**Literature Review Outline**

1. Introduction
   1. The practicum investigates the need and demand for evaluating the development of a methodology for efficiently recording the learning process of an AI model, focusing on convolutional neural networks. The vision will establish if existing CNN explainability methods can be used to define a novel temporal chain of major differences in the underlying structure of the network during training.

Once we have determined what elements of value in the process show what can be recorded, it will create a verification repository for the network (similar to a CV) and show how this can be used as either a *trust* chain when for establishing the provenance of the model, reviewing the network training in comparison with the post-network visualisation, or providing a mechanism to see if the training process is either optimal or biased.

* 1. The research area overlaps between the current drive to explain black box models to address trust issues with artificial intelligence, with the drive to be able to optimise and evaluate the process for creating and expanding existing models. The bulk of current research focuses on the behaviour of neural networks after training, examining how an input triggers various layers, and this has reduced as deeper networks become more commonplace. Starting to record and validate how networks are trained, rather than used may provide additional support to trustworthy AI, while benefiting training evaluations and visualisation of when neurons are activated during the learning process.

1. Theme A - What is the state of the art in relation to measuring the layers and understanding convolutional neural networks for explainability.
   1. Overview of characteristics of the theme (commonalities, differences, nuances)

In the process of understanding and trusting deep neural networks, researchers have been towards extensive evaluation to identify pathway activation, or by presenting input triggers and looking at activations between layers. There is a difference in approach regarding what has generated use on explainability, allowing multiple approaches to cover the effect of individual features, morphology characteristics or how pixels resonate various components of the network. The focus is to understand the network, rather than understand how it is trained. The closest research in understanding the training process is the use of classifier probes, and how these can be recorded in a provenance chain.

* 1. Examination
     1. The first step in our journey to evaluating the trustworthiness of AI systems is examining how fairness and bias is interpreted across all artificial intelligence systems. In *AI fairness: how to measure and reduce unwanted bias in machine learning*, Mahoney, et al, examine each phase of the AI deployment process to identify where bias is created, identified and countered. The resultant research created the IBM AI Fairness (AIF360) toolkit, which looks at all network types and rewires the network weights and biases to remove predefined bias from the outcome. This works well on business AI systems classifiers like *random forest* and *naive bayes,* and simple neural networks, but is highly limited in deep neural networks. The research does identify one important gap - there is no mechanism to alter a trained network, and expect it to behave as before. Part of the research will provide a prevalence trace to verify that the network being examined is mathematically verified to not be tampered with, potentially increasing the trust in deployments.
     2. While the above paper looked at basic networks, the journey is more difficult for convolutional neural networks due to the increasing depth of networks. A popular paper on understanding the decision process within ConvNets was discussed in the 2014 paper *Deep Inside Convolutional Networks* (*2014, Simonyan et al*) where the innards of a CNN when presented with an image to be classified was visualised using saliency maps. The method of saliency extraction with a view to presenting a single 2D image map gave a “last layer” backwards view of the network activations, as if you were looking through a telescope, while imagining the neurons. This research highlighted that aggregation of individual neuron activations can be utilised for explaining CNNs. What we would like to see is how to twist this view around, and instead of seeing the activations of neurons, instead see the neurons being trained.
     3. The final piece in the jigsaw is attempting to use this understanding to make a score of interpretability and the experiments of breaking down the components of CNNs are discussed in *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. In this 2017 paper, *Bau et al*, introduces the concept of scoring unit interpretability to give a segmentation threshold to an individual concept when the CNN is classifying. The attempt is to visualise and score diverse CNNs to evaluate discrimination and interpretability. The outcome of this research will be to examine if there is a link between the number of detectors of concepts in the outcome tests to the number of times a certain portion of the network is trained to classify or ignore that concept.

1. Theme B - Interpretation of the Learning Process
   1. Overview of the Theme

Once we understand that there is a gap in being able to visualise the learning process when training each convolutional neural pathway in an AI system, we then have to examine the state of the art in being to interpret this training process. In this part, research is quite light, as the focus has been on understanding a network when it is making a classification or using evaluation to interpret differences in classification. The approach that I will take is to examine how individual elements of each layer are tweaked during the training process. While a brute force approach would be to record the loss function changes in back-propagation for every trained step, this would result in a compute requirement of 12m parameter changes each step, which would require multi-terabyte storage for even a competition trained VGG-16 network. The approach has to be optimal to record only changes in neurons that are beyond an average activation threshold, and then only in a constant dimension depth per layer. The resulting delta will be configured every*X* steps or epochs to provide a consistent path of evolution during training.

* 1. Examination
     1. In assessing the *Understanding intermediate layers using linear classifier probes (Alain, G., Bengio, Y., 2018) ,* we finally get an appreciation of the value of examining the learning process as opposed to the post-trained model activations. Although still concerned with feature activation, it introduces a new method that sits in parallel to the traditional neural network library of tools. In this case, monitors of features at every layer of a model are evaluates suitability for classification. The approach is to have completely independent linear classifiers, which is needed for delivery of this theme. This was applied to two large models in Inception v3 and Resnet-50, and the resulting analysis of similar networks helped us focus on high performing and large networks like VGG-16. The issue with this research in line with the theme is the focus is on interoperability, the key outcome is how linear separability of features change throughout the model, although it provides the best research available to optimise the training for visualisation.
     2. A further exploration at examination of the internals of a CNN, and the introduction of identifying how low level concepts interact with high level features is put forward as a proposal for *Quantitative Testing with Concept Activation Vectors*. In this 2018 paper, *Been Kim, et al,* ,agree that explainability remains a significant challenge in the AI world, but is still important to attempt to de-opaque to make sure that a value based machine learning world is possible. While reference is paid to Alain and Bengio’s previous 2016 findings, this goes beyond and uses the CAVs for interpretability for testing. In particular, we can rework the proposed extensions to TCAVs by seeing if we can establish if the learning pattern grows with the shape, texture of repetition of the training images. The findings were also used to develop artistic style generative images based on the test outcomes. While we will not attempt to replicate the Empirical Deepdream, visualisation of the outcomes, similar to the Simonyon outcomes above should be available.

1. Theme C – Visualisation of Convolutional Network Values
   1. Overview of the theme

Focus on the exploration of data layers of a CNN - Explore the layers of a neural network with your camera: <https://experiments.withgoogle.com/what-neural-nets-see> and how this evolved from gradcam. More traditional visualisation techniques are needed to better understand CNNs, and we will use the same approach. More complex 3D imaging will be required due to the need to see temporal differences in the network over time.

* 1. Examination
     1. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7
     2. Palacio, S., Folz, J., Hees, J., Raue, F., Borth, D., Dengel, A., 2018. What do Deep Networks Like to See?, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, pp. 3108–3117. https://doi.org/10.1109/CVPR.2018.00328
     3. Zeiler, M.D., Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901 [cs].

1. Conclusion:
   1. While there has been continuous research and improvement into the field of explainability, there remains a disconnect between the understanding of *why* a neural network produces a specific outcome – of which research is plentiful; and a lack of investigation into establishing a trust quotient to provide verifiable classifications based on either its initial training or amendments provided after transfer.
   2. The benefit of the research is there is a clear metric and methodology to verify what components of a CNN are activated by a source image, and we can do this not only in an interesting way that can show how far down a deep network the feature or concept is important, but also to guide optimisations. CNN visualisation is also strong, but again based on the
   3. What is lacking is a guidance as to why the training of the network led to this behaviour. If we create an analogy of a human being required to make a classification, we look at their background, their education, their environment, the acquired competence – nearly all of this is used to build a body of evidence that this experience can be trusted to make a decision or classification. Like lifelong learning being recorded in a verifiable resumé, CNNs are trained on a repository of either a single domain, or pre-trained using a generic corpus of data, and while the outcomes may be similar, an anomaly in training or a tweak afterwards will yield untrustworthy results, which can be evaluated only extensive measurements.
   4. The missing link is being able to see the training process, similar to recording inputs that a human will see from birth to deployment in the workplace. While we do not need to see everything, we would benefit from seeing the important triggers over time that form and tweak the neuron. This is the aim - to record and store the *lifelog* of the network, to see what parts of the training process were not necessary and to optimise before an exponentially wasteful underfitting of pathways becomes an unnecessary overhead in deployment. What is also missing is being able to baseline this model alongside this record, which would give a provenance. In the event of the next generation of AIF360 manipulating bias for any rational like Giskard from Asimov’s Robots of Dawn, this would show if that manipulation was performed, and If the model can be trusted.
   5. How we get to the next step is to implement a system of measuring the delta within a tolerance at multiple steps over the training. This can be established during CNN training in any of the available CNN frameworks, which also include predefined transfer models to identify the difference in a ‘big bang’ early training to additions to pre-trained systems, which will be quieter, but identify the domain impact on existing models. The visualisation will be challenging, as we will require the model to be visualised in 3D, as 2D models will hide the detail we are hoping to be exposed like a nebula in front of a galaxy. Existing visualisation techniques, and abstraction of the visualisation to multiple 2D images, similar to GradCAM, will help identify the initial benefit of the data produced by the measurement process.

Use APA level headings 1, 3, & 4. See Creswell (page 112) or APA manual for formatting.

### References

1. Alain, G., Bengio, Y., 2018. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644.
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. arXiv:1704.05796 [cs].
3. Explore the layers of a neural network with your camera., n.d. URL https://experiments.withgoogle.com/what-neural-nets-see
4. Gatys, L.A., Ecker, A.S., Bethge, M., 2015. A Neural Algorithm of Artistic Style. arXiv:1508.06576 [cs, q-bio].
5. How convolutional neural networks see the world [WWW Document], n.d. URL https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html (accessed 1.20.21).
6. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R., 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat].
7. Mahoney, T., Varshney, K.R., Hind, M., 2020. AI fairness: how to measure and reduce unwanted bias in machine learning.
8. Palacio, S., Folz, J., Hees, J., Raue, F., Borth, D., Dengel, A., 2018. What do Deep Networks Like to See?, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, pp. 3108–3117. https://doi.org/10.1109/CVPR.2018.00328
9. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7
10. Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs].
11. Zeiler, M.D., Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901 [cs].