- The Transformer architecture was originally designed for sequence-to-sequence tasks like machine translation, but both the encoder and decoder blocks were soon adapted as standalone models.
- The representation computed for a given token in this architecture depends only on the left context. This is often called causal or autoregressive attention.
- s. The representation computed for a given token in this architecture depends both on the left (before the token) and the right (after the token) con- texts. This is often called bidirectional attention.
- encoder-only models like BERT can be applied to summarization tasks
- decoder-only models like those in the GPT family can be primed for tasks like translation
- attention is a mechanism that allows neural networks to assign a different amount of weight or "attention" to each element in a sequence. For text sequences,
- in practice, the meaning of a word will be better informed by complementary words in the context than by identical words—for example, the meaning of "flies" is better defined by incorporating infor- mation from "time" and "arrow" than by another mention of "flies"
- In practice, the self-attention layer applies three independent linear transformations to each embedding to generate the query, key, and value vectors. These transformations project the embeddings and each projection carries its own set of learnable parameters, which allows the selfattention layer to focus on different semantic aspects of the sequence
- But why do we need more than one attention head? The rea- son is that the softmax of one head tends to focus on mostly one aspect of similarity. Having several heads allows the model to focus on several aspects at once. one head can focus on subject-verb interaction, whereas another finds nearby adjectives.
- The feed-forward sublayer in the encoder and decoder is just a simple two-layer fully connected neural network, but with a twist: instead of processing the whole sequence of embeddings as a single vector, it processes each embedding independentlyFor this reason, this layer is often referred to as a position-wise feed-forward layer
- a GELU activation function is most commonly used in feed forward layer
- r, the Transformer architecture makes use of layer normalization and skip connections
- the difference between pre layer normalization(before) and post layer normalization (after which need learning rate warm up)
- Positional embeddings are based on a simple, yet very effective idea: augment the token embeddings with a position-dependent pattern of values arranged in a vector
- Transformer models can use static patterns consisting of modulated sine and cosine signals to encode the positions of the tokens. This works especially well when there are not large volumes of data available.
- we need to add classification head at the end of transformer encoder for example to build classification model
- the main difference between the decoder and encoder is that the decoder has two attention sublayers:
- The trick with masked self-attention is to introduce a mask matrix with ones on the lower diagonal and zeros above
- there are tree family of transformer (encoder only , decoder only and the whole structure)
- he first encoder-only model based on the Transformer architecture was BERT. At the time it was published, it outperformed all the state-of-the-art models on the pop- ular GLUE benchmark,7 which measures natural language understanding (NLU) across several tasks of varying difficulty (entity recognition , text classification and question answering)

- The progress on transformer decoder models has been spearheaded to a large extent by OpenAI. These models are exceptionally good at predicting the next word in a sequence and are thus mostly used for text generation tasks
-