Sequence Models

Dr.Venkataramana Veeramsetty
NVIDIA DLI Instructor & Faculty@CAIDL
Assistant Professor
Dept. of EEE
SR Engineering College

Email : dr.vvr.research@gmail.com
(Venkataramana Veeramsetty)

December 31, 2020

Problems dealing with sequential data

- ✓ Speech Recognition:
  - Both input "X" and output "Y" are in sequence
  - Able to convert speech into text
  - X is an audio clip and so that plays out over time and Y, the output, is a sequence of words.
- ✓ Music Generation:
  - Input "X" is an empty or single integer and output "Y" are in sequence
- ✓ Sentiment Classification:
  - Input "X" is sequential data like "Movie is not good" and output "Y" may be rating in terms of an integer value

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Problems dealing with sequential data

- ✓ DNA Sequence Analysis:
  - Input "X" is DNA sequence, and output "Y" is also sub sequence in DNA
- ✓ Machine Translation:
  - Input "X" is a sequence of words in one language and Output "Y" is a sequence of words in target language
- ✓ Video activity recognition:
  - Input "X" is sequence of frames in video, output "Y" tells the activity
- ✓ Name entity recognition:
  - Input "X" is a sentence which is sequence of words, output "Y" consists names in that sentence

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Representation of Sequence: Name entity identification

X(1)	Ram	is	a	good	boy
	$X^{1,1}$	$X^{1,2}$	$X^{,13}$	$X^{1,4}$	$X^{1,5}$
Y(1)	1	0	0	0	0
	$Y^{1,1}$	$Y^{1,2}$	$Y^{1,3}$	$Y^{1,4}$	$Y^{1,5}$

In$X^{i,t}$, "i" represents sample number and "t" represents time stamp.

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

One-Hot Vector

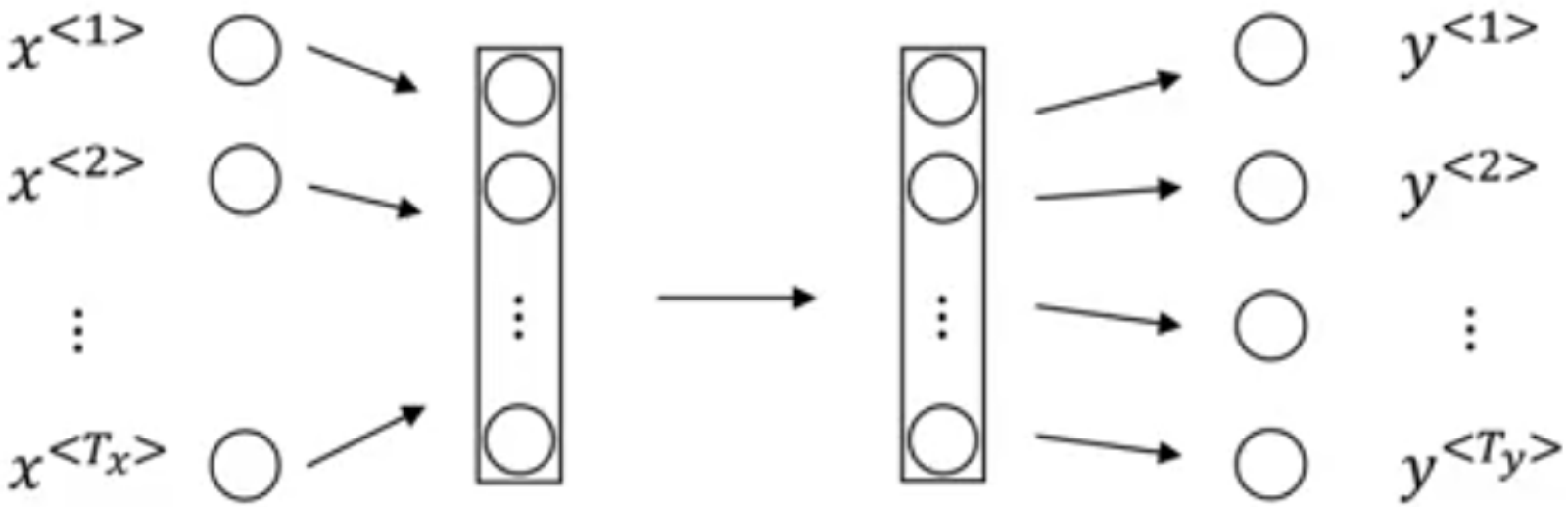
Let Assume a dictionary as follows

a	an	are	boy	girl	good	is	Ram	Sitha	Zero
1	2	3	4	5	6	7	8	9	10

Now X(1) can be represented as

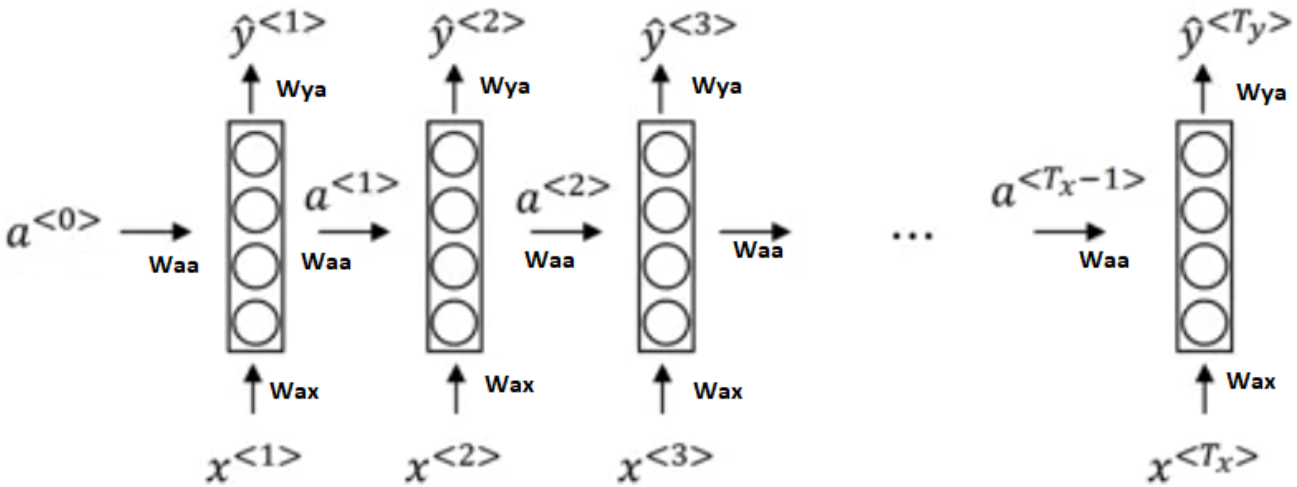
Ram	$X^{1,1}$	0	0	0	0	0	0	0	1	0	0
is	$X^{1,2}$	0	0	0	0	0	0	1	0	0	0
a	$X^{1,3}$	1	0	0	0	0	0	0	0	0	0
good	$X^{1,4}$	0	0	0	0	0	1	0	0	0	0
boy	$X^{1,5}$	0	0	0	1	0	0	0	0	0	0

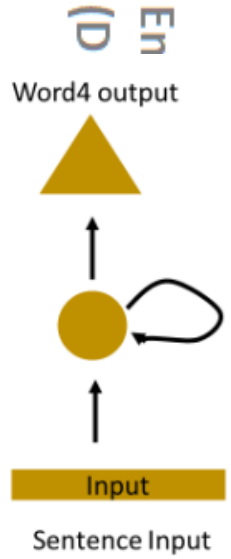
Can we ANN for Sequential Data Problems



- ✓ Not able to use for different sequential lengths
- ✓ Not able to share features learned across different positions of text
- ✓ Network will be more complex with size of

Recurrent Neural Network





$$a^{<0>} = 0 \tag{1}$$

$$a^{<t>} = g(Wax * X^{<1>} + Waa * a^{<t-1>} + b_a) \tag{2}$$

$$y^{<t>} = f(Wya * a^{<t>} + b_y) \tag{3}$$

Activation function to calculate $a^{<}t>$ is either Tanh or Relu. Where as for output it is depending on type of problem. For regression and binary classification can use sigmoid, for multiple classification can use softmax.

- Predict character "o" after feeding characters "h","e","l","l" sequentially
- one hot encoding dictionary

	h	e	l	o
h	1	0	0	0
e	0	1	0	0
l	0	0	1	0
0	0	0	0	1

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Assume 3 hidden neurons in hidden layer, size of W_{ax} is 3X4 and W_{aa} is 3X3 and size of bias matrix s 3X1

Initialize Weights and Bias Parameters and activation states

$$W_{ax} = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ -0.1 & -0.2 & 0.3 & 0.1 \\ 0.1 & 0.2 & -0.3 & -0.4 \end{bmatrix}$$

$$W_{aa} = \begin{bmatrix} 0.1 \\ -0.1 \\ 0.1 \end{bmatrix}$$

$$\begin{bmatrix} b_a & b_y \\ 0.1 & 0.1 \\ 0.2 & 0.2 \\ 0.3 & 0.3 \\ & 0.4 \end{bmatrix}$$

$$W_{ya} = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ -0.1 & -0.2 & 0.3 \end{bmatrix}$$

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Pass first input h=[1;0;0;0] and compute state of hidden layer using equation shown below.

$$a^{<t=1>} = g(Wax * X^{<t>} + Waa * a^{<t-1>} + b_a) \quad (4)$$

$$\tanh\left(\begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ -0.1 & -0.2 & 0.3 & 0.1 \\ 0.1 & 0.2 & -0.3 & -0.4 \end{bmatrix} * \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ -0.1 & -0.2 & 0.3 \\ 0.1 & 0.2 & -0.3 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}\right)$$

$$a(t=1) = \begin{bmatrix} 0.197 \\ 0.099 \\ 0.379 \end{bmatrix}$$

Email : dr.vramana.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Compute output using

$$y^{<t>} = f(Wya * a^{<t>} + b_y) \quad (5)$$

$$\text{softmax}\left(\begin{bmatrix} 0.1 & 0.2 & 0.3 \\ -0.1 & -0.2 & 0.3 \\ 0.1 & 0.2 & -0.3 \\ -0.1 & -0.2 & 0.3 \end{bmatrix} * \begin{bmatrix} 0.197 \\ 0.099 \\ 0.379 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix}\right)$$

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

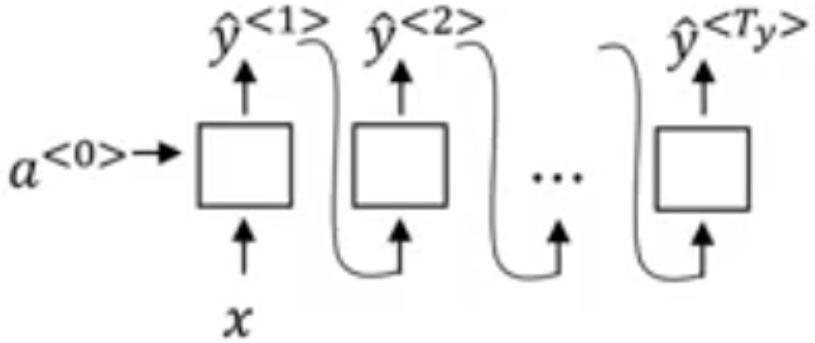
Back Propagation Through Time

$$L(Y^P, Y) = -Y log(Y^P) - (1 - Y) log(1 - Y^P) \quad (6)$$

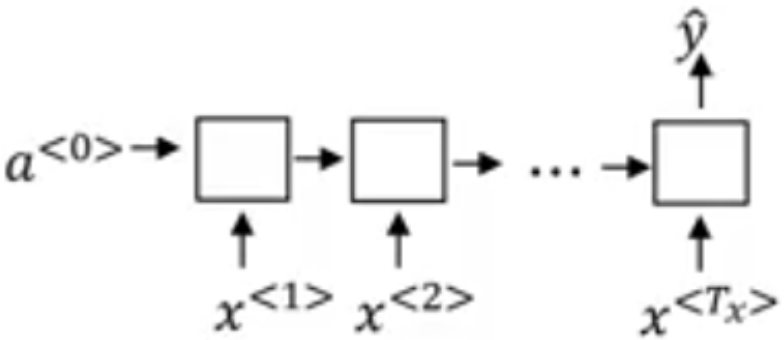
$$L(Y^P, Y) = \sum_{t=1}^{T_y} (-Y^t log(Y^{P,t}) - (1 - Y^t) log(1 - Y^{P,t})) \quad (7)$$

RNN Models

One To Many - Music Generation



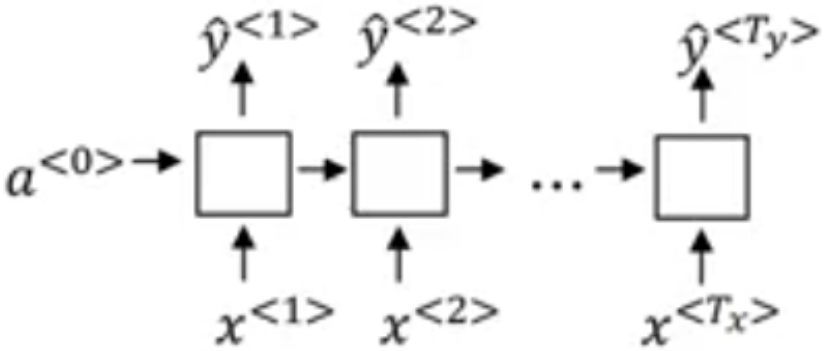
Many To One - Sentiment Classification



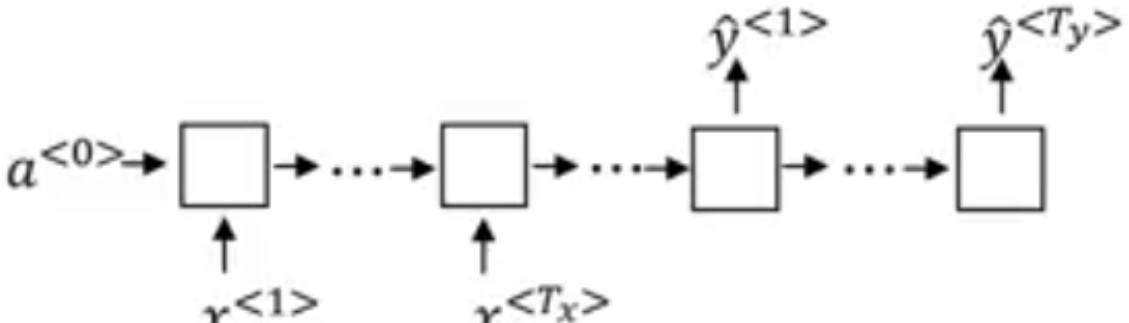
Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

RNN Models

Many To Many (Equal Length) - Name entity identification

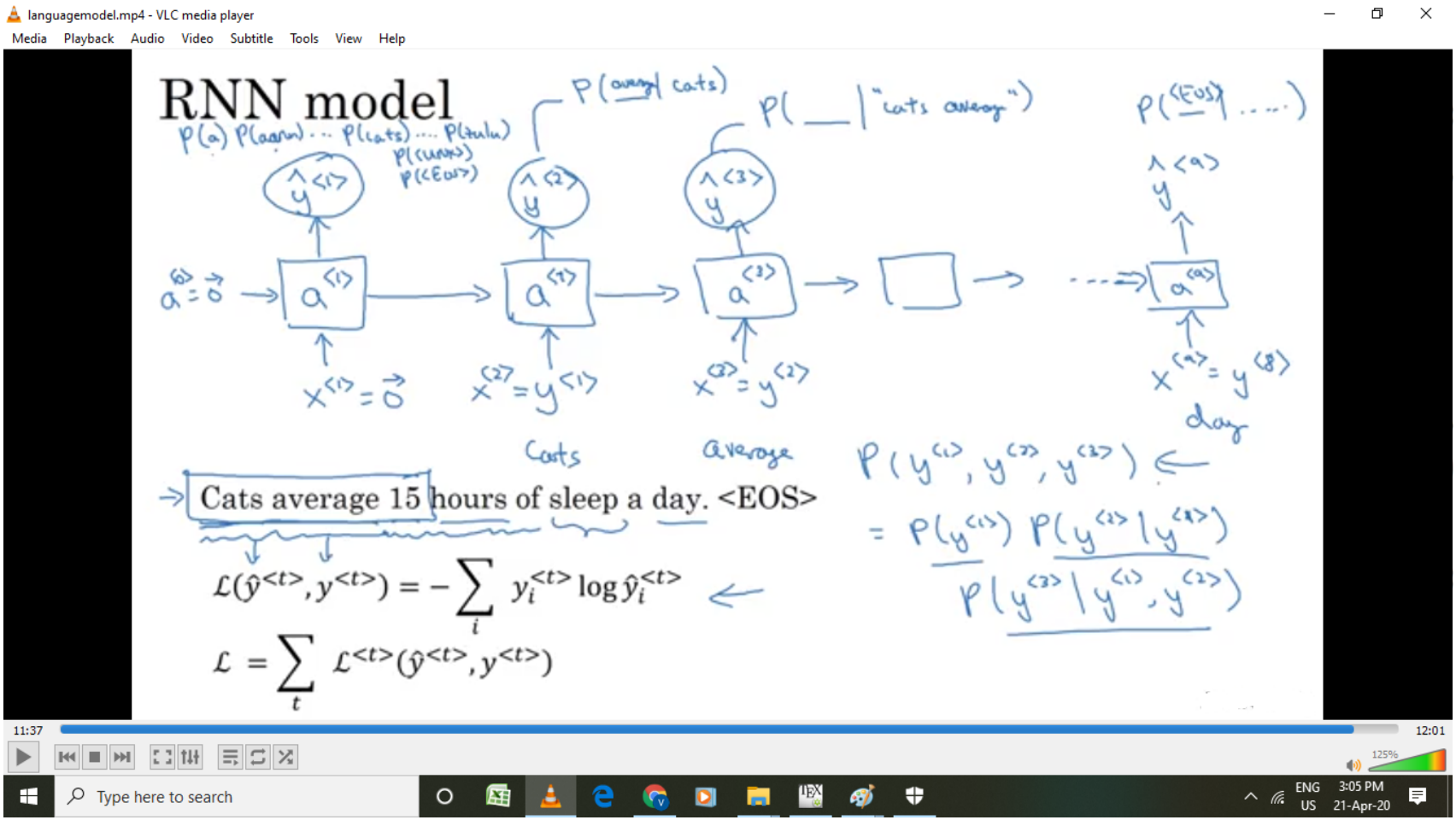


Many To One (Unequal Length) - Language Translation

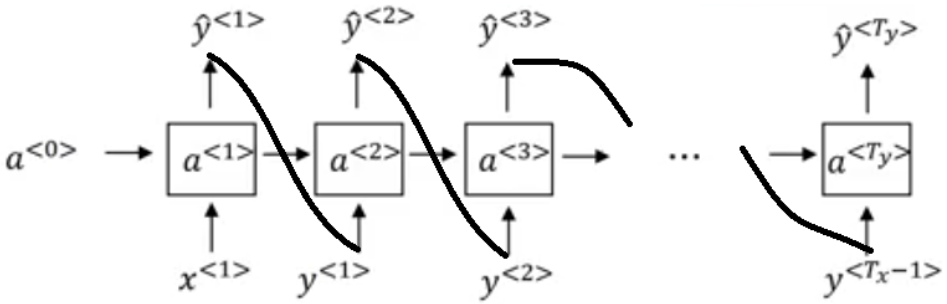


Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Language Modeling



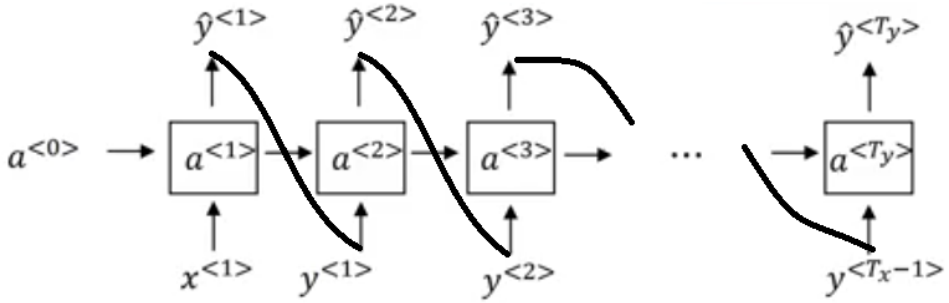
Word Level Sequence Generation using Trained RNN



- ✓ Vocabulary=[a,aaron,...,zuzu,..UNK,EOS]
- ✓ $y^{<1>}$=np.random.choice(p(a),p(aaron),..,p(zuzu),...,
- ✓ Based on T_y, sequence generation will stop
- ✓ Can generate until "UNK" generated

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Character Level Sequence Generation using Trained RNN



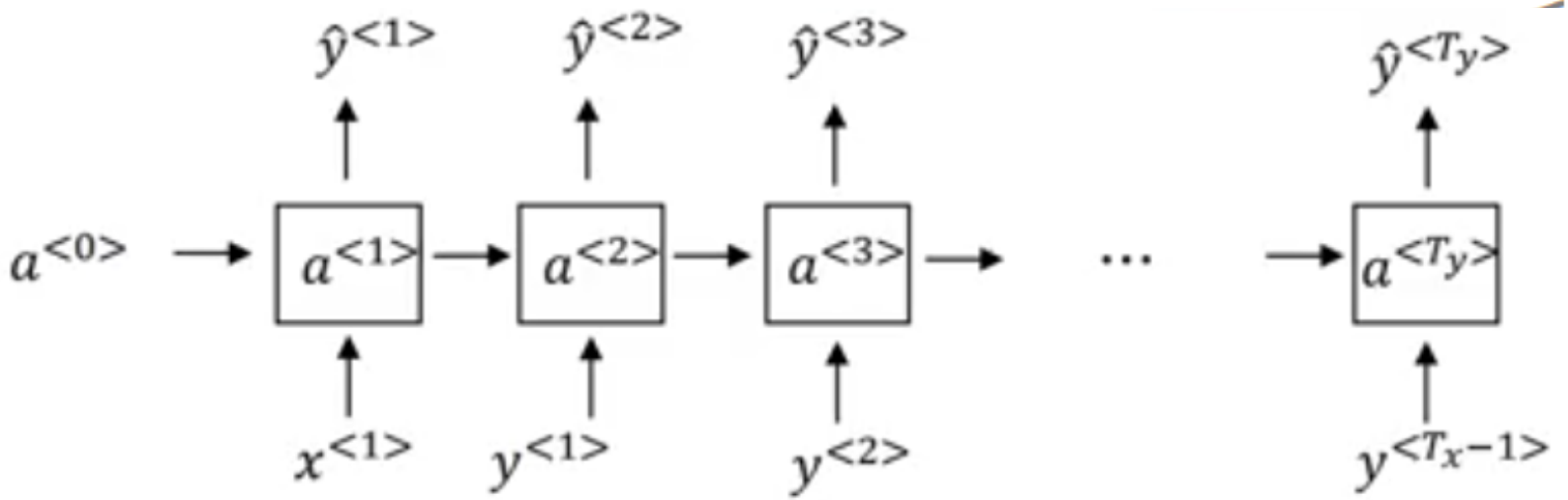
- ✓ Vocabulary=[a,b,..,z,-,*,..A,B,..Z]
- ✓ $y^{<1>}$=np.random.choice(p(a),p(b),..,p(*),...,p(Z))

Limitations:

- ✓ Not recognize unknown word
- ✓ High complex
- ✓ More computation time
- ✓ High dependent previous time step.

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Vanishing Gradient and Exploding Gradient



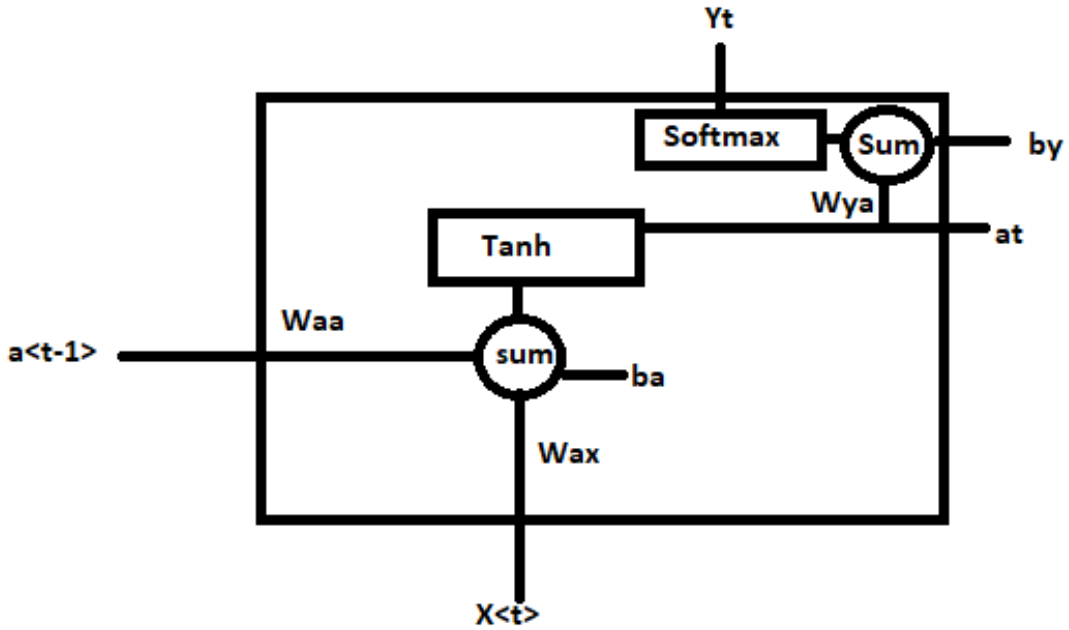
- ✓ RNN fails for long term dependency
  - The **cat**, which ate apple,banana,....,**was** full
  - The **cats**, which ate apple,banana,....,**were** full

Vanishing Gradient and Exploding Gradient

- ✓ Output at particular time stamp highly depends on near time stamps and weakly effected by initial time stamps.
- ✓ Gradient of error will be have neglisible impact on initial weights (Gradient vanishing)
- ✓ Some times gradients may increase to high value (Exploding gradient), will lead high values of parameters.
- ✓ Exploding gradient can be overcome by using gradient clipping method

Email: dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

RNN- Graphical Representation



$$a_t = g(W_{ax}x_t + W_{aa}a_{t-1} + b_a) \quad (8)$$

$$Y_t = f(W_{ya}a_t + b_y) \quad (9)$$

g: tanh/Relu

f:Sigmoid/Softmax

RNN:Limitations

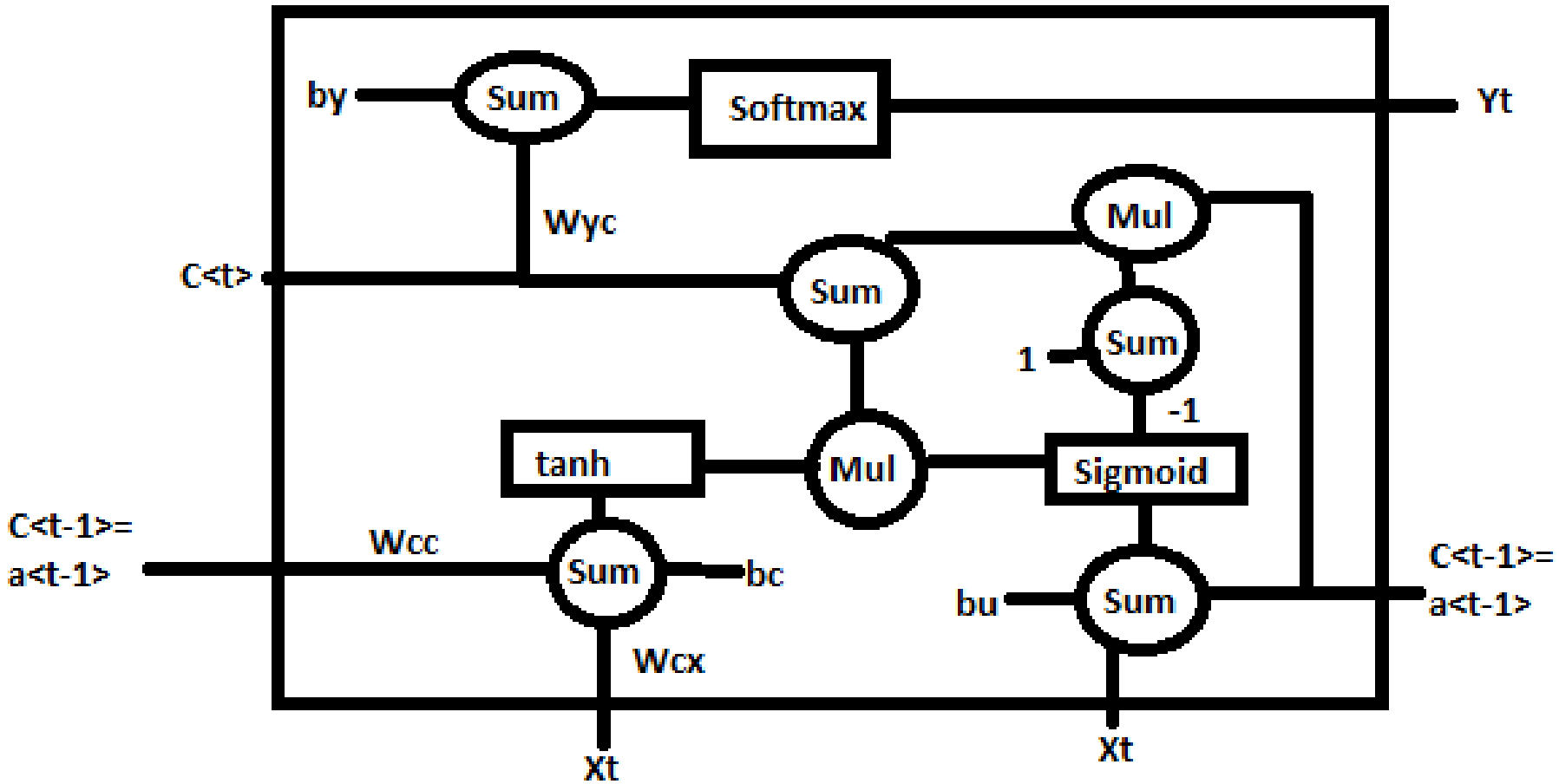
- ✓ RNN will estimate word/character/number in the sequence based on previous time stamps, not based on future dependency
- ✓ Due to deep structure, RNN has vanishing gradient problem
- ✓ Due to deep structure, RNN may has exploding gradient problem
- ✓ Due to deep structure, the value at particular time stamp may forget initial values in RNN

Email : dr.vvr.research@gmail.com
(Dr.Venkataramana Veeramsetty)

Gated Recurrent Unit (GRU)

- ✓ GRU can be helpful for long term dependency and also overcome vanishing gradient problem
  - The cat, which already ate apple,..., was full
  - The cats, which already ate apple,..., were full

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)


by
Sum
Softmax
Yt
Wyc
C<t>
Sum
Mul
1
Sum
-1
tanh
Mul
Sigmoid
C<t-1>=
a<t-1>
Wcc
Sum
bc
bu
Sum
C<t-1>=
a<t-1>
Wcx
Xt
Xt

$$C_t = a_t \tag{10}$$

$$C_t^{hat} = tanh(W_{cc} C_{t-1} + W_{cx} X_t + bc) \tag{11}$$

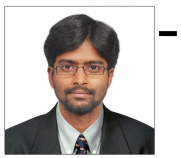
$$\Gamma_u = sigmoid(W_{uc} C_{t-1} + W_{ux} X_t + bu) \tag{12}$$

$$C_t = \Gamma_u * C_t^{hat} + (1 - \Gamma_u) * C_{t-1}^{hat} \tag{13}$$

C_t represents memory cell

C_t^{hat} represents candidate cell

Γ_u will be either 0 or 1



Full GRU

- ✓ It will have control over candidate cell of previous time stamp to generate current time stamp candidate cell.

$$C_t^{hat} = tanh(W_{cc} * \Gamma_r * C_{t-1} + W_{cx} X_t + bc) \quad (14)$$

$$\Gamma_u = sigmoid(W_{uc} C_{t-1} + W_{ux} X_t + bu) \quad (15)$$

$$\Gamma_r = sigmoid(W_{rc} C_{t-1} + W_{ux} X_t + br) \quad (16)$$

$$C_t = \Gamma_u * C_t^{hat} + (1 - \Gamma_u) * C_{t-1}^{hat} \quad (17)$$

Long short Term Memory (LSTM)

$$C_t^{hat} = tanh(W_{cc} * a_{t-1} + W_{cx}X_t + bc) \quad (18)$$

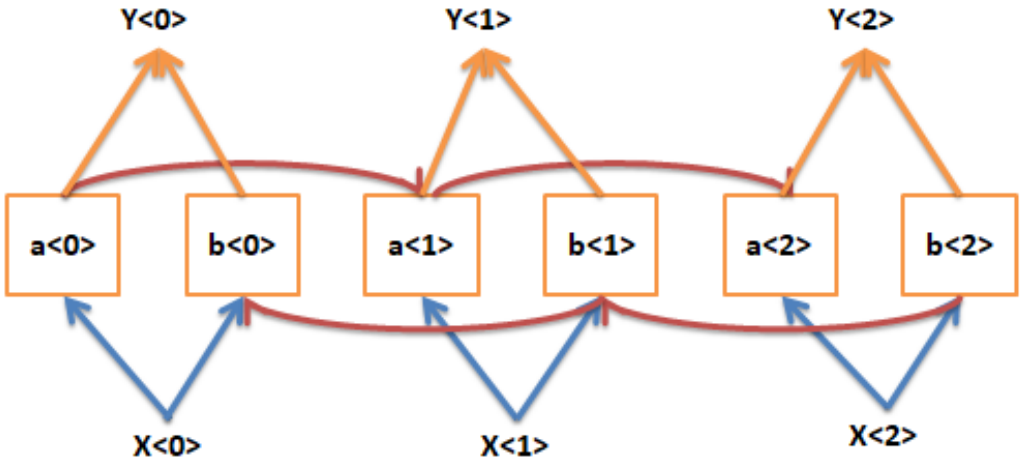
$$\Gamma_u = sigmoid(W_{ua}a_{t-1} + W_{ux}X_t + bu) \quad (19)$$

$$\Gamma_f = sigmoid(W_{fc}a_{t-1} + W_{fx}X_t + bf) \quad (20)$$

$$\Gamma_o = sigmoid(W_{oc}a_{t-1} + W_{ox}X_t + bo) \quad (21)$$

$$C_t = \Gamma_u * C_t^{hat} + \Gamma_f * C_{t-1}^{hat} \quad (22)$$

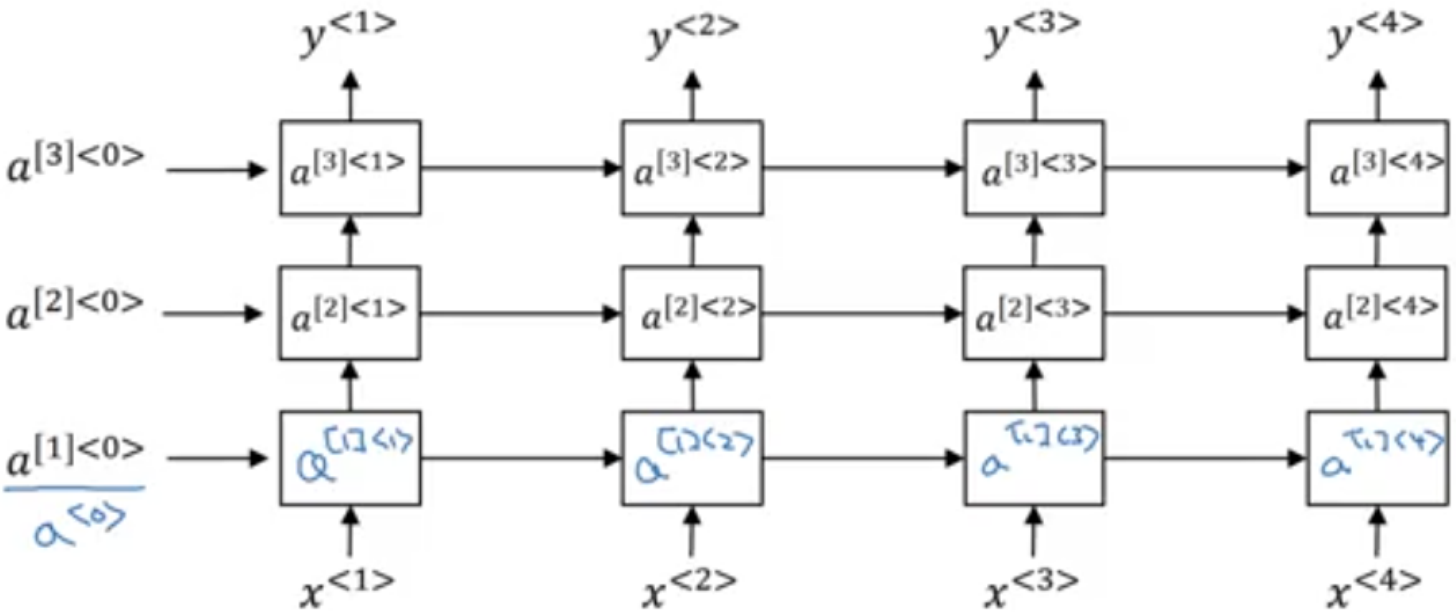
Bi-directional RNN



$$Y < t >= g(Wya * a < t > + Wyb * b < t > + by) \quad (24)$$

- It can predict based on previous and upcoming information

Deep-RNN



$$a < l >< t >= g(W_{l-1,t} * a < l-1, t > + W_{l,t-1} * a < l, t-1 > + b < l >) \quad (25)$$

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Word embedding

Let Assume a dictionary as follows

a	an	Arjun	boy	girl	good	is	Ram	Zero	Zoo
1	2	3	4	5	6	7	8	9	10

Now X(1) can be represented as

Ram	$X^{1,1}$	0	0	0	0	0	0	0	1	0	0
is	$X^{1,2}$	0	0	0	0	0	0	1	0	0	0
a	$X^{1,3}$	1	0	0	0	0	0	0	0	0	0
good	$X^{1,4}$	0	0	0	0	0	1	0	0	0	0

- Ram is a good ..... ? Laxman is a good ..... ?

Expected one hot vector for both is

Indexing of words in embedding matrix

Word	Index
A	O_1
An	O_2
Arjun	O_3
Boy	O_4
Girl	O_5
Good	O_6
Is	O_7
Ram	O_8
Zero	O_9
Zoo	O_{10}

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Feature matrix

Fe	a O_1	an O_2	Arjun O_3	boy O_4	girl O_5	good O_6	is O_7	Ram O_8	Zero O_9
Age	0	0	0.93	0.95	0.92	0	0	0.96	0
Gen	0	0	0.93	0.95	0.92	0	0	0.96	0
Place	0	0	0	0	0	0.1	0	0	0
Verb	0.9	0.9	0	0	0	0.1	0.9	0	0
Qual	0	0	0	0	0	0.9	0	0	0

As both Ram and Arjun have similar feature matrix. Will get same one hot vector for both below words

- Ram is a good boy
- Laxman is a good boy

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

Man-Woman :: King-?

e_man=[-1,0.01,0.01,0.09]

e_woman=[1,0.02,0.02,0.01]

e_man-e_woman=[-2,0,0,0]

Email : dr.vvr.research@gmail.com
(I eeramsetty)

Find e_w which suits in question mark in previous slide

$$arg^{w}_{max} = sim(e_w, e_{king} - e_{man} + e_{woman}) \quad (26)$$

Cosine Similarity

$$sim(u, v) = \frac{u^T v}{|u||v|} \quad (27)$$

Euclidean distance similarity

$$sim(u, v) = ||u - v||^2 \quad (28)$$

Applications

- Man:Woman as Boy:Girl
- Ottawa:Canada as Delhi:India
- Big:Bigger as Tall:Taller

Embedding matrix (E)

Fe	a O_1	an O_2	Arjun O_3	boy O_4	girl O_5	good O_6	is O_7	Ram O_8	Zero O_9
Age	0	0	0.93	0.95	0.92	0	0	0.96	0
Gen	0	0	0.93	0.95	0.92	0	0	0.96	0
Place	0	0	0	0	0	0.1	0	0	0
Verb	0.9	0.9	0	0	0	0.1	0.9	0	0
Qual	0	0	0	0	0	0.9	0	0	0

$$e_w = E.O_w \tag{29}$$

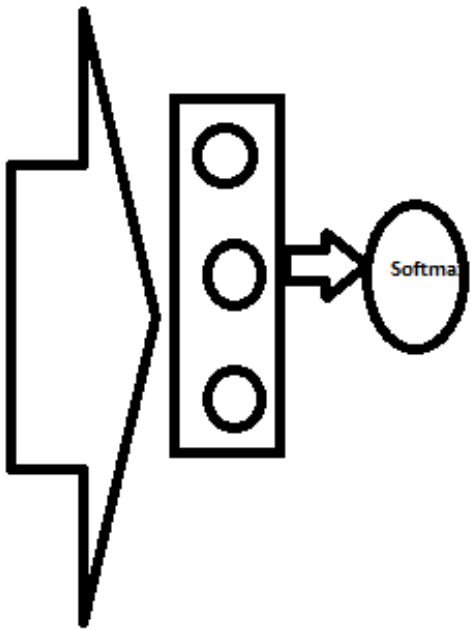
One-Hot Vector (O)

a	an	Arjun	boy	girl	good	is	Ram	Zero	Zoo
1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1

Natural Language Model

I	want	a	glass	of	orange	_____
4343	9665	1	3852	6163	6257	

I	o_{4343}	$\longrightarrow$ E $\longrightarrow$	e_{4343}
want	o_{9665}	$\longrightarrow$ E $\longrightarrow$	e_{9665}
a	o_1	$\longrightarrow$ E $\longrightarrow$	e_1
glass	o_{3852}	$\longrightarrow$ E $\longrightarrow$	e_{3852}
of	o_{6163}	$\longrightarrow$ E $\longrightarrow$	e_{6163}
orange	o_{6257}	$\longrightarrow$ E $\longrightarrow$	e_{6257}



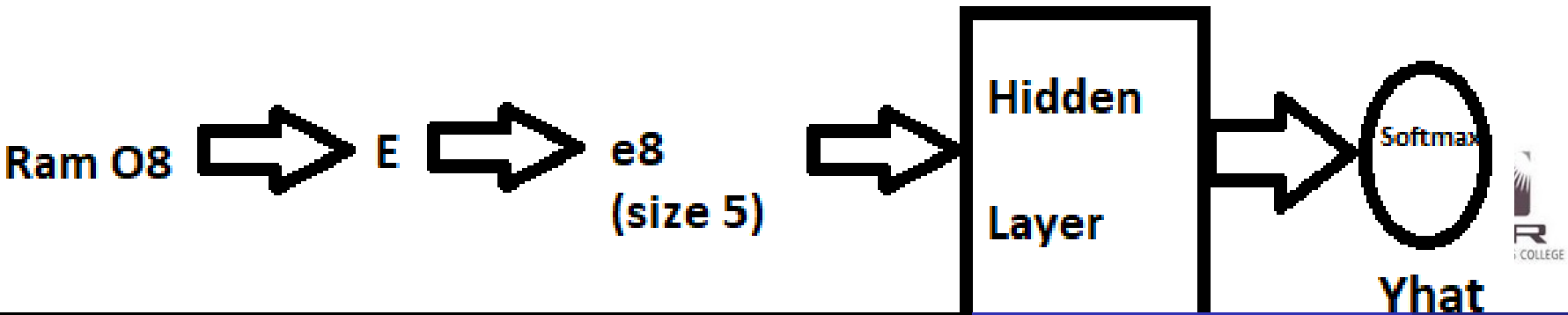
Skip-grams model

Ram is a good boy unlike rahul

- Choose context word and corresponding target word, And apply supervised learning, loss function categorical cross entropy

Context	Target
Ram	Is
Ram	good
Ram	boy

Email : dr.vvr.research@gmail.co
(Dr. Venkataramana Veeramsett



Negative Sampling

- Similar to Skip-gram model but more efficient

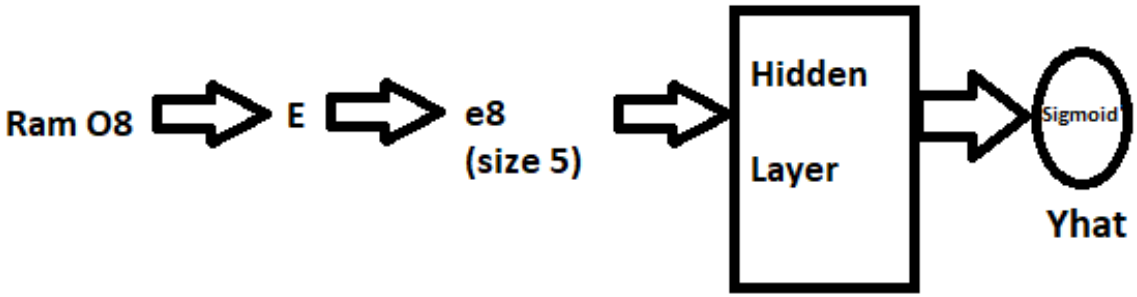
Ram is a good boy unlike rahul

- Choose context word and corresponding target word, and set output as 1, similarly choose negative target words for given sampling word and set output as zero

X		Y
Context	Target word	Output
Ram	Is	1
Ram	good	0
Ram	boy	0

- No. of negative sample generally between 5-10

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)



Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Note:

Number of neurons in output layer depends on size of vocabulary, however will update K+1 neurons. K is number of negative words. Negative examples will choose based on below equation

$$P(w_i) = \frac{f(w_i)^{\frac{3}{4}}}{\sum_{Voc.size} f(w_i)^{\frac{3}{4}}} \quad (30)$$

GloVe - Global Vectors for word representation

Minimize

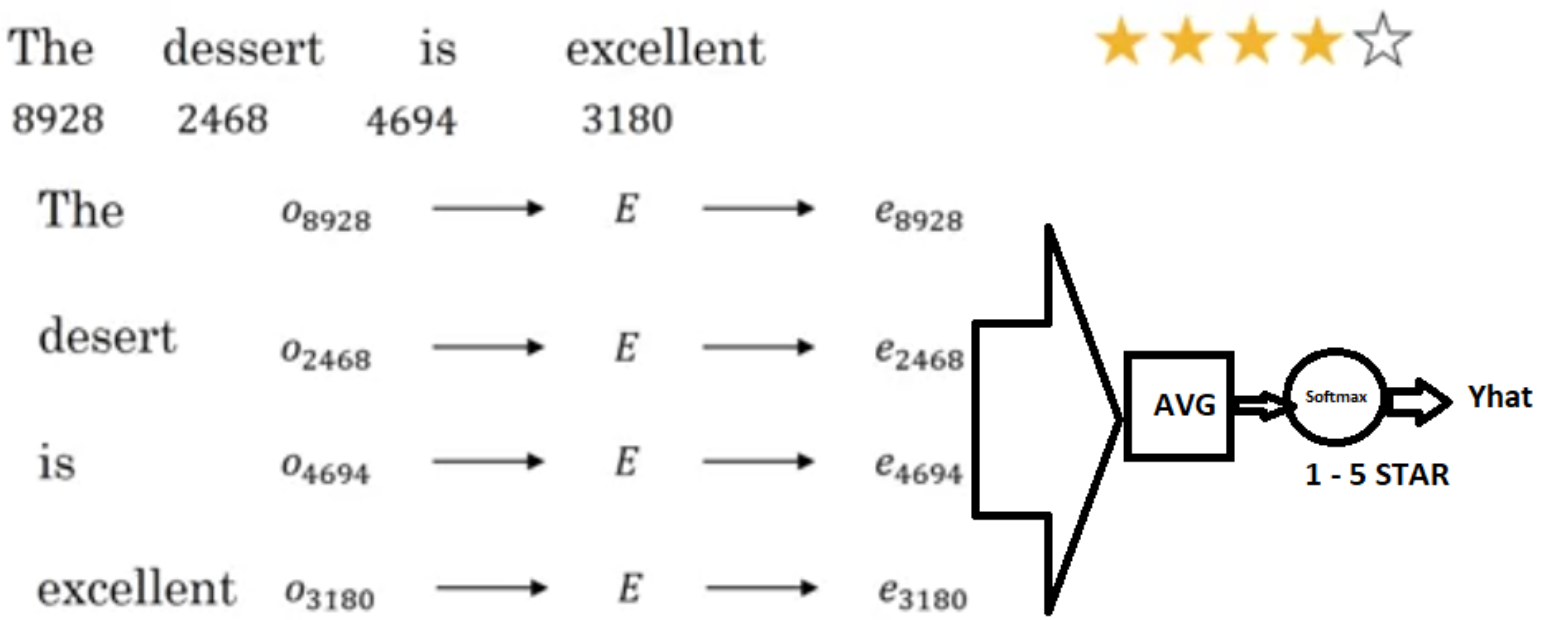
$$\sum_{i=1}^{Vocsize} \sum_{j=1}^{Vocsize} f(X_{ij})(\Theta_i^T e_j + b_i + b_j - log(X_{ij}))^2 \quad (31)$$

Where
X_{ij} represent how many number of time j^{th} target comes with i^{th} context
f(X_{ij})= if X_{ij}=0

Sentiment Classification

x	y
The dessert is excellent.	★★★★☆
Service was quite slow.	★★☆☆☆
Good for a quick meal, but nothing special.	★★★☆☆
Completely lacking in good taste, good service, and good ambience.	★☆☆☆☆

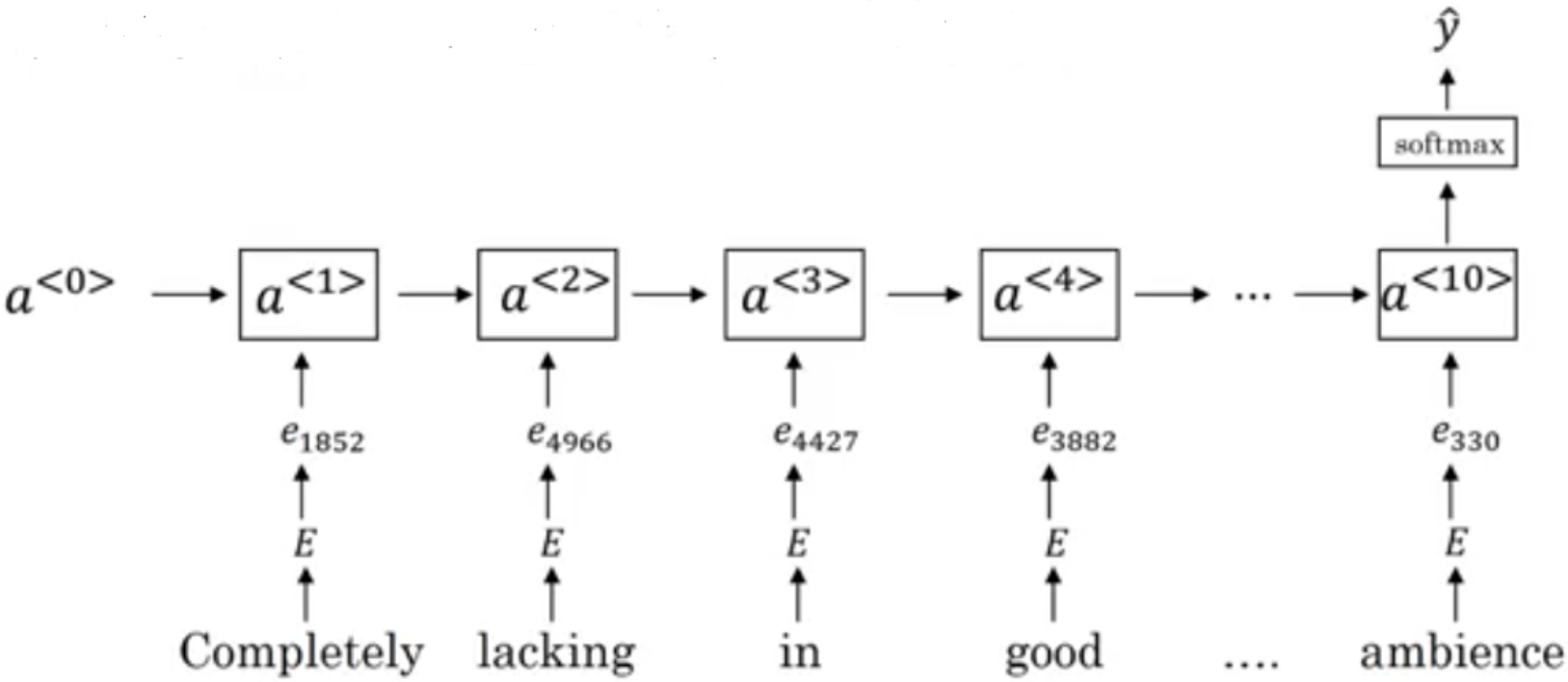
Sentiment Classification Model



Limitation:

Ram is not a good learner, good listener and good human

Sentiment Classification RNN model



Email : dr.vvr.research@gmail.com
(D

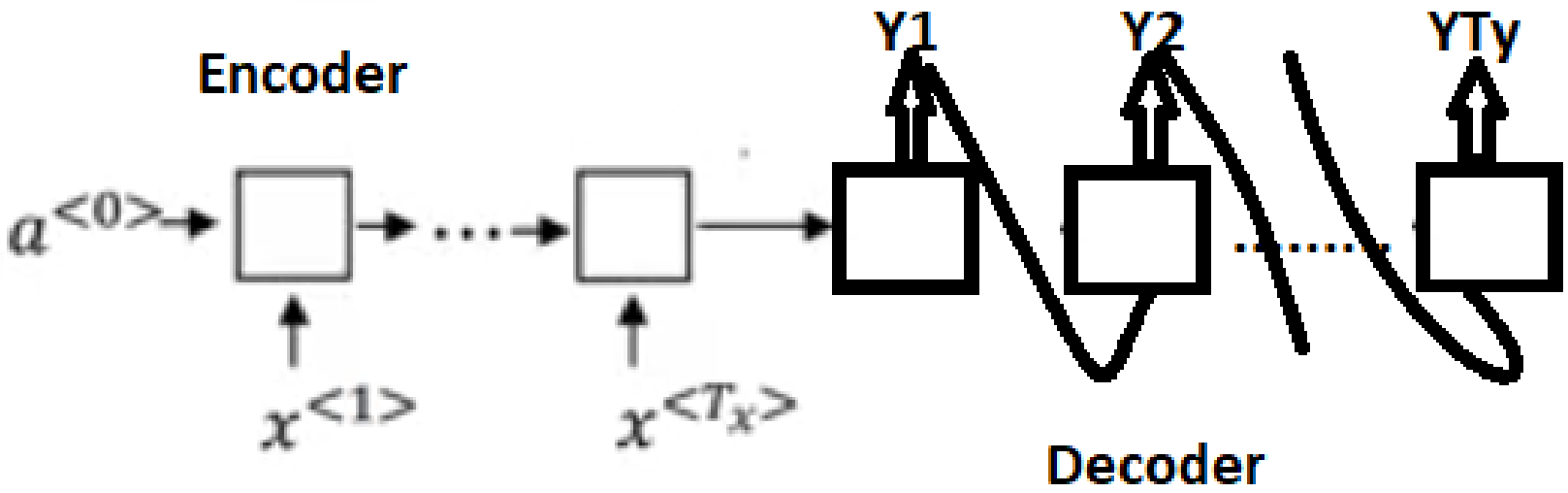
Sequence To Sequence Models: Machine Translation

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

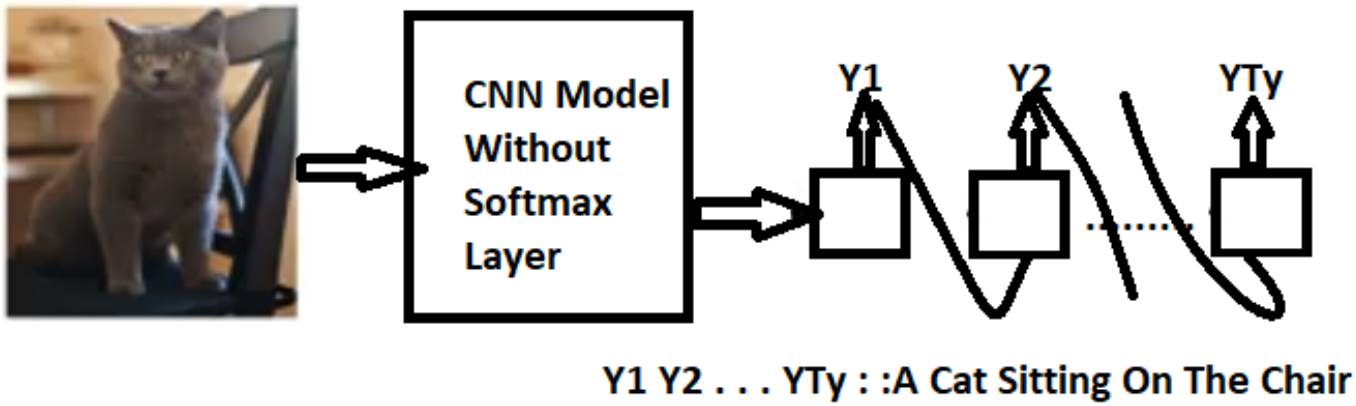
Jane visite l'Afrique en septembre

⟶ Jane is visiting Africa in September.

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$



Sequence To Sequence Models: Image Captioning



Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

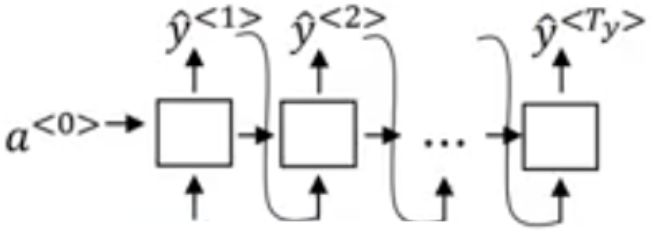
Machine Translations using Conditiona Load Modeling

Jane visite l'Afrique en septembre. $P(y^{<1>}, \ldots, y^{<T_y>} | x)$

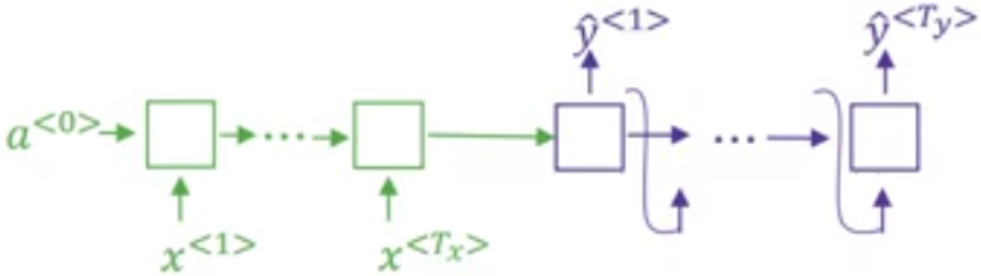
- → Jane is visiting Africa in September.
- → Jane is going to be visiting Africa in September.
- → In September, Jane will visit Africa.
- → Her African friend welcomed Jane in September.

Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)

Language model:



Machine translation:



Email : dr.vvr.research@gmail.com
(Dr. Venkataramana Veeramsetty)