DATS-6501 Data Science Capstone

Anomaly detection in Wood Fossil

Prof. Edwin lo

Mid-term Report

Pranay Bhakthula

Introduction:
The secret to humanity's success is probably in how well we are able to use tools to mediate environmental change. But practically all of our present understanding of early cultural evolution comes from research on stone tools and fossilized bones discovered in the archaeological record. Due to their inherent fragility, tools constructed of plants are almost never found in the earliest records of human material culture. The fact that plant materials are used for tools much more frequently than stone in contemporary human civilizations and among non-human primate species raises the possibility that a significant portion of ancient technology is being left out of current archaeological data.
Here, we provide methods for analyzing internal and external damage patterns in living primates' percussive wooden tools. Our research demonstrates that the harm done is irreversible and may endure throughout the fossilization processes. This study provides the opportunity to examine organic artifacts, a significant but overlooked part of the evolution of technology within the primate order.

Dataset:
Fieldwork to collect percussive tools used by chimpanzees was carried out during December 2017 and March 2018 in the North Group of the Tai Chimpanzee Research Project in the Tai National Park (Cote d'Ivoire).
Dataset link : http://cdna.eva.mpg.de/Organic_Tool_Data/

The dataset is also shared my Professor Chen Zeng. The link to it is shared in blackboard submission.

The dataset contains 8 high resolution images of wood fossil which are dated more the 1 million years. The target image we are performing our majority part of the project is "FW31_Damaged_14525_1206_8_18022019.png".

Methods:
In this project we aim to find a method to train a machine to detect anomalies in the wood fossil; i.e, it can detect the damaged part of the fossil from the undamaged part.

For this problem we will analyze a few "features" of the wood fossil images to see if they can separate the damaged from the undamaged portion via unsupervised linear PCA or nonlinear AE/VAE (autoencoder or variational autoencoder).

The features that we use in this capstone project will be:
1. Haralick features
2. VGG's 4096 features normally used in transfer learning
3. Resnet50's Gram matrix used in neural style transfer

Finally, for each of the above features, we run PCA and AE/VAE to visualize data clustering, asking if the damaged crops can be somewhat separated from the undamaged crops

Some of the limitations of the project are the damage on the surface and internal structure will most likely be impacted by the physical properties of wood. These properties vary widely depending on tree species and water content at the time of use. Our study has focused on the materials that were selected by chimpanzees for nut cracking in the Tai Forest. As a result, we only investigated the damage pattern of the most prevalent wood species (Coula edulis).

The required resources for the project are usage of python libraries and cloud computing. Significant progress can be made in the following 2 months and come out with insightful findings from wood fossils.

The one real time usage of this project when fully implemented would be useful for archeologists in remote part of the world to take pictures of fossils and anomalies/unusual/important part can be detected instantly, instead of transporting the delicate fossil to labs for analysis.

Experimental Setup:

First, we crop the wood fossil image in the sizes of 1 x 1mm, 3 x 3mm, 5 x 5mm, 7 x 7mm and store them separately in each folder. We label the damaged part of the image as "damaged" and undamaged part of image as "undamaged".

Since our aim is to use pre-trained models like Resnet50 and VGG19, we convert the image into size of 244x244x3 as it is the input sizes of mentioned pre trained models.

We then apply pre-trained models on the re-sized images and store outputs after applying gram matrix on 1st, 22nd, 23rd, 24th, 25th, 26th, 27th and 49th layers. We perform this step separately for all the various sizes of crops.

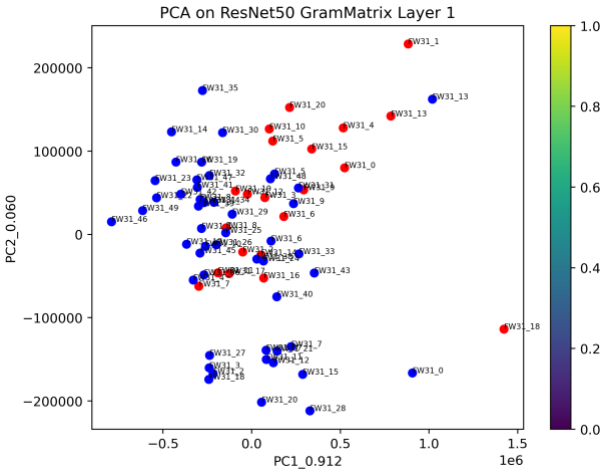
Finally, we apply PCA model on outputs of all the above-mentioned layers for each of the various cropped images and plot a scatter plot of PC2 vs PC1. We color code damaged as “red and undamaged as blue and check to see if the PCA model has separated them distinctively.

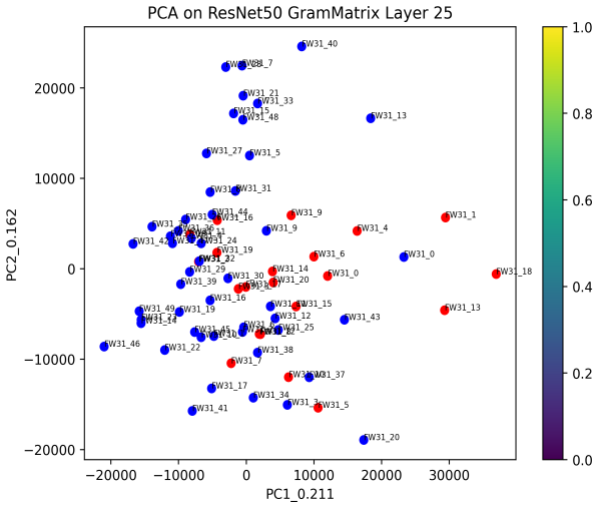
We also repeat the above process for other wood fossil images and check if the model still holds up by clustering these images in places other than our damaged image.

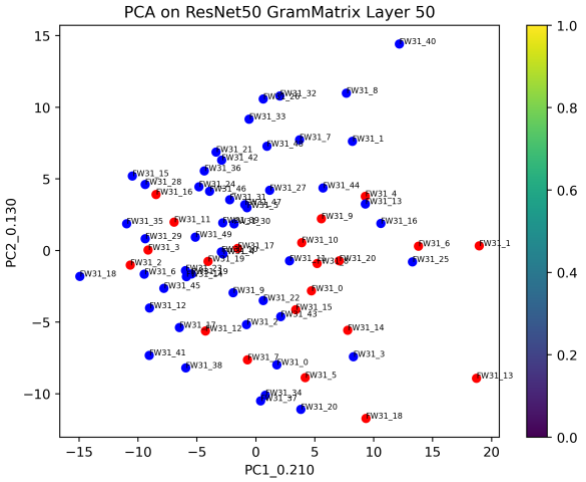
Results:

When applied PCA model on output of various layers of Resnet50 we get following graphs.

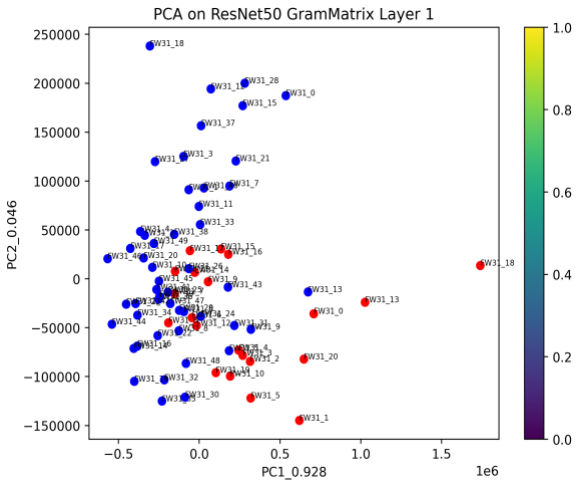
For 1mm cut:

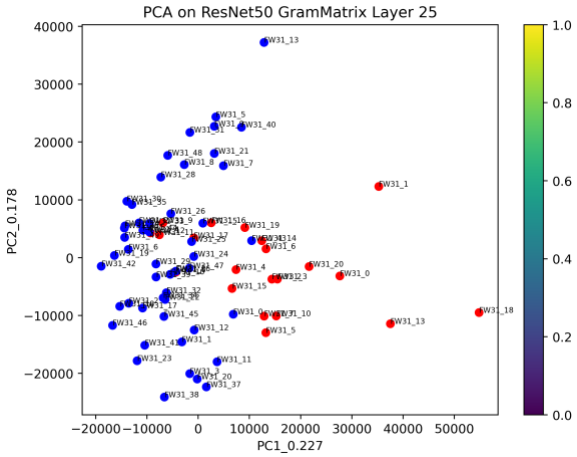


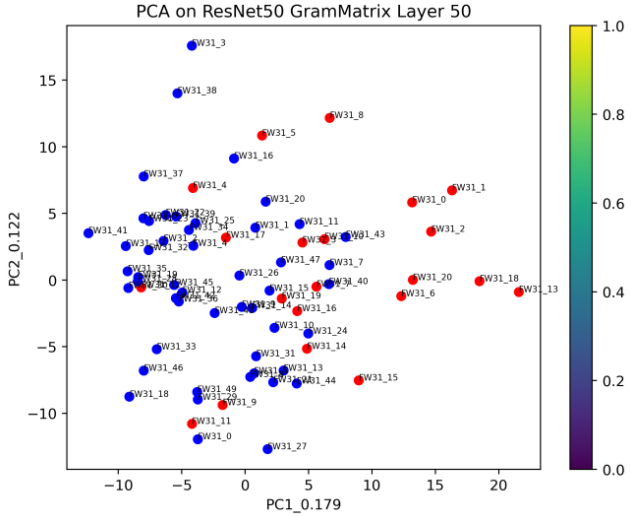




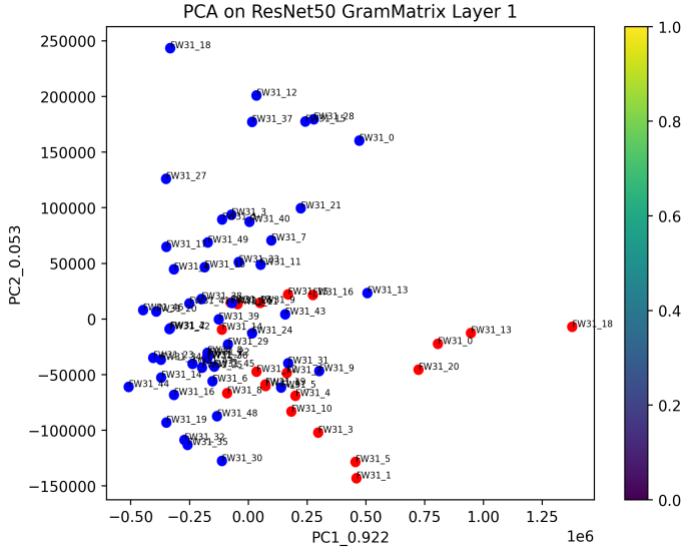
For 3mm cut:

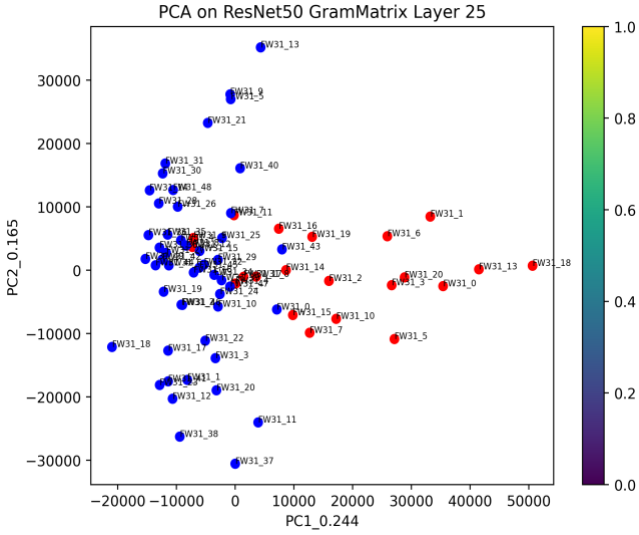


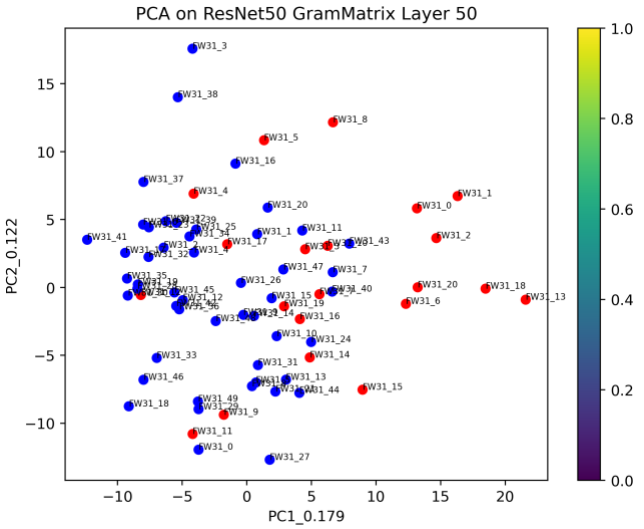




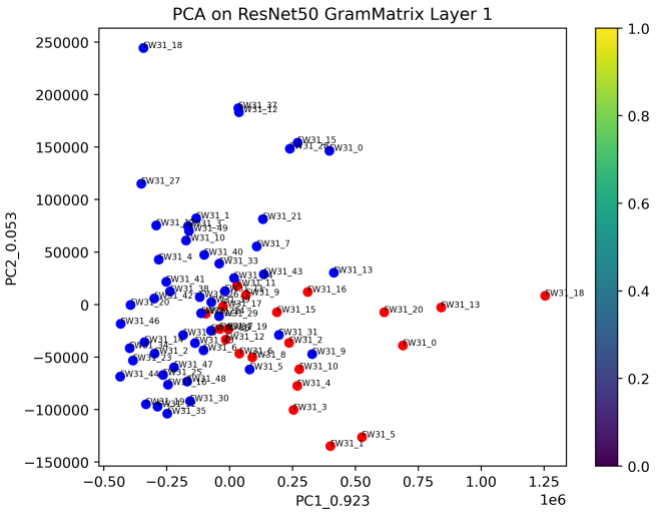
For 5mm cut:

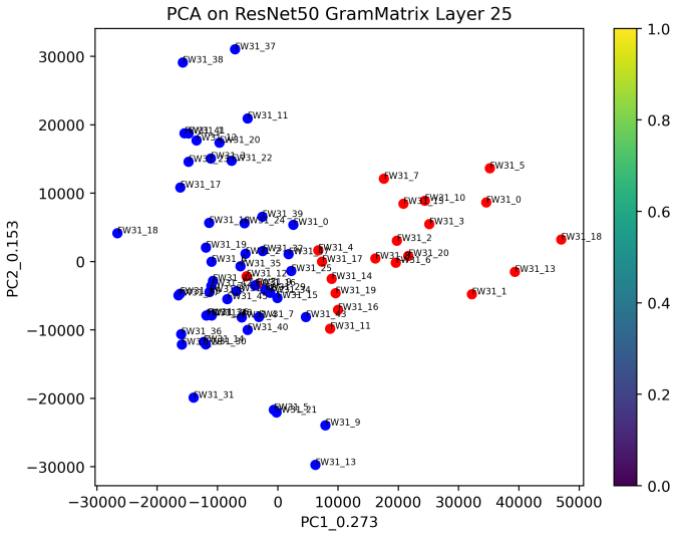


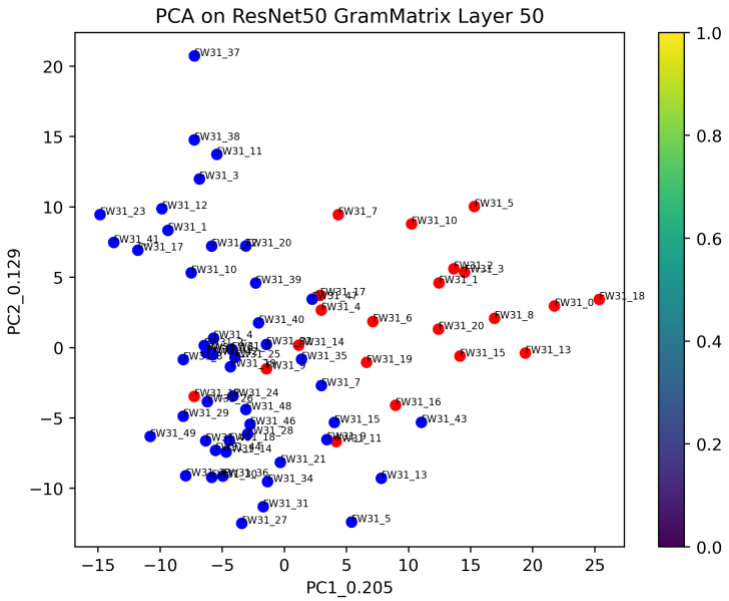



PCA on ResNet50 GramMatrix Layer 50
PC1_0.179
PC2_0.122

For 7mm cut:


PCA on ResNet50 GramMatrix Layer 1
PC1_0.923
1e6
PC2_0.053

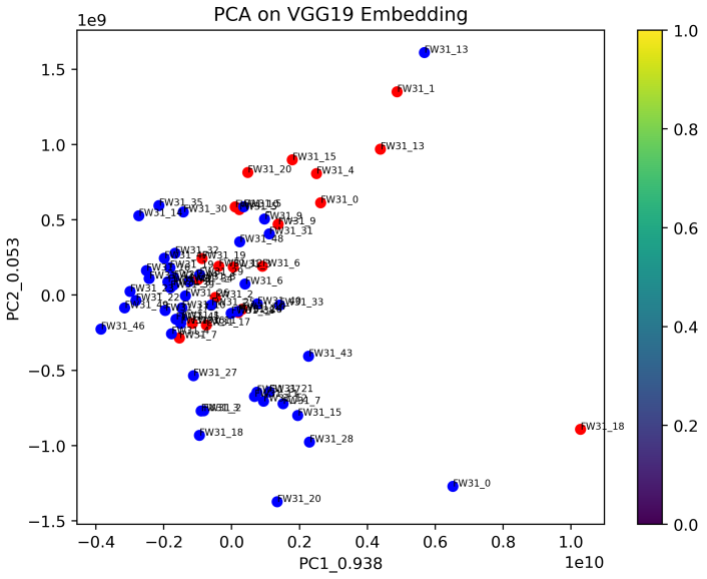

PCA on ResNet50 GramMatrix Layer 25
PC1_0.273
PC2_0.153


PCA on ResNet50 GramMatrix Layer 50
PC1_0.205
PC2_0.129

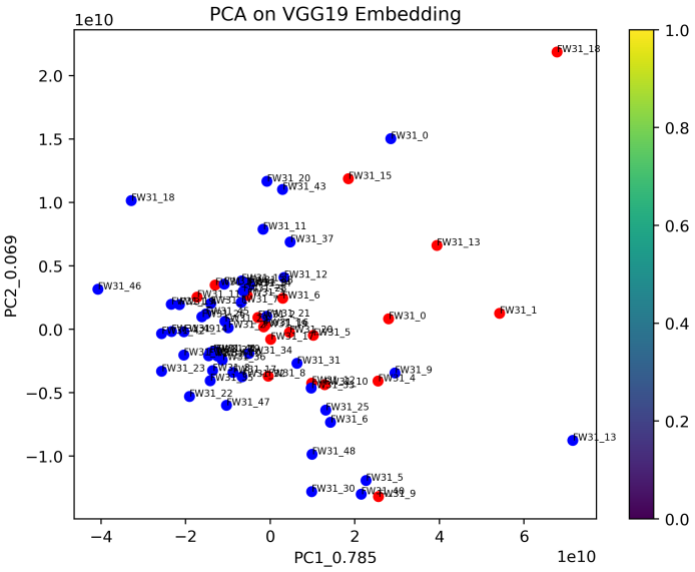
From above graphs we observe that the outputs of 1st and 49th layers are as expected all the various cuts i.e; the damaged part of the image is mixed up with the undamaged part of the image since the initial and final layers represent given image and not the patterns in it. When we look at outputs of 25th layer, we can observe that for cuts of 1mm and 3mm damaged part of the image is mixed up with the undamaged part of the image since it doesn't find proper patterns and conclude the models are looking at very zoomed in versions of image so it can't differentiate between damaged and undamaged part; which are essentially same i.e; image of wood. If we look at graphs of 5mm and 7mm cuts for 25th layer, we can see that majority points of the damaged and undamaged images are separated. This goes to prove that the process we used in this project holds up pretty good for seperating 2 parts of images if they are in different styled (damaged and undamaged).

When applied PCA model on output of various layers of VGG19 we get following graphs.

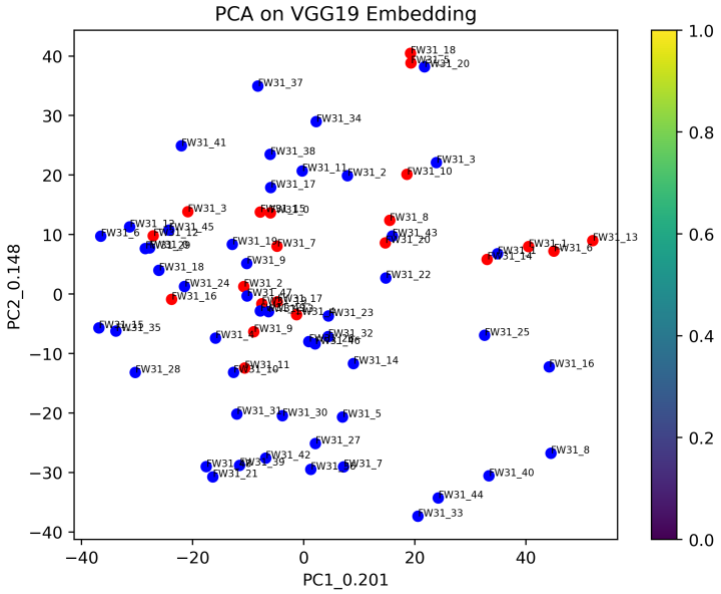
For 1mm cut:



1st layer

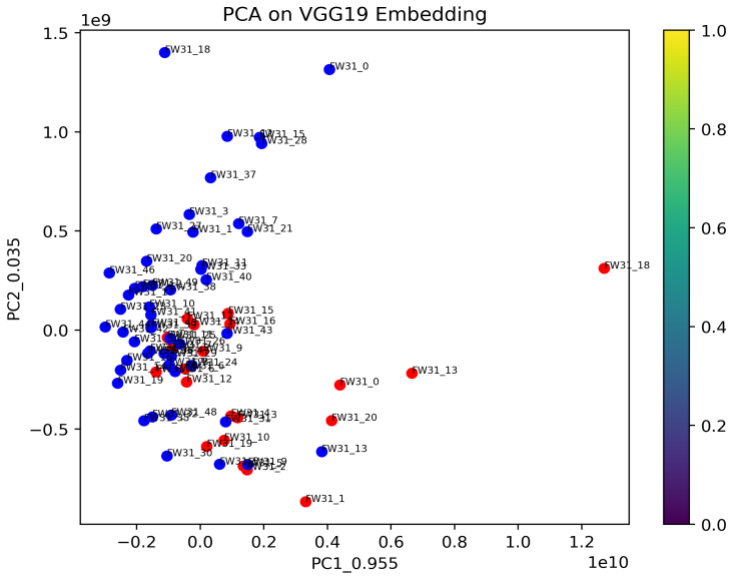


8th layer

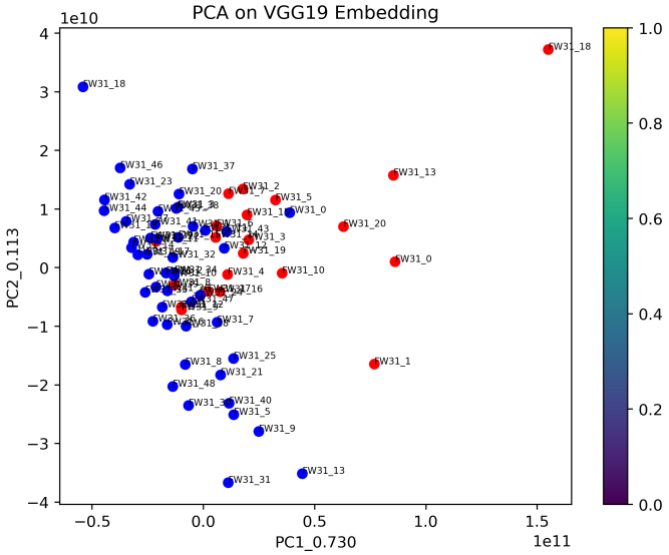


18th layer

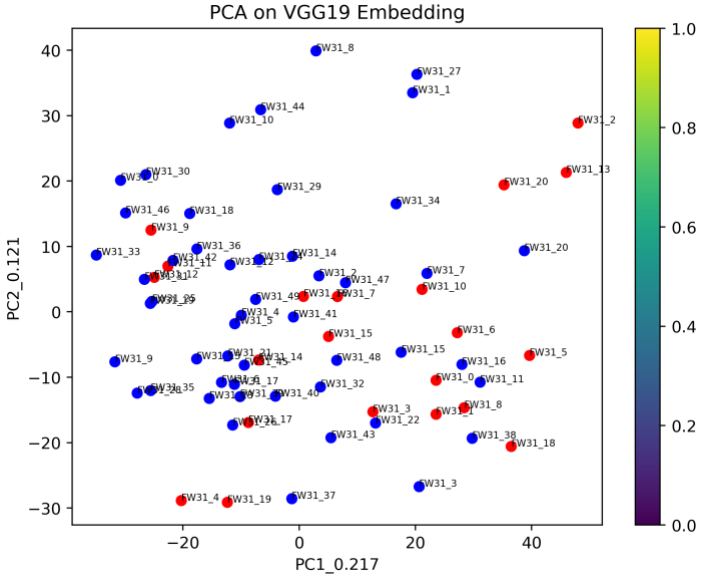
For 3mm cut:



1st layer

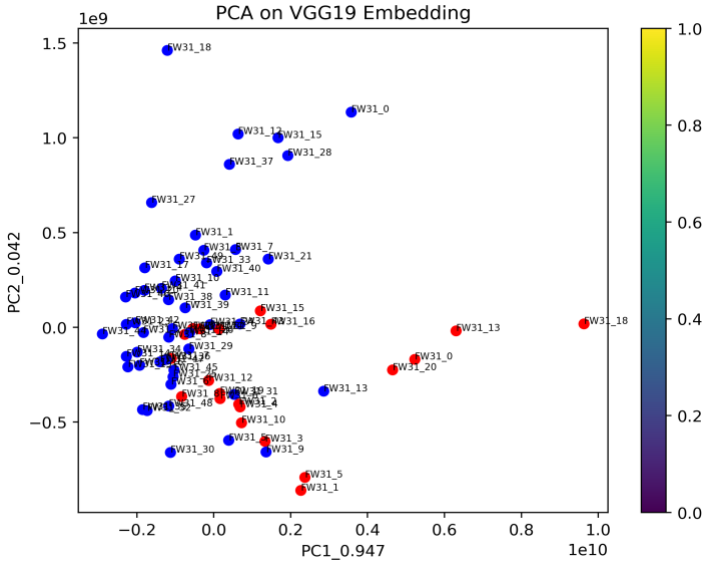


8th layer

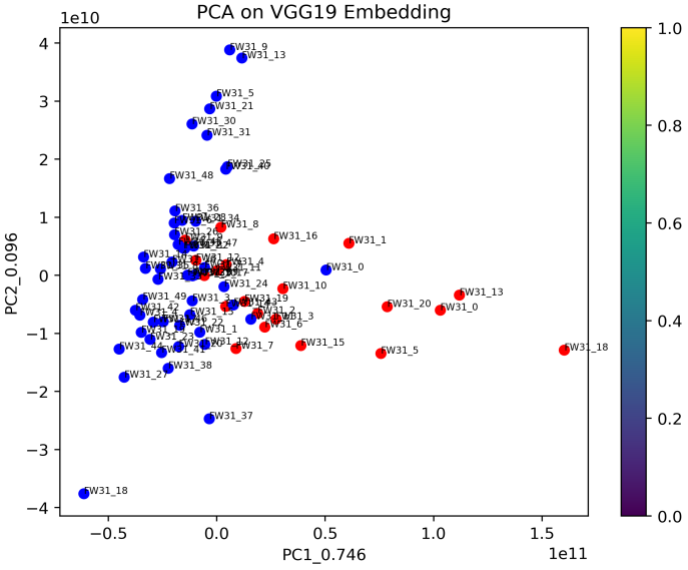


18th layer

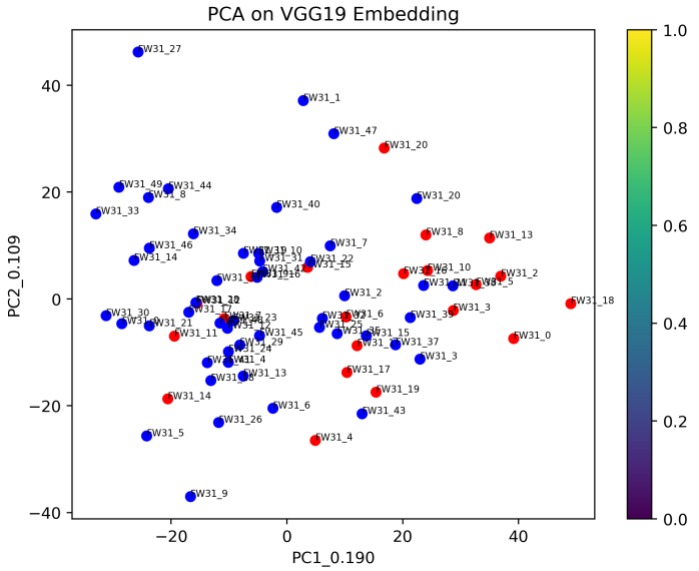
For 5mm cut:



1st layer

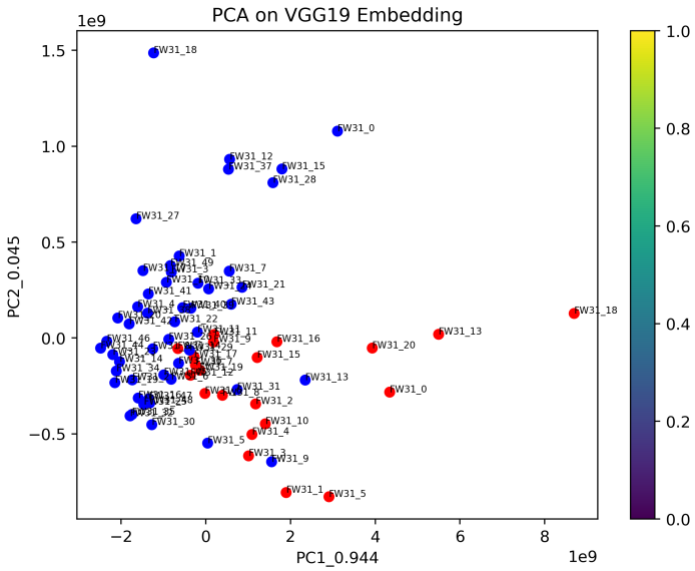


8th layer

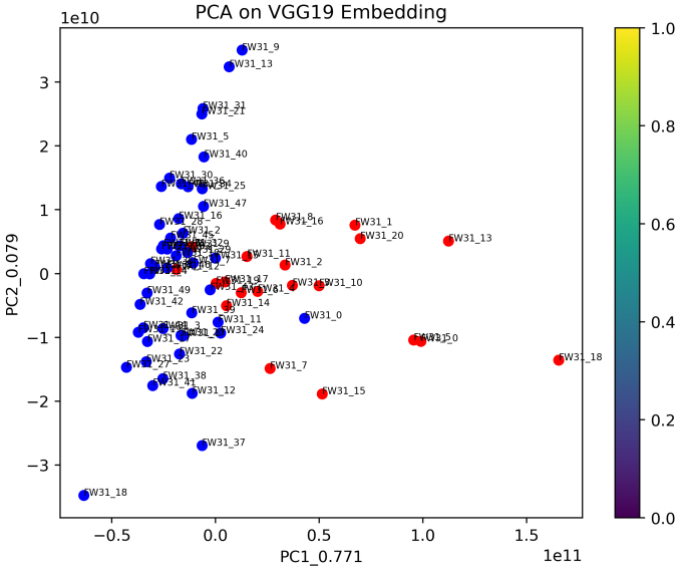


18th layer

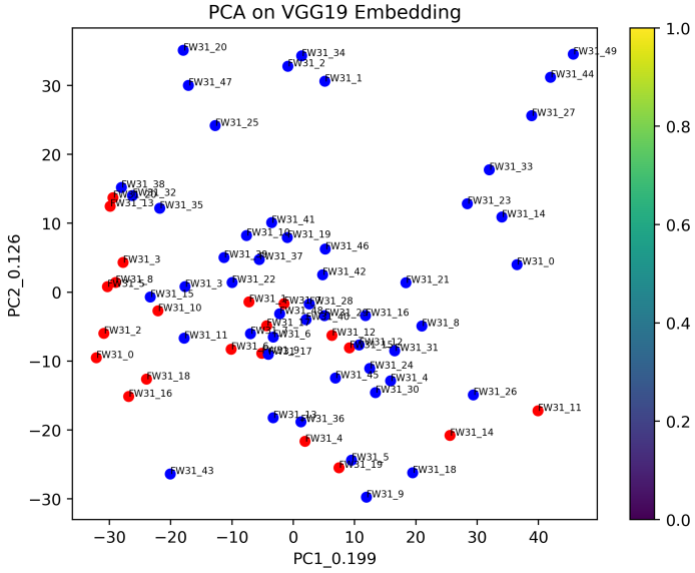
For 7mm cut:



1st layer



8th layer



18th layer

From above graphs we observe that the outputs of 1st and 18th layers are as expected all the various cuts i.e; the damaged part of the image is mixed up with the undamaged part of the image since the initial and final layers represent given image and not the patterns in it. When we look at outputs of 8th layer, we can observe that for cuts of 1mm and 3mm damaged part of the image is mixed up with the undamaged part of the image since it doesn't find proper patterns and conclude the models are looking at very zoomed in versions of image so it can't differentiate between damaged and undamaged part; which are essentially same i.e; image of wood. If we look at graphs of 5mm and 7mm cuts for 8th layer, we can see that majority points of the damaged and undamaged images are separated. This goes to prove that the process we used in this project holds up pretty good for separating 2 parts of images if they are in different styled (damaged and undamaged) by using gram matrix on output of middle layer of pre-trained model on image and using clustering process to separate different styled part of image.

But if we compare graphs of Resnet50 and VGG19, we can observe that the Resnet50 does slightly better job at separating damaged and undamaged images.

Future work:

1. To perform above analysis with other images (FW24, FW26, FW28, FW40, FW42, FW43) of wood fossils.
2. To work on project, to extract haralick features and perform analysis based on it.
3. To prepare video to document the process for future references.

My learnings:

1. By repeating the process of working on pre-trained models like resnet50 and VGG19, I have become comfortable in using them at my will, which is a big improvement considering I was intimidated on using them in the beginnning.
2. I learnt new technique of using neural style transfer to capture patterns in an image and using it to separate varied patterns in the same image.
3. Working with Professor Chen Zeng I learnt to interpret the results in more depth and not losing the main objective of the project while looking at the results.