Supplementary Material for Continual Learning Long Short Term Memory

Xin Guo*
University of Delaware
guoxin@udel.edu

Yu Tian*
Rutgers University
yt219@cs.rutgers.edu

Qinghan Xue
IBM
qinghan.xue@ibm.com

Panos Lampropoulos
IBM
panosl1@ibm.com

Steven Eliuk
IBM
steven.eliuk@ibm.com

Kenneth Barner
University of Delaware
barner@udel.edu

Xiaolong Wang†
IBM
xiaolong.wang@ibm.com

1 CL-LSTM$^+$ Model

We also implement a complex version CL-LSTM$^+$, as shown in Fig. 1. Compared to CL-LSTM in which each task $k > 1$ has an *unique* broadcast module M_k^b and collect module M_k^c, CL-LSTM$^+$ allocates *multiple* broadcast modules $M_{k,j}^b$ and collect module $M_{k,j}^c$ for every $j < k$. The intuition behind this design is to learn *specific* broadcast and collect information between every pair of tasks, instead of learning *general* broadcast and collect information as in CL-LSTM.

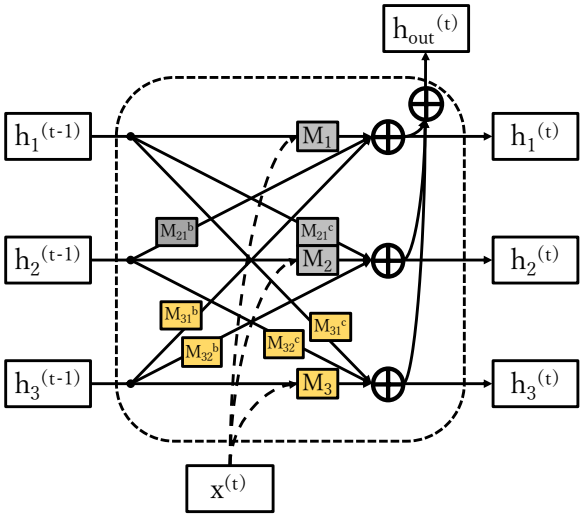

Figure 1: CL-LSTM$^+$ with three tasks. For the third task, old modules are *frozen* (grey) and $M_3, M_{31}^c, M_{31}^b, M_{32}^c, M_{32}^b$ (yellow) are trained for information sharing. $h_{out}^{(t)}$ is the aggregation of all hidden states.

Similar to the Eq. 10 in the main paper, for CL-LSTM$^+$, when current task is k, the hidden state update rule for $1 \leq j \leq k$ is given by:

$$h_j^{(t)} = M_j(x^t, h_j^{(t-1)}) + \sum_{1 \leq i < j} M_{j,i}^c(h_i^{(t-1)}) + \sum_{j < l \leq k} M_{l,j}^b(h_l^{(t-1)}), t \in \{1, 2, \cdots, T\}, \quad (1)$$

2 Different Orders of the Dataset

In order to make our experimental results more convincing, we also investigate different orders in the addressed datasets where we test CL-LSTM , LWF and finetune on Exp1 in the reverse order (WR→SNIPS→ATIS).

The results are shown in Table 1, while all methods have large performance drop, we found it is due to the forgetting on the largest WR dataset, probably longer training and parameter tuning can alleviate this problem. However, in this reverse order, our method still outperforms others.

3 The Proposed Models with Distillation Loss

We also evaluate the proposed CL-LSTM and CL-LSTM$^+$ with additional distillation loss (Hinton et al., 2015), named as CL-LSTM$_D$ and CL-LSTM$_D^+$. Experimental results of CL-LSTM$_D$ and CL-LSTM$_D^+$ on Exp1 and Exp2 are shown in Table 2∼4 and Table 5, respectively. From those results, we can see that distillation loss contributes only a little improvement over the proposed models, the proposed methods can work with or without

* indicates equal contributions. This work was done during Xin and Yu's internship at IBM.

† Corresponding author.

Method	**50, slot**	**50, indent**	**50, semantic**	**500, slot**	**500, indent**	**500, semantic**
CL-LSTM	**52.06**	**58.46**	**17.13**	**71.77**	**79.50**	**40.78**
LWF	49.75	56.89	14.48	67.22	78.82	36.94
Finetune	50.74	56.15	16.18	68.78	77.65	38.53

Table 1: Results of Exp1 on reverse order of the datasets with exemplar size of 50 and 500 samples.

Method	**50**	**100**	**200**	**300**	**500**
Joint Training	89.91	89.91	89.91	89.91	89.91
CL-LSTM	**74.74**	79.96	**83.97**	85.54	87.68
CL-LSTM$_D$	73.18	79.33	83.58	85.21	87.46
CL-LSTM$^+$	74.43	79.81	83.88	85.20	**87.73**
CL-LSTM^+_D	73.06	**80.17**	83.02	**85.65**	87.50

Table 2: Results of Exp1 on *F1-score* along with exemplar size from 50 to 500 samples, where D means model with distillation loss.

Method	**50**	**100**	**200**	**300**	**500**
Joint Training	95.05	95.05	95.05	95.05	95.05
CL-LSTM	**79.10**	82.49	86.48	87.91	91.15
CL-LSTM$_D$	77.80	81.86	86.17	87.65	91.21
CL-LSTM$^+$	78.84	81.79	**87.59**	**88.58**	91.23
CL-LSTM^+_D	78.19	**82.78**	85.79	88.24	**91.41**

Table 3: Results of Exp1 on *intent accuracy* along with exemplar size from 50 to 500 samples.

Method	**50**	**100**	**200**	**300**	**500**
Joint Training	76.92	76.92	76.92	76.92	76.92
CL-LSTM	**50.46**	**57.84**	**63.81**	65.36	70.99
CL-LSTM$_D$	47.36	55.96	62.76	66.02	70.63
CL-LSTM$^+$	50.36	56.96	63.67	64.91	**71.00**
CL-LSTM^+_D	48.41	56.10	61.60	**66.34**	69.75

Table 4: Results of Exp1 on *semantic accuracy* along with exemplar size from 50 to 500 samples.

Method	**50**	**100**	**200**	**300**	**500**
Joint Training	75.86	75.86	75.86	75.86	75.86
CL-LSTM	48.26	**53.34**	55.62	60.77	70.82
CL-LSTM$^+$	**49.49**	52.80	**55.85**	**61.84**	**71.75**
CL-LSTM^+_D	44.12	50.29	55.09	59.08	69.88

Table 5: Results of Exp2 on *F1-score* along with exemplar size from 50 to 500 samples.

it depending on the computational complexity in real scenarios.

References

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.