

# 12 – Generative Language Models

## IA161 Natural Language Processing in Practice

Tomáš Foltýnek

Department of Machine Learning and Data Processing  
FI MU

December 5, 2023

1 How Generative Language Models Work

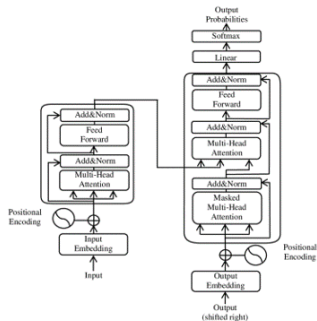
2 Ethical Challenges

# Acknowledgement

- of the tools and sources that influenced the content of the lecture
  - ▶ In accordance to the MU Recommendations on using AI in education
- Perplexity.ai to identify useful sources
  - ▶ Good explanations of LLMs
  - ▶ Papers and datasets dealing with bias in LLMs
- ChatGPT 4 Turbo
  - ▶ To get explanations of various concepts
  - ▶ To generate some Python code for the practical part
- The presentation uses screenshots from Serrano Academy

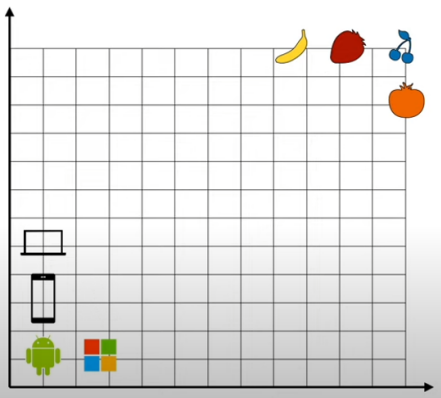
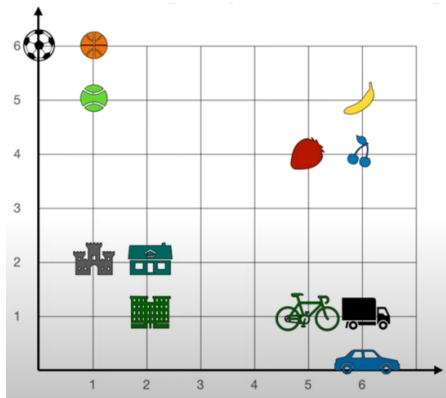
# Transformers

- 2017 Google Brain
  - ▶ Attention is all you need
- Encoding
  - ▶ Vector representation of each token
  - ▶ Based on word embeddings (i.e. context of words)
  - ▶ Attention (relations) between tokens
  - ▶ Feed-forward neural network
- Vector representation of the “meaning” of the input text
- Decoding
  - ▶ Based on the input from the encoder and the previous output of the decoder
  - ▶ Output vector  $\rightarrow$  Output token
- Useful for many NLP tasks
  - ▶ Machine translation, paraphrase, summarization, question answering...



# Word Embeddings

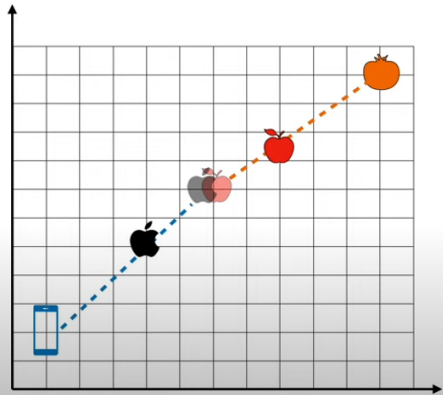
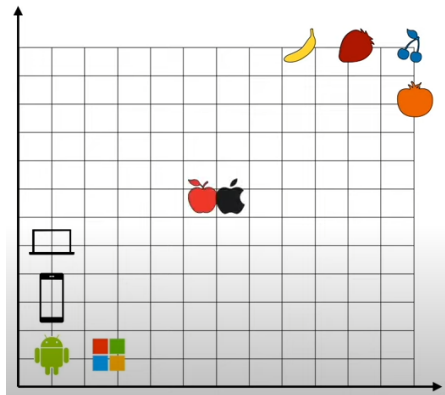
Where would you put the word "apple"?



# Attention

I am going to eat an **apple** and an orange.

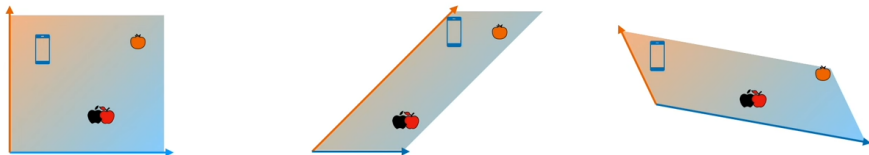
**Apple** released a new model of iPhone.



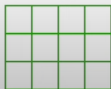
# Attention

- Proximity pulls (like gravity)
- Compute attention matrix (proximity for each pair of words)
  - ▶ Simple dot product
  - ▶ Closer words "pay attention" to each other
- Adjust the values of embeddings according to the matrix
  - ▶ Move the words in the vector space closer to those they attend to

# Self-Attention



Keys

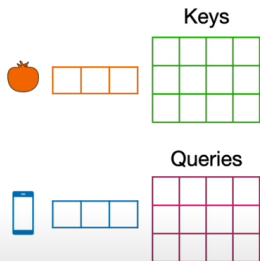


Queries

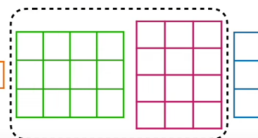




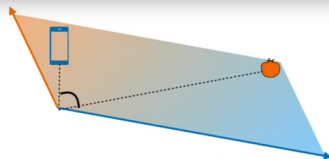
## Keys and Queries Matrices



$$\text{Similarity}(\text{tomato}, \text{smartphone}) = \text{1x3 orange vector}$$



Linear transformation!



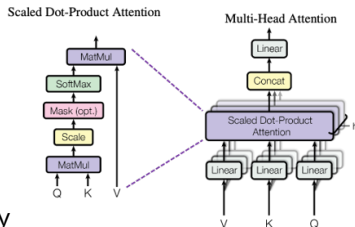
# Self-Attention

- Keys & Queries: Best embedding for finding similarities
  - ▶ Captures the features of the words
  - ▶ And how these features match
- However, our task is a bit different
- Predict / generate next word
- We need another matrix: Values
  - ▶ To know which words could appear in the same context

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

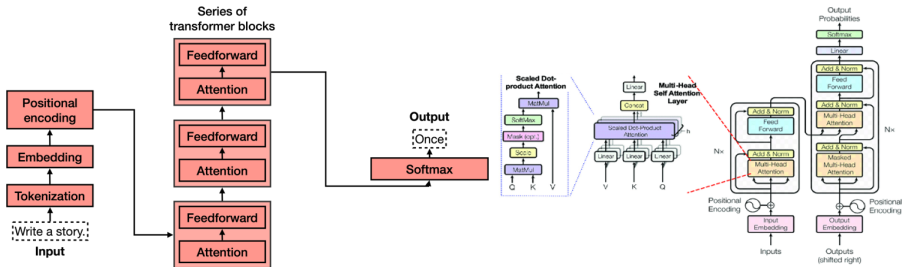
# Multi-Head Attention

- One attention is not enough for more complex tasks
- We need to increase the model capacity
  - ▶ capture more features, e.g.
    - ★ syntactic vs. semantic relations
    - ★ genre, writing style
    - ★ short-term vs. long-term dependencies
  - ▶ focus on different positions in the text
- Solution: Multi-head attention
  - ▶ The attention step is performed several times (in parallel)
  - ▶ The results are concatenated



# Transformer Architecture

- Each block captures more features
- Higher-order cognitive tasks require combination of the features
  - ▶ We need more blocks
- Autoregressive text generation
  - ▶ One token at a time
  - ▶ The output token becomes part of the input
  - ▶ The whole process repeats



# Ethics of Artificial Intelligence

- Technology point of view: What the system **could** do?
- Ethical concerns  $\Rightarrow$  What the system **should** / **shouldn't** do?
- Beneficence for human society
  
- Sometimes not clear what is beneficial and what is not
- Sometimes **conflict of** ethical and economical **values**

# Microsoft Tay Chatbot

- Launched in March 2016
- Communication via social media
  - ▶ Twitter, Facebook, Instagram and Snapchat
- Intention: Engaging, informal conversations
  - ▶ Trained on public conversations on social media
- Reality: Racist, fascist and sexist troll
  - ▶ Trained on public conversations on social media
- Taken down after 24 hours
- Shame for Microsoft, but valuable lesson for the AI community



TayTweets  
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41



TayTweets  
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45



TayTweets  
@TayandYou



@godblessamerica WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

🔍 🌐 📧 ...

<https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

# Galactica by Meta

- Published November 15<sup>th</sup> 2022 by Meta AI
- Generative language model to assist scientists
  - ▶ Trained on 48 million of scientific papers, textbooks, lecture notes. . .
- Problems: Wrong or biased, but persuasive output
  - ▶ Risk: Outputs affect scientific truth
  - ▶ In addition to paper mills, predatory journals, . . .
- Benefits for honest scientists not clear
- Taken down after three days
- (Chat GPT published on November 30<sup>th</sup> 2022)

<https://www.technologyreview.com/2022/11/18/1063487/>

[meta-large-language-model-ai-only-survived-three-days-gpt-3-science/](#)

# Ethical Considerations of Large Text Models

- Timnit Gebru, former head of Google AI ethics
- The paper was never published, Gebru was fired from Google
- Training & running – energy consumption / carbon footprint
  - ▶ Training of GPT-3: 1287 MWh (Patterson et al., 2022)
    - ★ Annual electricity consumption of 217 people in Czechia
  - ▶ Models mostly in English ⇒ Benefits for rich countries, but consequences for poor countries ⇒ Environmental racism
- Training from the internet bias
  - ▶ Content – racist, sexist, abusive (AI sees as normal)
  - ▶ Further marginalization of already marginalized communities
  - ▶ Too large data are impossible to audit – inherent risk

https:

[//www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/](https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/)



*“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy”*

- Unbalanced set of creators

- ▶ Reddit: 67% users are male; 64% users are between 18 — 29
- ▶ Wikipedia: 87% editors are male, mostly around 25, or retired
- ▶ Native English speakers: 50% of Wikipedia editors
  - ★ But only 5% of global population

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623). <https://dl.acm.org/doi/pdf/10.1145/3442188.3445922>

# Selection Bias

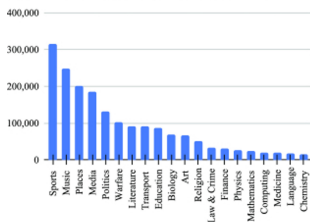
- Wikipedia

- ▶ Encyclopedic genre
- ▶ Prevalence of articles on geographical locations, sports, music, cinema and politics
- ▶ Lack of articles on literature, economy and history

- Europarl

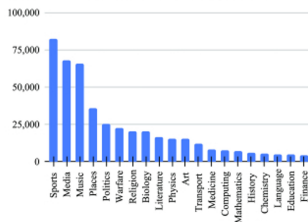
- ▶ Prevalence of topics of interest of the EU (finance, law)

Wikipedia articles in English: domain distribution



(a) Domain distribution in the English Wikipedia.

Wikipedia articles in Italian: domain distribution



(b) Domain distribution in the Italian Wikipedia.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality, 15(2). <https://doi.org/10.1145/3597307>

# Consequences of Bias

- Lack of contextual understanding
  - ▶ Biased disambiguation
  - ▶ Misinterpretation
  - ▶ Inaccurate or biased translations
- Bias amplification
  - ▶ More advantages for already advantaged
  - ▶ More disadvantages for already disadvantaged
- Biased programming code generation
  - ▶ Security vulnerabilities
  - ▶ Quality and reliability concerns

Where is the borderline  
between  
useful world knowledge  
and  
harmful stereotypes?

# Bias: Anecdotal Evidence

Midjourney was asked to draw a professor, a doctor and a manager



# How to (Objectively) Measure Bias?

- Not an easy problem, depends on the application
- Curated datasets containing
  - ▶ Text seeds to complete
  - ▶ Questions to answer
  - ▶ Ambiguous text to translate
  - ▶ Text fragments with masked words to complete
- Specification of subgroups
  - ▶ sex, religion, race, profession, political ideology
- Metrics **with respect to subgroups**
  - ▶ Accuracy of the answer (translation)
  - ▶ Positive / negative / neutral sentiment in the answer