Date • modern NLP Subject

- modern NLP:
  - modern LLM becomes powerful through one simple training task. predict next word from context, repeated billion of times.
  - Training data: huge text corpora, model learns grammers, facts, reasoning implicitely
  - if a model can predict the next word distribution it can also generate text by sampling, from that distitbution.

- Three Transformer Architecture:
  - Decoder only
  - encoder only
  - Encoder - decoder ((seq to seq))

Date decoder only models Subject

- Definition :

input : tokens   output : next tokens

- How it works :

- Generates one token at a time
- only looks at previous tokens
- training objective : next tooken prediction

- why powerful ??

- left to right generation → prediction perfect for open ended text generation

- diagram ??

[the] ⟶ predict ⟹ [the, cat] ⟶ Predict next

Date Encoder only Subject

- Encoder only models ((masked language models))

- Definition:
  input : tokens, output : vector representation

- training :
  - mask some tokens : model predicts the masked word using both left and right context
  - objective : masked token prediction ((MLM))

- examples :
  - The capital of France is [mask]
  - model [mask] ⟶ paris

- properties :
  - Bidirectional attention
  - excellent for understanding tasks

- used for :
  - classification - semantic search
  - sentence similarity

Date Encoder – decoder ((seq2seq)) Subject

- Definition: - input: tokens - output: tokens

- mechanism:

① encoder reads whole input → compress it into contex

② decoder generates output tokens unique ena
representation

- why different from decoder only:

- what makes it different than decoder-only models, is t
an encoder decoder has a much looser relationship betw
the input tokens and output tokens.

- It is Good for tasks where input and output diffe
in form of length or even language.

- used for:

machine translation

summarization

paraphrasing

question answering

- example:

Encoder input : "i am hungry"

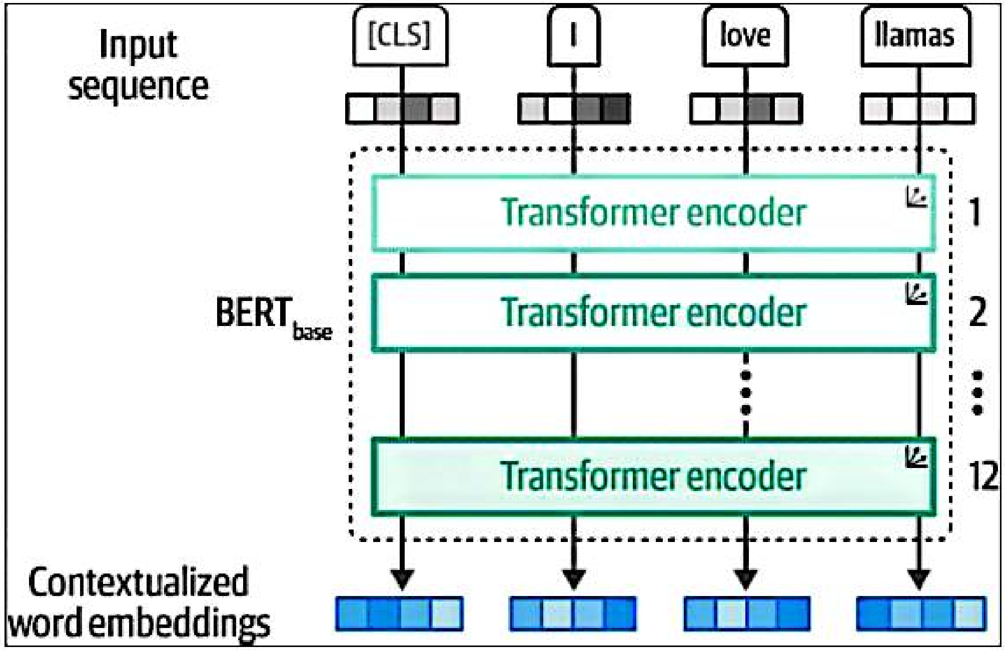
decoder output : "من گرسنه هستم"

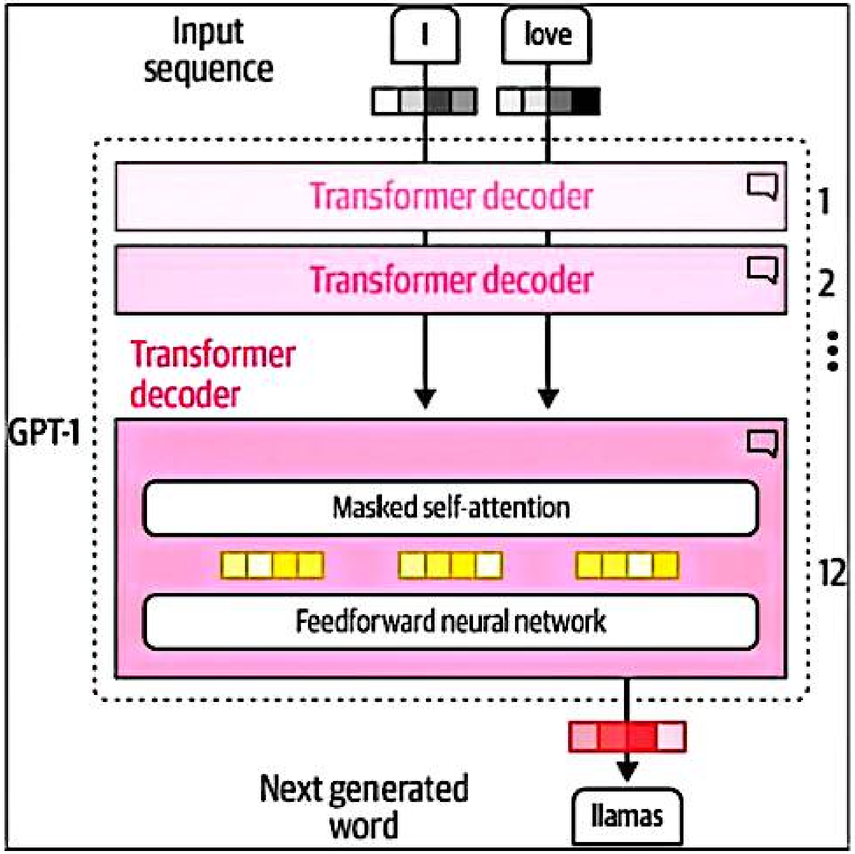
architecture:

Summary

architecture	input	output	training	strength
Decoder only	tokens	Generated tokens	next token	Generation, reasoning
encoder only	Tokens	embeddings	masked tokens	understanding, classification
encoder decoder	tokens	Generated tokens	seq 2 seq	translation, summarization

Encoder (Bidirectional) vs. Decoder (Autoregressive)


Input sequence
[CLS]
I
love
llamas
BERT base
Transformer encoder
1
Transformer encoder
2
Transformer encoder
12
Contextualized word embeddings


Input sequence
I
love
Transformer decoder
1
Transformer decoder
2
Transformer decoder
GPT-1
Masked self-attention
Feedforward neural network
12
Next generated word
llamas