Seminar at the Department of Business Analytics at the University of Iowa

Friday November 12, 2021. Time: 11:00am-12:00pm (US central time)

Title: What Makes Neural Networks So Expressive, and What Could Make Them Smaller? Some Answers Based on Polyhedral Theory and Mixed-Integer Linear Programming

Speaker: Thiago Serra, Assistant Professor of Analytics & Operations Management, Bucknell University

Abstract:
Neural networks have been successfully applied to complex predictive modeling tasks in areas such as computer vision and natural language processing. On the one hand, they have been shown to be a very powerful mathematical modeling tool: a neural network can model a piecewise-linear function with an exponential number of pieces with respect to its number of artificial neurons. On the other hand, we may still need an unreasonably large neural network in order to obtain a predictive model with good accuracy in many cases. How can we reconcile those two facts?

In this talk, we apply traditional tools from operations research to analyze neural networks that use the most common type of artificial neuron: the Rectified Linear Unit (ReLU). First, we investigate both theoretically and empirically the number of linear regions that networks with such neurons can attain, which reflect the number of pieces of the piecewise linear functions modeled by those networks. With respect to that metric, we unexpectedly find that sometimes a shallow network is more expressive than a deep network having the same number of neurons. Second, we show that we can use optimization models to remove units and layers of a neural network while not changing the output that is produced, which thus implies a lossless compression of the network. We find that such form of compression can be facilitated by training neural networks with certain types of regularization that induce a stable behavior on its neurons.

This talk is based on papers coauthored with Christian Tjandraatmadja (Google Research), Srikumar Ramalingam (Google Research), Xin Yu (The University of Utah), and Abhinav Kumar (Michigan State University) and published at ICML 2018 (https://arxiv.org/abs/1711.02114), AAAI 2020 (https://arxiv.org/abs/1810.03370), CPAIOR 2020 (https://arxiv.org/abs/2001.00218), and NeurIPS 2021 (https://arxiv.org/abs/2102.07804).

Speaker Bio:
Thiago Serra is an assistant professor of analytics and operations management at Bucknell University's Freeman College of Management. Previously, he was a visiting research scientist at Mitsubishi Electric Research Labs from 2018 to 2019, and an operations research analyst at Petrobras from 2009 to 2013. He has a Ph.D. in operations research from Carnegie Mellon University's Tepper School of Business, and received the Gerald L. Thompson Doctoral Dissertation Award in Management Science in 2018. During his PhD., he was also awarded the INFORMS Judith Liebman Award and a best poster award at the INFORMS Annual Meeting. His work on neural networks has been published at ICML, AAAI, CPAIOR, and NeurIPS; received a best poster award at the Princeton Day of Optimization; and the National Science Foundation (NSF) award 2104583 (CRII:

RI: RUI: Principled Methods for Compressing Neural Networks through Discrete Optimization and Polyhedral Theory).

Dr. Serra is the principal investigator of the Advanced Analytics Research Lab, in which undergraduate students from all Bucknell colleges have been developing research projects involving machine learning and mathematical optimization with the support of NSF and endowment-supported programs at Bucknell such as the Emerging Scholars and the Presidential Fellowship.