# Issues for Continual Learning in the Presence of Dataset Bias

**Anonymous submission**


### Abstract

While most of continual learning algorithms have focused on tackling the stability-plasticity dilemma, they have overlooked the effects of the knowledge transfer when it is *biased* by learning unintended spurious correlations that do not capture the true causal structure of the tasks. In this work, we designed systematic data experiments and show that such bias is indeed transferred, both forward and backward, during continual learning and argue that causality-aware design of continual learning algorithm is critical.


## Introduction

Continual learning (CL) is essential for a system that needs to learn (potentially increasing number of) tasks from sequentially arriving data in an online fashion. The main challenge of CL is to overcome the *stability-plasticity* dilemma (Mermillod, Bugaiska, and Bonin 2013) that is a trade-off where if a CL model focuses too much on the stability of learned tasks, it suffers from low plasticity for integration of a new task (and vice versa). Recent deep neural networks (DNNs) based CL methods (Kirkpatrick et al. 2017; Jung et al. 2020; Li and Hoiem 2017) attempted to address the dilemma by devising mechanisms to attain stability while improving plasticity thanks to the *knowledge transferability* (Tan et al. 2018), which is one of standout properties of DNNs. Namely, while maintaining the learned knowledge, the performance on a new task (resp. past tasks) are improved by transferring of knowledge of past tasks (resp. a new task). Such phenomena are called the *forward* and *backward transfer*, respectively.

Unfortunately, it is widely known that DNNs often dramatically fail to generalize to out-of-distribution data due to learning some unintended spurious correlations (*e.g.*, dataset bias (Torralba and Efros 2011)), not the true causal relations (Sagawa et al. 2020; Bahng et al. 2020). For instance, a DNN that classifies birds on the sky perfectly may fail on images in which birds are outside the typical sky background when the model has learned a *shortcut* strategy relying on the background (Geirhos et al. 2020). Furthermore, a recent work (Salman et al. 2022) shows that such *bias* in a model can be transferred; namely, the bias in pre-trained models are remained present even after fine-tuning them on downstream tasks. In CL, when a model becomes biased while learning a specific task, such bias transfer is likely to happen and may lead to unexpected failures continuously.

In this paper, we show that when the causal learning is not appropriately considered, naively applying CL methods would be problematic since they can maintain unwarranted knowledge (*e.g.*, background bias). In this end, we make a synthetic dataset with color bias, and systematically conduct extensive experiments on various two task scenarios with different degrees of the bias. We identify that bias of a specific task affects other tasks in CL due to *two sources*: the forward and backward transfer of bias. Specifically, a typical CL method preserves the knowledge such that the bias of the knowledge is reused to train on a new task (*i.e.*, forward transfer of bias). Furthermore, the biased knowledge learned from the current task affects the decision rules for the past tasks to be biased (*i.e.*, backward transfer of bias). To solve these, our experimental results show that it is necessary to learn causal relation of each task in CL. Furthermore, we demonstrate that, when learning the relation, CL should take the stability into account simultaneously, since otherwise it can lead to severe catastrophic forgetting. Therefore, our results strongly appeal the necessity of a novel method that can do causal learning while preventing forgetting.

## Case Studies of Bias Transfer in CL

### Experimental Settings

**Dataset.** We use Split CIFAR-100 (Zenke, Poole, and Ganguli 2017; Chaudhry et al. 2019; van de Ven, Siegelmann, and Tolias 2020), which divides CIFAR-100 into 10 tasks with 10 distinct classes. To study bias transfer, we modify Split CIFAR-100, such that half of the classes in each task are skewed toward the grayscale domain and the other half toward the color domain. Namely, given a skew-ratio $\alpha \geq 0.5$, the training images of each class are split into $\alpha$ and $1 - \alpha$ ratios for each domain. We set 6 bias levels by dividing the range from 0.5 to 0.99 evenly on a log scale for systematic control of the degree of bias.

**CL Scenario.** We consider a task-incremental learning scenario (Van de Ven and Tolias 2019) in which task identifier is given during inference time and further assumes the domain of images is known. For simplicity, we only considered the scenario of incrementally learning *two* tasks; we randomly chose 2 out of 10 tasks in every run and report the

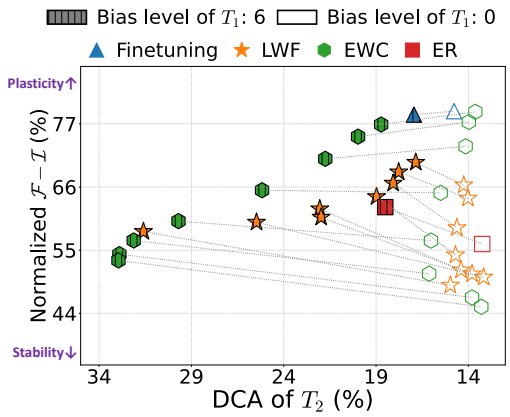


Figure 1: Forward transfer of bias. The higher DCA indicates a model is more biased. y-axis shows the level of focus on plasticity or stability. Dashed lines connect the points with the same learning strategy (hyperparameters).

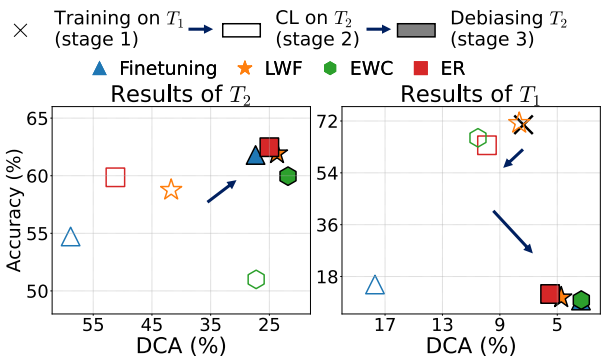


Figure 2: Backward transfer of bias. Blue arrows indicate the sequence of stages. Since all baselines are trained in the same way on $T_1$, we report the results with one cross marker.

averaged results over 4 different runs. We denote the first and second task as $T_1$ and $T_2$, respectively.

**Metrics.** We consider two metrics as evaluation metrics for CL performance and bias, accuracy and the difference of classwise accuracy (DCA) (Berk et al. 2021) for each task. Note DCA is defined to be the average (over class) of per-class accuracy difference between domains. We also compute forgetting ($\mathcal{F}$) and intransigence ($\mathcal{I}$) measure (refer to Section 3.1 in Cha et al. 2021 for details), which evaluate stability and plasticity, respectively, of a CL method. We use the normalized difference of them to evaluate the relative weight on plasticity and stability; *i.e.*, the lower the value is, the more the model focuses on stability (and vice versa).

**Baselines.** We adopt finetuning without any consideration of CL and three representative CL methods: LWF (Li and Hoiem 2017), EWC (Kirkpatrick et al. 2017), and ER (Chaudhry et al. 2019). LWF and EWC add regularization terms in their training objectives to penalize deviation from the past model and balance the stability-plasticity trade-off by controlling the regularization hyperparameter. On the other hand, ER stores some data from past tasks and replays them while learning current task. Finally, as a model debiasing technique, we employ MFD (Jung et al. 2021), a state-of-the-art method that trains a domain-independent model using a MMD-based feature distillation.

## Study 1: Forward Transfer of Bias

To investigate the influence of bias captured from $T_1$ in a CL scenario, we evaluated baseline methods by varying the bias level of $T_1$, with that of $T_2$ is fixed to level 2. Figure 1 shows DCA of $T_2$ along with $\mathcal{F}-\mathcal{I}$ after learning $T_2$ with two different bias levels of $T_1$, *i.e.*, level 0 & 6. The figure plots the results of LWF and EWC with various regularization strengths; namely, the upper the point is, the lower the regularization strength is. From the gap of blue triangles in the figure, we first observe that bias of $T_1$ adversely affects bias of $T_2$, *i.e.*, forward transfer of bias exists, even with simple finetuning, which is consistent with Salman et al. 2022. Second, we observe that when applying CL methods, the gaps between connected points get larger than finetuning. Moreover, when the bias level of $T_1$ is 6, DCA of $T_2$ for EWC and LWF increases more drastically as the focus on stability is larger. Thus, these results imply that CL methods promote the forward transfer of bias since they tend to remember the knowledge of past tasks for stability. Finally, we clearly observe that DCA of $T_2$ is always better when learned after $T_1$ with bias level 0 than with bias level 6, for similar $\mathcal{F}-\mathcal{I}$. Therefore, we argue that whenever a given task has a bias in CL scenario, its bias should be mitigated for learning future tasks.

## Study 2: Backward Transfer of Bias

Here, we set the bias level of $T_1$ and $T_2$ as 0 and 6 and assume a scenario that bias is detected after learning $T_1$ (stage 1) and continually learning $T_2$ (stage 2) by a CL method. In this situation, one may naively consider applying MFD (stage 3) for debiasing the models on $T_2$. Figure 2 shows accuracy and DCA of $T_1$ and $T_2$ on each stage for each baseline. In the right plot, we observe that points shift to bottom left as progressing from stage 1 to stage 2. It means that as the stability gets less focus, the bias obtained from $T_2$ is more transferred to $T_1$, *i.e.*, the backward transfer of bias. From results of the stage 3 in the left plot, we show that DCA of $T_2$ can be successfully reduced by employing a debiasing technique of a model with similar accuracy. However, we also identify that accuracy of $T_1$ significantly drops. Thus, when debiasing after learning each task as we argued in Study 1, one should consider forgetting issue of the learned tasks at the same time, *i.e.*, it is necessary to develop a novel method learning causal relation considering the stability for CL.

## Conclusion

We showed forward and backward transfer of bias in CL. Since bias can be transferred either way unexpectedly, we must pursue causal learning, but under the consideration of stability. In future works, we will investigate the bias transfer in more realistic scenario with long sequence of data stream.

## References


Bahng, H.; Chun, S.; Yun, S.; Choo, J.; and Oh, S. J. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 528–539. PMLR.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

Cha, S.; Hsu, H.; Hwang, T.; Calmon, F.; and Moon, T. 2021. {CPR}: Classifier-Projection Regularization for Continual Learning. In *International Conference on Learning Representations*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019. Continual Learning with Tiny Episodic Memories. *arXiv preprint arXiv:1902.10486, 2019*.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Jung, S.; Ahn, H.; Cha, S.; and Moon, T. 2020. Continual learning with node-importance based adaptive group sparse regularization. *Advances in Neural Information Processing Systems*, 33: 3647–3658.

Jung, S.; Lee, D.; Park, T.; and Moon, T. 2021. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12115–12124.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13): 3521–3526.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Mermillod, M.; Bugaiska, A.; and Bonin, P. 2013. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Salman, H.; Jain, S.; Ilyas, A.; Engstrom, L.; Wong, E.; and Madry, A. 2022. When does Bias Transfer in Transfer Learning? *arXiv preprint arXiv:2207.02842*.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, 270–279. Springer.

Torralba, A.; and Efros, A. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 1521–1528.

van de Ven, G. M.; Siegelmann, H. T.; and Tolias, A. S. 2020. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11: 4069.

Van de Ven, G. M.; and Tolias, A. S. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, 3987–3995. PMLR.