Internship Report

Intern Details:

Student Name	**SATISH G**
Roll Number	***CS14B042***
Company of Internship	***QUANTIPHI INC***
Internship start date	***08/5/2017***
Internship duration	***9 weeks (48 Working days)***
Location	***Mumbai***
Designation	***DECISION SCIENCE INTERN***
Internship Mentor	***Vishal Vaddina***
Project Name	***Caffe & Deep Compression***

Project Overview:

Name of the project: *CAFFE & DEEP COMPRESSION*

Languages used: PYTHON, C++

Frameworks used: Caffe

Environments used: SPYDER

Services used: GOOGLE CLOUD PLATFORM

ML Libraries used:

- TensorFlow
- And several other python libraries

PROJECT: Caffe

PURPOSE:

Caffe is a deep learning framework which is written in C++, with a Python interface. With the increase in popularity of this framework in recent days, I was required to research&explore Caffe and document the experience, issues, comparisons etc.

Project Details:

- Introduction to Caffe and getting familiar with the basic syntax.
- Understood CNN-Basics and explored per-trained models

- Retrained an existing model(CaffeNet) to perform a classification on ImageNet(cats, dogs) dataset and achieved an accuracy of **94.23%**[*]
- Built a basic CNN model(BasicNet) from scratch to perform a classification on ImageNet(cats, dogs) dataset and achieved an accuracy of **88.9%**[*]

*limited the models to 10k iterations for comparison

PROJECT: Deep Compression

PURPOSE:

Deep learning models generally include deep networks, having a large number of neurons and connections and hence, generally very large in size. This makes these models very difficult to be deployed in small devices like mobiles, smart watches etc. Hence, Deep-Compression comes into play. With recent developments, this process with certain compressing tasks in pipeline results in a reduction of the model size by about 35-50 times, making it very easy for deploying in small devices and faster computations. Follows a pipeline procedure involving Pruning, Quantization and Huffman Coding.

Project Details:

- Deep Compression – Introduction, familiarising with concept.
- Implemented pruning and Quantization on a tensorflow model

Model	**Parameters**	**Parameters Compression Ratio**	**Pb Size**	**Pb Size (gzipped)**	**Pb Compression**	**Acc***
Uncompressed	~60k		247 kb	229 kb	~18%	98.6
Uncompressed + Quantization	~60K		71 kb	55 kb	~80%	98.6
Compressed	~7.5K	~90%	247 kb	56 kb	~80%	98.7
Compressed + Quantization	~7.5K	~90%	71 kb	19 kb	~93%	98.7

[*]Accuracy(%)

- The protobuf file(file used for deployment) size is reduced by 80% after implementing the pruning and by 93% after quantizing the pruned model

Model	Download Size	In-app size (After extraction)(kb)	Compression
Uncompressed+Quantization	55 kb	71	
Compressed+Quantization	19 kb	70	65%

- The download bandwidth has been reduced for the compressed quantized model by 65% compared to the uncompressed quantized model and the size after extraction is same in both the models.

Core Skills used:

- Objective Oriented Programming in C++ learnt in Paradigms of Programing(CS3100) has helped in understanding the structure of caffe framework.
- The knowledge yielded in Data Structures and Algorithms(CS2800) paved an easy way in coding the Deep Learning Algorithms efficiently.
- All the algorithms are coded in Python (CS2810-Advanced Programming Lab)
- Coded following the guidelines(modularity, agile, commenting) learnt in Principles of Software Engineering(CS3400)

Learning to the student:

Technical:

- Learnt Caffe, which is a pioneering Deep Learning framework
- Worked in TensorFlow, a widely used machine learning library.
- Experienced the working in various machine learning python libraries(scikit,keras)
- Learnt the usage of virtual machines in Google Cloud Platform
- Learnt to host demos in open-source web applications such as Google Datalab, Jupyter Notebook
- Learnt the importance of Agile Software Development.

- Learnt good debugging practices using breakpoints appropriately

Personal Learnings:

- Learnt good **coding practices**
- Understood the importance of **modular code** and **proper commenting**, so that the next coder who takes up the project to continue will find it easier.
- It taught me that there are libraries for many algorithms and you just need to learn to apply them properly. This means that we don't always have to start from ground zero.
- Helped me in realising the importance of having good design architecture before coding.
- I learnt that debugging in appropriate way eases the work of developer.
- I have seen the applications in real case scenarios of the concepts learnt in our courses in college.
- I have observed the difference in the coding style in college and workplace.
- This internship helped me understand more deeply about the various concepts of working in INDUSTRY.
- It taught me things like Professionalism, Professional ethics, working in a team, interacting with people, the importance of communicating your thoughts clearly.
- I have had an exposure to the company's policies on data privacy, security, ethical practices and open source software usage.
- I learnt to reuse the code as much as possible by writing functions

---

Non-revelation of Confidential Information

☐The above report does not contain any confidential information from Quantiphi.

Vishal Vaddina	**Vivek Khemani**
(Project Mentor)	**(Co-founder of Quantiphi)**