Transformer Review

By Dequan Er

Outline

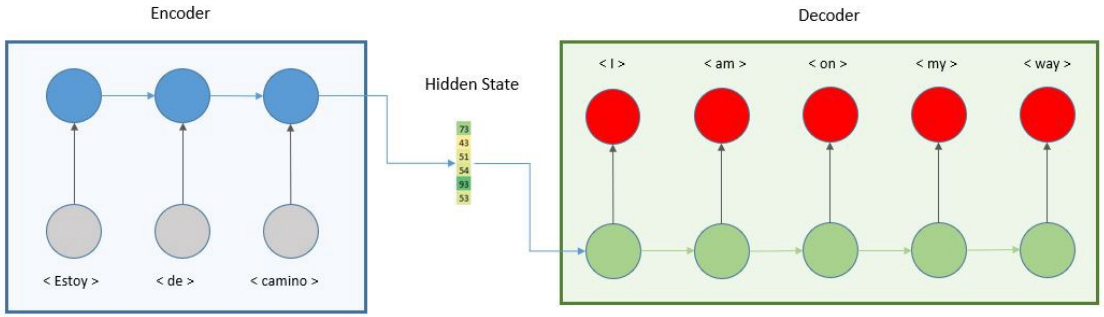
- What is Transformer

Introduction

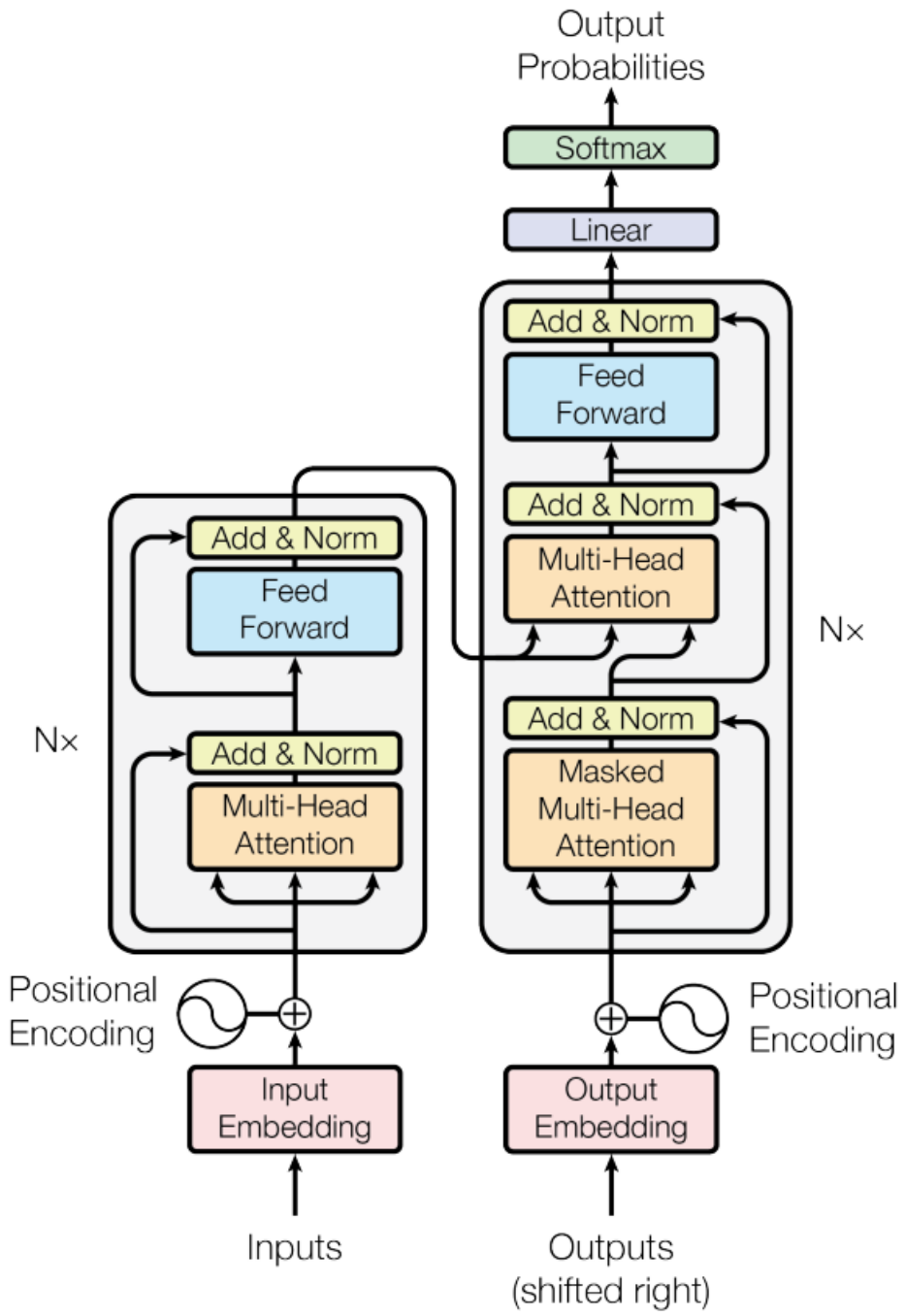
- Sequence Modeling
  - RNN, LSTM, GRU
  - From recurrent language models to encoder-decoder model
    - H_t-1 to h_t
    - Hard to paralleize
    - Long sequence memory lose, large H_t
  - Attention used from encoder to decoder

Model Architecture

- Structure of encoder-decoder

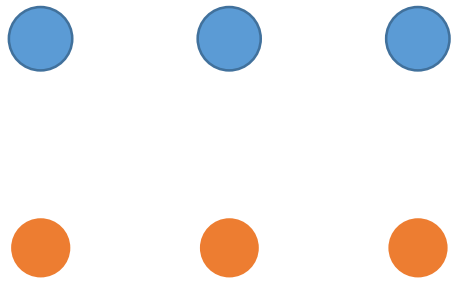


- Transformer



Detail of architect

- Attention: mapping a query and a set of key-value pairs to an output in vector form.



Detail of architect

- Scaled dot-product attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$