IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 12, 2021, accepted April 5, 2021, date of publication April 8, 2021, date of current version April 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071787

Lifelong Language Learning With the Most Forgotten Knowledge

HEEJEONG CHOI AND PILSUNG KANG
School of Industrial Management Engineering, Korea University, Seoul 02841, Republic of Korea
Corresponding author: Pilsung Kang (pilsung_kang@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant NRF-2019R1F1A1060338 and Grant NRF-2019R1A4A1024732.

ABSTRACT Lifelong language learning enables a language model to accumulate knowledge through training on a stream of text data. Recent research on lifelong language learning is based on samples of previous tasks from an episodic memory or generative model. LAMOL, a representative generative model-based lifelong language learning model, preserves the previous information with the generated pseudo-old samples, which are suboptimal. In this paper, we propose an improved version of LAMOL, MFK-LAMOL, which constructs a generative replay using a more effective method. When a new task is received, MFK-LAMOL replays sufficient previous data and retrieves important examples for training alongside the new task. Specifically, it selects the examples with the most forgotten knowledge learned from previous tasks based on the extent to which they include knowledge that has been forgotten after learning new information. We showed that the proposed method outperforms LAMOL on a stream of three different natural language processing tasks.

INDEX TERMS Lifelong language learning, natural language processing, catastrophic forgetting, a stream of text data, generative replay.

I. INTRODUCTION

Artificial neural networks have successfully exceeded human ability on narrow tasks such as image recognition, video game play, and voice recognition [1]–[3]. However, these models have had limited success when they are sequentially trained with a stream of tasks [4]. While humans can continually learn new skills and accumulate knowledge throughout their lifetimes [5], neural networks suffer from a catastrophic forgetting; that is, training a model with new information interferes with previously learned knowledge [6], [7]. To alleviate this problem, various approaches have been proposed for continual lifelong learning. Lifelong learning models are designed to accumulate interconnected knowledge between learned tasks and maintain their performance of those tasks.

Most research on lifelong learning applies to the vision domain, but not natural language processing (NLP). However, we focus on lifelong language learning (LLL), where a model addresses a stream of NLP tasks. LLL models have been carried out concerning how language models preserve previously learned information to accumulate knowledge and reuse it during training. To the best of our knowledge, research on LLL has been conducted for only a few NLP tasks. For sentiment analysis, [8] adopted a Bayesian optimization framework based on stochastic gradient descent. Reference [9] adopted a domain-independent neural conversational model and introduced a novel neural continual learning algorithm that allows a conversational agent to accumulate skills across different tasks in a data-efficient way. For lifelong word representation, [10] tried to expand the new in-domain corpus by exploiting the corpora from past domains via meta-learning. Reference [11] initialized sentence encoders with the help of corpus-independent features and then sequentially updated sentence encoders using Boolean operations of conceptor matrices to learn corpus-dependent features in lifelong sentence representation. For text classification, [12] was proposed in which replay is adopted to prevent the rapid change of the weights from the previous status. It investigated the performance of the episodic memory system on NLP problems. This method distills the knowledge from previous tasks into episodic

The associate editor coordinating the review of this manuscript and approving it for publication was Xi Peng.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

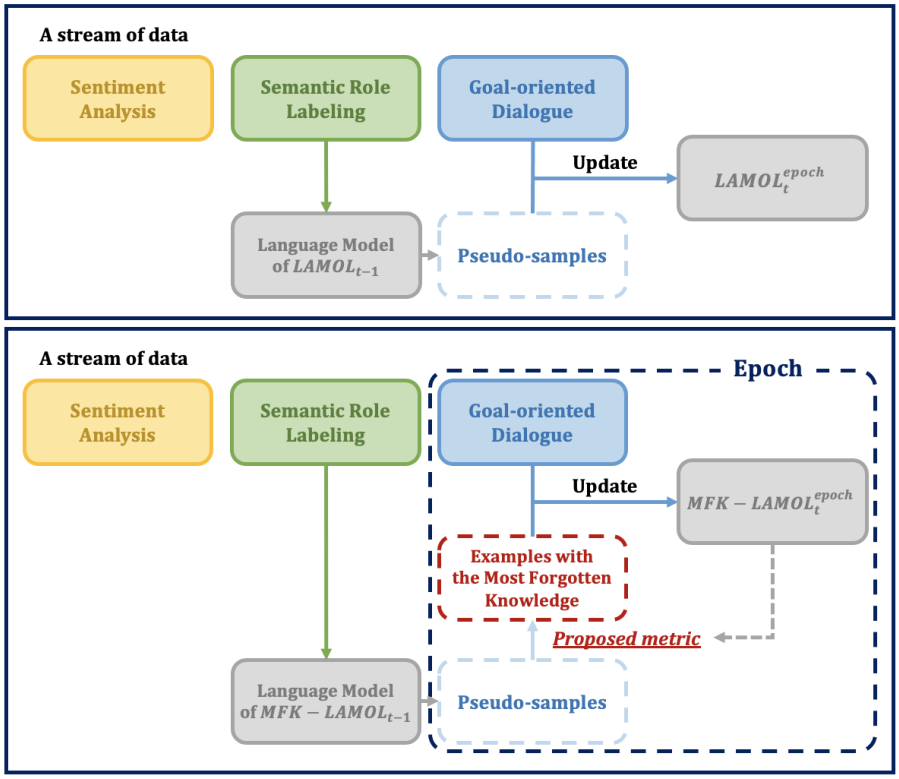


FIGURE 1. Comparison between the overall frameworks of LAMOL and MFK-LAMOL. LAMOL (top) depends on the pseudo-samples generated from the language model. It is in contrast with the construction of generative replay of MFK-LAMOL (bottom). It consists of the selected examples based on the proposed metric, which represents the amount of forgotten knowledge.

memory and replays it later. Finally, LAMOL [13] was proposed to achieve LLL on fundamentally different tasks.

In this paper, we focus on LAMOL, which can prevent catastrophic forgetting based on generative replay. It enables a single model to train multiple NLP tasks through training on a stream of data for different tasks and domains. Specifically, it achieves LLL on fundamentally different tasks in the form of a question-answering because many NLP tasks can be considered question-answering [14]. LAMOL is a method of training a single LM that learns not only to answer the question given the context but also to generate the context, the question, and the answer given a generation token. In other words, a model plays the role of both the language model and the question-answering model in LAMOL. Hence, answering questions and generating pseudo-old samples can be done using a single model. During LLL, these pseudo-old samples are trained with new samples from new tasks to mitigate catastrophic forgetting. However, LAMOL generates only a certain number of pseudo-samples and use them to train the next task, so these samples are suboptimal for previous tasks.

Therefore, we propose an MFK-LAMOL, an advanced LAMOL with an efficient generative replay construction method that ensures pseudo-old samples contain the most forgotten knowledge for previous tasks. Specifically, our method builds pseudo-samples by selecting examples based on the extent to which they include knowledge that has been forgotten after learning new information. Figure 1 illustrates the difference between the original LAMOL and proposed MFK-LAMOL. When the LLL model received a new task, a goal-oriented dialogue, after training the previous two tasks, new samples from new tasks are trained with pseudo-old samples generated by the previous language model in LAMOL. However, MFK-LAMOL generates sufficient pseudo-old samples and selects the examples with the most forgotten knowledge to train more optimal examples for retraining the forgotten knowledge efficiently. To select appropriate pseudo-old samples, we propose an MFK metric that measures the extent to which previous information is forgotten by the current model.

The remainder of this paper is organized as follows. In Section 2, we briefly review lifelong learning. In Section 3, we describe our proposed method to improve LAMOL with the most forgotten knowledge. In Section 4, the experimental settings are explained, followed by the experimental results. Finally, we conclude our study in Section 5.

II. RELATED WORK

In this section, we briefly review the main approaches of continual lifelong learning. Three approaches have been developed as follows: regularization, architecture, and replay-based approaches [4], [15]–[17].

A. REGULARIZATION-BASED APPROACHES

Regularization-based approaches mitigate catastrophic forgetting by consolidating the model parameters that are important to previous tasks. Elastic weight consolidation (EWC) [18] reduces the loss of previously learned knowledge by reducing learning on the weights important for previous tasks. It remembers previous tasks using a quadratic penalty on the difference between the parameters for the old and new tasks with the Fisher information matrix to represent the importance of parameters. Synaptic intelligence (SI) [19] rapidly stores new memories without forgetting old memories by accumulating task-relevant information and exploiting this information. In SI, the importance of each parameter is derived from its contribution to the change in the total loss. Memory-aware synapses (MAS) [20] effectively prevent overwriting essential parameters related to previous information by constraining changes in these parameters. In this process, the importance of the parameters is computed based on the gradients of the model outputs, that is, the sensitivity of the predicted output function in response to the changes in these parameters.

B. ARCHITECTURE-BASED APPROACHES

Architecture-based approaches remember previous tasks by expanding model architectures or allocating the capacity of a fixed-sized model to each task. A progressive neural network (PNN) [21] consists of multiple columns of a neural network for each task. PNN enables knowledge transfer by utilizing prior knowledge from connections to previously learned features. When a new task is received, a new column of the neural network for the task is added and connected to the frozen past trained columns to transfer knowledge from old tasks. A dynamically expanding network (DEN) [22] can transfer knowledge and train a new task by dynamically deciding its network capacity according to the difficulty of the task. DEN is efficiently trained online by performing selective retraining, dynamic network expansion, and network split/duplication. Specifically, selective retraining trains only the parameters related to the task, and dynamically expands network capacity using group sparse regularization, and network split/duplication effectively prevents semantic drift. PathNet [23] reuses subsets of a neural network to transfer knowledge between tasks. Unlike PNN and DEN, PathNet does not allow the model to expand. It builds a huge fixed-size model and trains the best combination of neural networks and paths between different layers of neural networks for each task. PackNet [24] preserves previously learned information by allocating the parameters employed to learn new tasks in a fixed-size network. By performing iterative pruning and network retraining, numerous tasks can be sequentially packed into a single network.

C. REPLAY-BASED APPROACHES

Replay-based approaches alleviate catastrophic forgetting by storing a few examples in old tasks or remembering the previous distribution and continually learning them with emerging new tasks. Gradient episodic memory (GEM) [25] preserves a subset of real samples from previous tasks that can transfer knowledge across tasks. GEM constrains the gradients of the parameters by training them with a new task. Averaged-GEM (A-GEM) [26] is an improved version of GEM that is more efficient in computation and memory while achieving the same or better performance than GEM. Unlike GEM and A-GEM, deep generative replay (DGR) [27] mitigates catastrophic forgetting by generating pseudo-samples based on the distribution of previous tasks. The DGR is a dual-model architecture consisting of a deep generative model and task solver. Deep generative model generates pseudo-old samples according to the previously learned distribution, and the task solver solves the task by learning the new data and the generated ones together. In this way, training data from previous tasks can be sampled based on the generated pseudo-data and interleaved with information from the new tasks.

Among the three main approaches, replay-based approaches have proven to be the most promising for NLP tasks. MbPA++ [12] is an encoder-decoder architecture with episodic memory. To achieve LLL, MbPA++ distills the knowledge of previous tasks into sparse episodic memory and replays it afterward. Local adaptation is performed during the inference. LAMOL [13] is also an encoder-decoder architecture, but it does not use any extra memory. LAMOL mitigates catastrophic forgetting by means of pseudo-sample generation, as it makes the model itself generate data from previous tasks and use it when the model is trained on a new task. Although it has the advantage of addressing the memory problem, its main limitation is that it uses suboptimal data generated from the model itself.

III. METHOD

In this section, we first demonstrate LAMOL [13], based on which our proposed method is built. Next, we demonstrate the proposed MFK-LAMOL, focusing on the generative replay of LAMOL consisting of suboptimal pseudo-old samples.

LAMOL is a single LLL model comprising a language model (LM) and question-answering model (QA). To address a stream of different NLP tasks, data for various tasks are transformed into a question-answering format used in decaNLP [14]. These data consist of context, questions, and answers. QA is trained to answer a question given a context, and the LM is trained to generate a context, question, and answer given a generation token.

To pursue LLL, LAMOL allows a single model not only to answer questions but also to generate pseudo-old samples that follow the distribution of previous tasks. Before training a new task, LM uses the generation token as input and generates pseudo-samples representing the distribution of previous tasks using top-k sampling. At this time, the size of the generated samples is the *size of new data* $\times$ *sampling ratio*. LAMOL learns a new task based on a mixture of pseudo-old samples and new data. During training, LM and QA are simultaneously optimized, and the loss function is a weighted sum of the text prediction loss for LM and the

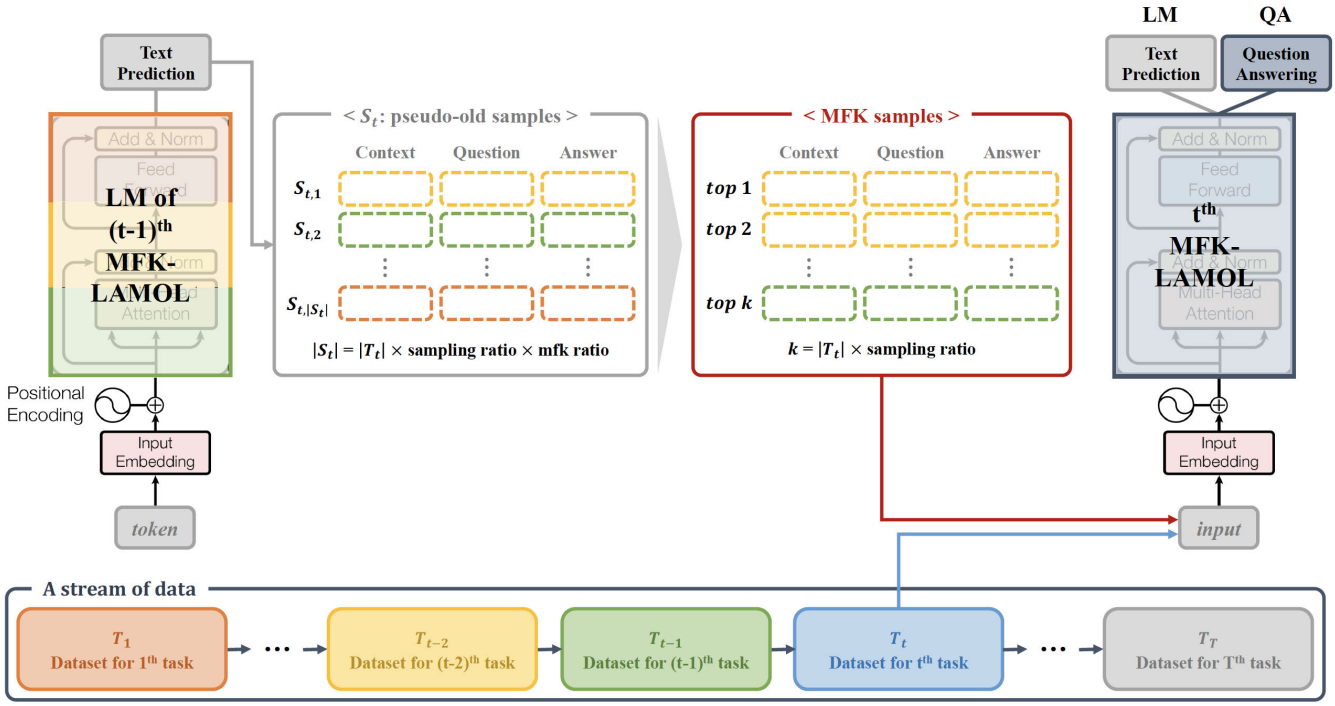


FIGURE 2. Overall framework of our proposed MKF-LAMOL method. When the t-th task is received, pseudo-old samples are sufficiently generated from LM of the $(t-1)$-th MFK-LAMOL trained on a stream of tasks of $t-1$, $\{T_1, T_2, \ldots, T_{t-1}\}$. From these samples, the model selects the top-k MFK samples based on the proposed metric. Specifically, MFK samples have the most forgotten knowledge after training a new task. Once MFK samples are determined, the model is trained on a mixture of new task and MFK samples.

question-answering loss for QA. Moreover, LM can use GEN or TASK tokens for generation. The GEN token means that the LM uses the same token for all tasks. In contrast, TASK tokens are task-specific tokens for each task, and thus all previous tasks have the same share of the generated pseudo-samples. However, these samples are likely to be suboptimal because LAMOL generates only a certain number of samples and uses all of them to train the next task. To alleviate this suboptimality, we propose a more effective version of LAMOL, called MFK-LAMOL.

MFK-LAMOL is LAMOL with effective pseudo-old samples, which enhances the way it constructs a generative replay. To use more important previous data, we generated sufficient pseudo-old samples and selected those with the most forgotten knowledge for previous tasks. As shown in Figure 2, when a new task is received, LM generates pseudo-samples of the *size of new data* $\times$ *sampling ratio* $\times$ *MFK ratio*. They are used to draw important samples to remember previously learned knowledge. In every epoch, only the examples of the *size of new data* $\times$ *sampling ratio*, which is the same as the previous samples used in LAMOL, are extracted from these pseudo-old samples based on how much important information is contained to preserve previous tasks. Once MFK samples are determined, MFK-LAMOL is trained on a mixture of new data and pseudo-old samples with the most forgotten knowledge.

The metric for selecting the MFK samples is as follows:

$$MFK\ metric = \mid L(f_\theta(x), y) - L(f_{\theta'}(x), y) \mid, \qquad (1)$$

where $\theta' = \theta - \alpha \nabla L(f_\theta(x), y)$, f is the QA of MFK-LAMOL, α is the learning rate in each epoch, and x and y are the pseudo-sample and its label, respectively. The metric represents the change in loss before and after the gradient update for each generated pseudo-sample x. The higher the value of the MFK metric, the more rapidly the MFK-LAMOL changes when it learns from the previous example. In other words, a pseudo-sample with a significant MFK metric contains substantial previous information that the model has forgotten while learning a new task. The absolute value is used for the proposed metric because the change in loss can be negatively affected by the learning rate. In this process, we improve LAMOL with an efficient method to construct a generative replay.

The key to alleviating catastrophic forgetting is the accumulation of knowledge from previous tasks. To remember previous tasks, the model should learn the previous data distribution while training on a new task. In this process, the importance of the information in generative replay determines the degree of preventing catastrophic forgetting. Therefore, MFK samples can improve the performance of the LAMOL. A complete view of MFK-LAMOL is presented in Algorithm 1.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASETS

We evaluated the proposed method on a stream of multiple NLP tasks. We conducted experiments on three disparate

TABLE 1. Summary of datasets for LLL. Each dataset has a different task, size, and its own metric: EM is an exact match between texts, nF1 represents the normalized version of the F1 score, and dsEM is a turn-based dialogue state exact match.

Dataset	Task	Train	Test	Metric
SST	Sentiment analysis	6,920	1,821	EM
QA-SRL	Semantic role labeling	6,414	2,201	nF1
WOZ	Goal-oriented dialogue	2,536	1,646	dsEM

TABLE 2. Summary of averaged metric scores for each method with permuted task orders using models at the last epoch of the last task. Multitask learning (bottom), in which all tasks are trained simultaneously, is an upper bound.

Methods	SST-SRL-WOZ	SST-WOZ-SRL	SRL-SST-WOZ	SRL-WOZ-SST	WOZ-SST-SRL	WOZ-SRL-SST	Average	Std
$LAMOL^{0.05}_{GEN}$	79.6	78.9	73.1	73.7	68.6	75.7	74.9	**3.4**
$MFK-LAMOL^{0.05}_{GEN}$	80.7	76.0	76.9	73.8	69.8	72.4	**74.9**	3.5
$LAMOL^{0.2}_{GEN}$	80.0	80.7	79.6	78.7	78.4	80.5	79.7	0.8
$MFK-LAMOL^{0.2}_{GEN}$	81.3	82.0	81.9	80.8	82.5	81.5	**81.7**	**0.6**
$LAMOL^{0.05}_{TASK}$	77.3	76.9	78.1	74.7	73.4	75.8	76.0	1.5
$MFK-LAMOL^{0.05}_{TASK}$	80.7	77.4	79.0	78.2	80.0	79.5	**79.3**	**1.1**
$LAMOL^{0.2}_{TASK}$	79.4	79.9	80.1	78.7	79.8	79.0	79.5	**0.5**
$MFK-LAMOL^{0.2}_{TASK}$	81.0	80.7	80.4	79.6	81.6	79.7	**80.5**	0.7
Multitask				81.5				

Algorithm 1 MFK-LAMOL

Input: A stream of tasks $\{T_1, T_2, \ldots, T_T\}$, Sampling ratio γ, MFK ratio ρ
Output: f_{θ_T}

```
for t = 1 to T do
  if t = 1 then
    Train f_{θ_t} on T_t
  else
    Generate pseudo-old samples S_t from f_{θ_{t-1}}, where | S_t |= ρ(γ | T_t |)
    for epoch = 1 to n do
      Calculate the criterion for all pseudo-samples
      Sort samples in descending order by criterion
      Select top-k S_t for MFK examples S_t^{epoch}, where k = γ | T_t |
      Train f_{θ_t}^{epoch} on T_t ∪ S_t^{epoch}
    end for
  end if
end for
```

tasks introduced in decaNLP [14]: sentiment analysis, semantic role labeling, and goal-oriented dialogue with a dataset for each task. For each task, there is a corresponding evaluation metric because they are fundamentally different tasks. Table 1 presents a description of the datasets used in our experiments. Moreover, we follow the same procedure as LAMOL [13] to produce equal-sized datasets.

2) MODELS

We evaluated the effect of our proposed method by comparing LAMOL and MFK-LAMOL. We conducted experiments under the same experimental settings as LAMOL [13]. Both methods use the smallest pre-trained GPT-2 [28] for LM. $MFK-LAMOL^{\gamma}_{token}$ denotes our proposed method with a *token* and a sampling ratio of γ. *token* is chosen between *GEN* and *TASK*, indicating that a task-general or task-specific token, respectively. The sampling ratio γ is set to 0.05 or 0.2, corresponding to the experimental settings for LAMOL. In all experiments on MFK-LAMOL, the MFK ratio ρ is set to 3. Moreover, multitask learning is used to be compared. It is the method where all tasks are trained simultaneously. Multitask learning is often seen as an upper bound of LLL. All baseline numbers are reported in [13].

B. RESULTS

1) LIFELONG LANGAUGE LEARNING

We conducted a LLL experiment on three small datasets SST, QA-SRL, and WOZ in the form of a question-answering. We trained our proposed method on all six permutations of the task order and all four combinations of the two tokens and two sampling ratios. The final score for each order was averaged metric scores for all three datasets because each task has a different metric. At this time, we evaluated the model at the last epoch of the last task after training on a stream of all three datasets. The results are shown in Table 2.

When the proposed method was applied to the LAMOL, the performance improved significantly in the overall setting. In the case of $MFK-LAMOL^{0.05}_{TASK}$, the proposed method produced the greatest improvement in the performance of LAMOL at 3.3%. From these results, we can conclude that the use of the most forgotten knowledge is significant in preventing catastrophic forgetting. $MFK-LAMOL^{0.2}_{GEN}$, our best performing method, outperformed multitask learning, which is considered as the upper bound of LLL, under many task sequence combinations. This implies that $MFK-LAMOL^{0.2}_{GEN}$ has almost no catastrophic forgetting during LLL. Moreover, MFK-LAMOL showed similar results to LAMOL with respect to the change in γ and *token*. MFK-LAMOL exhibited improved performance

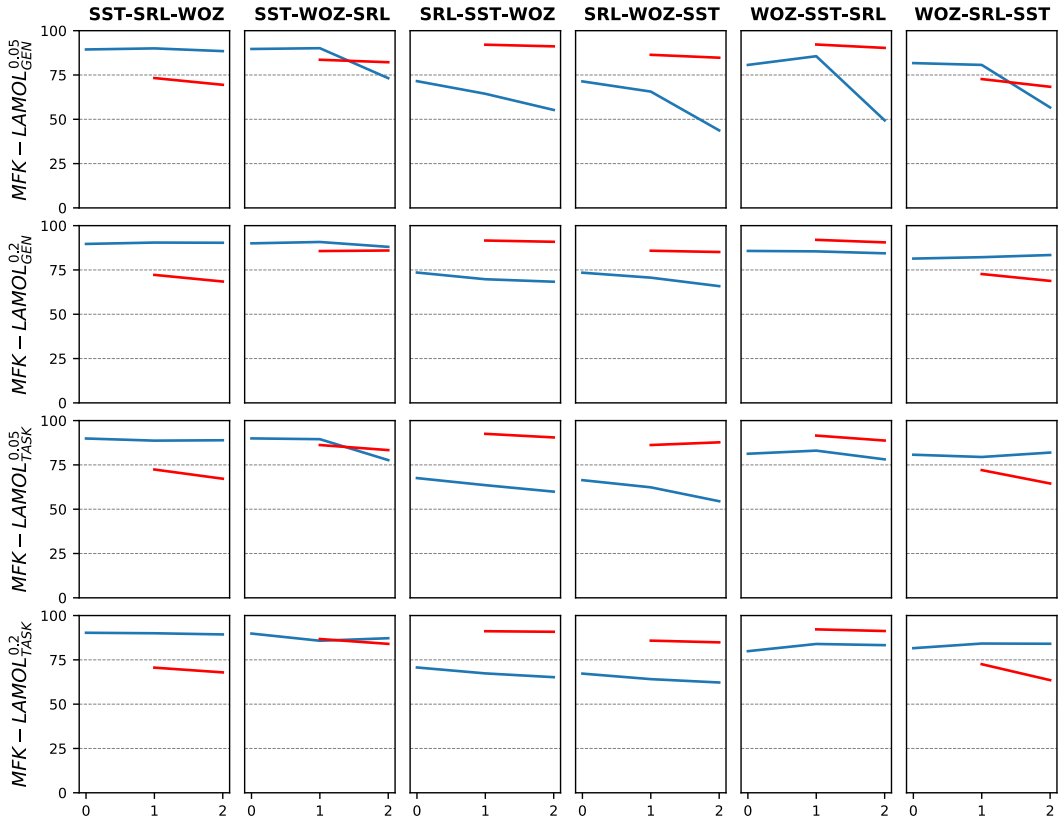


FIGURE 3. Overall catastrophic forgetting progress of the proposed method on the permuted task orders with different settings. The blue line indicates the scores for the first task after training each task sequentially. The red line corresponds to the scores for the second task.

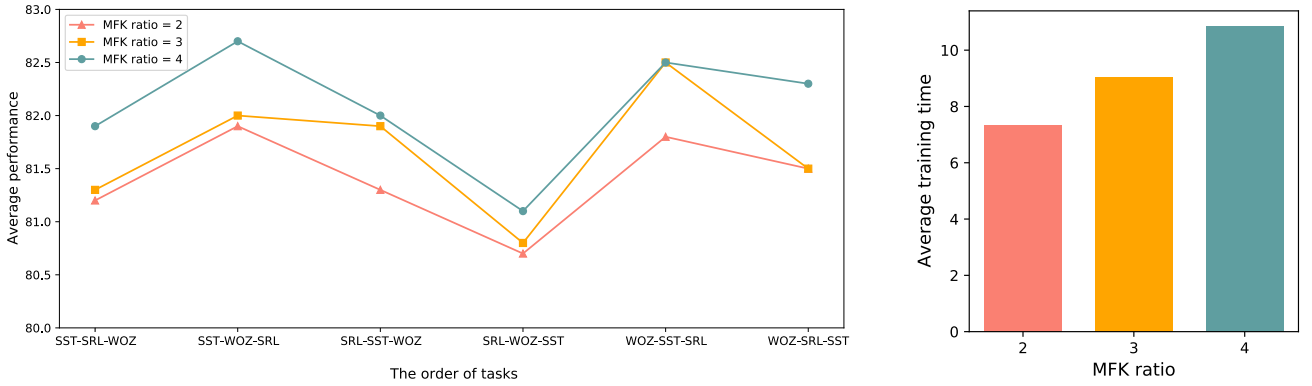
at higher values of γ. *token* showed different results according to γ in MFK-LAMOL. When γ is large, the proposed method is likely not to depend on *token*. In contrast, the task-specific tokens had a positive effect on both methods because the tokens caused the generative replay to retrain a substantial portion of the first task.

2) CATASTROPHIC FORGETTING

We confirmed the progress in mitigating catastrophic forgetting with MFK-LAMOL and compared it with LAMOL. The results are shown in Figure 3, and those of the LAMOL are shown in Figure 5 in [13]. These figures show how the performance of the first and second tasks changed after learning the next tasks. Overall, MFK-LAMOLs tended to significantly preserve previously learned knowledge. When WOZ was the first task, $MFK-LAMOL^{0.05}_{GEN}$ faced catastrophic forgetting similar to LAMOL. However, $MFK-LAMOL^{0.05}_{TASK}$, whose performance was significantly higher than that of $LAMOL^{0.05}_{TASK}$, considerably alleviated catastrophic forgetting. Furthermore, the order of tasks is crucial to the performance of LAMOL. For instance, the WOZ score dropped significantly after training in other tasks when WOZ was not the last task. In contrast, MFK-LAMOL solved this problem, except in the setting of $MFK-LAMOL^{0.05}_{GEN}$. It is considered that selecting the MFK samples from sufficiently generated samples can enhance the suboptimality of LAMOL. Finally, backward knowledge transfer was shown in the case of WOZ-SRL-SST for all models except $MFK-LAMOL^{0.05}_{GEN}$. Backward knowledge transfer means that the model achieves improved performance on previous tasks after learning a new task. Therefore, the WOZ scores improved after learning other tasks.

3) EFFECT OF PARAMETER

In this section, we investigate the effects of the MFK ratio ρ, which is the only hyperparameter introduced by MFK-LAMOL. All experiments were performed using $MFK-LAMOL^{0.2}_{GEN}$, which is the best performing method. The main objective of this ablation study is to investigate how MFK-LAMOL responds to different ρ values: the larger the ρ value, the more pseudo-old samples are selected in every epoch. Note that if we increase the number of pseudo-old samples, it is more likely to select near-optimal samples, but the computational cost, that is, training time, is also increased. Figure 4 shows the results obtained using the three different MFK ratios. These results show that a larger MFK ratio improves the performance of the proposed method by increasing the possibility that the optimal samples should be retrained. In our experiments, we used $\rho = 3$, considering both performance and training time.



(a) Average Performance

(b) Average training time

FIGURE 4. Effect of MFK ratio ρ based on $MFK - LAMOL^{0.2}_{GEN}$. MFK ratio is set to three different value {2, 3, 4}. Each point in (a) is averaged metric scores for each task inferenced from the model with a different MFK ratio. Average training times of the six orders for each MFK ratio are shown in (b).

V. CONCLUSION

Lifelong continual learning is an important research area to reach artificial intelligence of human level. Therefore, numerous studies have been conducted, and recently great progress has been made owing to the rapid development in images and games. However, only a few researches have been proposed on LLL. Therefore, we focus on LLL and improve LAMOL, the representative LLL model. LAMOL is an effective LLL method that does not require additional memory or model capacity based on language modeling. LAMOL is a language model that simultaneously learns to solve the tasks and generate training samples. When the model is trained on a new task, LAMOL replays pseudo-samples of the old tasks and uses them with the data for the new task. However, these generated samples are likely to be suboptimal because LAMOL generates only as many samples as needed and uses all of them for training new tasks.

Therefore, we propose MFK-LAMOL, which uses pseudo-old samples that can effectively preserve previous tasks based on LAMOL. These samples contain a lot of old knowledge that has been forgotten while training for new task. Specifically, when MFK-LAMOL begin to train new task, it sufficiently generates pseudo-samples and selects the best examples that can efficiently remember the forgotten knowledge based on the proposed MFK metric. From this process, we improve LAMOL by alleviating the suboptimality problem of LAMOL. Moreover, we showed that the proposed method outperforms LAMOL on a stream of three different NLP tasks. It implies that retrieving the examples can alleviate catastrophic forgetting efficiently.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*. [Online]. Available: http://arxiv.org/abs/1312.5602

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," 2016, *arXiv:1611.01599*. [Online]. Available: http://arxiv.org/abs/1611.01599

[4] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.

[5] J. Fagot and R. G. Cook, "Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 46, pp. 17564–17567, Nov. 2006.

[6] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychol. Learn. Motivat.*, vol. 24, pp. 109–165, Dec. 1989.

[7] J. L. McClelland, B. L. McNaughton, and R. C. O'reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," *Psychol. Rev.*, vol. 102, no. 3, p. 419, 1995.

[8] Z. Chen, N. Ma, and B. Liu, "Lifelong learning for sentiment classification," 2018, *arXiv:1801.02808*. [Online]. Available: http://arxiv.org/abs/1801.02808

[9] S. Lee, "Toward continual learning for conversational agents," 2017, *arXiv:1712.09943*. [Online]. Available: https://arxiv.org/abs/1712.09943

[10] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Lifelong domain word embedding via meta-learning," 2018, *arXiv:1805.09991*. [Online]. Available: http://arxiv.org/abs/1805.09991

[11] T. Liu, L. Ungar, and J. Sedoc, "Continual learning for sentence representations using conceptors," 2019, *arXiv:1904.09187*. [Online]. Available: http://arxiv.org/abs/1904.09187

[12] C. de Masson d'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13122–13131.

[13] F.-K. Sun, C.-H. Ho, and H.-Y. Lee, "LAMOL: LAnguage MOdeling for lifelong language learning," 2019, *arXiv:1909.03329*. [Online]. Available: http://arxiv.org/abs/1909.03329

[14] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," 2018, *arXiv:1806.08730*. [Online]. Available: http://arxiv.org/abs/1806.08730

[15] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," 2019, *arXiv:1909.08383*. [Online]. Available: http://arxiv.org/abs/1909.08383

[16] M. Biesialska, K. Biesialska, and M. R. Costa-jussà, "Continual lifelong learning in natural language processing: A survey," 2020, *arXiv:2012.09823*. [Online]. Available: http://arxiv.org/abs/2012.09823

[17] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," 2021, *arXiv:2101.10423*. [Online]. Available: http://arxiv.org/abs/2101.10423

[18] K. James, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, and D. Hassabis, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[19] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.
[20] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 139–154.
[21] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," 2016, *arXiv:1606.04671*. [Online]. Available: http://arxiv.org/abs/1606.04671
[22] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," 2017, *arXiv:1708.01547*. [Online]. Available: http://arxiv.org/abs/1708.01547
[23] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "PathNet: Evolution channels gradient descent in super neural networks," 2017, *arXiv:1701.08734*. [Online]. Available: http://arxiv.org/abs/1701.08734
[24] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.
[25] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," 2017, *arXiv:1706.08840*. [Online]. Available: http://arxiv.org/abs/1706.08840
[26] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2018, *arXiv:1812.00420*. [Online]. Available: http://arxiv.org/abs/1812.00420
[27] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," 2017, *arXiv:1705.08690*. [Online]. Available: http://arxiv.org/abs/1705.08690
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.



HEEJEONG CHOI received the B.S. degree in statistics from Ewha Womans University. She is currently pursuing the Ph.D. degree in industrial and management engineering with Korea University. Her main research interests include image processing and anomaly detection.



PILSUNG KANG received the B.S. and Ph.D. degrees in industrial engineering from Seoul National University. He is currently an Associate Professor with the School of Industrial Management Engineering, Korea University, Republic of Korea. His main research interest includes developing machine learning algorithms for both structured data and unstructured data (image, video, and text) and applying them to solve engineering and business problems, such as fault classification in manufacturing, abnormal behavior detection from system logs, and sentiment classification from news and review texts.