Faculty of Natural and Mathematical Sciences
Department of Informatics

The Strand
Strand Campus
London WC2R 2LS
Telephone 020 7848 2145
Fax 020 7848 2851

7CCSMDPJ

Individual Project Submission

2016/17

Name:	Inzamamul Haque
Student Number:	1664749
Degree Programme:	MSc Data Science
Project Title:	Using Google's TensorFlow to Classify Stem Cells
Supervisor:	Dr. Nishanth Sastry, Dr. Amos Folarin, Dr. Davide Danovi
Word Count:	0

RELEASE OF PRODUCT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

? I **agree** to the release of my project

? I **do not** agree to the release of my project

Signature: **Date:** August 17, 2017

Abstract

This thesis looks at utilising Google TensorFlow's machine learning library to try and classify stem cells. Images were recorded hourly over 24 hours of cell wells containing stem cells from anonomous donors that had been exposed to various concentrations of protein. The raw images of the cell wells were processed and masked in order to segment the individual stem cells from the raw images. Stem cells that were segmented from the first raw image captured in the 24 hour period were labelled as 'normal', i.e. cells which had not yet been affected by the protein concentrations.

These individual stem cells were then trained upon by a convolutional neural network, in order to be able to predict and classify whether a new unseen stem cells was 'normal' or 'abnormal', or in other words, whether or not a stem cell had been affected by the protein concentrations. Using the count of 'normal' and 'abnormal' stem cells in a specific hour of the experiment, we are able to identify interesting metrics to assign to the activity levels of the cells exposed to varying levels of protein.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors Dr. Davide Danovi, Dr. Amos Folarin and of course Dr. Nishanth Nastry, for their incredible support and patience that was conveyed throughout this project process. I would also like to thank my close friends and family for supporting me throughout this project and my life thus far, and I hope for their continued warm support going forward.

1 Introduction

The Introduction is the first content section of your report. You should describe the general area (e.g., application domain) in which your project research is conducted, the motivation for conducting the research and the overall aims of the research. Be sure to outline your research questions and give a brief summary of the conclusions drawn, though the conclusions will be detailed later in the report. With the Introduction, you want to interest your reader and tell them why they should care about your research and why they should read the rest of the report. The report will be read (marked) by examiners with a technical Computer Science background, but not necessarily any knowledge of your domain, so make sure that you provide enough information for a naive reader. [?] [?] [?] [?] [?] [?] [?] [?]

MARKING FOR INTRODUCTION

- A clear context for the work is provided.
- The problem to be addressed is clearly defined.
- The work is well motivated.
- The relevance of the work, with respect to Data Science, is established.

My project is titled ?Using Google's TensorFlow to train the Inception Deep Learning Convolutional Neural Network in order to classify images of cells from drug and stem cell differentiation screens?. As part of Google?s efforts to grow its presence in the deep learning space, the company built DistBelief in 2011 as a proprietary machine learning system. The DistBelief system picked up traction among computer scientists wishing to develop and build larger neural networks to answer large data problems. In 2015, Google released TensorFlow, a successor to the DistBelief system, [?] which was made to be more flexible, portable.

In 2015, TensorFlow was built upon the DistBelief system and has the added ability to

be able to compute any computation that can be expressed by a computation flow graph. The system performs computations on multidimensional arrays, tensors, which are then passed on through neural networks. The system also improves upon DistBelief?s speeds, and scalability. TensorFlow has been made open source from and has been forked now over 25,000 times on the version control repository, GitHub. A Python API is available for developers, scientists and researchers who would like to develop on the project on virtually any domain they see fit.

Introduction of the Domain and Research Motivation

The domain is centred around the work being done at the Human Induced Pluripotent Stem Cells Initiative (HipSci) at King?s College London. The group at HipSci are developing assays to image cells in artificial microenvironments to study its responses to various stimuli in order to develop disease models. The HipSci group has state of the art equipment that is able to take pictures of a variety of stem cells for use in training these disease models, and brings together researchers and scientists working in the field of genomics, proteomics and cell biology. This topic is very interesting to me as the developments in machine learning are slowly coming to present itself in this domain to assist in the creation of more powerful and precise disease models.

Work done in this domain could see the improvements in the way that we diagnose or treat diseases that affect millions of us daily, could see quicker turnaround times that drugs are developed and produced for public health and could help us understand more about how stem cells (specifically induced pluripotent stem cells) behave over time. This is also of great relevance to my domain supervisors, Dr. Amos Folarin and Dr. Davide Danovi, Director of HipSci at King?s College London, and the work done using TensorFlow is of interest to my supervisor, Dr. Nishanth Sastry. Another attraction for this topic is the use of deep learning to provide the analytics and results.

Deep learning has only recently become a technology trend that more and more domains are turning to, for providing solutions to ever growing complicated questions. I do hope that my experience with this project, especially after using deep learning techniques will allow me to use my skills in other domains as well as the work I will do in the bio-informatics sphere.

2 Background

BACKGROUND MARKING CRITERIA::

- The discussion of relevant related literature is structured and coherent.
- Key technical issues are summarised.

The Background section of your report should provide the reader with enough technical background so that they understand the area in which your research is conducted. This should be the kind of information that you might find in a textbook that teaches someone about the area. The next section of the report ("Related Work") is where you describe new research in your area, so think of this Background section as where you provide enough information so that the reader will be able to understand the important details contained in the Related Work section.

For the project, "Using Google's TensorFlow to train the Inception Deep Learning Convolutional Neural Network" in order to classify images of cells from drug and stem cell differentiation screens.?, the data to be used for the research will be images of well plates (using 96 welled plates, of which there are 8 rows and 12 columns) of stem cells that have been collected at the King?s Centre for Stem Cells and Regenerative Medicine. On each image of the plates are stem cells that have been exposed with various levels of Fibronectin (FN) protein of 1 microgram/ml, 5 microgram/ml and 25 microgram/ml.

Each row of the well plate has a different cell line, and each row contains three wells of the same cell line exposed to the above concentrations of FN. The data is comprised on 50k+ image of 101 cell lines which were imaged once per hour for 24 hours. The lines were imaged multiple times in multiple experiments ? with each cell line being imaged first using a live-imaging device, where the snapshots were taking every 24 hours. At the end of the 24 hours, the same cell line was then stained and imaged using the Operetta High Content Imaging System.

The Operetta system computes summary plate files, where the cell line from the plate wells and the FN concentration are recorded. The metadata from the saved images then get annotated with the metadata provided from the operetta plate files. They may exist possible issues of batch effects due to errors in plating, and potential variation in the concentration used during the cell assays. The cell line behavior may also differ slightly across multiple experiments, even though they are the exact same cell line. The data was obtained from the following individuals (Domain supervisors) who kindly provided the data privately.

Dr. Davide Danovi Director,
HipSci Cell Phenotyping Centre for Stem Cells and Regenerative Medicine
King's College London,
28th floor, Tower Wing,
Guy's Hospital, Great Maze Pond,
London
SE1 9RT, UK
davide.danovi@kcl.ac.uk

Dr. Amos Folarin
Informatics Software Development Group Leader
SLaM/Kings College London
amos.folarin@kcl.ac.uk

Maximilian Kerz, PhD Candidate
Lead Developer of RADAR-CNS
Front-end Ecosystem
Dept. Biostatistics and Health Informatics
King?s College London

Tel.: +44 (0) 207 848 0924

Example of glossary entry is SVM. CNN. Example of bibliography entry is given by Johnstone [**?**]. Further information can be found at: [**?**].

3 Related Work

MARKING CRITIERA FOR LITERATURE ::

- Literature sources primarily come from published, peer-reviewed venues.
- Multiple literature sources, from different authors/labs, have been critically analysed, compared and contrasted.
- The relevant literature has been comprehensively covered, both in terms of the Data Science techniques and technologies, as well as the particular domain from which the student's data set(s) and research question(s) originate.

The Related Work section of your report should provide a review of recent literature in the area of your research. This is distinguished from the Background section because it is typically newer and more experimental. If there are standard terms or techniques mentioned in the literature, then you can define what these are in the Background and use the Related Work section to explain how researchers have used the standard techniques as benchmarks or fundamental methodologies for their research. For example, if you review an article that describes using k-means clustering for finding appropriate groups of patients with similar sets of symptoms, then you could describe what k-means clustering is in your Background section and describe how researchers used that technique on patient data in your Related Work section. When you review literature, be sure to explain how the articles you cite are relevant to your project. Be critical—outline pros and cons of the work you are reviewing. Be clear to explain how the work you review is different from your own work. Note that you may find it easier to compare and contrast others' techniques with yours later in the report, after you have explained your own work. That is fine—just be sure to forward reference in the Related Work where you will compare to your own work (and backward reference in the later sections back to the Related Work). This can include information that you had in your Project Proposal report that was due in April, but should typically be substantially expanded from what you had in your proposal.

Literature Review An Overview of Data Science Uses in Bio-image Informatics

As processing power and efficiency has increased over the recent years, it has become more feasible to conduct large scale analytics on data across various domains. Bioinformatics is one of the domains that has benefited from the statistical methods that can be applied to large sets of bio image data in order to provide both quantitative and qualitative relationships between biological concepts and the data presented. As the usage of deep learning has been increasing in recent times, its applications are starting to appear in bioinformatics. An application of deep learning includes using convolutional neural networks (CNNs) to compare classical features found in bio-images and to then predict the behaviour or actions of cells or small molecules (Chessel, 2017).

Utilising deep learning methods in bioinformatics will continue to expand as bio imaging provides for higher resolution images and biological research questions become more detailed. This is an area of research that I anticipate to explore more of in order to develop applications for the above domains.

TensorFlow: Biology's Gateway to Deep Learning?

Following on from previous discussions involving the rapid rise of deep learning across various domains, especially in bio informatics. The introduction of TensorFlow to the deep learning community has provided much needed toolset that can better assist in pushing research further. TensorFlow?s ability to provide graphical visualisations as well as quicker speeds for learning (Rampasek and Goldenberg, 2017). Currently the TensorFlow environment is still a low level system, requiring its users and developer to be comfortable working with its Python API.

There are however low level wrappers available that could make it easier for newcomers

to begin experimenting with the platform and test it with already acquired computational biological data. Along with its support for parallelization over multiple machines running on either CPUs or GPUs, the bioinformatics community believes that TensorFlow can provide a powerful path to mix deep learning with bioinformatics.

Machine Learning Predicts the Look of Stem Cells

Induced pluripotent stem cells, or iPS cells, are cells that were previously normal cells but have been reprogrammed to behave and appear similar to embryonic stem cells. These cells now have the capability to transform into any type of working cell type used in the body (Scudellari, 2016). Upon its discovery by Shinya Yamanaka at Kyoto University, Japan in 2006, much work has been done in order to try and use these stem cells to cure or treat cell abnormities and diseases found in humans today, among other uses. As these cells provide the ability to be used as a form of regenerative medicine, due to its properties to change into a variety of different cells, it has garnered a lot of interest in the regenerative medical community.

Machine learning techniques are starting to make a presence in the modelling and prediction of the behaviour of iPS cells. At the Allen Institute in Seattle, USA a team have been training deep learning algorithms to (Maxmen, 2017) predict what the structure of a cell would look like with minimal data, such as just the location of the nucleus. They were able to do this by identifying relationships between the locations of the cell?s structures of the 6,000 or so iPS cell images that they have in their library.

A Novel Automated High-Content Analysis Workflow Capturing Cell Population Dynamics from Induced Pluripotent Stem Cell Live Imaging Data

One of the main obstacles of performing machine learning techniques on iPS cells is the initial segmentation of the cell from other cells that are found on the cell plates.

Using phase-contrast and confocal microscopy to capture bio images of these cells require that many cells be placed on a shared well plate. Images are then taken of the well to capture the various conditions and structures that the cells may take form of. By introducing time as another dimension, i.e. to track the cell structure evolution over time as well as still capturing a high enough resolution image, this may be tricky to complete.

A solution proposed is to use a widely used application in the bio informatics community, CellProfiler, to enable identification of iPS cells with highly dynamic structures in single-channel, phase contrast images. The cell segmentation can be challenging at times using conventional methods because of the sometimes poor edge contrasts between the background and foreground, as well as the non-standard structures that the iPS cells can take (Kerz et al., 2016). CellProfiler, built by the Broad Institute at Cambridge, Massachusetts, provides for the ability to conduct the image segmentation and then outputting a pipeline that can be alter used for image analysis applications, such as for input into machine learning methods.

Examples of articles we might cite are [**?**] and [**?**].

4 Approach

MARKING CRITERIA FOR APPROACH::

- Sufficient research question(s) is/are presented.
- Approaches to answering the research questions have been described adequately and alternative approaches are described and considered.
- A rationale for the chosen approach to addressing the research question(s) is given.
- A rationale for the data set(s) chosen is given.
- Selected issues derived from a synthesis of the relevant literature are included (which may include rationale for the data set(s) and research question(s) chosen).
- Implementation choices have been justified.
- Careful planning and understanding of methodology are demonstrated.
- Methodology is sound and has been clearly presented.
- Adequate and effective testing of methodology has been performed (e.g., verified with sample data/results).

The Approach section of your report should describe what you did. You should discuss your research questions in detail here, explaining for each question how you addressed each question (i.e., what techniques you used) and how you evaluated the success (or failure) of your investigation. This should include a description of the data set(s) that you used for your research (e.g., what you included in your Data Acquisition report that was due in March).

Research Question 1: Is it Possible to Perform Image Segmentation Using the Data Set?
The data does not to be cleaned as such, but the images of the cell lines will need to

undergo image segmentation in order to extract the individual cell objects so that they can be fed into a neural net pipeline in order to achieve the goals stated in the project description. This can be done a number of ways. Currently one method is to build upon existing open source software that is already built for cell profiling and image analysis (CellProfiler) as demonstrated by Kerz et al., and to include the functionality that allows the user to save individual cell pipelines for feeding into neural networks. Another alternative is to use a U-Net convolutional network for the image segmentation (Ronneberger, et al., 2015) that has seen success in other biomedical image segmentation challenges. The plan is to have the images segmented and ready to fed into a neural net pipeline by the end of April 2017.

I used MatLab to segment the images. The process is outlined below.

1. The images have been taken every hour for 24 hours using the incucyte imaging device. I will use the cells contained in the first image of the 24 images for training on the model. This will give me the cells that have had the protein added to them.

2. Read in the image using MatLab. Identify thresholds of the image.

3. Identify a mask of the image to get get a binary image from the gray scale image. This then locates the individual cells, or blobs, in the raw image.

4. Save all the cells or blobs into a directory, to be used for training further in the downstream process.

Research Question 2: Is it possible to train a model using TensorFlow to perform anomaly detection on the images, training with normal cells from the iPS image data set?

For this research question I propose to implement a neural network method for identifying cells that have deviated from the normal trained cells and to then classify as these as abnormalities. The number of cells that will have deviated from normal will increase over time as the cell evolves. It would be useful to keep a scorecard of the ratio of ?normal? to ?abnormal? cells at any time to provide statistics or metrics to identify a rate of change from ?normal? to ?abnormal? that could be useful in learning more about how iPS cells differ from one another. These differences may be due to a variation of genes

in the cells or due to some other reason which would be of interest to find out for the purposes of this research.

5 Results

MARKING CRITERIA RESULTS::

- Results have been clearly presented (visually), analysed and discussed.
- Evaluation and Analysis procedures have been explained clearly.
- Results have been compared and contrasted.
- An assessment of the results with respect to the original research question(s) has been performed.
- Potential extensions (i.e., future work) are highlighted.
- Limitations of the work are clearly enumerated and discussed.
- Strengths and limitations of the research are identified and discussed, including Lessons Learned.
- Reflective and justified conclusions have been drawn.
- The research contribution (i.e., approach to addressing research questions) is sophisticated or non-trivial.

The Results section of your report basically contains the answers to your research questions. This section should present the results of your evaluation, provided quantitatively, qualitatively and/or visually, as appropriate, followed by an analysis of the results. Discuss with your project supervisor(s) and/or domain advisor(s) how best to present your results. The main point is to make sure that it is clear to the reader what the answers to your research questions are and how you arrived at these answers.

6 Conclusion

I'm fucked lol.

The Conclusion is the last section of your report (other than Appendixes). In this section, you can revisit the research questions and summarise your answers. Clearly explain how your investigation and your answers are a contribution—why your work is worthy of a passing mark. Also in the Conclusion section, it is good to have subsections that highlight (a) Future Work, in case you were going to keep working on the same line of research or you wanted to recommend follow-up investigation for another student to pursue next year; and (b) Lessons Learned, where you can explain how you might do things differently if you started over, because you've learned valuable things along the way (these could be technical, but they could also be personal, such as organising your time better or listening to the project coordinator who told you to BACK UP your work frequently).

FURTHER RESEARCH QUESTIONS THAT CAN BE DONE

Research Question 3: Is it possible to retrain the model for other image data sets specifically including images captured from drug screen tests?
Research Question 4: Is it possible to retrain the model for other image data sets, specifically including images captured from other cell lines collected from HIPSCI?
For these two research questions, providing that the model works for the original datasets mentioned in the 2nd research question, it would be interesting to see if we are able to identify similar rate of change metrics for iPS cells collected from different environments, or iPS cells that were intended for different research purposes. If the model does work for other iPS cells then it would assist in the research that is currently being undertaken in the drug screening community, as well as the King?s HIPSCI community.

Research Question 5: Is it possible to train the model to perform unsupervised class discovery from the iPS cell images?
For this research question, this builds on improving the existing model, provided that it works with a range of different iPS cells. It would involve building a feature extraction model on top of, or alongside the existing model that would identify other classes from the population of individual stem cell images. This could even be a clustering problem, I propose that after separating the normal and abnormal cells, to then go back to the normal cells and identify if there are subgroups of cells that exist that can be classified into its own class.

Research Question 6: Is the model able to work on more complicated stem cell images such as embryo cells or cells collected from PhaseFocus equipment?
This is also a follow up from the previous research questions, to see if the model works with other types of iPS cells that are may not have the same classes as the ones trained on previously, but would it be possible to implement the feature extraction or clustering problem from the previous question on the new dataset to try and identify a rate of change from one ?normal? or time=0 class and any of its evolutionary stages. This would require retraining the original model with the new classes obtained

References

[1] F. Inc., "Phage lambda: description & restriction map," November 2008.

[2] J. Doe, *The Title*. PhD thesis, University of Mars, 2011.

[3] I. M. Johnstone, *Gaussian estimation: Sequence and multiresolution models*. 2011.

[4] I. Johnstone and B. Silverman, "Ebayesthresh: R programs for empirical bayes thresholding," *Journal of Statistical Software*, vol. 12, no. 8, pp. 1–38, 2005.

[5] I. M. Johnstone, *Gaussian estimation: Sequence and multiresolution models*. kfc, 2011.

[6] M. M. Johnstone, *Gaussian*. kfc, 2015.

[7] wallah wallah., *wallah wallah*. waaaalllaahh, 1993.

A Appendix: Review of Data Mining Techniques

A.1 Classification

A.2 Clustering