Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

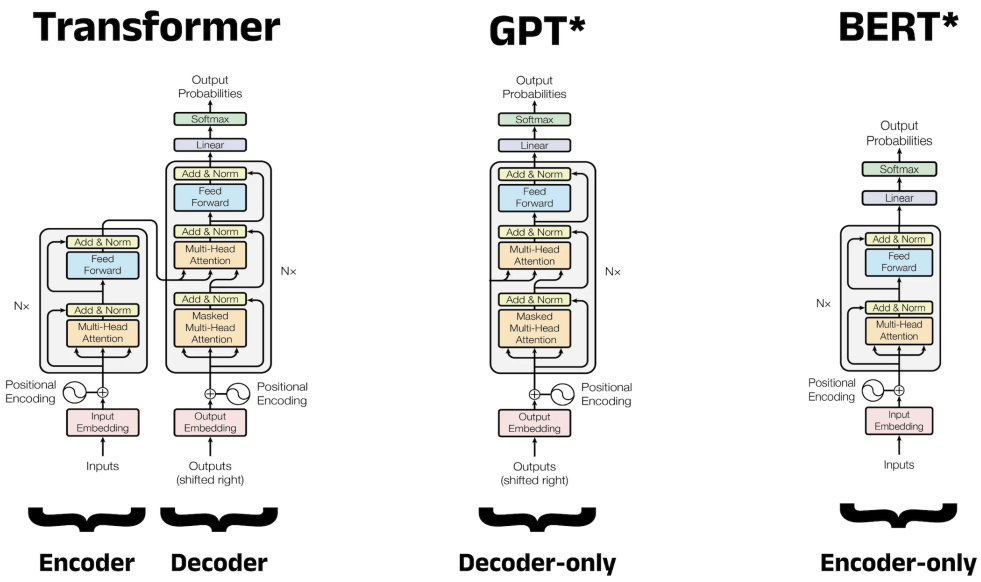
Different Transformers

BERT

BERT language model is an open source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT framework was pretrained using text from Wikipedia and can be fine-tuned with question-and-answer data sets.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection.

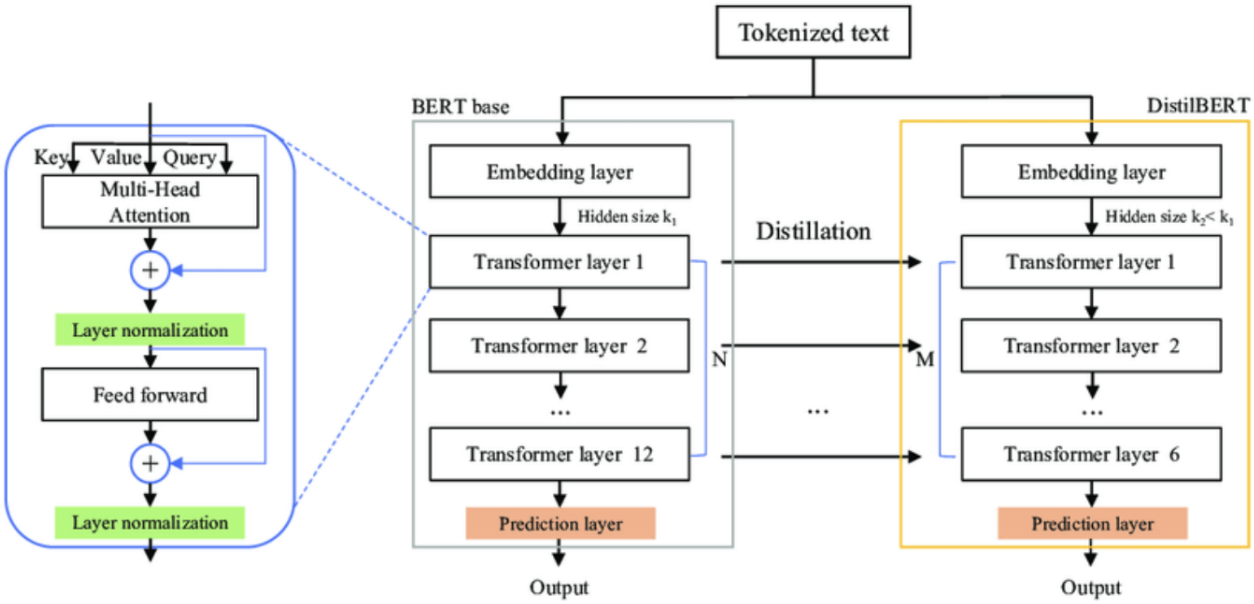
Historically, language models could only read input text sequentially – either left-to-right or right-to-left – but couldn't do both at the same time. BERT is different because it's designed to read in both directions at once. The introduction of transformer models enabled this capability, which is known as bidirectionality. Using bidirectionality, BERT is pretrained on two different but related NLP tasks: masked language modeling (MLM) and next sentence prediction (NSP).



*Illustrative example, exact model architecture may vary slightly

DistilBERT

DistilBERT is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40 %. To leverage the inductive biases learned by larger models during pre-training, the authors introduce a triple loss combining language modeling, distillation and cosine-distance losses.



BETO

BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique.

RoBERTa

RoBERTa (short for "Robustly Optimized BERT Approach") is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, which was developed by researchers at Facebook AI. Like BERT, RoBERTa is a transformer-based language model that uses self-attention to process input sequences and generate contextualized representations of words in a sentence.

One key difference between RoBERTa and BERT is that RoBERTa was trained on a much larger dataset and using a more effective training procedure. In particular, RoBERTa was trained on a dataset of 160GB of text, which is more than 10 times larger than the dataset used to train BERT. Additionally, RoBERTa uses a dynamic masking technique during training that helps the model learn more robust and generalizable representations of words.

RoBERTa has been shown to outperform BERT and other state-of-the-art models on a variety of natural language processing tasks, including language translation, text classification, and question answering. It has also been used as a base model for many other successful NLP models and has become a popular choice for research and industry applications.

Overall, RoBERTa is a powerful and effective language model that has made significant contributions to the field of NLP and has helped to drive progress in a wide range of applications.
