

* HW03 -

* HW04 -

Q Re: Transformer
vs.

Previous models

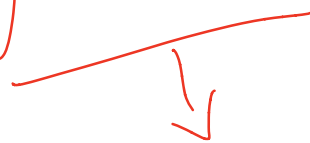
- In general, Transformer
more powerful.
- Take into account
hyperparameters:

* 12 heads, dataset is small \rightarrow overfit

- Make sure it is not overfitting.

Large \rightarrow BiLSTM
Transformer

Regularization



dropout \uparrow 0.7
0.5

underfit



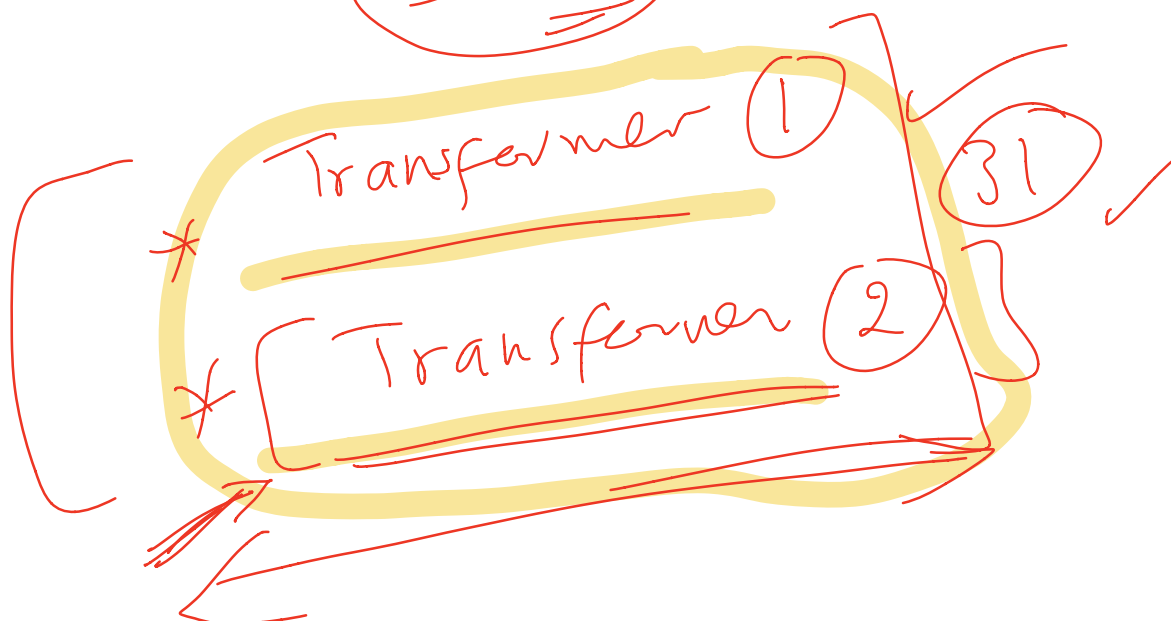
Survey

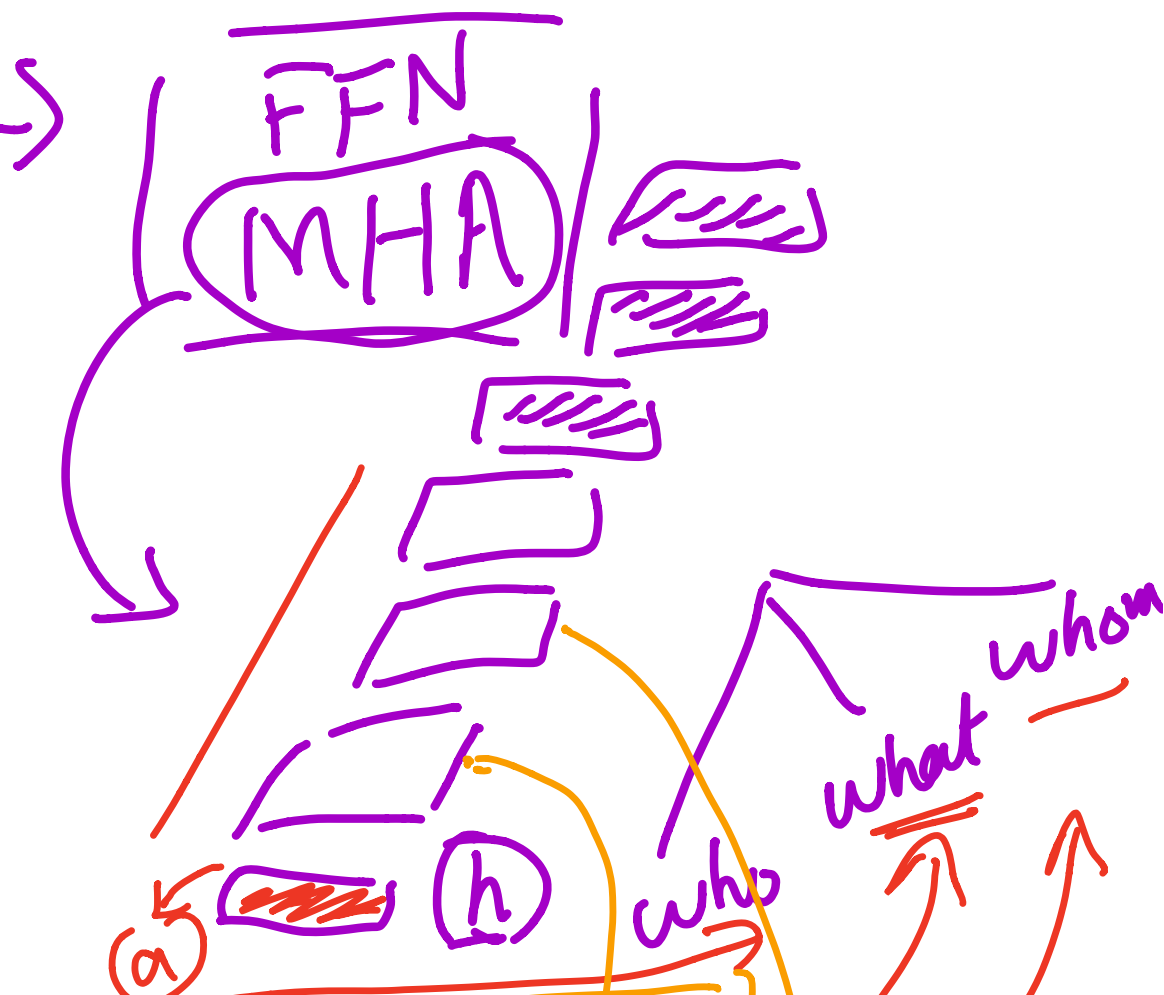
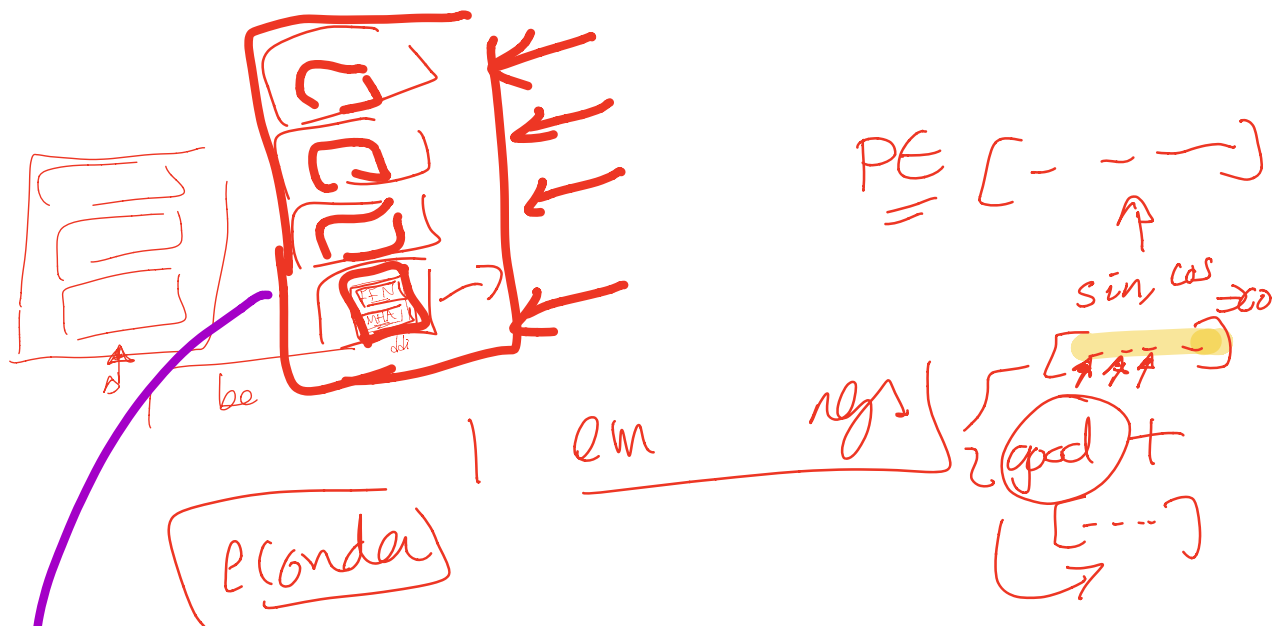
Lecture

- ① - Language models & ^{MT} models
- ② Beam Search
- ③ Multilingual Neural MT

* Recorded a lecture

1:38





John kicked the ball (b) (c)

H₁

Who

Turdan
John
Amy asked a q

H₂

What
pizza
ball
coffee

$H_3 \rightarrow$ (when)

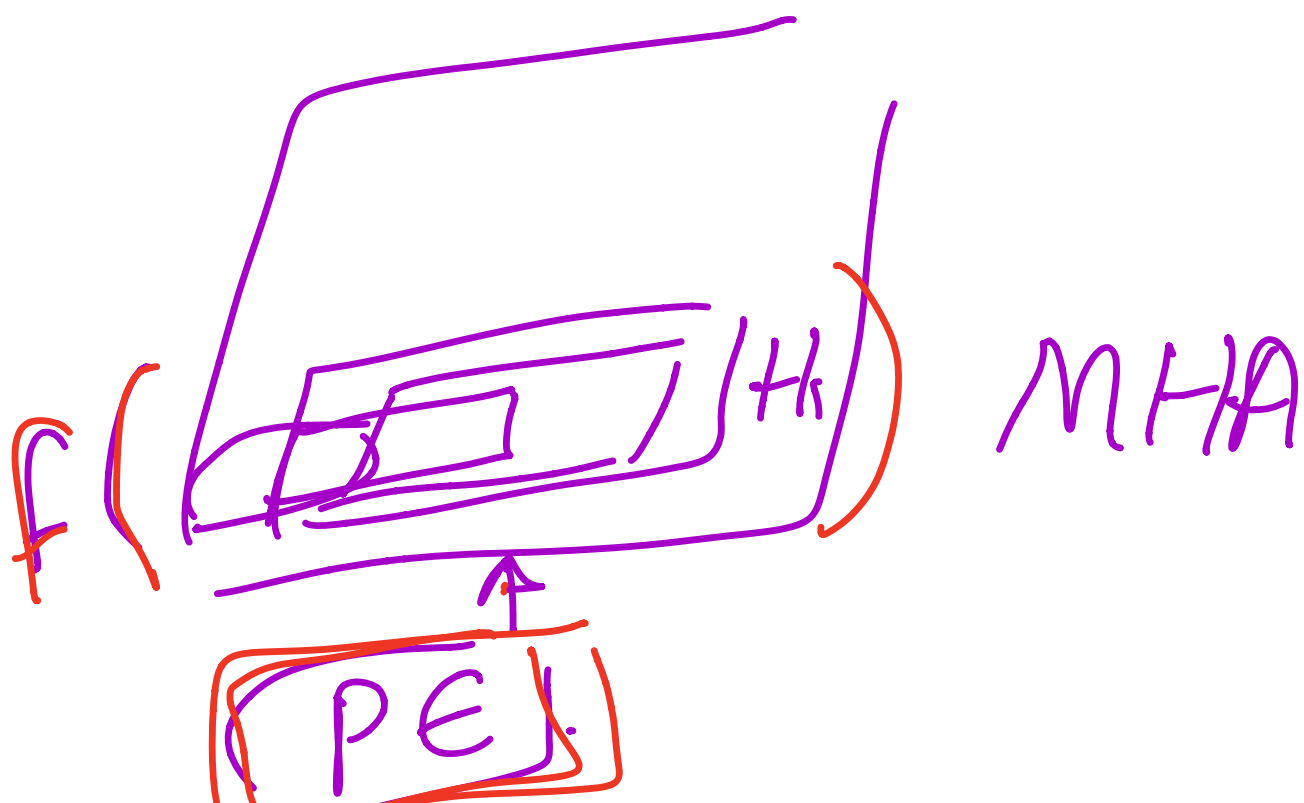
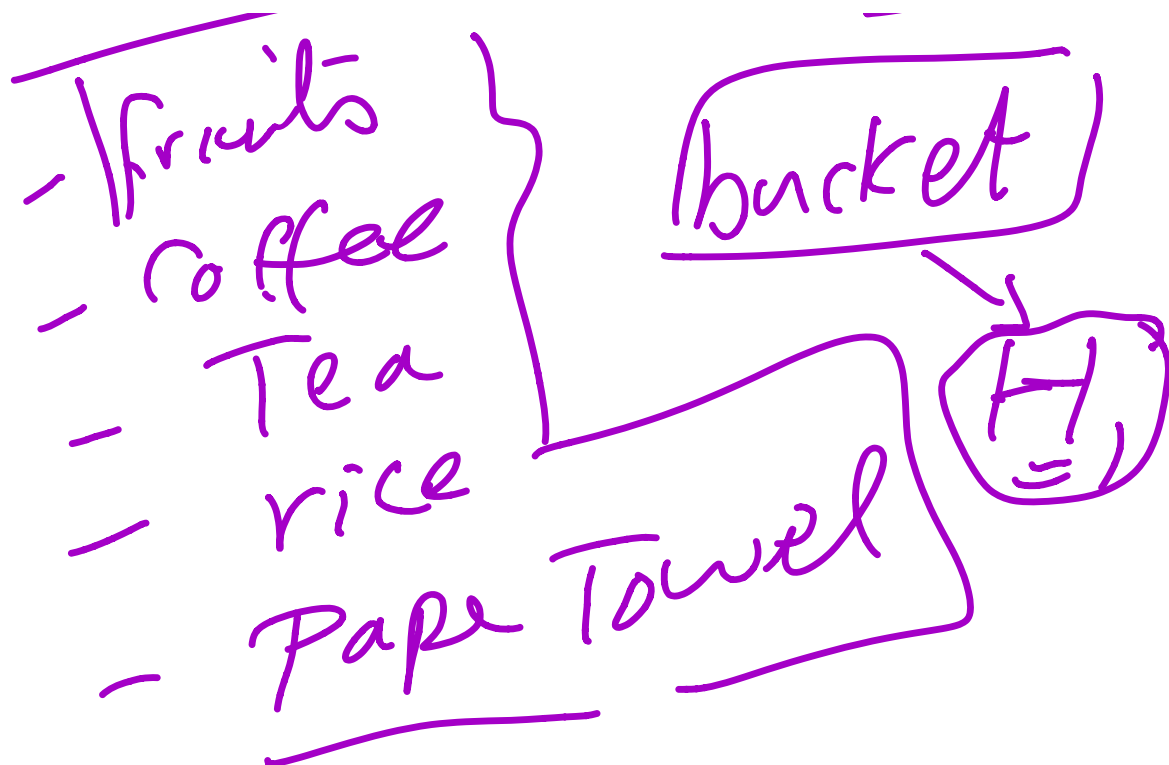
$H_4 \rightarrow$ (How)

who did what
to whom when
& how

($H_1 \dots N$) who

(H_1)





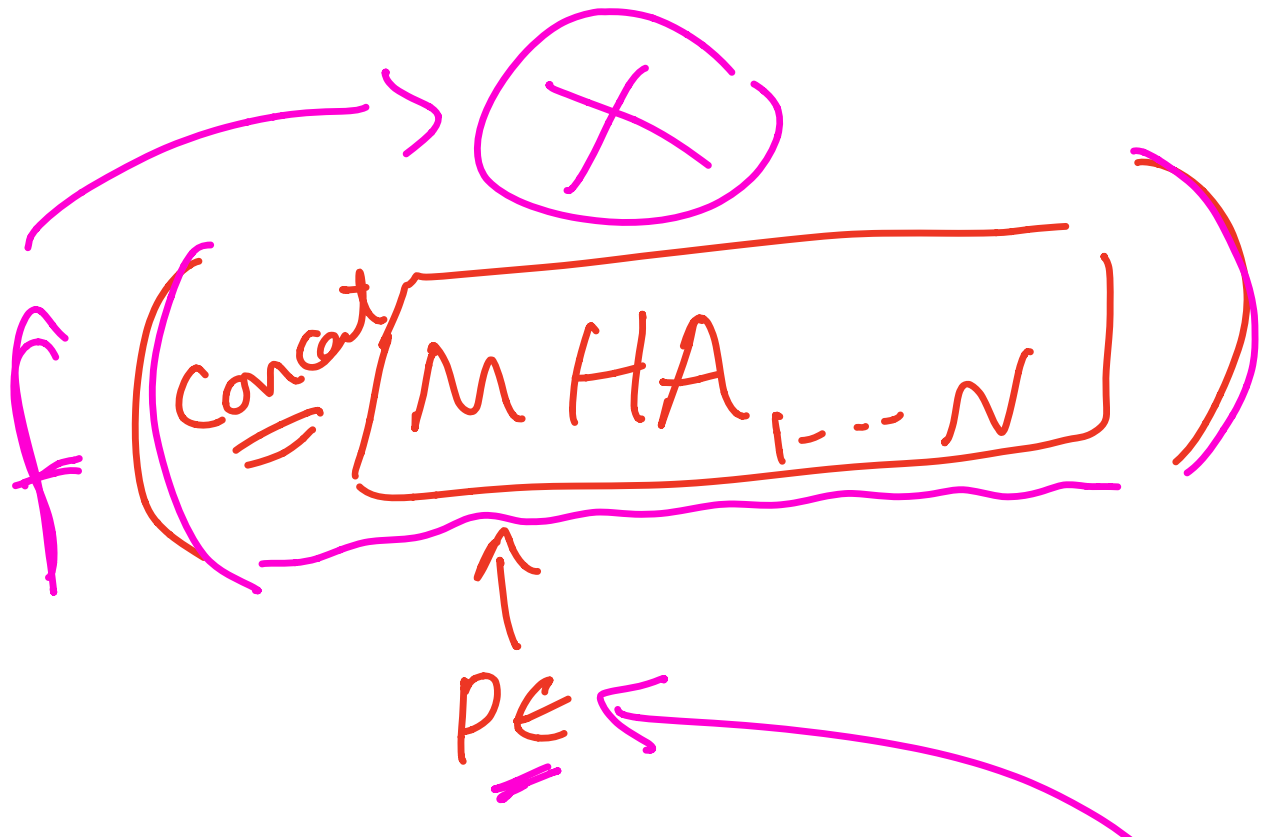
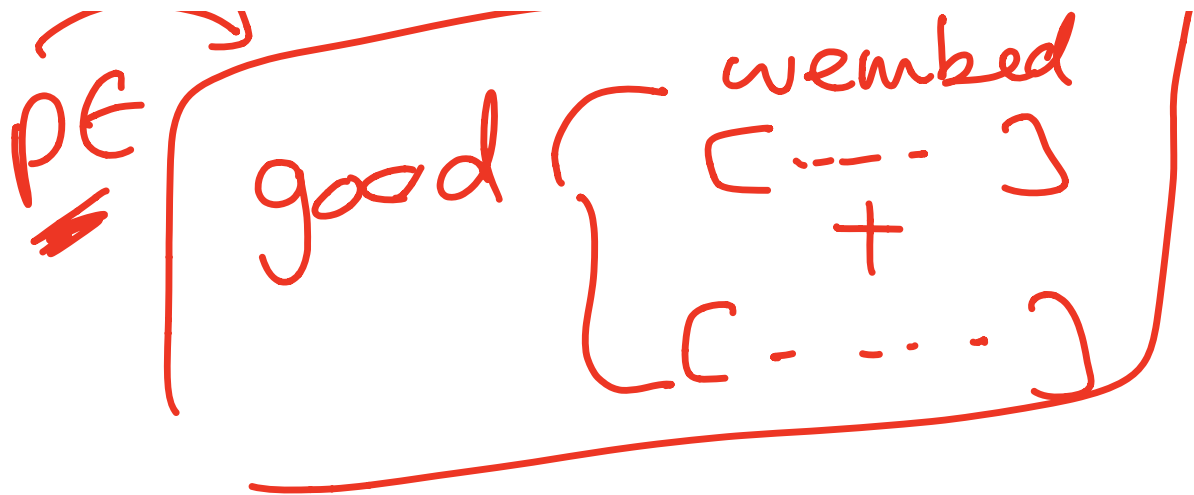
$$H_1 + H_2 + \dots + H_n$$

$$f(\text{concat}(H_{1..N}))$$

$$f(x)$$

$$\frac{MHA}{\uparrow}$$

PE



$$\boxed{f(\underline{x})} + \textcircled{\underline{x}}$$

output + x
~~DES~~

Residual
paper

+ x
 f(layer)
 (x)

✓

$$f(x) + x$$

Stabilize

~~Residual~~ weights

backpropagation

2016

Skip vector

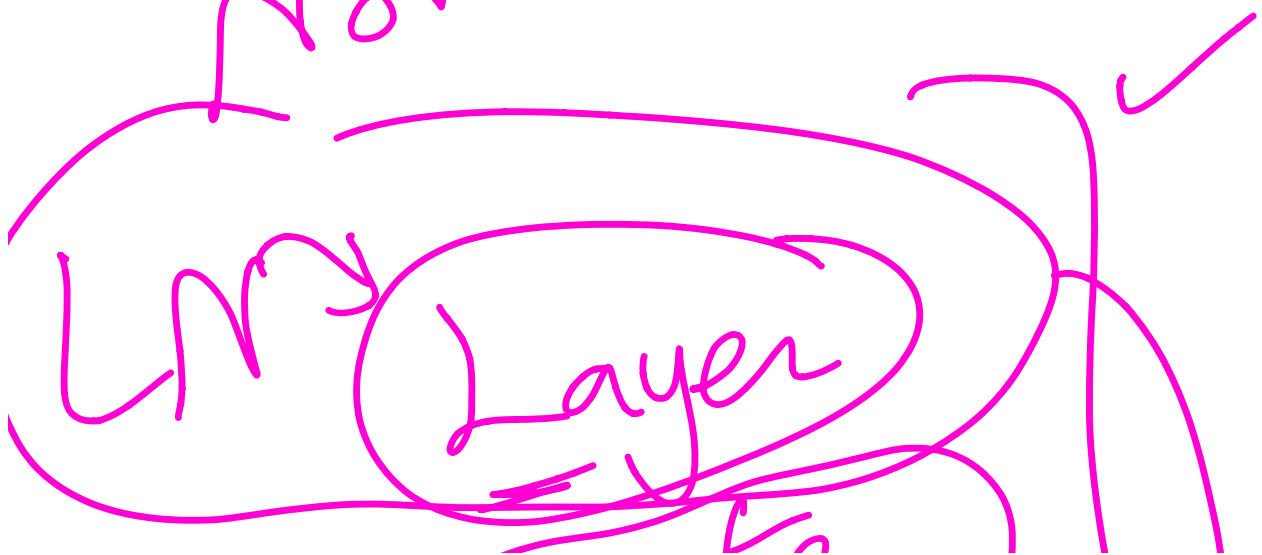
≡

|




Transformer

layer Normalization

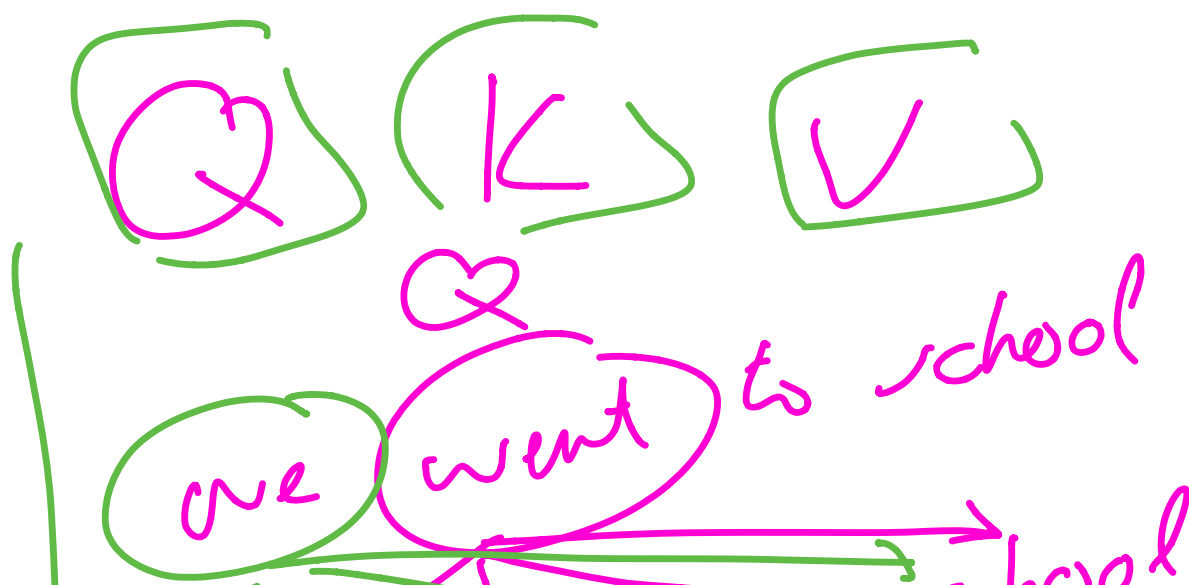
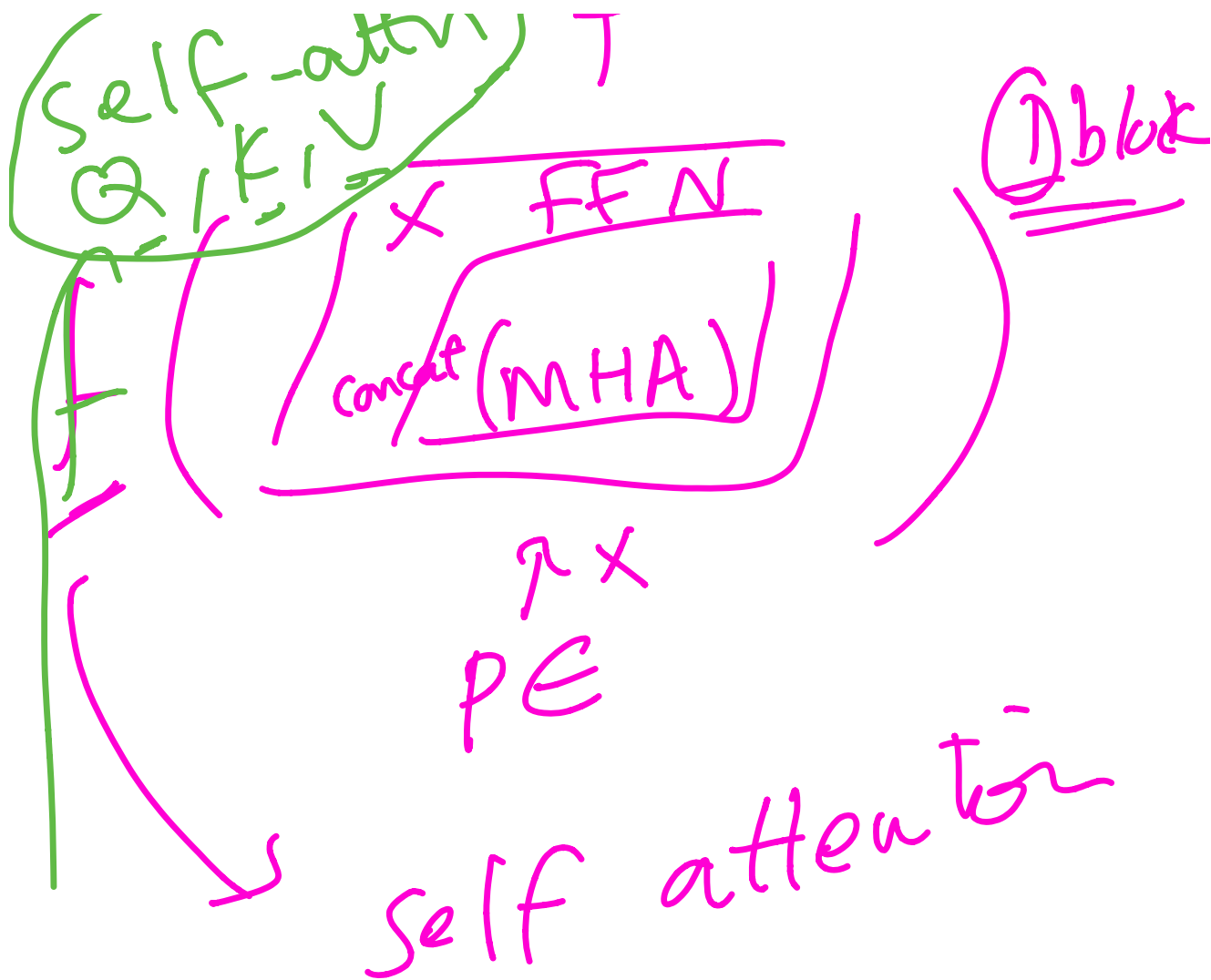


Covariate
Shift



| | |
|--|--|
| | |
| | |
| | |
| | |
| | |
| | |





| ~~we~~ went to school

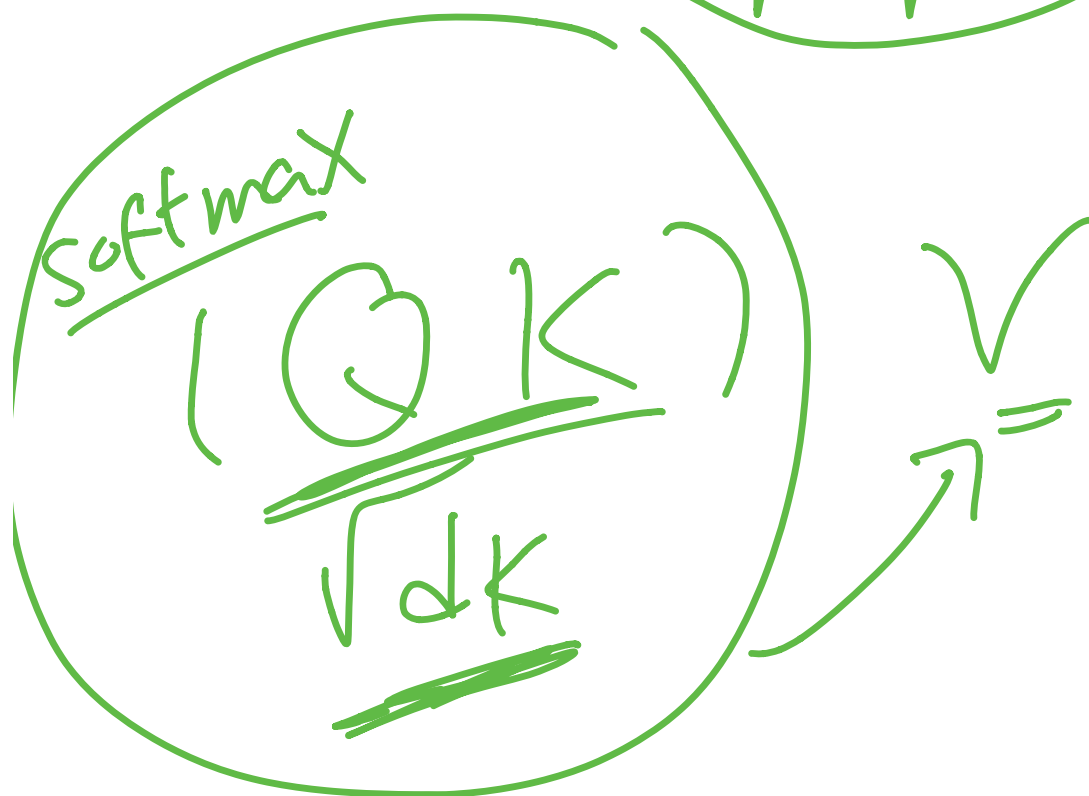
[encoder]

[decoder]

→ + ←
[enc/dec]

Attention is
all you need

Additive
dot-product



g → vector

