Neural Computation

Introduction to the Transformer

Attention Is All You Need

Ashish Vaswani* Noam Shazeer* Niki Parmar* .Jakob Uszkoreit* Google Brain Google Brain Google Research Google Research nikip@google.com usz@google.com avaswani@google.com noam@google.com Aidan N. Gomez* † Llion Jones* Łukasz Kaiser* Google Research University of Toronto Google Brain llion@google.com aidan@cs.toronto.edu lukaszkaiser@google.com

Illia Polosukhin* † illia.polosukhin@gmail.com

Sequence-to-Sequence Architecture



- Hugely influential
- Basis for ChatGPT (Generative Pre-trained Transformer)
- Also for vision

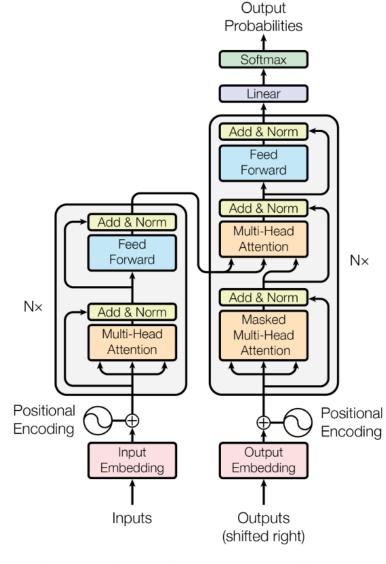


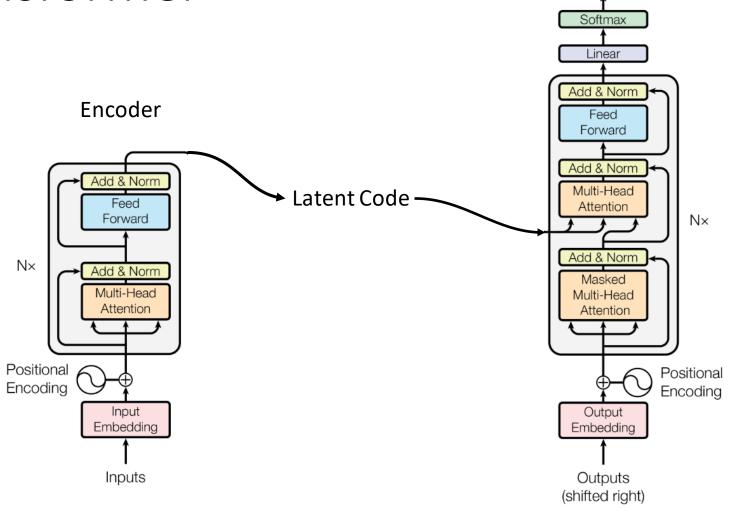
Figure 1: The Transformer - model architecture.

The Transformer Output Probabilities Softmax Linear Add & Norm Feed Forward Add & Norm Add & Norm Multi-Head Feed Attention Forward $N \times$ Add & Norm $N \times$ Add & Norm Masked Multi-Head Multi-Head Attention Attention Positional Positional Encoding Encoding Input Output Embedding Embedding Inputs Outputs (shifted right)

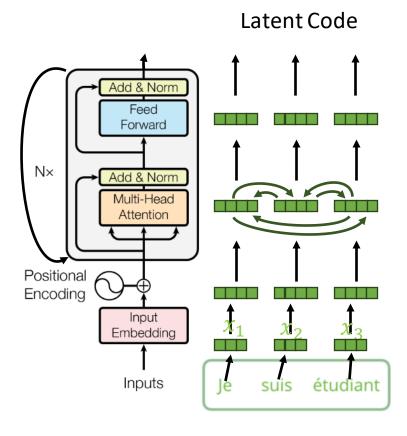
Decoder

Output Probabilities

The Transformer



The Encoder



- Process set of tokens
- Tokens remain separate
 - (except for attention layer)
- Tokens don't have order
 - (except for positional encoding)

