PNAS LETTER

Reply to Dablander and Bury: Dealing with the unknown unknowns of deep learning

Chris T. Bauch[a,1] and Madhur Anand[b]

Dablander and Bury (1) make a valuable clarification that the approach of Bury et al. (2) requires detrended time series—not trendless time series per se—and point out that they are not the same thing for this type of deep learning algorithm (DLA). The algorithm learned not only the generic features of bifurcations but also features of the time series that are specific to the choice of detrending filter. This behavior of DLAs is also vividly illustrated by the "single-pixel attack," where an image-recognizing DLA can be fooled by replacing a single pixel. This causes the DLA to misclassify a cat as a dog, in cases where the human eye would not be misled (3). In both cases, DLAs learn features that are specific to the training set, but some of the learned features are neither intended nor known by the programmer. And, when the DLA is exposed to something slightly outside of its training set (by changing a pixel or a preprocessing filter), these features cause the DLA to misclassify something that—to a human—is obviously still a cat or a trendless time series. Dablander and Bury flag a problem that all DLAs share, and the scientific community should certainly be reminded of, as these methods become more mainstream. We only have a few comments to add. We think the expectation for DLAs to be robust to phenomena like single-pixel attacks may reflect a tendency to anthropomorphize DLAs, perhaps on the assumption that artificial neural networks should operate in ways similar to biological neural networks. But they differ in many fundamental aspects. Hence, when these algorithms are challenged with something slightly outside of their training set, we are surprised when they produce answers a human brain would not think of. Given that DLAs can exhibit such "unknown unknown" limitations on their operational parameters, what should researchers do? Dablander and Bury highlight one important aspect, which is to be aware of the differences between DLAs and other types of classifiers such as traditional early warning indicators or human expert opinion: Every scientist needs to understand their instrument, and DLAs are no different. In the context of Bury et al., researchers should practice consistent use of the same filter for training, testing, and application, and devote precious computer memory to including a greater variety of bifurcations or noise types (instead of multiple filter types) in the training set. Detrending may still be valuable for developing early warning signals of tipping points, even if future DLAs do not require it as a preprocessing step: These methods exploit the properties of critical phenomena near tipping points (4, 5), and detrending should help any classifier detect the information that is present in the noisy fluctuations around a slowly changing equilibrium (6). Like many readers, we are excited to see what the coming years will bring for data-driven dynamical systems, and especially the interface between complex systems and deep learning algorithms as exemplified by Bury et al. (2).

Author affiliations: [a]Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada; and [b]School of Environmental Sciences, University of Guelph, Guelph, ON N1G 2W1, Canada

Author contributions: C.T.B. and M.A. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: cbauch@uwaterloo.ca.

Published August 16, 2022.

1. F. Dablander, T. M. Bury, Deep learning for tipping points: Preprocessing matters. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2207720119 (2022).
2. T. M. Bury *et al.*, Deep learning for early warning signals of tipping points. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2106140118 (2021).
3. J. Su, D. V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **23**, 828–841 (2019).
4. I. Farahbakhsh, C. T. Bauch, M. Anand, Best response dynamics improve sustainability and equity outcomes in common-pool resources problems, compared to imitation dynamics. *J. Theor. Biol.* **509**, 110476 (2021).
5. R. Biggs, S. R. Carpenter, W. A. Brock, Turning back from the brink: Detecting an impending regime shift in time to avert it. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 826–831 (2009).
6. C. Boettiger, From noise to knowledge: How randomness generates novel phenomena and reveals information. *Ecol. Lett.* **21**, 1255–1267 (2018).