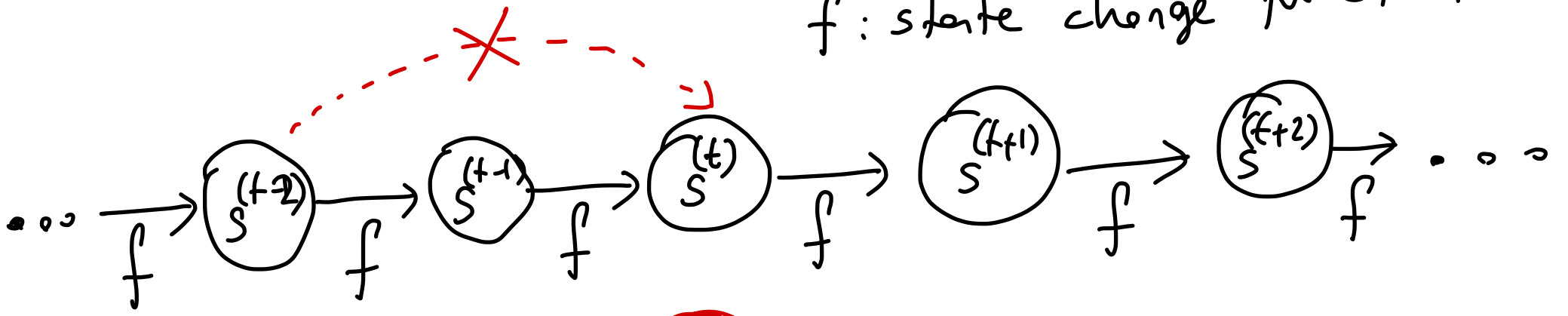
Recurrent Neural Networks (RNNs)

→ Dynamic Systems

$s^{(t)}$: state vector at time t.

f : state change function



$$f(s^{(t)}) = \underbrace{W}_{D \times D} \cdot \underbrace{s^{(t)}}_{D \times 1} + \underbrace{b}_{D \times 1}$$

$$\underbrace{s^{(t+1)}}_{\in \mathbb{R}^{D \times 1}}$$

$$s^{(t+1)} = f(s^{(t)}) \quad \forall t.$$

$s^{(t+1)}$ ↑ next state.

$s^{(t)}$ ↑ current state

parameters : $[W, b]$

We are going to learn the parameters of f function.

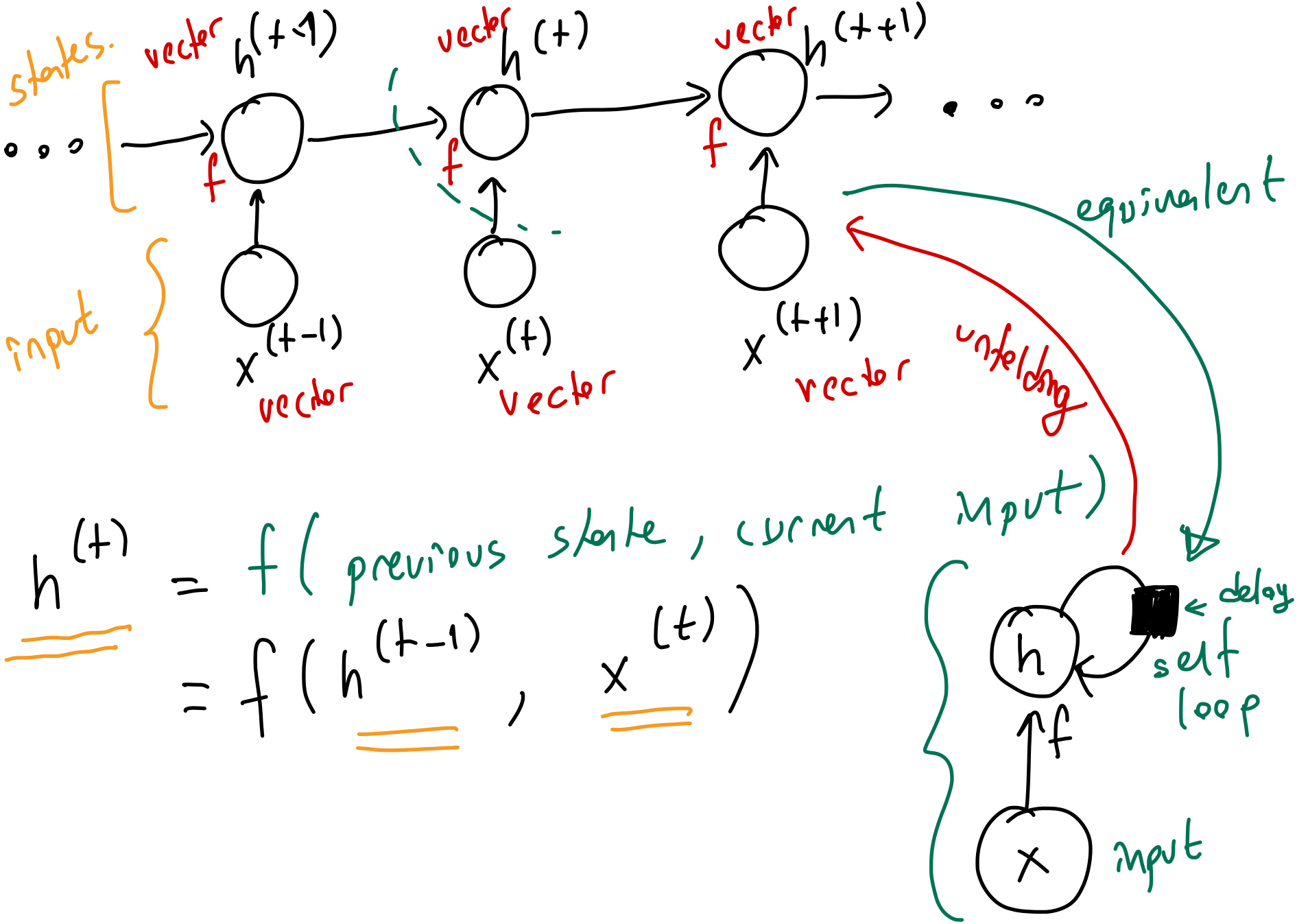
$$s^{(2)} = f(s^{(1)}) \Rightarrow s^{(2)} = W \cdot s^{(1)} + b$$

$$s^{(1)} = f(s^{(0)}) \Rightarrow s^{(1)} = W \cdot s^{(0)} + b$$

$$s^{(2)} = W\left[W s^{(0)} + b\right] + b$$

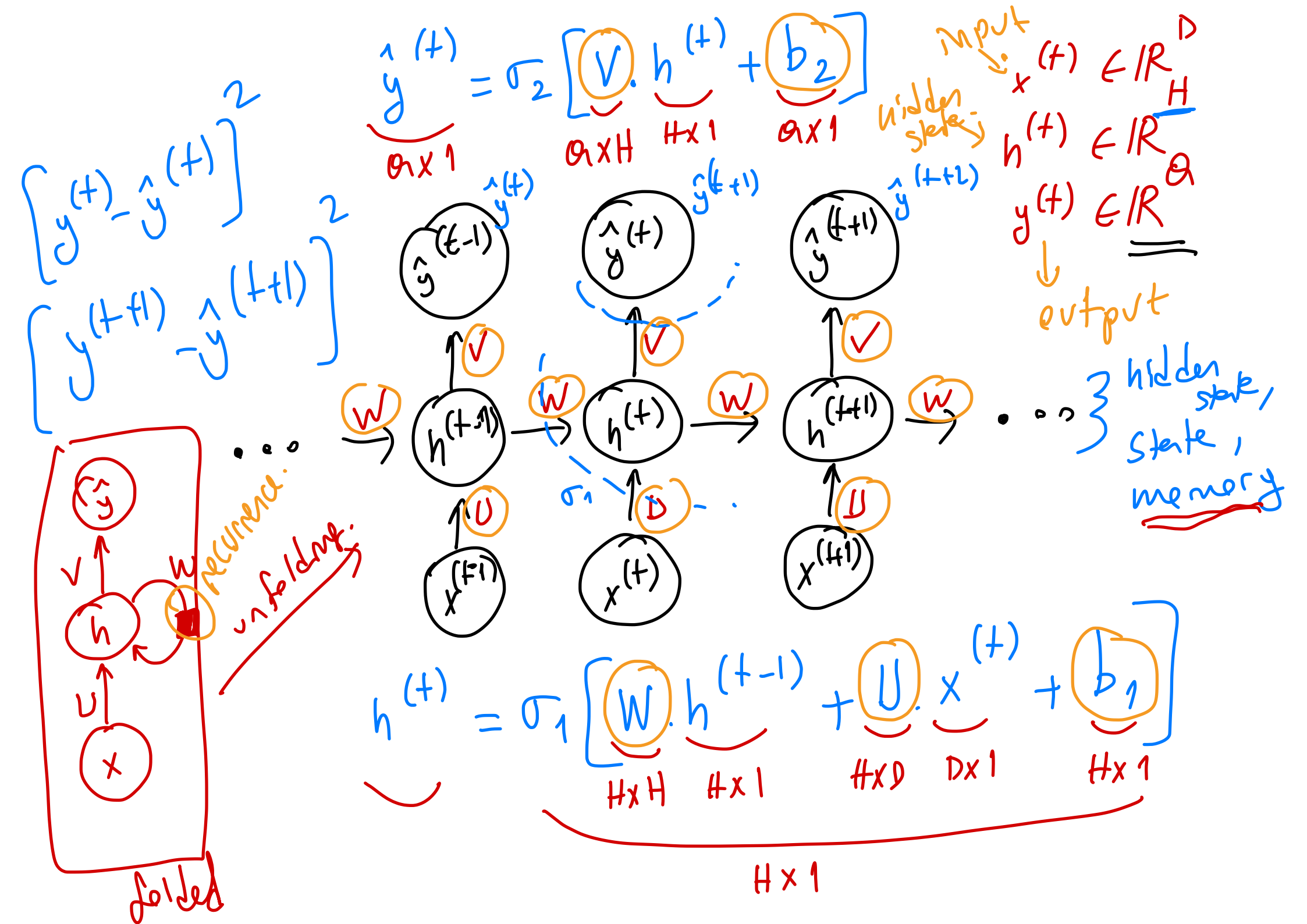
$$= \underbrace{W \cdot W}_{\tilde{W}} \cdot s^{(0)} + \underbrace{W \cdot b + b}_{\tilde{b}}$$

$$= \tilde{W} \cdot s^{(0)} + \tilde{b}$$



$$h^{(t)} = f(\text{previous state}, \text{current input})$$

$$= f(h^{(t-1)}, x^{(t)})$$

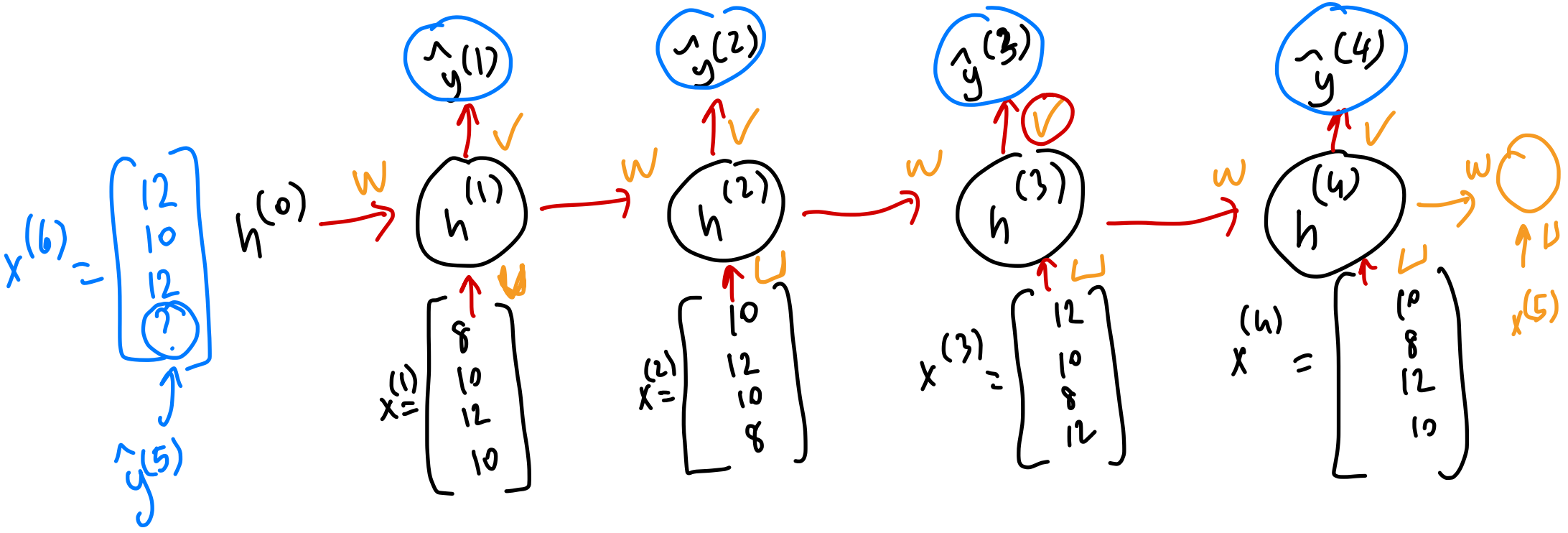

$\hat{y}^{(t)} = \sigma_2\left[V \cdot h^{(t)} + b_2\right]$
$\hat{y}^{(t)}$: $Q \times 1$
V : $Q \times H$
$h^{(t)}$: $H \times 1$
b_2 : $Q \times 1$
input
$x^{(t)} \in \mathbb{R}^D$
hidden state
$h^{(t)} \in \mathbb{R}^H$
$y^{(t)} \in \mathbb{R}^Q$
output
$\left[y^{(t)} - \hat{y}^{(t)}\right]^2$
$\left[y^{(t+1)} - \hat{y}^{(t+1)}\right]^2$
$\hat{y}^{(t-1)}$
$\hat{y}^{(t)}$
$\hat{y}^{(t+1)}$
$\hat{y}^{(t+2)}$
$h^{(t-1)}$
$h^{(t)}$
$h^{(t+1)}$
$x^{(t-1)}$
$x^{(t)}$
$x^{(t+1)}$
V
W
U
σ_1
hidden state, state, memory
recurrence
unfolding
folded
$h^{(t)} = \sigma_1\left[W \cdot h^{(t-1)} + U \cdot x^{(t)} + b_1\right]$
W : $H \times H$
$h^{(t-1)}$: $H \times 1$
U : $H \times D$
$x^{(t)}$: $D \times 1$
b_1 : $H \times 1$
$H \times 1$

t=1	t=2	t=3	t=4	t=5	t=6	t=5	t=8	?	?
8	10	12	10	8	12	10	12		

window size = 4

$$x^{(1)} = \begin{bmatrix} 8 \\ 10 \\ 12 \\ 10 \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} 10 \\ 12 \\ 10 \\ 8 \end{bmatrix} \quad x^{(3)} = \begin{bmatrix} 12 \\ 10 \\ 8 \\ 12 \end{bmatrix} \quad x^{(4)} = \begin{bmatrix} 10 \\ 8 \\ 12 \\ 10 \end{bmatrix} \quad X^{(5)} = \begin{bmatrix} 8 \\ 12 \\ 10 \\ 12 \end{bmatrix}$$

$$y^{(1)} = [8] \quad y^{(2)} = [12] \quad y^{(3)} = [10] \quad y^{(4)} = [12] \quad y^{(5)} = ?$$



$$\text{Loss [MSE]} = \frac{1}{4}\left[(y^{(1)} - \hat{y}^{(1)})^2 + \cdots + (y^{(4)} - \hat{y}^{(4)})^2 \right]$$

$$h^{(5)} = \sigma_1 \left[\underbrace{W}_{20\times 20} \cdot \underbrace{h^{(4)}}_{20\times 1} + \underbrace{U}_{20\times 4} \cdot \underbrace{x^{(5)}}_{4\times 1} + b_1 \right]$$

$$\underbrace{\hat{y}^{(5)}}_{1\times 1} = \sigma_2 \left[V \cdot \underbrace{h^{(5)}}_{20\times 1} + b_2 \right]$$

assume.

$H = 20$

$(H \times H)$ (Hxwindowsize)

$\hat{y}^{(6)} = ?$

521 parameters

$W \in \mathbb{R}^{20\times 20}$ $(H \times H)$

$U \in \mathbb{R}^{20\times 4}$ $(\theta \times H)$ ←

$V \in \mathbb{R}^{1\times 20}$

$b_1 \in \mathbb{R}^{20\times 1}$ (H)

$b_2 \in \mathbb{R}^{1}$ (θ) ←

8, 10, 12, 10, 8, 12, 10, 12, ?, ...

$$x^{(1)} = \begin{bmatrix} 8 \\ 10 \\ 12 \\ 10 \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} 10 \\ 12 \\ 10 \\ 8 \end{bmatrix} \quad x^{(3)} = \begin{bmatrix} 12 \\ 10 \\ 8 \\ 12 \end{bmatrix}$$

$$y^{(1)} = \begin{bmatrix} 8 \\ 12 \end{bmatrix} \quad y^{(2)} = \begin{bmatrix} 12 \\ 10 \end{bmatrix} \quad y^{(3)} = \begin{bmatrix} 10 \\ 12 \end{bmatrix}$$

$$x^{(4)} = \begin{bmatrix} 10 \\ 8 \\ 12 \\ 10 \end{bmatrix}$$

$$y^{(4)} = \begin{bmatrix} 12 \\ ? \end{bmatrix}$$

20×20

20×4

2×20

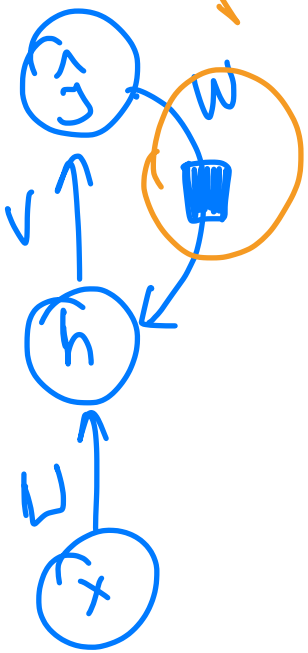
20×1

2×1

542 parameters

folded version

recurrence



$$h^{(t)} = \sigma_1\left[W \cdot \hat{y}^{(t-1)} + U \cdot x^{(t)} + b_1 \right]$$

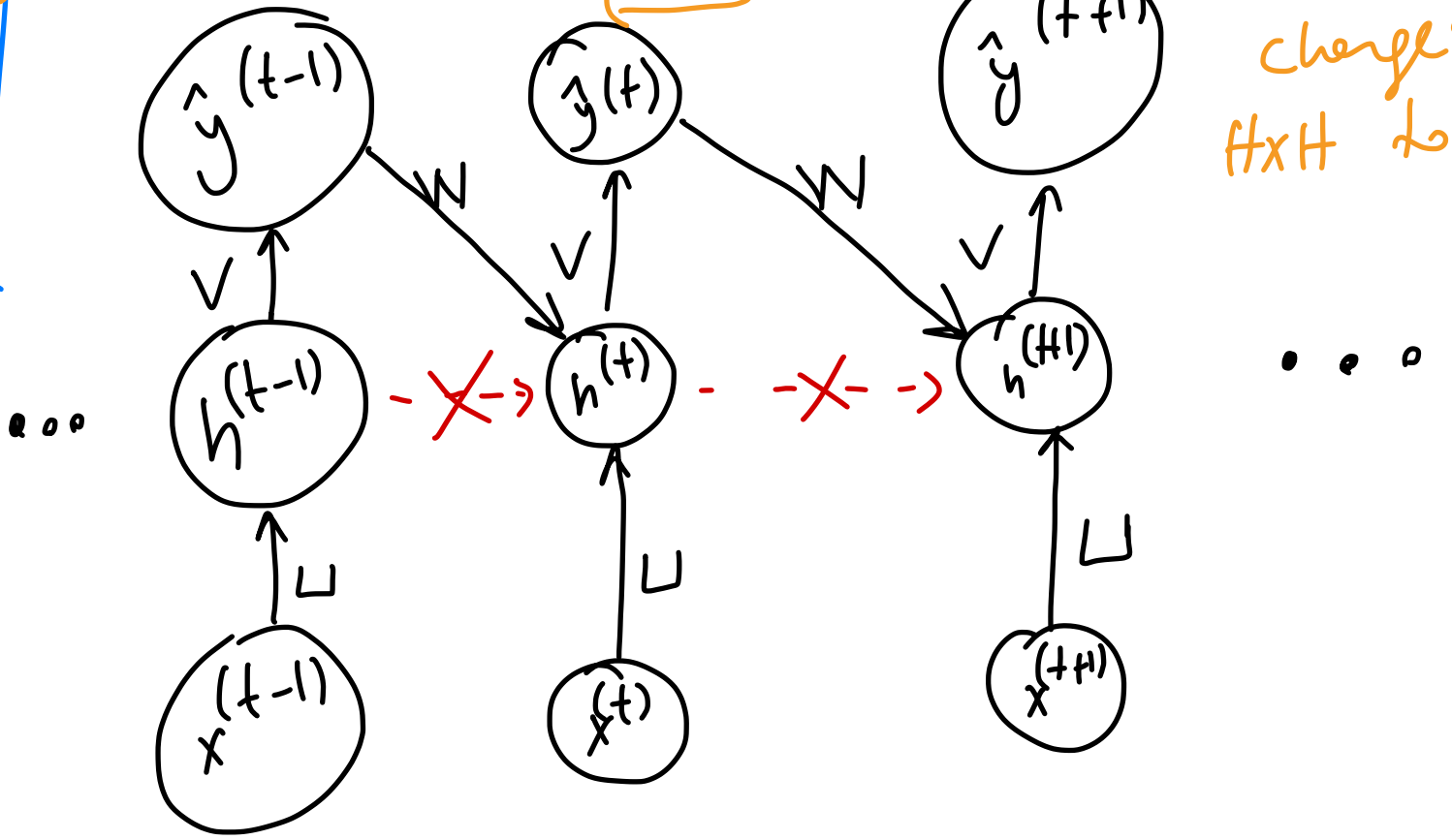
different

some

some

$H \times Q$ $Q \times 1$

W's size changes from $H \times H$ to $H \times Q$.



$$\hat{y}^{(t)} = \sigma_2\left[V \cdot h^{(t)} + b_2 \right]$$

Sequence Input / Single Output Models

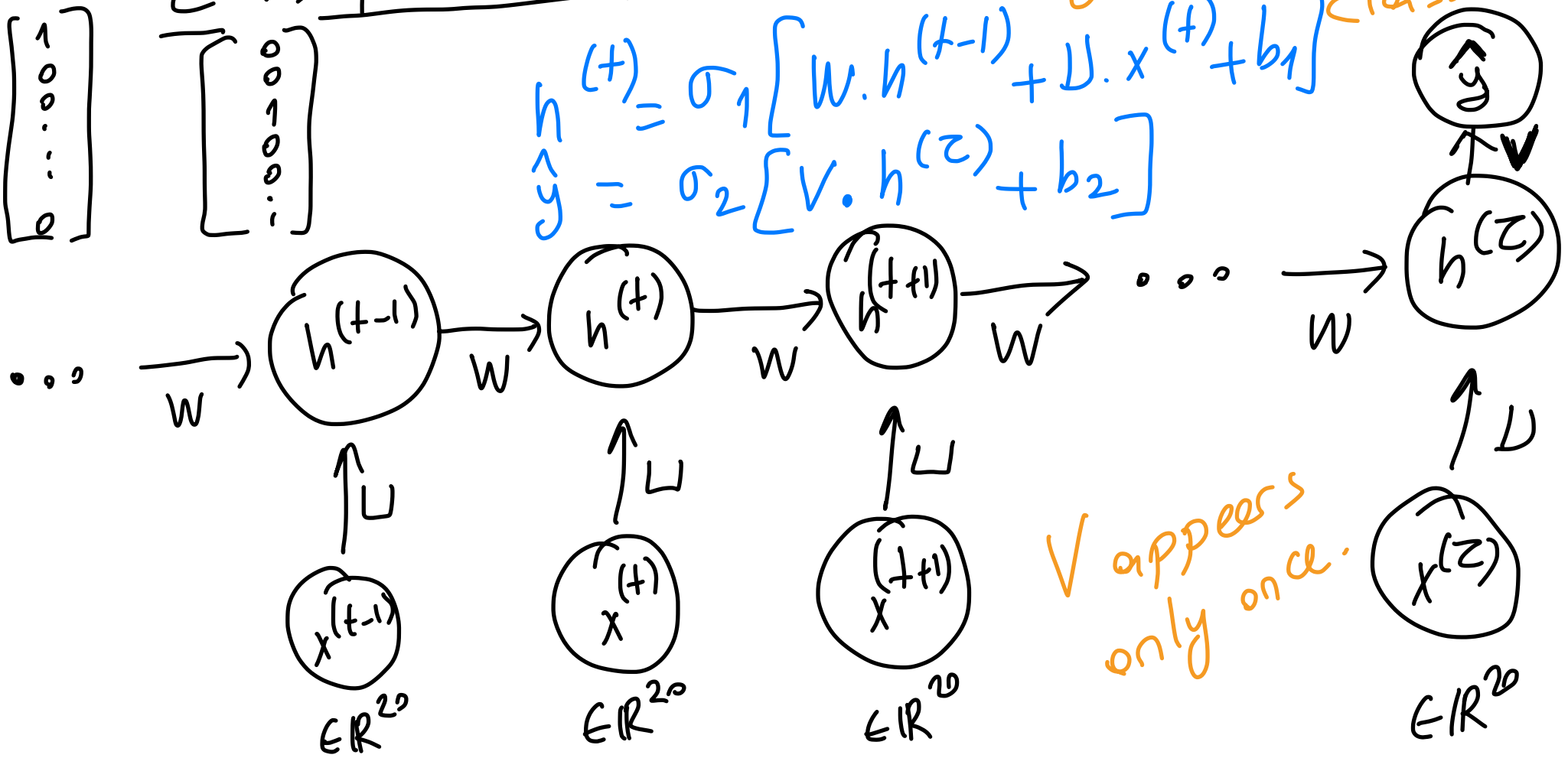
↳ Sentiment Analysis

↳ signal (time series) classification

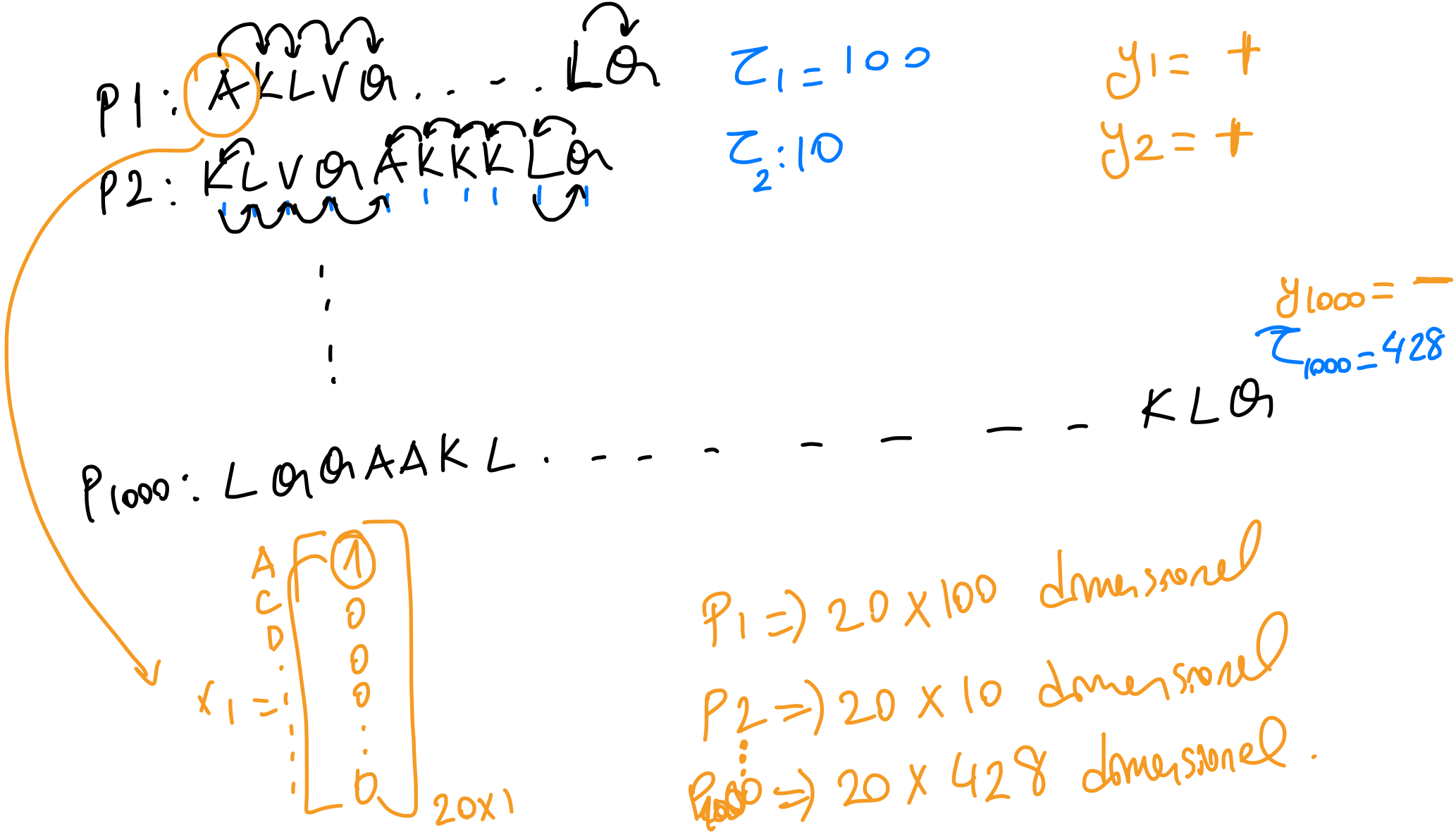
τ : sequence length.

$$h^{(t)} = \sigma_1\left[W \cdot h^{(t-1)} + U \cdot x^{(t)} + b_1\right]$$

$$\hat{y} = \sigma_2\left[V \cdot h^{(\tau)} + b_2\right]$$

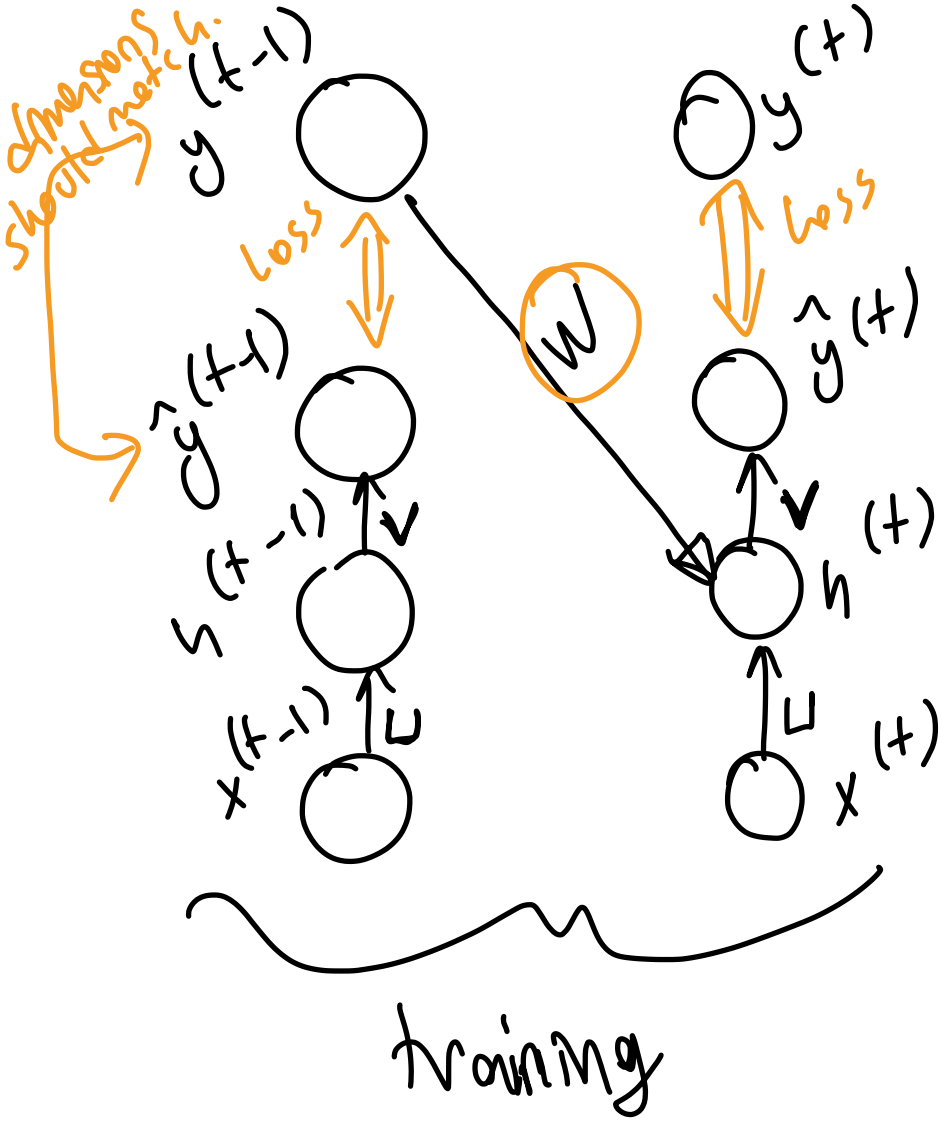


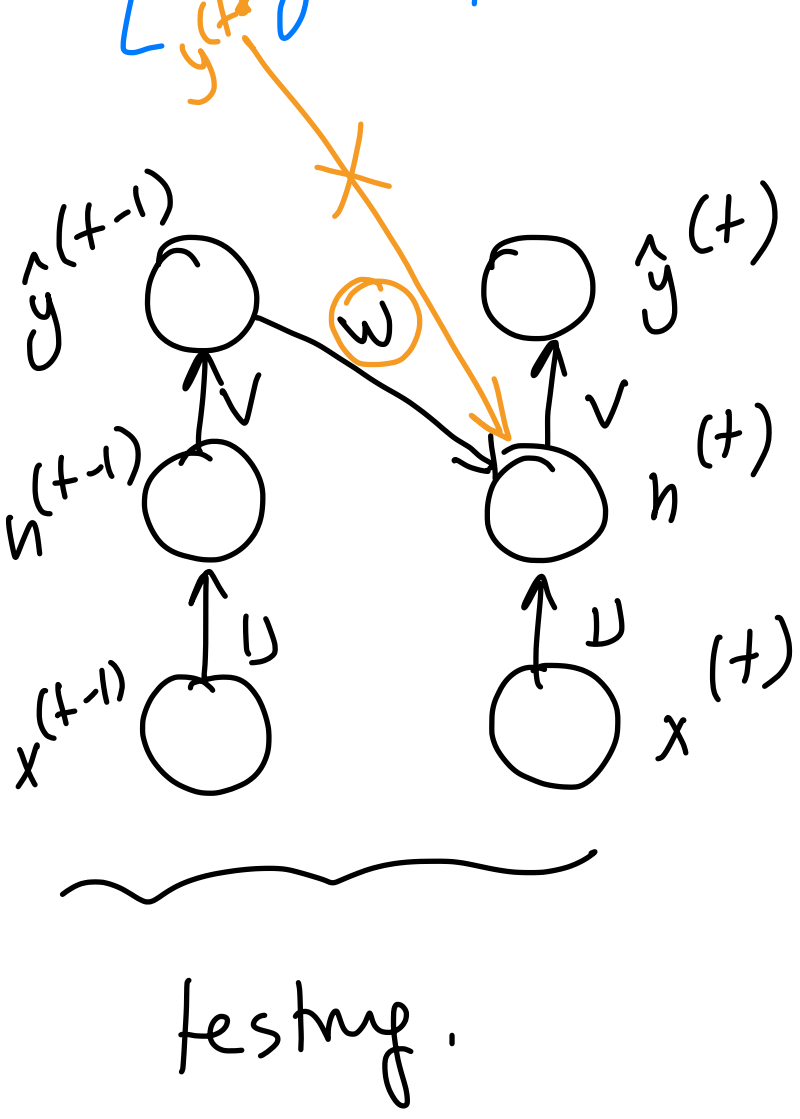
V appears only once.

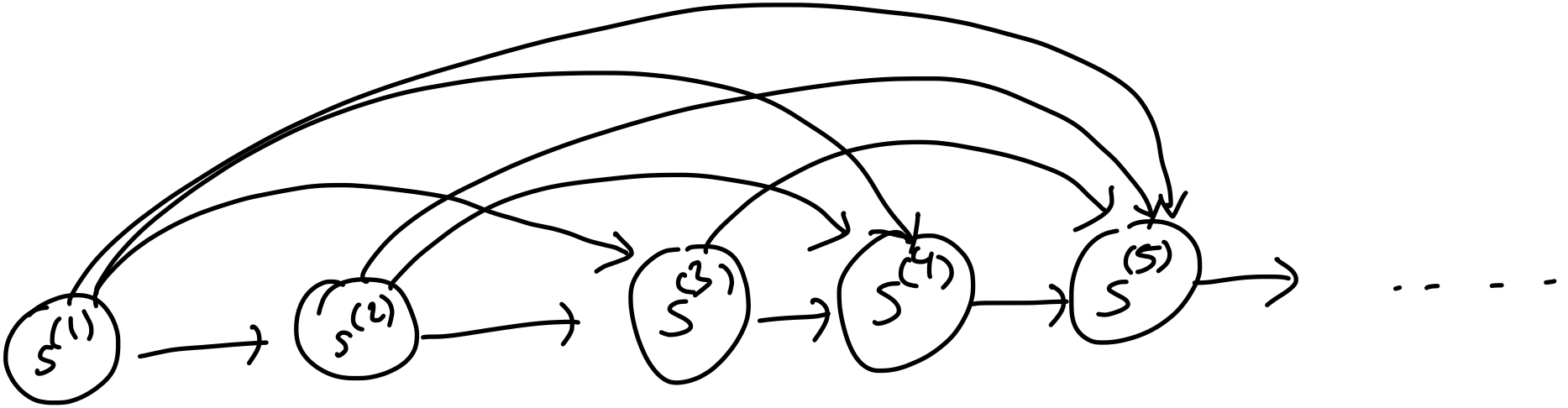

P1: AKLVQ . . . . LQ
P2: KLVQAKKKLQ
$\tau_1 = 100$
$\tau_2 : 10$
$y_1 = +$
$y_2 = +$
$y_{1000} = -$
$\tau_{1000} = 428$
P_{1000}: LQQAAKL . . . . . KLQ
A
C
D
$x_1 =$
1
0
0
0
0
20x1
$P_1 \Rightarrow 20 \times 100$ dimensional
$P_2 \Rightarrow 20 \times 10$ dimensional
$P_{1000} \Rightarrow 20 \times 428$ dimensional.

Teacher Forcing

$$h^{(t)} = \sigma_1\left[W y^{(t-1)} + U \cdot x^{(t)} + b_1\right]$$





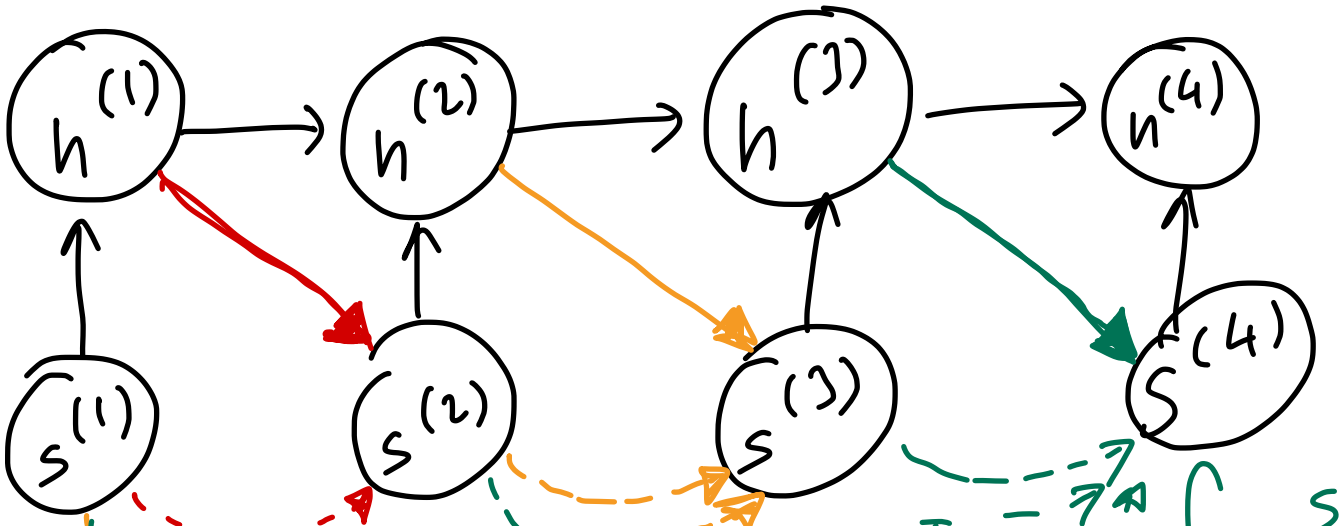


$$s^{(2)} = f(s^{(1)})$$

$$s^{(3)} = f(s^{(2)}, s^{(1)})$$

$$s^{(4)} = f(s^{(3)}, s^{(2)}, s^{(1)})$$

$$s^{(5)} = f(s^{(4)}, s^{(3)}, s^{(2)}, s^{(1)})$$



$h^{(1)}$ is a function of $s^{(1)}$.

$h^{(2)}$ is a function of $h^{(1)}$ and $s^{(2)}$

$s^{(3)}$ is a function of $h^{(2)}$

$s^{(4)}$ is a function of $h^{(3)}$.

$h^{(3)}$ is a function of $h^{(2)}$ and $s^{(3)}$.

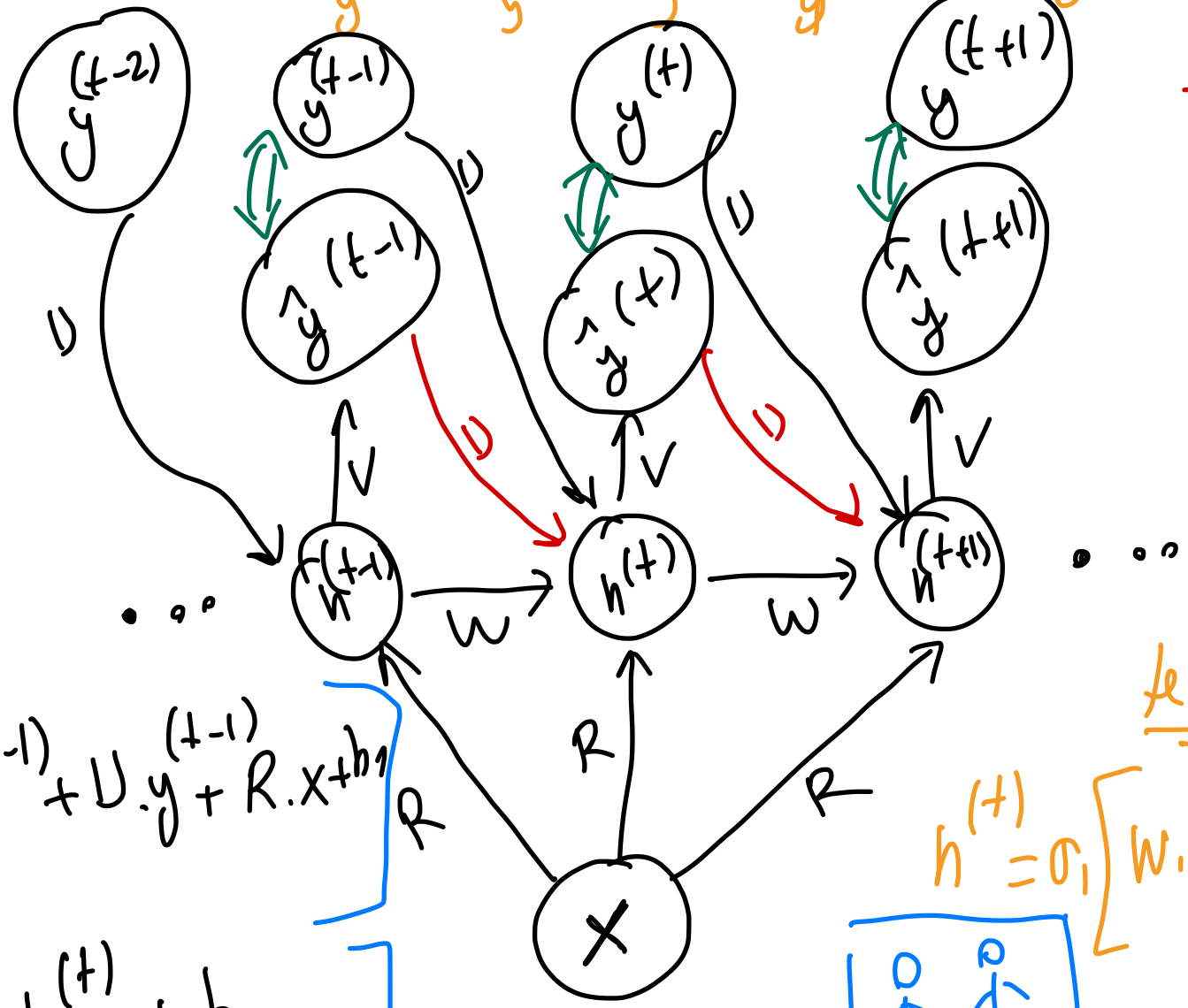
Vector to Sequence

image captioning.

Two kids are playing basketball

$y^{(1)}$ $y^{(2)}$ $y^{(3)}$ $y^{(4)}$ $y^{(5)}$ $y^{(6)}$

→ only for testing.



$$h^{(t)} = \sigma_1\left[W.h^{(t-1)} + U.y^{(t-1)} + R.x + b_1\right]$$

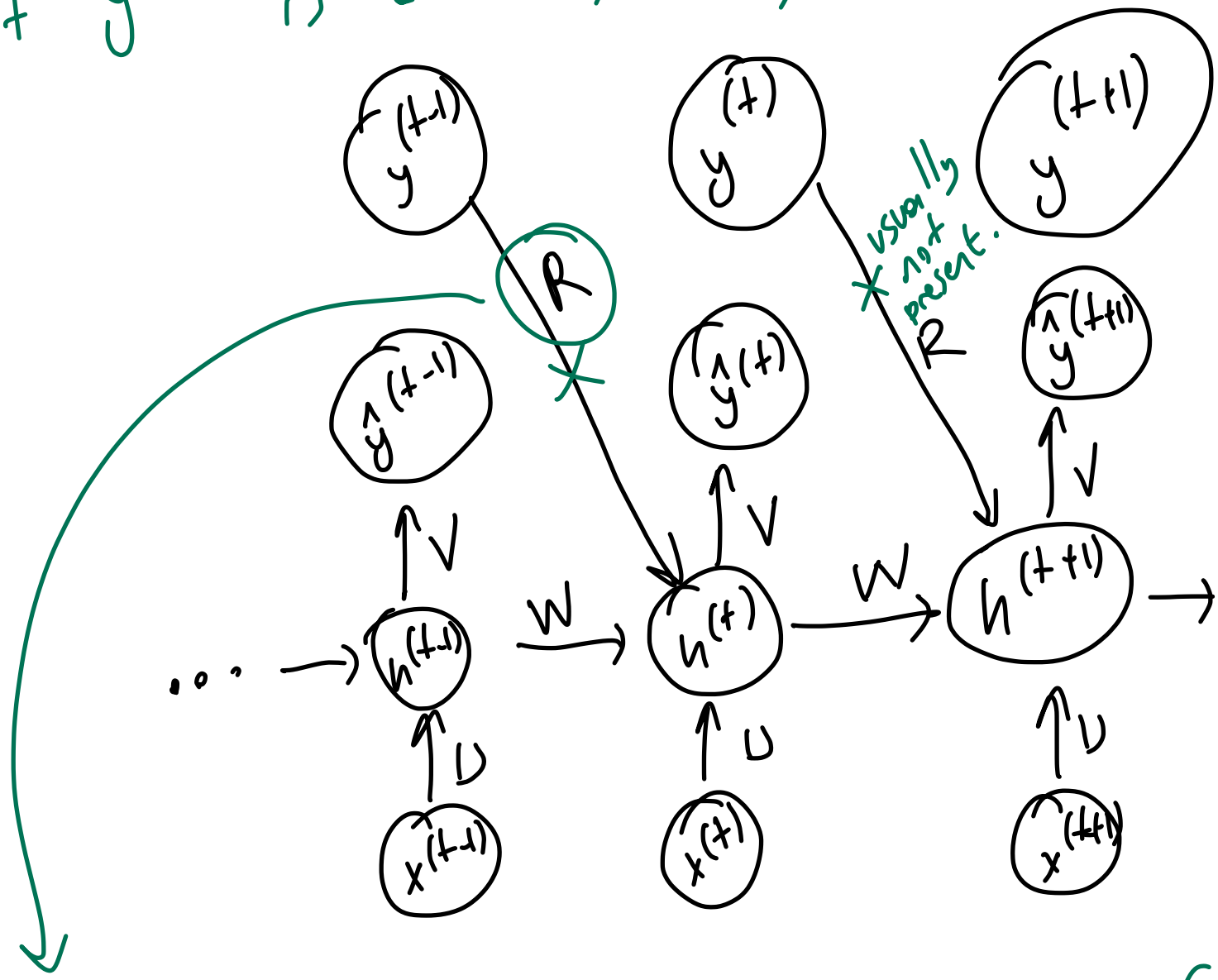
$$\hat{y}^{(t)} = \sigma_2\left[V.h^{(t)} + b_2\right]$$

testing.

$$h^{(t)} = \sigma_1\left[W.h^{(t-1)} + U.\hat{y}^{(t-1)} + Rx + b_1\right]$$

$$\hat{y}^{(t)} = \sigma_2\left[V.h^{(t)} + b_2\right]$$

if $y^{(t-1)}$ is an ordpective, the model will force $\hat{y}^{(t)}$ to be a noun.



← observed output

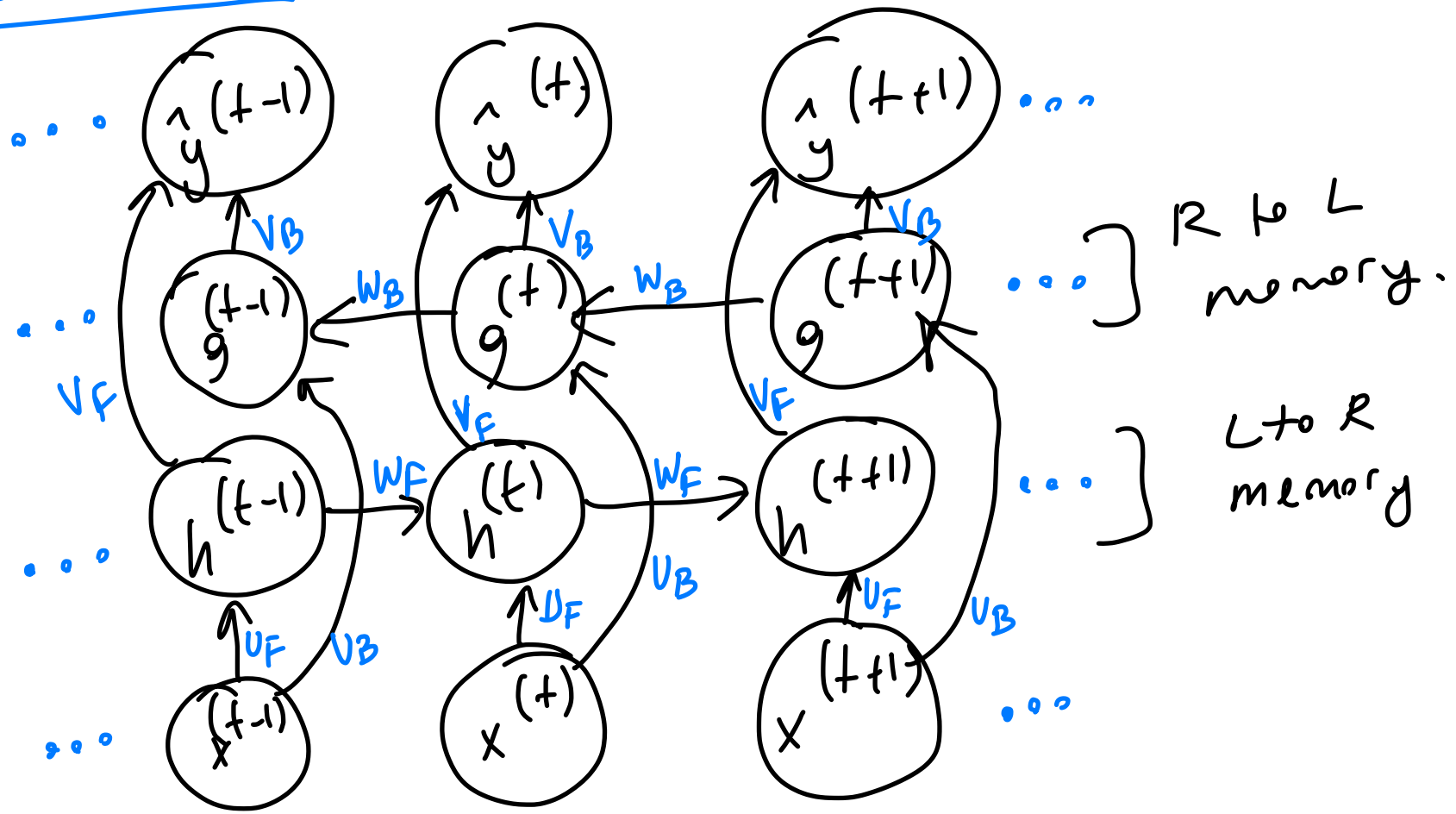
← predicted output

← memory

← input

,, going to learn grammatical rules from training data.

Bidirectional RNNs



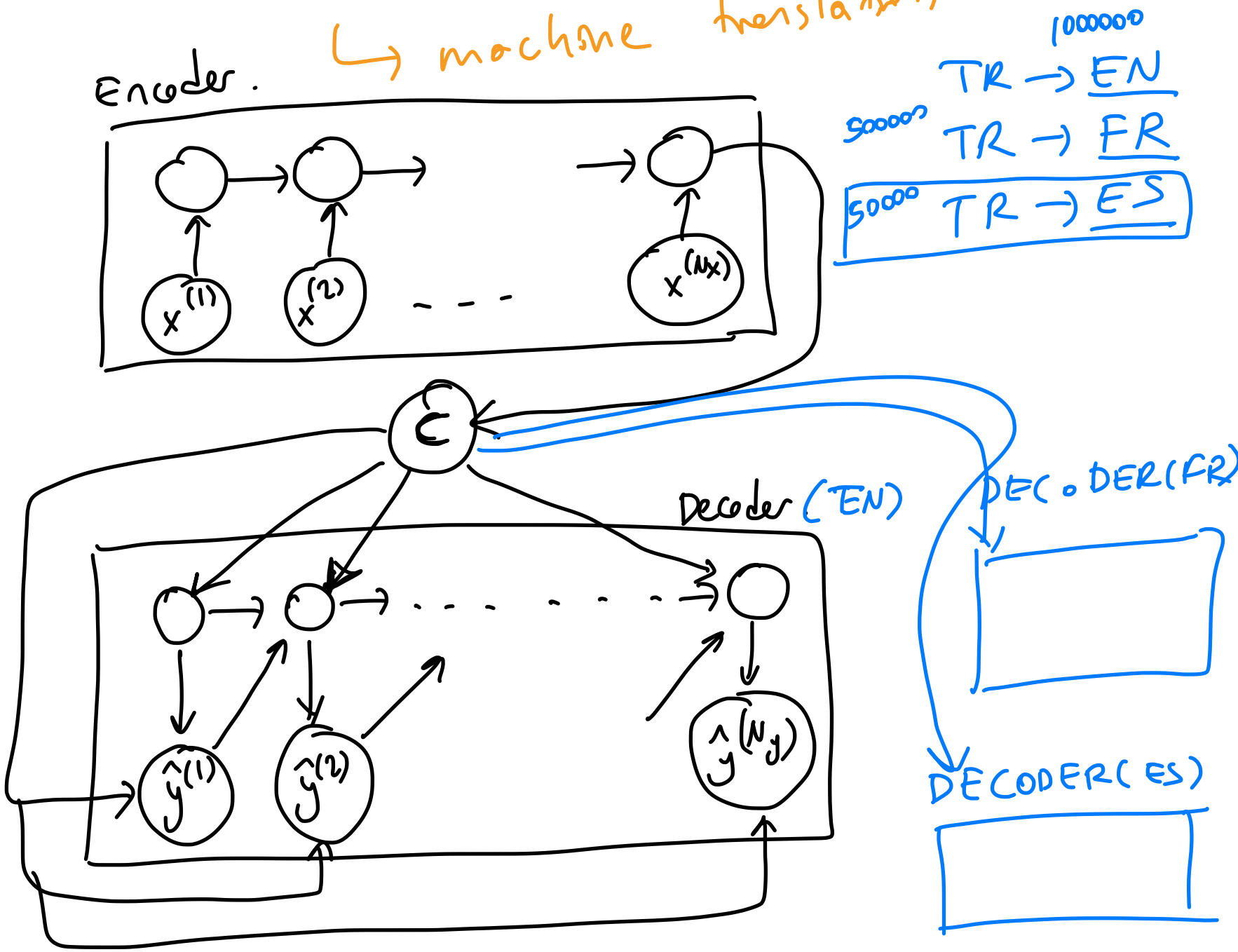
$$h^{(t)} = \sigma_1\left[W_F \cdot h^{(t-1)} + U_F x^{(t)} + b_F\right]$$

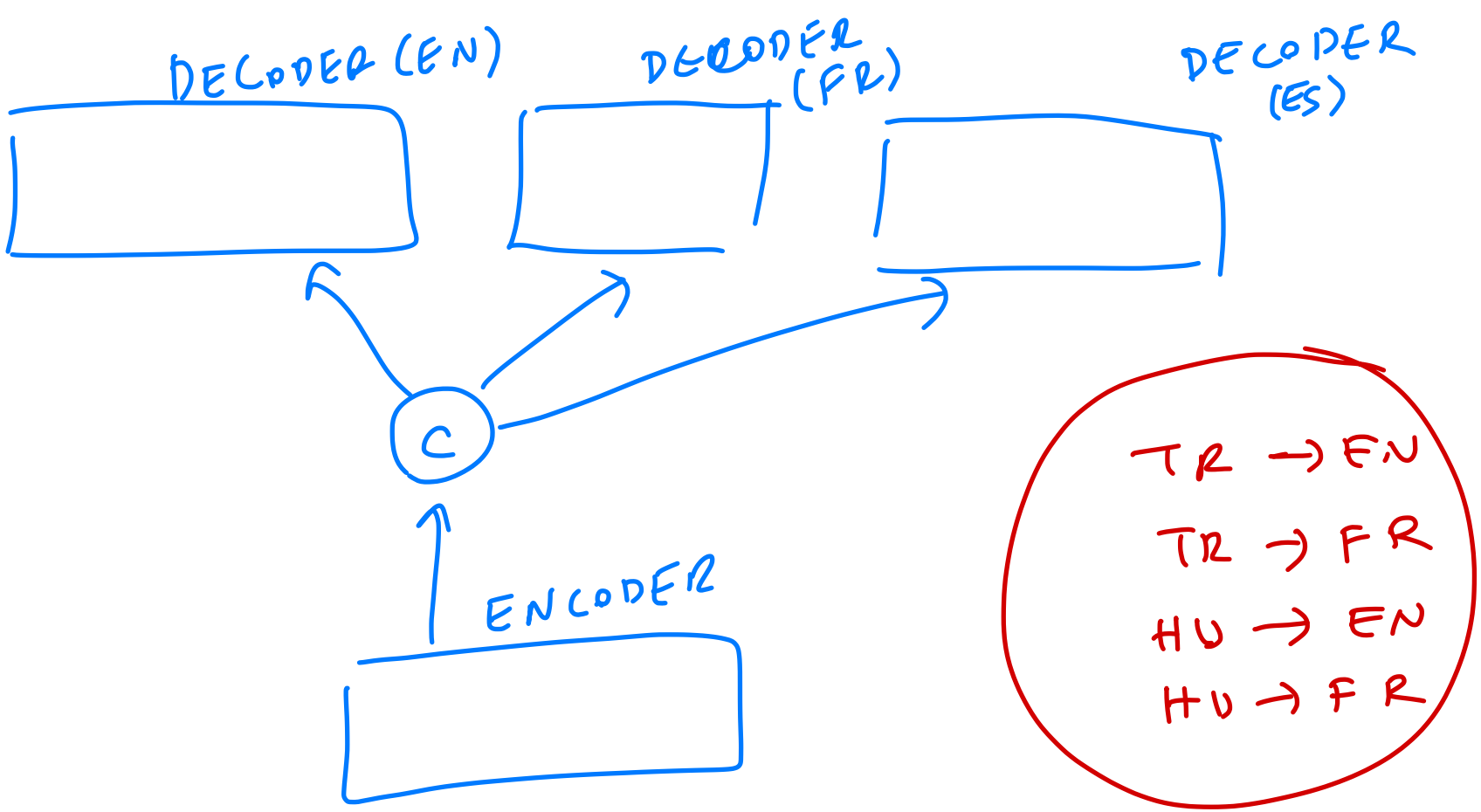
$$g^{(t)} = \sigma_2\left[W_B \cdot g^{(t+1)} + U_B \cdot x^{(t)} + b_B\right]$$

$$\hat{y}^{(t)} = \sigma_3\left[V_F \cdot h^{(t)} + V_B \cdot g^{(t)} + b\right]$$

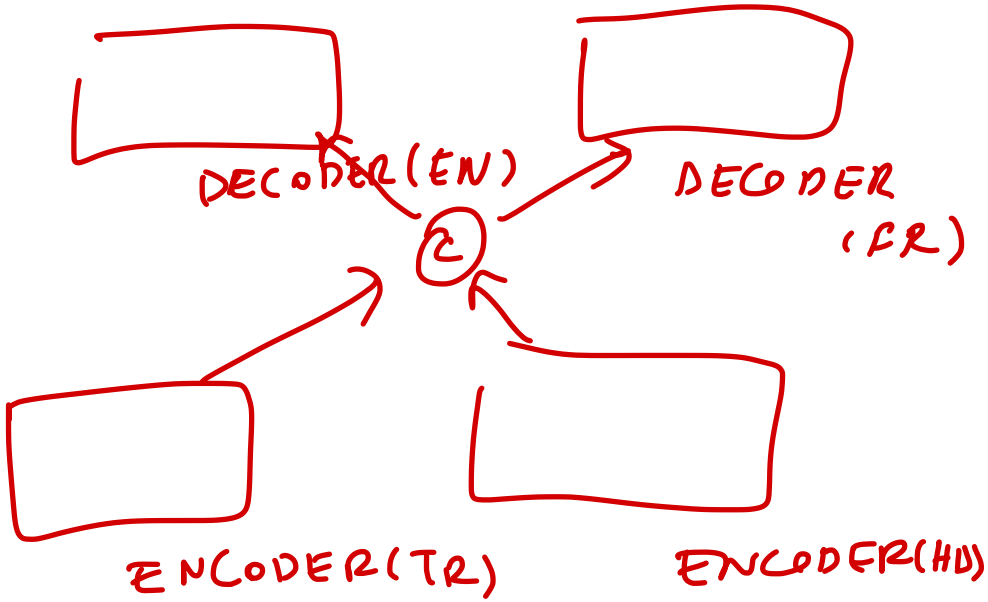
(Sequence to Sequence) Models.

↳ machine translation

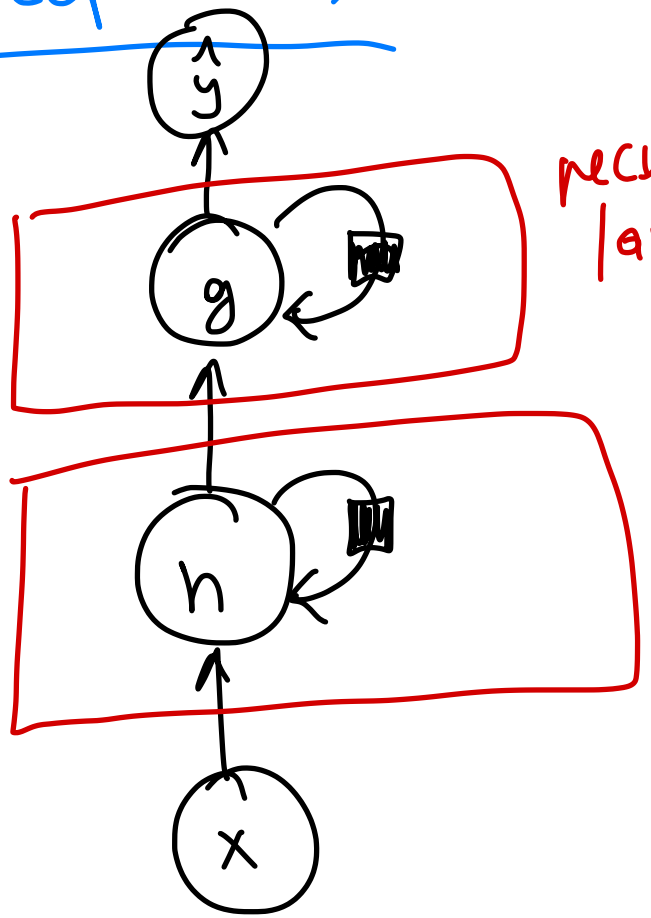




TR → EN
TR → FR
HU → EN
HU → FR



Deep RNNs



recurrent layer.

recurrent layer

↓

{ GRU / LSTM / RNN