Personalized Perception Project

Student: Sheng Kai Liao

Date:9/9/2021

American University

Project goal:

As human beings, we have our own preference on what we see, hear, and feel. It can be constructed based on personality, personal experience, unique context, and the situation at the time. For instance, beauty is a personalized perception, and everyone has their own criteria. Some people might prefer small dogs with lots of hairs and others might prefer bigger dogs with floppy ears. So, there is an assumption we try to research with. Are we able to capture an individual's personalized perceptual preference and predict it based on limited data using machine learning? Based on the question assumption, this project focuses on researching what and how different machine learning models and different computer vision technologies can predict human preferences well with limited data.

Key words: Computer Vision, Machine learning, Personalized Perception

Experiment:

We set up the experiments with Psychopy which is a very powerful tool for designing variant experiments and collecting data from observers by publishing the experiment on Pavlovia. All experiments have the same format, we give participants a pair of images and they may choose one they prefer.

For now, we focus on finding the attributes of the images which may really affect how people define cuteness. Thus, in our experiment, we offer participants two cats' images and ask them to choose the one they think is much cuter.

Fig1. The interface of our experiment, the participant is able to choose the left or right image regarding his preference.

There are two types of our experiment, one we call a random sample experiment in which we randomly picked up 2000 unrepeated images (like Dog, Cat, House…. etc.) to construct the pair

decision experiment with 1000 trials. In this case, we can easily obtain balanced binary results (1000 labeled data for like, 1000 labeled data for dislike). The other one is a so-called Fully sampled experiment in which we offer 50 unrepeated images within one category, and all images will fully compare to each other. Therefore, the experiment has 1225 trials. The result for this experiment will be decimal data for each image which we called its score, and the score is according to how many times the image is selected during the experiment by the same participant.

Feature extraction:

For each image in the experiments, we use pre-trained VGG-19 to extract their own features. For this experiment, we try to extract the features from FC1 layers and FC2 layers from random sampled data and fully sampled data, and two layers provide different perspectives from VGG-19 (You may refer Fig 3. and Fig 4.). Also, the extracted features from VGG-19 will be 4096 long vectors which can't be easily visualized. Thus, to find the relationship between images based on the features, we descend the gradient of features by using t-SNE methods which will turn the multidimensional data into 2/3D coordinates (here we turn the features into 2D). For the following figures, they illustrate the image distribution regarding the t-SNE transformed 2D coordinates.

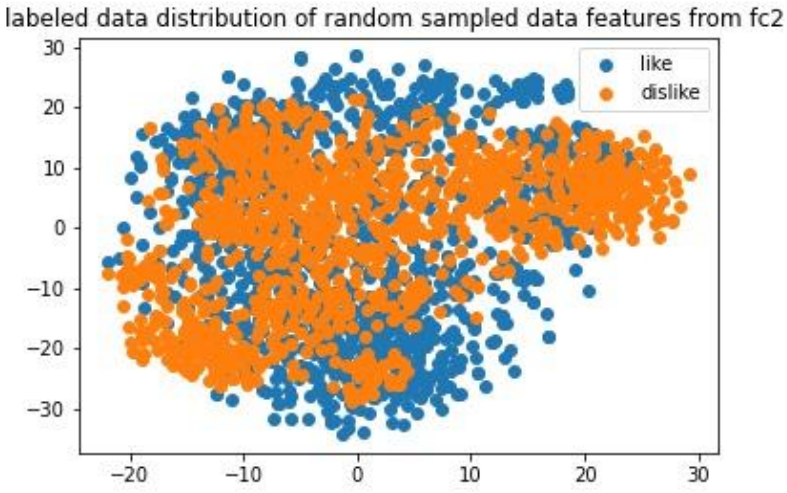


Fig 2. The labeled data distribution includes 2000 images

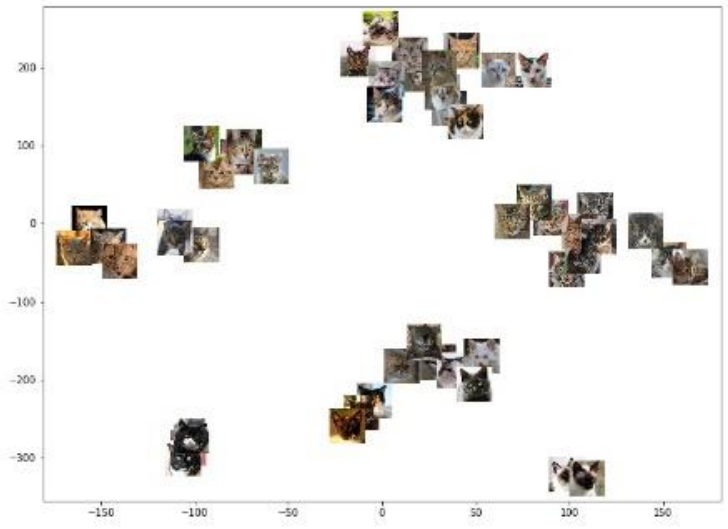


Fig 3. FC1 features tSNE plot distribution with images

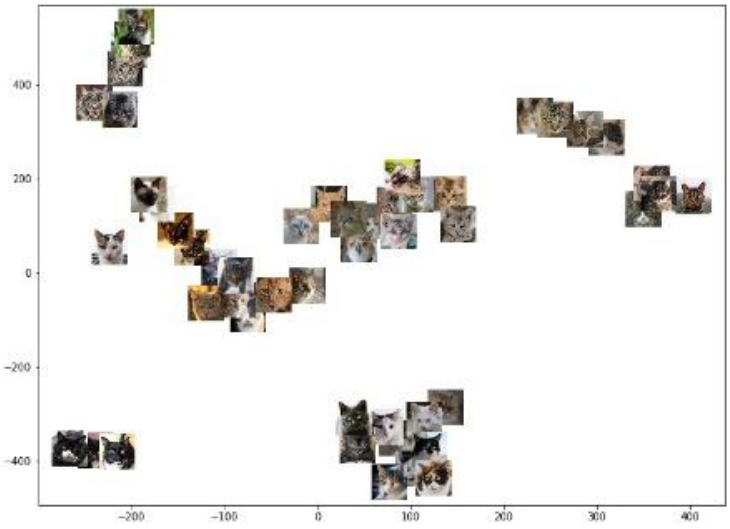


Fig 4. FC2 features tSNE plot distribution with images

Data analysis:

- Random sampled experiment:
    - Regression result:

        Our first idea for analyzing the data is trying to fit them into a regression model. And the result matches our expectations. Since there is no clear trend of the

distribution (according to Fig 2.), the prediction of the models are not really good in linear regression model, polynomial regression model and logistic regression model.

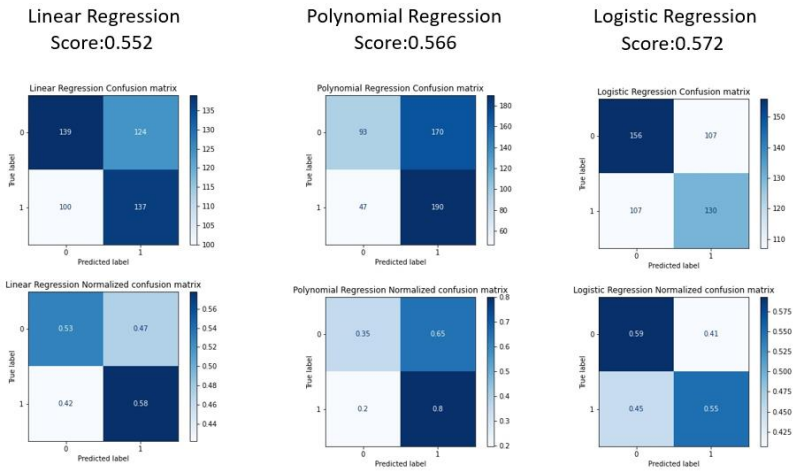


Fig 5. Regression result with random sampled data

- VGG-19 Fine-tuning result:

In this study, we decided to fine tune VGG-19 as the baseline for this experiment. For the dataset, we did the data augmentation for extension, such as rotation, flip the image from left to right and top and down and gaussian blur. The reason why we chose this method for augmentation is that we think it is an unlikely way to affect the participant's initial decision. For testing and validating, we split the data into 80% and 20%. Regarding the right side charts of Fig 6, we can observe that the prediction accuracy is only 54.4% which is really low and the training loss and the validation loss while each epoch doesn't decrease continuously, which means the model isn't trained well according to the dataset.

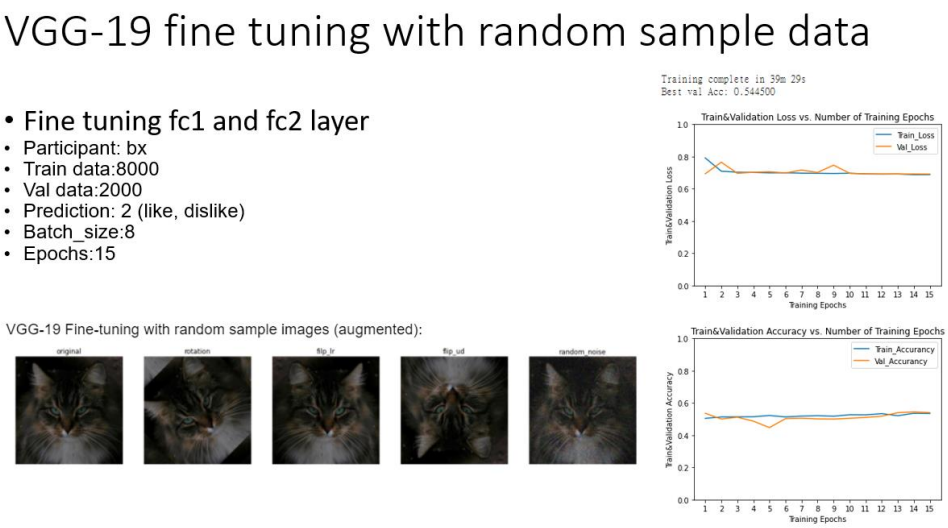


Fig 6. Vgg-19 fine tuning result with random sampled data

- Fully sampled experiment:

Compared to the random sampled experiment, we believe the fully sampled experiment is able to obtain more completed perception data from participants because every image is compared in this experiment. The result of this experiment is a sequence decimal score for each image. For easier observing the trend, I plot the color gradient figure for different participants (Fig 7. and Fig 8.). As you can see, different participants have different highlighted data points in the figures and several data points have similar color gradients between each participant.

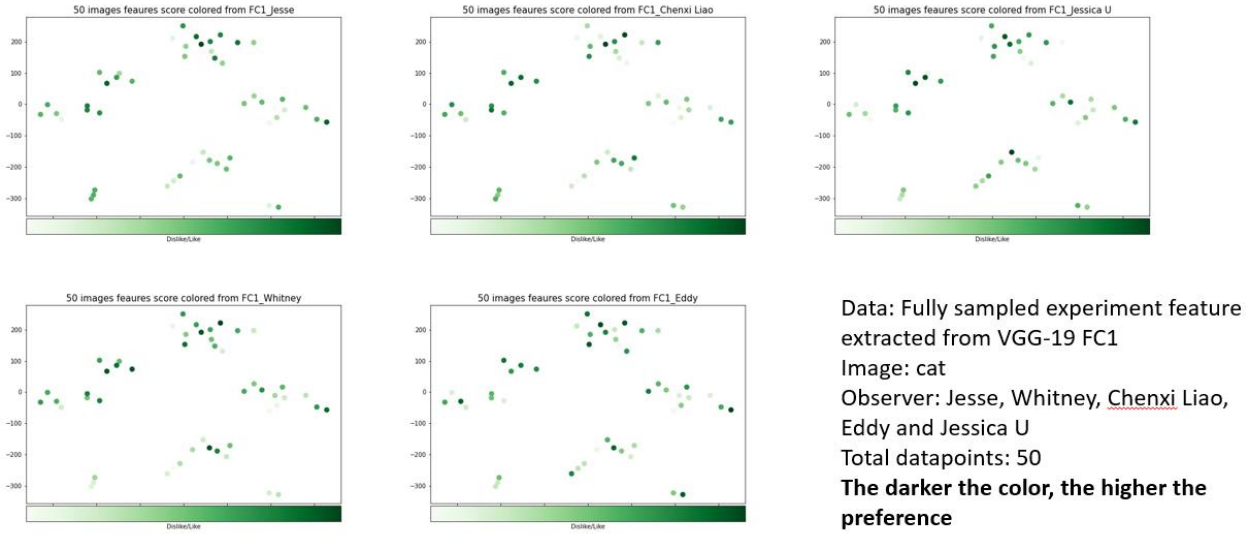


Fig 7. Fully sampled data distribution with fc1 layer's feature

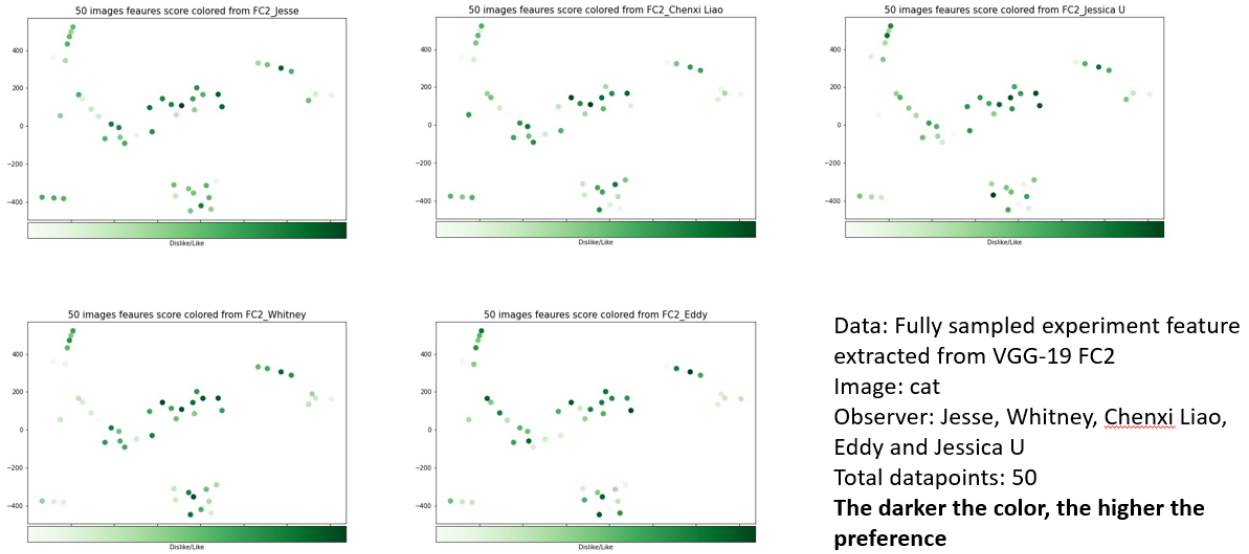


Fig 8. Fully sampled data distribution with fc2 layer's feature

In order to classify the data, we separated the result in two ways. One is to separate the data in 5 groups according to the score of the image, such as 0-10, 11-20, 21-30, 31-40, 41-50 (Fig 9. and Fig 10.). In this way, the data is categorized for group discovery and also it allows us to fine tune the VGG-19 with consistent classes. Unfortunately, by classifying data into 5 classes, the data in each group is imbalanced

and has low quantity. Therefore, we also try the other way is to divide the data into two groups based on the median of each participant's result (Fig 11. and Fig 12.). If we look at Fig 11. You can observe that most participants like the top group and the slightly upper left group (the red circles), and mostly dislike the button and the slightly lower right group (the lime circles). For Fig 12., it is much easier to observe that most of the participants like the center group (the red circle) and dislike the rightest group (the lime circle). Even though there is much noise in the features from VGG-19 since it is trained with different purposes, the model can still distinguish some attributes that may affect most people's perception (in this case, the definition of cuteness).

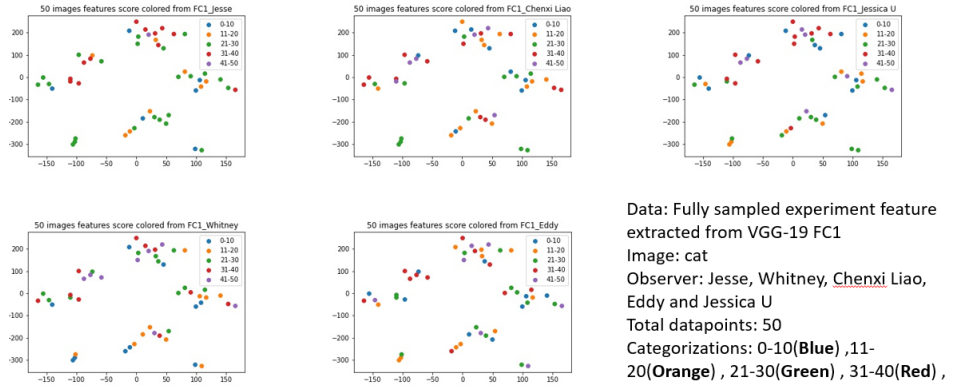


Fig 9. The tSNE plotting with features from VGG-19 fc1

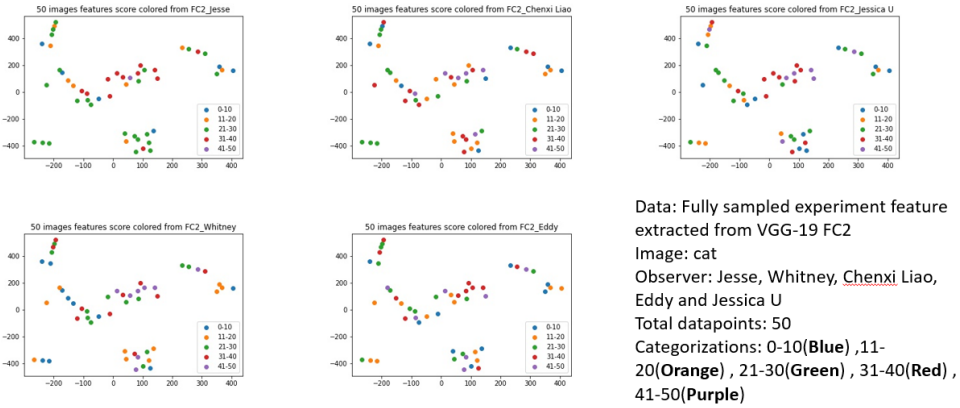
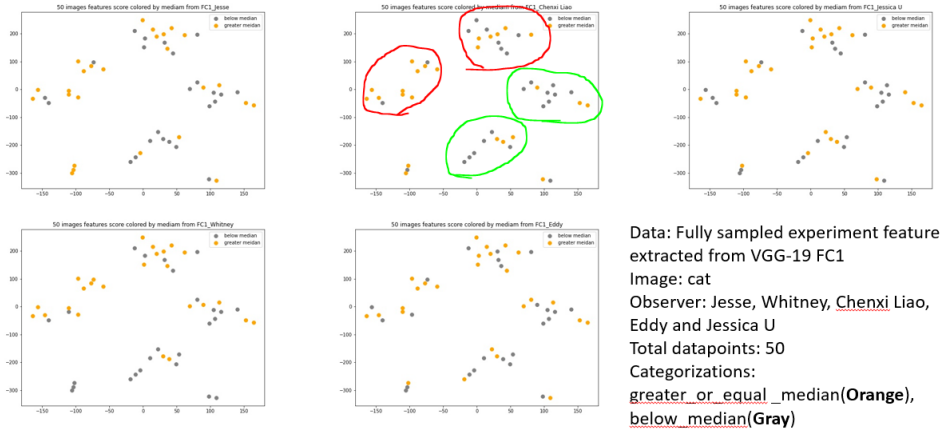
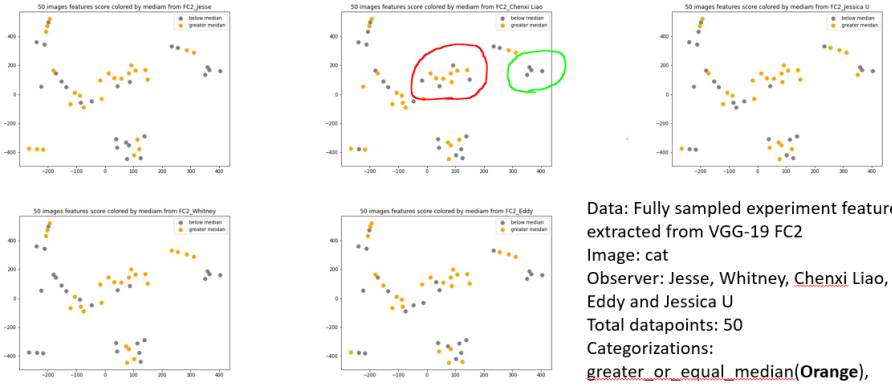


Fig 10. The tSNE plotting with features from VGG-19 fc2



Data: Fully sampled experiment feature extracted from VGG-19 FC1
Image: cat
Observer: Jesse, Whitney, Chenxi Liao, Eddy and Jessica U
Total datapoints: 50
Categorizations: greater_or_equal _median(**Orange**), below_median(**Gray**)

Fig 11. Fc1 features tSNE plotting separated by median

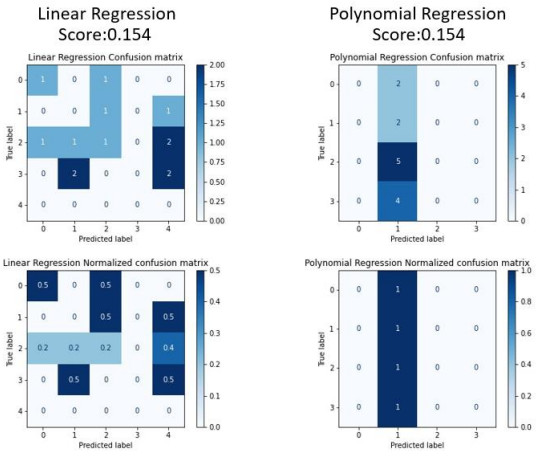


Data: Fully sampled experiment feature extracted from VGG-19 FC2
Image: cat
Observer: Jesse, Whitney, Chenxi Liao, Eddy and Jessica U
Total datapoints: 50
Categorizations: greater_or_equal_median(**Orange**), below_median(**Gray**)

Fig 12. Fc2 features tSNE plotting separated by median

- Regression result:

Here I input the fully sampled data with 5 classes as fc2 features. Unfortunately, the prediction is super low which means the regression models are unable to nicely fit into the groups. For this case, we believe that is due to the low quantity of the data.



Data: **Fully sampled experiment** feature extracted from VGG-19 **FC2**
Image: cat
Observer: Eddy
Train/test ratio: 0.75/0.25
Total datapoints: 50
Prediction: 0-10 (0),11-20 (1), 21-30 (2), 31-40 (3), 41-50 (4)

Fig 13. Regression result with fully sampled data with 5 classes

We also try to implement the regression models with the fully sampled data

separated by its median. For this case, the prediction accuracy is improved since the classes of data are down to 2. However, the prediction accuracy is still less than the regression results with random sampled experiment's data. The only thing can be considered is the quantity of the dataset. But we do think there has more to explore, like how the experiment difference affect regression result and how the samples' difference affect the result.

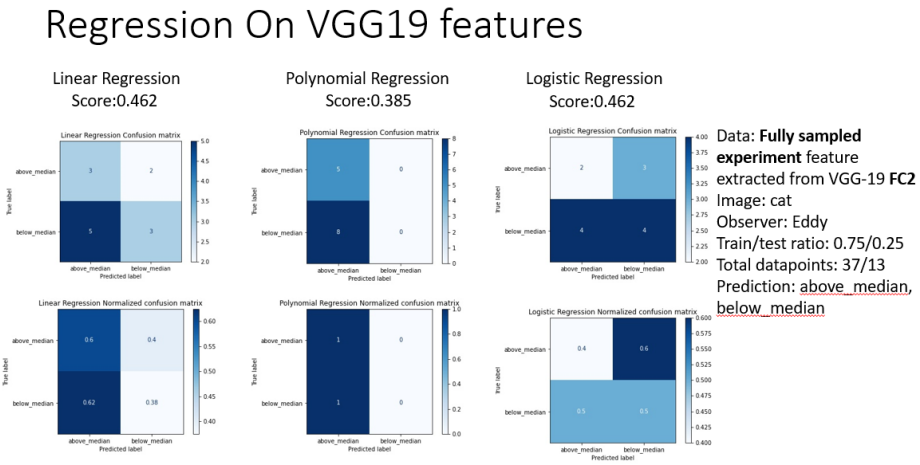


Fig 14. Regression result with fully sampled data separated with median

- Correlation and Coefficient analysis:

In this section, we present the correlation and coefficient between the result from each participant.

- Image score analysis:

We use Spearman's correlation method to analyze the score data due to its statistical characteristics is more suitable for the non-linear data such as we have in the fully sampled experiment. According to the result, we can discover that some people do have similar preferences for each other, such as Jesse, the correlation of this result with others is over 50%. However, there are some cases illustrated that they had very different tastes in cats, like Eddy and Chenxi Liao and Eddy and Whitney. However, by this test, we can tell that there has a chance to find a tendency of preference between participants.

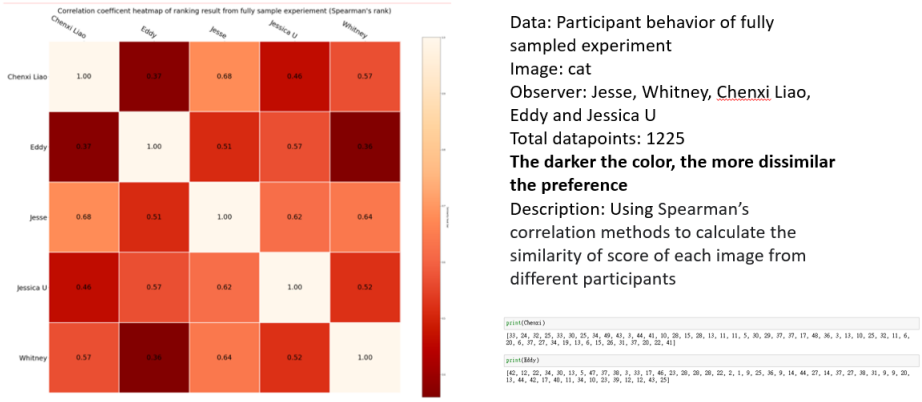


Fig 15. Correlation matrix between observers

▪ Trial-by-trial decision analysis:

We test the result of every trial decision from each participant. In this case, everyone has a significant difference of preference between pairs in the experiment. It means that everyone has their own perspective concerns while comparing the images in each trial in the experiment.

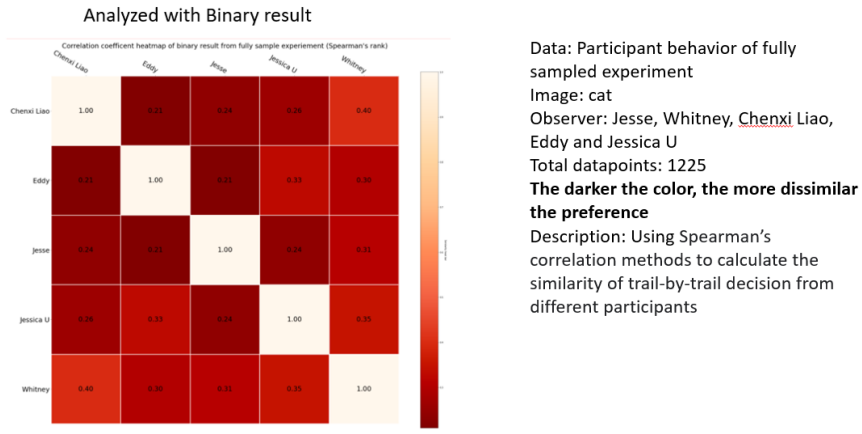


Fig 16. Coefficient matrix between observers' trial-by-trial decision

- VGG-19 Fine-tuning result:

In this part, we classified the fully sampled dataset into five groups (0-10,11-20,21-30,31-40,41-50) and two groups (data separated by median). For augmentation, we used rotation, flipping from left to right, flipping from bottom to top, random_noise and gaussian blur. For fine tuning with 5 classes, you can observe that according to the right side of Fig 17, the prediction accuracy is really low, and the validation loss doesn't even decrease. We believed that it is due to the limited dataset since we overall only have 300 for this training.

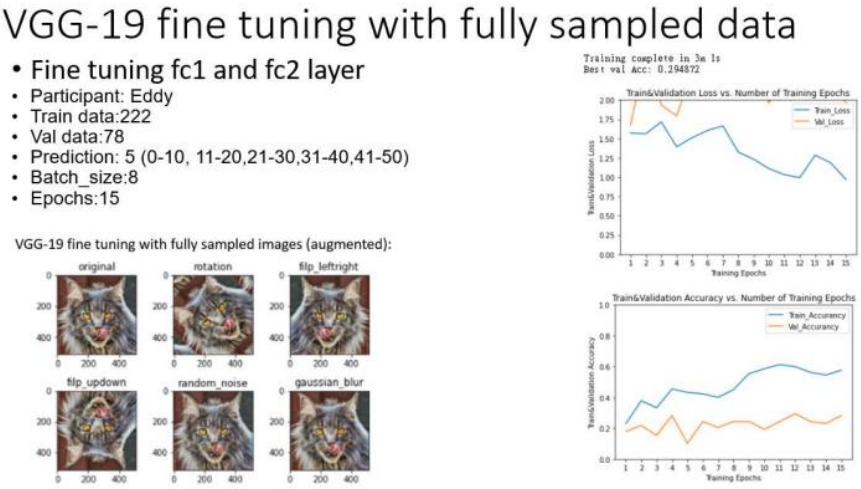


Fig 17. Vgg19 fine tuning result with fully sampled data with 5 classes

According to the right-side figure of Fig 18, the prediction accuracy is

surprisingly increased to 64.1% based on the limited dataset. In addition, for the loss function, the training and validation loss seems to have a trend to stabilize while the training. We think by classifying the fully sampled data with its own median, we may obtain a good prediction accuracy from fine tuning VGG-19. But unfortunately, fully sampled data is very hard to collect, the procedure is time-consuming but the samples we can get are few.

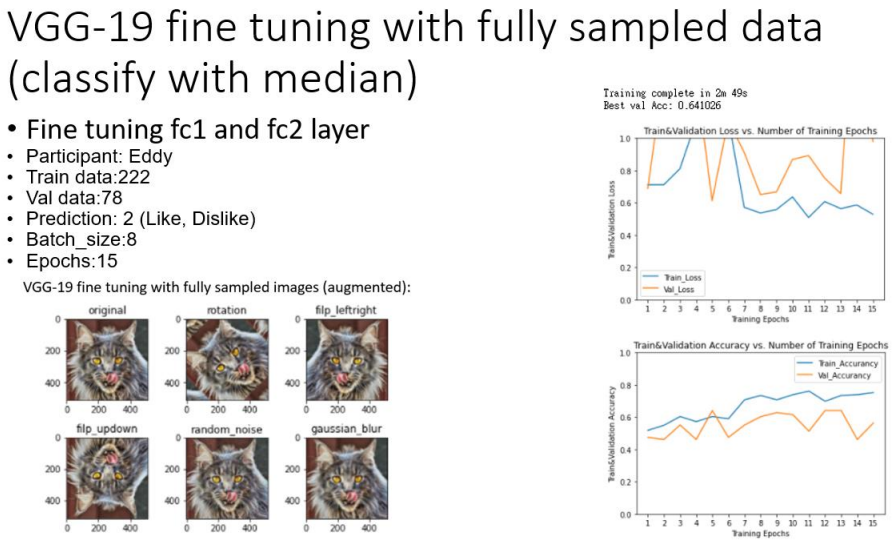


Fig 18. Vgg-19 fine tuning result with fully sampled data separated with median

Conclusion:

In this report, we have several ideas. For experiment-wise, we don't have 100% confident about which experiment is able to collect more useful data into this study. However, obtaining this two experiments' data provide different perspectives and flexibility for human perception analysis. Moreover, ever though each participant has their own preference in different images, we still have a chance to catch the attributes from the images that may affect the most of their own perception. Additionally, the result of VGG fine tuning showcases that even though the prediction accuracies are not really satisfied with all experiment's data, the model is learning with the fully sampled data which we think is because it contains clear perception of individuals, especially separated with median. Moreover, we think there is a lot of valuable information in participant trial-by-trail decisions. Thus, instead of inputting labeled data individually into the signal input machine learning model (like VGG, Restnet), implementing data pair-by-pair with labels into pair-comparison learning (like Contrastive learning) might have a good opportunity for getting surprising results.

Furthermore, we had met once with Prof. Pieter Peers. We have some conclusions about the features of this project. The first is to keep focusing on random sample experiment, extend it and fine tuin other machine learning models (like RestNet or ImageNet) with the data features extracted from VGG-19. The second is trying to classify the difference between the feature vectors of the image by referring to this paper (Zhang et al. 2018) [1]. For the paper, instead of using L2 distance, we should also try to implement a small network for computing the distance between images.

Also, there are many ideas worth exploring for this study, like how time affects a human's perception, how data augmentation techniques affect the machining learning model, and how to more precisely catch the attributes from the subjects that affect a human's perception most. In short, this project is really ambitious and requires lots of work in data collecting, analysis and management, computer vision, machine learning and some psychology-wise professionals. Nonetheless, the ultimate goal of this project is still exciting, and its application range is extremely wide. It is a project worth carrying on.

Reference List

[1] Richard, Z. & Phillip, I. & Alexei, A.E. & Eli, S. & Oliver, W. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Computer Vision and Pattern Recognition. DOI: 10.1109/CVPR.2018.00068