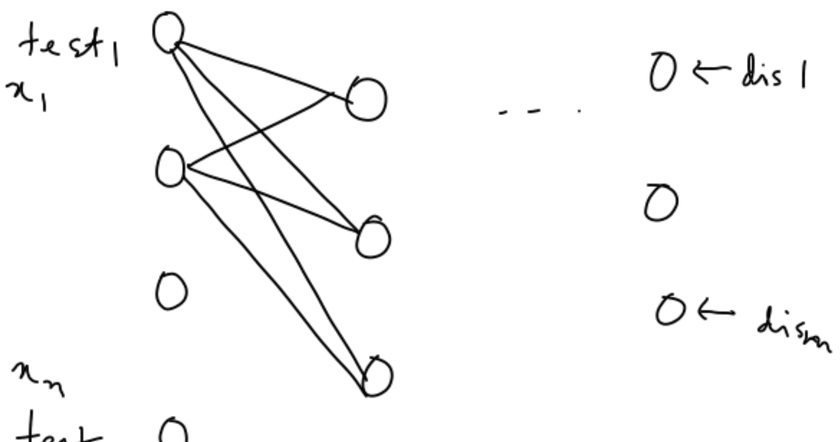
Lec 13: Recurrent Neural Networks

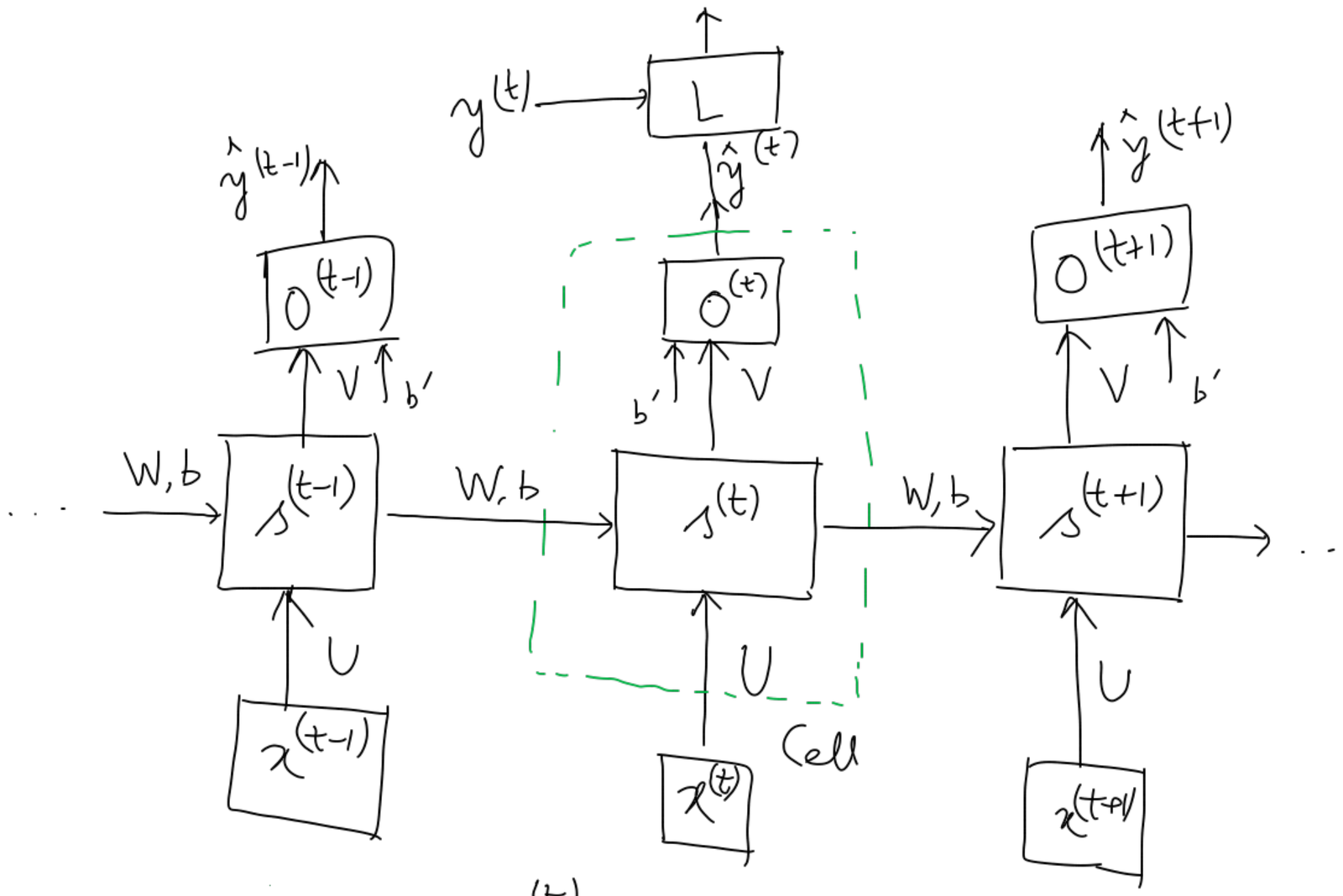


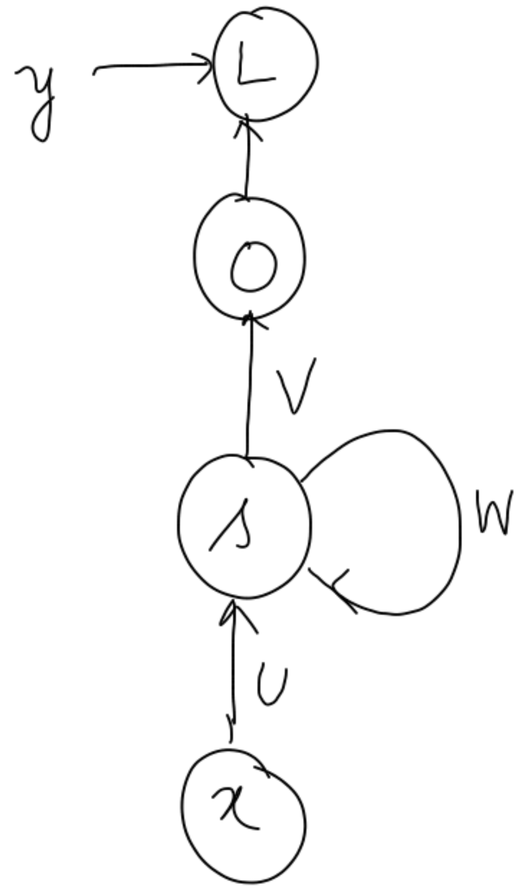
Data is sequential. ← Ordering matters
← Length is variable (input/output)

Sentence:

Examples: autocomplete, grammar error, virtual assistants, language translation.

RNN is a way to get around this → changes the architecture.





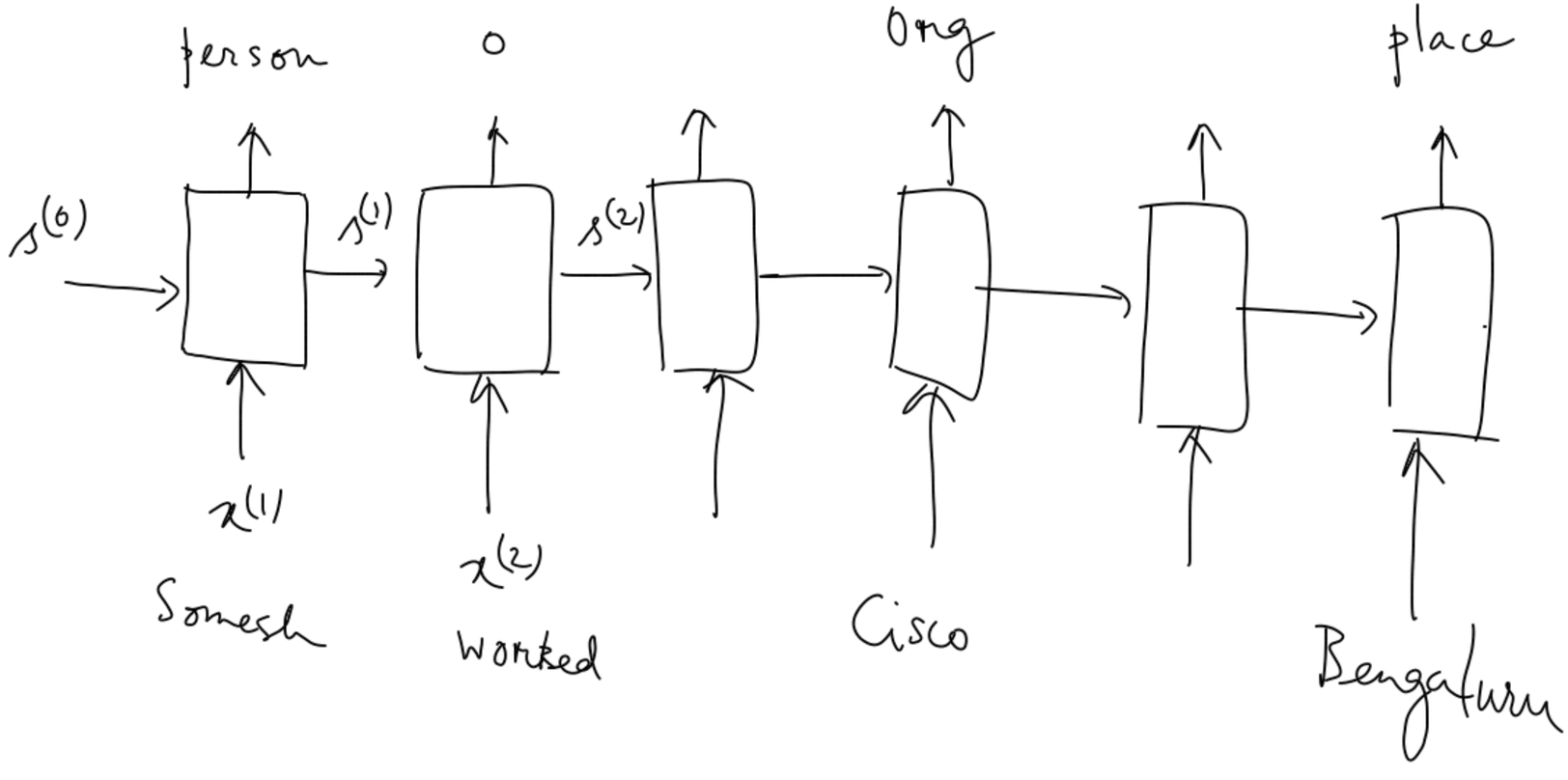
$$s^{(t)} = g\left(W s^{(t-1)} + U x^{(t)} + b\right), \quad g \equiv \sigma, \tanh, \text{ReLU}$$

$$o^{(t)} = V s^{(t)} + b' \qquad \hat{y}^{(t)} = \text{softmax}\left(o^{(t)}\right)$$

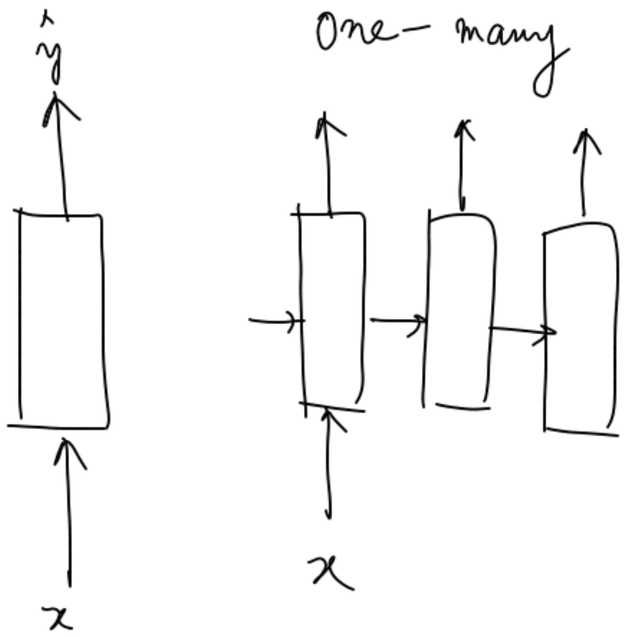
Somesh worked at Cisco in Bengaluru

person org place

Ex: Named Entity Recognition

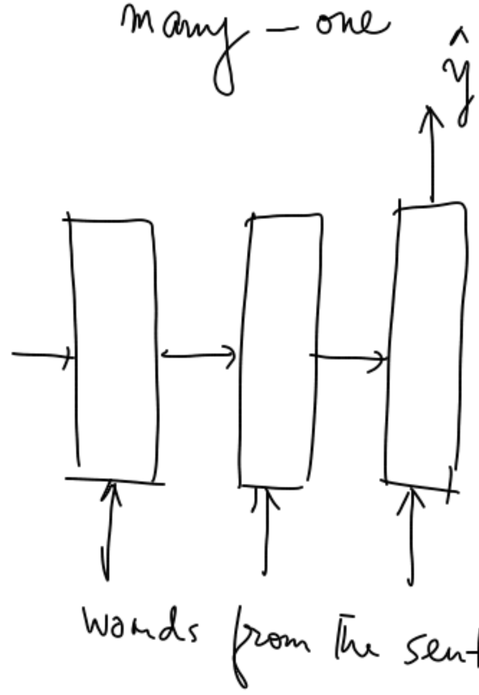


RNN possible structures

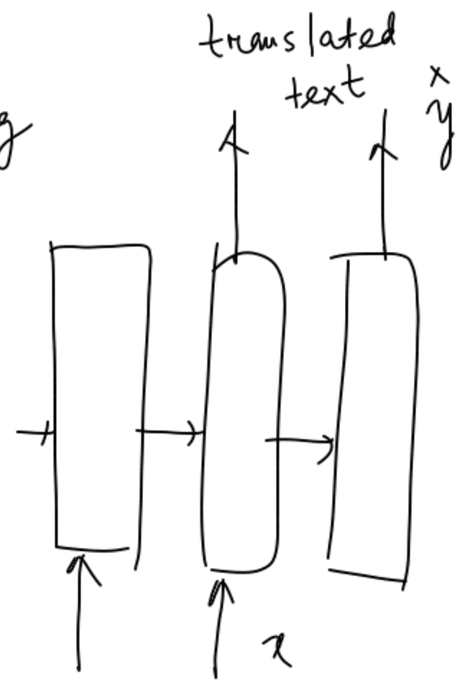


FFNN

Image caption



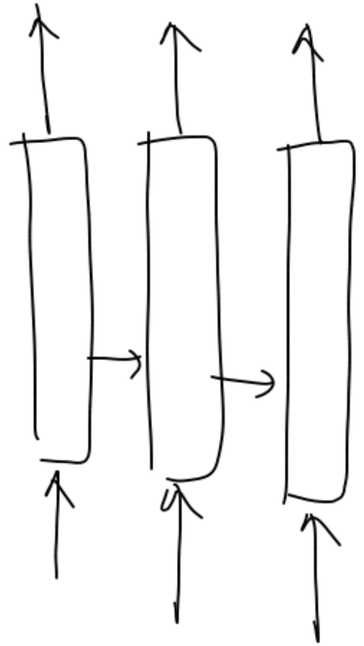
rating



text translation

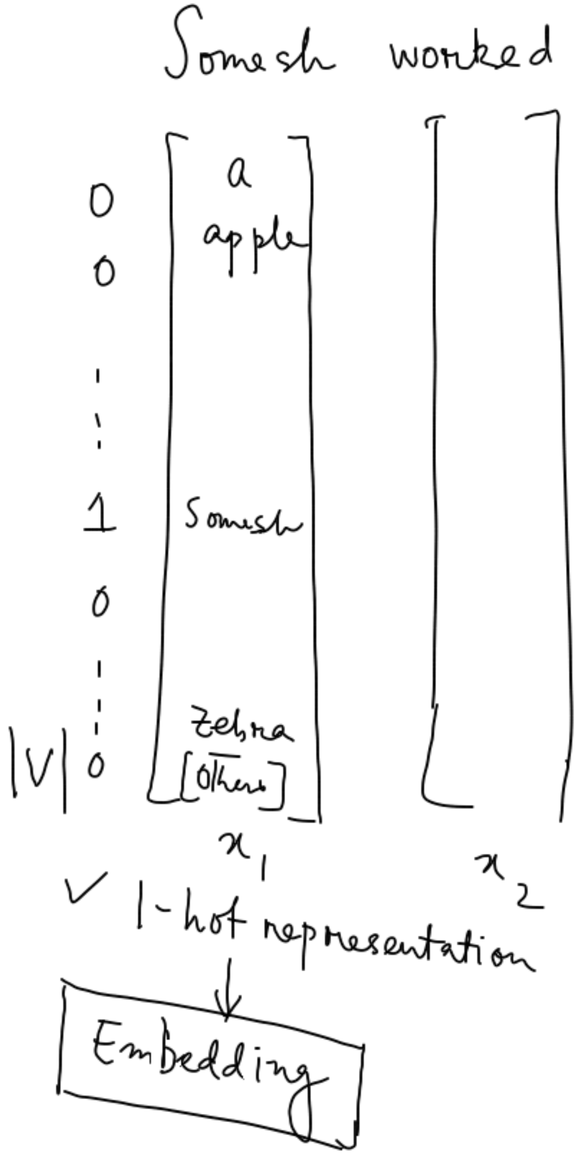
many-many staggered

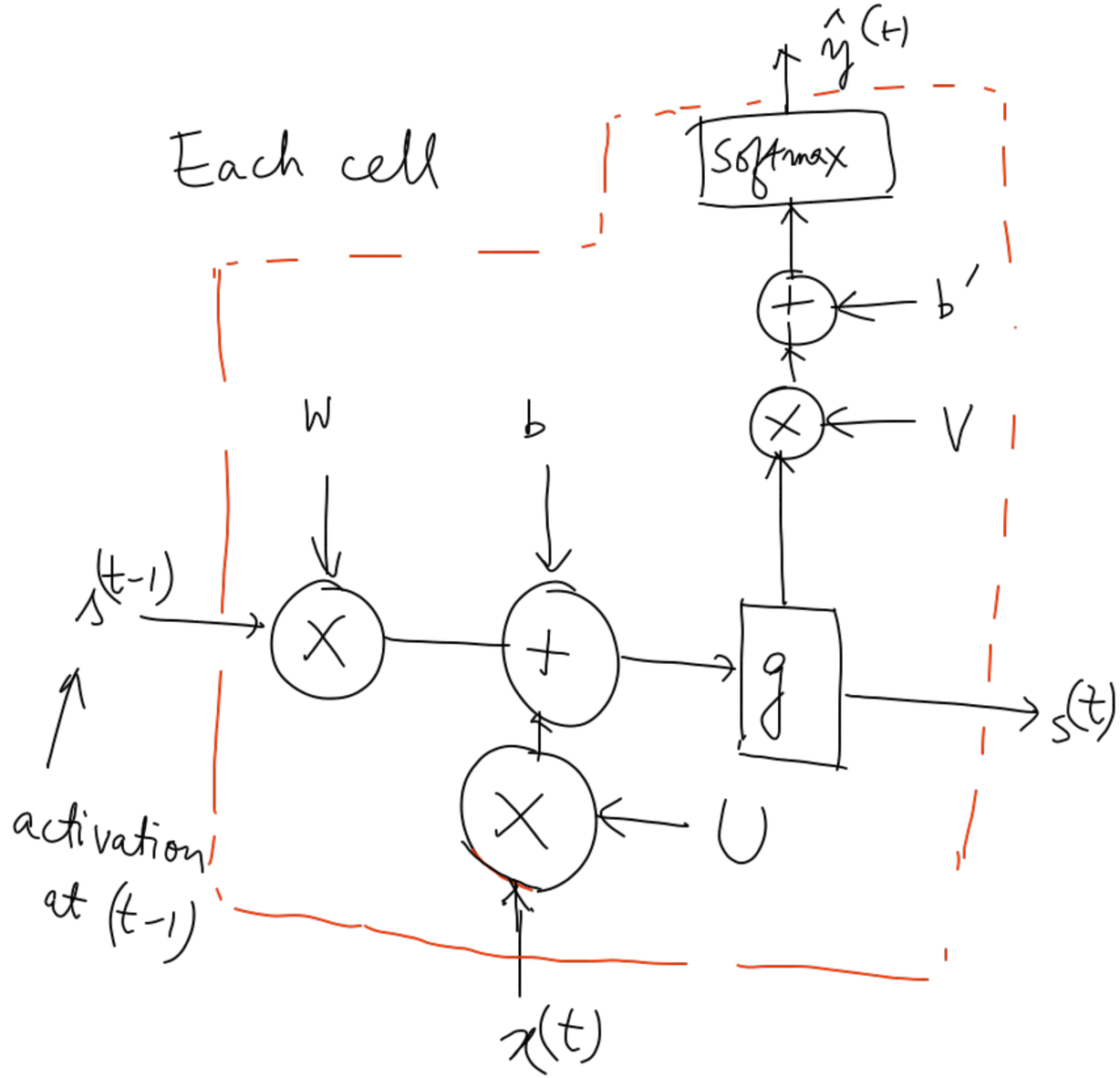
$len(x) \neq len(\hat{y})$



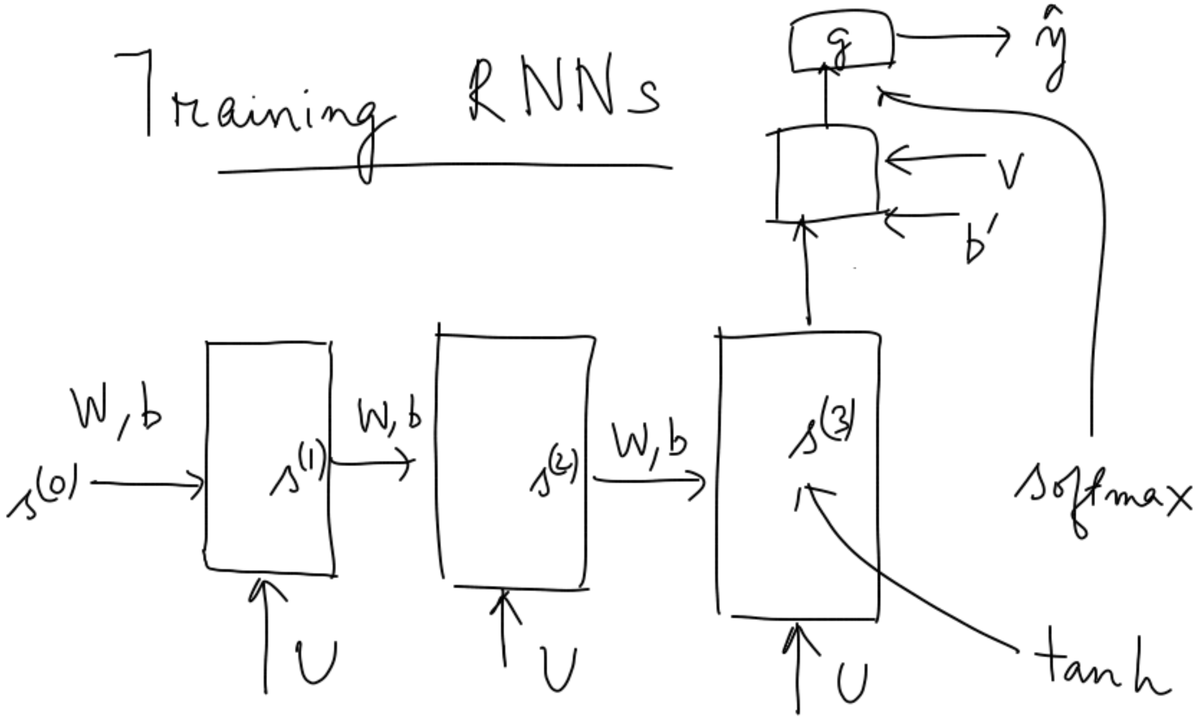
many-many synchronous (NER)

Somesh worked at Cisco in Bengaluru





Training RNNs



$$\text{Loss}(y, \hat{y}) = -\sum y_k \log \hat{y}_k = L$$

↓ Categorical cross entropy fn.

$$\text{softmax} = -\sum_{k=1}^{K} I\{y = k\} \log\left(\hat{y}_k\right)$$

$$\hat{y}_k = \frac{\exp\{w^T \dots\}}{\sum \exp\{\dots\}}$$

BPTT: Backprop through time

Forward pass: compute $s^{(1)}, s^{(2)}, \dots$

$$s^{(t)} = \tanh\left(W s^{(t-1)} + U x^{(t)} + b\right)$$

$$\hat{y} = \text{softmax}\left(V s^{(T)} + b'\right)$$

$$\frac{\partial L}{\partial W}, \frac{\partial L}{\partial b}, \frac{\partial L}{\partial U}, \frac{\partial L}{\partial V}, \frac{\partial L}{\partial b'} \leftarrow \text{find}$$

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial V} \quad ; \quad \frac{\partial L}{\partial b'} = \dots$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} \left(-\sum y_k \log \hat{y}_k \right)$$

$L \leftarrow \hat{y} \leftarrow s^{(3)} \leftarrow U$

$s^{(2)} \leftarrow U$

$s^{(1)} \leftarrow U$

$\frac{\partial L}{\partial U}$ in a similar way

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W}$$

$\hat{y} \leftarrow s^{(3)}$, $s^{(3)} \leftarrow W, s^{(2)}$, $s^{(2)} \leftarrow W, s^{(1)}$, $s^{(1)} \leftarrow W$

$$= \frac{\partial L}{\partial \hat{y}} \left(\frac{\partial \hat{y}}{\partial s^{(3)}} \right) \frac{\partial s^{(3)}}{\partial W}$$

$$+ \frac{\partial L}{\partial \hat{y}} \left(\frac{\partial \hat{y}}{\partial s^{(3)}} \cdot \frac{\partial s^{(3)}}{\partial s^{(2)}} \right) \frac{\partial s^{(2)}}{\partial W}$$

$$+ \frac{\partial L}{\partial \hat{y}} \left(\frac{\partial \hat{y}}{\partial s^{(3)}} \cdot \frac{\partial s^{(3)}}{\partial s^{(2)}} \times \frac{\partial \hat{y}}{\partial s^{(1)}} = \frac{\partial s^{(2)}}{\partial s^{(1)}} \right) \frac{\partial s^{(1)}}{\partial W}$$

$$= \sum_{i=1}^{T} \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial s^{(i)}} \cdot \frac{\partial s^{(i)}}{\partial W}$$

$\frac{\partial L}{\partial b}$ in a similar way

$f(g_1(x), g_2(x))$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g_1} \frac{\partial g_1}{\partial x} + \frac{\partial f}{\partial g_2} \cdot \frac{\partial g_2}{\partial x}$$

Disadvantages: Vanishing gradient problem

Isha went for a vacation, ... He was overjoyed

→ In traditional RNN, grammar checker won't be able to detect this.

→ To mitigate, a modification to the cell structure. Long Short Term Memory (LSTM)