Transformer and GPT Models

An Overview of Their Architecture

Mingzhi Ye

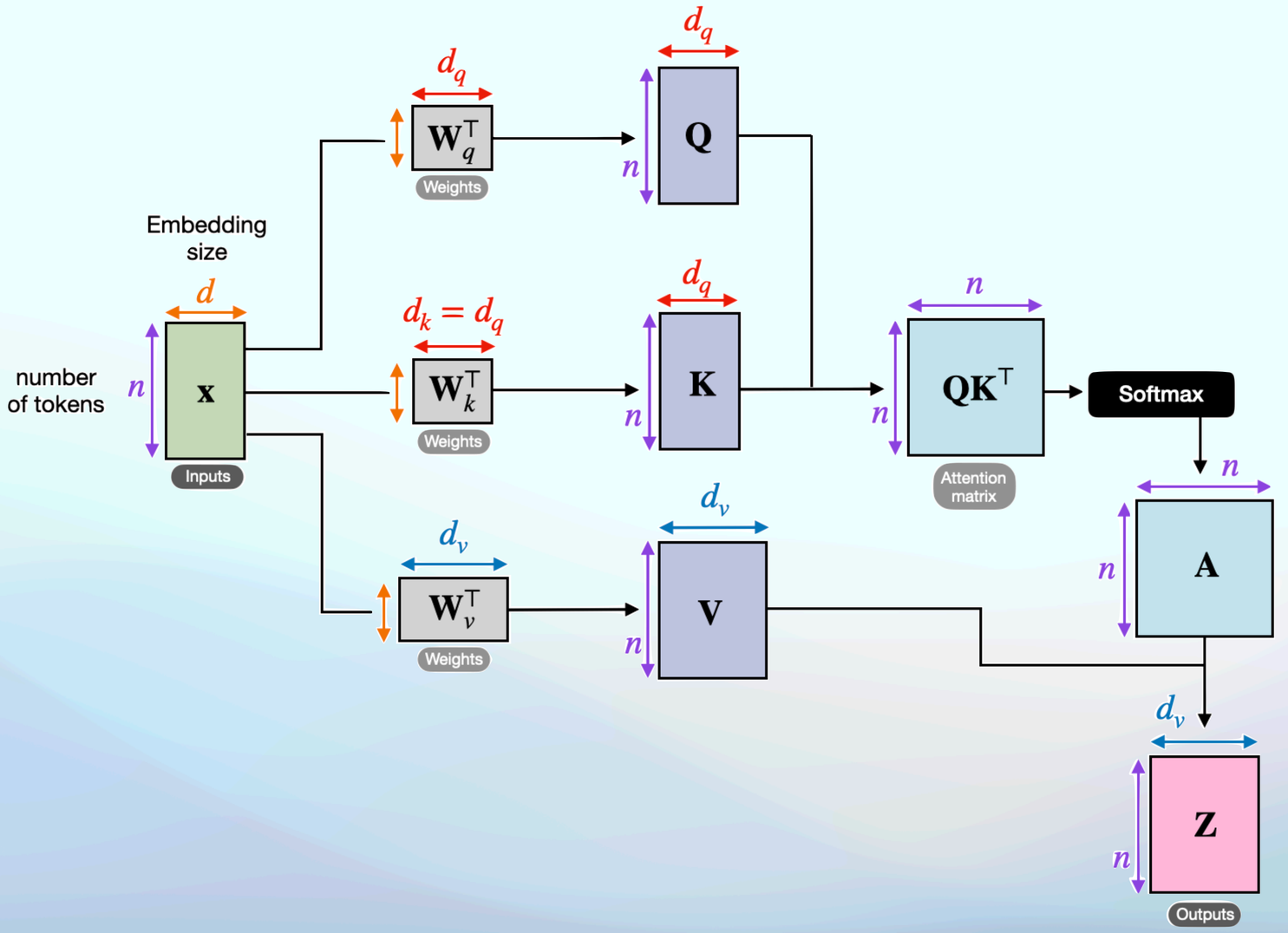
Introduction to Transformers

Large and Powerful

- Transformers have revolutionized NLP tasks by using the self-attention mechanism.
- In Enginsh-to-French translation task, the BLUE score(measuring translation quality) reaches 41.8, much better than former models.
- The original Vanilla Transformer model has 213 million parameters.
- The model processes input data in parallel, unlike the sequential nature of RNNs, providing computing efficiency.

Self-Attention

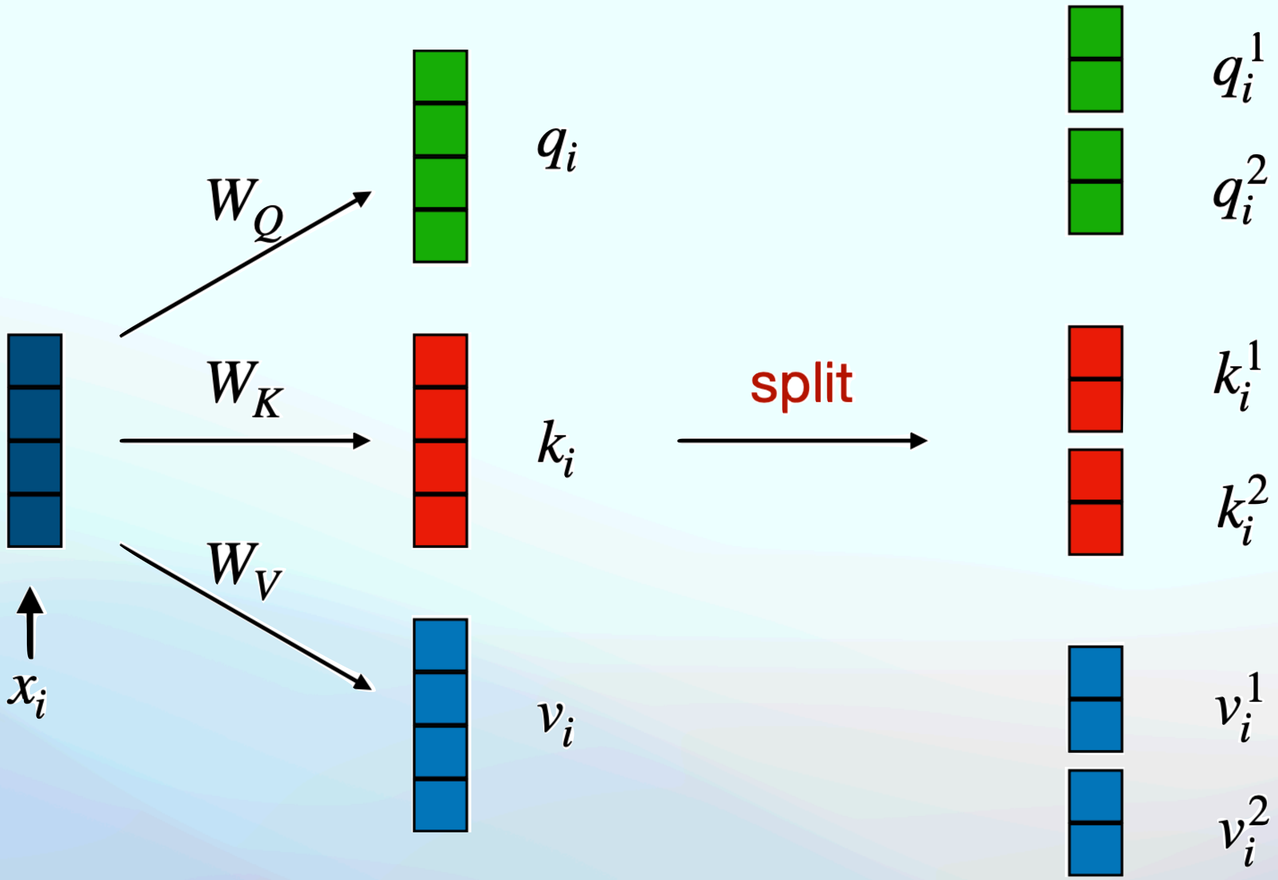
Essential Mechanism for Transformer



- The input is X with n x_i
- Query q_i = Wq * x_i
- Key k_i = W_k*x_i
- Value v_i = Wv * x_i
- Attention $a_{ij} = \mathrm{softmax}(\frac{\mathbf{q}_i \mathbf{k}_j{}^\top}{\sqrt{d_k}})$
- Output $y_i = \sum_{j=1}^{n} a_{i,j} v_j$

Multi-head Attention

Split and Process Separately



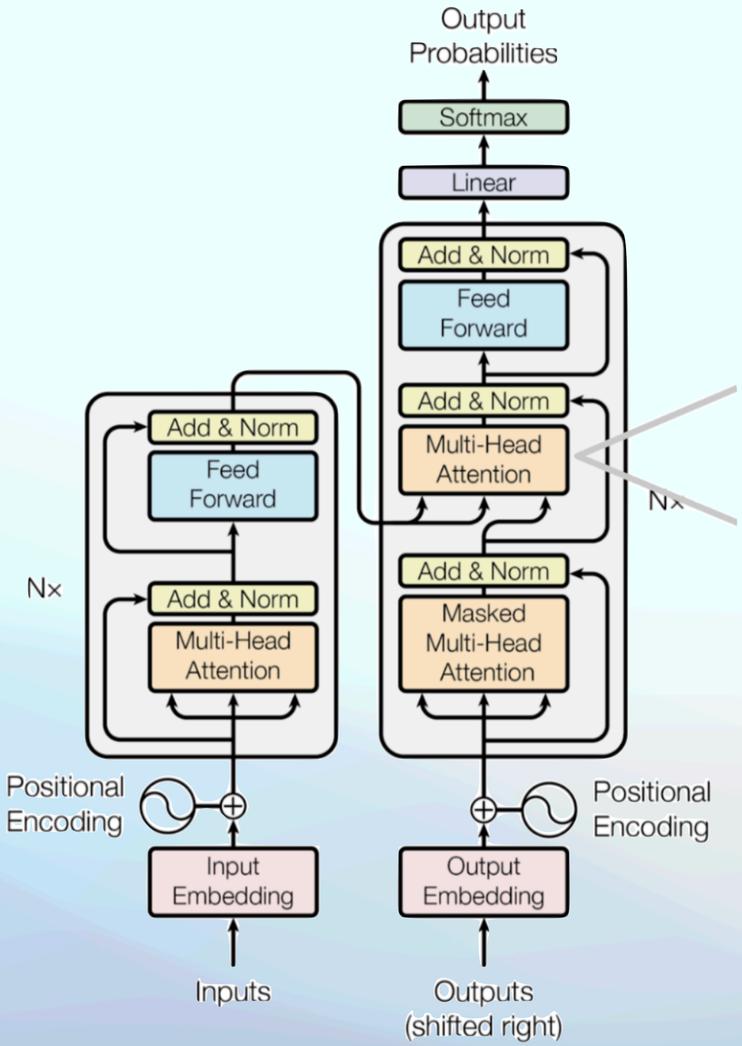
$$h_1 = \text{attn}(Q_1, K_1, V_1) = \text{softmax}(\frac{Q_1 K_1^T}{\sqrt{d/2}})V_1$$

$$h_2 = \text{attn}(Q_2, K_2, V_2) = \text{softmax}(\frac{Q_2 K_2^T}{\sqrt{d/2}})V_2$$

$$Y = \text{concat}(h_1, h_2)W_O$$

Architecture of Transformer

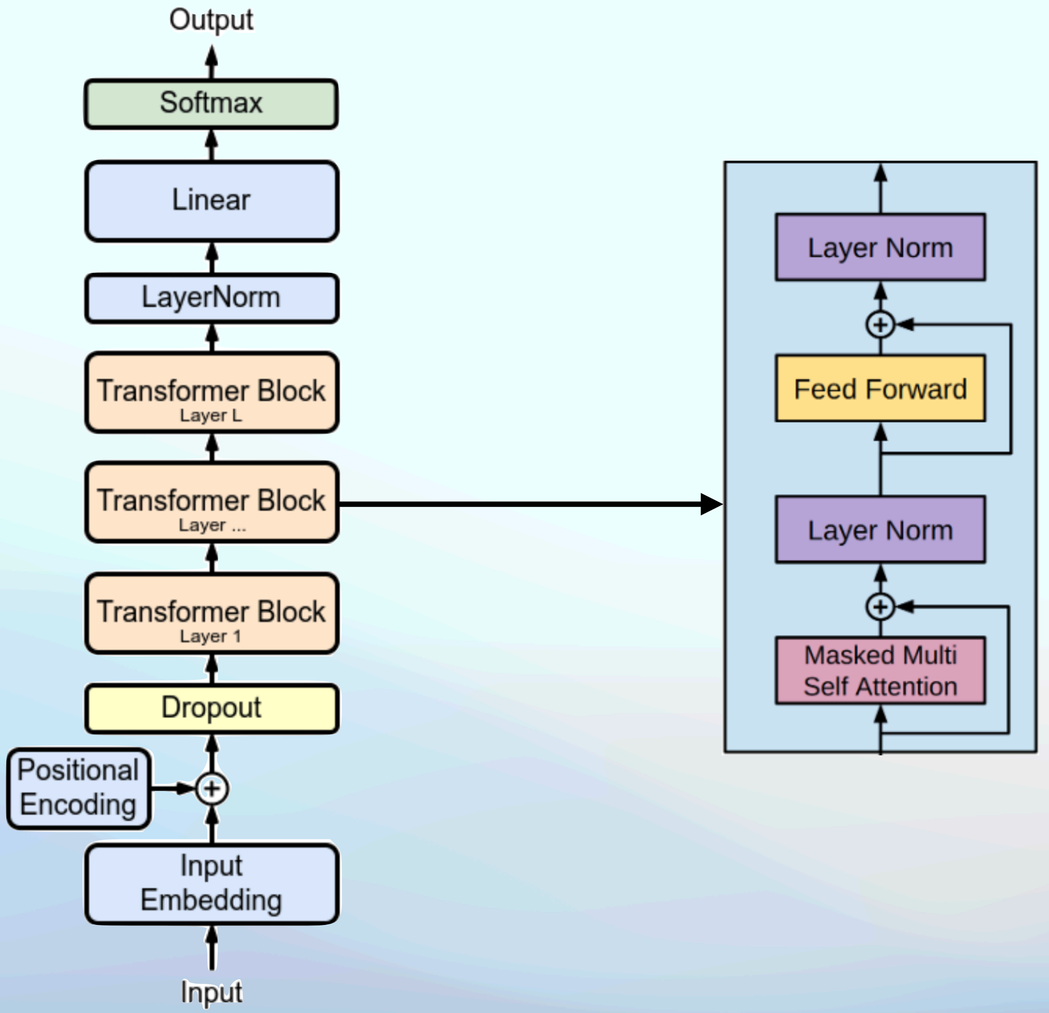
Including Multiple Multi-head Attention



- This is the architecture of Vanilla-Transformer model, the first Transformer model
- This picture shows only one encoder block and one decoder block, but actually has 6 encoder blocks and 6 decoder blocks.
- Multiple blocks lead to many parameters and big computing pressure. But only this can cope with the high complexity and variability of natural language.

GPT Model

Generative Pre-trained Transformer



- Only uses the decorder block of Transformer in last page.
- But has much more blocks, so have much more parameters
- The newest GPT-4 has 100 trillion parameters, around $5*10^5$ (half million) times of Vanilla-Transformer model
- Copilot use GPT-3.5, having 1.3 billion parameters

Thanks