OpenAi's APIs

Souvik Ghosh

OpenAi's APIs

The Founders

OpenAI was founded in 2015 by a group of high-profile individuals in tech, including:

Elon Musk
Sam Altman
Greg Brockman
Ilya Sutskever
John Schulman
Wojciech Zaremba

What is GPT?

- Full from of GPT: Generative Pre-Trained Transformer.
- A language model that uses deep learning to generate text
- We give it initial text (the prompt) and it produces text to continue the prompt.

e.g.

Prompt	Output
Why did the chicken cross the	road?
My favorite food is:	pizza!
Translate from English to French: snails =>	escargots
Write a tagline for a doggy day care	Cozy, Safe and Fun for Your Furry Friend!
Write 2 sentences explaining acid rain to a toddler	Acid rain is rain that has acids in it, which makes it a little bit yucky. Acid rain can hurt trees and other plants, and even give us coughs if the air is full of it.
Correct this to standard English: She no have no money	She doesn't have any money.

The 2 ways we can "talk" to GPT models

Completion	Chat
• Expects a single text prompt	• Expects a list of messages in a chat-based format
• Does not support GPT-4	• Supports GPT-4

How?

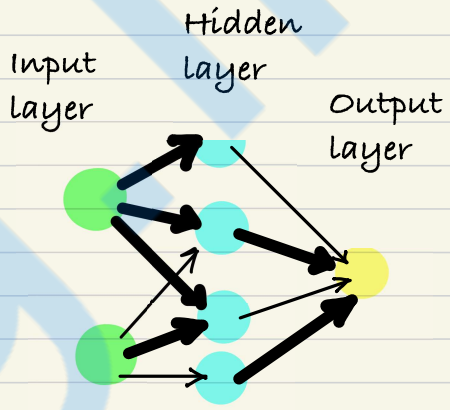
GPT-4 is based on a type of neural network called a transformer.

• Transformers are a deep learning model that excel at processing sequential data. (like natural language text!)

Neural Networks

• Neural networks are a category of models that are very good at analyzing complicated data types.

• They consist of layers of connected nodes that can "fire" like neurons - passing data to other nodes.



A Simple Neural Network

There are many different types of neural networks

• Convolutional Neural Networks work great for analyzing images

• Recurrent Neural Networks work well at text processing and translation

• Transformers (what we care about)

Recurrent Neural Networks

RNNs work sequentially, processing text one word at a time, but they have some problems:

• They're not great at analyzing large pieces of text.

• They're slow to train. This means they can't be trained on huge amounts of data easily.

• Training can't be parallelized because they process sequentially.

Transformers

- Transformers are a relatively recent approach (2017)
- Transformers process the entire input at once, rather than sequentially, meaning there is less risk of "forgetting" previous context.
- This means they can be trained in parallel!
- Transformers introduced a couple key innovations

Positional & Self Encoding Attention

Positional Encoding

- Instead of dealing with each word sequentially, one at a time, transformers encode positional data
- Before feeding each piece of the input into the neural network, we label it with positional information
- Word-order information is stored in the actual data itself rather than in the network structure
- The network learns the significance of word-order from the data itself

Attention

- Attention in a neural network is a mechanism that allows the network to selectively focus on certain parts of input data, while ignoring others
- Think of how humans focus our attention on certain aspects of the world, while filtering out irrelevant information
- Attention allows the network to focus on parts of the input data and dynamically asjust its focus as it processes the data
- There are different attention mechanisms but most involve computing an attention score for each piece of input data
- These scores are then used to compute a weighted sum or average of the input elements

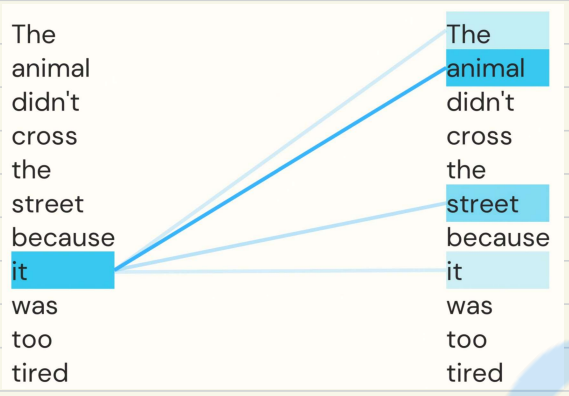
But How??

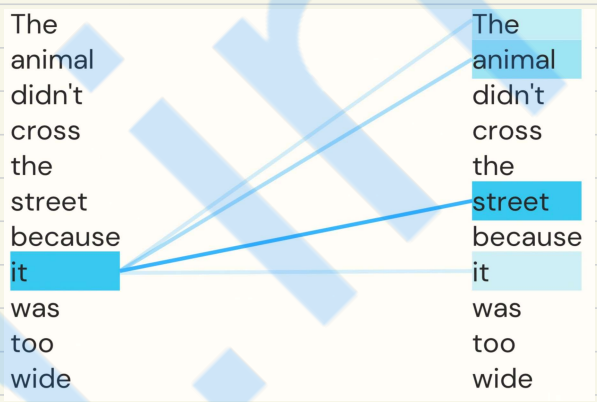
How does the model "know" what words it should "attend" to?

It's learned over time from lots and lots of data. With enough data, the model learns the basics of grammar, word order, etc.

Self-Attention

- Self-attention is one of the key innovations that makes the transformer model so effective
- In self-attention, each element in the input sequence is compared to every other element in the sequence, and a set of attention weights is computed based on the similarity between each pair of elements
- The "self" in self-attention refers to the the same sequence which is currently being encoded.





openAi Models

DALL-E: generates and edits images
Whisper: converts audio to text
Moderation: detects safe and unsensitive text
GPT-3: understands and generates natural language
GPT-3.5: set of models of that improve upon GPT-3
GPT-4: The latest and most advanced version of OpenAi's large language model

Training

- GPT-3 was trained on over 45TB of text data
- Nearly 500 billion tokens of training data

Open AI has not released information on the training of GPT-4

Size

- GPT-3 is absolutely massive compared to GPT-2 GPT-3 has 175 billion parameters and takes 800gb just to store the model itself
  @ That's 800gb of basically numeric data that forms the model.
- It cost over $4.6 million in GPU costs to initially train GPT-3
- OpenAI has not released the technical details of GPT-4

GPT-3.5

GPT-3.5 is not an entirely new model. It's a finely-tuned version of GPT-3 developed in 2022 and trained on data through 2021.

GPT-4

GPT-4 is a HUGE new model that is currently in beta. You must join a waitlist and be approved to gain access to GTP-4 via the APIs.

Tokens

- GPT doesn't work with words, but instead uses a system of tokens.
- Tokens are essentially pieces of words (though some tokens are full words)
- A token on average is ~4 characters of English text.

Pricing

- Open AI charges based on tokens.
- It adds together the tokens in your prompt plus the tokens in the output it returns.
- Different models are priced differently:
    - text-davinci-003: $0.02 / 1K tokens
    - text-curie-001: $0.002 / 1K tokens
    - text-babbage-001: $0.005 / 1K tokens
    - text-ada-001: $0.0004 / 1K Tokens
    - GPT-3.5-turbo: $0.002 / 1K tokens
    - GPT-4 Models: $0.06 - $0.12 / 1K Tokens

Prompt Design

Main Instructions - a task you want the model to perform
Data - any input data (if necessary)
Output Instructions - what type of output do you want? What format?

Provide clear instructions
Complete the sentence:
Humans are

Use a separator to designate instructions and input
Instruction
Translate the text below to French:
Text: "I am a huge idiot"

Reduce "fluffy" language. Be precise
In 3-4 sentences, explain the role of the Supreme Court in US politics. The explanation should be geared towards middle-schoolers

Be specific about your desired output
Extract the place names in the following
Desired format:
Places: <comma_separated_list_of_places>
Input: " Airbnb, Inc.is an American San Francisco-based company. The company is credited with revolutionizing the tourism industry however it has also been the subject of intense criticism by residents of tourism hotspot cities like Barcelona, Venice, etc. for enabling an unaffordable increase in home rents, and for a lack of regulation."

Some API Parameters

Temperature

- A value from 0-2, though most often between 0 and 1
- Its default value is 1
- Controls the randomness of the output. Higher values are more random, lower values are more deterministic

Top P

- An alternative to sampling with temperature, called nucleus sampling Its default value is 1
- Like temperature, top p alters the "creativity" and randomness of the output.

Frequency Penalty

- A number from -2 to 2
- Defaults to 0
- Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.
- If you want to reduce repetitive samples, try a penalty from 0.1 - 1
- To strongly suppress repetition, you can increase it further
- BUT this can lead to bad quality outputs Negative values increase the likelyhood of repetition

Presence Penalty

- A number from -2 to 2
- Defaults to 0
- Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.
- Presence penalty is a one-off additive contribution that applies to all tokens that have been sampled at least once
- Frequency penalty is a contribution that is proportional to how often a particular token has already been sampled

ChatGPT API

• The ChatGPT API allows us to use gpt-3.5-turbo and gpt-4.
It uses a chat format designed to make multi-turn conversations easy.
• It also can be used for any single-turn tasks that we can with the completion API.

Completion API

```
openai.Completion.create(
model="text-davinci-003"
prompt="tell me a joke"
)
```

Chat API

```
openai. ChatCompletion.create(
model="gpt-3.5-turbo",
messages=[
{"role": "user", "content": "tell me a joke"}
]
)
```

Messages

• The chat API expects a list of messages rather than a single text prompt.
• Messages must be an array of message objects, where each object has:

- @ a "role", set to:
  - "system",
  - "user",
  - "assistant"
- @ "content" (the content of the message)

DALL-E

- DALL-E is a neural network-based image generation system.
- It generates images from text prompts.
- To train DALL-E, OpenAI used a dataset of over 250 million images and associated text descriptions.

```
response = openai.Image.create(
prompt="a dancing pickle",
n=1,
size="1024×1024"
)
image_url = response['data'][0]['url']
```

Image Sizes

- The DALL-E API only supports square images of the following sizes:
  - 256x256 pixels
  - 512x215 pixels
  - 1024x1024 pixels

The smaller the image, the faster it is to generate

Whisper

- Whisper is OpenAI's speech recognition model.
- It can perform speech recognition, translation, and language identification
- It costs $0.006 / minute (rounded to the nearest second)

Embeddings

- Embeddings are numerical representations of text concepts converted to number sequences
- They make it easy for computers to understand the relationships between those concepts.
- OpenAI has an embedding model called text-embedding-ada-002.
- Given some input text, it returns an embedding as a 1536 dimension vector.
- We can store these embeddings and then use them to perform searches, recommendations, and more.