RNN : Unidirectional, weakness, does not take account the words after current timestamp.

$$a^{<t>} = g(W_{aa} a^{<t-1>} + W_{ax} x^{<t>} + b_a)$$
$$\hat{y}^{<t>} = g(W_{ya} a^{<t>} + b_y)$$

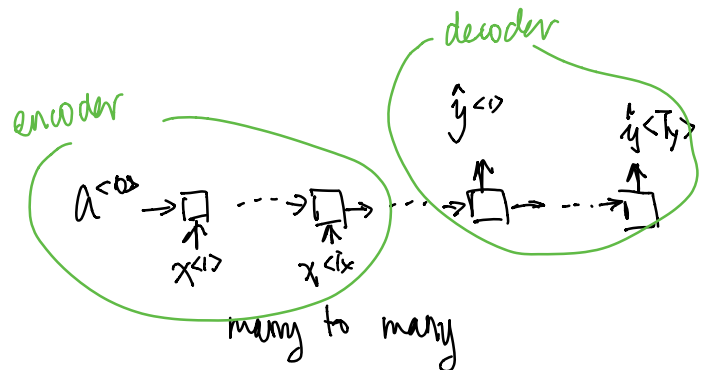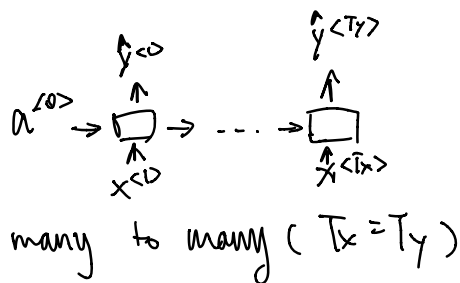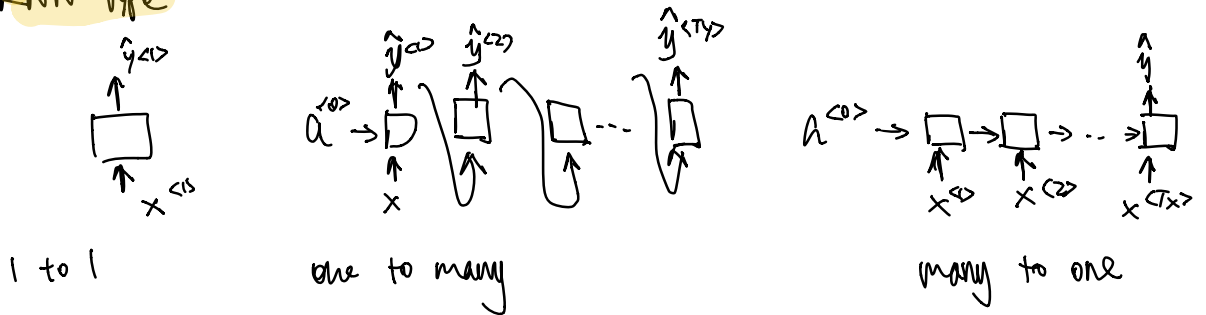usually stacked as $\underline{W_a}$

$\rightarrow a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}]$

loss

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

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_x} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

Basically sum of $\mathcal{L}^{<t>}$ at all t.

RNN Type -

1 to 1

one to many

many to one

many to many ($T_x = T_y$)

encoder

decoder

many to many

Language Model

→ fundamentally $P(sentense)$ $P(\_\_|cats\ average)$

$$a^{<0>}=0 \rightarrow \boxed{a^{<1>}} \rightarrow \boxed{a^{<2>}} \rightarrow \boxed{a^{<3>}}$$

with outputs $\hat{y}^{<1>}$, $\hat{y}^{<2>}$ $P(\_\_|cats)$, $\hat{y}^{<3>}$

$x^{<1>}=\vec{0}$ $x^{<2>}=y^{<1>}$ $x^{<3>}=y^{<2>}$
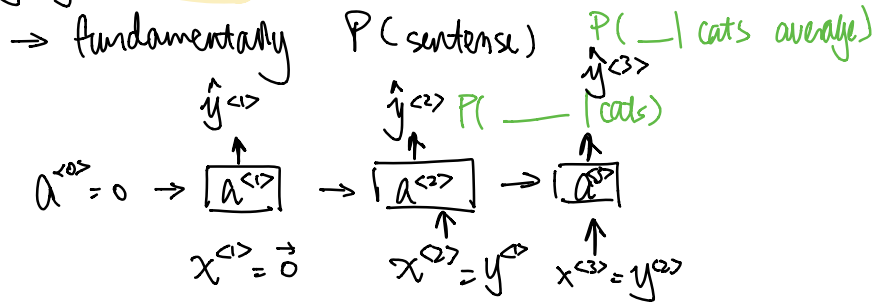
cats averages 15 hours of sleep per day. <EOS>
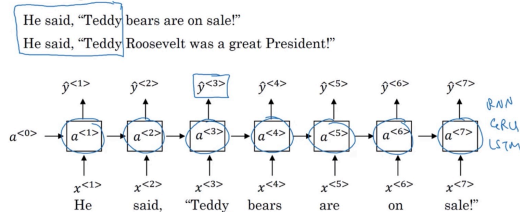
Sample Novel Sequence.

Basically feed in a random word, propogate until <EOS>

Char level RNN : Do not need to worry about < unknown word>
harder to train and capture long-term dependency.

Bidirectional RNN

Bidirectional RNN (BRNN)

Getting information from the future

He said, "Teddy bears are on sale!"
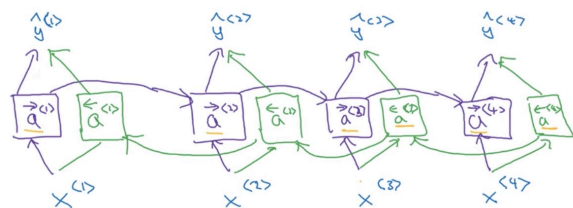He said, "Teddy Roosevelt was a great President!"

Problem

Acyclic graph

BRNN with LSTM commonly
used for NLP Problem

Deep RNN. usually not that deep due
to horizontal and vertical connection, it
is common for sequence model followed by normal NN.

normal NN

$$a^{[2]<3>} = g\left(W_a^{[2]}\left[a^{[2]<2>}, a^{[1]<3>}\right] + b_a^{[2]}\right)$$

Deep RNN