Lecture 35: ~~Intro~~ Transformers & GPT

... review of previous lecture

-- attention (solve the problem of long sequence - which information is present important)

-- Token

--

Attention based Model

↓ Deep Neural Network

@ Transformers (∵ inspired by autoencoders

we only decoder → GPT

use only encoder → BERT family of models

Transformers

generative model that learn to understand and generate human like text.

⇒ when some sentences are translated from one to another language - the sequence length is different

⇒ Transformers were developed to solve sequence transduction or neural machine tasks.

• They transform an input sentences to an output sequence. That's why called "Transformers"

Transformer Architecture

encoder & decoder

encoder takes input and output matrix representation

decoder Takes that encoded representation and iteratively generates an output

Encoder

positional encoding :- (encode the position of token in sequence)

allows to handle sequential data

⇒ enables transformers to capture order & structure of input data.

⇐ لفظوں کو پوزیشن کی معلومات دینا یعنی اُن کی ترتیب۔ sine, cosine

" provides information about the position of tokens in a sequence for transformer models, which process tokens in parallel "

اب اگر لفظ زیادہ لمبا ہو (sentence) تو اس کی پوزیشنز اور ویلیو زیادہ بڑھ جائے گی۔ تو اس لیے استعمال Sine & Cosine کرتے ہیں تاکہ ویلیو نہ بڑی ہو جائے۔

Attention → Self Attention → Multi Head attention
(Advance) (Advance)
basic moderate advanced

Self Attention (focus on different part of same sentences)

after previous step (vector) now → self attention

→ feed forward network → key, value, query

query × key = ...

Processes whole sentences at once, like reading a full page instead of line by line.

Multi-Head Attention

extends self attention by splitting the input multiple parts (heads) & processing them independently.

Summarized Comparison

features	Attention	Self-attention	Multi-Head Attention
i) focus	weighted part of input	part of same input	multiple aspects of same input
ii) Use Case	General focusing mechanism	understanding relationships within single input	enhanced understanding through diverse focus

encoder block

- takes input source language
- add positional encoding
- pass to multi-Head attention
- Then normalize the layer
- pass to feed forward network
- output embedding (vector with position)

Decoder

- Masking

attention matrix + masked matrix ⇒ Resultant Matrix

during training we use this technique to avoid data leakage

- Decoder
  - takes input target language
  - hides some words (mask)
  - add positional encoding
  - pass to masked multi-Head attention

Transformer Model

encoder, decoder, attention layer

BERT
(Google)
↓
encoder

forward + backward

- Bidirectional Encoder Representation from Transf
- reads in both direction simultaneously

output embeddings of

+ dimension
+ input length + type

GTP
(open AI
↓
decoder

- Generative Pre-Trained Transform
- uni-directional (next word predicts)
- abilities like summarize, comprehen...
- generate human-like text based on input-prompts

why
benefits
understand concepts