ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Continual Improvement of Deep Neural Networks in The Age of Big Data

Alexander Gepperth[1] and Timothée Lesort[2]

1- University of Applied Sciences Fulda - Department of Applied Computer Science
Leipzigerstraße 123, 36093 Fulda - Germany

2- Aignostics GmbH
Alt-Moabit 73/73a, 10555 Berlin, Germany

Abstract.
Many applications of deep learning are set in an environment with perpetual change or at least with an ever-growing amount of data. In practice, deep neural network (DNNs) and large language models (LLMs) are continually trained and evaluated. They need to incorporate new data or new annotations, where one typical issue is the extensive availability of unannotated or low-quality data, coupled with a bottleneck concerning annotations and/or curated samples. In such setups, the scaling behavior of continual learning (CL) algorithms w.r.t. training time becomes critical, which is in contrast to the standard CL setting operating on small databases like MNIST, CIFAR or ImageNet. Annotations or curated samples become available progressively, e.g., because they are created by humans, or due to an ongoing exploration of the environment, and need to be progressively incorporated into models. This article explores how advancement in continual learning can improve the scalability and performance of DNNs and LLMs in such setups. One interesting aspect is to leverage dedicated (small-scale) CL techniques to achieve advantageous trade-offs between computational cost and accuracy, or how such CL methods can maintain advantageous scaling behavior w.r.t. continuous re-training on all data.

1 Introduction

In many applications, new data arrive continuously and are added to an existing data pool, and there is a need for a continual adaptation of models to all of the existing data. Typical examples are large language models (LLMs), but classification tasks(e.g., vehicle classification) can serve as a useful example as well, since a model needs to be up-to-date w.r.t. new vehicle brands and types. If we simplify this to a scenario where the amount of new data per time unit $d(t)\ equivd$ is constant over time, the total amount of data at time t is unbounded and grows linearly. Since training time is ideally proportional to the amount of data used for training, this implies unbounded linear scaling of training time and, above all, cost. Please see fig. 1 for a visualization of this fundamental problem in machine learning.

This is a very serious and fundamental issue, the solution of which would enable really large-scale learning over long time periods. Some inspiration can

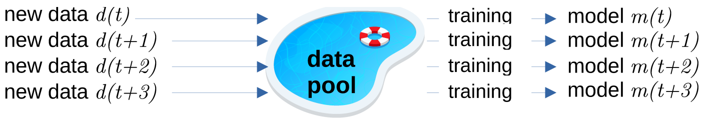

Fig. 1: Continuous retraining as it is commonly performed in applications. New data are continuously fed into an ever-growing data pool, which is continuously used to train new models. The computational cost associated with this procedure is unbounded and linear in time.

come from the field of continual learning (CL, see, e.g., [33]) which studies machine learning from non-stationary data distributions.

2 Continual Learning at Scale

CL is by now a quite broad sub-field of machine learning, with many new avenues still being explored, see [33]. Traditionally, CL has focused on classification tasks and supervised learning, but there are contributions in unsupervised and reinforcement learning as well.

In supervised CL, data distributions are simplified from being completely non-stationary to being non-stationary only at a finite set of transition points between *tasks* where data distributions are taken to be stationary, see fig. 2. CL in this scenario is required to accumulate knowledge of all tasks, typically without any forgetting. Conflicting tasks, or deliberate forgetting, are not considered.

This simple "default scenario" maps quite well to the continuous retraining scenario, where every set of newly arriving data defines a CL task. Similarly, typical application requirements assume that data is non-contradictory, so an accumulation of knowledge without forgetting is the desired outcome.

However, this is just one aspect of CL research. Other scenarios addressed in CL research include:

- Data growth: The scenario where more and more data becomes available is

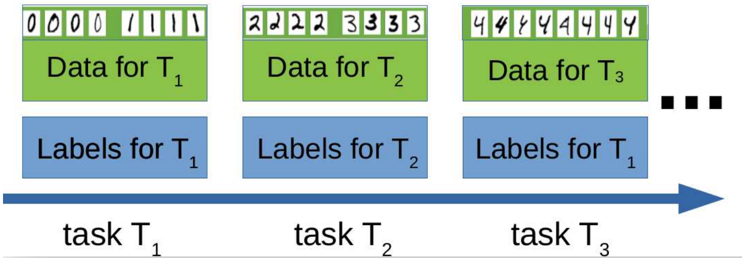

Fig. 2: The "default scenario" in supervised CL. Data are assumed to be stationary except at tasks onsets. Often, such non-stationarities are modeled by introducing a new class to the model.

the most common one in continual learning as in class incremental learning, domain incremental learning [32]. There might be constraints on accessing past data, however the total amount of data grows over time.

- Objective shift: The objective shift [16], where the loss to optimize changes through time has also been investigated, for example in reinforcement learning [31], but also in a classification where the same image can have a more and more precise label [1].
- Annotation growth: Continual learning can also be applied if all the data is available since the beginning but annotation grows through time as they are gathered. The model continually learns to be on top of the current state of annotation and maximize possible performance. [35, 34]
- Compute growth: Something that has not been specifically investigated in the literature, however, we could imagine specific approaches to handle a change in compute availability. Indeed some hyper-parameters are set for hardware specification [4], therefore a change in available compute resources can lead to a change in optimal hyper-parameters for training and change the optimization process.

3 Continual Learning as a solution for training at scale

Continual learning mainly focuses on acquiring new knowledge without forgetting already acquired knowledge. When training at scale, avoiding to reprocess data usually helps to improve efficiency, i.e. avoiding forgetting is the way to scale further. Since the potential types of non-stationary data distributions, and the possible application constraints, are endless, many different aspects are addressed by CL research:

- **Forgetting and knowledge accumulation:** Continual learning aims at training efficiently machine learning models in dynamic data distribution scenarios. The community has mostly focused on primarily preventing models from forgetting while maintaining some plasticity [18, 13, 19, 26, 8], however recently some approaches aim at a relaxed and more general approach consisting of ensuring that the model forgets less than it learns and accumulates knowledge [17, 6].
- **CL at constant time complexity:** An in principle very obvious solution is to make the learning complexity of CL proportional to the amount of *new* data, not all data. After all, the old data has already been learned, so most of the effort invested in continual re-training is actually wasted. Two principled approaches have been put forward: Maximally Interfered Retrieval [3] and Adiabatic Replay [15]. Both use new data to generate synthetic samples, the incorporation of which will protect the underlying model from forgetting when including them in the training process. Hence, training complexity will be constant even over long time scales. A similar,

scalable strategy is believed to be employed by biological agents, see, e.g., [14, 20]. In the line of simple and scalable approaches, when data distribution reoccurs through time, SGD and Adam showed interesting knowledge retention and accumulation properties [17] in particular in self-supervised learning [29, 12, 11, 10].

- **Connection to federated learning:** While continual learning aims at training on different data distributions in a sequential manner, federated learning aims at training on multiple data distributions in a parallel manner [21, 27, 9, 28]. Numerous examples such as model merging [24] show that methods for both can be identical and having both would be a significant advantage to scale up machine learning.

4 Discussion: The annotation bottleneck

Annotations are a necessary step to define what is the expected behavior of any artificial intelligence system.

If data gathering practice scrapping the internet or other large existing databases had allowed the creation of huge datasets [cite], gathering labels to train a model for a specific application remains complicated and costly. Several strategies exist to make this easier and more label-efficient.

- **Self-supervised training and priors:** One of the recent improvements in label efficiency is the development of general foundation models that can easily be adapted for downstream tasks in language and vision [cite]. Those models incorporate general knowledge and prior that can later be beneficial for continual learning downstream tasks [7, 23, 22]
- **Automatic label extraction:** Beyond making the model easier to fine-tune, some strategies create labels automatically, either by crossing databases, automatically using caption of images for example from Instagram, or by generating labels automatically. This strategy can, for example, be used in Vision-Language Models (VLMs) such as PaliGemma [5], CLIP [25], or Flamingo [2].
- **Designing a reward function:** RL has also recently helped to fine-tune models in an efficient way to make models more aligned with our objectives like RLHF for language models [36], or reward function for vision models [30].
- **Active learning:** Historically active learning is the research field that aims at gathering labels in a cost-efficient way and some methods are often used to guide annotations [cite].

Continual learning can considerably help with training at scale and improving performance, however, as one still has to define the expected behavior of the models, developing an efficient annotation system and fine-tuning either for

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

training or evaluating stays a bottleneck for developing and assessing AGI models.

5 Conclusion

The recent interest in scalable machine learning underlines the need for a principled solution to the problem of continuous retraining. Such a scalable solution will almost certainly include techniques from the CL domain and pave the way to truly long-term machine learning, with a significant impact on energy consumption, training time, and cost. To this end, CL methods must be adapted for scalability which has not hitherto been a significant issue. In supervised CL, the number of tasks (or re-learning iterations) must be assumed to be very large to make scaling effects apparent in future CL research.

Apart from this principled problem, large-scale learning will still require appropriate strategies to improve label efficiency, for better leveraging the available data and the (usually low) amount of available annotations.

References

[1] M. Abdelsalam, M. Faramarzi, S. Sodhani, and S. Chandar. Iirc: Incremental implicitly-refined classification. *CVPR*, pages 11038–11047, 2021.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] R. Aljundi, L. Caccia, E. Belilovsky, M. Caccia, M. Lin, L. Charlin, and T. Tuytelaars. Online continual learning with maximally interfered retrieval. In *Neural Information Processing Systems*, 2019.

[4] Q. Anthony, J. Hatef, D. Narayanan, S. Biderman, S. Bekman, J. Yin, A. Shafi, H. Subramoni, and D. Panda. The case for co-designing model architectures with hardware. *arXiv preprint arXiv:2401.14489*, 2024.

[5] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[6] M. Caccia, P. Rodriguez, O. Ostapenko, F. Normandin, M. Lin, L. Caccia, I. Laradji, I. Rish, A. Lacoste, D. Vazquez, and L. Charlin. Online fast adaptation and knowledge accumulation: a new approach to continual learning. *NeurIPS*, 2020.

[7] A. Cossu, A. Carta, L. Passaro, V. Lomonaco, T. Tuytelaars, and D. Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024.

[8] T. Doan, M. Abbana Bennani, B. Mazoure, G. Rabusseau, and P. Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR, 13–15 Apr 2021.

[9] A. Douillard, Q. Feng, A. A. Rusu, R. Chhaparia, Y. Donchev, A. Kuncoro, M. Ranzato, A. Szlam, and J. Shen. Diloco: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

[10] K. Fujii, T. Nakamura, M. Loem, H. Iida, M. Ohi, K. Hattori, H. Shota, S. Mizuki, R. Yokota, and N. Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*, 2024.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

[11] E. Gogoulou, T. Lesort, M. Boman, and J. Nivre. A study of continual learning under language shift. *arXiv preprint arXiv:2311.01200*, 2023.

[12] A. Ibrahim, B. Thérien, K. Gupta, M. L. Richter, Q. Anthony, T. Lesort, E. Belilovsky, and I. Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

[13] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proc. of the national academy of sciences*, 2017.

[14] M. Klasson, H. Kjellström, and C. Zhang. Learn the time to learn: Replay scheduling in continual learning. *Transactions on Machine Learning Research*, 9, 2023.

[15] A. Krawczyk and A. Gepperth. Adiabatic replay for continual learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2024.

[16] T. Lesort, M. Caccia, and I. Rish. Understanding continual learning settings with data distribution drift analysis. *arXiv preprint arXiv:2104.01678*, 2021.

[17] T. Lesort, O. Ostapenko, D. Misra, M. Rifat Arefin, P. Rodríguez, L. Charlin, and I. Rish. Challenging Common Assumptions about Catastrophic Forgetting. *arXiv e-prints*, page arXiv:2207.04543, July 2022.

[18] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[19] D. Lopez-Paz and M.-A. Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6467–6476. Curran Associates, Inc., 2017.

[20] J. L. McClelland, B. L. McNaughton, and A. K. Lampinen. Integration of new information in memory: new insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B*, 375(1799):20190637, 2020.

[21] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[22] S. V. Mehta, D. Patil, S. Chandar, and E. Strubell. An empirical investigation of the role of pre-training in lifelong learning (2021). In *URL https://openreview. net/forum*, 2021.

[23] O. Ostapenko, T. Lesort, P. Rodríguez, M. R. Arefin, A. Douillard, I. Rish, and L. Charlin. Continual learning with foundation models: An empirical study of latent replay, 2022.

[24] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, and F. Piccialli. Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, 150:272–293, 2024.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[27] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[28] M. Ryabinin, E. Gorbunov, V. Plokhotnyuk, and G. Pekhimenko. Moshpit sgd: Communication-efficient decentralized training on heterogeneous unreliable devices. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18195–18211. Curran Associates, Inc., 2021.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[29] T. Scialom, T. Chakrabarty, and S. Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, 2022.

[30] A. Susano Pinto, A. Kolesnikov, Y. Shi, L. Beyer, and X. Zhai. Tuning computer vision models with task rewards. *arXiv e-prints*, page arXiv:2302.08242, Feb. 2023.

[31] R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, G. Cai, N. D. Rodríguez, and D. Filliat. Discorl: Continual reinforcement learning via policy distillation. *CoRR*, abs/1907.05855, 2019.

[32] G. M. van de Ven and A. S. Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[33] E. Verwimp, R. Aljundi, S. Ben-David, M. Bethge, A. Cossu, A. Gepperth, T. L. Hayes, E. HÃ¼llermeier, C. Kanan, D. Kudithipudi, C. H. Lampert, M. Mundt, R. Pascanu, A. Popescu, A. S. Tolias, J. van de Weijer, B. Liu, V. Lomonaco, T. Tuytelaars, and G. M. van de Ven. Continual learning: Applications and the road forward. *Transactions on Machine Learning Research (TMLR)*, 2024.

[34] Y. Zhang, P. Zhao, S. Niu, Q. Wu, J. Cao, J. Huang, and M. Tan. Online adaptive asymmetric active learning with limited budgets. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2680–2692, 2019.

[35] S. Zhou, H. Zhao, S. Zhang, L. Wang, H. Chang, Z. Wang, and W. Zhu. Online continual adaptation with active self-training. In *International Conference on Artificial Intelligence and Statistics*, pages 8852–8883. PMLR, 2022.

[36] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.