EECS 445: Project 2 – Tiffany's Tensor Tourism Writeup
Alan Yang

1. Data Preprocessing

a) Training Mean and Standard Deviation

i) Results

Mean:

- R: 115.942
- G: 115.351
- B: 115.663

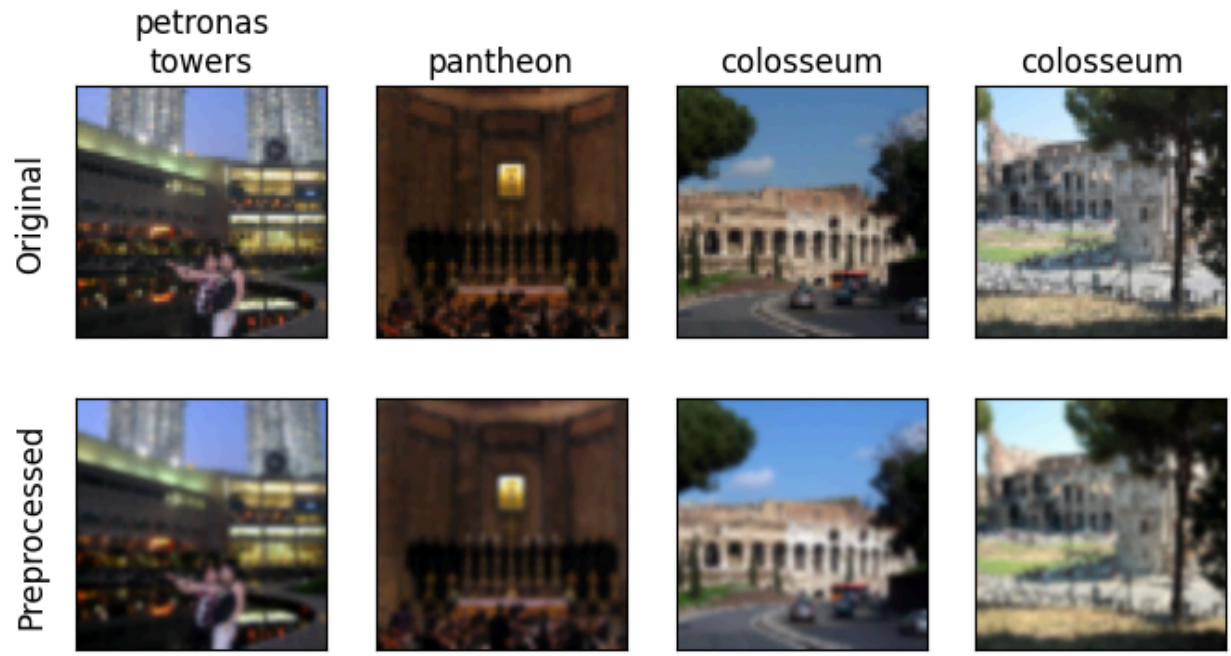
Standard Deviation:

- R: 63.827
- G: 66.688
- B: 72.982

ii) Training Partition

We extract the mean and standard deviation from the training data as opposed to other partitions so that all images can be centered around the same data that we already know.

b) Images



2. Convolutional Neural Networks

a) Target Architecture

Convolutional Parameters = (Filter Size * # Filters in Previous Layer + 1) * # of Filters
Fully Connected Parameters = (Input * Output) + Output

Layer 1:

- Input = 3 x (64 x 64)
- Output = 16 x (32 x 32)

- Filter Size = 5 * 5
- # of Filters in Previous Layer = 3
- # of Filters = 16

Parameters = ((5 * 5 * 3) + 1) * 16 = 1,216

Layer 3:

- Input = 16 x (16 x 16)
- Output = 64 x (8 x 8)

- Filter Size = 5 * 5
- # of Filters in Previous Layer = 16
- # of Filters = 64

Parameters = ((5 * 5 * 16) + 1) * 64 = 25,664

Layer 5:

- Input = 64 x (4 x 4)
- Output = 8 x (2 x 2)

- Filter Size = 5 * 5
- # of Filters in Previous Layer = 64
- # of Filters = 8

Parameters = ((5 * 5 * 64) + 1) * 8 = 12,808

Layer 6:

- Input = 32
- Output = 2

Parameters = (32 * 2) + 2 = 66

Total # of Parameters = 1,216 + 25,664 + 12,808 + 66 = 39,754

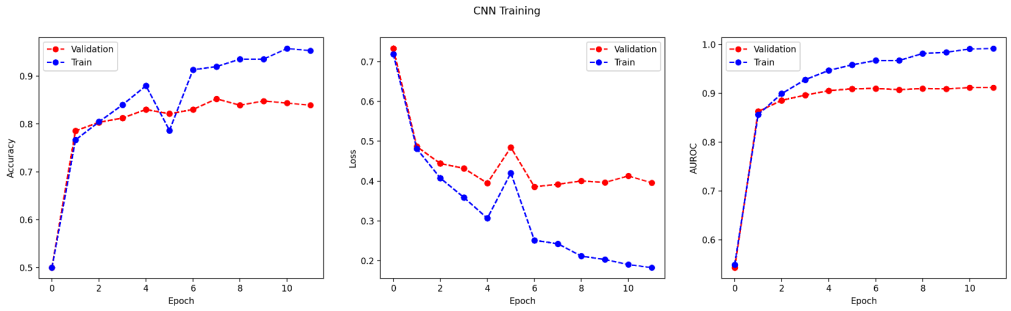
f) Training Plots

i) Increasing Validation Loss

Increasing validation loss can likely be attributed to overfitting and a difference between validation and test data. With overfitting, we may focus too much on our validation data when training our model, which can cause it to treat outliers as normal data. Similarly, our validation dataset may be slightly different from the test data, which may cause us to test our model on data that is not consistent with the entire dataset.

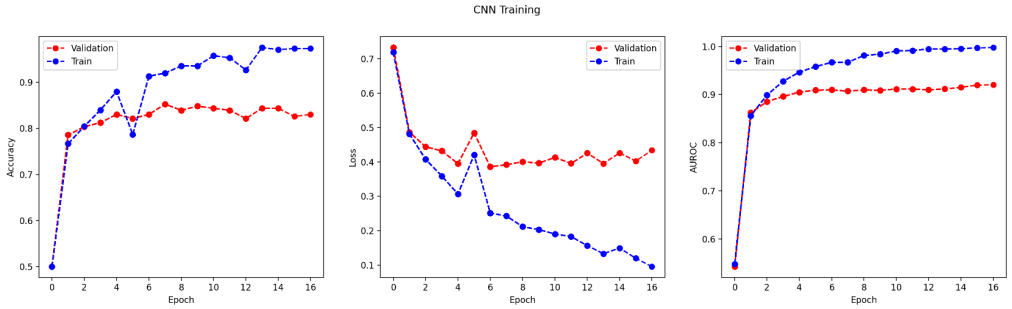
ii) Patience

Patience: 5	**Epoch: 11**



Validation Accuracy	0.8393
Validation Loss	0.3955
Validation AUROC	0.9116
Train Accuracy	0.9533
Train Loss	0.1828
Train AUROC	0.9916

Patience: 10	Epoch: 16



Validation Accuracy	0.8304
Validation Loss	0.434
Validation AUROC	0.9204
Train Accuracy	0.9733
Train Loss	0.0954
Train AUROC	0.9978

Having a patience of 10 gives marginally better results than when the model has a patience of 5. When examining the Training Accuracy, we see an increase from 0.9533 to 0.9733, as well as a slight increase in training AUROC from 0.9916 to 0.9978. However, the Training Loss benefitted the most, with it decreasing from 0.1828 to 0.0954. Increased patience is often better when there are adequate computational resources to handle more epochs, as well as when the data is not going through enough iterations and is possibly underfitting. There is no exact patience that works for every model.

iii) Changing Model Architecture

Change in Input Size to Fully Connected Layer: (8 x 2 x 2) => (128 x 2 x 2)
New Input Size to Fully Connected Layer: (128 * 2 * 2) = 512

Filters	**Epoch**	**Training AUROC**	**Validation AUROC**
8	6	0.967	0.9098
128	5	0.9913	0.9067

When increasing the number of filters from 8 to 128, we see that our model with 8 filters took one epoch longer to reach its lowest validation loss with a lower Training AUROC. Despite the slight decrease in Validation AUROC, the increase in Training AUROC is certainly more significant. This may be due to the greater number of filters increasing the capacity of our model, allowing it to better fit the training data. This may also have decreased the number of epochs, as the model was able to learn more during each iteration, allowing it to reach its lowest validation loss earlier.

g) Test Data

i) Results

	Training	**Validation**	**Testing**
AUROC	0.967	0.9098	0.6745

ii) Overfitting

Our model may be overfitting, as the Training AUROC score is quite a bit higher than the Validation AUROC score. The Validation AUROC likely remains so high as the two datasets are still relatively similar.

iii) Datasets

Our training and validation datasets are likely very different from our testing data, as both the Training and Validation AUROC scores are over 0.9, whereas the Testing AUROC is noticeably lower, being lower than 0.7. This shows that our model is relatively accurate for our training and validation datasets, but is not for our testing dataset, which must differ from the previous two datasets.

3. Visualizing what the CNN has learned

a) Grad-CAM

$$L^{c}_{Grad-CAM} = ReLU(\sum_{k} a^{c}_{k} A^{k})$$

$$L^{1} = ReLU(\sum_{k} a^{1}_{k} A^{k}) = ReLU(a^{1}_{1} A^{(1)} + a^{1}_{2} A^{(2)})$$

$$a^{1}_{k} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\delta y^{1}}{\delta A^{k}_{ij}} = \frac{1}{16} * \frac{\delta y^{1}}{\delta A^{(k)}}$$

$k = 1$:

$$a^{1}_{1} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\delta y^{1}}{\delta A^{1}_{ij}} = \frac{1}{16} * \frac{\delta y^{1}}{\delta A^{(1)}}$$

$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot -1$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 1$
$\frac{1}{16} \cdot -2$	$\frac{1}{16} \cdot -1$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 0$
$\frac{1}{16} \cdot -1$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 1$
$\frac{1}{16} \cdot 1$	$\frac{1}{16} \cdot 1$	$\frac{1}{16} \cdot 2$	$\frac{1}{16} \cdot 2$

a^{1}_{1}:

0	$-\frac{1}{16}$	0	$\frac{1}{16}$
$-\frac{1}{8}$	$-\frac{1}{16}$	0	0
$-\frac{1}{16}$	0	0	$\frac{1}{16}$
$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$

$k = 2$:

$$a_2^1 = \frac{1}{Z}\sum_i\sum_j \frac{\delta y^1}{\delta A_{ij}^2} = \frac{1}{16} * \frac{\delta y^1}{\delta A^{(2)}}$$

$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 1$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 0$
$\frac{1}{16} \cdot 2$	$\frac{1}{16} \cdot 2$	$\frac{1}{16} \cdot 2$	$\frac{1}{16} \cdot 2$
$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 0$	$\frac{1}{16} \cdot 1$	$\frac{1}{16} \cdot 1$
$\frac{1}{16} \cdot - 1$	$\frac{1}{16} \cdot - 2$	$\frac{1}{16} \cdot - 1$	$\frac{1}{16} \cdot 0$

a_2^1:

0	$\frac{1}{16}$	0	0
$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
0	0	$\frac{1}{16}$	$\frac{1}{16}$
$-\frac{1}{16}$	$-\frac{1}{8}$	$-\frac{1}{16}$	0

$a_1^1 A^{(1)}$:

$-\frac{1}{16}$	$-\frac{1}{16}$	$-\frac{3}{16}$	$-\frac{1}{8}$
$-\frac{3}{16}$	$-\frac{1}{4}$	$-\frac{5}{16}$	$-\frac{1}{8}$
$-\frac{1}{16}$	0	$-\frac{1}{4}$	$-\frac{3}{16}$
$\frac{1}{8}$	$\frac{7}{16}$	$-\frac{1}{16}$	$-\frac{5}{16}$

$a_2^1 A^{(2)}$:

$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{3}{8}$
$\frac{1}{16}$	$\frac{1}{16}$	0	0
$-\frac{7}{16}$	$-\frac{7}{16}$	$-\frac{3}{8}$	$-\frac{5}{16}$

$a_1^1 A^{(1)} + a_2^1 A^{(2)}$:

$\frac{1}{16}$	$\frac{1}{16}$	$-\frac{1}{16}$	0
$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{4}$
0	$\frac{1}{16}$	$-\frac{1}{4}$	$-\frac{3}{16}$
$-\frac{5}{16}$	0	$-\frac{7}{16}$	$-\frac{5}{8}$

$L^1 = ReLU(a_1^1 A^{(1)} + a_2^1 A^{(2)})$:

$\frac{1}{16}$	$\frac{1}{16}$	0	0
$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{4}$
0	$\frac{1}{16}$	0	0
0	0	0	0

b) CNN Features

The CNN seems to identify pictures of the Hofburg Imperial Palace using the base palace features, including the walls, pillars, arches, and the courtyard in front, with those features being much darker in the GradCAM. The pillars and sky, on the other hand, are much lighter, showing that they are not used as much to identify the Hofburg Palace.

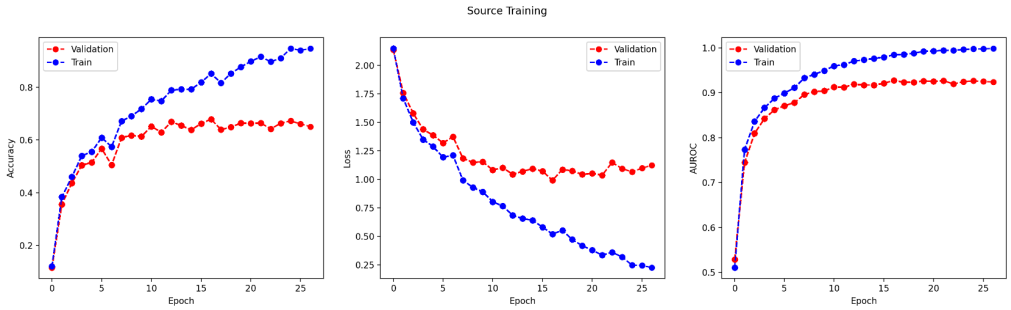
c) Overfitting and Datasets

One thing to note, is that in some pictures, the clouds are also darker, showing that they are also being used to help classify the Hofburg Imperial Palace, although they are significantly lighter than many of the actual palace features (as mentioned above). The fact that clouds are being used to help classify, however, may show that there is overfitting or perhaps a disconnect between datasets, as the validation and testing datasets may not have clouds. Additionally, pictures of certain parts of the palace are darker than others, meaning that the model might be emphasizing certain locations and wings of the palace. This could also create issues if the training and validation datasets have more images of certain areas than the testing dataset. Overall, the visualizations somewhat confirm my hypothesis from 2(g).

4. Transfer Learning & Data Augmentation

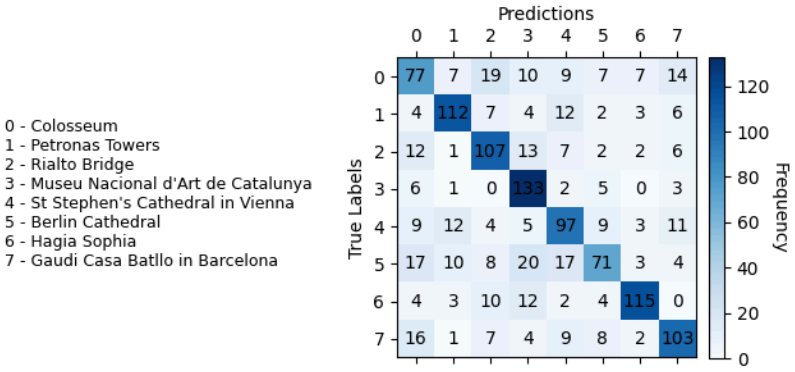
4.1 Transfer Learning

c) Source Model Architecture



Epoch with the Lowest Validation Loss: 16

d) Prediction Accuracy



The Museu Nacional d'Art de Catalunya is by far the most accurately predicted landmark, while the Colosseum and Berlin Cathedral are the least accurately predicted. This is likely because the Museu Nacional d'Art de Catalunya has the most unique features of the dataset, which makes it less likely for the images to be misclassified as something else. The Colosseum and Berlin Cathedral, on the other hand, likely have very general features, which may be similar to the other landmarks, causing them to be misclassified as something else, as well as causing other landmarks to be misclassified as them. This is especially true for the Colosseum, which is both misclassified as something else and classified incorrectly for another landmark relatively frequently.

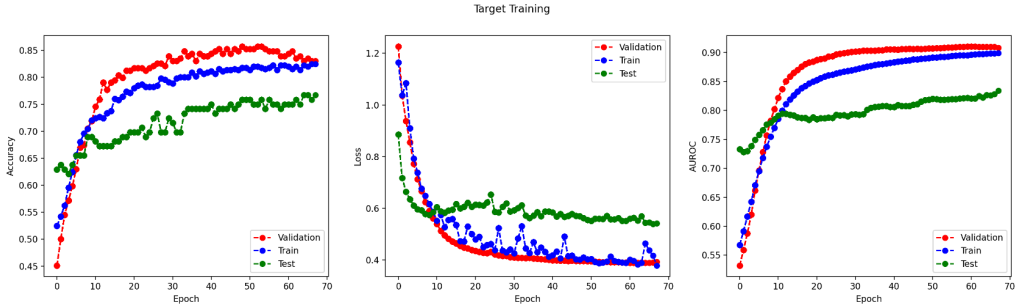
f) Freezing Layers

	AUROC		
	Training	**Validation**	**Testing**
Freeze All CONV	0.8971	0.9099	0.8249
Freeze First Two CONV	0.9756	0.9298	0.8255
Freeze First CONV	0.9949	0.9382	0.8184
No Freeze	0.9868	0.9397	0.7818
No Pretraining or Transfer Learning	0.967	0.9098	0.6745

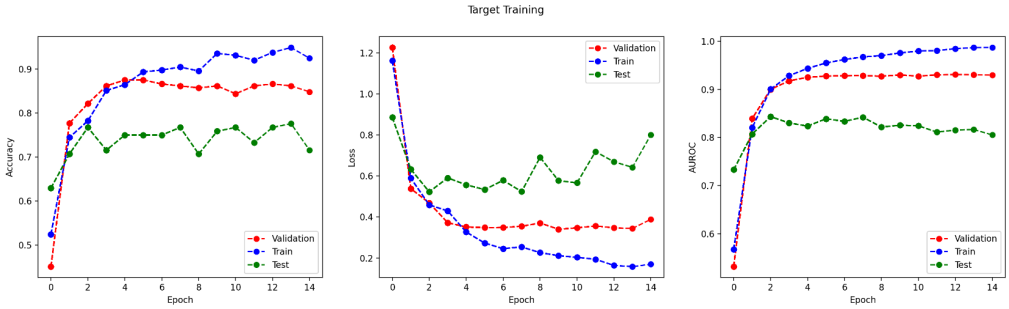
Yes, the source task is helpful to the target task, as shown by significant increases in performance for all models that used transfer learning. This comes even with the model that froze all convolution layers, which resulted in a much lower Training AUROC score, but a Testing AUROC score greater than all but one model. This is likely due to the models with transfer learning being able to better distinguish which features were useful, and which may be more commonplace.

As we increase the number of layers frozen, we can see that the Training and Validation AUROC scores generally trend downwards, while the Testing AUROC score generally increases. This is likely due to less overfitting. By freezing layers, the model remains more generalized and is less likely to correctly predict the training and validation datasets. However, by not overfitting to the training and validation data, it becomes better at predicting the more general test dataset. We can also see this from the graphs below, where the models with fewer frozen layers tend to increase drastically in performance in fewer epochs and die down soon afterward. In some cases, the Testing AUROC score even slowly trends downwards. However, with all layers frozen, we can see the models slowly increasing in performance, peaking at a much later epoch at a much lower AUROC score. However, after a brief stagnation in performance, it continues to slowly increase in performance, showing that it is still improving to the testing datasets, rather than having already overfitted to it.

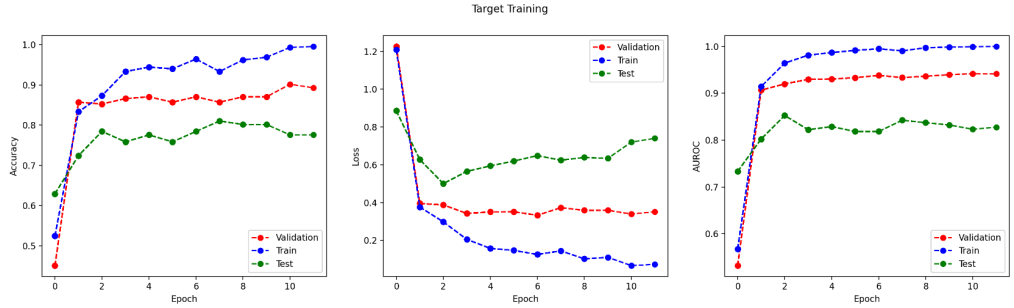
Freeze All CONV:



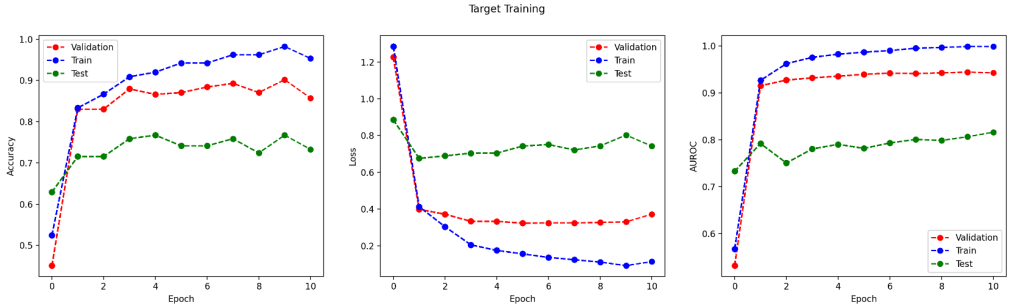
Freeze First Two CONV:



Freeze First CONV:



No Freeze:

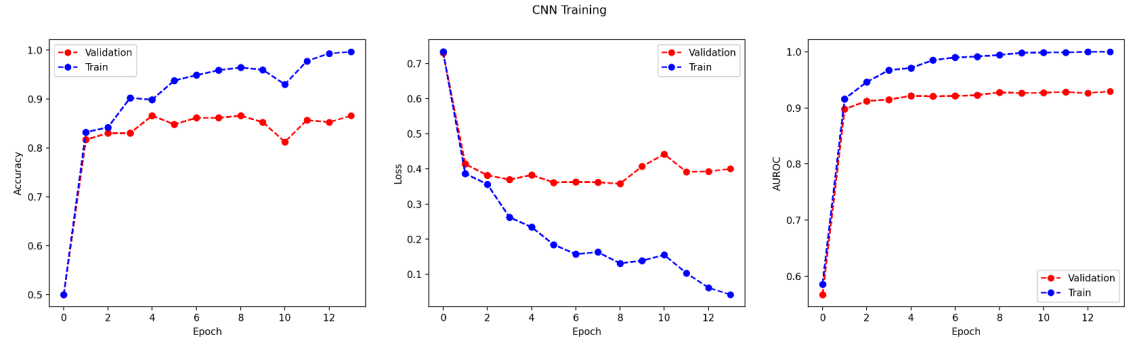


4.2 Data Augmentation

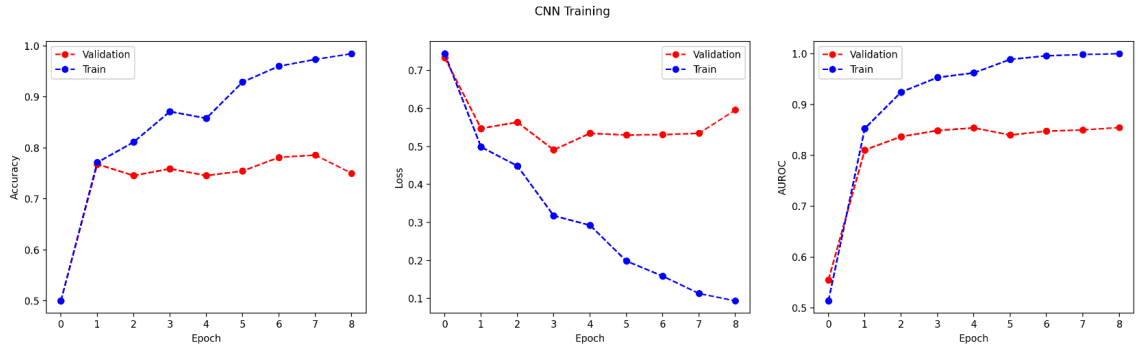
b) Augmented Data Results

	AUROC		
	Training	**Validation**	**Testing**
Rotation (keep original)	0.9949	0.9288	0.7096
Grayscale (keep original)	0.948	0.8493	0.7851
Grayscale (discard original)	0.9738	0.8492	0.8148
No Augmentation	0.967	0.9098	0.6745

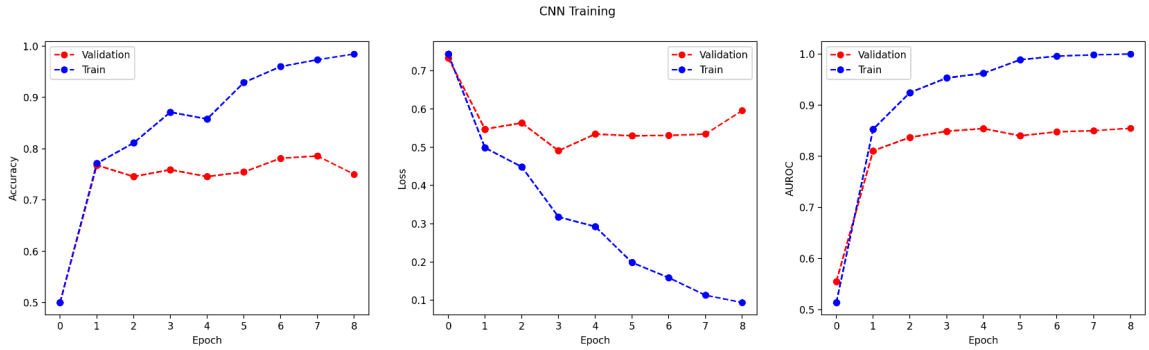
Rotation (keep original):



Grayscale (keep original):



Grayscale (discard original):



c) Evaluation

We can clearly see that both rotation and grayscale augmentations result in greater training, validation, and testing AUROC scores. This makes sense, as the variation in the images helps prevent the model from emphasizing certain features and the way the images are taken. The training and validation AUROC score was greatest with rotation, as the images are often very similar to the originals. This is further corroborated by the significant drop in loss by the first epoch.

On the other hand, the testing AUROC score was greatest with grayscale images (especially after discarding the original). This could be attributed to the lack of colors making the models generalize more, perhaps removing an emphasis on certain features based on color schemes. This would likely emphasize the shapes and structural features of the landmarks more. This greater generalization helped the testing AUROC score continue to increase even after the initial jump in the first couple of epochs.

5. Challenge

:(