Class-Incremental Continual Learning for General Purpose Healthcare Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Healthcare clinics regularly encounter dynamic data that changes due to variations in patient populations, treatment policies, medical devices, and emerging disease patterns. Deep learning models can suffer from catastrophic forgetting when fine-tuned in such scenarios, causing poor performance on previously learned tasks. Continual learning allows learning on new tasks without performance drop on previous tasks. In this work, we investigate the performance of continual learning models on four different medical imaging scenarios involving ten classification datasets from diverse modalities, clinical specialties, and hospitals. We implement various continual learning approaches and evaluate their performance in these scenarios. Our results demonstrate that a single model can sequentially learn new tasks from different specialties and achieve comparable performance to naive methods. These findings indicate the feasibility of recycling or sharing models across the same or different medical specialties, offering another step towards the development of general-purpose medical imaging AI that can be shared across institutions.

1 Introduction

Deep Neural Networks (DNNs) have recently exhibited remarkable achievements in various tasks, surpassing human expertise in some cases (7; 12; 14). However their dependence on fixed, balanced datasets within stable environments presents a significant constraint. The ever-changing nature of the real world requires networks capable of sequential learning over time and adapting to shifting data distributions. This shortcoming is especially pronounced in healthcare and medical imaging. The emergence of new diseases, changes in patient population, treatment policies, disease distribution, imaging hardware, or image acquisition techniques can significantly impact the model's performance. Fine-tuning exclusively on new data, adapts models to the latest targets, resulting in a rapid loss of previously acquired knowledge. Techniques such as Joint Training (JT) are used to overcome this, where the model is trained on both old and new data. However, healthcare data can't always be shared due to safety concerns and regulation differences across geographies. On the contrary, the NAIVE approach trains an independent model for each task, increasing computational resources required for training, deployment, and missing performance gain due to shared representation. Continually learning is an active area of research, allowing efficient training and adaptation of algorithms to new data without losing prior knowledge. This can improve the model updating, sharing, and resource optimization for healthcare institutions. Cross-sharing models between multiple hospitals can benefit both institutions. Finally, healthcare professionals will be able to screen and detect patients more effectively by quickly identifying and analyzing new biomarkers as they emerge with disease or population shift.

In this work, we show the feasibility of training continually learning models for medical imaging. Our contributions are as follows:

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

1. We explore the potential of continual learning for sharing medical imaging AI algorithms across changes in hospitals/geographies, medical specialties, and imaging modalities.
2. We create 4 continual learning scenarios to assess cross-sharing (inter-hospital scenario and one inter-specialty) and intra-specialty model recycling (pathology, radiology) use cases.
3. We show that continual learning methods can gain performance on par with naive and joint learning approaches while remembering previous tasks.

2 Methods

We implement 6 variants of continual learning methods, namely memory aware synapses (MAS), replay using memory indexing (REMIND), MAS with replay (MAS+r), Neuro-inspired stability-plasticity adaptation (NISPA), dark experience replay (DER), and DER++; with prior two not requiring data access to previous tasks. The final four require data access, generally solved by local storage in replay buffers of fixed sizes. MAS (3) and its replay variant MAS+r are regularization methods that calculate the importance of the model parameters in an online fashion. REMIND (9) stores compressed low-level feature representations instead of actual input data, making it well-suited when past data is not feasible. NISPA (8) uses a rewiring mechanism inspired by the structural plasticity of biological neurons and driven by local activations of units similar to Hebbian learning in the brain. DER (4) and DER++ are replay methods that selectively choose examples with high uncertainty to replay. We train naïve learner, and joint learner for baseline comparison. The naive learner corresponds to training an independent model for each data, and the joint learner is trained on all data together. We use a 5-layered convolutional neural network (CNN) followed by a linear layer classifier as a backbone model. No task labels are provided during testing. To succeed, the model needs to learn inter-task differences to predict task ID on testing and learn intra-task differences to predict the correct class. We measure task accuracy percentage after every episode, average accuracy on seen classes after completing an episode, and backward transfer.

2.1 Datasets and scenarios

Figure 1 provides a snapshot of scenarios and datasets used. We devise four continual learning scenarios, divided into 3 major categories, to simulate learning new tasks inside the same specialty, different specialties, and hospitals. **Inter-hospital scenario** simulates model sharing across hospitals from different geographies. We used x-rays for pleural effusion, cardiomegaly, atelectasis, and consolidation from the Chexpert dataset, CXR-14 dataset, and VinBig dataset; in total representing 3 hospitals and 2 countries. **Inter-Specialty scenarios** simulate model sharing between different medical specialties inside the same hospital and can help specialties with fewer data to benefit from those with more. We combine three specialties; pathology(10), radiology(11), and dermatology(5; 6). **Intra-Specialty scenarios** simulate model rotation within a specialty, mimicking learning new disease finding that appears later on. The pathology scenario contains three subtasks: histology of colorectal cancer (11), blood cells (1), and kidney cortex cells (13). The radiological scenario contains three subtasks: computed tomography (CT) (15)), ultrasound (2), and chest X-ray (11) images.

3 Results and Discussion

In this section, we discuss the performance of continual learning methods with comparison at the scenario and algorithm level. Figure 1 shows the average accuracy of methods on test data, with every point representing average accuracy on current and all previous tasks. As expected accuracy for NAIVE method takes a sharp dip. Continual learning methods perform on par or better, with NAIVE on the current task, with little to no drop in their performance on the previous task. Across all continual learning methods, replay methods perform better than regularization methods across all scenarios. MAS with replay and DER++ get high accuracies compared to other methods. MAS+r is consistently the best performer in all scenarios, achieving an average accuracy of 88,82,75,79, on inter-hospital, inter-specialty, pathology, and radiology respectively at the end of each scenario. MAS with replay had the highest backward transfer with a value of -2, -5, +3, and -5 on the respective scenarios. Among non-replay methods, MAS has a huge performance drop. REMIND achieves an average accuracy of 83,77,75,80 on inter-hospital, inter-specialty, pathology, and radiology respectively. It's important to note that replay methods have access to a subset of previous data stored for future

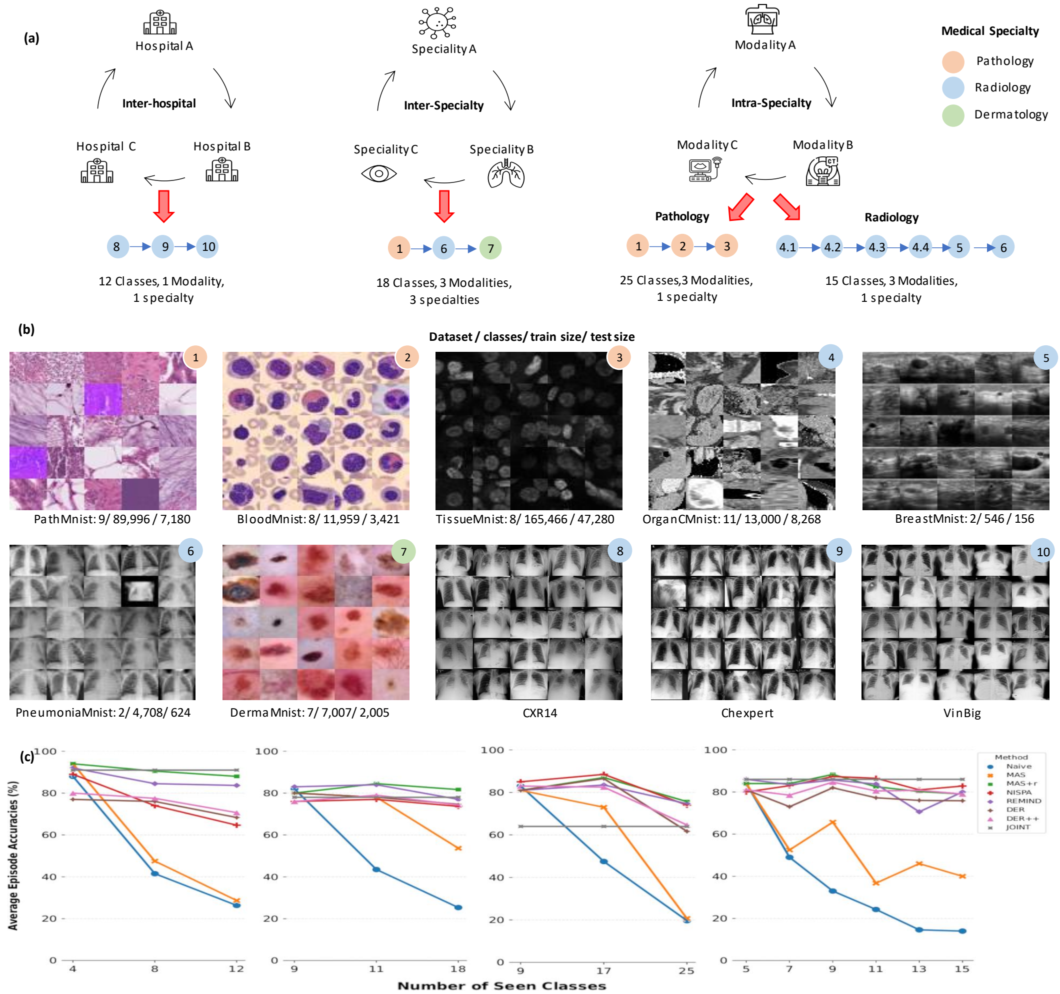

Figure 1: Different learning scenarios and dataset. (a) shows scenarios and sequences of learning tasks used to train the algorithms. (b) shows snapshots of images from different datasets, color-coded (as per their specialty) sequence index on the right top of images. Each dataset has two or more classification labels. Note the visual similarities inside each dataset and the diversity among the different tasks in the scenario. Dataset 4 is split into 4 sub-tasks with unique classes, named 4.1,4.2,4.3 and 4.4

retraining. On the other hand, REMIND stores compressed representations, instead of the actual data, allowing memory efficiency and data privacy. Another critical point to note is REMIND's dependence on its feature extractor. Since the initial feature extractor is frozen after its initialization, the model is less flexible for learning tasks unrelated to initial studies. Using pre-trained weights from bigger datasets can provide a boost in model performance.

Limitations and Future work We don't explore the effect of the sequence of tasks on learning, as this can impact quality of features learned. We used 32*32*3 image size on a small CNN architecture. Using higher resolution medical imaging and larger pre-trained models can help boost performance further. This can be an important future direction.

4 Conclusion

In this work, we explore the performance of continuous learning methods in varying specialties, modalities, and geographies. We show that continual learning algorithms can learn new tasks while maintaining performance on previous tasks, even while changing modalities specialties, and hospitals.

This shows potential in developing general-purpose medical imaging AI that can be shared across institutions, with the ability to adapt to new tasks.

5 Potential Negative Societal Impacts

Continual learning models may inherit biases present in the data on which they are trained. If the training data is not representative, these biases can lead to disparities in medical diagnoses and treatment recommendations. While these risks are inherent in deep learning models, automatic unsupervised training of continual learning can exacerbate these biases when deployed, often going unnoticed. Furthermore, due to concerns related to quality control and disparities in deployment regions, continual learning models may inadvertently generate incorrect or misleading medical images or interpretations, which can have detrimental consequences for patients if not adequately monitored and controlled. AI systems are highly sensitive, potentially leading to overdiagnosis and overtreatment if not properly calibrated, thereby resulting in unnecessary medical interventions and increased healthcare costs. Lastly, as with any technological advancement, AI systems are susceptible to hacking and cybersecurity threats. Breaches of medical AI systems could result in unauthorized access to sensitive patient data or manipulation of diagnostic results, posing significant privacy and security risks.

References

[1] Acevedo, A., Merino, A., Alférez, S., Molina, , Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief **30**, 105474 (jun 2020). https://doi.org/10.1016/j.dib.2020.105474, `http://dx.doi.org/10.1016/j.dib.2020.105474`

[2] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (feb 2020). https://doi.org/10.1016/j.dib.2019.104863, `http://dx.doi.org/10.1016/j.dib.2019.104863`

[3] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: The European Conference on Computer Vision (ECCV) (2018)

[4] Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: Advances in Neural Information Processing Systems. vol. 33 (2020)

[5] Chen, E.Z., Dong, X., Wu, J., Jiang, H., Li, X., Rong, R.: Lesion attributes segmentation for melanoma detection with deep learning. https://doi.org/10.1101/381855, `https://www.biorxiv.org/content/10.1101/381855v4`, pages: 381855 Section: New Results

[6] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). https://doi.org/10.48550/arXiv.1902.03368, `http://arxiv.org/abs/1902.03368`

[7] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing **187** (2016)

[8] Gurbuz, M.B., Dovrolis, C.: Nispa: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. arXiv preprint arXiv:2206.09117 (2022)

[9] Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C.: Remind your neural network to prevent catastrophic forgetting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 466–483. Springer (2020)

[10] Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., Jansen, L., Reyes-Aldasoro, C.C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., Halama, N.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Medicine **16**(1), e1002730 (jan 2019). https://doi.org/10.1371/journal.pmed.1002730, `http://dx.doi.org/10.1371/journal.pmed.1002730`

[11] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M.Y.L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A.N.,

Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131.e9 (feb 2018). https://doi.org/10.1016/j.cell.2018.02.010, http://dx.doi.org/10.1016/j.cell.2018.02.010

[12] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015)

[13] Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature Methods **9**(7), 637 (jun 2012). https://doi.org/10.1038/nmeth.2083, http://www.nature.com/doifinder/10.1038/nmeth.2083

[14] Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529** (2016)

[15] Xu, X., Zhou, F., Liu, B., Fu, D., Bai, X.: Efficient multiple organ localization in CT image using 3D region proposal network. IEEE transactions on medical imaging (jan 2019). https://doi.org/10.1109/TMI.2019.2894854, http://dx.doi.org/10.1109/{TMI}.2019.2894854