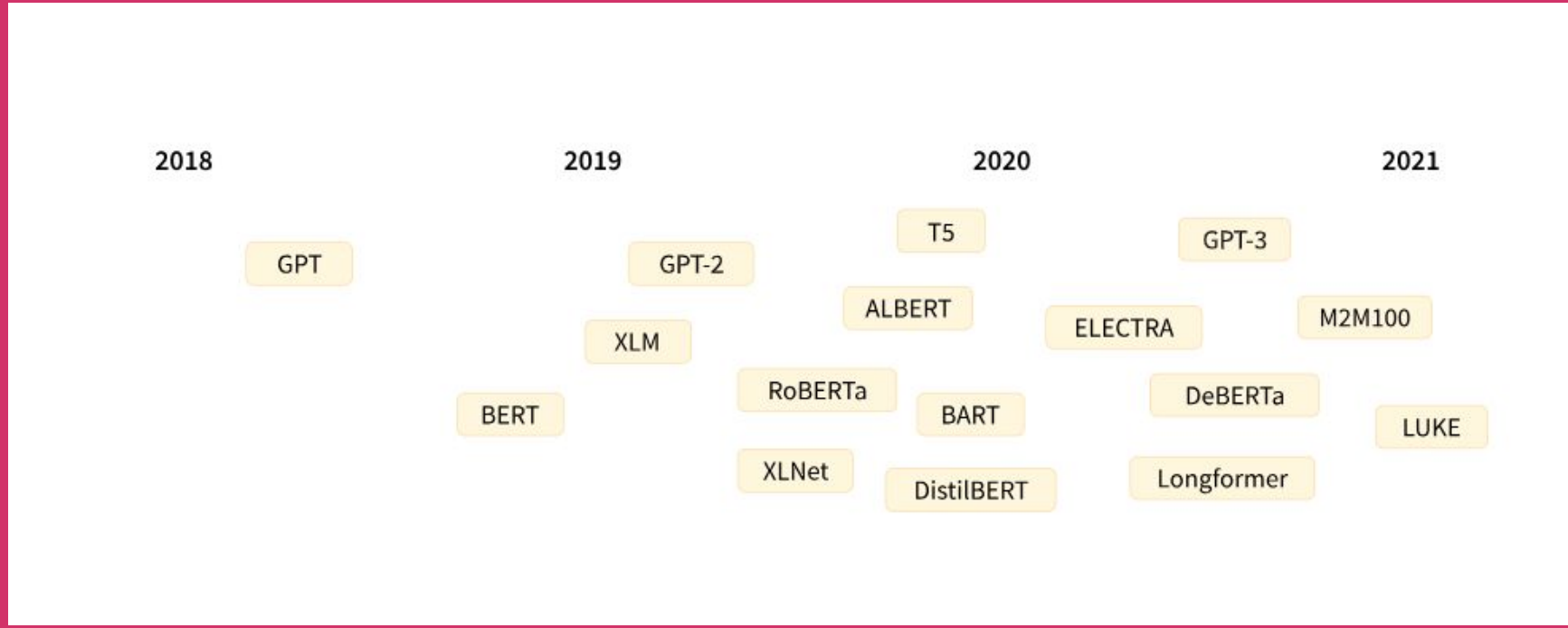
NLP INTERESTSHIP

MILESTONE 1 : Transformers, Attention, Transfer Learning, BERT Model, Fine Tuning.

By Pallavi Pannu

A bit of Transformer history!!



Transformers are language models…

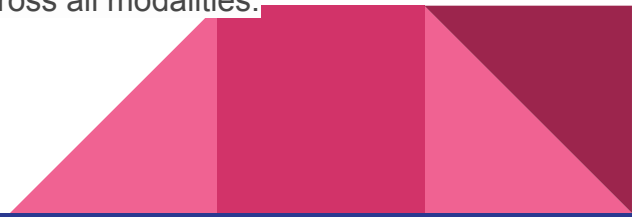
All the Transformer models mentioned before (GPT, BERT, BART, etc.) have been trained as *language models*. This means they have been trained on large amounts of raw text in a self-supervised fashion.

Introduction to Transformer

The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease.

Why should I use transformers?

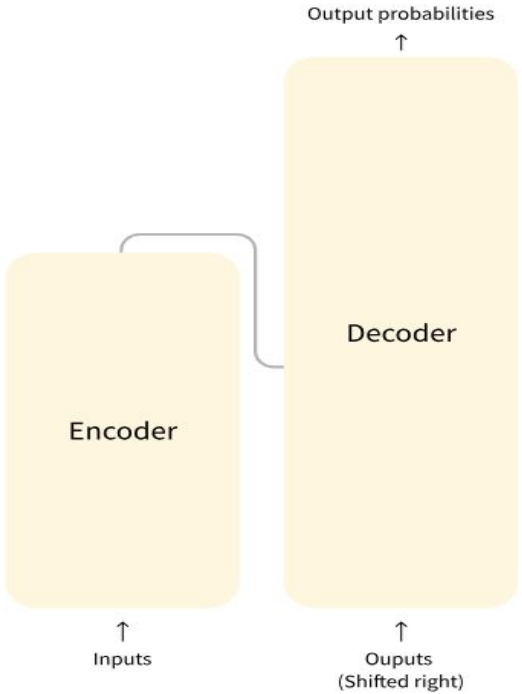
1. Easy-to-use state-of-the-art models:
    - High performance on natural language understanding & generation, computer vision, and audio tasks.
2. Lower compute costs, smaller carbon footprint:
    - Researchers can share trained models instead of always retraining.
    - Practitioners can reduce compute time and production costs.
    - Dozens of architectures with over 60,000 pretrained models across all modalities.



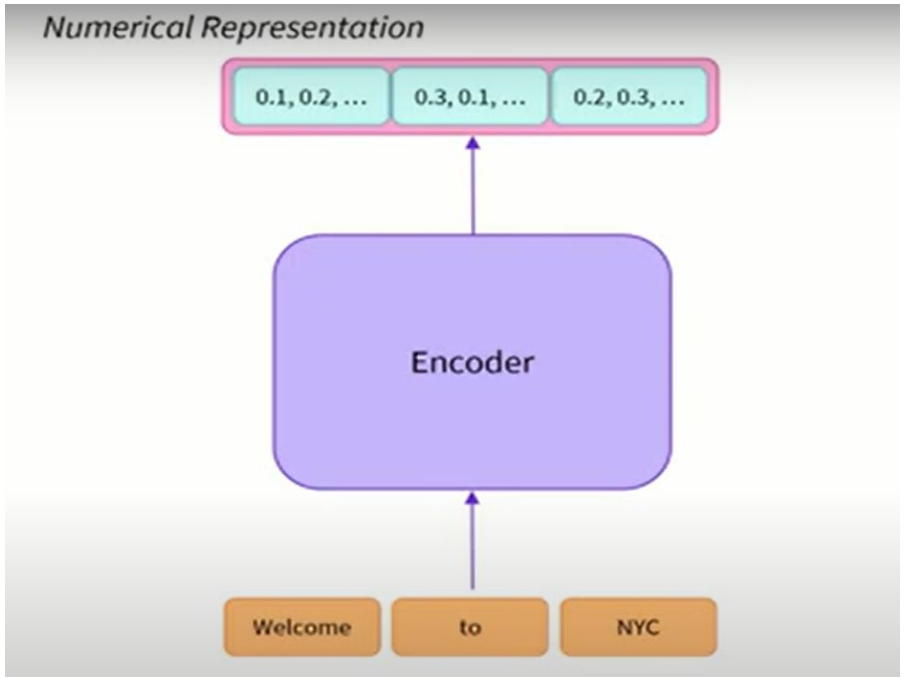
General Architecture of Transformer model

- **Encoder (left):** The encoder receives an input and builds a representation of it (its features).

- **Decoder (right)**: The decoder uses the encoder's representation (features) along with other inputs to generate a target sequence.



Let’s check in detail….



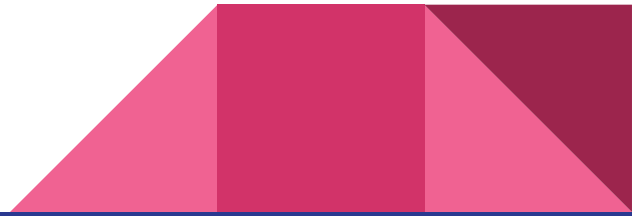
Pretraining

Pretraining is the act of training a model from scratch: the weights are randomly initialized, and the training starts without any prior knowledge.

Fine Tuning

Fine Tuning is the training done after a model has been pretrained.

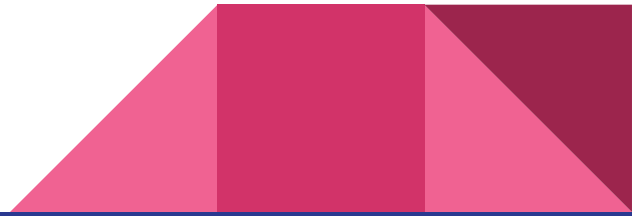
To perform fine-tuning, you first acquire a pretrained language model, then perform additional training with a dataset specific to your task.



Transfer learning

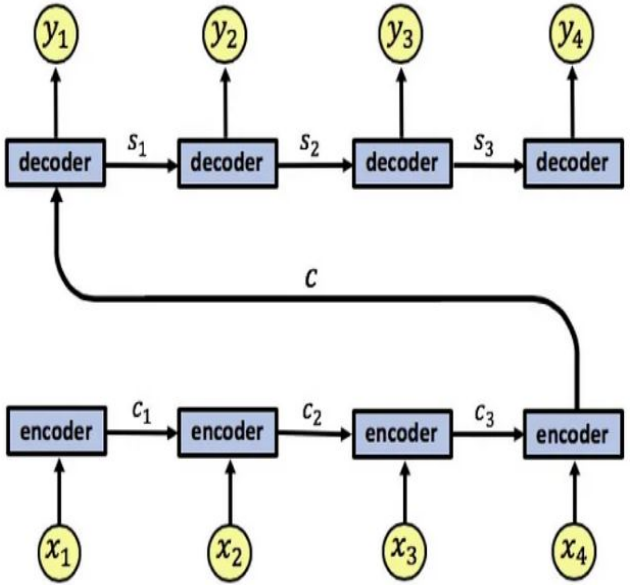
Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task.

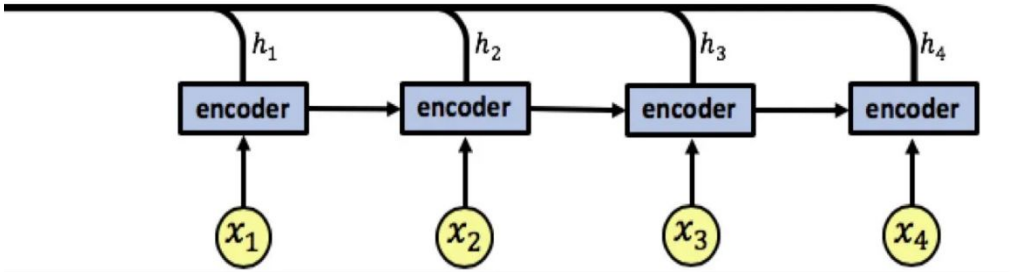
The fine-tuning will only require a limited amount of data: **the knowledge the pretrained model has acquired is “transferred,” hence the term *transfer learning*.**



Attention

Then came the concept of **attention**, that is during translation or in sequential tasks we not only need the output of the final encoder but instead we will have a output coming from every encoder, and we are **paying attention to every word** .





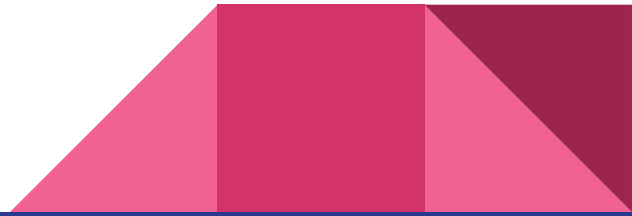
BERT

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once.

Therefore, BERT is considered bidirectional, though it would be more accurate to say that it's non-directional.

This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

BERT is a pretrained model.



Thank You !!