In: *Proceedings of the 30th International Conference on Artificial Neural Networks (ICANN 2021)*, Part II, LNCS 12892, pp. 409-420, Bratislava, Slovakia, September 14–17, 2021.

DRILL: Dynamic Representations for Imbalanced Lifelong Learning

Kyra Ahrens (✉), Fares Abawi, and Stefan Wermter

Knowledge Technology, Department of Informatics, University of Hamburg
{kyra.ahrens,fares.abawi,stefan.wermter}@uni-hamburg.de
www.knowledge-technology.info

Abstract. Continual or lifelong learning has been a long-standing challenge in machine learning to date, especially in natural language processing (NLP). Although state-of-the-art language models such as BERT have ushered in a new era in this field due to their outstanding performance in multitask learning scenarios, they suffer from forgetting when being exposed to a continuous stream of non-stationary data. In this paper, we introduce DRILL, a novel lifelong learning architecture for open-domain sequence classification. DRILL leverages a biologically inspired self-organizing neural architecture to selectively gate latent language representations from BERT in a domain-incremental fashion. We demonstrate in our experiments that DRILL outperforms current methods in a realistic scenario of imbalanced classification from a data stream without prior knowledge about task or dataset boundaries. To the best of our knowledge, DRILL is the first of its kind to use a self-organizing neural architecture for open-domain lifelong learning in NLP.

Keywords: Continual Learning · NLP · Imbalanced Learning · Self-organization · BERT.

1 Introduction

Humans possess the ability to continuously acquire, reorganize, integrate, and enrich linguistic concepts throughout their lives. As early as infancy and based on an innate inference capability, the conventional symbols of language are learned within a socio-communicative context. The underlying neuro-cognitive mechanisms involved in human language acquisition are still far from being fully understood. However, they offer great potential for computational models that are inspired by the neuroanatomical mechanisms in the mammalian brain, enabling the continual integration of consolidated linguistic knowledge with current experience [27].

With the advent of deep learning and the surge of computational resources and data collection, state-of-the-art transformer-based language models (LM) such as BERT [5] and OpenAI GPT [20] have gradually moved away from the symbolic level and given way to *isolated learning* solutions revealing an outstanding performance on downstream NLP tasks in a multitask set-up [4]. Yet when

being exposed to a sequence of tasks, *catastrophic forgetting* or *catastrophic interference* of previously learned concepts was observed [17]. As re-training on all prior data would be inefficient both in terms of computational cost and memory capacity, this observation motivated the introduction of *continual* or *lifelong language learning* (LLL).

Despite recent advances in LLL, most current methods make overly simplistic assumptions that are in stark contrast to realistic, biologically inspired learning settings. This includes enabling multiple passes over the input data stream instead of single-epoch training, resulting in a surge of computational cost. Such methods further rely on perfectly balanced and annotated data, arranged in a way that the assumption of independent and identically distributed samples holds. As a consequence, they are poorly applicable to few-shot, unsupervised, or self-supervised learning scenarios [2].

Thus, striving for a more biologically grounded model architecture and training set-up, we introduce DRILL, a text classification model applicable to CL settings that involve the presence of a continuous stream of imbalanced data without prior knowledge about task boundaries or probability distributions. DRILL is a hybrid architectural and rehearsal-based CL method that uses meta-learning and a self-organizing neural architecture to enable rapid adaptation to novel data while minimizing catastrophic forgetting.

Due to the lack of a continual text classification benchmark of imbalanced data, we introduce two sampling strategies to induce class imbalance artificially. These strategies are evaluated on five text classification datasets presented by Zhang et al. [30], commonly used as CL benchmarks in NLP. With this setting, we show in our experiments that our model outperforms current baselines while better generalizing to unseen data.

2 Related Work

2.1 Continual Learning

Striving for a balance between memory consolidation and generalization to new input data from non-stationary distributions, also referred to as the *stability-plasticity dilemma* [7], paved the way for various LLL approaches in recent years. These approaches can be fully or partially categorized into regularization, rehearsal, and dynamic architectures:

Regularization-based approaches constrain the plasticity of a learning model either by introducing additional loss terms for weight adaptation at a fixed model capacity [13,29], or by setting an additional constraint on prior tasks' predictions to be kept invariant using *knowledge distillation* [15]. This fixed-capacity paradigm contrasts with *architecture-based* CL models that assign some model capacity to each task and therefore dynamically expand in response to novel input [18,22]. Inspired by the concept of memory consolidation, *rehearsal-based* (or *memory replay*) approaches maintain performance on prior tasks by storing and retraining the model on old training samples from an episodic memory [3,16,21].

To limit the associated memory overhead with an increasing number of tasks, *pseudo-rehearsal* approaches employing generative network architectures have been proposed. Such models rely on experience replay of task-representative samples or latent representations based on statistical properties learned from old training data [11,19]. Two such generative replay approaches based on GPT-2 [20], i.e. Language Modelling for Lifelong Language Learning (LAMOL) [24] and Distill and Replay (DnR) [25] view LLL through the lens of question answering. DnR deviates from LAMOL in that it bounds model complexity through knowledge distillation following a teacher-student strategy. Although both approaches are the current performance leaders on datasets benchmarked in this work, they require multiple epochs of training and explicit knowledge about task boundaries. Such preconditions deviate from the realistic CL scenario we advocate for in this work.

2.2 Meta-Learning

Meta-learning [26] has become increasingly popular in recent years as it paved the way for sophisticated algorithms capable of quickly adapting to new data. Online aware Meta-Learning (OML) [10] combines the common meta-learning objective of maximizing fast adaptation to new tasks with the CL objective of minimizing catastrophic interference during training. A Neuromodulated Meta-Learning Algorithm (ANML) [1] extends OML by an independent representation learning stream to selectively gate latent activations. Holla et al. [8] introduce a sparse experience replay mechanism to OML and ANML, denoting their two novel methods by OML-ER and ANML-ER respectively. Both extensions outperform state-of-the-art methods for text classification and question answering benchmarks under a training set-up in which data becomes only available over time and a lack of information about when a dataset or task boundary is crossed.

2.3 Growing Memory and Self-Organization

In an attempt to mimic the explicit memory formation in the mammalian brain, early artificial neural networks based on competitive learning mechanisms and self-organization have been developed and refined [6,14]. One more recent extension to such topology learning methods is the Self-organizing Incremental Neural Network (SOINN) algorithm, which regulates plasticity in unsupervised learning tasks by means of dynamically creating, adapting, and deleting neurons [23]. SOINN+ [28] extends the original SOINN algorithm by introducing a novel node deletion mechanism based on (i) idle time, (ii) trustworthiness, and (iii) non-usage of a network unit. Given that SOINN+ successfully demonstrates its resilience to noisy data and its ability to learn a high-quality topology from the input domain while keeping the number of nodes small, we utilize it as a semantic memory component in our DRILL architecture.

3 Methods

With the challenge of achieving LLL from unbalanced data in mind, we lay the theoretical foundation for our proposed DRILL method.

3.1 Task Formulation

Consider an ordered sequence of tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$, where we observe n_k annotated input samples from the k-th task, i.e. $T_k = \{(\boldsymbol{x}_k^i, y_k^i)\}_{i=1}^{n_k}$ drawn from the distribution $P_k(\mathcal{X}, \mathcal{Y})$. Assuming a realistic scenario of missing task and dataset descriptors, we have no knowledge about which task each input sample belongs to. Following prior work [8], we define task in terms of text classification domain, i.e. sentiment, news topic, question-and-answer, and ontology. Our objective is to learn a model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ with parameters θ to minimize the negative log-likelihood averaged across all N tasks

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{k=1}^{N} \ln P(\boldsymbol{x}_k \,|\, y_k \,;\, \theta) \tag{1}$$

3.2 Progressive Imbalancing

Prior work on lifelong text classification [8,16,24,25] has traditionally deployed the perfectly balanced version of the five NLP datasets by Zhang et al. [30]. Following the idea of d'Autume et al. [16], we introduce two sampling techniques called *progressive reduction* (R) and *progressive expansion* (E), which exponentially increase or decrease the number of samples for each incoming task, such that

$$n_{k+1}^R \leftarrow \left\lfloor \frac{n_k^R}{2} \right\rfloor \tag{2}$$

with progressive reduction and

$$n_{k+1}^E \leftarrow 2 \cdot n_k^E \tag{3}$$

with progressive expansion respectively, and $k \in \{1, \ldots, N\}$. Both sampling techniques allow us to simulate two opposite LLL settings in which data at an early or late stage are significantly less present.

3.3 Episode Generation

For the construction of training episodes and experience rehearsal from episodic memory, we follow a commonly adopted set-up [8,16]:

Under the assumption that samples arrive in batches of size s and are written into episodic memory module $\mathcal{M}_\mathcal{E}$ with probability $p_\mathcal{E}$, we construct the i-th episode from b batches, where the first $b-1$ batches denote support set $\mathcal{S}_i$ and the b-th batch denotes query set $\mathcal{Q}_i$.

After having observed R_I samples from the stream, $\lfloor r \cdot R_I \rfloor$ samples from $\mathcal{M}_\mathcal{E}$ are randomly being drawn for rehearsal, where $r \in [0, 1]$ denotes the predefined replay ratio.

Aligned with the episodic fashion of meta-learning, we calculate the replay frequency

$$R_F = \left\lceil \frac{R_I/s + 1}{b} \right\rceil \tag{4}$$

Thus, every R_F-th episode can be considered as replay episode in a way that its query set does not consist of data from the stream, but from the episodic memory module $\mathcal{M}_\mathcal{E}$.

3.4 DRILL

The DRILL architecture comprises four main elements, namely a dual-memory system of (1) an episodic memory module $\mathcal{M}_\mathcal{E}$ and (2) a semantic memory module $\mathcal{M}_\mathcal{S}$, and, following the original OML algorithm [10], (3) a representation learning network (RLN) h_ϕ as well as (4) a prediction learning network (PLN) $g_{\boldsymbol{W}}$.

We use the SOINN+ [28] algorithm as semantic memory module $\mathcal{M}_\mathcal{S}$. Each neural unit is a d-dimensional real-valued vector. The network is parameterized by a pull factor η denoting the influence of a new observation on neighboring nodes. For the sake of simplicity and as proposed by Wiwatcharakoses and Berrar [28], we set the pull factor to a constant value $\eta = 50$. Thus, our model f_θ optimizes for the set of parameters $\theta = \phi \cup \boldsymbol{W}$, consisting of parametrization ϕ from $\mathcal{X} \to \mathbb{R}^d$ of the RLN h_ϕ and $\boldsymbol{W}$ from $\mathbb{R}^d \to \mathcal{Y}$ of the PLN $g_{\boldsymbol{W}}$ respectively.

Taking inspiration from the mammalian thalamus as a 'gate to consciousness', we propose two DRILL variants that differ in how latent representations from the RLN are integrated with signals from the semantic memory $\mathcal{M}_\mathcal{S}$, translating its internal selective plasticity to the entire learning process.

The first variant, called *Integration by Multiplication* (DRILL_M), can be described as follows: On receiving input $\boldsymbol{x}$, the model gates the activations $h_\phi(\boldsymbol{x})$

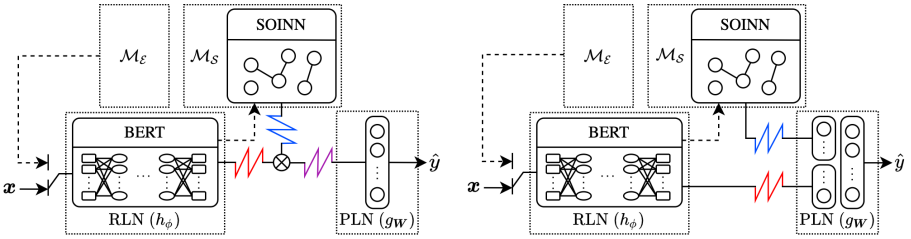

Fig. 1. Overview of the two variants DRILL_M (left) and DRILL_C (right). Latent representation signals retrieved from RLN are integrated with neural weight signals from $\mathcal{M}_\mathcal{S}$ either by multiplication (DRILL_M) or concatenation (DRILL_C). Input to the model is either an new observation $\boldsymbol{x}$ from the stream or an episodic replay sample from $\mathcal{M}_\mathcal{E}$.

arriving from the RLN by multiplying them element-wise with a set of neural weights $\boldsymbol{w}_{\mathcal{S}}$ drawn from $\mathcal{M}_{\mathcal{S}}$ in a procedure described in subsection 3.5. We express the model variant DRILL$_{\mathrm{M}}$ as

$$f_{\theta}^{M}(\boldsymbol{x}) = g_{\boldsymbol{W}}(\boldsymbol{w}_{\mathcal{S}} \cdot h_{\phi}(\boldsymbol{x})) \tag{5}$$

For the second variant called *Integration by Concatenation* (DRILL$_{\mathrm{C}}$), each of the d-dimensional signals $\boldsymbol{w}_{\mathcal{S}}$ and $h_{\phi}(\boldsymbol{x})$ retrieved from $\mathcal{M}_{\mathcal{S}}$ and the RLN respectively are reduced to half of their dimension $\frac{d}{2}$ and subsequently concatenated in a d-dimensional linear layer that is allocated to the PLN, as shown in Figure 1. Thus, we derive the following model

$$f_{\theta}^{C}(\boldsymbol{x}) = g_{\boldsymbol{W}}(\,[\boldsymbol{w}_{\mathcal{S}}, h_{\phi}(\boldsymbol{x})]\,) \tag{6}$$

where $[\cdot,\cdot]$ denotes the concatenation operator. The meta-learning procedure for both DRILL variants works as follows: During inner-loop optimization of the i-th episode, the RLN is kept frozen while the PLN is fine-tuned using SGD with an inner-loop learning rate α, such that

$$\boldsymbol{W}' \leftarrow \mathrm{SGD}(\mathcal{L}_i(\phi, \boldsymbol{W}), \mathcal{S}_i, \alpha) \tag{7}$$

Subsequently, both RLN and PLN are fine-tuned on the query set $\mathcal{Q}_i$ during outer-loop optimization, such that all model parameters are updated using the Adam optimizer [12] with an outer-loop learning rate β to give

$$\theta' \leftarrow \mathrm{Adam}(\mathcal{L}_i(\phi, \boldsymbol{W}'), \mathcal{Q}_i, \beta) \tag{8}$$

For the RLN, we use the state-of-the-art transformer-based language model BERT$_{\mathrm{BASE}}$ [5] with 12 transformer layers and $d = 768$ hidden dimensions. With DRILL$_{\mathrm{M}}$, the PLN is a single linear layer with softmax activation that outputs the class probabilities, while a linear concatenation layer additionally precedes this layer with DRILL$_{\mathrm{C}}$.

3.5 Self-Supervised Sampling

In contrast to the episodic memory $\mathcal{M}_{\mathcal{E}}$ that we solely use for experience replay, we use the semantic memory $\mathcal{M}_{\mathcal{S}}$ for generating high-quality representations, which influence the fine-tuning of the PLN. For every input sample $(\boldsymbol{x}_i, y_i)$, we initiate a competitive voting mechanism among all nodes in $\mathcal{M}_{\mathcal{S}}$ to determine the two neurons with neural weights $\boldsymbol{w}_{\mathcal{S}}^1$ and $\boldsymbol{w}_{\mathcal{S}}^2$ that have most frequently been best-matching units (BMUs) for class y_i. According to the original SOINN+ algorithm [28], the network node that lies closest to the input in Euclidean space is denoted as BMU.

The two winners are then either multiplied element-wise (DRILL$_{\mathrm{M}}$) or concatenated (DRILL$_{\mathrm{C}}$) with the activations of the latent representation $h_{\phi}(\boldsymbol{x}_i)$ coming from the RLN, thus generating two inputs to the PLN from one output of the RLN. During the evaluation phase, only one signal from the winning node $\boldsymbol{w}_{\mathcal{S}}$ is retrieved from $\mathcal{M}_{\mathcal{S}}$ for the purpose of unambiguous label prediction by the PLN.

4 Experiments

4.1 Benchmark Datasets

We train our model sequentially on five text classification datasets by Zhang et al. [30] covering four different tasks: Sentiment analysis, news topic detection, question-and-answer classification, and ontology categorization. We summarize them in Table 1. Following d'Autume et al. [16], the datasets are arranged in four randomized permutations reflecting the significant impact of task ordering on evaluation results.

Table 1. The five balanced text classification datasets as in Zhang et al. [30], each containing 7,600 test samples randomly drawn from the original datasets. The number of training samples differs depending on order position and imbalanced sampling strategy.

Classification Domain	Dataset	Classes	Order Position			
			I	II	III	IV
Sentiment	Amazon	5	4	4	3	3
	Yelp	(merged)	1	5	1	2
News Topic	AGNews	4	2	3	5	1
Question Topic	Yahoo	10	5	2	2	4
Ontology	DBPedia	14	3	1	4	5
	Total:	**33**				

For evaluation, we follow prior work [8,16,24,25] and randomly draw 7,600 samples from each of the five datasets, yielding a total test size of 38,000. However, we depart from the perfectly balanced and thus poorly realistic scenario of 115,000 training samples per dataset and instead apply progressive imbalancing as described in subsection 3.2 with $n_0^R = 115,000$ and $n_0^E = 7,187$, thus providing a total training size of $222,812$ for either sampling strategy.

4.2 Baselines

For performance evaluation, we compare our two proposed model variations $\textbf{DRILL}_{\textbf{M}}$ and $\textbf{DRILL}_{\textbf{C}}$ with the two performance leaders given a realistic single-epoch set-up without prior task-specific knowledge, i.e. **ANML-ER** and **OML-ER** [8]. Just like our method, they use a pretrained $\text{BERT}_{\text{BASE}}$ language encoder. We further implement the lower bound for CL model performance, **SEQ**, in which we fine-tune both RLN and PLN on all tasks sequentially without any rehearsal. We also compare our methods with **REPLAY**, an extension of SEQ towards experience rehearsal with samples stored in an episodic memory. Finally, we train RLN and PLN jointly in a multitask set-up **MTL**, which we

consider as an upper bound for CL model performance. For a fair comparison, we choose the same memory-write and rehearsal policies for REPLAY, ANML-ER, and OML-ER, as well as our two proposed DRILL variants.

4.3 Implementation Details

Our experimental set-up consists of three independent runs on seeds 42-44, each run performed on the four order permutations and two sampling strategies respectively. Accordingly, the comparison results are averaged over all three runs.

Due to computational limitations, we train all baseline models on normalized batches of size $s = 8$ following the procedure of Ioffe and Szegedy [9] and optimize based on the cross-entropy loss on all 33 classes. We truncate the $\text{BERT}_{\text{BASE}}$ input sequences to length 448 and set the buffer size $b = 6$. The inner-loop and outer-loop learning rates of the four meta-learning-based models DRILL_{M}, DRILL_{C}, OML-ER, and ANML-ER are set to $\alpha = 8\text{e}{-3}$ and $\beta = 1.5\text{e}{-5}$ respectively. The learning rate of all remaining baselines SEQ, REPLAY, and MTL is set to $1\text{e}{-5}$.

All models are trained for a single epoch, whereas MTL is trained for two epochs. The probability of storing an observation in the episodic memory module $\mathcal{M}_{\mathcal{E}}$ is governed by the maximum write probability $p_{\mathcal{E}} = 0.8$. The $p_{\mathcal{E}}$ is inversely proportional to the expansion or reduction for all rehearsal-based models, restoring class balance within $\mathcal{M}_{\mathcal{E}}$. The learning rates and $p_{\mathcal{E}}$ are derived using a Parzen–Rosenblatt estimator [1]. The hyperparameter optimization is applied to OML-ER as the representative model for all meta-learning-based approaches and SEQ for inferring the learning rate of the remaining models. Both OML and SEQ are trained on the full dataset (without expansion or reduction) with order I and random seed 42. With both DRILL architectures, the unsupervised SOINN+ algorithm is performed as described in the original paper [28], including setting the pull factor $\eta = 50$.

We follow the rehearsal and evaluation strategies adopted by Holla et al. [8], setting $R_I = 9,600$ and $r = 1\%$, such that we draw 96 samples from $\mathcal{M}_{\mathcal{E}}$ after observing 9,600 samples from the data stream. The evaluation of the four meta-learning models is performed by generating five episodes, each containing the test datasets as query sets. All baseline models were trained on an NVIDIA TITAN RTX with 24GB VRAM and 64GB RAM. The training time ranges between 1 and 7 hours, depending on the model and number of observations.

5 Results

5.1 Imbalanced Lifelong Text Classification

Unlike prior work, we report F_1 scores rather than macro-averaged classification accuracy due to the unbalanced nature of the training data. Our main results are summarized in Table 2.

[1] CometML Hyperparameter Optimizer: https://www.comet.ml/

Table 2. Text classification F_1 scores on four permutations of task orders and progressive expansion (E) and progressive reduction (R) sampling respectively. The two rightmost columns denote the macro-average and standard deviation across all orderings and sampling strategies.

	Order (E)				**Order (R)**					
Method	**I**	**II**	**III**	**IV**	**I**	**II**	**III**	**IV**	μ	σ
SEQ	17.4	27.6	26.6	21.0	23.7	32.7	28.8	25.0	25.4	4.9
REPLAY	55.3	67.9	58.6	**65.7**	44.2	57.7	53.5	37.0	55.0	10.5
ANML-ER	66.7	**70.5**	55.0	62.9	57.0	58.6	62.7	45.2	59.8	8.8
OML-ER	**70.2**	64.9	52.2	64.4	56.0	**62.0**	**66.5**	48.7	60.6	8.1
DRILL$_{\text{M}}$	23.2	36.5	37.1	37.6	58.0	51.7	41.0	**50.4**	41.9	13.7
DRILL$_{\text{C}}$	68.4	68.1	**59.1**	65.5	**61.8**	61.6	62.9	49.5	**62.1**	6.2
MTL	77.9	78.7	76.2	76.7	77.7	76.4	78.3	78.2	77.5	1.0

Our DRILL$_{\text{C}}$ variant outperforms existing methods in terms of higher overall average performance and higher median under equal conditions (the latter is depicted in Figure 2). In addition, it has a significantly smaller variance than all other replay-based comparison methods, thus demonstrating its robustness to the order of training data and the imbalancing strategy. Consequently, it narrows the gap to the upper bound of multitask learning.

Interestingly, the DRILL$_{\text{M}}$ method is trailing the current models with respect to absolute performance. Yet, it provides a smaller variance for progressively expanded data than all other baselines except SEQ, exhibiting robustness against undersampled classes at the beginning of training. The enormous performance difference of our two DRILL variants motivates a more detailed analysis of the impact of knowledge integration mechanisms from RLN and $\mathcal{M}_{\mathcal{S}}$.

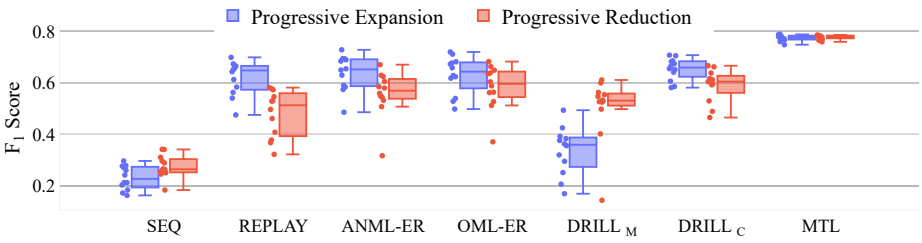

Fig. 2. F_1 scores of all comparison models aggregated across three seeds and four orderings. Sequential (SEQ) and multitask (MTL) learning can be viewed as lower and upper bound for model performance respectively.

5.2 Knowledge Integration Mechanisms

Although the introduction of class-representative signals drawn from semantic memory yields greater robustness under a realistic training scenario, the overall model performance varies greatly depending on how the latent signals retrieved are integrated during training. The relatively poor performance of DRILL_{M} could be attributed to the multiplicative gating mechanism that we adopted from the original ANML algorithm [1]. The ANML is designed so that 'gating parameters' of preceding layers are learned in a supervised fashion, which is in contrast to the unsupervised nature of SOINN+.

Conversely, with DRILL_{C}, signals from RLN are enriched with those from the SOINN rather than fused, allowing for better linear separation, thus resulting in an increase of model performance. From this, we conclude that the concatenation of modalities in our training scenario provides a better knowledge retention strategy.

5.3 Self-Organized Networks in NLP

A generally known problem of self-organizing networks is that they capture the entire evolution of hidden representations in feature space along with obsolete knowledge and are therefore unsuitable for training on shifting latent distributions. With the DRILL architecture, we overcome this problem by freezing the RLN parameters during inner-loop optimization and by the choice of our retrieval strategy for neural weight signals coming from the SOINN. The former leads to a more stable latent data distribution over a longer period. The latter ensures that neural units residing in the current input distribution are more likely to be considered as high-quality class representatives.

As this is the first work to combine a self-organizing neural architecture with a transformer-based language model in a CL setting, we advocate further exploring such set-ups in future work. This is due to the intrinsic ability of the SOINN and its various extensions to be applicable in an infinite learning setting with an unlimited number of tasks. The model can additionally handle partially annotated data, setting the basis for semi-supervised LLL scenarios.

6 Conclusion and Future Work

In this work, we introduce a novel, more challenging continual learning set-up with imbalanced data. We further propose Dynamic Representations for Imbalanced Lifelong Learning (DRILL), a neuroanatomically inspired CL method which combines a state-of-the-art language model with a self-organizing neural architecture. It outperforms current baselines, yet is more stable against data ordering and imbalancing. Thus, the fusion of supervised language models with unsupervised clustering algorithms has proven effective for lifelong learning methods, further narrowing the gap to multitask learning approaches. DRILL achieves the best results on imbalanced data, with the least overall variance in

comparison to other meta-learning-based lifelong learning approaches. For future work, we plan to extend our model towards infinite learning of an unknown number of tasks as well as sequence-to-sequence learning.

Acknowledgements

We would like to thank Dr. Cornelius Weber (University of Hamburg) and Katja Kösters (University of Hamburg) for their feedback and suggestions. The authors gratefully acknowledge partial support from the German Research Foundation (DFG) under Project CML (TRR-169).

References

1. Beaulieu, S., Frati, L., Miconi, T., Lehman, J., Stanley, K.O., Clune, J., Cheney, N.: Learning to Continually Learn. In: 24th European Conference on Artificial Intelligence. vol. 325, pp. 992–1001. IOS Press (2020)
2. Biesialska, M., Biesialska, K., Costa-jussà, M.R.: Continual Lifelong Learning in Natural Language Processing: A Survey. In: 28th International Conference on Computational Linguistics. pp. 6523–6541. International Committee on Computational Linguistics (2020)
3. Chaudhry, A., Marc'Aurelio, R., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with A-GEM. In: 7th International Conference on Learning Representations (2019)
4. Chen, Z., Liu, B.: Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **12**(3), 1–207 (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. pp. 4171–4186. Association for Computational Linguistics (2019)
6. Fritzke, B., et al.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems **7**, 625–632 (1995)
7. Grossberg, S.: How does a brain build a cognitive code? Studies of mind and brain pp. 1–52 (1982)
8. Holla, N., Mishra, P., Yannakoudakis, H., Shutova, E.: Meta-Learning with Sparse Experience Replay for Lifelong Language Learning. arXiv preprint arXiv:2009.04891 (2020)
9. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: 32nd International Conference on Machine Learning. vol. 37, pp. 448–456. PMLR (2015)
10. Javed, K., White, M.: Meta-Learning Representations for Continual Learning. Advances in Neural Information Processing Systems **32**, 1820–1830 (2019)
11. Kemker, R., Kanan, C.: FearNet: Brain-inspired model for incremental learning. In: 6th International Conference on Learning Representations (2018)
12. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (2015)
13. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America **114**(13), 3521–3526 (2017)

14. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
15. Li, Z., Hoiem, D.: Learning without Forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
16. de Masson d'Autume, C., Ruder, S., Kong, L., Yogatama, D.: Episodic Memory in Lifelong Language Learning. Advances in Neural Information Processing Systems **32** (2019)
17. McCloskey, M., Cohen, N.J.: Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In: Bower, G.H. (ed.) Psychology of Learning and Motivation, Psychology of Learning and Motivation, vol. 24, pp. 109–165. Academic Press (1989)
18. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
19. Parisi, G.I., Tani, J., Weber, C., Wermter, S.: Lifelong Learning of Spatiotemporal Representations With Dual-Memory Recurrent Self-Organization. Frontiers in neurorobotics **12**, 78 (2018)
20. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. OpenAI (2018)
21. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental Classifier and Representation Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
22. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
23. Shen, F., Hasegawa, O.: Self-Organizing Incremental Neural Network and Its Application. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) International Conference of Artificial Neural Networks. vol. 6354, pp. 535–540. Springer, Berlin, Heidelberg (2010)
24. Sun, F.K., Ho, C.H., Lee, H.Y.: LAMOL: LAnguage MOdeling for Lifelong Language Learning. In: 8th International Conference on Learning Representations (2020)
25. Sun, J., Wang, S., Zhang, J., Zong, C.: Distill and Replay for Continual Language Learning. In: 28th International Conference on Computational Linguistics. pp. 3569–3579. International Committee on Computational Linguistics, Barcelona, Spain (2020)
26. Thrun, S., Pratt, L.: Learning to learn: Introduction and overview. In: Learning to learn, pp. 3–17. Springer (1998)
27. Tomasello, M.: The social bases of language acquisition. Social development **1**(1), 67–87 (1992)
28. Wiwatcharakoses, C., Berrar, D.: SOINN+, a Self-Organizing Incremental Neural Network for Unsupervised Learning from Noisy Data Streams. Expert Systems with Applications **143**, 113069 (2020)
29. Zenke, F., Poole, B., Ganguli, S.: Continual Learning Through Synaptic Intelligence. In: 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3987–3995. PMLR (2017)
30. Zhang, X., Zhao, J., LeCun, Y.: Character-Level Convolutional Networks for Text Classification. In: 28th International Conference on Neural Information Processing Systems. pp. 649–657. MIT Press (2015)