Continual Learning Techniques for Image Classification: Literature Survey

Student Name: Finlay Boyle
Supervisor Name: Dr. Donald Sturgeon

1 Related Work

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 [1]. 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.

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 the ability to retain knowledge whereas plasticity is the ability to learn new knowledge [2]. In 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.

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 assumed infeasible 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.

There are two key setups: Task-Incremental (Task-IL) and Class-Incremental (Class-IL) [3]. 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. A third scenario known as Domain-Incremental (Domain-IL) is also presented in [4], this is a different scenario where the model only needs to solve the task without identifying which task it has been presented with. The literature typically overlooks this scenario in favour of Class-IL.

1.1 Algorithm Paradigms

Continual learning literature has been heavily focused on the algorithms used to train models rather than the architecture of the models themselves. Within the literature there are three primary classes of approach: regularisation, memory-based, and meta-learning. There also exists some approaches that do not fit concretely into these categories.

Finetuning is the most naive approach to continual learning where the model is optimised on the incoming data once and then disregarded without concern about forgetting leading to very poor results [5]. On the other end of the spectrum is offline training which is the traditional approach to training models by repeatedly optimising on mini-batches.

1.1.1 Regularisation

Regularisation approaches were the initial focus of continual learning literature. Inspired by the stability-plasticity dilemma, the aim of this class of approaches is to control the weights in the network with the purpose of preventing significant changes to the weights in the network which could accelerate catastrophic forgetting. The general principle of regularisation approaches is to penalise weight changes through loss functions relative to the change in the weight [6]. This enables models to balance learning new knowledge without forgetting what they have already seen.

Elastic Weight Consolidation (EWC) was one of the first, and most influential, techniques in the continual learning domain. The proposed solution uses the Fisher information matrix to calculate maximum-likelihood estimates that are used to apply a weighting to the importance of parameters with the purpose of reducing the variance of parameters vital to previously seen tasks [7]. This importance weighting limits the plasticity of essential parts of the network for acquired knowledge by penalising weight changes with a quadratic constraint. The aim of this is to alleviate the risk of forgetting previously learnt tasks to maintain the

models performance. EWC set the initial groundwork for the continual learning domain and remains a prominent baseline comparison for proposed solutions [5].

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 [8]. 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. Then the network is jointly optimised across all sets of parameters using regularisation to prevent the overwriting of previous knowledge. Significantly, it was the first to use more complex datasets for evaluation such as ImageNet laying the foundations for future literature.

Memory aware synapses is another regularisation technique based on the ideas of Hebbian Learning [9], this is a concept from biology that 'cells that fire together, wire together'. The technique proposes that, due to limited model capacity, it may be necessary to erase rarely used knowledge from the network to create space for new knowledge. To achieve this, the importance of parameters in the network are estimated by measuring the gradient of the output with respect to specific data points. This is used as a weighting for how significant a specific neuron is within the network. Similarly to previously seen regularisation methods, changes to these parameters is penalised according to a quadratic loss function [10]. The defining characteristic of this technique 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 similar to EWC.

Ultimately, regularisation approaches have become less prevalent in the literature due to their poor performance as the number of classes to identify in the problem increases compared to other solution paradigms. As regularisation approaches were some of the first attempts to overcome the catastrophic forgetting problem, they do not address the Class-IL scenario but rather the simpler Task-IL scenario making them less relevant as the literature has progressed but they do provide a strong starting point that has influenced the domain.

1.1.2 Memory-Based

Memory-based approaches, also known as rehearsal based approaches, have been the subject of the majority of recent research in the continual learning domain. The underlying approach consists of storing a subset of incoming data, or a representation thereof, to be used throughout training to prevent forgetting. Due to the constraints on the continual learning problem, it is infeasible to store the entirety of the incoming data. As such, the aim is to remind the network of previously seen examples to limit the decay of existing knowledge.

The strategy used to sample from the incoming data to select exemplars is a significant part of the literature in this paradigm. The focus is to store exemplars that reflect the structure of the classes in the data and provide definition of the class boundaries.

One of the first prominent algorithms was iCaRL, it focused on prioritised exemplar selection [11]. Exemplars are selected using a technique called herding where exemplars are removed in a fixed order such that each exemplar is itself a near approximation of the mean of the set resulting in low variation in the exemplar set [12]. 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 uses the nearest-mean-of-exemplars rule where it computes the average feature vector for each class using the exemplars stored in memory and the extracted feature vector for the sample to classify. The average feature vector of each class is compared to the sample feature vector using a norm, the smallest norm implies the classification class. The results of this method are competitive and it appears regularly in other literature within the domain to be compared against. It outperformed finetuning by a factor of 5 and LwF [8] by a factor of 2.

Gradient Episodic Memory (GEM) is an early memory-based technique. This technique maintains an episodic memory (a memory buffer) that stores a set of exemplars. It stores the latest samples seen from each class in the buffer. In contrast to other methods, it uses loss functions as inequality constraints setting an upper bound for the loss. One significant downside to GEM is its computational inefficiency as it constrains the loss for each task seen so far to prevent knowledge decay. Averaged GEM (A-GEM) is a proposed alternative that instead constrains the average loss across all tasks reducing the wall-clock time by around 100 times with a small trade-off in effectiveness [13].

GDumb presented a far simpler memory-based approach than seen previously. 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 [3]. The purpose of this naive approach is to highlight flaws in the literature in continual learning. Despite the more complex approaches taken in previously discussed techniques, GDumb outperforms iCaRL [11] and other state of the art techniques by significant margins. One of the main concerns that raised was the impractical and unrealistic constraints used to formulate the continual learning problem. Primarily, it raised concerns around the focus of the literature on the Task-IL formation rather than the more realistic Class-IL scenario. 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 overly complicated with misleading results; it raises important questions regarding the direction of the field.

Mnemonics training is a state of the art technique that focuses on selecting exemplars via solving a Bilevel Optimisation Problem [14]. This technique extracts exemplars automatically, called mnemonics, which are optimised by stochastic gradient descent as more samples arrive, this is achieved by formulating an optimisation problem where there are two models that are optimised in alternating phases (i.e. the classification model and exemplar optimisation model) [15]. It proposes this strategy as an alternative to herding, which was used in iCaRL [11]. The authors 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.

Gradient-based sample selection uses a similar 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 [16]. 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 techniques 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.

Rainbow is a another recently successful sampling strategy. It selects 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 to increase the diversity of the stored samples in the feature space [17]. 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 with over twice the accuracy on ImageNet.

Adversarial Shapley Value Experience Replay (ASER) is underpinned by cooperative game theory. The premise for ASER's sampling strategy 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 [18]. 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 approaches to enhance the performance of existing ideas. One such approach is compressing exemplars via product quantisation which reduces the exemplars into their hidden representations from the neural network [19]. 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 technique 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 at both 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 to represent them. The idea behind using generative networks instead of directly storing exemplars is rooted in nature as humans and other species are not believed to store direct representations of events in our memories [6]. Generative models also have the potential to reduce the memory footprint associated with memory-based techniques as they store weights rather than exemplars directly. 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 as 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 [20]. 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 StyleGAN3 [21].

One of the first generative approaches to the continual learning problem was to use a generative adversarial network (GAN) [22] that 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 classifies the samples [23]. 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 in this technique was the proof of concept that GANs were a potential, viable avenue for further exploration. Whilst this technique 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 represents long-term and short-term memory by neural networks and a third network, the selector network, predicts whether a class is in long-term or short-term memory [24]. 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 that achieves this by compressing samples using an auto-encoder into feature representations to reduce the dimensionality of the samples which leads to approximately a 50x reduction in

memory consumption [25]. 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 methods that focus on using generative feature replays such as in [26] 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 [27]. 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 suggesting there are significant limitations in generative approaches for continual learning.

Another proposed generative replay approach is to combine the main network with the generator to mimic recall. This technique uses feedback connections in the network and allows knowledge to be consolidated by reactivating neurons to remind the network of the samples that it has seen [28]. 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.

Another recent development in the literature is a technique known as Learning to Prompt [29]. This proposes a different form of memory-based system. This method introduces prompts [30] which are small, learnable parameters that are prepended to the input features prior to passing the input to the model. The prompts are used to condition the model without needing to infer the exact task when classifying a sample, instead, the prompts are learnt during the training process. This method performed well and was compared to recent techniques such GDumb where it outperformed all of the other methods that it was compared against on complex datasets such as CIFAR-100 and CORe50. It also used sensible evaluation metrics such as average accuracy and forgetting. It was capable of performing near to the upper bound of offline training.

1.1.3 Meta-Learning

Meta-learning techniques are distinct in that they have a dual training procedure. Methods in this paradigm typically have two training loops. The inner training loop trains the model to classify data and the outer meta-training loop optimises the training procedure for the inner loop [6].

Online aware Meta-Learning (OML) is a technique that aims to learn sparse representations of the data to generalise the model such that it is able to support learning in the future while maintaining the existing knowledge [31]. To achieve this, OML has two networks: the Representation Learning Network (RLN) is updated in the meta-training phase, where the learning procedure is optimised, and the Prediction Learning Network (PLN), that aims to classify samples. To classify samples, they are passed through the RLN to give a representation of the sample using the learnt model. This representation is then passed into the PLN which predicts the output. The PLN is updated in the inner loop of the meta-learning process.

Look-ahead Meta Learning (La-MAML) is an optimisation-based meta-learning method that utilises a replay buffer to store exemplars, the inner training loop uses the incoming data samples and the exemplars whereas the outer training loop optimises the exemplar set that is stored [32]. This method also takes inspiration from regularisation approaches by having learnable learning rates on a neuron level to prevent rapid change in the weights. It is a hybrid regularisation, memory-based, and meta-learning approach. The authors used sufficiently complex datasets but the technique was only compared against historic techniques rather than the state of the art making it difficult to draw conclusions on the actual effectiveness.

Dataset Distillation using Neural Feature Regression is a meta-learning approach that aims to synthesise a small dataset that preserves the original information in the actual dataset [33]. The outer meta-training loop optimises the synthesised dataset and the inner training loop uses the synthesised dataset to update the classification model. In order to prevent overfitting of the synthetic data, this technique maintains a set of models each trained on the synthetic dataset but with random initialisations which feeds back into the meta-training loop. This technique achieves good results on MNIST and CIFAR-10, but similarily to the generative model based solutions, it struggles to scale up to CIFAR-100 and other complex datasets.

Meta-Experience Replay (MER) proposes a combined meta-learning and memory-based technique. It introduces two concepts, transfer and interference that are trade-offs against one another. The goal of MER is to learn parameters with the purpose of reducing interference and increasing transfer on future samples [34]. To achieve this, MER optimises a loss function using the dot product between gradients of different samples to encourage parameter sharing when gradients are similar and discourage interference between samples with poor transfer. The exemplars in the experience replay are selected via reservoir sampling which overcomes the problem of sampling with equal probability from a dataset of unknown size [35].

Another proposed meta-learning technique is MERLIN. This takes a different perspective by proposing that the parameters of a network can be sampled from a meta-distribution that is learnt during the meta-learning process [36]. This is achieved using varational auto encoders on a task-by-task basis which suggests it may not produce results that can be applied in realistic scenarios. The results of this method were generated using CIFAR-100 and Mini-ImageNet, however they are difficult to accurately compare because the authors compare them predominantly against other meta-learning techniques rather than state of the art from other continual learning paradigms.

1.1.4 Miscellaneous Approaches

There are a few approaches that lie outside of the three main categories discussed so far. They present possible alternative directions for future exploration in continual learning research.

Reinforced Continual Learning uses reinforcement learning to adapt the architecture of the network being trained to optimise it each time a new task is introduced [37]. The proposed solution consists of three seperate networks, the first is the controller which is a LSTM neural network that generates policies, the second is the value network that is a fully-connected network that estimates the value of each state, and the task network that is the network to be optimised. The task network is optimised by an actor-critic strategy, with the accuracy and complexity of the network determining the reward. The technique was not compared against high quality models and the results declined rapidly when CIFAR-100 was used over MNIST suggesting this technique struggles with scaling up to complex datasets.

Distillation is another technique that appears in the literature. Ensembling models together is seen as a simple way to increase the performance of machine learning models, however this is expensive and can require large amounts of computational resources [38]. Distillation is an alternative to ensembling, the much larger ensembled models are reduced into smaller, more manageable models that are quick at inference time and do not require as much memory as ensemble models. They can also be used to prevent variation in existing knowledge by consolidating it into a distilled form.

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 [39]. Instead of preventing the drift of classes, 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 [11] 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 relative to 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 [40]. 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 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 [41]. It achieves this by applying the Laplacian to the loss function computed at a specific node. It is used to measure how much a specific node 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 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 compete against EWC it is unlikely it would have challenged the state of the art memory-based methods.

1.2 Architecture Considerations

As discussed previously, continual learning literature has primarily focused on the algorithms used to train a model to prevent catastrophic forgetting and the architecture of the models being trained is sidelined. Typically, ResNet [42], convolutional neural networks, or vision transformers [43] are used as the underlying model for these training algorithms.

A comprehensive study of how the width and depth of a neural network affects catastrophic forgetting was recently published [44]. It found that two networks based on the same architecture but one with the parameters providing width to the model significantly outperformed the same architecture with the parameters providing depth to the model in terms of both accuracy and forgetting. This raises interesting questions about the surprising effect of the width of a neural network and provides a potential avenue for future exploration during the project.

A further study was also conducted by the same authors that measured the impact of different architecture types and the impact of specific layers within a network used for continual learning [45]. The authors of the study found that ResNet had greater capability to learn new tasks whereas convolutional neural networks and vision transformers were better at retaining information. They also found that the impact of batch normalisation layers was dependent on the data distribution, if it was relatively stationary then they were beneficial, otherwise they were detrimental. In addition to this, it was clear that global pooling layers negatively impacted the performance as they narrow the network which further confirmed the results of [44] whereas max pooling layers improved the performance because they did not narrow the network. Importantly, the authors noted that using a high quality architecture can be as impactful as a high quality training algorithm but the best performance can be achieved by using both.

1.3 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 techniques were not compared against the state of the art or created their own

metrics in order to quantify the performance of their algorithms [46].

Gradually, key metrics have been identified to compare different algorithms fairly. For example, the average forgetting rate, which is unique to online training methods, measures the drop in performance of the network caused by learning new classes. Forgetting is measured as the difference between the maximum knowledge about a specific task seen so far and the current knowledge of the model on the same task [47]. Intransigence is another important metric that was also introduced in [47] that quantifies the inability of a model to learn by comparing it against an offline trained version.

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. In continual learning it is also common to measure the accuracy throughout the training process although this is more relevant in the Task-IL scenario to evaluate the impact of new distinct tasks on the accuracy.

It is also important to consider the decay of knowledge as more tasks are learnt. To measure this, [48] proposes forwards and backwards transfer. Backwards transfer measures the influence of the current task on the model's knowledge of previous tasks by measuring the mean difference between the knowledge known prior to training on the current task compared to the model's knowledge of the previous tasks after training on the current task. Forwards transfer is similar except it measures how much influence the current task has on future tasks in a similar way. This has greater application in the Task-IL scenario as in the Class-IL scenario the tasks are not disjoint.

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) [3]. 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.

Interestingly, there does exist a concrete lower bound and upper bound for continual learning. Any continual learning algorithm should outperform finetuning and on the other end of the spectrum is offline training as discussed previously.

1.4 Datasets

Datasets are important to assess how well the continual learning algorithms perform on increasingly complex tasks as well as ensuring consistency during the comparison. Early literature tended to focus on using simple datasets such as CIFAR-10 [49] and MNIST [50] which are useful as basic examples but do not reflect the complexity of data from real scenarios and thus are not true indicators of how these algorithms would perform in real world environments [3]. As the literature has progressed, it is clear that the trend is to use more complicated datasets to better represent and compare algorithms.

CIFAR-100 is a prominent dataset in general image classification literature, it 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 [49]. It commonly features in the continual learning literature as outlined in the algorithm paradigms section.

ImageNet (and Mini-ImageNet), which has been used for visual recognition challenges, is a dataset containing 1000 classes with over 1.3 million images [51] making it the most challenging of the datasets but also the most realistic as it has been used in offline training to create some of the best classification models such as ResNet [42].

A popular evaluation dataset in the literature is Permuted MNIST [52]. This is a variation of the MNIST dataset where a fixed, random permutation is applied to all of the images in the dataset to create a different dataset of similar difficulty. It is designed to test the ability of continual learning models to identify items of the same class from different datasets. However, a review of its use found that it creates an unrealistically perfect scenario for continual learning [53].

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 20th century [54]. 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 continual learning on an optical character recognition task 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 are comparable against future techniques.

1.5 Summary

Overall, there is substantial recent literature in the continual learning domain and active research continues to be carried out to identify potential avenues of exploration both from an algorithm and architectural point of view. Further to this, the literature is beginning to mature but it remains difficult to compare and evaluate across different methods due to the inconsistencies in datasets and metrics used.

REFERENCES

[1] 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.

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

[3] 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.

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

[5] 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.

[6] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in cognitive sciences*, vol. 24, no. 12, pp. 1028–1040, 2020.

[7] 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.

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

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

[10] 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.

[11] 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.

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

[13] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," *arXiv preprint arXiv:1812.00420*, 2018.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. NeurIPS*, 2021.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.

[30] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[31] K. Javed and M. White, "Meta-learning representations for continual learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[32] G. Gupta, K. Yadav, and L. Paull, "La-maml: Look-ahead meta learning for continual learning. 2020," *URL https://arxiv. org/abs*, 2020.

[33] Y. Zhou, E. Nezhadarya, and J. Ba, "Dataset distillation using neural feature regression," *arXiv preprint arXiv:2206.00719*, 2022.

[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] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, vol. 11, no. 1, pp. 37–57, 1985.

[36] J. KJ and V. N Balasubramanian, "Meta-consolidation for continual learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 374–14 386, 2020.

[37] J. Xu and Z. Zhu, "Reinforced continual learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[38] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] S. I. Mirzadeh, A. Chaudhry, D. Yin, H. Hu, R. Pascanu, D. Gorur, and M. Farajtabar, "Wide neural networks forget less catastrophically," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 699–15 717.

[45] S. I. Mirzadeh, A. Chaudhry, D. Yin, T. Nguyen, R. Pascanu, D. Gorur, and M. Farajtabar, "Architecture matters in continual learning," *arXiv preprint arXiv:2202.00275*, 2022.

[46] 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.
[47] 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.
[48] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.
[49] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Citeseer*, 2009.
[50] 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.
[51] 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.
[52] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.
[53] S. Farquhar and Y. Gal, "Towards robust evaluations of continual learning," *arXiv preprint arXiv:1805.09733*, 2018.
[54] "Kuzushiji recognition." [Online]. Available: https://www.kaggle.com/c/kuzushiji-recognition