Extensions of the Methods for Understanding Neural Network Generalization

Thomas Walker

Contents

Part I
Exploring Stiffness

Here I extend the work of [1] to start to understand how stiffness relates to the architecture of the neural network.

1 Neighbourhoods Stiffness

The stiffness in neighbourhoods can be used to identify features of the dataset. Points along the boundary of a class for example will experience differing gradients to its neighbours. Performing weighted averages of the stiffness between points in a neighbourhood can therefore identify the boundary of features within a dataset.

In Figure 2 we compare how the stiffness of samples in relation to other examples varies through the layers of a trained and untrained network.

In Figure 3 look at how stiffness between examples varies through the layers for networks of different architecture.

Affiliation Imperial College London
Correspondence thomas.walker21@imperial.ac.uk
Date July 2023

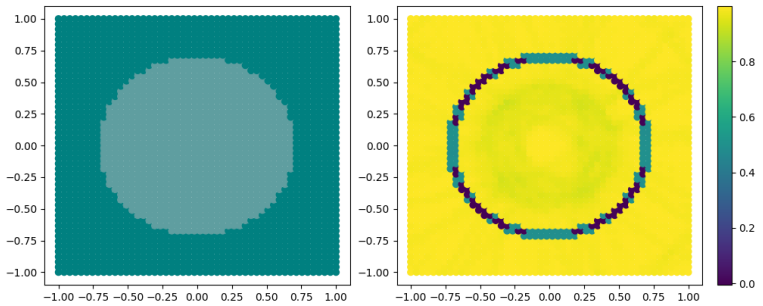

Figure 1: The stiffness of points within a neighbourhood can be used to identify the learned boundary of a classification problem.

References

[1] Stanislav Fort, Pawel Krzysztof Nowak, and Srini Narayanan. “Stiffness: A New Perspective on Generalization in Neural Networks”. In: *CoRR* (2019).

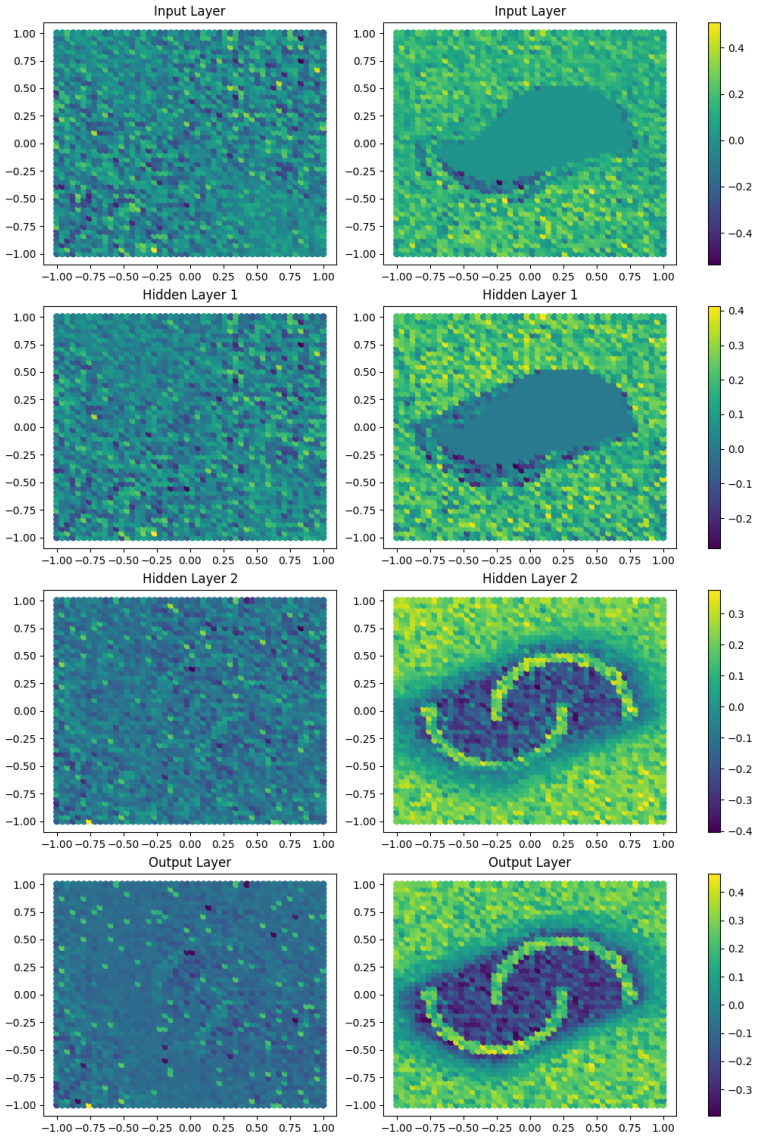

Figure 2: Average stiffness between an example and a random sample from the training data containing samples from both categories.

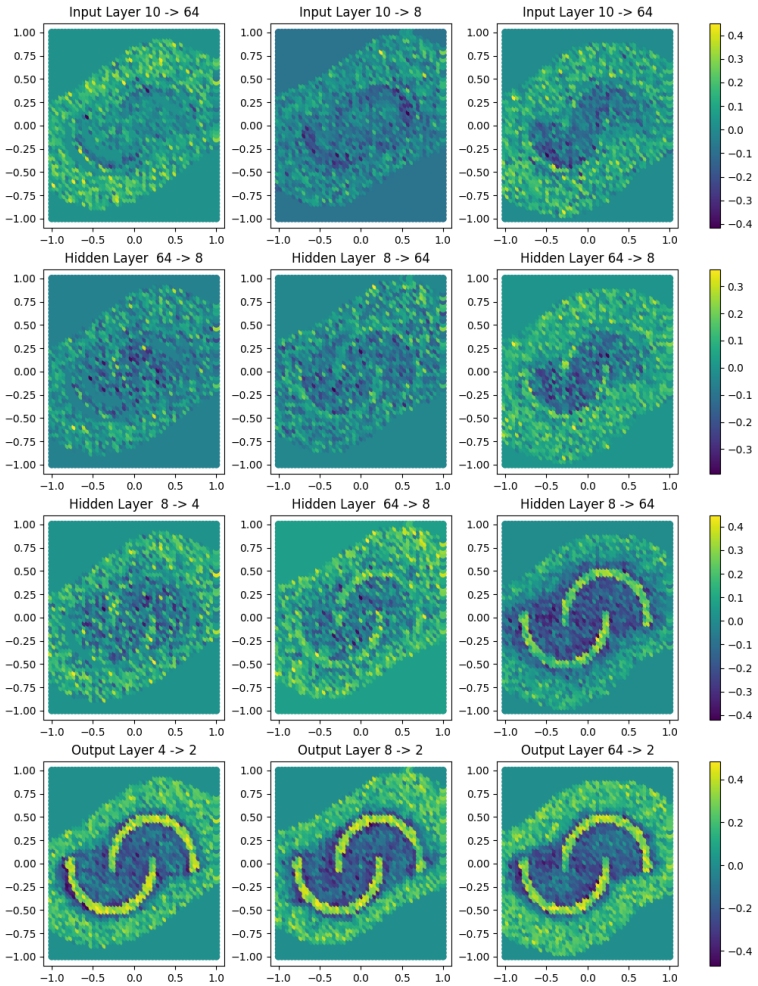

Figure 3: This shows how the stiffness of examples changes through the layers of networks with different numbers of hidden units.