healthware international

Communicators
Connectors
Builders of Future Health

Transformers

A Deep Learning Revolution

NLP Gentle Introduction

healthware international

NLP Gentle Introduction

- Natural language processing (NLP) is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis. The purpose of these techniques is to achieve human-like language processing for a range of tasks or applications.
- Although it has gained enormous interest in recent years, research in NLP has been going on for several decades dating back to the late 1940s. This review divides its history into two main periods: NLP before and during the deep learning era.

NLP History

1950-90
Symbolic NLP

Rule based emulation of Natural Language Understanding and Generation

1990-2010
Statistical NLP

Statistical based method like bag-of-words, and n-grams

2010-now
Neural NLP

The Deep Learning revolution

2013
Word Embedding

Words with similar meaning have similar representation

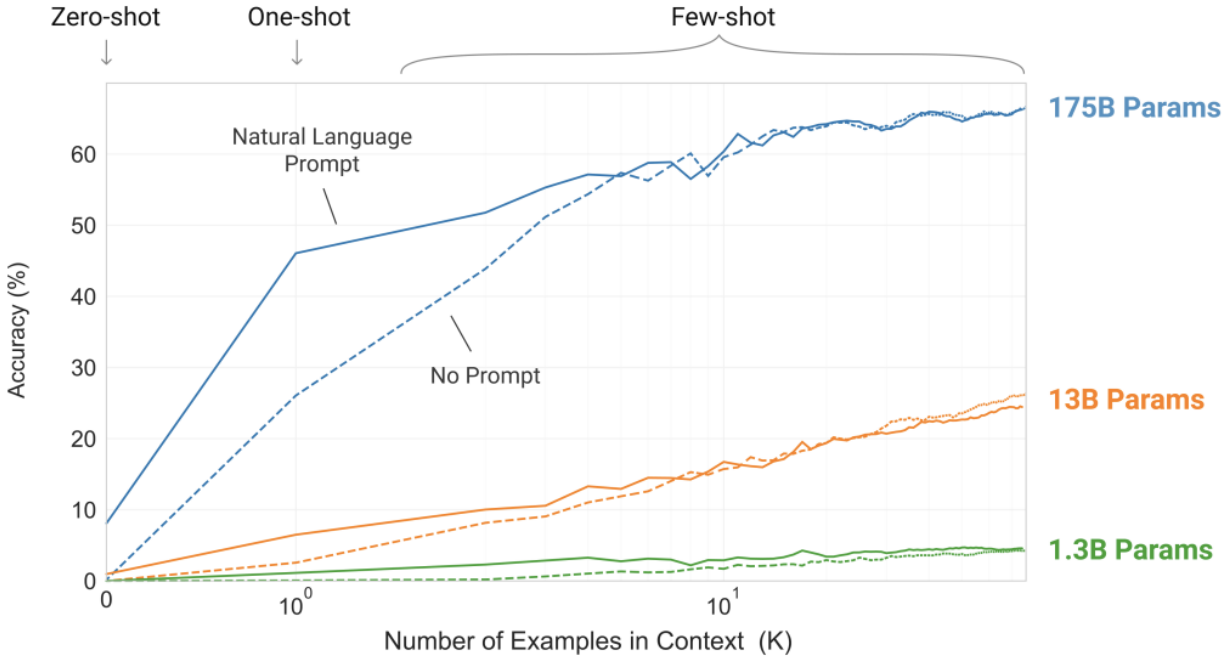
2014
Seq-2-Seq

Encoder-Decoder architectures start being impemented

2018-now
Transformes

Extremely large and powerful models

Powerful of Deep Learning

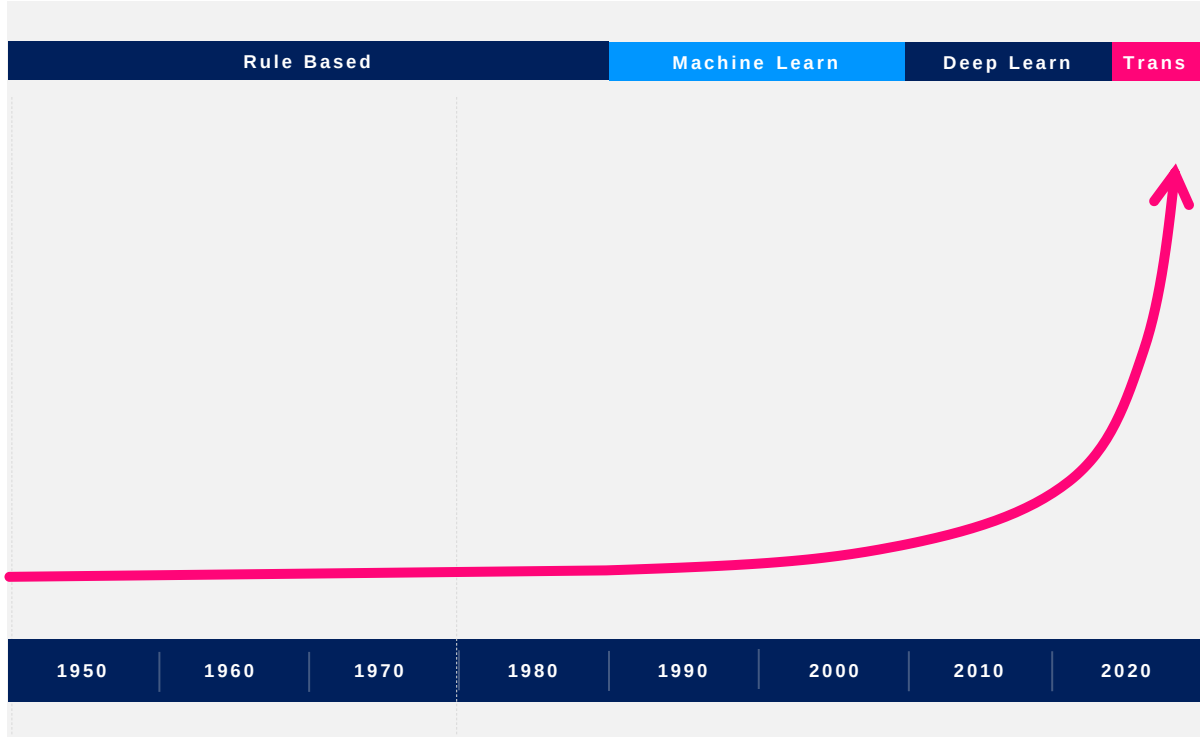


Recent studies have demonstrated the powerful of models with a very high number of parameters.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. *Language models are few-shot learners*. In Advances in Neural Information Processing Systems (NeurIPS).

healthware international

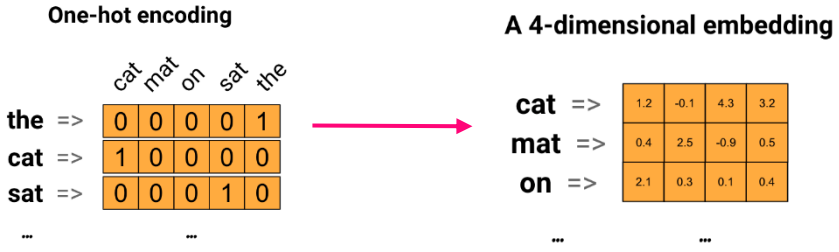
NLP powerful evolution

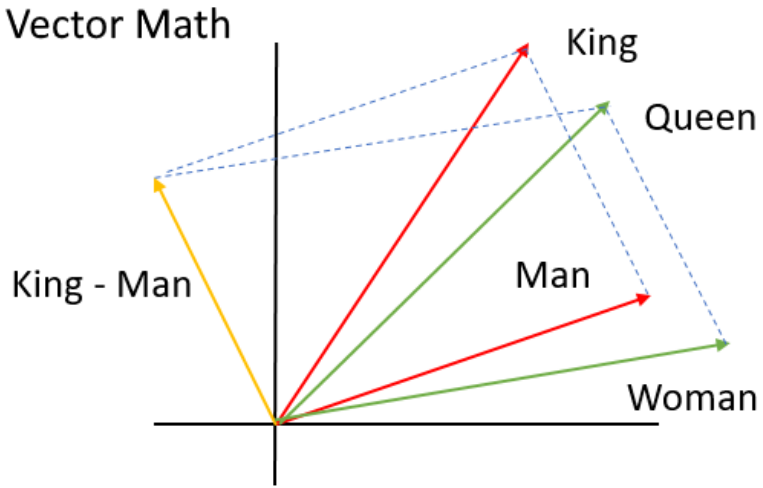


- We can experience the powerful evolution of NLP techniques every day.
- In recent years, our daily experience of interacting with machines has changed: we can talk on cell phones, converse with chatbots, translate texts, and in general, we can interact by speaking our natural language that will be interpreted and understood.

healthware international

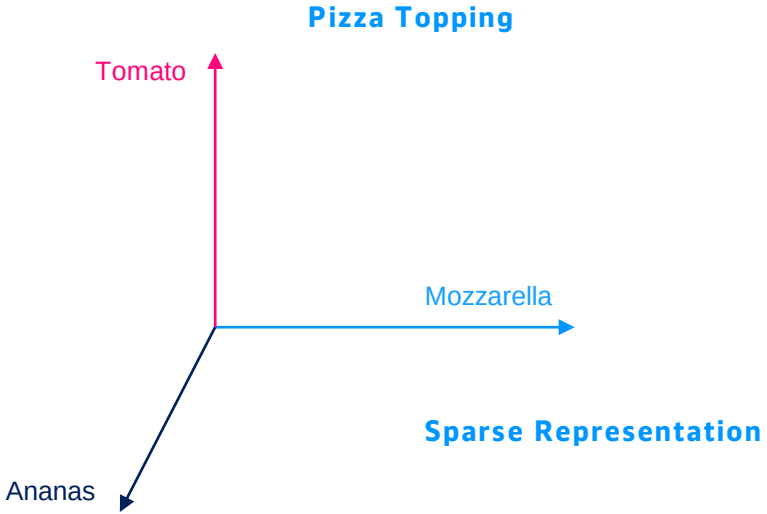
Word Embeddings

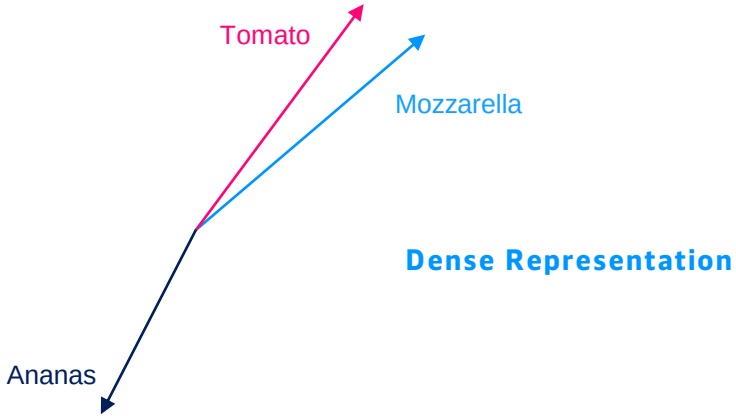




Representing Text as numbers

- Classic NLP techniques for encoding text, such as OneHot Encoding or dictionary creation, have two main problems:
  - **Arbitrary coding** (does not take relationships into account)
  - **Sparseness**, thus difficult interpretability for models
- In 2013, Word2vec, a set of templates that are used to produce word embedding, was developed at Google.
- Word2vec is a two-layer ANN designed to process natural language; the algorithm requires a corpus as input and returns a set of vectors representing the semantic distribution of words in the text.





What are Embeddings?

- *An embedding is a relatively low-dimensional space into which you can translate high-dimensional vectors.*
- Embeddings make it easier to do machine learning on large inputs like **sparse vectors representing words.**
- Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. An embedding can be learned and reused across models.

Word2vec: pros & cons

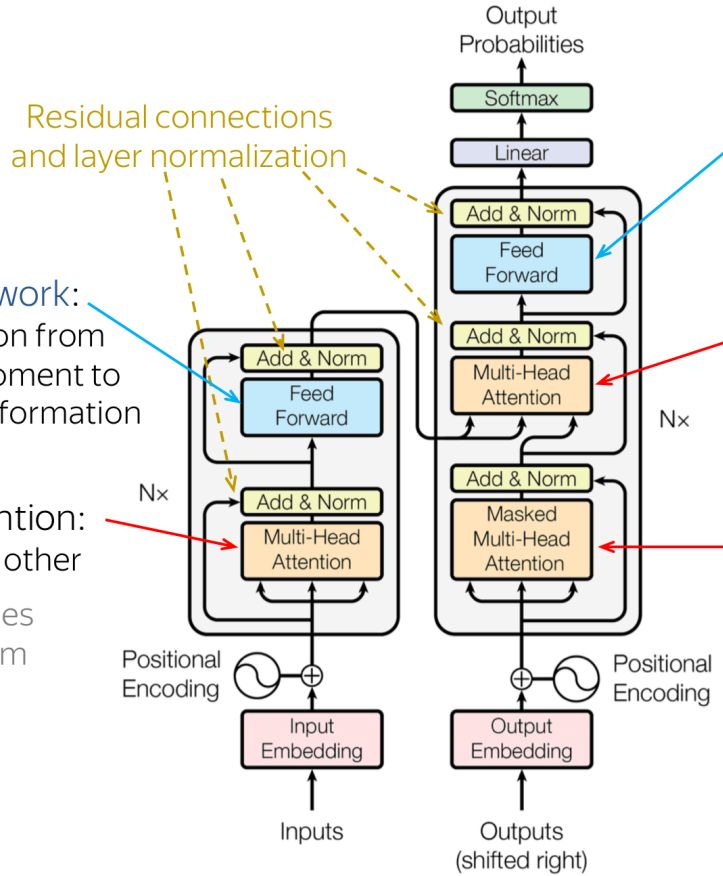
- Pros
  - The idea is very intuitive
  - The data can be fed into the model in an online way and needs little preprocessing, thus requires little memory.
  - You can perform algebraic operations on the vectors
- Cons
  - Require many examples to give generalizable, quality results
  - The sub-linear relationships are not explicitly defined. There is little theoretical support behind such characteristic.
  - How to separate some opposite word pairs.

Transformers

Deep Learning Revolution

- Transformers were introduced in 2017 and revolutionized deep learning.
- Some pros:
  - processing of long input sequence
  - high computational parallelization
  - do not suffer from the vanishing (or exploding) gradient problem
  - speed in the training phase

Transformers: Architecture



Residual connections and layer normalization

Feed-forward network:
after taking information from other tokens, take a moment to think and process this information

Encoder self-attention:
tokens look at each other

queries, keys, values are computed from encoder states

Feed-forward network:
after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention:
target token looks at the source

queries – from decoder states; keys and values from encoder states

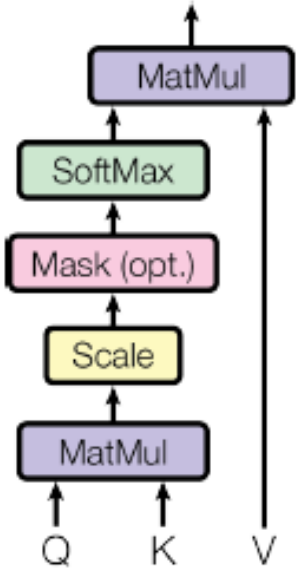
Decoder self-attention (masked):
tokens look at the previous tokens

queries, keys, values are computed from decoder states

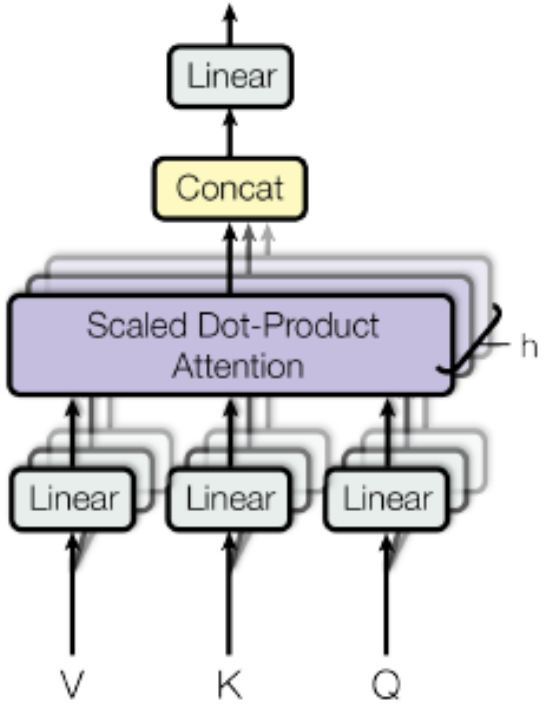
healthware international

Transformers: Attention

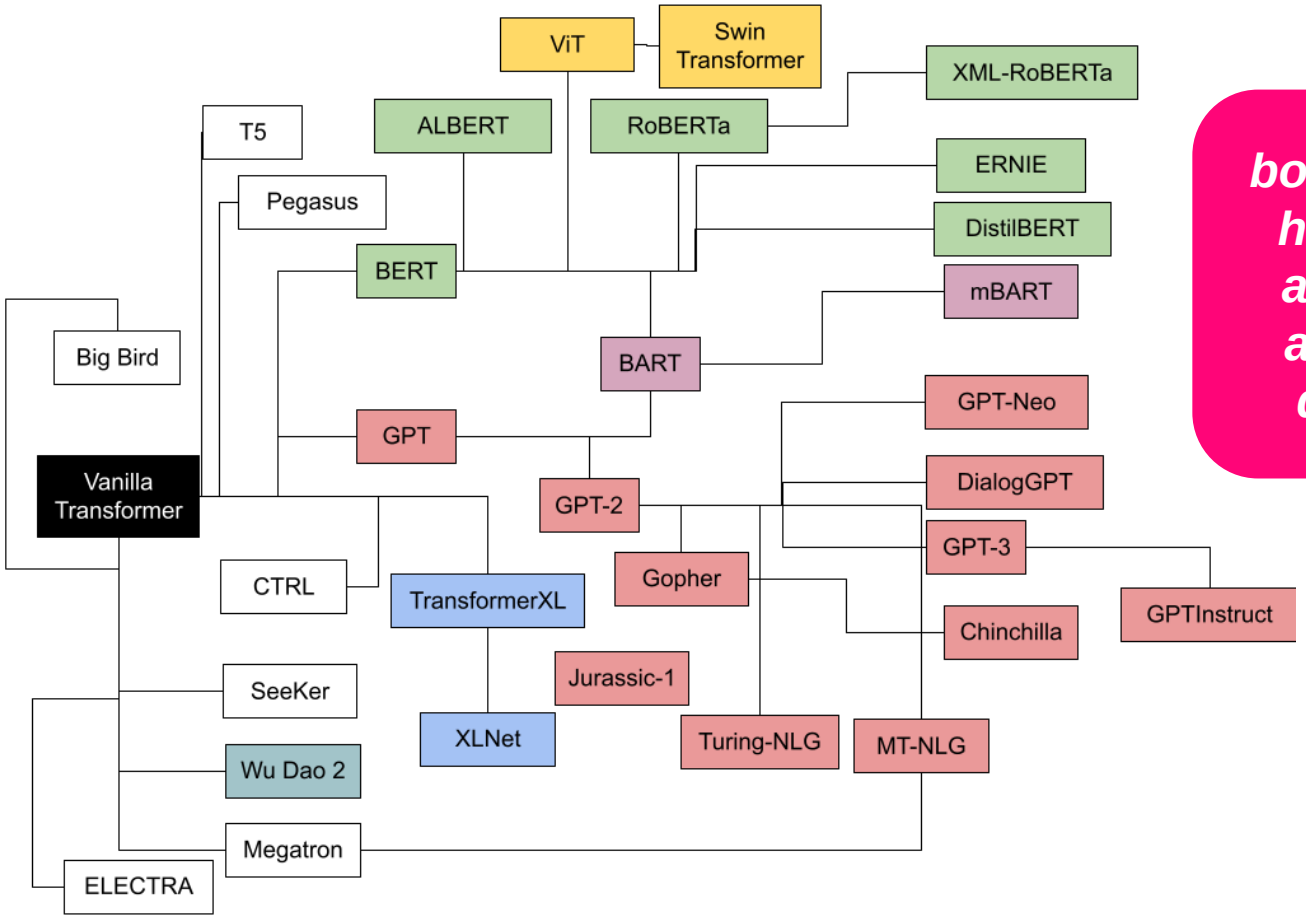
Scaled Dot-Product Attention



Multi-Head Attention



Transformers: Family Tree



born for NLP have been applied to almost all domains

healthware international

Hugging Face

Large Models Availability

Journal of Machine Learning Research 23 (2022) 1-40 Submitted 8/21; Revised 3/22; Published 4/22

Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity

William Fedus*
LIAMFEDUS@GOOGLE.COM
Barret Zoph*
BARRETZOPH@GOOGLE.COM
Noam Shazeer
NOAM@GOOGLE.COM
Google, Mountain View, CA 94043, USA

NEWS

Google Open-Sources Trillion-Parameter AI Language Model Switch Transformer

Anthony Alford

AI, ML AND DATA ENGINEERING

InfoQ

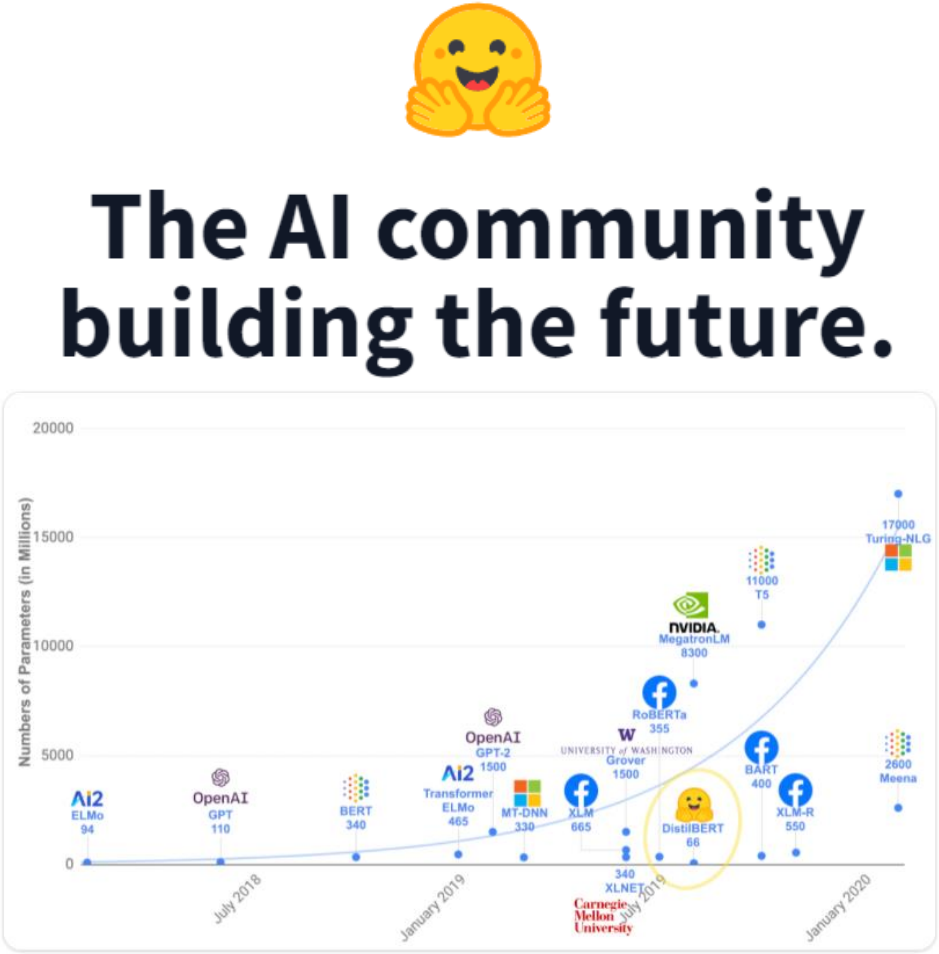
- Some months ago Google releases in open source a **trillion** parameter language model
- Training such vast models requires enormous computational resources: probably millions of dollars in electricity, and the availability of thousands and thousands of CPU/GPUs
- There is a sort of **competition** among Google, Microsoft, Amazon, OpenAI, etc. to prove they have the best resources for NLP

healthware international

Hugging Face

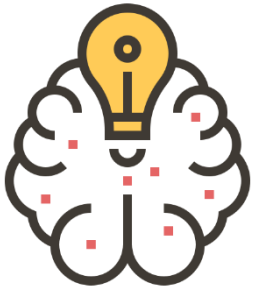
The AI community building the future

- **Hugging Face** is an open-source and platform provider of machine learning technologies.
- It was launched in 2016 and is headquartered in New York City.
- **Hugging Face** allows users to build, train, and deploy art models using the reference open source in machine learning.
- It aims to be the GitHub of machine learning
- It makes available the state of the art of large language models



healthware international

Limit: Efficiency



- 90 x 10^9 neurons firing 10^3 time/s each 10^4 connections
- 2 x 10^9 Mflops (ops/s)
- Parallel
- **Energy: 20 watt**

- 10^9 operations/s
  5 x 10^9 transistors/cpu
- 8 x 10^9 Mflops (ops/s)
- Serial
- **Energy: 2.5 x 10^7watt**

Thank You!

@healthware_intl | @healthware_ita
linkedin.com/company/healthwaregroup/
facebook.com/healthwareintl
instagram.com/healthware_intl

www.healthwaregroup.com

healthware international

Communicators
Connectors
Builders of Future Health