



Figure 1: SentenceAx Bayesian network. 2 copies of dashed box are connected in series. 5 copies of plain box are connected in series. We display the tensor shape superscripts in the Linear Algebra R2L order. (PyTorch uses a L2R order instead). All tensor shape superscripts have been simplified by omitting a  $[s_{ba}]$ , where  $s_{ba} = 24$  is the batch size.  $D = dn_{\underline{h}}$  where  $d = 768$  is the hidden dimension per head, and  $n_{\underline{h}} = 12$  is the number of heads.

$$\begin{aligned}
\underline{a}^{[86]} &: \text{ll\_greedy\_ilabel} \\
\underline{B}^{[121],[768]} &: \text{lll\_hidstate} \\
\underline{d}^{[121],[768]} &: \text{lll\_hidstate} \\
\underline{E}^{[86],[768]} &: \text{lll\_pred\_code} \\
\underline{G}^{[86],[768]} &: \text{lll\_word\_hidstate} \\
\underline{L}^{[86],[6]} &: \text{lll\_word\_score} \\
\underline{M}^{[86],[300]} &: \text{lll\_word\_hidstate} \\
\underline{n}^{[121],[768]} &: \text{lll\_hidstate} \\
\underline{S}^{[86],[768]} &: \text{lll\_word\_hidstate} \\
A^{[121],[D]} &= \text{Attention}(Q^{[121],[D]}, K^{[121],[D]}, V^{[121],[D]})
\end{aligned} \tag{1a}$$

$$\begin{aligned}
a^{[86]} &= \text{argmax}(G^{[86],[768]}; \text{dim} = -1) \\
&: \text{ll\_greedy\_ilabel}
\end{aligned} \tag{1b}$$

$$\begin{aligned}
B^{[121],[768]} &= \text{BERT}() \\
&: \text{lll\_hidstate}
\end{aligned} \tag{1c}$$

$$\begin{aligned}
d^{[121],[768]} &= \text{dropout}(n^{[121],[768]}) \\
&: \text{lll\_hidstate}
\end{aligned} \tag{1d}$$

$$\begin{aligned}
E^{[86],[768]} &= \text{embedding}(a^{[86]}) \\
&: \text{lll\_pred\_code}
\end{aligned} \tag{1e}$$

$$\begin{aligned}
G^{[86],[768]} &= \text{gather}(d^{[121],[768]}; \text{dim} = -2) \\
&: \text{lll\_word\_hidstate}
\end{aligned} \tag{1f}$$

$$K^{[121],[D]} = B^{[121],[768]} W_{\underline{k}}^{[768],[D]} \tag{1g}$$

$$\begin{aligned}
L^{[86],[6]} &= M^{[86],[300]} W_{il}^{[300],[6]} \\
&: \text{lll\_word\_score}
\end{aligned} \tag{1h}$$

$$\begin{aligned}
M^{[86],[300]} &= G^{[86],[768]} W_{il}^{[768],[300]} \\
&: \text{lll\_word\_hidstate}
\end{aligned} \tag{1i}$$

$$n^{[121],[768]} = A^{[121],[D]} W_{\underline{a}}^{[D],[768]} \quad (1j)$$

`: lll_hidstate`

$$Q^{[121],[D]} = B^{[121],[768]} W_{\underline{q}}^{[768],[D]} \quad (1k)$$

$$S^{[86],[768]} = E^{[86],[768]} + G^{[86],[768]} \quad (1l)$$

`: lll_word_hidstate`

$$V^{[121],[D]} = B^{[121],[768]} W_{\underline{v}}^{[768],[D]} \quad (1m)$$