CAN A CONVEX PARTITION CAUSED BY A CPWL NEURAL NETWORK BE USED FOR DENSITY ESTIMATION?

Anonymous authors
Paper under double-blind review

The increasing popularity of deep learning methods has led to a growing interest in understanding the theoretical properties of neural network. The use of information theory, as proposed by Tishby & Zaslavsky (2015); Shwartz-Ziv & Tishby (2017), provides a promising framework for studying the behaviour of these models. However, calculation of information theoretic (IT) quantities rely on estimating high-dimensional probability densities, which is challenging. Geiger (2022) show that the numerical result of different IT-measures varies depending on which estimator is used. In attempt to combat this, we want to use the convex partitioning properties of Continous Piecewise Linear (CPWL) neural network, such as ReLU networks, in order to create "natural bins" of the models input and latent spaces.

A ReLU network will partition its input space into convex regions (Serra et al., 2018; Hanin & Rolnick, 2019a). The number of regions is upper bounded by 2^H where H is the total number of hidden units in the network. However, only a small number of these are considered feasible regions Hanin & Rolnick (2019b). Finding these regions for the whole network is previously explored by Liu et al. (2023); Sattelberg et al. (2023); Humayun et al. (2024). However, all of these approaches are limited to specific cases, or weak numerical implementations. Our goal is to create a more general and scalable algorithm for finding the feasible regions of ReLU network, both for the whole network and for each layer. The ultimate goal is to use these regions as bins when performing density estimation for IT-measures.

If successfull, this could provide a more principled way of estimating IT-measures for neural networks, free from arbitrary choices of binning or kernel sizes and other hyperparamaters. In addition, we would be able to compute the mutual information between the input and output, based on how the network partitions input space. The hypothesis is that this should increase as the network learns how to partition input space in a feasible way. Furthermore, we should be able to spot bottlenecks in the network, by looking at how the number of regions changes from layer to layer, and which layers that causes the most reduction in number of regions.

REFERENCES

Bernhard C. Geiger. On Information Plane Analyses of Neural Network Classifiers – A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (12):7039–7051, December 2022. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2021.3089037.

Boris Hanin and David Rolnick. Complexity of Linear Regions in Deep Networks, June 2019a.

Boris Hanin and David Rolnick. Deep ReLU Networks Have Surprisingly Few Activation Patterns, October 2019b.

Ahmed Imtiaz Humayun, Randall Balestriero, Guha Balakrishnan, and Richard Baraniuk. SplineCam: Exact Visualization and Characterization of Deep Network Geometry and Decision Boundaries, June 2024.

Yajing Liu, Christina M. Cole, Chris Peterson, and Michael Kirby. ReLU Neural Networks, Polyhedral Decompositions, and Persistent Homology. June 2023.

Ben Sattelberg, Renzo Cavalieri, Michael Kirby, Chris Peterson, and Ross Beveridge. Locally linear attributes of ReLU neural networks. *Frontiers in Artificial Intelligence*, 6, November 2023. ISSN 2624-8212. doi: 10.3389/frai.2023.1255192.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and Counting Linear Regions of Deep Neural Networks, September 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information, April 2017.

Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle, March 2015.