**MSc. in Computing**

**Practicum Approval Form**

# Section 1: Student Details

|  |  |
| --- | --- |
| Project Title: | A method for recording provenance in the development status of CNN Model Training |
| Student ID: | 19214652 |
| Student name: | Brendan Bonner |
| Student email | brendan.bonner2@mail.dcu.ie |
| Chosen major: | Artificial Intelligence |
| Supervisor | Dr. Alessandra Mileo |
| Date of Submission | 05/12/2020 |

# Section 2: About your Practicum

Please answer all questions below. Please pay special attention to the word counts in all cases.

**What is the topic of your proposed practicum? (100 words)**

The topic covers the recording of CNN layer metrics during the training sequence. The hypothesis is to examine trust mechanisms for explaining AI, and providing an immutable record of the training sequences prior to a model being utilised or customised.

The practicum focuses on two approaches to this; recording and visualisation of the sequence of changes during the training of a model. During model fitting at configurable intervals, layer characteristics are measured and recorded, alongside mathematical verification, and stored. This can be visualised, supporting modern explainable AI measures, to verify model behaviour according to training, and can be *trusted*.  
  
**Please provide details of the papers you have read on this topic (details of 5 papers** expected).

1. Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

This paper gives an overview of visualising the internals of convolutional neural networks, with the approach of abstracting the CNN with saliency maps, post training.

2. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. and Nagar, S., 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

The topic of this paper is to showcase and provide the basis for the IBM developed toolkit for measuring the fairness metrics or various artificial intelligence frameworks, and introducing tweaks to the network to remove bias.

3. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T. and Torralba, A., 2018. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

This paper is an advancement on the CNN dissection paper, identifying the process of explaining the parts of RNN and CNN when used to generate data based on a trained model. The paper includes methods of abstracting the important components of layers when generating certain outputs.

4. Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. In *Advances in neural information processing systems* (pp. 4066-4076).

The background to the practicum proposal is to investigate when certain bias traits are perceived as fair or not in the case of neural networks. Building upon this paper, we would try to identify elements of training that are not currently investigated to counter the case that black box AI systems cannot be trusted.

5. Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), pp.2579-2605.

This paper builds on techniques far variations of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) for graphically representing high dimension arrays, calculating the selection of neighbours based on gaussian distances of elements.

**How does your proposal relate to existing work on this topic described in these papers?** (200 words)

The proposal builds upon the attempts to provide additional metrics to support the establishment of the concept of fairness in AI. In comparison with human trust, which is established through verification and measurement of academic and professional credentials, Fairness in AI is an addendum to the model being examined, with attempts to quantify via measures such as Thiel scores for certain bias tests. This is further developed by IBMs AI360 Fairness toolkit paper which show how this is measured and corrected, but not being able to identify where that behaviour was acquired and why a model is trained in a certain way to develop these outcomes.

There is currently a gap in identifying if there is a *provenance of trust* when using deployed AI systems, and if sibling and inherited models retain these behaviours and if this can be visualised. While attempts to visualise the internals of decisions in the Simonyon and Bau papers, these focus on trying to visualise the pathways being activated in a 2 dimensional map. The dissection and visualisation papers provide a key how to measure and record this aspect of models, and our research will examine if there is a potential to securely record neural network snapshots.

What are the research questions that you will attempt to answer? (200 words)

The question being asked is to determine if we develop a novel mechanism and metric for representing the evolution of knowledge within a neural network during the training and development stages of creating a model. By analysing the state of the art in terms of AI fairness and methods of visualising the dissected layers common networks, the research will establish if creating an immutable chain of layer delta changes over time will provide an insight into the trustworthiness of the underlying model.

The second part of the research will ask if the information being stored can be visualised based on the elements of explainable AI to be able to identify which iteration and epoch of the training process is responsible for the manifestation of the behaviour. Ultimately, the outcome of the research should ask if providing verifiable additional data on how an artificial intelligence system evolves over time, if this can provide a foundation into further research towards comparisons of human trust and artificial intelligence trust.

How will you explore these questions? (Please address the following points. Note that three or four sentences on each will suffice.)   
  
- What software and programming environment will you use?

The majority of the research will be based on the AI360 models for validation. The core of this is based on a high performance AI platform utilising Tensorflow and python access to it’s libraries.  
- What coding/development will you do?

The research will develop a framework in either C or Python to extend the Tensorflow fit() command to create the method for calculating and recording the current layer status. It is envisaged that an AlexNet module will be used to provide a balance between layer size and performance. Modules will be developed in the demonstrators to write the outcome to either a cloud based data store or local database. The visualisation activity will be developed in javascript front and backend using node.js for discovering data and d3.js for presentation of information.

- What data will be used for your investigations?

Initially, the proof of concept will use the publicly available German Credit Scoring *Statlog*, to train the model and examine how the system will operate, as well as the value of the metrics. Once this model is tested, we will apply the data on a large UTK facial image based dataset.

- Is this data currently available, it not, where will it come from?

Both datasets are already available, and are used for testing bias and performance of CNNs.

- What experiments do you expect to run?   
I expect to initially train the German Credit Scoring using a Random Forest to identify localised changes in smaller decision trees, and test which methodologies work best for explaining the difference over training iterations. Once this is completed, we will run multiple training sessions; again with either a modified fit() function, or temporal snapshot function; to generate a large dataset identifying the structure of the training. Finally, we will experiment with visualising this information in a method that creates a chain between a blank slate brain and a trained one.

- What output do you expect to gather?

I expect the output to be a multidimensional array covering time, iterations, layers and additional changeable data that is pushed onto an immutable shared data store that can be accessed by the subsequently created systems. This, in turn, will generate a series of visualisation outputs that will be used to test the usability hypothesis.

- How will the results be evaluated?

The results will be evaluated by using the outcome of fairness comparison solutions against the above output. If it is possible to obtain new knowledge as to when a neural network obtained specific traits, then the evaluation of the outcome will be if this can be used as evidence of increased or decreased trust in a system.