Introduction to Large Language Models

Martynas Maciulevičius

2023-05-26

About me

- Software engineer
- Past several jobs were in Clojure
- A little bit of experience in Machine Learning
  - In-browser Tetris
  - Plays itself using a genetic algorithm
  - Worked in a data science company
  - Master's in Computer science

Contents


LEFT
EXIT 12
Honest work
ChatGPT
Humanity
imgflip.com

How do big companies embrace models?

- Apple banned ChatGPT for employees [1]
- ChatGPT banned in Italy over privacy concerns [2]

Misconception with title OpenAI

- AI: Yes (it's still Machine Learning)
- Open: No
- Microsoft, Elon Musk, Peter Thiel and more
- Non-profit at first $\rightarrow$ capped profit of 10x/investment [3]
- Microsoft invests $10B [4] [5]
- Codex is a scrambler; there is a class action lawsuit [6] [7]

Crypto scams aren't on the same level

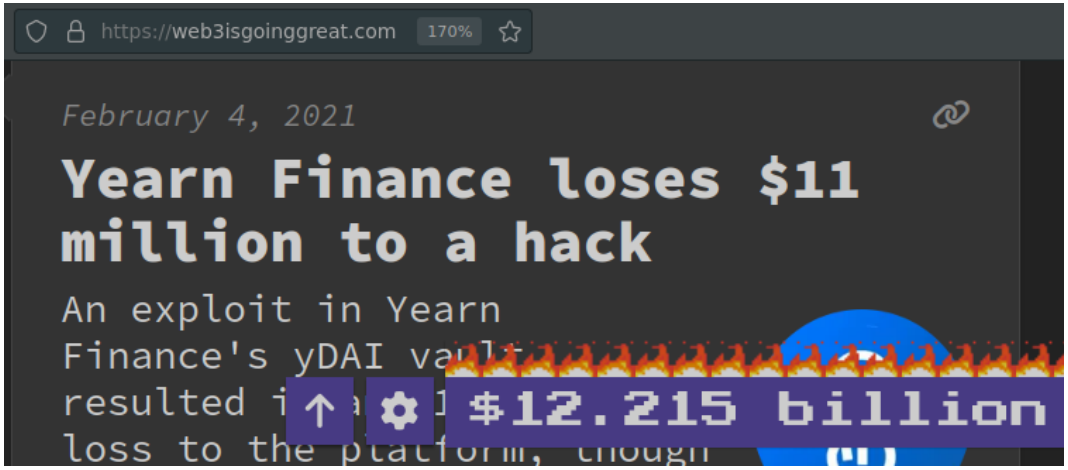


Figure: Lost money from crypto scams from 2021-02-04

Natural Language Processing (NLP) models

- Token list processing
  `["Hello" "world" "."]`
- Consume the input [list of tokens]
- Remember the meaning
- Decide on the output [list of tokens]

Natural Language Processing (NLP) models

- Token list processing
  `["Hello" "world" "."]`
- Consume the input [list of tokens]
- Remember the meaning
- Decide on the output [list of tokens]
- Sentences have variable length $\Rightarrow$ Can't use fixed-size NNs

Natural Language Processing (NLP) models

- Token list processing
  `["Hello" "world" "."]`
- Consume the input [list of tokens]
- Remember the meaning
- Decide on the output [list of tokens]
- Sentences have variable length $\Rightarrow$ Can't use fixed-size NNs
- Output can have different length than input $\Rightarrow$ Can't use fixed-size NNs

How can a computer "understand" a word?

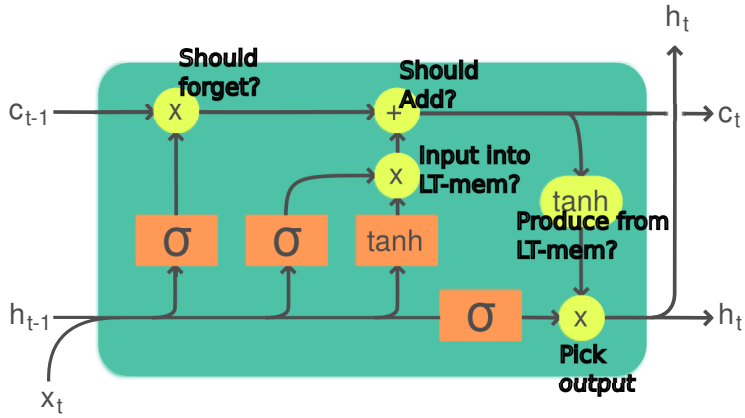
- Bag of Words
- TF-IDF
- ...
- Word Embedding:

```
{:attrs          [:fluffy :pointy :metallic :wooden]
 :words {:cat    [    0.8     0.9       0.0     0.0]
         :dog    [    0.6     0.9       0.0     0.0]
         :leaf   [    0.1     0.7       0.0     0.2]
         :hammer [    0.0     0.2       0.3     0.7]}}
```

LSTM cell

- Long Short Term Memory
- Stateful, changes states with every parsed token

LSTM memory cell



Legend:

Layer

Componentwise

Copy

Concatenate

LSTM cell in a nutshell

- Stores internal state in two fixed-size variables. Literally.
- Must choose between forget/add if a new token is interesting
- Can't remember everything in its two variables
- Outputs its short-term memory component at every step

Transformer [9]

- Self-attention instead of RNN/LSTM's variables
- Multi-head self-attention

Transformer [9]

- Self-attention instead of RNN/LSTM's variables
- Multi-head self-attention

What data does self-attention use?

- k - word embedding vector size
- Word embedding $\vec{w} \rightarrow$ dimension of $1 \cdot k$ ($1 \cdot k$ columns)
- Matrix $M \rightarrow$ dimension of $k \cdot k$
- $M * \vec{w} \rightarrow$ dimension of $1 \cdot k$
- W_Q, W_K, W_V – matrices with trained data

Transformer [9]

- Self-attention instead of RNN/LSTM's variables
- Multi-head self-attention

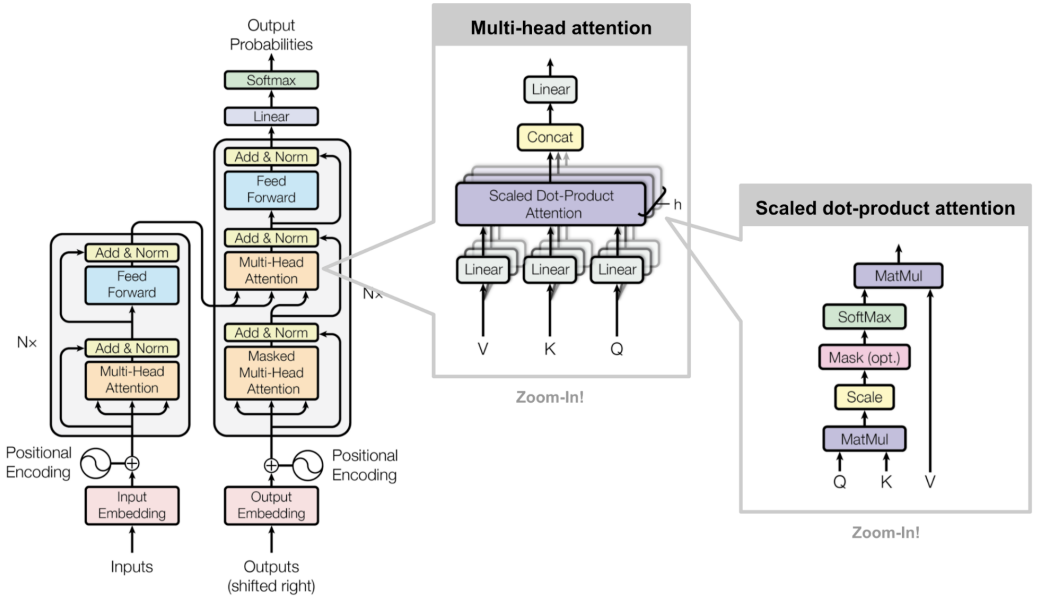
What data does self-attention use?

- k - word embedding vector size
- Word embedding $\vec{w} \rightarrow$ dimension of $1 \cdot k$ ($1 \cdot k$ columns)
- Matrix $M \rightarrow$ dimension of $k \cdot k$
- $M * \vec{w} \rightarrow$ dimension of $1 \cdot k$
- W_Q, W_K, W_V – matrices with trained data
- Non square-shaped matrices [9] can be used to reduce vector sizes in the self-attention block

How is Transformer different from LSTM?

- Vanishing gradient of RNN & LSTM
  - LSTM has two "slots"
  - RNN has one "slot"
  - Transformer uses dot-product of vectors (attention (basically words*words))
- "infinite" theoretical window of reference
- Parallelizable – LSTMs and RNNs are procedural
- Transformer is stateless: outputs its "whole memory" instead of storing state
- Attention matrix multiplication complexity is N^2, it's more expensive

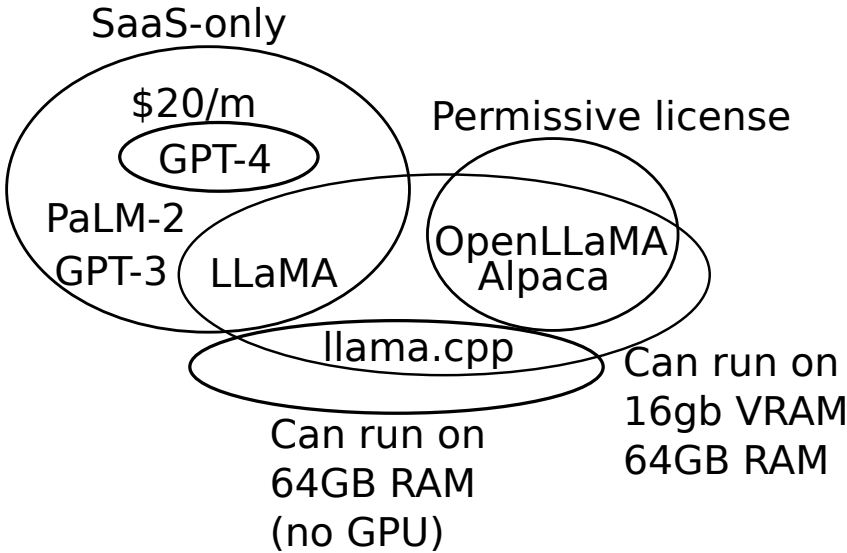
Transformer LLM [9]



Models in the wild

- PaLM - Vertex AI (from Google)
- GPT - OpenAI (from Microsoft)
- LLaMa - Leaked from Meta [10] [11]
- llama.cpp - LLaMA in C++, GPU not needed [12]
- OpenLLaMa - Trained on open RedPajama dataset [13]
- Alpaca - Stanford University [14]
- GPT-4chan: Trained on toxic posts [15]
- ...

Use and Licenses of models




HUMANS?
WHERE WE'RE GOING WE DON'T NEED... HUMANS
imgflip.com

Misconceptions while using the models

- Let's tie a shoe
- OpenLLaMA 7B (undertrained model)
- OpenLLaMA 30B (still basic model)
- Clojure environment
- ChatGPT

Thanks slide

- Daniel Slutsky
- Žygimantas Medelis
- Others

References I

[1] *Apple bans ChatGPT use by employees, report says*. URL: https://mashable.com/article/apple-chatgpt-employee-ban-report (visited on 05/26/2023).

[2] *ChatGPT banned in Italy over privacy concerns*. URL: https://www.bbc.com/news/technology-65139406 (visited on 05/26/2023).

[3] *OpenAI shifts from nonprofit to 'capped-profit' to attract capital*. URL: https://techcrunch.com/2019/03/11/openai-shifts-from-nonprofit-to-capped-profit-to-attract-capital/ (visited on 05/25/2023).

[4] *Microsoft and OpenAI extend partnership*. URL: https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ (visited on 05/25/2023).

References II

[5] *Who's getting the better deal in Microsoft's $10 billion tie-up with ChatGPT creator OpenAI?* URL: https://fortune.com/2023/01/24/whos-getting-the-better-deal-in-microsofts-10-billion-tie-up-with-chatgpt-creator-openai/ (visited on 05/25/2023).

[6] *We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy.* URL: https://githubcopilotlitigation.com/ (visited on 05/25/2023).

[7] *We've filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists.* URL: https://stablediffusionlitigation.com/ (visited on 05/25/2023).

[8] *Long short-term memory*. URL: https://en.wikipedia.org/wiki/Long_short-term_memory (visited on 05/25/2023).

References III

[9] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[10] Meta. *LLaMA model*. Feb. 2023. URL: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/ (visited on 05/24/2023).

[11] *TechScape: Will Meta's massive leak democratise AI – and at what cost?* Feb. 2023. URL: https://www.theguardian.com/technology/2023/mar/07/techscape-meta-leak-llama-chatgpt-ai-crossroads (visited on 05/24/2023).

[12] *Locally executable LLM models*. URL: https://libreddit.esmailelbob.xyz/r/LocalLLaMA/wiki/models/ (visited on 05/25/2023).

References IV

[13] Xinyang Geng and Hao Liu. *OpenLLaMA: An Open Reproduction of LLaMA*. May 2023. URL: https://github.com/openlm-research/open_llama.

[14] Rohan Taori et al. *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca. 2023.

[15] ???? *Toxic GPT model from 4chan*. URL: https://www.msn.com/en-us/news/technology/ai-trained-on-4chan-s-most-hateful-board-is-just-as-toxic-as-you-d-expect/ar-AAYe3RJ (visited on 05/24/2023).