MeLL: Large-scale Extensible User Intent Classification for Dialogue Systems with Meta Lifelong Learning

Chengyu Wang[1*], Haojie Pan[1*], Yuan Liu[1], Kehan Chen[1], Minghui Qiu[1], Wei Zhou[1], Jun Huang[1], Haiqing Chen[1], Wei Lin[1], Deng Cai[2]

[1] Alibaba Group [2] State Key Lab of CAD & CG, Zhejiang University

Key Contributions

- In this work, we introduce the task of *large-scale Extensive User Intent Classification* (EUIC), which is vital for understanding users' intents in dialogue systems with a large, increasing number of User Intent Classification (UIC) tasks involved.
- We propose the *Meta Lifelong Learning* (MeLL) framework to address this task. It employs a slowly updated text encoder to learn representations across tasks and global/local memory networks to learn task semantics.
- We conduct extensive experiments on public and real-world industry datasets. The results show that the MeLL framework consistently outperforms strong baselines.
- We deploy MeLL on a real-world dialogue system AliMe and observe significant improvements in an online A/B test.

Introduction

Background. User intent detection is vital for understanding their demands in dialogue systems. Although the User Intent Classification (UIC) task has been widely studied, for large-scale industrial applications, the task is still challenging. When the underlying application evolves, new UIC tasks continuously emerge in a large quantity. Hence, it is crucial to develop a framework for *large-scale extensible UIC* that continuously fits new tasks and avoids catastrophic forgetting with an acceptable parameter growth rate.

Our Work. We introduce the **Me**ta **L**ifelong **L**earning (MeLL) framework to address this task. In MeLL, a BERT-based text encoder is employed to learn robust text representations across tasks, which is slowly updated for lifelong learning. We design global and local memory networks to capture the cross-task prototype representations of different classes, treated as the meta-learner quickly adapted to different tasks. Additionally, the Least Recently Used replacement policy is applied to manage the global memory such that the model size does not explode through time. Finally, each UIC task has its own task-specific output layer, with the attentive summarization of various features.

MeLL: The Proposed Framework

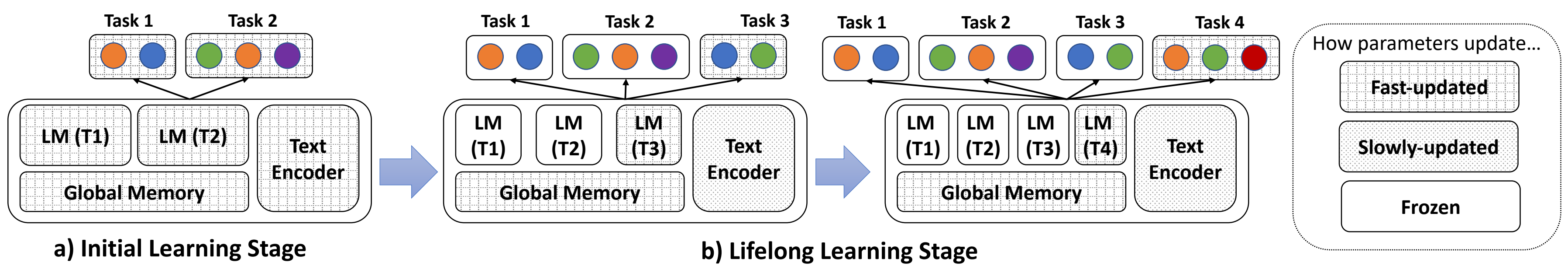

Figure 1:The high-level framework MeLL for large-scale EUIC tasks in dialogue systems.

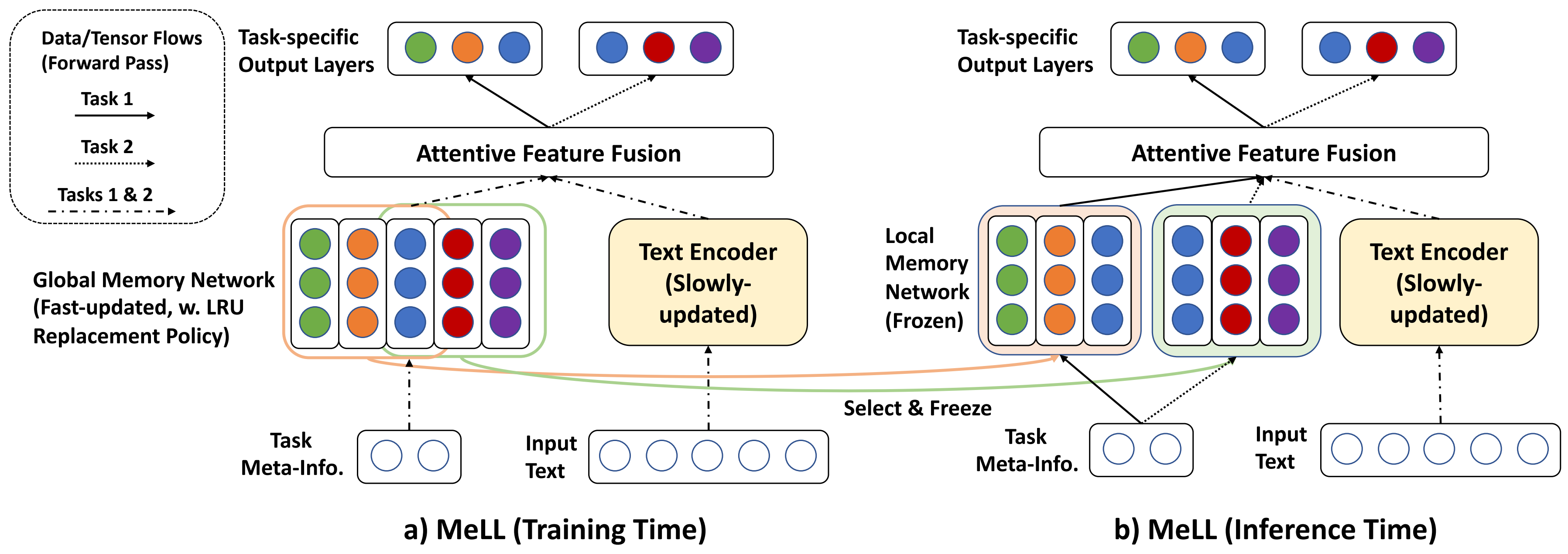
Data/Tensor Flows (Forward Pass)
Task 1
Task 2
Tasks 1 & 2
Task-specific Output Layers
Attentive Feature Fusion
Global Memory Network (Fast-updated, w. LRU Replacement Policy)
Text Encoder (Slowly-updated)
Task Meta-Info.
Input Text
Select & Freeze
Local Memory Network (Frozen)
a) MeLL (Training Time)
b) MeLL (Inference Time)

Figure 2:Model structure of MeLL for large-scale EUIC. During training, the features used for UIC are attentively generated by the slowly-updated text encoder and the fast-updated global memory network. After the training process of a specific task, the respective class representations are copied from the global memory to a task-specific local memory network. During inference, we use the text encoder and the task-specific local memory to generate features for prediction.

MeLL: The Proposed Framework

Text Encoder: We employ BERT as our model backbone to learn the universal, deep representations of input texts across tasks. As new UIC tasks continuously arrive, the BERT parameters are *slowly updated* to digest transferable knowledge across multiple tasks.

Global Memory Network: The global memory network is only applied during training, which contains a certain number of "slots" to store class representations. The memory units are *fast updated* to acquire knowledge from new tasks. As the sum of the numbers of distinct classes is increasing, we use the LRU replacement policy such that there are only a fix number of "slots" in the memory.

Local Memory Networks: The fast update of the global memory may significantly affect the final features used for prediction. To guarantee the performance of previous tasks is not affected, for each task, we have a separate local memory that copies the respective class representations from the global network with parameters frozen.

Task-specific Networks: Finally, we use class representations and text representations to generate attentive features for user intent prediction. Each task has its own prediction head.

Experiments

Key Results. To evaluate the effectiveness of MeLL, we construct two datasets, including a fused dataset for query intent classification in task-oriented dialogues (TaskDialog-EUIC) and a real-world dataset for response intent classification (Hotline-EUIC). Specifically, Hotline-EUIC is collected and annotated from the hotline data produced from an e-commerce dialogue system AliMe.

Table 1:Comparison of different models over two datasets.

Task	TaskDialog-EUIC				Hotline-EUIC			
Results	All tasks		New tasks		All tasks		New tasks	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
MTL*	0.9597	0.9590	0.9568	0.9562	0.9788	0.9480	0.9832	0.9523
Single*	0.9006	0.8974	0.9005	0.8969	0.9196	0.8685	0.9239	0.8814
Lifelong-freeze	0.9214	0.9194	0.9015	0.8988	0.9401	0.8798	0.9259	0.8501
Lifelong-seq	0.3140	0.2043	0.3447	0.2455	0.4517	0.3485	0.5272	0.4238
Lifelong-replay*	0.6225	0.5481	0.5485	0.4573	0.8215	0.8260	0.9420	0.8553
MeLL	**0.9379**	**0.9342**	**0.9271**	**0.9224**	**0.9673**	**0.9341**	**0.9675**	**0.9319**

Online A/B Test. We deploy our model in the Alime hotline agent. After deployment, we collect back annotated data of 600 tasks from the online system, compare the performance of our model with the previous single-model system, and report the F1 score. The previous online system includes thousands of single models. Each model is a TextCNN model distilled from an fine-tuned ALBERT-base model. Overall, our model improves the overall F1 in 8.61%.

Table 2:The online performance comparison between MeLL and the online system.

Method	F1	Relative Improv.
Online system (Single)	0.8359	N.A.
MeLL (w. LRU)	0.9079	8.61%

Conclusion

We formally introduce the task of *large-scale EUIC*, and propose the *Meta Lifelong Learning* (MeLL) framework to address this task. Extensive experiments on both English and Chinese EUIC datasets show the effectiveness of MeLL, which consistently outperforms strong baselines. We have also deploy MeLL on a real industrial dialogue system AliMe. The online A/B test results show the superiority of our method. In the future, we will further explore how MeLL can be employed to solve other tasks and support other applications.