**COMPARATIVE CASE STUDY BETWEEN GATED GRAPH NEURAL NETWORKS**

**VERSUS**

**RELATIONAL GRAPH CONVOLUTIONAL NETWORKS FOR THE VARIABLE MISUSE TASK**

**USING PYTHON AND DEEP LEARNING**

**By**

**Mukund Raghav Sharma**

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in

Data Science

University of Wisconsin – La Crosse

La Crosse, Wisconsin

May 2020

**DEDICATION**

**To my grandparents and my favorite planet, Saturn.**

**ABSTRACT**

**GATED GRAPH NEURAL NETWORKS VS RELATIONAL GRAPH CONVOLUTIONAL NETWORKS FOR THE VARIABLE MISUSE TASK**

**MUKUND RAGHAV SHARMA**

Identifying bugs in source code has been an extremely important part of software development since the inception of the industry. The majority of static analysis, the analysis of software without actually executing programs, is rule based without much involvement of deep learning until fairly recently. This paper engages in a comparative study of determining the more performant graph neural network model on the basis of test accuracy between Gated Graph Neural Network (GGNN) models and Relational Graph Convolutional Network (RGCN) models on the Variable Misuse Task, a prediction task involving choosing the correct variable based on all the variables of the same type in a particular scope.

The data is of source code from the files of 24 trending C# repositories that are converted into a modified Abstract Syntax Tree to represent a directed graph whose vertices that represent the tokens and relationships between the tokens are represented by edges. Each of these vertices are associated with one of the aforementioned type of networks for the training phase after a particular embedding is computed for each token.

The comparison to decide the more efficient model is based on the test accuracy of all the repositories, an esoteric repository and an extremely popular repository to cover the spectrum of different types of repositories. The results show that the RGCN based models outperformed the GGNN models for all cases, albeit, within < 5% range.

*Keywords: Deep Learning, Graph Neural Networks, Tensorflow, Sequence Models, Convolutional Models, Learning from Code, Static Analysis.*

**TABLE OF CONTENTS**

**[TODO: FIX]**

Abstract

Introduction

Literature Review

Project Overview

Graph Neural Networks 101

Preliminary Steps

Data Exploration

Methodology and Outcomes

Results

Conclusion

**CHAPTER 1**

**INTRODUCTION**

**Project Background**

Since the inception of the computing industry more than 70 years ago, the need for correct and efficient verification tools has been extremely desirous in the software engineering world. Static Analysis tools or tools that analyze the code without executing the program boost the productivity of the developer by highlighting some bugs at development time rather than during when the code is running in production.

Deep learning application in the field of learning from source code and static analysis is still in its nascent phase and a large part of the current work doesn’t take advantage of the representational power of both the syntactic and semantic nature of the code. For example, shallow representations of source code is prevalent in recent research such a simple sequence of tokens such as by work done by Hindle et al. (2012) [1] or flat dependency networks of variables by Raychev et al. (2015) [2].

Recent work from Microsoft Research involves inculcating both the semantic and syntactic nature of source code by representing programs as directed graphs [3] and applying deep learning via sequential or convolutional neural networks to create graph neural network models. Using graph neural network models has proved to more easily solve state-of-the-art problems in the space of learning from code as the representational capacity of these models reduces the need for a large training set as well as meticulously encapsulates the semantic relationships between variables and types of the source code.

In this paper, I use the work by Microsoft Research in the field of Graph Neural Networks to conduct a comparative study between two different graph neural networks namely, Gated Graph Neural Networks (GGNN) and Relational Graph Convolutional Networks (RGCN), on the Variable Misuse Task to discern the better performing model on the basis of test accuracy. The Variable Misuse Task is a prediction based task on source code involving predicting the variable that most accuracy fits the current context from all the variables of the same type in a particular scope.

The Variable Misuse Task is a simple yet an extremely important task that has its uses in the world of static analysis. Microsoft Research’s application of Graph Neural Networks on the Variable Misuse task has caught bugs that had been deployed in production for important repositories such as RavenDB and Rosyln [3]. Further application of Graph Neural Networks on newer tasks show great promise and have already started changing the way code is tested and validated.

The input data used for this model is that of the top 24 trending C# repositories on Github. There are three main experiments conducted to discern the better performing model on the basis of test accuracy and these are training and testing on the source code of all the repositories, an esoteric repository and an extremely popular repository. The rationale here was to gain insight about how the GGNN and RGCN models would fare in different cases and which one of the two would prove to be the better predictor.

**Objectives**

The main objective of this project is to compute the test accuracies for the GGNN and RGCN models for three experiments namely, training and testing on data from 27 of the most popular C# repositories on Github, an esoteric repository and a popular repository, in an effort to discern the more performant model.

This process of computing the test accuracies involved enumerating through some minor, more granular, objectives and these are:

1. Obtaining, Cleaning and Understanding the Input Data.
2. Choosing, from the trending C# repositories, which one would characterize as the esoteric and popular one to conduct our experiments on.
3. Establishing a working pipeline through which the experiments will be conducted.
4. Interpreting results i.e. test accuracy of the experiments to land on a conclusion.

**Rationale and Inspiration for undertaking the project**

The constant strive to build better development tools that will improve the overall user experience is one I am starting to familiarize myself with more since I joined Microsoft’s Developer Division as a software engineer where I work on the Performance and Reliability of Visual Studio. My role involves creating processes that improve the prevention of performance regressions by early detection before they reach the end user. My experience with performance engineering and testing on a large code base coupled with my strong inclination towards deep learning got me going down a path to choose a topic for my capstone that would fit the said intersection.

After discovering Microsoft Research’s ground breaking work, I was convinced I needed to know the ins and outs of the details of how they improved state-of-the-art research models. And as a result to add more specificity to my topic, I decided to choose between a sequence based model (GGNN) and one convolutional based model (RGCN) on a the Variable Misuse Task. Also, I was lucky enough to have all my questions answered within a couple of hours by the good people at Microsoft Research and this was another inspiration booster i.e. to work on ground breaking work with individuals who were extremely eager to help out in improving the understanding of their work.

The choice of repositories were amongst the top trending C# repositories that I have used professionally which made pursuing the data exploration all the more exciting as it provided me with a reason to dive deep into the code, as well. Conducting the comparison on repositories like Newtonsoft.Json was enticing as these are libraries I use on a daily basis and have used all my career being a .NET developer.

**CHAPTER 2**

**LITERATURE REVIEW**

**Previous Work**

Graph Neural Networks

2005

2016

Gated Graph Neural Networks

Graph Convolutional Networks

2017

Neural Message Passing

Programs As Graphs

2018

**Fig. 1: Timeline of Graph Neural Networks Research**

Researching through a lot of the related past work in the field of Graph Neural I was able to narrow down the timeline of key developments in this field details of which are as follows:

**2005**: Gori et al. proposed the notion of Graph Neural Networks that was the seminal research paper in this field that laid down the foundation of future work. Although, deep learning wasn’t at the forefront of statistical research when this paper was written at it is now, important fundamental concepts such as training of neural networks on graphs as well message passing emerged.

**2016**: Li et al. took the core concept of Graph based networks and applied sequence based concepts such as gated cells to neural networks to prevent loss of importance of information in large, extremely spread out sequences represented as graphs. The most important aspect of this model was the propagation model that took the messaging passing logic and applied recurrent gated cells at each vertex.

**2017**: Liao et al proposed idea of Neural Messaging Passing where information between the neighbors of vertices is shared and therefore, not only does each vertex in a graph, through training, have information about its initial state but also other vertices. I will be expounding on this a lot more in a later chapter as this concept has be explained thoroughly for full effect but to summarize, neural message passing allows graph neural network models to make use of their initial state but as well as the state of other vertices in an effort to increase overall representational capacity that requires fewer training data points.

**2017**: Kipf et al. introduced the concept of Graph Convolutional Networks that took plain vanilla graph neural networks and in state update training step used convolution of the state of the neighbors as the mechanism of message passing. The premise here is that the Convolution operation that is typically used in the case of computer vision can be applied in the case of graph neural networks in the case of message passing where each vertex can gain further insight about it’s position with respect to other vertices.

**2018:** Allamanis et al. applied concepts from previous research in the context of learning from programs following the simple yet powerful assertion that source code can be represented as graphs where the vertices represent the variables and it’s associated information and the edges represent the relationship between the variables. Additionally, taking advantage of the semantic and syntactic nature of source code adds more detail to the input of the graphs that can result in quicker training. The majority of this report is built on this paper and its applications and further details will be elucidated upon in future chapters.

The aforementioned concepts at this point might seem difficult to grasp as they aren’t fully elucidated on; I plan to cover the basics of all the concepts in one of the future chapters that’ll simplify the subject matter. Needless to say, the extensive research on this topic of graph neural networks is truly inspiring and exciting as well as growing exponentially with developments in multiple dimensions.

**CHAPTER 3**

**PROJECT OVERVIEW**

**General Approach / Layout**

1. The basic steps undertaken to approach this project include:
2. Defining the Experience, Task and Performance Metric in the style of the formal machine learning once the problem statement was identified that we’ll base the comparison on.
3. Acquisition and exploration of the input data source.
4. Deep diving into the source code associated with the Learning to Represent Graphs library to gain insight about how the code is executed.
5. Once the insight is gained to run the code, customize the code to accommodate for ease of use for my experiments by constructing a working pipeline that’s based on ease of use.
6. Defining the 3 experiments and which data will be used for them.
7. Training models with the training and validation input data on GGNN and RGCN models for the experiments.
8. Training with different hyperparameters that can be used to improve the overall validation accuracy.
9. Using the weights from the training process to compute the test accuracy for RGCN and GGNNs.
10. Aggregate results to discern which model was the more performant one.
11. Conclude with possible next steps.

Now that the goals and outline are set, I want to shed light on what’s actually happening under the hood during the training and evaluation of these models and define the task at hand with more detail.

**CHAPTER 5**

**GRAPH NEURAL NETWORKS 101**

**Goals**

In this section I plan to achieve the following goals:

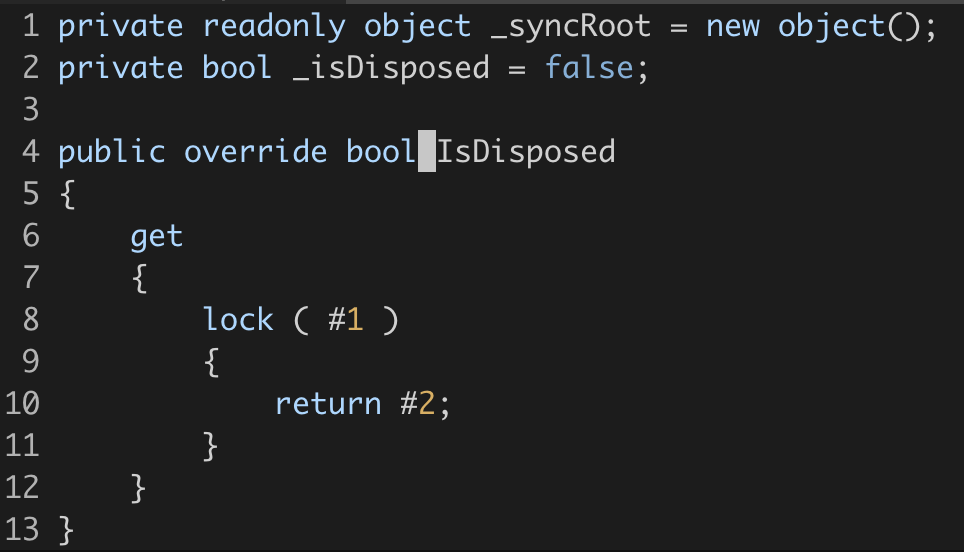
1. Define and Describe the Variable Misuse Task in detail.
2. Explain how Graph Neural Networks work.
3. Deep Dive into Gated Graph Neural Networks.
4. Deep Dive into Relational Convolutional Neural Networks.

By the end of this chapter, the reader should have gained some knowledge about graph neural networks and their details with relation to the comparative study between GGNNs and RGCNs.

**The Variable Misuse Task**

Properly defining and understand the task associated with our statistical learning algorithm is of paramount importance. The Variable Misuse Task, in a nutshell, is a task that requires the algorithm to predict the most fitting variable in a particular scope among all the variables that are of the same type. If there is a discrepancy between what the code highlights and our confidence level of the prediction from our algorithm, the intended action to flag that line of code as a potential bug. To get a better understanding of this task, two examples would be helpful:

**Fig 2: Example 1 of Application of the VarMisuse Task in C#**



This snippet of code is locking on a variable to provide synchronized access and then returning another variable. As a part of this task, predicting which variable is most suited for #1 and #2. And if there is a discrepancy between our confidence levels from the model and what the code indicates, that there is a case of variable misusage. To be more specific, if we were to run our GGNN model on this code that’s been trained on other repositories, we get the following results:

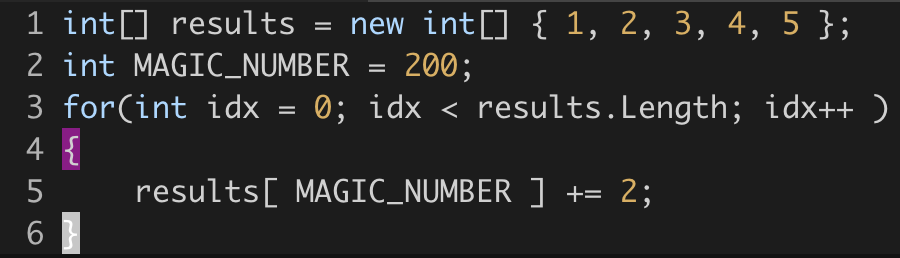
For #1: **\_syncRoot** as the correct variable with a confidence of 95% and **\_isDisposed** as 5%.

For #2: **\_isDisposed** as the correct variable with a confidence of 99% and **\_syncRoot** as the correct variable as 1%.

and these results imply there was no variable misuse bug in this case.

As a counterexample where we do detect a discrepancy between what’s predicted and the actual variable used in the code is:

**Fig 3: Example 2 of Application of the VarMisuse Task in C#**



This simple snippet of code essentially adds two to all the numbers in the array while looping through it. The fact that the code is trying to add 2 to the 200th index of the results array wouldn’t be highlighted as a compilation error and therefore, would require other less rule based methods to detect this at develop or compile time. The results from a well trained hypothetical model would illuminate the at the low confidence of the actual variable used in the source code and complain to the user that there could potentially be a bug with the code while suggesting the variable i.e. idx in this case, with the highest confidence.

I’d like to point out here that this task would be extremely difficult to achieve with high accuracy with a plain vanilla sequence model. The representative capacity of the graph neural network helps with the inference of the role and functions of the variables of the program and allows the efficient learning and retaining of pertinent features. Additionally, the variable misuse task can be used as a seminal example as it is a proxy for a considerable number of static analysis related tasks that can be similarly incorporated due to its similarity to code completion.

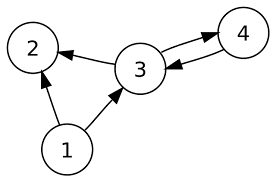
**Graph Neural Networks**

Now that we have defined the task at hand and gained intuition about what we are trying to achieve, it is important to shift gears and talk about Graph Neural Network Models by first going over what a graph is and then how neural networks can be applied for prediction for any feasible task.

Graphs and Representation of Programs

A graph is defined by a set a nodes or vertices and a list of edges connecting the nodes. A simple example of a graph is as follows where 1,2,3 and 4 are the nodes and the arrows are the edges connecting these nodes.

**Fig. 4: Graph Example**

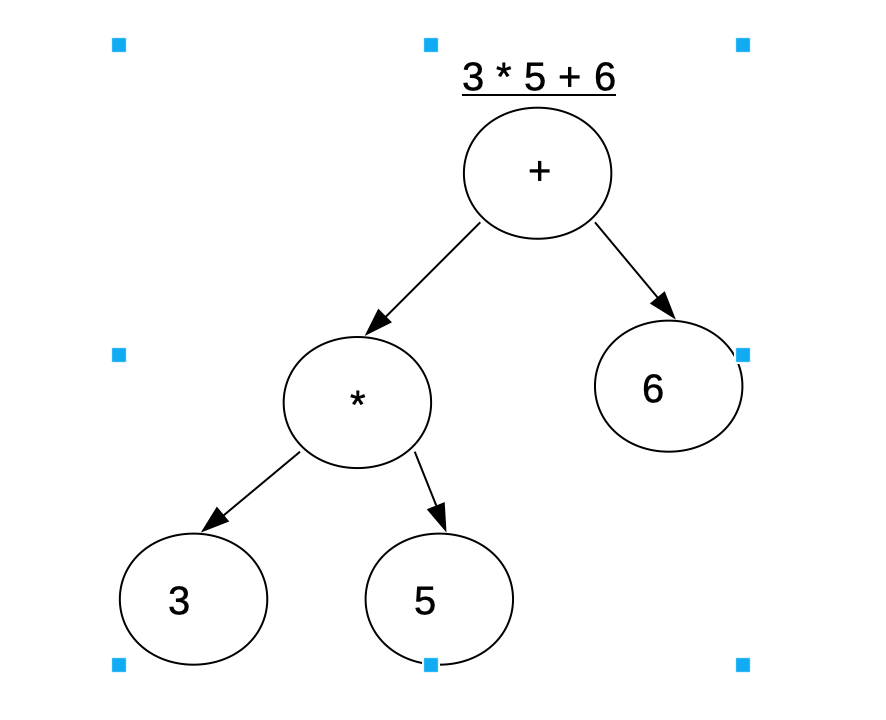


In the context of representation of programs as graphs, we specifically consider directed graphs or graphs with edges that can be pointing in either or both directions between the nodes contrasted with undirected graph where all edges are bi-directional. We further add more characteristics to the plain vanilla directed graph to accommodate the representational capacity we want to achieve. For each node or vertex, we include associated features that will represent the state of the node associated with a token that could be a variable. Additionally, we introduce the concept of edge types that will come in handy later when we want to add different types of edges that represent disparate characteristics of the program such as a different edge type for representing if a variable is reading from another variable and a different one if the variable is writing to it.

Graph Construction

The construction of the graphs that represent programs is a modified version of the Abstract Syntax Tree represented by Rosyln, the .NET compiler framework. The Abstract Syntax Tree is a representation of the abstract syntactic structure of the code where the **syntax nodes** correspond to the programming language’s grammar while the **syntax tokens** correspond to the leafs of the tree that represent the string from the source code.

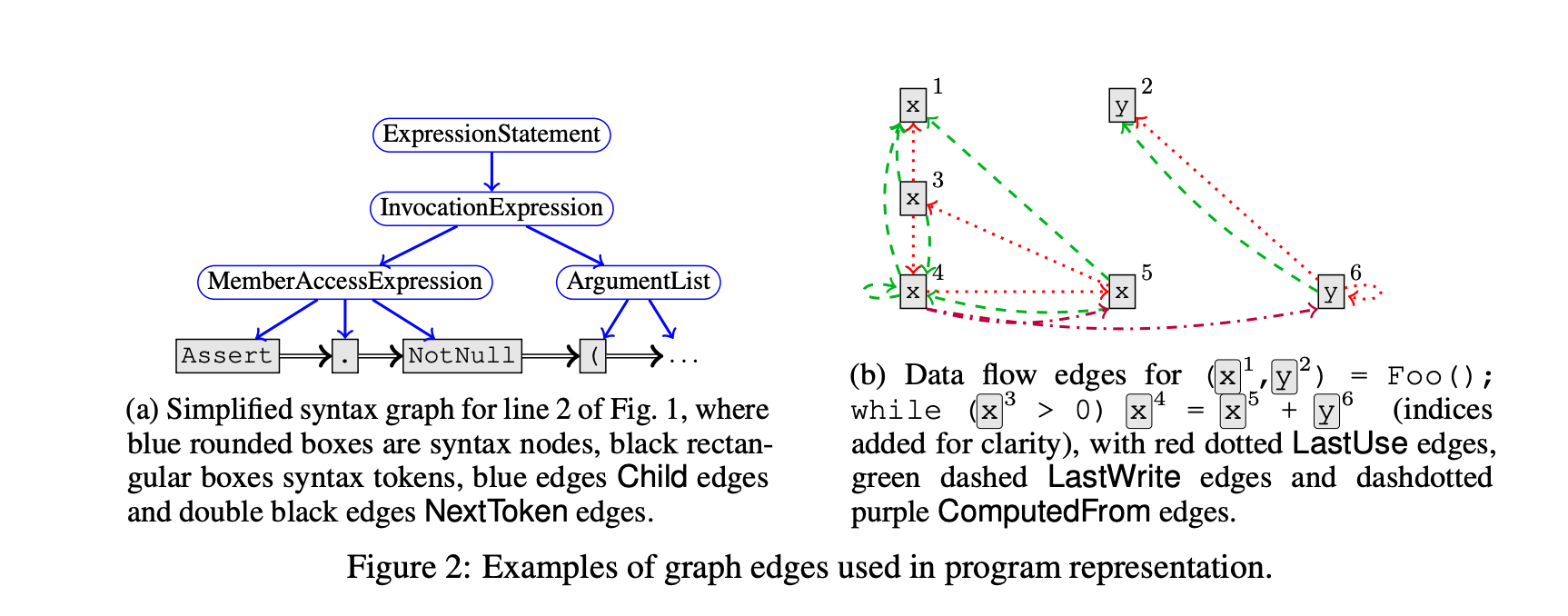
**Fig 5: Simple Abstract Syntax Tree**



The example highlights that for the line of code 3 \* 5 + 6, we have the syntactic tokens i.e. 3,5 and 6 are leafs of the tree and \* and + that are the corresponding arithmetical operators.

It is worth noting that the type of modified syntax trees we will be working with will be a lot more complex. As an example of how these graphs can be visualized, we borrow an example from Allamanis et al. (2018):

**Fig 6: Modified Abstract Syntax Tree that Represent Programs**

[TODO]: Fix the image

From the figure on the left, we observe the corresponding Program Graph for a line of code that asserts if a particular variable isn’t null. The figure on the left highlights the modifications done to the edges among the vertices that are based on the type of relationship exemplified by the code. This modification step is what transforms the abstract syntax tree into a graph by adding different types of directed edges between components.

Node Feature Embeddings

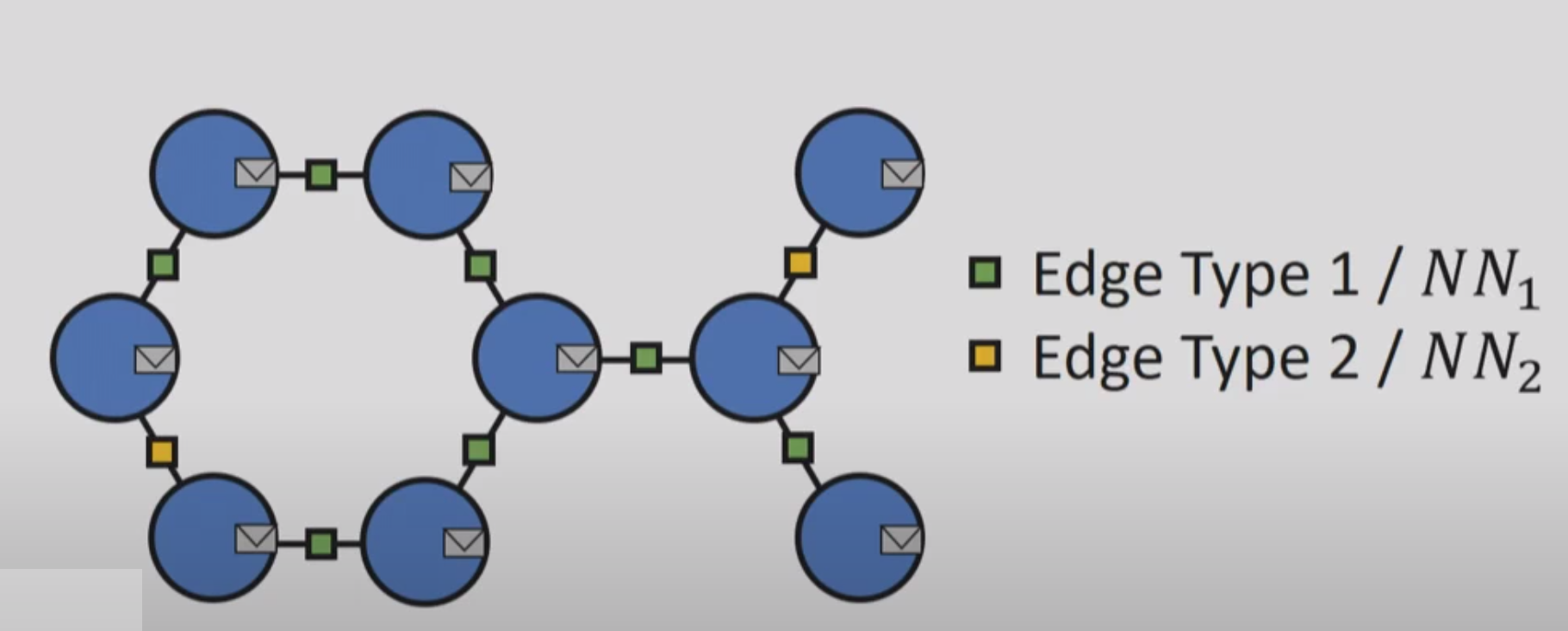
Each node feature is initialized by taking the initial embedding of the token that’s obtained as a function of the embedding of the textual representation of the token and it’s associated type. To be more specific, the embeddings of subtokens of the token are averaged and then the embedding of type of the node is concatenated with the averaged value and passed through a linear layer.

Training

The training is described as follows:

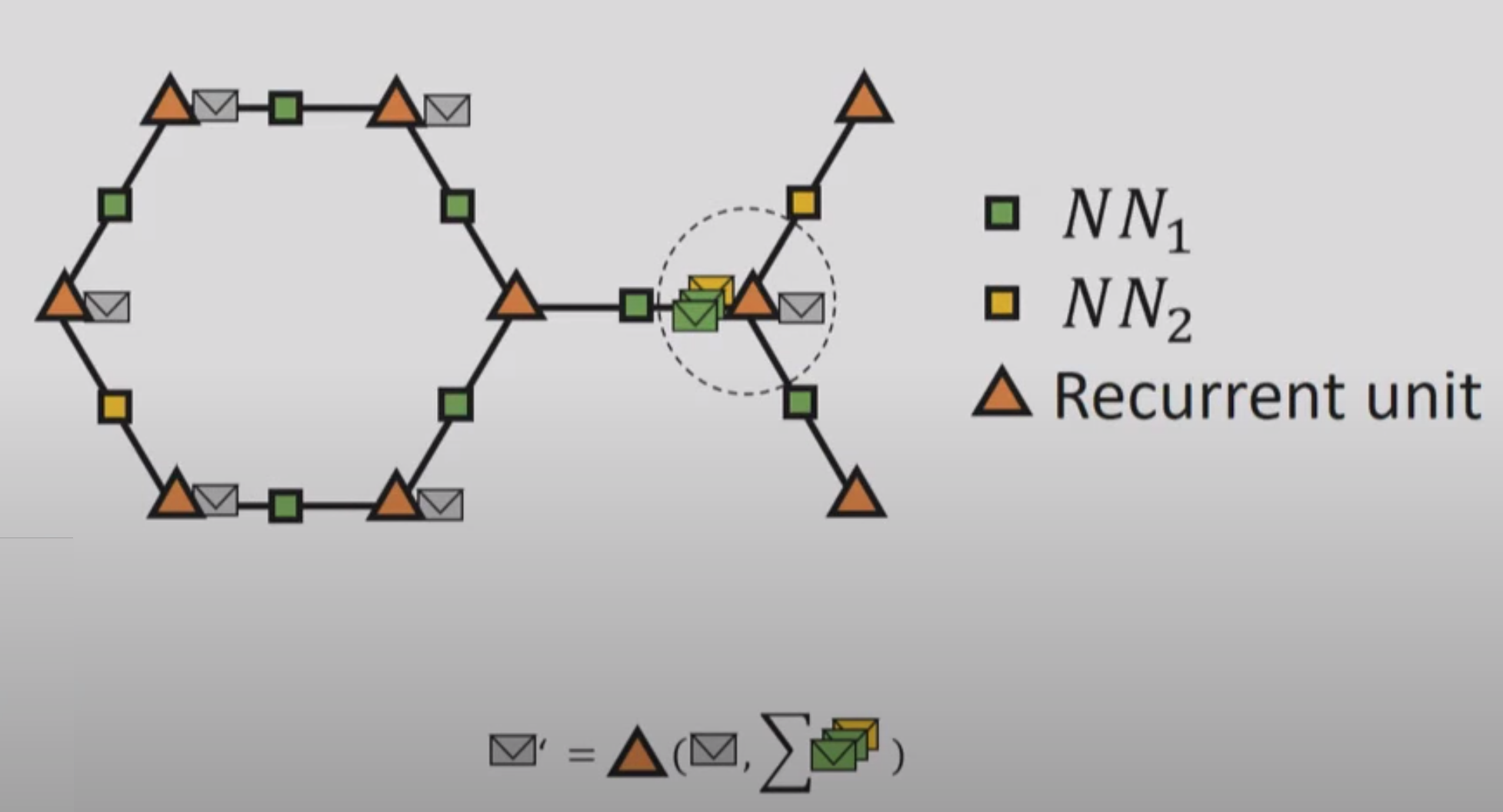
1. Each node is associated with a state i.e. its features that will be sent to each of its neighbors based on the type of edge. There is a unique neural network for each of the type of edges associated with a particular node that demarcates the learning behavior.

**Fig 7: Figure Highlighting the Messages For Each Node**



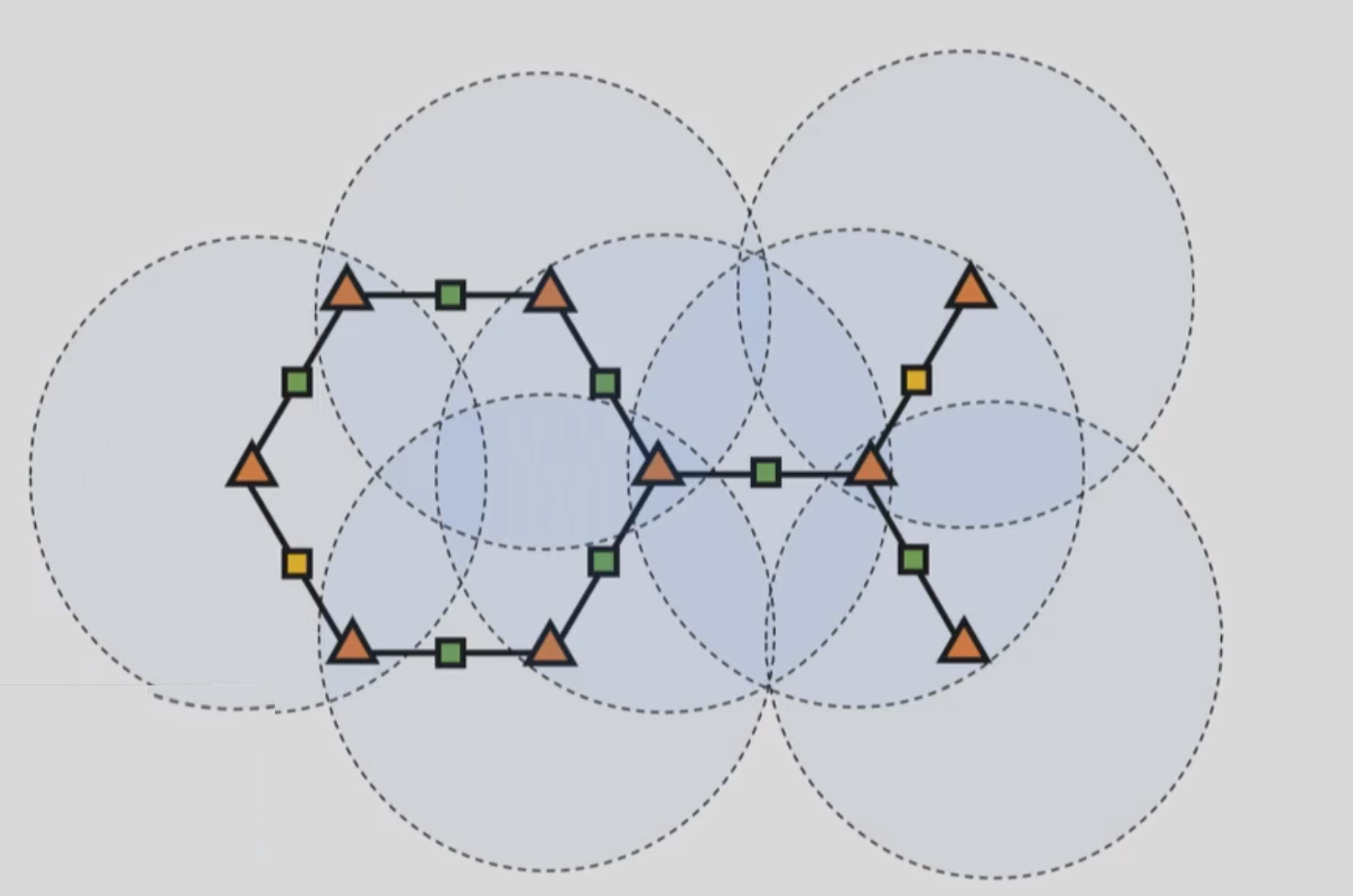
1. Once the data is sent from one node to its neighbors, it is aggregated using a summing function and is subsequently passed through an additional recurrent unit. This unit applies an additional mathematic function based on the previous state and the summed up messages from its neighbors and eventually other nodes. The intuition here is that the features of a node after each time step become are some mathematical function of its original state and states of other nodes of the graph

**Fig 8: Message Passing And Aggregation by Nodes via a Recurrent Unit**



1. The unrolling stage is when, after some number of time steps that can be a hyperparameter, the individual nodes are “aware” of not only the features of its neighbors but also other nodes in further locations on the graph and these are indicated by the large dotted circles in the figure below. By the end of a certain number of time steps, all nodes should be acquainted with their own position as well as all other positions on the graph.

**Fig 9: Unrolled Effect Of the Graph Where Nodes Become Positionally Aware**



The objective function used for the training is that of a max-margin objective on the concatenation of the state of a slot variable i.e. the state of an empty spot we are trying to predict and the state of all possible same typed variables in the same scope. In other words, using the state representations, we are trying to maximize the similarity of the state of the nodes representing a variable of the same type in the scope matches the ones that would be characteristic of the slot.

This trained graph neural network can also be fed to higher layers for more complicated representations for predictions based on more involved tasks. Specific to our comparative study, we’ll be making use of the trained graph neural networks with different units and aggregation functions to discern the more performant model and therefore, the two sections build on the concepts.

The difference in the models is based on the mathematic functions involved in generating the next time step’s state.

**Gated Graph Neural Networks**

The GGNNs take advantage of a Recurrent Cell function of the Gated Recurrent Unit as the mathematical function to compute the next state and is responsible for keeping important features as the distance between nodes increase in a large graph.

**Fig 10: Formula for the GGNN State Update Per Node**



To explain the formula, for node ‘v’, we compute the state by running the previous state of ‘v’ and the aggregated messages of all the neighbors of ‘v’ multiplied by the weights related to edge type ‘l’ through a gated recurrent unit. We apply this same formula to compute the state of all nodes to get our gated graph neural network based model to a trained state for inference.

**Relational Graph Neural Networks**

Rather than a gated recurrent unit, the relational graph neural network makes use of the convolution function using a non-linear function on the previous state of the neighbors, weights and problem specific normalization constant that serves as a hyperparameter.

**Fig 11: Formula for the RGCN Node State Update Per Node**



Like before, the next state of the node ‘v’ is computed by passing the normalized convolved aggregated states of all neighbors of ‘v’ through a non-linear unit such as a Rectified Linear Unit (ReLu) for all graph edge types ‘l’.

The intuition here is that accumulation of the transformed feature vectors of neighboring nodes through the normalized sum provide us with apt representations. Additionally, the Relational aspect of the RGCN is closely associated with the different edge types that have been covered before as per the research paper [TODO]

Now that we have expounded on the basics of the models we’ll be comparing and providing a bird eyes view of what’s happening under the hood, we proceed with getting into details about the comparative case study.

**CHAPTER 6**

**PRELIMINARY STEPS IN CASE STUDY**

**Problem Definition**

Defining the problem as meticulously as possible is an important step in Data Science and furthermore, even more important in the case of a comparison study to avoid ambiguity in the criteria of comparison.

Our problem statement is to demonstrate which model between the Gated Graph Neural Network and Relational Graph Convolutional Network based models has a higher predictive power on the basis of the test accuracy based on data from:

1. All 27 top trending C# Repositories
2. Esoteric Repository namely “CommonMark.NET”.
3. A Popular Repository namely “Dapper”.

The general goal is to identify if there is a better model out of the two i.e. which model has a higher representational capacity, one associated with a gated recurrent unit and one with a convolutional operation. Once the better model is identified, doing further research to apply the model to more sophisticated tasks by pushing the boundaries of the intersection of deep learning and static analysis.

The impetus behind choosing the GGNN and RGCN models was to prove a definitive hypothesis about the better model between two very disparate models from sequence models and convolutional models respectively. In a sense, *caeteris parbirus*, what type of mathematical construct generalizes and predicts better.

**Interview Notes**

As a part of the interview process for this project, I was lucky enough to exchange emails with one of the authors of the paper my comparative study is based off of, Marc Brockschmidt who extremely graciously answered all my very specific questions about the paper and methodology. Additionally, I spoke to Anmol Joshi, a data scientist from C-1 at Boeing to get a more general perspective about deploying models in production. The main takeaways from the conversations were:

1. There is no silver bullet. The evaluation depends on the data.
2. Keep the code simple and highly decoupled from the data.
3. Treat the code that generates the initial representations of your data with as much importance as your modeling code.
4. The code that can be used for initial representation can be made decoupled from our modeling code if the pipeline is setup correctly.
5. Experiment as much as you can. You will never get the right answer right away as there is no right answer. It's only improvement of representation.
6. Keep researching and reading white papers to learn ideas from across other subfields.

These conversations helped me gain a good foundation as to how to go through with my experimentation. The insight I gained about the methodology was extremely helpful as well.

The choice of the experiments will be highlighted in the next chapter where the source data will be described.

**CHAPTER 7**

**SOURCE DATA EXPLAINED**

**Introduction**

The input data for this project is very different from the typical data frame based exploratory data analysis approach as it involves making use of the unstructured source code from repositories.

The source code is transformed into an Abstract Syntax Tree using the Rosyln, the .NET Compiler, converting that result into a customized directed graph with the appropriate type information.

**Fig 12: Data Transformation Pipeline**

Source Code

Directed Graph

Splitting Data

AST

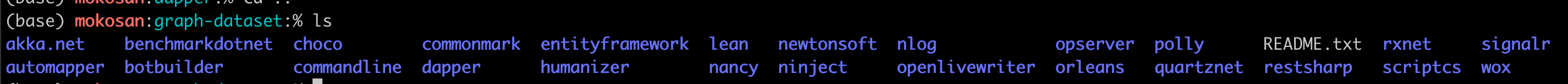
Type Hierarchy

Zip + JSON

The graphs are converted into JSON and then zipped up. The code involved with acquiring the source data from the repositories till we get all the JSON files can be found here: <https://github.com/microsoft/dpu-utils/tree/master/dotnet>.

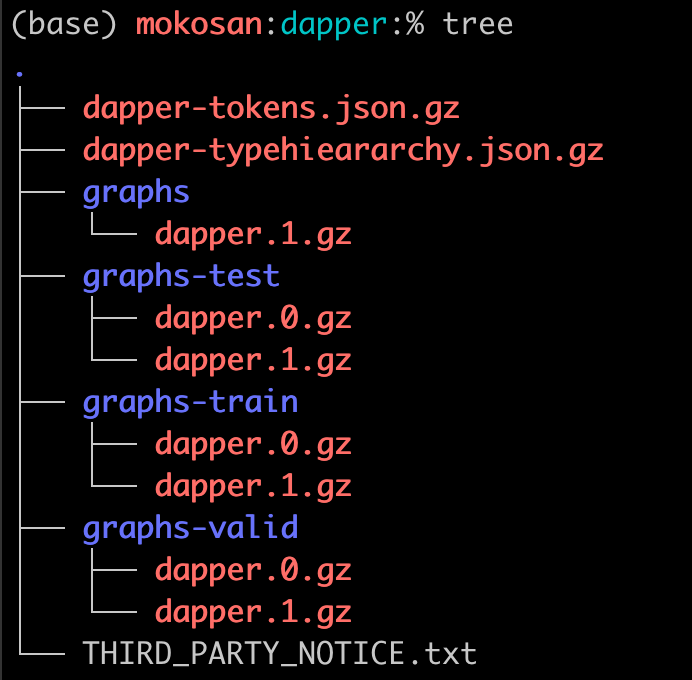
Thankfully, the transformed data for the top 27 trending C# repositories were easily available through Microsoft Research. The data obtained was a zip file consisting of 27 directories corresponding to each of the top trending repositories.

**Fig 13: Listing of all the Directories From the Graph Data Set.**



The structure of each of the directories is very similar where we have 2 zipped JSON files consisting of the token and type hierarchy information for the entire repository. Additionally, the data is split into testing, training and validation sets. The training and validation sets are used primarily for the training and hyperparameter validation and as suggested by the name, the testing files will be used to compute the test accuracy. This split is based on a percentage that’s typically set to 60-20-20 for the training, validation and testing respectively.

**Fig 14: Typical Directory Structure of Data From A Repository**



Next, I’ll be going over the contents and structure of the zipped JSON files above. The three types of files that can be found in the data folder are the tokens file, type hierarchy and the actual graph.

**The Token JSON**

The tokens files consist of metadata associated with all the tokens as well as the data edges associated with the customized directed graph. This metadata along with the type hierarchy gives enough background information to the program to construct the graph in memory to conduct the training and testing. Examples of the type of properties are included in the token files include the types of the tokens, locations in terms of row and columns where the tokens were used, other variables of the same type in scope with a particular token and so on.

**Type Hierarchy**

The type hierarchy file consists of the type hierarchy associated with types i.e. the literal type used in the source code as well as the base classes, if any, that they were derived from. Additionally, an array of the type hierarchy where the indices correspond to the types used in the source code were added.

**Structure of Typical Graph File**

The graph file consists of the context directed graph that contains all the information needed for both the training and testing. The main part of the file consists of a representation of the graph as a dictionary of the different edges, nodes and combinations of the cases where each variable we are trying to predict is replaced by a “slot” or empty spot for which we will maximizing the match on with all the other variables of the same type and in the same scope.

Studying these files in depth gave me a considerable insight as to how meticulously the input data is structured. Verbosity in the data seems to be the prevailing trend here where there is considerable amount of emphasis on is made on over specification of the details so that there is minimal transformation during training and testing. The data itself made me considerably aware of . It is a bit difficult to specify all the properties in each of these classes and for that reason

**CHAPTER 8**

**METHODOLOGY AND OUTCOMES**

Choice Of Experiments

Details About The Repositories

Conducting The Experiments

Results

**CHAPTER 9**

**CONCLUSION**

**REFERENCES**

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The graph neural network model. IEEE Transactions on Neural Networks 20(1):61–80.

[1] Abram Hindle, Earl T Barr, Zhendong Su, Mark Gabel, and Premkumar Devanbu. On the naturalness of software. In International Conference on Software Engineering (ICSE), 2012.

[2] Benjamin Bichsel, Veselin Raychev, Petar Tsankov, and Martin Vechev. Statistical deobfuscation of android applications. In Conference on Computer and Communications Security (CCS), 2016.

[3] Miltiadis Allamanis, Marc Brockschmidt, and Mahmoud Khademi. 2017. Learning to Represent Programs with Graphs. arXiv:cs.LG/1711.00740