On Developing Efficient and Robust Neural Networks for Healthcare using Condensa Model Compression System

Keaton Rowley
University of Utah

UUCS-21-015

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

18 May 2021

Abstract

This thesis describes a study of how compressing neural networks used to identify malaria cells and respiratory diseases affects network accuracy, and system performance metrics. Its focus is on a state-of-the-art framework for neural network (NN) compression called Condensa to compress network size and improve network performance according to different compression schemes. It details the impact malaria and lung disease have on a worldwide level each year. It then describes previous research in automating medical image classification. It also gives a background on what research has been applied towards network compression. The study also describes work in developing a CNN for the Malarial and Chest-X-ray datasets. It details the results of compressing the CNN using Condensa's Filter, StructPrune, Prune, and Quantization schemes. This thesis provides a complete software implementation to help reproduce our results and facilitate tool adoption. It also indicates a plan for future research in applying Condensa towards the problem of developing an efficient system of disease identification for different medical dataset problems.