① Transformer

$$Z = \text{softmax}\left(\frac{1}{\sqrt{d}}\, Q K^T\right) \cdot V$$

$L \quad M$

$(L \times d')\,(d' \times L)$

$\boxed{L \times L}$

$L \times M$

$\frac{L \times d''}{L \times d''}$

$\frac{L \times d''}{M \times d''}$

$L\,[\,Q = X W^Q$

$M\,[\,K = X W^K$

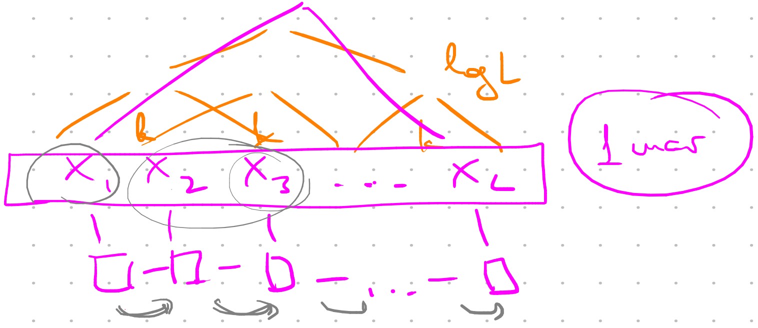
$\quad\;\, V = X W^V$

$L \times d' \qquad L \times d \quad d \times d'$

$K = y \cdot W^K$

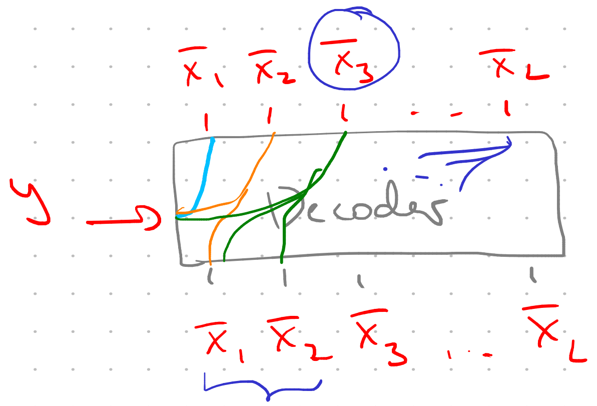
$V = y \cdot W^V$

encoder-decoder attention

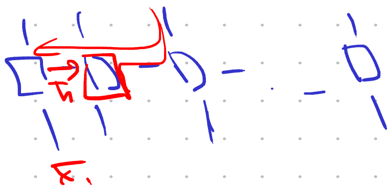


$L \times d \qquad d \times d_{ff} \qquad d_{ff} \times d$

$$X \rightarrow W_1 \rightarrow \text{ReLU} \rightarrow W_2 \rightarrow \text{ReLU} \rightarrow X'$$


$\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ $\bar{x}_L$
y
Decoder
$\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ ... $\bar{x}_L$

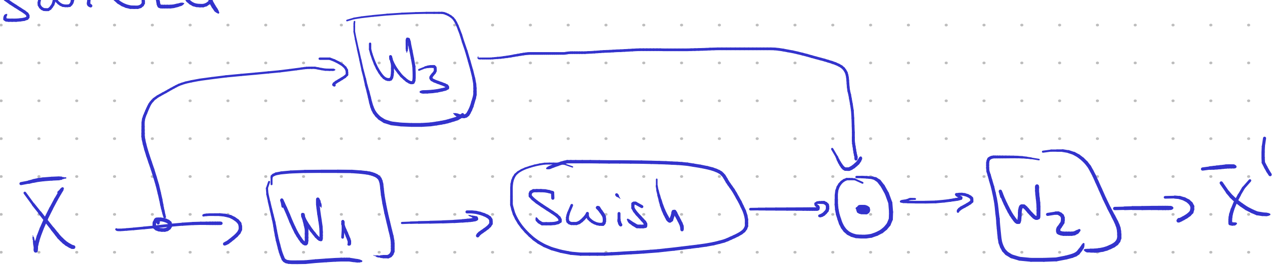
teacher forcing



② The Modern Transformer

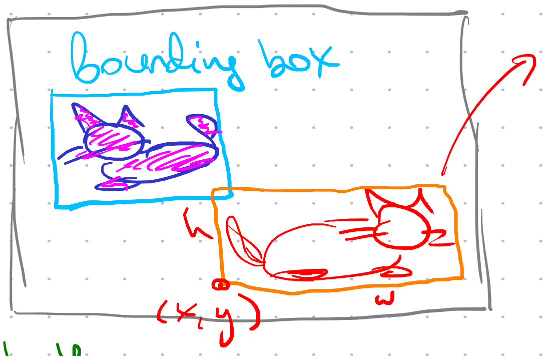
$$\bar{X} \rightarrow \bigcirc\!\!\!\!W_1 - \text{ReLU} - W_2 - \text{ReLU} - \bar{X}'$$

SwiGLU



③ Vision Transformers

④ Object detection



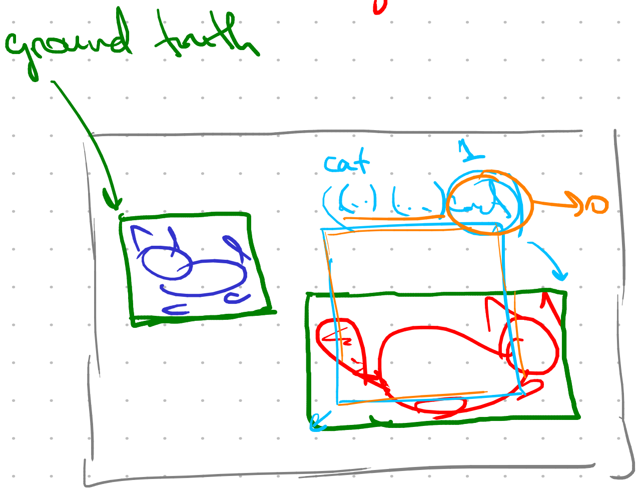
C

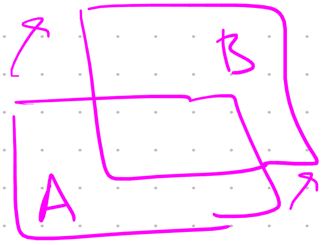
(- - - , p("cat"), - - -)

4

("cat", (x, y, w, h))

(p(c) , (- - -) , conf)
(p(c) , (- - -) , conf)
(p(c) , (- -) , conf)





$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}$$