EEC: LEARNING TO ENCODE AND REGENERATE IMAGES FOR CONTINUAL LEARNING

Ali Ayub, Alan R. Wagner
The Pennsylvania State University
State College, PA, USA, 16803
{aja5755,alan.r.wagner}@psu.edu

ABSTRACT

The two main impediments to continual learning are catastrophic forgetting and memory limitations on the storage of data. To cope with these challenges, we propose a novel, cognitively-inspired approach which trains autoencoders with Neural Style Transfer to encode and store images. During training on a new task, reconstructed images from encoded episodes are replayed in order to avoid catastrophic forgetting. The loss function for the reconstructed images is weighted to reduce its effect during classifier training to cope with image degradation. When the system runs out of memory the encoded episodes are converted into centroids and covariance matrices, which are used to generate pseudo-images during classifier training, keeping classifier performance stable while using less memory. Our approach increases classification accuracy by 13-17% over state-of-the-art methods on benchmark datasets, while requiring 78% less storage space.

1 INTRODUCTION

Humans continue to learn new concepts over their lifetime without the need to relearn most previous concepts. Modern machine learning systems, however, require the complete training data to be available at one time (batch learning) (Girshick, 2015). In this paper, we consider the problem of continual learning from the class-incremental perspective. Class-incremental systems are required to learn from a stream of data belonging to different classes and are evaluated in a single-headed evaluation (Chaudhry et al., 2018). In single-headed evaluation, the model is evaluated on all classes observed so far without any information indicating which class is being observed.

Creating highly accurate class-incremental learning systems is a challenging problem. One simple way to create a class-incremental learner is by training the model on the data of the new classes, without revisiting the old classes. However, this causes the model to forget the previously learned classes and the overall classification accuracy decreases, a phenomenon known as catastrophic forgetting (Kirkpatrick et al., 2017). Most existing class-incremental learning methods avoid this problem by storing a portion of the training samples from the earlier learned classes and retraining the model (often a neural network) on a mixture of the stored data and new data containing new classes (Rebuffi et al., 2017; Hou et al., 2019). Storing real samples of the previous classes, however, leads to several issues. First, as pointed out by Wu et al. (2018b), storing real samples exhausts memory capacity and limits performance for real-world applications. Second, storing real samples introduces privacy and security issues (Wu et al., 2018b). Third, storing real samples is not biologically inspired, i.e. humans do not need to relearn previously known classes.

This paper explores the “strict” class-incremental learning problem in which the model is not allowed to store any real samples of the previously learned classes. The strict class-incremental learning problem is more akin to realistic learning scenarios such as a home service robot that must learn continually with limited on-board memory. This problem has been previously addressed using generative models such as autoencoders (Kemker & Kanan, 2018) or Generative Adversarial Networks (GANs) (Ostapenko et al., 2019). Most approaches for strict class-incremental learning...
use GANs to generate samples reflecting old class data, because GANs generate sharp, fine-grained images (Ostapenko et al., 2019). The downside of GANs, however, is that they tend to generate images which do not belong to any of the learned classes, hurting classification performance. Autoencoders, on the other hand, always generate images that relate to the learned classes, but tend to produce blurry images that are also not good for classification.

To cope with these issues, we propose a novel, cognitively-inspired approach termed Encoding Episodes as Concepts (EEC) for continual learning, which utilizes convolutional autoencoders to generate previously learned class data. Inspired by models of the hippocampus (Renoult et al., 2015), we use autoencoders to create compressed embeddings (encoded episodes) of real images and store them in memory. To avoid the generation of blurry images, we borrow ideas from the Neural Style Transfer (NST) algorithm proposed by Gatys et al. (2016) to train the autoencoders. For efficient memory management, we use the notion of memory integration, from hippocampal and neocortical concept learning (Mack et al., 2018), to combine similar episodes into centroids and covariance matrices eliminating the need to store real data.

This paper contributes: 1) an autoencoder based approach to strict class-incremental learning which uses Neural Style Transfer to produce quality samples reflecting old class data (Sec. 3.1); 2) a cognitively-inspired memory management technique that combines similar samples into a centroid/covariance representation, drastically reducing the memory required (Sec. 3.2); 3) a data filtering and a loss weighting technique to manage image degradation of old classes during classifier training (Sec. 3.3). We further show that EEC outperforms state-of-the-art (SOTA) approaches on benchmark datasets by significant margins while also using far less memory.

2 RELATED WORK

Most recent approaches to class-incremental learning store a portion of the real images belonging to the old classes to avoid catastrophic forgetting. Rebuffi et al. (2017) (iCaRL) store old class images and utilize knowledge distillation (Hinton et al., 2015) for representation learning and the nearest class mean (NCM) classifier for classification of the old and new classes. Knowledge distillation uses a loss term to force the labels of the images of previous classes to remain the same when learning new classes. Castro et al. (2018) (EEIL) improves iCaRL with an end-to-end learning approach. Wu et al. (2019) also stores real images and uses a bias correction layer to avoid any bias toward the new classes.

To avoid storing old class images, some approaches store features from the last fully-connected layer of the neural networks (Xiang et al., 2019; Hayes & Kanan, 2020; Ayub & Wagner, 2020). These approaches, however, use a network pretrained on ImageNet to extract features, which gives them an unfair advantage over other approaches. Because of their reliance on a pretrained network, these approaches cannot be applied in situations when new data differs drastically from ImageNet (Russakovsky et al., 2015).

These difficulties have forced researchers to consider using generative networks. Methods employing generative networks tend to model previous class statistics and regenerate images belonging to the old classes while attempting to learn new classes. Both Shin et al. (2017) and Wu et al. (2018) use generative replay where the generator is trained on a mixture of generated old class images and real images from the new classes. This approach, however, causes images belonging to classes learned in earlier increments to start to semantically drift, i.e. the quality of images degrades because of the repeated training on synthesized images. Ostapenko et al. (2019) avoids semantic drift by training the GAN only once on the data of each class. Catastrophic forgetting is avoided by applying elastic weight consolidation (Kirkpatrick et al., 2017), in which changes in important weights needed for old classes are avoided when learning new classes. They also grow their network when it runs out of memory while learning new classes, which can be difficult to apply in situations with restricted memory. One major issue with GAN based approaches is that GANs tend to generate images that do not belong to any of the learned classes which decreases classification accuracy. For these reasons, most approaches only perform well on simpler datasets such as MNIST (LeChun, 1998) but perform poorly on complex datasets such as ImageNet. Conditional GAN can be used to mitigate the problem of images belonging to none of the classes as done by Ostapenko et al. (2019), however the performance is still poor on complex datasets such as ImageNet-50 (see Table 1) and...
model (classifier for class-incremental learning)

Our formal presentation of continual learning follows Ostapenko et al. (2019), where a task solver increments. In each increment, the model is evaluated on all the classes seen so far. Each task can contain one or multiple classes and data for different tasks is available to the model in different samples. Following the notation of Chaudhry et al. (2018), we consider a pre-trained network to extract features gives FearNet an unfair advantage over other approaches. The autoencoder parameters are optimized using a combination of reconstruction and content losses: $L_{r} = \|x - f_{r}(x)\|_{2}$ (1) Although autoencoders are suitable for dimensionality reduction for complex, high-dimensional data like RGB images, the reconstructed images lose the high frequency components necessary for correct classification. To tackle this problem, we train autoencoders using some of the ideas that underline Neural Style Transfer (NST). NST uses a pre-trained CNN to transfer the style of one image to another. The process takes three images, an input image, a content image and a style image. The three images are passed through the pre-trained CNN generating convolutional feature maps (usually from the last convolutional layer) and $l_2$ distances between the feature maps of the input image and content image (content loss) and style image (style loss) are calculated. These losses are then used to update the input image.

Intuitively, our intent here is to create reconstructed images that are similar to the real images (in the pixel and convolutional space), thereby improving classification accuracy. Hence, we only utilize the idea of content transfer from the NST algorithm, where the input image is the image reconstructed by the autoencoder and content image is the real image corresponding to the reconstructed image. The classifier model, $D_{t}$, is used to generate convolutional feature maps for the NST, since it is already trained on real data for the classes in the increment $t$. In contrast to the traditional NST algorithm, we use the content loss ($L_{cont}$) to train the autoencoder, rather than updating the input image directly. Formally, let $f_{r} : \mathcal{X} \rightarrow \mathcal{F}$ be the classifier pipeline that converts input images into convolutional features. For an input image, $x_{t}^{i}$ of task $t$, the content loss is:

$$L_{cont} = \|f_{c}(x_{t}^{i}) - f_{c}(f_{r}(x_{t}^{i}))\|_{2}$$ (2)

The autoencoder parameters are optimized using a combination of reconstruction and content losses:
Figure 1: Complete architecture of EEC. (a) For each new task, a convolutional autoencoder is trained on real images and a combination of reconstruction loss $L_r$ and content loss $L_{cont}$. The encoded episodes are stored in memory and converted into centroids and covariance matrices when the system runs out of memory. (b) The classifier is trained on a combination of real, reconstructed and pseudo-images in each increment.

Figure 2: Reconstructed images generated by autoencoders trained (a) using NST and (b) not using NST.

$$L = (1 - \lambda)L_r + \lambda L_{cont}$$  \hspace{1cm} (3)

where, $\lambda$ is a hyperparameter that controls the contribution of each loss term towards the complete loss. During autoencoder training, classifier $D$ acts as a fixed feature extractor and its parameters are not updated. This portion of the complete procedure is depicted in Figure 1(a).

To provide an illustration of our approach, we perform an experiment with ImageNet-50 dataset. We trained one autoencoder on 10 classes from ImageNet-50 with NST and one without NST. Figure 2 depicts the reconstructed images by the two autoencoders. Note that the images generated by the autoencoder trained without using NST are blurry. In contrast, the autoencoder trained using NST creates images with fine-grained details which improves the classification accuracy.

3.2 MEMORY INTEGRATION

As mentioned in the introduction, continual learning also presents issues associated with the storage of data in memory. For EEC, for each new task $t$, the data is encoded and stored in memory. Even though the encoded episodes require less memory than the real images, the system can still run out of memory when managing a continuous stream of incoming tasks. To cope with this issue, we propose a process inspired by memory integration in the hippocampus and the neocortex [Mack et al., 2018]. Memory integration combines a new episode with a previously learned episode summarizing the information in both episodes in a single representation. The original episodes themselves are forgotten.

Consider a system that can store a total of $K$ encoded episodes based on its available memory. Assume that at increment $t - 1$, the system has a total of $K_{t-1}$ encoded episodes stored. It is now required to store $K_t$ more episodes in increment $t$. The system runs out of memory because $K_t + K_{t-1} > K$. Therefore, it must reduce the number of episodes to $K_t = K_{t-1} + K_t - K$. Because each task is composed of a set of classes at each increment, we reduce the total encoded episodes belonging to different classes based on their previous number of encoded episodes. Formally, the reduction in the number of encoded episodes $N_y$ for a class $y$ is calculated as (whole number):

$$N_y(\text{new}) = N_y\left(1 - \frac{K_t}{K_{t-1}}\right)$$  \hspace{1cm} (4)

To reduce the encoded episodes to $N_y(\text{new})$ for class $y$, inspired by the memory integration process, we use an incremental clustering process that combines the closest encoded episodes to produce cen-
troids and covariance matrices. This clustering technique is similar to the Agg-Var clustering proposed in our earlier work (Ayub & Wagner, 2020b). The distance between encoded episodes is calculated using the Euclidean distance, and the centroids are calculated using the weighted mean of the encoded episodes. The process is repeated until the sum of the total number of centroids/covariance matrices and the encoded episodes for class $y$ equals $N_y^{(\text{new})}$. The original encoded episodes are removed and only the centroid and the covariances are kept (see Appendix A for more details).

3.3 Rehearsal, Pseudorehearsal and Classifier Training

Figure 1 (b) depicts the procedure for training the classifier $D$ when new data belonging to task $t$ becomes available. The classifier is trained on data from three sources: 1) the real data for task $t$, 2) reconstructed images generated from the autoencoder’s decoder, and 3) pseudo-images generated from centroids and covariance matrices for previous tasks. Source (2) uses the encodings from the previous tasks to generate a set of reconstructed images by passing them through the autoencoder’s decoder. This process is referred to as rehearsal (Robins, 1995).

Pseudorehearsal: If the system also has old class data stored as centroids/covariance matrix pairs, pseudorehearsal is employed. For each centroid/covariance matrix pair of a class we sample a multivariate Gaussian distribution with mean as the centroid and the covariance matrix to generate a large set of pseudo-encoded episodes. The episodes are then passed through the autoencoder’s decoder to generate pseudo-images for the previous classes. Many of the pseudo-images are noisy. To filter the noisy pseudo-images, we pass them through classifier $D$, which has already been trained on the prior classes, to get predicted labels for each pseudo-image. We only keep those pseudo-images that have the same predicted label as the label of the centroid they originated from. Among the filtered pseudo-images, we select the same number of pseudo-images as the total number of encoded episodes represented by the centroid and discard the rest (see Appendix A for the algorithm).

To illustrate the effect of memory integration and pseudorehearsal, we performed an experiment on MNIST dataset. We trained an autoencoder on 2 classes (11379 images) from MNIST dataset with the embedding dimension of size 2. After training, we passed all the training images for the two classes through the encoder to get 11379 encoded episodes (see Figure 3 (a)). Next, we applied our memory integration technique on the encoded episodes to reduce the encoded episodes to 4000, 2000, 1000 and 200 centroids (and covariances). No original encoded episodes were kept. Pseudorehearsal was then applied on the centroids to generate pseudo-encoded episodes. The pseudo-encoded episodes for different numbers of centroids are shown in Figure 3 (b-e). Note that the feature space for the pseudo-encoded episodes generated by different number of centroids is very similar to the original encoded episodes. For a larger number of centroids, the feature space looks almost exactly the same as the original feature space. For the smallest number of centroids (Figure 3 (e)), the feature space becomes less uniform and more divided into small dense regions. This is because a smaller number of centroids are being used to represent the overall concept of the feature space across different regions resulting in large gaps between the centroids. Hence, the pseudo-encoded episodes generated using the centroids are more dense around the centroids reducing uniformity. Still, the overall shape of the feature space is preserved after using the centroids and pseudorehearsal. Hence, our approach conserves information about previous classes, even with less memory, contributing to classifier performance while also avoiding catastrophic forgetting. The results presented in Section 4.3 on ImageNet-50 confirm the effectiveness of memory integration and pseudorehearsal.
Sample Decay Weight: The reconstructed images and pseudo-images can still be quite different from the original images, hurting classifier performance. We therefore weigh the loss term for reconstructed and pseudo-images while training $D$. For this, we estimate the degradation in the reconstructed and pseudo-images. To estimate degradation in the reconstructed images, we find the ratio of the classification accuracy of the reconstructed images ($c^r$) to the accuracy of the original images ($c^o$) on network $D$ trained on previous tasks. This ratio is used to control the weight of the loss term for the reconstructed images. For a previous task $t-1$, the weight $\Gamma_{t-1}^r$ (the sample decay weight) for the loss term $L_{t-1}^r$ of the reconstructed images is defined as:

$$\Gamma_{t-1}^r = e^{-\gamma_{t-1} \alpha_{t-1}}$$

(5)

where $\gamma_{t-1} = 1 - \frac{c_{t-1}^r}{c_{t-1}^o}$ (sample decay coefficient) denotes the degradation in reconstructed images of task $t-1$ and $\alpha_{t-1}$ represents the number of times an autoencoder has been trained on the pseudo-images or reconstructed images of task $t-1$. The value of sample decay coefficient ranges from 0 to 1, depending on the classification accuracy of the reconstructed images $c_{t-1}^r$. If $c_{t-1}^r = c_{t-1}^o$, $\gamma_{t-1} = 0$ (no degradation) and $\Gamma_{t-1}^r = 1$. Similarly, the sample decay weight $\Gamma_{t-1}^p$ for loss term $L_{t-1}^p$ for the pseudo-images is based on the classification accuracy $c_{t-1}^p$ of pseudo-images for task $t-1$. Thus, for a new increment, the total loss $L_D$ for training $D$ on the reconstructed and pseudo-images of the old tasks and real images of the new task $t$ is defined as:

$$L_D = L_t + \sum_{i=1}^{t-1} (\Gamma_i^r L_i^r + \Gamma_i^p L_i^p)$$

(6)

4 Experiments

We tested and compared EEC to several SOTA approaches on four benchmark datasets: MNIST, SVHN, CIFAR-10 and ImageNet-50. We also report the memory used by our approach and its performance in restricted memory conditions. Finally, we present an ablation study to evaluate the contribution of different components of EEC. Experiments on CIFAR-100 and comparison with additional methods are presented in Appendix D. Other generative memory approaches by previous authors did not test on the CIFAR-100 dataset.

4.1 Datasets

The MNIST dataset consists of grey-scale images of handwritten digits between 0 to 9, with 50,000 training images, 10,000 validation images and 10,000 test images. SVHN consists of colored cropped images of street house numbers with different illuminations and viewpoints. It contains 73,257 training and 26,032 test images belonging to 10 classes. CIFAR-10 consists of 50,000 RGB training images and 10,000 test images belonging to 10 object classes. Each class contains 5000 training and 1000 test images. ImageNet-50 is a smaller subset of the iLSVRC-2012 dataset containing 50 classes with 1300 training images and 50 validation images per class. All of the dataset images were resized to 32×32, in concordance to Ostapenko et al. [2019].

4.2 Implementation Details

We used Pytorch [Paszke et al., 2019] and an Nvidia Titan RTX GPU for implementation and training of all neural network models. A 3-layer shallow convolutional autoencoder was used for all datasets (see Appendix B), which requires approximately 0.2MB of disk space for storage. For classification, on the MNIST and SVHN datasets the same classifier as the DCGAN discriminator (Radford et al., 2015) was used, for CIFAR-10 the ResNet architecture proposed by Gulrajani et al. (2017) was used and for ImageNet-50, ResNet-18 (He et al., 2016) was used.

In concordance with Ostapenko et al. (2019) (DGMw), we report the average incremental accuracy on 5 and 10 classes ($A_5$ and $A_{10}$) for the MNIST, SVHN and CIFAR-10 datasets trained continually with one class per increment. For ImageNet-50, the average incremental accuracy on 3 and 5 increments ($A_{30}$ and $A_{50}$) is reported with 10 classes in each increment. For a fair comparison,
Table 1: Comparison of EEC and EECS to approaches that store and use real samples (denoted with an S) and those that generate samples, for class-incremental learning on MNIST, SVHN, CIFAR-10 and ImageNet-50 datasets. The difference between EEC and the SOTA approach is shown in bold.

we typically compare against approaches with a generative memory replay component that do not use a pre-trained feature extractor and are evaluated in a single-headed fashion. Among such approaches, to the best of our knowledge, DGMw represents the state-of-the-art benchmark on these datasets which is followed by MeRGAN (Wu et al., 2018a), DGR (Shin et al., 2017) and EWC-M (Seff et al., 2017). Joint training (JT) is used to produce an upperbound for all four datasets. We compare these methods against two variants of our approach: EEC and EECS. EEC uses a separate autoencoder for classes in a new increment, while EECS uses a single autoencoder that is retrained on the reconstructed images of the old classes when learning new classes. For both EEC and EECS, results are reported when all of the encoded episodes for the previous classes are stored. The results for EEC under restricted memory conditions are also presented.

Hyperparameter values and training details are reported in Appendix C. We performed each experiment 10 times with different random seeds and report average accuracy over all runs.

4.3 Comparison with SOTA Methods

Table 1 compares EEC and EECS against SOTA approaches on the MNIST, SVHN, CIFAR-10 and ImageNet-50 datasets. We compare against two different types of approaches, those that use real images (episodic memory) of the old classes and those that generate previous class images when learning new classes. Both EEC and EECS outperform EWC-M and DGR by significant margins on the MNIST on $A_5$ and $A_{10}$. MeRGAN and DGMw perform similarly to our methods on the $A_5$ and $A_{10}$ experiments. Note that MeRGAN and EEC approach the JT upperbound on $A_5$. Further, accuracy for MeRGAN, DGMw and EEC changes only slightly between $A_5$ and $A_{10}$, suggesting that MNIST is too simple of a dataset for testing continual learning using generative replay.

Considering SVHN, a more complex dataset consisting of colored images of street house numbers, the performance of our method remains reasonably close to the JT upperbound, even though the performance of other approaches decreases significantly. For $A_5$, EEC achieves only 0.27% lower accuracy than JT and 11.36% higher than DGMw (current best method). For $A_{10}$, EEC is 4.55% lower than JT but still 15.21% higher than DGMw and achieves 5.66% higher accuracy on $A_{10}$ than DGMw did on $A_5$. EECs performs slightly lower than EEC on $A_5$ but the gap widens on $A_{10}$. However, EEC also beats the SOTA approaches on both $A_5$ and $A_{10}$.

Considering the more complex CIFAR-10 and ImageNet-50 datasets, only DGMw reported results for these datasets. On CIFAR-10, EEC beats DGMw on $A_5$ by a margin of 20.18% and on $A_{10}$ by a margin of 10.7%. Similar to the SVHN results, the accuracy achieved by EEC on $A_{10}$ is even higher than DGMw’s accuracy on $A_5$. In comparison with the JT, EEC performs similarly on $A_5$ (0.38% lower), however on $A_{10}$ it performs significantly lower. For ImageNet-50, again we see similar results as both EEC and EECS outperform DGMw on $A_{30}$ and $A_{50}$ by significant margins (13.29% and 17.42%, respectively). Similar to SVHN and CIFAR-10 results, accuracy for EEC on $A_{50}$ is even higher than DGMw’s accuracy on $A_{30}$. Further, EEC also beats iCaRL (episodic
Figure 4: Reconstructed Images for CIFAR-10 (top) and ImageNet-50 (bottom) after all tasks

memory SOTA method) with margins of 16.01% and 6.26% on $A_{30}$ and $A_{50}$, respectively, even though iCaRL has an unfair advantage of using stored real images.

Discussion: Our method performed consistently across all four datasets, especially on $A_5$ for MNIST, SVHN and CIFAR-10. In contrast, DGMw (the best current method) shows significantly different results across the four datasets. The results suggest that the current generative memory-based SOTA approaches are unable to mitigate catastrophic forgetting on more complex RGB datasets. This could be because GANs tend to generate images that do not belong to any of learned classes, which can drastically reduce classifier performance. Our approach copes with these issues by training autoencoders borrowing ideas from the NST algorithm and retraining of the classifier with sample decay weights. Images reconstructed by EEC for CIFAR-10 after 10 tasks and for ImageNet-50 after 5 tasks are shown in Figure 4 (more examples in supplementary file).

Memory Usage Analysis: Similar to DGMw, we analyze the disc space required by our model for the ImageNet-50 dataset. For EEC, the autoencoders use a total disc space of 1 MB, ResNet-18 uses about 44 MB, while the encoded episodes use a total of about 66.56 MB. Hence, the total disc space required by EEC is about 111.56 MB. DGMw’s generator (with corresponding weight masks), however, uses 228MB of disc space and storing pre-processed real images of ImageNet-50 requires disc space of 315MB. Hence, our model requires 51.07% ((228-111.56)/228 = 0.5107) less space than DGMw and 64.58% less space than the real images for ImageNet-50 dataset.

EEC was tested with different memory budgets on the ImageNet-50 dataset to evaluate the impact of our memory management technique. The memory budgets (K) are defined as the sum of the total number of encoded episodes, centroids and covariance matrices (diagonal entries) stored by the system. Figure 5 shows the results in terms of $A_{30}$ and $A_{50}$ for a wide range of memory budgets. The accuracy of EEC changes only slightly over different memory budgets. Even with a low budget of K=5000, the $A_{30}$ and $A_{50}$ accuracies are only 3.1% and 3.73% lower than the accuracy of EEC with unlimited memory. Further, even for K=5000, EEC beats DGMw (current SOTA on ImageNet-50) by margins of 10.17% and 13.73% on $A_{30}$ and $A_{50}$, respectively. The total disc space required for K=5000 is only 5.12 MB and the total disc space for the complete system is 50.12 MB (44 MB for ResNet-18 and 1 MB for autoencoders), which is 78.01% less than DGMw’s required disc space (228 MB). These results clearly depict that our approach produces the best results even with extremely limited memory, a trait that is not shared by other SOTA approaches. Moreover, the results also show that our approach is capable of dealing with the two main challenges of continual learning mentioned earlier: catastrophic forgetting and memory management.

4.4 Ablation Study

We performed an ablation study to examine the contribution of using: 1) the content loss while training the autoencoders, 2) pseudo-rehearsal and 3) the sample weight decay. These experiments were performed on the ImageNet-50 dataset. Complete encoded episodes of previous tasks were used for ablations 1 and 3, and a memory budget of K=5000 was used for ablation 2. We created hybrid versions of EEC to test the contribution of each component. Hybrid 1: termed as EEC-noNST does not use the content loss while training the autoencoders. Hybrid 2: termed as EEC-CGAN uses a conditional GAN instead of NST based autoencoder. Hybrid 3: termed as EEC-VAE uses a variational autoencoder and Hybrid 4: termed as EEC-DNA uses a denoising autoencoder instead of our proposed NST based autoencoder. Hybrid 5: termed as EEC-noPseudo simply removes the extra encoded episodes when the system runs out of memory and does not use pseudo-rehearsal. Hybrid 6: termed as EEC-noDecay does not use sample weight decay during classifier training on new
old tasks. Except for the changed component, all the other components in the hybrid approaches were the same as used in EEC.

All of the hybrids show inferior performance as compared to the complete EEC approach (Table 2). Evaluated on $A_{30}$ and $A_{50}$, EEC-noNST, EEC-CGAN, EEC-VAE and EEC-DNA all result in $\sim$5.8% and $\sim$6.5% lower accuracy, respectively than EEC. EEC-noDecay results in 3.77% and 4.77% lower accuracy on $A_{30}$ and $A_{50}$ respectively, than EEC. As a fair comparison with EEC using K=5000, EEC-noPseudo results in 4.98% and 6.27% lower accuracy for $A_{30}$ and $A_{50}$, respectively. The results show that all of these components contribute significantly to the overall performance of EEC, however training the autoencoders with content loss and pseudo-rehearsal seem to be more important. Also, note that the conditional GAN with the other components of our approach (such as sample weight decay) produces 7.80% and 11.3% higher accuracy on $A_{30}$ and $A_{50}$, respectively, than DGMw. Further note that all the hybrids that do not use the NST based autoencoder (Hybrids 1-4) produce similar accuracy, since they use the sample weight decay to help manage the degradation of reconstructed images. These results show the effectiveness of sample weight decay to cope with image degradation.

4.5 Experiments on Higher Resolution Images

As mentioned in Subsection 4.2, all the experiments on the four datasets were done using resized images of size 32×32. To show further insight into our approach, we performed experiments on ImageNet-50 dataset with higher resolution images of size 256×256 and randomly cropped to 224×224. Evaluated on $A_{30}$ and $A_{50}$, EEC achieved 70.75% and 56.89% accuracy, respectively using images of size 224×224. These results show that using higher resolution images improve the performance of EEC by significant margins (25.36% and 21.65% increase on $A_{30}$ and $A_{50}$). Note that storing real high resolution images requires a much larger disk space than storing images of size 32×32. For ImageNet-50, storing original images of size 224×224 requires 13.04GB, while images of size 32×32 require only 315MB. In contrast, EEC requires the same disk space regardless of the size of the original images (only 111.56MB, 99.14% less than the real images of size 224×224) because EEC only stores the encoded episodes and their size is independent of the resolution of the images. These results further show the effectiveness of our approach to mitigate catastrophic forgetting while using significantly less memory compared to other approaches.

5 Conclusion

This paper has presented a novel and potentially powerful approach (EEC) to strict class-incremental learning. Our paper demonstrates that the generation of high quality reconstructed data can serve as the basis for improved classification during continual learning. We further demonstrate techniques for dealing with image degradation during classifier training on new tasks as well as a cognitively-inspired clustering approach that can be used to manage the memory. Our experimental results demonstrate that these techniques mitigate the effects of catastrophic forgetting, especially on complex RGB datasets, while also using less memory than SOTA approaches. Future continual learning approaches can incorporate different components of our approach such as the NST-based autoencoder, pseudo-rehearsal and sample decay weights for improved performance.
REFERENCES


A  EEC ALGORITHMS

The algorithms below describe portions of the complete EEC algorithm. Algorithm 1 is for autoencoder training (Section 3.1 in paper), Algorithm 2 is for memory integration (Section 3.2 in paper), Algorithm 3 is for rehearsal, pseudo-rehearsal and classifier training (Section 3.3 in paper) and Algorithm 4 is for filtering pseudo-images (Section 3.3 in paper).

Algorithm 1: EEC: Train Autoencoder

Input: $S_t = \{(x^t_i, y^t_i)\}_{i=1}^{n^t}$  \(\triangleright\) data points with ground truths for task $t$
require: $\lambda$
require: $f_r = (g_{dec} \circ g_{enc})$  \(\triangleright\) autoencoder pipeline
require: $f_c : \mathcal{X} \rightarrow \mathcal{F}_c$  \(\triangleright\) convolutional feature extractor from classifier $D$
require: $N_{epoch}$  \(\triangleright\) Number of epochs
Output: $F_t = \{(f^t_i, y^t_i)\}_{i=1}^{n^t}$  \(\triangleright\) encoded episodes with ground truths for task $t$

1: for $j = 1; j < N_{epoch}$ do
2: $X^*_t = f_r(X_t)$  \(\triangleright\) get reconstructed images for task $t$
3: $L_r = ||X_t - X^*_t||_2$  \(\triangleright\) Reconstruction loss
4: $L_{cont} = ||f_c(X_t) - f_c(X^*_t)||_2$  \(\triangleright\) Content loss
5: $L = (1 - \lambda)L_r + \lambda L_{cont}$  \(\triangleright\) complete autoencoder loss
6: $\theta_r \leftarrow \minimize(L)$  \(\triangleright\) update autoencoder parameters such that it minimizes $L$
7: $F_t = g_{enc}(X_t)$  \(\triangleright\) get encoded episodes for images of task $t$

Algorithm 2: EEC: Memory Integration

Input: $F_{\text{new}} = \{F_i\}_{i=1}^{t-1}$, where $F_i = \{(f^i_j, y^i_j)\}_{j=1}^{n^i}$  \(\triangleright\) encoded episodes set for previous $t-1$ tasks
require: $K$  \(\triangleright\) maximum number of episodes
require: $K_t$  \(\triangleright\) number of encoded episodes for task $t$
Output: $F_{\text{new}}$  \(\triangleright\) reduced encoded episodes for previous tasks
Output: $C = \{C_i\}_{i=1}^{t-1}$  \(\triangleright\) centroids for previous tasks
Output: $\Sigma = \{\Sigma_i\}_{i=1}^{t-1}$  \(\triangleright\) covariances for previous tasks

Initialize: $K_{t-1} = \sum_{i=1}^{t-1} n^i$  \(\triangleright\) number of encoded episodes for $t-1$ previous tasks

1: $K_r = K_t + K_{t-1} - K$
2: for $i = 1; i \leq t - 1$ do
3: for $j = 1; j \leq y^i$ do
4: repeat  \(\triangleright\) for each class in task $i$ find centroids and covariances
5: $c_j, \sigma_j \leftarrow \text{combine}(x^i_l, x^i_q) \forall y^i_l = y^i_q = j$  \(\triangleright\) Combine closest points for class $j$ in task $i$
6: $N_j(\text{new}) = 2N_j(\text{cent}) + \sum_{i=1,y^i_j=1}^{n^i} 1$  \(\triangleright\) $N_j(\text{cent})$: number of centroids for class $j$
7: until $N_j(\text{new}) = N_j(1 - \frac{K_r}{K_{t-1}})$
Algorithm 3: EEC: Train Classifier

Input: $S_t = \{(x^t_i, y^t_i)\}^n_t, \quad \triangleright$ data points with ground truths for task $t$
Input: $F = \{F_i\}^t_{i=1}$, where $F_i = \{(f^i_j, y^i_j)\}^n_{j=1} \quad \triangleright$ encoded episodes for previous tasks
Input: $C = \{C_i\}^t_{i=1}$, where $C_i = \{(c^i_j, y^i_j)\}^{N^i(cent)}_{j=1}$ \quad \triangleright centoids for previous tasks
Input: $\Sigma = \{\Sigma_i\}^t_{i=1}$, where $\Sigma_i = \{\sigma^i_j, y^i_j\}^{N^i(cent)}_{j=1}$ \quad \triangleright covariances for previous tasks
Input: $M = \{M_i\}^t_{i=1}$, where $M_i = \{m^i_j\}^{N^i(cent)}_{j=1} \quad \triangleright$ no. of episodes clustered in each centroid
require: $c^i_{t-1}$ \quad \triangleright accuracy of original training images of previous $t-1$ tasks
require: $D$ \quad \triangleright Classifier model
require: $\{g^j_{dec}\}^t_{j=1}$ \quad \triangleright decoder part of autoencoders for each of the previous tasks
require: $N_{epoch}$ \quad \triangleright Number of epochs

1: for $i = 1; i \leq t-1$
2:   for $j = 1; j \leq N^i(cent)$ do
3:     $F^i_t(large).append(multivariate_normal(c^i_j, \sigma^i_j))$ \quad \triangleright generate pseudo-samples for each concept
4:     $X^i_1(large) = g^i_{dec}(F^i_t(large))$ \quad \triangleright generate large no. of pseudo-images from pseudo-samples
5:     $X^i_1 = FILTER(X^i_1(large), M_i)$ \quad \triangleright Algorithm 4
6:     $X^i_1 = g^i_{dec}(F_i)$ \quad \triangleright get reconstructed images for encoded episodes
7:   for $i = 1; i \leq t-1$ do
8:     $Y^i_j = D(X^i_1)$ \quad \triangleright predictions for reconstructed images for task $i$
9:     $Y^i_j = D(X^i_1)$ \quad \triangleright predictions for pseudo-images for task $i$
10:    $c^p_i = \frac{\sum_{j=1}^{N^i(cent)} [Y^i_j; Y^t_j]==Y^i(j)}{N^i}$ \quad \triangleright accuracy for reconstructed images of task $i$
11:    $c^p_i = \frac{\sum_{j=1}^{N^i(cent)} [Y^i_j; Y^t_j]==Y^i(j)}{N^i}$ \quad \triangleright accuracy for pseudo images of task $i$
12:    $\gamma^i_t = 1 - \frac{c^p_i}{c^p}$ \quad \triangleright sample decay coefficient for reconstructed images of task $i$
13:    $\gamma^p_i = 1 - \frac{c^p_i}{c^p}$ \quad \triangleright sample decay coefficient for pseudo images of task $i$
14:    $\Gamma^i_t = e^{-\gamma^i_t \alpha^i_t}$
15:    $\Gamma^p_i = e^{-\gamma^p_t \alpha^p_t}$
16:   for $j = 1; j < N_{epoch}$ do
17:     for $i = 1; i \leq t-1$ do
18:        $Y^i_1 = D(X^i_1)$ \quad \triangleright predictions for reconstructed images for task $i$
19:        $Y^i_1 = D(X^i_1)$ \quad \triangleright predictions for pseudo-images for task $i$
20:        $\mathcal{L}^i_t = \text{CrossEntropy}(Y^i_1 - Y^i_t)$ \quad \triangleright loss for reconstructed images of task $i$
21:        $\mathcal{L}^p_i = \text{CrossEntropy}(Y^i_1 - Y^i_t)$ \quad \triangleright loss for pseudo-images of task $i$
22:        $\mathcal{L}^i_D = \mathcal{L}^i_t + \sum_{i=1}^{t-1} \Gamma^i_t \mathcal{L}^i_t$ \quad \triangleright loss for task $t$
23:        $\mathcal{L}^p = \mathcal{L}^p_t + \sum_{i=1}^{t-1} \Gamma^p_t \mathcal{L}^p_t$ \quad \triangleright loss for task $t$
24:        $\theta_D \leftarrow \text{minimize}(\mathcal{L}^i_D)$ \quad \triangleright update classifier parameters such that it minimizes $\mathcal{L}^i_D$

Algorithm 4: EEC: Filter Pseudo-images

Input: $S_t(large) = \{(x^t_i, y^t_i)\}{n^t(large)}_{i=1}$, where $X_t(large) = \{x^t_i\}{n^t(large)}_{i=1}$,
$Y_t(large) = \{y^t_i\}{n^t(large)}_{i=1}$
Input: $M_t = \{m^t_i\}^{N^t(cent)}_{i=1}$
require: $D$
Output: $S_t = \{(x^t_i, y^t_i)\}{n^t}_{i=1}$ \quad \triangleright filtered pseudo-images for task $t$

1: $Y^*_t(large) = D(X_t(large))$
2: $X_t = \{x^*_i \in X_t(large); y^*_i == y^t_i\}^{n^t(large)}_{i=1}$ \quad \triangleright pseudo-images with correct predicted label
3: $X_t = \{x^*_i \in X_t\}^{M_t}_{i=1}$ \quad \triangleright keep $M_t$ pseudo-images
B Autoencoder Architectures

<table>
<thead>
<tr>
<th>Name</th>
<th>Layer Type</th>
<th>Filter Size</th>
<th>Stride</th>
<th>Output Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>Conv2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>batch_norm1</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>relu1</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>conv2</td>
<td>Conv2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>batch_norm2</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>relu2</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>conv3</td>
<td>Conv2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,16,4,4)</td>
</tr>
<tr>
<td>batch_norm3</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,16,4,4)</td>
</tr>
<tr>
<td>relu3</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,16,4,4)</td>
</tr>
</tbody>
</table>

Encoder

<table>
<thead>
<tr>
<th>Name</th>
<th>Layer Type</th>
<th>Filter Size</th>
<th>Stride</th>
<th>Output Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv_trans3</td>
<td>ConvTranspose2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>batch_norm3</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>relu3</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,32,8,8)</td>
</tr>
<tr>
<td>conv_trans2</td>
<td>ConvTranspose2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>batch_norm2</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>relu2_</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,64,16,16)</td>
</tr>
<tr>
<td>conv_trans1</td>
<td>ConvTranspose2D</td>
<td>3×3</td>
<td>2</td>
<td>(None,1,32,32)</td>
</tr>
<tr>
<td>batch_norm1</td>
<td>BatchNorm2d</td>
<td>-</td>
<td>-</td>
<td>(None,1,32,32)</td>
</tr>
<tr>
<td>relu1_</td>
<td>ReLU</td>
<td>-</td>
<td>-</td>
<td>(None,1,32,32)</td>
</tr>
</tbody>
</table>

Decoder

Table 3: Convolutional Autoencoder Architecture for MNIST, SVHN, CIFAR-10 and ImageNet-50 datasets

C Hyperparameters

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of epochs</td>
<td>100</td>
</tr>
<tr>
<td>Starting Learning Rate (LR)</td>
<td>0.001</td>
</tr>
<tr>
<td>LR Decay Milestones</td>
<td>{50}</td>
</tr>
<tr>
<td>LR Decay Factor</td>
<td>1/10</td>
</tr>
<tr>
<td>miniBatch Size</td>
<td>128</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Weight Decay</td>
<td>0.0005</td>
</tr>
<tr>
<td>[ \lambda ]</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 4: Hyper-parameters for EEC autoencoder training

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of epochs (1st increment)</td>
<td>200</td>
</tr>
<tr>
<td>No. of epochs (next increments)</td>
<td>45</td>
</tr>
<tr>
<td>Starting Learning Rate (LR)</td>
<td>0.1</td>
</tr>
<tr>
<td>LR Decay Milestones</td>
<td>{60,120,160}</td>
</tr>
<tr>
<td>LR Decay Factor</td>
<td>1/5</td>
</tr>
<tr>
<td>miniBatch Size</td>
<td>128</td>
</tr>
<tr>
<td>Optimizer</td>
<td>SGD</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.9</td>
</tr>
<tr>
<td>Weight Decay</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Table 5: Hyper-parameters for EEC classifier training
Table 6: Comparison of EEC against episodic memory approaches and approaches that use a pre-trained CNN in terms of average incremental accuracy (%) with 10 classes per increment. EEC-P uses a pre-trained CNN for feature extraction.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iCaRL (Rebuffi et al., 2017)</td>
<td>59.3</td>
</tr>
<tr>
<td>EEIL (Castro et al., 2018)</td>
<td>64.0</td>
</tr>
<tr>
<td>BiC (Wu et al., 2019)</td>
<td>64.3</td>
</tr>
<tr>
<td>LUCIR (Hou et al., 2019)</td>
<td>63.4</td>
</tr>
<tr>
<td>EEC (Ours)</td>
<td>65.6</td>
</tr>
<tr>
<td>FearNet (Kemker &amp; Kanan, 2018)</td>
<td>66.2</td>
</tr>
<tr>
<td>CAN (Xiang et al., 2019)</td>
<td>67.1</td>
</tr>
<tr>
<td>EEC-P (Ours)</td>
<td>76.7</td>
</tr>
</tbody>
</table>

D EXPERIMENTS ON CIFAR-100

For a fair comparison, the main set of approaches that we compared to are generative memory approaches. These approaches were only tested on the four datasets (Section 4). However, many episodic memory approaches and approaches that use pre-trained CNN features did not present results on the four datasets in Section 4. Hence, we test our approach (EEC) on CIFAR-100 to compare against more episodic memory approaches and approaches that use a pre-trained CNN as a feature extractor.

CIFAR-100 consists of 60,000 $32 \times 32$ images belonging to 100 object classes. There are 500 training images and 100 test images for each class. We divided the dataset into 10 batches with 10 classes per batch. During training, in each increment we provided the algorithm a new batch of 10 classes. We report the average incremental accuracy over the 10 increments as the evaluation metric. For evaluation against approaches that use a pre-trained network, we first extracted features from a pre-trained ResNet-34 on ImageNet and then applied EEC on the feature vectors instead of raw RGB images. This version of EEC is denoted as EEC-P. The hyperparameter values and training details for this experiment are presented in Appendix C. We performed this experiment 10 times with different random seeds and report the average accuracy over all the runs.

Among the episodic memory approaches, we compare against iCaRL (Rebuffi et al., 2017), EEIL (Castro et al., 2018), BiC (Wu et al., 2019) and LUCIR (Hou et al., 2019). Among the approaches that use a pre-trained CNN, we compare against FearNet (Kemker & Kanan, 2018) and the method proposed by Xiang et al. (2019) (CAN). These approaches have been introduced in Section 2.

Table 6 shows the comparison of EEC against all the above mentioned approaches. EEC beats all the state-of-the-art (SOTA) episodic memory approaches by a significant margin, even though these approaches have an unfair advantage of using stored real images of the old classes. Although the difference between EEC and other approaches (1.3%) is lower than the difference on the other four datasets. EEC-P also beats the SOTA approaches that use a pre-trained CNN by a much greater margin (9.6%). These results further show the effectiveness of our approach for mitigating catastrophic forgetting.

E COMPARISON WITH FEARNET

FearNet (Kemker & Kanan, 2018) is one of the few approaches that uses some ideas similar to EEC. In particular, FearNet uses autoencoders to reconstruct old data and stores a single centroid and covariance matrix for each of the old classes. FearNet also uses pseudorehearsal on the centroids and covariances of old classes to regenerate pseudo-encoded episodes of the old classes, which is the same as our approach. However, unlike EEC, FearNet uses a pre-trained feature extractor which is a limitation. Further, for pseudorehearsal, FearNet only stores a single centroid and covariance matrix for each of the old classes, which is not enough to capture the entire feature space of the classes. In contrast, EEC uses memory integration based clustering technique to store multiple centroids and covariance matrices per class to better capture the overall feature space of the old classes.

To illustrate the difference between the two approaches, we performed the same experiment on MNIST dataset as in Section 3.3. For FearNet, we allowed the model to store a single centroid (and
Figure 6: Visualization of the encoded episodes (a), pseudo-encoded episodes generated using 200 centroids and covariance matrices (b) and pseudo-encoded episodes generated using 2 centroids and covariance matrices (c).

Table 7: Performance of EEC for different values of $r$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$A_{30}$</th>
<th>$A_{50}$</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JT</td>
<td>57.38</td>
<td>49.88</td>
<td>315</td>
</tr>
<tr>
<td>EEC ($r = 0.25$)</td>
<td>54.85</td>
<td>45.92</td>
<td>161.48</td>
</tr>
<tr>
<td>EEC ($r = 0.2$)</td>
<td>54.04</td>
<td>45.14</td>
<td>151.496</td>
</tr>
<tr>
<td>EEC ($r = 0.15$)</td>
<td>53.38</td>
<td>44.26</td>
<td>141.412</td>
</tr>
<tr>
<td>EEC ($r = 0.1$)</td>
<td>51.67</td>
<td>42.94</td>
<td>131.53</td>
</tr>
<tr>
<td>EEC ($r = 0$)</td>
<td>45.39</td>
<td>35.24</td>
<td>111.56</td>
</tr>
</tbody>
</table>

covariance matrix) for each of the two classes and then used pseudorehearsal to generate pseudo-encoded episodes. Figure 6 shows the comparison between the original feature space (Figure 6(a)), pseudo-encoded episodes generated with 200 centroids using memory integration (Figure 6(b)) and pseudo-encoded episodes generated using a single centroid per class (2 centroids) as in FearNet (Figure 6(c)). The feature space generated by FearNet is almost circular and does not resemble the original feature space. Also, there is an overlap between the feature spaces of the two classes which will lead to similar images for the two classes which will eventually hurt classifier performance. In contrast, the pseudo-encoded episodes generated by memory integration capture the overall concept of the original feature space without any overlap between the feature spaces of the two classes. Results in Table 6 on CIFAR-100 dataset also confirm that EEC is significantly superior to FearNet in terms of the average incremental accuracy (10.5% improvement in accuracy).

F EXPERIMENT ON IMAGENET-50 WITH ORIGINAL AND RECONSTRUCTED IMAGES

In this experiment, we allow EEC to partially store original images and the rest as encoded episodes. We define a ratio of original images and encoded episodes as $r = n_{o}/N$, where $n_{o}$ is the number of original images stored per class and $N$ is the total number of images (original and reconstructed) per class. This experiment was performed with images of size $32 \times 32$. Table 7 shows the accuracy for EEC in terms of $A_{30}$ and $A_{50}$ for different values of $r$. $r = 0$ shows EEC with no original images and only the encoded episodes. By increasing $r$ to 0.1, we see a dramatic increase in accuracy, however the increase in memory is minimal (131.53 MB for $r = 0.1$ as compared to 111.56 MB for $r = 0$). As the value of $r$ continues to increase, EEC’s accuracy gets closer and closer to the JT upperbound.