Advancements and Challenges in Continual Learning for Natural Language Processing: Insights and Future Prospects

Asma Kharrat[1][a], Fadoua Drira[1][b], Franck Lebourgeois[2][c] and Bertrand Kerautret[3][d]

[1]ReGIM-Lab, University of Sfax, ENIS, BP1173, 3038, Sfax, Tunisia

[2]LIRIS, University of Lyon, INSA-Lyon, CNRS, UMR5205, F-69621, Lyon, France

[3]LIRIS, University of Lyon, Université Lumière Lyon2, F-69365, Lyon, France

Keywords: Deep Learning, Continual Learning, Natural Language Processing, Catastrophic Forgetting.

Abstract: Deep learning-based Natural Language Processing (NLP) has advanced significantly over the past decades, in light of static learning's remarkable performance across a range of text datasets. However, this method heavily relies on static surroundings and predefined datasets, making it difficult to manage ongoing data streams without losing track of previously acquired knowledge. Continual learning provides a more effective and adaptable framework. It tries to make it possible for machine learning models to learn from an ongoing data stream while maintaining their prior knowledge. In the context of NLP, continual learning presents unique challenges and opportunities due to its dynamic and diversity. In this paper, We shall provide a thorough analysis of CL's most recent advancements in the NLP disciplines in which major challenges are illustrated. We also critically review the existing CL evaluation solutions and benchmarks in NLP. Finally, we present open issues that we consider need further investigations and our outlook on future research directions.

1 INTRODUCTION

Continual Learning is a machine learning subfield that deals with non-iid (identically and independently distributed) data. Its purpose is to gain insight into global optima for an optimization problem with changing data distribution over time. This is common in databases that are routinely updated with fresh data or when data is streamed to algorithms with limited storing capabilities. Continual learning (CL) seeks alternate techniques to iid training in order to avoid retraining with entire data each time new data becomes available. In particular, when learning data samples with different distributions or progressing through a sequence of tasks, a parametric model eventually reaches a point where no more knowledge can be stored. At this point, either the capacity of the model is increased or an override (i.e. forgetting) happens, which will most likely result in a decrease in performance. Models lose accuracy over time due to shifting data distribution. With the introduction of deep learning, the problem of continual learning (CL) in Natural Language Processing (NLP) has become even more serious, as present techniques are unable to efficiently retain previously gained knowledge while also adapting to new information. Numerous approaches have been presented over the years to solve the problem known as catastrophic forgetting (CF) or catastrophic interference. CL algorithms present several memory storage systems for storing information from previous learning experiences, as well as learning algorithms to continue learning with this memory and fresh data.

When it comes to CL and unlike computer vision or robotics, NLP is still in its infancy. It has recently been investigated since it highly contributes and is useful in many applications. Main applications include digital libraries, document processing, Language Modeling, Sentiment Analysis and Text Classification, Question Answering, Word and Sentence Representations, etc. Accordingly, many researchers are now interested in developing applications that embrace change and continually learn from new text images. Therefore, in this study, we conducted an expanded deep overview of recent research from different perspectives with a special emphasis on classical static learning and continual learning within deep ar-

[a] https://orcid.org/0000-0002-3964-2313
[b] https://orcid.org/0000-0001-6706-4218
[c] https://orcid.org/0000-0002-1917-0347
[d] https://orcid.org/0000-0001-8418-2558

Kharrat, A., Drira, F., Lebourgeois, F. and Kerautret, B.
Advancements and Challenges in Continual Learning for Natural Language Processing: Insights and Future Prospects.
DOI: 10.5220/0012462400003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 1255-1262
ISBN: 978-989-758-680-4; ISSN: 2184-433X

chitectures. We survey hundreds of papers on NLP for text recognition and Continual Learning (CL). We categorize these papers into popular topics according to our Research Questions:

- ***RQ1****: What approaches can be developed to handle shifts and changes in NLP models for text recognition in continual learning settings?*

  In settings of ongoing learning, the statistical features of the incoming data alter with time, which presents a problem for handling shifts and changes. This question addresses the investigated strategies like domain adaptation, transfer learning, or dynamic model adaptation to guarantee that text recognition algorithms work robustly and accurately even when data distributions change.

- ***RQ2****: How can natural language processing techniques be leveraged to enhance the accuracy and efficiency of information extraction from unstructured textual data?*

  This question investigates the potential of NLP methods in extracting meaningful information from unstructured text, including handwritten or historical documents. It explores techniques like named entity recognition and domain adaptation to enable more effective and efficient extraction of structured information from unstructured textual data.

- ***RQ3****: What strategies can be employed to handle imbalanced or evolving class distributions in text classification and recognition tasks during continual learning?*

  This question addresses the challenge of handling imbalanced or evolving class distributions that may arise in real-world scenarios during continual learning. It investigates techniques such as class balancing methods, importance weighting, prototype-based approaches, or data augmentation strategies to ensure fair representation and accurate classification for all classes, even as the data distribution evolves over time.

The search results were subsequently filtered by reviewing the abstracts and only papers published from 2016 to 2023 were considered. However, we acknowledge the importance of including papers published before 2016 that are valuable contributions to the field.

With the growth of interest in CL, other terminology and concepts were revealed. Table 1 presents the most common related paradigms.

To the best of our knowledge, only our work investigates research progress related to static learning in NLP and Continual Learning in NLP. We further discuss gaps and challenges in this domain.

2 STATIC LEARNING WITH DEEP ARCHITECTURES FOR NLP: TEXT RECOGNITION

DNN-based architectures could be performed using two different strategies: The first strategy proceeds by the design of the overall model following a training process from scratch. The second strategy is based on transfer learning and is characterized by the use of a pre-trained CNN model already trained on bigger datasets (Raffel et al., 2020) (Houlsby et al., 2019) (Ruder et al., 2019). Often, the advancement of NLP is bound by adequate language modeling (Dogra et al., 2022) (Esmaeilzadeh and Taghva, 2021). The probabilistic representation of word sequences in a language is one of statistical language modeling's objectives, and it is a challenging task because of dimensionality (Mittal et al., 2020). The research reported in (Bengio and Senécal, 2008) was a breakthrough for language modeling with neural networks that intended to overcome the difficulty of dimensionality by training a distributed representation of words and providing a probability function for sequences. The difficulties of establishing an in-depth representation of language using statistical models appear to be a key problem in NL Presearch when compared to other fields such as Computer Vision. The basic aim of NLP applications is to provide textual representations, such as documents. This entails feature learning extracting relevant information from raw data to allow for additional processing and analysis. Traditional approaches begin with the time-consuming handcrafting of features through meticulous human examination of a specific application and are followed by the development of algorithms to extract and use such features. To connect word-level and character-level representations using convolutional neural networks for OStagging, a deep neural network architecture called CharWNN has been developed in (Santos and Zadrozny, 2014). They mainly focus on the applicability of character-level feature extraction because their experimental findings show that, in the absence of character-level features, hand-crafted features must be used to attain cutting-edge performance. LSTM networks, bidirectional LSTM networks, LSTM networks with a CRF8 layer, and other neural network-based models for sequence tagging tasks have all been presented in (Huang et al., 2015). Furthermore, in the NLP domain, Generative adversarial networks (GAN) are used for text generation (Li et al., 2017) (Yu et al., 2016). Deep supervised feature learning approaches are extremely data-driven and can be employed in more general efforts aiming at delivering robust data representation. Because of the large vol-

Table 1: Continual Learning related Paradigms.

Related Concept	**Details**
Meta-learning	The goal is to increase learning efficiency and generalization performance on previously unknown problems by leveraging existing knowledge. While meta-learning isn't directly tackling the issue of catastrophic forgetting (learning without forgetting), it does focus on increasing a learner's ability to adapt and learn effectively over various tasks. Although meta-learning can aid in the development of more efficient and adaptive models for continuous learning scenarios, it is not fundamentally related to the problem of learning without forgetting.
Active learning	Is an iterative kind of supervised learning in which the learner (an algorithm or a model) actively selects a subset of unlabeled instances and seeks labels from a human user or an oracle. The goal is to select examples that are likely to produce the most valuable or informative labels to gain the largest knowledge for the model.
Curriculum Learning	Is a training procedure where a set of tasks are given in a specific order to tackle more challenging problems effectively. The learning is accomplished by starting with simpler or easier tasks and gradually increasing their difficulty. Contrary, continual learning does not require sorting or choosing tasks according to their level of difficulty. The objective is to retain knowledge, adapt to new tasks, and minimize interference with prior knowledge rather than just completing the last or most difficult assignment.
Online learning	Is a particular instance of Continual Learning, where model updates are made on a per-data-point basis with a batch size of one. Without the need to store or batch data, the model is updated gradually as each new data point is received in online learning. The main benefit of online learning is its capacity to adjust to shifting input in real-time, enabling the model to pick up knowledge as quickly as possible and to make decisions or predictions right away. However, because the model is constantly updating and adjusting based on specific data points, online learning presents difficulties in terms of stability, managing concept drift, and preventing catastrophic forgetting.
Multi-task Learning	Is a machine learning technique where a single model is taught to carry out several connected tasks at once. The model can benefit from shared knowledge and dependencies by pooling learning from several tasks, which improves efficiency and allows for generalization across all tasks. It increases the model's ability to learn by enabling it to draw on task-specific knowledge while also identifying common patterns across tasks.
Domain Adaptation	Refers to the process of adopting a model from a source domain to a target domain. It entails reducing domain variances between the source and target domains in order to boost the model's performance on target domain tasks. Domain adaptation approaches in CL are intended to transmit knowledge, retain learned information, and adjust the model to new domains or tasks. It aims to reduce the differences in data distribution or feature space between the two domains to improve the model's performance on the target domain.
Transfer Learning	Is the process of using knowledge from a previously trained model to improve learning and performance on a target domain. The objective is to improve generalization and learning efficiency in the target domain by utilizing prior knowledge. Transfer learning involves using knowledge from a source domain to improve learning on a target domain, while domain adaptation specifically addresses the domain shift between a source and target domain to enhance the model's performance on the target domain.

umes of unlabeled data, unsupervised feature learning is seen as a critical problem in NLP (Fatima et al., 2017) (Le et al., 2019). Several approaches include K-means clustering and principal component analysis to this goal, analysis has been presented and effectively executed. In (Al-Anzi and AbuZeina, 2016), the authors propose using Latent Semantic Indexing (LSI), a method singular value decomposing (SVD) and clustering techniques, for Arabic text categorization, to group similar unlabeled documents into a pre-specified number of topics. The generated groups are then categorized using a suitable label With the advent of deep learning and a plethora of unlabeled data, unsupervised feature learning has become a critical problem for representation learning, a prerequisite in NLP applications.

3 CONTINUAL LEARNING WITH DEEP ARCHITECTURES FOR NLP

Deep learning has made the issue of continual learning (CL) in natural language processing (NLP) even more urgent because existing methods do not properly preserve previously learned information while simultaneously adapting to new information.

3.1 Continual Learning: Challenges, Strategies and Related Paradigms

Continual learning system is also known as incremental and lifelong learning system. It is an adaptive algorithm that can gradually learn the new flow of data while maintaining the knowledge acquired from previously seen tasks even when the number of tasks to be learned is not predetermined (Parisi et al., 2018). For data scientists, CL will increase models' accuracy, performance, and save retraining time by making models auto-adaptive. Unlike static learning algorithms, CL systems can build up knowledge over a range of tasks without the need to retrain from scratch. However, a quick update and fitting of new information equally cause a major issue known as 'Catastrophic Forgotten' (CF). CF happens precisely when the new instances to be learned diverge tremendously from previously seen examples and overwrite old knowledge. When the network is trained sequentially, the weights that are essential for antecedent tasks are shifted to meet the objectives of the new ones (McCloskey and Cohen, 1989). To overcome CF, a vast range of methods and techniques have been suggested (Farquhar and Gal, 2019) (Delange et al., 2021) (Lee et al., 2018) (van de Ven and Tolias, 2019). These methods can be categorized into three main classes (Kirkpatrick et al., 2017) based on how task-specific knowledge is retained and used during the learning process (Mai et al., 2022) (Belouadah et al., 2020):

(1) Regularization-based methods: Attempt to achieve an equilibrium between conserving previously learned representations and allowing enough flexibility to encode new information. It alleviates CF by setting limits on neural weight updates. This is accomplished by either including additional penalty terms in the loss function or changing the gradient of parameters during optimization.

(2) Replay-based methods: Also known as memory-based or rehearsal-based methods. As it implies, these methods aim to replay data from already learned tasks. The training process typically involves inserting samples of former tasks along with the actual data for the present task. It either stores samples in raw format and tends to replay them or generates pseudo-data that mimics past data 'Generative Replay'. A crucial aspect of these techniques is to select a suitable subset of data, known as exemplars, that approximates the entire observed data distribution. Such exemplars used as training data for previous tasks are often referred to as pseudo-data.

(3) Parameter isolation methods: To avoid forgetting anything, this family assigns different model parameters to each task. One can generate new branches for new activities while freezing the parameters of older tasks when there are no restrictions on architectural size. These types of methods can be subdivided into Fixed Architecture (FA) which only activates relevant parameters for each task without modifying the architecture and Dynamic Architecture (DA) which adds new parameters for new tasks while keeping old parameters unchanged.

Memory-based methods are robust to distribution shifts, meaning that they are not sensitive to changes in task distribution over time, unlike regularization methods. They also preserve old task information by storing a subset of past examples and using them to prevent forgetting during training on new tasks. Regularization methods, on the other hand, encourage the model to retain old knowledge by adding a penalty term to the loss function, but they do not explicitly store past information. Besides, regularization methods can sometimes result in overfitting to the current task, while replay methods can prevent overfitting by providing a diverse set of examples from past tasks. Architectural-based methods tend to provide better results than both replay and regularization methods. Architectural-based methods can provide a structural bias towards the desired behavior in the continual learning scenario, which can be difficult to achieve with replay and regularization methods. They can scale to handle an increasing number of tasks over time, whereas replay and regularization methods may become less effective as the number of tasks or classes increases. Moreover, they allow the model to independently process information from different tasks, unlike replay and regularization methods (Kharrat et al., 2023).

It is worth noting that all of these methods do not contradict each other. As each strategy has some limitations, researchers tend to combine various strategies to achieve continual learning. Therefore, we refer the readers to explore mainly these references among many others (Mundt et al., 2023) (Qu et al., 2021) (Parisi et al., 2018).

3.2 Exploring Continual Learning in NLP

In the realm of Natural Language Processing (NLP), a diverse array of tasks is encompassed. This section delves into how Continual Learning (CL) approaches are applied to address the most prevalent NLP tasks. The training process for numerous neural-based NLP systems involves two primary steps: initially utilizing a large unlabeled text dataset to train a Neural Network (NN)-based language model, and subsequently employing the pre-trained language model in supervised downstream tasks. Notably, Language Model-based approaches for CL in NLP have gained substantial research interest recently. A huge Language Model (LM) trained on diverse corpora can, in theory, perform perfectly across numerous datasets and domains (Radford et al., 2019). In a study by (Ke and Liu, 2023), the focus is directed toward knowledge transfer, which holds particular importance in NLP. This is attributed to the fact that words and phrases used in texts from various activities or domains typically convey similar meanings or shapes and many NLP tasks share a common knowledge base. Additionally, major advances in CL are oriented toward alleviating Catastrophic Forgetting (CF). In addition, (Gururangan et al., 2020) investigate the impact of task- and domain-adaptation on the transfer ability of pre-trained LMs across domains and tasks. In the study, the authors concluded that ongoing domain- and task-adaptive pretraining improves downstream NLP performance. To better comprehend the internal behavior of Pre-trained Language Models (PLMs) in extracting knowledge, the researchers in (Wang et al., 2022) first establish knowledge-bearing (K-B) and knowledge-free (K-F) tokens for unstructured text, engaging expert annotators to label some samples manually. As a result, PLMs are more likely to provide inaccurate predictions on K-B tokens, owing to diminished attention inside the self-attention module. Based on these findings, two methods are proposed to assist the model in learning more knowledge from unstructured text in a completely self-supervised manner, demonstrating effectiveness in knowledge-intensive tasks. The usefulness of the proposed strategies is demonstrated by experimental results on knowledge-intensive tasks. In (Escolano et al., 2020), the authors presented a language-specific encoder/decoder architecture, in which languages are embodied into a shared space and either the encoder or the decoder is frozen when training on a new language, to eliminate and minimize the necessity for retraining the entire system. Class imbalance is a pertinent aspect in NLP, although limited surveys address it compared to the computer vision field (Johnson and Khoshgoftaar, 2019). Taking into consideration strategies to correct class imbalance, can result in up to 20% improvement in performance. However, NLP Research frequently fails to highlight how crucial this is in real-world scenarios where minority classes may be of particular relevance. (Jang et al., 2021) describe imbalanced classification as a k-stage continuous learning problem, progressively achieving a more balanced dataset (sequential targeting). The stage that exhibits the greatest degree of imbalance is considered as the first stage, and the stage that exhibits the least degree of imbalance is the final one. Good performance on the present level and retention of knowledge from prior stages are both encouraged by the training aim. They have conducted ternary and binary text categorization studies in Korean and English. To also deal with imbalanced data the authors in (Kim et al., 2020) come up with Partitioning Reservoir Sampling (PRS), a novel sampling technique for the replay-based method that enables the model to retain an even understanding of the head and tail classes. They jointly address the two independently solved problems, Catastropic Forgetting and the long-tailed label distribution by first empirically showing a new challenge of destructive forgetting of the minority concepts on the tail. Then, they curate two benchmark datasets, COCOseq and NUS-WIDEseq, that allow the study of both intra- and inter-task imbalances. Research has demonstrated that Active Learning (AL), by definition comprising many phases, can enhance the performance of BERT models, particularly for minority groups (Buda et al., 2018). While domain adaptation approaches are extensively employed in the context of Neural Machine text recognition and translation, there have been additional attempts to adapt and joint learning on several languages. Multilingual might be viewed as a multi-task learning issue (Dong et al., 2015). The goal of the multilingual translation model is to employ a single model to translate across several different languages. Such systems are useful not only because they are capable of handling several translation directions with a single model as joint training with high-resource languages enhances performance on low- and zero-resource languages (Arivazhagan et al., 2019). (Kim and Rush, 2016) also investigated knowledge distillation, in which the student model learns to match the teacher's behaviors at the word and sequence levels. (Wei et al., 2019) suggested an online knowledge distillation method that uses the best checkpoints as the instructor model. Contextual models learned by unsupervised pre-trained text classification, such as ULMFIT (Howard and Ruder, 2018) in which the authors effectively investigated and present a transfer learning method that can be employed to any task in NLP, and introduce techniques for fine-tuning. (Devlin et al., 2019) employed the newly introduced BERT Transformer model across 26 distinct text classification tasks, encompassing the GLUE benchmark. The adapters not only attain state-of-the-art performance but also contribute only a marginal increase in parameters specific to the task. Specifically, on the GLUE benchmark, their approach achieves nearly indistinguishable results, falling short by just 0.4% compared to complete fine-tuning, all while introducing a mere 3.6% increase in parameters per task. In contrast, fine-tuning mandates the training of 100% of task-specific parameters. It is important to mention that the authors preserve the original network's parameters intact, facilitating an extensive level of parameter sharing. (de Masson d'Autume et al., 2019) introduced MBPA++, an episodic memory-based model that complements the encoder/decoder architecture. MBPA++ also does sparse experience replay and local adaptation to learn continuously. According to the researchers, MBPA++ trains quicker than A-GEM and takes no longer to train than an encoder/decoder model. The synergistic benefits of leveraging sparse experience replay and localized adaptation are highlighted through their experimental results in text categorization and question-answering tasks. They demonstrate that by selecting cases to store in memory at random, the space complexity of the episodic memory module may be lowered dramatically (by 50 to 90%) with a negligible impact on performance. The results show that, while the model's performance diminishes as the number of stored examples drops, the model can still retain a reasonably good performance even with only 10% of the whole model's memory capacity. Except for a few papers, a CL scenario for word and sentence representations has received little attention so far. To address this issue, (Wei et al., 2019) suggests a meta-learning method that utilizes information from previous multi-domain corpora to enhance new domain embeddings. (Liu et al., 2019) presented a sentence encoder that is constantly updated utilizing matrix conceptors to learn corpus-dependent features. Importantly, (Wang et al., 2019) presents a sentence embedding alignment for Lifelong learning by suggesting that when an NN model is trained on a new task, the embedding vector space undergoes undesirable modifications, rendering embeddings for earlier tasks infeasible. In the context of handling unstructured data and distribution shifts. The authors of (Vijay and Priyanshu, 2022) Proposed NERDA-Con, a pipeline for training Named Entity Recognizers (NERs) with Large Language Model (LLM) bases. They incorporate Elastic Weight Consolidation (EWC) into the NER fine-tuning NERDA pipeline particularly for integrating distinct tasks and updating distribution shifts.

4 DISCUSSION AND OPEN ISSUES

We examined current developments in CL across various situations and NLP tasks. Currently, most NLP tasks rely on annotated data, whereas a preponderance of unannotated data encourages research toward deep data-driven unsupervised approaches. Given the potential superiority of deep learning approaches in NLP applications, it becomes critical to conduct a detailed review of various deep learning methods and architectures with a focus on NLP applications. CL in NLP currently lacks benchmark datasets. The majority of articles create baselines and run tests on their own datasets. This renders it challenging to measure and compare progress in the field. In addition, the performance is affected by the order of tasks in the task sequence. Furthermore, learning on a small number of samples (e.g., using few-shot learning) is a significant issue for existing models, as is performing out-of-distribution generalization. Widely utilized in NLP sequence-to-sequence models, in particular, continue to struggle with systematic generalization (Lake and Baroni, 2018), unable to acquire general principles and reason about high-level language ideas. Additionally, researchers face the challenge of dealing with historical documents, which pose specific difficulties. Ancient papers and documents pose significant challenges due to the presence of unwanted elements such as noise, and some characters may be missing entirely or partially. Decoding ancient papers goes beyond noise and missing characters. It involves historical context, language evolution, preservation issues, and interdisciplinary collaboration. Technological advances like computational linguistics and machine learning help unveil the mysteries within these texts, offering insights into ancient civilizations. Exploring these complexities enriches our understanding of this endeavor's importance. When we delve deeper into these complexities, we uncover the need for interdisciplinary collaboration involving historians, linguists, computer scientists, and experts in various domains. Techniques such as computational linguistics, machine learning, and image analysis can be harnessed to enhance our understanding of these documents. It's fascinating how advances in technology can aid us in unraveling the mysteries embedded in these texts, shedding light on ancient civilizations and cultures. Moreover, historical databases are limited in number. The dataset used for training is often generated from modern handwriting, rendering word recognition and character recognition models ineffective in deciphering historical texts. This calls for new areas of research such as domain adaptation and active learning.

5 CONCLUSIONS

In this paper, we provided a comprehensive overview of contemporary NLP research encompassing static learning and continual learning. We delved into the application of continual learning methods to mitigate concept forgetting in diverse NLP tasks. In closing, we underscored persistent challenges and identified research gaps that warrant further exploration, signaling areas demanding more in-depth investigation and attention within the dynamic landscape of natural language processing.

REFERENCES

Al-Anzi, F. S. and AbuZeina, D. (2016). Big data categorization for arabic text using latent semantic indexing and clustering.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges.

Belouadah, E., Popescu, A., and Kanellos, I. (2020). A comprehensive study of class incremental learning algorithms for visual tasks. *CoRR*, abs/2011.01844.

Bengio, Y. and Senécal, J.-S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*.

de Masson d'Autume, C., Ruder, S., Kong, L., and Yogatama, D. (2019). Episodic memory in lifelong language learning.

Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J., Ijaz, M. F., et al. (2022). A complete process of text classification system using state-of-the-art nlp models. *Computational Intelligence and Neuroscience*.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Escolano, C., Costa-jussà, M. R., Fonollosa, J. A. R., and Artetxe, M. (2020). Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders.

Esmaeilzadeh, A. and Taghva, K. (2021). Text classification using neural network language model (nnlm) and bert: An empirical comparison. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys)*. Springer.

Farquhar, S. and Gal, Y. (2019). Towards robust evaluations of continual learning.

Fatima, S., Srinivasu, B., et al. (2017). Text document categorization using support vector machine. *International Research Journal of Engineering and Technology (IRJET)*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*. PMLR.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv:1801.06146*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*.

Jang, J., Kim, Y., Choi, K., and Suh, S. (2021). Sequential targeting: A continual learning approach for data imbalance in text classification. *Expert Systems with Applications*, 179:115067.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Ke, Z. and Liu, B. (2023). Continual learning of natural language processing tasks: A survey.

Kharrat, A., Drira, F., Lebourgeois, F., and Garcia, C. (2023). Toward digits recognition using continual learning. In *IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*.

Kim, C. D., Jeong, J., and Kim, G. (2020). Imbalanced continual learning with partitioning reservoir sampling. In *Computer Vision–ECCV 2020: 16th European Conference, UK, Proceedings, Part XIII 16*. Springer.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Lake, B. and Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *CoRR*.

Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., and Zhang, B.-T. (2018). Overcoming catastrophic forgetting by incremental moment matching.

Li, J., Monroe, W., Shi, T., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *CoRR*, abs/1701.06547.

Liu, T., Ungar, L., and Sedoc, J. (2019). Continual learning for sentence representations using conceptors. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*. Association for Computational Linguistics.

Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., and Sanner, S. (2022). Online continual learning in image classification: An empirical survey. *Neurocomputing*.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Elsevier.

Mittal, V., Gangodkar, D., and Pant, B. (2020). Exploring the dimension of dnn techniques for text categorization using nlp. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE.

Mundt, M., Hong, Y., Pliushch, I., and Ramesh, V. (2023). A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *Neural Networks*.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2018). Continual lifelong learning with neural networks: A review. *CoRR*.

Qu, H., Rahmani, H., Xu, L., Williams, B., and Liu, J. (2021). Recent advances of continual learning in computer vision: An overview.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota. Association for Computational Linguistics.

Santos, C. D. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, Bejing, China. PMLR.

van de Ven, G. M. and Tolias, A. S. (2019). Three scenarios for continual learning.

Vijay, S. and Priyanshu, A. (2022). Nerda-con: Extending ner models for continual learning–integrating distinct tasks and updating distribution shifts.

Wang, C., Luo, F., Li, Y., Xu, R., Huang, F., and Zhang, Y. (2022). On effectively learning of knowledge in continual pre-training. *arXiv preprint arXiv:2204.07994*.

Wang, H., Xiong, W., Yu, M., Guo, X., Chang, S., and Wang, W. Y. (2019). Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Wei, H.-R., Huang, S., Wang, R., Dai, X.-y., and Chen, J. (2019). Online distilling from checkpoints for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2016). Seqgan: Sequence generative adversarial nets with policy gradient. *CoRR*.