



Figure 1: Sax bnet. 2 copies of dashed box are connected in series. 5 copies (5 depths) of plain box are connected in series. However, in the first of those 5 plain box copies, the dotted box is omitted and node \underline{G} feeds directly into node \underline{M} (indicated by red arrow). We display the tensor shape superscripts in the PyTorch L2R order. All tensor shape superscripts have been simplified by omitting a $[s_{ba}]$ from their left side, where $s_{ba} = 24$ is the batch size.

$$\begin{aligned}
 \underline{a}^{[86]} &: \text{ll_greedy_ilabel} \\
 \underline{B}^{[121],[768]} &: \text{lll_hidstate} \\
 \underline{d}^{[121],[768]} &: \text{lll_hidstate} \\
 \underline{E}^{[86],[768]} &: \text{lll_pred_code} \\
 \underline{G}^{[86],[768]} &: \text{lll_word_hidstate} \\
 \underline{I}^{[121],[768]} &: \text{lll_hidstate} \\
 \underline{L}^{[86],[6]} &: \text{lll_word_score} \\
 \underline{M}^{[86],[300]} &: \text{lll_word_hidstate} \\
 \underline{S}^{[86],[768]} &: \text{lll_word_hidstate} \\
 \underline{X}^{[86],[6]} &: \text{lll_word_score} \\
 \underline{a}^{[86]} &= \text{argmax}(\underline{X}^{[86],[6]}; \text{dim} = -1) \\
 &: \text{ll_greedy_ilabel}
 \end{aligned} \tag{1a}$$

$$B^{[121],[768]} = \text{BERT}() \quad (1b)$$

: l1l_hidstate

$$d^{[121],[768]} = \text{dropout}(I^{[121],[768]}) \quad (1c)$$

: l1l_hidstate

$$E^{[86],[768]} = \text{embedding}(a^{[86]}) \quad (1d)$$

: l1l_pred_code

$$G^{[86],[768]} = \text{gather}(d^{[121],[768]}; \dim = -2) \quad (1e)$$

: l1l_word_hidstate

$$I^{[121],[768]} = [B^{[121],[768]} \mathbb{1}(\text{depth} = 0) + M^{[86],[300]} \mathbb{1}(\text{depth} > 0)] \quad (1f)$$

: l1l_hidstate

$$L^{[86],[6]} = M^{[86],[300]} W_{il}^{[300],[6]} \quad (1g)$$

: l1l_word_score

$$M^{[86],[300]} = [G^{[86],[768]} \mathbb{1}(\text{depth} = 0) + S^{[86],[768]} \mathbb{1}(\text{depth} > 0)] W_{me}^{[768],[300]} \quad (1h)$$

: l1l_word_hidstate

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

: l1l_word_hidstate

$$X^{[86],[6]} = L^{[86],[6]} \mathbb{1}(\text{depth} > 0) \quad (1j)$$

: l1l_word_score