Approximating Localized Functions using Shallow Neural Networks

Julia Nakhleh, Robert Nowak
University of Wisconsin-Madison
jnakhleh@wisc.edu

ETI Annual Workshop, February 8 - 9, 2023

5

Introduction and Problem Overview

Fully-connected neural networks are compositions of linear functions, parameterized by learned weights, with nonlinear activation functions such as the rectified linear unit, or ReLU (see Figure 1). The representational cost of a neural network can be captured by a norm on its weights, i.e., an indicator of how "large" the network is. The norm of the network can be viewed as a measure of inductive bias; hence the goal of training a neural network is to learn a function that fits the training examples reasonably well, but whose norm is not too large.

Functions which are bounded and compactly-supported (i.e., "localized") are difficult to approximate using *shallow* ReLU neural networks of finite norm. For example, not all compactly-supported piecewise linear functions in greater than one dimension are exactly expressible as a single-hidden layer ReLU network of bounded norm, even if the network has infinite width [1]. In contrast, many localized functions are exactly expressible as finite-norm ReLU network of two or more hidden layers (see Figure 2). This fact seems to be consistent with the empirical finding that deeper networks often perform better in practice; however, the exact nature of the relationship between network depth, width, and approximation accuracy for broader classes of functions is not yet well understood.

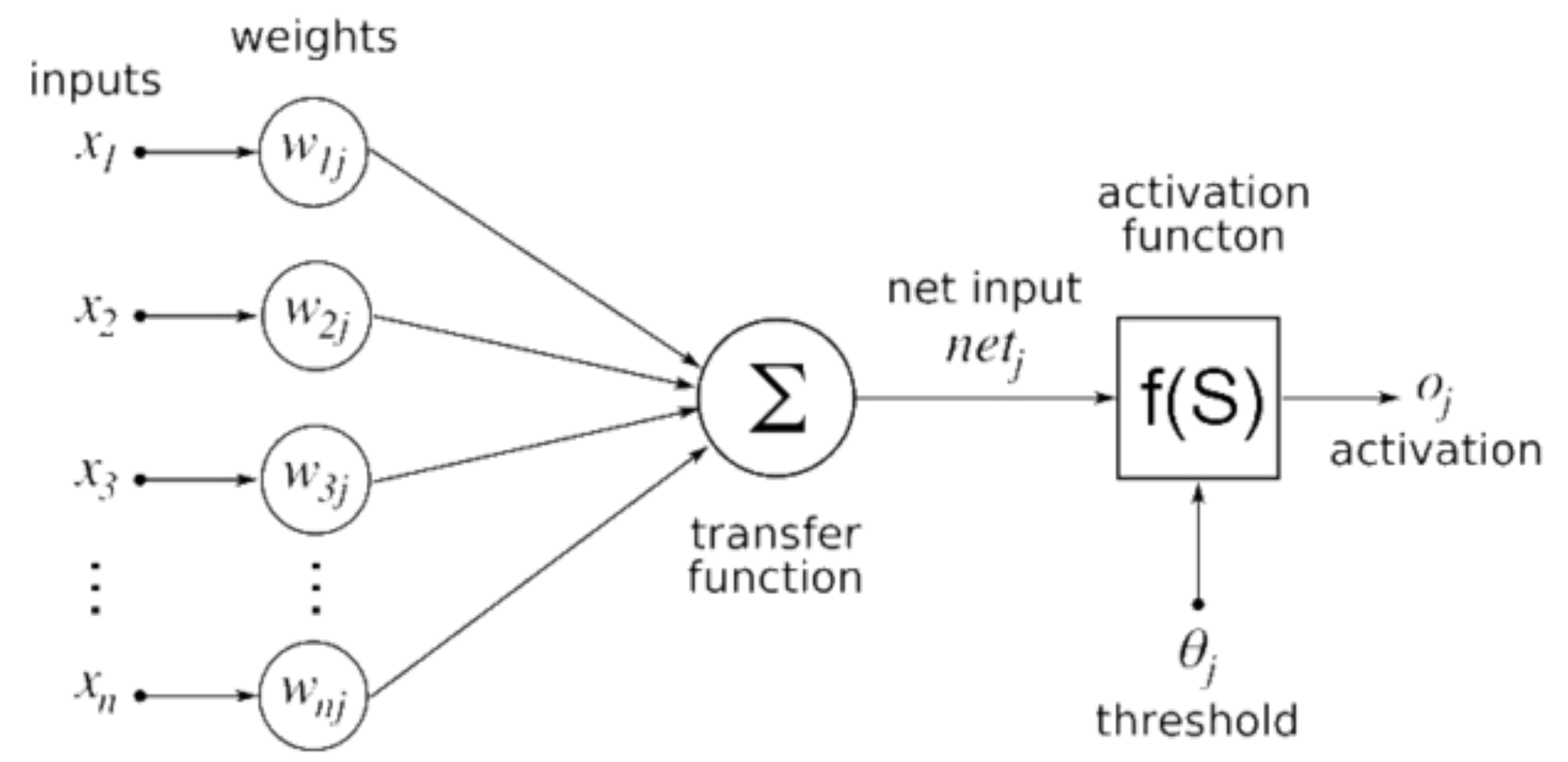

Figure 1: architecture of a single-hidden-layer, single-output neural network.

Existing Results, Methods and Next Steps

Preliminary existing results indicate that for some localized functions, the mathematical properties of interpolations spaces can be used to bound the approximation rate of single-layer network approximations in terms of the number of hidden-layer units. Bounds of this nature may be useful since they can help indicate whether reduced depth can be mitigated by increased width, even if an exact representation is unachievable; they also may provide insight as to how the dimension of the input data affects approximation accuracy. The existing bound attained with this method is loose and does not depend on input dimension, but it may be possible to achieve a tighter bound which includes dimensionality using more careful mathematical analysis. It also remains to be seen whether similar bounds can be obtained for broader classes of localized functions, which may require more difficult analytical techniques.

Additionally, it may be possible to provide useful bounds on how well certain non-smooth localized functions (which are not always representable by a finite-norm single-layer network) can be approximated using smooth localized functions (which are representable in this manner), and how large the norms of these smooth approximations are. Both the smoothness and "spikiness" of localized functions appear to affect their representational cost using shallow networks, but the limits of this finding are not known. For example, smooth "radial bump" functions can be represented using finite-norm shallow networks whose representational cost is inversely proportional to the radius of the bump [1]. For non-smooth localized functions which can be reasonably well-approximated using bump functions of this type, it may be possible to use this finding to generate similar approximation rates.

Relevance to Program Objectives

Neural networks exhibit state-of-the-art performance on a wide variety of pattern-recognition tasks, and can be trained as effective anomaly detectors across multiple problem domains. However, their demonstrated success is largely empirical, and many of their mathematical properties are not yet well-understood. Improved understanding of the approximation properties of shallow vs. deep neural networks is relevant for developing a mathematically rigorous theory of deep learning, and it may help guide machine learning practitioners in their choice of model architecture when applying deep learning to real-world problems.

Localization via Depth

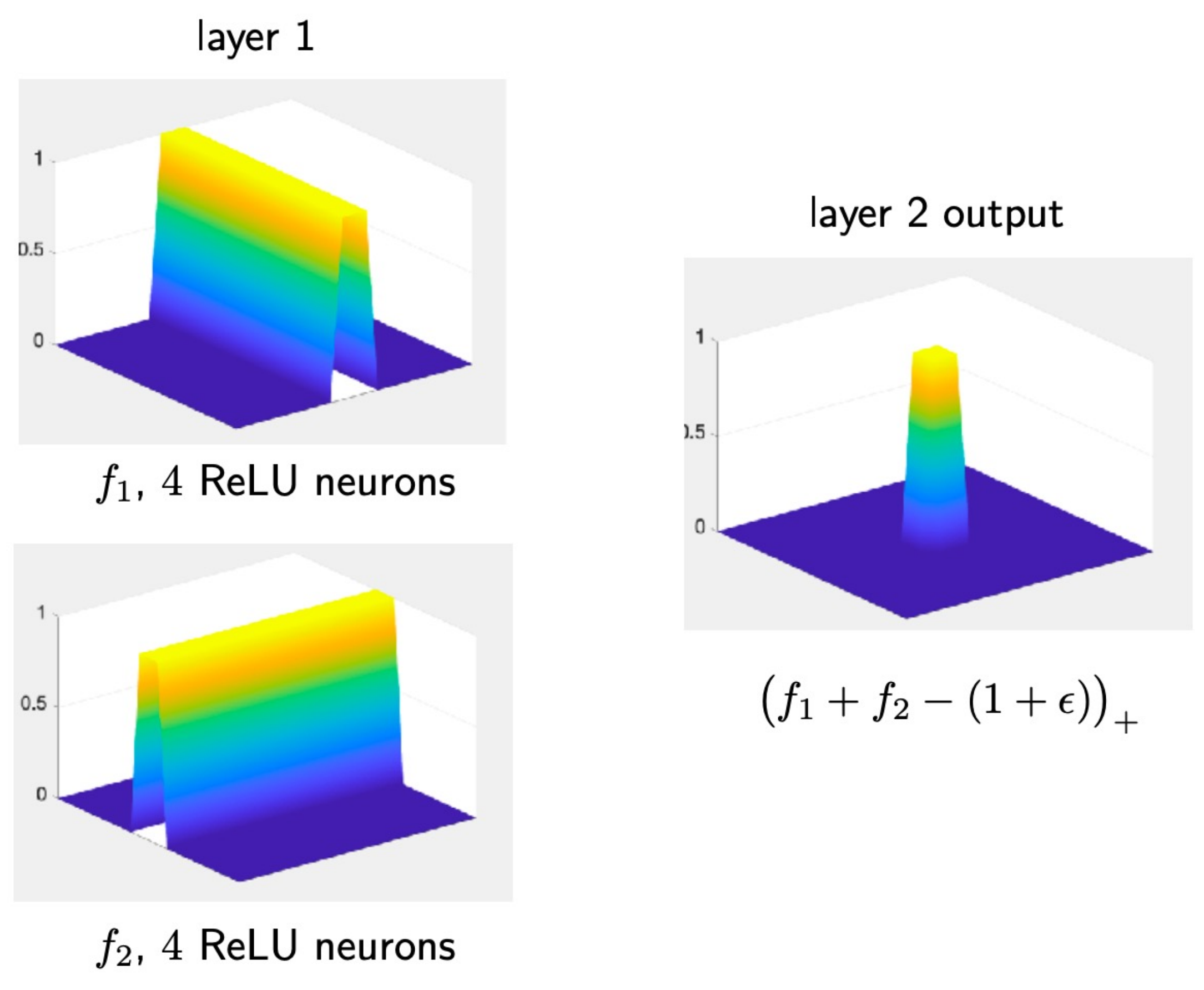

Figure 2: illustration of localization for single- vs double-layer ReLU networks.

References

[1] G. Ongie, R. Willett, D. Soudry, and N. Srebro, "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case," presented at the Eighth International Conference on Learning Representations, Apr. 2020.

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.