What are the most effective strategies to improve incremental learning performance for image classification in neural networks?

Student Name: Finlay Boyle
Supervisor Name: Dr. Donald Sturgeon
Submitted as part of the degree of MSci Mathematics and Computer Science to the Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—The catastrophic forgetting phenomenon remains a severe challenge for neural networks. Continual learning techniques aim to overcome this phenomenon by providing an alternative to the traditional offline training methods with the objective of overcoming catastrophic forgetting and facilitating incremental learning. This project explores the different solution paradigms within continual learning with the aim to evaluate and compare the existing techniques when applied to the domain of image classification and optical character recognition. This project aims to provide a unified comparison of the state of the art approaches used to achieve continual learning and evaluate the progress made towards attaining similar results to offline training.

Index Terms—Computer vision, machine learning, optical character recognition, performance evaluation

1 Introduction

Training a neural network typically involves having an entire dataset available. This dataset is split into batches and repeatedly used to train the network. Collectively, these techniques are known as offline learning. This is not always feasible or ideal; it is common that the data that a neural network is being trained on, or the task that it has been designed for, is not static. As such, some neural networks need to be able to learn continually in response to the changing data distribution and update the hyperparameters of the network accordingly.

Neural networks draw inspiration from the brains of animals, typically humans, to model the network based on how neurons interact in the brain [1]. However, implementations of neural networks often ignore that humans are capable of learning continually throughout their lifetime. For continual learning it is important to retain existing knowledge whilst being able to learn new knowledge and adapt to changes in the training data. This is difficult to achieve using standard offline training methods due to a phenomenon known as catastrophic forgetting.

Catastrophic forgetting occurs when a neural network is trained to solve one task and then later the same neural network is trained on new data causing the model to forget the knowledge it had acquired in the previous training phase [2]. This is perhaps an unsurprising result as the existing weights of the neural networks can be viewed as an initialisation rather than knowledge it has already gained. Offline training techniques do not account for existing knowledge stored in the neural network and thus retraining the neural network causes catastrophic forgetting. Continual learning, also known as incremental learning, aims to overcome this problem by proposing different approaches to training which are designed to help the network to avoid forgetting the knowledge that it has previously acquired.

Being able to train a neural network in an online manner (where the whole dataset is not available immediately) is important. This is because training neural networks is computationally expensive and time-consuming with high-quality models typically taking many hours to train at a minimum with the best models, such as Microsoft and Nvidia's Megatron-Turing Natural Language Generator [3], costing millions of dollars [4]. Hence, if high-quality models want to retain their success when new data becomes available or the distribution of the data changes over time, this change in the underlying data needs to be incorporated into the network. Using offline training methods this would require a complete retraining of the network despite it having already acquired a significant portion of the knowledge. If a full-network retrain can be avoided then this can extensively save on the computational and time costs.

Retaining knowledge and learning new knowledge presents a trade-off in biology as well as in neural networks, this is known as the stability-plasticity dilemma. Stability is associated with retaining knowledge while plasticity is associated with learning new knowledge [5]. For humans, the brain is initially heavily tipped towards plasticity as we learn new knowledge while growing up. Later in life, the brain shifts towards stability but the ability to learn new knowledge is still present. In artificial neural networks, the learning rate of optimisers do partially reflect this trade-off although this affects the whole network rather than the individual neurons making it difficult to retain knowledge while also maintaining the ability to learn new knowledge.

Continual learning offers many benefits in terms of reducing the costs associated with training a neural network from scratch. It also offers real benefits to the quality of the neural network. One key area that incremental learning could be beneficial for is computer vision tasks, namely im-

age classification and optical character recognition. Ideally, image classification needs to be able to adapt to changing input in order to classify new objects correctly and identify when it has not seen objects similar to those that it is classifying. This is not possible using offline training, if new objects are presented then the accuracy of the network would decrease due to misclassification rather than have the potential to adapt. Incremental learning has potential to support this type of behaviour by continuing to train the network on incoming data.

For optical character recognition, being able to adapt to handwriting on the fly is important especially if the handwriting being provided as input is not from the same person or if the person's handwriting changes over time but still needs to be classified by the same model. The goal of incremental learning in this case is for the neural network to adapt to the differences while attempting to retain a reasonable rate of success. In addition to this, optical character recognition has applications in extracting text from historic documents. For example, in digital humanities, extracting, parsing, and transcribing text from dead or extinct languages may be a difficult or tedious task for humans to do directly from images but something that machine learning can assist with. Specifically, consider the Kuzushiji writing style which was used in Japan from around the 8th century up until the start of the 1900s when it was phased out in favour of the modern Japanese writing style [6]. This style of writing is no longer used but many potentially historically significant books and documents are written in this style making it a worthwhile task to attempt to transcribe this rich literature.

1.1 Setup for Continual Learning

The setup for the continual learning scenario is important. Data is assumed to arrive sequentially in a continuous stream and the amount of unseen data is unknown - and could potentially be infinite (e.g. real-time data). This is key because otherwise it would be possible to retain all of the data and follow traditional learning techniques instead. However, this is infeasible (as there may be a massive amount of data yet to arrive) and waiting for the data would sacrifice the benefits of training continuously on the fly to support and classify the rest of the data that is unseen as it arrives.

The structure of the data is also important to consider. There are two key setups: Task-Incremental (Task-IL) and Class-Incremental (Class-IL) [7]. In Task-IL, data is segmented into predefined blocks with disjoint, limited and known class labels. In Class-IL, data arrives randomly and there are no limitations on the class labels and the corresponding data. Task-IL is less challenging and less realistic than Class-IL because in real applications of continual learning the data that arrives continuously to the neural network is unlikely to be disjoint or controllable. As a result of this, the applications of Task-IL are limited compared to Class-IL which is able to handle the inherent randomness of the sequential data and does not impose unrealistic and restrictive constraints on the data format or its underlying distribution.

Within this project I intend to evaluate high-performance, quality strategies for applying continual learning to neural networks in the domain of image classification and optical character recognition. This will be achieved by comparing historically significant papers, in the context of continual learning, as well as recent state of the art to determine the most viable strategies. The project will primarily focus on two attributes of continual learning algorithms: computational efficiency - as previously discussed this is a large benefit to using these training methods compared to offline training - and classification success.

In order to draw useful conclusions in the context of this project I will compare the continual learning methods on both standard datasets from the literature and specific digital humanities datasets with the focus on optical character recognition such as the Kuzushiji style of writing previously discussed to assist with recovery of historic documents.

2 Related Work

There has been significant research into overcoming catastrophic forgetting and applying continual learning recently with a variety of different solution paradigms emerging. The research into continual learning primarily focuses on how to train the network rather than the exact architecture of the networks. The architecture is important to ensuring that a neural network is successful in the task that it was designed to fulfil but the priority within this field is the training of models rather than the design of their architecture. As a result, researchers have typically used tried and tested neural network architectures such as ResNet [8] or AlexNet [9] as they have been shown to have great success in the offline training format for image classification. Using these architectures ensures that the literature focuses on the learning process rather than the network design.

2.1 Solution Paradigms

2.1.1 Regularisation

Regularisation approaches focus on the stability-plasticity dilemma by attempting to control the weights in the network and their individual rates of change. The general premise is to avoid drastically altering weights which are important to previously learnt knowledge by penalising weight changes through loss functions [10]. They also balance this against the plasticity of the network to ensure that it retains capacity to continue to learn more from the incoming data.

One of the first approaches to tackle the continual learning problem was Elastic Weight Consolidation (EWC). This uses a regularisation approach to prevent samples from other classes from drastically affecting weights that are important to previously learnt knowledge, thereby helping the network to consolidate previously learnt knowledge [11]. It does this through the Fisher information matrix which is used to compute the importance of weights. This paper was incredibly influential and presented a lot of the groundwork for the continual learning domain as it was the first paper to truly tackle the problem of online continual learning. It was compared against finetuning, a naive online training approach that does not attempt to prevent catastrophic forgetting, and successfully outperformed it. As EWC was the first real attempt to solve this problem it could only

be compared against finetuning and it is often used as a benchmark in the later literature.

Memory aware synapses is another regularisation technique based on the ideas of Hebbian Learning [12] which is a concept that 'cells that fire together, wire together' in biology. This paper argues that due to limited model capacity it may be necessary to erase rarely used knowledge from the network to create space for new knowledge. It does this by calculating how important parameters are to the network and preventing the overwriting of those that are most important using loss functions [13]. The important distinction for this paper compared to previous regularisation techniques was the potential for it to be used on unlabelled data in an unsupervised manner while retaining competitive performance. The results for this method were approximately the same as EWC.

Learning without forgetting (LwF) is another influential continual learning method. In this method, the network has a set of shared parameters and each task has a set of specific parameters [14]. It is a hybrid approach between finetuning the network and using regularisation. The shared parameters and other task-specific parameters are frozen while the new task parameters are warmed up using the incoming data and then the network is jointly optimised across all sets of parameters using regularisation to prevent the overwriting of previous knowledge. As this was one of the first methods, the results are primarily compared to finetuning which it outperforms. Significantly, it was the first paper to use more complex datasets for evaluation such as ImageNet laying the foundations for future literature.

Ultimately, regularisation approaches have become less prevalent in the literature due to their failure to address the class incremental scenario and their poor performance as the number of classes to identify in the problem increases compared to memory-based approaches.

2.1.2 Memory Based

Memory-based approaches are a popular paradigm in continual learning and they have been the most active area of research within the field recently. As the algorithms for continual learning cannot store the entirety of the incoming data, this class of approaches allocates a significantly smaller amount of memory to store samples which are often referred to as exemplars. In general, these stored samples are then used to remind the network of previously seen knowledge to prevent decay of previously acquired knowledge. How the exemplars are selected, the frequency of memory replay, and how to effectively store exemplars, or representations thereof, continue to be heavily researched.

Increasing the diversity of exemplars available is a key focus in the literature. This focuses on the strategy used to draw samples from the incoming data to store as exemplars in order to better reflect the structure of the classes in the data and provide greater definition of the class boundaries.

The algorithm iCaRL was the first prominent sampling strategy, it focused on prioritised exemplar selection [15]. This is where it manages its exemplars that are stored using a technique called herding which is where exemplars are removed in a fixed order such that each exemplar is itself a near approximation of the mean of the set so there is low variation [16]. During training, this method also learns a feature map allowing it to represent images as vectors of a lower dimension. To classify samples the model then uses the nearest-mean-of-exemplars rule which is where it computes the average feature vector for each class using the exemplars stored in memory and the feature vector for the sample to classify. It then calculates which average class feature vector the sample feature vector is nearest to using a norm and classifies the sample accordingly. The results of the paper are competitive and this method appears regularly in other literature within the field to be compared against. It outperformed finetuning by a factor of 5 and learning without forgetting [14] by a factor of 2.

Gradient-based sample selection uses a different approach, it treats the selection of exemplars as a constrained optimisation problem which is useful when the data is imbalanced or task boundaries are unclear [17]. It aims to optimise the loss on the current sample without negatively impacting the loss on the previously stored exemplars. These requirements impose a set of constraints on the optimisation of the current sample and the goal is to solve this optimisation problem. This approach was not tested on complex datasets, only CIFAR-10 and MNIST, and it was of the simpler Task-IL formulation. It also did not compare itself to state of the art papers making it difficult to draw any significant conclusions or assess its impact on the literature and thus it is difficult to evaluate its effectiveness despite offering an interesting alternative sampling strategy.

Mnemonics training is a state of the art technique that focuses on selecting exemplars via solving an optimisation problem [18]. This extracts exemplars automatically, called mnemonics, which are optimised by stochastic gradient descent as more samples arrive. This is achieved by formulating this approach as a Bilevel Optimisation Problem [19] which is where there are two models that are optimised in alternating phases (i.e. the classification model and exemplar optimisation model). It proposes this strategy as an alternative to herding, which was used in iCaRL [15]. The paper demonstrated that this method finds exemplars on the boundaries of classes intrinsically allowing for greater separation which helps consolidate the knowledge in the network. It performed well on CIFAR-100 and ImageNet in terms of both forgetting rate and average accuracy where it outperformed iCaRL by approximately 10% as well as other state of the art models.

GDumb is a recent paper that presents a far simpler approach than previous papers. The authors propose a technique consisting of a greedy sampler, that aims to keep the number of exemplars in memory balanced between classes, and a dumb learner that is trained from scratch at inference time [7]. The purpose of this naive approach to sampling is to highlight flaws in the literature in continual learning. Despite the more complex approaches taken in previously discussed techniques, GDumb outperforms iCaRL [15] and other state of the art techniques by significant margins. One of the main concerns that this paper raised was the impractical and unrealistic constraints used to formulate the continual learning problem. This relates back to the task incremental vs class incremental dilemma with many of the previously published papers focusing on the simpler task incremental version. While this technique does have poor wall-clock time performance (as it is repeatedly training a

network from scratch) it is an important proof of concept that existing algorithms are perhaps overly complicated and poorly designed with misleading results; it raises important questions regarding the direction of the field.

Another recently successful sampling strategy is to select samples that are representative of their class as well as discriminative against the other classes in the dataset. This was achieved by taking samples judged to be near the centre of their class as well as samples that are judged to be near the class boundaries using a technique known as Rainbow [20]. The aim of this strategy is to define the boundaries of the class allowing samples to be accurately classified. This strategy was tested on simple datasets such as MNIST and CIFAR-10 as well as more complex datasets such as CIFAR-100 and ImageNet. It outperformed both iCaRL and EWC significantly as well as more recent memory-based methods such as GDumb [7] with over twice the accuracy on ImageNet.

Game theory has also served as inspiration for a sampling technique. Adversarial Shapley value experience replay (ASER) draws on cooperative game theory to estimate how important exemplars are to the overall performance of the continual learning procedure [21]. The idea is to avoid blurring the boundaries between classes by carefully selecting exemplars to reduce interference at the boundaries of other classes while defining the boundaries of their own class. The Shapley value from game theory is used in this method as a measure of the contribution of each exemplar to the task. The technique is fairly effective and when tested on CIFAR-100 and Mini-ImageNet it performs competitively but it did not make meaningful improvements over other memory-based approaches.

In addition to different exemplar selection strategies, some papers take approaches to enhance the performance of existing ideas. One such idea was using an experience replay with a review step. Prior to the final testing of the model, the network is reminded of knowledge that it has acquired by completing a pass over the network using the stored exemplars [22]. This leads to minor, but effective, improvements that can boost other techniques and it was successfully used to win the CVPR2020 continual learning challenge.

Another idea to enhance existing memory-based techniques is compressing exemplars via product quantisation which reduces the exemplars into their hidden representations from the neural network [23]. This is effective at allowing a greater number of exemplars to be stored in the same amount of memory as other memory-based techniques while preserving the effectiveness of memory replay techniques. This paper was able to outperform many existing methods on both the ImageNet and CORe50 datasets when using the same amount of memory suggesting that this type of compression technique is effective both at maintaining the quality of the exemplars and increasing the number of exemplars stored.

An alternative approach to storing exemplars explicitly is to use generative networks such as generative adversarial networks (GANs) [24]. The purpose of this is to reduce the memory footprint of the algorithms since they are storing weights representing the exemplars instead. Unsurprisingly, there exists significant drawbacks to these approaches because they do not store complete representations and thus introduce an additional layer of error into the networks since they have to produce outputs from their generative networks. In addition to this, these algorithms inherit issues surrounding generative networks such as the need for a large amount of training data, mode collapse, and the vanishing gradient problem [25]. Generating convincing image samples using a neural network is itself a difficult problem and significant resources and research effort has been focused on this area producing state of the art such as Nvidia's StyleGAN2 [26]. This also raises the question: does using generative models shift the difficulty of continual learning on to generating convincing fake samples? This is explored by the literature.

One of the first generative approaches to the continual learning problem was to use a GAN which was trained simultaneously. In this approach, the authors define a scholar as a set of two models, the GAN (consisting of a generator and a discriminator) and a solver (which aims to classify the samples) [27]. The scholars are trained sequentially in the continual learning format. The GAN is trained on the incoming data whereas the solver is trained on both the incoming data and the output from the GAN. The main contribution of this paper was the proof of concept that GANs were a potential, viable avenue for further exploration. Whilst this paper only tested the model on MNIST it was still a significant milestone and showed that GANs could potentially compete against pure exemplar methods.

FearNet is an innovative algorithm which uses a dual-memory system inspired by the human brain. It uses a long-term and short-term memory represented by neural networks and a third network, the selector network, which predicts whether a class is in long-term or short-term memory [28]. The aim of the selector network is to predict the probability that the long-term memory contains the correct class to classify the incoming sample. Samples are consolidated during sleep phases where data is consolidated from the short-term memory to the long-term memory and the selector network is updated to reflect these changes. To consolidate the data, FearNet uses a generative model since it does not store samples but rather representations of them in the short and long-term networks. This algorithm was evaluated on CIFAR-100 and performed well against state of the art at the time of publication such as iCaRL (which it outperformed by approximately 10% throughout the duration of the training in terms of accuracy) and used significantly less memory than other compared models due to the generative approach. It attained performance within 90% of offline training methods.

An intermediary approach between storing samples explicitly and attempting to generate samples from a separate neural network is to generate feature representations. ACAE-REMIND is an algorithm which does this by compressing samples using an auto-encoder into feature representations to reduce the dimensionality of the samples which leads to around a 50x reduction in memory consumption [29]. While this algorithm did not convincingly outperform previously discussed methods it is capable of reducing memory consumption while alleviating some of the concerns around using generative models. This approach has also been seen in other papers which focus on using

generative feature replays such as in [30] where they are able to obtain similar high-quality results.

Maximally interfered retrieval is another technique which aims to replay exemplars that will be most negatively affected by the parameter updates caused by the incoming samples [31]. It can be used either with a regular memory buffer or a generative model (or a combination of both using an offline trained auto-encoder) to produce samples that can be replayed through the network to minimise the interference caused by the incoming data. The results suggest it was successful on MNIST but struggles on CIFAR-10 and highlights the previously mentioned issues that can arise when using a generative model as the authors deemed the approach non-viable for CIFAR-10 whereas using a memory buffer continues to be successful. This is significant because CIFAR-10 is an easier datasets compared to CIFAR-100 and ImageNet and thus suggests there are significant limitations in generative approaches for continual learning.

Another proposed generative replay approach is to combine the main network with the generator through the use of context which is inspired by the human brain where context is important for recall, this can be modelled via feedback connections in the network [32]. This allows knowledge to be consolidated by reactivating neurons to remind the network of the samples that it has seen. This is a form of feature replay since the model is not fully generating samples but rather replaying extracted features with context. This technique was able to outperform regular generative approaches on CIFAR-100 although critically it did not include a substantial comparison against regular replay methods making it difficult to gauge the success in the context of the wider literature.

Some papers also combine techniques from across machine learning. One such paper uses the idea of a meta-experience replay which combines techniques from meta-learning (which is where the training procedure is learnt over time instead of being fixed [33]) with a memory-based approach. The aim is to reduce interference between classes in the gradients (i.e. will learning about a new sample negatively affect a sample already learnt about) while aiming to increase transfer (i.e. will learning about a new sample positively affect a sample already learnt about) [34]. This leads to improvements in memory consumption but the results were not conclusive and it was not tested on complex datasets.

Overall, memory-based approaches show significant promise for tackling the continual learning problem but there still exists outstanding issues within the literature predominantly surrounding the direction of literature as highlighted in [7] and the impact that early papers had as they were focused on the simpler task incremental scenario with low complexity datasets. The literature continues to move in a positive direction and has many avenues for potential solutions but one key problem is the lack of comparability between the existing state of the art which impedes decision making around the success of different approaches within this paradigm.

2.1.3 Other Approaches

Outside of the two primary solution paradigms, regularisation and memory-based techniques, are approaches that apply other abstract ideas. For example, architectural approaches consider ways to dynamically change the structure of the neural network. These approaches are rare in the literature but they do provide an interesting avenue to explore.

Drift compensation is a technique that utilises embedding networks. These are networks that are able to map data into a lower dimension where simple metrics such as L2-norm can be used to compute similarities between embedding representations [35]. Instead of preventing the drift of classes, which is where they move around in the embedding space, this method aims to compensate for the drift by estimating how much each class has drifted and accounting for it during classification. This technique was shown to outperform iCaRL [15] on CIFAR-100 and a subset of ImageNet but it was not compared to the latest state of the art memory-based methods making it difficult to fully evaluate its performance.

LUCIR is another technique consisting of three main components: cosine normalisation for the probabilities in the final layer of classification (instead of the traditional softmax layer), the 'less-forget constraint' which is a regularisation constraint that considers the position of previously acquired knowledge against the new data in an embedding space, and inter-class separation which is a ranking loss that compares previous samples (used as anchoring points) against the new samples to attempt to separate out classes [36]. This technique outperforms iCaRL which it is primarily compared to on CIFAR-100, however it only considers the accuracy of the best class which does not provide the whole picture since it is important for a classifier to classify all (or at least most) classes well rather than just a small fraction of them.

Laplace operator based node-importance dynamic architecture (LNIDA) is a hybrid approach that combines a dynamically changing architecture with a regularisation-based technique to evaluate the importance of nodes in the network [37]. This applies the Laplace operator (also known as the Laplacian) to the loss function computed at a specific node. It is used to measure how much a specific nodes affects and interferes with other nodes around it. The architecture of the network is dynamically updated using this information. After learning for a fixed period of time the interfering connections are removed and inconsequential nodes are reinitialised to attempt to restore their value to the network. This technique was evaluated on a few datasets including the CIFAR-100. It performed well for small numbers of classes but it struggled as the number of classes increased and its results were similar to EWC. It was not compared to memory-based solutions making it difficult to evaluate its true effectiveness although as it struggled to outcompete EWC it is unlikely it would have challenged the state of the art memory-based methods such as mnemonics training [18].

2.2 Benchmarking

It is important to be able to compare methods described in the literature in order to evaluate their success in the continual learning domain. Early literature was plagued with issues regarding metrics. Many papers did not compare to state of the art or created their own metrics in

order to quantify the performance of their algorithms [10]. Gradually, key metrics have been identified to fairly compare different algorithms such as the average forgetting rate which is unique to online training methods and measures the drop in performance of the network caused by learning new classes [38]. Another important metric is the overall accuracy, this is typically measured on a per class basis as well as overall which is standard with respect to training neural networks.

Interestingly, there does exist a concrete lower bound and a soft upper bound for continual learning. Any continual learning algorithm should outperform finetuning. This is where we do not attempt to do anything special but instead just train the neural network on the incoming sample and then move on to the next one. As such, this leads to catastrophic forgetting because the network simply overwrites the existing knowledge. Unsurprisingly, this leads to exceptionally poor performance with low accuracy rates [39]. On the other end of the spectrum is offline training which is the collection of methods used to train typical neural networks where we have access to the whole dataset instead of receiving it in a sequential stream. As such, the performance of a neural network trained continually as samples arrive from the data stream can be compared against the performance of a neural network trained on the same data but in an offline, batched approach. This presents an upper bound as ultimately the goal of continual learning is to match or surpass the performance of typical neural network training procedures since these obtain high performance while being able to retain knowledge from multiple classes due to the batch training procedure [10].

In addition to the accuracy and forgetting metrics, for comparison of continual learning algorithms there are other important considerations. As many of the previously covered methods have a focus on storing exemplars, or representations thereof, it is important to consider other factors in the practicality of these approaches. Other important metrics in the literature are computational efficiency, memory consumption, and wall-clock time (how long it actually takes to run the training procedure). These must be considered when comparing methods because a method may have exceptional classification performance but require an unrealistic amount of memory or computation time to achieve these results. This risks nullifying the benefits of continual learning and thus it requires significant consideration.

2.3 Datasets

Datasets are important to assess how well the continual learning algorithms perform on increasingly complex tasks. Early literature tended to focus on using simple datasets such as CIFAR-10 [40] and MNIST [41] which are useful as toy examples but do not reflect the complexity of data from real scenarios, and the manifolds that they are subsets of, and thus are not true indicators of how these algorithms would perform in real world environments [7].

As the literature has progressed, it is clear that the trend is to use more complicated datasets to better represent and compare algorithms. This is highlighted when using the metrics previously discussed since an algorithm may have acceptable performance on CIFAR-10 or MNIST but when more realistic datasets are used their performance drops off. Clearly, this is important for realistic scenarios as it is unlikely that they are as simplistic as these easier datasets.

One commonly used dataset now is CORe50, this consists of 50 classes of images with varying lighting and occlusion making it more suited to realistic scenarios [42]. Another prominent dataset is CIFAR-100 which is an expanded version of CIFAR-10 and contains 100 classes (which are classified into 20 higher-level superclasses) of images each containing 600 images [40]. Finally, ImageNet (and Mini-ImageNet), which have been used for visual recognition challenges, is a dataset containing 1000 classes with over 1.3 million images [43]. This makes ImageNet the most challenging of the datasets but also the most realistic as it has been used for offline training to create some of the best classification models such as ResNet [8].

In addition to standard datasets, I will be evaluating the performance of the algorithms on the Kuzushiji dataset. This consists of characters from the Japanese Kuzushiji style of writing which was widely used in Japan from the 8th century until the 1900s [6]. It poses a significant challenge due to containing over 4300 character types in a long tailed distribution as some characters are rarely used. This makes it suitable for testing the application of incremental learning on optical character recognition as it provides enough classes to make meaningful comparisons between the efficacy of different solutions. It is important to use standard datasets on top of the Kuzushiji dataset so that the results that I find are comparable against future papers which are most likely to use the standard datasets.

2.4 Summary

In summary, there has been substantial research into the continual learning domain. There exists two prominent solution paradigms, namely memory-based and regularisation, as well as some other promising solutions outside of these. Memory-based solutions have moved to the forefront of the literature and there has been significant progress made. Concerns have been raised regarding the comparability of results [10] and around the complexity of solutions and whether the literature has been moving in the correct direction [7]. These concerns continue to persist as evidenced by the lack of comparisons made by proposed techniques and the existing state of the art as I have previously mentioned. A lack of comparability leads to confusion and difficulty in evaluating the progress made in the continual learning field. The trends of the literature are promising in some respects as the datasets used in the literature have been trending towards more complex, realistic examples and there is a strong chance that the literature will continue to improve our understanding of how to overcome catastrophic forgetting.

3 Methodology

The primary focus of this project is to create a balanced, informative comparison of strategies for incremental learning in image classification and optical character recognition. As outlined in the related works, the literature in this field is fragmented and proposed techniques are not always fairly

compared to existing techniques. This makes it difficult to separate out high-quality techniques that are effective in real scenarios and low-quality techniques that are only effective in carefully constructed toy examples.

3.1 Comparing Approaches

In order to compare approaches it is essential that the techniques are implemented in a standardised way. If a technique that is implemented in a favourable way is compared to other techniques that are not then this will nullify the purpose and quality of the comparison being conducted. PyTorch [44] is a common machine learning framework that provides the necessary tools to implement neural networks and train them efficiently. As such, I will use this to implement the algorithms for comparison. Alternatively, a recent framework called Avalanche [45] is a purpose built incremental learning framework that has shown promise. It is not currently widely used in the literature as it has only recently been released but it does offer a codebase to construct incremental learning algorithms on making it a potentially viable framework to assist in the implementation of the different techniques.

In order to have a comprehensive evaluation of the different incremental learning strategies it is important to include the baselines previously discussed: finetuning (where samples are directly passed to the network and it trains continuously as each sample arrives) and offline training (using the whole dataset and splitting it into batches). It is expected that the results from finetuning will provide an informative lower bound, and offline training - which is traditionally used for training neural networks - will act as the tentative upper bound for incremental learning techniques.

As well as these baselines, GDumb [7] is an important technique to include because it highlights the difficulties and inconsistencies in the existing literature while also providing a practical method that can be used for comparisons. Their proposed technique was relatively primitive but capable of outperforming other more complex techniques. As such, I will include their proposed technique in my evaluation of the different strategies because effective continual learning techniques should be able to outperform it as suggested by GDumb's original authors. Further to this, I will include an implementation of iCaRL [15] for historical reasons because many papers in the literature, both old and new, use this as a target for comparison. By comparing techniques to iCaRL it will serve as a boundary between the effectiveness of older literature and recent state of the art.

Memory-based replay methods have shown significant promise in the literature and it is important to include these in my comparison as they could prove to be the most effective solutions. As discussed previously, much of the literature for memory-based methods focuses on the techniques used to select the exemplars. Hence, I will include state of the art methods with different sampling techniques. The specific papers that are state of the art, and that I will implement, are Rainbow [20], and Mnemonics [18]. I have selected these (on top of GDumb and iCaRL) because they are recent state of the art techniques that have shown excellent results.

Furthermore, generative replay methods have become increasingly popular in the recent literature. Although they do not appear to be fully fledged out yet, I will include these in my comparison to compare the progress made in this sub-paradigm. Feature-replay methods, which are an intermediate between pure memory-based and pure generative replay methods, were shown to be a successful compromise and could be promising, especially in datasets with many classes. Specifically, I will include ACAE-REMIND [29] which was successful at reducing the memory footprint of these types of memory-based approaches. Another technique that I will include is FearNet which was an interesting concept and had promising results that suggested it could be capable on datasets with many classes [28] making it suitable for complex optical character recognition and complex image classification datasets.

In addition to these methods, I will include regularisation techniques. They are still prominent in the literature (although they are primarily used for comparisons or in hybrid approaches) and they were responsible for providing initial effective solutions for incremental learning. Regularisation techniques have been eclipsed in the literature in favour of memory-based techniques and hybrid approaches. As such, I will include elastic weight consolidation [11] in order to highlight the progress that has been made with recent methods compared to the historical literature. It is important to have at least one technique from each class of solutions because they may thrive in different setups - especially with respect to the amount of memory available and the wall-clock time. For example, regularisation techniques may outperform memory-based solutions in low memory environments, such as on embedded devices, or where there are many classes and a fixed memory budget which limits the performance of memory-based solutions.

Architecture-based solutions have also shown some success recently. The need to adapt to the number of classes is less relevant for optical character recognition because the number of characters in an alphabet or writing style are often known in advance. However, not all characters in an alphabet are used equally and as such the most commonly used characters may dominate, especially if the data provided for training is heavily imbalanced as a result (e.g. E appears far more commonly than Z does in English). This is emphasised by Kuzushiji having over 4300 characters but a long-tailed frequency distribution suggesting some characters are used significantly more than others [6]. As a result, architecture-based solutions cannot be preemptively excluded from the comparison and it is important that they are included to provide a comprehensive overview of the state of the art and their application to this setting. There have been recent developments which I will aim to implement, namely LNIDA [37] which combines regularisation with dynamic architecture. This was not compared to memory-based approaches in the original paper so providing a direct comparison will be useful for the future.

In addition to these algorithms already detailed there will inevitably be further papers released throughout the course of 2022 and into 2023. I will continue to follow relevant conferences, such as CVPR 2022 and ICLR 2022, for upcoming research papers that present interesting ideas and form the latest state of the art.

It is also paramount to follow the scientific method when conducting comparisons by controlling as many variables as possible to reduce the influence of external factors on the results. This should be realistically achievable in this environment because, ultimately, I will have near-complete control over the implementations and thus be able to ensure fairness in the comparisons. One important factor is the network architecture, this is especially relevant for memory-based and regularisation methods. Each different technique will be tested on the same network architecture with ResNet [8] forming the base feature extraction layers where necessary and any additional layers will be dependent on the specific technique being implemented. It is important when conducting the comparison to ensure that the number of nodes is similar across techniques.

Further to this, the order of the sequence of the data that is fed to the networks will be controlled so that it is the same for each incremental learning technique to ensure that the order of the sequence of data does not impact the results. This is important because if classes are uniformly distributed as they arrive for one model but for another they are split into disjoint sets then this will severely impact the fairness of the comparison of the approaches used. This could also potentially be an avenue worth exploring in the future as the project progresses, or as future work within the domain, to investigate the performance of the models based on the distribution of the sequential data arriving for training. The experiments will also be repeated with different data orders in order to reduce the effect that this has on the results. The hyperparameters used for each technique will be those recommended by their corresponding paper except in the case of things such as memory budgets which will be varied throughout the comparison. Furthermore, the hardware that the model will run on will also be the same for each continual learning method implemented to ensure that each model has access to the same computational resources to ensure fairness.

In terms of data analysis the majority of data collected will be experimental data. This will be collated over a series of repeat runs to average out uncertainty in the results and to provide a measure of the range of error in the results. Using tools such as TensorBoard, it will be possible to store telemetry data related to the running of each algorithm. The flexibility this provides makes it suitable for storing data related to all metrics previously discussed. This will allow analysis of the accuracy over time during training which is particularly relevant for continual learning. These results will be quantitative data that can then be summarised and visualised graphically and compared in tabular formats against other algorithms allowing comparisons to be made as desired. Statistical analysis using, for example, hypothesis testing is unnecessary and it is non-standard in the literature to do so. As such, a more descriptive approach to comparison will be undertaken and supported by the quantitative results obtained.

For memory-based methods, extra consideration needs to be given to the constraints placed on the amount of memory that they can occupy. There are two primary approaches to constraining the memory: each class is allocated a fixed amount of memory or there is a fixed amount of memory for all the classes to share between them. In following with the suggestions of [7], it is important to avoid imposing assumptions that restrict the applicability of the incremental learning techniques to real-world environments. Thus, it is more realistic to have a fixed amount of memory for all the classes to share. This is especially important for the Kuzushiji dataset because there are over 4300 classes and if they each use a fixed amount of memory for their exemplars then this can quickly add up to a substantial amount of total memory used.

3.2 Evaluating

During the training process, data will be continually collected to evaluate the different approaches to continual learning. This is important because unlike traditional offline learning, the accuracy and forgetting rates of the network being trained during the training process is important because one of the core aims of continual learning is to retain strong classification results while simultaneously training on incoming data. By assessing the accuracy, and other metrics, during the training process this will provide a clear picture of the strengths and weaknesses of each of the different approaches.

In addition to this, evaluating the final performance of the online trained model is crucial. To do this, unseen data will be used - as is standard in machine learning - which will be a subset of the dataset used to train the model. This subset will be selected prior to training and will be kept the same for each different training approach to ensure fairness in the evaluation and comparison of the approaches.

Average accuracy is an important metric to use to compare techniques because it provides a summary of the whole model which is simple to calculate. This reflects the network's overall ability to correctly classify samples. In the literature, average accuracy is typically recorded throughout the training process as more and more classes are introduced in the sequential data stream for training which reflects the impact of the number of classes on the accuracy [10]. While it is a useful metric it can cause issues when considering imbalanced datasets such as those seen in optical character recognition. As such, top-5 error and top-5 accuracy are frequently used in the literature for incremental learning because of the importance that all classes are accurately classified. This provides greater insight into the training process and ensures a fairer comparison between techniques.

Another useful metric is average forgetting. Forgetting for a class is defined as the difference between the maximum amount of knowledge that the model has known about the class and the current knowledge that the model has about the class [38]. This difference represents the amount of knowledge forgotten by the network. The average forgetting is calculated as the average over the forgetting for each class. The aim of this metric is to measure the effects of catastrophic forgetting on the network so far.

As well as measuring the classification performance of the model it is also important to measure the resource consumption of a model. If a model is too slow or too memory-hungry then it may not be suitable for certain applications and this is important to quantify and consider when evaluating different methods. As such, one of the other metrics that I will use to compare the solutions is

wall-clock time which is a measure of the time taken for the model to train on all of the data arriving sequentially. For example, GDumb repeatedly retrains a fresh model on the exemplars [7] whereas EWC controls the weights of the model [11], as such EWC would expect to have better wall-clock time whereas GDumb has better classification performance overall.

Further to this, the memory consumption of the training is also important. The success of memory-based methods is dependent on the availability of memory space to store the exemplars, or representations of them, in. As a result, monitoring the memory usage of all the models is important for evaluation. For example, these methods may not be suitable for low memory environments such as embedded devices but may be best if a large amount of memory is available. This all needs to be considered when evaluating and comparing different methods against each other.

In the related works section I highlighted the distinction between the two main types of datasets: there are the toy examples, which are simpler datasets such as MNIST and CIFAR-10, and the realistic examples, which are typically run on CIFAR-100 and ImageNet. For initial testing purposes I will evaluate the approaches on MNIST and CIFAR-10 to verify that the implementations are working as intended. The results of this should reflect those found by the authors and this will help to highlight any flaws in the experimental setup prior to training on the complex datasets which will take greater resources.

After using the toy datasets, I will evaluate the approaches on both CIFAR-100 and ImageNet. These are complex datasets that better reflect realistic scenarios. The purpose of using these standard datasets for the initial comparison is due to the standardisation in the literature. The majority of the literature makes use of these datasets and it is most likely that future methods will also use these datasets. This will increase the validity of my results and add value to my project for the future, it is important that my results are useful in conjunction with future literature hence enabling comparison against future literature is essential.

Finally, I will use the Kuzushiji dataset to evaluate the performance of the continual learning techniques on a real scenario. This is an optical character recognition dataset which may present additional challenges to the continual approaches because of the long-tailed distribution of the classes in the dataset [6].

4 Validity

To ensure that the literature that I have read is relevant, I have included mainly papers from recent years (typically 2020 onwards) as well as papers of historical significance which outline the foundations of continual learning and offer some insight into early approaches and the direction that the field has taken. The most recent literature, some of which has been published this year, shows that the problem remains open and relevant providing concurrent validity for the problem. Further to this, I have used papers from reputable conferences and journals which ensures that I have covered the cutting edge of the field. The use of three major review papers [10], [46], and [39] have guided my research and ensure that I have surveyed the breadth of the field.

In order to ensure that my methodology produces valid results I will make use of metrics that have been defined previously in the existing literature and have been used by widely-cited review papers in the field. This will ensure that the metrics I am using are concurrently valid since the techniques will be compared using standardised metrics allowing direct comparison to existing algorithms and literature and thus allowing me to draw conclusions against current and future literature and allow others to trust my results. Further to this, there are a variety of benchmarks that have been used on multiple existing algorithms and algorithms that perform well in one metric typically perform well in others as seen in [10] which supports the predictive validity of these.

The metrics and tests outlined in the methodology are clearly designed to measure the performance of the network and the level of catastrophic forgetting. For example, average forgetting is measuring the retention of the information learnt by the network throughout training. This helps to ensure that the tests have construct validity as they are testing what they are designed to test. Further to this, measuring the information retention at various stages of the training via top-5 accuracy, overall accuracy, and the average forgetting rate highlights the face validity of these tests as they are clearly designed to ensure that the training procedures are doing what they are expected to do.

Furthermore, the datasets that I will be using in my experiments are widely used and have not been specifically designed for continual learning tasks but rather general machine learning which ensures they are sufficiently complex and extensively tested. This makes them appropriate as they range from toy datasets such as MNIST to more complex, realistic datasets such as CIFAR-100 and ImageNet. These datasets have been used throughout the existing literature as explored through my related works sections which suggests that they are widely accepted as being suitable. This also allows comparison against other published results, such as by the original authors of techniques that I will implement, as these datasets are standard in the literature.

5 Results

6 Evaluation

7 Conclusion

References

[1] R. Hecht-Nielsen, "Neurocomputing: picking the human brain," *IEEE spectrum*, vol. 25, no. 3, pp. 36–41, 1988.

[2] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

[3] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti *et al.*, "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model," *arXiv preprint arXiv:2201.11990*, 2022.

[4] K. Wiggers, "Ai weekly: Ai model training costs on the rise, highlighting need for new solutions," Oct 2021. [Online]. Available: https://venturebeat.com/2021/10/15/ai-weekly-ai-model-training-costs-on-the-rise-highlighting-need-for-new-solutions/

[5] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in psychology*, vol. 4, p. 504, 2013.

[6] "Kuzushiji recognition." [Online]. Available: https://www.kaggle.com/c/kuzushiji-recognition

[7] A. Prabhu, P. H. Torr, and P. K. Dokania, "Gdumb: A simple approach that questions our progress in continual learning," in *European conference on computer vision*. Springer, 2020, pp. 524–540.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[10] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[11] 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*, vol. 114, no. 13, pp. 3521–3526, 2017.

[12] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

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

[14] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[15] 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*, 2017, pp. 2001–2010.

[16] M. Welling, "Herding dynamical weights to learn," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1121–1128.

[17] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, "Gradient based sample selection for online continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[18] Y. Liu, Y. Su, A.-A. Liu, B. Schiele, and Q. Sun, "Mnemonics training: Multi-class incremental learning without forgetting," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 245–12 254.

[19] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, pp. 235–256, 2007.

[20] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8218–8227.

[21] D. Shim, Z. Mai, J. Jeong, S. Sanner, H. Kim, and J. Jang, "Online class-incremental continual learning with adversarial shapley value," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9630–9638.

[22] Z. Mai, H. Kim, J. Jeong, and S. Sanner, "Batch-level experience replay with review for continual learning," *arXiv preprint arXiv:2007.05683*, 2020.

[23] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "Remind your neural network to prevent catastrophic forgetting," in *European Conference on Computer Vision*. Springer, 2020, pp. 466–483.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[25] L. Wang, W. Chen, W. Yang, F. Bi, and F. R. Yu, "A state-of-the-art review on image synthesis with generative adversarial networks," *IEEE Access*, vol. 8, pp. 63 514–63 537, 2020.

[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[27] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *Advances in neural information processing systems*, vol. 30, 2017.

[28] R. Kemker and C. Kanan, "Fearnet: Brain-inspired model for incremental learning," *arXiv preprint arXiv:1711.10563*, 2017.

[29] K. Wang, J. van de Weijer, and L. Herranz, "Acae-remind for online continual learning with compressed feature replay," *Pattern Recognition Letters*, vol. 150, pp. 122–129, 2021.

[30] X. Liu, C. Wu, M. Menta, L. Herranz, B. Raducanu, A. D. Bagdanov, S. Jui, and J. v. de Weijer, "Generative feature replay for class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 226–227.

[31] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia, "Online continual learning with maximal interfered retrieval," *Advances in neural information processing systems*, vol. 32, 2019.

[32] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[33] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.

[34] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," *arXiv preprint arXiv:1810.11910*, 2018.

[35] L. Yu, B. Twardowski, X. Liu, L. Herranz, K. Wang, Y. Cheng, S. Jui, and J. v. d. Weijer, "Semantic drift compensation for class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6982–6991.

[36] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[37] Z. Li, M. Meng, Y. He, and Y. Liao, "Continual learning with laplace operator based node-importance dynamic architecture neural network," in *International Conference on Neural Information Processing*. Springer, 2021, pp. 52–63.

[38] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–547.

[39] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[40] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[41] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[42] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," in *Conference on Robot Learning*. PMLR, 2017, pp. 17–26.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[45] 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.

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