# Adversarial Robustness and Catastrophic Forgetting

**Benjamin Black** [*] **David Li** [*] **Soumik Mukhopadhyay** [*]

## Abstract

Two major undesirable characteristics of deep neural networks are their vulnerability to adversarial examples and to catastrophic forgetting. While both characteristics have been independently studied in recent literature, few have examined the relationship between them. In this paper, we empirically evaluate the relationship between robustness to adversarial examples and robustness to catastrophic forgetting. We examine whether there is any interaction between some of the most common techniques in each area: projected gradient descent (PGD) and elastic weight consolidation (EWC). We also evaluate whether combining these two strategies by using PGD adversarial examples in EWC, a technique we call adversarial EWC (AEWC), can help mitigate catastrophic forgetting of adversarial robustness. Our initial results suggest that PGD adversarially robust models are equally vulnerable to catastrophic forgetting, EWC can be combined with PGD to preserve PGD adversarial robustness in the continuous learning scenario, and AEWC does not provide a consistent benefit for mitigating catastrophic forgetting of adversarial robustness.

## 1. Introduction

Typical human learning follows a life-long learning scenario, as humans learn to do numerous different things throughout their whole life. In machine learning however, artificial neural networks are expected to perform only a single task and often are unable to learn how to do multiple different tasks without explicit supervision. In particular, training neural network models in a sequential scenario for a set of different tasks causes them to forget how to perform early tasks in the sequence, a phenomenon known as catastrophic forgetting. This continuous learning scenario is particularly prevalent in reinforcement learning, were it may be desirable for to train models for multiple different environments without continual supervision on all environments. To address catastrophic forgetting, many methods have been proposed from altering the network architecture to adding regularization to the loss function. Currently, one of the most well-known methods to mitigate catastrophic forgetting is elastic weight consolidation (EWC) (Kirkpatrick et al., 2017), which calculates the contribution of each network parameter using Fisher information and applies a regularization term to the loss function on successive tasks.

In a separate line of work, neural networks have been found to be vulnerable to adversarial examples, carefully modified images which are indistinguishable to humans yet are able to cause neural networks to create grave misclassifications. Many proposals have been made to address the problem of adversarial examples, such as custom training procedures (Wong & Kolter, 2018; Gowal et al., 2019) or custom inference processes which can even be probably robust (Cohen et al., 2019; Levine & Feizi, 2019).

While the vulnerability of neural networks to catastrophic forgetting and adversarial examples have each been independently studies in numerous works, barely any research has been performed to determine if there is a link between these two areas. In adversarial machine learning, it has been found that low-level features learned by adversarially robust training can transfer between different datasets to preserve adversarial robustness (Davchev et al., 2019). Additionally, learning general features which are common between tasks aids efficient learning in the continuous learning scenario. Based on these observations, we seek to determine whether adversarially robust models are able to avoid catastrophic forgetting in the continuous learning scenario due to learning more general features and whether there is any synergistic relationship between adversarial examples and catastrophic forgetting.

In this paper, we evaluate the relationship between adversarial training techniques and catastrophic forgetting prevention techniques, focusing on elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) and projected gradient descent (PGD) (Madry et al., 2018). Specifically, we seek to answer the following questions:

1. Are PGD-adversarially trained networks equally vulnerable to catastrophic forgetting compared to nor-

[*]Equal contribution . Correspondence to: David Li <david@davidl.me>.

mally trained networks?

2. Is standard EWC able to preserve adversarial robustness for existing tasks in sequential learning?

3. Does estimating Fisher information from PGD adversaries improve EWC's performance at mitigating catastrophic forgetting for adversarial accuracy?

To answer these questions, we conduct experiments using 6 baselines and 2 datasets on the class-incremental learning scenario. For each baseline, we evaluate accuracy against catastrophic forgetting, adversarial robustness, and forgetting of adversarial robustness using PGD and EWC.

## 2. Related Works

Our work builds upon fundamental techniques from adversarial examples and catastrophic forgetting. We briefly discuss the fundamental research from both of these areas. We refer interested readers to (Parisi et al., 2019) for a review of continuous learning and (Chakraborty et al., 2018) for a survey on adversarial attacks and defenses.

### 2.1. Adversarial Examples

Initially discovered by (Goodfellow et al., 2014) and (Szegedy et al., 2013), adversarial examples are perturbed inputs which cause neural networks to misclassify their inputs. Previous research have shown that adversarial examples are transferable between models (Goodfellow et al., 2014).

Two of the earliest and most well known methods for generating adversarial examples are the fast-gradient sign method (FGSM) (Goodfellow et al., 2014) and the projected gradient descent method (PGD) (Madry et al., 2018).

The fast-gradient sign method (FGSM) (Goodfellow et al., 2014) aims to create adversarial examples by taking the largest $l_\infty$-bounded step in the direction of the positive gradient:

$$x^{(adv)} = x + \epsilon \operatorname{sign}(\nabla_x L(\theta, x, y))$$

Projected gradient descent (Madry et al., 2018) extends this idea by taking multiple steps in this direction and projecting each step to be within the $\epsilon$-neighborhood $x + S$ of the original example:

$$x^{t+1} = x^t + \Pi_{x+S} \alpha \operatorname{sign}(\nabla_x L(\theta, x^t, y))$$

To train an adversarially robust model, adversarial examples are generated during training time and the model is trained to correctly classify adversarial exampels.

Since the discovery of these two attack methods, many additional attack and defence methods have been proposed. Recently, provably robust methods have also appeared for various subsets of attacks (Wong & Kolter, 2018) and the range of attacks have broadened with attacks being generated along perceptual manifolds (Laidlaw et al., 2020).

### 2.2. Catastrophic Forgetting

When training deep neural networks sequentially on multiple tasks, they often forget earlier tasks during the training of subsequent tasks, a property known as *catastrophic forgetting*. In general, existing techniques for mitigating catastrophic forgetting can be classified as: architectural techniques, regularization techniques, and replay techniques.

Architectural techniques in continuous learning focus on expanding the neural network architecture by adding new layers or neurons as tasks accumulate. The intuition behind this idea is that learned knowledge from previous tasks can be encoded in the existing network and frozen while new knowledge can be learned in new portions of the neural network. In *Progressive Neural Networks* (Rusu et al., 2016), new columns of layers are added adjacent to the neural network as tasks accumulate while the existing columns are frozen. Via adapter layers, new layers can leverage existing knowledge learned in previous tasks.

Regularization techniques focus on constraining neural network updates such that important learned weights from previous tasks are preserved during training on new tasks. "Elastic weight consolidaton" (Kirkpatrick et al., 2017) is one of the most popular regularization methods which uses Fisher information to estimate the importance of each weight. A similar technique known as Synaptic Intelligence (Zenke et al., 2017) estimate parameter weights in an online way during training along the entire learning trajectory.

Replay techniques, or rehersal techniques, attempt to mitigate catastrophic forgetting by ensuring that the response of neural networks to specific saved inputs remains the same during training of new tasks. (Robins, 1995) propose a pseudo-rehersal strategy which performs replay to ensure that the model continues to produce the same output on random inputs when training on new tasks. Since then, new replay methods have been proposed which use generative adversarial networks (GANs) to sample training data for replay purposes (Shin et al., 2017; Wu et al., 2018).

Methods have also been proposed which apply multiple techniques to mitigate catastrophic forgetting. (Wen & Itti, 2019) propose combining EWC with task-dependent memory units into their network to capture adversarial subspaces. While they achieve a high accuracy on a 3-task split of MNIST and CIFAR-10, their approach requires knowing the task-ID at inference time. Additionally, their network scales linearly with the number of tasks due to the addition of memory units.

## 3. Training Methods

We provide a brief overview of projected gradient descent (PGD) and elastic weight consolidation (EWC). Then we discuss adversarial elastic weight consolidation (AEWC), an extension to EWC which uses adversarial examples to estimate the Fisher information.

### 3.1. PGD Training

One of the most popular methods to train models robust against adversarial examples is using projected gradient descent (PGD) adversarial examples (Madry et al., 2018). PGD adversarial examples are generated from differentiating through the model to calculate the gradient of the loss with respect to the input images. Then for each gradient descent step, a projection operation $\Pi$ is applied such that the adversarial example generated is within some $\epsilon$ ball $x + S$ around the original example $x$. For $l_\infty$ adversaries, Madry *et al.* extend the fast gradient sign method (FGSM) by using steps with size $\eta$ based on only the sign of the gradient:

$$x^{t+1} = \Pi_{x+S}\left(x^t + \eta \operatorname{sgn}\left(\nabla_x L\left(\theta, x, y\right)\right)\right) \tag{1}$$

By training on $l_\infty$ adversarial examples generated using PGD, the trained model acquires some degree of robustness against adversarial examples generated using the same method and norm.

In other norm-bounded adversaries such as $l_1$ and $l_2$ adversaries, PGD is performed using $l_1$ or $l_2$ normalized gradient steps with step-size $\eta$ rather than steps based on the sign of the gradient:

$$x^{t+1} = \Pi_{x+S}\left(x^t + \eta \frac{\nabla_x L(\theta, x, y)}{\|\nabla_x L(\theta, x, y)\|}\right) \tag{2}$$

### 3.2. EWC Training

A simple and popular way of mitigating catastrophic forgetting in a sequential learning setup is EWC (Kirkpatrick et al., 2017). EWC is a regularization method based upon a Baysian approach. In Bayesian learning, we wish to minimize the posterior $p(\theta|D_A, D_B)$ for two tasks $A$ and $B$ and their associated distributions $D_A$ and $D_B$. Using Baye's rule, this can be written as

$$\begin{aligned}\log p(\theta|D_A, D_B) = &\log p(D_B|\theta) \\ &+ \log p(\theta|D_A) - \log p(D_B)\end{aligned} \tag{3}$$

Here, $\log p(D_B|\theta)$ is our standard loss function $L_B$ for task $B$ and $\log p(D_B)$ is a constant which can be disregarded for optimization purposes. Kirkpatrick *et al.* approximate $\log p(\theta|D_A)$ using the Fisher information $F$ to arrive at the final loss function:

$$L(\theta) = L_B(\theta) + \frac{\lambda_{EWC}}{2} \sum_i F_i * (\theta_i - \theta^*_{A,i})^2 \tag{4}$$

Online EWC (Schwarz et al., 2018) extends this idea by accumulating the Fisher information over several tasks. For task $t$, online EWC first estimates the new Fisher information $\hat{F}_t$ and accumulates it as follows:

$$F_t = \hat{F}_t + \lambda_f F_{t-1} \tag{5}$$

By doing so, only one set of Fisher information weights and network weights needs to be preserved in memory during training regardless of how many tasks are trained.

### 3.3. Adversarial EWC Training

**Algorithm 1** PGD + AEWC Training Loop

Initialize model $f_\theta$
$fisher = 0 * \theta$
$optimizer = Adam()$
$N_0 = 0$
**for** task $t$ **do**
  $\theta' = \theta$
  $N_t = N_{t-1} + size$(task $t$ dataset)
  **for** iteration $i$ from 1 to $max\_iterations$ **do**
    $x$, $y$ = getMinibatch($t$)
    $x'$ = getPGDAdversary($x$, $y$, $f_\theta$)
    $logits$ = $f_\theta(x')$
    $loss$ = crossEntropy($logits$, $y$)
      $+\lambda_{EWC} * fisher * |\theta - \theta'|^2 / N_t$
    $loss$.backwards()
    $optimizer$.step()
  **end for**
  *// Compute Fisher information using PGD*
  $new\_fisher = 0$
  **for** $x$, $y$ in task $t$ dataset **do**
    $x'$ = getPGDAdversary($x$, $y$, $f_\theta$)
    $logits$ = $f_\theta(x')$
    $loss$ = crossEntropy($logits$, $y$)
    $new\_fisher = new\_fisher + loss.grad^2$
  **end for**
  $fisher$ = $\lambda_f$ * $fisher$ + $new\_fisher$
**end for**

Whereas standard EWC estimates Fisher information using gradients from standard training examples, we also evaluate whether estimating Fisher information from PGD adversarial examples can help improve adversarial accuracy in the continuous learning scenario. We formulate adversarial EWC, or AEWC, by estimating the Fisher information as follows:

$$\tilde{F} \approx \sum_{i=1}^{N} \left(\frac{\partial}{\partial \theta} \ell(\{x_{i,adv}, y_i\}|\theta)\right)^2 \tag{6}$$

*Figure 1.* Examples of CIFAR-100 images (sorted in order of labels 1-100). Each row represents examples for a single task. For each task, the network needs to classify objects from 10 categories.

Here, $\{x_{i,adv}\}$ represents adversarial examples computed from our training examples using PGD and $\ell$ represents the negative log-likelihood (cross-entropy) loss function. Our final loss will be similar to EWC, with $F$ replaced by $\tilde{F}$:

$$L(\theta) = L_B(\theta) + \frac{\lambda}{2}\sum_{i=1}^{p}\tilde{F}_{A,i} * (\theta_i - \theta^*_{A,i})^2 \qquad (7)$$

Our full training loop including PGD and AEWC regularization is outlined in Algorithm 1.

## 4. Evaluation

To evaluate the interaction and relationship between adversarial learning and catastrophic forgetting, we conduct experiments training different models with and without PGD adversarial training and EWC catastrophic forgetting regularization. For our continuous learning scenario, we use a class-incremental learning setup where a neural network is trained sequentially on new disjoint sets of image classes. For each of our baselines, we measure the classification accuracy and adversarial classification accuracy on the test set after training each task. An overview of our testing setups is shown in Table 2.

*Table 1.* Test Accuracy for Ordinary Baseline

| DATASET | TOTAL ACCURACY |
|---|---|
| CIFAR-100 | 44.23% |
| CORE50 | 67.24% |

### 4.1. Datasets

We conduct our continuous learning experiments on two datasets: CORe50 (Lomonaco & Maltoni, 2017) and CIFAR-100 (Krizhevsky, 2009). For each dataset, we use a class-incremental learning scenario where different classes of images are shown to the classifier for each task. In this scenario, the task boundaries are known at training time since we train using the ground truth classes. At test time, we assume that we do not know the task associated with each input. While there are other continuous learning scenarios, such as task-incremental learning where the task-ID is provided at test time (van de Ven & Tolias, 2019), we focus on the class-incremental learning scenario as CORe50 and CIFAR-100 are single-incremental task datasets (Maltoni & Lomonaco, 2019).

The CIFAR-100 dataset (Figure 1) consists of 32x32 pixel sized natural images of 100 object categories like beaver, ray, bottles, apples, clocks, etc. The dataset consists of 500 training and 100 testing images per category. We use 5% of the training images for validation and the rest for training. Here as well, we use a class-incremental learning setup. We divided the 100 categories into 10 tasks where each task contains 10 categories each. This division was done based on sorted class labels, ie., label 1 to 10 belong to task 1, label 11 to 20 belong to task 2, and so on. During training, the order of images may change within each task but the order of tasks remain fixed.

The CORe50 dataset consists of video recordings of 50 household objects. Each object has 11 video recordings, or sessions, 300 frames long and is considered its own class. There are 10 types of objects including plug adapters, mobile phones, scissors, light bulbs, cans, glasses, balls, markers, cups, and remote controls each with 5 instances. We use sessions 1, 2, 4, 5, 6, and 8 for training with sessions 9 and 11 for validation. Sessions 3, 7, and 10 are used for testing. For our experiments, we use an class-incremental learning setup where each task is to classify images of 10 objects from 2 types. That is, the first task is to identify images of the 5 plug adaptors and 5 mobile phones. The second task is to identify images of the 5 scissors and 5 light bulbs. During training, the order of images within each task are randomized but the order of tasks is fixed.

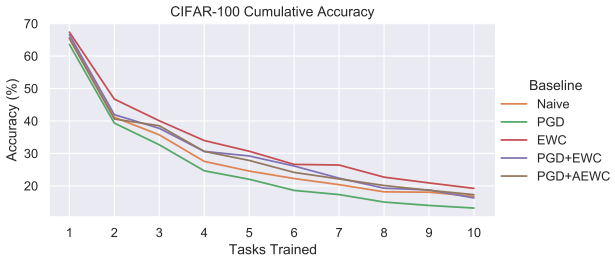


*Figure 2.* Cumulative accuracy results on the Cifar-100 dataset.

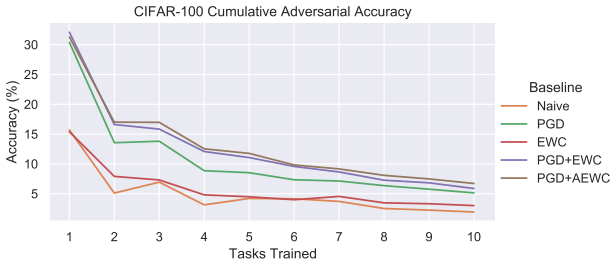


*Figure 3.* Cumulative PGD adversarial accuracy results on the Cifar-100 dataset.

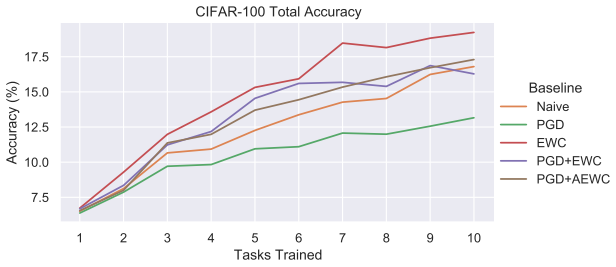


*Figure 4.* Total accuracy results on the Cifar-100 dataset.

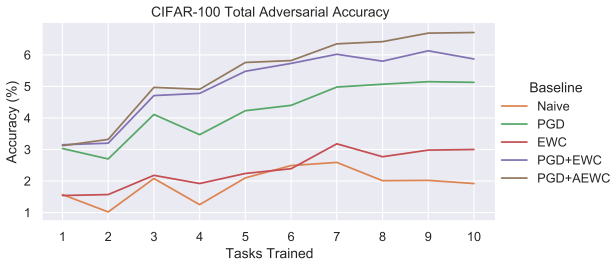


*Figure 5.* Total PGD adversarial accuracy results on the Cifar-100 dataset.

### 4.2. Models

Following (Maltoni & Lomonaco, 2019), we conduct our evaluations using variations of existing image classification networks. Each model has individual outputs for each class and hence each task. Because logit outputs are disjoint between classes and we consider logits from all possible classes, we do not need to know the class of each input at test time.

For CIFAR-100, as in (Maltoni & Lomonaco, 2019), we use the default Keras CIFAR-10 CNN as described in (Zenke et al., 2017). It is a small network that accepts $32 \times 32$ images and has 4 convolutional layers. As opposed to (Maltoni & Lomonaco, 2019), we do not pre-train the model with CIFAR-10 dataset as we could achieve comparable accuracy without pre-training.

For CORe50, we modify GoogLeNet (Szegedy et al., 2015) to accept $128 \times 128$ images by setting the first convolutional layer to have stride $1$ and padding $0$ as done in (Maltoni & Lomonaco, 2019). During fine-tuning of the pretrained model, we ignore the $2$ intermediate outputs and only consider the logits produced at the end of the network. We train for $4$ epochs per task and with the Adam optimizer's learning rate set to $0.001$.

### 4.3. Baselines

To evaluate our hypotheses, we train multiple classifiers according to the following baselines:

1. Ordinary training with all classes
2. Sequential training (Naive)
3. Sequential EWC training (EWC)
4. Sequential PGD training (PGD)
5. Sequential PGD and EWC training (PGD+EWC)
6. Sequential PGD training and Adversarial EWC training (PGD+AEWC)

Ordinary training represents the scenario where images of all classes are presented to the network during training. This provides an optimal baseline for all sequential models. Our primary baseline for catastrophic forgetting is (EWC) and our primary baseline for adversarial training is (PGD). In PGD+EWC, the model is trained using adversarial examples but the Fisher information from EWC is computed using clean examples. In PGD+AEWC, the Fisher information is computed from only adversarial examples.

For all sequential baselines we train using cross-entropy loss taken with respect only to the classes within each task. For instance, if task $2$ involves classifying images from classes $10 - 19$, then only the logits for those classes are input into the cross-entropy loss during training for task $2$. However, at test time, we use all logits. Unless otherwise stated, accuracy measures are computed using all logits, including those represented classes from other tasks.

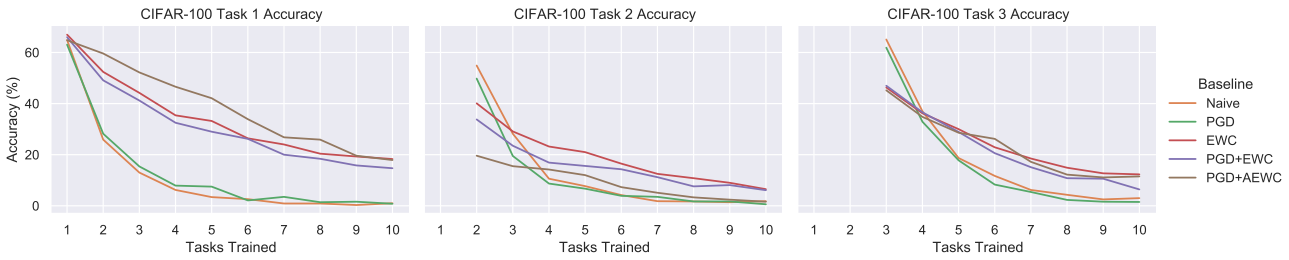


*Figure 6.* Task-wise accuracies for Task 1, 2 and 3 during sequential training on the Cifar-100 dataset.

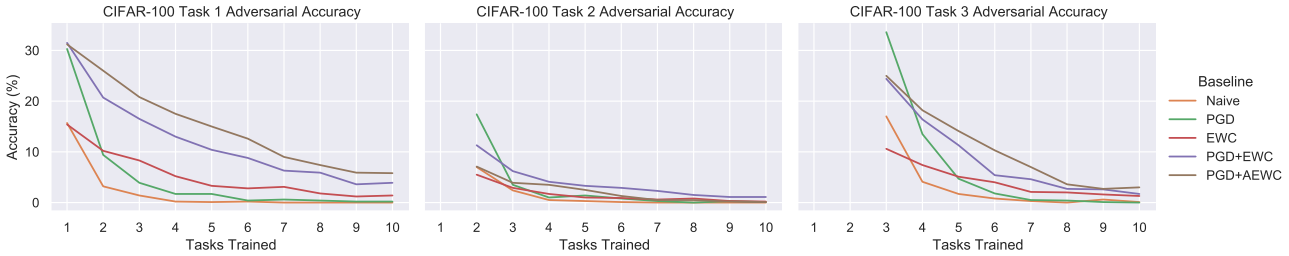


*Figure 7.* Task-wise PGD adversarial accuracies for Task 1, 2 and 3 during sequential training on the Cifar-100 dataset.

For PGD adversarial training, we present results on training and testing using $l_\infty$ adversarial examples. We use a step size of 1.5 and an $\epsilon$ size of $8/255 \approx 0.031$ following (Madry et al., 2018). For EWC training, we accumulate Fisher information weighing by the number of examples in each task which is roughly equal. That is, we use $\lambda_f = 1$ to accumulate Fisher information and then divide the Fisher by the total number of examples in the loss function as shown in Algorithm 1.

### 4.4. Metrics

We present our catastrophic forgetting results in three formats: cumulative accuracy on a growing test set, total accuracy on a fixed test set, and per-task accuracy. For our cumulative baseline, we present only the accuracy of the final model since training is not sequential. Cifar-100 results are presented in the main text while CORe-50 results are presented in the supplementary material.

Our first format is cumulative accuracy on a growing test set. In this format, after each task, we compute the accuracy only on test examples from classes we have already trained on. Furthermore, the predicted class is computed only using trained logits. This accuracy represents the scenario where new classes or tasks are accumulated over time and added into the model in the form of new neurons in the last layer, however training examples from previous classes are no longer available for replay. Our results for cumulative accuracy are shown in Figure 2 with adversarial cumulative accuracy in Figure 3.

*Table 2.* Testing Setups

| | CIFAR-100 | CORE50 |
|---|---|---|
| CLASSES | 10 * 10 | 10 * 5 |
| TASKS | 10 | 5 |
| MODEL | CIFAR-10 CNN | GOOGLENET |
| LAYERS | 6 | 22 |
| EPOCHS | 50 | 4 |
| BATCH SIZE | 256 | 32 |
| OPTIMIZER | ADAM(0.001) | ADAM(0.001) |
| $\lambda_{EWC}$ | 5E5 | 5E5 |

Our second format is the total accuracy on a fixed test set. In this format, we present results on the entire dataset including test examples from yet-to-be seen tasks. The prediction is computed from all logits in the final layer at the end of each task. This represents accuracy for the scenario where the input data is improperly sorted. Our results for total accuracy are shown in Figure 4 with adversarial total accuracy in Figure 5. Total accuracy for the ordinary training baseline is shown in Table 1.

Our final format is per-task accuracy. Per-task accuracy allows us to visualize how the network forgets earlier tasks as it is trained on new tasks. Each plot shows the results on all test set examples for a single task. Results are calculated without slicing logits to only previously trained tasks so performance on existing tasks may increase as weights for the final layer represented new classes are trained. Our results for per-task accuracy are shown in Figure 6 with adversarial per-task accuracy Figure 7.

## 5. Discussion

### 5.1. Results

Looking at Cifar-100 and CORe50 results, we see that EWC alone achieves the best cumulative and total accuracy between all sequential baselines as shown for Cifar-100 in Figure 2, Figure 4. Our PGD baseline yields lower performance than our naive baseline during training of the first task which is shown in task 1 of Figure 2. This is expected as adversarial training comes with a small penalty to clean accuracy. We observe that our PGD robust baseline is equally vulnerable to catastrophic forgetting with PGD forgetting at the same rate as naive baseline in Figure 6.

When it comes to adversarial accuracy, i.e. robustness against adversarial examples, the baselines trained with PGD are yield better accuracies than baselines trained without PGD. However without EWC, the PGD baselines' adversarial accuracy quickly drops below that of our EWC baseline which is not trained for adversarial robustness as seen in Figure 7. When comparing catastrophic forgetting of our PGD baseline and our PGD+EWC baseline, we see that computing fisher information based on clean examples allows EWC to preserve adversarial robustness while training on new tasks, as shown in Figure 7. Furthermore, we observe that adversarial accuracy declines at the same rate as clean accuracy in the continuous learning scenario for our PGD+EWC baseline. This suggests that EWC on clean examples is sufficient to preserve adversarial accuracy.

From our results, the PGD+AEWC baseline does not yield a consistent benefit over the PGD+EWC baseline. For Cifar-100, our PGD+AEWC baseline yields slightly better accuracies on both clean examples and adversarial examples as shown in Figure 3 and Figure 5. However, the opposite is true for CORe50 as shown in the partial and total adversarial accuracy plots for CORe50.

### 5.2. Limitations

Although we evaluate with two different datasets and two models, our evaluation currently focuses only on $l_\infty$ PGD adversaries. Furthermore we only consider PGD and EWC as they are the most popular methods in the areas of adversarial examples and catastrophic forgetting respectively. We leave further exploration with additional methods to future work.

In our experiments, we also observe a trade-off between EWC and naive training. As EWC regularization constrains the optimization process in additional tasks, the initial clean accuracy for subsequent tasks is reduced compared to the naive baseline. This trade-off is tuned in the weight $\lambda_{EWC}$ of EWC regularization in the loss function. Furthermore, the empirical Fisher approximation is not always a good estimate of the true Fisher information as noted in (Kunstner et al., 2019), meaning the final EWC regularization may not be a good estimate for the posterior $p(\theta, D_A)$.

## 6. Conclusion

In this paper we evaluate the relationship between adversarial robustness and catastrophic forgetting. To accomplish this, we compare 6 training methods based on various combinations of projected gradient descent and elastic weight consolidation in the class-incremental learning scenario. Based on our empirical experiments, we observe the following:

1. PGD adversarially robust models are equally vulnerable to catastrophic forgetting. Furthermore, adversarial accuracy decays at the same rate as clean accuracy.
2. EWC when computed on clean examples is able to preserve adversarial accuracy in PGD-trained models in sequential learning.
3. Performing adversarial EWC by computing Fisher information from adversarial examples does not yield a consistent benefit over performing EWC on clean examples.

While we examine one of the most popular methods for adversarial training and one of the most popular regularization methods for mitigating catastrophic forgetting, our evaluations are not comprehensive. Future work may focus on examining whether there are any interactions between other forms of adversarial robustness and other forms of mitigating catastrophic forgetting.

## References

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cohen19c.html.

Davchev, T., Korres, T., Fotiadis, S., Antonopoulos, N., and Ramamoorthy, S. An empirical evaluation of adversarial robustness under transfer learning, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Dvijotham, K. D., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P.

Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/content/114/13/3521.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Kunstner, F., Hennig, P., and Balles, L. Limitations of the empirical fisher approximation for natural gradient descent. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 4156–4167. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/46a558d97954d0692411c861cf78ef79-Paper.pdf.

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models, 2020.

Levine, A. and Feizi, S. Robustness certificates for sparse adversarial attacks by randomized ablation, 2019.

Lomonaco, V. and Maltoni, D. Core50: a new dataset and benchmark for continuous object recognition. volume 78 of *Proceedings of Machine Learning Research*, pp. 17–26. PMLR, 13–15 Nov 2017. URL http://proceedings.mlr.press/v78/lomonaco17a.html.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Maltoni, D. and Lomonaco, V. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56 – 73, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.03.010. URL http://www.sciencedirect.com/science/article/pii/S0893608019300838.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54 – 71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL http://www.sciencedirect.com/science/article/pii/S0893608019300231.

Robins, A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. doi: 10.1080/09540099550039318. URL https://doi.org/10.1080/09540099550039318.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks, 2016.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/schwarz18a.html.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 2990–2999. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/0efbe98067c6c73dba1250d2beaa81f9-Paper.pdf.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

van de Ven, G. M. and Tolias, A. S. Three scenarios for continual learning, 2019.

Wen, S. and Itti, L. Overcoming catastrophic forgetting through weight consolidation and long-term memory, 2019. URL https://openreview.net/forum?id=BJlSHsAcK7.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. volume 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/wong18a.html.

Wu, C., Herranz, L., Liu, X., wang, y., van de Weijer, J., and Raducanu, B. Memory replay gans: Learning to generate new categories without forgetting. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 5962–5972. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a57e8915461b83adefb011530b711704-Paper.pdf.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/zenke17a.html.