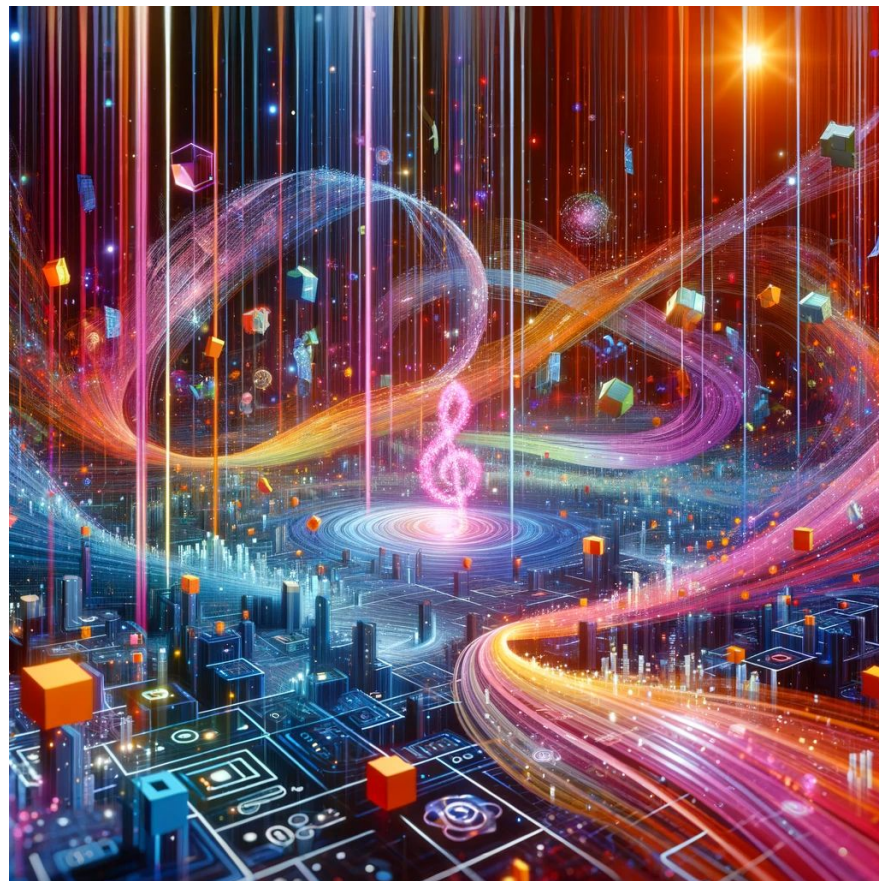


SPRING 2024
DSE I2450 - 3GG
BIG Data &
Scalable Computation

Week 7
AI Pipelines + the Cloud




Dall-E 3: "inside a big data system, use your imagination"

plan for today!

- housekeeping, logistics
- lecture + demo: Spark ML, Transformers
- (break)
- group work
- back together, until 7:20 🖐️



housekeeping:

-  **grades + feedback** by end of this week
- **next week: data center tour!**
 - 4pm, 395 Hudson St. (TriBeCa, near Canal/Spring)
 - Get there early if you can, ~3:50pm for front desk
 - After we meet @ 5th floor

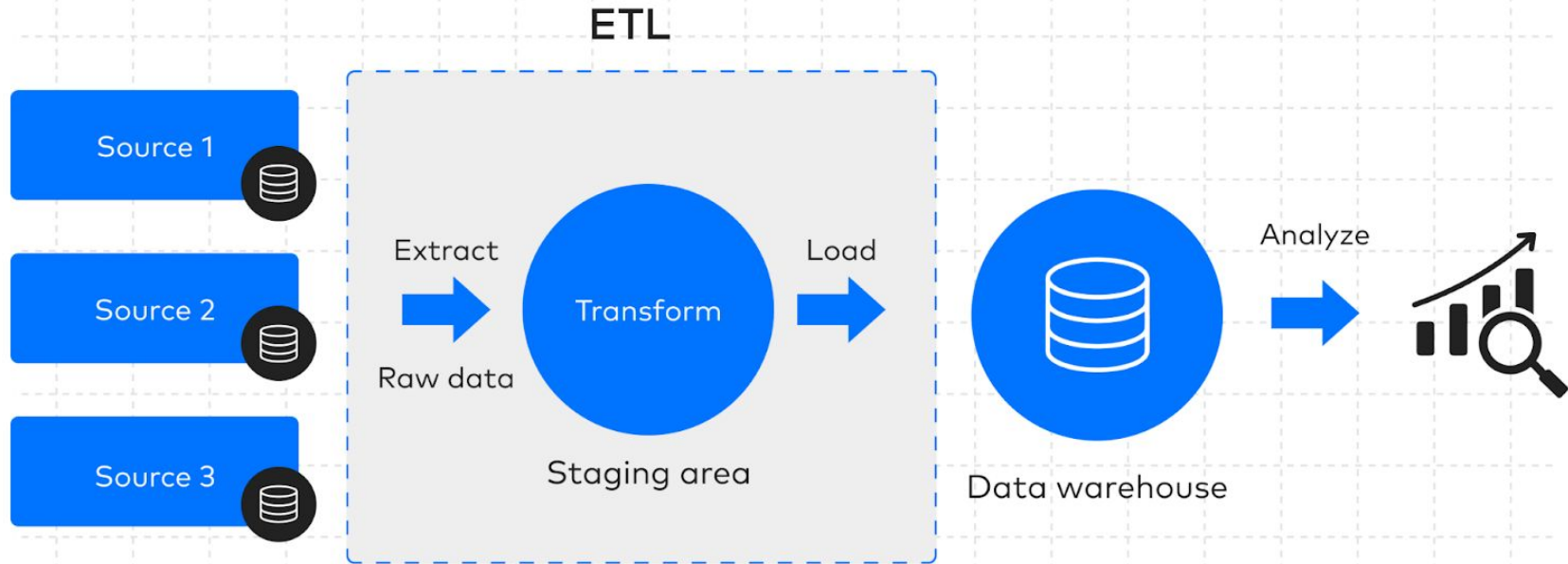
AI PIPELINES IN THE CLOUD



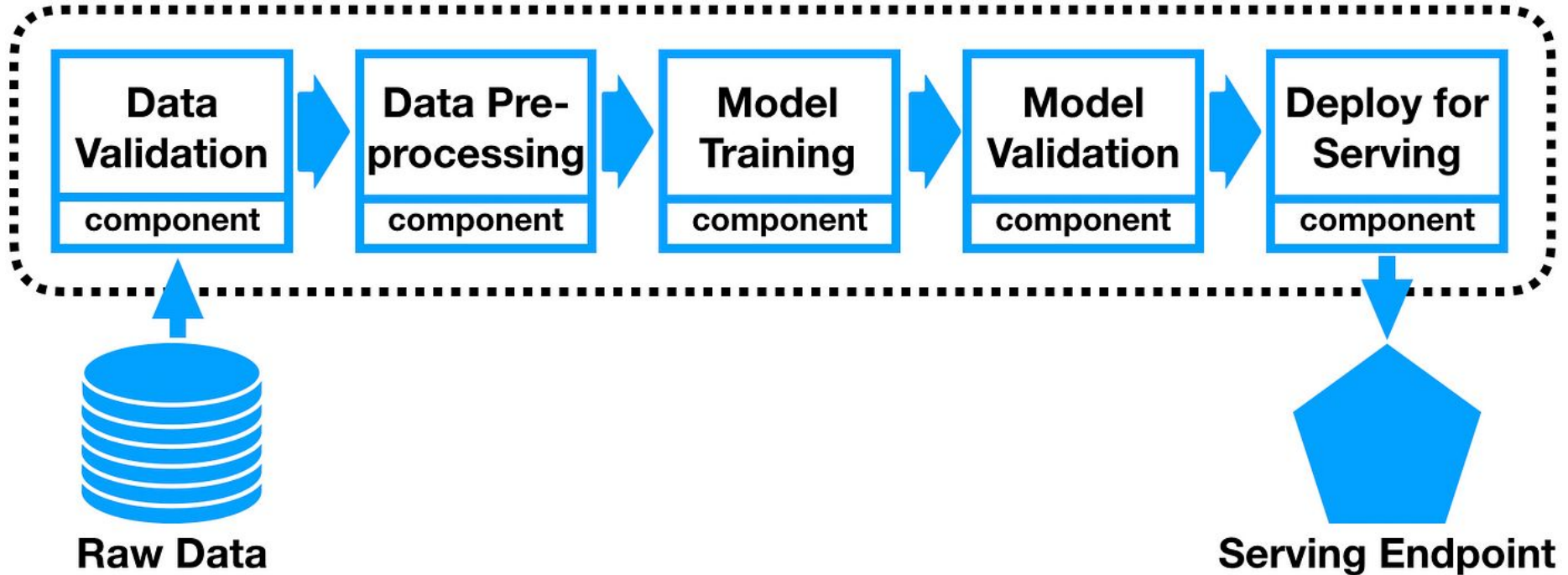
WHAT'S A PIPELINE?



ETL in data pipeline



ML Model Training Pipeline







CHINSTRAP!



GENTOO!



ADÉLIE!



@allison_horst

Remember, the cloud is
HARDWARE ENABLED!





<https://www.nvidia.com/en-us/deep-learning-ai/solutions/data-science/apache-spark-3/>

<https://www.databricks.com/blog/contributing-spark-loader-for-hugging-face-datasets>



Hugging Face

transformers: a machine learning architecture, 2017

"The transformer is a way to capture interaction **very quickly all at once between different parts of any input.** It's a general method that captures interactions between pieces in a sentence, or the notes in music, or pixels in an image, or parts of a protein. It can be purposed for any task." - Vaswani (Google)

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

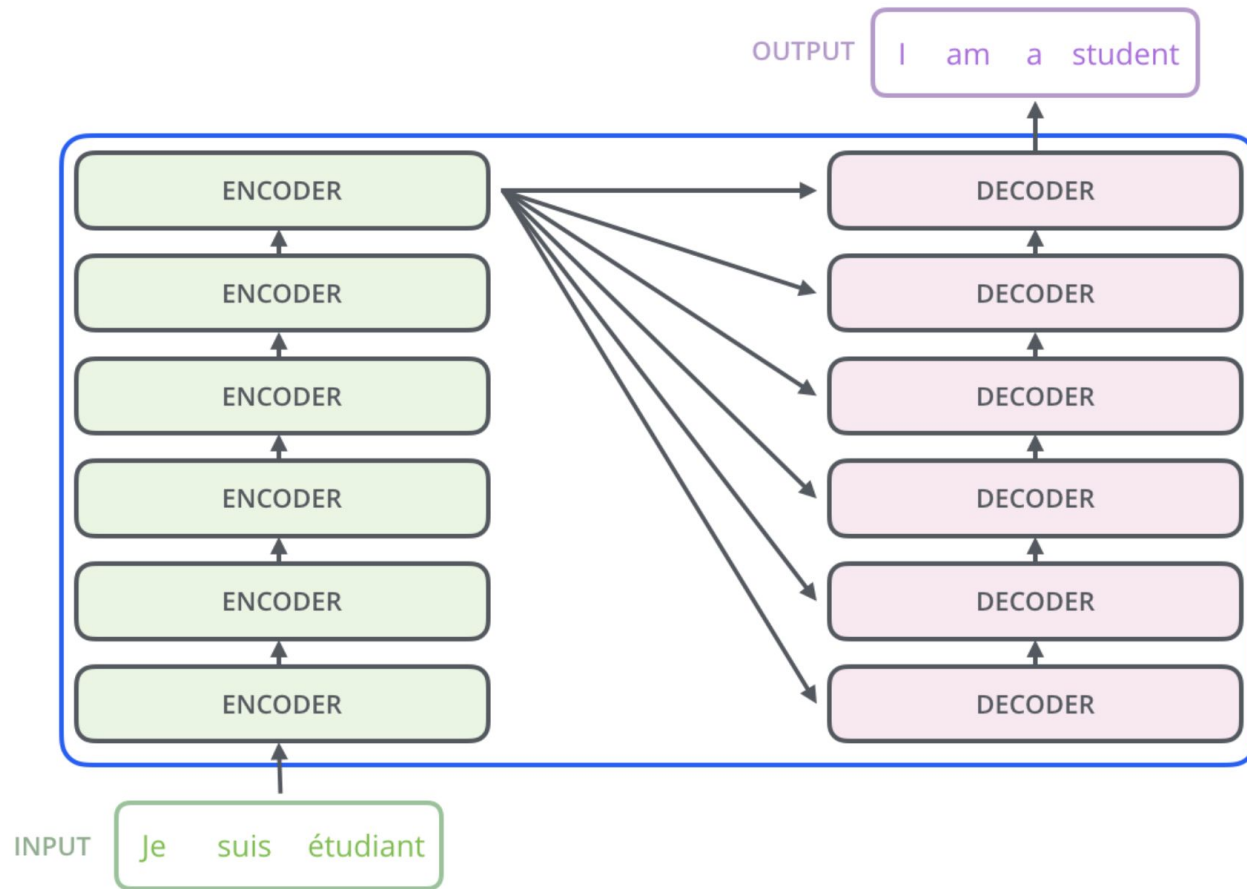
Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

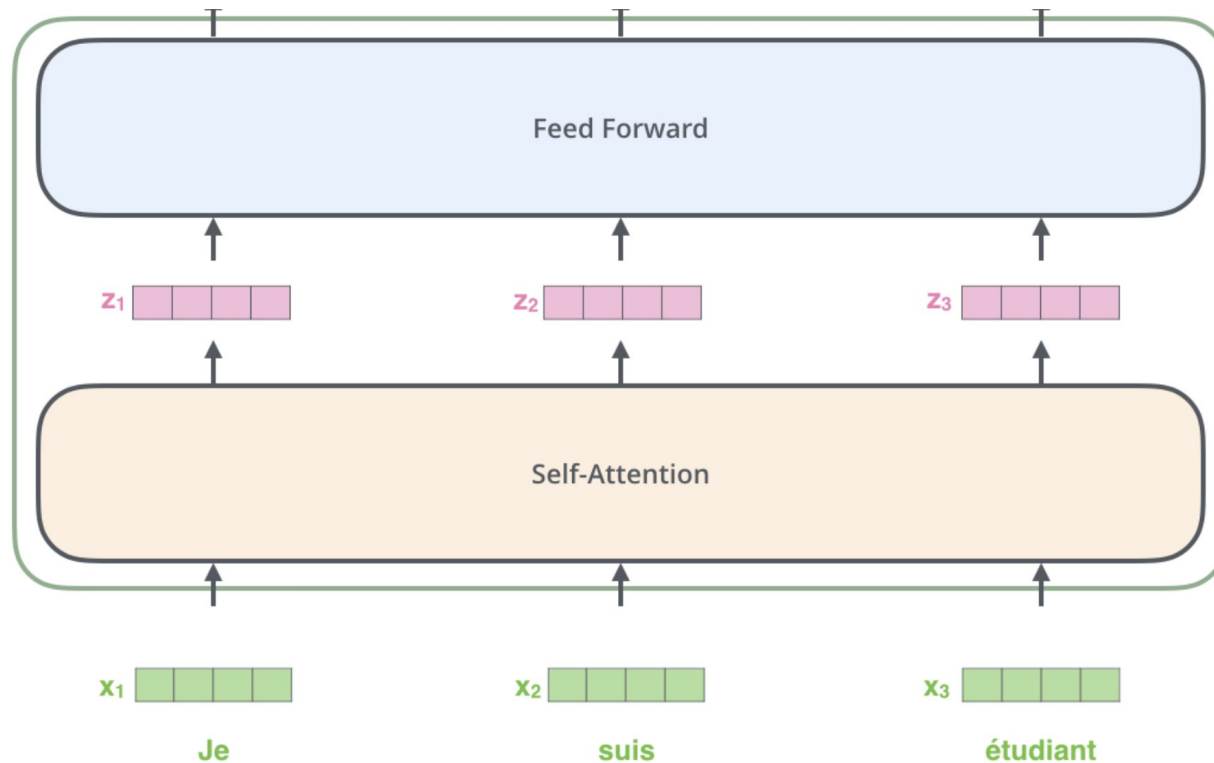
Illia Polosukhin* †
illia.polosukhin@gmail.com

Abstract

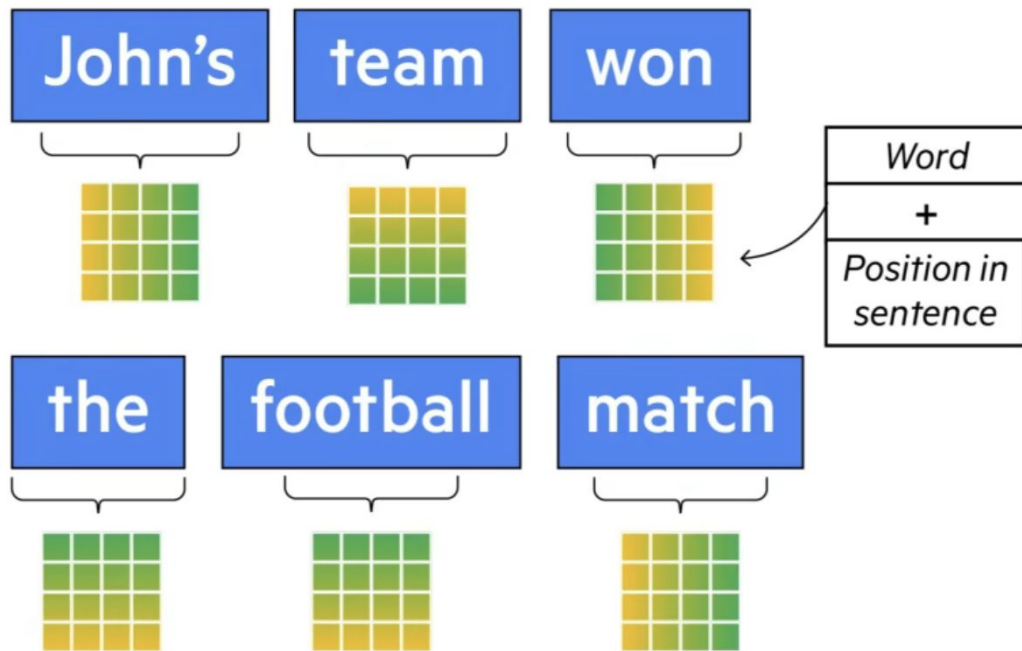
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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 English-to-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 other tasks by applying it successfully to English constituency parsing both with large and limited training data.



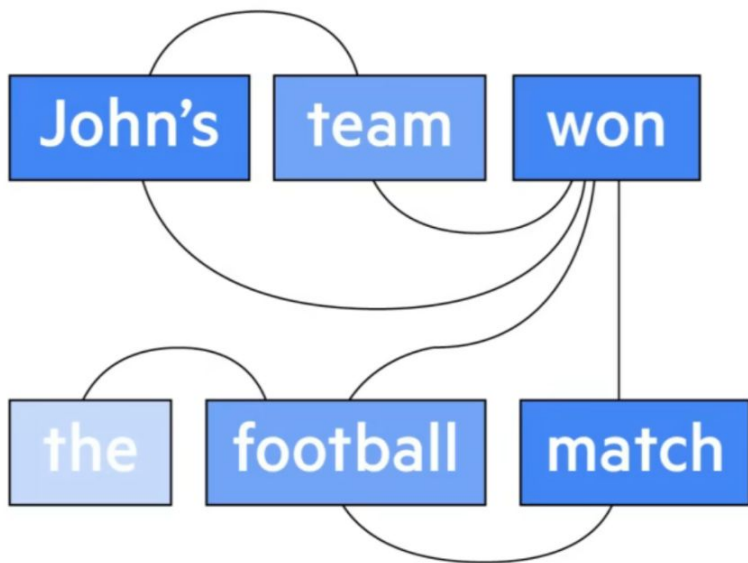




Here we begin to see one key property of the Transformer, which is that the word in each position flows through its own path in the encoder. There are dependencies between these paths in the self-attention layer. The feed-forward layer does not have those dependencies, however, and thus the various paths can be executed in parallel while flowing through the feed-forward layer.



Tokenize, embeddings

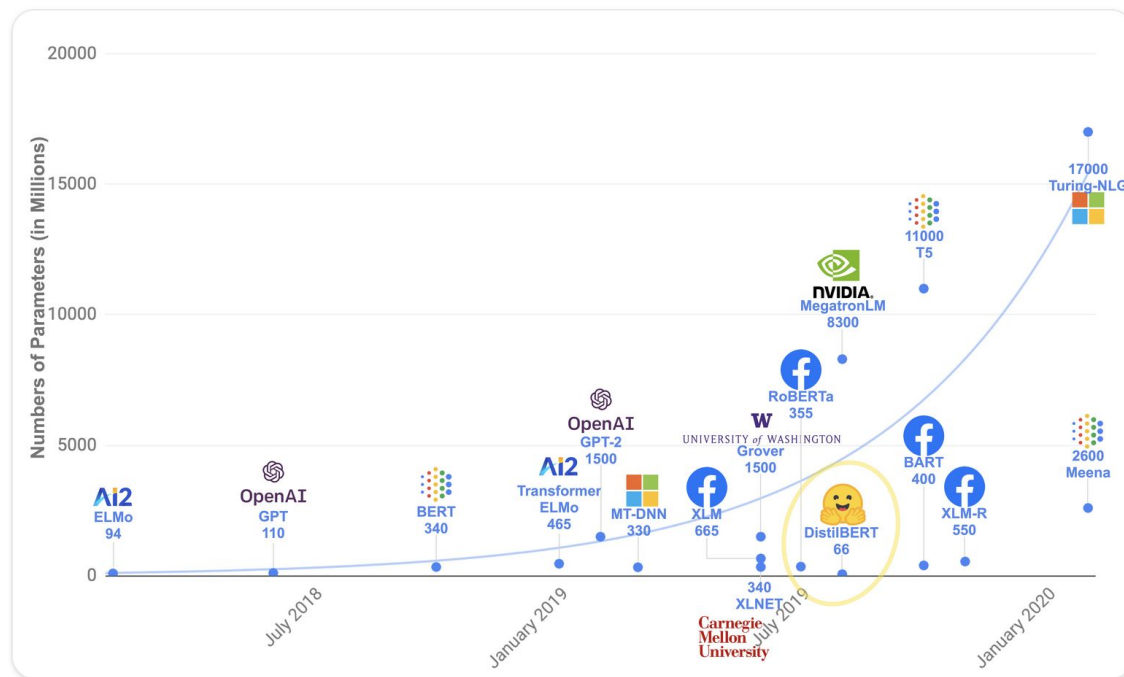


But it also uses self-attention, which is when the model looks at all the words in the sentence at the same time, capturing their connections and dependencies. It works out meaning based on context in the sentence.

"Meaning is a result of relationships between things, and self-attention is a general way of learning relationships" - Vaswani

Transformers are big models

Apart from a few outliers (like DistilBERT), the general strategy to achieve better performance is by increasing the models' sizes as well as the amount of data they are pretrained on.



- the transformer architecture
scales quadratically with sequence length
- self-attention compares every single word in a sequence to every other word in that sequence ...
- how to keep performance low on time?
- **PARALLELIZE + SCALE OUT**

CO2 emissions for a variety of human activities

