Week 3.

Various Sequence to Sequence Architectures

Video 1: Basic Models

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

Jane visite l'Afrique en septembre

Jane is visiting Africa in September

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

Many to many models are used in machine translation & speech recognition

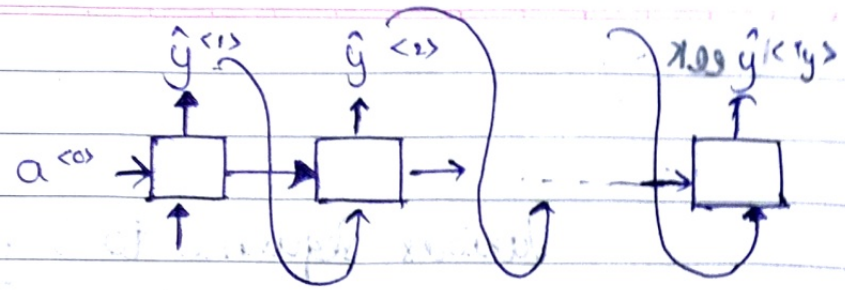
encoder - RNN/LSTM/GRU

~~inputs~~ ~~the~~ takes as input the whole sentence and outputs a vector that should represent the whole input.

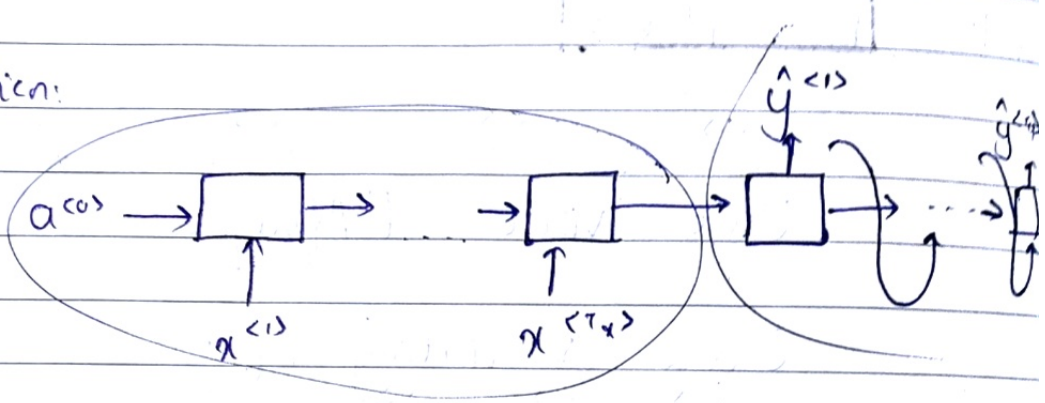
decoder - takes in sequence built by encoder and outputs new sequence

Video 2: Picking the most likely sentence

Language Model:



Machine Translation:



language model helps us estimate the probability of a sentence & generate sentences too

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

Machine Translation:

$$P(y^{<1>}, \ldots y^{<T_y>} \mid x^{<1>} \ldots, x^{<T_x>})$$

Example:

X: "Jane visite l'Afrique en Septembre"

Y maybe:

- Jane is visiting Africa in September
- Jane is going to visit Africa in September
- In September Jane will visit Africa.

So for best output,

$$\underset{y^{<1>} \ldots y^{<T_y>}}{\text{argmax}} \left(P y^{<1>}, \ldots, y^{<T_y>} \mid x\right)$$

Greedy Search

first word - most likely first for word according to conditional language model.

↓

second most likely

↓

third most likely

X working

example "Jane is" most likely beginning
after is going is most common word
∴ "Jane is going to be visiting Africa in September". P(Jane is going ...|x) > P(Jane is visit...|x)

~~Jane is~~ In September, Jane will visit Africa.

Her African friend welcomed Jane in September

Video 3: **Beam Search**

B = 3
↑
beam width

input french sentence → encoder → decoder
↓
retain top 3 possibilities ← softmax output (10000 possibilities)

In
Jane
September } most likely beginning words

Then for each of the above words the most likely 2nd word will be generated according to B=3 : for 'In' there will 3 most likely upcoming words & so on. for the rest.

Out of these again top 3 best possibilities are selected.

& so on

Video 4: **Refinements in Beam Search**

length normalisation

$$\arg\max \prod_{t=1}^{T_y} P(y^{<t>} | x, y^{<1>}, \ldots, y^{<t-1>})$$

i.e $P(y^{<1>} \ldots y^{<y>} | x)$

$= P(y^{<1>} | x) \cdot P(y^{<2>} | x, y^{<1>}) \ldots$

Probabilities are too small

∴ we use log

$$\arg\max \sum_{y=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \ldots, y^{<t-1>})$$

$$\frac{1}{T_y^{\alpha}} \sum_{t=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \ldots, y^{<t-1>})$$

$\alpha = 1$ (full normalisation)
$\alpha = 0$ (no normalisation)
$\alpha = 0.7$ ✓

B ↑, time ↑

Videos: **Error Analysis in Beam Search**

To segregate if error is due to B / RNN

x = 'Jane visite L'Afrique en Septembre'
y^* = "Jane visits Africa in September" ✓
$\hat{y}$ = "Jane visited Africa last September" ✓

Case ①

$P(y^*|x) > P(\hat{y}|x)$

beam search chose $\hat{y}$ although y^* has higher $P(y|x)$

beam search is at fault.

Case ②

$P(y^*|x) \leq P(\hat{y}|x)$

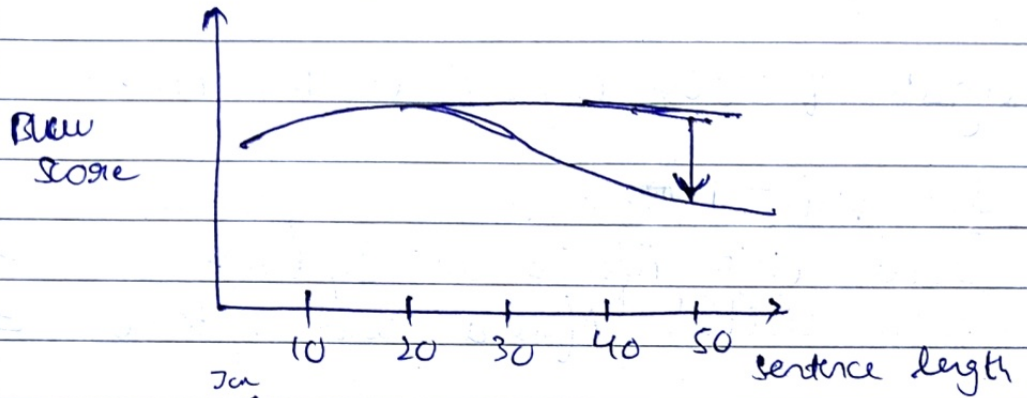
RNN predicted $P(\hat{y}|x) > P(y^*|x)$

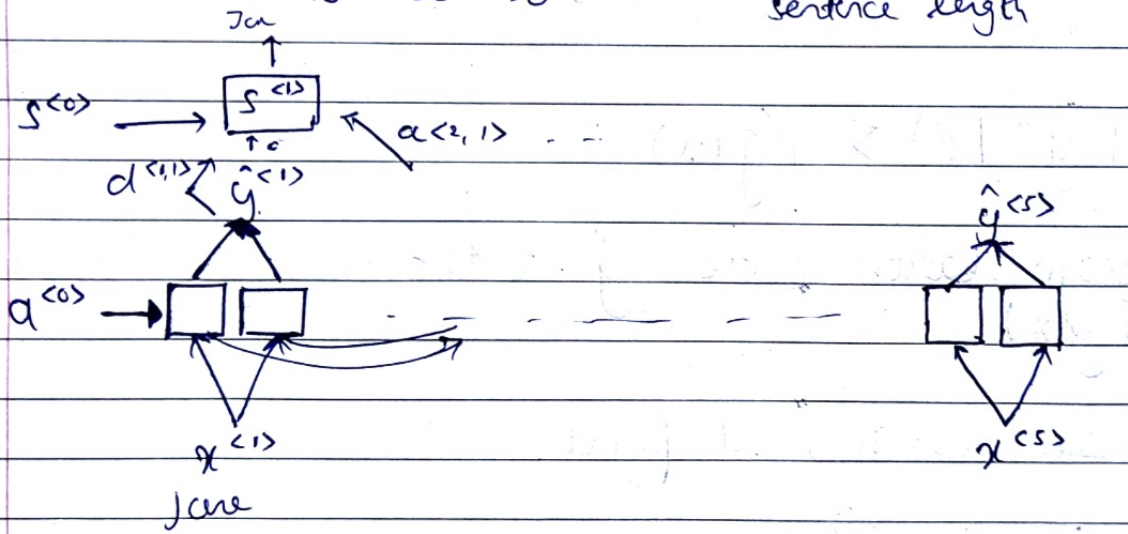
RNN - faulty

Video 6. **Attention Model Intuition**

Given a long sentence and the job of translation ~~the~~ humans would it memorise the sentence and translate it but they would translate in parts.

Model's performance ↓ if sentence is long

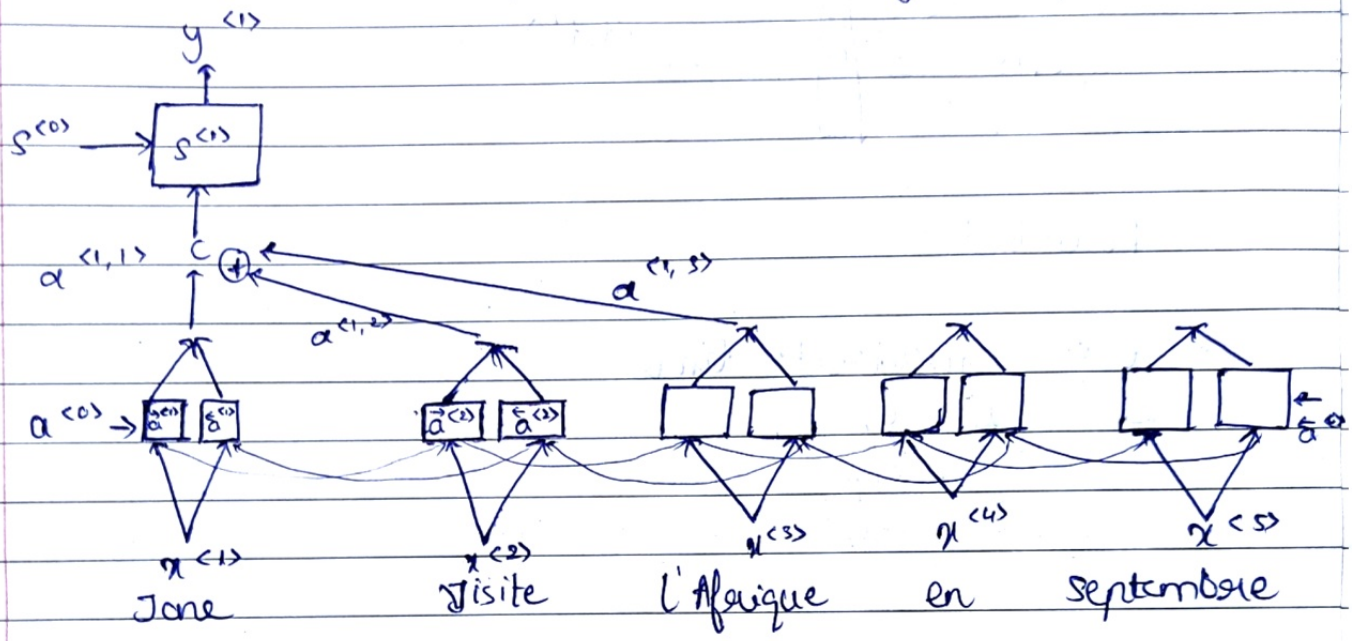




encoder is bidirectional RNN
it generates a vector representing inputs.

Video 7: **Attention Model**

* input sentence is processed by bidirectional RNN/GRU/LSTM to compute features on every word.



* α tells us how much the context would depend on the features we are getting on the activations we are getting from different time steps.

$$a^{<t'>} = [\vec{a}^{<t'>}, \overleftarrow{a}^{<t'>}]$$

$$\sum_{t'} \alpha^{<1,t'>} = 1$$

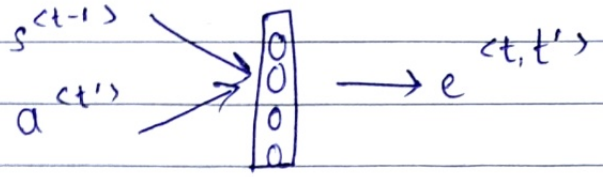
$$c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$$

(α: weights; a: activation)

$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

$$\alpha^{\langle t,t' \rangle} = \frac{\exp(e^{\langle t,t' \rangle})}{\sum_{t'=1}^{T_x} \exp(e^{\langle t,t' \rangle})}$$

$$= 1$$



Drawback:

- takes quadratic time

eg. T_x - i/p

T_y - o/p

time = $T_x \cdot T_y$

Speech Recognition - Audio Data

Video Speech Recognition