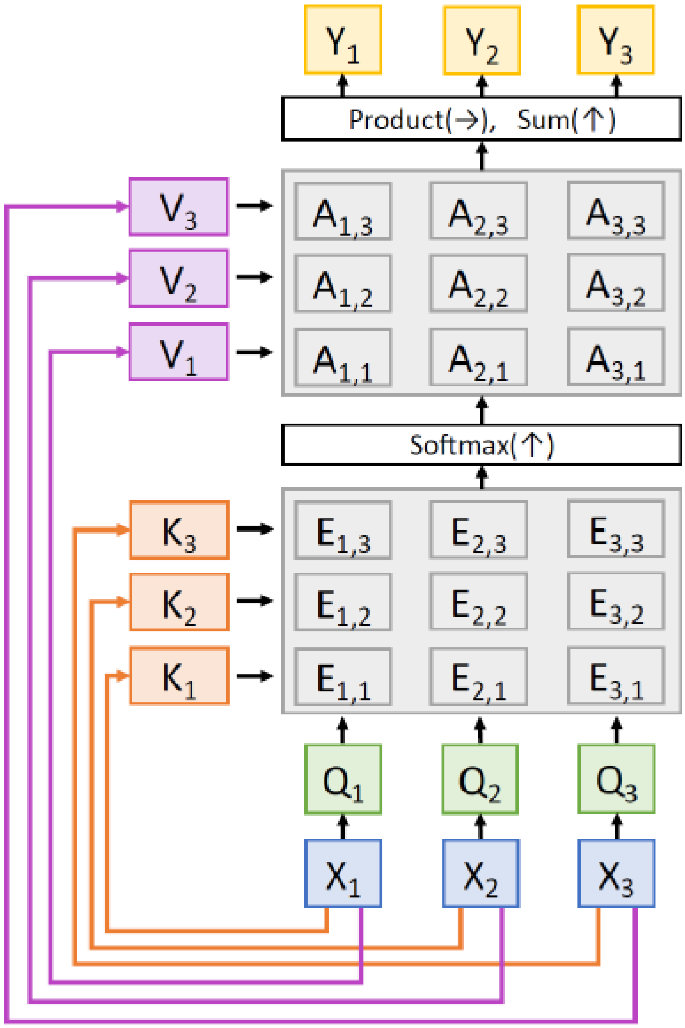
1. Self-attention



(a) Prepare Inouts：

X_1 shap： $1 \times D_x$

$X = [X_1^T, X_2^T, X_3^T]^T$, shape：$N_x \times D_x, N_x = 3$

(b) Initialize weights:

Key matrix W_k, shape：$D_X \times D_Q$

Query matrix W_Q, shape：$D_X \times D_Q$

Value matrix W_V, shape：$D_X \times D_V$

(c) Derive key, query and value:

Key $K = XW_k$, shape：$N_X \times D_Q$

Query $Q = XW_Q$, shape：$N_X \times D_Q$

Value $V = XW_V$, shape：$N_X \times D_V$

(d) Calculate attention scores for input：

$E^T = QK^T/\sqrt{D_Q}$, shape：$N_X \times N_X$

$E_{i,J} = Q_iK_i^T/\sqrt{D_Q}$, shape：$1 \times 1$

(e) Calculate softmax:

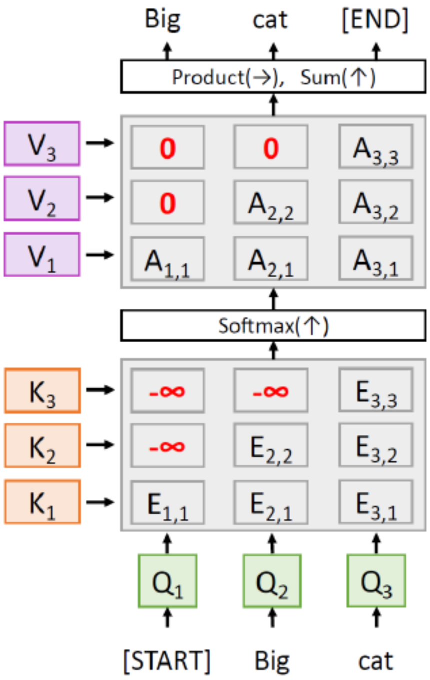
$A = softmax(E, dim = 1)$, shape：$N_X \times N_X$

(f) Multiply scores with values

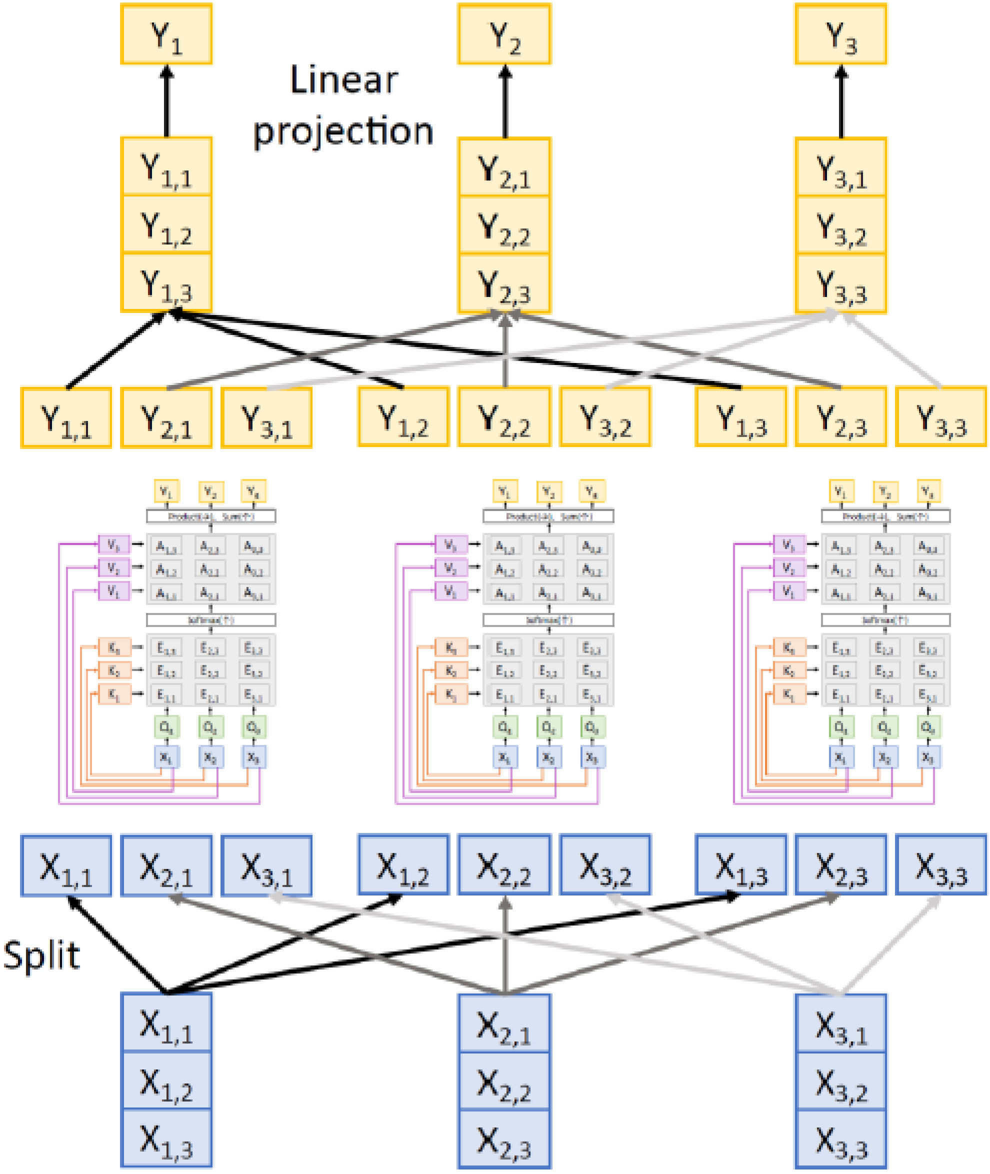
(g) Sum weighted values to get output

$$Y_i = \sum_i A_{i,j}, V_j$$

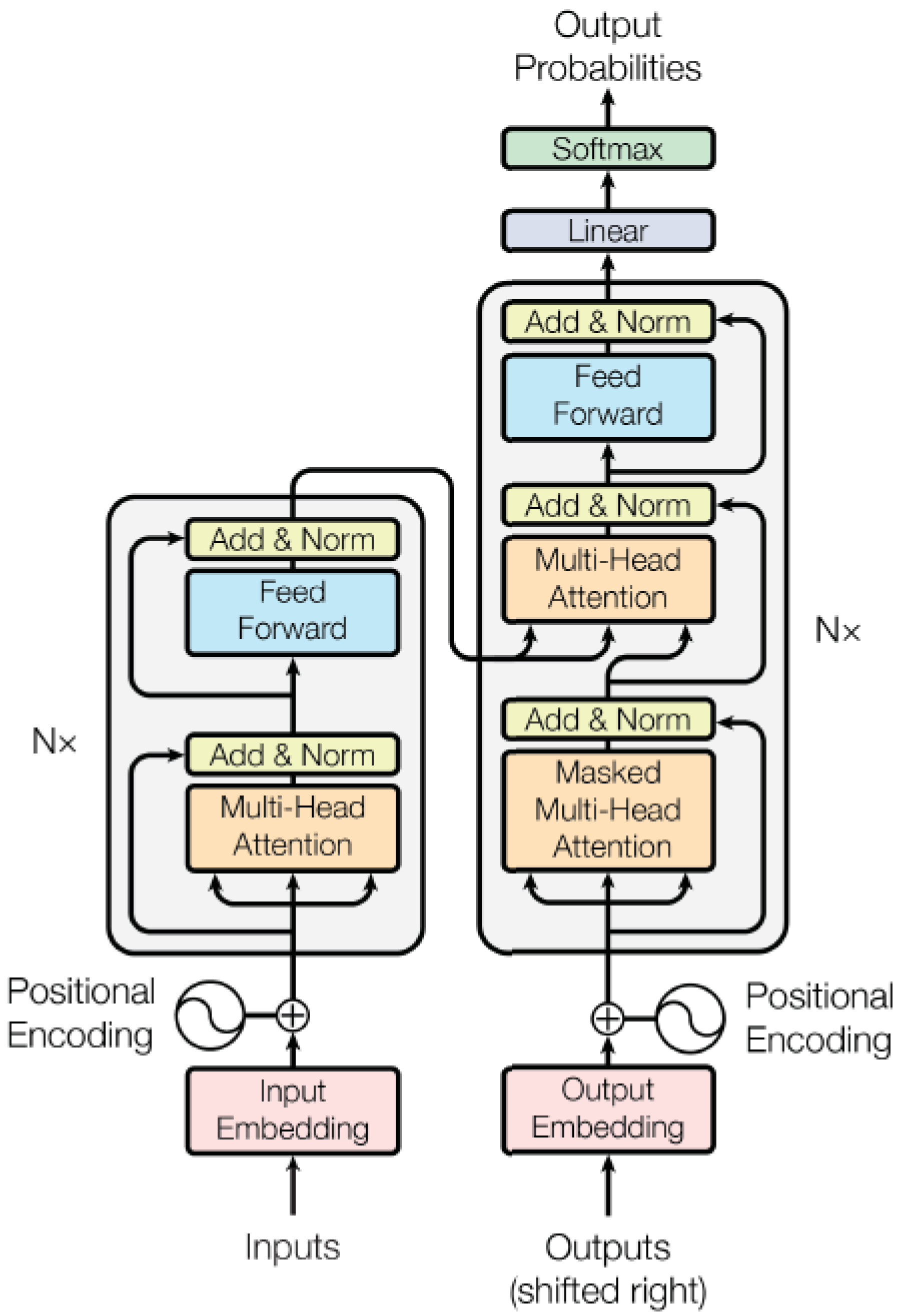
2. Masked self-attention



3. Multihead self-attention



4. Transformer


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)