Personalized Perception Project

Student: Sheng Kai Liao

Date:9/9/2021

American University

Project goal:

The project is about testing how different machine learning model and different computer vision technics can predict humans' preference better with limited data and trying to find the represents from images which affect the individual's preference most. For example, we present a cat image with two different people, one may like it because it has white hair and tip ears. However, the other one may dislike it not just because the color of hair or the shape of ears but also the face of the cat. There are so many attributes which affect a person's feeling in particular subject's images (house (exterior, interior), dog, cat… etc.), and this project will try to find the trend of those attributes from different images for each person and make prediction for the observers which match their own preference.

Key words: Computer Vision, Machine learning, Personalized Perception

Experiment:

We setup 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 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 type of our experiment, one we called random sampled experiment which we randomly picked up 2000 unrepeated but same class of images (like Dog, Cat, House…. etc.) to construct the pair decision experiment with 1000 trials. In this case, we can easily obtain balanced binary result (1000 labeled data for like, 1000 labeled data for dislike). The other one is we so called Fully sampled experiment which we offer 50 unrepeated but same class of images and fully compared the 50 images with each other. Therefore, the experiment has 1225 times of trials. The result for this experiment will be decimal data for each image which we called it score, and the score is according to how many times the image be selected during the experiment by 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 provides different perspective 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 multi-dimension data into 2/3D coordinate (here we turn the features into 2D). For following figures, they illustrate the images distribution regarding the t-SNE transformed 2D coordinates.

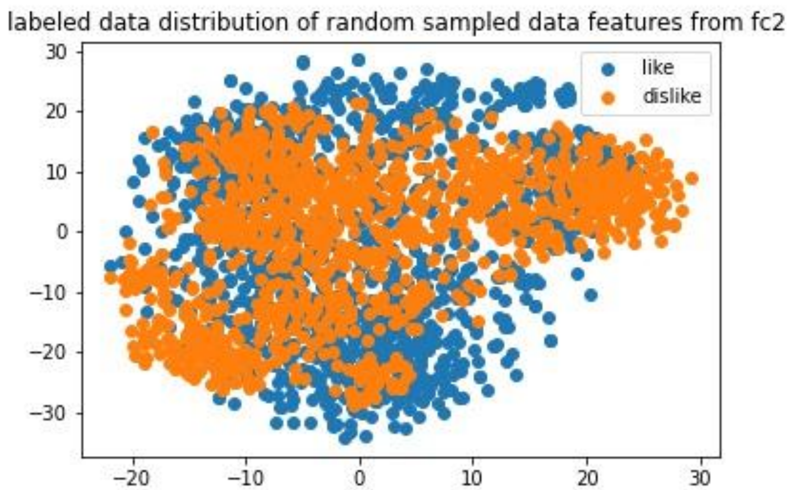


Fig 2. The labeled data distribution includes 2000 images

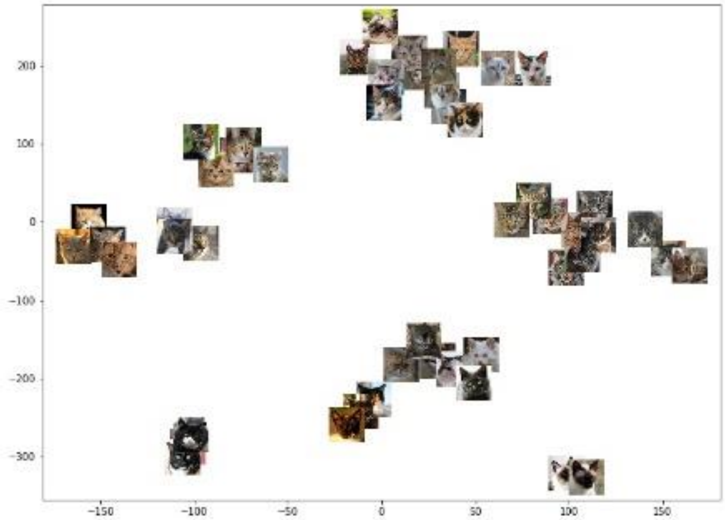


Fig 3. FC1 features tSNE plot distribution with images

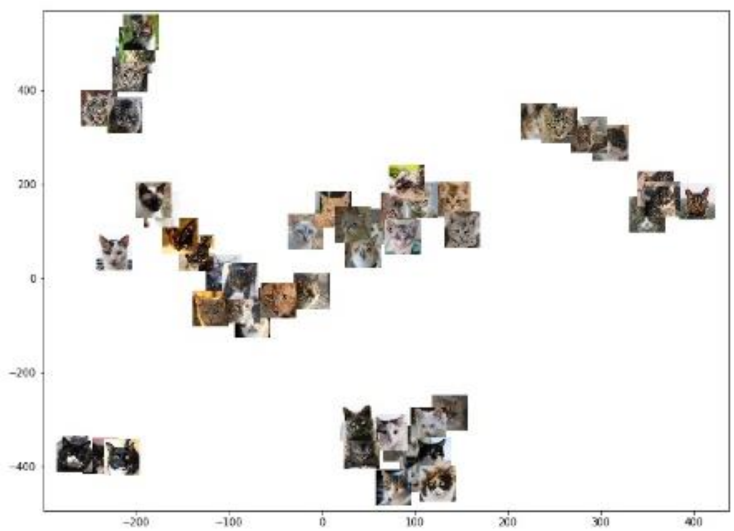


Fig 4. FC2 features tSNE plot distribution with images

Data analysis:

- Random sampled experiment:
  - Regression result:

    Our first idea for analizing the data is trying to fit them into regression model. And the result match our expected. Since there is no clear trend of the distribution (according to Fig 2.), the prediction of the models are not really good in linear regressoin model, polynomial regrssion model and logistic regression model.

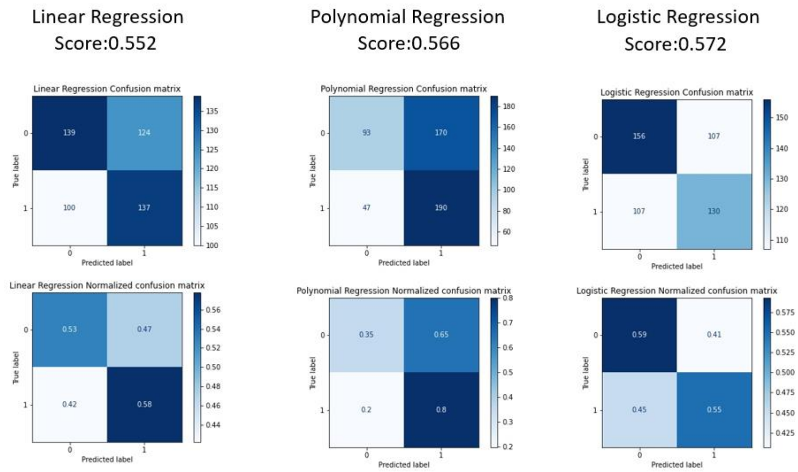


Fig 5. Regression result with random sampled data

- 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 extendtion, such as rotation, flip the mage from left to right and top and down and gussian blur. The reason why we chose this methods for augmentation is that we think it is the unlikely way to affect the participant's initial decision. For testing and validating, we splite the data into 80% and 20%. Regarding the right side charts of Fig 6, we can observer that the prediction accurancy is only 54.4% which is really low and the training loss and the validation loss while each epoch dosen't decrease continuously, which means the model isn't train well according the dataset.

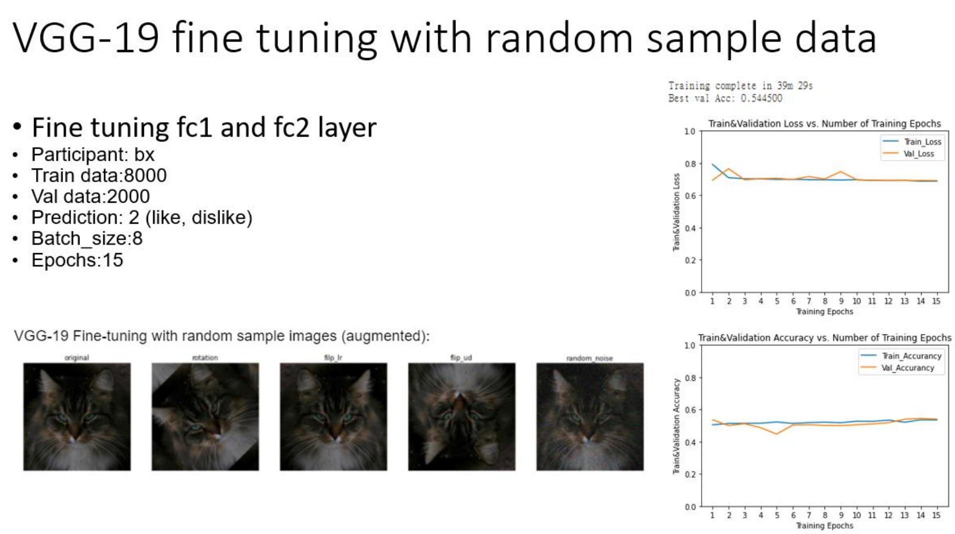


Fig. 6. Vgg-19 fine tuninf result with random sampled data

- Fully sampled experiement:

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

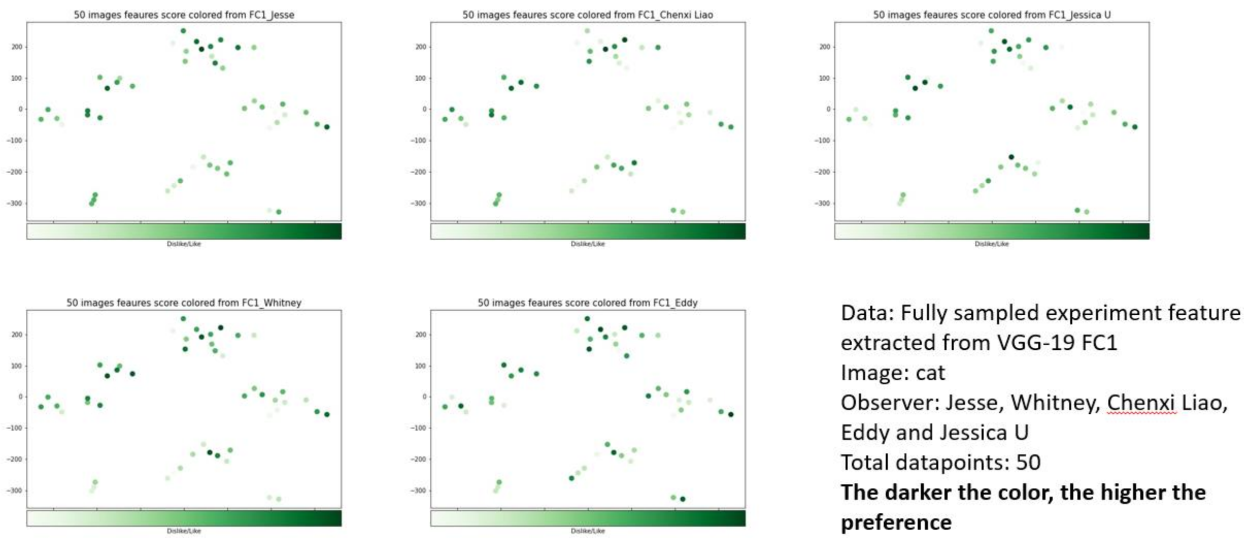


Fig 7. Fully sampled data distribution with fc1 layer's feature

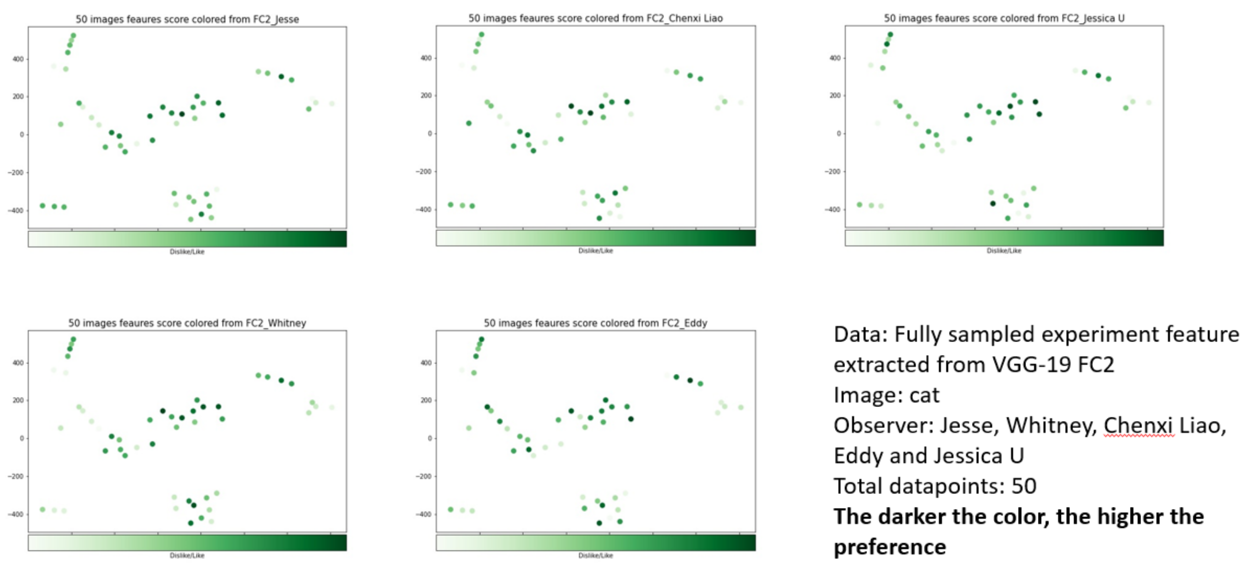


Fig 8. Fully sampled data distribution with fc2 layer's feature

In order to classify the data, we seperated the result with two ways. One is to seperate the data in 5 groups according 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 catalogrized for group discovering and also it allows us to fine tuning the VGG-19 with consistent calsses. Unfortuenly, data by classifing data into 5 classes, the datapoints in each group is inbalance and have low quantity. Therefore, we also try the other way is to devide the data into two groups based on the median of each participant's result (Fig 11. and Fig 12.). If we look at the Fig 11. You can observe that most participants like top group and the slight upper left group (the red circles), and mostly dislike the botten and the slight lower right group (the lime circles). For Fig 12., it is much easier to observe that most of participants like the center group (the red circle) and dislike the most right group (the lime circle). Even though there is many noise in the features from VGG-19 since it is trained with different purpose, but the model can still distinguish some attributes that may affect most people's perception (in this case, the definition of cuteness).

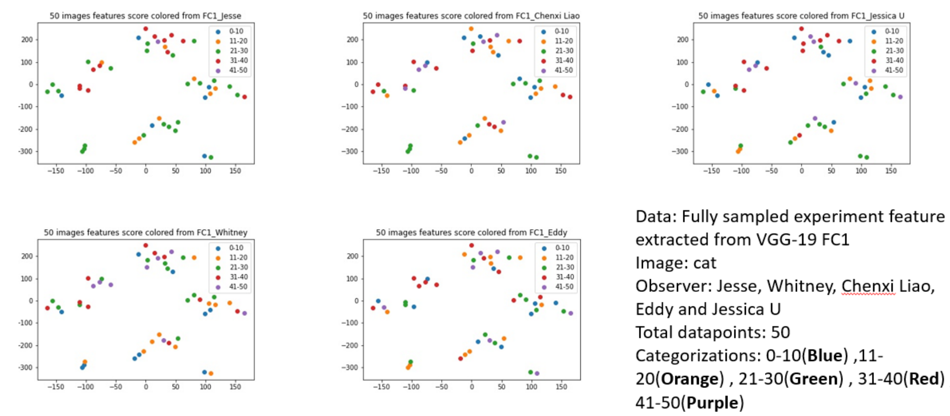


Fig 9. The tSNE plotting with features from VGG-19 fc1

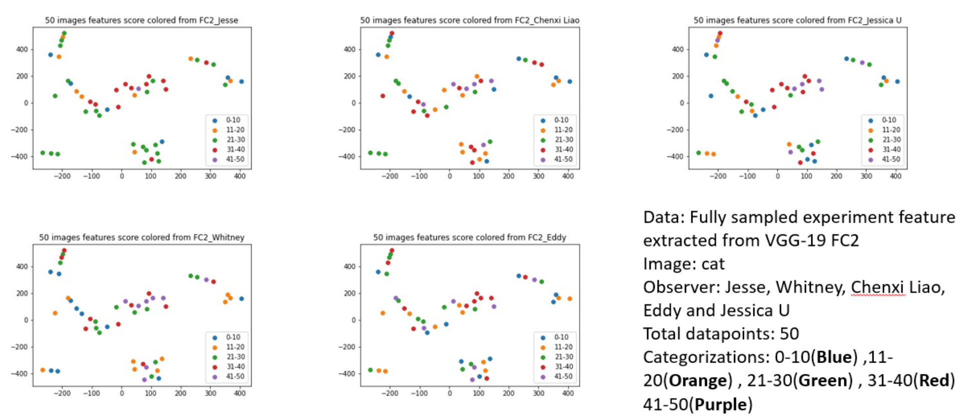
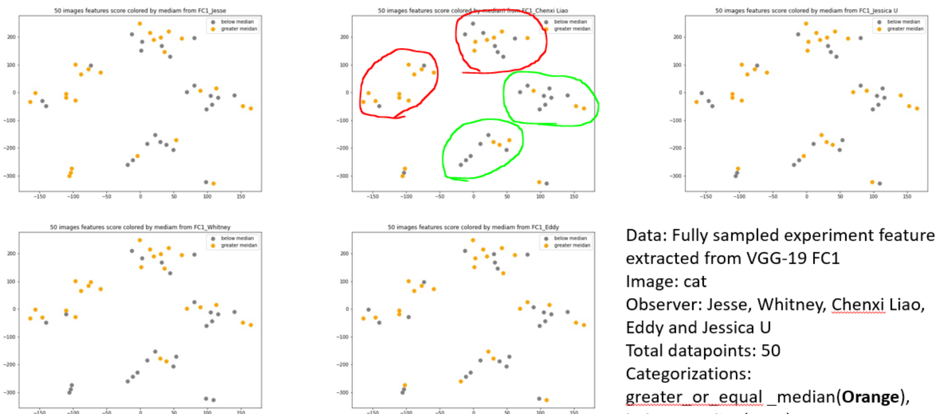
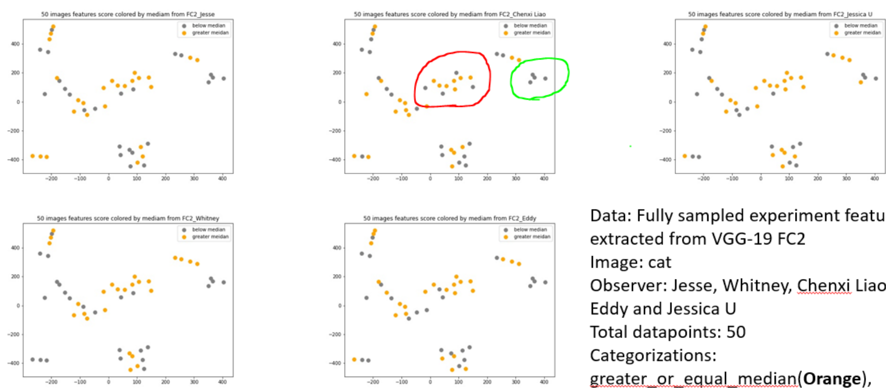


Fig 10. The tSNE plotting with features from VGG-19 fc2



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**)

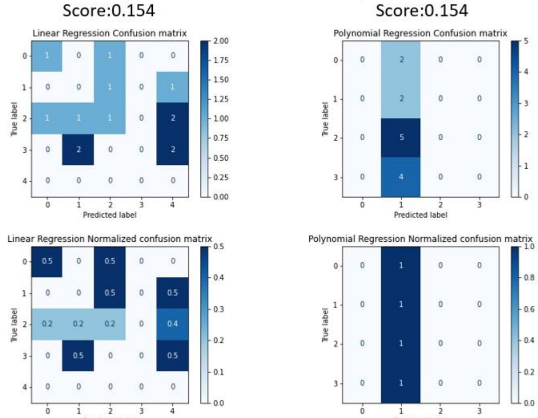
Fig 12. Fc2 features tSNE plotting separated by median

- Regression result:

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

Linear Regression
Score:0.154

Polynomial Regression
Score:0.154



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 which seperated by its median. For this case, the prediction accuency is improved since the classes of data is downed to 2. However, the prediction accurancy 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.

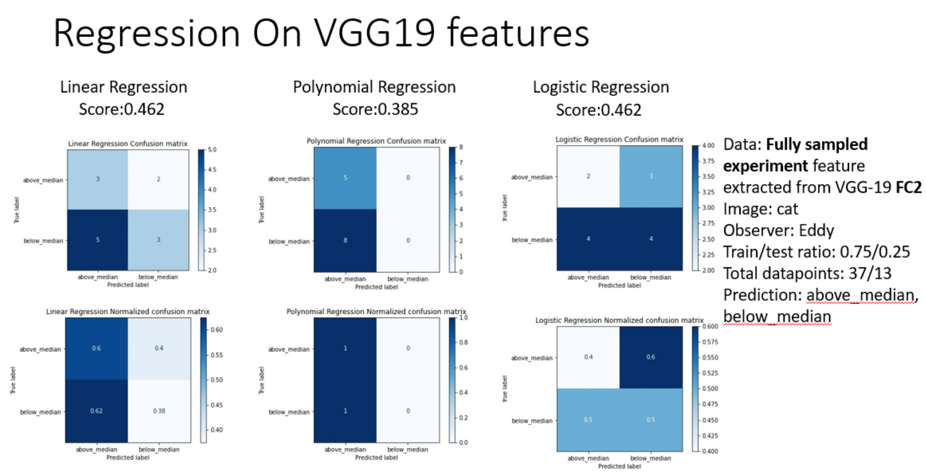


Fig 14. Regression result with fully sampled data seperated with median

- Correlation and Coefficent analysis:

In this section, we will presents the correlation and coefficent between the result from each participants.

  - Image score anaylsis:

    We use Spearman's correlation methods to anaylze the score data due to it's statistic characteristics is more suitable for the non-liner data such as we have in the fully sampled experiement. According to the result, we can discover that the some people do have similar preference for each other, such as Jesse, the correlation of this result with others all over 50%. However, there are some case illustrated that they had very different tastes on cat, 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 perference between participant.

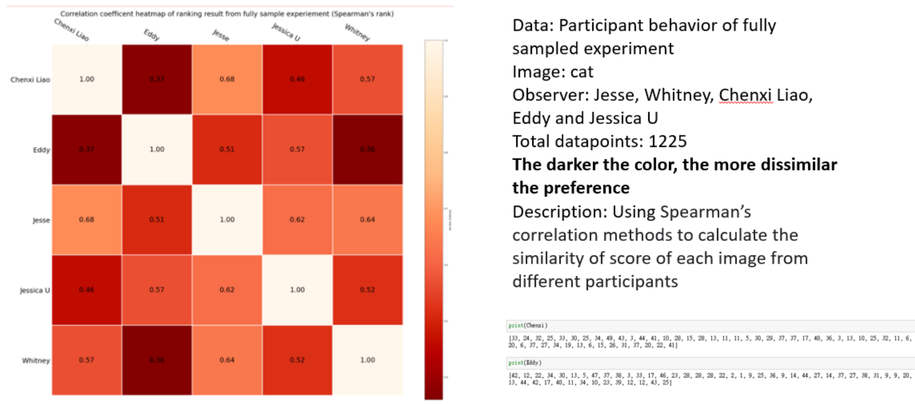


Fig 15. Correlation matrix between observers

- Trial-by-trial decision analysis:

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

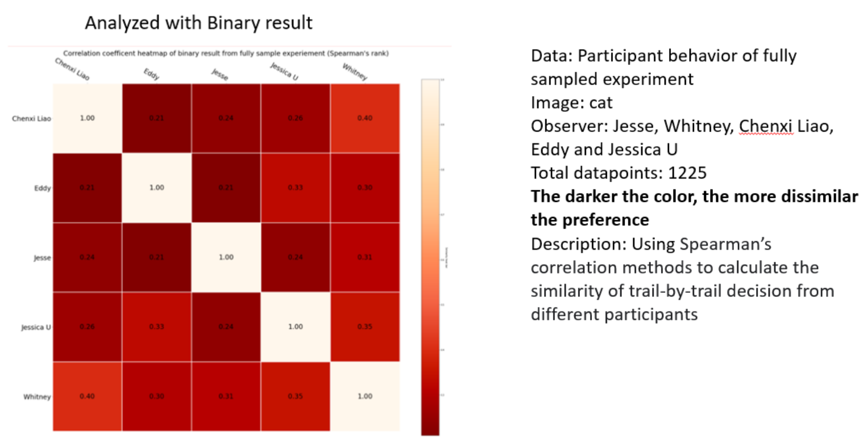


Fig 16. Coefficient matrix between observers' trial-by-trial decision

- 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 seperated by median). For augmentation, we used rotation, filping from left to right, fliping from bottom to top, random_noise and gaussian blur. For fine tuning with 5 classes, you can observe that according the right side of Fig 17, the prediction accurency is really low and the validation loss isn't even decress. We belived that it is due the the limited datatset since we overall only have 300 for this traning.

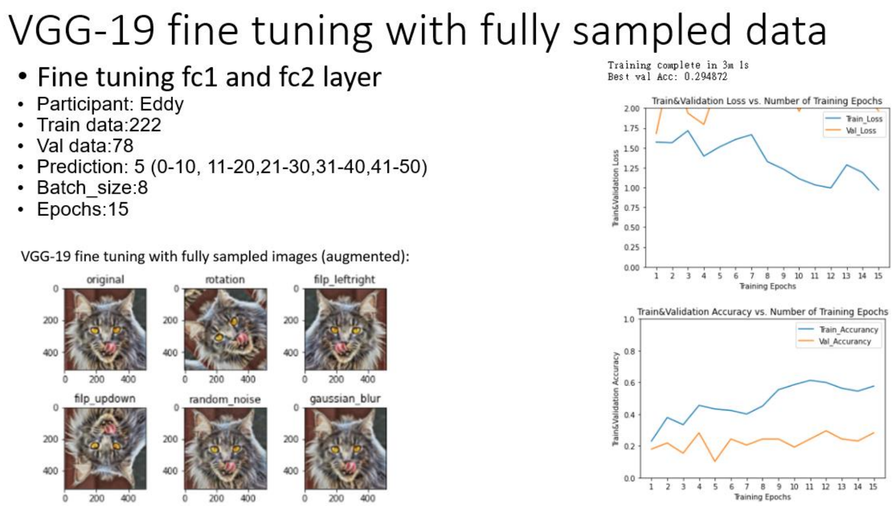


Fig 17.Vgg19 fine tuning result with fully sampled data with 5 classes

According to the right side figure of Fig 18, the prediction accurency is supprisely increased to 64.1% based on the limited dataset. In addition, for the loss function, the training and validation loss seems having a trend to stablize while the training.We think by classified the fully sampled data with its own median, we may obtain a good prediction accurency form fine tuning VGG-19. But unfortuenly, fully sampled data is very hard to collect, the precedure is time-comsuming but the samples we can get are faw.

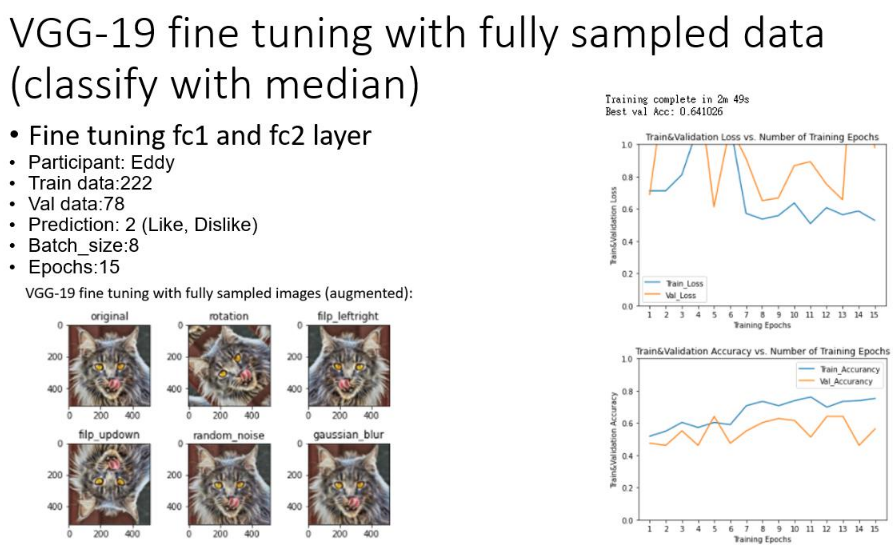


Fig 18. Vgg-19 fine tuning result with fully sampled data seperated with median

Conclusion:

In this report, we have serveral idea. 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 provids different perspectives and flexibility for human perception analysis. Moreover, ever though each participant have 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 accurencise are not really satisfied with all experiment's data, but the model is learing with the fully sampled data which we think it is because it contains cleaier perception of individual participant, especially sepeated with median. Moreover, we think there have many valueable information in participant trial-by-trail decisions. Thus, instead of inputting labeled data individually into the singal input machine learning model (like VGG, Restnet), implementing data pair-by-pair with labels into pair-comparision learning (like Contrastive learning) might have good opportunity for getting surprised result.

Futhermore, we had met once with Prof. Pieter Peers. We have some conclusion about the feautrue of this project. The first is keep focusing on random sample experment, extend it and fine tuin others macine learning model (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 refering 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 has many idea worth to be exlored for this study, like how the time affect a human's perception, how data augmentation technic affect the machining learning model, and how to more precisely catch the attributes from the subjects that affect a human's preception 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 prefessionals. Nonetheless, the ultimate goal of this project is still exciting, and its application range is extremely wide. It is the project worth for 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