Appendix of "Lattice CNNs for Matching Based Chinese Question Answering"

Yuxuan Lai[1], Yansong Feng[1,*], Xiaohan Yu[1]
Zheng Wang[2], Kun Xu[3], Dongyan Zhao[1]
[1]Institute of Computer Science and Technology, Peking University, China
[2]School of Computing and Communications, Lancaster University, UK [3]Tencent AI Lab
[1]{erutan, fengyansong, yuxiaohan, zhaodongyan}@pku.edu.cn
[2]z.wang@lancaster.ac.uk [3]syxu828@gmail.com

Appendix

Formula Derivation

$$F_w = g\{f(\boldsymbol{W}_c(\boldsymbol{v}_{\boldsymbol{w}_1} : ... : \boldsymbol{v}_{\boldsymbol{w}_n}) + \boldsymbol{b}_c^T)| \\ \forall i, w_i \in V, (w_i, w_{i+1}) \in E, w_{\lceil \frac{n+1}{2} \rceil} = w\} \tag{1}$$

Lattice based CNN layer with average pooling, derived from Eq.1 (the same as Eq.3 in the paper). Omit the condition $w_i \in V$, set $n = 3$ and mark $\boldsymbol{W_c}$ as $(\boldsymbol{W_{c1}} : \boldsymbol{W_{c2}} : \boldsymbol{W_{c3}})$, where the size of $\boldsymbol{W_{ci}}$ is (m', m) $\forall i \in \{1, 2, 3\}$, and suppose the pooling mode is average, and ignore the activation function.

$$\begin{aligned} F_w &= \texttt{Pool}\{f(\boldsymbol{W_c}(\boldsymbol{v_{w_1}} : ... : \boldsymbol{v_{w_n}}) + \boldsymbol{b_c^T})| \\ &\qquad \forall i, w_i \in V, (w_i, w_{i+1}) \in E, w_{\lceil \frac{n+1}{2} \rceil} = w\} \\ &= \texttt{ave}\{(\boldsymbol{W_{c1}} : \boldsymbol{W_{c2}} : \boldsymbol{W_{c3}})(\boldsymbol{v_{w_1}} : \boldsymbol{v_{w_2}} : \boldsymbol{v_{w_3}}) + \boldsymbol{b_c^T}| \\ &\qquad \forall i, (w_i, w_{i+1}) \in E, w_2 = w\} \\ &= \texttt{ave}\{\boldsymbol{w_{c1}v_{w_1}} + \boldsymbol{w_{c2}v_w} + \boldsymbol{w_{c3}v_{w_3}} + \boldsymbol{b_c^T}| \\ &\qquad (w_1, w), (w, w_3) \in E\} \\ &= \boldsymbol{w_{c2}v_w} + \boldsymbol{b_c^T} + \frac{\sum_{w_1} \boldsymbol{w_{c1}v_{w_1}}}{t_1} + \frac{\sum_{w_3} \boldsymbol{w_{c3}v_{w_3}}}{t_2} \end{aligned} \tag{2}$$

, where t_1 is the number of previous words of w and t_2 is the number of next words.

The same as Eq.2, but have maximum pooling.

$$\begin{aligned} F_w &= \texttt{Pool}\{f(\boldsymbol{W_c}(\boldsymbol{v_{w_1}} : ... : \boldsymbol{v_{w_n}}) + \boldsymbol{b_c^T})| \\ &\qquad \forall i, w_i \in V, (w_i, w_{i+1}) \in E, w_{\lceil \frac{n+1}{2} \rceil} = w\} \\ &= \texttt{max}\{(\boldsymbol{W_{c1}} : \boldsymbol{W_{c2}} : \boldsymbol{W_{c3}})(\boldsymbol{v_{w_1}} : \boldsymbol{v_{w_2}} : \boldsymbol{v_{w_3}}) + \boldsymbol{b_c^T}| \\ &\qquad \forall i, (w_i, w_{i+1}) \in E, w_2 = w\} \\ &= \texttt{max}\{\boldsymbol{w_{c1}v_{w_1}} + \boldsymbol{w_{c2}v_w} + \boldsymbol{w_{c3}v_{w_3}} + \boldsymbol{b_c^T}| \\ &\qquad (w_1, w), (w, w_3) \in E\} \\ &= \boldsymbol{w_{c2}v_w} + \boldsymbol{b_c^T} + \max\{\boldsymbol{w_{c1}v_{w_1}}|\forall w_1\} + \max\{\boldsymbol{w_{c3}v_{w_3}}|\forall w_3\} \end{aligned} \tag{3}$$

Directed graph convoluational networks (DGCs), derived from *Eq.2* in (Vashishth et al. 2018). We also suppose the pooling mode is average(not sum, which is in original paper and the first line of our derivation) and ignore the activation function.

$$\begin{aligned} h_v^{k+1} &= f(\sum_{u \in \mathcal{N}} (\boldsymbol{W}_{l(u,v)}^k h_u^k + \boldsymbol{b}_{l(u,v)}^k)) \\ &= \texttt{ave}\{(\boldsymbol{W}_{l(u,v)}^k h_u^k + \boldsymbol{b}_{l(u,v)}^k)|u \in \mathcal{N}\} \\ &= \texttt{ave}\{\boldsymbol{W_{c1}V_{w_1}} + \boldsymbol{b_1^T}, \boldsymbol{W_{c3}V_{w_3}} + \boldsymbol{b_3^T}, \boldsymbol{W_{c2}V_w} + \boldsymbol{b_2^T}| \\ &\qquad (w_1, w), (w, w_3) \in E\} \\ &= \frac{1}{1 + t_1 + t_2}(\boldsymbol{w_{c2}v_w} + (\boldsymbol{b_2^T} + \boldsymbol{b_1^T} * \boldsymbol{t_1} + \boldsymbol{b_3^T} * \boldsymbol{t_2}) \\ &\qquad + \sum_{w_1} \boldsymbol{w_{c1}v_{w_1}} + \sum_{w_3} \boldsymbol{w_{c3}v_{w_3}}) \end{aligned} \tag{4}$$

, where $\boldsymbol{W_{c1}}, \boldsymbol{b_1^T}$ are parameters for forward connections, $\boldsymbol{W_{c3}}, \boldsymbol{b_3^T}$ are parameters for backward connections, and $\boldsymbol{W_{c2}}, \boldsymbol{b_2^T}$ are parameters for self-loops.

The same as Eq.4, but have maximum pooling.

$$\begin{aligned} h_v^{k+1} &= f(\sum_{u \in \mathcal{N}} (\boldsymbol{W}_{l(u,v)}^k h_u^k + \boldsymbol{b}_{l(u,v)}^k)) \\ &= \max\{(\boldsymbol{W}_{l(u,v)}^k h_u^k + \boldsymbol{b}_{l(u,v)}^k)|u \in \mathcal{N}\} \\ &= \max\{\boldsymbol{W_{c1}V_{w_1}} + \boldsymbol{b_1^T}, \boldsymbol{W_{c3}V_{w_3}} + \boldsymbol{b_3^T}, \boldsymbol{W_{c2}V_w} + \boldsymbol{b_2^T}| \\ &\qquad (w_1, w), (w, w_3) \in E\} \\ &= \max\{\max\{\boldsymbol{W_{c1}V_{w_1}} + \boldsymbol{b_1^T}|\forall w_1\}, \\ &\qquad \boldsymbol{W_{c2}V_w} + \boldsymbol{b_2^T}, \max\{\boldsymbol{W_{c3}V_{w_3}} + \boldsymbol{b_3^T}|\forall w_3\}\} \end{aligned} \tag{5}$$

References

Vashishth, S.; Dasgupta, S. S.; Ray, S. N.; and Talukdar, P. 2018. Dating documents using graph convolution networks. In *ACL 2018*, 1605–1615.