Continual Learning Benchmarks for Antipodal Grasping

Sayantan Auddy[1] Jakob Hollenstein[1] Matteo Saveriano[1,3] Antonio Rodríguez-Sánchez[1] Justus Piater[1,2]

Abstract— **A continual learning robot that can repeatedly learn from new data without forgetting past knowledge is surely preferable to a robot that cannot learn incrementally. As most continual learning methods focus on image classification tasks, it is not clear if or how they scale to more complicated vision problems such as robotic grasp prediction. To fill this gap, we propose a set of benchmarks that can be used to evaluate continual learning methods on the problem of antipodal grasping. We adapt a state-of-the-art grasp prediction model for continual learning and evaluate multiple baselines using our benchmarks. Our preliminary findings indicate that replay-based methods may be suitable for the grasp prediction task.**

I. INTRODUCTION

Antipodal robot grasping is arguably a challenge that has already been met, considering the performance of state-of-the-art methods [8], [11], [14]. All such methods either rely on large scale grasp datasets [3], or use domain randomization [16] to diversify the training data [3], [10]. However, it may not be possible to completely anticipate every possible situation in advance. A better strategy is to augment the grasp learning methods with *continual learning* (CL) [13] to make them capable of learning from a sequence of multiple small, disjoint, and non-IID datasets that are encountered over time.

Barring a few works which address continual learning for robotics [2], [5], most of the current CL research [7], [9], [15], [18] focuses on image classification problems using relatively simple network architectures. It is not apparent how well these methods scale to robotics applications, which typically involve the use of more complex network components and architectures. If properly applied to robot learning, CL can help achieve open-ended robot learning, and this would be especially useful for a ubiquitous robotics problem such as vision-based grasp prediction. With this motivation, we present preliminary work on a set of continual learning benchmarks that can be used to evaluate the effectiveness of continual learning methods on the problem of grasp learning. We adapt a state-of-the-art grasp learning method [8] for continual learning and evaluate multiple baselines on our benchmarks. Our initial results indicate that replay-based CL strategies outperform regularization-based CL.

II. CONTINUAL LEARNING FRAMEWORK

A. Benchmarks

The Cornell dataset [14] is a widely used dataset for grasp learning [1]. It consists of 885 RGB images of 280 different household objects, where each image is annotated with multiple top-down grasp rectangles. We create the following 4 benchmarks by partitioning the images of this dataset into multiple *tasks* (sub-datasets), where each benchmark is learned independently by learning its tasks sequentially:

Shape: We manually partition the 885 images of the Cornell dataset into 5 tasks, where each task contains images of objects with one of these shapes: *rectangular* (e.g. boxes, TV remotes), *rim* (objects with a graspable rim, e.g. bowls, frisbees), *long* (elongated objects, e.g. stick), *round* (objects with a circular symmetry, e.g. apple, potato), and *handle* (objects with a handle, e.g. spatula, toothbrush).

Width 5: For each image we compute the average of the 5 largest grasp widths. Then, 5 tasks are defined by partitioning the images according to their average grasp width such that each task has roughly the same number of images.

Width 10: We follow the same process as Width 5, but here we partition the 885 images into 10 tasks.

Object: We compute the number of images for each distinct object and then choose 10 objects with the most number of images. Each of these 10 objects corresponds to a task.

Shape has highly disbalanced tasks, whereas the *Width* benchmarks have balanced tasks. *Object* has very few images per task. For each task, we create training and validation sets in the ratio 75:25 and use image augmentation (random rotations, translations, crops) to learn from such small datasets.

B. Baselines

As our base architecture, we choose the fully convolutional network proposed in [8]. This network has 1.8×10^6 parameters and produces heatmaps for the grasp center, orientation and width. Using this, we implement the following baselines:

Finetuning (FT): For each benchmark, a network is initialized at the beginning and is then successively finetuned on each task. This forms the lower performance baseline, as here we would expect that only the last task is remembered.

Replay 20% (RE20): This setting is similar to FT, but for each task 20% of the training data is randomly cached and then combined with the training data of the next task. For example, while training for task 2, we use all the data of task 2 and 20% of the data each from tasks 1 and 0.

Replay 100% (RE100): This is the same as RE20, except that here all the data from past tasks is cached and replayed.

Synaptic Intelligence (SI): This setup is similar to FT, but the grasp prediction network's parameters are protected from *catastrophic forgetting* [13] using a regularization term according to the formulation of Synaptic Intelligence [18].

Joint Training (JT): This forms the upper performance baseline. For learning each task, we use a freshly initialized network and the data for all tasks up to that point. Note that

[1] Department of Computer Science, University of Innsbruck, Technikerstrasse 21a, Innsbruck, Austria. {name.surname}@uibk.ac.at

[2] Digital Science Center (DiSC), University of Innsbruck, Austria.

[3] Department of Industrial Engineering, University of Trento, Italy.

all the data of previous tasks is also used for RE100 but it does not reinitialize the network for each task.

III. Results

We train and evaluate the baselines on each benchmark independently, where the tasks of a benchmark are learned in sequence. Each training run is repeated with 3 independent seeds. To compute the accuracy of the predicted grasp, we compare the intersection over union (IoU) of the prediction with the grasp annotations, and if any of these comparisons have an IoU of more than 25%, the prediction is considered to be accurate [8]. After training on each task, each baseline is evaluated on the validation sets of the current and past tasks. For example, after training on task 3, the network is evaluated on tasks 0, 1 and 2. The average of these validation accuracies is depicted in Fig. 1. The upper baseline JT maintains an average accuracy close to 90% after all tasks for all benchmarks. On the other hand, the performance of the lower baseline FT drops as more tasks are learned, clearly exhibiting *catastrophic forgetting* [13]. The regularization-based CL method (SI [18]) is not able to avoid forgetting past tasks and performs similar to FT. The replay-based baselines perform better than FT and SI but worse than JT.

In Fig. 2, we show how the validation accuracy for each of the 10 tasks of the *Width 10* benchmark changes as newer tasks are learned. The drop in accuracy for the oldest tasks can be seen for all the baselines except JT. This drop is severe for FT and SI, for which the accuracy for task 0 (which corresponds to objects needing the smallest and most precise grasps) drops to around 20% after all tasks are learned. It can also be seen that for FT and SI, whenever a new task is learned, its accuracy starts around 90%, but starts dropping sharply as newer tasks are learned.

Using the validation accuracies we also compute continual learning metrics [4] in Tab. I (for *Width 10*). In terms of accuracy (ACC) over all tasks, and remembering (REM) past tasks, JT is the best, followed by the two replay-based baselines. Since joint training and replay involve the storage of data from past tasks, they achieve low scores on the storage size efficiency metric (SSS). Interestingly, FT has the highest forward transfer (FWT) score, indicating that it is the best at using past knowledge to learn newer tasks.

IV. Summary and Outlook

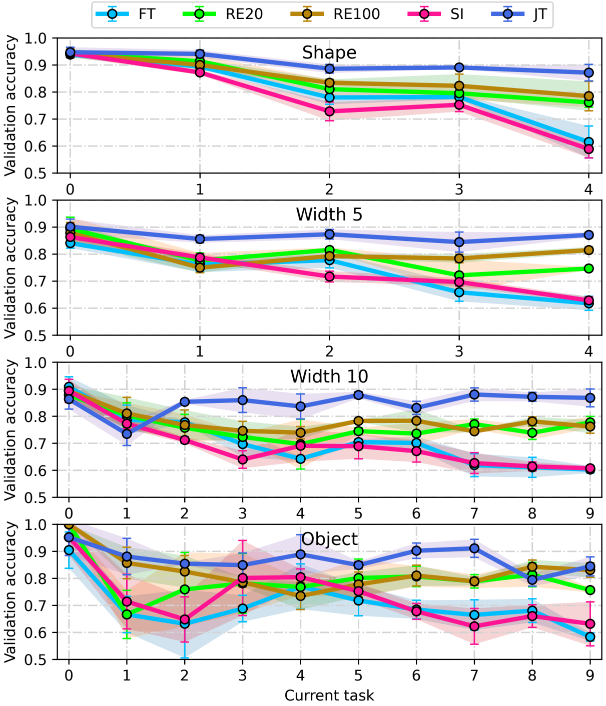

Fig. 1. Cumulative validation accuracies for current and past tasks. Each data point shows the mean accuracy of all tasks till that point on the $x-$axis.

The preliminary results presented in this short paper indicate that (i) parameter regularization may not be as effective as replay-based CL for continual grasp learning, and (ii) it is possible to learn grasps using small datasets. To further expand these findings, we will evaluate more continual learning methods [6], [7], [15], [17] in the future. We also plan to perform similar evaluations for other robot vision applications such as affordance detection [12]. Our goal for this future work will be to identify areas where current continual learning methods can be improved to make them more suitable for robotics tasks.

TABLE I

CL metrics for Width 10 (1-best, 0-worst).

Benchmark	Baseline	ACC	REM	FWT	SSS
Width 10	FT	0.663	0.733	**0.732**	**1.000**
	RE20	0.752	0.827	0.711	0.912
	RE100	0.769	0.839	0.681	0.550
	SI	0.655	0.717	0.698	**1.000**
	JT	**0.858**	**0.943**	0.702	0.550

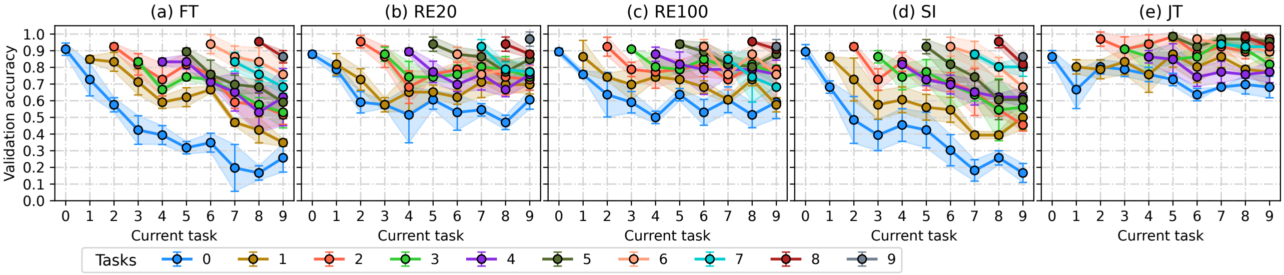

Fig. 2. Per-task validation accuracies for *Width 10*, showing how the accuracy of each task changes as newer tasks are learned.

References

[1] "Robotic Grasping on Cornell Grasp Dataset," https://paperswithcode.com/sota/robotic-grasping-on-cornell-grasp-dataset-1, accessed: 2022-04-22.

[2] S. Auddy, J. Hollenstein, M. Saveriano, A. Rodríguez-Sánchez, and J. Piater, "Continual learning from demonstration of robotic skills," *arXiv preprint arXiv:2202.06843*, 2022.

[3] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3511–3516.

[4] N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni, "Don't forget, there is more than forgetting: new metrics for continual learning." [Online]. Available: http://arxiv.org/abs/1810.13166

[5] C. Gao, H. Gao, S. Guo, T. Zhang, and F. Chen, "Cril: Continual robot imitation learning via generative and prediction model," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6747–5754.

[6] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "REMIND your neural network to prevent catastrophic forgetting." [Online]. Available: http://arxiv.org/abs/1910.02509

[7] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks." [Online]. Available: http://arxiv.org/abs/1612.00796

[8] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," version: 3. [Online]. Available: http://arxiv.org/abs/1909.04810

[9] T. Lesort, "Continual learning: Tackling catastrophic forgetting in deep neural networks with replay processes." [Online]. Available: http://arxiv.org/abs/2007.00487

[10] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[11] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach." [Online]. Available: http://arxiv.org/abs/1804.05172

[12] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1374–1381.

[13] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," vol. 113, pp. 54–71. [Online]. Available: http://arxiv.org/abs/1802.07569

[14] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[15] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," p. 10.

[16] J. Tobin, L. Biewald, R. Duan, M. Andrychowicz, A. Handa, V. Kumar, B. McGrew, A. Ray, J. Schneider, P. Welinder, *et al.*, "Domain randomization and generative models for robotic grasping," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3482–3489.

[17] J. von Oswald, C. Henning, B. F. Grewe, and J. Sacramento, "Continual learning with hypernetworks," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://arxiv.org/abs/1906.00695

[18] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence." [Online]. Available: http://arxiv.org/abs/1703.04200