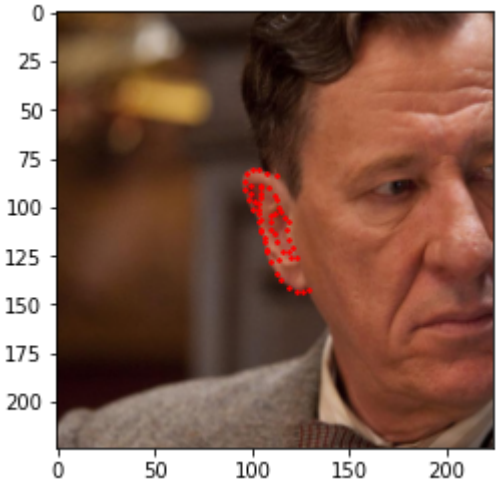
Final Project Report

Tianye Chen, tianyech@buffalo.edu
Bohui Nong, bohuinon@buffalo.edu
Shuwei Hou shuweiho@buffalo.edu

Dataset

Our dataset is from the Intelligent Behaviour Understanding Group (ibug), which contains images of people with at least one ear visible. Each image has a points file with 55 coordinates identifying the unique landmark features of the outer human ear. Since this dataset has a unique structure, with each record having an image having a pts file, we've created a custom dataset class for the purpose of transforming the images and points accordingly. Below is a sample image that has undergone resize, grayscale, and normalized transformations, where the magenta dots are the 55 ear landmark coordinates.



Planning

Our goal is to train a model that can accurately plot the 55 ear landmark points on any image given that the ear is visible in the input image. To get a better understanding of what type of model architecture works best on this task, we have decided to use some of the predefined models in Pytorch to do some basic training. We used ResNet18, and VGG16 with both pre-trained and non-pretrained variations for a total of 4 models. All four models were using L1Loss with SGD and trained using the same hyperparameters. Below are the test results of each model:

ResNet18 pre-trained

Test loss: 63.55239295959473, Test mse: 4716.59326171875, Test r2:

-8.692088360295886

VGG16 pre-trained

Test loss: 72.98396682739258, Test mse: 6002.18115234375, Test r2:

-11.35760177043982

ResNet18 Non-pre-trained

Test loss: 64.26624298095703, Test mse: 4796.69775390625, Test r2:

-8.856666271650335

VGG18 Non-pretrained

Test loss: 72.97989273071289, Test mse: 6001.54833984375, Test r2:

-11.356361563565114

Based on the result, we would be using ResNet18 as our model with a convolutional block attention mechanism and no pre-trained weights initialized.

Training

We started by training our model without any optimization techniques and observed its performance. As expected, using the r2 value as the performance metric, the result is in the negative range. Our goal is to achieve an r2 value of at least 0.8 for our model to have a decent prediction accuracy. We experimented with various optimization techniques such as different optimizers, loss functions, dropout ratios, learning rate scheduler, learning rate, gradient clipping, and early stopping. The optimization setup that performed the best involved using the following hyperparameters:

- Optimizer: Adam with smooth L1 loss
- Dropout: 0.2
- Learning rate: 0.0035,
- Learning rate scheduler
- Early stopping.

However, the r2 value is still below 0.5 using this setup. To further optimize the model we decided to flatten the y_pred and targets before passing them to the criterion. This would ensure that each landmark point is evaluated on a one-by-one basis instead, which will provide more accurate results. After applying flattening, the mode results improved greatly with Test Loss: 20.8889 and Test r2: 0.8159. However, the test loss is too high compared to train loss and

validation loss with 3.8 and 5.1. We've concluded that the cause of this disparity is due to either overfitting the model or an insufficient number of training samples.

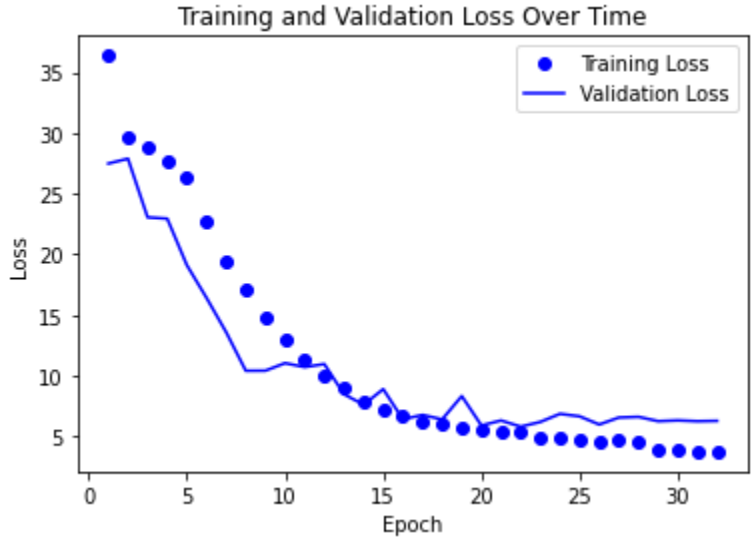
To address this issue, we proposed to either obtain more images online or augment our existing images in the dataset. Considering that we would need to manually label the landmarks of online images due to the scarcity of already labeled images, we've decided to augment our existing images instead. Our augmentation process involved applying vertical and horizontal flip on the origin image to produce a set of 3 training samples. Since the image contains other elements apart from the human ear, such as the person, background, etc., we decided to also center the images with the ear as the focal point by zooming and cropping the images. After this process, we would have a set of 6 training samples consisting of the original, vertical flip, horizontal flip, crop, crop vertical flip, and crop horizontal flip. These augmentations will help the model to more accurately capture the pattern around the ear to identify ear landmarks.

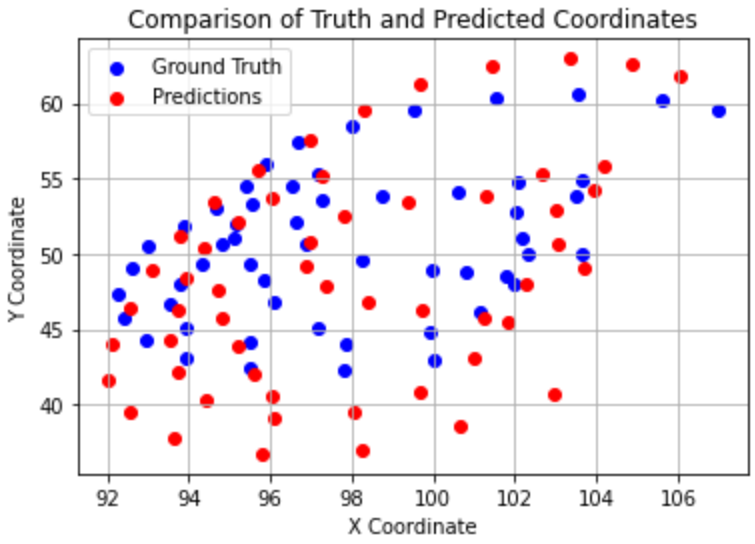
Evaluation

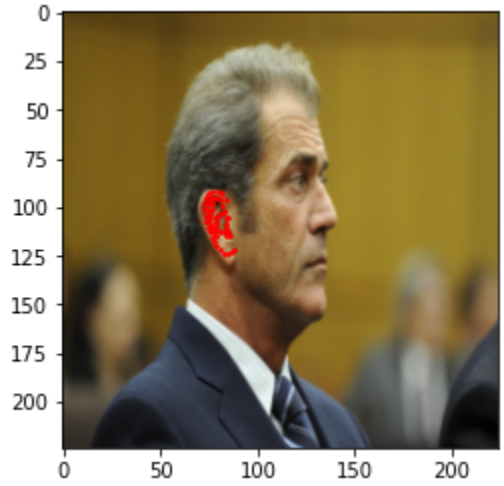
We would use the measurement of the loss value and r2 value to determine the performance of our model. The information revealed by the loss and r2 values is very limited as we can only approximately deduct the overall performance of the model. To gain further insights into our model's performance, we implemented a comparison graph between the true and the predicted landmarks, in which we can visualize the model's ability to accurately determine the position of the landmarks. Below are the final results that we obtained from using the parameters that were defined in the previous step as well as a sample of an augmented image. These results would not be possible without expanding our dataset through augmentations in previous steps.

```
Epoch 00028: reducing learning rate of group 0 to 3.5000e-04.
Epoch 29/50: train loss: 3.954746890575328, train r2: 0.9757773960125793
               validation loss: 6.244681894779205, validation r2: 0.569366253673987

Epoch 30/50: train loss: 3.8544111315240253, train r2: 0.9766814930749234
               validation loss: 6.308944821357727, validation r2: 0.5502190473830346

Epoch 31/50: train loss: 3.774909735994136, train r2: 0.9774760235861352
               validation loss: 6.241860270500183, validation r2: 0.5532622613243978

Epoch 32/50: train loss: 3.7339012001423124, train r2: 0.9779119765287448
               validation loss: 6.270541489124298, validation r2: 0.5495155572607274

Early stop at Epoch: 32
Test loss: 12.484852075576782, Test r2: 0.956828380723906
```


Training and Validation R-Squared Value Over Time
R-Squared Value
Epoch
Training R2
Validation R2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
5
10
15
20
25
30


Training and Validation Loss Over Time
Loss
Epoch
Training Loss
Validation Loss
35
30
25
20
15
10
5
0
5
10
15
20
25
30





Conclusion

There are still certain cases where the model can be inaccurate in the prediction of ear landmarks. We suspect the cause of these inaccuracies is a result of the noises that are present within real-world images factored in the innate differences in the shape of each human ear. As such, further improvements can be made through a larger dataset that contains a more diverse set of ear shapes. Due to the insufficient amount of data we were able to obtain, our validation set also suffered, as a result, the criterion results were less accurate.

Through this project, we were able to gain valuable insights into how to optimize the performance of models when faced with a limited dataset. We learned about the techniques that

can be applied to artificially increase the number of samples of a dataset through different augmentation strategies.

Contributions

Team Members	Project Part	Contribution (%)
Bohui Nong	• Implementation of Convolutional Block Attention Module and base model • Image augmentation • training/optimizing/evaluation of the model • Trello management	40
Tianye Chen	• Implementation of custom dataset • Refine model architecture • Writing the report • Training/optimizing/evaluation of the model	40
Shuwei Hou	• Project proposal • Research papers • Dataset	20

References

- https://ibug.doc.ic.ac.uk/resources/ibug-ears/ (Dataset)
- https://blog.paperspace.com/attention-mechanisms-in-computer-vision-cbam/
- https://openaccess.thecvf.com/content_ECCV_2018/papers/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.pdf
- https://medium.com/analytics-vidhya/resnet-understand-and-implement-from-scratch-d0eb9725e0db
- https://pytorch.org/tutorials/beginner/data_loading_tutorial.html
- https://github.com/harshavardhan183/Ear-Landmark-Detection-AND-Convex-Hull-Extraction
- https://github.com/kbulutozler/ear-landmark-detection-with-CNN
- https://github.com/Dryjelly/Face_Ear_Landmark_Detection