# Interpretation of an Inception Convolutional Neural Network Model through Concept Whitening Layers

Megana Reddy Boddam

## INTRODUCTION

Convolutional neural networks (CNN) have been very successful in image classification since their first creation in 2012 [1]. There have been many further variations and improvements made on the classic CNN. One such high performing CNN variant is one that incorporates Google's Inception Architecture [2]. This model is made up of multiple neural network layers that include convolutions with different kernel sizes, max pooling, concatenations, dropouts, and inception modules as shown in figure 1 [3]. Batch normalization is used extensively throughout this model and applied to activation inputs. The final probabilities of the input image being in a specific category is evaluated using Soft-Max.

In 2020, M. Chen et. al. used this model to automatically classify the presence or absence of Hepatocellular Carcinoma (HCC) in histopathology H & E (Hematoxylin-Eosin stained) images from the Genomic Data Commons [4]. They were able to achieve 96 accuracy, which is equivalent to the ability of a pathologist with 5 years of experiential medical knowledge. Despite the accuracy and efficacy of this pathology use case of ML, as of the current day, there are very few pathology tools that use ML as standard medical practice [5]. One of the biggest reasons is the lack of transparency in ML models. For example, in this model, the interactions between layers of a CNN are not easily understood by Pathologists. They have years of experience in reading histology images and wouldn't have the confidence that the ML model is using medical field standard practices in its predictions. Improving the interpretability of ML models would go a long way towards increasing confidence in them [1].

Concept Whitening is a technique developed by Z. Chen et. al. that improves the interpretability of an existing neural network by aligning that network to specific concepts at specific layers [6]. It does not provide a full explanation of the network's computations, rather it imposes a specific concept on one of the batch normalization steps of ResNet50 CNN developed by He. et. al. [10]. This proved to improve the understanding of and confidence in the neural network. The researchers were able to show that the concepts related to the data domain knowledge were being learned by the image classifier CNN.

The project plan is to apply concept whitening layers on the Inception CNN model used by M. Chen et. al. on histology images to detect liver cancer. The concepts I plan to use are standard medical image patterns that are commonly present in liver histology

slides if they display liver cancer. The HCC-present images have micro-trabecular patterns (or unconnected and disjoint endothelial lines), multi-nucleated cells, and frequent vascular (blood vessel) invasion [14]. These concepts will be converted into matrices and computationally applied to the Inception CNN model used by Chen et. al. This will help non-cs experts understand that the concepts important to HCC detection are being learned by the Inception CNN. Hopefully, this leads to better practical reception of AI tools and techniques in the medical field and wider community.

# GOALS

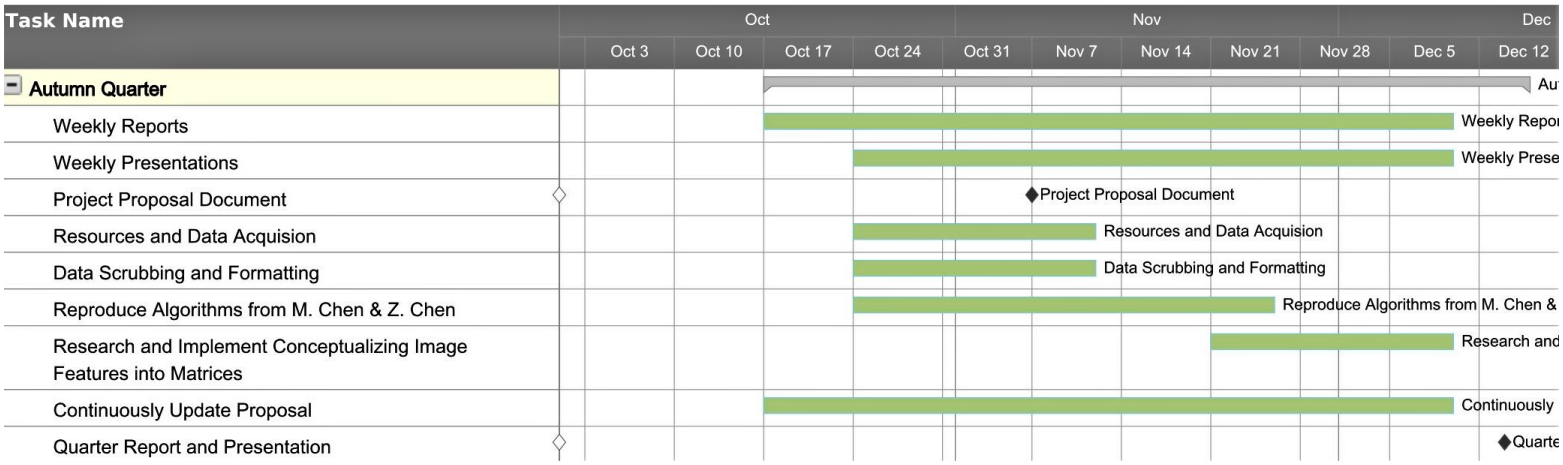
The minimum goals of this project are the following. Reproduce the M. Chen et. al. article from Precision Oncology Journal in its use of Google's Inception CNN model to categorize liver histology images into HCC cancer presence or absence with 96 accuracy. Reproduce the Z. Chen et. al. article from Nature Machine Intelligence in its use of Concept Whitening on a 50-layer CNN (ResNet50) model on Places365 MIT-collected data. The expected goal is to apply medical concept whitening modules to the Inception CNN Version 3 model and achieve the same or better classification accuracy as M. Chen et. al. The medical concepts are micro-trabecular patterns, multi-nucleated cells, and frequent vascular invasions. The aspirational goal is to apply the above modules to Google's Inception-ResNet CNN model to create a novel algorithm that can achieve similar or better precision and performance compared to the metrics in M. Chen et. al. [16] The metrics used to evaluate the success of the new algorithm will be accuracy, performance, concept importance maps, concept purity, intra-concept similarities, inter-concept similarities.

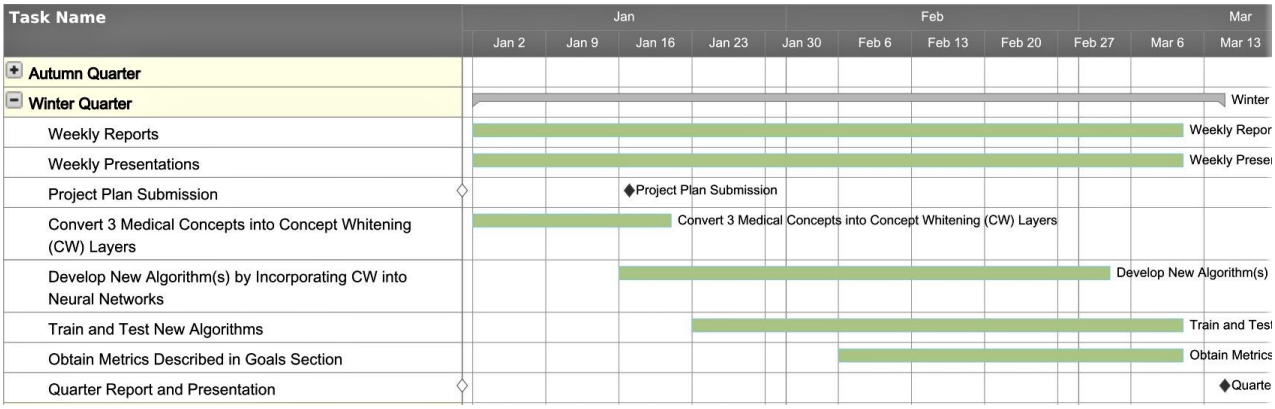
# POSITIONING OF THE CAPSTONE

Incorporating interpretability into artificial intelligence is an important aspect of the technique that is being continuously researched. There are two schools of thought in these efforts: introducing interpretability through the AI model is one and the other is interpreting the model after it is created or the post-hoc explanation methods. The difficulty with the second type of school of thought is that the methods are more a summary or comparison of performance or accuracy statistics rather than actual explanations [6]. Saliency maps are one example of interpretation after model creation. They assign importance weights to each pixel to describe its connection with the image's category [7]. These are proven to be problematic and unreliable because they provide similar explanations, for example: edge distinction, for multiple categories [8].

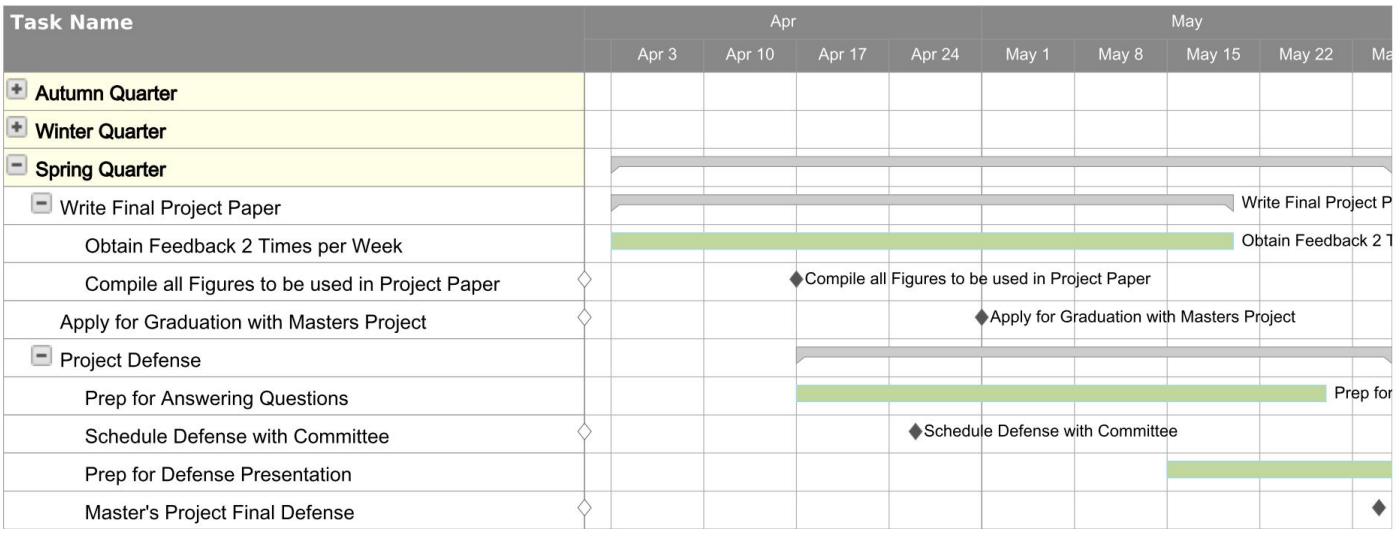
To address problem, researchers moved towards techniques where they tested whether layers or combination of layers in a CNN could be described as predefined concepts [9]. These show some positive alignment but not always because there is no guarantee the CNN layers have trained themselves to recognize these predefined concepts. So, the CNN models often fail to achieve the expected concept separation.

This project proposes to use the first school though in interpretability where AI model




trains with the concepts deterministically shaping a part of the model: i.e., inherent interpretability. Current research in this field has been into networks that have the following incorporated concepts: case-based reasoning, grammatical sentence structure rules, rules to decompose images into sections, and more (11, 12, 13). This project is unique in terms of the domain, concept type, and the CNN being utilized. I plan to use Inception CNN version 3 described in Google Guides and in M. Chen et. al. to categorize publicly available liver histology images into presence or absence of cancer. However, I plan to alter this CNN model with biochemical concepts that describe the presence of liver cancer in histology images, namely multi-nucleated cells, disjoint endothelial lines, and an abundance of blood vessels. To help with the translation of medical concepts into concept matrices, I plan to use the concept whitening techniques described by Z. Chen et. al. [6] This project's research of the incorporation of medical concepts into a complex computational neural network is its uniquely interesting aspect in the interpretable machine learning field.




# SCHEDULE of TASKS and DELIVERABLES

# CONSTRAINTS, RISKS, and RESOURCES

The most important difficulty I will face with this project is to understand how to convert concepts such as broken lines (micro-trabecular patterns), single nucleated cells (opposite of multi-nucleated cells), and high/low numbers of blood vessels into numerical matrices. Next, it will be time consuming to check which concepts or combination of concepts would be the most effective and efficient for the Inception CNN. Lastly, I will need access to UW Bothell computing resources because I will need to analyze approximately 100 Gigabytes of image data files. I will also need to use this big data to train CNN models which also requires more computing power than my laptop. To sanity check my algorithms, I will be using Google's Co-laboratory to handle smaller batches of data, train the algorithms and test them.

# SCOPE and REQUIREMENTS

I will be utilizing the Inception CNN V3 described by Google researchers and M. Chen et al. in Figure 1 [3].

I will be utilizing the medical concepts described in histology and oncology journal articles to edit the above CNN. The concepts I plan to use are single nucleus cells in normal liver cell images, broken endothelial lining in cancerous liver cell images, and high numbers of blood vessels in cancerous liver cell images [14]. One of the concepts, broken endothelial lining, is described below in figure 2.

I plan to convert these concepts into concept whitening layers that can replace any normalization layer in any CNN using Z. Chen's CW layer methodology [6].

Finally, I plan to experimentally incorporate the CW layers into Inception CNN to either test the capabilities of CW-CNN algorithms or create my own novel conceptual

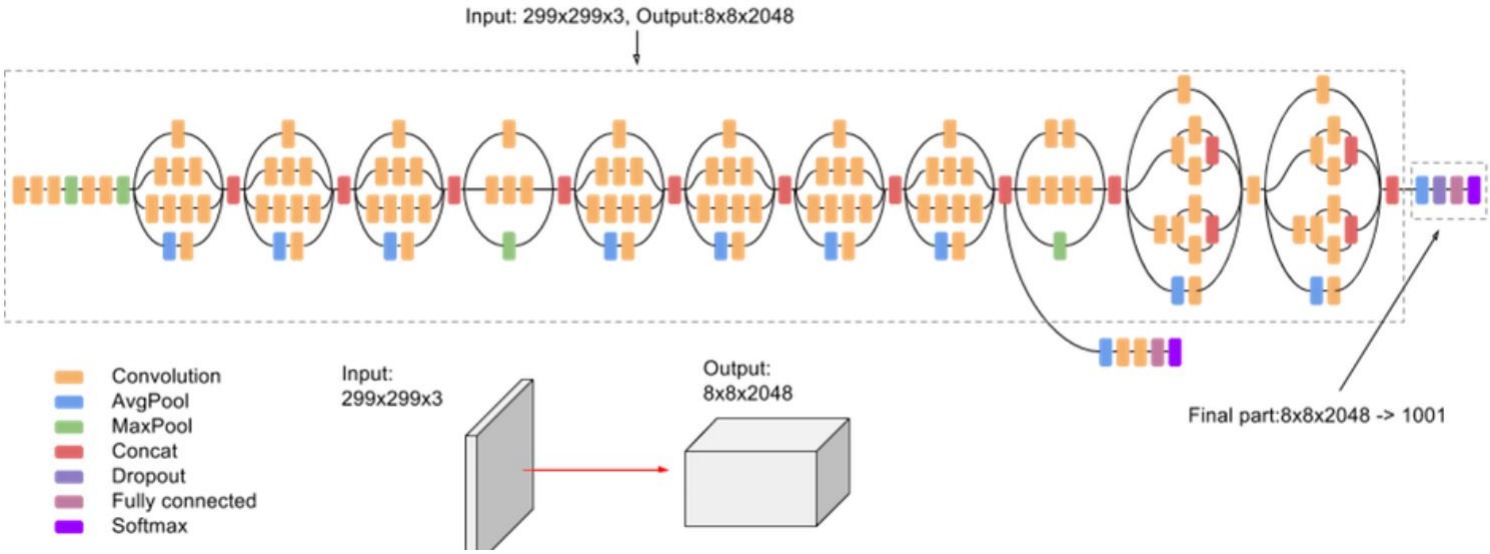



Figure 1: Google's Inception CNN V3 Architecture [3]

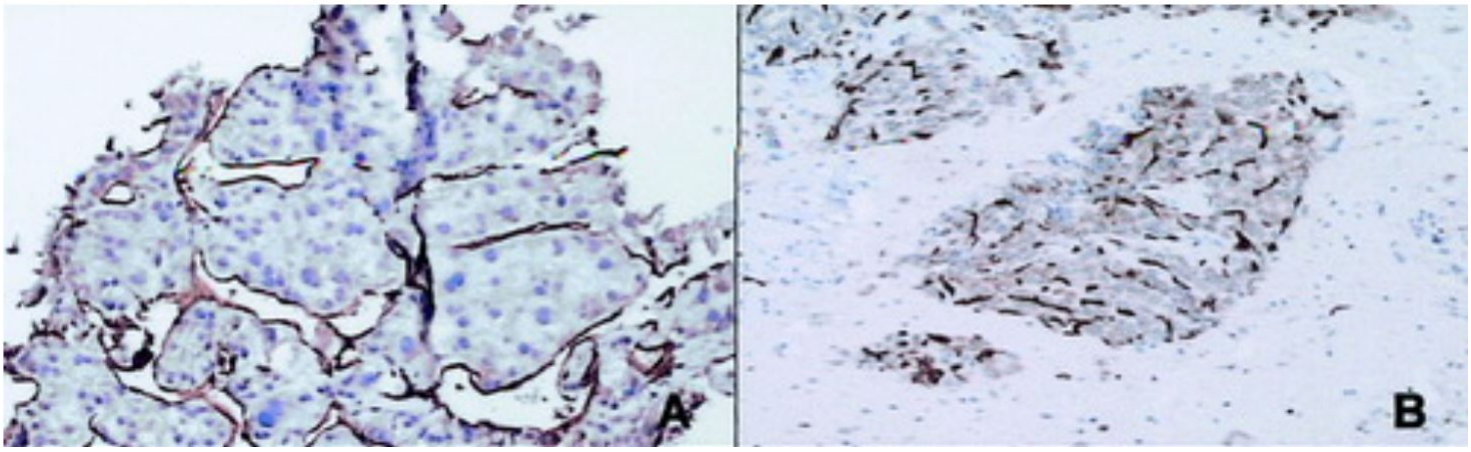



Figure 2: On the left is a normal liver cell histology; it has linear, connected endothelial cell lining represented by black lines. On the right is HCC tissue cells with broken endothelial cell lining represented by disconnected black lines [14].

layering methodology.

# References


[1] C. Molnar, Interpretable machine learning. Lulu. com, 2020.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," CoRR, vol. abs/1512.00567, 2015.

[3] "Advanced guide to inception v3 — cloud tpu — google cloud." [Online]. Available: https://cloud.google.com/tpu/docs/inception-v3-advanced

[4] C. M. B. W. J. H. S. Q. H. X;, "Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32550270/

[5] M. G. Hanna and M. H. Hanna, "Current applications and challenges of artificial intelligence in pathology," Human Pathology Reports, vol. 27, p. 300596, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772736X22000081