Mitigating Catastrophic Forgetting Using Improved Clustering-based Episodic Memory

Owen Beabout, Salisbury University; Abigail Dodd, University of Virginia;
Titus Murphy, Sattler College; Dr. Enyue Lu, Salisbury University

Background.

In continual learning, data is given to the model in a series of tasks. Because task distribution changes over time, models trained on new tasks tend to have high losses when retested on old tasks, also known as **catastrophic forgetting.** With real-world data, task identity is often not known at training time, and many current models require it to be known. Mixed Stochastic Gradient (MEGA) and Task-Agnostic Averaged Gradient Episodic Memory (TA-A-GEM) are algorithms that use episodic memory to counteract catastrophic forgetting in continual learning.

Proposed Model.

STRATA (**St**ochastic G**ra**dient with **T**ask-**A**gnosticity) is based on the TA-A-GEM and MEGA models mentioned. We improve the clustering memory by removing the sample that is furthest from the mean. Combining our improved clustering-based memory task-agnostic extensions on the MEGA-I and MEGA-II algorithms provides improved performance for domain-incremental online learning.

Neither the episodic memory handlers nor the model itself are given access to any task labels.

For each training iteration, STRATA recomputes reference gradients to control the model's parameters in relation to the current batch and random samples from the model's episodic memory. First, the model's loss on the incoming batch is computed and back-propagated to obtain the flat gradient vector. If memory samples exist, we stack up past examples and obtain both the reference loss and reference gradient without affecting the optimizer state. After this, like MEGA-I, STRATA-I balances the current and reference gradients through a loss-based weighted sum. Similarly, following MEGA-II, STRATA-II rotates the flat gradient vector toward the reference gradient vector using a loss-based ratio calculation.

Experiments.

We used the MNIST, Fashion MNIST, and CIFAR-10 datasets to evaluate our model. Our task types were permutations, rotations, and class-split. We used two frameworks of task introduction: sequential and continual. Regardless of framework, each task was training over 20 epochs with a random sample order every epoch.

We used three different baselines to test our models: TA-A-GEM and BGD as previous state-of-the-art models for this task, and randomly adding and removing from clusters in episodic memory with TA-A-GEM.

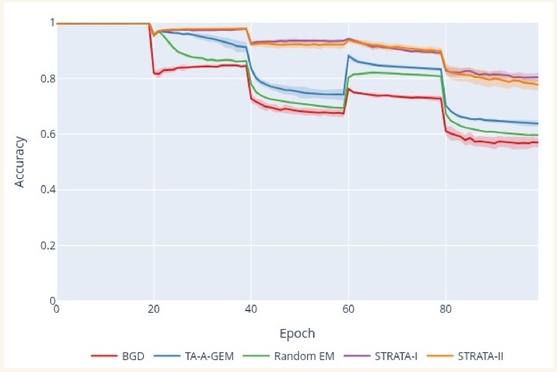
Results.

	Class Split	Permutation	Rotation
Random	0.8068	0.7537	0.6027
TA-A-GEM	0.8412	0.7805	0.6162
BGD	0.7684	**0.8609**	**0.6973**
STRATA-I	**0.9260**	0.7953	**0.7008**
STRATA-II	**0.9222**	0.7923	**0.6913**

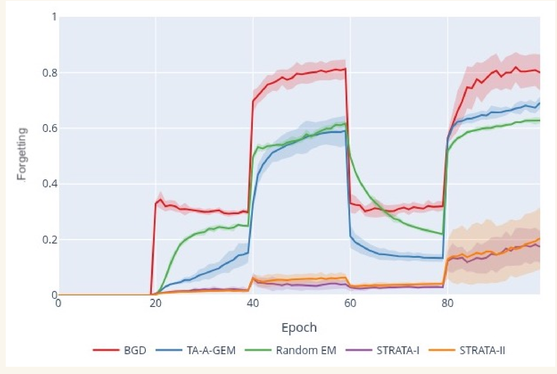
Table 1. Average overall **accuracy** for **sequential** task introduction, **MNIST**. The best result in each column, plus any result within 99% confidence of the best result, is written in bold.

	Class Split	Permutation	Rotation
Random	0.41350	0.16803	0.46465
TA-A-GEM	0.35401	**0.06275**	0.44032
BGD	0.54056	0.17244	0.47983
STRATA-I	**0.05917**	**0.06541**	**0.33136**
STRATA-II	**0.06718**	**0.06169**	**0.33670**

Table 2. First-task **forgetting** for **sequential** task introduction, **MNIST**. The best result in each column, plus any result within 99% confidence of the best result, is written in bold.



Left: Average **accuracy** of each of the 5 models. MNIST, **class-split** task, sequential. Averaged over all epochs, then over 5 runs. Shaded area shows 1 standard deviation.



Right: First-task **forgetting** of each of the 5 models. MNIST, **class-split** task, sequential. Averaged over all epochs, then over 5 runs. Shaded area shows 1 standard deviation.

Discussion.

STRATA-I has statistically significant improvements on class split tasks across all datasets. Not only does STRATA achieve incredibly high task-agnostic accuracies on MNIST and Fashion MNIST datasets (clearing 92% accuracy on both sequential and continual task introduction frameworks), we also see roughly 16, 14, and 2.5 percentage point increases in accuracy on datasets MNIST, Fashion MNIST, and CIFAR-10, respectively, on sequentially-introduced class-split tasks when compared to BGD. STRATA retains its superior performance on continual tasks, with roughly 8.5, 9.5, and 2 percentage point increases in accuracy on datasets MNIST, Fashion MNIST, and CIFAR-10, respectively, on continual task introduction, when compared to BGD.

Both graphs shown to the left exemplify STRATA's superior performance over all three baseline models. Both STRATA models have significantly higher accuracy and lower forgetting than any of the three baseline models.

BGD consistently outperforms STRATA-I and -II on MNIST permutation tasks, but STRATA-I's overall accuracy is higher than TA-A-GEM's with 99% confidence. STRATA-II outperforms TA-A-GEM's average accuracy in a continual task introduction environment with 99% confidence for continual introduction, and with 98% confidence for sequential introduction.

With the exception of permutation tasks on the CIFAR-10 dataset, STRATA performs statistically indistinguishably from baseline or significantly better on first-task accuracy. This, combined with its superior performance on overall accuracy, suggests that STRATA has successfully walked the line between flexibility for newer tasks and retention of early tasks.

References.

1. C. Lamers, R. Vidal, N. Belbachir, N. van Stein, T. Baeck, and P. Giampouras, "Clustering-based domain-incremental learning," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 3384–3392.
2. Y. Guo, M. Liu, T. Yang, and T. Rosing, "Improved schemes for episodic memory-based lifelong learning," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1023–1035, 2020.
3. C. Zeno, I. Golan, E. Hoffer, and D. Soudry, "Task agnostic continual learning using online variational Bayes," arXiv preprint arXiv:1803.10123, 2019.

Acknowledgment.

This work was funded by NSF CCF-2447041 under the Research Experiences for Undergraduates Program.