Artificial Intelligence
Advanced Topics in AI & ML
Transformers: BERT, GPT, LLM

Aleksandr Petiushko

ML Research

SOFIA UNIVERSITY

Content

1. Transformers

Content

1. Transformers
2. BERT

Content

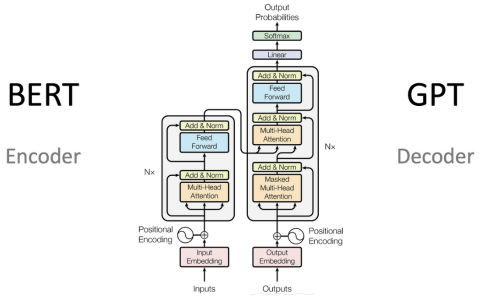
1. Transformers
2. BERT
3. GPT

Content

1. Transformers
2. BERT
3. GPT
4. LLM

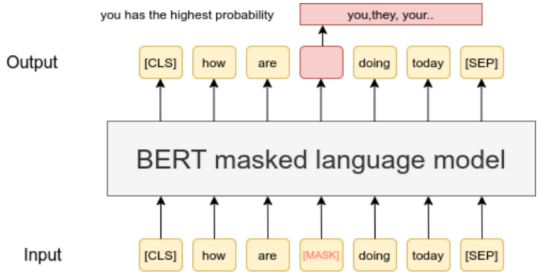
Transformer Architecture

- Transformer is the architecture with the minimal inductive bias (e.g., in comparison to CNN or RNN): we are looking at the correlation ("attention") between every pair of input elements
- Super efficient architecture for literally every ML direction: CV, ASR, LLM, etc
- Can be encoder-decoder (machine translation), encoder (language modeling), and decoder (text generation - `http://jalammar.github.io/illustrated-gpt2/`)
- Read material: link



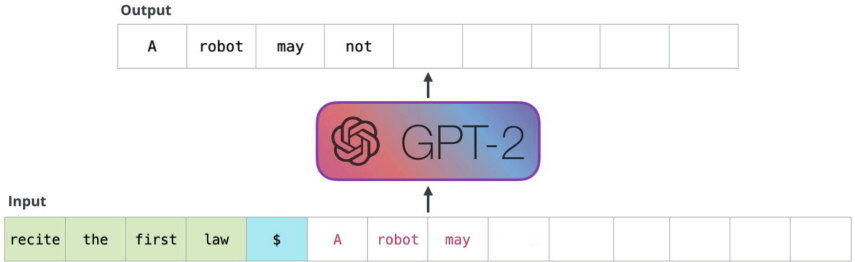
Transformer: BERT

- BERT = Bidirectional **Encoder** Representations from **Transformers**
- It is an encoder-based Transformer, only self-attention, trained by Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)
- For each input there is a vectored output that can be used for a variety of tasks: e.g., text classification, or the key/value parts of cross-attention
- Read material: link



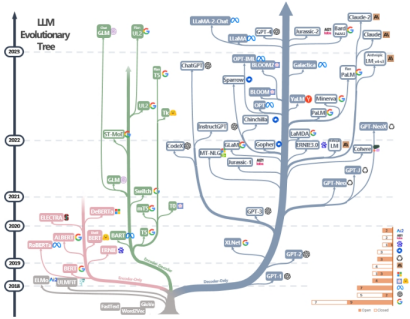
Transformer: GPT

- GPT = **Generative** Pre-trained **Transformer**
- It is a decoder-based Transformer, trained with **masked** self-attention (temporal causality) by Language Modeling (LM) — likelihood of the next predicted token based on the previous ones
- Generative because it generates autoregressively the new token based on statistical assumptions
- Read material: link



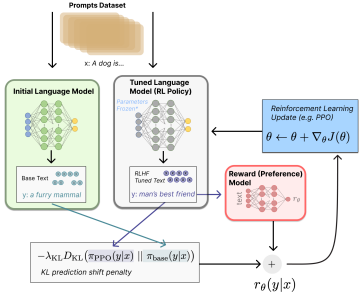
LLMs

- Large Language Model: A **huge language** model: encoder (e.g., BERT), decoder (e.g., GPT), or an encoder-decoder (e.g. for Machine Translation; although decoder models also can be used for it) pre-trained on a **massive** language corpus
- Main problems with LLMs right now: how to restrict the outputs (e.g., obscene language), how to ensure fairness, and how to avoid hallucinations
- Read material: link



ChatGPT

- ChatGPT is an auto regressive decoder model trained on a huge amount of textual data (like a half of the Internet) using the RLHF
- The architecture is not known but can be guessed on one of the OpenAI papers - InstructGPT[1]
- Read material: link



[1] Training language models to follow instructions with human feedback

Hallucinations

- Hallucination: one of the problems with LLMs
- It's the output that is not a part of the LLM training data
- Can be overcome with chain-of-thoughts, iterative output adjustment, and using external memory (retrieval augmented generation)
- Read material: link

Passage	Scenario #1 - Hallucination
Original Answer	Philip Hayworth was an English barrister and politician who served as Member of Parliament for Thetford from 1859 to 1868.
Sample 1	*Philip Hayworth was a **British politician** who served as the Member of Parliament for **Bolton West** from **1931 to 1945**. He was also a member of the Free Trade Union and served on several government committees.*

Takeaway notes

1. Read all the mentioned links

Takeaway notes

1. Read all the mentioned links
2. Transformers with their self-attention block is the key architectural ingredient of LLMs and beyond

Takeaway notes

1. Read all the mentioned links
2. Transformers with their self-attention block is the key architectural ingredient of LLMs and beyond
3. Encoder, encoder-decoder, and decoder variants can be used for LLMs; decoder is now the dominant one

Takeaway notes

1. Read all the mentioned links
2. Transformers with their self-attention block is the key architectural ingredient of LLMs and beyond
3. Encoder, encoder-decoder, and decoder variants can be used for LLMs; decoder is now the dominant one
4. ChatGPT and successors added human feedback in the loop to make the output more plausible for the human user

Takeaway notes

1. Read all the mentioned links
2. Transformers with their self-attention block is the key architectural ingredient of LLMs and beyond
3. Encoder, encoder-decoder, and decoder variants can be used for LLMs; decoder is now the dominant one
4. ChatGPT and successors added human feedback in the loop to make the output more plausible for the human user
5. It is still unknown how far we are from AGI

Thank you!