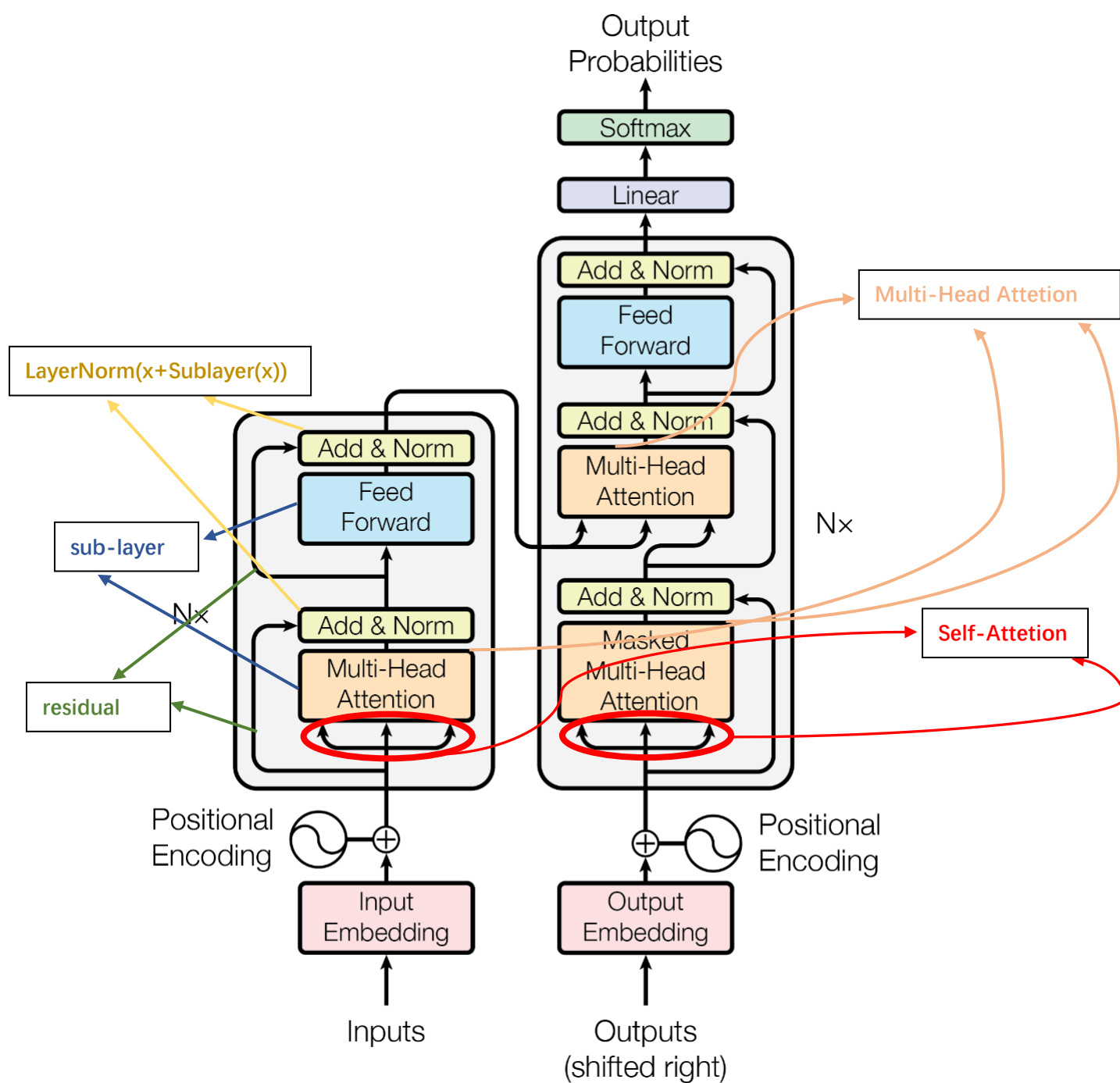
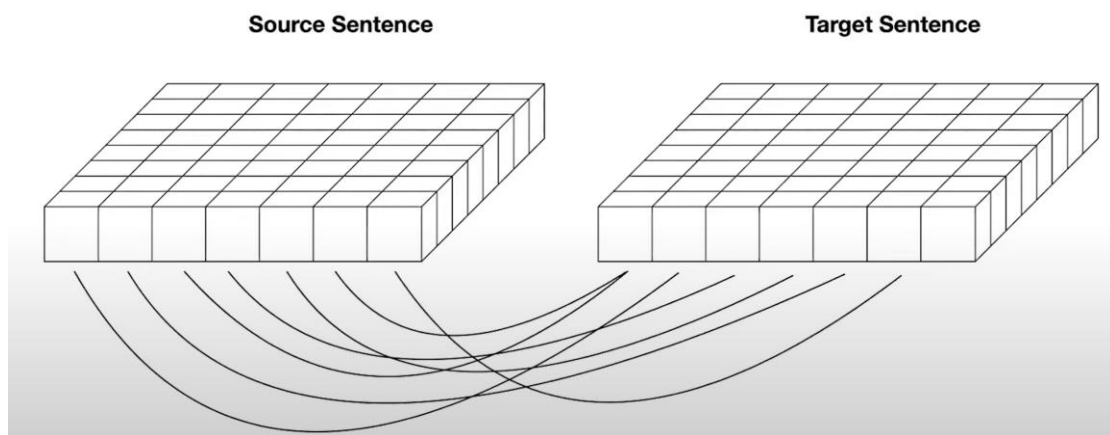


Attention Is All You Need

提出了 multi-head attention, self-attention

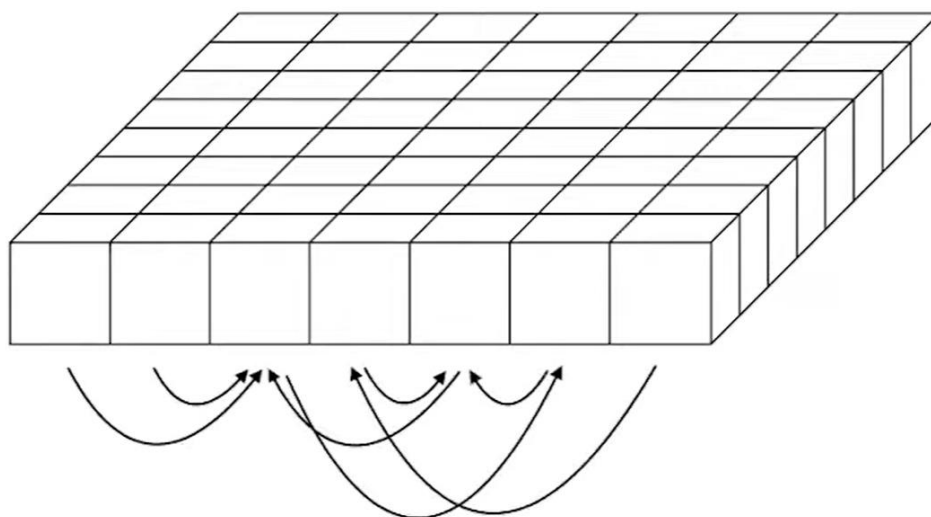
combines attention with fast autoregressive decoding





- **Attention relates** elements in a **source** sentence **to a target** sentence

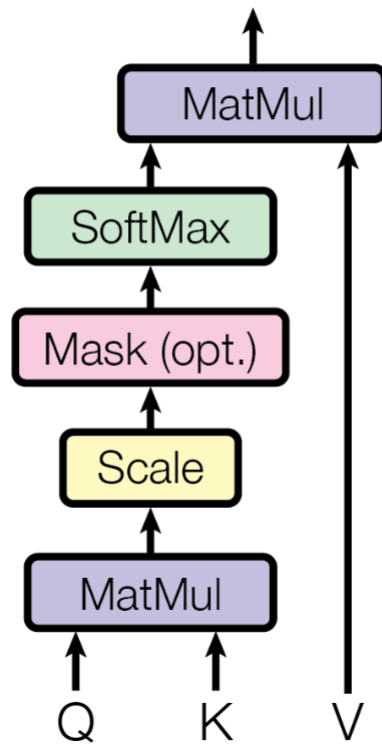
Sentence



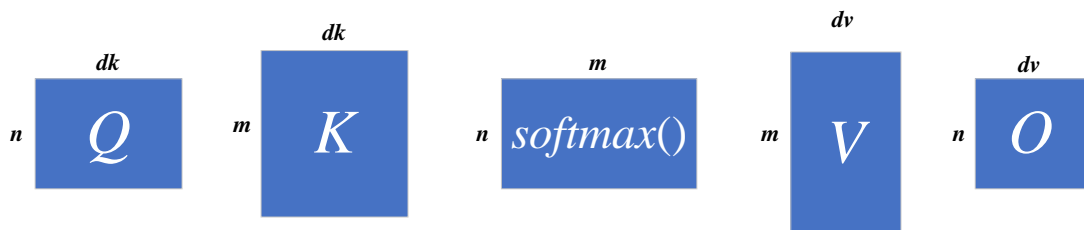
- **Self-Attention** is when your **source and target** are the same

学习到的是一个句子中单词之间的联系

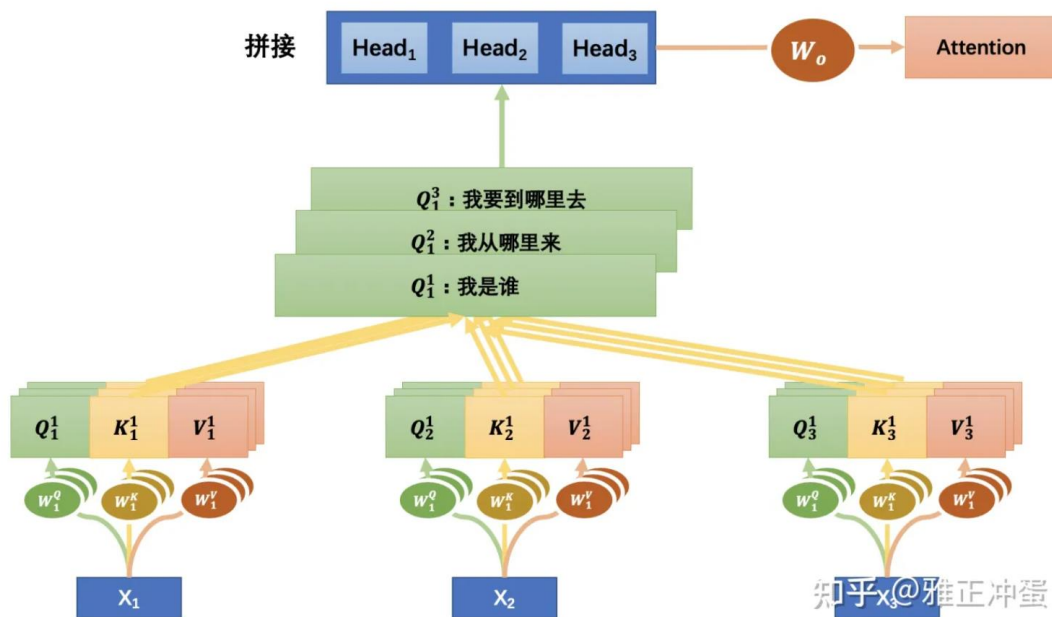
Scaled Dot-Product Attention



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Multi-headed Attention



- Compute k attentions in parallel
- Allows more than one relation

Layer Normalization

$$\mu = \frac{1}{T_x} \sum_{i=1}^{T_x} A_i, \quad A_i \in R^{[1,512]}$$

$$\sigma = \sqrt{\frac{1}{T_x} \sum_{i=1}^{T_x} (A_i - \mu)^2}$$

$$LayerNorm(A) = \frac{\mathbf{g}}{\sigma} \odot (A - \mu) + \mathbf{b}$$

全连接层

$$FFN(A) = \max\{0, AW_1 + b\}W_2 + b_2$$

Autoregressive decoding:

- condition each output on all previously generated outputs:

$$\begin{aligned}\hat{y}_0 &= \text{decoder}(\text{encoder}(x_0, \dots, x_N)) \\ \hat{y}_1 &= \text{decoder}(\hat{y}_0, \text{encoder}(x_0, \dots, x_N)) \\ &\vdots \\ \hat{y}_{t+1} &= \text{decoder}(\hat{y}_0, \hat{y}_1, \dots, \hat{y}_t, \text{encoder}(x_0, \dots, x_N))\end{aligned}$$

At train time we have access to the true target outputs: y_0, \dots, y_T

**But we still need to
run decoder T times**

$$\begin{aligned}\hat{y}_0 &= \text{decoder}(\text{encoder}(x_0, \dots, x_N)) \\ \hat{y}_1 &= \text{decoder}(y_0, \text{encoder}(x_0, \dots, x_N)) \\ &\vdots \\ \hat{y}_{t+1} &= \text{decoder}(y_0, y_1, \dots, y_t, \text{encoder}(x_0, \dots, x_N))\end{aligned}$$

注意力机制每次可以看到全部输入，用 mask 来屏蔽 t 之后的输入，从而保证预测和训练的行为一致