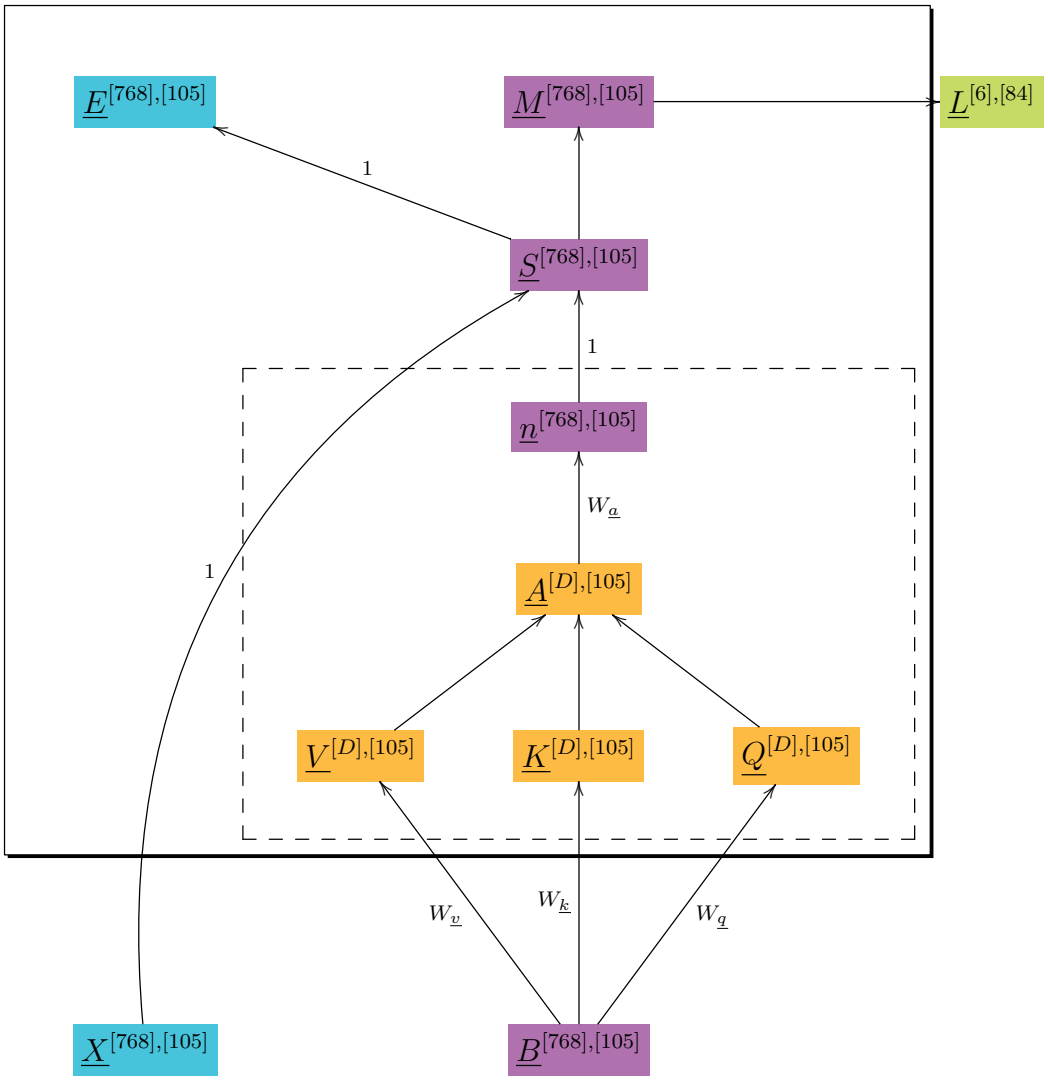

Figure 1: Sentence Ax tranet. $D = dn_{\underline{h}}$ where $d = 768$ is the hidden dimension per head, and $n_{\underline{h}} = 12$ is the number of heads. 2 copies of dashed box connected in series. 5 copies of plain box connected in series.

$$A^{[D],[105]} = \text{Attention}(Q^{[D],[105]}, K^{[D],[105]}, V^{[D],[105]}) \tag{1a}$$

$$B^{[768],[105]} = \text{BERT}() \tag{1b}$$

$$E^{[768],[105]} = X^{[768],[105]} \tag{1c}$$

$$K^{[D],[105]} = W_{\underline{k}}^{[D],[768]} B^{[768],[105]} \tag{1d}$$

$$L^{[6],[84]} = \text{ilabel}(M^{[768],[105]}) \tag{1e}$$

$$M^{[768],[105]} = \text{merge}(S^{[768],[105]}) \tag{1f}$$

$$n^{[768],[105]} = W_{\underline{a}}^{[768],[D]} A^{[D],[105]} \tag{1g}$$

$$Q^{[D],[105]} = W_{\underline{q}}^{[D],[768]} B^{[768],[105]} \tag{1h}$$

$$S^{[768],[105]} = X^{[768],[105]} + n^{[768],[105]} \tag{1i}$$

$$V^{[D],[105]} = W_{\underline{v}}^{[D],[768]} B^{[768],[105]} \tag{1j}$$

$$X^{[768],[105]} = 0 \tag{1k}$$