THE UNIVERSITY OF WINNIPEG

Thesis Proposal
Machine Learning Data Pipeline

Author
Reid Lowdon

Supervisor
Dr. Christopher Henry,
P. Eng.

June 13, 2019

1 Introduction

The idea for my thesis is to create a machine learning data pipeline for the current research project I am working on, which would include the ability to "plug in" multiple different neural network architectures to the end of this pipeline.

I am currently part of a team that is working on a way to automate the creation of significantly large labeled data sets in order to improve the role of artificial intelligence in the agriculture industry.

2 Problem Statement

Agriculture is one of the most important industries in Manitoba and with the increase in both computing power and resources, we have the opportunity to increase productivity immensely with machine learning.

My part in this project is to create a data pipeline between the cameras taking images of the plants out in the field to loading those images into an artificial neural network for training. This will involve writing Python scripts to move data and creating a database to store the images and associated metadata.

I will also be responsible for initially setting up the database so I will have to be mindful of maintaining referential integrity and setting up the relations between tables.

A continual responsibility of mine will be to write SQL queries that other team members require in order for them to obtain the data that they require for their work.

Later in the project I will also be responsible for ensuring that our data sets are set up in a way that agrees with the biology community in order to be put in long term storage at the Federated Research Data Repository (FRDR).

I would also like to setup a website so that people could visit it and obtain their own data sets by entering in whatever parameters are appropriate to them, then the website would query the database and return a zipped data set with associated labels.

3 Statement of Overall Purpose

The purpose of this project is to ultimately automate the creation of significantly large labeled data sets and have these data sets be used in the training of automated agricultural equipment.

4 Project Significance

I believe that this research is critical to a better future in agriculture, for example, this project could drastically improve farming efficiency and throughput by allowing an automated robot to continually maintain and care for entire crops.

5 Timeline

This project has been underway since early 2019 and it is broken up into several sub projects:

- Subproject 1.0
  - 1.1 - Develop camera system
  - 1.2 - Generate labeled data sets
  - 1.3 - Develop plug-and-play camera system
  - 1.4 - Explore generation of multiple plant images
- 2.0 - Develop deep neural networks and deep learning platforms
- 3.0 - Sensor development and exploration of new tools
- 4.0 - Field tests with autonomous vehicles

All of these sub projects are laid out over the course of 24 months, but the parts that I am directly involved in are set to last 18 months. My hope is to graduate around the end of December 2019, or at the latest, spring of 2020.

6 Goals and Next Steps

The main goal of this project is to create an automated way of creating large labeled agricultural data sets and a secondary goal is to have this technology implemented in real world applications.

During this project I hope to personally obtain a deeper understanding of relational databases, large data sets, setting up data pipelines, and machine learning.

At the current point in the project, we are working on smaller "proof of concept" situations and I would love to see this technology implemented in real world agriculture fields.

7 References

[1] Bidinosti, Christopher *Mitacs Accelerate Proposal* 2019