Date

GRU

Gated Relu units, useful for solving the vanishing gradient Problem.

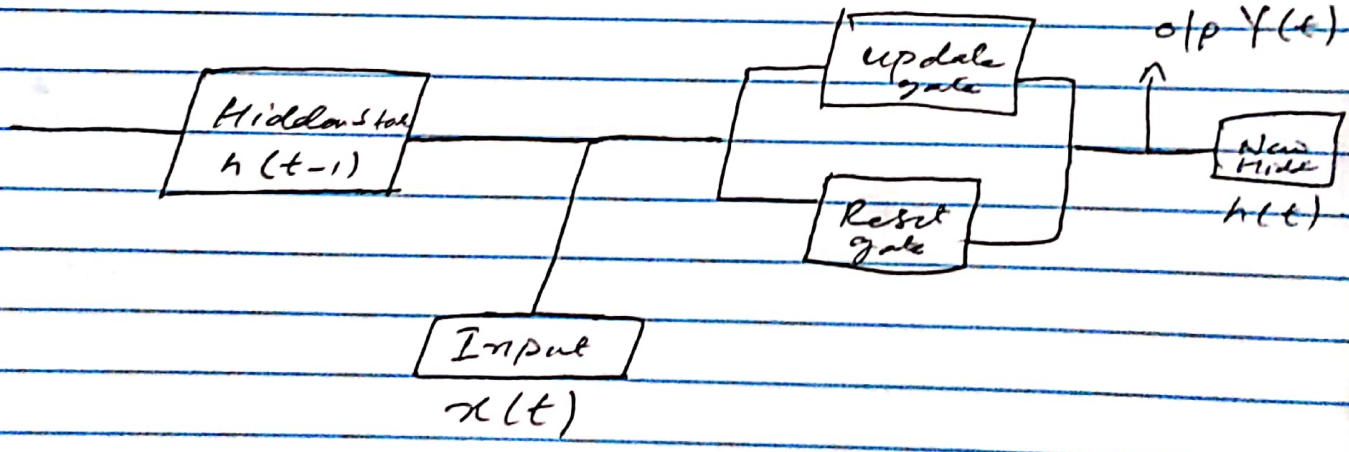
* Architecture Similar to LSTM, while there are some diffrences.

1- LSTM - 3 gates (forget, input and output gates)

2- GRU - 2 gates (update, reset)

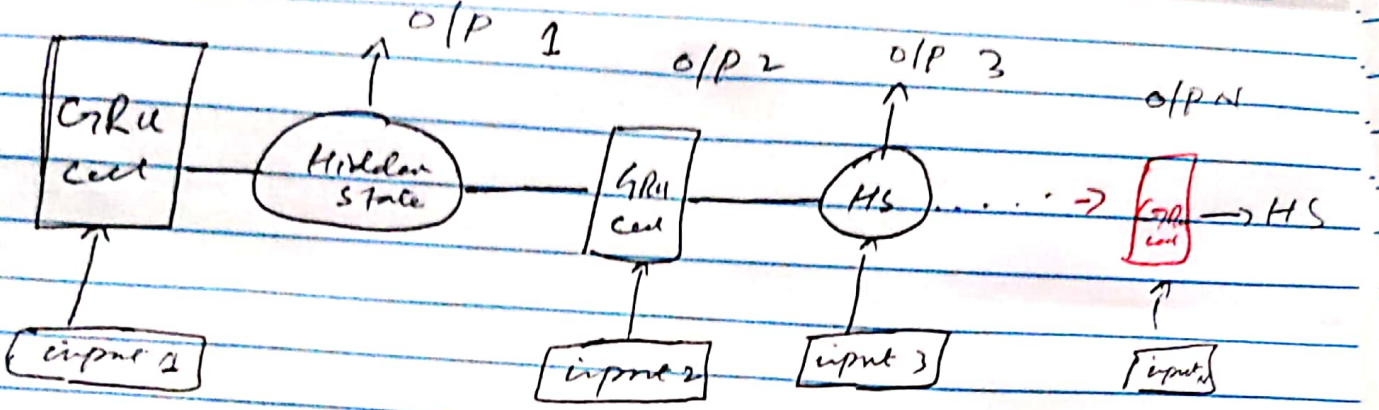
* GRU IS less complex, well suited for small datasets

* Gates Allow you to control the flow of information. How much information to be kept and how much information to forget. This helps in tracking the Long term dependency problem.



GRU Architecture.

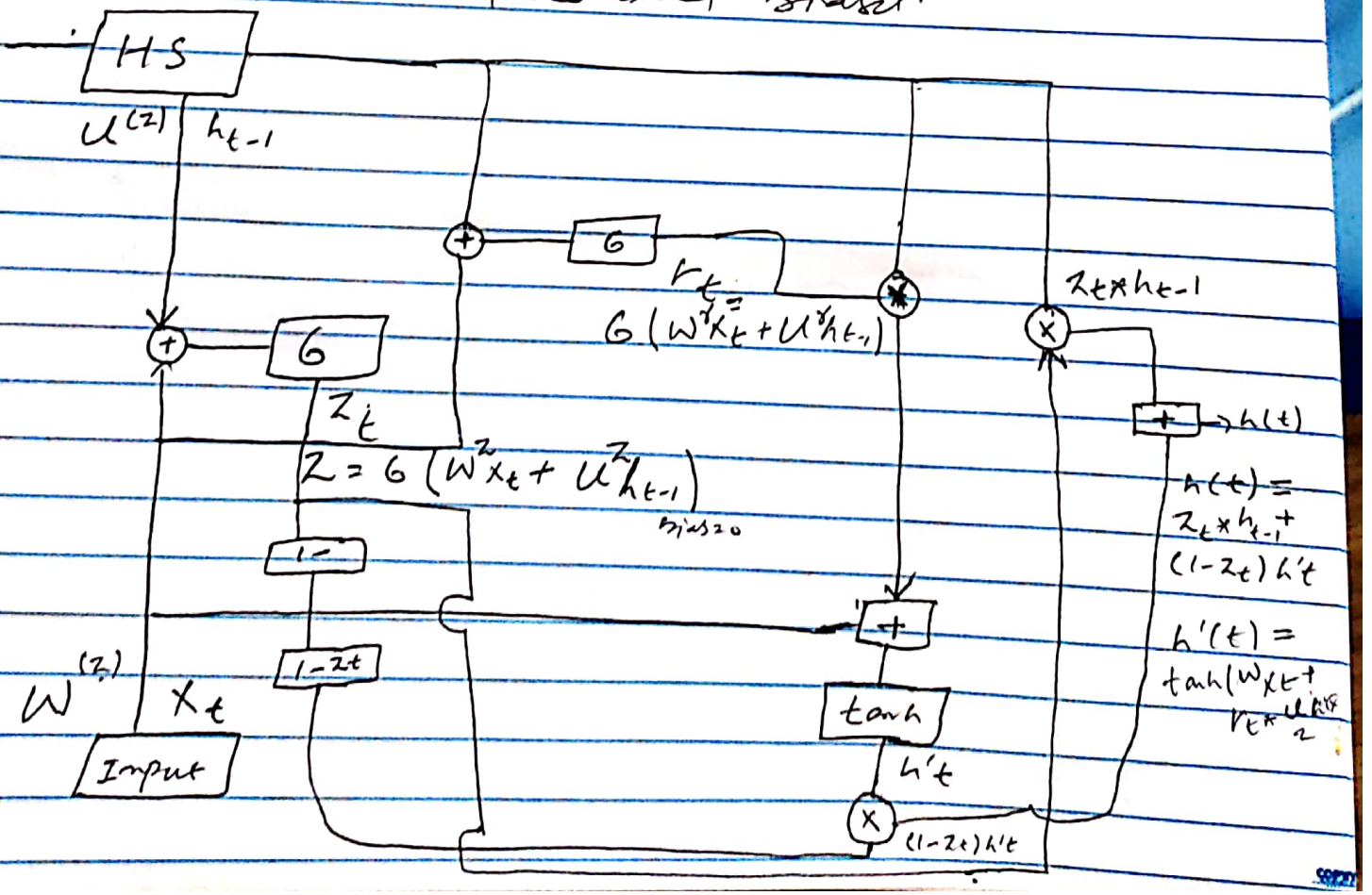
Scanned with CamScanner



Update Gate:- [How much to keep the information]
Reset Gate:- [How much to forget].

* weights are associated at every stage.

* Update gate :- z_t works on i/p $x(t)$, $h(t-1)$ hidden state and biases.



Scanned with CamScanner

$z_t = \sigma(W^z x_t + U^z h_{t-1})$ (update) bias = 0

$r_t = \sigma(W^r x_t + U^r h_{t-1})$ (Reset)

$h'_t = \tanh(W x_t + r_t * U h_{t-1})$ (Intermediate)

$h_t = (1 - z_t) h'_t + z_t * h_{t-1}$ (current unit)

A worked example

'text = MathMathMathMathMathMath'

Preprocessing steps:-

S-1 Convert text into numeric values and form a dictionary.

Dictionary: {'h': 0, 'a': 1, 't': 2, 'M': 3}

our encoded O/P

MathMath = [3, 1, 2, 0, 3, 1, 2, 0].

S-2 Create batches of data

Let's put the following settings

Batch Size (B) = 2

Sequence Size (S) = 3

Vocabulary (V) = 4

output (O) = 4

Scanned with CamScanner

Date

[3, 1, 2], [0, 3, 1], [2, 0, 3], [1, 2, 0], ...

Batch 1 Batch 2

Create Mini-batches of 3

[[[3, 1, 2
0, 3, 1]] [[2, 0, 3
1, 2, 0]] [[3, 1, 2
0, 1, 3]]]

Transpose Sequences

[[[3, 1, 2
0, 1, 3]] . [[2, 0, 3
1, 2, 0]] [[3, 1, 2
0, 1, 3]]] = [[[3] [1] [2]]
[0] [3] [1]
[2] [0] [3]]
[3] [1] [2]
[0] [3] [1]]

5- One hot encoding

[[[3] [1] [2]]
[0] [3] [1]
[2] [0] [3]]
[3] [1] [2]
[0] [3] [1]]
= [[0001] [0100] [0010]
[1000] [0001] [0100]
[0010] [1000] [0001]
[0100] [0010] [1000]
[0001] [0100] [0010]
[1000] [0010] [0100]]

x(t-1) x(t) x(t+1)

Scanned with CamScanner

Dimension of our dataset

Number of batches x Number of letters in each sequence x Size of batch x Vocabulary

$3 \times 3 \times 2 \times 4$ => Shape

Vocabulary = 4

Batch size

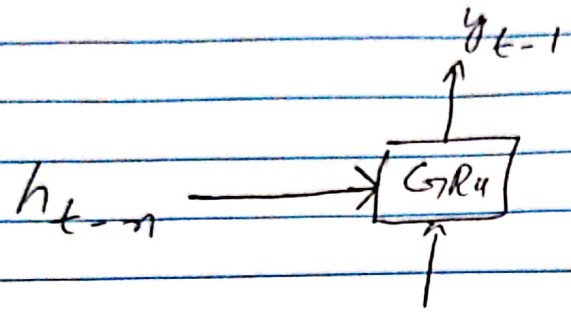
Batches

Sequence = 3

GRU GRU GRU

Scanned with CamScanner

Date



Batch 1 Sequence 1 $\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

S-2 Define weight Matrices :- W_{ax}

Hidden Size = 2, $\Rightarrow 4 \times 2$

Random initialization

Vocabulary = 4 $\begin{bmatrix} 0.6614 & 0.2669 \\ 0.0617 & 0.6213 \\ 0.4519 & -0.1661 \\ -1.5228 & 0.3817 \end{bmatrix}$

W_2

Scanned with CamScanner