Transformer

Google brain => attention is all you need

{RNN, LSTM, GRU}

Processing the sequence

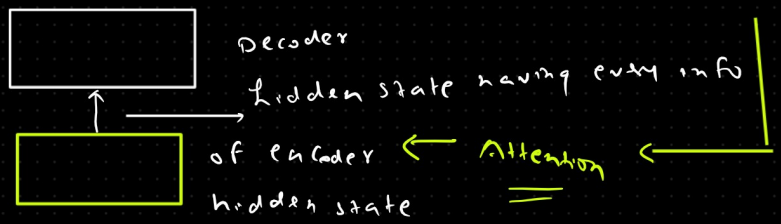
RNN => LSTM => GRU => {Sequence to sequence mapping}

2014

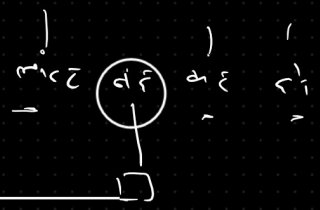
(encoder and decoder)

(many to many seq mapping)

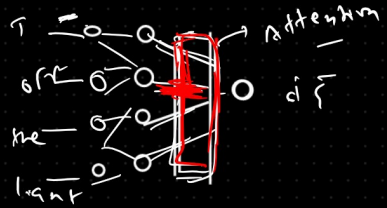
Machine translation



Attention



Neural netw



Turn

off

the

lights

Transformer => RNN, LSTM, GRU

encoder & decoder

embedding

Sentence -> {Self attention} -> key, value, query => NN => Softmax

feed forward

{multi headed}

Sequential ⇒ RNN, LSTM, GRU

timestamp

Transformer ⇒ Parallel fashion ⇒ it was able to sustain longer sentences

Positional encoding ⇒ Fast Processing

Datatime ⇒ sequential ⇒ Process ⇒ transformer

Encoder Decoder ⇒ this was the original

Encoder & decoder

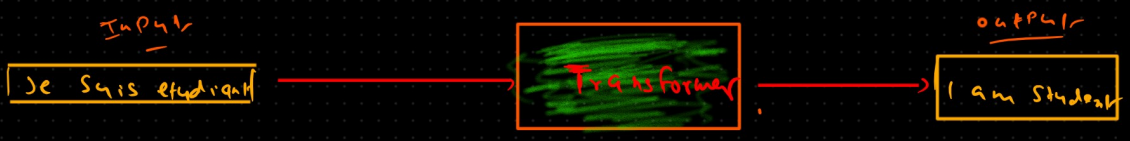
Encoder - Decoder

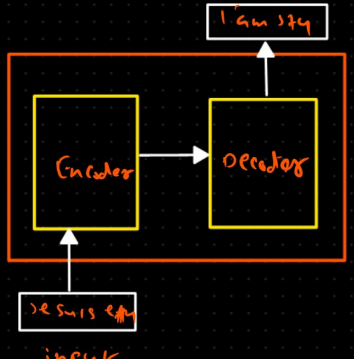
stack stack

1. Architecture
2. Attention
3. multihead attention
4. Embedding
5. Positional encoding
6. Softmax
7. {training / evaluation / conclusion}

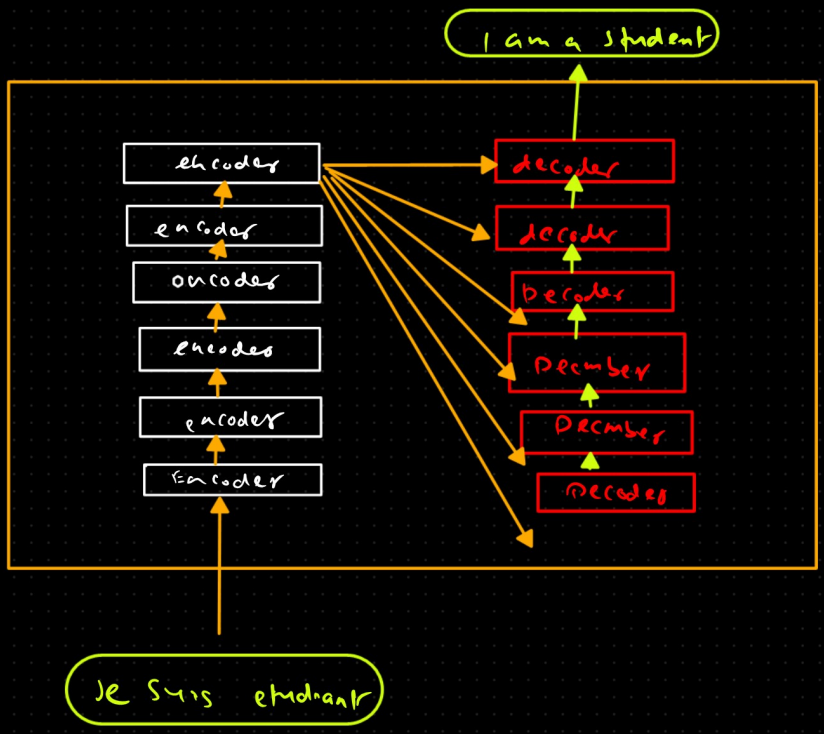
GPU ⇒ {Some interview ⇒ Train → transform}

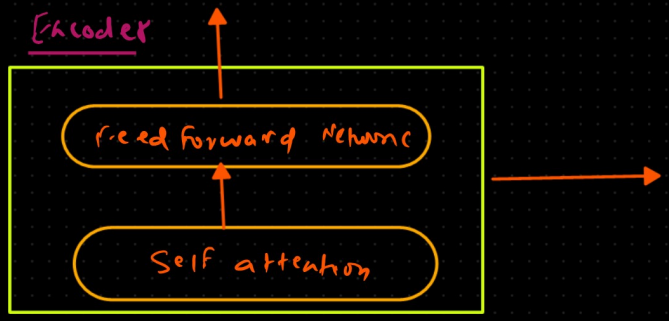
Machine translation

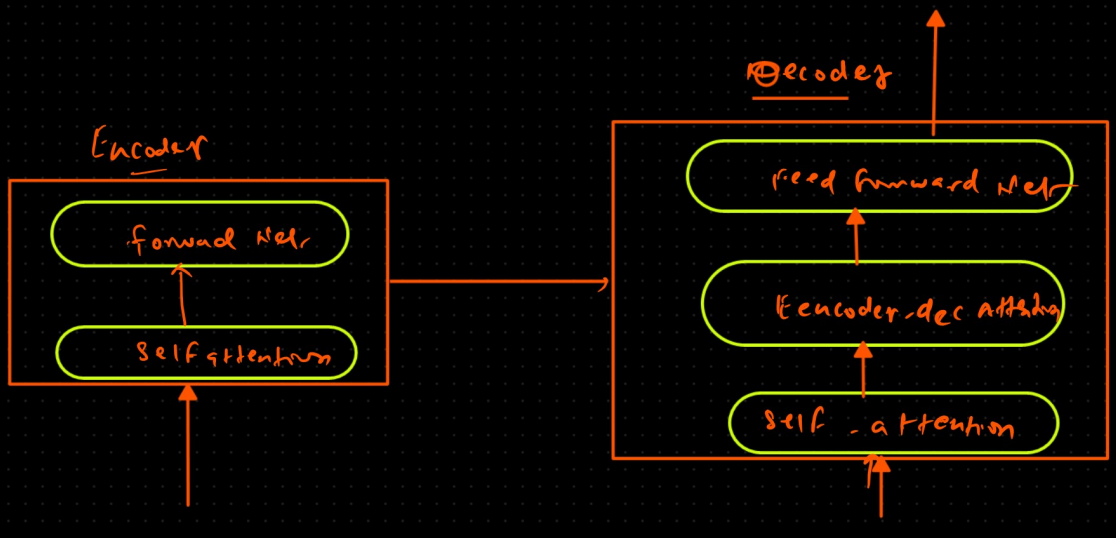




Research Paper







Sentence

French sentence

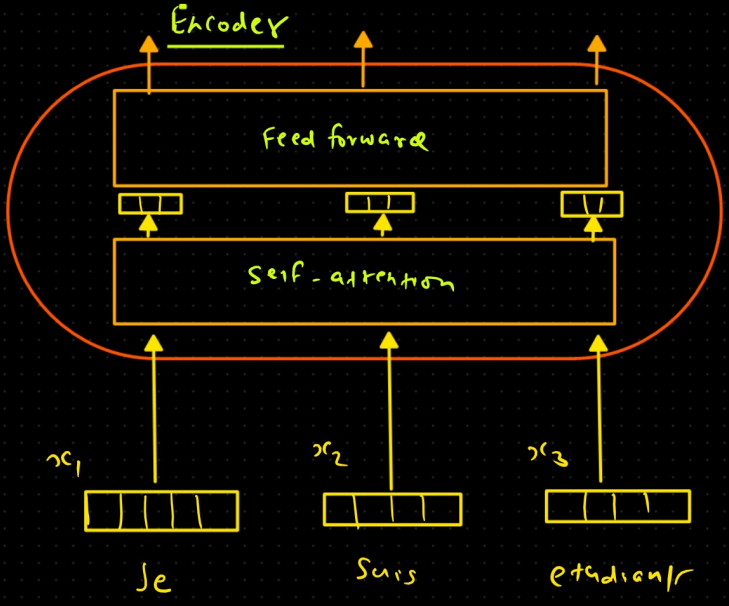
Je Suis étudiant

embedding

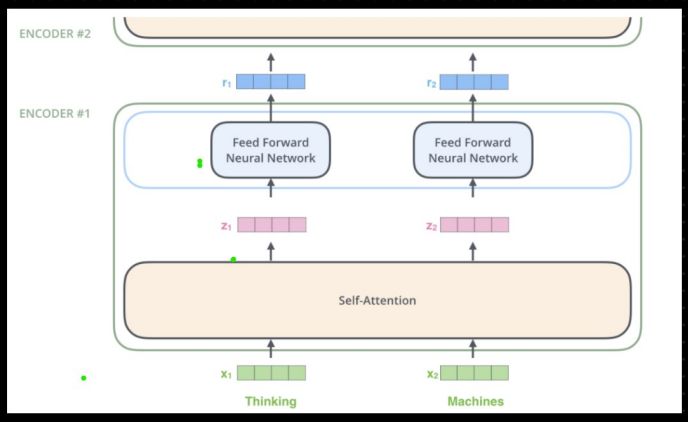
Vectors

x_1 x_2 x_3

Supervised



Self attention



Attention

one language — translation → another language

turn off the light

মাইট বাতি অফ কৰা

Connection

Input → output

মাইট বাতি অফ কৰা

decoder hidden state

Attention layer

encoder hidden state

turn off the light

Self-attention

Reference => The animal didn't cross the street because it was too tired

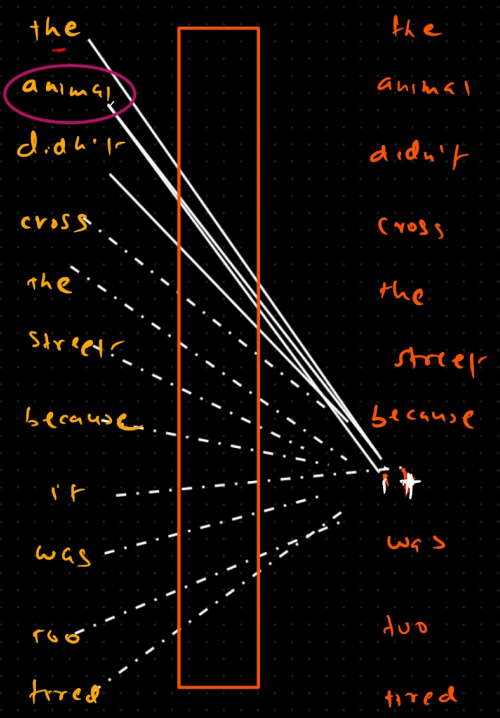
attention

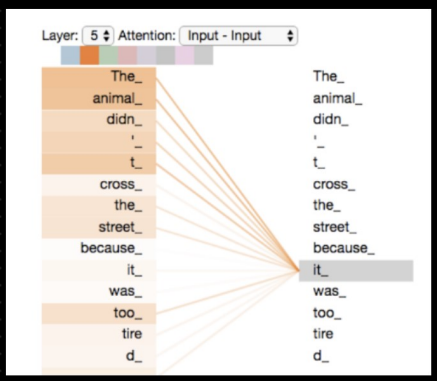
Sequence

(RNN, LSTM

Processing

t=1, t-2, t-3 ...





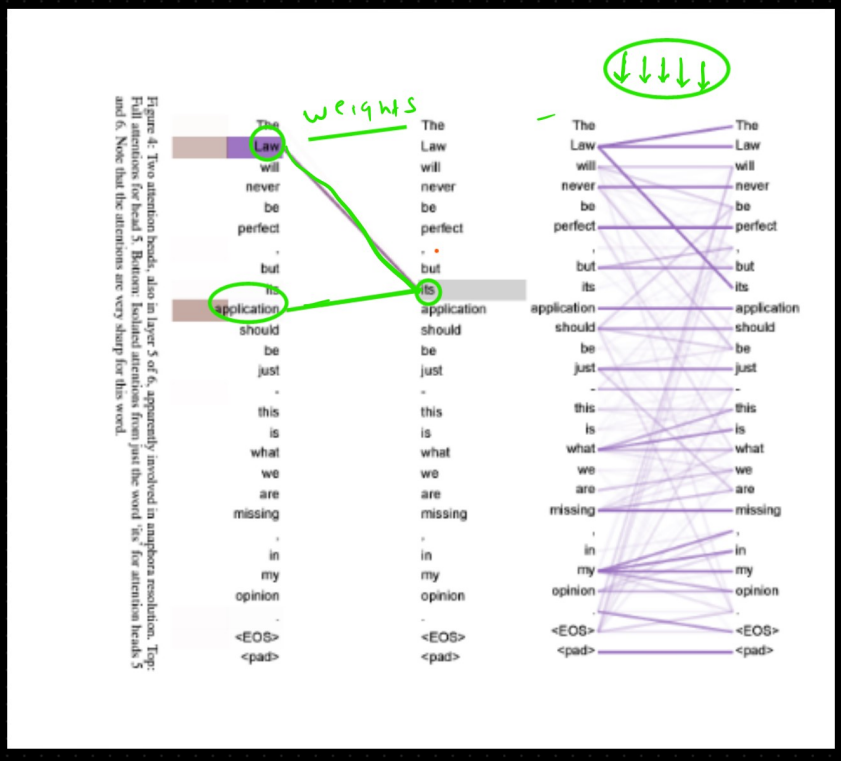
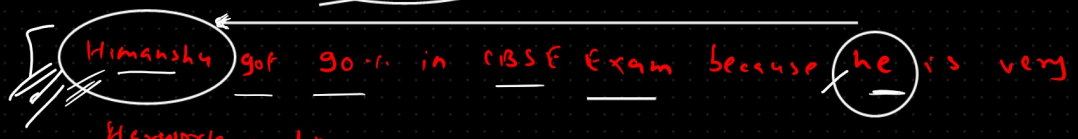


Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

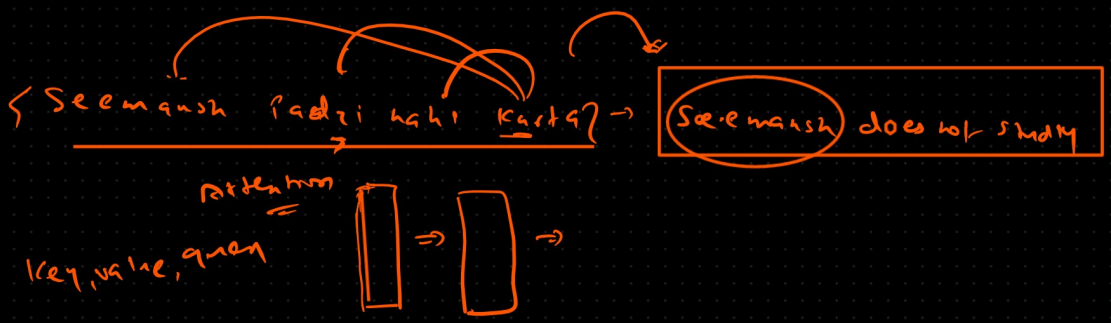
{ Key , query , value } → Preserving the context

Self attention ⇒ decoding the sequence

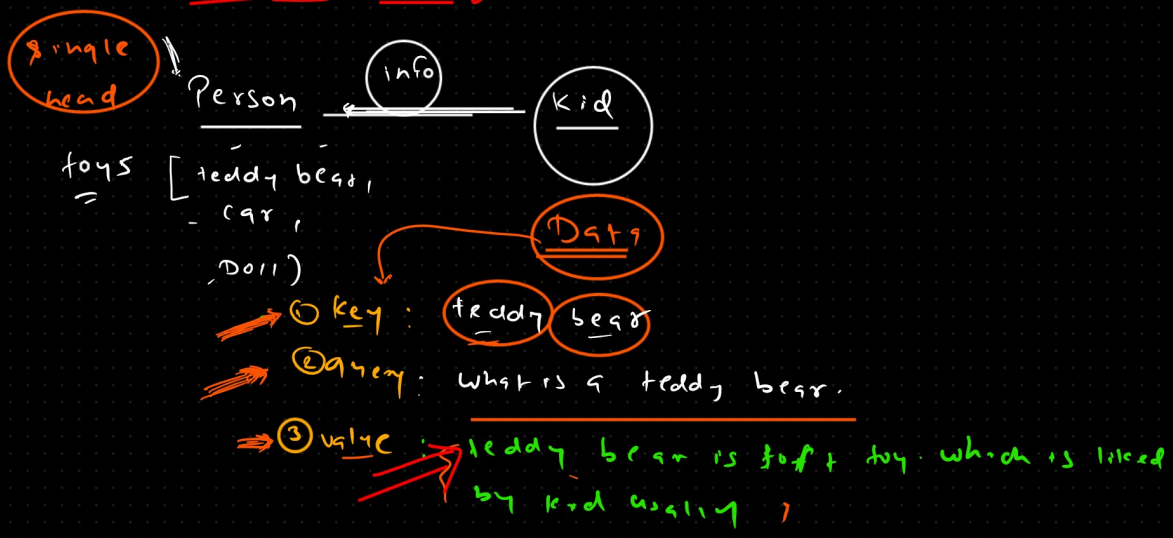


RNN / LSTM

RNN



Key value query



Single Person one head } Attention

Multiple Person multihead } multihead

Attention all you need

- self attention
- self attention

(Single-head attention)

{key, query, value} ⇒ pair

multihead-attention

{key, query, value} ⇒ (8)

$v_1, v_2, v_3, v_4, v_5 \ldots v_8$

vector

Softmax

Normalize

(5 × 5)

output

Sunday

Wednesday ⇒ trans, Bert, gpt

Saturday ⇒ code, Language, trans.

Sunday ⇒ text classification bert

BERT encoder

gpt

Decoder

13th Anglythey

13th Eng-Data

= (9) - (8)

1. Postion encoding
2. self attention
3. multihead
4. key, value, query

⇒ revise ⇒ Bert, gpt