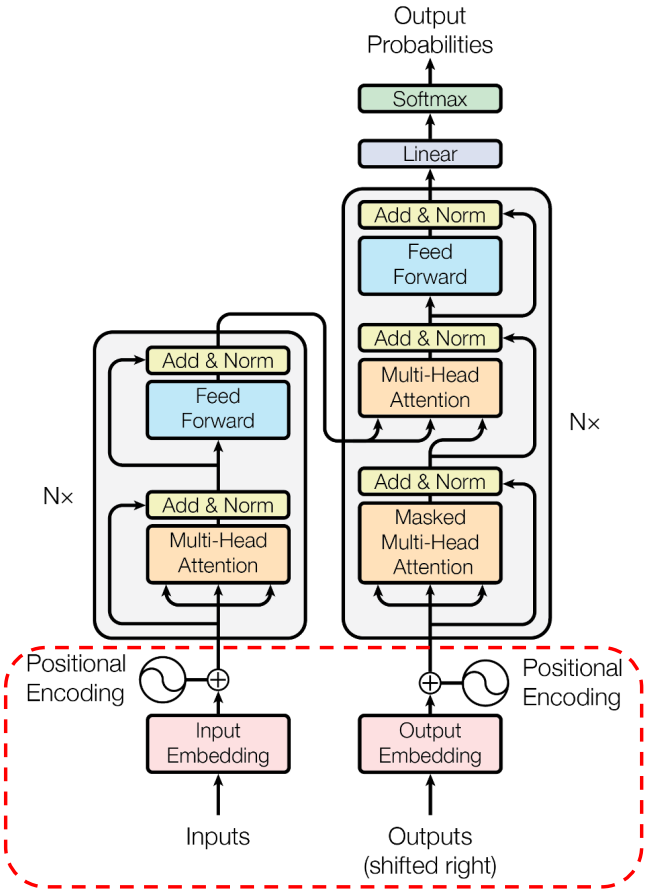
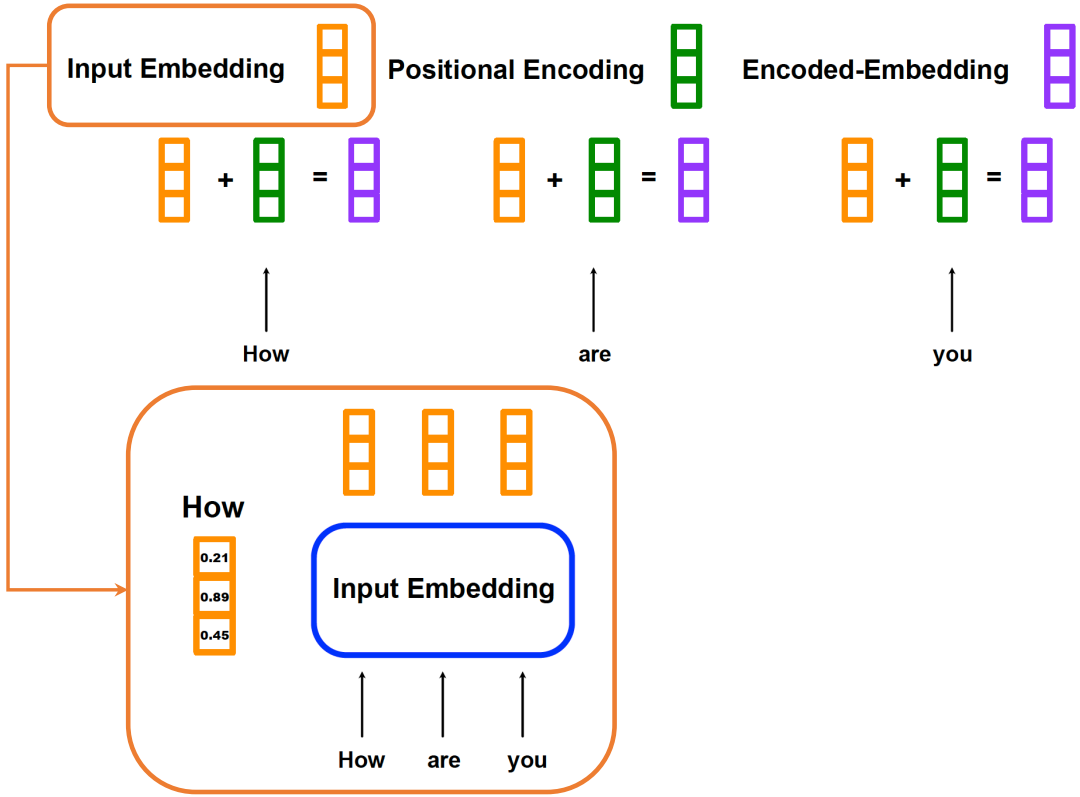
Summary

Mingyu (Jerry) Liu 4th Sep

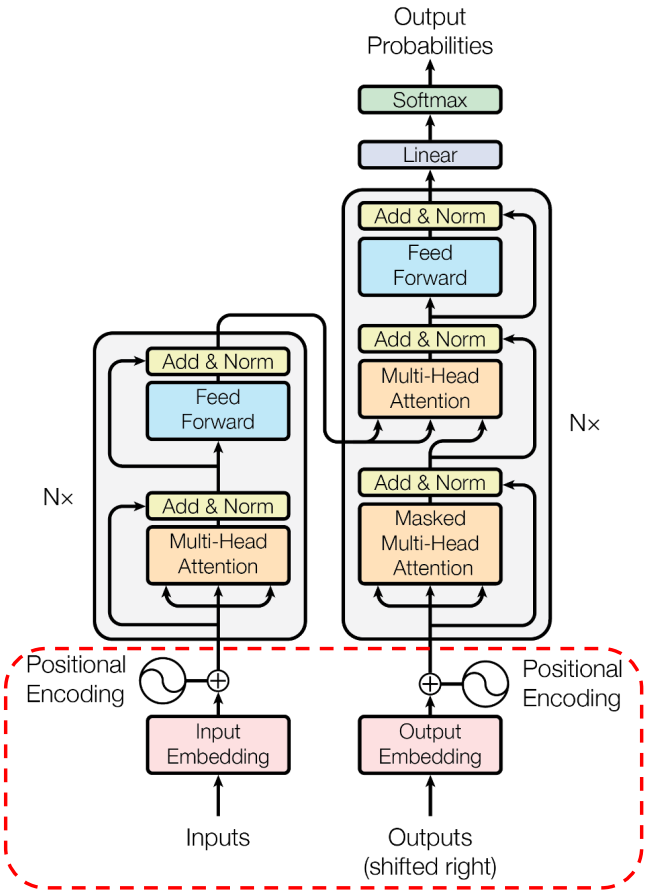
1. Embedding



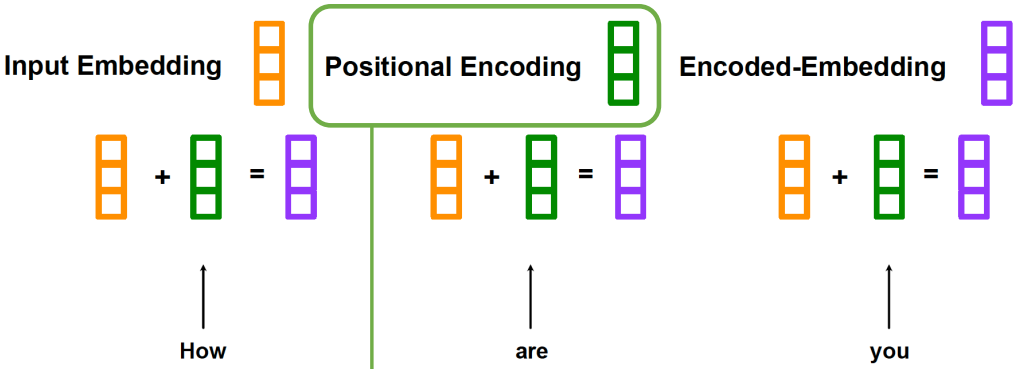
1.2 Input Embedding



1. Embedding



1.2 Positional Encoding

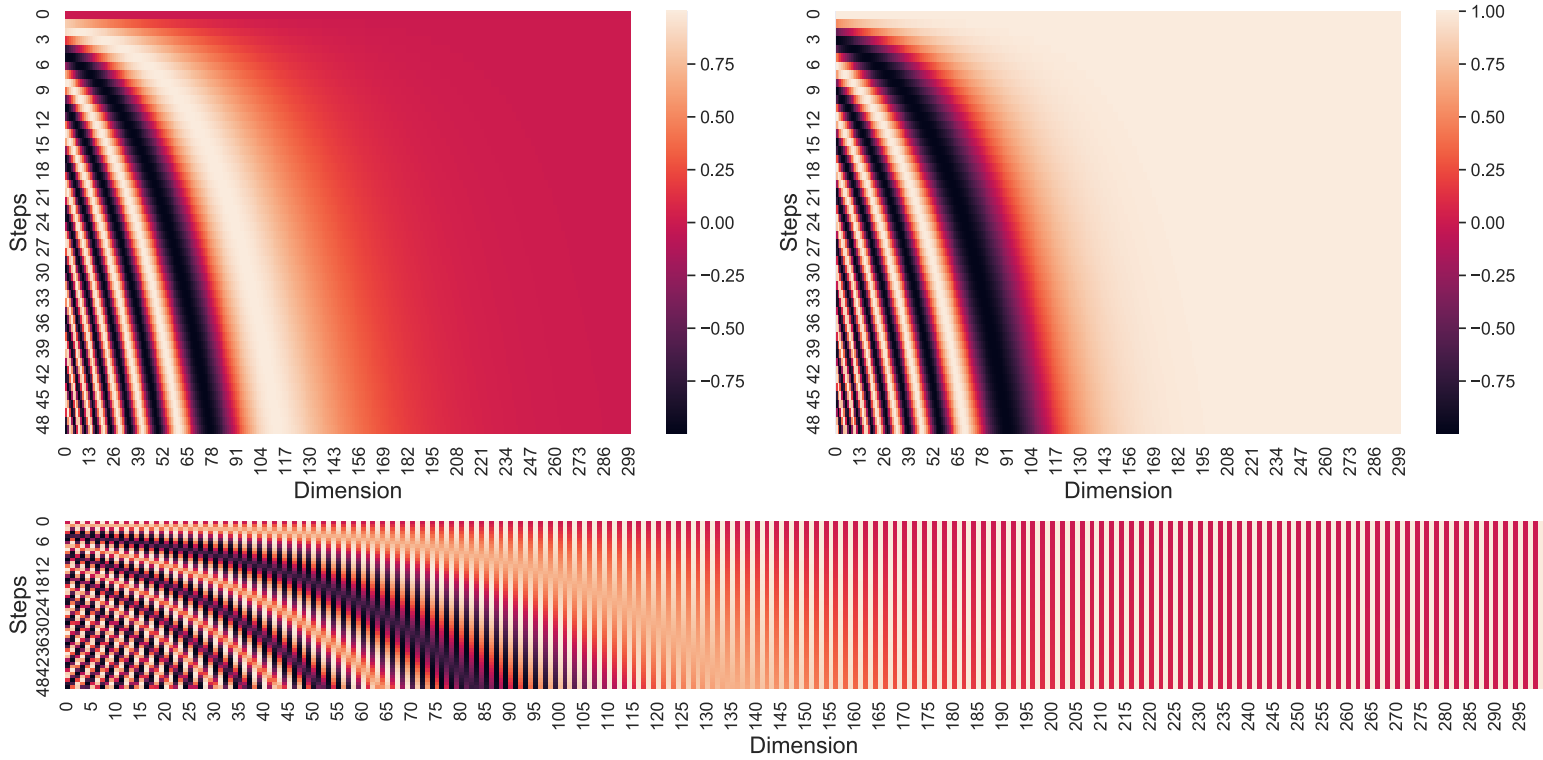


$$\vec{p}_t = \begin{bmatrix} \sin(\omega_0 \cdot t) \\ \cos(\omega_0 \cdot t) \\ \sin(\omega_1 \cdot t) \\ \cos(\omega_1 \cdot t) \\ \vdots \\ \sin(\omega_{d_{\text{emb}}/2} \cdot t) \\ \cos(\omega_{d_{\text{emb}}/2} \cdot t) \end{bmatrix}_{d_{\text{emb}} \times 1} \quad \text{where} \quad \omega_i = \frac{1}{10000^{2i/d_{\text{emb}}}}$$

$$\vec{p}_{(t,2i+1)} = \cos\left(\frac{t}{10000^{2i/d_{\text{emb}}}}\right) \qquad \vec{p}_{(t,2i)} = \sin\left(\frac{t}{10000^{2i/d_{\text{emb}}}}\right)$$

1. Embedding

1.2 Positional Encoding – Visualized

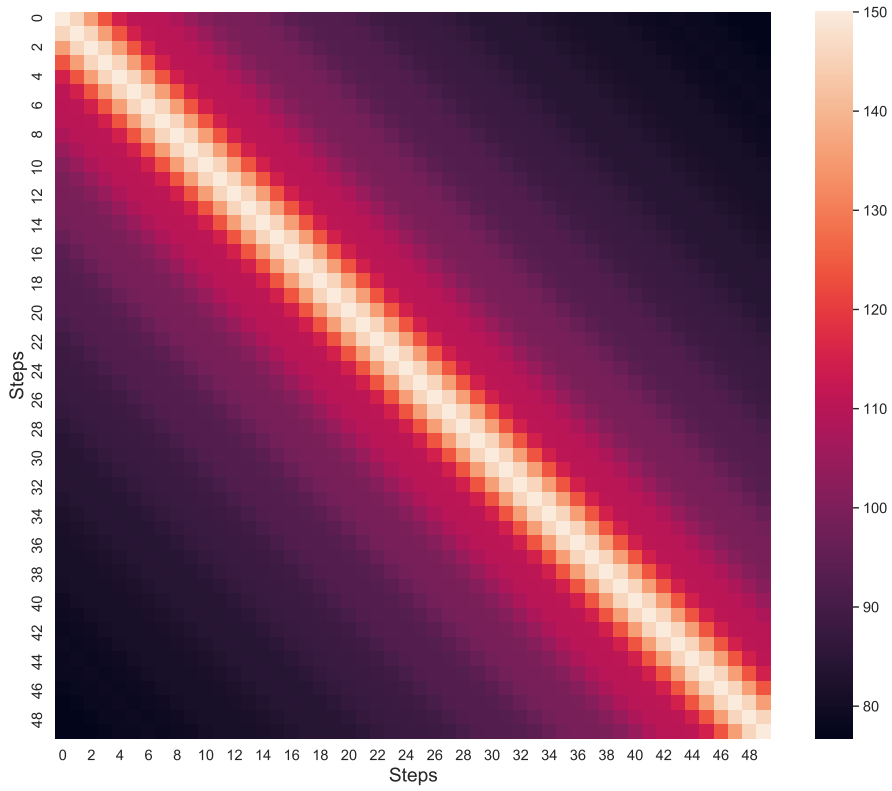


1. Embedding

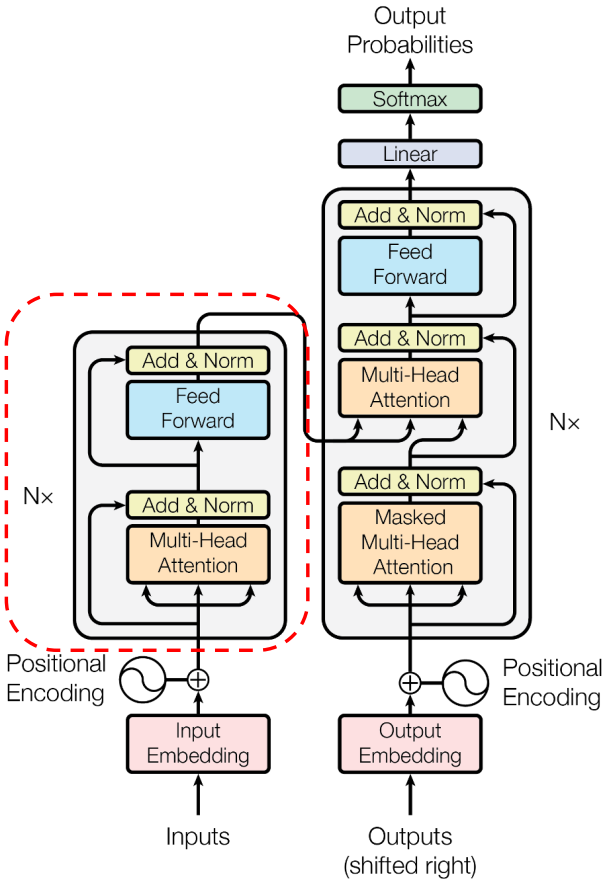
1.2 Positional Encoding – Properties

Property 1: The encoding $\vec{p}(t+\phi)$ can be represented as a time-independent linear transformation of $\vec{p}(t)$ for any offset ϕ

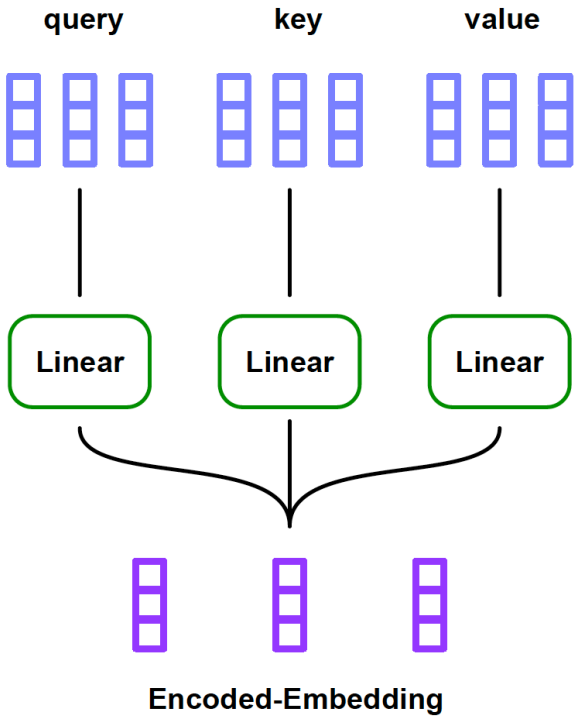
Property 2: Neighbouring time-steps' dot-products are symmetrical and decay nicely with time



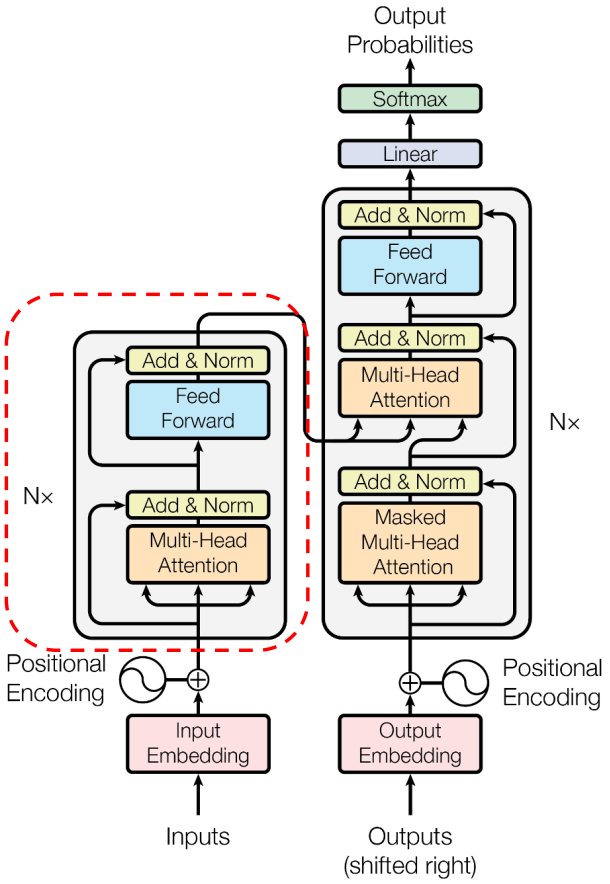
2. Encoder



2.1 Creation of query/key/value vector

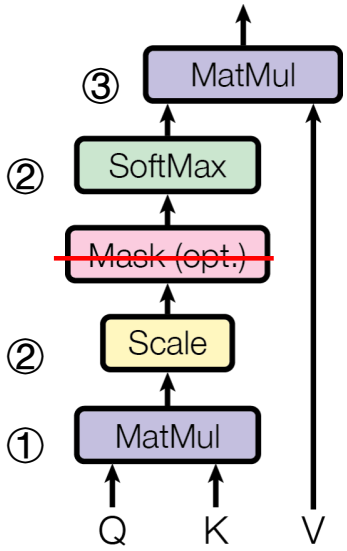


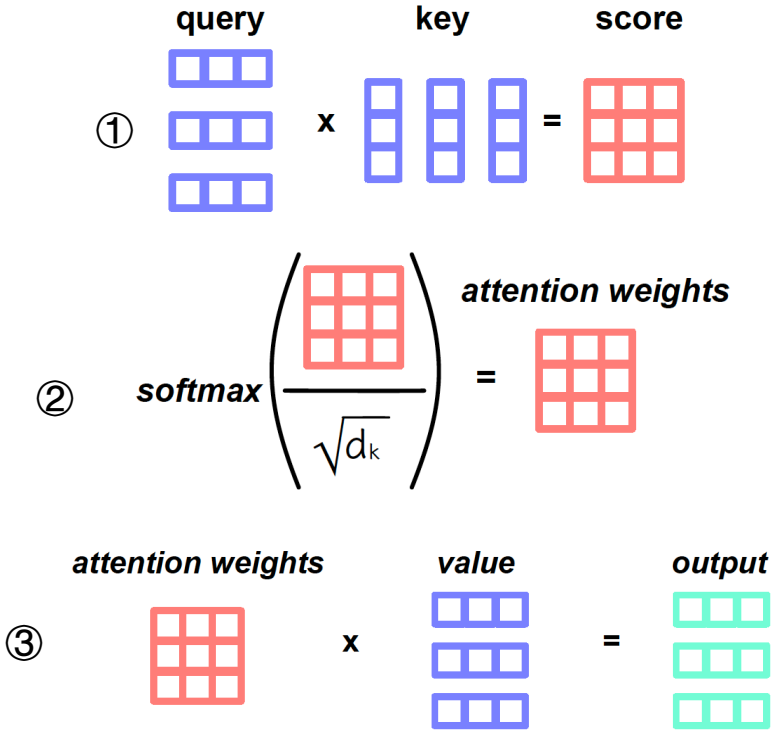
2. Encoder



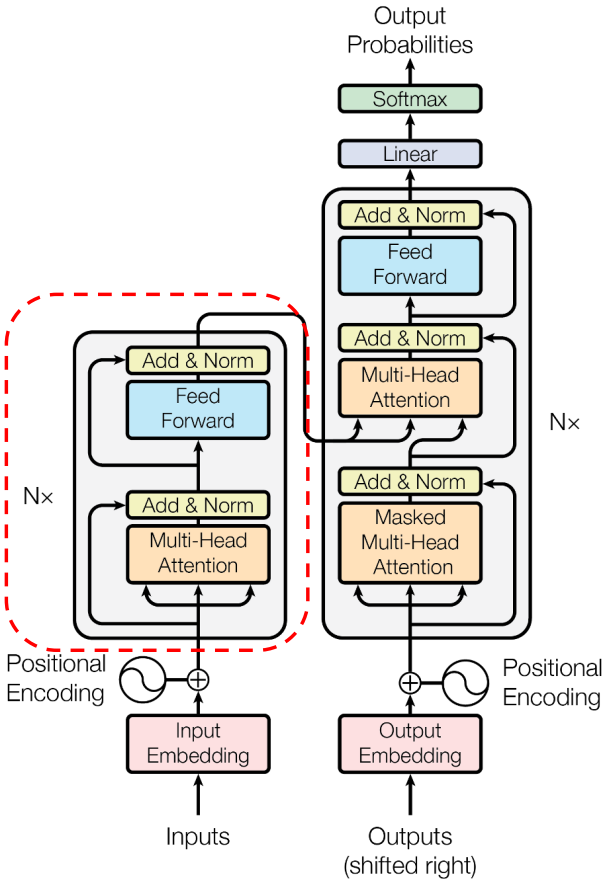
2.2 Multi-Headed Attention

- Self Attention



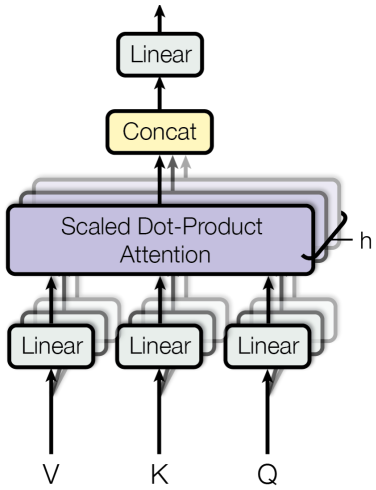


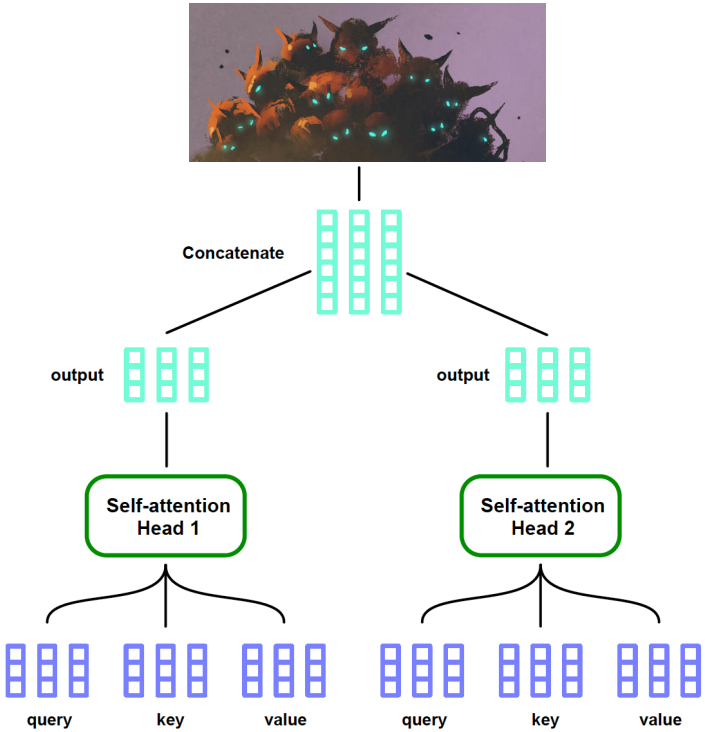
2. Encoder



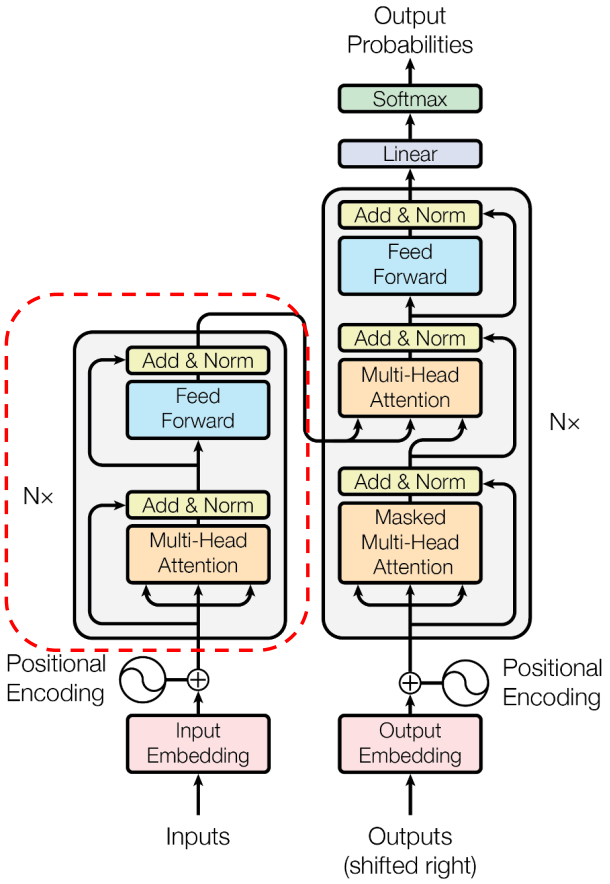
2.2 Multi-Headed Attention

- Multi-head

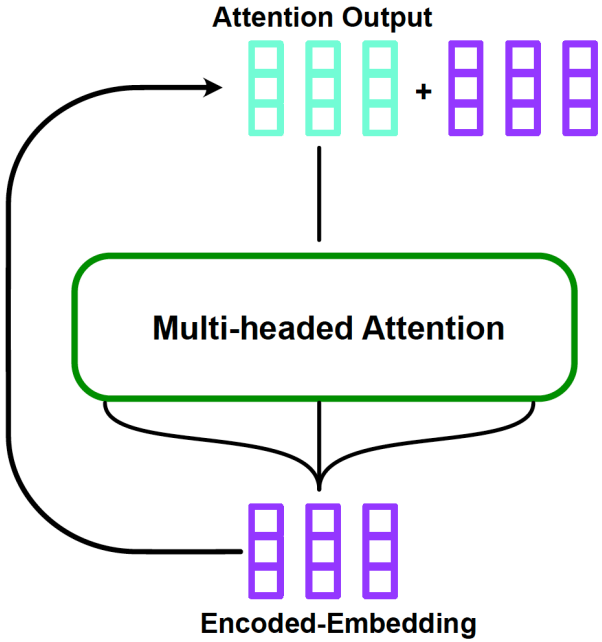




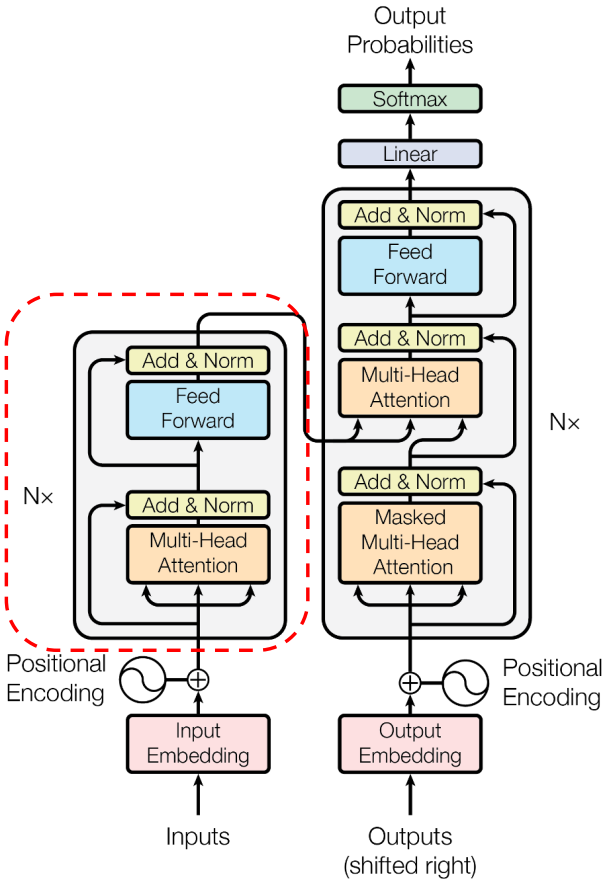
2. Encoder



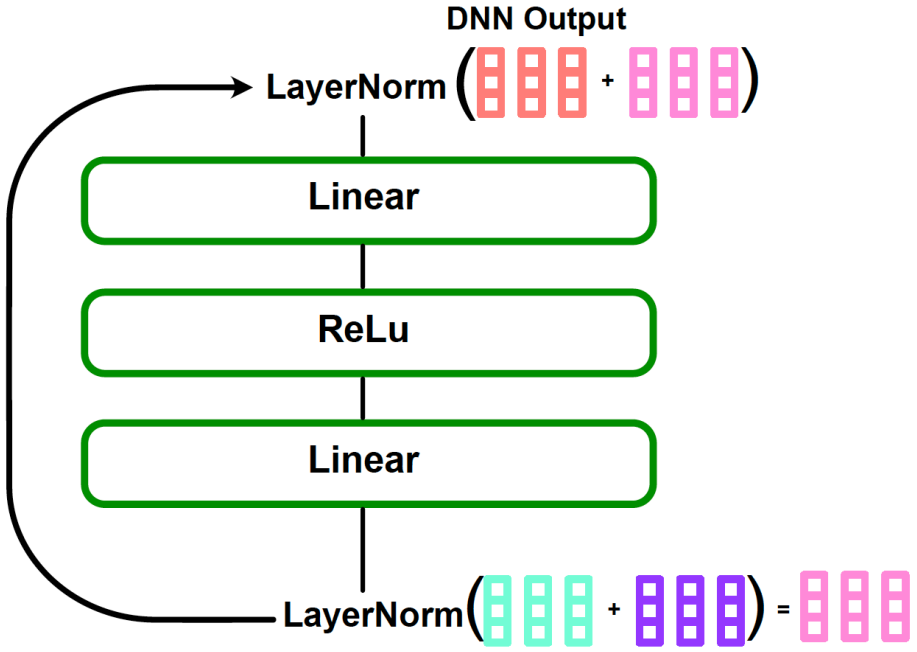
2.3 Residual Connection



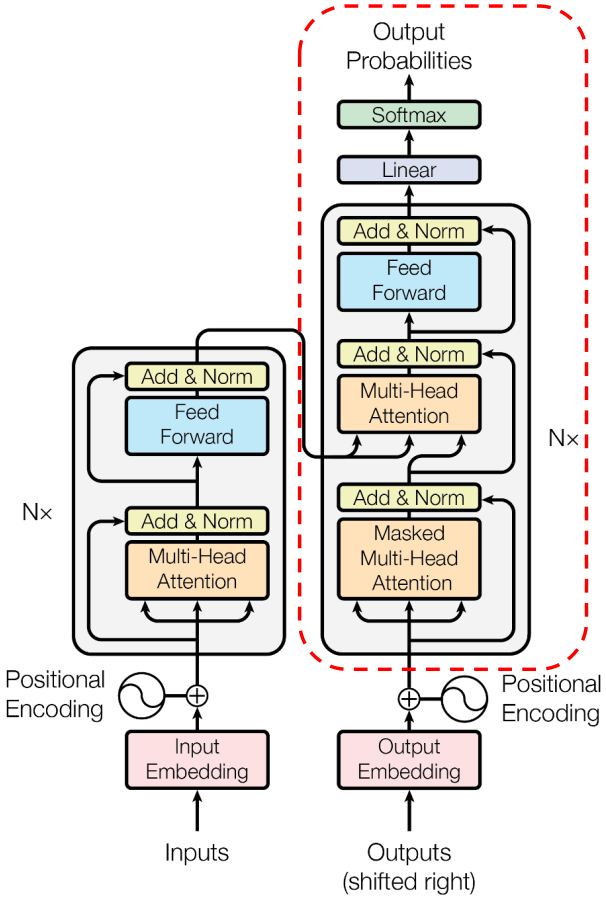
2. Encoder



2.4 Layer Normalization & Feed-forward



3. Decoder



3.1 Masked Self-Attention