Incremental Prompting: Episodic Memory Prompt for Lifelong Event Detection

Anonymous ACL submission

Abstract

Lifelong event detection aims to incrementally update a model with new event types and data while retaining the capability on previously learned old types. One critical challenge is that the model would catastrophically forget old types when continually trained on new data. In this paper, we introduce **E**psodic **M**emory **P**rompts (**EMP**) to explicitly preserve the learned task-specific knowledge. Our method adopts continuous prompt for each task and they are optimized to instruct the model prediction and learn event-specific representation. The EMPs learned in previous tasks are carried along with the model in subsequent tasks, and can serve as a memory module that keeps the old knowledge and transferring to new tasks. Experiment results demonstrate the effectiveness of our method. Furthermore, we also conduct a comprehensive analysis of the new and old event types in lifelong learning.

1 Introduction

Class-incremental event detection is a challenging setting in lifelong learning, where the model is incrementally updated on a continual stream of data for new event types while retaining the event detection capability for all the previously learned types. The main challenge of class-incremental event detection lies in the *catastrophic forgetting* problem, where the model's performance on previously learned types significantly drops after it is trained on new data. Recent studies (Wang et al., 2019; Lopez-Paz and Ranzato, 2017) have revealed that replaying stored samples of old classes can effectively alleviate the catastrophic forgetting issue. However, simply fine-tuning the entire model on the limited stored samples may result in overfitting, especially when the model has a huge set of parameters. How to effectively leverage the stored examples still remains an important question.

Prompt learning, which is to simply tune a template-based or continuous prompt appended to the input text while keeping all the other parameters freezed, has recently shown comparable or even better performance than fine-tuning the entire model in many NLP tasks (Brown et al., 2020; Jiang et al., 2020; Gao et al., 2021; Li and Liang, 2021; Lester et al., 2021; Hambardzumyan et al., 2021). It is especially flavored by the lifelong learning setting since it only tunes a small amount of parameters, thus has the potential to alleviate the catastrophic forgetting and exemplar memory overfitting issues. Moreover, the prompts can also be used to store task-specific knowledge.

In this work, we propose a simple but effective incremental prompting framework that introduces **E**psodic **M**emory **P**rompts (**EMP**) to store the learned type-specific knowledge. At each training stage, we adopt a learnable prompt for each new event type added from the current task. The prompts are initialized with event type names and fine-tuned with the annotations from each task. To encourage the prompts to always carry and reflect type-specific information, we entangle the feature representation of each event mention with the type-specific prompts by optimizing its type distribution over them. After each training stage, we keep the learned prompts in the model and incorporate new prompts for next task. In this way, the acquired task-specific knowledge can be carried into subsequent tasks. Therefore, our EMP can be considered as a soft episodic memory that preserves the old knowledge and transfers it to new tasks. Our contributions can be summarized as follows:

• We propose **E**psodic **M**emory **P**rompts (**EMP**) which can explicitly carry previously learned knowledge to subsequent tasks for class-incremental event detection. Extensive experiments validate the effectiveness of our method.

• To the best of our knowledge, we are the first to adopt prompting methods for class-incremental event detection. Our framework has the potential to be applied to other incremental learning tasks.

2 Problem Formulation

Given an input text $x_{1:L}$ and a set of target spans $\{(x_i, x_j)\}$ from it, an event detection model needs to assign each target span with an event type in the ontology or label it as *Other* if the span is not an event trigger. For class-incremental event detection, we aim to train a single model f_θ on a sequence of T tasks $\{\mathcal{D}_1, ..., \mathcal{D}_T\}$ that consist of non-overlapping event type sets $\{\mathcal{C}_1, ..., \mathcal{C}_T\}$[1]. In each t-th task, the model needs to classify each mention to any the types that have seen so far $\mathcal{O}_t = \mathcal{C}_1 \bigcup ... \bigcup \mathcal{C}_t$. The training instances in each task $\mathcal{D}_t$ consist of tuples of an input text $x^t_{1:L}$, a target span $\bar{x}^t$, and its corresponding label $\mathbf{y}^t$ where $\mathbf{y}^t \in \mathcal{C}_t$. For convenience, the notations are for the t-th training stage by default unless denoted explicitly in the following parts of the paper.

3 Approach

3.1 Span-based Event Detection

Given an input sentence $x^t_{1:L}$ from task $\mathcal{D}_t$, we first encode it with BERT (Devlin et al., 2019) to obtain the contextual representations $\mathbf{x}^t_{1:L} = \text{BERT}(x^t_{1:L})$. Note that we freeze BERT's parameters in our method and all baselines. For each span $\bar{x}^t$, we concatenate its starting and ending token representations and feed them into a multilayer perceptron (MLP) to get the span representation $\mathbf{h}^t_{span}$. Then, we apply a linear layer on $\mathbf{h}^t_{span}$ to predict the type distribution of the span $p^t = linear(\mathbf{h}^t_{span})$. We use cross-entropy loss to train the model on $\mathcal{D}_t$:

$$\mathcal{L}_C = -\sum_{(\bar{x}^t, y^t) \in \mathcal{D}_t} \log p^t. \quad (1)$$

3.2 Episodic Memory Prompting

To overcome the catastrophic forgetting and exemplar memory overfitting issues, we design a continuous prompting approach with Episodic Memory Prompts (EMPs) to preserve the knowledge learned from each task and transfer to new tasks.

Given an incoming task $\mathcal{D}_t$ and its corresponding new event type set $\mathcal{C}_t = \{c^t_1, ..., c^t_{n_t}\}$, we first initialize a sequence of new *prompts* $\mathbf{C}^t = [\mathbf{c}^t_1, ..., \mathbf{c}^t_{n_t}]$ where $\mathbf{c}^t_i \in \mathbb{R}^{1\times e}$ is a type-specific prompt for type c^t_i. e is the embedding dimension size. In our experiments, we use the event type name to initialize each event type prompt $\mathbf{c}^t_i$ (see Appendix A for details). Note that we always preserve the prompts learned from previous tasks, thus the accumulated prompts until the t-th task are represented as $\mathbf{I}^t = [\mathbf{C}^1, ..., \mathbf{C}^t]$. Given a particular sentence $x^t_{1:L}$ from $\mathcal{D}_t$, we concatenate it with the accumulated prompts $\mathbf{I}^t$, encode the whole sequence with BERT, and obtain the sequence of contextual representations $[\tilde{\mathbf{x}}^t_{1:L}; \tilde{\mathbf{I}}^t]$, where $\tilde{\mathbf{x}}^t_{1:L}i$ and $\tilde{\mathbf{I}}^t_i$ denote the sequence of contextual embeddings of $x^t_{1:L}$ and $\mathbf{I}^t$ respectively. $[;]$ is concatenation operation. Then, similar as Section 3.1, we obtain a representation $\tilde{\mathbf{h}}^t_{span}$ for each span based on $\tilde{\mathbf{x}}^t_i$, and predict the logits over all target event types $\tilde{p}^t = linear(\tilde{\mathbf{h}}^t_{span})$.

We expect the EMPs to be specific to the corresponding event types and preserve the knowledge of each event type from previous tasks. So we design an entangled prompt optimization strategy to entangle the feature representation of each span with the event type-specific prompts by computing an event type probability distribution over them. Specifically, given a span representation $\tilde{\mathbf{h}}^t_{span}$ and EMP representations $\tilde{\mathbf{I}}^t$, we compute the probability distribution over all prompts as $\tilde{p}^t_c = \text{MLP}(\tilde{\mathbf{I}}^t) \cdot \tilde{\mathbf{h}}^t_{span}$, where $\cdot$ is the dot product. Finally, we combine the original logits $\tilde{p}^t$ and $\tilde{p}^t_c$ to predict the event type label for each span:

$$\tilde{\mathcal{L}}_C = -\sum_{(\bar{x}^t, y^t) \in \mathcal{D}_t} \log (\tilde{p}^t + \tilde{p}^t_c). \quad (2)$$

At the end of each training stage, we keep the learned prompts from the current task $\mathbf{C}^t$ in the model, and then initialize a new prompt $\mathbf{C}^{t+1}$ for the next task incrementally: $\mathbf{I}^{t+1} = [\mathbf{I}^t; \mathbf{C}^{t+1}]$.

3.3 Lifelong Learning with Experience Replay and Knowledge Distillation

To alleviate the catastrophic forgetting issue, a common strategy is to store a limited amount of data from old tasks in a memory buffer and pass them to later tasks. We follow this strategy and adopt two popularly used methods: (1) Experience Replay which is to repeatedly optimize the model on the stored data in subsequent tasks; and (2) Knowledge Distillation (KD) that is to ensure the output probabilities and features from the current and previous models to be matched, respectively.

Specifically, after training on $\mathcal{D}_t$, we apply the herding algorithm (Welling, 2009) to select 20 training samples for each type into the memory buffer, denoted as $\mathcal{M}$. Similar as Equation 2, the objective

[1]Though the type sets from all tasks contain *Other*, they have distinct meanings given different seen types.

for experience repaly is:

$$\mathcal{L}_{ER} = -\sum_{(\bar{x}^r, y^r)\in\mathcal{M}} \log\,(\tilde{p}^t + \tilde{p}^t_c). \quad (3)$$

For knowledge distillation, following (Cao et al., 2020), we apply both *prediction-level* and *feature-level* distillation, and use a temperature parameter to rescale the probabilities of prediction-level KD. The objectives for prediction-level KD and feature-level KD are computed as:

$$\mathcal{L}_{PD} = -\sum_{(\bar{x}^r, y^r)\in\mathcal{M}} (\tilde{p}^{t-1} + \tilde{p}^{t-1}_c) \log\,((\tilde{p}^t + \tilde{p}^t_c)).$$

$$\mathcal{L}_{FD} = \sum_{(x^r,(x^r_i,x^r_j),y^r)\in\mathcal{M}} 1 - g(\bar{\mathbf{h}}^{t-1}_{span}, \bar{\mathbf{h}}^{t}_{span}),$$

where g is the cosine similarity function. $\bar{\mathbf{h}}^{t-1}_{span}$ and $\bar{\mathbf{h}}^{t}_{span}$ are l_2-normalized features from the model at $t-1$ and t stages, respectively.

Optimization We apply a weighting factor λ to control how much loss from experience replay and knowledge distillation to use in each batch. The final loss is computed as:

$$\mathcal{L} = \tilde{\mathcal{L}}_C + \lambda(\mathcal{L}_{ER} + \mathcal{L}_{PD} + \mathcal{L}_{FD}).$$

4 Experiments and Discussion

Experiment Settings We conduct experiments on two benchmark datasets: ACE05-EN (Doddington et al., 2004) and MAVEN (Wang et al., 2020), and construct the class-incremental datasets following the *oracle negative* setting in (Yu et al., 2021). We divided the ontology into 5 subsets with distinct event types, and then use them to constitute a sequence of 5 tasks denoted as $\mathcal{D}_{1:5}$. We use the same partition and task order permutations in (Yu et al., 2021). During the learning process from $\mathcal{D}_1$ to $\mathcal{D}_5$, we constantly test the model on the entire test set (which contains the whole ontology) and take the mentions of unseen event types as negative instances. More implementation details, including parameters, initialization of prompts as well as baselines are shown in Appendix A.

Results We present the main results in Table 1. We have following observations: (1) by comparing the performance of various approaches on Task 1 which are not affected by any catastrophic forgetting, our prompting based approach improves 4.1% F-score on MAVEN and 1.3% F-score on ACE05, demonstrating that by incorporating task-specific prompts, event detection itself can be significantly improved. EMPs even provide more improvement on MAVEN which contains a lot more event types than ACE05, suggesting the potential of incorporating EMPs for fine-grained event detection; (2) **KCN** can be viewed as an ablated version of our approach without EMPs. Our approach consistently outperforms **KCN** on almost all tasks on both datasets, demonstrating the effectiveness of EMPs on improving class-incremental event detection; (3) Comparing with **BERT-ED**, **KCN** adopts experience replay and knowledge distillation. Their performance gap verifies that these two strategies can dramatically alleviate the catastrophic forgetting problem. (4) There is still a large gap between the current lifelong learning approaches and the upperbound, indicating that catastrophic forgetting still remains a very challenging problem. Note that for fair comparison, for all approaches, we set the exemplar buffer size as 20, and allow one exemplar instance to be use in each training batch instead of the whole memory set, thus most results in our paper cannot be directly compared with the results reported in (Yu et al., 2021). We also analyze the effect of exemplar buffer size in Appendix B.

Analysis of New and Old Types in Lifelong Learning Figure 1 shows the F-score on old and new event types in each training stage for both our approach and **KT** (Yu et al., 2021) on the MAVEN dataset. Our approach consistently outperforms **KT** by a large margin on both old types and new types, demonstrating that our EMPs effectively preserve learned knowledge from old event types and significant improve event detection when the annotations are sufficient. Interestingly, comparing the F-score on new types in Task 1 and old types in Task 2, both methods improve the performance on the types of Task 1, indicating that both methods have the potential of leveraging indirect supervision to improve event detection.

Ablation Study For ablation study, we consider three ablated models based on our EMPs: (1) change the prompt initialization from using event type name representations[2] to using random distribution; (2) remove the knowledge distillation loss $\mathcal{L}_{PD}$ and $\mathcal{L}_{FD}$; (3) use completely fixed prompts to replace the trainable soft prompts. From Table 2, we observes that: (1) using event type names

[2]Details of using event type name to initialize prompts are shown in Appendix A

	MAVEN					ACE05-EN				
Task	1	2	3	4	5	1	2	3	4	5
BERT-ED	63.51	39.99	33.36	23.83	22.69	58.30	43.96	38.02	21.53	25.71
iCaRL* (Rebuffi et al., 2017)	18.08	27.03	30.78	31.26	29.77	4.05	5.41	7.25	6.94	8.94
EEIL (Castro et al., 2018)	63.51	50.62	45.16	41.39	38.34	58.30	54.93	52.72	45.18	41.95
BIC (Wu et al., 2019)	63.51	46.69	39.15	31.69	30.47	58.30	45.73	43.28	35.70	30.80
KCN (Cao et al., 2020)	63.51	51.17	46.80	38.72	38.58	58.30	54.71	52.88	44.93	41.10
KT (Yu et al., 2021)	63.51	52.36	47.24	39.51	39.34	58.30	**55.41**	53.95	45.00	42.62
EMP (Ours)	**67.62**	**58.33**	**54.53**	**47.70**	**44.30**	**59.60**	53.19	**55.20**	**45.64**	**43.28**
Upperbound (Ours)	/	/	/	/	66.68	/	/	/	/	68.22

Table 1: Comparison between our approach and baselines in terms of micro F-1 (%) on 5 class-incremental tasks. We report the averaged results on 5 permutations of tasks to alleviate the affect of task order.

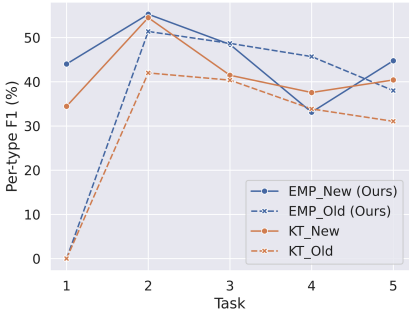

Figure 1: Performance on old types and new types in each lifelong task on MAVEN (best viewed in color).

to initialize the prompts is helpful in most tasks. We leave how to incorporate more effective prior knowledge into prompts for future work; (2) switching the continuous prompts to discrete prompts degrades the performance significantly, suggesting that the continuous prompts is generally more promising than the discrete prompts.

Task	1	2	3	4	5
EMP (Ours)	**70.57**	**57.87**	**54.33**	48.39	**45.82**
- w/o EInit	70.26	54.78	50.56	**48.42**	42.28
- w/o KD	70.57	54.82	53.24	45.37	41.22
- Discrete	67.57	54.86	49.99	45.51	39.08

Table 2: Ablation study on event-specific prompt initialization (EInit), knowledge distillation (KD), and switching to discrete prompts (Discrete) on MAVEN.

5 Related Work

Lifelong Event Detection Recent deep neural networks have shown state-of-the-art performance on conventional supervised event detection (Chen et al., 2015; Nguyen et al., 2016; Feng et al., 2016; Lu et al., 2019). However, when moving to lifelong learning setting, the performance significantly drops (Kirkpatrick et al., 2017; Li and Hoiem, 2016; Aljundi et al., 2019; Cui et al., 2021). Episodic memory replay (EMR) (Lopez-Paz and Ranzato, 2017; Guo et al., 2020; de Masson d'Autume et al., 2019; Wang et al., 2019; Han et al., 2020) and knowledge distillation (Chuang et al., 2020; Cao et al., 2020; Yu et al., 2021) have been the two most effective techniques to overcome the catastrophic forgetting problem. However, they highly rely on the stored data from old tasks, which is not the most realistic setting for lifelong learning.

Prompt Learning Conditioning on large-scale pre-trained language models, prompt learning (Brown et al., 2020; Lester et al., 2021; Chen et al., 2021; Liu et al., 2021; Wang et al., 2021a) have shown comparable performance as language model fine-tuning. Several recent studies explore prompt learning in lifelong learning setting. Qin and Joty (2021) use prompt tuning to train the model as a task solver and data generator in their proposed Lifelong Few-shot Language Learning problem. Wang et al. (2021b) propose L2P for continual learning in the vision area. To the best of our knowledge, we are the first work to adopt prompt learning for class-incremental event detection.

6 Conclusion

In this paper, we propose a novel prompting framework, namely Episodic Memory Prompts (EMP), for class-incremental event detection. During each training stage, EMP learns type-specific knowledge via a continuous prompt for each event type. The EMPs trained in previous tasks are kept in the model, such that the acquired task-specific knowledge can be transferred into the following new tasks. Experimental results validate the effectiveness of our method comparing with competitive baselines. In addition, our extensive analysis shows that by employing EMPs, both event detection itself and the incremental learning capability of our approach are significantly improved.

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11816–11825.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pengfei Cao, Yubo Chen, Jun Zhao, and Taifeng Wang. 2020. Incremental event detection via knowledge consolidation networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 707–717, Online. Association for Computational Linguistics.

Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-end incremental learning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216, pages 241–257. Springer.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *CoRR*, abs/2104.07650.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2020. Lifelong language knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2914–2924, Online. Association for Computational Linguistics.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 232–243, Online. Association for Computational Linguistics.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, pages 13122–13131.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. 2020. Improved schemes for episodic memory-based lifelong learning. In *Advances in Neural Information Processing Systems*.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 6429–6440, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 614–629. Springer.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2019. Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4366–4376, Florence, Italy. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Chengwei Qin and Shafiq Joty. 2021. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of T5. *CoRR*, abs/2110.07298.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021a. TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2792–2802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 796–806, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. 2021b. Learning to prompt for continual learning. *CoRR*, abs/2112.08654.

Max Welling. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1121–1128, New York, NY, USA. Association for Computing Machinery.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large scale incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2019, Long Beach, CA, USA, June 16-20, 2019, pages 374–382. Computer Vision Foundation / IEEE.

Pengfei Yu, Heng Ji, and Prem Natarajan. 2021. Lifelong event detection with knowledge transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5278–5290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experimental Details

Baselines We consider the following baselines for comparison: (1) **BERT-ED**: simply trains the BERT based event detection model on new tasks without prompts, experience replay or knowledge distillation. It's the same as the span-based event detection baseline in Section 3.1. (2) **KCN** (Cao et al., 2020): use a prototype-based example sampling strategy and hierarchical distillation. As the original approach studied a different setting, we adapt their prediction-level and feature-level distillation as the baseline. (3) **KT** (Yu et al., 2021): transfer knowledge between old types and new types in two directions. (4) **iCaRL*** (Rebuffi et al., 2017): use nearest-mean-of-exemplars rules to perform classification combined with knowledge distillation. iCaRL adopts different strategies for classification, experience replay, and distillation. We directly report the result in (Yu et al., 2021) for reference. (5) **EEIL** (Castro et al., 2018): use an additional finetuning stage on the balanced dataset. (6) **BIC** (Wu et al., 2019): use a bias correction layer after the classification layer. (7) **Upperbound**: trains the same model on all types in the datasets jointly. For **iCaRL**, **EEIL**, and **BIC**, we use the same implementation in (Yu et al., 2021). For fair comparison, our approach and all baselines (except for the Upperbound baseline) are built upon **KCN** and use the same experience replay and knowledge distillation strategies described in Section 3.2.

Implementation Details During training, we use AdamW (Loshchilov and Hutter, 2019) optimizer with the learning rate set to $1e-4$ and weight decay set to $1e-2$. Different from previous work (Yu et al., 2021), we set the batch size to 1 as we encode each sentence once and consider all target spans in the sentence at the same time. We adopt gradient accumulation with the step set to 8. As the number of batches is large, we apply a periodic replay strategy with the interval set to 10 to reduce computational cost. For each lifelong task $\mathcal{D}_t$, we set the maximum number of training epochs to 20. We adopt the early stopping strategy with patience 5, i.e., the training stops if the performance on the development set does not increase for 5 epochs. We set the weighting factor $\lambda = k/(s+k)$, where s is the number of predicted spans and k is set to 50. The temperature parameter used in prediction-level distillation is set to 2.

The parameters of each prompt in EMPs are initialized with the corresponding event type name. Specifically, there are three cases in the initialization: (1) If the type name is *single-token* and it is contained in BERT's vocabulary, we directly use the pre-trained embedding of this token to initialize the prompt; (2) If the type name is *multiple-token* and the tokens are contained in BERT's vocabulary, we take the average of the pre-trained embeddings of these tokens to initialize the prompt; (3) If the type name contains *Out-of-Vocabulary (OOV)* tokens, we replace the OOV tokens with the synonyms that are contained in BERT's vocabulary.

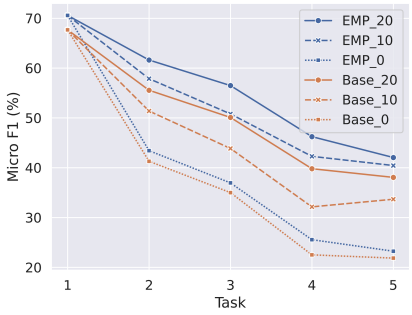

Figure 2: Performance with different buffer size in each task on MAVEN (best viewed in color).

B Effect of Exemplar Buffer Size

We conduct an analysis on the effect of exemplar buffer size. We explore the buffer size for each type in {0, 10, 20}. Note that although we reduced the buffer size, we did not modify the replay frequency, as we want to investigate the effect of data diversity in memory buffer. We use **KT** as the baseline when buffer size is 20 and 10. Note that when buffer size is 0, we do not adopt either experience replay or knowledge distillation and thus use **BERT-ED** as the baseline. We plot the results on Figure 2. We observed that: (1) Decreasing the buffer size for each type from 20 to 10 degrades the performance of both models. This indicates that reducing data diversity may result in the overfitting on example data, and thus deteriorates the performance; (2) The performance of our method is not affected as much as the baseline, demonstrating our prompting framework is more tolerant to smaller

buffer size and remains very competitive performance when less data are available; (3) When the buffer size decreases to 0, the performance of both methods drops significantly. This shows that both approaches highly rely on the stored data to overcome the catastrophic forgetting problem. This calls for developing more advance techniques to reduce the dependence on stored examples, as storing past data could result in data leakage in real-world applications.