Attention Mechanism

as the cat that was sitting on the mat, which was next to the dog, jumped

LLM should understand to pay most ATTENTION to jumped (meaningful word)

① 4 types of attention mechanisms

Simplified self-attention
- A simplified self attention technique to introduce broad idea

Self attention
- Trainable weights, that form the basis of the mechanism used in LLMs

Causal Attention
- A type of self attention used in LLMs that allows model to only consider previous and current inputs

Multi-headed Attention
- An extension of self attention and causal attention that enables the model to simultaneously attend to information from different representation subspaces
- LLM attends to various input data in parallel)

② Problem w/ modeling long sequences

Q) What is problem w/ architecture w/o attention mechanism which came b4 LLMs?

* Let's consider a language translation model (German)

kannst du mir helfen diesen Satz zu übersetzen
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
can you me help this sentence to translate

word-by-word translation does not work!

Should be, can you help me to translate this sentence

* translation process requires contextual understanding and grammar alignment.

b) To address this issue that we cannot translate text word by word, it is common to use a NN w/ two submodules

German → Encoder → Decoder → English

Encoder: reads and processes text

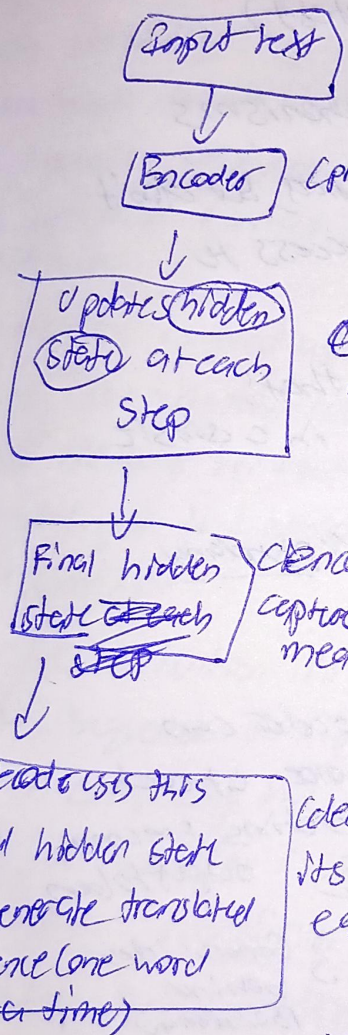
Decoder: translates text

A) Encoder processes each item in input, it compresses the info it captures into a vector (context). After processing the entire input sequence, encoder sends the context over to decoder, which begins producing the output sequence item by item

C) Before transformers, Recurrent Neural Networks (RNNs) were the most popular encoder-decoder architecture for language translation

D) RNN: output from previous step is fed as input to current step

E) Here's how encoder-decoder RNN works



(S) Encoder processes the entire input text into a hidden state (memory cell). Decoder takes this hidden state to produce an output
↳ Think of this as an embedding vector

* BOTTLENECK
- RNN can't directly access earlier hidden states from encoder, during the decoding phase
↓
Relies on solely current hidden state
↓
Leads to a loss of context, especially in complex sentences where dependencies might span long distances

Encoder compresses entire input sequence into a single hidden state vector.
↳ If the sentence is very long, it becomes very difficult for RNN to capture all information in a single vector

Input: "The cat that was sitting on the mat, which was next to the dog, jumped"

↓

"French translation"

Here, the action 'jumped' depends on the subject 'cat', but also on understanding the longer dependencies ("that was sitting on the mat, next to the dog")

↳ The RNN decoder might struggle w/ this class of context)

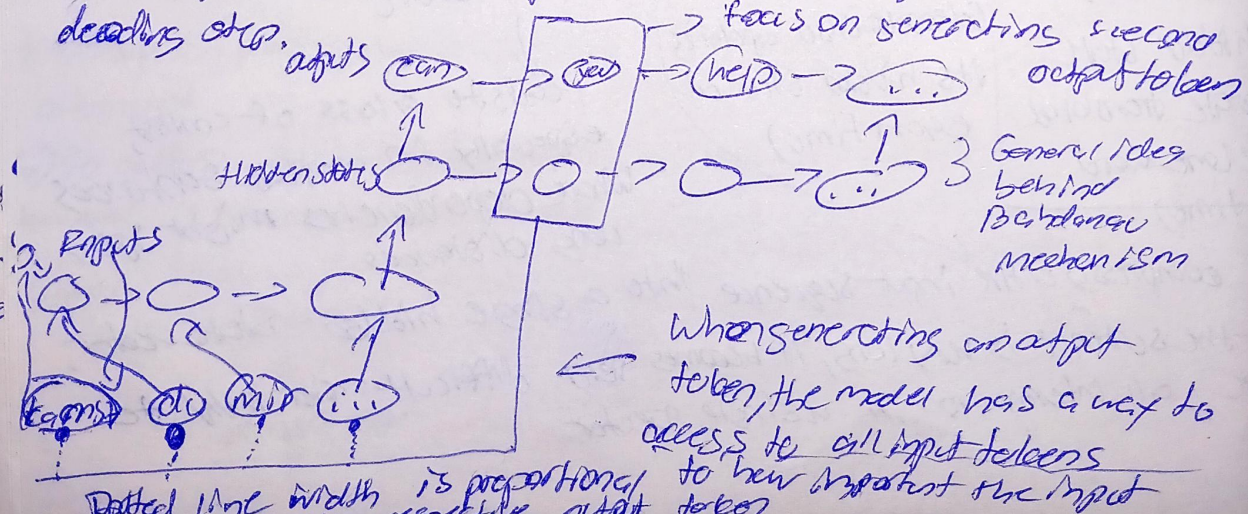
③ Capturing data dependencies w/ attention mechanisms

(a) RNNs work fine for translating short sentences, but don't work for long texts as they don't have direct access to previous words in the input.

(b) One major shortcoming in this approach is that: RNN must remember the entire encoded input in a single hidden state before passing it to the decoder

(c) In 2014, researchers developed the so called "Bahdanau attention mechanism for RNN's"

↳ Modifies encoder-decoder RNN such that the decoder can selectively access different parts of the input sequence at each decoding step.



When generating an output token, the model has a way to access to all input tokens to how important the input token

Dotted line width is proportional to the respective output token

- using an attention mechanism, the text generating decoder part of the network can access all input tokens selectively.
- This means that some input tokens are more important than others for generating a given output token.
- This importance is determined by the so called attention weights

(d) 2017, researchers found that RNN architectures are not required for building deep neural networks for NLP and proposed the original (transformer architecture) w/ a self attention mechanism inspired by the Bahdanau attention mechanism

At each decoding step the model can look back at the entire input sequence and decide which parts are most relevant to generate current word

When decoder is predicting "sate," the attention mechanism allows decoder to focus on part of input that corresponds to "jumped"

"The cat that was sitting on the mat, which was next to the dog, jumped"

"Le chat qui était assis sur le tapis, qui était à côté du chien, a sauté"

Dynamic focus on different parts of input sequence allows models to learn long range dependencies more effectively ☆

(e) Self attention is a mechanism that allows each position of the input sequence to attend to all positions in the same sequence when computing the representation of a sequence.

(f) Self attention is a key component of contemporary LLMs based on the transformer architecture, such as the GPT series

① Attending to different parts of the input with self attention.

(a) In "self attention", the "self" refers to the mechanisms ability to compute attention weights by relating different positions in a single input sequence.

(b) It learns the relationship between various parts of the input itself such as words in a sentence

(c) This is in contrast to traditional attention mechanisms where the focus is on relationships between elements of 2 different sequences.