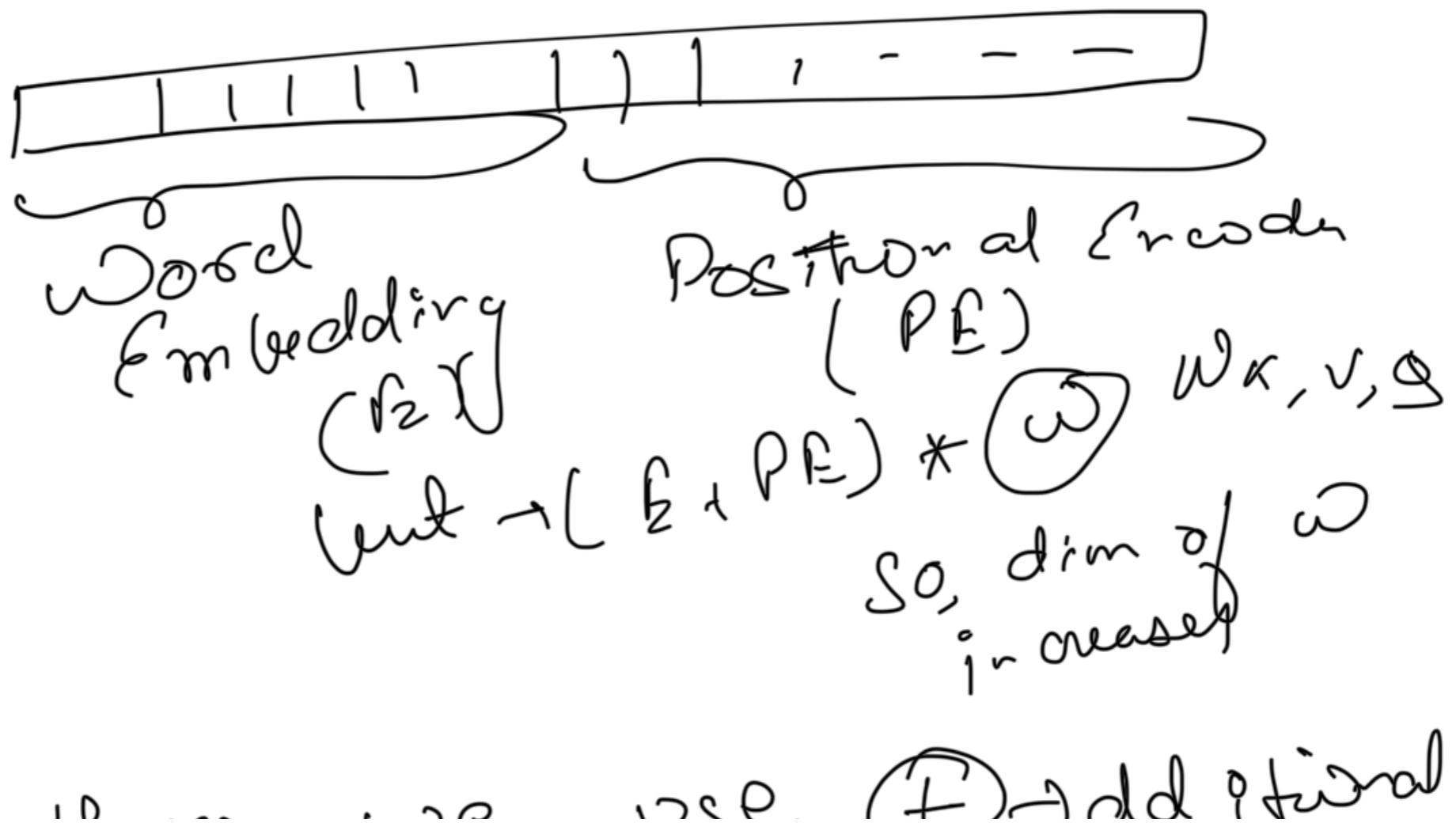
$$PE(pos, 2i) = \sin(pos / 1000 \text{ mod } 2\pi)$$

$$PE_{\text{pos}}(2i+1) = \cos(pos / 1000 \text{ mod } 2\pi)$$

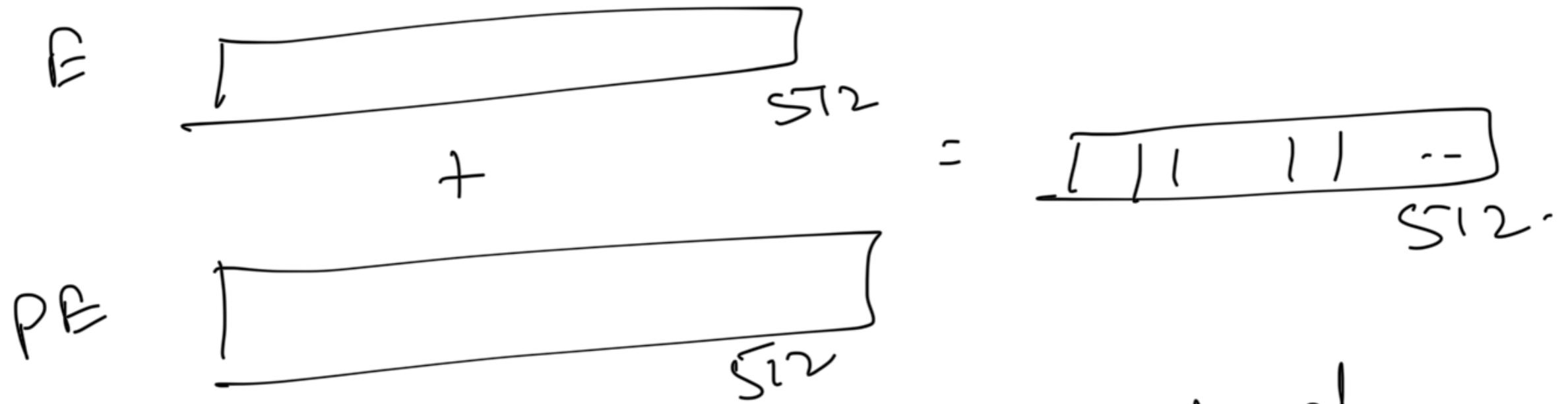
If we know

$\text{PrePos} \rightarrow \text{PE Pos}(k) \rightarrow \text{offset}$
 can be calculated
 ↓
 as there is a transformation
 matrix T_K .

$$\text{PE Pos}_K = T_K \times \text{PE Pos}$$



Hence we can skip the operation



④ Reduces computational overhead

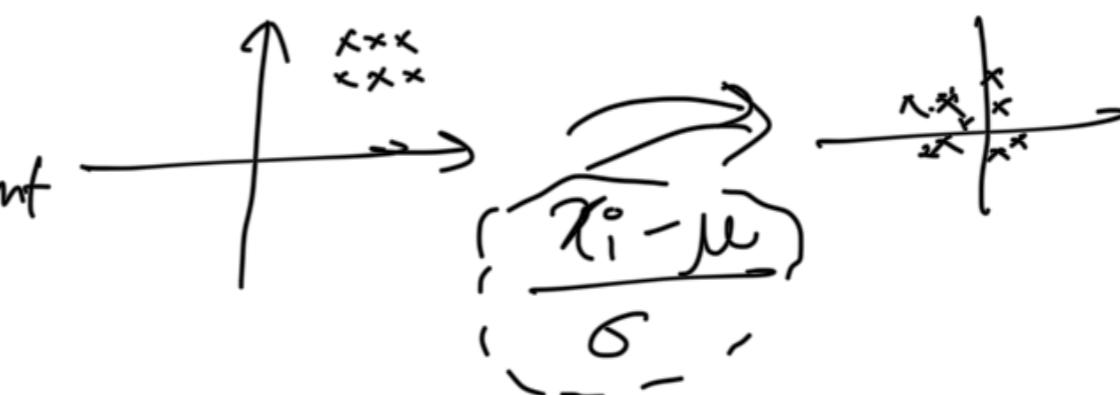


Layer: Normalization
 $\xrightarrow{\text{---}} \mu = 0 \}$
 $\sigma = 1 \}$

→ Stabilizes training of model

↓ by
avoiding exploding or
vanishing gradient
prob.

→ Faster convergence.



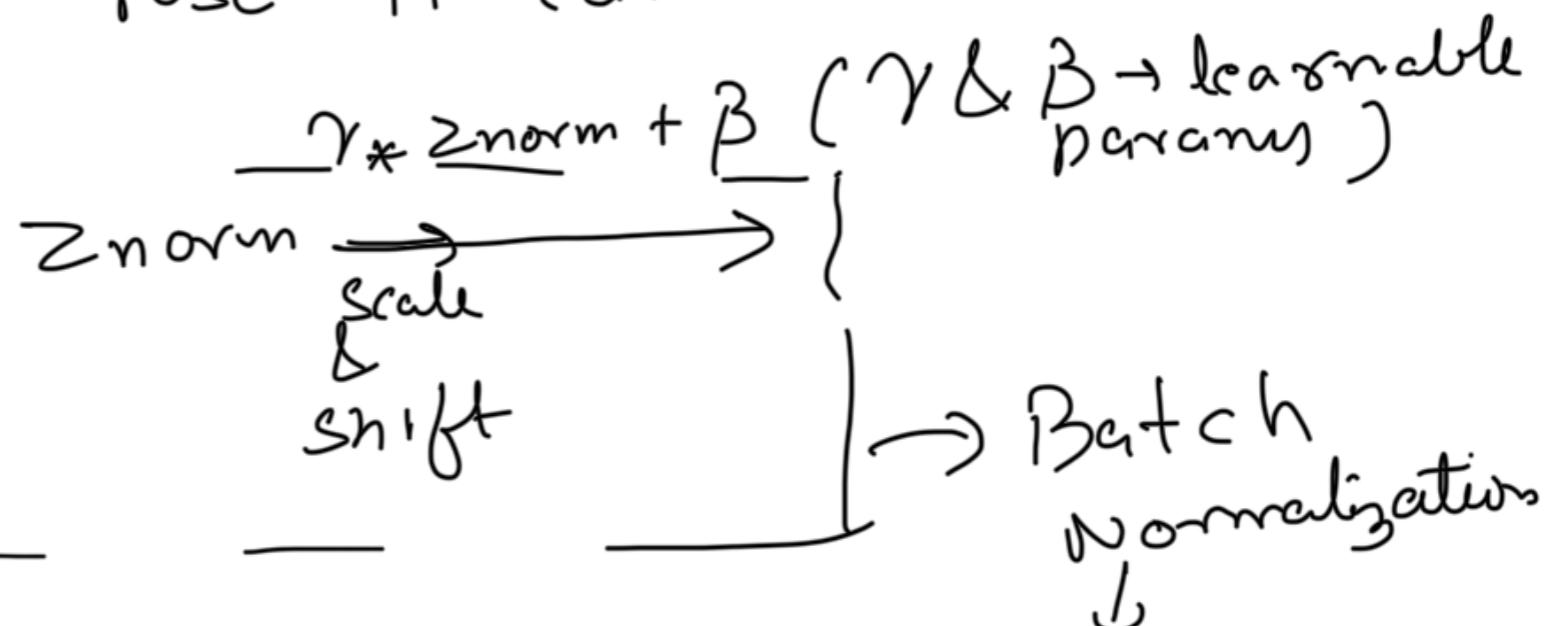
Internal Covariate Shift

Eg → training done to recognize red rose
so it will recognize it in testing phase

but if we give white rose it can't.

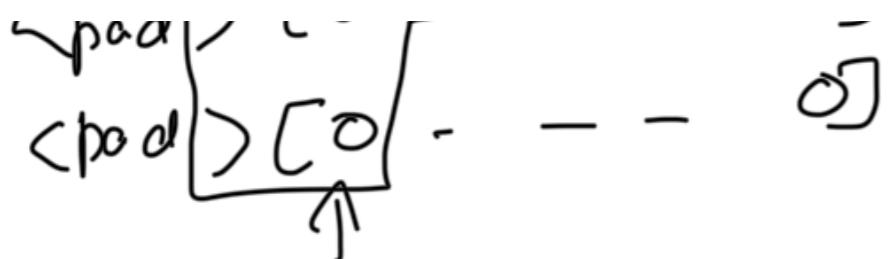
Scale & shift,

$$z \xrightarrow{\text{normalization}} \left(\frac{z - \mu_i}{\sigma_i} \right)$$

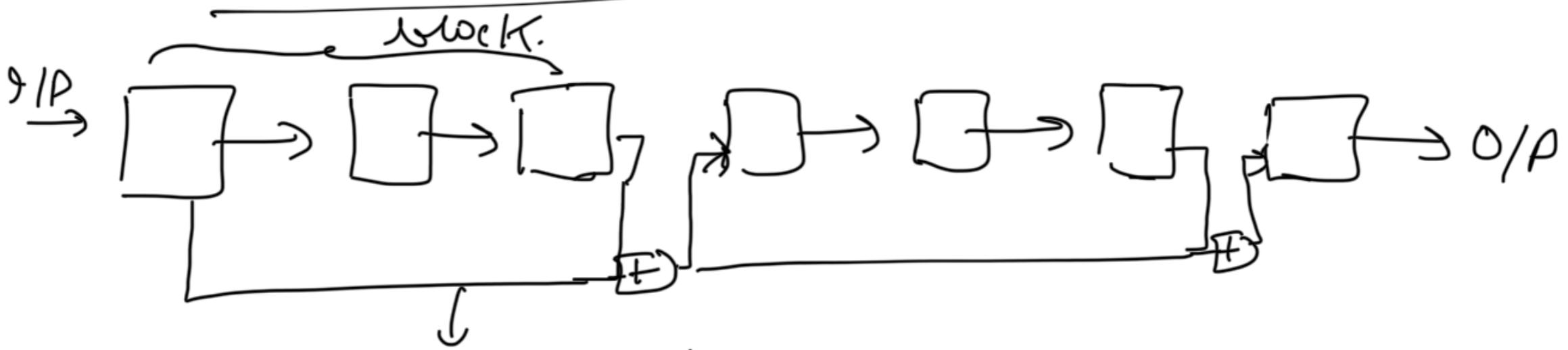


how	[0.8]	...	1.7]	$\leftarrow \mu_1, \sigma_1$
are	[0.5]	...	0.7]	
you	[0.2]	...	0.0]	
good	[0.1]	...	-0.2]	
you	[0.1]	...	0.7]	

does not work on transforms as instead of batch we take each data point



* Residual Connection

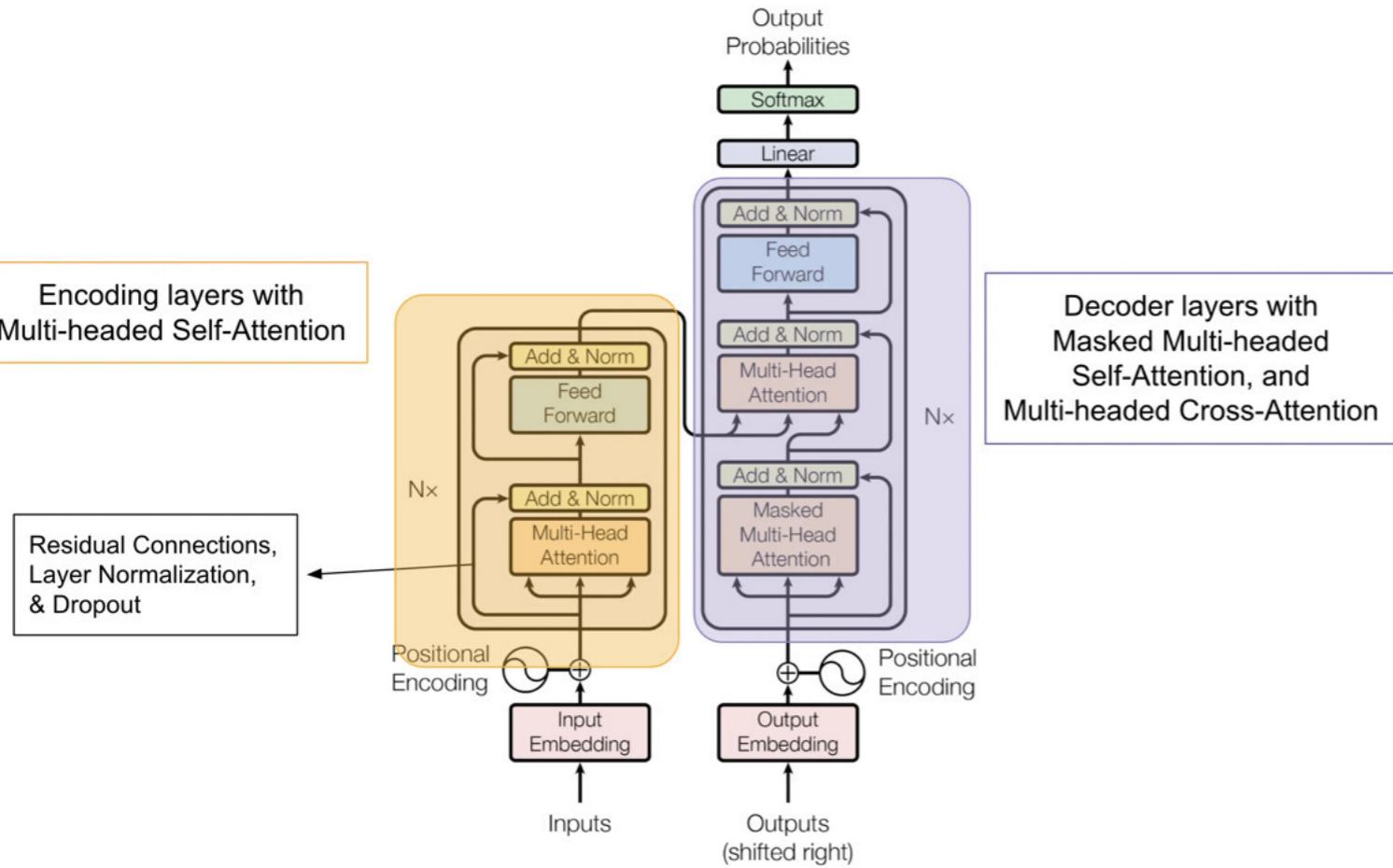


Skip connection

→ Better gradient flow

→ Reduced burden on layers

→ Enables Deeper N/w



Title: Annotated Transformer Architecture