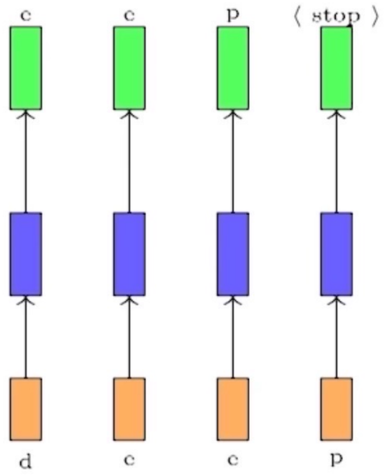
Sequence Learning Problems

1. Successive inputs may not be independent of each other.
2. The input size is not fixed.

Text Completion

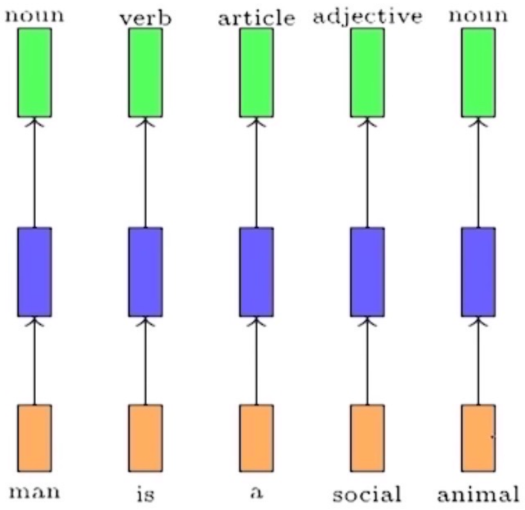


(i) Input is no longer independent.

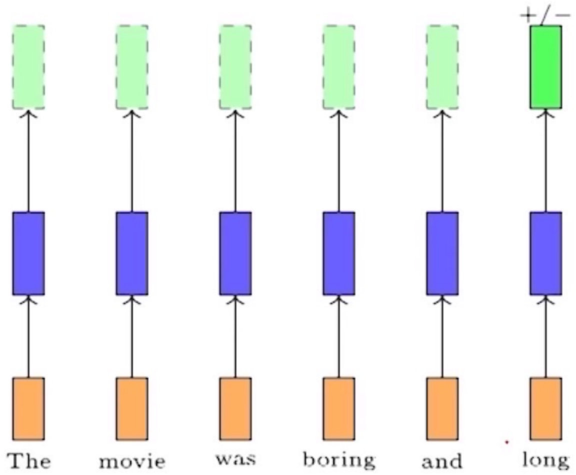
(ii) Length of word (i/p) is not fixed.

(iii) Each network is performing the same task.

Part of Speech Tagging



Sentiment Analysis



Objectives:

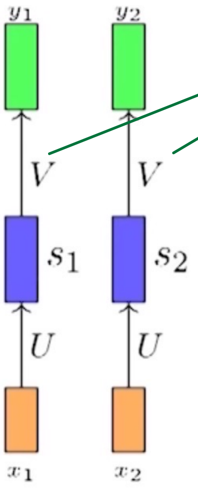
1. Account for the dependence between inputs.
2. Account for variable size of input
3. Make sure that the function executed at each time step is the same.

bias
↓

$s_1 = \sigma(Ux_1 + b)$

$y_1 = o(Vs_1 + c)$ ↖ bias

o: output function



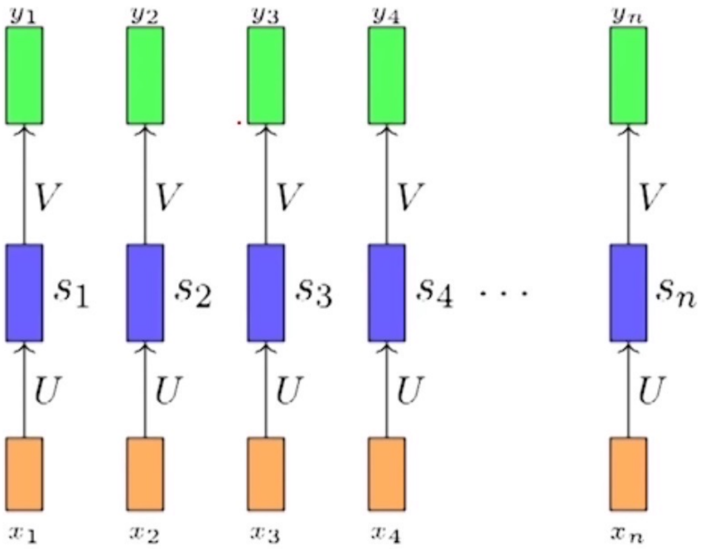
Share the same weight accross different time-steps.

$s_2 = \sigma(Ux_2 + b)$

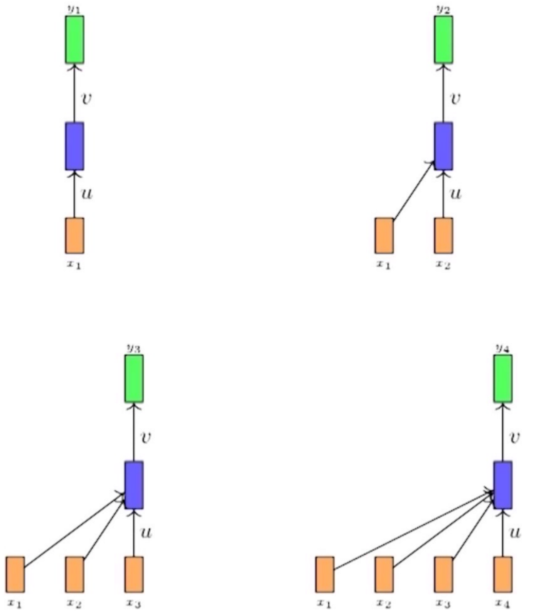
$y_2 = o(Vs_2 + c)$

Since we want the same function to be implemented at each step, we should share the same network parameters.

Variable size of the input can be taken care of by replicating the network at each time step.

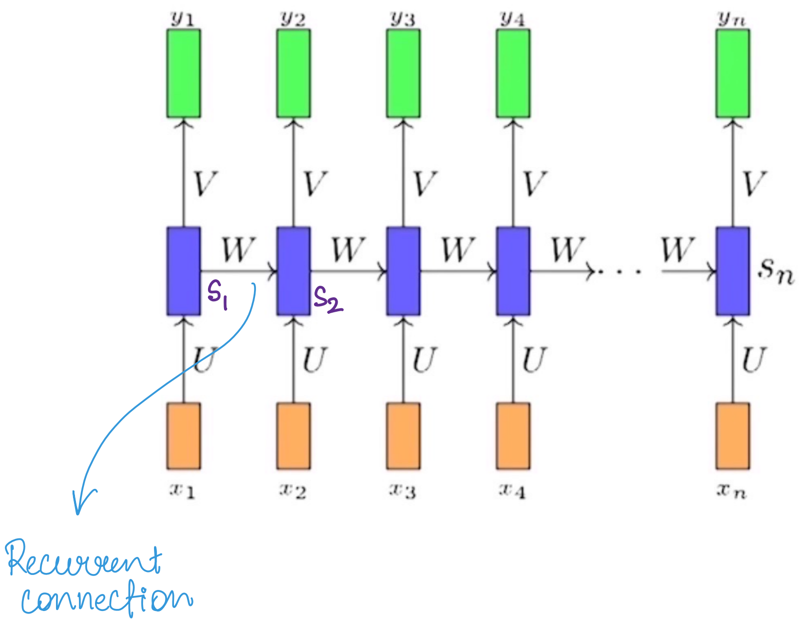


How do we account for the dependence between the inputs?



Not efficient

Solution: Add a recurrent connection

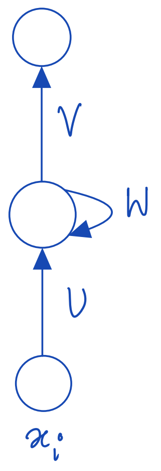


$$s_2 = \sigma(Ux_2 + Ws_1 + b)$$
$$y_2 = O(Vs_2 + c)$$

(Unrolled version)

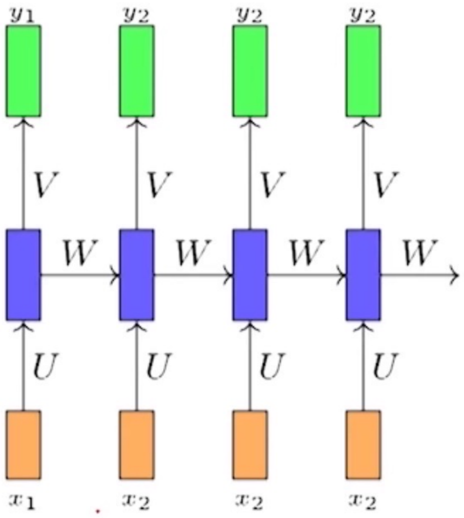
$$\boxed{\begin{aligned} s_i &= \sigma(Ux_i + Ws_{i-1} + b) \\ y_i &= O(Vs_i + c) \end{aligned}}$$

s_i : State of the network



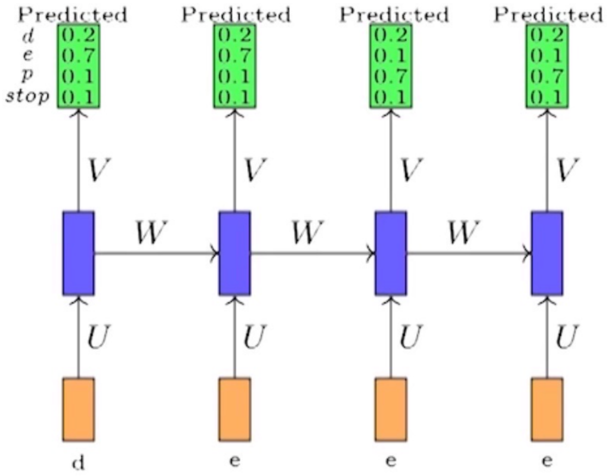
How to train an RNN?

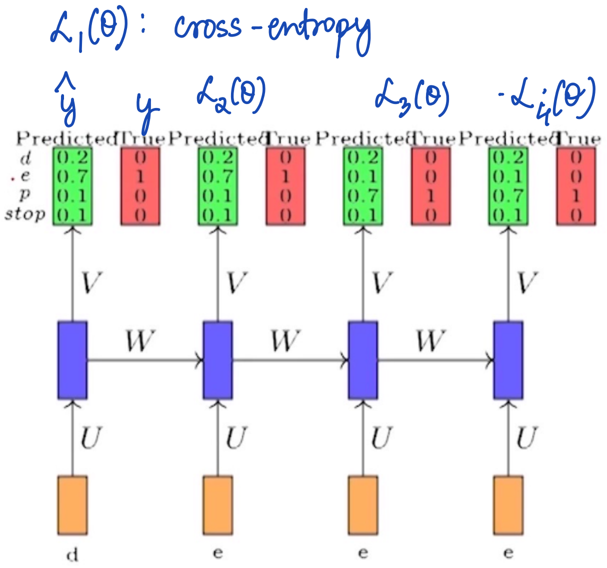
Backpropagation through time.



What is a suitable output function

In this case: softmax



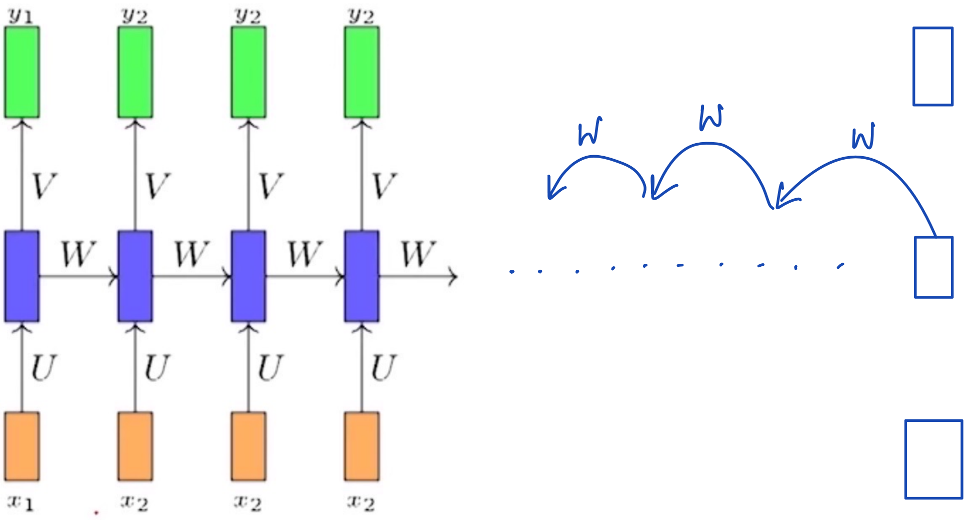


$$\min \mathcal{L}(\theta) = \sum_{t=1}^{4} \mathcal{L}_t(\theta)$$

↓

Do backpropagation

Problem of Vanishing and Exploding gradients



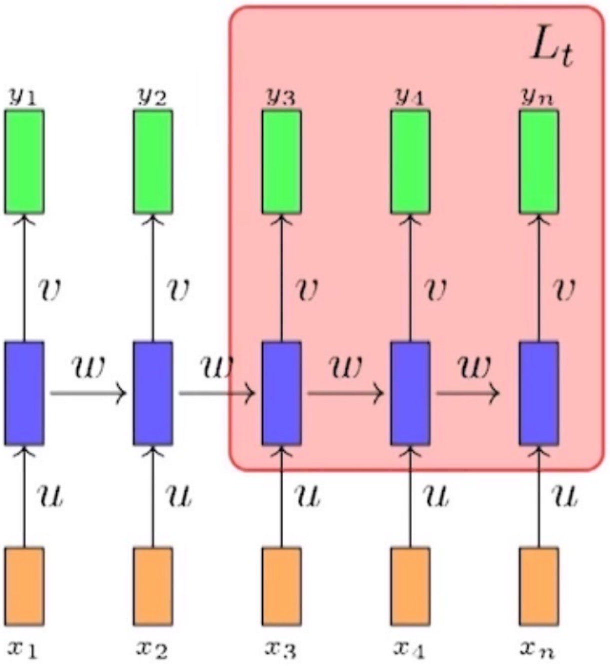
When gradient information is passed backward it's multiplied with W.

(1) Exploding Gradients : Gradient value starts to increase very rapidly roughly when $\|W\| > 1$ (eg. 2^t)

(2) Vanishing gradients: Gradient value starts to decrease very rapidly.

roughly when $\|W\| < 1$ (eg. $\frac{1}{2^t}$)

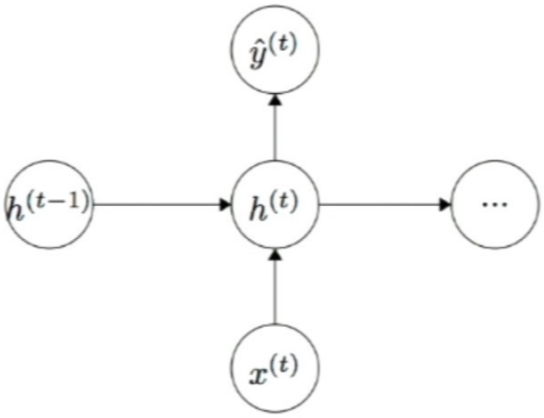
One simple solution could be Truncated Backpropagation restrict the product to τ-terms.

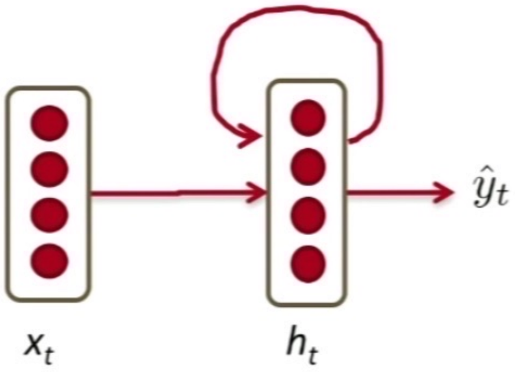


Another solution: Clipping

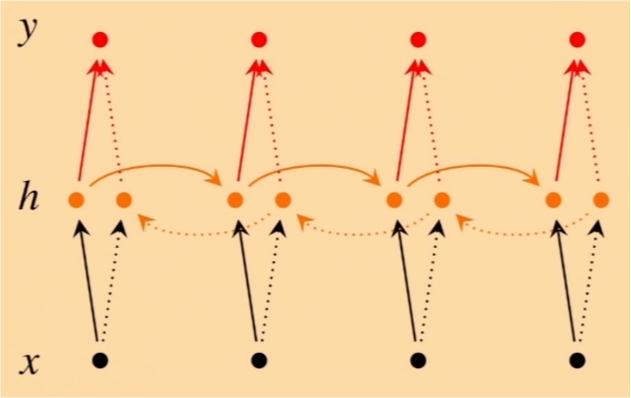
If gradient > Threshold T then
gradient = T

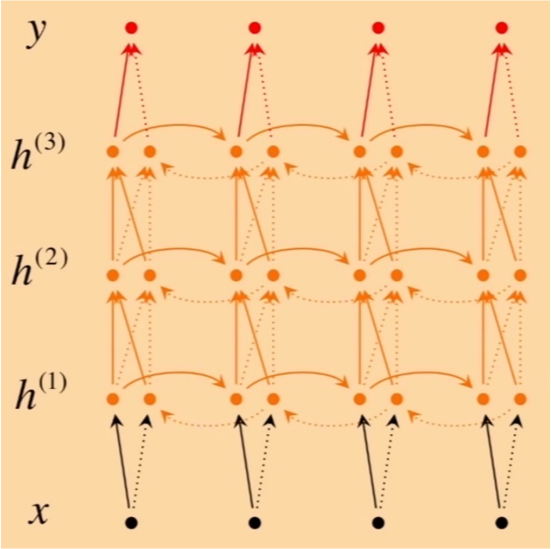
Other Representations





Bi-directional RNN





Multiple hidden layers