Sequence Models

Examples of Sequence Data

1) Speech Recognition
2) music generation
3) Sentiment classification
4) DNA Sequence Analysis
5) Machine Translation
6) Video Activity Recognition
7) Named entity Recognition

Can be addressed as supervised learning

$X \rightarrow Y$

i/p   o/p

X and Y, can have same or diff length.

Notation

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<9>}$

x: Harry Potter and Hermione Granger invented new spell.

NER - People name, Company's Name, Time, Location, Currency Names, Country Names.

y : 1 1 0 1 1 0 0 0

$T_x = 9$ (length of i/p sequence, x)

→ Individual word representation ⟶ Vocabulary / Dictionary

One-hot vectors will be generated from the sentence.

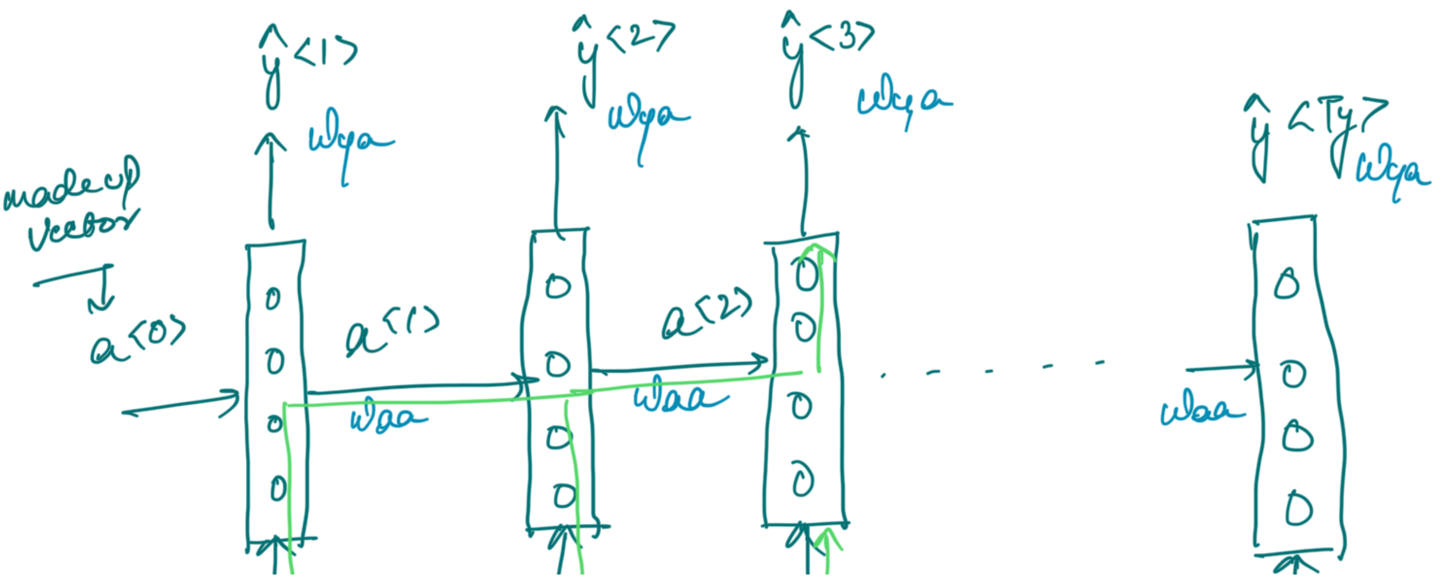
RNN

why we cannot use standard NNs:-

① I/p & o/p can be different lengths.

② Doesn't share features learned across different position of text.

⇒ What is RNN



$\boxed{x^{<1>}}$ W_{aa} (first word) $\boxed{x^{<2>}}$ W_{ax} (2nd word) $\boxed{x^{<3>}}$ W_{ax} . . . . . $x^{<T_x>}$ W_{ax} (9th word)

→ at every step activation will be passed on to the next layer.

→ RNN scans through the data from left to right.

→ Parameters at each time step are shared.

→ When predicting $\hat{y}^{<3>}$

$$\uparrow$$

$$x^{<3>} + x^{<2>} + x^{<1>}$$

→ Weakness : uses words before it, not after.

$\hat{y}^{<3>}$ will only use till $x^{<3>}$ not $x^{<4>}$ or so on.

Forward Propagation

usually tanh / ReLU

$$a^{<0>} = \vec{0}$$

$$a^{<1>} = g_1(W_{aa}\, a^{<0>} + W_{ax}\, x^{<1>} + b_a)$$

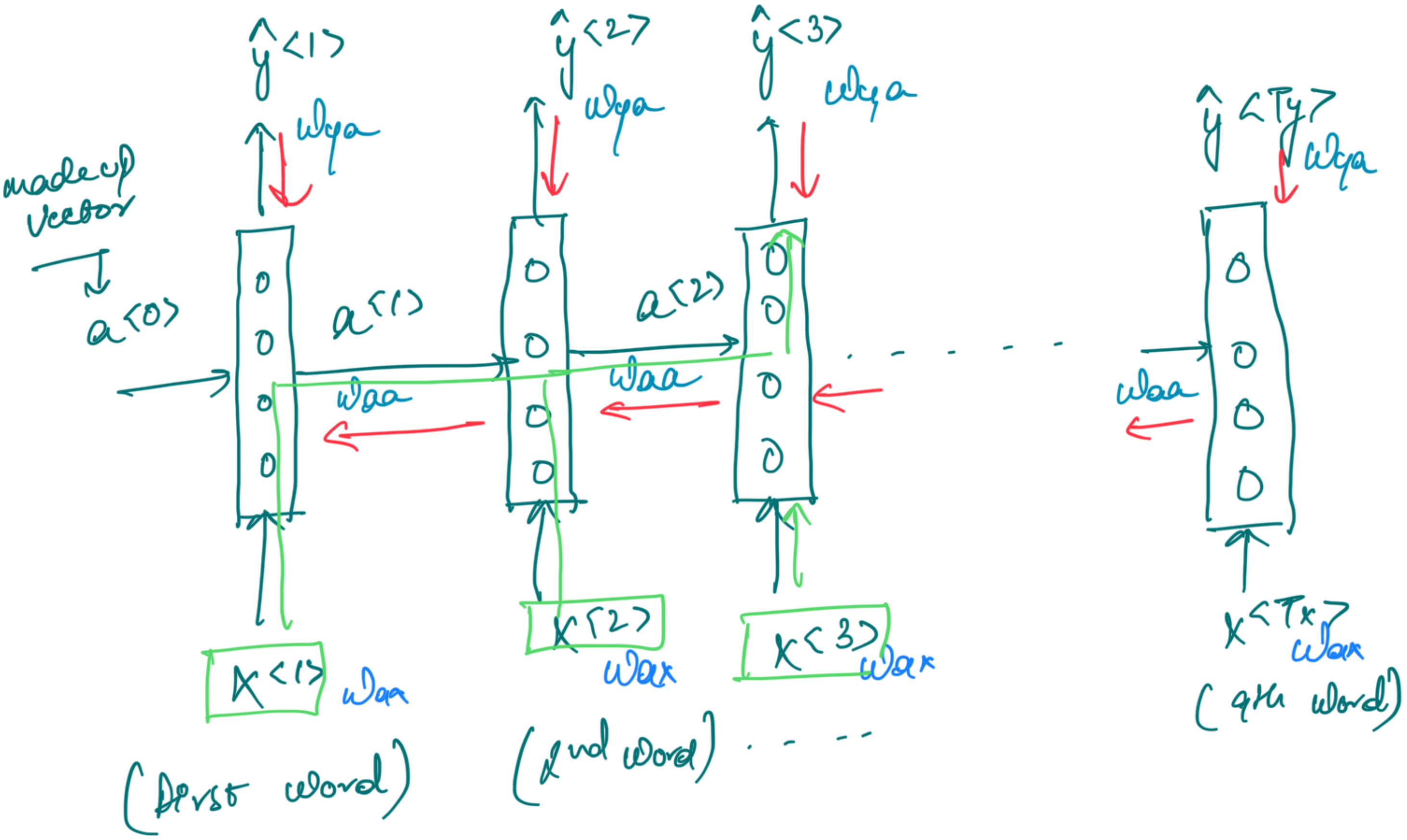
$$\hat{y}^{<1>} = g_2(W_{ya}\, a^{<1>} + b_y)$$

o/p sigmoid activation

$\underline{FP} \Rightarrow \begin{cases} a^{<t>} = g_1\left(W_{aa}\, a^{<t-1>} + W_{ax}\, X^{<t>} + b_a\right) \\ \hat{y}^{<t>} = g_2\left(W_{ya}\, a^{<t>} + b_y\right) \end{cases}$

$a^{<0>}, \; x^{<1>} \Longrightarrow a^{<1>} \rightarrow \hat{y}^{<1>} \; \ldots\ldots$

Backward Propagation :-



You need loss function for back Prop.

Standard logistic Regression loss

$$L^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1-y^{<t>}) \log(1-\hat{y}^{<t>})$$

for 1 time-step

$$L(\hat{y}, y) = \sum_{t=1}^{T_y} L^{<t>}(\hat{y}^{<t>}, y^{<t>}) \Rightarrow \text{loss of all}$$

Types of RNNs:

$T_x = T_y$

i/p o/p

→ many-to-many (NER)

→ many-to-one (Sentiment classification)

→ one-to-many (music generation)

→ many-to many ($T_x \neq T_y$) like machine Translation.
also called as <u>encoder - decoder</u>.

<u>Language Model and Sequence Generation.</u>

→ Most basic and imp task in NLP → Language Modelling.

<u>What is Language Model</u> :-

Speech recognition example

The apple and pair salad. ①

✓ The apple and pear salad ②

... using Lang Model

way ② is picked ⇒ by using

tells what is prob of each sentences.

① ⟶ 3.2×10^{-13}

② ⟶ 5.7×10^{-10} ✓

A Language Model's job

→ What is the probability of the sentence.

$$P(\text{Sentence}) = ?$$

$$P(y^{<1>}, y^{<2>}, \ldots\ldots y^{<T_y>})$$

→ How to build LM using RNN :-

Training set : large corpus of english text.

Step 1 :- Tokenize (Vocab)

Step 2 :- One-hot vectors / indices, Also,

add extra token <EOS> : end of sentence.

<UNK> : for unknown token

eg <u>sentence</u>:- Cats average 15 hours of sleep a day. <EOS>

$P(a)\ P(aa)\ P(aaron) \ldots P(cats)$ $\quad$ $P(average \mid cats) \ldots$

$\hat{y}^{<1>}$ $\quad$ $\hat{y}^{<2>}$

$P(\langle EOS\rangle \mid \ldots\ldots)$

$y^{<1>}\ y^{<2>}\ y^{<3>}$: 3 word sentence.

$$= P(y^{<1>}) * P(y^{<2>} \mid y^{<1>}) * P(y^{<3>} \mid y^{<1>}, y^{<2>})$$

Probability of 3 word sentences.