CSC-680 Intro of Data Mining

Final Project:
Personalized Perception: Cat Face Cuteness Preference Prediction with LPIPS

Student: Sheng Kai Liao

Supervisor: Dr.Xiao

Fall 2021

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, cuteness is a personalized perception, and everyone has their own criteria. Some people might prefer small cats with fluffy hair and others might prefer bigger cats with tip 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 by using machine learning? In order to achieve this goal, we constructed two types of experiments, both will present two cat images in each trial and the observer would pick one that he or she thinks is cuter than the other one. Instead of analyzing human intuition of each image, we believe that by offering one comparison, human perception can be more possible to be analyzed by focusing on two image differences. According to Zhang et al., they proposed Learned Perceptual Image Patch Similarity (LPIPS) matrix which is able to measure the similarity of two images in a way that conforms to human judgment. In other words, by comparing two images, LPIPS calculate human perceptual distance between two images which helps machine learning classifiers accomplish human judgment predictions. Therefore, inspired by their work, we assume that by using LPIPS, we have a chance to capture human-like perception from two images in order to train the machine learning model to predict one person's preference. In this research, we extract the embeddings of the difference between two images and train those embeddings with a simple CNN model in order to predict one person's preference of cuteness perception of cats.

Key words: Computer Vision, LPIPS, Machine learning, Personalized Perception, Cat

Cat Face 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 think is cuter than the other one. The format of our experiments are shown below:

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

As I mentioned before, we create two types of experiments. One is what we call the random sampled experiment. In this experiment, the image samples are randomly picked from our cat face dataset, and 2000 unrepeated cat images are used in this experiment. The purpose for this experiment is to collect the cuteness perception preference from the images that the person has never seen before. In this case, 1000 perceptual distance data would be created by implementing LPIPS with labels according to the number of experiment trials. The label represented by 0 and 1, 1 means in one trial, the observer chooses the right one and 0 means the observer chooses the left one. In further research, this experiment has been extended into 2000 trials as a bigger dataset for training.

The other one is a so-called fully sampled experiment in which we gathered 50 unrepeated cat images, and all images will fully compare to each other. The reason for designing this experiment is that we believed by having a full comparison with limited images, the observer could be able to provide more comprehensive perception information of cuteness. In this experiment, we will have 1225 perceptual distance data from each trial with labels.

What is LPIPS:

LPIPS is a machine learning algorithm that is used to calculate the perceptual distance which is similar to the human perspective from two images. According to Zhang et al., LPIPS model will firstly extract the features of image from pre-trained models which are implemented in visual reorganization, such as VGG-16, Alexnet and SqueezeNet. Afterward, regarding Fig2, five convolutional layers are added after the model (seven convolutional layers are added in SqueezeNet case) in order to normalize the activations in channel dimension. As the default, perceptual distance would be represented as one value representing the average L2 norm distance across the layers. However, in order to obtain more information about the difference between input images, LPIPS allows users to extract the average spatial distance data across layers and the spatial distance data from each layer. Hence, in my research, I implemented those two kinds of data I mentioned above as training data with a simple CNN models in order to predict one person's cuteness preference from cat images.

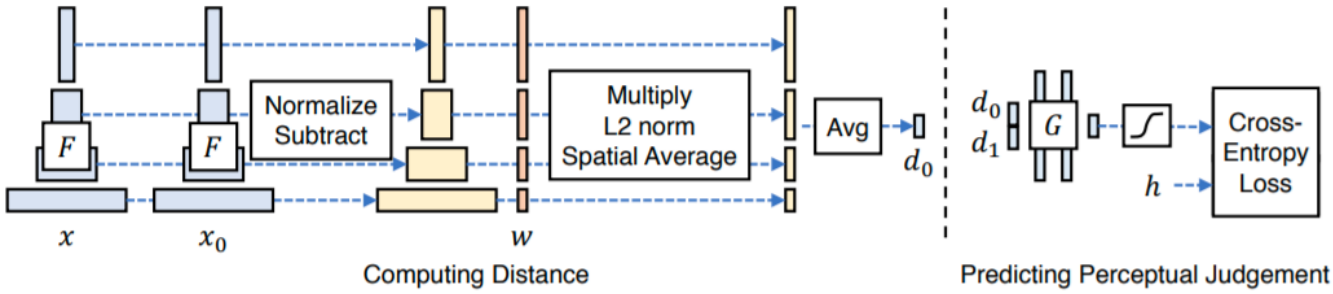


Fig2. Illustration of the research idea from Zhang et al. The left figure showcases the calculation of the average L2 norm distance value across all layers. And the right figure means a small network G is trained to predict perceptual judgment h from distance pair (d0, d1).

Embedding extraction from LPIPS:

For each experiment trial, we input the two images into LPIPS in order to obtain the average spatial distance data across layers of the inputs. All of input images are 512*512*3. And the spatial distance data is 512*512*1. An example is shown below (Fig 3): The top two images are the images within one trial and the below green image represents the average spatial distance data across layers from LPIPS. In this research, the average special distance data used for my CNN model.

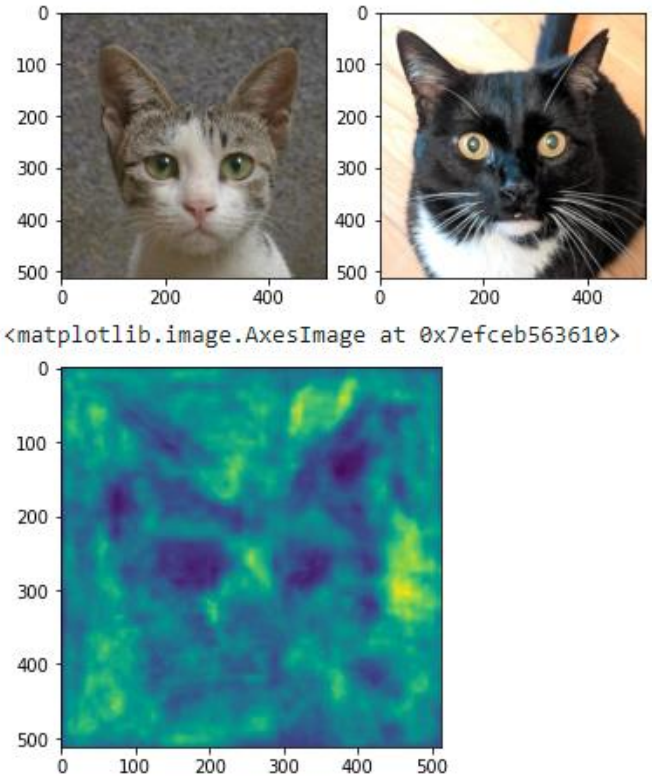


Fig 3. The examples of input images and the average spatial distance data across each layer from LPIPS for one trial

Furthermore, as I mentioned, LPIPS allows users to extract the spatial distance data from each layer. The outputs are shown in Fig 4, the input images are the same as Fig 3. As you can see, each layer focuses on different details of the difference of the images. In our research, we also use the embeddings extract from the first layer since we think that human consider much detailed information from the cat faces in order to determine their feeling of cuteness.

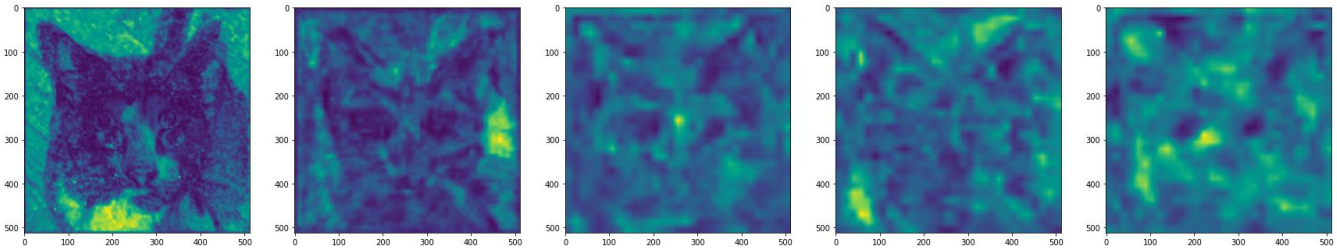


Fig 4.The examples of the spatial distance data from each layer of LPIPS for one trial

The CNN model,

In this research, I have created a simple CNN model with three convolutional layers which has 32,64 and 128 filter size respectively and the kernel settings are all 3*3, and one maxpool layers with 2*2 pooling size come after each convolutional layer. Consequently, I add a flatten layer to integrate the information we have and add three fully connected hidden layers with 128,64,32 hidden nodes with relu activation. Finally, an output layer is added with 2 nodes to represent the left/right choice prediction.

```
from keras.models import Sequential
from keras.layers import Dense,Dropout,Flatten,Conv2D,MaxPooling2D
model = Sequential()

model.add(Conv2D(32,kernel_size=(3,3),input_shape=(512,512,1)))
model.add(MaxPooling2D((2,2)))
model.add(Conv2D(64,kernel_size=(3,3)))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Conv2D(128,kernel_size=(3,3)))
model.add(MaxPooling2D(pool_size=(2,2)))
#now flaten the layer and aldd hiden layers
model.add(Flatten())
# First hiden layer:
model.add(Dense(128,activation = 'relu'))
model.add(Dropout(0.2))
# Second hiden layer:
model.add(Dense(64,activation = 'relu'))
model.add(Dropout(0.2))
# Third hiden layer:
model.add(Dense(32,activation = 'relu'))
model.add(Dropout(0.2))
# Output layer:
model.add(Dense(2,activation = 'softmax'))
```

Fig 5. CNN model architecture

Experiment:

The purpose of the experiment is to train the CNN model to mimic an observer to make the decision for the cat face experiment. To achieve it, the pair image of each trial would be inputted into LPIPS in order to calculate the spatial distance data and the data would be labeled with the observer behavior data for each trial. For the experiment, I trained the CNN model from two kinds of experiment data labeled with my behavior data.The datasets are contracted by the average spatial distance data across layers from LPIPS of random sampled experiment and the fully sampled experiment and the first layer spatial distance data from LPIPS of the fully sampled experiment.

In the first experiment, we used the average spatial distance data from LPIPS from each trial in the random sampled experiment for training. The overall dataset contains 1000 data points with 1000 labels which represent 0 and 1. 0 means in the trial, the observer thinks the left one is cuter than the right one. 1 means the opposite. 80% of the dataset was used for training and 20% of the dataset was used for testing. By observing the trend of loss function and accuracy (Fig 6), the loss function didn't decrease following the epochs and the validation accuracy did not increase as expected. Also, according to the prediction, the model only has one predict in all testing samples. Therefore, we think the model is not

able to distinguish the major difference between training data in order to make the prediction.

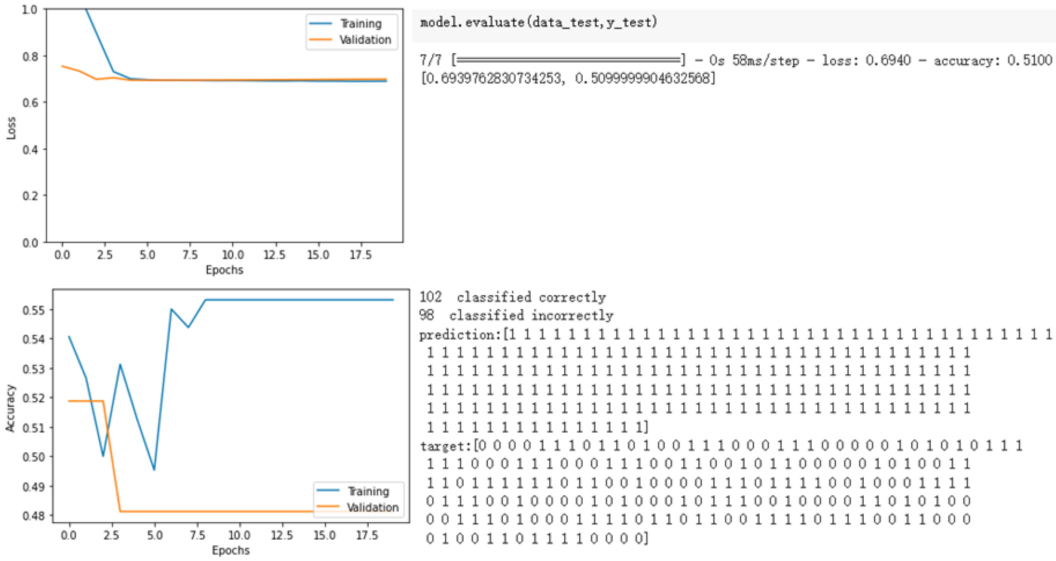


Fig 6. Loss and accuracy history diagram for the CNN model trained with 1000 data points from random sampled experiment

At first thought, we believed it was because there was not enough data for the CNN model to catch the insight of the data. Thus, we extended the random sampled experiment into 200 trials and used it for training. As you can see at Fig 7, the loss function is held in some point of epochs which means the model can't not find the further information from the dataset to achieve the goal. In addition, the model is still predicting 1 in all testing samples. Therefore, by this experiment, we have an assumption that maybe the random sampled experiment can't deliver the information of the person's perception clearly since there is no consistent comparison between samples. Regarding the conclusion, we think that the fully sampled experiment data would benefit the model.

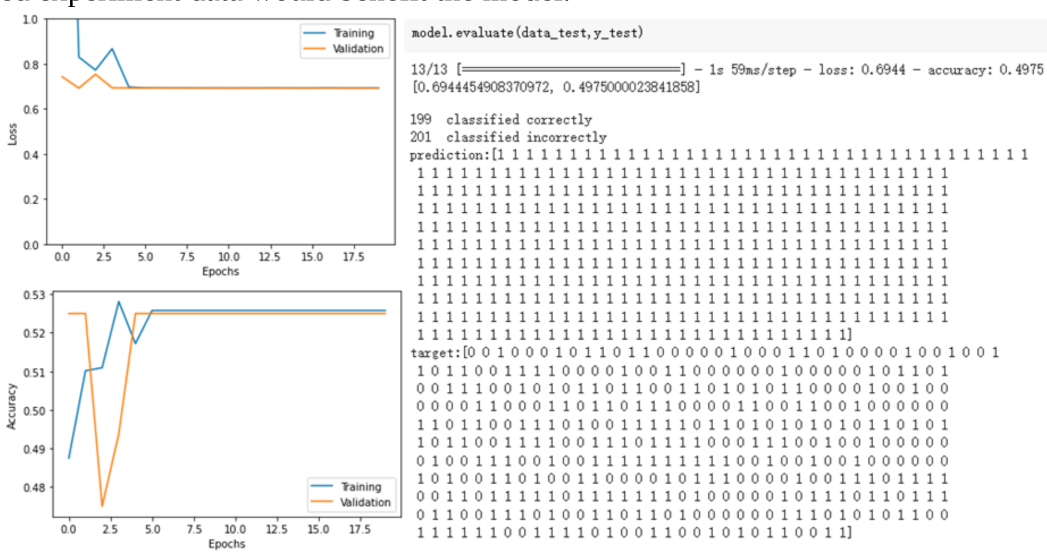
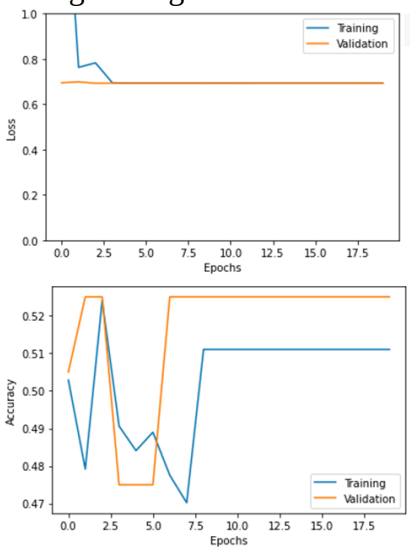


Fig 7. Loss and accuracy history diagram for the CNN model trained with 2000 data points from random sampled experiment

Thus, for our second experiment, we used the average spatial distance data of the fully sampled experiment which has 1225 data points due to 1225 experiment trials and the labels according to my behavior data for each trial. Unfortunately, according to Fig 8, you can see that the loss function is held at some point of epochs and the model keeps predicting the same prediction as the result of when we

trained the model with 2000 data points from random sampled experiment. At this point, we assume that in order to determine the cuteness preference of cat image, observers may consider more detailed information such as face contour, eye color or ear shape. Hence, we decided to use the spatial distance data from the first layer of LPIPS for training since it contains more detailed information, you may refer to the botten image of Fig 3 and the first image of Fig 4.



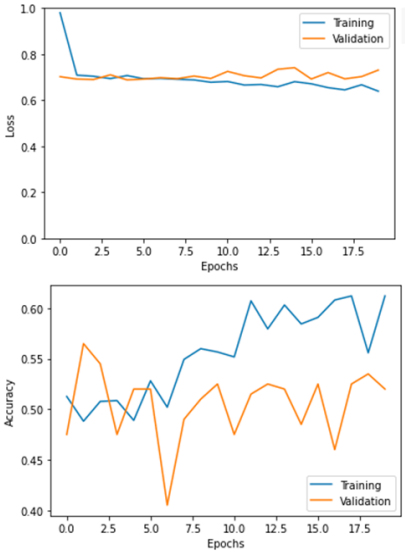

```
model.evaluate(data_val,y_val)

4/4 [==============================] - 0s 65ms/step - loss: 0.6916 - accuracy: 0.5700
[0.6915742754936218, 0.5699999928474426]

57  classified correctly
43  classified incorrectly
prediction:[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
target:[1 1 1 0 0 1 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 1 0 1 0
 1 0 0 1 1 1 1 1 1 0 1 1 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 0 1
 1 0 0 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 0 1 0 0 0 1]
```

Fig 8. Loss and accuracy history diagram for the CNN model trained with 1225 data points from fully sampled experiment

For the third experiment, we trained the CNN model with the spatial distance data from the first layer of LPIPS from the fully sampled experiment. Surprisingly, the model starts learning from the dataset and making predictions from testing samples. According to Fig 9, even though the loss function does not have significant change, it does decrease simultaneously following the epochs. In addition, the training and validation accuracy keep increasing while the training process. Therefore, we think the model is able to learn the information from the spatial distance data from the first layer of LPIPS in order to predict the observer preference of cat cuteness.




```
model.evaluate(data_val,y_val)

4/4 [==============================] - 1s 62ms/step - loss: 0.7241 - accuracy: 0.5000
[0.7241355180740356, 0.5]

50  classified correctly
50  classified incorrectly
prediction:[1 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 0 1 0 1 1 1 0 1 0 1 1 1 0
 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0
 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 1 0 1 0 0]
target:[1 1 1 0 0 1 0 1 1 1 1 0 1 0 1 1 1 0 1 1 1 1 1 0 1 1 0 0 0 0 0 0 1 0 1 0
 1 0 0 1 1 1 1 1 1 0 1 1 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 0 1
 1 0 0 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 0 1 0 0 0 1]
```

Fig 9. Loss and accuracy history diagram for the CNN model trained with 1225 data points from fully sampled experiment

Conclusion:

According to my research, we discover that by implementing the spatial distance data from the first layer of LPIPS, our CNN model is able to catch the insight from the data that may affect the observer's preference of cat most and make predictions from different samples. However, the performance of the model is still low. Therefore, in order to improve the performance of the model, some strategies are considered to be implemented in further research. For example, extend the size of fully sampled experiments in order to collect more data. Furthermore, data augmentation is another way to extend the dataset we have. But one thing needs to be taken into consideration is that the data augmentation image should not be able to affect the observer's perception preference. Finally, we believe that by implementing a pre-trained model which trained for cat face reorganization relevant purpose into LPIPS, we should be able to get the perceptual distance from LPIPS which reflects more details of cat face, and with more details of information, machine learning model should be easier to catch the tendency of the observer perceptual preference of cat face cuteness.

Reference List

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