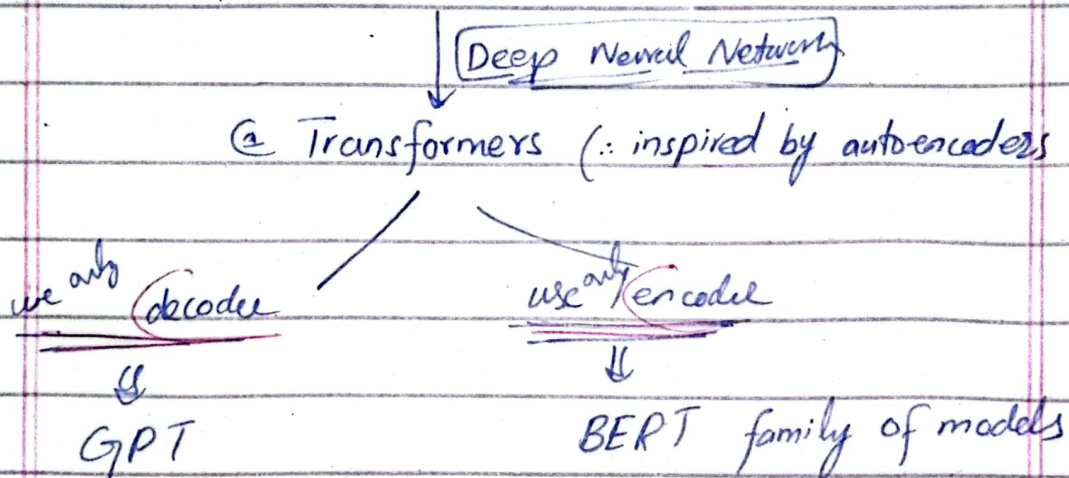


# Lecture 35: ~~Intro~~ Transformers & GPT

- ... review of previous lecture
- ... attention (solve the problem of long sequence - which information is present in past)
- ... Token

## Attention-based Model



## Transformers

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

⇒ when some sentences are translated from one to another language - the sentence, 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 representations  
decoder Takes that encoded representations and iteratively generates an output etc

### Encode

positional encoding:- (encode the position of token in sequence)  
 allow to handle sequential data  
 ⇒ enables transformers to capture order & structure of input data.

سینکس، کوسینکس  
 لفظوں کی پوزیشن کی طرف سے ان کی ترتیب۔

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

اب اگر ہم زیادہ سے زیادہ (continuous) سے تو اس کی پوزیشنز  
 اور ویلیمز (discrete) سے تو اس کے سینکس و کوسینکس  
 کے ساتھ ساتھ ویلیمز بھی ہو جائے۔



Attention <sup>advance</sup> → Self Attention <sup>advance</sup> → Multi-Head Attention <sup>advanced</sup>  
 basic moderate

Self Attention (focus on different part of same sentence)  
 after previous step (reader) now → self attention

→ 3 feed forward network → key  
 → value  
 → query

query = query × key

→ 1 output

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.

1. 2, 3, 6 (nodes) → 3 (split) = 3

→ 3 (split) = 3

→ 3 (split) = 3

→ 3 (split) = 3

Summary: real comparison

features	Attention	Self-attention	Multi-Head <sup>attention</sup>
i) focus	selected part of input	part of same input	multiple aspects of same input
ii) Use Case	General focusing mechanism	understands 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  $\Rightarrow$  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

encoders, decoder, attention layer

BERT  
(Google)  
↓  
encoder

forward + backward  
↑  
Bidirectional Encoder  
Representator from Transf

- reads in both directions simultaneously

output embeddings

+ dimension  
+ input length + type

GPT  
(OpenAI)  
↓  
decoder

• Generative / Pre-Trained Transform

- uni-directional (next word predicts)
- abilities like summarizing, comprehension
- generate human-like text based on input-prompts

why  
benefits  
// understand concepts //

