Transfomers

Attention is all you need

- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. (2017)

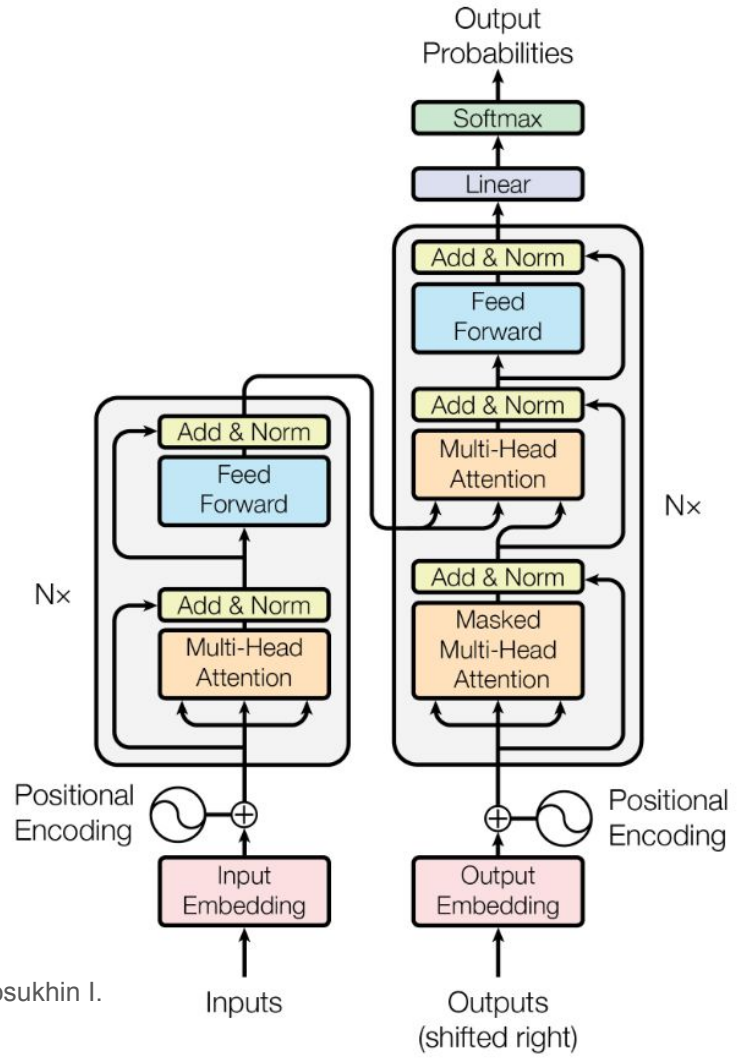
Idea:

- No more recursive processing (no LSTM, RNN)
- Use only attention
- Still has an encoder - decoder architecture

Transformer

- Left: Encoder
- Right: Decoder

- Multi head attention
- Self attention in encoder
- Masked attention in decoder



Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. (2017)

Encoder only

- Make predictions about the sequence or its parts

Tasks

- Text classification
- Predict missing words
- Document embeddings
- …

E.g. Bert

Decoder only

- Predict new output sequence in response to an input sequence

Tasks

- Chat
- Question answering
- Translation

E.g. GPT

Foundation models

- Models trained on huge text corpora
- Without a specific task in mind

- Fine tune to task at hand
- Reuse the same “expensive” foundation model