Language modeling Using transformers → they are Large language models

MLM
- BERT, RoBERTa
- predictive, comprehension
- look at both preceding & succeeding text
- Thus, uses bidirectional full context

Causal LM (More on these at the very end)
- BART, PLBART, T5 GPT, LLama, CodeLlama
- generative
- predict the next token in a sequence based only on the tokens that came before it
- Thus, only uses uni-directional context (left-to-right) here, (autoregressive) (more on this later)

both use the transformer model.

- evolved from RNN-based seq. to seq. model
- key feature → new attention mechanisms

attention → the importance of each component in a sequence relative to other components in that seq.

also referred to as "self-attention" from the perspective of a token in a sequence.

formalized by Vaswani, to have 3 components Q, K, V, & computed as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

d_k → dimensionality of the key

contd.

In transformers.

You have multi-head attention
↓
runs several attentions in parallel,
↓
each can focus on different relationships in text,
Syntax, semantics

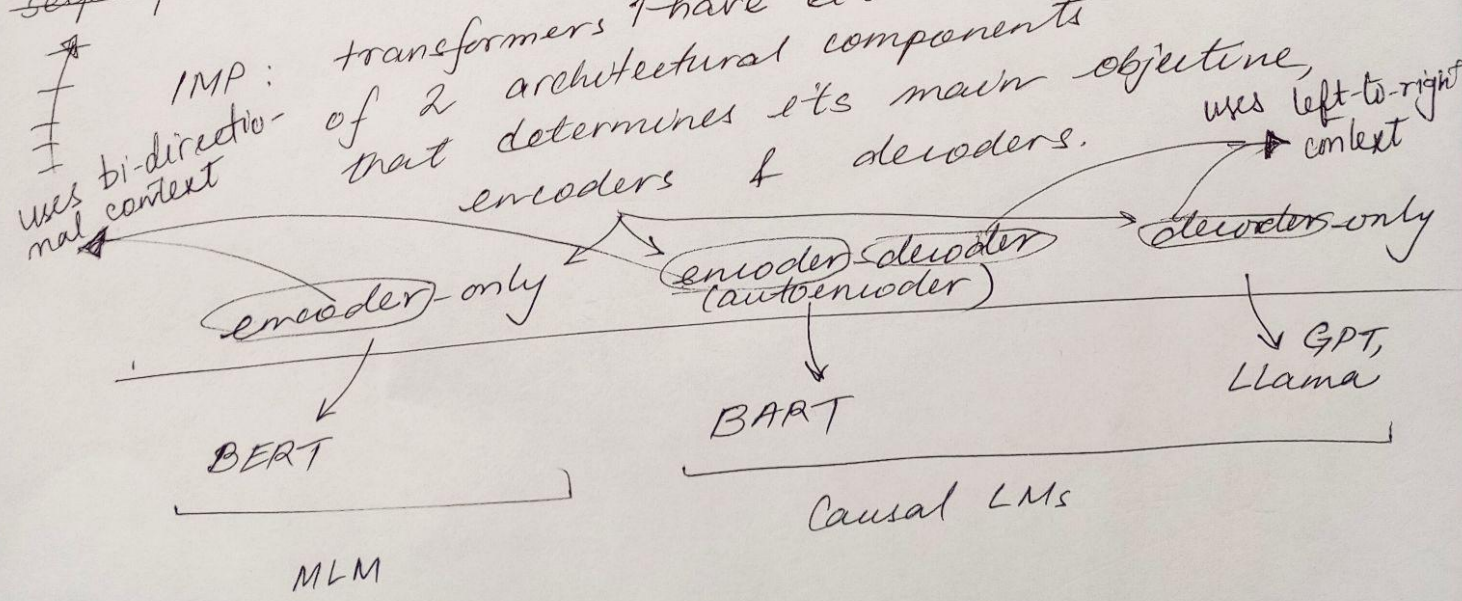
other components of transformers,

FFNs built by FCs

Residual conns. (a type of skip conn.)
↓
any bypassing conn.
↓
additive skip
(lets each layer make a small edit to the representation instead of rewriting from scratch)
(also exists in RNNs)

~~employs [For later] self-supervision~~

IMP: transformers can have either or both of 2 architectural components that determines it's main objective, encoders & decoders.



Masked lang. modeling (MLM) & causal lang. modeling (CLM) are language modeling techniques distinguished by the pre-training objectives they optimize, which in turn determine how context is used during prediction.

See more

Also note the denoising process in BART.