VNU Journal of Science: Comp. Science & Com. Eng. Vol. 40, No. 1 (2024) 60–70

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Contrastive Learning for Boosting Knowledge Transfer in Task-Incremental Continual Learning of Aspect Sentiment Classification Tasks

Thanh Hai Dang*, Quynh-Trang Pham Thi, Duc-Trong Le, Tri-Thanh Nguyen

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 29 November 2023;
Revised 04 May 2024; Accepted 15 May 2024

Abstract: Continual learning (CL) aims to learn a sequence of tasks, with task datasets emerging incrementally over time, and without a predetermined number of tasks. CL models strive to achieve two primary objectives: preventing catastrophic forgetting and facilitating knowledge transfer between tasks. Catastrophic forgetting refers to the sharp decline in the performance of CL models on previously learned tasks as new ones are learned. Knowledge transfer, which leverages acquired knowledge from previous tasks, empowers the CL model to adeptly tackle new tasks. However, only a few CL models proposed by far successfully achieve those two objectives simultaneously. In this paper, we present a task-incremental CL based model that leverages a pre-trained language model (i.e., BERT) with injected CL-plugins to mitigate catastrophic forgetting in continual learning. Additionally, we propose the utilization of two contrastive learning-based losses, namely contrastive ensemble distillation (CED) and contrastive supervised learning of the current task (CSC) losses, to enhance our model's performance. The CED loss improves the knowledge transferability of our continual learning model, while the CSC loss enhances its performance for the current learning task. Experimental results on benchmark datasets demonstrate that our proposed model outperforms all existing continual learning models in the task-incremental learning setting for continual aspect sentiment classification.

Keywords: Continual Learning, Contrastive Learning, Aspect-Sentiment Classification.

1. Introduction

Continual learning (CL) is designed to learn a sequence of tasks where task datasets become available progressively over time, and the number of tasks is not predetermined [1]. This learning paradigm allows leveraging acquired knowledge from past tasks without storing previously trained

*Corresponding author
E-mail address: hai.dang@vnu.edu.vn
https://doi.org/10.25073/2588-1086/vnucsce.1833

data, thereby addressing security concerns. There are three popular settings for continual learning systems [2], namely Task incremental learning (TIL), Domain incremental learning (DIL) and Class incremental learning (CIL). TIL needs to know which task it is handling in the testing phase. TIL builds a prediction component specific for each task within one unified network. During the inference, it is routed to the task-specific component within the learned model based on the provided task identity. In contrast to TIL, DIL does not require knowing the task identity when testing. DIL aims to learn a single task, but in different contexts. Both TIL and DIL demand that when learning a new task the data labels must be seen in the previous tasks. CIL relaxes this strict requirement. It can automatically learn incoming data of a new class/label that did not appear when learning the previous tasks [3].

CL encounters a significant obstacle known as catastrophic forgetting, wherein the model's parameters being already learned for previous tasks must be readjusted to optimize performance for a new task. As a consequence, this significantly diminishes the model's effectiveness on previously learned tasks. Early CL models [4] based on fine-tuning a pre-trained language model BERT [5] subsequently for coming tasks are severe victims of this parameters' readjustment [1, 6–8]

To mitigate catastrophic forgetting, a natural approach is to minimize reconfiguration of models' parameters as much as possible. To this end, various CL models have been proposed, of which each task has its own specific module (or subset of the model's parameters) to be optimized. When learning a new task, only the module specific for this task is tuned while others remain untouched. These modules can be task-specific adapters [9, 10] or continual learning plugins (CL-plugin) [11] or a relatively small 2-layer fully connected network (so-called a Capsule Networks) [12, 13], which are injected into the pre-trained BERT.

The isolation of task-specific modules helps CL models mitigate catastrophic forgetting. However, it hinders (or even prohibit from) the shared knowledge utilization (update included) among tasks, which can be useful for enhancing the model's performance on related tasks. Knowledge utilization enables CL models to leverage insights acquired from past learned tasks in one domain and apply them effectively to new tasks in other domains. Without effective knowledge exploitation mechanisms, CL models may struggle to adapt to the nuances of different domains, resulting in reduced performance and reliability. This contradiction causes CL hard to achieve both objectives together, namely preventing catastrophic forgetting and facilitating knowledge transfer. They are still two significant challenges for continual learning.

To our knowledge, only a few proposed CL models successfully achieve those two objectives at the same time. Among them, CTR [11] and CLASSIC [14] recently proposed are two well-established CL models. The former belongs to the TIL setting while the latter is of the DIL setting. Both are dedicated to the Aspect Sentiment Classification (ASC) tasks. ASC is a task that identifies a sentiment about the aspect of an object whether positive, negative, or neutral. For example, the sentence "the mic quality is quite nice" is an opinion about the aspect of "mic quality" of a mobile phone object. This sentence should be classified as a "Positive" sentiment by an ASC model.

Given a text, Aspect Sentiment Classification (ASC) involves determining whether the sentiment towards a specific aspect of an object is positive, negative, or neutral. For instance, the sentence "the mic quality is quite nice," expresses the sentiment pertaining to the aspect "mic quality" of a mobile phone. An ASC model should correctly classify this sentence as conveying a "Positive" sentiment. In the realm of Aspect Sentiment Classification (ASC) continual

learning, discerning sentiment nuances are about various aspects and/or objects of totally different evolving contexts and domains.

In the realm of Aspect Sentiment Classification (ASC) continual learning, the challenges of catastrophic forgetting and insufficient knowledge transfer are particularly pronounced due to the intricate nature of sentiment analysis tasks. ASC models, tasked with discerning sentiment nuances across various aspects and domains, confront a relentless stream of new data and evolving contexts. Inevitably, this causes ASC continual learning models to suffer from severe catastrophic forgetting.

Even when domains for ASC tasks largely differ, sentiment expressions inherent to one domain still share some commonalities with those in other domains, as they are all conveyed using the same language. Therefore, the importance of knowledge transfer in Continual Learning for ASC becomes even more pronounced and crucial for ensuring robust adaptation and accurate sentiment analysis, particularly when dealing with disparate domains. This process allows a CL model to grasp domain-specific sentiment patterns and nuances more efficiently, leading to improved accuracy and generalization across diverse domains.

In this work, we explore how much knowledge should be transferred in continual learning and empirically demonstrate its effectiveness for sequential aspect sentiment classification (ASC) tasks in the TIL setting. To this end, we propose a continual ASC model in the TIL setting that selectively transfers knowledge across ASC tasks. This is achieved by utilizing a normalized cross-attention contrastive loss in the classification head of the model. In addition, our model enhances its performance on the current task through contrastive supervised learning (CSC).

To evaluate our model, we conduct extensive experiments on 19 benchmark ASC data sets from various domains. Experimental results indicate that our model outperforms recent state-of-the-art continual learning ASC models in the TIL setting, thus demonstrating the effectiveness of our model in transferring selective distilled knowledge between tasks for continual learning.

The rest of this paper is organized as follows: Section 2 reviews the related work; Section 3 presents our proposed method; Section 4 details the experimental results; and the final section concludes with future directions.

2. Related Work

Various advanced approaches have been proposed for continual learning to mitigate CF [1], such as Regularization-based approach [15] and Memory (or Replay) based approach [16–18]. The former estimates the importance of parameters for previously learned tasks to penalize changes in these parameters to avoid catastrophic forgetting. The latter stores a few samples of all previous tasks and then exert these in-memory data to update the parameters of the current task. As a result, it helps CL models retain the knowledge of the previous tasks, thereby mitigating catastrophic forgetting.

Recent years have seen a surge of research efforts focusing on the integration of advanced pre-trained models, such as BERT (Bidirectional Encoder Representations from Transformers) [5], into continual learning approaches. BERT adapters, lightweight task-specific neural network components, have emerged as a promising solution [9], which chooses to freeze the parameters of the pre-trained BERT, only tuning a small number of parameters of injected task-specific adapters [9–11]. A BERT adapter can be a relatively small 2-layer fully connected network or a Capsule Networks (CapsNet) [12, 13] that uses vector capsules instead of scalar feature detectors. These task-specific adapters enable CL models to learn new tasks while preserving knowledge from previous tasks. Researchers have explored combining regularization and

memory-based techniques with BERT adapters to further enhance their performance in continual learning scenarios. Regularization methods are applied to adapter parameters to mitigate forgetting, while memory-based strategies such as Experience Replay and Generative Replay are integrated to retain knowledge from past tasks. By leveraging the strengths of both pre-trained models and continual learning techniques, these approaches aim to develop robust and adaptable models across diverse tasks and domains.

Although the CL capability of existing advanced models have been empirically demonstrated they still struggle with CF issues when tasks lack substantial shared knowledge [11, 14].

Isolating task-specific modules in CL models aids in mitigating catastrophic forgetting to some extent. However, it poses a challenge by limiting the utilization and updating of shared knowledge across tasks, which could enhance the model's performance on related tasks. This contradiction makes it difficult for CL to simultaneously achieve both objectives: preventing catastrophic forgetting and facilitating knowledge transfer.

As far as we know, only a handful of continual learning (CL) models have effectively achieved both objectives simultaneously. Notably, CTR [11] and CLASSIC [14] are two prominent CL models that have recently demonstrated success in this regard. Both models employ task masks to isolate task-specific knowledge for dealing with CF. While CTR operates within the TIL setting, CLASSIC is designed for the DIL scenario. Remarkably, both models are specifically tailored for Aspect Sentiment Classification (ASC) tasks.

CLASSIC introduces an approach for knowledge transfer though contrastive learning, focusing on domain continual learning (i.e. DIL setting). Its key innovation lies in a contrastive continual learning method facilitating both knowledge transfer across tasks and distillation from old tasks to the new task. CLASSIC utilizes BERT-Adapter [9], maintaining BERT parameters unchanged while achieving performance comparable to BERT fine-tuning. By proposing task masks to isolate task-specific knowledge and a contrastive continual learning method for knowledge transfer and distillation, CLASSIC significantly enhances the accuracy of all tasks.

CTR utilizes a key component known as the CL-plugin that is injected into BERT at two locations. Comparing with BERT-Adapter, CL-plugins differ significantly. While an adapter is a simple 2-layer fully-connected network inserted into BERT (to be fine-tuned) for each specific end task, the CL-plugin functions as a capsule network, which deviates from traditional neural networks by utilizing vector-output capsules instead of scalar activations, thereby preserving additional information in a more nuanced manner. CL-plugin integrates a novel transfer routing mechanism that facilitates knowledge exchange across tasks while safeguarding task-specific information to prevent interference. CTR can learn all tasks using only one pair of CL-plugin modules inserted into BERT. By this strategical integration, CTR eliminates the need for individual fine-tuning of BERT for each task, which often leads to catastrophic forgetting [11].

3. Method

Taking inspiration from [11], we employ a continual learning plugin (CL-plugin) into our model within the TIL setting. The CL-Plugin comprises two essential modules, namely Knowledge Sharing Sub-Module (KSM) and Task-Specific Sub-Module (TSM). KSM identifies and transfers shareable knowledge from analogous previous tasks to the new task and TSM focuses on learning task-specific neurons and their associated masks. These masks serve to protect the neurons from updates by future tasks, thereby addressing the challenge of catastrophic forgetting (CF). Further, we

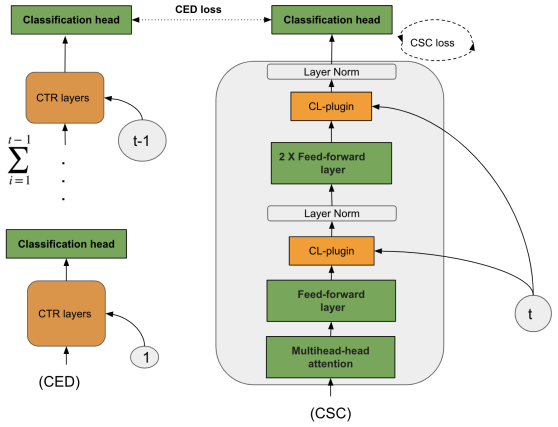

Figure 1. Overall architecture of our proposed model.

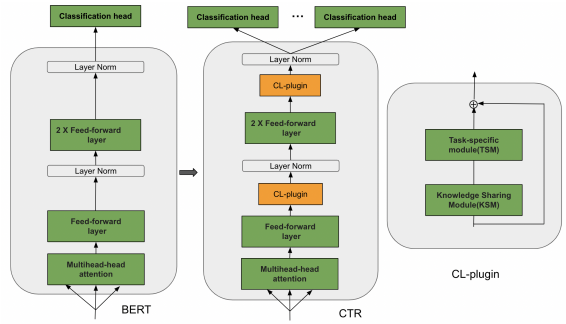

Figure 2. General architecture of BERT (left) and the CTR system (middle) and the CL-plugin (right) that are integrated into BERT [11].

propose to boost the TIL performance of our CL-plugin based model by integrating contrastive learning losses [14] into classification heads of our model, including contrastive ensemble distillation (CED) loss to capture shared knowledge across tasks and contrastive supervised learning on the current task (CSC) loss to further enhance task-specific performance. The CED loss encourages knowledge transfer further by distilling knowledge from all previous tasks into the current task. The CSC loss aims to improve the plasticity of the current task.

Fig. 1 illustrates the overall architecture of our proposed method. The detailed components are presented as follows.

3.1. Continual Learning Plugin (CL-plugin)

Given hidden states $h(t)$ extracted from the feed-forward layer within a transformer layer, and the task ID t, the CL-plugin yields outputs that comprise hidden states with informative features specific to the t^{th} task. Finally, CL-plugin employs the cross-entropy loss to optimize the model's parameters, expressed as follows:

$$\mathcal{L}_{CE} = \sum_{i=1}^{N} -y_i \log \hat{y}_i \quad (1)$$

where $\hat{y}_i$ is the probability score of the i^{th} sample in the batch data after passing through the model.

Inside the CL-plugin, the KSM consists of two capsule layers, namely the task capsule layer and the knowledge-sharing capsule layer, equipped with a dynamic routing algorithm. This configuration effectively clusters similar tasks and shared knowledge, enabling knowledge transfer among tasks with commonalities. In contrast, the TSM comprises differentiable fully-connected layers, with each layer's output undergoing additional processing using a task-specific mask. This mask indicates which neurons need protection for the specific task to address CF, preventing gradient updates on these neurons masked for the specific task during back-propagation when adapting to a new task (see Fig. 2). Specially, for a specific task ID t, $e_l^{(t)}$ denotes its embedding trained for the l^{th} layer within the TSM. This embedding encompasses differentiable parameters that can be learned concurrently with other components of the network. The task mask $m_l^{(t)}$, akin to a "soft" binary mask, is generated by employing a pseudo-gate function denoted as σ and incorporating a positive scaling hyper-parameter denoted as s throughout the training procedure. The calculation of $m_l^{(t)}$ is outlined as follows:

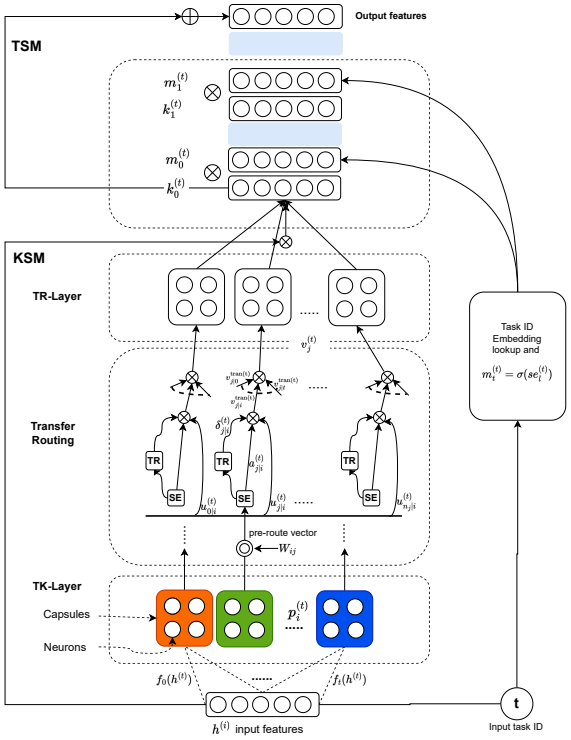

Figure 3. The architecture details of CL-plugin within our model.

$$m_l^{(t)} = \sigma(se_l^{(t)}) \tag{2}$$

For tasks sharing some common knowledge, their masks may have some neurons that coincide with each other. Each layer output $k_l^{(t)}$ in TSM when learning the task t is element-wise multiplied with $m_l^{(t)}$. To this end, the masked output $k^{(t)}$ from the last layer is then passed to the subsequent BERT model layer through a skip connection (see TSM in Fig. 3). The final $m_l^{(t)}$ for all layers after learning the task t are saved for further used.

3.2. Contrastive Ensemble Distillation (CED) Loss

This loss employs contrastive learning to distill knowledge from all previous tasks into the current task t. For each previous task i, the Contrastive Ensemble Distillation (CED) is calculated as follows:

$$\mathcal{L}_{\text{CED}}^{(i)} = \sum_{n=1}^{N} -\log \frac{\exp((z_n^{(i)}.z_n^{(t)})/\tau)}{\sum_{j=1,j\neq n}^{N} \exp((z_n^{(i)}.z_j^{(t)})/\tau)} \tag{3}$$

where N is the batch size; τ is an adjustable temperature parameter controlling the separation of classes; n is the index of the data sample in the batch; $z_n^{(i)}$ and $z_n^{(t)}$ are the logits for the same input data at the index n, generated from our model for the previous task i and current task t, respectively. This logit pair is treated as the positive pair and all of the other pairs are considered as negative pairs.

Since the model for each previous task remains fixed during the current task's training process, it effectively serves as a teacher for the current task (acting as the student). For $t-1$ models corresponding to all previous tasks, the final CED is formulated as Formula 4.

$$\mathcal{L}_{\text{CED}} = \sum_{i=1}^{t-1} \mathcal{L}_{\text{CED}}^{(i)} \tag{4}$$

3.3. Contrastive Supervised Learning of Current Task (CSC) Loss

Taking inspiration from CLASSIC [14], we formulate the Contrastive Supervised Learning of Current Task (CSC) Loss to enhance the performance of the current task, as follows:

$$\mathcal{L}_{\text{CSC}} = \sum_{n=1}^{N} \frac{1}{N_{y_n}-1} \sum_{j=n,y_j=y_n}^{N} \log \frac{\exp\left(\frac{h_n^{(t)}.h_j^{(t)}}{\tau}\right)}{\sum_{k=1,y_k\neq y_n}^{N} \exp\left(\frac{h_n^{(t)}.h_k^{(t)}}{\tau}\right)} \tag{5}$$

Where N_{y_n} is the number of samples in the data batch that have the same label as y_n, while $h_j^{(t)}$ is the masked output of the TSM for the j^{th} sample in the batch, derived from the layer before the classification head of the task t model.

3.4. Model's Loss

We combine three losses described above to optimize our CL model in the TIL setting, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CED}} + \mathcal{L}_{\text{CSC}} \quad (6)$$

where L_{CE} denotes the cross-entropy loss, L_{CED} represents the contrastive ensemble distillation loss calculated using Formula 4 and L_{CSC} indicates the contrastive supervise learning of current task as expressed in Formula 5.

4. Experimental Results

We conducted experimental evaluations, comparing existing non-continual learning and task-incremental continual learning models for a sequence of ASC tasks on benchmark datasets. The reported results are averaged across five random task orders.

4.1. Experiment Datasets

Like other state-of-the-art continual learning models in the TIL setting for ASC tasks (such as CTR [11]), we employ four benchmark datasets, namely: (1) HL5Domains [19], comprising review sentences for 5 products; (2) Liu3Domains [20], focusing on 3 products; (3) Ding9Domains [21], covering 9 products; and (4) SemEval14 [22], relating to 2 products. These datasets collectively encompass sentiments about 19 aspects, each treated as an individual aspect-based sentiment classification task. Detailed statistics for each dataset are provided in Table 1.

In alignment with previous studies, we partitioned 10% of the original data for validation and another 10% for testing, for each dataset (1), (2), and (3). However, for dataset (4), we opted for a validation set consisting of only 150 examples from the training set. To maintain methodological consistency with prior research, we present a comprehensive breakdown of the sample numbers for each task in Table 1.

4.2. Hyperparameters

Our approach utilises a two-layer fully connected network with 768 dimensions in the capsule layers and three transfer capsules. Concerning the task-specific modules, we employ 2000 dimensions for the final states. To optimize our model, we use the Adam optimizer and set the learning rate to $3e-5$. We conduct training for ten epochs on the SemEval datasets and extend it to 30 epochs for other datasets. For all the baseline models, we utilize the code provided by their respective authors and adapt it for classification purposes.

4.3. Results and Analysis

Firstly, we employ multi-task learning, an approach that incorporates comprehensive knowledge of all preceding tasks and is trained on the entire dataset, to evaluate the upper bound of this problem. Then, we employ non-continual learning baselines independently for each task. We have three baselines of such, namely BERT, BERT (frozen), and Adapter-BERT. Continuously, we set the foundation based on "no forgetting handling" (NFH). To this end, we compare our model against 12 state-of-the-art continual learning models in the TIL setting, including KAN [23], SRK [24], HAT [25], CAT [6], UCL [26], EWC, L2 [15], OWM [27], A-GEM [28], DER++ [29], BCL [10], LAMOL [30].

Due to imbalanced classes, we calculate both Accuracy and Macro-F1 (MF1) scores to account for potential biases in performance evaluation. As shown in Table 2, our model demonstrates strong continual learning ASC performance, achieving an Accuracy of 88.5% and a Macro-F1 score of 82.35%. Notably, our model outperforms all baseline models, including state-of-the-art (SOTA) continual learning models, in terms of Macro-F1. While our model's Accuracy ranks second, trailing behind LAMOL by only 0.41%, it's worth noting that LAMOL is based on GPT-2, whereas our model is based on

Table 1. Number of sentences per dataset, referred to as tasks in the context of continual learning [11].

Data source	Task/domain	Train	Validation	Test
Liu3domain	Speaker	352	44	44
	Router	245	31	31
	Computer	283	35	36
HL5domain	Nokia6610	271	34	34
	Nikon4300	162	20	21
	Creative	677	85	85
	CanonG3	228	29	29
	ApexAD	343	43	43
Ding9domain	CanonD500	118	15	15
	Canon100	175	22	22
	Diaper	191	24	24
	Hitachi	212	26	27
	Ipod	153	19	20
	Linksys	176	22	23
	MicroMP3	484	61	61
	Nokia6600	362	45	46
	Norton	194	24	25
SemEval14	Rest.	3452	150	1120
	Laptop	2163	150	638

BERT. To ensure a fair comparison, we replicate experiments with CTR using the same task orders as those employed in our model experiments (Results denoted by the label CTR*). However, CTR* yields inferior results compared to our model in terms of both Accuracy and Macro-F1 metrics. It's worth noting that the Accuracy and Macro-F1 values reported in their paper are obtained from different task orders, which are not disclosed. Additionally, compared to the non-continual learning method Adapter-BERT, our continual learning model demonstrates a higher performance of 3% in both Accuracy and Macro-F1 metrics. In comparison to the non-continual learning approach of Adapter-BERT, our continual learning model achieves a superior performance, with a 3% increase in both Accuracy and Macro-F1 metrics.

4.4. Ablation Study

We conducted experiments by removing specific components from our method, and the results in Table 3 highlight the effectiveness of the CL-plugin, CED, and CSC. The outcomes, averaged across two evaluation metrics (Accuracy and MF1) from 5 random task orders, demonstrate notable improvements. Specifically, when the CL-plugin is omitted, a two-layer fully connected adapter is employed. The results reveal that the inclusion of the CL-plugin enhances the model's performance by approximately 2.6% in both accuracy and MF1 score. Additionally, Table 3 illustrates the positive impact of incorporating two contrastive learning-based losses (CED and CSC) on the continual ASC performance of our model. We also conducted experiments with our model, incorporating another contrastive learning-based loss known as the Knowledge Sharing Loss (CKS), inspired by [14]. However, experimental results revealed that CKS did not yield any improvement for our model (data not shown). This could be attributed to the fact that our model already includes the KSM module, which facilitates knowledge sharing in a manner similar to CKS.

Table 2. Average Accuracy (Acc.) and Macro-F1 (MF1) over five random sequences of 19 tasks.

Scenario	**Category**	**Model**	**Acc.(%)**	**MF1(%).**
Non-continual Learning (SDL)	BERT	MTL	91.91	88.11
	BERT	SDL	85.84	76.35
	BERT (Frozen)	SDL	78.14	58.13
	Adapter-Bert	SDL	85.96	78.07
	W2V	SDL	77.01	51.89
Continual Learning	BERT	NFH	49.60	43.08
	BERT (Frozen)	NFH	85.51	76.64
	Adapter-BERT	NFH	85.51	76.64
	W2V	NFH	82.69	73.56
	BERT (frozen)	L2	56.04	38.40
		A-GEM	86.06	78.44
		DER++	84.27	75.08
		KAN	85.49	77.38
		SRK	84.76	78.52
		EWC	86.37	74.52
		UCL	83.89	74.82
		OWM	87.02	79.31
		HAT	86.74	78.16
		CAT	83.68	68.64
	Adapter-BERT	L2	63.97	52.43
		A-GEM	45.88	28.21
		DER++	47.63	35.54
		EWC	56.30	49.58
		UCL	64.46	36.64
		OWM	72.99	66.51
		HAT	86.14	78.52
		BCL	88.29	81.40
		LAMOL	**88.91**	80.59
		CTR	89.47	83.62
		CTR*	88.21	81.19
		Ours	88.50	**82.35**

Table 3. Ablation experimental results of our model.

Model	Acc.(%)	MF1(%)
Ours	**88.50**	**82.35**
without CL-plugin	85.83	79.75
without CSC+CED	86.48	79.32
without CSC	87.52	79.93
without CED	87.98	80.10

5. Conclusion

This paper investigates the task-incremental learning (TIL) paradigm for the continual learning of a sequence of ASC tasks. Our approach leverages a CL-plugin to capitalize on the acquired knowledge from a pre-trained language model, i.e. BERT, mitigating catastrophic forgetting in continual learning. Additionally, we propose the utilization of two contrastive learning-based losses, namely contrastive ensemble distillation (CED) and contrastive supervised learning of the current task (CSC) losses, to enhance our model's performance. The CED loss improves the knowledge transferability of our continual learning model, while the CSC loss enhances its performance for the current learning task. Our proposed model outperforms all existing continual learning models in the task-incremental learning setting for continual aspect sentiment classification. For future research directions, exploring the application of memory-based methods to address catastrophic forgetting could further enhance the effectiveness of our model.

References

[1] Z. Chen, B. Liu, Lifelong machine learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, Vol. 12, No. 3, 2018, pp. 1–207.

[2] G. M. van de Ven, A. S. Tolias, Three scenarios for continual learning, CoRR, Vol. abs/1904.07734, (2019). arXiv:1904.07734.
URL http://arxiv.org/abs/1904.07734

[3] S. Rebuffi, A. Kolesnikov, C. H. Lampert, icarl: Incremental classifier and representation learning, CoRR, Vol. abs/1611.07725, (2016). arXiv:1611.07725.
URL http://arxiv.org/abs/1611.07725

[4] H. Xu, B. Liu, L. Shu, P. S. Yu, BERT post-training for review reading comprehension and aspect-based sentiment analysis, CoRR, Vol. abs/1904.02232, (2019). arXiv:1904.02232.
URL http://arxiv.org/abs/1904.02232

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR, Vol. abs/1810.04805, (2018). arXiv:1810.04805.
URL http://arxiv.org/abs/1810.04805

[6] Z. Ke, B. Liu, X. Huang, Continual learning of a mixed sequence of similar and dissimilar tasks, CoRR, Vol. abs/2112.10017, (2021). arXiv:2112.10017.
URL https://arxiv.org/abs/2112.10017

[7] Y. Huang, Y. Zhang, J. Chen, X. Wang, D. Yang, Continual learning for text classification with information disentanglement based regularization, CoRR, Vol. abs/2104.05489, (2021). arXiv:2104.05489.
URL https://arxiv.org/abs/2104.05489

[8] M. Biesialska, K. Biesialska, M. R. Costa-jussà, Continual lifelong learning in natural language processing: A survey, CoRR, Vol. abs/2012.09823, (2020). arXiv:2012.09823.
URL https://arxiv.org/abs/2012.09823

[9] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, CoRR, Vol. abs/1902.00751, (2019). arXiv:1902.00751.
URL http://arxiv.org/abs/1902.00751

[10] Z. Ke, H. Xu, B. Liu, Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks, CoRR, Vol. abs/2112.03271, (2021). arXiv:2112.03271.
URL https://arxiv.org/abs/2112.03271

[11] Z. Ke, B. Liu, N. Ma, H. Xu, L. Shu, Achieving forgetting prevention and knowledge transfer in continual learning, CoRR, Vol. abs/2112.02706, (2021). arXiv:2112.02706.
URL https://arxiv.org/abs/2112.02706

[12] S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, CoRR, Vol. abs/1710.09829, (2017). arXiv:1710.09829.
URL http://arxiv.org/abs/1710.09829

[13] G. E. Hinton, A. Krizhevsky, S. D. Wang, Transforming auto-encoders, in: International conference on artificial neural networks, Springer, 2011, pp. 44–51.

[14] Z. Ke, B. Liu, H. Xu, L. Shu, CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6871–6883. doi:10.18653/v1/2021.emnlp-main.550.

[15] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, CoRR, Vol. abs/1612.00796, (2016). arXiv:1612.00796.
URL http://arxiv.org/abs/1612.00796

[16] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, CoRR, Vol. abs/1606.04671, (2016). arXiv:1606.04671.
URL http://arxiv.org/abs/1606.04671

[17] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, CoRR, Vol. abs/2110.11334, (2021). arXiv:2110.11334.
URL https://arxiv.org/abs/2110.11334

[18] S. Hou, X. Pan, C. C. Loy, Z. Wang, D. Lin, Learning a unified classifier incrementally via rebalancing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 168–177. doi:10.1145/1014052.1014073.
URL https://doi.org/10.1145/1014052.1014073

[20] T. Liu, L. Ungar, J. Sedoc, Continual learning for sentence representations using conceptors, ArXiv, Vol. abs/1904.09187, (2019).

[21] X. Ding, B. Liu, P. S. Yu, A holistic lexicon-based approach to opinion mining, in: Web Search and Data Mining, 2008.

[22] Q. Jiang, L. Chen, R. Xu, X. Ao, M. Yang, A challenge dataset and effective models for aspect-based sentiment analysis, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 6280–6285.

[23] Z. Ke, B. Liu, H. Wang, L. Shu, Continual learning with knowledge transfer for sentiment classification, CoRR, Vol. abs/2112.10021, (2021). arXiv:2112.10021.
URL https://arxiv.org/abs/2112.10021

[24] G. Lv, S. Wang, B. Liu, E. Chen, K. Zhang, Sentiment Classification by Leveraging the Shared Knowledge from a Sequence of Domains, 2019, pp. 795–811. doi:10.1007/978-3-030-18576-3-47.

[25] J. Serrà, D. Surís, M. Miron, A. Karatzoglou, Overcoming catastrophic forgetting with hard attention to the task, CoRR, Vol. abs/1801.01423, (2018). arXiv:1801.01423.
URL http://arxiv.org/abs/1801.01423

[26] H. Ahn, D. Lee, S. Cha, T. Moon, Uncertainty-based continual learning with adaptive regularization, CoRR, Vol. abs/1905.11614, (2019). arXiv:1905.11614.
URL http://arxiv.org/abs/1905.11614

[27] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2514–2523. doi:10.18653/v1/P18-1234.
URL https://aclanthology.org/P18-1234

[28] A. Chaudhry, M. Ranzato, M. Rohrbach, M. Elhoseiny, Efficient lifelong learning with A-GEM, CoRR, Vol. abs/1812.00420, (2018). arXiv:1812.00420.
URL http://arxiv.org/abs/1812.00420

[29] P. Buzzega, M. Boschini, A. Porrello, D. Abati, S. Calderara, Dark experience for general continual learning: a strong, simple baseline, ArXiv, Vol. abs/2004.07211, (2020).

[30] F.-K. Sun, C.-H. Ho, H. yi Lee, Lamol: Language modeling for lifelong language learning, in: International Conference on Learning Representations, 2019.