Neural Computation

Introduction to the 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

- Sequence-to-Sequence Architecture

- Hugely influential
- Basis for ChatGPT (**G**enerative **P**re-trained **T**ransformer)
- Also for vision

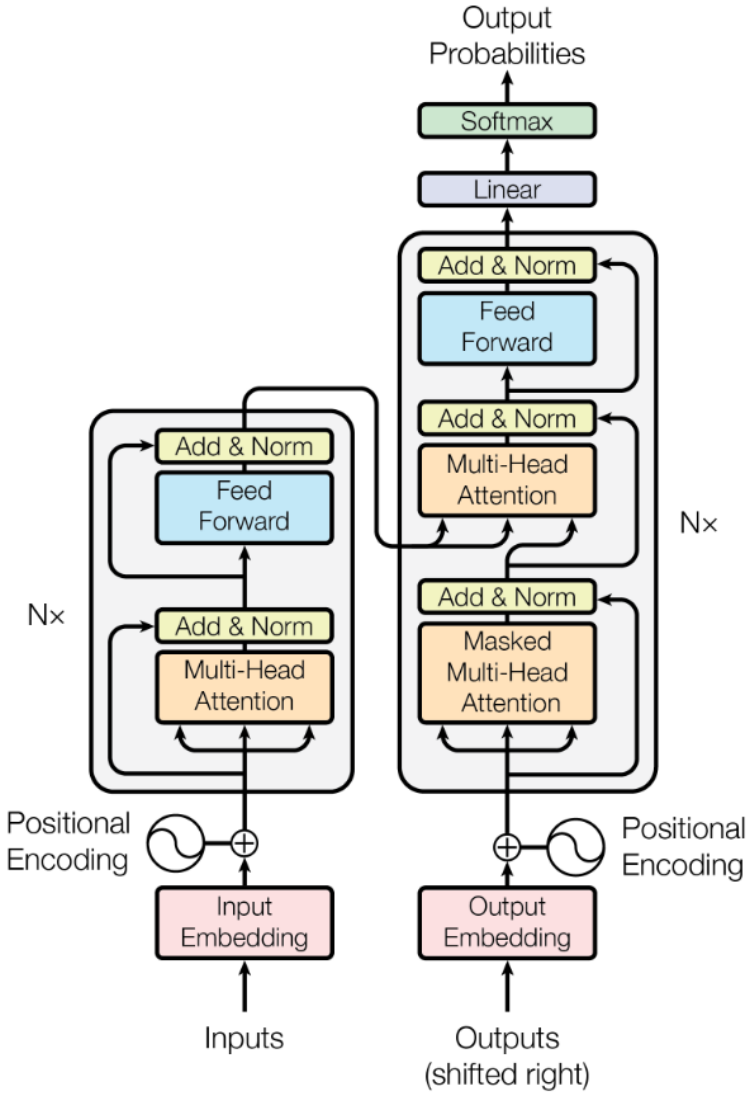
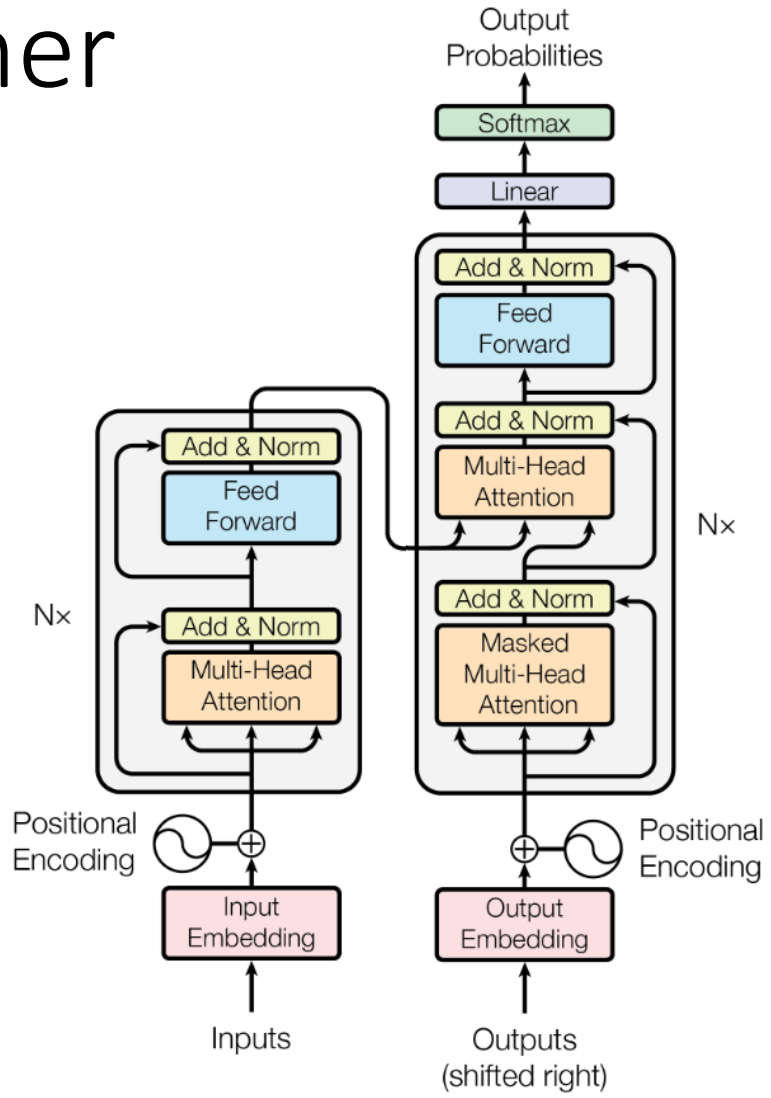
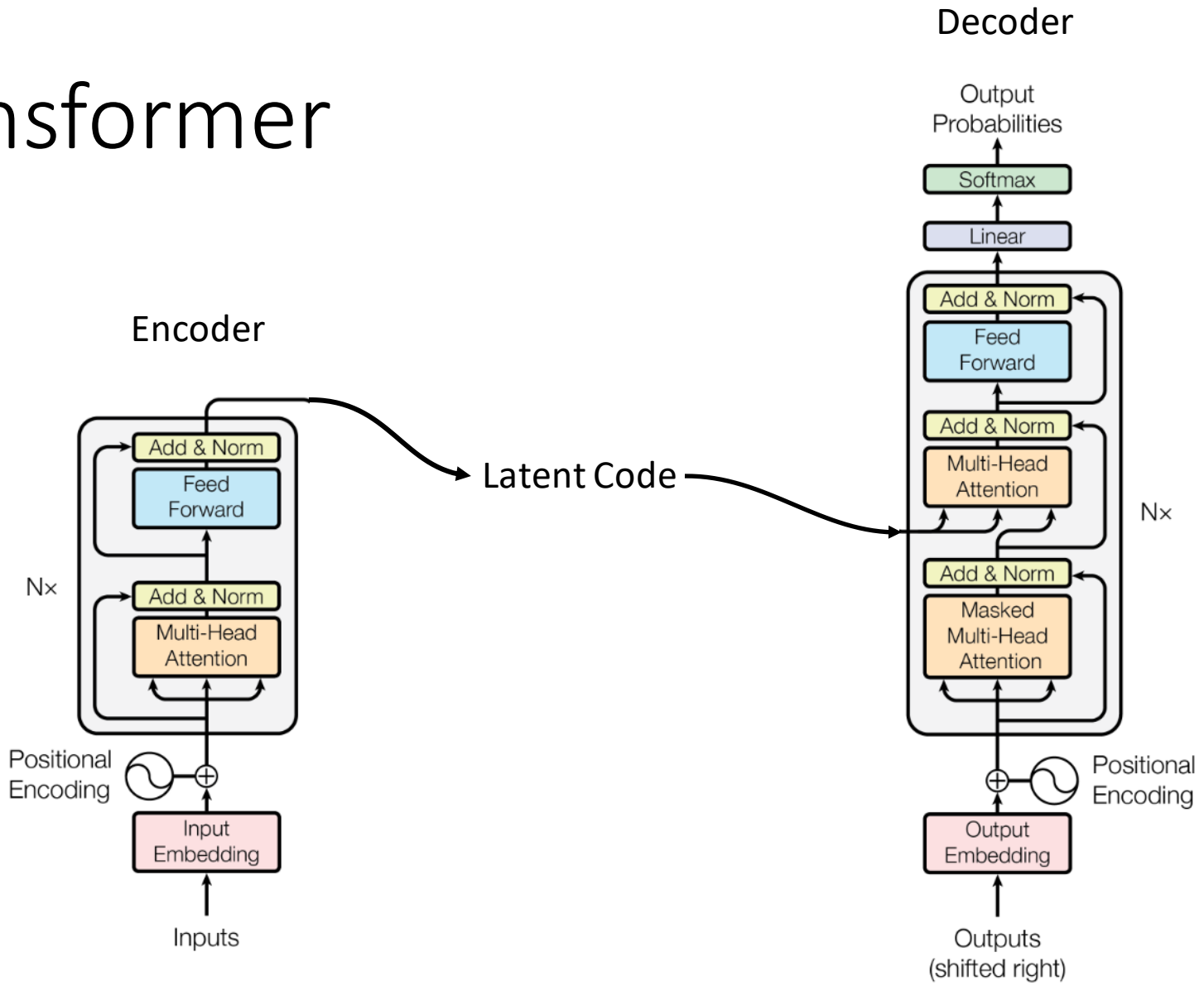


Figure 1: The Transformer - model architecture.

The Transformer

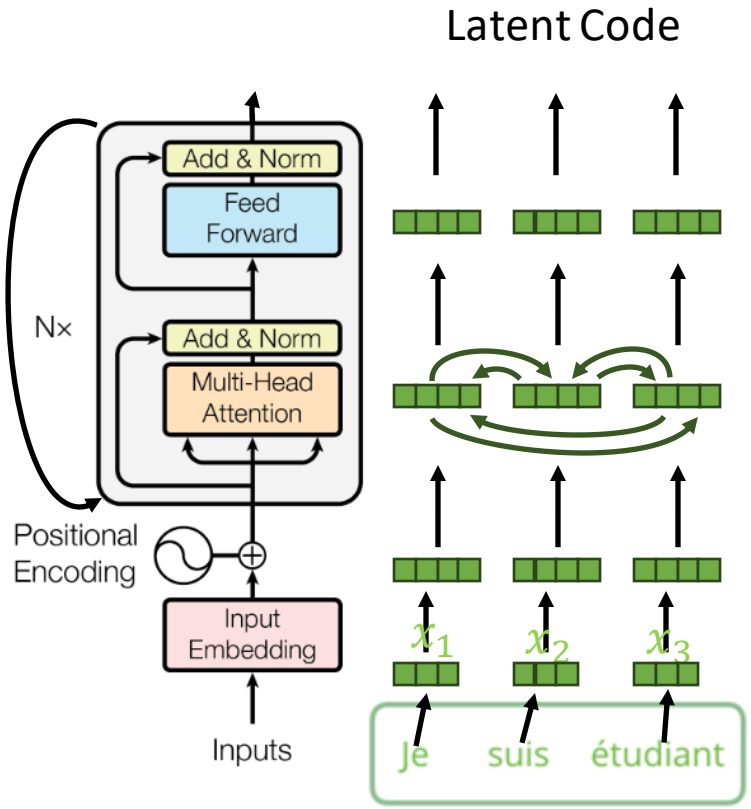


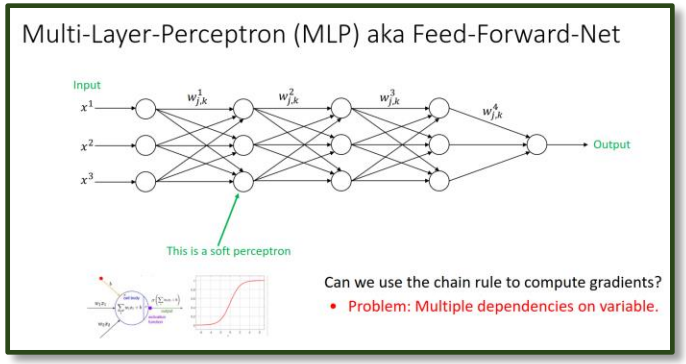
The Transformer

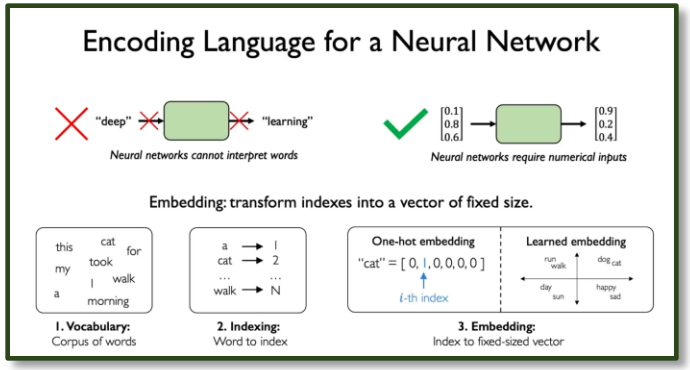


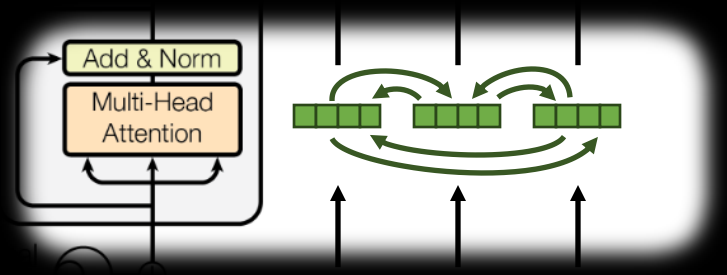
The Encoder

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








Add & Norm
Multi-Head
Attention