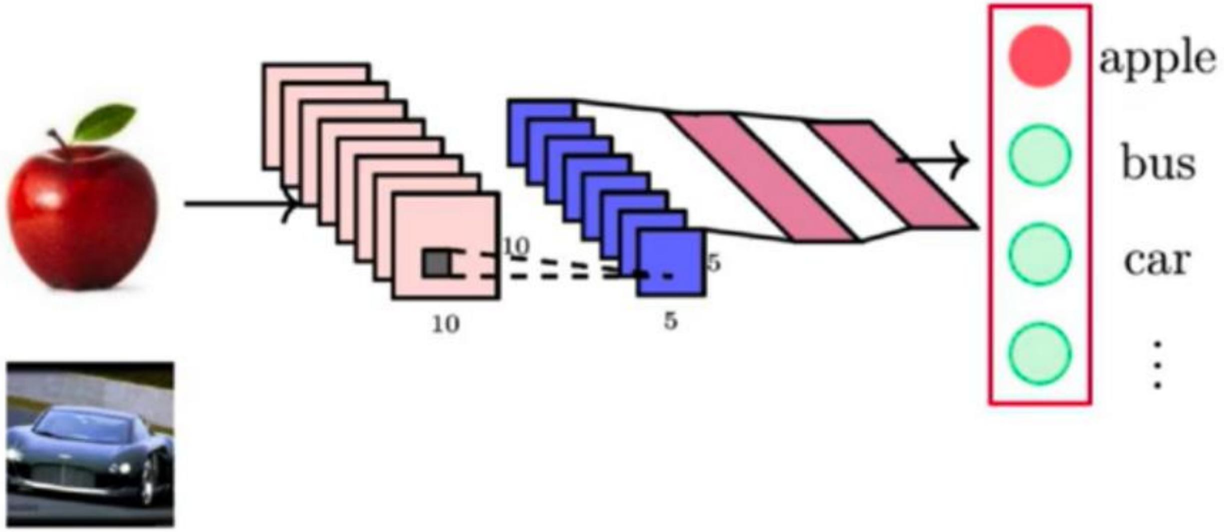
AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Introduction to Deep Learning

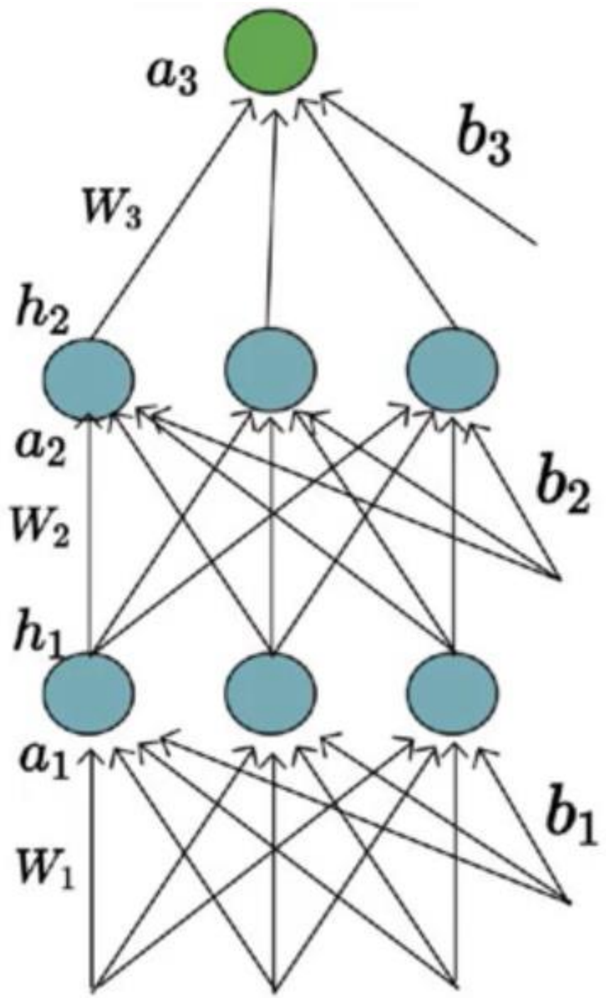
Sequential Models

Recurrent Neural Network (RNN)

Introduction

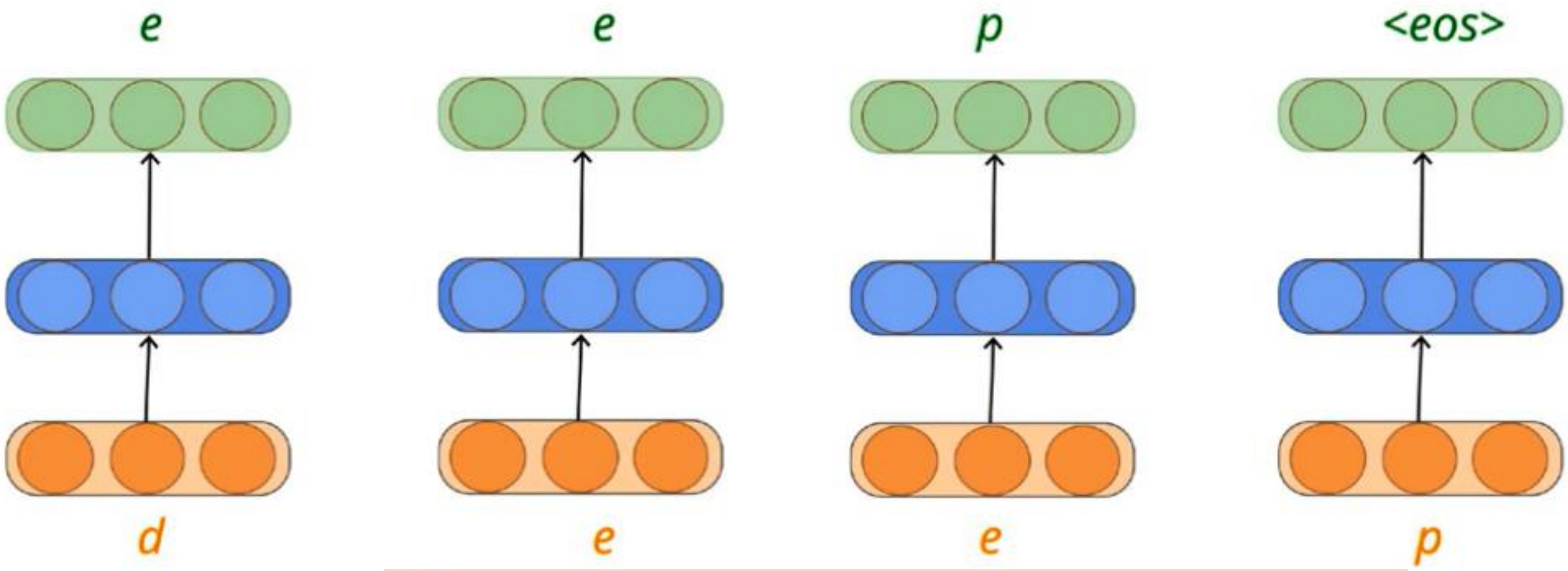


- Outputs are independent of previous inputs
- Input is of a fixed length



x1 height weight ... . sugar bp ECG
x2 height weight ... . sugar bp ECG

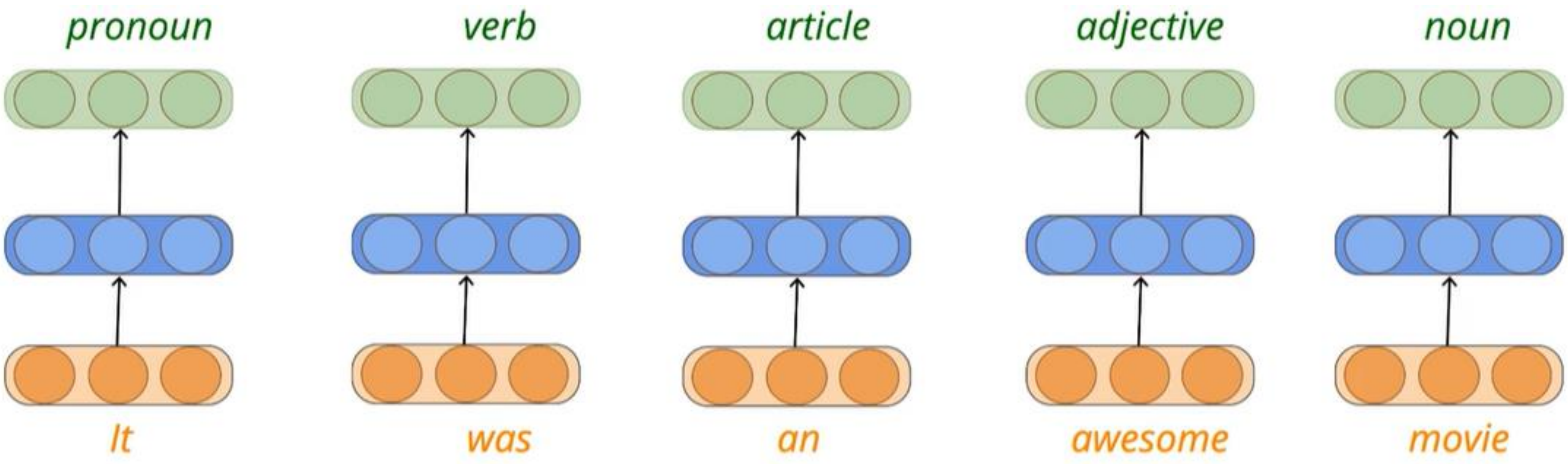
Sequence Learning Problems



- Outputs depend on previous inputs also
- The length of the input is not fixed

Sequence Learning Problems

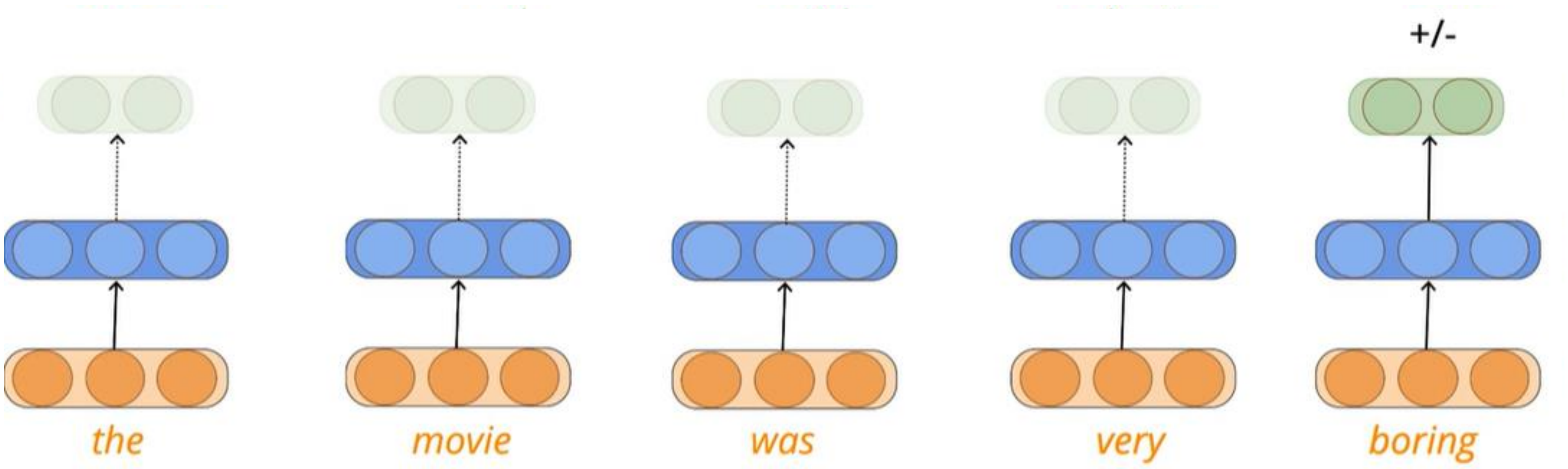
Sequence of words



1 hot encoding used to convert input to number

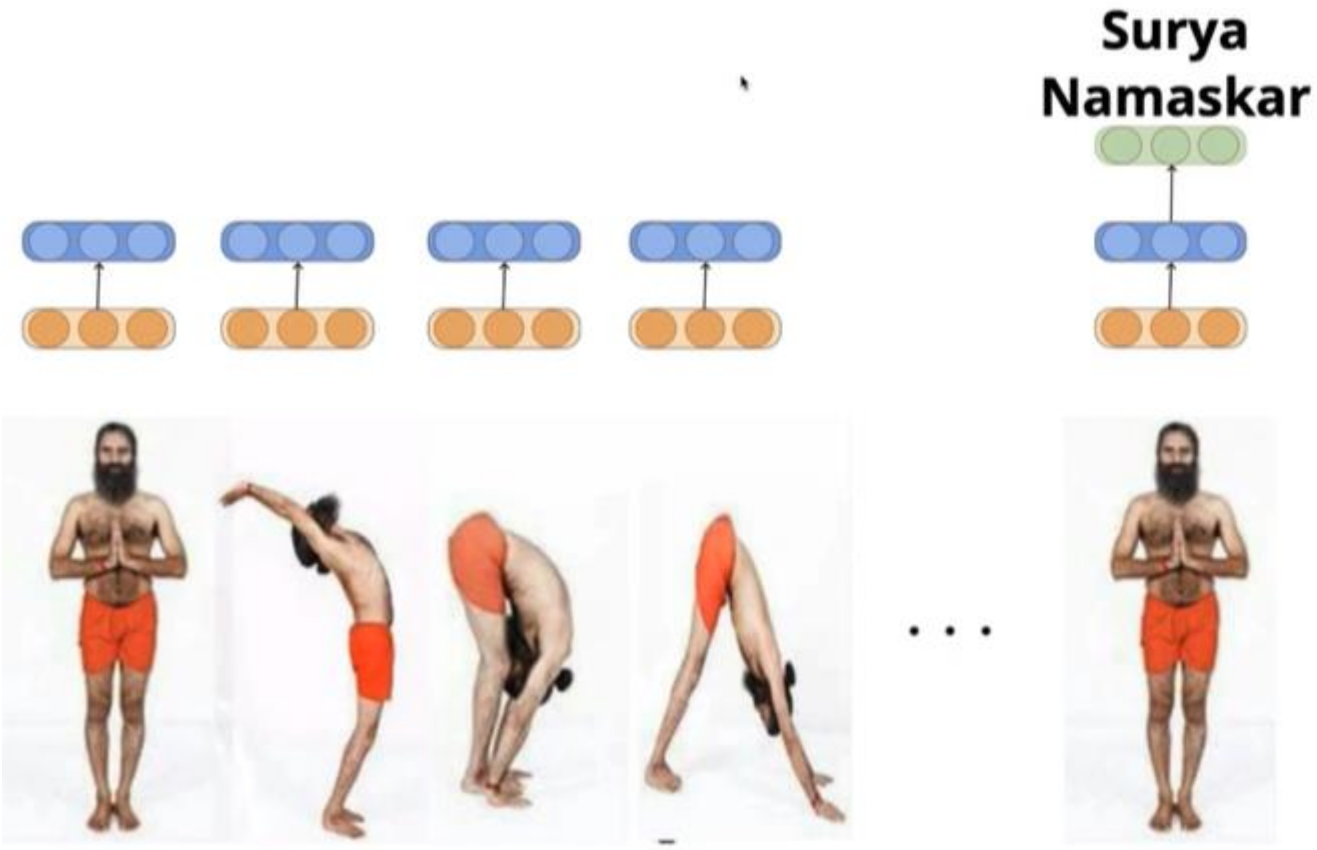
Awesome being an adjective has helped in identifying **movie** as noun
Some words like **bank**- may be used as noun or verb. So the context matters

Sequence Learning Problems- Predict the polarity of the sentence



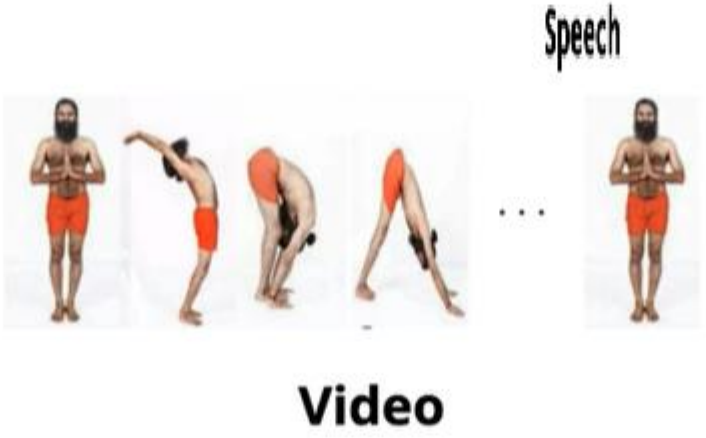
The no: of input and output may not be same. Here the input is a sequence of words and output is single which classifies the sentence polarity- positive or negative.

Other sequence learning problems



Classify the sequence of yoga posters as one Yoga exercise name - Input is a sequence of frames, and output is a single Yoga exercise name. challenges- Variable no of frames based on speed of action

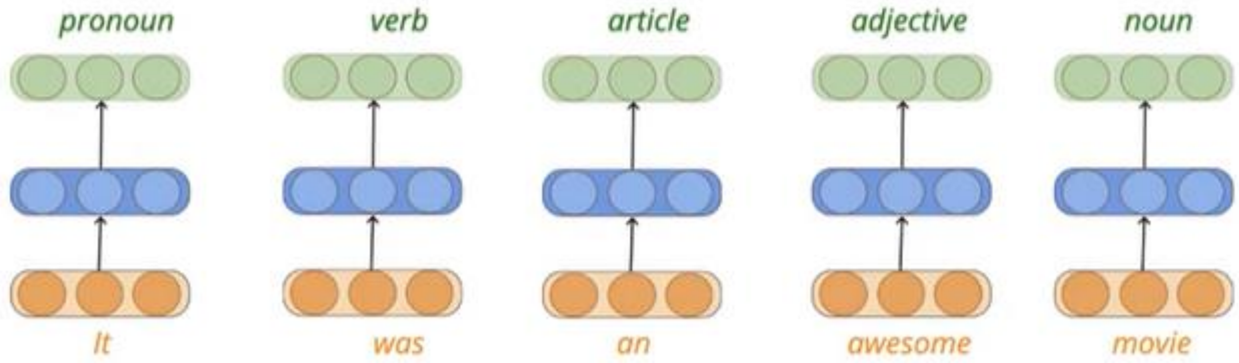
Speech Processing- another sequence learning problem. Take audio signals as input and classify each of them as phonemes



Yoga Video classification :. Each frame in the video correspond to a pose and we want to classify each of the frames into one pose resulting in a sequence of poses

What can be a solution?

- Ensure that y_t is dependent on previous inputs also
- Ensure that the function can deal with variable number of inputs
- **Ensure that the function executed at each time step is the same**



$$h_i = \sigma(W_1 x_i + b_1)$$

$$y_i = O(W_2 h_i + b_2)$$

$$i = timestep$$

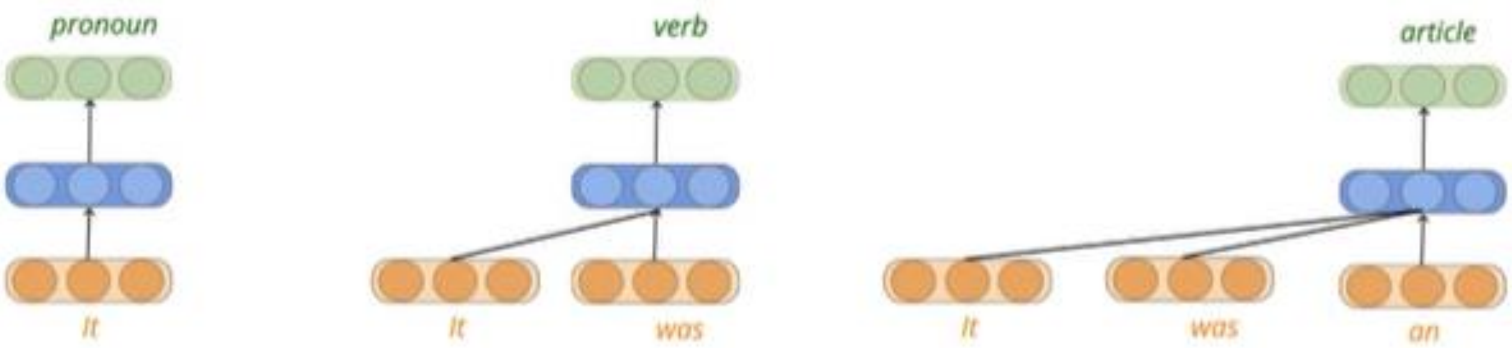
$$s_i = \sigma(U x_i + b)$$

$$y_i = O(V s_i + c)$$

Parameter Sharing

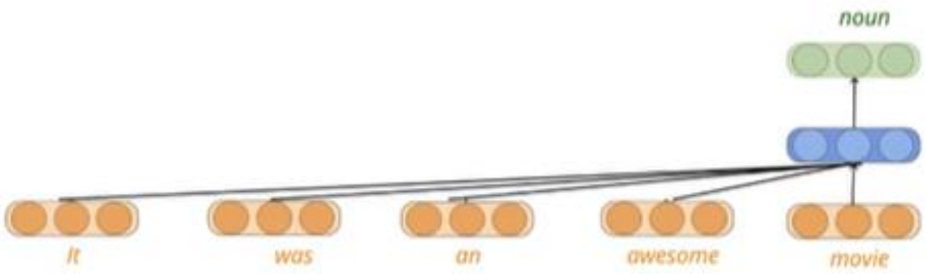
$$y_t = \hat{f}(x_1, x_2, ..., x_t)$$

This solution also does not satisfy all 3 criterias



$$s_i = \sigma(Ux_i + b)$$
$$y_i = O(Vs_i + c)$$

✓ **Parameter Sharing**



- ✓ Ensure that y_t is dependent on previous inputs also
- ✗ Ensure that the function can deal with variable number of inputs
- ✗ Ensure that the function executed at each time step is the same

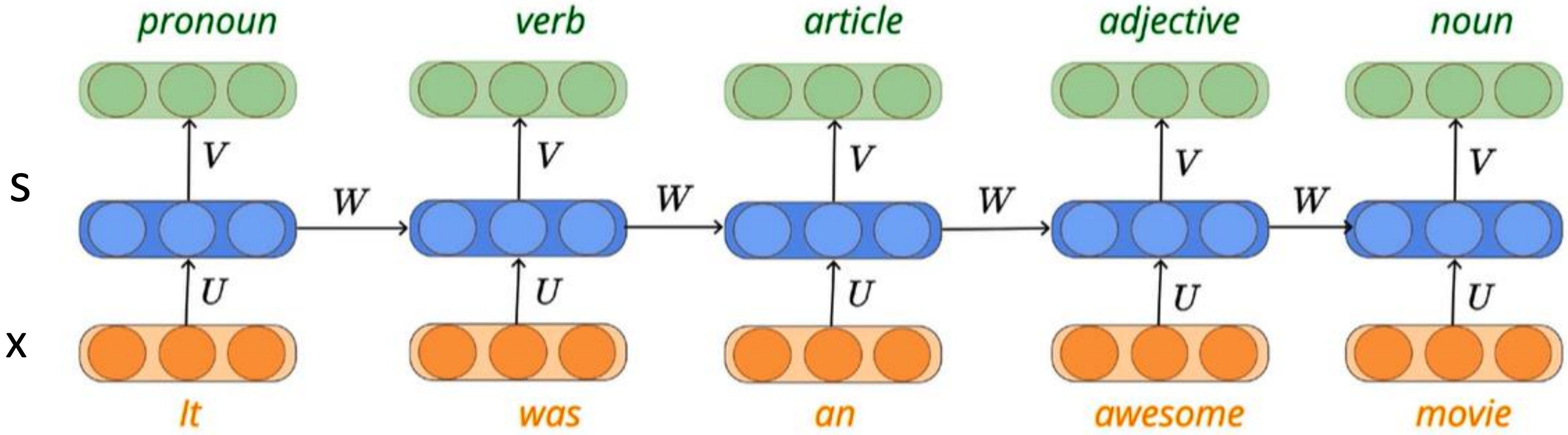
$$y_1 = f(x_1)$$
$$y_2 = f(x_1, x_2)$$
$$y_3 = f(x_1, x_2, x_3)$$

Still this solution does not satisfy all the 3 criterias

$$y_n = f(x_1, x_2, x_3, \ldots, x_n)$$

A solution – Recurrent Neural Networks (RNN)

RNN satisfies all the 3 criterias



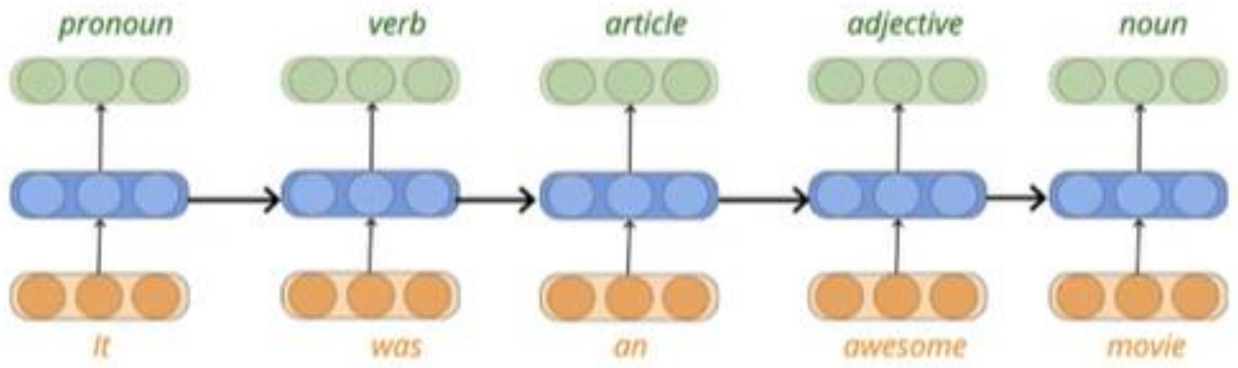
$$s_i = \sigma(Ux_i + Ws_{i-1} + b)$$
$$y_i = O(Vs_i + c)$$

$$y_i = \hat{f}(x_i, s_{i-1}, W, U, V, b, c)$$

RNN – Types of problems

$$y_i = \hat{f}(x_i, s_{i-1}, W, U, V, b, c)$$

- How do you represent words and characters as numbers ? **(data and tasks)**
- What is an appropriate loss function ? **(loss)**
- How do you train the model ? **(learning algorithm)**



No of inputs	No of Outputs	Appln
n	**1**	**Classification**
n	**n**	**Parts of Speech tagging**
n	**m**	**Machine translation**

- **Sequence Classification** (sentiment classification, video classification)
- **Sequence Labelling** (part of speech tagging, named entity recognition)
- **Sequence Generation** (machine translation, **transliteration**)

Data and Task

<sos> start of sequence- to indicate that sentences is starting

<eos> end of sequence- to indicate that sentences is ending. Sometimes sentence end with.,?! Or nothing. Hence we give this

<pad> artificial word to make sure all sentences of equal length

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
<sos>	The	first	half	was	very	boring	.	<eos>	<pad>
<sos>	Great	performance	by	all	the	lead	actors	.	<eos>
<sos>	The	bacground	music	was	awesome	.	<eos>	<pad>	<pad>
<sos>	The	movie	was	a	waste	of	time	.	<eos>

y
0
1
1
0

- lower case all words
- compute the total number of unique words across all sentences (say, L --> 24 in the above case)
- Assign a unique id to each word (between 1 to L)
- Represent each word using a L dimensional binary vector with only the bit corresponding to the word id set to 1

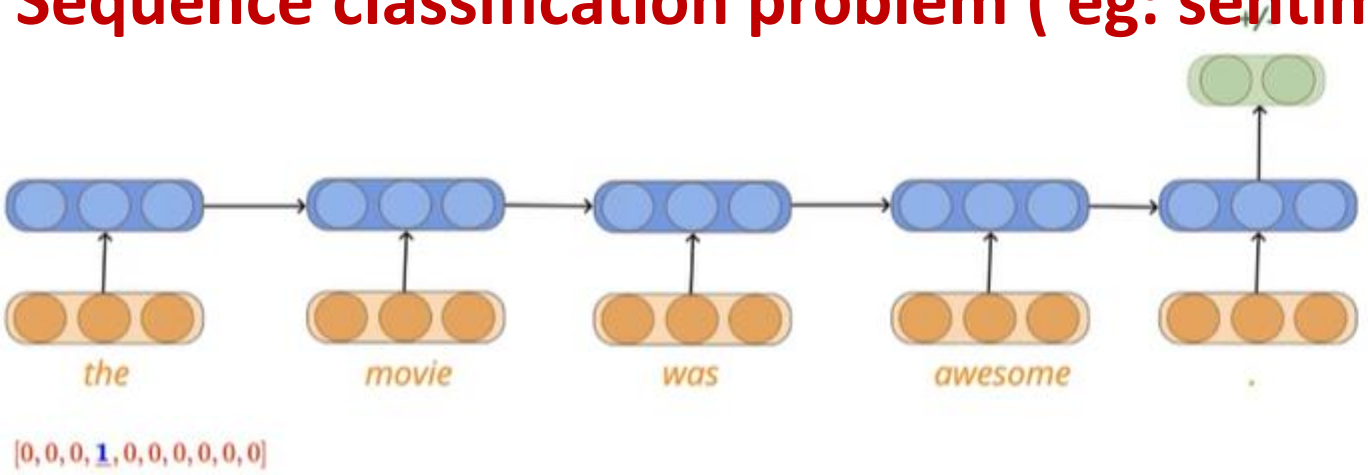
word	id
<sos>	1
<eos>	2
<pad>	3
the	4
first	5
half	6
...	
...	
time	24

1-hot vector representation

[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

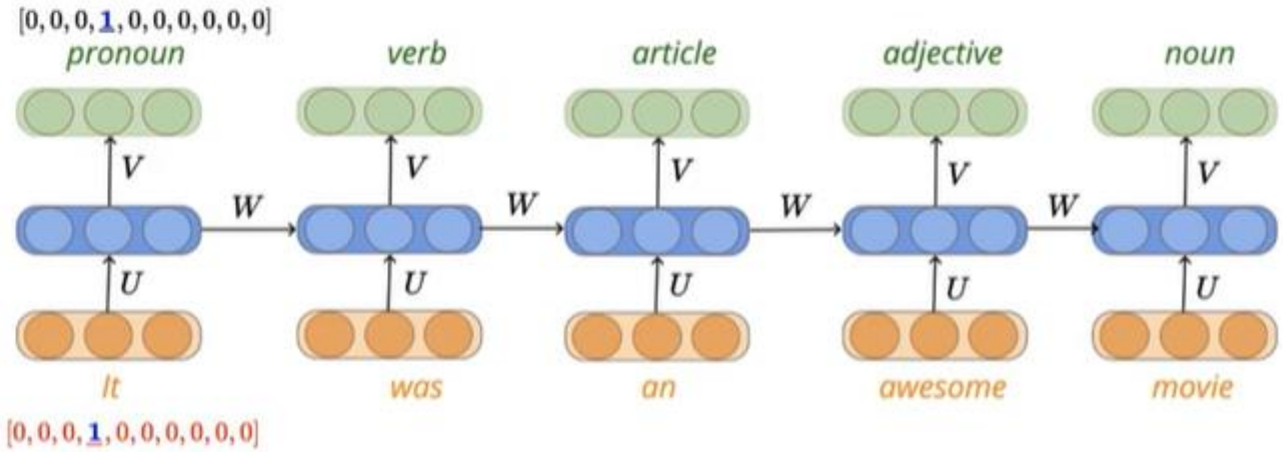
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Sequence classification problem (eg: sentiment analysis-Polarity)



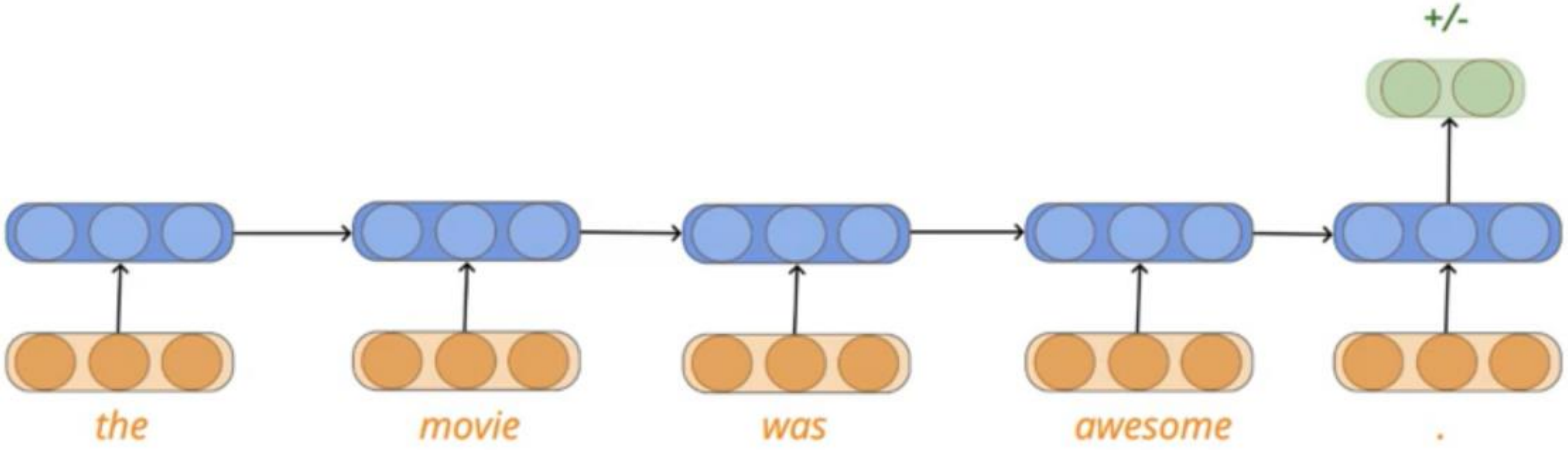
$$s_i = \sigma(Ux_i + Ws_{i-1} + b)$$
$$\hat{y} = O(Vs_T + c)$$
$$T = sequence\ length$$

Sequence Labelling problem (eg: parts of speech tagging)



$$s_i = \sigma(Ux_i + Ws_{i-1} + b)$$
$$\hat{y}_i = O(Vs_i + c)$$

Loss function for sequence classification problem

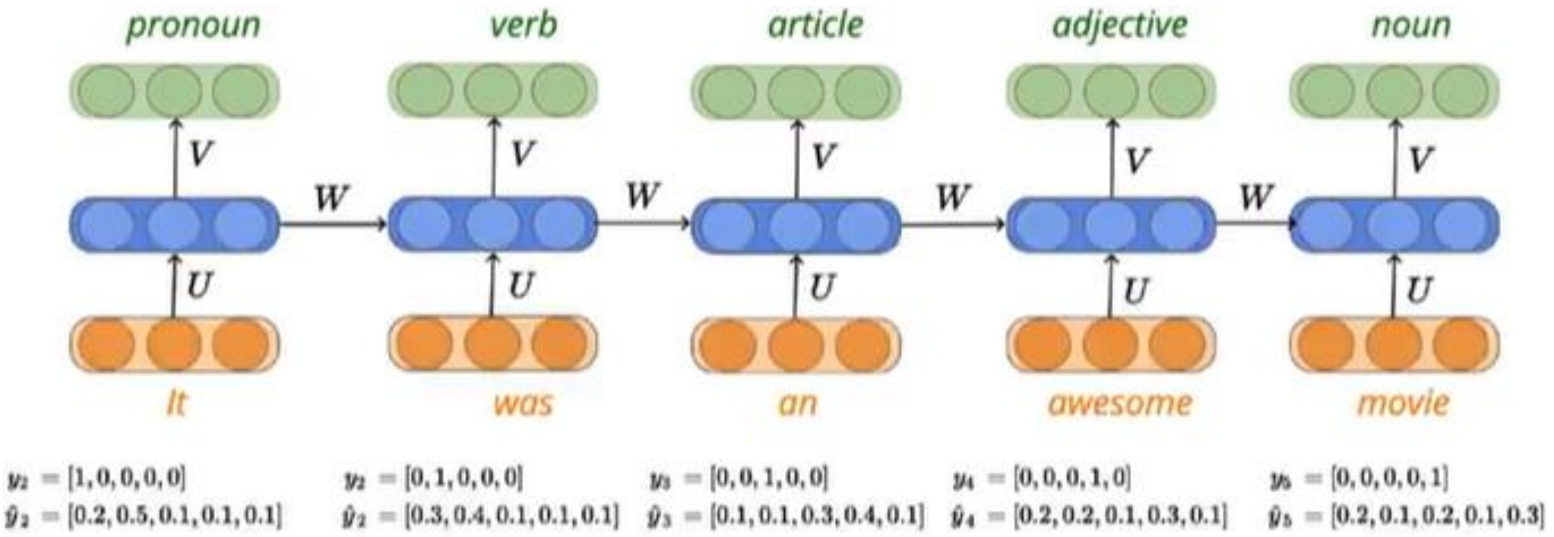


$$\mathcal{L}(\theta) = -\sum_{i=0}^{1} y[i] \log \hat{y}[i]$$
$$= -\log y_c$$
$$= -\log 0.7$$

$$y = [1, 0]$$
$$\hat{y} = [0.7, 0.3]$$

Only one output at the end
Possible output classes [1,0]
Y[i]=0 for one hence -log yc

Loss function for sequence labelling problem



$$\mathscr{L}(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{T}\log y_{ijc}$$

Every time step has an output

Sum up for each time step (T)for each data samples (m) and average

Training Algorithm – Back propagation

Initialise w, b

Iterate over data:

$compute\ \hat{y}$

$compute\ \mathscr{L}(w, b)$

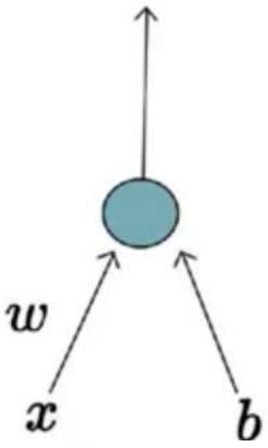
$w_{11} = w_{11} - \eta \Delta w_{11}$

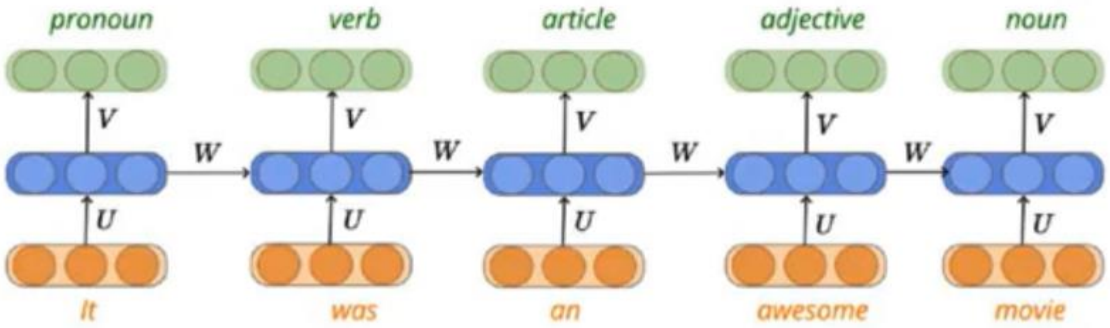
$u_{12} = u_{12} - \eta \Delta u_{12}$

....

$v_{13} = v_{13} - \eta \Delta v_{13}$

till satisfied





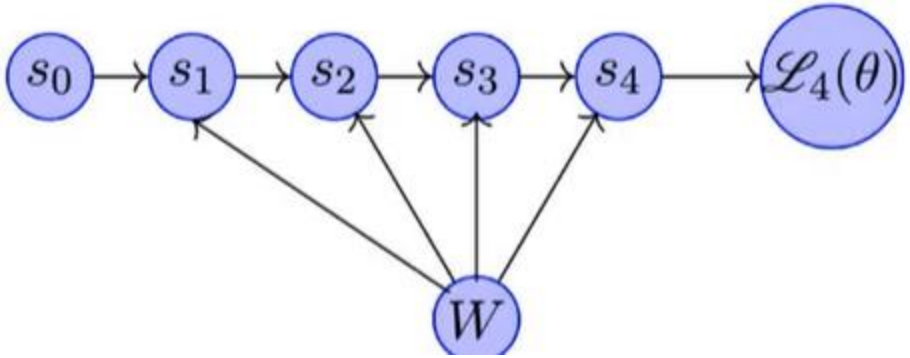
$$Earlier : w, b$$

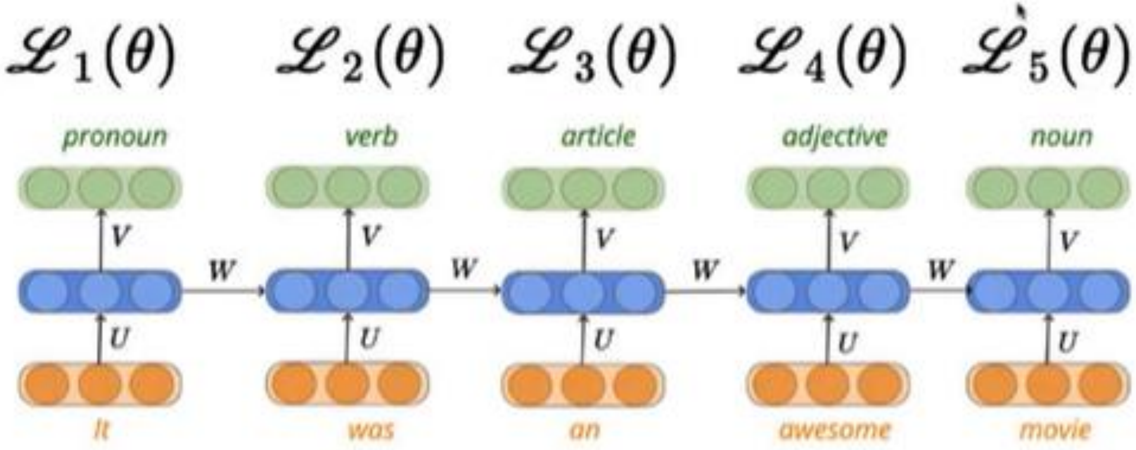
$$Now : w_{11}, w_{12}, ..., u_{11}, u_{12}, ..., v_{11}, v_{12}$$

$$Earlier : L(w, b)$$

$$Now : L(W, U, V)$$

Learning Algorithm (Derivative of loss function w.r.t w)





$$\frac{\partial \mathscr{L}_4(\theta)}{\partial W} = \frac{\partial \mathscr{L}_4(\theta)}{\partial s_4}\frac{\partial s_4}{\partial W}$$

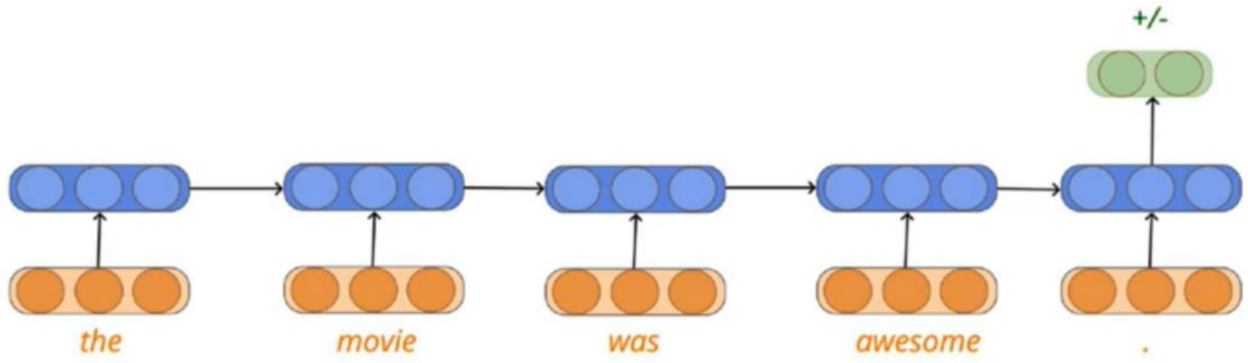
$$\frac{\partial s_4}{\partial W} = \frac{\partial s_4}{\partial W} + \frac{\partial s_4}{\partial s_3}\frac{\partial s_3}{\partial W} + \frac{\partial s_4}{\partial s_3}\frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial W} + \frac{\partial s_4}{\partial s_3}\frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial W}$$

$$\frac{\partial s_4}{\partial W} = \frac{\partial s_4}{\partial s_4}\frac{\partial s_4}{\partial W} + \frac{\partial s_4}{\partial s_3}\frac{\partial s_3}{\partial W} + \frac{\partial s_4}{\partial s_2}\frac{\partial s_2}{\partial W} + \frac{\partial s_4}{\partial s_1}\frac{\partial s_1}{\partial W}$$

$$\frac{\partial s_4}{\partial W} = \sum_{k=1}^{4}\frac{\partial s_4}{\partial s_k}\frac{\partial s_k}{\partial W}$$

Similarly derivative w.r.t V and U

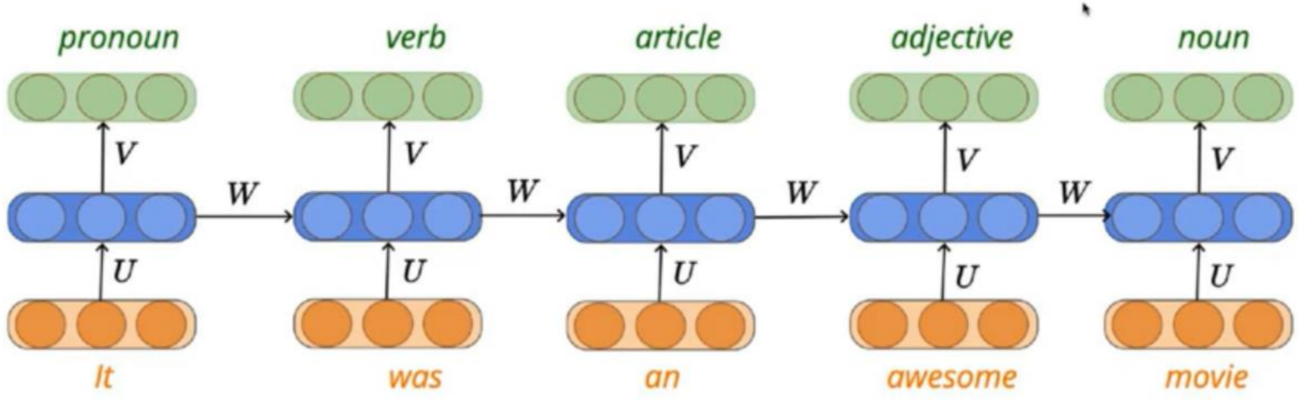
Evaluation- Sequence classification



Accuracy $= \dfrac{\boldsymbol{No:of\ correctly\ classified}}{\boldsymbol{Total\ samples}}$

Predicted y cap	Ground Truth y	Correct/incorrect
1(P)	1(P)	correct
0(N)	0(N)	incorrect
1(P)	0(N)	incorrect
0(N)	0(N)	correct
1(P)	1(P)	correct
0(N)	1(P)	incorrect

Evaluation Sequence labelling



Overall accuracy /Accuracy per class

Data sample	Pronoun	verb	article	adjective	noun
D1	P	V	Ar	Ad	N
D2					
D3					
D4					
D5					

- **Overall Accuracy**
- **Accuracy Per Class**
- **Confusion Matrix**

Confusion Matrix

	Pronoun	verb	article	adjective	noun
Pronoun	3	2	3	5	6
verb	2	7			
article			3		
adjective				2	
noun					1

Namah Shivaya

Courtesy : Video lectures of Dr.Mitesh Kapra