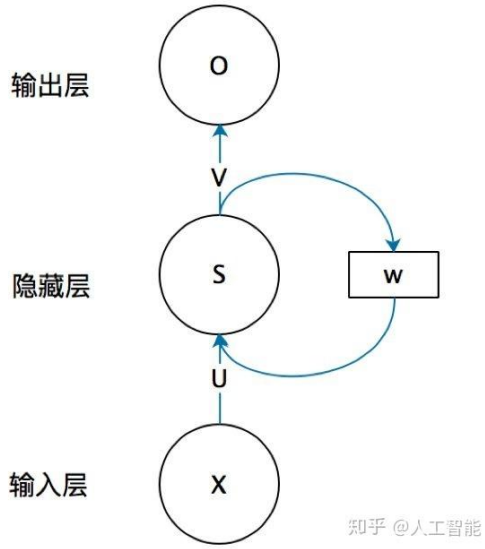
Transformer

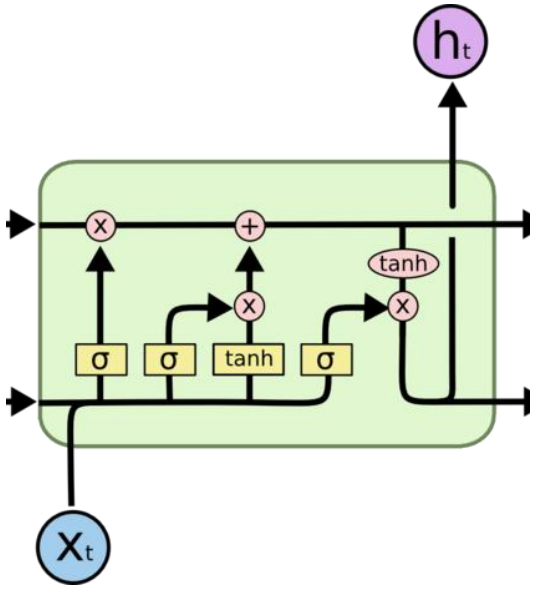
Reporter：Bowen Xu

Background

Traditional translation model



RNN



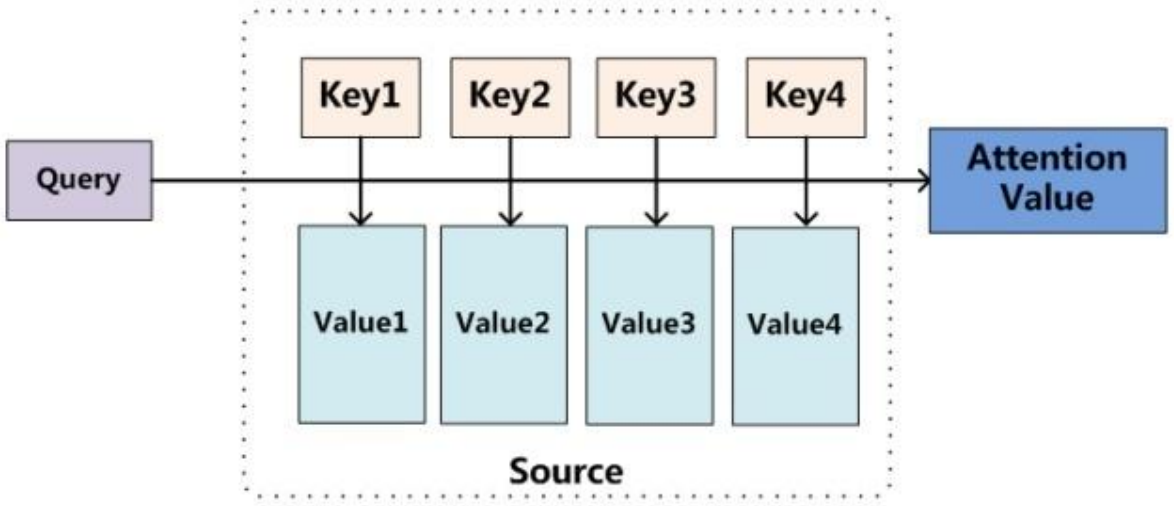
LSTM

Drawbacks:

- Sequential nature precludes **paralleliazation**
- **Information loss** in long term

Introduction

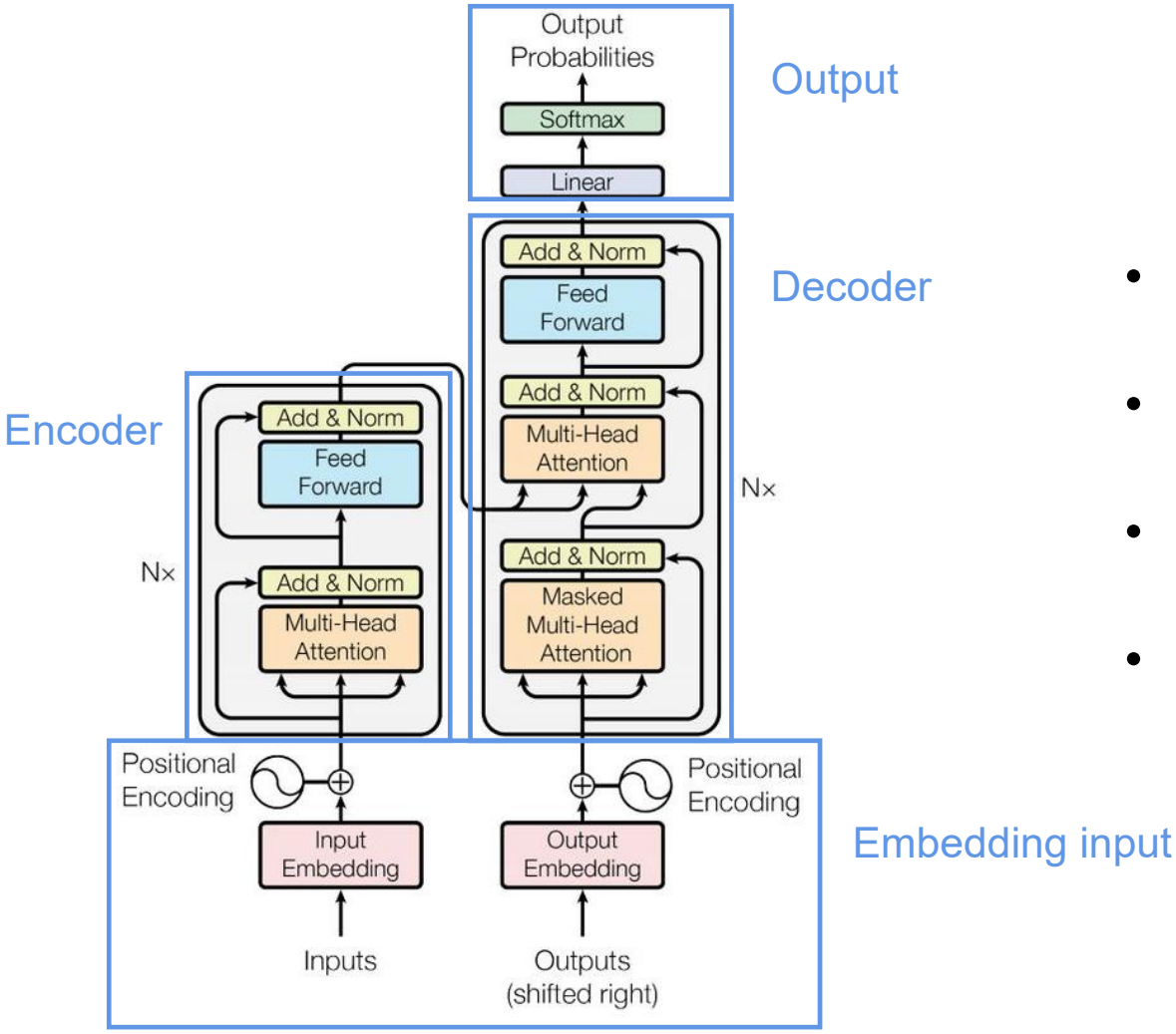
Based entirely on Attention Mechanism



Advantages:

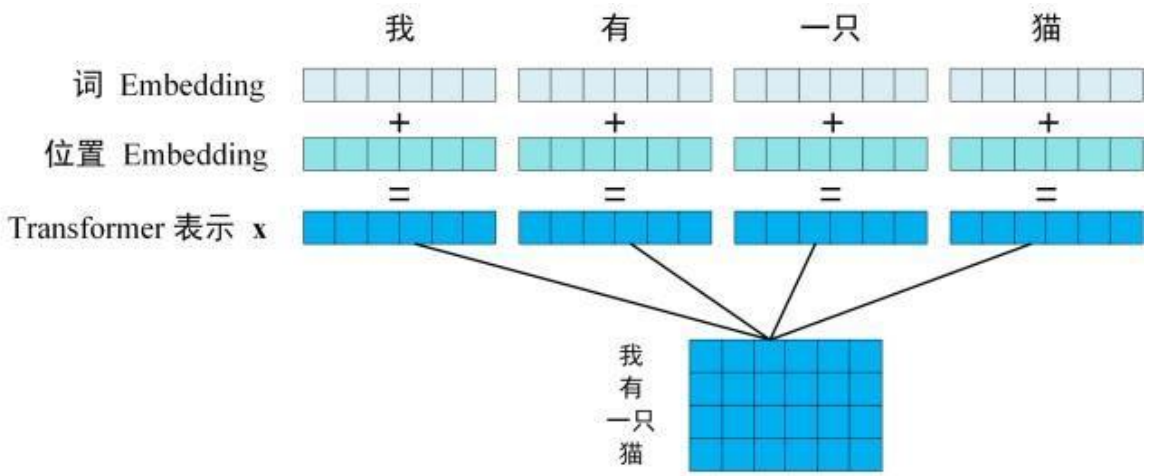
- Allowing for **parallelization**
- Allowing modeling of dependencies without regard to distance

Model Architecture



- Embedding input
- **Encoder**
- **Decoder**
- Output

Embedding Input



- Word Embedding: word2vec, GloVe, one-hot, etc.
- Positional Embedding:

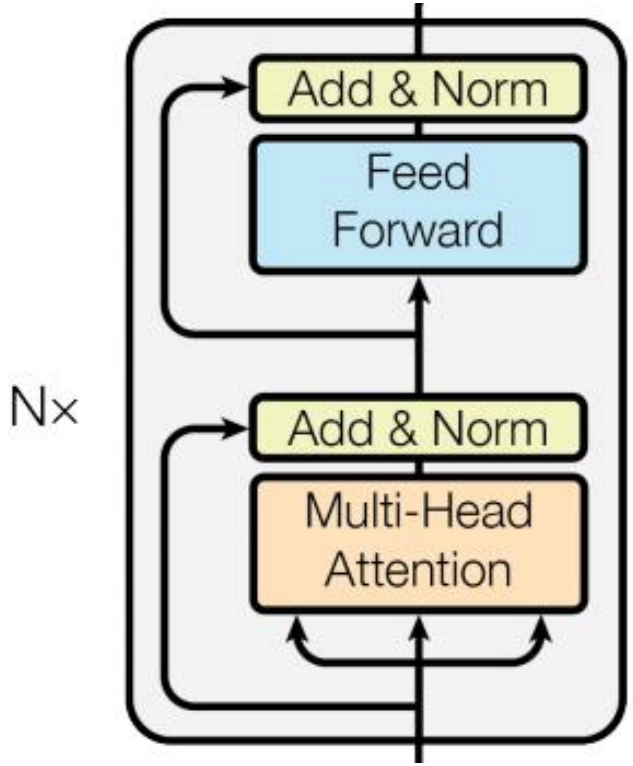
$$PE(pos, 2i) = sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

$$PE(pos, 2i+1) = cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

  - d_{model}: dimension of input
  - pos: position of words

Encoder

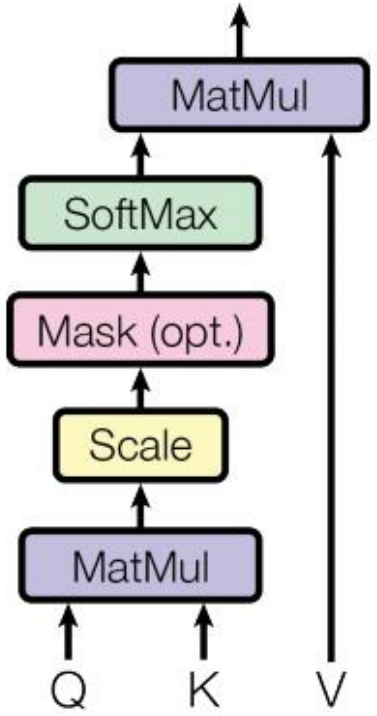
Encoder has N=6 identical layers



Structure of one layer:

- **Multi-Head Attention**
- Feed Forward
- Add & Norm

Self-Attention Mechanism

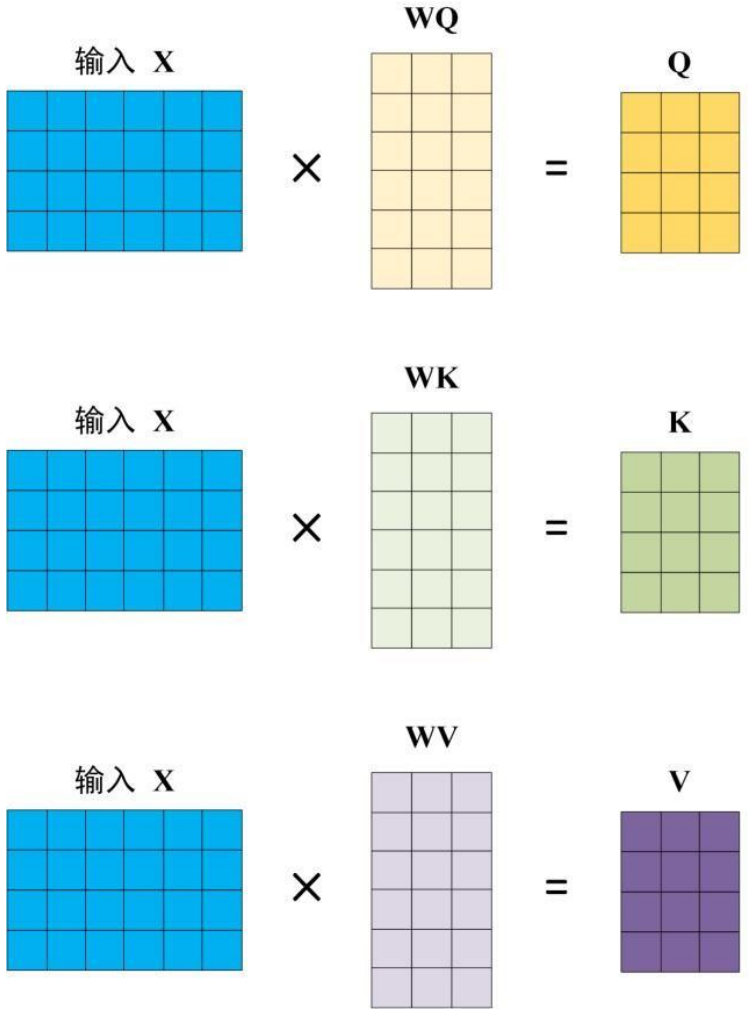


Self-Attention(without mask):

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

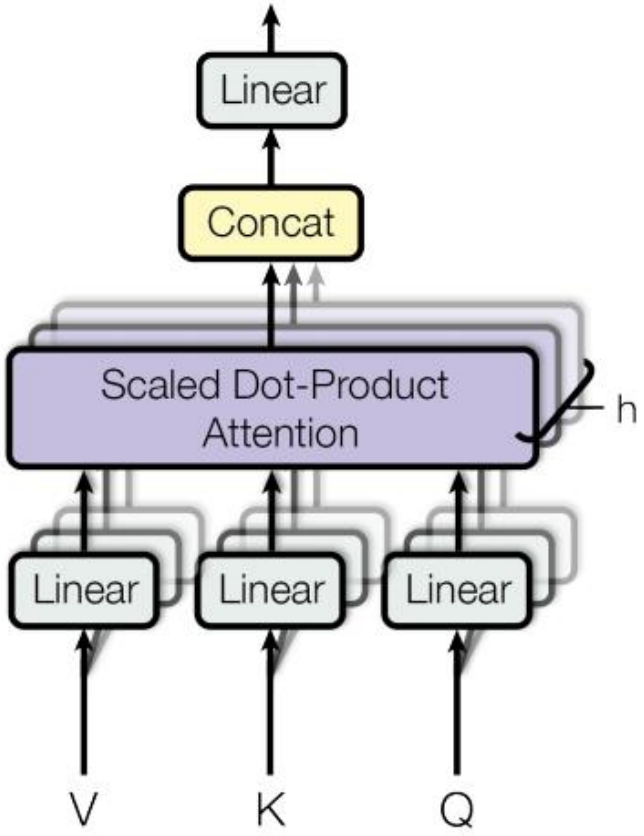
- d_k: Number of columns (dimension) of Q and K

Self-Attention Input



- X: Input of each layer of encoder
- W^Q, W^K, W^V: Linear transformation matrix

Multi-Head Attention



$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

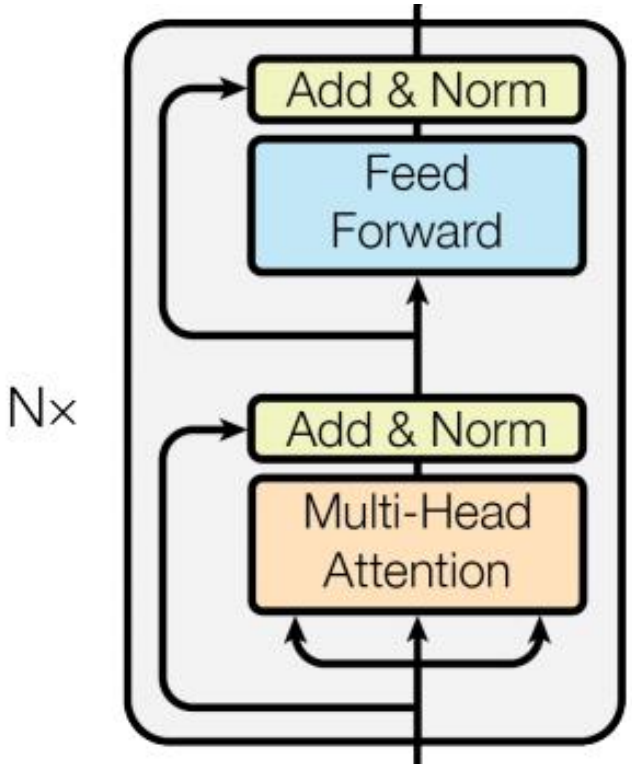
- where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

Multi-Head attention ensures the **parallelism**

Feed Forward

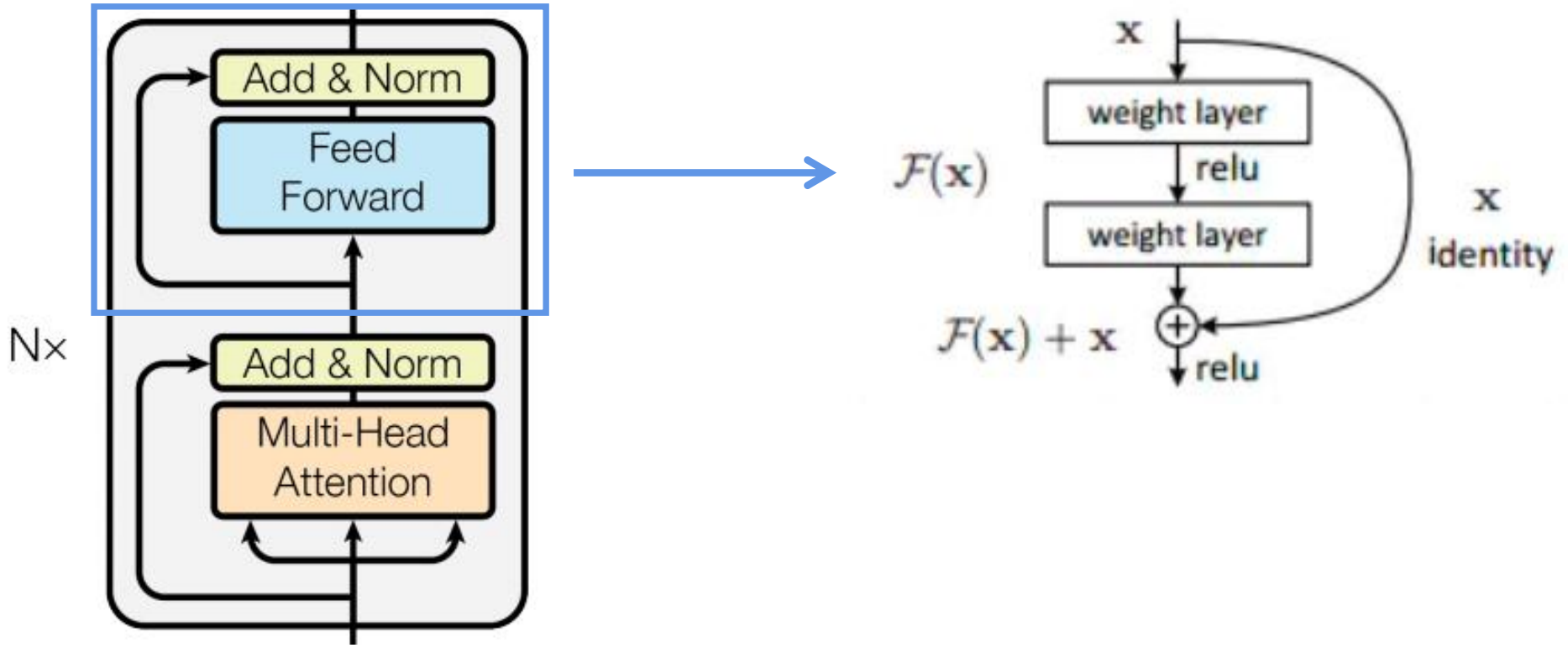
$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

Add & Norm

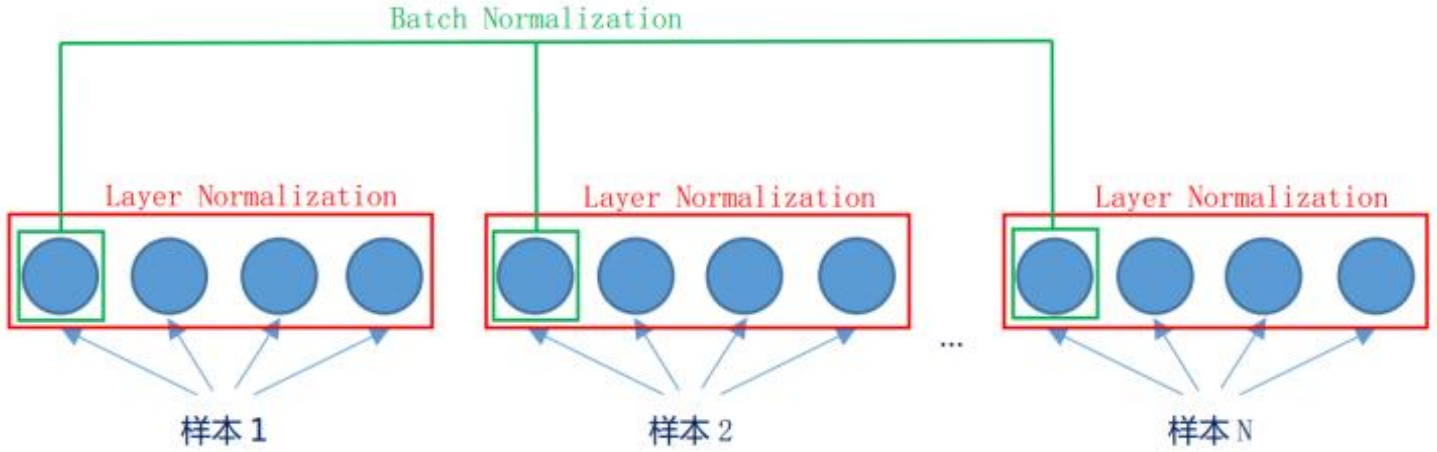


- Add: Residual connection
- Norm: Layer normlization

Residual Connection



Layer Normalization

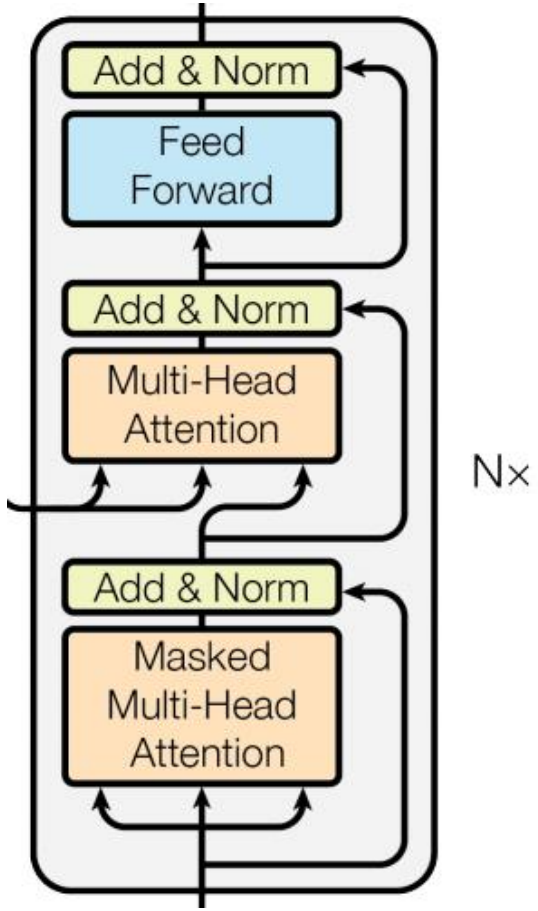


- **Layer norm**: Normalization **between different dimensions** in the same sample

- Batch norm: Normalization between different samples in the same dimension

Decoder

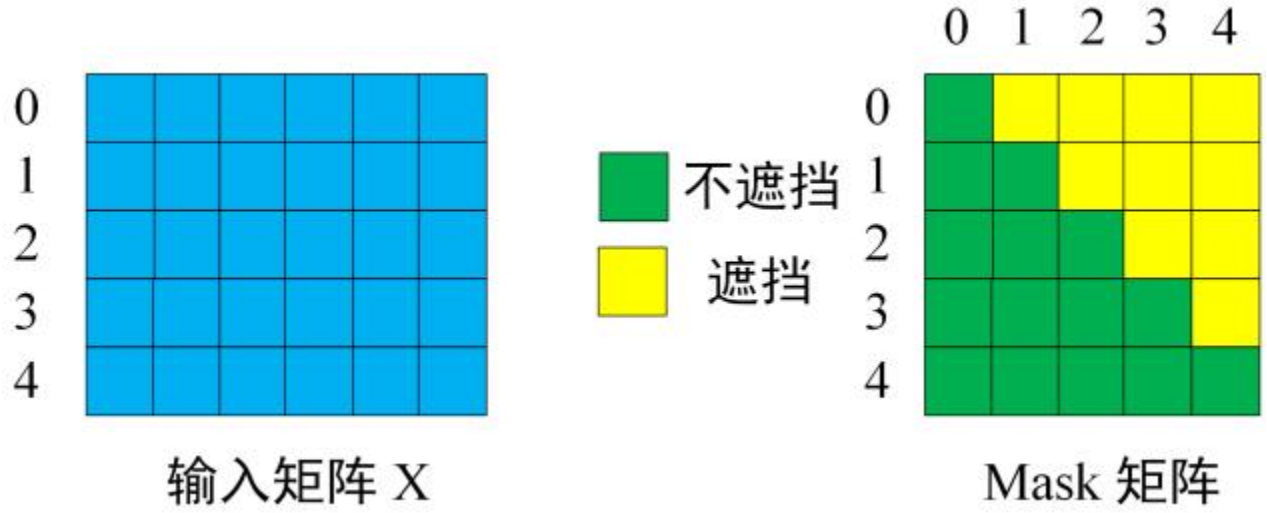
Decoder has N=6 identical layers



Structure of one layer:

- Masked Multi-Head Attention
- Multi-Head Attention
- Feed Forward
- Add & Norm

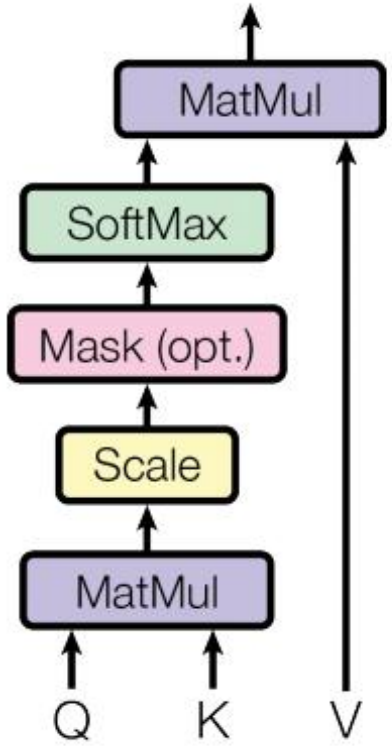
Mask Operation

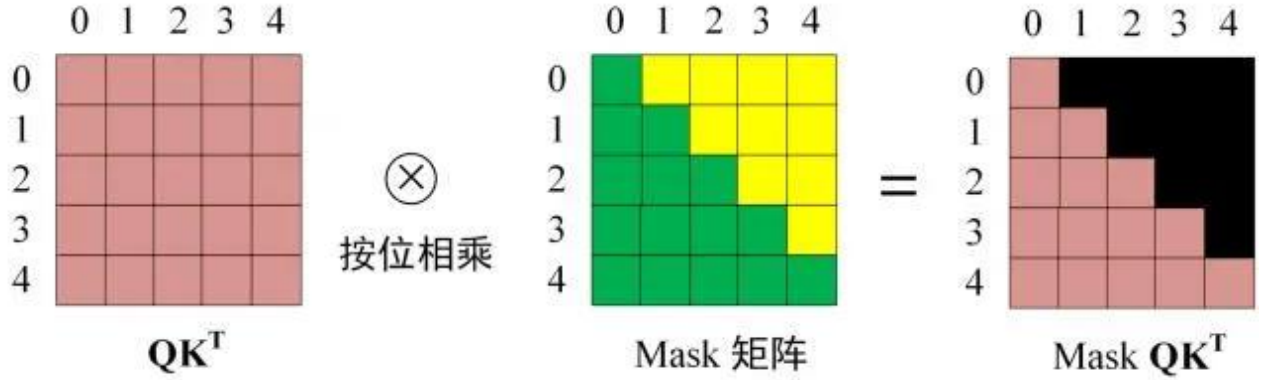


Sequence Mask:

- Why: The decoder cannot see the future information.
- How: Setting masked position to $-\infty$.

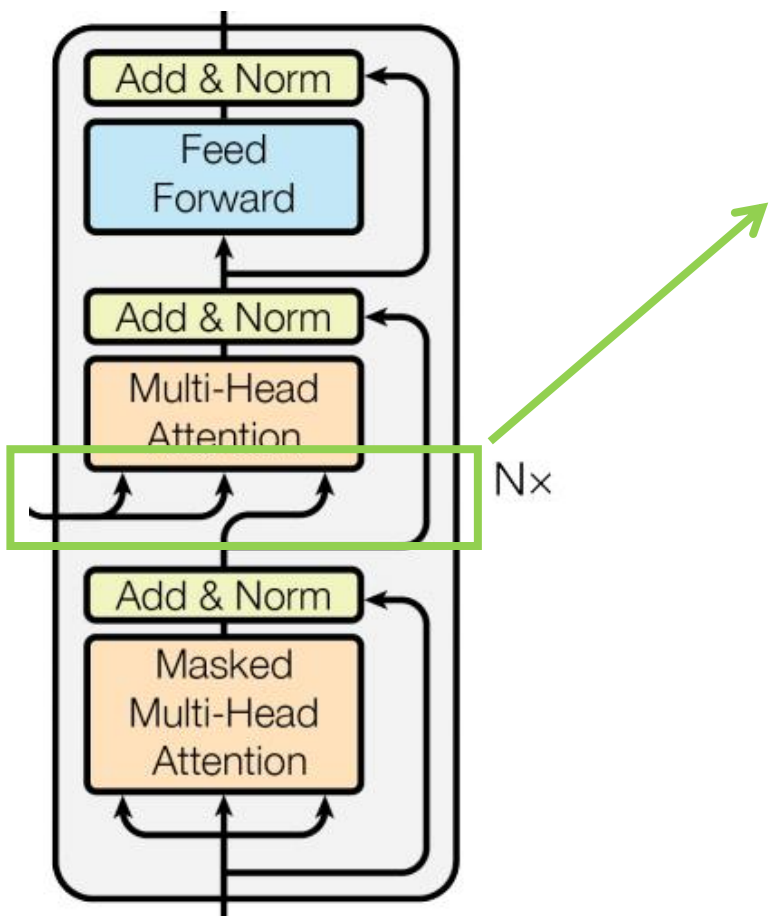
Masked Multi-Head Attention

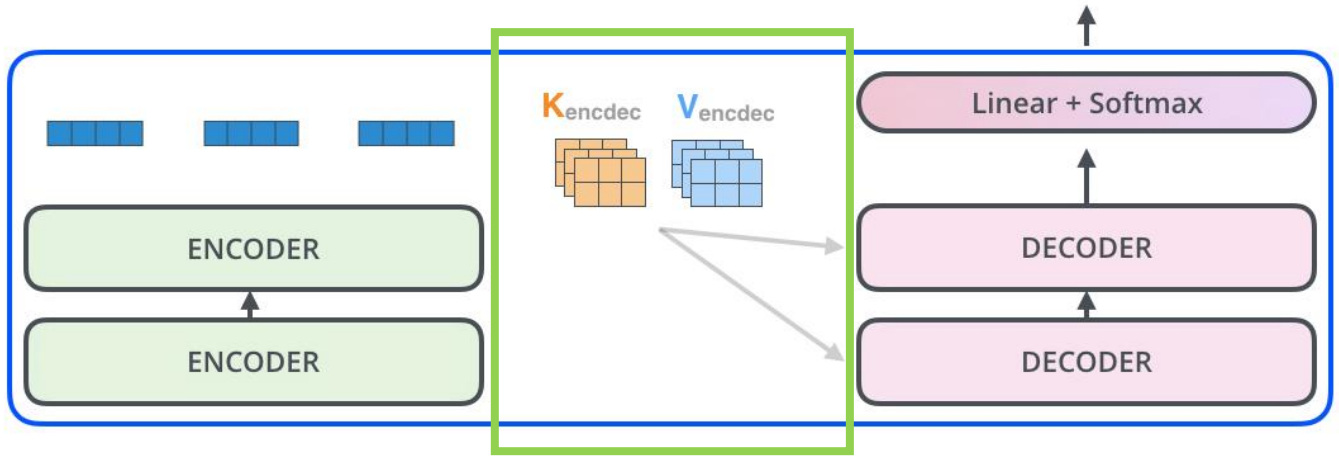




- Mask matrix is a lower triangular matrix (non-0 elements are all 1).

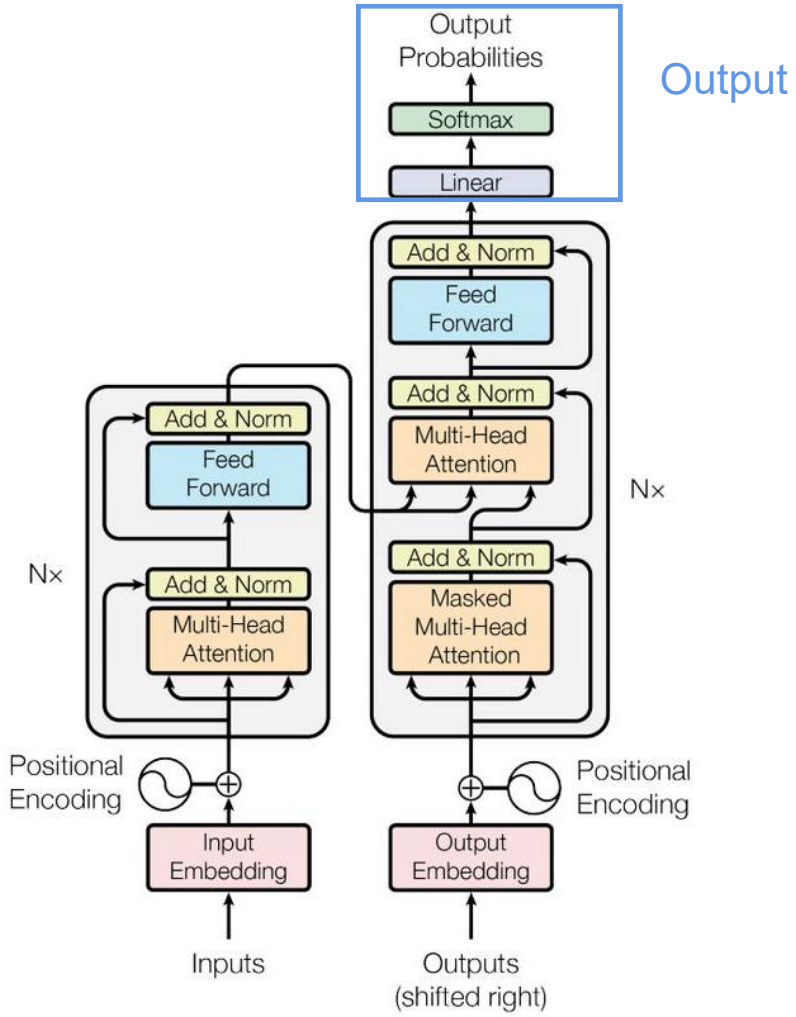
Multi-Head Attention In Decoder





- K, V: From the output of encoder
- Q: From the output of the previous decoder block

Output

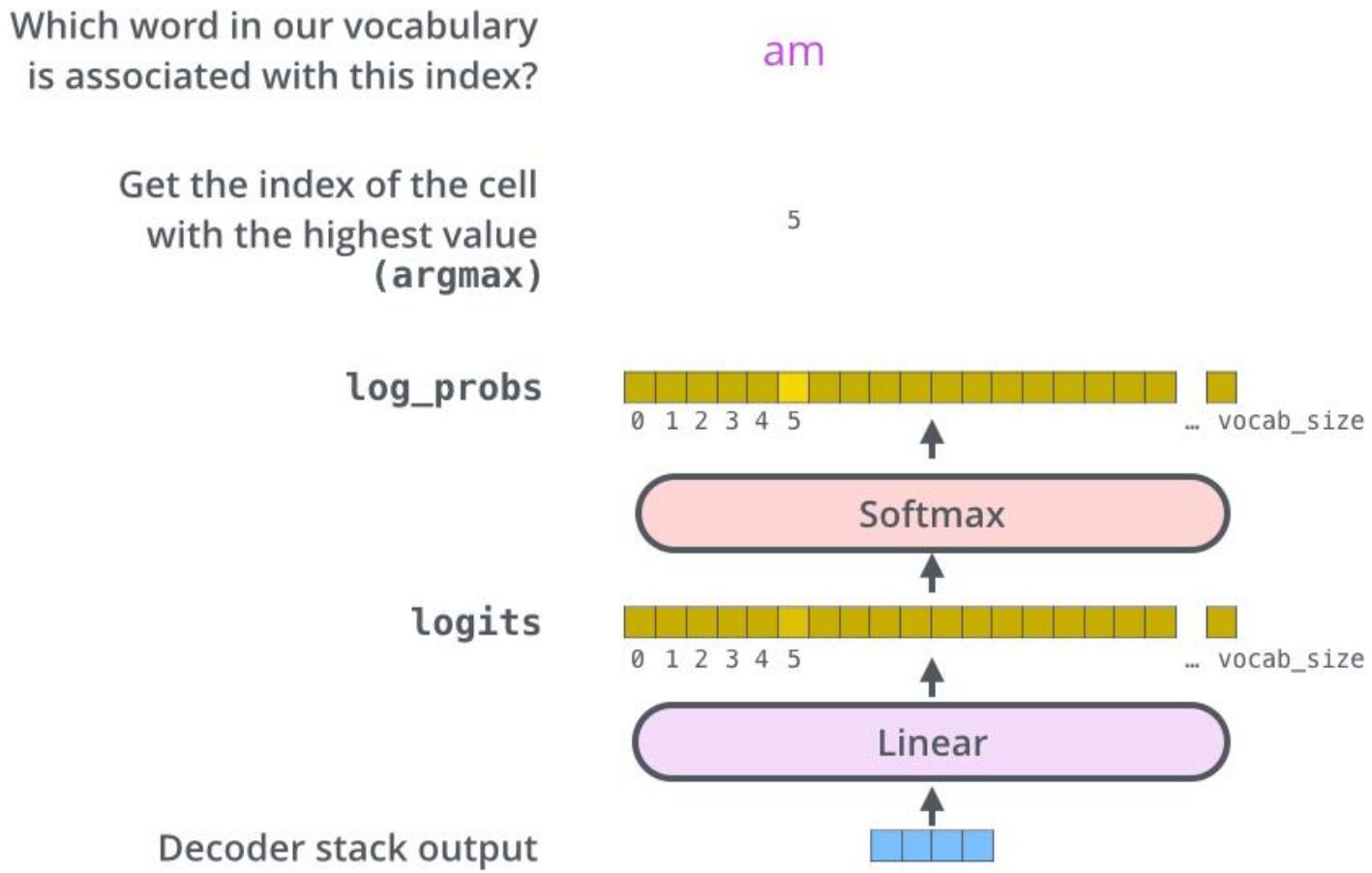


Structure of output block:

- Linear layer
- Softmax

Output

- How to predict a word:



Conclusion

Advantages：

- Avoid sequence model and can realize **parallelization**.
- Solves the problem of **long distance dependence**.
- Self-attention can produce more **interpretable models**.

Disadvantages:

- Loss of **position information** (Compare with RNN)
- Lost the ability to capture **local features** (Compare with CNN)

Transformer is widely used in Deep Learning fields (NLP, CV, etc.).


Thank
you