Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Maximilian Bernhard, Niklas Strauß

Munich, 25.11.2024

Deep Learning and Artificial Intelligence

WS 2024/25

Exercise 6: Attention & Transformer

Exercise 6-1 Implementation

In Moodle you find a Jupyter notebook in which you should implement the multihead attention mechanism in PyTorch.

Exercise 6-2 Additional Questions

(a) Discuss the relation between self-attention and cross-attention in terms of their similarities and dissimilarities, and where they are used in the transformer architecture.

(b) The figure below depicts the original transformer architecture as proposed in Attention Is All You Need. Which of the attention layers use self-attention and which ones use cross-attention?

(c) What are positional encodings? Why do we need them? What two kinds of positional encodings exist?

(d) Compare convolutional, recurrent, and self-attention layers in terms of their computational complexity.

(e) Compare transformers and RNNs in terms of their capabilities for handling long sequences.

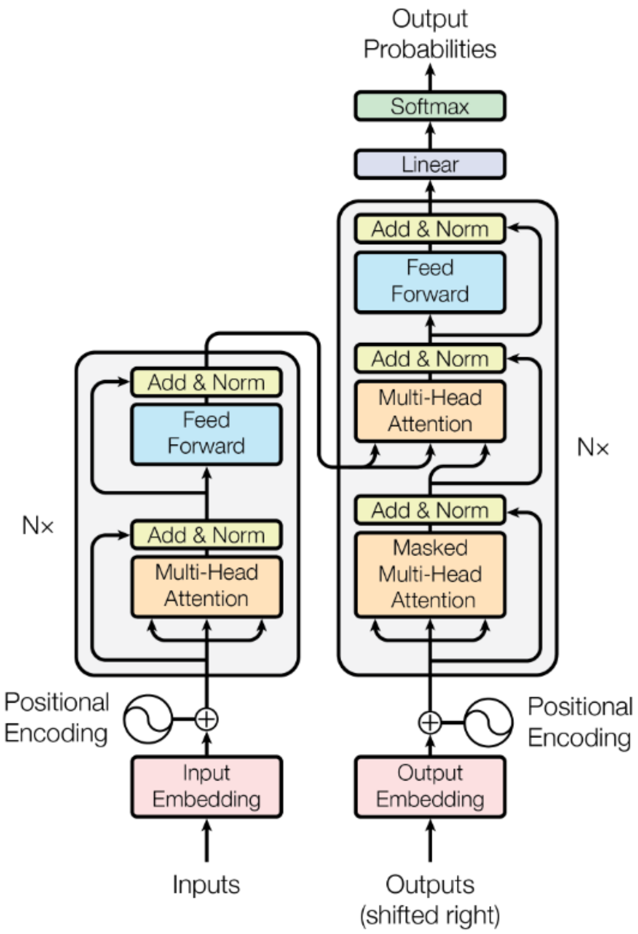

Figure 1: **Transformer architecture** from Attention Is All You Need.