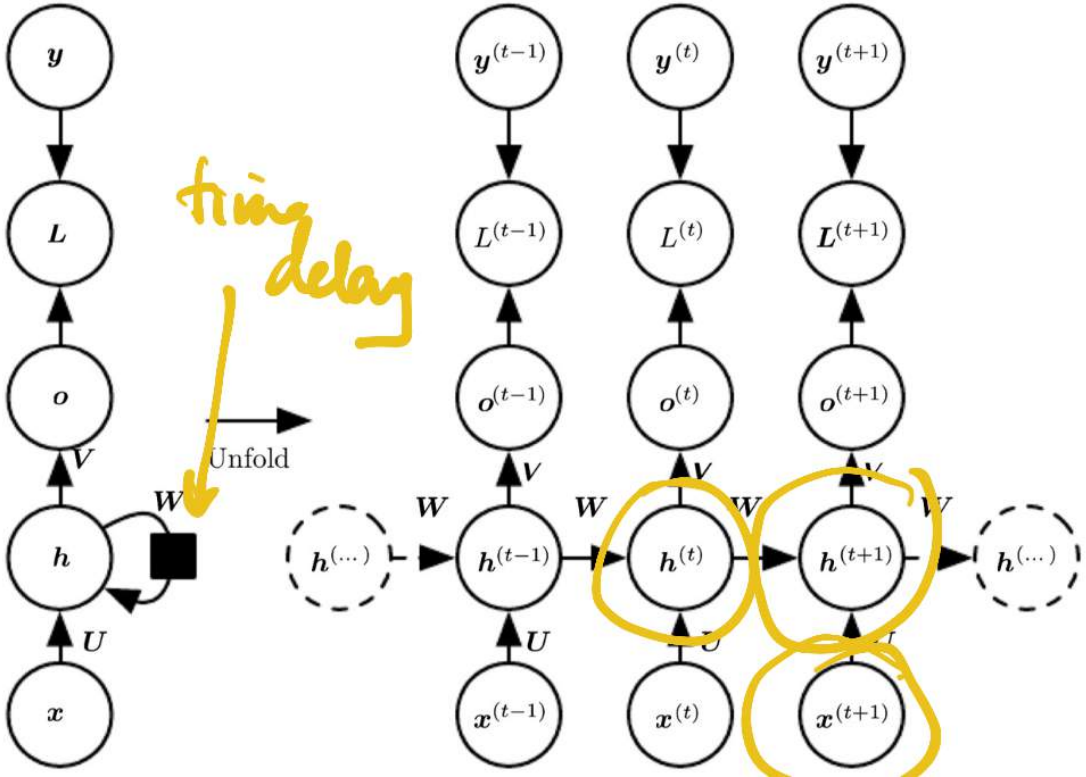
Advanced Machine learning, 3 Sep 2019

Sequence input – RNNs

Speech processing, handwriting, bioinformatics

Have feedback from earlier inputs

Hidden layer at time t+1 depends on $h^{(t)}$

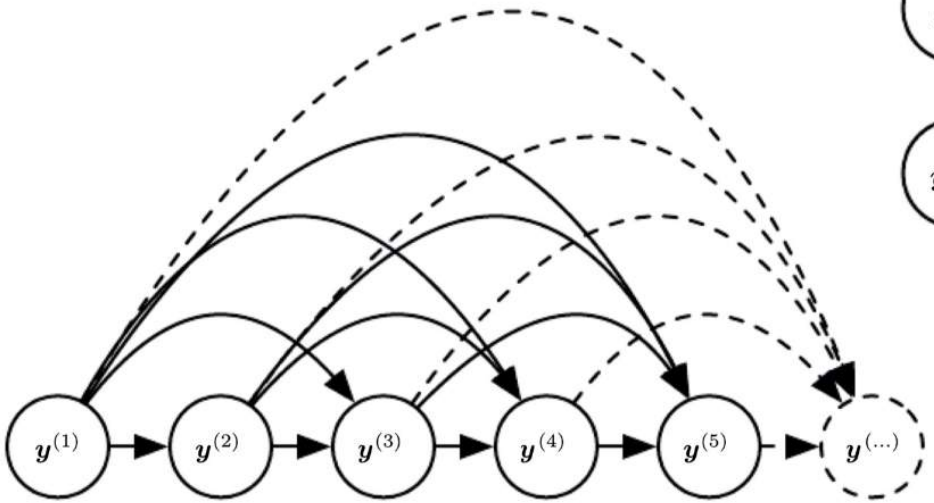


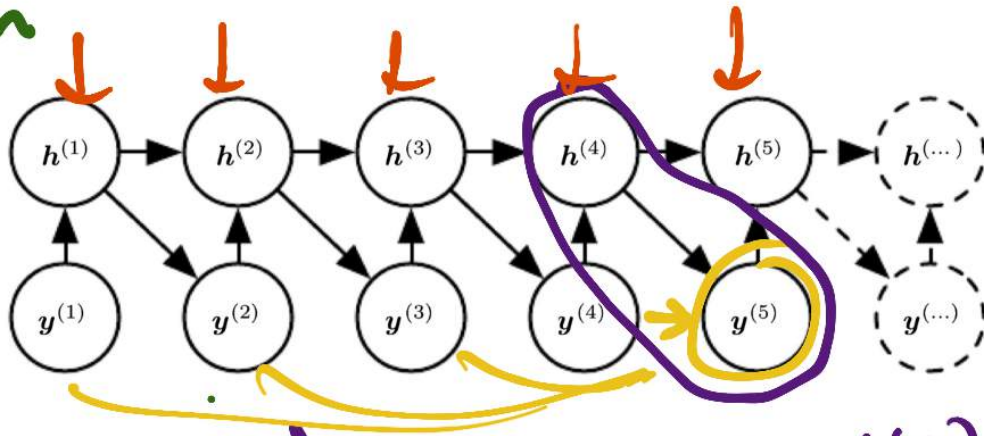
Unfolded network

Forward & backward propagation

Backward Propagation through time

Graphical interpretation





Dependence of $y^{(t+1)}$ on $y^{(t)}$ is factorized via $h^{(t)}$

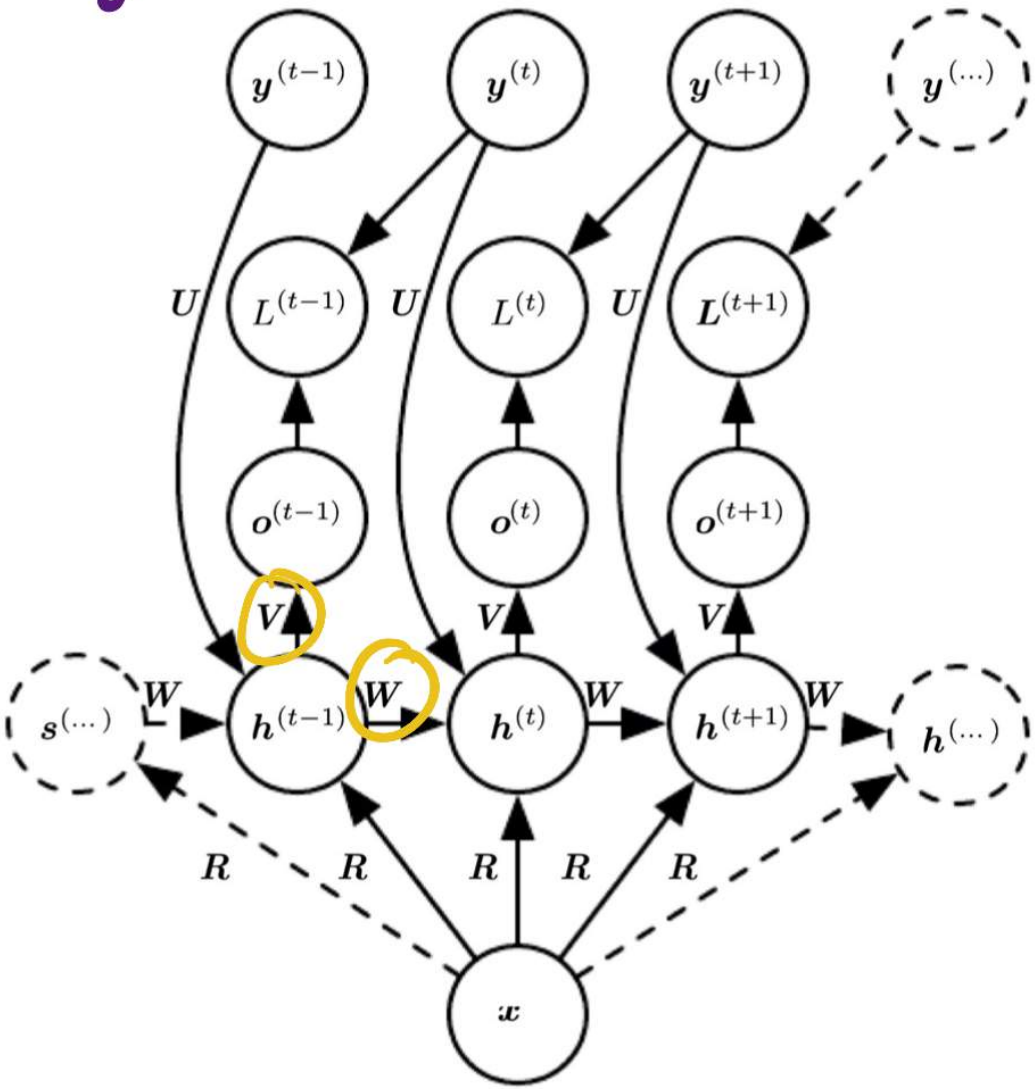
$y|x \rightarrow y|h \qquad h = f(x)$

Input is generically called "Context"

Feeding context to an RNN

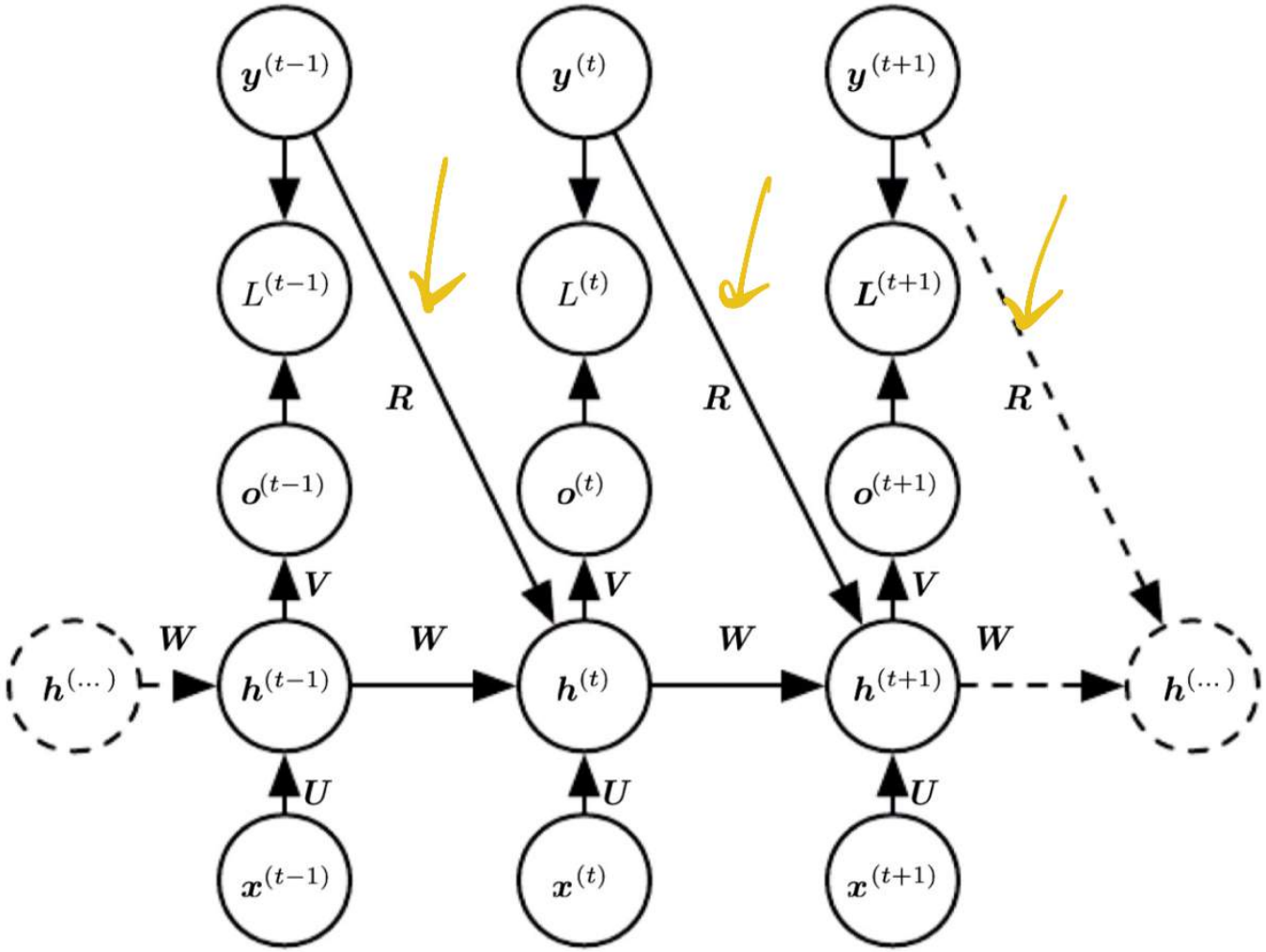
Option 1

All nodes get same x



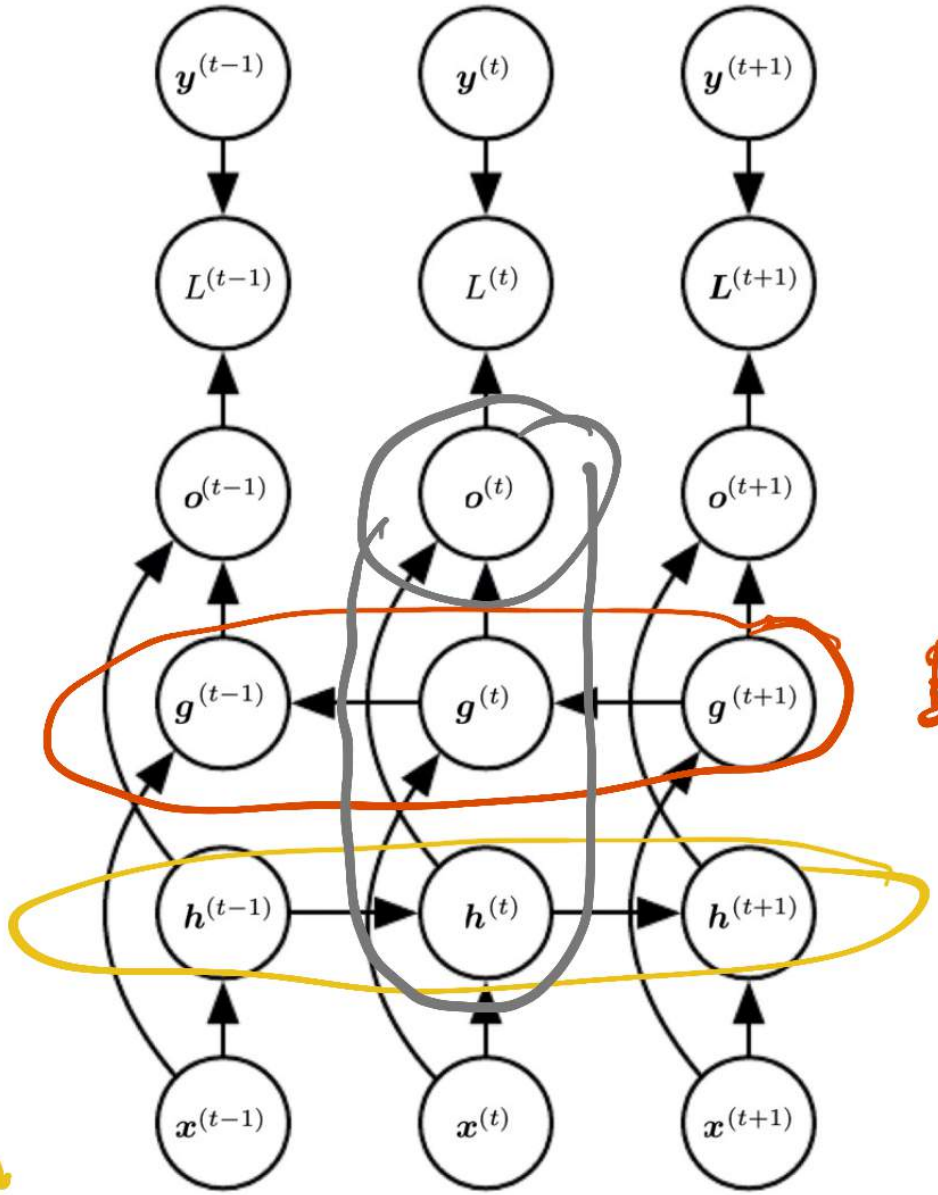
More "obvious" way is one input per time instant

Explicit influence of $y^{(t)}$ on $y^{(t+1)}$



Sometimes we need to see the future

Bidirectional RNNs



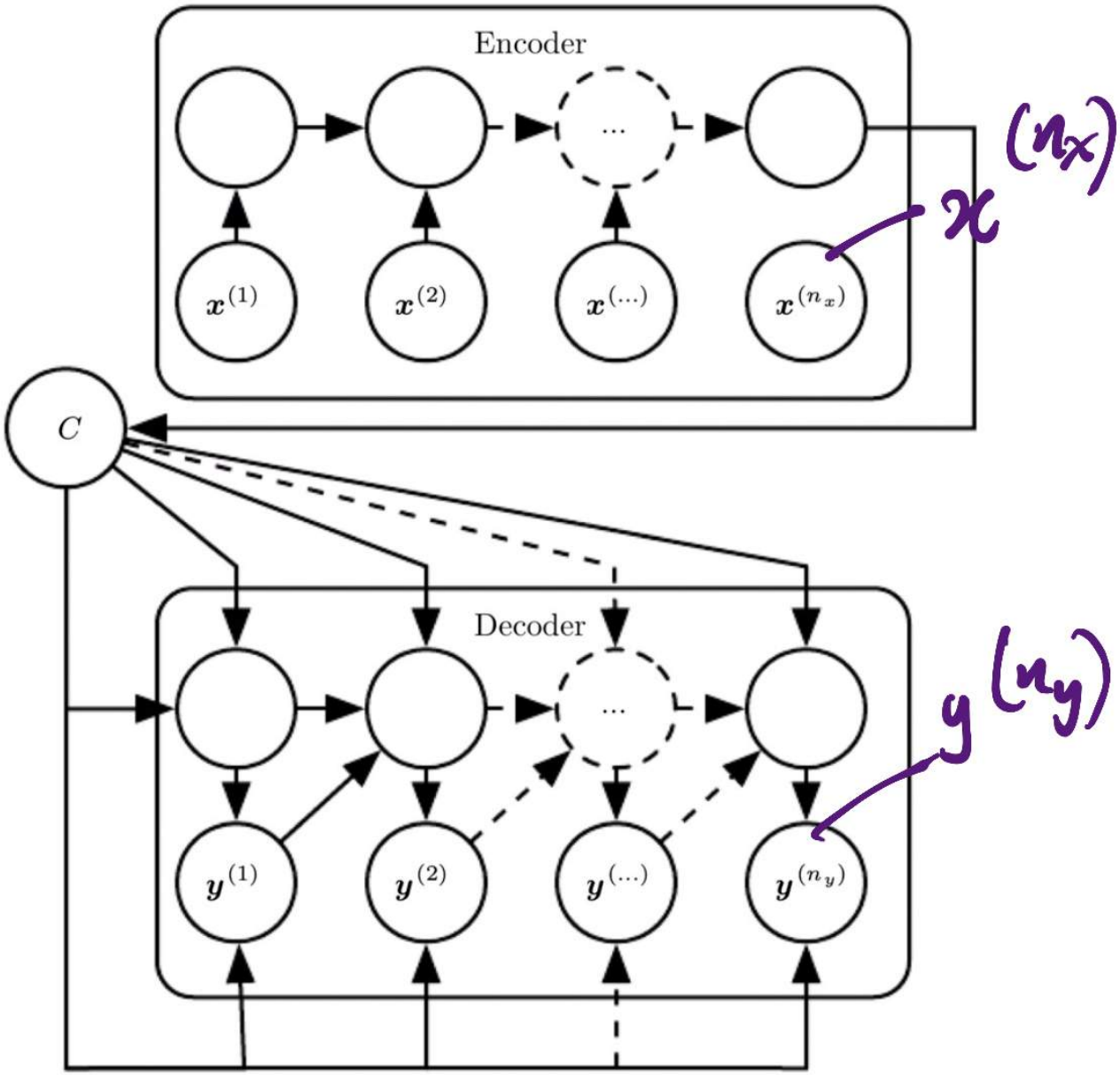
Backwards hidden layer

Forward connected hidden layer

What if output sequence length is different from input sequence length?

Separate input from output

Encoder-Decoder



The problem of "long term" memory

Gradients vanish (or explode)

Hard to model story dependency or distant past

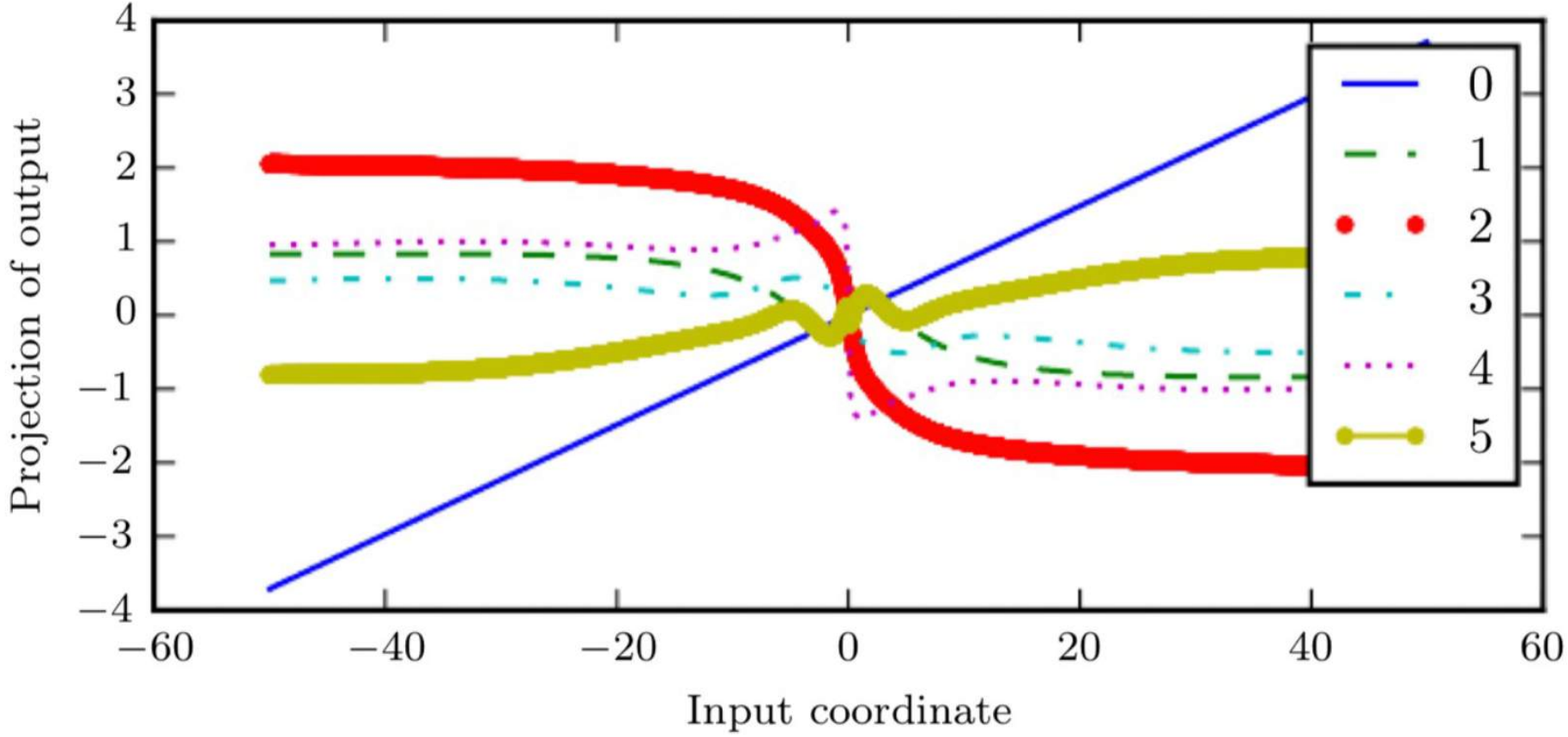
Increasing non linearity as time progresses

$$h^{(t)} = W^T h^{(t-1)}$$

↓ decompose as eigen decomposition

$$W = Q \Lambda Q^T$$

- Q: orthogonal
- Λ: eigenvalue diagonal matrix



Fix tanh activation & explore outputs as function of input after many layers

$$h^{(t+1)} = Q \Lambda Q^T h^{(t)}$$

Compose across n layers

$$h^{(n+1)} = \underbrace{(Q \Lambda Q^T)(Q \Lambda Q^T)}_{I} \underbrace{\cdots (Q \Lambda Q^T)}_{I} h^{(1)}$$

$$= Q \Lambda^n Q^T h^{(1)}$$

↳ eigenvalues away from 1 will vanish

How to remember "long term"?

1. Vary W across time. How?

2. Skip connections

   Connect layer t to t+d directly

3. "Leaky" connection

   Make the feedback linear

Moving average

$$h^{(t)} = \alpha h^{(t-1)} + (1-\alpha)\, i^{(t)}$$

Vary α:

$\alpha \rightsquigarrow 1$: More influence of past

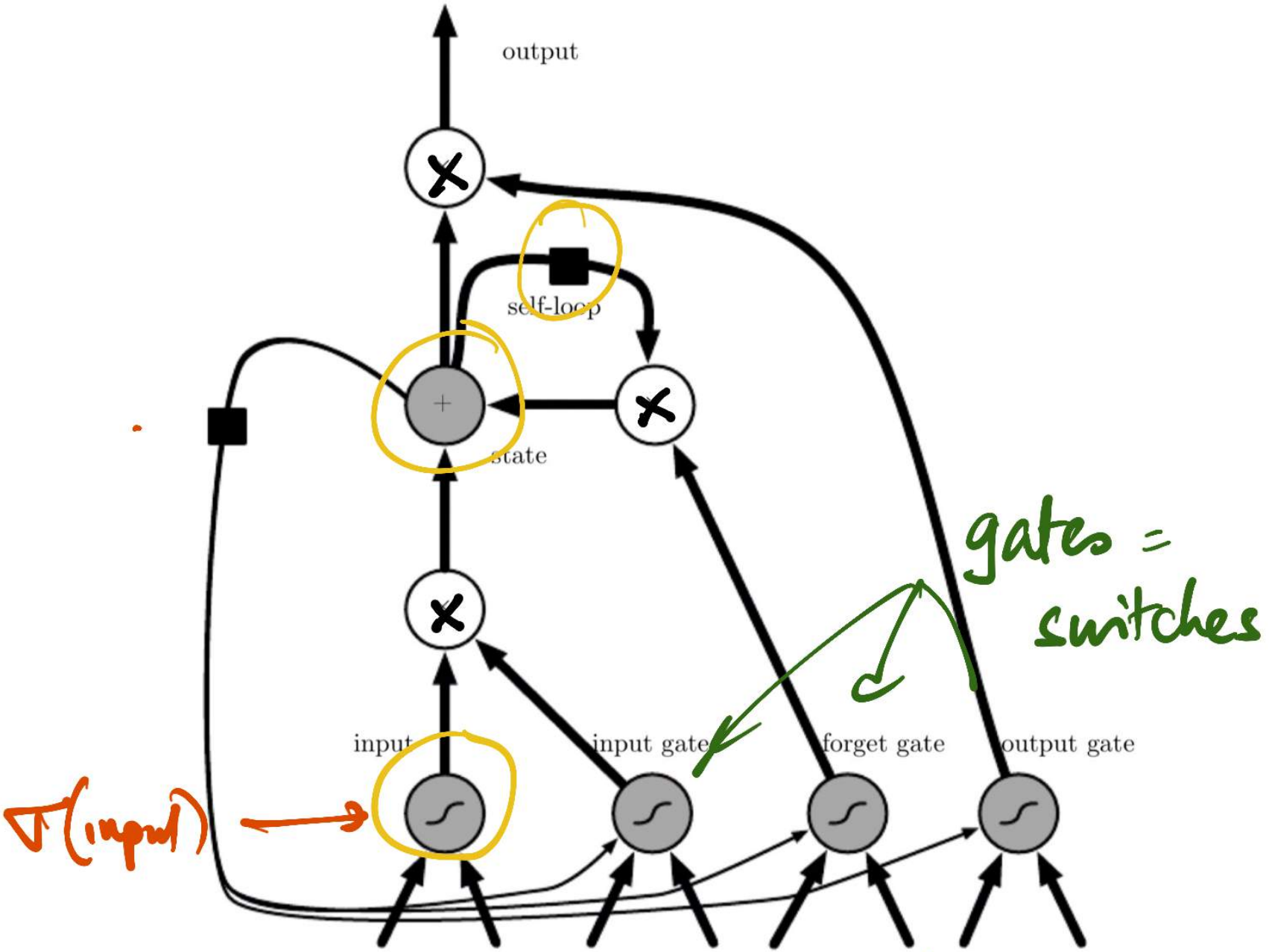
$\alpha \rightarrow 0$: More influence of present

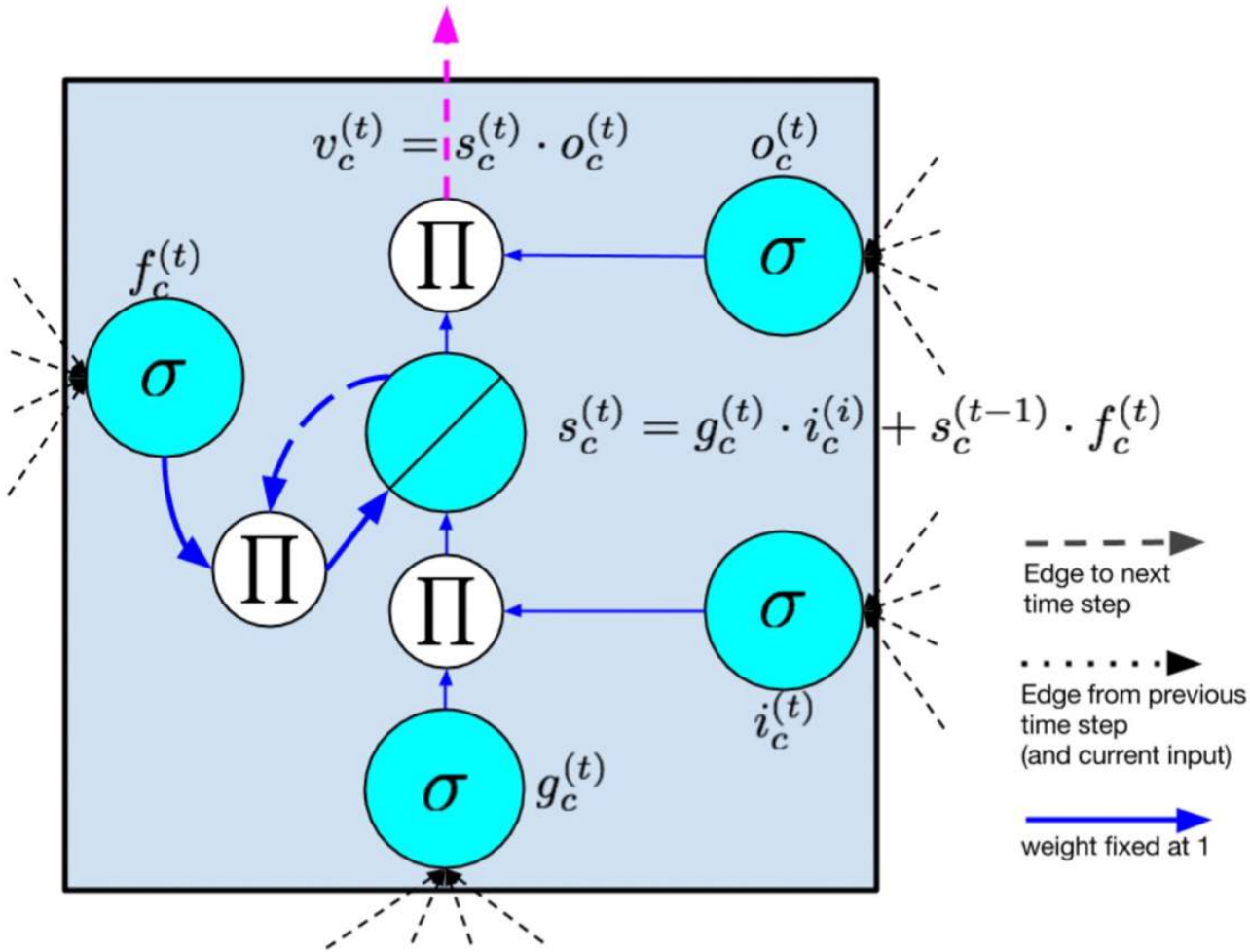
New architecture using "gates"

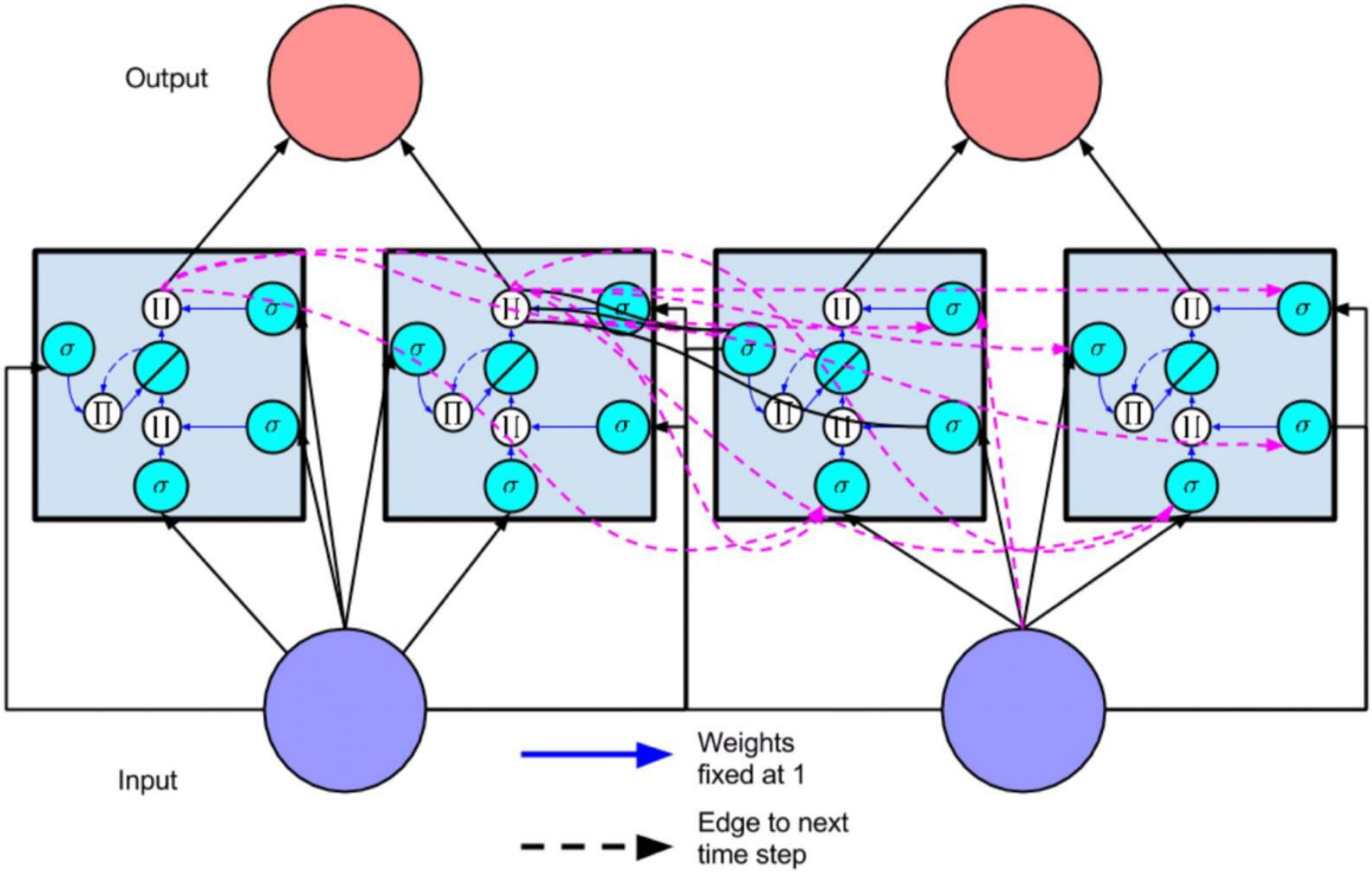
1. Is the input still valid ?
2. Is the past of h still valid ?
3. Should I continue to emit output ?

These decisions should be time varying

LSTM Long short term memory




$v_c^{(t)} = s_c^{(t)} \cdot o_c^{(t)}$
$o_c^{(t)}$
$f_c^{(t)}$
$s_c^{(t)} = g_c^{(t)} \cdot i_c^{(i)} + s_c^{(t-1)} \cdot f_c^{(t)}$
$i_c^{(t)}$
$g_c^{(t)}$
Edge to next time step
Edge from previous time step (and current input)
weight fixed at 1


Output
Input
Weights
fixed at 1
Edge to next
time step