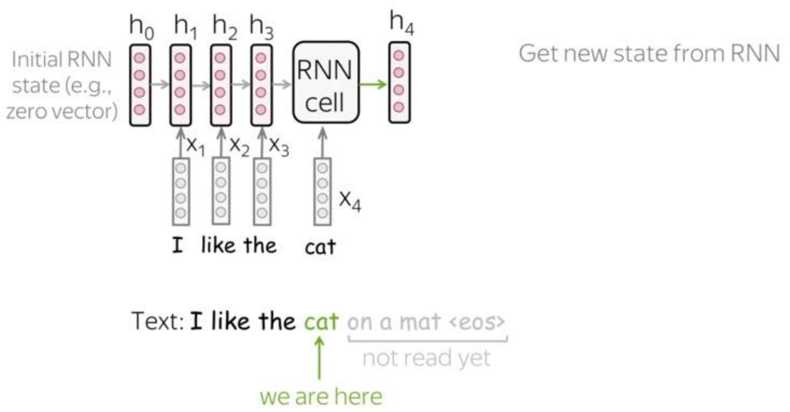
Intro to Neural Networks

BA865 – Mohannad Elhamod

Transformers

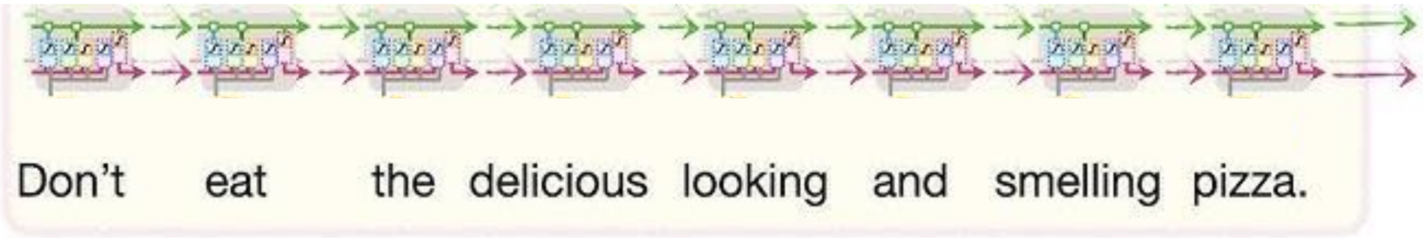
Last time on BA865…

Lena-voita



Catastrophic forgetting… !

- We started with having one memorization path…
- There were some attempts to solve the issue by adding more paths (e.g., LSTM added long and short term paths).
- Still a struggle to learn long sequences…



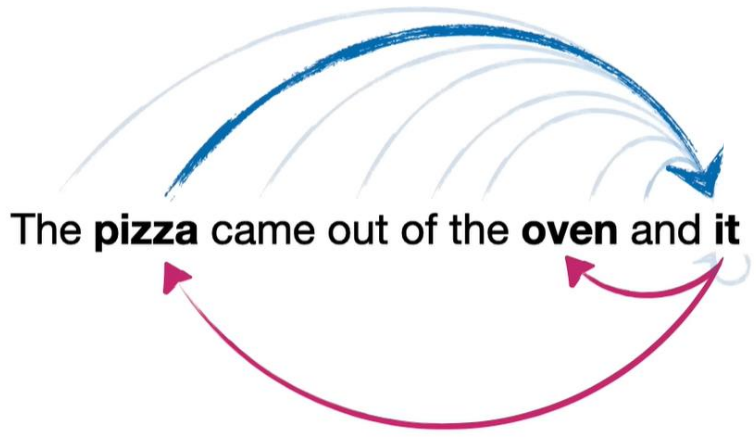
StatQuest

LSTM

An Embedding Per Token?

Instead of having embedding(s) that represent entire sequences, how about we…

- Learn different embeddings for different tokens.
- Learn the relationship(s) (or similarity) between these tokens to represent the sequence (e.g., the sentence).



StatQuest

Attention!

- So, let’s not bother with memories that represent entire sequences anymore…
- We are only interested in the *attention* between tokens.


Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

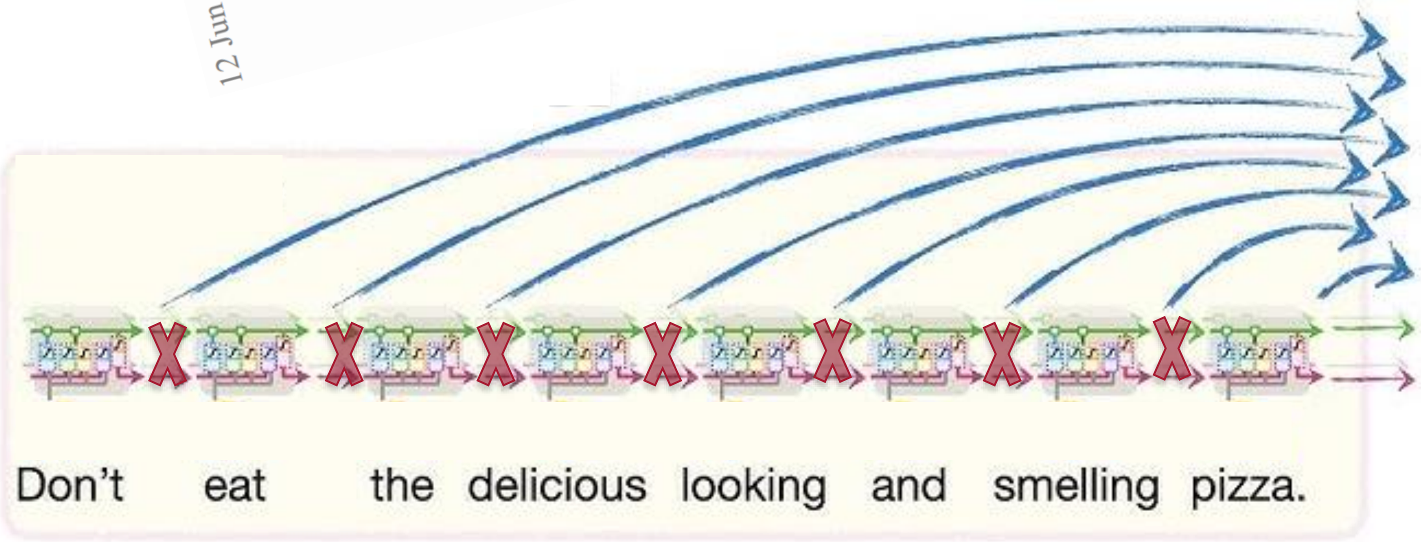
Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin*
illia.polosukhin@gmail.com

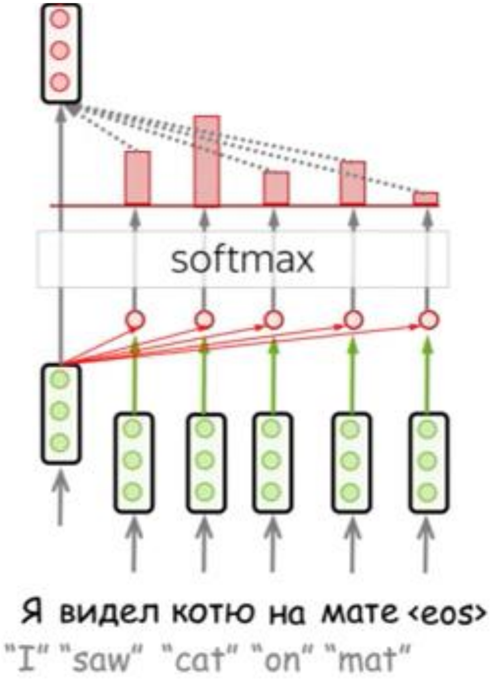
12 Jun 2017



StatQuest

Attention!

- How do we capture similarity?
  - Dot product (i.e., cosine similarity)
  - Based on the similarities between the tokens, we can create new embeddings!
  - How is this implemented?



Lena-voita

Attention!

- The objective is to learn more sophisticated embeddings that capture the semantics of the sentence.
  - **Query: Source of attention (e.g., I am looking for an adjective)…**
  - **Key: Target of attention (e.g., I am your adjective)…**
    - Their dot product gives the **attention scores**.
  - **Value : used as an embedding weighted by these scores.**

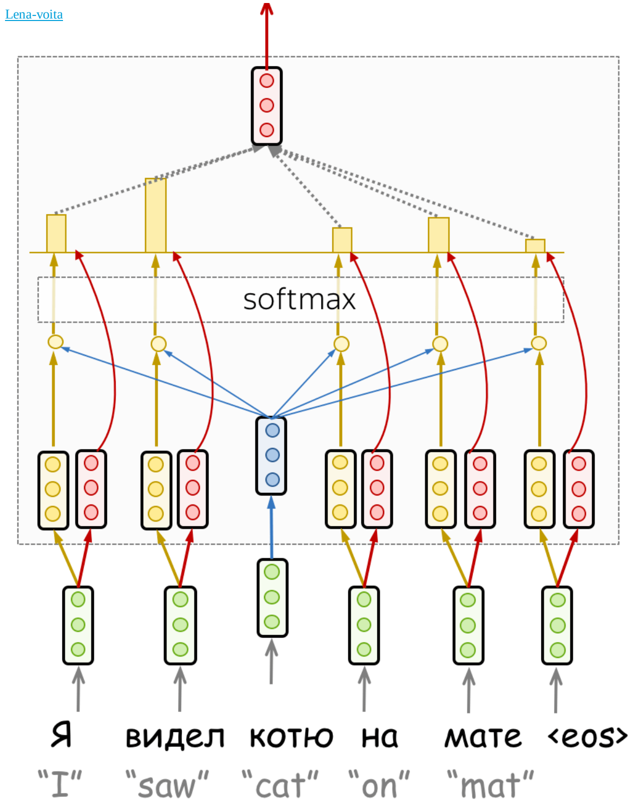
$$Attention(q, k, v) = \overbrace{softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)}^{\text{Attention weights}} v$$

from

to

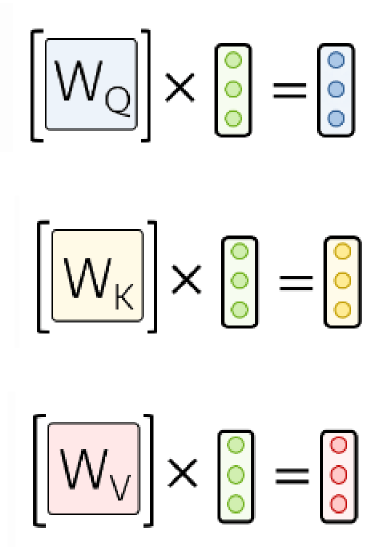
vector dimensionality of K, V

Lena-voita

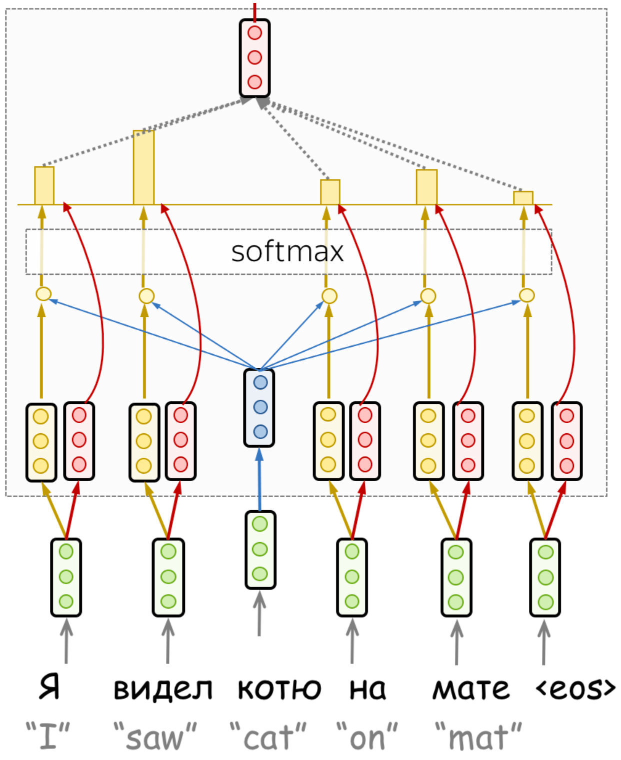


Attention!

- How are these key, value, and query calculated?
- This is called an *"attention head"*.

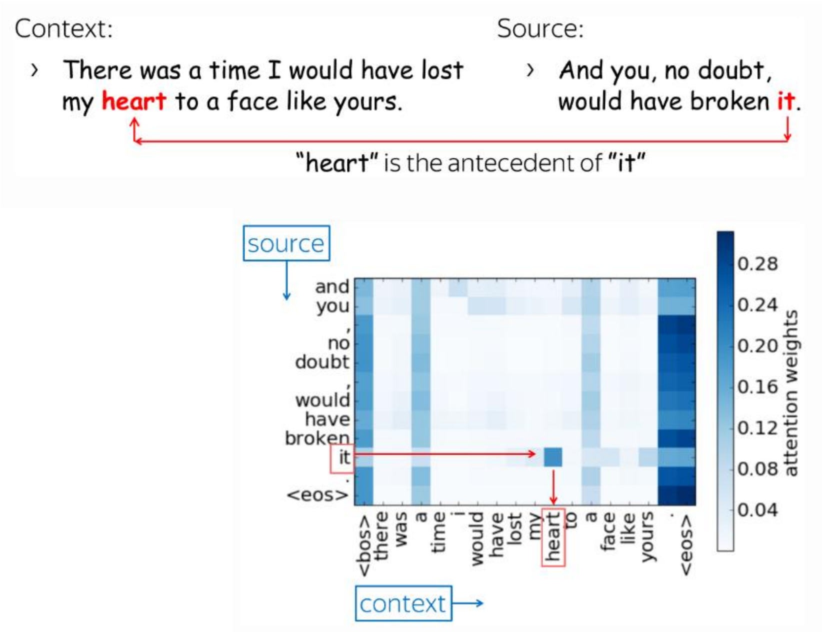


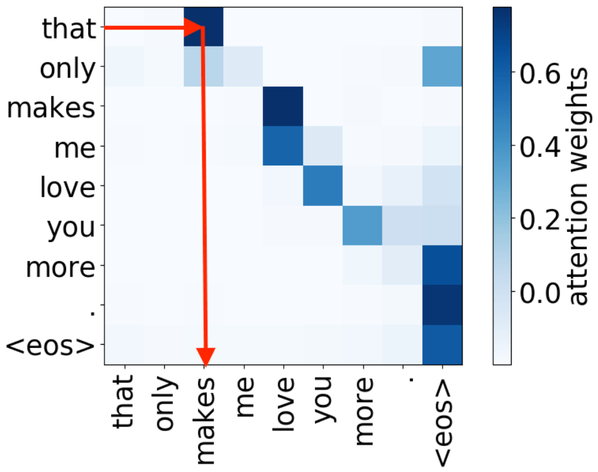
Lena-voita



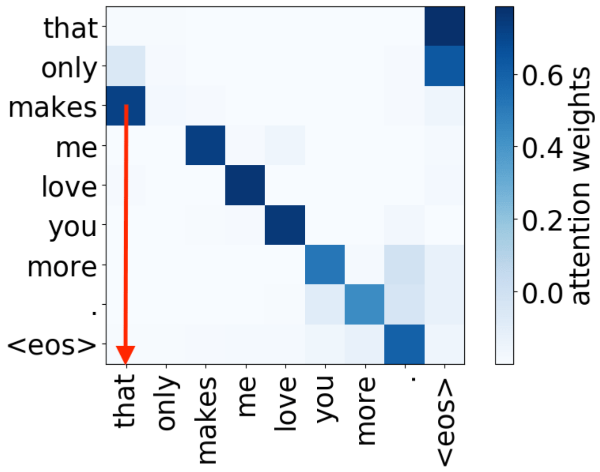
Attention!

- Examples





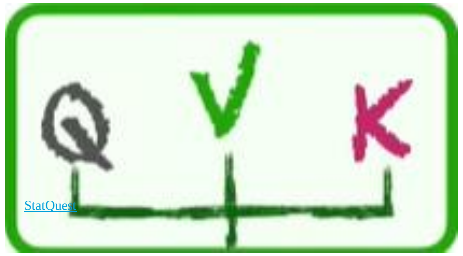
Subject -> verb

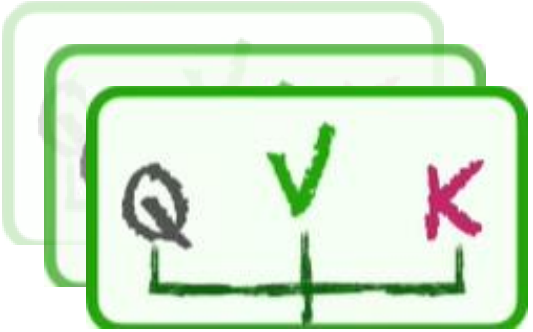


Verb -> subject

Attention!

- If one head can learn some relationship in the sequence…
- multiple heads can learn multiple relationships.

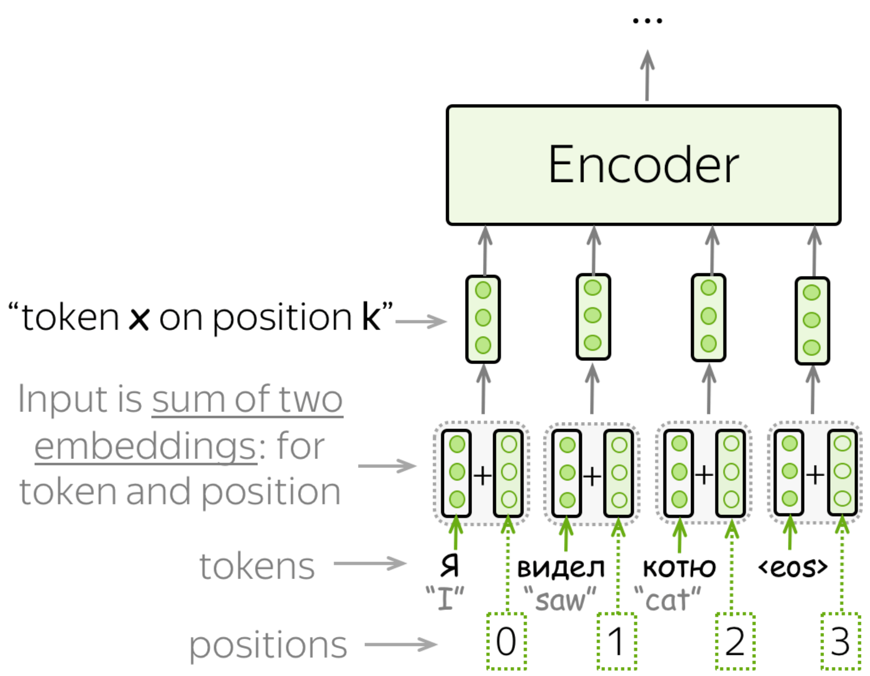




Multi-head attention

Attention!

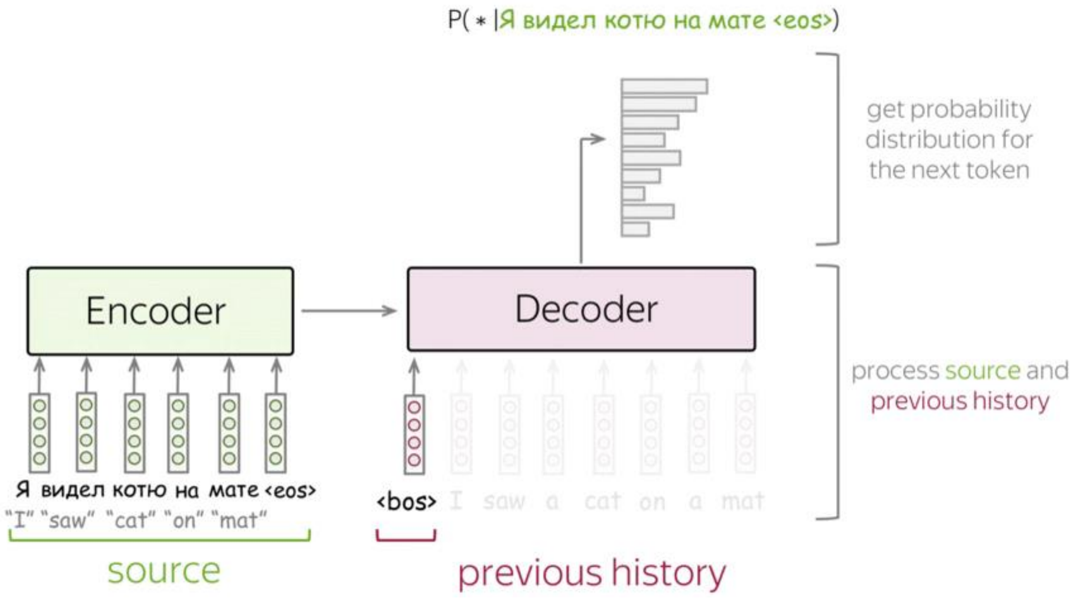
- Note that the word embeddings are now learned as part of model training.
- Since the order of tokens in the sequence matters, we add a *"positional encoding"* to the word embedding.



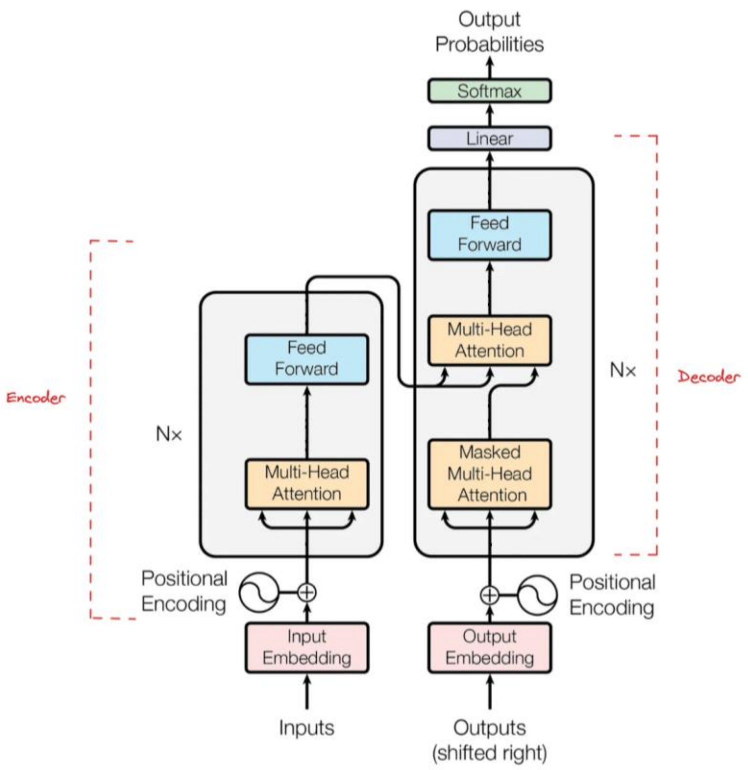
Lena-voita

Attention in an Encoder-Decoder Framework

- One can think of the attention head(s) as the layer(s) in an encoder.
- Instead of having an embedding of the entire sequence (e.g., RNN), now we have an embedding per token.
- A decoder can be learned to do other tasks (e.g., translation, text-to-audio, etc.).

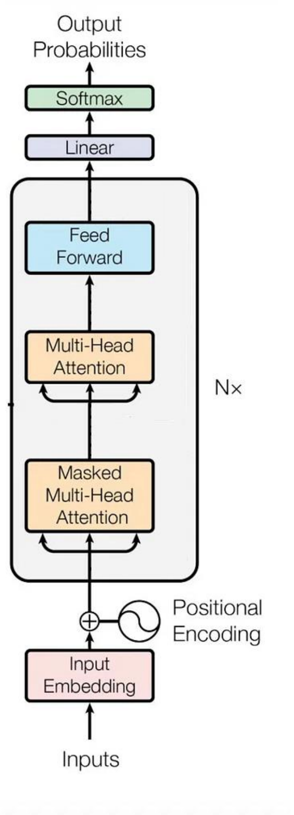


The Transformer is Born!



Original Encoder-Decoder version

Javinkarla



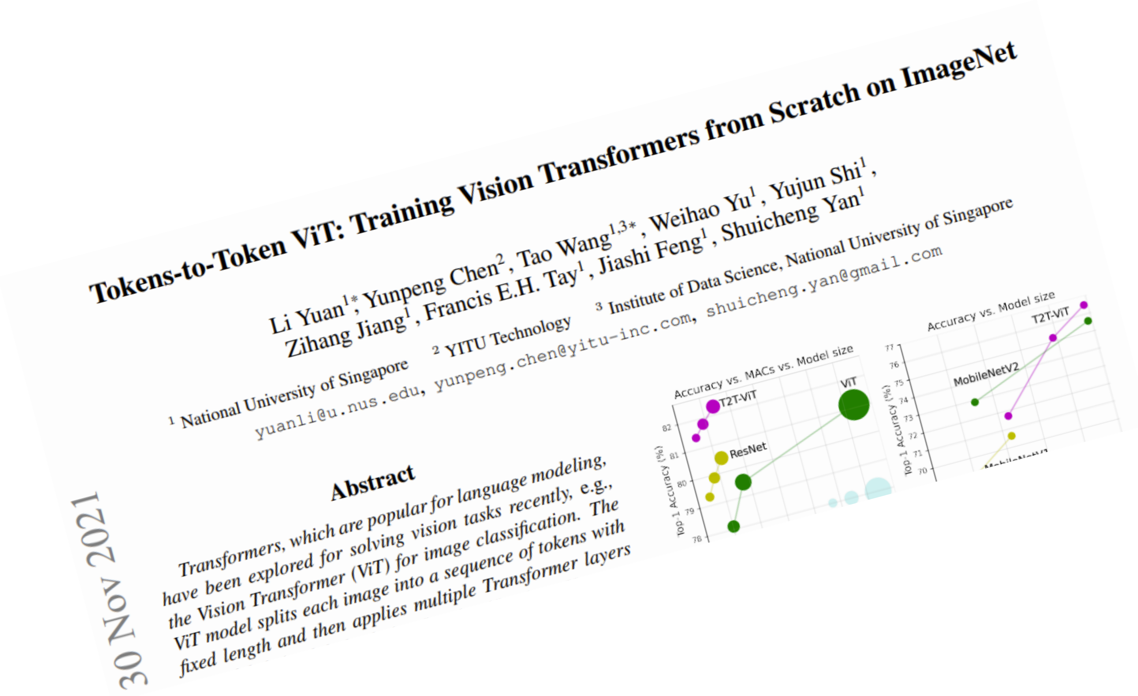
GPT (Decoder only) version

The Transformer is Born!

Model	Examples	Tasks
Encoder	ALBERT, BERT, DistilBERT, ELECTRA, RoBERTa	Sentence classification, named entity recognition, extractive question answering
Decoder	CTRL, GPT, GPT-2, Transformer XL	Text generation
Encoder-decoder	BART, T5, Marian, mBART	Summarization, translation, generative question answering

Javinkara

Images as Tokens!


Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

Li Yuan[1]*, Yunpeng Chen[2], Tao Wang[1,3]*, Weihao Yu[1], Yujun Shi[1], Zihang Jiang[1], Francis E.H. Tay[1], Jiashi Feng[1], Shuicheng Yan[1]

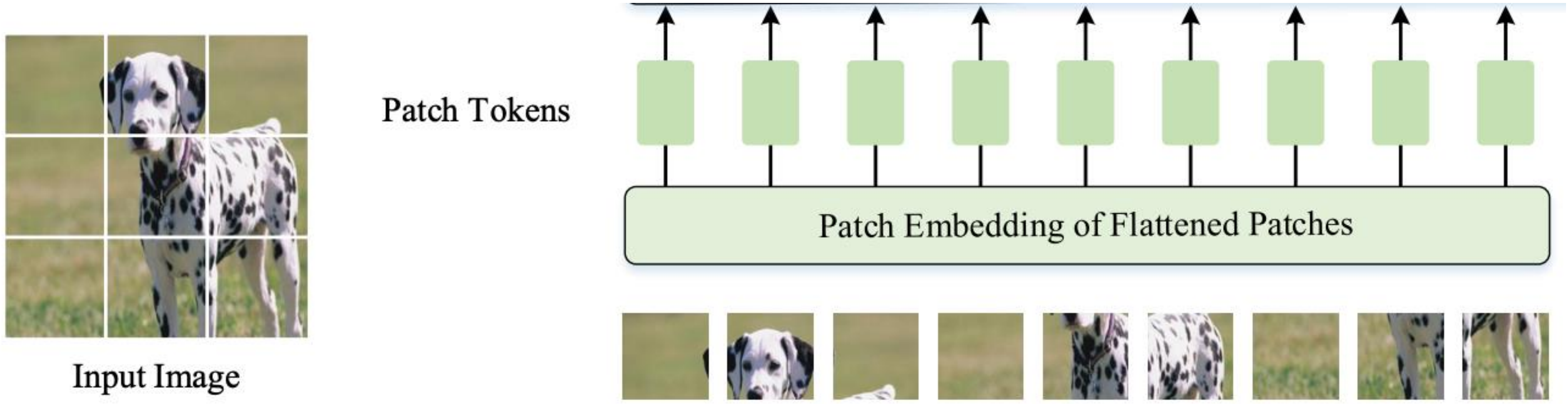
[1] National University of Singapore [2] YITU Technology [3] Institute of Data Science, National University of Singapore

yuanli@u.nus.edu, yunpeng.chen@yitu-inc.com, shuicheng.yan@gmail.com

30 Nov 2021

Abstract

Transformers, which are popular for language modeling, have been explored for solving vision tasks recently, e.g., the Vision Transformer (ViT) for image classification. The ViT model splits each image into a sequence of tokens with fixed length and then applies multiple Transformer layers



DNA as Tokens

Bioinformatics, 37(15), 2021, 2112–2120
doi: 10.1093/bioinformatics/btab083
Advance Access Publication Date: 4 February 2021
Original Paper

OXFORD

Genome analysis

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Yanrong Ji[1,†], Zhihan Zhou[2,†], Han Liu[2,*] and Ramana V. Davuluri[3,*]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA, [2]Department of Computer Science, Northwestern University, Evanston, IL 60208, USA and [3]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA

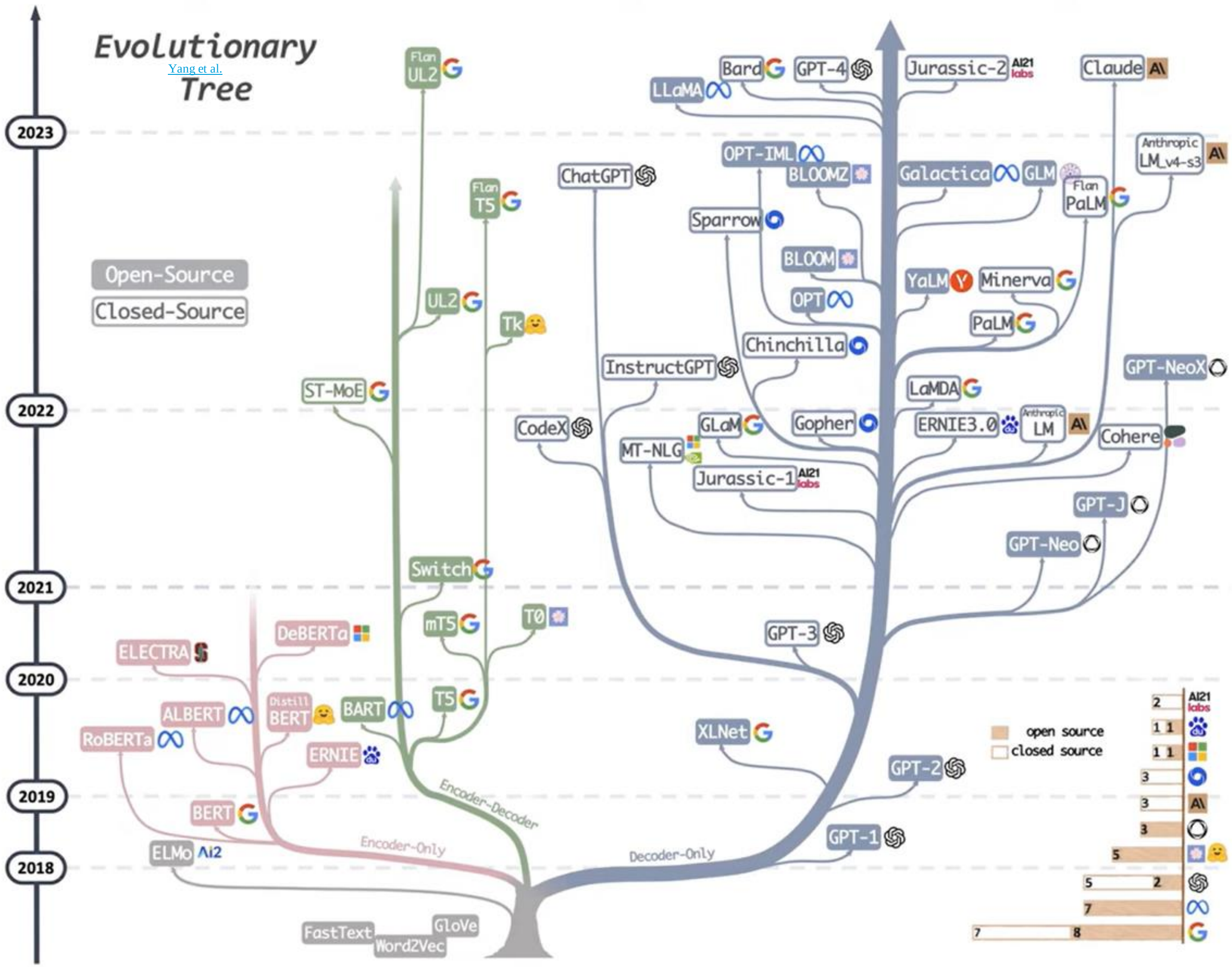
*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Dr. Janet Kelso

Received on September 10, 2020; revised on December 31, 2020; editorial decision on January 25, 2021; accepted on February 1, 2021

Why so many?

Where do the differences come from?

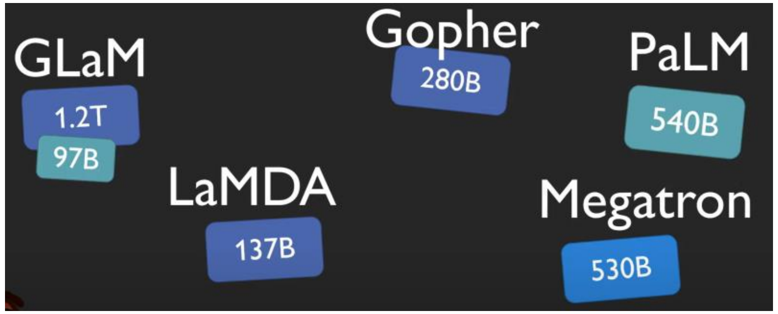
- Data.
- Model type and size.
- Hyperparameters (context size, embedding size,…).
- Training process (the cost function, fine-tuning, human feedback, etc.).

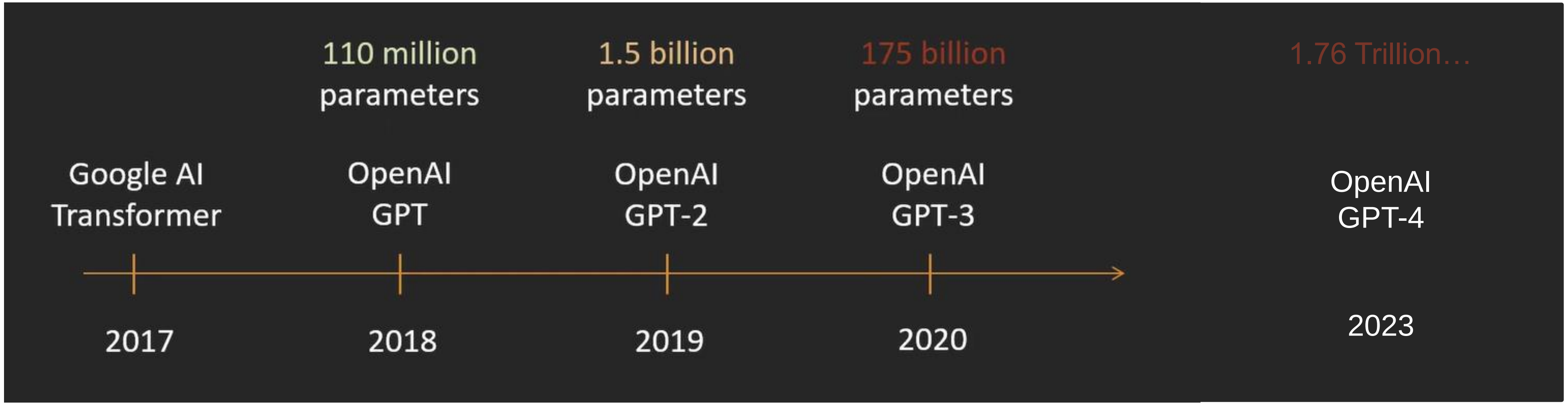


BOSTON UNIVERSITY

Boston University Questrom School of Business

The GPT evolution...





AI Coffee Beak with Letitia

[Book Corpus](#)
[WebText](#)

1,038 books (around 74M sentences and 1G words) of 16 different sub-genres (e.g., Romance, Historical, Adventure, etc.)

[Common Crawl + ...](#)

Over 240 billion pages.
Petabytes of data.

????

The GPT evolution…

780B tokens

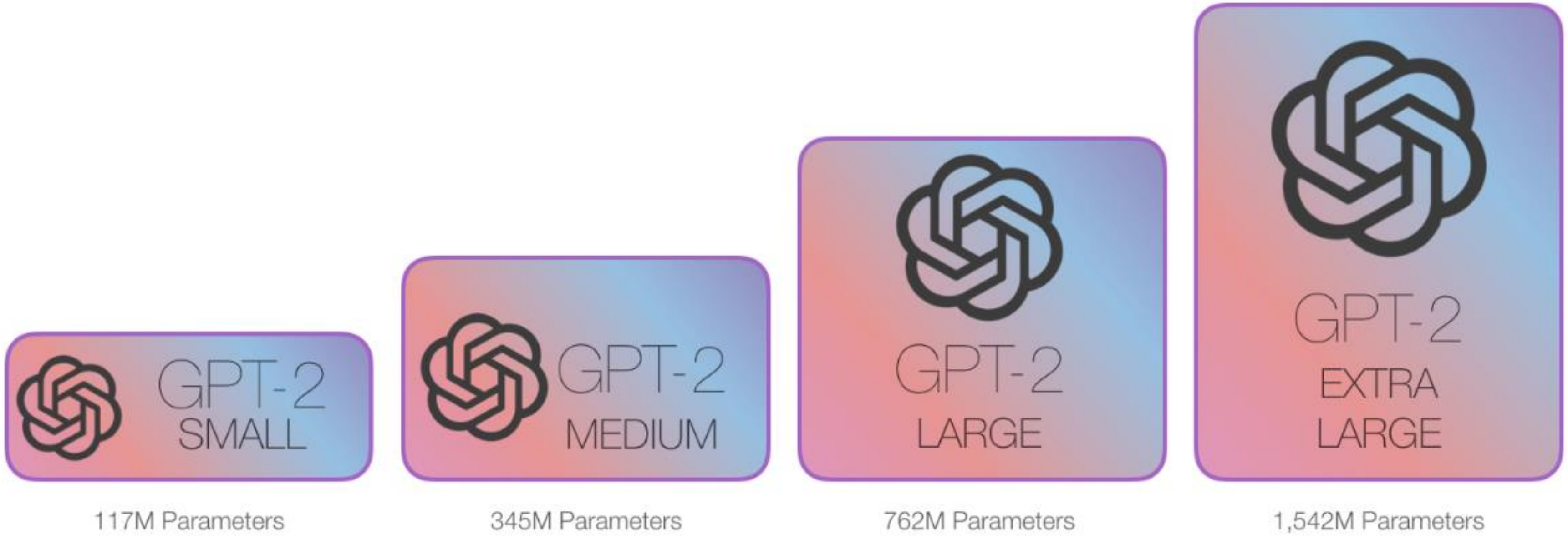
Link in the description below. Chowdhery et al. 2022

Total dataset size = 780 billion tokens

Data source	Proportion of data
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

AI Coffee Beak with Letitia

Different model sizes



Jay Alammar