Introduction to Generative A.I and LLMs

by, Umair M.Imam

DAVE3625- INTRO TO A.I. BY UMAIR M.I

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Generative AI

python Copy code

```python
def add_numbers(a, b):
    return a + b

# Example usage:
result = add_numbers(5, 7)
print(result)  # Output: 12
```

This function, named `add_numbers`, accepts two parameters `a` and `b` and returns the result of adding them together.

K Now make it so this same function takes in a list of numbers and adds them together

To modify the function to accept a list of numbers and return their sum, you can use the built-in `sum()` function in Python. Here's the updated function:

python Copy code

```python
def add_numbers(numbers):
    return sum(numbers)

# Example usage:
result = add_numbers([1, 2, 3, 4, 5])
print(result)  # Output: 15
```

Generative AI

Branch of AI which creates new content
(in text, images, audio, or other forms of media)

- Generative AI is a subset of traditional machine learning
- **Creativity and Innovation**:
  - Unlike traditional AI that is often about recognizing patterns or making predictions, generative AI can create novel outputs, like a new poem, a piece of music, or a unique image.
- **Learn from data:**
  - This kind of machine learning finds statistical patterns in massive amounts of data sets
- **Diverse Applications**:
  - Such as writing, art creation, drug discovery, synthetic data generation etc etc.

Brief history of Generative AI models: GAN

- First there was GAN for Generative AI
- An important subset of Generative AI is the Generative Adversarial Network (also called GAN)
- They consist of two neural networks, a generator and a discriminator, that compete against each other, leading to the generation of highly realistic data.
- the generator creates data, while the discriminator evaluates it. The generator aims to produce data so realistic that the discriminator can't distinguish it from real data.
- used to generate a wide range of data types, including images, music and text.


Network & latent
Pickle
stylegan2_ffhq_512_pytorch.pkl
Recent...
Browse...
Latent
7
Seed
0.001
Step Size
Reset
w
w+
Drag
Drag
Add point
Reset point
Start
Stop
Steps: 0
Mask
Movable area
Fixed area
Reset mask
Show mask
112
Radius
20
Lambda
Capture
Capture
_screenshots
Save image
Save GUI

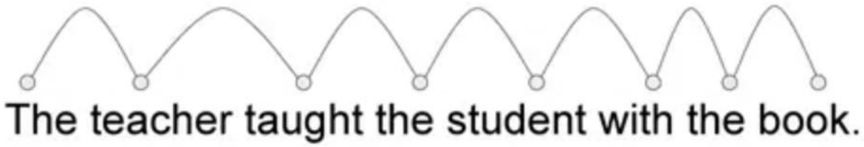
Brief history of Generative AI models

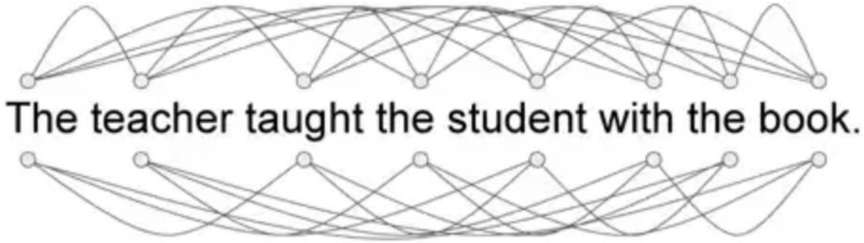
- Then Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) networks were the go-to models for handling sequential data like text. However, they struggled with long-range dependencies and were computationally intensive.
  - a "network" refers to a specific architecture of neural networks designed for processing sequences of data, like time series, sentences, or audio.

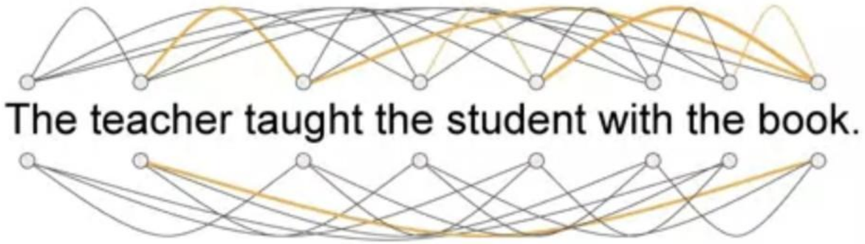
Transformer Models

The transformer model, introduced in the paper “Attention is All You Need” (2017), brought a significant shift. It relies on an attention mechanism, allowing it to process input data in parallel and handle long-range dependencies efficiently.

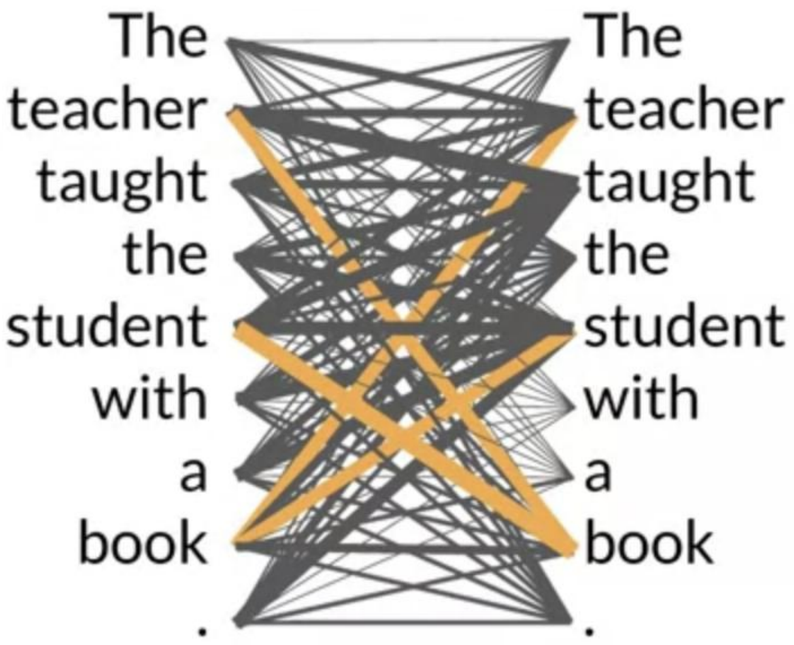
The development of transformer models marked a turning point in generative AI, particularly in natural language processing. This led to the creation of Large Language Models (LLMs) like GPT and BERT, capable of understanding and generating human-like text.


The teacher taught the student with the book.


The teacher taught the student with the book.


The teacher taught the student with the book.

Self attention



Large Language Models

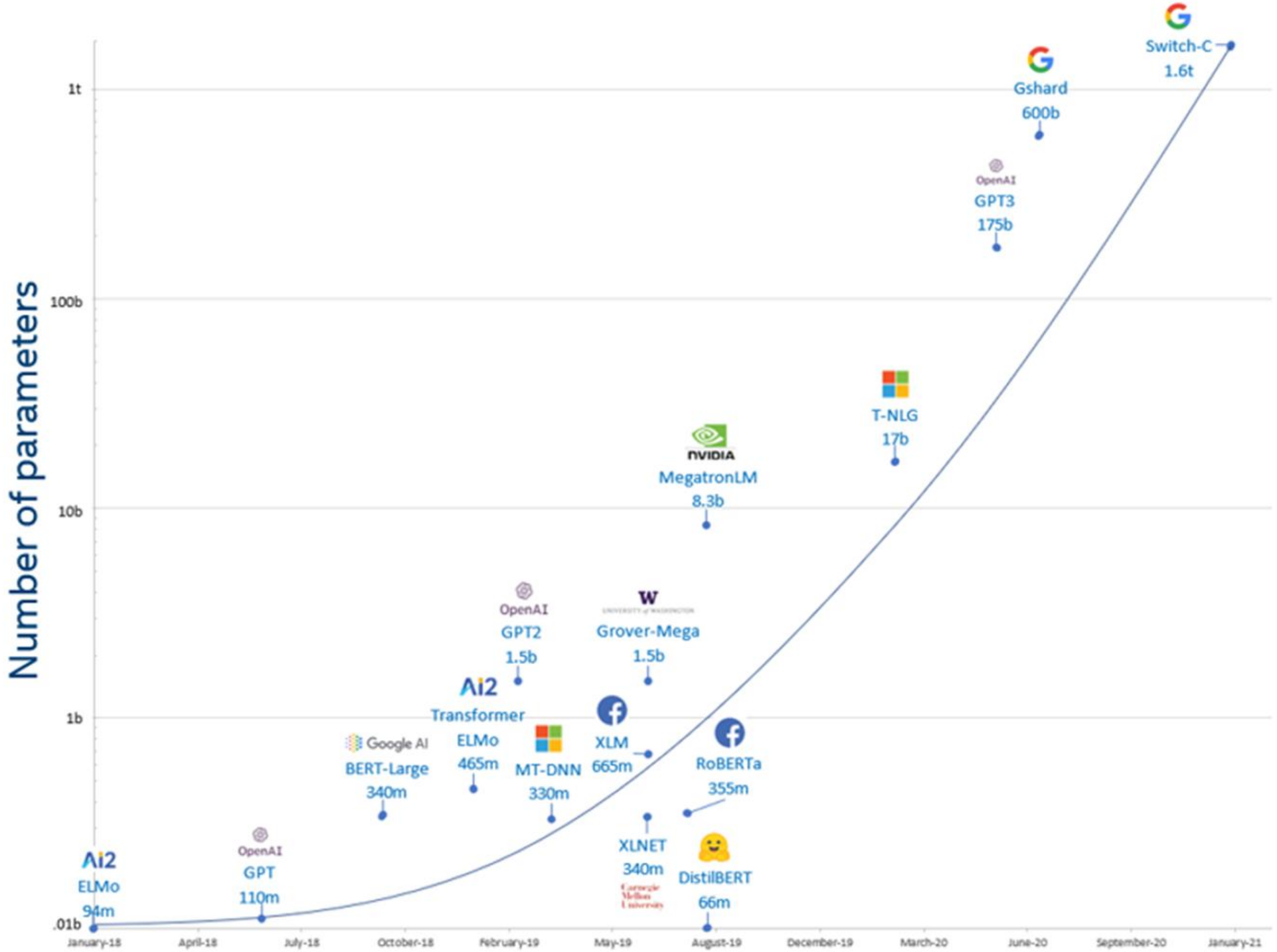
- Large Language Models are advanced AI models trained on vast datasets to understand and generate human language. Their large size allows them to capture intricate nuances of language.
- Examples: GPT, BERT, T5 and so many more
- Training these models requires substantial computational resources and carefully curated datasets. Challenges include handling biases in training data and ensuring ethical use.

OSLOMET

LLMs are also called foundation models which have a transformer architecture

- Transformer model refers to the architecture of a model
- Whereas Foundation model refers to the category of the model (big model)

- **Foundation Model**: This is like a big, versatile toolbox that can be used for many different things. It's got tools for all sorts of jobs and can be adapted to do lots of tasks.
- **Large Language Model (LLM)**: This is like a special set of tools in that big toolbox, but these tools are specifically for understanding and using language. This means they're really good for jobs like writing stories, having conversations, answering questions, and even making jokes.


Number of parameters
1t
100b
10b
1b
.01b
January-18
April-18
July-18
October-18
February-19
May-19
August-19
December-19
March-20
June-20
September-20
January-21
Ai2
ELMo
94m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
Ai2
Transformer
ELMo
465m
OpenAI
GPT2
1.5b
MT-DNN
330m
XLM
665m
W
UNIVERSITY of WASHINGTON
Grover-Mega
1.5b
NVIDIA
MegatronLM
8.3b
XLNET
340m
Carnegie Mellon University
RoBERTa
355m
DistilBERT
66m
T-NLG
17b
OpenAI
GPT3
175b
Gshard
600b
Switch-C
1.6t

- Parameters in a Large Language Model (LLM) refer to the internal variables that the model uses to process and generate text.
- For example
  - They can also be referred to as the amount of memory of a model
  - The more parameters, the higher the memory

- The more parameters, the more sophisticated task they can perform

Terminologies

Token

- Is a basic unit of text, which can be a word, part of a word, or punctuation
- For example,
  - in the sentence "The cat sat on the mat," each word is a separate token.
  - the word "unbelievable" might be split into "un-", "believ-", and "-able”
  - Punctuation marks like commas, periods, or question marks are often treated as separate tokens.

- Context Window:
  - refers to the maximum span of text the model can consider at any given time when processing input and generating responses.
  - For instance, GPT-3 has a context window of 2048 tokens. This means it can consider up to 2048 tokens at a time when reading an input or generating a response.

Prompting