Two-argument activation functions learn soft XOR operations like cortical neurons

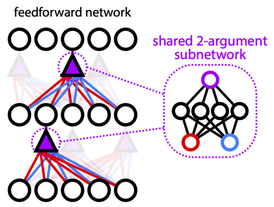
Kijung Yoon[1 *], Emin Orhan[2], Juhyun Kim[1], and Xaq Pitkow[3]

[1]Department of Electronic Engineering, Hanyang University [2]Center for Data Science, NYU [3]Department of Neuroscience, BCM; Department of ECE, Rice University [*]kiyoon@hanyang.ac.kr

HANYANG UNIVERSITY SEOUL, KOREA

Introduction

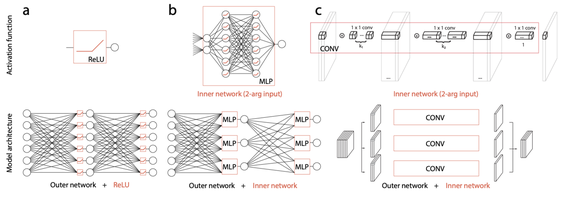
- Neurons in the brain are not simply linear filters followed by a half-wave rectification, and exhibit properties like divisive normalization, coincidence detection, and history dependence.
- **We are particularly interested in multivariate nonlinearities** like $f(w_1^\top x, w_2^\top x, \ldots)$ which could allow inputs that arise from multiple distinct pathways such as feedforward, lateral, or feedback connections, or different dendritic compartments.
- Recent work showed that a single dendritic compartment of a single neuron can compute the exclusive-or (XOR) operation. The finding highlights the possibility that individual neurons may be much more sophisticated than is often assumed in machine learning.
- Many single-argument nonlinearities permit universal computation, but **the right nonlinearity could allow faster generalization during inference, both for the brain and for artificial networks.**



Multi-arg nonlinearities in artificial neurons.

Model structure

- The **inner network** aims to learn an arbitrary multivariate nonlinear function $f(x_1, \ldots, x_n)$ with n inputs and a single output.
- The **outer network** refers to the rest of the model architecture aside from the activation function.
- Our framework, composed of two disjoint networks, is flexible and general since diverse neural architectures can be used as outer networks, such as MLPs, CNNs, ResNets, etc.

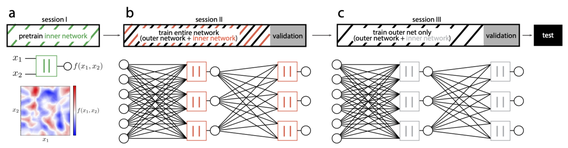


Overview of the proposed model structures.

Experiments

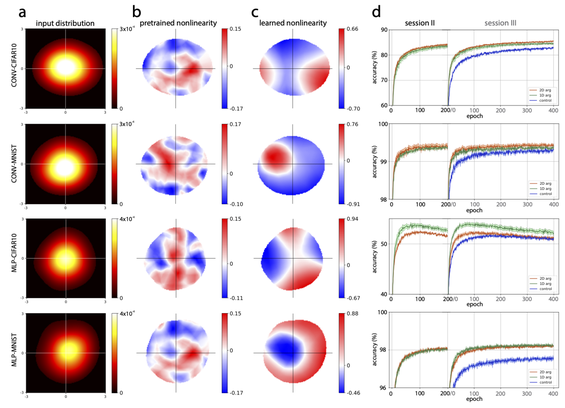
Training procedure

- **Pretraining:** Schematic of two-input argument inner network (green) trained to predict a smoothed random initial activation map (bottom).
- **Training inner and outer networks:** Simultaneously training inner (red) and outer (black) networks.
- **Training outer network for fixed inner network:** Retraining outer network (black) with frozen inner networks (gray).



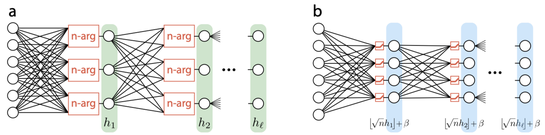
The learned two-input argument nonlinearities

- If two-argument nonlinearities learned what is essentially a one-argument structure, we would see parallel bands of constant color.
- **Instead, notably, all the examples show nontrivial two-dimensional structure, reflecting interactions between two input arguments.**



Comparing to other nonlinearities

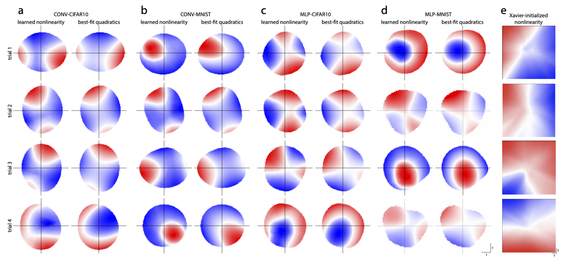
- We use comparable numbers of learnable parameters in the classification tasks by systematically adjusting the number of hidden units or feature maps in each layer.
- The models with learned activation functions learn to achieve an **overall strong performance.**



- The steeper rise in the learning curve shows that our **proposed network learns faster than the ReLU network.**

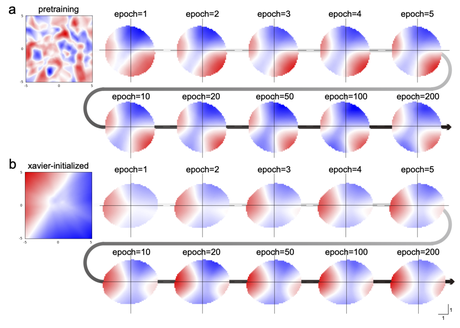
Explicit polynomial nonlinearities

- Multiple trials demonstrates that learned two-argument nonlinearities are reliably shaped like quadratic functions, varying by shifts and/or rotations.
- We fit an algebraic quadratic function and find that **the learned nonlinearity and its best-fit quadratics have extremely similar structure.**
- It indicates a **multiplicative interaction** between the input features, and is **consistent with a gating interaction or soft XOR.**



Evolution of learned two-input nonlinearities

- Similar training curves in session II & III **indicate that the most of the learning is attributable to the change of parameters in outer network.**

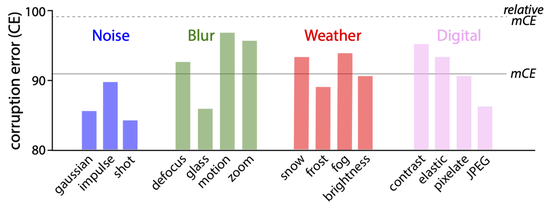


- They evolve to display the typical two-dimensional structure in similar fashion, **strongly suggesting the third-order inferences are an emergent property from learnable multi-argument nonlinear activation functions.**

Generalization

Robustness against common image corruptions

- The robustness gain achieved by the network with two-input argument nonlinearities is significantly larger than the baseline model with ReLU (mCE = 91.3%).



Adversarial robustness

- For each classifier we report the accuracy on the robustness test, at the ε specified in the table, on the whole test set obtained by the ensemble AutoAttack.
- In all cases, AutoAttack reveals a greater robustness in the networks with two-argument nonlinearities than the baseline with ReLU.

		AutoAttack		
Dataset	Outer-Net	2-arg	ReLU	increment
MNIST ($l_\infty, \varepsilon = 0.3$)	MLP	28.79	22.46	6.33
MNIST ($l_\infty, \varepsilon = 0.3$)	Conv	18.65	9.69	8.96
CIFAR-10 ($l_\infty, \varepsilon = 0.031$)	MLP	5.24	2.97	2.27
CIFAR-10 ($l_\infty, \varepsilon = 0.031$)	Conv	2.56	2.51	0.05

Conclusion

The learned two-argument nonlinearities reliably produce soft XOR functions, consistent with recent experimental observations about interactions between inputs in human cortical neurons.

The multiplicative interaction between the input features suggests that the gating interaction or soft XOR is an emergent property from learnable multi-argument nonlinear activation functions.