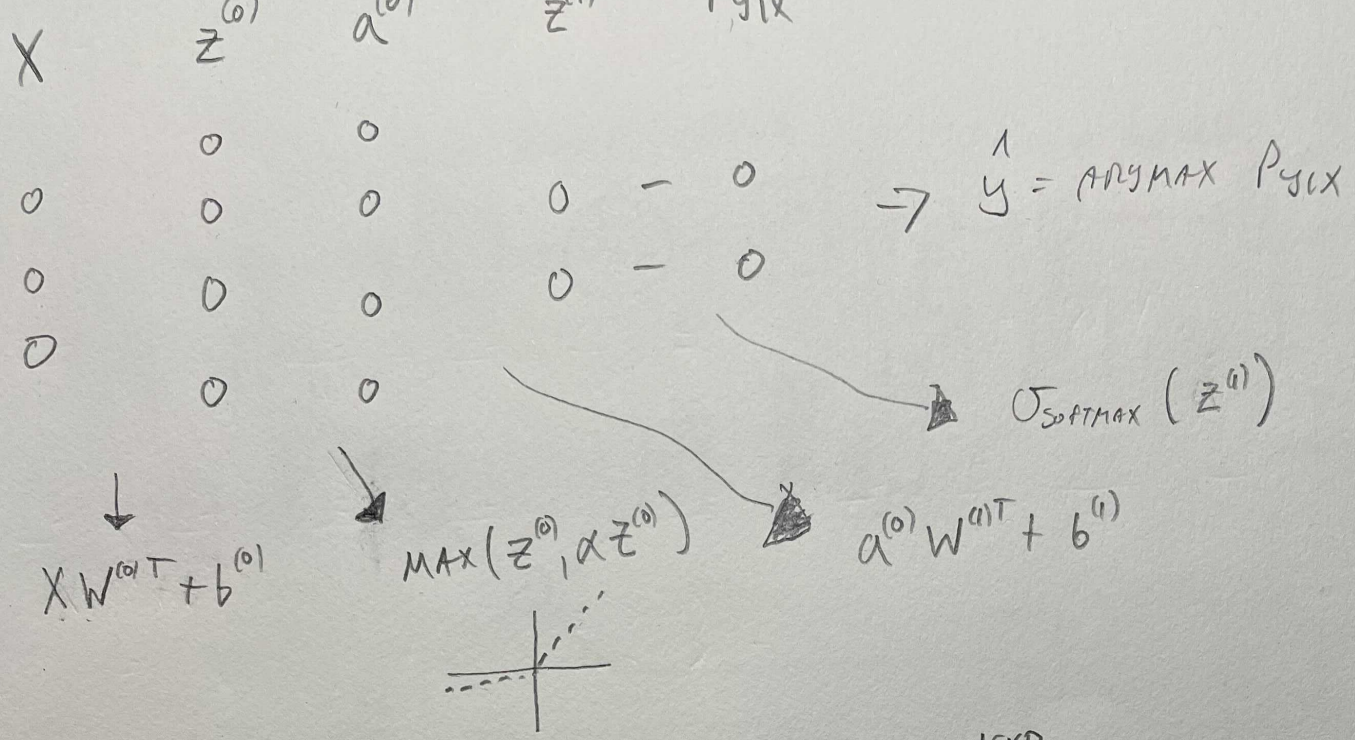
NEURAL NETWORKS AS COMPUTATIONAL GRAPHS



$X \in \mathbb{Z}_+^N$

$z^{(0)} \in \mathbb{R}^D$

$z^{(1)} \in \mathbb{R}^K$

$P_{y|x} \in [0,1]^K$

$W^{(0)} \in \mathbb{R}^{D \times N}$

$b^{(0)} \in \mathbb{R}^D$

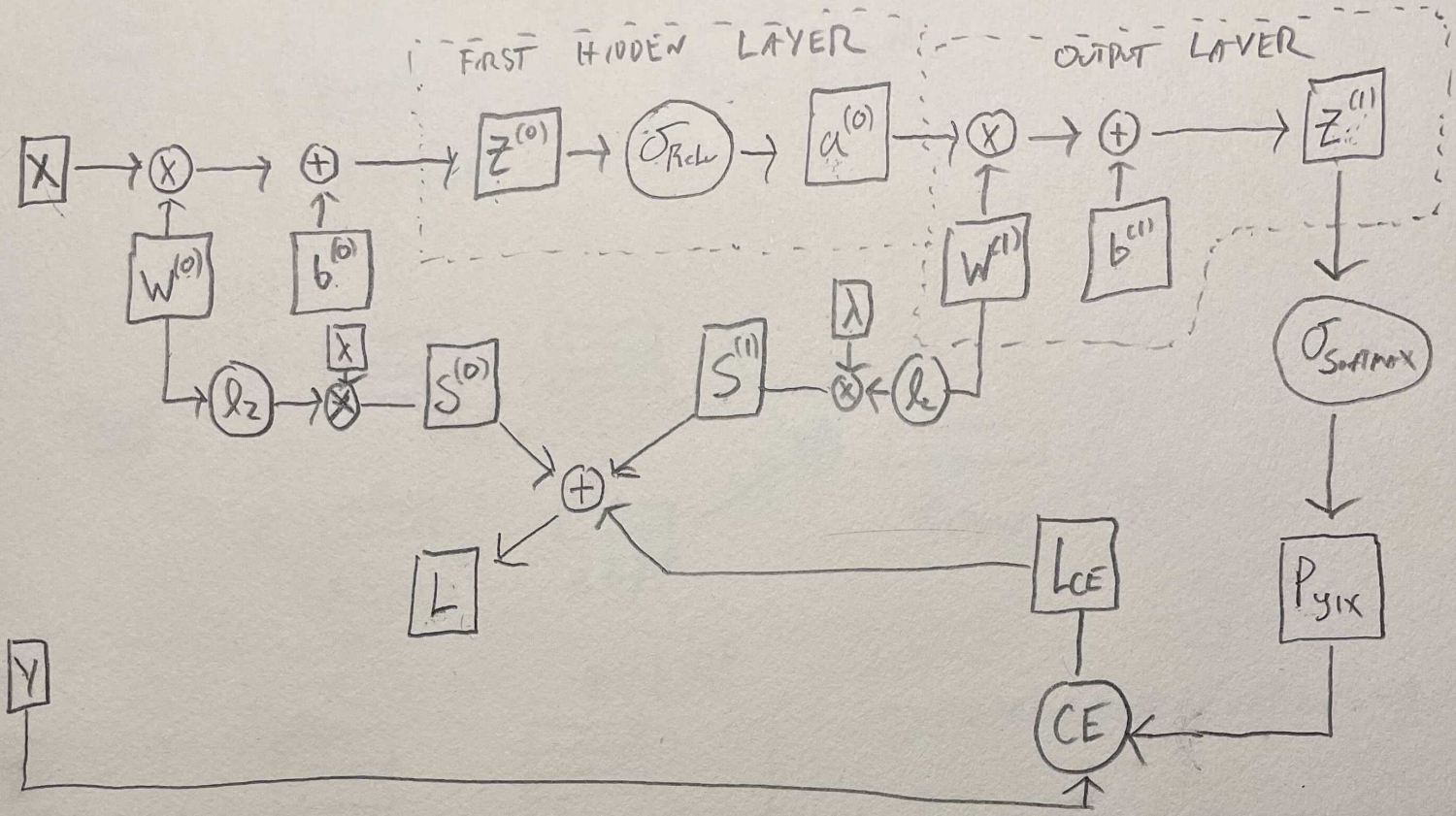
$W^{(1)} \in \mathbb{R}^{K \times D}$

$b^{(1)} \in \mathbb{R}^K$

CHAIN RULE REFRESHER

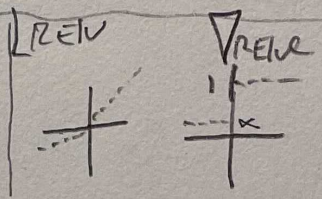
$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial X}$$

Computational Graph



From Lecture 05 - Softmax Regression

$$\frac{\partial L}{\partial z^{(1)}} = P_{y|x} - Y \quad \in \mathbb{R}^K$$



$$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial b^{(1)}} = P_{y|x} - Y \quad \in \mathbb{R}^K$$

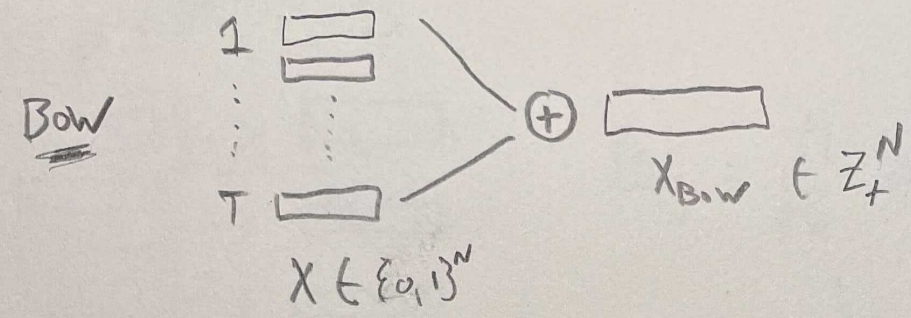
$$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial W^{(1)}} + \frac{\partial L}{\partial S^{(1)}} \cdot \frac{\partial S^{(1)}}{\partial W^{(1)}} = (P_{y|x} - Y) \cdot a^{(0)} + \lambda W^{(1)} \quad \in \mathbb{R}^{K \times D}$$

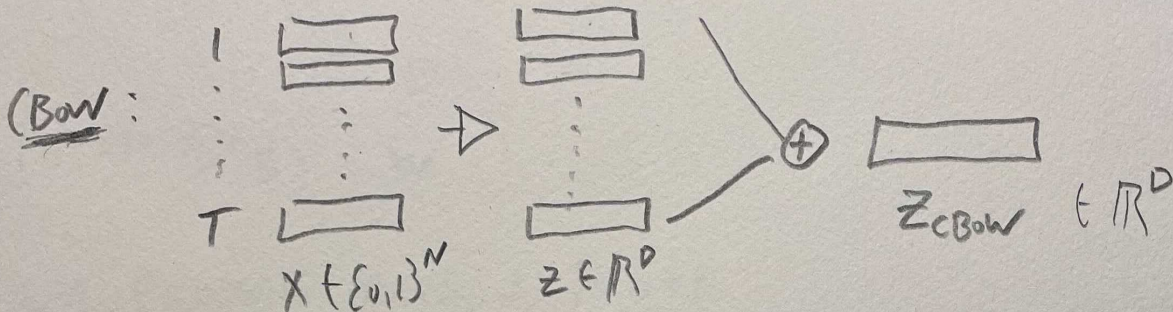
$$\frac{\partial L}{\partial b^{(0)}} = \frac{\partial L}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial a^{(0)}} \cdot \frac{\partial a^{(0)}}{\partial z^{(0)}} \cdot \frac{\partial z^{(0)}}{\partial b^{(0)}} = \left(P_{y|x} - Y\right) \cdot W^{(1)} \cdot \begin{cases} 1 & z^{(0)} >= 0 \\ \alpha & z^{(0)} < 0 \end{cases} \cdot 1 \quad \in \mathbb{R}^D$$

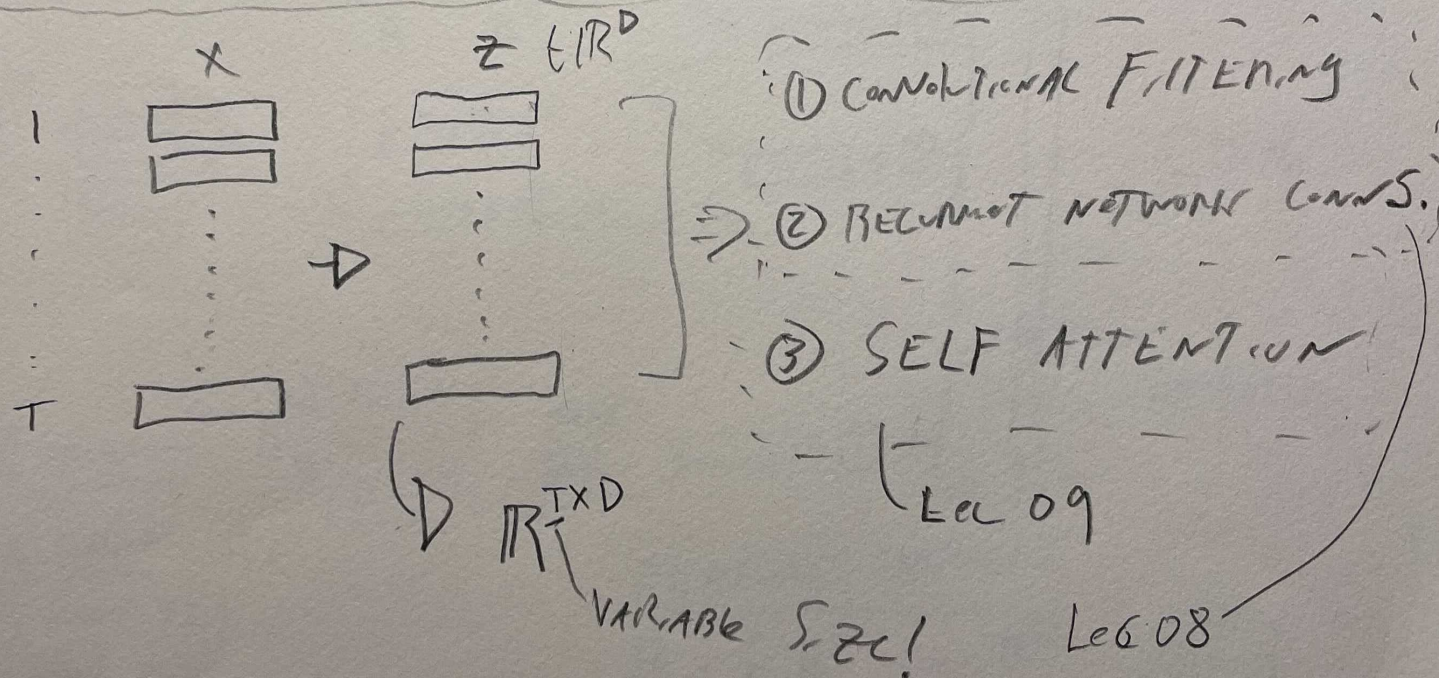
$$\frac{\partial L}{\partial W^{(0)}} = \frac{\partial L}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial a^{(0)}} \cdot \frac{\partial a^{(0)}}{\partial z^{(0)}} \cdot \frac{\partial z^{(0)}}{\partial W^{(0)}} + \frac{\partial L}{\partial S^{(0)}} \cdot \frac{\partial S^{(0)}}{\partial W^{(0)}} = \left(P_{y|x} - Y\right) \cdot W^{(1)} \cdot \begin{cases} 1 & z^{(0)} >= 0 \\ \alpha & z^{(0)} < 0 \end{cases} \cdot X + \lambda W^{(0)}$$

$$\in \mathbb{R}^{D \times N}$$

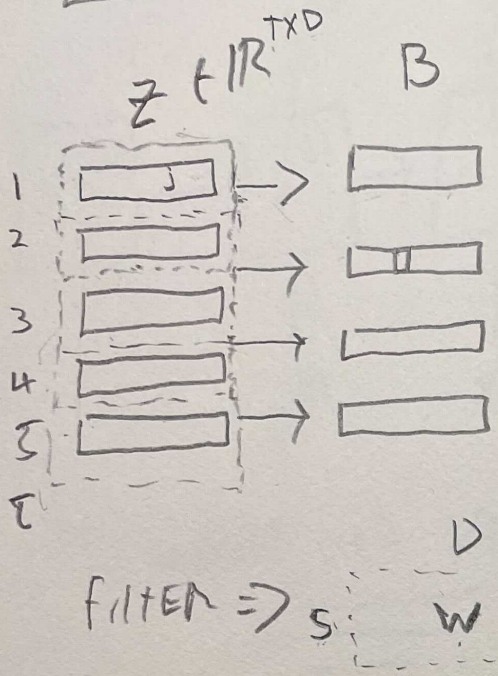
SEQUENCE PROCESSING WITH NN's







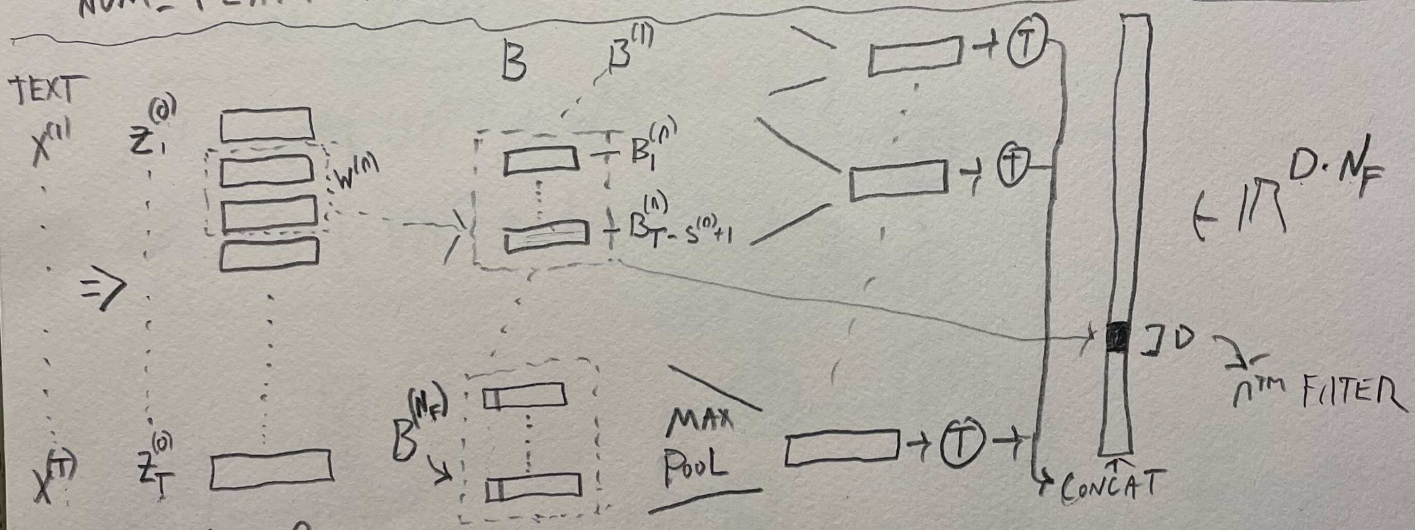
Convolutional Filters



$$B_{tj} = \sum_{t'=1}^{S} W_{t'j} \cdot Z_{t+t'-1,j}$$

$\Rightarrow \forall_{t \in \{1 \dots T-S+1\}}$

NUM_FEATURE_MAPS = $T-S+1$



EMBEDDING LOOKUP

$Z_t^{(i)} \in \mathbb{R}^D$ INDEXED BY j

D = WORD EMB DIM

CONVOLUTIONAL FILTERS

$W = \{ W^{(1)} \dots W^{(n)} \dots W^{(N_F)} \}$

FILTER SIZES

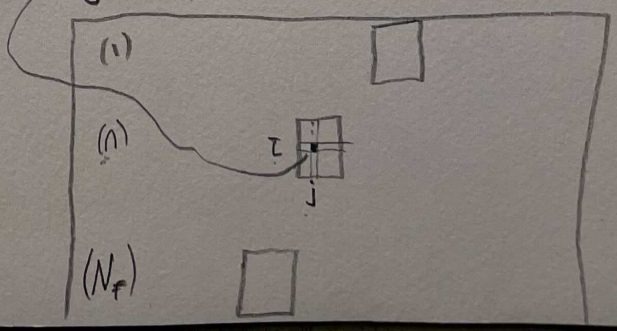
$S = \{ S^{(1)} \dots S^{(n)} \dots S^{(N_F)} \}$

$\Rightarrow W^{(n)} \in \mathbb{R}^{S^{(n)} \times D}$

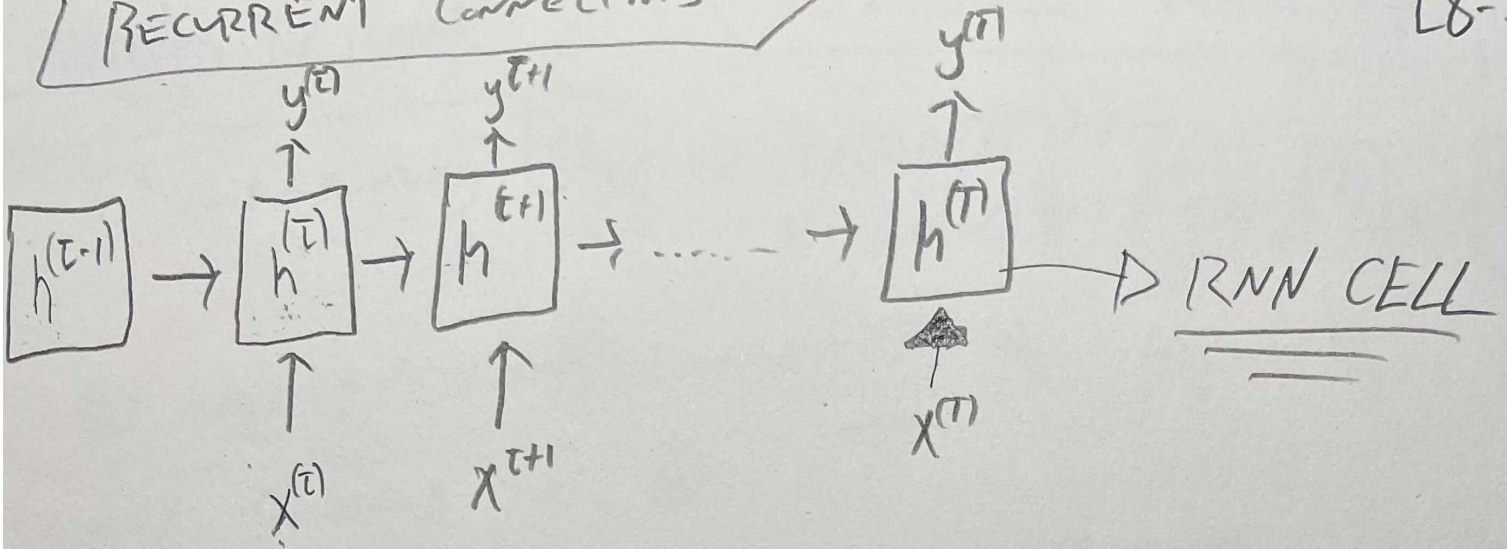
FEATURE MAP

$B = \{ B^{(1)} \dots B^{(n)} \dots B^{(N_F)} \}$

$$B_{tj}^{(n)} = \sum_{t'=1}^{S^{(n)}} W_{t'j}^{(n)} Z_{t,j}^{(0)}$$



RECURRENT CONNECTIONS

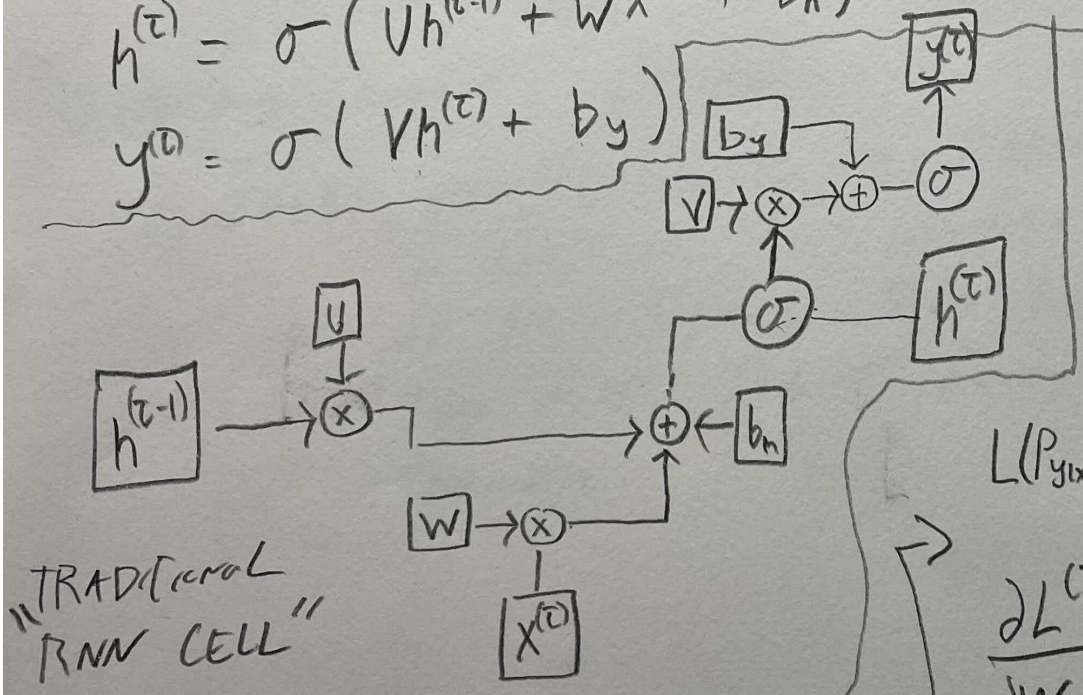


↳ What is $h^{(\tau)}$?

TRADITIONAL RNN

$$h^{(\tau)} = \sigma(Uh^{(\tau-1)} + Wx^{(\tau)} + b_h)$$

$$y^{(\tau)} = \sigma(Vh^{(\tau)} + b_y)$$



"TRADITIONAL RNN CELL"

$$L(P_{y|x}, y) = \sum_{\tau=1}^{T} L(P_{y|x}^{(\tau)}, y^{(\tau)})$$

$$\frac{\partial L^{(T)}}{\partial W} = \sum_{\tau=1}^{T} \frac{\partial L^{(T)}}{\partial W}\Big|_{(\tau)}$$

How are RNNs OPTIMIZED?

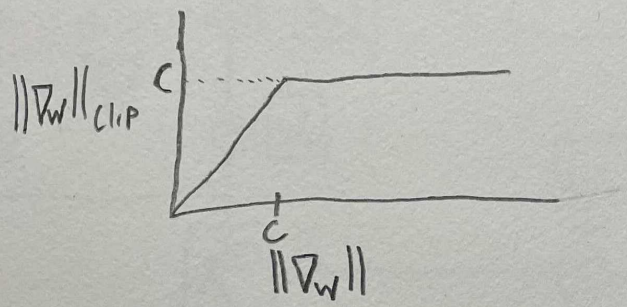
PROBLEMS W/ RNNs

① COMPUTATIONALLY SLOW => BACKPROP THROUGH NETWORK + TIME!

② VANISHING / EXPLODING GRADIENTS

COMMON SOLUTIONS TO ②

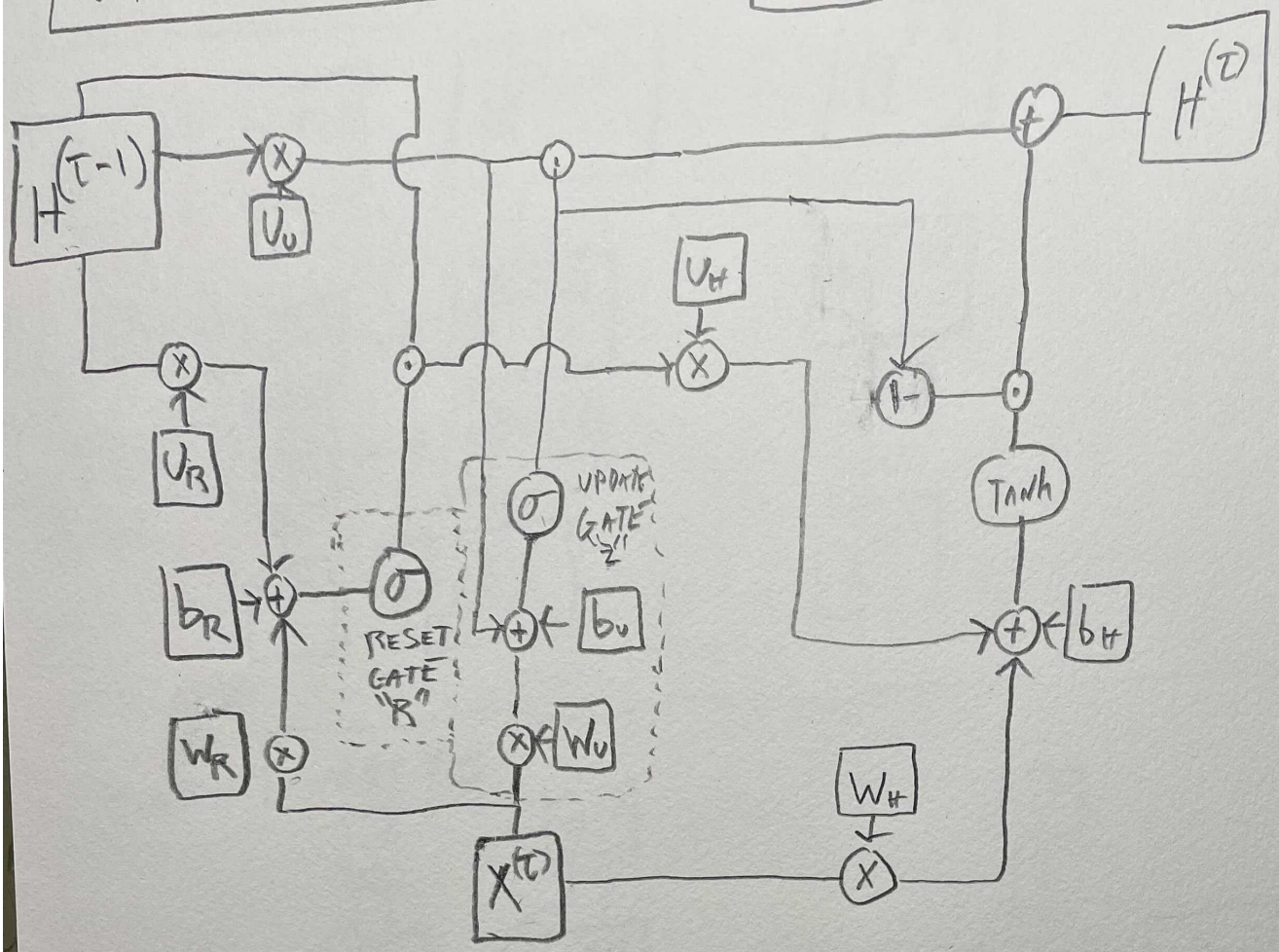
↳ EXPLODING GRADIENTS: GRADIENT CLIPPING



↳ VANISHING GRADIENT: GATING

↳ IDEA: APPLY GATES TO THE HIDDEN STATE TO LIMIT THE FWD/BACKWARD PROPAGATION OF THE SIGNAL SUCH THAT ONLY THE MOST SALIENT FEATURES PASS THROUGH.

GATED RECURRENT UNIT | GRU



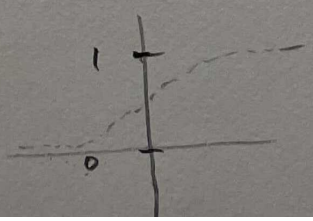
$H \in \mathbb{R}^H$
$X \in \mathbb{R}^D$
$R \in \mathbb{R}^H$
$Z \in \mathbb{R}^H$

$U_H \in \mathbb{R}^{H \times H}$
$U_R \in \mathbb{R}^{H \times H}$,
$U_Z \in \mathbb{R}^{H \times H}$

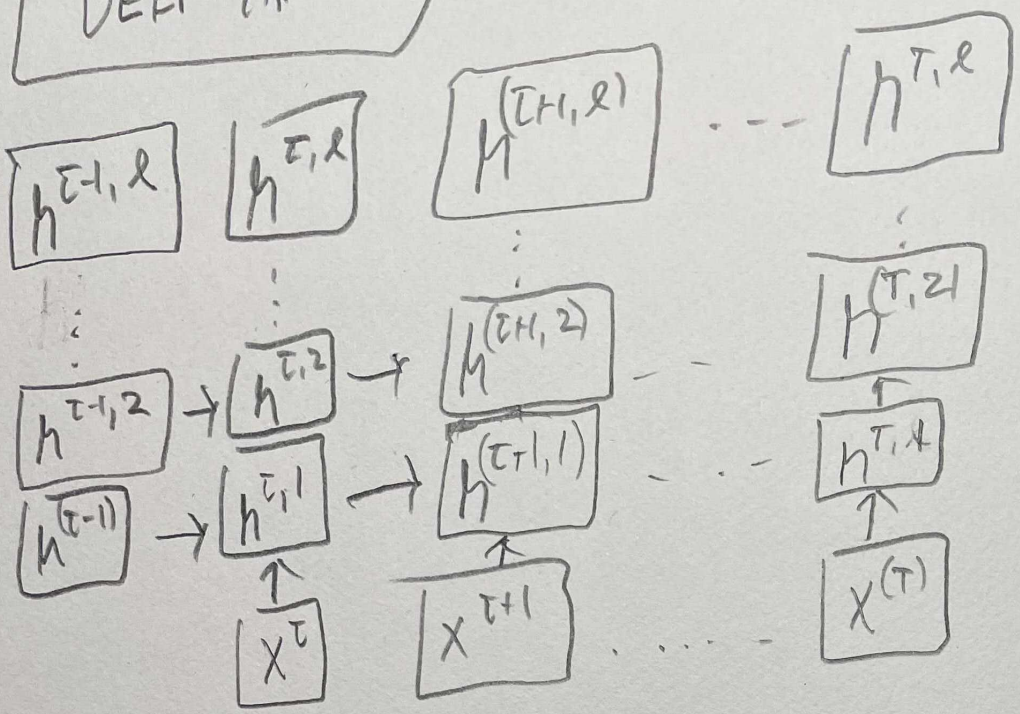
$b_H \in \mathbb{R}^H$
$b_R \in \mathbb{R}^H$,
$b_Z \in \mathbb{R}^H$

$W_H \in \mathbb{R}^{H \times D}$
$W_R \in \mathbb{R}^{H \times D}$
$W_Z \in \mathbb{R}^{H \times D}$

SIGMOID FUNCTION IS IS A GATE:



DEEP RNNS



BIDIRECTIONALITY