TASK AGNOSTIC REPRESENTATION CONSOLIDATION: A SELF-SUPERVISED BASED CONTINUAL LEARNING APPROACH

Prashant Bhat, Bahram Zonooz*, Elahe Arani*
Advanced Research Lab, NavInfo Europe, The Netherlands
{prashant.bhat, elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

ABSTRACT

Continual learning (CL) over non-stationary data streams remains one of the long-standing challenges in deep neural networks (DNNs) as they are prone to catastrophic forgetting. CL models can benefit from self-supervised pre-training as it enables learning more generalizable task-agnostic features. However, the effect of self-supervised pre-training diminishes as the length of task sequences increases. Furthermore, the domain shift between pre-training data distribution and the task distribution reduces the generalizability of the learned representations. To address these limitations, we propose Task Agnostic Representation Consolidation (TARC)[1], a two-stage training paradigm for CL that intertwines task-agnostic and task-specific learning whereby self-supervised training is followed by supervised learning for each task. To further restrict the deviation from the learned representations in the self-supervised stage, we employ a task-agnostic auxiliary loss during the supervised stage. We show that our training paradigm can be easily added to memory- or regularization-based approaches and provides consistent performance gain across more challenging CL settings. We further show that it leads to more robust and well-calibrated models.

1 INTRODUCTION

Computational systems that operate in the real world are exposed to the continuous stream of non-i.i.d data and are required to learn multiple tasks sequentially. Learning from a stream of non-i.i.d data causes the new information to overwrite the previously learned knowledge in the neural network leading to catastrophic forgetting. Several approaches have been proposed in the literature to address the problem of catastrophic forgetting in CL. Replay-based methods (Robins, 1995; Buzzega et al., 2020; Ratcliff, 1990b) store and replay a subset of samples belonging to previous tasks along with the current batch of samples. Regularization-based methods (Schwarz et al., 2018; Zenke et al., 2017) insert a regularization term to consolidate the previous knowledge when training on new tasks. These methods avoid using memory buffer altogether alleviating the memory requirements (Delange et al., 2021).

Although aforementioned approaches have been partially successful in mitigating the catastrophic forgetting, they still suffer from several shortcomings. Since CL methods rely extensively on cross-entropy loss for classification tasks, they are prone to lack of robustness to noisy labels (Sukhbaatar et al., 2015) and the possibility of poor margins (Elsayed et al., 2018) affecting their ability to generalize across tasks. Furthermore, the optimization objective in cross-entropy loss encourages learning of representations optimal for the current task sidelining the representations that might be necessary for the future tasks, resulting in prior information loss (Zhang et al., 2020). Also, the representations of the observed tasks drift when new tasks appear in the incoming data stream exacerbating the backward interference (Caccia et al., 2021). Therefore, we assume task-specific learning is the root cause of these problems and is not well equipped to deal with catastrophic forgetting.

We hypothesize that learning task-agnostic representations in addition to task-specific representations can potentially mitigate aforementioned problems by improving forward facilitation while reducing backward interference in CL. Self-supervised pre-training has been widely regarded to learn task-agnostic generalizable representations (He et al., 2020; Grill et al., 2020). In many real-world CL scenarios however, the data distribution of the future tasks is not known beforehand. Thus, pre-training on a different data distribution often leads to domain shift subsequently reducing the generalizability of the learned representations. Furthermore, longer task sequences diminish the effect of self-supervised pre-training as the learned representations are repeatedly overwritten to maximize the performance on the current task. Domain shift, non-availability of pre-training data and computational costs associated with pre-training can be avoided if task-agnostic learning is integrated into CL training.

* Shared last author.
[1] Code can be found at: https://github.com/NeurAI-Lab/TARC.

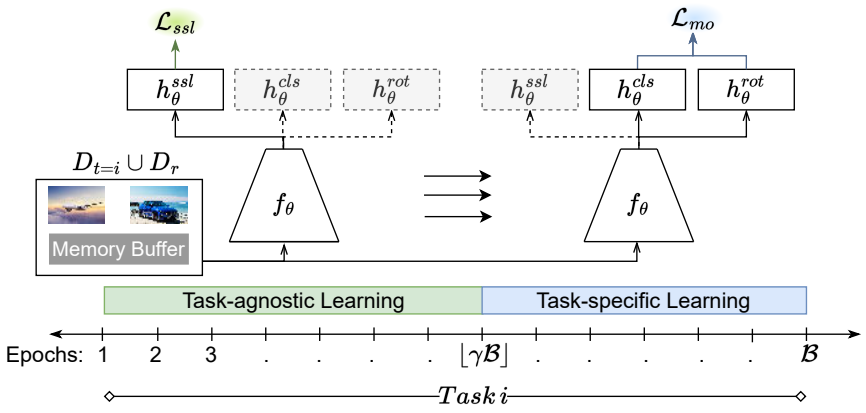


Figure 1: Proposed method. The training budget $\mathcal{B}$ for each task i is divided into task-agnostic and task-specific learning phases.

We therefore propose an online Task-Agnostic Representation Consolidation (TARC), a two-stage generic CL training paradigm that intertwines task-specific and task-agnostic learning through self-supervised learning and multi-objective learning. Our method integrates task-agnostic learning into CL training thereby avoiding problems associated with self-supervised pre-training. Our contributions are as follows:

- We propose a two-stage generic CL framework that intertwines task-agnostic and task-specific learning to train robust, well-calibrated models.
- We extend our method to state-of-the-art replay-based and regularization-based methods. Our method outperforms the baseline in the more challenging CL scenarios.
- We provide extensive analyses including the bias towards recent tasks and robustness to noisy labels to highlight the additional benefits our method brings without any explicit constraints.

2 Related Works

2.1 Continual Learning

Early works attempted to mitigate the effect of catastrophic forgetting by replaying a subset of training data from the previous tasks stored in the replay buffer alongside samples from the new task (Robins, 1993). Samples in the replay buffer are used as model inputs for rehearsal and/or for constraint optimization of the new task (Delange et al., 2021; Arani et al., 2022). Experience Replay (ER) (Ratcliff, 1990a; Robins, 1995) explicitly interleaves the old samples from the replay buffer with the current batch of samples while training on the new tasks.

Regularization-based methods on the other hand, avoid storing raw inputs, thus alleviating the memory requirements (Delange et al., 2021). Instead, a regularization term is introduced to consolidate the previous knowledge when learning a new task. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) estimates the importance of all parameters of a neural network and penalizes the changes to the important parameters in the later tasks. However, EWC is not scalable with large number of tasks as the number of regularization terms grow linearly with the number of tasks. Online-EWC (oEWC) (Schwarz et al., 2018) modified the original algorithm to avoid the linear growth in the computational requirements. Synaptic Intelligence (SI) (Zenke et al., 2017) introduced brain-inspired intelligent synapses that accumulate task relevant information over time, and exploit this information to rapidly store new memories without forgetting old ones.

Remembering for the Right Reasons (RRR) (Ebrahimi et al., 2020) proposed a training paradigm that additionally stores model explanations for each example and encourages explanation consistency throughout CL training. RRR is a generic framework that can be extended to any memory or regularization-based approaches, similar to our proposed framework. Therefore, we consider RRR as a competing approach in our analysis.

2.2 SELF-SUPERVISED LEARNING

Self-supervised learning methods broadly fall into one of two categories: generative and contrastive methods (Liu et al., 2021). Context-instance contrastive methods learn the representation of local features associative to the representation of the global context. More recently, InstDisc (Wu et al., 2018), MoCo (He et al., 2020) and SimCLR (Chen et al., 2020) leverage instance discrimination as a pretext task. SimCLR learns representations by contrasting semantically similar (positive) and dissimilar (negative) pairs of data samples such that similar pairs have the maximum agreement via a contrastive loss in the latent space (Chen et al., 2020). However, as with metric learning, contrastive learning benefits from hard negatives (Cai et al., 2021; Robinson et al., 2020). SupContrast (Khosla et al., 2020) extends SimCLR to a fully supervised setting to effectively leverage the label information. It eliminates the need for hard negative mining by using several positive and negative samples per anchor sample. Also, SupContrast consistently outperforms cross-entropy loss on large scale classification problems (Khosla et al., 2020).

Each task in CL is usually constrained by a training budget, limiting model's ability to consolidate the representations for the current task. Pre-training has been traditionally used to offset the limited data/training time through transfer learning. Gallardo et al. (2021) empirically showed that self-supervised pre-training yields representations that generalize better across tasks than supervised pre-training in CL. Owing to additional computational effort, some of the approaches, e.g. (Zhang et al., 2020; Caccia et al., 2021; Mazumder et al., 2021; Su et al., 2020), relinquished pre-training altogether and employed auxiliary pretext task to boost task-agnostic learning. However, self-supervised learning as an auxiliary loss improved the baseline only marginally.

In this work, we adapt ER, oEWC and SI to our CL training paradigm. We employ SupContrast as a task-agnostic learning objective and rotation prediction as an auxiliary loss alongside cross-entropy loss. Our method differs from the above methods in two ways: Firstly, we do not employ any pre-training, instead achieve the same objective through task-agnostic learning during CL training. Secondly, the use of auxiliary self-supervised loss in our training paradigm is necessitated by the need to preserve task-agnostic representations. Section 6 provides a comparison of different variants of self-supervision aided CL methods.

3 PROPOSED METHOD

Continual learning normally consists of T sequential tasks. During each task, the input samples and the corresponding labels (x_t, y_t) are drawn from the task-specific data distribution $\mathcal{D}_t$. Our continual learning model consists of a backbone network f_θ and three heads h_θ^{ssl}, h_θ^{cls} and h_θ^{rot}. These heads correspond to task-agnostic head, classification head and rotation head respectively. The continual learning model $g_\theta = \{f_\theta, h_\theta^{ssl}, h_\theta^{cls}, h_\theta^{rot}\}$ is sequentially optimized on one task at a time up to the current one $t \in 1, ..., T_c$. The objective function is therefore as follows:

$$\mathcal{L}_{T_c} = \sum_{t=1}^{T_c} \mathop{\mathbb{E}}_{(x_t, y_t)\sim\mathcal{D}_t} \left[l_{ce}(\sigma(h_\theta^{cls}(f_\theta(x_t))), y_t) \right], \tag{1}$$

where σ is a softmax function and l_{ce} is a classification loss, generally a cross-entropy loss. Continual learning is especially challenging since the data from the previous tasks are unavailable i.e at any point during training, the model g_θ has access to the current data distribution $\mathcal{D}_t$ alone. As the objective function in Eq. (1) is solely optimized for the current task, it leads to overfitting on the current task and catastrophic forgetting of older tasks. Replay-based methods sought to address this problem by storing a subset of training data from previous tasks and replaying them alongside current task samples. For replay-based methods, Eq. (1) can thus be rewritten as:

$$\mathcal{L}_{cls} = \mathcal{L}_{T_c} + \mathop{\mathbb{E}}_{(x,y)\sim D_r} \left[l_{ce}(\sigma(h_\theta^{cls}(f_\theta(x))), y) \right], \tag{2}$$

where D_r represents the distribution of samples stored in the buffer. Although cross-entropy loss is widely used for classification tasks in continual learning, it suffers from several shortcomings such as lack of robustness to noisy labels (Sukhbaatar et al., 2015) and the possibility of poor margins (Elsayed et al., 2018), affecting the ability to generalize across tasks. As evidenced in (Gallardo et al., 2021), self-supervised learning offers an alternative by learning task-agnostic, robust, and generalizable representations. We hypothesize that a two-stage training consisting of task-agnostic learning followed by task-specific learning can help bridge the aforementioned shortcomings without the need for pre-training.

3.1 TASK-AGNOSTIC LEARNING

We aim to learn task-agnostic representations that are robust and generalizable across multiple tasks. Solving pretext tasks created from known information can help in learning representations useful for downstream tasks. Inspired by

Algorithm 1 The Proposed Method

input: training budget $\mathcal{B}$ and ratio $0 < \gamma < 1$, data streams $\mathcal{D}_t$ and $\mathcal{D}_r$
for all tasks $t \in \{1, 2, .., T\}$ **do**
  for $e = 0 : \lfloor \gamma \mathcal{B} \rfloor$ **do** ▷ **Task-agnostic Learning**
    for minibatch $(X_m, Y_m)_{m=1}^{M} \in \mathcal{D}_t \cup \mathcal{D}_r$ **do**
      Draw augmentation functions $a', a'' \sim A$
      $X = \{a'(X_m), a''(X_m)\}$
      $Z = h_\theta^{ssl}(f_\theta(X))$
      $\mathcal{L}_{ssl} = \frac{1}{2\mathcal{N}} \sum_{k=1}^{\mathcal{N}} [\, l_{ssl}(2k-1, 2k) + l_{ssl}(2k, 2k-1)]$
      Update the networks f_θ and h_θ^{ssl}
  for $e = \lfloor \gamma \mathcal{B} \rfloor : \mathcal{B}$ **do** ▷ **Task-specific Learning**
    for minibatch $(X_m, Y_m)_{m=1}^{M} \in \mathcal{D}_t \cup \mathcal{D}_r$ **do**
      Draw rotation $a \sim \{0^0, 90^0, 180^0, 270^0\}$
      $F = f_\theta(a(X_m))$
      $Z^{cls}, Z^{rot} = h_\theta^{cls}(F), h_\theta^{rot}(F)$
      Compute L_{mo}
      Update the networks f_θ, h_θ^{cls} and h_θ^{rot}
      Update replay buffer $\mathcal{D}_r$
return model f_θ with h_θ^{cls}

the recent advancements in contrastive self-supervised representation learning, we cast task-agnostic learning as an instance-level discrimination task. For a set of $\mathcal{N}$ randomly sampled images, each image is passed through two sets of augmentations $a', a'' \sim \mathcal{A}$ producing $2\mathcal{N}$ images per minibatch. Therefore, each image within $2\mathcal{N}$ samples will have a unique positive pair and $2(\mathcal{N}-1)$ negative samples. Let $Z = h_\theta^{ssl}(f_\theta(.))$ be a projection matrix of $2N$ augmented samples and $sim(.,.)$ denote cosine similarity. The self-supervised contrastive loss (Chen et al., 2020) for a positive pair of examples (i, j) is defined as:

$$l(i,j) = -log \frac{e^{sim(z_i, z_j)/\tau_c}}{\sum_{k=1}^{2\mathcal{N}} \mathbb{1}_{[k \neq i]} e^{sim(z_i, z_k)/\tau_c}}. \tag{3}$$

Contrastive learning in Eq. (3) learns visual representations by contrasting semantically similar (positive) and dissimilar (negative) pairs of data samples such that similar pairs have the maximum agreement via a contrastive loss in the latent space through Noise Contrastive Estimation (NCE) (Gutmann & Hyvärinen, 2010). Given a limited training time for each task, it is pertinent to learn task-agnostic features that are in line with the class boundaries to avoid interference in the downstream tasks. Following Khosla et al. (2020), we adapt Eq. (3) to leverage label information. Within each minibatch, normalized embeddings belonging to the same class are pulled together while those belonging to other classes are pushed away in the latent space.

$$l_{ssl}(i,j) = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} log \frac{e^{sim(z_i, z_p)/\tau_c}}{\sum_{k=1}^{2\mathcal{N}} \mathbb{1}_{[k \neq i]} e^{sim(z_i, z_k)/\tau_c}}, \tag{4}$$

where $\mathcal{P}(i)$ is a set of indices of samples belonging to the same class as the positive pair and $|\mathcal{P}(i)|$ is its cardinality. While Eq. (4) is a simple extension to contrastive loss, it eliminates the need for hard negative mining.

3.2 Task-specific learning

To align the task-agnostic representations to the current task, we train the classification head h_θ^{cls} with cross-entropy objective defined in Eq. (2). However, the interplay between task-agnostic and task-specific learning objectives can lead to sharp drift in the feature space and erode the generic representations learned during task-agnostic learning.

Multi-objective learning offers a viable solution to address this trade-off. Multi-objective learning can be thought of as a form of inductive transfer and is known to improve generalization (Caruana, 1997). It is also data efficient as multiple objectives are learned simultaneously using shared representations. However, simultaneous learning of multiple objectives poses new design and optimization challenges, and choosing which objectives that should be learned together is in itself a non-trivial problem (Crawshaw, 2020). We employ rotation prediction as an auxiliary loss to preserve the task-agnostic features. During task-specific stage of each task, input samples $x \in \mathcal{D}_t \cup \mathcal{D}_r$ are

Table 1: Comparison of CL models across various continual learning scenarios. We provide the average top-1 accuracy (%) across all tasks after continual learning training.

Bufer size	Method	Class-IL				Domain-IL	General-IL
		CIFAR-10	CIFAR-100	TinyImageNet	STL-10	R-MNIST	MNIST-360
-	Joint	91.55 ±0.51	70.56 ±0.28	59.77 ±0.47	78.51 ±2.98	96.52 ±0.12	82.05 ±0.62
	SGD	19.55 ±0.06	17.49 ±0.28	8.00 ±0.14	17.35 ±1.52	70.76 ±5.61	21.09 ±0.21
-	oEWC	-	-	-	-	76.06 ±4.51	-
	oEWC + TARC	-	-	-	-	**89.57** ±1.18	-
-	SI	-	-	-	-	81.39 ±4.22	-
	SI + TARC	-	-	-	-	**87.91** ±1.19	-
200	ER	47.71 ±2.69	21.28 ±0.84	8.57 ±0.16	44.33 ±1.12	85.18 ±0.83	52.36 ±1.92
	ER + TARC	**53.23** ±1.00	**23.48** ±0.10	**9.57** ±0.12	**60.61** ±0.96	**90.10** ±0.10	**69.54** ±1.94
500	ER	61.97 ±1.77	27.85 ±0.39	10.33 ±0.24	58.07 ±1.06	88.03 ±0.94	66.27 ±2.82
	ER + TARC	**67.41** ±0.94	**31.50** ±0.40	**13.77** ±0.17	**67.96** ±1.84	**90.00** ±0.27	**72.25** ±3.25
5120	ER	**83.99** ±0.52	**53.64** ±0.55	27.48 ±0.58	74.83 ±0.94	**93.42** ±0.77	69.55 ±1.57
	ER + TARC	82.21 ±0.38	52.30 ±0.20	**32.04** ±0.37	**75.28** ±1.06	92.29 ±0.28	**73.01** ±1.74

rotated by a fixed angle in addition to other transformations. The learning objective is to match task-specific ground truths $y \in \mathcal{D}_t \cup \mathcal{D}_r$ as well as auxiliary ground truths $y^a \in \{0^0, 90^0, 180^0, 270^0\}$, i.e.

$$\mathcal{L}_{mo} = \alpha \mathcal{L}_{cls} + \beta \mathop{\mathbb{E}}_{x \sim \mathcal{D}_t \cup \mathcal{D}_r} \left[l_{ce}(\sigma(h_\theta^{rot}(f_\theta(x))), y^a) \right] \tag{5}$$

where α and β are hyperparameters for adjusting the magnitudes of two losses. Algorithm 1 summarizes the proposed method in detail.

4 Experimental Results

4.1 Empirical results

Table 1 provides a comparison of replay-based baseline ER (Ratcliff, 1990a; Robins, 1995) versus our method across various CL scenarios. We also provide a lower bound, *SGD*, without resorting to any measures to address catastrophic forgetting and an upper bound, *Joint*, where all tasks are trained together. We can make several observations from Tab. 1: (i) Our method shows a strong performance in majority of the benchmarks, especially when the buffer size is small. We argue that our method is able to retain information more efficiently than ER by learning task-agnostic representations. (ii) In case of a larger buffer size, our method is behind ER by a small margin on smaller datasets. This is also true for other leading methods such as Buzzega et al. (2020) where the gap in performance diminishes for large buffer sizes. (iii) Due to high class-to-buffer ratio (200/200, 200/500), TinyImageNet dataset is one of the hardest benchmarks under Class-IL setting. Our method outperforms the baseline across all buffer sizes. (iv) The relative improvement compared to the baseline improves with the increase in image size. Task-agnostic learning is able to discriminate better for STL-10 and TinyImageNet when the images are comparatively larger in size.

The benefits our training paradigm is not limited to replay-based methods alone. To prove the generalizability of our method, we adapt two regularization-based approaches (oEWC (Schwarz et al., 2018) and SI (Zenke et al., 2017)) and provide their results in Domain-IL scenario. We employ the same training schedule as in Algorithm 1 except that the number of epochs is limited to 2 and task-specific learning also includes a regularization term both in oEWC and SI. We also do not use any replay for these methods. R-MNIST under Domain-IL consists of 20 subsequent tasks and is more challenging dataset as it tests forward facilitation across longer task sequence. On R-MNIST, Our method outperforms the baseline by a large margin indicating that our framework works across different CL approaches.

Table 1 also presents the comparison of ER and TARC in more challenging General-IL scenario. It is worth noting that General-IL setting presents both sharp (changes in class) and smooth distribution (rotation) shifts, and involves recurring classes in subsequent sequences which makes the transfer of knowledge from previous occurrences important. Our method is able to leverage positive transfer when revisiting the previous task and outperforms the ER across different buffer sizes. Similar to prior results, the improvement is highest among low buffer regimes indicating the superiority of our method.

Table 2: Forward transfer and backward transfer on R-MNIST dataset.

Buffer size	Method	Forward Transfer	Backward Transfer	Method	Forward Transfer	Backward Transfer
200	ER	62.33±4.47	-12.30 ±0.89	oEWC	52.41±6.24	-21.55 ±4.68
	ER + TARC	**75.65**±2.06	**-1.37**±0.09	oEWC + TARC	**74.05**±2.02	**-1.56** ±1.22
500	ER	65.92±3.02	-9.26±1.04	SI	52.79 ±6.76	-17.82 ±3.75
	ER + TARC	**76.01**±1.52	**-1.51**±0.20	SI + TARC	**71.81** ±0.99	**-1.40** ±1.09
5120	ER	73.07±1.37	-3.61±0.73			
	ER + TARC	**78.79**±1.50	**-0.11**±0.14			

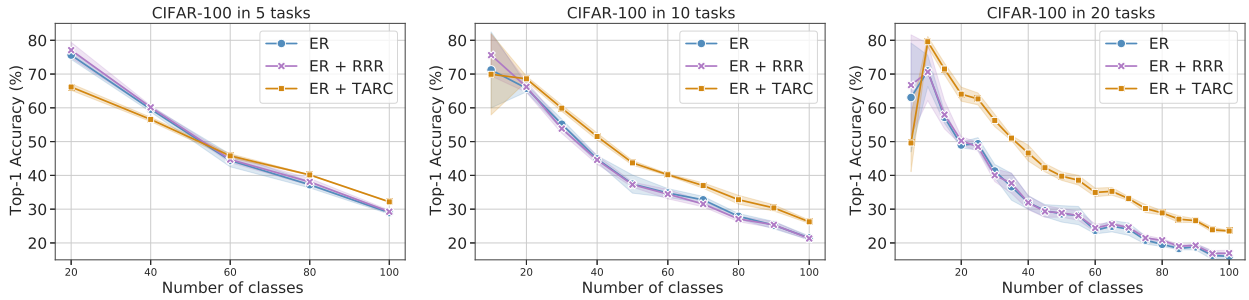


Figure 2: Comparison of our method against RRR across different number of tasks in CIFAR-100.

4.2 Forward and Backward Transfer

Following Lopez-Paz & Ranzato (2017), we compute forward transfer as the difference between the accuracy just before starting training on a given task and the one of the random-initialized network. While one can argue that learning to classify unseen classes is desirable, Class-IL shows distinct classes in distinct tasks, which makes transfer impossible. On the contrary, forward transfer can be relevant for Domain-IL scenarios, provided that the input transformation is not disruptive. As is the case with the R-MNIST, it requires the CL model to classify all digits for 20 subsequent tasks, with images rotated by a random angle in the interval $[0, \pi)$. Therefore, the positive forward transfer is not only plausible but also a highly desirable property in this setting. As far as the backward transfer is concerned, we compute the difference between the current accuracy and its best value for each task. Table 2 presents the forward and backward transfer results on R-MNIST, averaged across all task. As can be seen, TARC exhibits strong forward transfer and improved backward transfer when compared to the respective baselines.

4.3 Comparison with RRR

Similar to our method, RRR proposed a training paradigm that is extensible to existing CL approaches. Figure 2 presents comparison of RRR against our method on CIFAR-100 under different number of task sequences. Following RRR, we tweak our training schedule and update the replay buffer once after each task. Gradient Class Activation Mapping (Selvaraju et al., 2017) is used for generating explanations for the buffered images. As the length of task sequence increases, our method clearly outperforms both RRR and ER. Our method is able to consolidate the generalizable features better and mitigate the effect of catastrophic forgetting. Unlike pre-training, the effect of task-agnostic learning does not diminish with longer task sequences. Since both TARC and RRR are generic frameworks, a combination of them might improve the our proposed method even further.

5 Model Analyses

We attempt to gain insights into the working mechanism of our method and elaborate on the additional advantages it brings without any explicit constraints.

Robustness to natural corruptions: Autonomous CL agents operating in the real world are exposed to ever changing environments, often influenced by illumination and weather changes. Therefore, it is pertinent for CL agents to be robust to data distributions with natural corruptions. In this section, we evaluate our method against common image

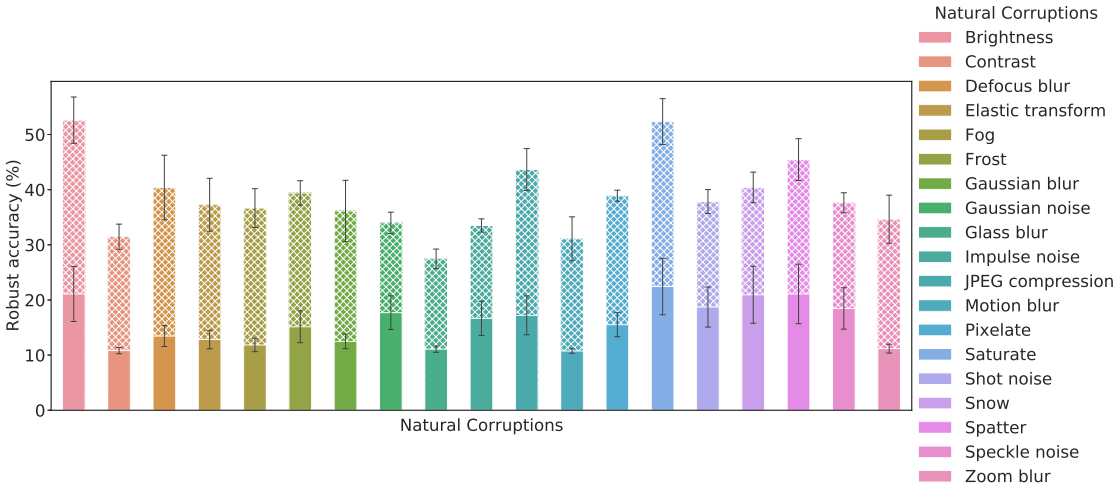


Figure 3: Robustness to natural corruptions. The unshaded part of each bar represents the robust accuracy of ER against 19 natural input corruptions. The shaded part of each bar represents the improvement our method (ER + TARC) achieves over the ER baseline.

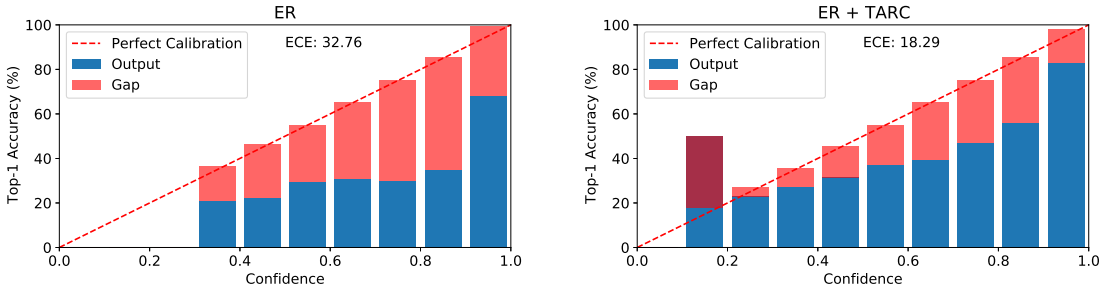


Figure 4: Confidence estimates and corresponding Expected Calibration Error (ECE) of CIFAR-10 trained CL models. Lower ECE is better. Our method is well calibrated with confidence estimates closer to the perfect calibration compared to ER.

corruptions using CIFAR-10-C (Hendrycks & Dietterich, 2018). Models are trained using clean CIFAR-10 with a buffer size of 500 and tested on CIFAR-10-C. Mean Corruption Error (mCE) and Relative Mean Corruption Error (Relative-mCE) are commonly used metrics to evaluate the performance under natural corruptions (Hendrycks & Dietterich, 2018). Figure 3 shows robustness to 19 different corruptions averaged over five severity levels. The shaded region in each bar in Fig. 3 indicates an improvement over the baseline (ER). Our method has a lower mCE across all different corruptions, achieving 61.53% while ER achieves 84.24%. When comparing against their respective natural accuracies, Relative-mCE is a better measure to compare models with different top-1 accuracy. Our method achieves 28.94% Relative-mCE while ER settles for 46.21%. Evidently, task-agnostic learning when coupled with task-specific learning brings discernible benefits in terms of robustness to natural corruptions.

Model calibration: In safety-critical applications, it is pertinent for a model to possess an adequate sense of uncertainty about its predictions. A model is said to be miscalibrated when it tends to be overconfident or under-confident about its predictions compared to ground truth accuracy (Guo et al., 2017). In this section, we evaluate how well the models are calibrated. Expected Calibration Error (ECE) (Naeini et al., 2015) is the most common metric to determine miscalibration in classification. ECE computes a weighted average over the difference between the absolute accuracy and average confidence. A lower ECE value indicates a better calibrated model. We report ECE along with a reliability diagram using a calibration framework (Küppers et al., 2020). Figure 4 shows the reliability diagram of our method versus the baseline. Result shows that ER predictions are more overconfident compared to our method resulting in ECE values 32.76 and 18.29 respectively. In addition to improvement in natural accuracy, the right combination of task-agnostic and task-specific learning can effectively improve calibration, thereby improving reliability in safety critical environments.

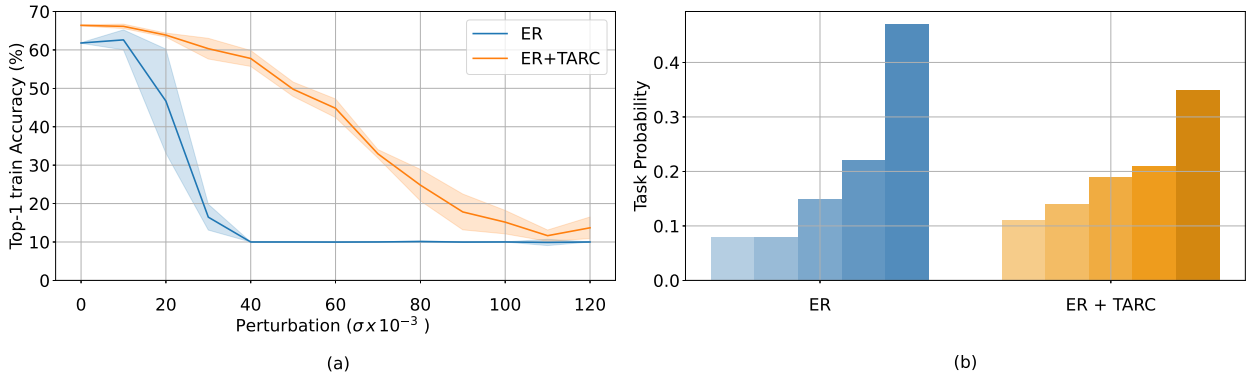


Figure 5: (a) Robustness of CL models to varying degrees of Gaussian noise added to the model weights. Our method is considerably robust to Gaussian perturbations and loses performance gradually suggesting convergence to flatter minima. (b) Average task probabilities of CL models trained on CIFAR-10 with 500 buffer size. Within each bar group, right most bar indicates the most recent task. Our method reduces the bias towards most recent task.

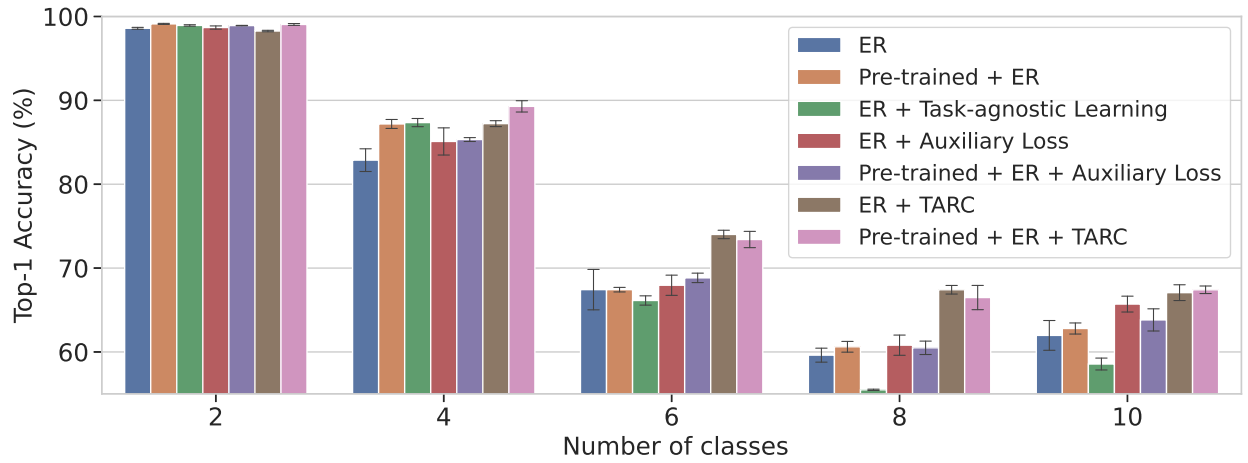


Figure 6: Effect of different components added to ER. TARC has the highest performance gain compared to individual components over ER. As number of tasks increases, pre-training has little to no effect on different CL methods.

Convergence to flatter minima: A CL model that converges to flatter minima in a loss landscape has more flexibility to adapt to a new task without drifting too far from the optimal parameters for the previous tasks. Furthermore, the solutions that reside in a flatter minima are more robust as the predictions do not change drastically with small perturbations (Buzzega et al., 2020). Following the analysis in (Zhang et al., 2018; Buzzega et al., 2020), we add independent Gaussian noise to all parameters of the model trained on CIFAR-10 with 500 buffer size. Figure 5-(a) shows the change in accuracy over different perturbation levels for the training set. Compared to the ER baseline, our method is significantly less sensitive to perturbations and the performance drops smoothly. We argue that task-agnostic learning followed by task-specific learning guides the solution to a wider valley which could better explain the ability of our model to consolidate generalizable features.

Bias towards recent tasks: Due to the sequential nature of continual learning, predictions are biased towards the recent task as the number of samples for the current task are far more than the buffered samples. Explicit techniques such as cosine normalization (Hou et al., 2019), weight aligning (Zhao et al., 2020) have been employed in the past to reduce the bias towards recent tasks. Following the analysis in (Buzzega et al., 2021), the normalized probability of each task of a CIFAR-10 trained model is computed by averaging probabilities of all samples belonging to the associated classes in Class-IL setting. Figure 5-(b) shows the normalized probability of each task being predicted at the end of training. Our method reduces the bias towards the most recent task and task probabilities are more evenly distributed compared to ER. Intertwining of task-agnostic learning and task-specific learning implicitly reduces this

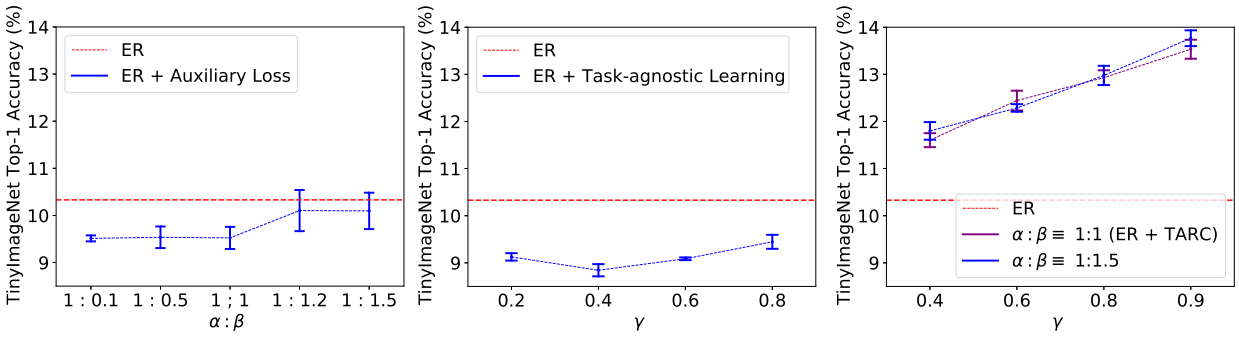


Figure 7: Hyperparmeter tuning for optimal CL performance. Graphs (a) and (b) consider our two main components of TARC separately with ER. As can be seen, the results are not satisfactorily above the ER baseline. The right most graph (c) presents our method TARC with different values for the hyperparamters. The results reaffirm our earlier hypothesis that both task-agnostic learning and auxiliary loss are essential components of our method.

bias without any additional constraints as the model spends more time on learning generalizable features than aligning itself with the current task.

6 Ablation Study

We attempt to shed light on the working of each component in our method. One might wonder whether task-agnostic learning or multi-objective learning alone can reap the aforementioned benefits for CL. We hypothesize that both task-agnostic learning and auxiliary loss are critical for our method since the former helps in learning generalizable features while the latter reduces the risk of these features being overwritten from task-specific learning. Figure 6 presents the evaluation of different variants of ER. We also extend the ablation study with more combinations of pre-training, auxiliary loss, and TARC. In line with our understanding, task-agnostic learning and auxiliary loss complement each other when coupled together. Our method is computationally less expensive and superior to pre-training on TinyImageNet. As number of tasks increases, pre-training has little to no effect on different CL methods. On the contrary, TARC provides discernible performance improvement over all other combinations with/without pre-training. We argue that this performance improvement over pre-training is essentially due to our method's ability to diminish the domain shift.

7 Hyperparameter tuning

We explore the effect of hyperparameters α, β and γ on our method. Figure 7 presents a study on the effect of these parameters on the the performance of our method on TinyImageNet. To reap the full benefits of our method, we note that it is crucial to have both task-agnostic learning and auxiliary loss paired with the baseline. As can be seen from the graph, when these components are used in isolation with ER, they do not perform satisfactorily above the ER baseline. Contrary to earlier methods that show nominal improvement over the ER baseline by using one of these components, TARC shows significant performance improvement by effectively combining both these components. Unless otherwise specified, the optimal parameters 1, 1 and 0.9 are used for α, β and γ respectively in all the experiments presented in this paper.

8 Conclusion and Future work

We proposed a novel two-stage CL training paradigm that intertwines task-agnostic and task-specific learning. Unlike self-supervised pre-training, our method does not suffer from domain shift, non-availability of pre-training data and additional computational costs. Furthermore, the effect of task-agnostic learning in our method does not diminish with longer task sequences. Our method can be easily added to memory- or regularization-based approaches with consistent performance gain across more challenging CL settings. We further provided an extensive analyses to shed light on the superiority of our method in terms of robustness, model calibration and bias towards recent tasks. Extending our method to algorithms oblivious of task boundaries and/or knowledge distillation based approaches are some of the useful future research directions for this work.

REFERENCES

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2180–2187. IEEE, 2021.

Lucas Caccia, Rahaf Aljundi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. Reducing representation drift in online continual learning. *CoRR*, abs/2104.05025, 2021.

Tiffany Cai, Jonathan Frankle, David J. Schwab, and Ari S. Morcos. Are all negatives created equal in contrastive instance discrimination?, 2021.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Michael Crawshaw. Multi-task learning with deep neural networks: A survey, 2020.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, et al. Remembering for the right reasons: Explanations reduce catastrophic forgetting. In *International Conference on Learning Representations*, 2020.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jhair Gallardo, Tyler L. Hayes, and Christopher Kanan. Self-supervised training enhances online continual learning, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

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

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Ya Le and X. Yang. Tiny imagenet visual recognition challenge. 2015.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30:6467–6476, 2017.

Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2337–2345, 2021.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308, 1990a. doi: 10.1037/0033-295X.97.2.285.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990b.

A. Robins. Catastrophic forgetting in neural networks: the role of rehearsal mechanisms. *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pp. 65–68, 1993.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537. PMLR, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *European Conference on Computer Vision*, pp. 645–666. Springer, 2020.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 06–11 Aug 2017.

Song Zhang, Gehui Shen, and Zhi-Hong Deng. Self-supervised learning aided class-incremental lifelong learning. *arXiv preprint arXiv:2006.05882*, 2020.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2020.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5871–5880, June 2021.

A Experimental setup

Following (Hsu et al., 2019; Buzzega et al., 2020), we evaluate on the following CL scenarios:

Class Incremental Learning (Class-IL): In this setting, the model encounters a new set of classes in each task and must learn to distinguish all classes encountered thus far after each task. In practice, we split CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009), TinyImageNet (Le & Yang, 2015), and STL-10 (Coates et al., 2011) into partitions of 2, 20, 20, and 2 classes per task, respectively.

Domain Incremental Learning (Domain-IL): In this setting, the number of classes remain the same across subsequent tasks. However, a task-dependent transformation is applied changing the input distribution for each task. Specifically, R-MNIST (Lopez-Paz & Ranzato, 2017) rotates the input images by a random angle in the interval $[0; \pi)$. R-MNIST requires the model to classify all 10 MNIST (Lecun et al., 1998) digits for 20 subsequent tasks.

General Incremental Learning (General-IL): In this setting, MNIST-360 (Buzzega et al., 2020) models a stream of MNIST data with batches of two consecutive digits at a time. Each sample is rotated by an increasing angle and the sequence is repeated six times. General-IL exposes the CL model to both sharp class distribution shift and smooth rotational distribution shift.

B Implementation details

To ensure a fair comparison between different methods in incremental learning, we build upon the *mammoth* CL framework (Buzzega et al., 2020) in PyTorch. Given a training budget $\mathcal{B}$ for each task, we adapt the original training schedule to accommodate self-supervised learning and multi-objective learning during the CL. Rotation prediction is used as an auxiliary task alongside supervised learning during multi-objective learning.

We employ ResNet-18 (He et al., 2016) for Class-IL and a 2-layer fully-connected network of 100 neurons each with ReLU activation for Domain-IL tasks. The underlying network consists of three different heads $h_\theta^{ssl}, h_\theta^{rot}$, and h_θ^{cls} each for supervised contrastive learning, rotation prediction and classification. h_θ^{ssl} consists of a linear layer

followed by a projection head of two-layer MLP with a ReLU non-linearity and 1-dimensional batch norm. We use ADAM (Kingma & Ba, 2017) optimizer with a learning rate of $3e^{-4}$. During task-agnostic learning, input images are transformed using a stochastic augmentation module consisting of a random resized crop, random horizontal flip followed by random color distortions. Since images are smaller in size, we leave out gaussian blur. During multi-objective learning, input images are rotated by one of $\{0^0, 90^0, 180^0, 270^0\}$ degrees. A linear layer is used for each rotation prediction h_θ^{rot} and classification h_θ^{cls} with cross entropy as a learning objective.

C Evaluation under noisy labels

Given the limited training budget for each task, quality of annotations play a crucial role in the success of CL methods. Robust training procedures need to be put in place to offset the impact of noisy labels as deep neural networks are known to memorize noisy labels (Arpit et al., 2017). We hypothesize that the representations learned using our method are more robust to noisy labels. To test our hypothesis, we simulate label corruption on CIFAR-10 by randomly sampling labels from a uniform distribution with a given probability (noise rate). A linear layer is trained on top of the backbone f_θ in presence of noisy labels and evaluated on the clean test set. Figure 8 presents robust accuracy for different noise rates. our method is less sensitive to noisy labels across different noise rates. By intertwining task-agnostic and task-specific learning, our method decouples representation learning from classifier. Since task-agnostic learning dominates the majority of the training budget $\mathcal{B}$, CL model has less chance to overfit to the noisy labels.

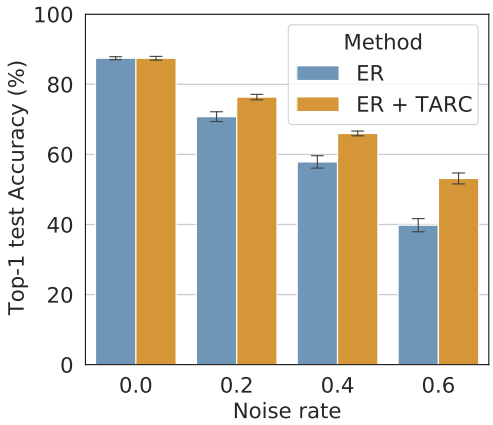


Figure 8: Robustness of CL models trained on CIFAR-10 with 500 buffer size against symmetric noisy labels.

D Analyses of our method on oEWC

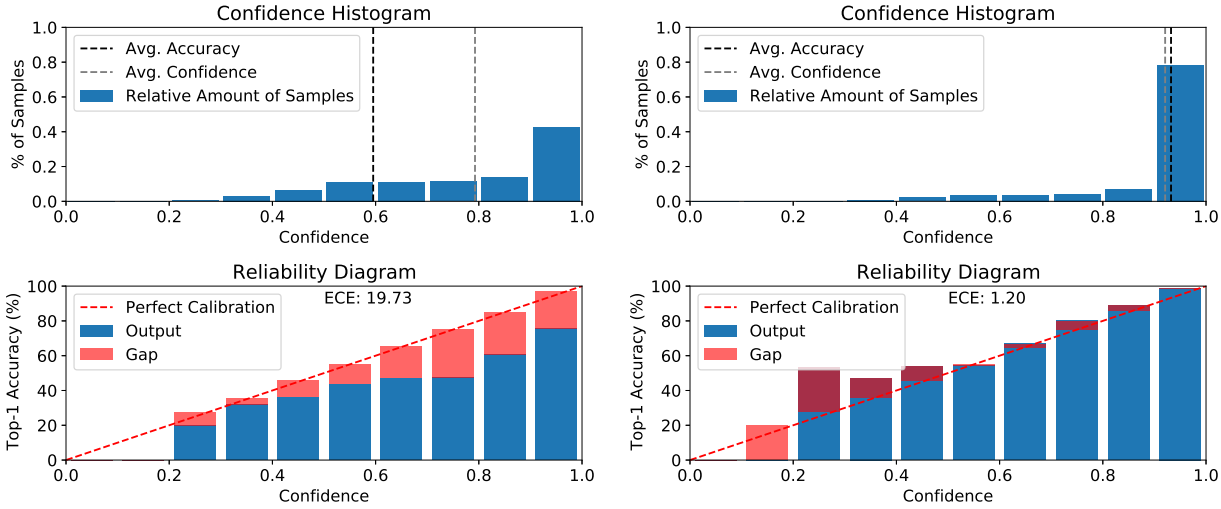
We extend some of the analyses in section 5 to oEWC. We use the same mechanism explained in the Algorithm 1 and extend oEWC to Domain-IL setting. As pointed out earlier, our method outperforms the oEWC baseline by almost 13% in Domain-IL setting.

D.1 Model calibration

A model is said to be miscalibrated when it tends to be overconfident or under-confident about its predictions compared to ground truth accuracy. Figure 9 presents the reliability diagram of our method versus the baseline. Result shows that ER predictions are more overconfident compared to our method resulting in ECE values 19.73 and 1.20 respectively. In addition to improvement in natural accuracy over 13% in Domain-IL setting, TARC effectively improves model calibration, thereby improving reliability in safety critical environments.

D.2 Flatter minima analysis

We extend the analysis conducted in section 5 to Domain-IL scenario. We add independent noise to all parameters of models trained on R-MNIST. Figure 10 presents change in accuracy over different perturbation levels for MNIST training set. Compared to oEWC baseline, our method is significantly less sensitive to perturbations and loses performance gradually suggesting convergence to a flatter minima in the loss landscape.



(a) Model calibration - oEWC (b) Model calibration - oEWC + TARC

Figure 9: Confidence estimates and corresponding ECE of R-MNIST trained CL models on clean MNIST. Lower ECE is better. Our method is well calibrated with confidence estimates closer to the perfect calibration compared to oEWC.

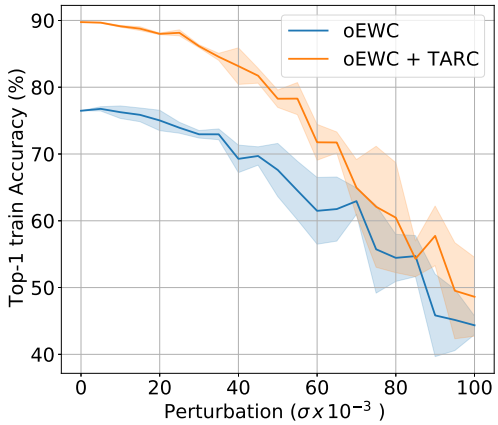


Figure 10: Robustness of CL models trained on R-MNIST to varying degrees of Gaussian noise added to the model weights. Our method is considerably robust to Gaussian perturbations and loses performance gradually suggesting convergence to flatter minima.

E SELF-SUPERVISED PRE-TRAINING FOR CONTINUAL LEARNING

Table 3 provides a comparison of the proposed training paradigm with pre-trained baselines across different datasets in Class-IL scenario. Pre-training is carried out on TinyImageNet using SimCLR Chen et al. (2020), a self-supervised learning objective. Similar to our main results in Table 1, our method shows a strong performance in majority of the benchmarks, especially when the buffer size is small. We argue that the effect of pre-training diminishes with the onset of CL training. On the other hand, TARC is able to retain information more efficiently than pre-training baselines by learning task-agnostic representations during CL training. In case of larger buffer size, our method is behind pre-trained baselines by a small margin on smaller datasets. The relative improvement compared to the baseline improves with the increase in image size and longer task sequences. TARC is able to discriminate better for STL-10 and TinyImageNet when the images are comparatively larger in size. For buffer size 500, the relative improvement in CIFAR-10 is close to 8% while it is approximately 35% in TinyImageNet.

Table 3: Comparison of TARC and SSL pre-training on CL models across various scenarios. We provide the average top-1 accuracy (%) across all tasks after continual learning training.

Bufer size	Method	Class-IL			
		CIFAR-10	CIFAR-100	TinyImageNet	STL-10
200	Pretrained + ER	49.27 ±0.87	22.57 ±0.39	8.67 ±0.10	56.71 ±0.62
	ER + TARC	**53.23** ±1.00	**23.48** ±0.10	**9.57** ±0.12	**60.61** ±0.96
500	Pretrained + ER	62.32 ±1.19	29.05 ±0.14	10.17 ±0.15	60.90 ±0.33
	ER + TARC	**67.41** ±0.94	**31.50** ±0.40	**13.77** ±0.17	**67.96** ±1.84
5120	Pretrained + ER	**82.73** ±0.07	**52.66** ±0.26	27.98 ±0.20	72.18 ±7.85
	ER + TARC	82.21 ±0.38	52.30 ±0.20	**32.04** ±0.37	**75.28** ±1.06

F Ablation study of robustness to natural corruptions

We evaluate the individual components of our method against common image corruptions using CIFAR-10-C (Hendrycks & Dietterich, 2018). Models are trained using clean CIFAR-10 with a buffer size of 500 and tested on CIFAR-10-C. We report the mean robust accuracy (%) to evaluate the performance under natural corruptions (Hendrycks & Dietterich, 2018). Figure 11 shows the mean robustness to 19 different corruptions averaged over five severity levels. ER with task-agnostic learning is considerably more robust to natural corruptions than with auxiliary task loss. However, as outlined in Section 6, ER with one of these components in isolation has a lower or marginally better natural accuracy than the baseline. On the other hand, the combination of these two components yields the best of both the worlds i.e. TARC achieves better natural accuracy while still being significantly robust to natural corruptions than ER baseline.

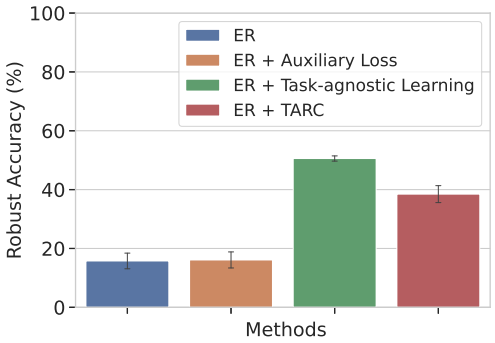


Figure 11: Mean robust accuracy (%) of CL models trained on CIFAR-10 with 500 buffer size against 19 natural corruptions.

G Comparison of different auxiliary tasks

TARC framework consists of a task-specific learning phase wherein the CL model is trained using an auxiliary loss alongside cross-entropy objective. To better understand the intuition behind the choice of the auxiliary loss function, we compare two primitive and two sophisticated auxiliary losses in the TARC framework. Figure 12 compares different CL models trained on CIFAR10 with buffer size 500 in the TARC framework. We also provide ER and ER+Task-agnostic Learning baselines for better comparison. As can be seen, Jigsaw prediction has the lowest performance while self-supervised algorithms such as SimCLR and SupCon have performances close to or slightly better than ER. Since CIFAR-10 images are small in size, jigsaw prediction turns out to be too difficult of a task to achieve.

On the other hand, rotation prediction helps preserve the generalizable features learned in the task-agnostic phase better, thereby leading to a significant improvement over ER baseline. The ability of rotation prediction to consolidate generalizable features has also been shown in other contemporary works (e.g. (Zhu et al., 2021)). We chose rotation prediction for its simplicity, limited computational overhead, and similarity to classification. In addition, rotation as an

augmentation does not change the underlying semantics of the input and does not need any additional forward passes through the CL model.

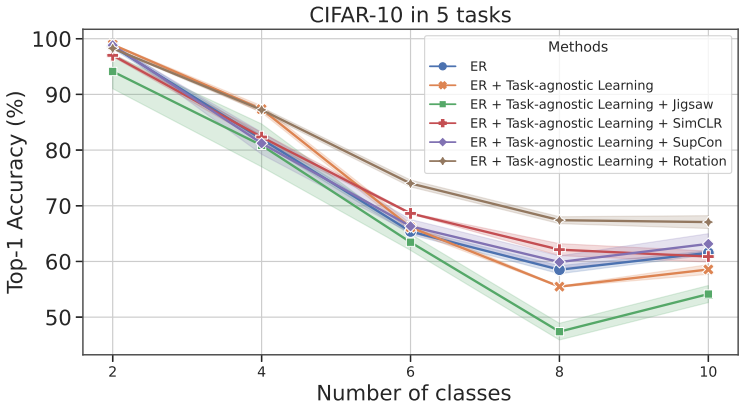


Figure 12: Comparison of different auxiliary tasks in TARC framework. We also provide ER and ER + Task-agnostic learning baselines. Rotation prediction is best suited as an auxiliary task due to higher performance and low computational overhead.