ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗

November 30, 2022
13 minute read

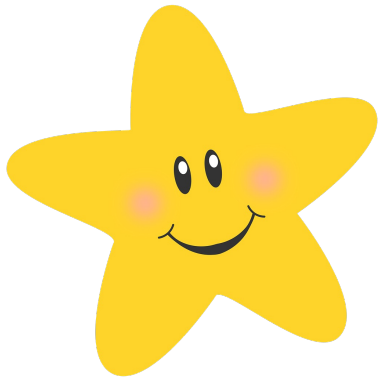
ChatGPT

How it works

Presented by:
John Tan Chong Min

What comes next?

- Twinkle Twinkle Little _____



- The wheels on the bus go round and _____

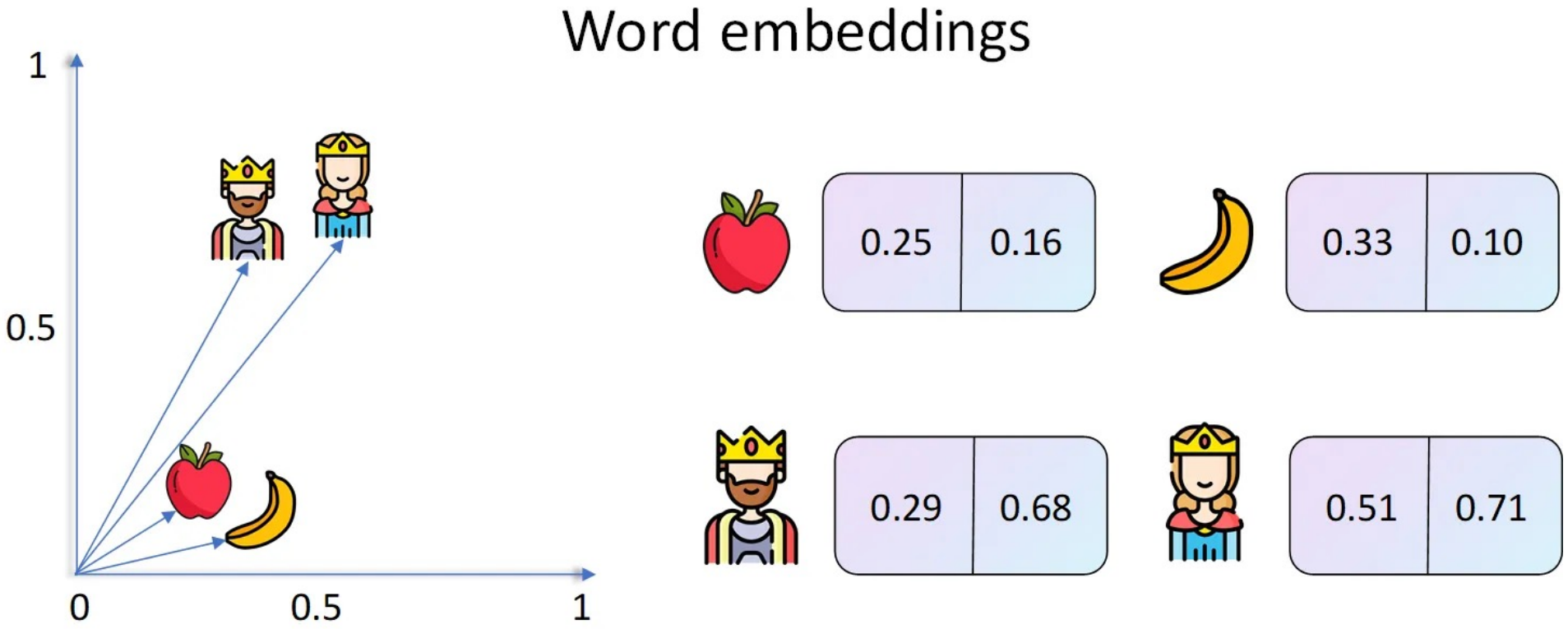
What comes next?

- The best FIFA player is _____

We can do these in our brains.

Now how do we do these in a computer?

Word Embeddings



https://towardsdatascience.com/deep-learning-for-nlp-word-embeddings-4f5c90bcdab5

What is the meaning of this word?

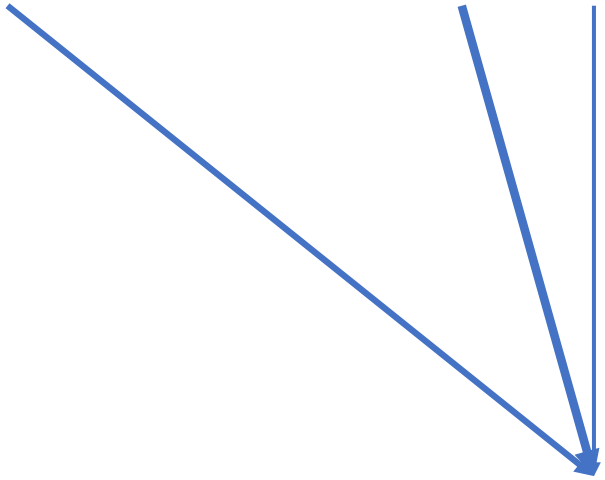
- Bank

We want to update meaning of each word iteratively

- Mary went to the river bank

We want to update meaning of each word iteratively

- Mary went to the river bank

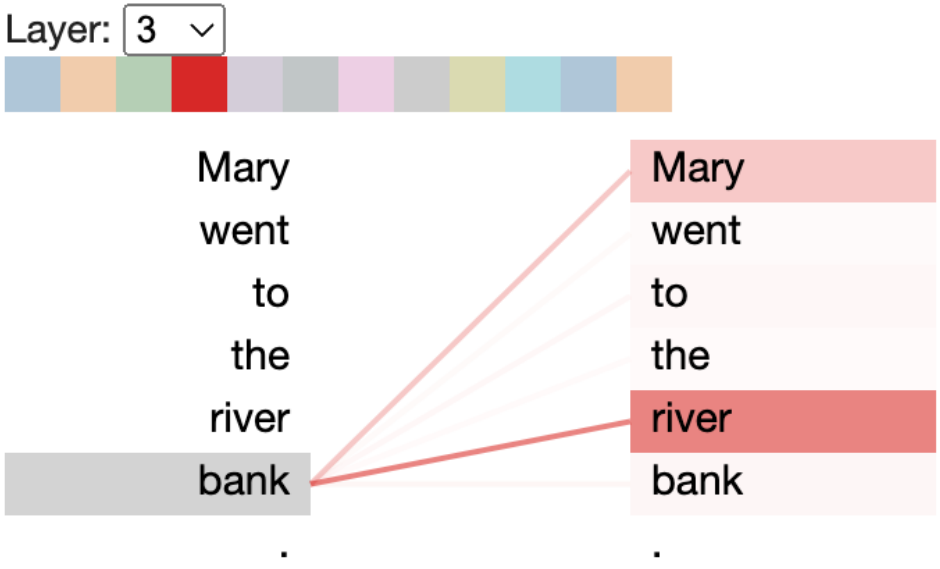


- Mary went to the river bank

Pay attention differently for each word!

Visualizing what “bank” pays attention to

```
from IPython.core.display import HTML
display(HTML('<script src="/static/components/requirejs/require.js"></script>'))
# Above two lines only needed when running in Colab
head_view(attention, tokens)
```



GPT generates differently each time!

- Samples the next token according to the probability distribution

- Example:
  - **Input: Mary had a**
  - Probability of next tokens: little (80%), big (20%)
  - Output: Mary had a little

  - **Input: Mary had a little**
  - Probability of next tokens: lamb (80%), dog (20%)
  - Output: Mary had a little lamb

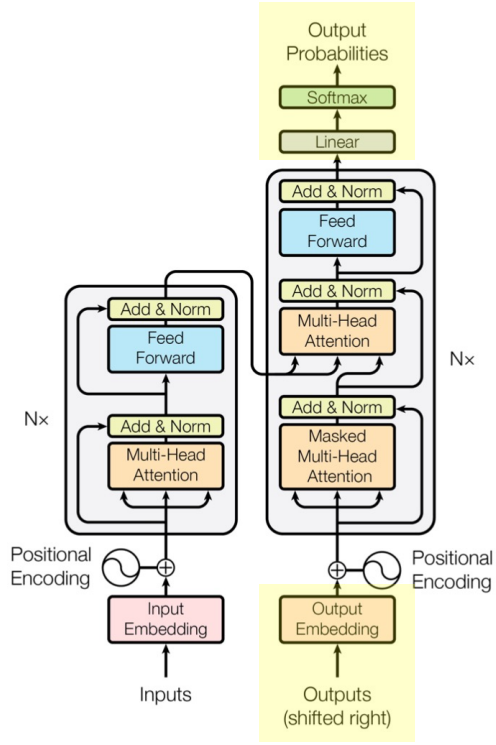


Figure 1: The Transformer - model architecture.

Strengths of ChatGPT

- Good for combining human knowledge in novel ways
  - Creative
  - Adaptable to small changes in input prompt
- Great for coding and summarization
- Good for doing homework requiring regurgitation of concepts
- Can be a fun AI to chat with on any topic

Weaknesses of ChatGPT

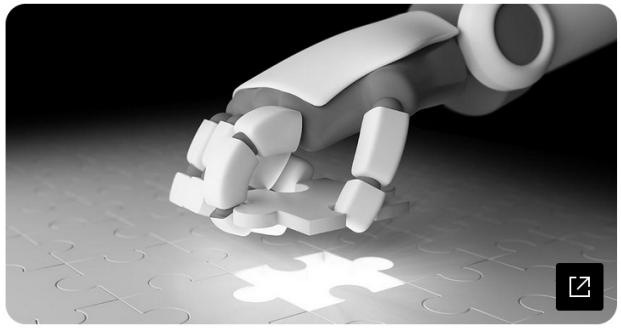
- Can give inconsistent output
- Confident even when incorrect
- Can be ambiguous and hedge multiple options even when there is one clear option
- Not good at math (no longer any more with calculator functionality!)

ChatGPT as an interface

My own video on advanced explanation of ChatGPT

how chatgpt works john

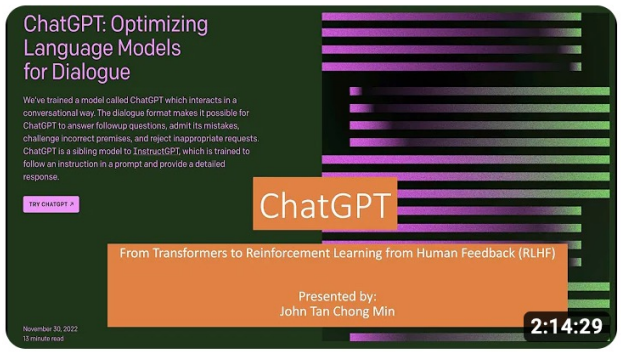
Filters



Using ChatGPT At Work - Subscribe To The Best Of Udemy

Take a ChatGPT course and learn how to use it as a tool when creating SEO and sales copy. Learn how ChatGPT can help you...

Ad · https://www.udemy.com/chatgpt/coding_courses



How ChatGPT works - From Transformers to Reinforcement Learning with Human Feedback (RLHF)

14K views • 4 months ago

John Tan Chong Min

ChatGPT has recently been released by OpenAI, and it is fundamentally a next token/word prediction model. Given the prompt ...

Introduction | Embedding Space | Overall Transformer Architecture | Transformer (Details) | GPT... 10 chapters

Discussion