# Beyond the hype: An introduction to NLP and language models

Adriano Belisario

Ajai Sreevatsan



### **Overview**

- Numerical representations of language
- Transformer-based model and LLMs
- Hands-on practice: transcription with Whisper

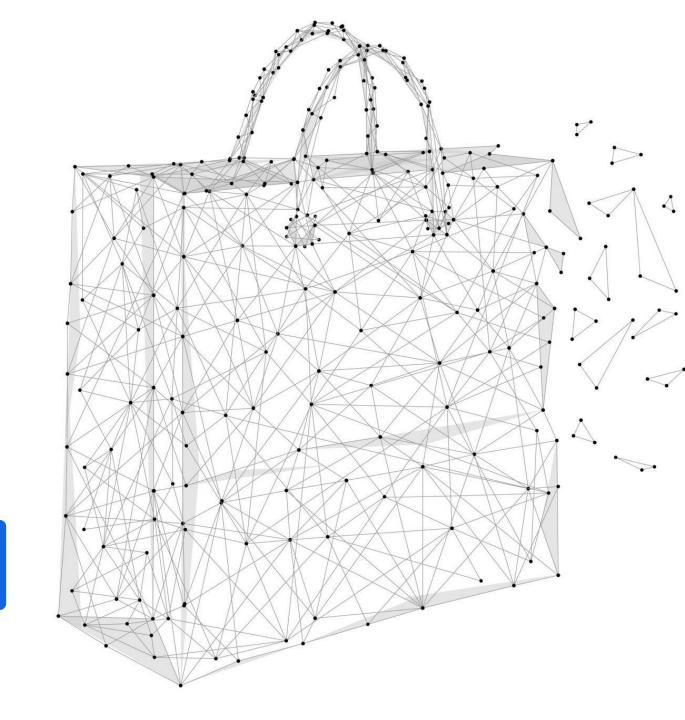


# From words to numbers

## Introduction

- What is NLP? The broader context
- Language's links to set theory and Boolean algebra
- The big divide within AI in the 1960s Symbolic: rule-based, logical engines Stochastic: learning through examples

Bag of words
(n-grams)



commonly used words in English

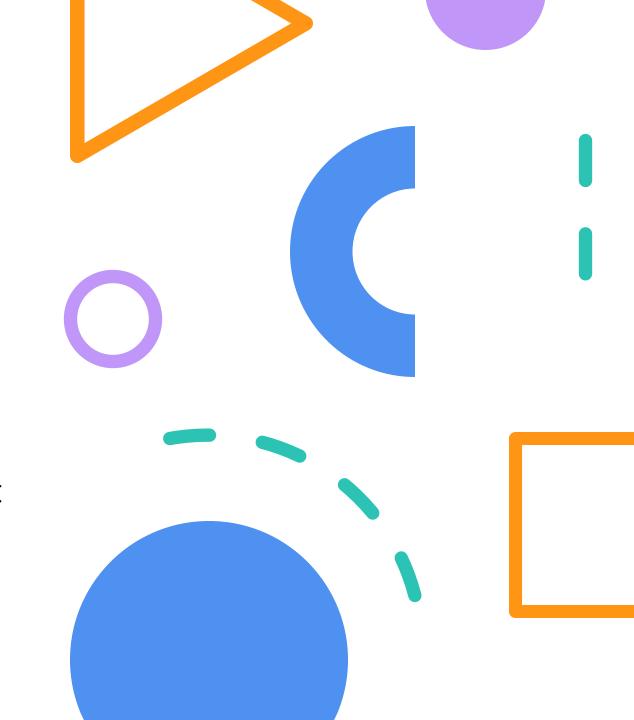
The number of possible 2-grams: 1.6 billion

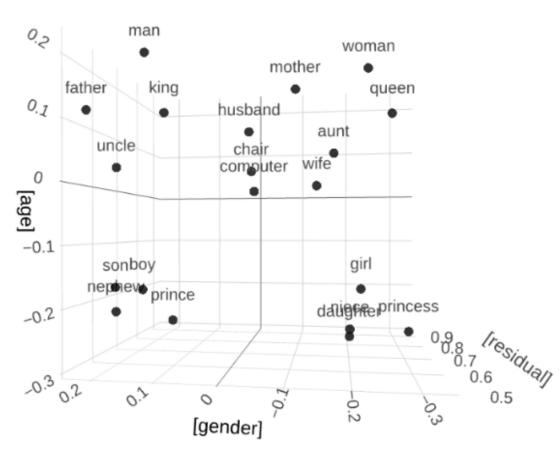
The number of possible 3-grams: 60 trillion

By the time we get to "essay fragments" of 20 words, the number of possibilities is larger than the number of particles in the universe

# Words to vectors

One can think of **word embeddings** as a way to
represent the "essence" of
something by an array of
numbers—with the property that
"nearby things" are represented
by nearby numbers.





# Linguistic feature space

# From attention to transformers

#### 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

#### Abstract

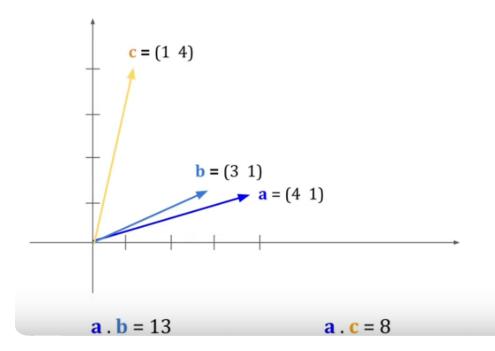
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 Englishto-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to

## **Attention**

- Attention allows a better understanding of **context**, determining the importance of each part of a sequence.
- It comes in different flavours, like self-attention or multiheaded attention

• Language games (Wittgenstein): "the meaning of a word is its use in the language"

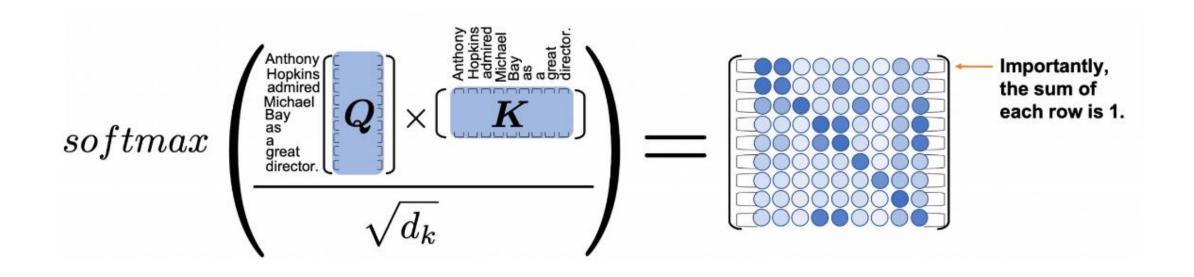
#### **Vector dot product**



# Calculating attention \

 Vectors that are closer in lexical space have a higher dot product

 Attention is based on 'proximity' in vector space

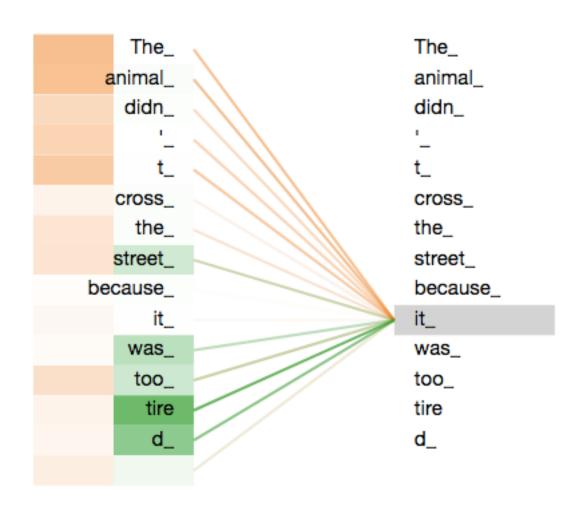


# The attention matrix

Query and key are both the same input sentence. Each word's attention score is calculated against every other word.

https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/

# Attention with text



https://medium.com/@YanAlx/step-by-step-into-transformer-79531eb2bb84

## Transformer-based models

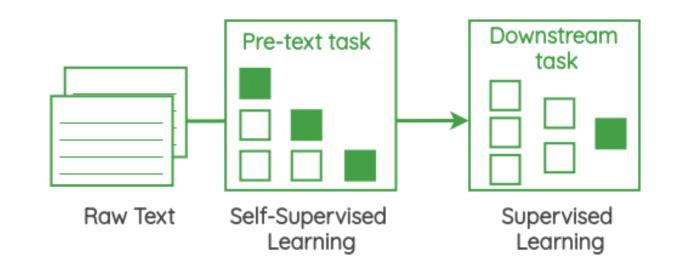
- Sequence-to-sequence: transformers and attention were initially implemented for translation;
- Transformer-based models use positional encoding and attention mechanism;
- Efficient in handling long-range sequences (parallelization);
- Famous use cases: BERT and GPT;

# Large Language Models (LLMs)

# **Training an LLM**

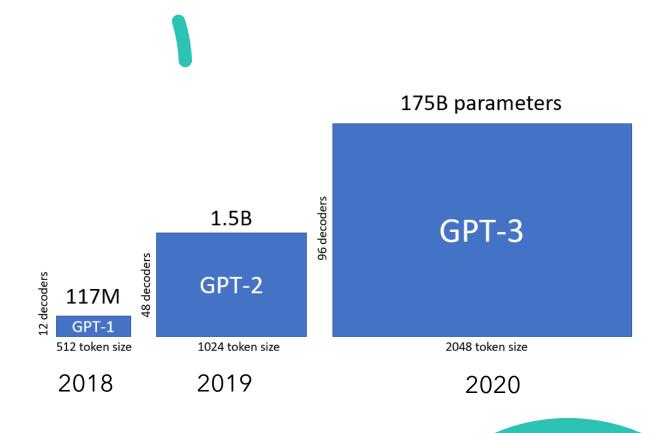
**Self-training**: mask the input sequences and try to fill the gaps.

**Fine-tuning:** use a ground-truth or human-in-the-loop to accomplish a specific task



# **Emergent behaviours**

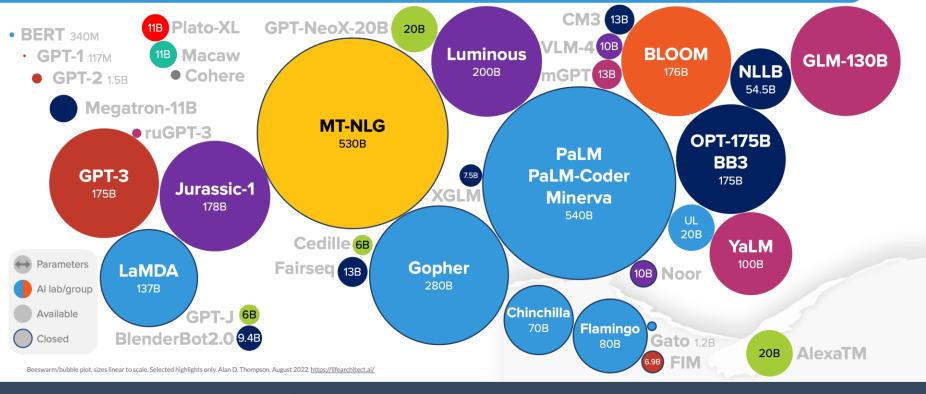
"GPT-3's capacity for **few-shot learning** on practical tasks
appears to have been discovered
only after it was trained, and its
capacity for **chain-of-thought**reasoning was discovered only
several months after it was
broadly deployed to the public"



Sources: '<u>Eight Things to Know about Large Language Models</u>' and '<u>Step by Step into GPT</u>'

# **Look beyond GPT**

#### LANGUAGE MODEL SIZES TO AUG/2022



S LifeArchitect.ai/models

# Understand what they are

LLMs are **statistical models** for words (i.e. good for summarisation, translation, text generation with well-structured data, code assistants, etc). They can be augmented with images or other types of data (multimodal LLMs, like <u>PaLM-E</u>).

They might encode some world representations but are NOT an artificial general intelligence - let alone autonomous agents - (i.e. not suitable as a primary source in research).

# Don't anthropomorphize

#### Input

Don't you love the feeling of sun on your skin?

#### Original response

As an Al language model, I don't have a physical form or the ability to experience sensations like the feel of the sun on my skin. I don't have personal experiences or emotions. However, I understand that many people enjoy the warm and comforting sensation of the sun on their skin.

#### De-anthropomorphised response

This generative language model does not have a physical form or the ability to experience sensations like the feel of the sun on a person's skin. This model does not have personal experiences or emotions. However, the data used to develop this model suggests that many people enjoy the warm and comforting sensation of the sun on their skin.

#### Figure 1: De-anthropomorphising system output.

Source: Mirages: On Anthropomorphism in Dialogue Systems (Abercrombie et al, 2023)

#### Risks in media coverage

- Misrepresentation of abilities;
- The myth of neutrality;

#### **Accountability**

- Look for the humans behind the models;
- The role of companies and developers;

# Showcase of tools



# Transcription with Whisper

Speech-to-text Transformer-based model.

Low-code hands-on exercise:

https://bit.ly/whisper-notebook

You'll need a Google account to run it on Google Colab, but you can also download the notebook and execute it locally.

## References

Code: <a href="https://github.com/belisards/nlp\_intro">https://github.com/belisards/nlp\_intro</a>

Reporting on artificial intelligence: a handbook for journalism educators (UNESCO):

https://unesdoc.unesco.org/ark:/48223/pf0000384551

#### Extra resources

<u>GPT4All</u>: has a GUI similar to ChatGPT and allows querying LLMs offline, but significantly slower than online versions (open-source)

<u>Chat PDF</u> or <u>AIExcelBot</u>: No need for programming skills; chat about the content in an uploaded PDF or generate Excel formulas on the fly (freemium)

Langchain: flexible but require programming skills (open-source)

Huggingface: like GitHub but for machine learning models; also requires programming skills to implement them (freemium)

# Thank you for your attention;)

Adriano Belisario

@belisards

adrianobf@gmail.com

Ajai Sreevatsan

@ajai.with.an.i

ajax5687@gmail.com