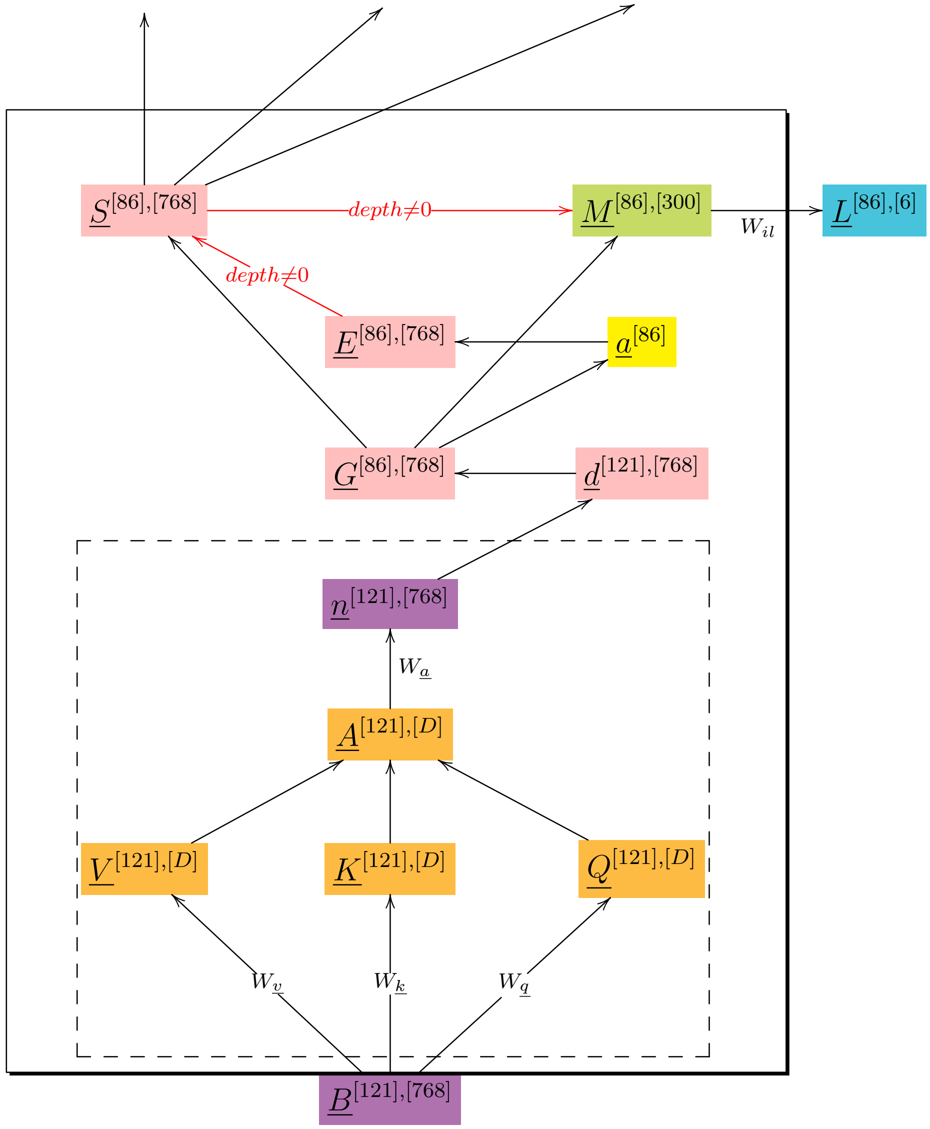

Figure 1: SentenceAx Bayesian network. 2 copies of dashed box are connected in series. 5 copies of plain box are connected in series. We display the tensor shape superscripts in the Linear Algebra R2L order. (PyTorch uses a L2R order instead). All tensor shape superscripts have been simplified by omitting 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.

$\underline{a}^{[86]}$: `ll_greedy_ilabel`
$\underline{B}^{[121],[768]}$: `lll_hidstate`
$\underline{d}^{[121],[768]}$: `lll_hidstate`
$\underline{E}^{[86],[768]}$: `lll_pred_code`
$\underline{G}^{[86],[768]}$: `lll_word_hidstate`
$\underline{L}^{[86],[6]}$: `lll_word_score`
$\underline{M}^{[86],[300]}$: `lll_merge_hidstate`
$\underline{n}^{[121],[768]}$: `lll_hidstate`
$\underline{S}^{[86],[768]}$: `lll_word_hidstate`

$$A^{[121],[D]} = \text{Attention}(Q^{[121],[D]}, K^{[121],[D]}, V^{[121],[D]}) \tag{1a}$$

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

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

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

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

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

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

$$L^{[86],[6]} = M^{[86],[300]} W_{il}^{[300],[6]} \\ : \texttt{lll\_word\_score} \tag{1h}$$

$$M^{[86],[300]} = G^{[86],[768]} W_{il}^{[768],[300]} \\ : \texttt{lll\_merge\_hidstate} \tag{1i}$$

$$n^{[121],[768]} = A^{[121],[D]} W_{\underline{a}}^{[D],[768]} \tag{1j}$$
$$: \texttt{lll\_hidstate}$$

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

$$S^{[86],[768]} = (E^{[86],[768]} + S^{[86],[768]}) \mathbb{1}(depth \neq 0) \tag{1l}$$
$$: \texttt{lll\_word\_hidstate}$$

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