Transformer Review

By Dequan Er

Outline

- Introduction
- Model architecture
- Result
- Derived models
- Applications

Introduction

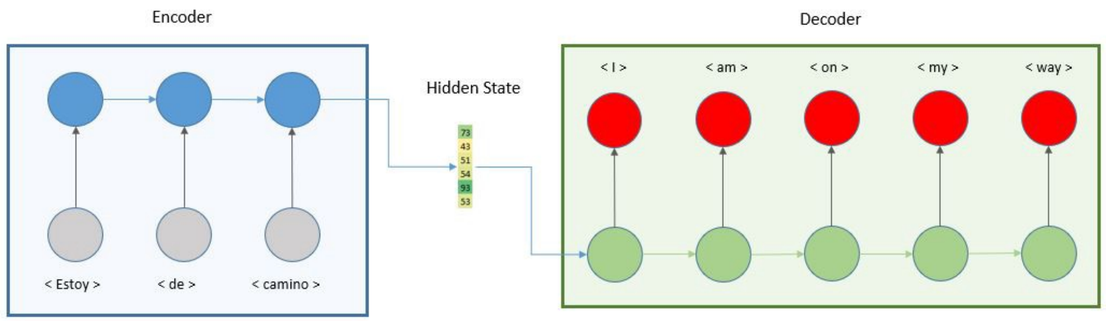
- From Sequence Modeling to Attention
  - RNN, LSTM, GRU
  - From recurrent language models to encoder-decoder model
    - H_{t-1} to h_t
    - Hard to parallelize
    - Long sequence memory lose, large H_t
  - Attention used from encoder to decoder

- Derivatives of Transformers
  - BERT (2018), GPT3 (2020)
  - ChatGPT(2022)

- From NLP to audio, video, etc.

Model Architecture

- Structure of encoder-decoder



- Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

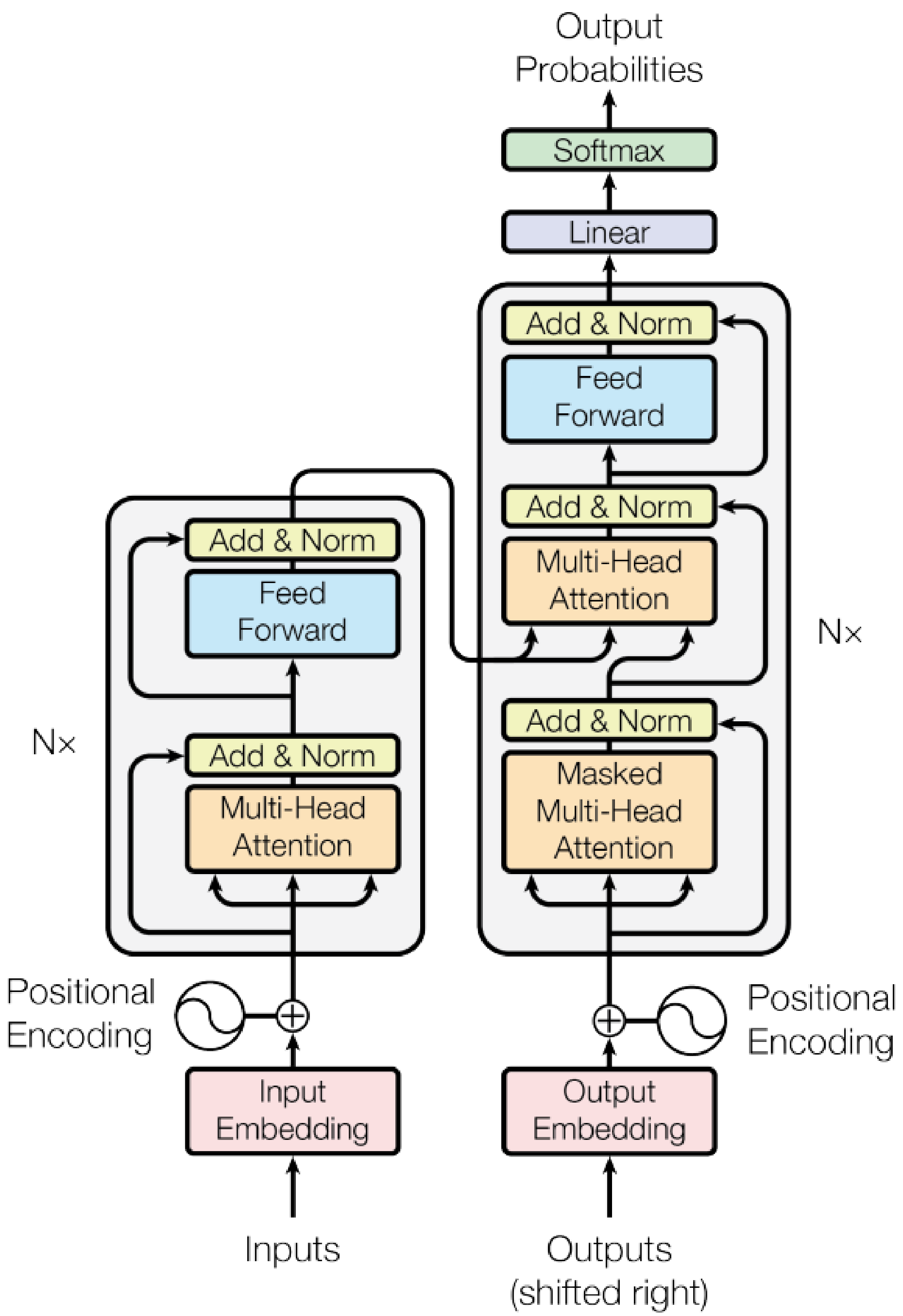
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

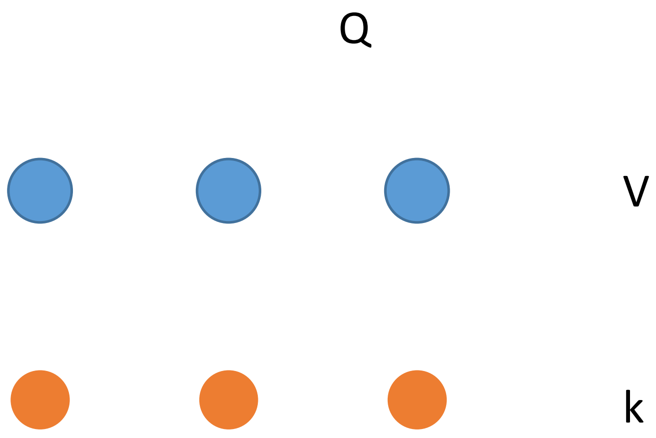
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

31st Conference on Neural Information Processing Systems (**NIPS 2017**), Long Beach, CA, USA.



Model Architecture

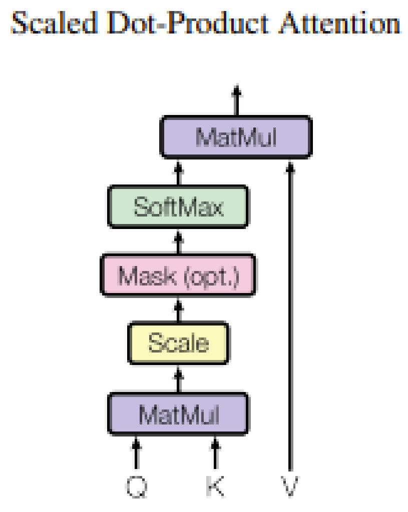
- Attention layer: mapping a query and a set of key-value pairs to an output in vector form.

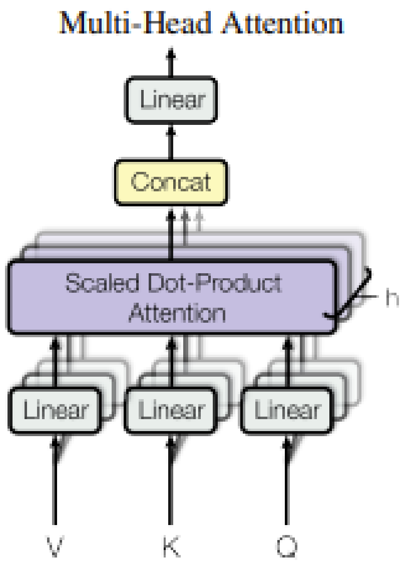


Model Architecture

- Scaled dot-product attention

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





Model Architecture

- Position-wise Feed-Forward Networks

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- Positional Encoding
  - Ensure sequence of tokens