$$y_{cls}, \; y_1, \; y_2 \cdots \cdots \qquad\qquad\qquad y_n$$

$$\boxed{\text{Transformer Block}}$$

$$X_{cls}, \; X_1, \; X_2, \; \cdots\cdots \quad \cdots\cdots \quad X_n$$

$m < M$,   For all [mask], [replace]

$$\quad\quad\quad\quad\quad\quad [\text{no change}]$$

$$z_i = \text{Softmax} \left(A_{MLM} \cdot y_i + b_{MLM} \right)$$

$$y_i \in \mathbb{R}^d \qquad y_i = BERT(X)$$

$$A_{MLM} \in \mathbb{R}^{|V| \times d}$$

$$\sum_{i \in Mask} - \log(z_{w_i})$$

$$y_{cls} \quad \cdots \quad y_n \ y_{sep} \ \text{-----} \quad y_{hem} \ y_{sep}$$

$$BERT$$

$$X_{cls} \cdots \quad X_n, \ X_{sep} \ \cdots \quad X_{hem} \ X_{sep}$$

$$NSP \qquad \delta(A_{NSP} \cdot y_{cls} + b_{NSP})$$

$$CE - Loss$$