The Expressivity of Random Neural Networks with Learned Inputs

Ezekiel Williams* (ezekiel.williams@mila.quebec)
Department of Mathematics and Statistics, Université de Montréal & Mila, 2920, chemin de la Tour
Montréal, Québec H3T 1J4 Canada

Avery Hee-Woon Ryoo* (hee-woon.ryoo@mila.quebec)
Department of Computer Science and Operations Research, Université de Montréal & Mila, 2920, chemin de la Tour
Montréal, Québec H3T 1J4 Canada

Thomas Jiralerspong* (thomas.jiralerspong@mila.quebec)
Department of Computer Science and Operations Research, Université de Montréal & Mila, 2920, chemin de la Tour
Montréal, Québec H3T 1J4 Canada

Matthew G. Perich (matthew.perich@umontreal.ca)
Department of Neuroscience, Université de Montréal & Mila, 2960 Chemin de la Tour
Montréal, Québec, H3T 1J4 Canada

Luca Mazzucato (lmazzuca@uoregon.edu)
Departments of Biology, Physics, and Mathematics, & Institute of Neuroscience, University of Oregon, 1425 E 13th Ave
Eugene Oregon, 97405 United States

Guillaume Lajoie (guillaume.lajoie@mila.quebec)
Department of Mathematics and Statistics, Université de Montréal & Mila, 2920, chemin de la Tour
Montréal, Québec H3T 1J4 Canada
'*' denotes co-first author

Abstract

The expressivity of a neural network where all weights are initialized randomly and only constant inputs (biases) are learned is not well-studied and of interest in two domains. In neuroscience, the contribution of inputs from upstream regions, versus local plasticity, to learning in neural circuits (e.g. motor cortex) is poorly understood. In artificial intelligence (AI), recent empirical work has shown that fine-tuning biases alone can yield efficient multi-task learning. However, both fields lack a thorough understanding of the limits of input-only learning. Here, we provide theoretical and empirical evidence that a wide class of functions and finite trajectories from many dynamical systems can be well approximated by randomly initialized networks where only biases are optimized. These results extend our understanding of neural network models, providing guidance for future AI development and models of inter-region learning in the brain.

Keywords: random networks; deep learning; recurrent neural networks; applied mathematics; neural dynamics; learning

Introduction

The diversity of behaviour that can be generated by learning only the inputs to a neural network is a relevant question for both neuroscience and machine learning. In the brain, it is unknown to what extent the output dynamics of a neural sub-network are determined by local synaptic changes in that network or adaptation in the sub-network's inputs (Feulner et al., 2022). If only highly sophisticated inputs will result in useful changes in dynamics then this might suggest that local plasticity is critical for learning; conversely, if very simple, e.g. constant, inputs can radically reshape output dynamics then input-driven learning might be an important element of biological learning. Recent work (Ogawa, Fumarola, & Mazzucato, 2023) has shown that diverse dynamics can occur from changes in constant input to a Recurrent Neural Network (RNN), but characterization of the total set of output dynamics that one can span simply by input-only learning is lacking.

In Machine Learning (ML), empirical research has begun to explore multi-task methods where a single set of synaptic weights are pre-trained and then biases are adapted on a per-task basis (Zaken, Ravfogel, & Goldberg, 2021). An understanding of the expressivity of networks where only the inputs, or biases, are learned would provide a theoretical backbone for this work. Related to bias-only learning, a rich literature in ML has explored the set of functions on *iid* data, and dynamical systems on temporal data, that can be learned by randomly initializing a network and then learning other subsets of the parameters. This research has focused on learning output weights only–referred to as random neural networks in the *iid* case (Rahimi & Recht, 2008; Scardapane & Wang, 2017) and reservoir networks in the temporal case (Gonon, Grigoryeva, & Ortega, 2023; Hart, Hook, & Dawes, 2021). It has been found that very general classes of functions/dynamical system trajectories can be learned by adapting only the outputs. Recent work in AI and neuroscience has also explored other sets of parameters, for example those involved in batch normalization (Giannou, Rajput, & Papailiopoulos, 2023; Burkholz, 2023) and neuron gain parameters (Stroud, Porter, Hennequin, & Vogels, 2018). An answer to whether similar results hold when you randomly initialize weight matrices and learn only biases would not only complement this work but help us understand the limits of what can and cannot be learned with random systems.

The current study characterizes the expressivity of neural networks where only the inputs are learned. We first provide theoretical results showing that random networks with learned inputs can approximate a wide variety of functions and dynamical system trajectories in the *iid* and temporal cases respectively, and then show experimental validation. By demonstrating the extent to which one can solve problems through tuning inputs alone, our work provides an important perspective on learning in biological and artificial neural networks.

Theory

Our theory builds off classic results on universal function approximation (Hornik, Stinchcombe, & White, 1989). For concreteness we state results for ReLU activations only. Note, however, that several key results apply more generally. In the interest of space we state theorems without proof. Let ν be the number of non-bias parameters associated with a single hidden unit in the given neural network. E.g. in the feed-forward case ν will be the dimension of the input plus that of the output. Let p_α be a uniform distribution on the zero-centered ball of radius α in a ν dimensional real space.

Lemma 0.1 *Consider a single hidden layer, feed-forward, ReLU activation neural network whose weight parameters for each hidden unit are sampled from p_α. We can find a hidden layer width and bias vector such that, with a probability that is arbitrarily close to 1, the random-weight neural network approximates any continuous function on compact support with any strictly positive degree of accuracy.*

We study the (partially observable) dynamical system $z_{n+1} = F(z_n, x_n),\ y_n = g(z_n),\ z_0 \in U$ where z_n, x_n, and y_n are real vectors of finite dimension, g is an arbitrary continuous function, and U is compact. Assume further that F is continuous and arbitrary except for the constraint that for any input x (in a set of interest), and any $z \in U$, $F(z,x) \in U$.

Theorem 0.1 *Consider a single hidden layer RNN with ReLU activations whose input, output, and recurrent weight parameters for each hidden unit are sampled from p_α. We can find a hidden layer width, a bias vector, and a hidden-state initial condition for the RNN such that, with a probability that is arbitrarily close to 1, the RNN approximates finite trajectories from the above dynamical system with any strictly positive degree of accuracy.*

Remark: a key element of our proofs is that the random network hidden layers are much larger than neural networks which have fully-tuned weights.

Experiments

Feedforward networks

We first show that the same randomly initialized weight matrices can be frozen and used to solve multiple tasks by changing only input biases. We train a feedforward network with 1 hidden layer of 32 000 neurons and no output layer biases. Individual weights are sampled uniformly on $[-0.1, 0.1]$ and reused with only input biases trained via backpropagation for 20 epochs separately on 7 different tasks: MNIST (Deng, 2012), KMNIST (Clanuwat et al., 2018), Fashion MNIST (Xiao, Rasul, & Vollgraf, 2017), Ethiopic-MNIST, Vai-MNIST, and Osmanya-MNIST from Afro-MNIST (Wu, Yang, & Prabhu, 2020), and Kannada-MNIST (Prabhu, 2019). All tasks involve classifying 28x28 grayscale images into 10 classes. For MNIST, Ethiopic-MNIST, Vai-MNIST, Osmanya-MNIST, and Kannada-MNIST, the classes are digits used by different cultures; for Fashion MNIST, classes are clothing types; for KMNIST, they are Japanese Hiragana characters.

We run 5 random seeds for each dataset and compare with a fully-trained network of the same architecture. The results are shown in Figure 1. We see that a network with frozen weights and only trained biases achieves comparable performance to the fully-trained network on all tasks, indicating the potential for multi-task learning on the same set of weights.

Finally, we have also observed that one requires larger hidden layers to achieve similar results to fully trained networks (figures not shown). Whether one requires more or less total trained parameters is currently being investigated.

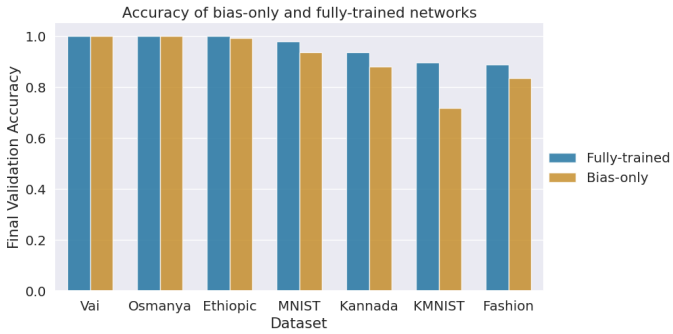

Figure 1: Accuracy of feedforward networks with hidden layer width of 32000 after training on different datasets for bias-only (same randomly initialized weights) and fully trained approaches with the same architecture. We use 5 runs of 20 epochs with different random seeds for the bias-only networks. Error bars have been omitted as the standard errors are of order 10^{-3}. We see that bias-only networks achieve comparable performance to the fully trained networks, demonstrating their effectiveness and flexibility.

Recurrent networks

We extend this framework to a single layer vanilla RNN where the input, recurrent, and readout weights are kept frozen and only the input biases are learned. As in the feedforward case, each weight is sampled from a uniform distribution on $[-0.1, 0.1]$ and 5 random seeds are used.

For the first experiment, networks with increasing numbers of hidden units–256, 512, and 1024–are trained to predict the value of the next time-step in a sum of sine waves given a sequence of the five previous values in time. We find that learning the biases is enough to reconstruct the signal in an equally-sized time window that does not overlap with the one on which it was trained, and importantly, the R^2 coefficient of the model increases as a function of width.

We also experiment with predicting future timesteps of the chaotic Lorenz attractor. Once again, each value is predicted given a history of the previous five time-steps, and the network is evaluated on a non-overlapping window. To increase the difficulty of the reconstruction task, the testing window is corrupted with Gaussian noise. We find that, despite the added noise, a network of width 1024 is sufficient to predict the general trajectory of the system in all three dimensions.

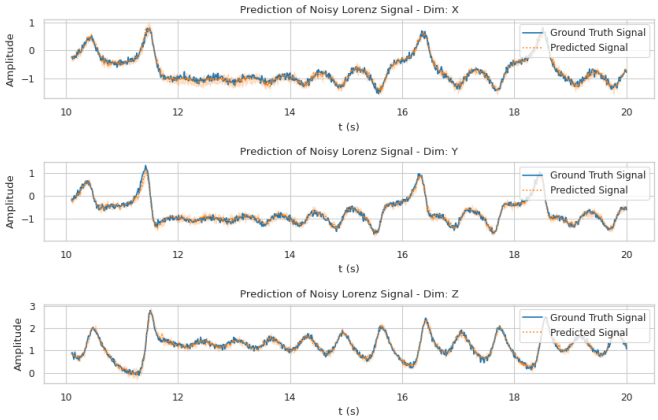

Figure 2: Reconstructing a Lorenz attractor using next-step prediction. The signal was corrupted with Gaussian noise sampled from $\mathcal{N}(0, 1)$ and then normalized. The parameters of the signal ($\sigma = 10, \rho = 28, \beta = \frac{8}{3}$) were chosen such that the system exhibits chaotic behaviour.

Conclusion

In this work, we demonstrate, analytically and empirically, that bias-only learning in feedforward and recurrent networks is more expressive than one might have anticipated. However, theory and experiments suggest the requirement, in networks with only biases trained, of a hidden layer that is larger–sometimes by a massive margin–than one would need if training weights. In this way we view our work as a proof of the existence of solutions in bias-only networks. Going forward we aim to explore how the computational efficiency of bias-only learning may be improved.

Acknowledgments

EW is funded by a NSERC CGS-D scholarship; TJ is funded by a NSERC CGS-M scholarship; MP is funded by Fonds de recherche du Quebec - Santé (chercheurs-boursiers en intelligence artificielle) and the Brain Canada Foundation Future Leaders in Canadian Brain Research Program; LM acknowledges NSF CAREER 2238247; GL acknowledges CIFAR and Canada chair programs.

References

Burkholz, R. (2023). Batch normalization is sufficient for universal function approximation in cnns. In *The twelfth international conference on learning representations.*

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). *Deep learning for classical japanese literature.*

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Feulner, B., Perich, M. G., Chowdhury, R. H., Miller, L. E., Gallego, J. A., & Clopath, C. (2022). Small, correlated changes in synaptic connectivity may facilitate rapid motor learning. *Nature communications*, *13*(1), 5163.

Giannou, A., Rajput, S., & Papailiopoulos, D. (2023). The expressive power of tuning only the normalization layers. *arXiv preprint arXiv:2302.07937*.

Gonon, L., Grigoryeva, L., & Ortega, J.-P. (2023). Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, *33*(1), 28–69.

Hart, A. G., Hook, J. L., & Dawes, J. H. (2021). Echo state networks trained by tikhonov least squares are l2 (μ) approximators of ergodic dynamical systems. *Physica D: Nonlinear Phenomena*, *421*, 132882.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Ogawa, S., Fumarola, F., & Mazzucato, L. (2023). Multitasking via baseline control in recurrent neural networks. *Proceedings of the National Academy of Sciences*, *120*(33), e2304394120.

Prabhu, V. U. (2019). Kannada-mnist: A new handwritten digits dataset for the kannada language. *arXiv preprint arXiv:1908.01242*.

Rahimi, A., & Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, *21*.

Scardapane, S., & Wang, D. (2017). Randomness in neural networks: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(2), e1200.

Stroud, J. P., Porter, M. A., Hennequin, G., & Vogels, T. P. (2018). Motor primitives in space and time via targeted gain modulation in cortical networks. *Nature neuroscience*, *21*(12), 1774–1783.

Wu, D. J., Yang, A. C., & Prabhu, V. U. (2020). *Afro-mnist: Synthetic generation of mnist-style datasets for low-resource languages.*

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.*

Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2021). Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.