Diagram

Description automatically generatedAttention Is All You Need

# 3. Model Architecture

* Most competitive neural sequence transduction models have an encoder-decoder structure.
  + **Encoder**:
    - Maps an input sequence of symbol representations
    - To a sequence of continuous representations
  + **Decoder**:
    - Given **z**, the decoder then generates an output sequence of symbols one element at a time.
    - At each step the model is auto regressive, consuming the previously generated symbols as additional input when generating the next.
* The Transformer follows this overall architecture using stacked self-attention and pointwise, fully connected layers for both the encoder and decoder.

## 3.1. Encoder and Decoder Stacks

Figure 1: The Transformer - model architecture

* **Encoder**:
  + Is composed of a stack of N = 6 identical layers.
  + Each layer has two sub-layers:
    - First: multi-head self-attention mechanism.
    - Second: position-wise fully connected feed-forward network.
  + We employ a residual connection around each of the two sub-layers, followed by layer normalization.
  + That is, the output of each sub-layer is:
  + Where is the function implemented by the sub-layer itself.
  + To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce output of dimension:
    - dmodel = 512