DS 785: Capstone Final Documentation

By Mukund Raghav Sharma

COMPARATIVE CASE STUDY BETWEEN GATED GRAPH NEURAL NETWORKS VERSUS RELATIONAL GRAPH CONVOLUTIONAL NETWORKS FOR THE VARIABLE MISUSE TASK USING PYTHON AND DEEP LEARNING

Author Introduction

Mukund Raghav Sharma who goes by Moko and currently resides in Bellevue, WA and is originally from New Delhi, India. He graduated with a Bachelor's of Science in Electrical and Computer Engineering from Carnegie Mellon University in 2015.

He currently works as a software developer at Microsoft's Developer Division on the Performance and Reliability of Visual Studio, a popular Integrated Development Environment. He has previously worked at a Proprietary Trading Firm called Susquehanna International Group near Philadelphia as a Statistical Options Trading developer where he heard about the University of Wisconsin's Master's Program in Data Science. Prior to working at Susquehanna International Group, he has also worked at a hedge fund called AVM in South Florida and Spot Trading LLC in Chicago.

He is extremely passionate about software engineering and its intersection with data science, value investing and has been playing the guitar for more than 15 years.

Executive Summary

The purpose of the comparative case study based project was to discern the better model between models based on Gated Graph Neural Networks (GGNN) and Relational Graph Convolutional Networks (RGCN) on the Variable Misuse Task, a prediction task involving discerning the correct variable to be used in a particular spot amongst all variables of the same type in the particular scope. The comparison between GGNN and RGCN models involved computing the test accuracy on three experiments the source data of which is obtained by downloading the source code of the top 25 trending C# repositories on Github. These three experiments involved training and obtaining the test accuracy of all the repositories, an esoteric and popular repository to deduce which model was more performant across different types of source code.

The overarching goal for this project was to discover which model would generalize and perform better in the Static Analysis tooling space that's typically rule based by inculcating the representational power of Deep Learning to solve more state-of-the-art problems.

The Data Science concepts related to this capstone are machine learning (specifically deep learning), hyperparameter optimization, visualization techniques, graph theory and natural language processing. Also, putting together the final report involved a lot of the writing techniques to present the data in a simple yet compelling manner.

The results from the data highlighted that the Relational Graph Convolutional Network outperformed the Gated Graph Neural Network across all experiments, although, within a margin of 5%. Training on more data resulted a higher test accuracy for both models and a smaller difference between the two.