Continual Learning for medical image classification

Alessandro Quarta[1,2,3], Pierangela Bruno[1] and Francesco Calimeri[1,3]

[1]Department of Mathematics and Computer Science, University of Calabria, Rende, Italy
[2]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy
[3]DLVSystem Srl, Rende, Italy

Abstract

Continual Learning (CL) is a novel paradigm in which the trained model is computed via a stream of data where tasks and data are only available over-time. Indeed, such approaches are able to learn new skills and knowledge without forgetting the previous ones: no access to previously encountered data and mitigate catastrophic forgetting.

In this work, we propose a comparison of different CL algorithms in performing the classification of medical images. In particular, we aim to highlight the potential and ability of current methods in preventing catastrophic forgetting of the previous tasks when a new one is learned.

CL-based methods have been tested for the classification of medical images showing the viability and effectiveness of these approaches.

Keywords

Continual Learning, Deep Learning, Medical Imaging, Lifelong Learning, Incremental Learning, Online Learning

1. Introduction

Artificial Intelligence, and especially techniques based on Machine learning, have sped up the automation of many processes, achieving performance comparable to humans in some specific tasks. In particular, in the last decades, supervised neural networks have shown a great deal of potential in dealing with numerous tasks such as natural language processing [1], object detection [2] and medical imaging [3].

In this context, in order to properly deploy a predictive model, huge amounts of data complying with high-quality standards are needed. However, such a requirement is not always satisfiable, depending on several factors. For instance, data frequently come from heterogeneous sources, might feature a large number of missing or null values, and may also be subject to usage restrictions because of particular agreements or privacy concerns. For example, the data could be stored in predefined servers without the possibility of extraction, thus making it difficult to implement integration mechanisms between sources, or the data may be available for the training phases of the various models only for a very limited period of time.

In recent years, these problems have attracted a lot of attention from the scientific community which proposed and studied different techniques to overcome such limitations.

1st AIxIA Workshop on Artificial Intelligence For Healthcare
*Corresponding author.

Continual Learning (CL) – referred to also as continuous learning, incremental learning, lifelong learning, and online learning – aims at defining models that continuously learn and evolve according to new amounts of data, retaining previously learned knowledge [4]. In this way, the model is able to incrementally learn and autonomously change its behaviour without forgetting the original task [4].

In general, humans learn concepts sequentially, and while humans may gradually forget old information, a complete loss of previous knowledge is rarely attested. By contrast, artificial neural networks work differently: they suffer from catastrophic forgetting of old concepts as new ones are learned [5]. This is a direct result of a more widespread issue with neural networks known as the stability-plasticity dilemma [6]. Stability refers to the ability to preserve prior knowledge while encoding it, while plasticity refers to the capacity to integrate new knowledge.

Furthermore, other potential duties can exist, especially in healthcare. For example, it is not easy to train specialized models from scratch, so that they can accomplish and predict all possible tasks and activities. It would require a huge annotated dataset containing, for example, all possible disease or image modalities that make such approaches non-scalable. Instead, more realistic scenarios include physicians that receive a model already trained on typical tasks; when a (completely) new activity occurs, the model can be trained again to solve new problems. In particular, the model should first swiftly master the new task with little guidance by drawing on prior knowledge from completed tasks. Additionally, it must be able to combine new information with existing knowledge to enhance itself for both the current task and any upcoming ones. Eventually, it should be possible to learn a new task without using training data for previous tasks, which might no longer be accessible.

In this work, we propose the use of CL approach to support Deep Learning (DL) architectures in classifying medical images; furthermore, we aim to show the feasibility and viability of such approaches via careful experimental analysis.

The remainder of the paper is structured as follows: In Section 2 we describe the related work; then, we report about a careful experimental activity in Section 3, which is discussed in Section 4; eventually, we draw our conclusions in Section 5.

2. Related works

Early studies discovered the catastrophic forgetting problem when learning samples of various input patterns sequentially. There have been several avenues investigated, and the majority of approaches to continuous learning do not rely on a single method to tackle catastrophic forgetting. Each strategy has benefits and drawbacks [7] but in many cases combining methods enables the discovery of the best answers.

Lesort et al. [7] describe several continual learning methods, classifying them into four different classes, as shown in Figure 1.

Baweja et al. [8] investigate the potential of Elastic Weight Consolidation (EwC) [9] algorithm in biomedical imaging to prevent the catastrophic forgetting of a neural network that aims to learn two different tasks sequentially. The goal is the multi-class segmentation of cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM) and the segmentation of white matter lesions (WML) in MRI images. This investigation on bio-medical data shows that the technique

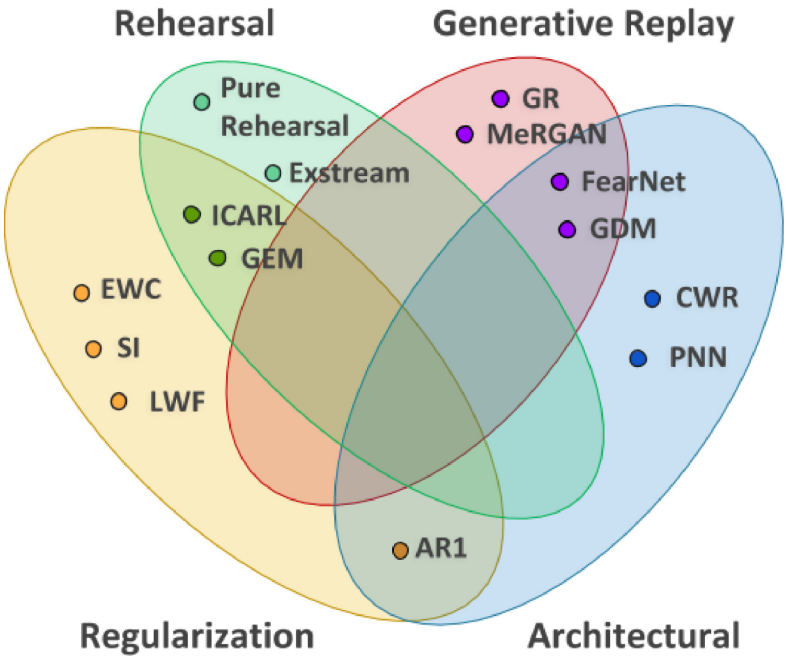

Figure 1: The four different classes defined by [7]. Figure extracted from [7]

is promising for alleviating catastrophic forgetting [8].

Wachinger et al. [10] suggest a continuous learning method for segmenting the brain in which a single network is successively trained on samples from various domains. The authors extend an importance-driven methodology to segment medical images. Specifically, they included learning rate regularization to stop the network's information from being lost. In particular, using Memory Aware Synapses (MAS) [11], for each network parameter, a weight is given to reflect its significance for a particular task. The important weights are calculated by MAS using an unsupervised method, where the importance weights represent the output of the network's sensitivity to changes in its parameters. Wachinger et al. [10] implemented MAS-LR, a parameter-specific learning rate based on the parameter's significance. In order to do this, the authors offered MAS-Fix, which only tweaks irrelevant factors while freezing crucial parameters throughout the training of a new assignment. As a result, the learning rate for critical parameters will be reduced while remaining constant for unimportant parameters.

Gonzalez et al. [12] proposed a fair multi-model benchmark. They distinguished between whether domain knowledge information is available during inference because it is uncertain in the incremental domain scenario. In the more straightforward scenario, which we refer to as Domain Knowledge, test inputs have the form (x, I), where I declares that x belongs to X_i. The capacity to keep domain-specific parameters that are not shared across domains is the core advantage of having domain identification information. As a result, only the shared parameters of the model need to be trained repeatedly. In contrast to the feature extraction part of the model, the final network layers in a classification model are frequently domain-specific and set during inference based on the domain precedence of the test instance

Therefore, the authors proposed a fair evaluation method. A comparison with the benchmark is possible if domain identity data can be employed during inference. They suggested the use of an oracle to determine the closest domain for an incoming image if domain information is not available. If the suggested approach needs domain knowledge, then a comparison of both the continual learning method and the benchmark that uses domain knowledge extrapolated from the oracle should be done. Only the benchmark would use oracle if the procedure does not employ domain knowledge. They make use of two different CL algorithms: EwC [9] and LwF [13].

A multi-domain learner can include new domains into lifelong learning with just a few labeled instances, maintaining performance on earlier domains [14]. In this study, Karani et al. provided strategies for segmentation across scanning protocols in a lifetime learning environment based on adaptive Batch Normalization (BN) layers. For each protocol/scanner, they specifically train a CNN with standard convolutional filters and unique BN parameters. To teach the network the proper convolutional filters, images from a few scanners are used as training data at first. By fine-tuning the BN parameters with a small number of labeled images, it can then be adjusted to work with different protocols and scanners.

Perkonigg et al. [15] proposed the Dynamic Memory (DM) and the DM method with pseudo-domain detection (DM-PD) approach that operate without domain knowledge, representing a more realistic assumption in clinical practice. The DM technique kept M-samples diversified and indicative of the visual variances across all domains. Selecting which image-target pairs to preserve in memory, without explicit domain knowledge, is a crucial stage in this method. Indeed, the pseudo-domain module distinguishes between multiple domains, acting as a hold for the unidentified, actual domains. During ongoing training, these pseudo-domains are utilized to balance the memory M and training-mini-batch T.

3. Methodologies and Experimental Activities

In this work, we proposed the use of different CL algorithms to support and improve the performance of the neural network in medical image classification. In particular, we compared Naive, Replay, CWR*, ICaRL, and Cumulative approaches. Naive and Cumulative methods are used as the lower and upper bound of our comparison. These approaches are tested in the task of multi-classification using a medical dataset that contains several pathologies, human districts, and image modalities.

3.1. Dataset

In this work, we used a collection of standardized biomedical photos, MedMNIST v2 [16]. MedMNIST v2 has 12 datasets for 2D images and 6 datasets for 3D. It includes several data types (binary/multi-class, multi-label, and ordinal regression), dataset sizes (ranging from 100 to 100,000), and tasks. An example of the dataset is shown in Figure 2. All datasets in MedMNIST v2 are subdivided into train-validation-test folders.

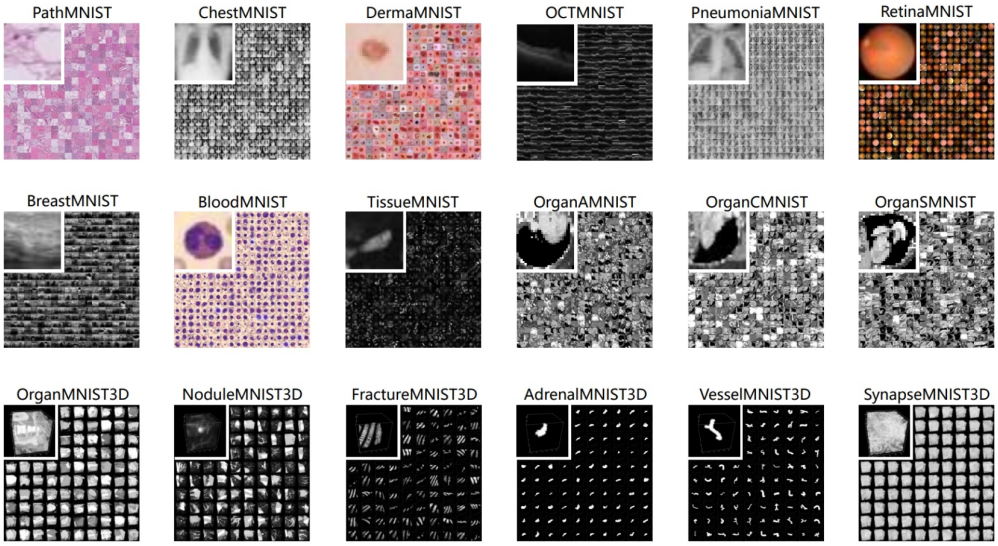

Figure 2: Example of images in MedMNIST dataset. Source: https://medmnist.com/

3.2. CL architecture

The methodology of continuous learning algorithms would be used in our study to handle both task-incremental difficulties (since it manages various types of images with various protocol acquisitions) and class-incremental problems (because we have many new classes for each experience). Nevertheless, for a first evaluation, even though each dataset relates to a different task, it is possible to think of it as a class-incremental scenario study since practically all tasks are a multi-class problem. Hence, it is as though we observed a single dataset with multiple classes each corresponding to a different pathology.

- *Naive*: it consists in applying a backpropagation algorithm every time a new stream of data is available [17]. It represents the lower bound from a CL point of view.
- *Replay*: random selection from the historical experience data; the Random Memory (RM) size is fixed, and filled with previous data. Keep roughly an equal number of instances for experience, and replace examples randomly [18].
- *CWR**: designed for the fully connected linear classifier (and perhaps extended to several layers); using two distinct memory systems, one for memory consolidation and the other for improved plasticity; very straightforward and effective approach, regardless of the circumstance or experience content (NI, NC, NIC) [19].
- *ICaRL*: a hybrid continual learning strategy. Combining the replay and regularization methods; distillation for the regularization of representation learning (feature extractor); template matching with the closest prototype (as a classifier); and more sophisticated example management through herding. However, this method is difficult to scale and has inefficient example management with large memory sizes [20].
- *Cumulative*: for every experience, store all data and re-train from scratch [17]. It represents the upper bound from a CL point of view.

3.3. Training phase

For an objective assessment of the CL approach employed throughout the model's training phase, we solely used 2D images. According to the work in [16], we made use of ResNet-18 [21] to identify various pathologies for each dataset because it performs well enough in average w.r.t. the other models for 28x28 images that were examined in the research. In particular, we focused on the following data sets: Colon Pathology, Dermatoscope, Retinal OCT, Blood Cell Microscope and Kidney Cortex Microscope [16].

We applied the same model to all datasets, as we first aimed at assessing whether an approach based on continuous learning techniques can be applied to the context of medical images.

The images are first pre-processed with standard procedures, and then transformed to RGB and normalized so that all images could be processed by the same neural network.

All experiments have been performed on a GNU/Linux machine equipped with a NVIDIA Quadro P6000 GPU with 24 GB of RAM and up to 250W. The training was performed with the same hyper-parameters for all approaches, except for the ICaRL (that follows the parameters proposed in the original paper [20]. Specifically, in order to subdivide the dataset into a training-validation-test set, we followed the split ratio provided in [16] that differs for each dataset; for example, for the Dermatoscope dataset, the authors suggested a 7:1:2 (training:validation:testing) split ratio (please refer to [16] for more detail on the other dataset distributions).

Then, we set the learning rate (lr) equal to 0.0001, the mini-batch size to 64 images, and the number of epochs equal to 20; indeed, considering that the images are small (28x28) the model was capable to achieve good performance (over 90% for every single experience/stream of data) in a small number of epochs. According to the CL principles and strategies, we used a small set of pathologies for each experience.

The implementation of CL algorithms was made by the use of Avalanche, a novel framework released by ContinualAI [22].

3.4. Evaluation metrics

We assessed the ability of continuous learning to improve performance on previously seen domains by adding new domains backward transfer (BWT) [23]. BWT measures how learning a new domain affects performance on prior tasks. Avoiding negative BWT is especially crucial for CL since negative BWT values signify catastrophic forgetting.

Mathematically the BWT is defined as,

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} R_{T,i} - R_{i,i}$$

where $R_{i,i}$ is the test classification accuracy of the model on task t_i after observing the last sample from task t_i and $R_{T,i}$ is the test performance on all T tasks.

By definition, it is possible to evaluate the forgetting of a neural model as $Forgetting = -BWT$, therefore lower the measure better the performance. In fact, this measure shows the capability of a Neural Network to maintain knowledge about previous experiences.

Additionally, we used also the top-1 accuracy metrics, in order to make a consideration on the overall performance of the model.

Table 1
Results obtained using different CL strategies. The approaches are evaluated in terms of BWT and Top-1 Accuracy.

	Naive	Replay	CWR*	ICaRL	Cumulative
BWT	-0.9247	-0.7305	-0.1928	-0.2465	0
Top-1 Acc. avg.	0.2334	0.3790	0.5431	0.5154	0.6285

4. Results and discussion

In this section, we discuss the experimental results and effectiveness of the continuous learning methodologies herein considered.

One of the crucial issues is how to establish boundaries for continuous learning methodologies. Hence, to the best of our knowledge, and according to the survey on CL approaches provided in [7] the current continuous learning approaches still fall short of matching the performance of a group of models (or a single model, depending on the problem to be solved) which is trained with all data. The cumulative approach refers to training a model with all the data at once. One of the goals of the CL method is to match and, if feasible, even outperform the performance of a model that has been trained using all the data.

However, it is also possible to establish a lower bound, which is illustrated by the so-called Naive technique, which uses back-propagation without considering a new domain.

The BWT and accuracy of the models, which are two helpful measures for assessing the CL performance from the initial stage, are shown in Table 1.

We started the performance evaluation of the models from the Cumulative approach, pointing out that compared to the work in [16] we have obtained lower accuracy (i.e. 0.63 Top-1 Accuracy average): this is mainly due to the fact that performance of a single model responsible for the classification of all the different pathologies is lower than the performance of multiple models that focus on the classification of pathologies only related to a specific dataset at each time. However, as previously mentioned, the main goal of this work was to compare the CL models with a possible (upper) limit case to identify the best approach usable in medical imaging; therefore, our primary objective is not to obtain a model that could outperform the state-of-the-art.

CL approaches can be evaluated according to the accuracy and forgetting value.

In the following, we propose a detailed analysis of the different strategies of CL. In spite of the fact that they are widely used in the literature, in our experiments random replay's accuracy and BWT results look inadequate, scoring 0.38 and -0.73, respectively. This could be caused by the unbalanced distribution of images across classes and a high number of images. Another motivation could be the memory capacity limit of the stored data, which was equal to 1500 while in [18] appears to be the limit above which there is no real advantage compared to the problems of an excessive cost of storage and privacy preserve.

Thus, it becomes evident that remembering only a few examples of previous experiences is insufficient in a scenario where the number of images per dataset is relevant, or huge. Instead, we found promising results for both CWR* and ICaRL according to each metric taken into account. Indeed, CWR* achieved the best results (i.e., Top-1 Accuracy average of 0.54 and BWT

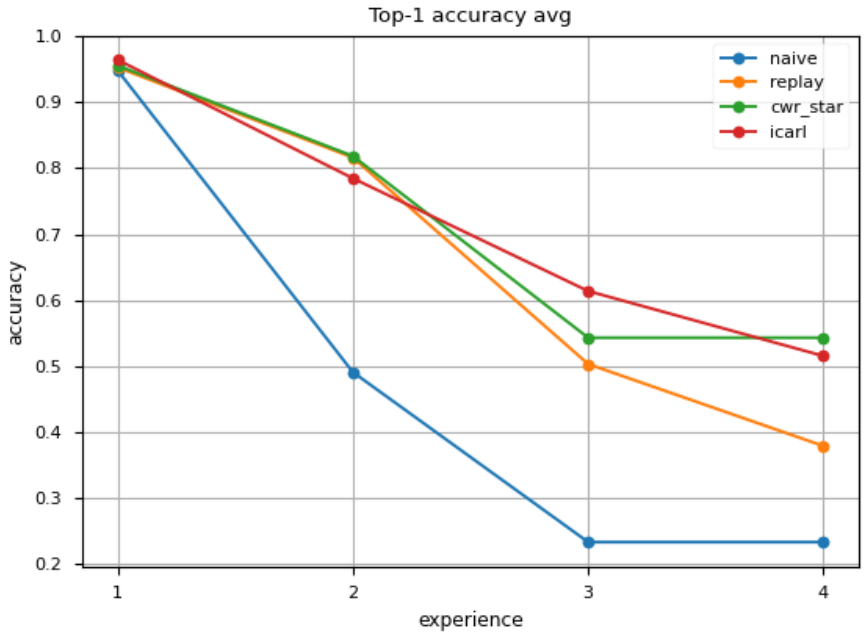

Figure 3: The figure shows the decrease in performance in terms of accuracy of different CL approaches against the increasing number of experiences. This is due to the fact that for every new experience the model has to recognize the new and older classes. Nevertheless, the figure does not show the cumulative, as, in our study, cumulative is not evaluated in terms of different experiences (all data sets are available at the same time).

of −0.19), even if slightly below the Cumulative strategy.

Figure 3 illustrates how the model performance degrades over experiences due to catastrophic forgetting, which is anyway mitigated by the CL strategies. It is worth noting that the cumulative strategy was intentionally omitted from the plot, given that, in this study, it was assessed towards a model that used all available datasets simultaneously instead of storing data from previous experiences and re-train the model from scratch. In a nutshell, with the cumulative strategy the model received all data together, in a sort of "single big experience"; interestingly, as shown in Table 1, it outperforms other strategies in accuracy. This was made possible since the problem can be traced back to a class-incremental problem.

5. Conclusion

In this work, we presented a comparison of different CL strategies used to support neural networks during the classification of medical images. We reported the results of an ad-hoc experimental analysis showing that the best results are obtained using CWR* and ICaRL. Although the results are slightly lower than the upper bound (i.e. cumulative strategy), the overall performance results are promising. Hence, our approach, featuring a comparative evaluation of CL means, can be of help to support the proper choice and use of these methods

in medical imaging.

As future work is concerned, we plan to include other metrics in our experimental comparison (e.g., forward transfer) and to evaluate the variation of the training time of different approaches.

Acknowledgements

This work has been partially funded by PON "Ricerca e Innovazione" 2014-2020, CUP: H25F21001230004.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.

[3] M. H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, Journal of digital imaging 32 (2019) 582–596.

[4] C. S. Lee, A. Y. Lee, Clinical applications of continual learning machine learning, The Lancet Digital Health 2 (2020) e279–e281.

[5] R. M. French, Catastrophic forgetting in connectionist networks, Trends in cognitive sciences 3 (1999) 128–135.

[6] S. T. Grossberg, Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control, volume 70, Springer Science & Business Media, 2012.

[7] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Information fusion 58 (2020) 52–68.

[8] C. Baweja, B. Glocker, K. Kamnitsas, Towards continual learning in medical imaging, arXiv preprint arXiv:1811.02496 (2018).

[9] 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, Proceedings of the national academy of sciences 114 (2017) 3521–3526.

[10] C. Wachinger, Importance driven continual learning for segmentation across domains, in: Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings, volume 12436, Springer Nature, 2020, p. 423.

[11] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, T. Tuytelaars, Memory aware synapses: Learning what (not) to forget, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 139–154.

[12] C. Gonzalez, G. Sakas, A. Mukhopadhyay, What is wrong with continual learning in medical image segmentation?, arXiv preprint arXiv:2010.11008 (2020).

[13] Z. Li, D. Hoiem, Learning without forgetting, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 2935–2947.
[14] N. Karani, K. Chaitanya, C. Baumgartner, E. Konukoglu, A lifelong learning approach to brain mr segmentation across scanners and protocols, 2018.
[15] M. Perkonigg, J. Hofmanninger, C. J. Herold, J. A. Brink, O. Pianykh, H. Prosch, G. Langs, Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging, Nature Communications 12 (2021) 1–12.
[16] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification, arXiv preprint arXiv:2110.14795 (2021).
[17] K.-H. Thung, C.-Y. Wee, A brief review on multi-task learning, Multimedia Tools and Applications 77 (2018) 29705–29725.
[18] L. Pellegrini, G. Graffieti, V. Lomonaco, D. Maltoni, Latent replay for real-time continual learning, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 10203–10209.
[19] V. Lomonaco, D. Maltoni, L. Pellegrini, Rehearsal-free continual learning over small non-iid batches., in: CVPR Workshops, volume 1, 2020, p. 3.
[20] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, icarl: Incremental classifier and representation learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[22] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. L. Hayes, M. De Lange, M. Masana, J. Pomponi, G. M. Van de Ven, et al., Avalanche: an end-to-end library for continual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3600–3610.
[23] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, Advances in neural information processing systems 30 (2017).