Enhancing Emotion Recognition in AI: A CLIP-Dissect Approach

Joon Cha jhc123@ucsd.edu
Sungjin Choi suc074@ucsd.edu
Jasmine Lo j2lo@ucsd.edu
Hieu Luu hhluu@ucsd.edu

Lily Weng lweng@ucsd.edu

Abstract

This paper looks at how CLIP-Dissect can make emotion recognition systems better. Currently, the best method to classify emotions based on an image is with deep learning, a rapidly growing field with state of the art performance in visual tasks. However, it is unclear what is happening within deep learning models that leads to such strong performance, often being nicknamed a "black box." With the development of CLIP-Dissect, which is a tool to easily interpret the role of deep neural network neurons, we seek to understand how a model classifies emotions and how to improve its performance by manipulating neurons with specific functions.

Code: https://github.com/kier0813/emotion_clip

1 Introduction

CLIP-Dissect (Oikarinen and Weng 2022) is a method derived from the principles of Network Dissect but tailored for OpenAI's CLIP model, which is designed to concurrently process and comprehend both image and text information through extensive contrastive learning techniques. The fundamental aim of CLIP-Dissect is to delve into and elucidate how CLIP functions internally, specifically examining how its individual neurons react to mixed inputs of images and texts. This exploration uncovers the 'multimodal neurons' within the model that activate in response to intricate combinations of visual and textual stimuli. This analytical technique is crucial for gaining a deeper understanding of the complex mechanisms of CLIP, showing how it embodies and conveys a wide variety of abstract ideas that connect visual and verbal elements. By using CLIP-Dissect, researchers can highlight the effectiveness and adaptability of multimodal neural networks in performing a broad spectrum of cognitive functions.

We seek to take advantage of CLIP-Dissect's ability to annotate internal neurons with ambiguous concepts in order to expand Dr. Lily Weng's project for emotion detection. We will improve emotion recognition systems by improving the interpretability of hidden neurons in vision networks by assigning emotional labels to them. This method takes advantage of CLIP-Dissect's adaptability, efficiency, and model agnosticism, providing a promising path for improving and developing emotion detection in the context of our project.

By giving emotional and facial labels to hidden neurons in vision networks, we can make these neurons easier to understand and use in emotion recognition systems. This approach offers a viable route forward for enhancing and expanding emotion detection within the framework of our project.

2 Data

2.1 Probing Data

The FER2013 dataset is a comprehensive collection of facial expression images, pivotal for advancing research in emotion recognition. This dataset comprises approximately 30,000 grayscale images, each standardized to a resolution of 48x48 pixels. These images encapsulate the complexity and variance inherent in human facial expressions, making it an invaluable resource for training and evaluating machine learning models in the field of computer vision and affective computing. Some examples are shown in Figure 1.

- FER2013
  - Description: The data consists of 48x48 pixel grayscale images of faces. The

Figure 1: FER2013 example images

faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.

- Task: Categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).
- Training Set: 28,709 examples.
- Public Test Set: 3,589 examples.
- Link: FER2013

2.2 Concept Set

Two concept sets were used: (1) the seven expression categories as above, and (2) a facial expression emotion concept set with twenty five concepts created with GPT-4 and based on those seven emotions.

- FER2013 Classes - 7 concepts
  - Angry
  - Disgust
  - Fear
  - Happy
  - Sad
  - Surprise
  - Neutral

- Facial Expressions - 25 concepts
  - Furrowed Brows
  - Raised Brows
  - Lowered Brows
  - Drooping Brows
  - Relaxed Brows
  - Wide Open Eyes
  - Narrowed Eyes
  - Twinkling or Crinkled Eyes
  - Tearful Eyes

- Closed Eyes
- Relaxed Eyes
- Tense Mouth and Jaw
- Relaxed Mouth and Jaw
- Smile
- Frown
- Downturned Mouth
- Slightly Open Mouth
- Wide Open Mouth
- Slight Chin Raise
- Slightly Dropped Jaw
- Flared Nostrils
- Tightened Facial Muscles
- Compressed Lips
- Elevated Upper Eyelids
- Relaxed Facial Muscles

3 Methods

In our project, we employed a comprehensive approach to understand and interpret facial expressions through machine learning models, focusing on the Facial Expression Recognition 2013 (FER-2013) dataset. The methodology involved several key steps, beginning with obtaining a baseline model, analyzing important features for emotion through dissection, and modifying the network. The baseline model we used is a VGGNet architecture (Figure 2) trained on FER2013 (Khaireddin and Chen 2021), with a test accuracy of 69.42%. We seek to analyze the features being used in the model to understand how the model works. Our goal with modifying the network is to improve accuracy.

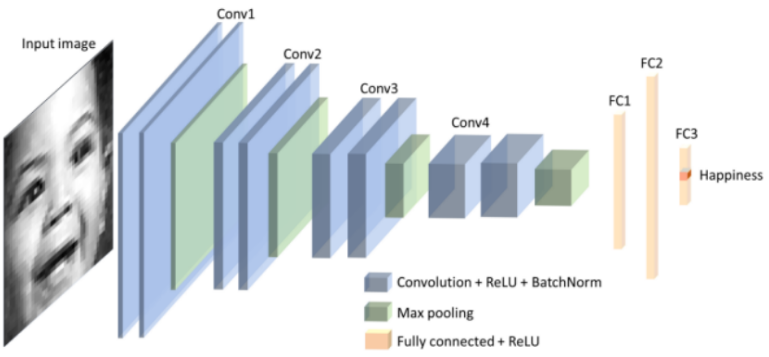

Figure 2: VGGNet architecture.

3.1 Data Processing

Within the FER2013 dataset, there were corrupted images, as shown in Figure 3. We ended up removing 42 of them.

Figure 3: Corrupted images in FER2013.

3.2 Part I: Emotion Feature Analysis with Neural Network Dissection

Our first step was to generate and label our new concept set with GPT4. This required asking several questions about emotions and the human face such as, ”What facial features are important in determining human emotion?”. Next, we applied CLIP-Dissect with FER2013 as our probing set and with the new concept set, hoping to understand what features are being used in the network and what neurons are activated by them. We calculated the accuracy of the neuron descriptions by seeing if the concepts matched the images that highly activated those neurons. Taking only the results that were predicted accurate, we counted the sampled concepts that were used and they were ready for visualization and analysis.

3.3 Part II: Neuron Dissection and Modification

As part of the network dissection performed in procedure I, we obtain the max similarity value between each neuron activation and concept label. The max similarity of 4096 neurons of layer ”lin2” are shown in Figure 4.

We choose the similarity cutoff tau = 0 to determine the set of interpretable neurons. We then choose the layer with the most interpretable neurons to modify. Data about layer interpretability can be seen in Table 1.

To modify the layer, we multiply the interpretable neuron weights by 1.5 and divide the uninterpretable neuron weights by 1.5. After modifying the layer weights, we re-evaluate the model on the test data.

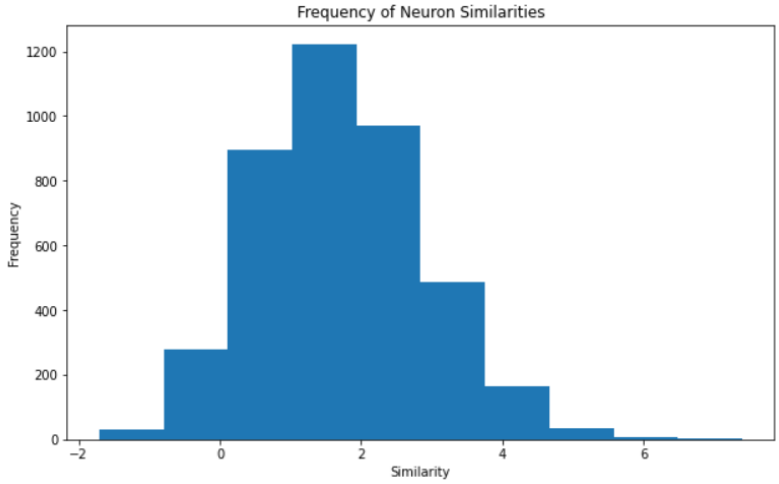

Figure 4: Distribution of neuron max similarities of layer ”lin2”.

Table 1: Interpretability Data

Layer	num_interp	num_uninterp	overall_acc	interp_acc	uninterp_acc	prop_interp
conv1a	64	0	0.22	0.22	0.00	1.00
conv1b	64	0	0.30	0.30	0.00	1.00
conv2a	127	1	0.21	0.21	0.30	0.99
conv2b	127	1	0.28	0.27	0.90	0.99
conv3a	254	2	0.26	0.26	0.05	0.99
conv3b	255	1	0.23	0.23	0.30	1.00
conv4a	492	20	0.28	0.29	0.21	0.96
conv4b	494	18	0.28	0.28	0.32	0.96
lin1	3489	607	0.36	0.38	0.25	0.85
lin2	3835	261	0.43	0.45	0.24	0.94

4 Results

4.1 Part I: Emotion Feature Analysis with Neural Network Dissection

We were able to use VGGnet and apply CLIP-dissect. After the application, we generated another concept set that includes more detailed facial expression for further analysis of how our machine makes its decision. Taking a portion of sample from dissected lin2 layer with 4096 neurons, we picked only the facial expression that were predicted correct and get results as shown in the visualization in Figure 5.

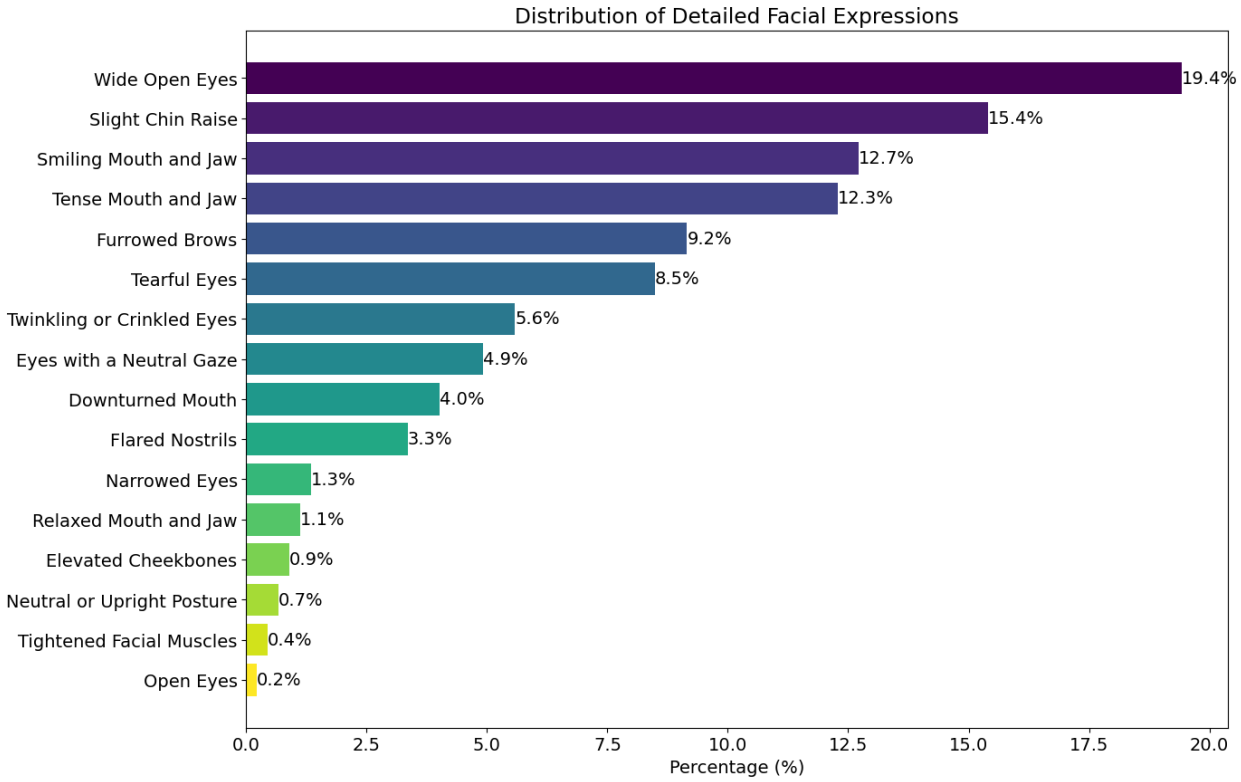

Figure 5: Distribution of concept labels.

4.2 Part II: Neuron Dissection and Modification

We analyzed neurons in the lin2 layer of a VGG-like network to distinguish between interpretable and uninterpretable ones based on the sum of absolute values of their weights. Using 0.0 as the similarity cutoff, we identified neurons above this threshold as interpretable and those below as uninterpretable. We modified the weights by multiplying those of interpretable neurons and dividing those of uninterpretable neurons by a factor of 1.5. This modification resulted in a slight increase in model accuracy (about 0.21%), demonstrating

that such targeted weight adjustments can potentially influence model performance.

5 Discussion

In our work, we analyzed the dissected neuron labels with an emotion concept set generated by GPT-4. We looked into which concept labels and images matched up best at each neuron. From 4096 neurons, we filtered out the correct predictions made by the model, and we were able to get distribution of facial expressions. Concepts such as “Wide Open Eyes” recorded the highest percentage with 19.4%, suggesting that such concept is effective for our model to predict class “surprise”. The next four highest rated concepts like ”Smiling Mouth and Jaw” helped predict the ”happy” class, ”Tense Mouth and Jaw” helped predict the ”anger” and ”disgust” classes, ”Slight Chin Raise” helped predict the ”surprise” and ”neutral” classes, and ”Furrowed Brows” helped predict the ”anger” class. By adding this set of more descriptive concepts, we better understood how the vision network was classifying the images in the probing dataset, which classes these concepts were predicting, and which neurons were most useful in this process.

Although we figured out a process to increase model interpretability and accuracy through looking at what intermediate level neurons were classifying, we need to be fully aware of what potential this holds for the future.

The classified and more interpretable networks for emotion recognition from computer vision could be dangerous on some points. One of our concerns was potential threats for privacy due to the better recognition of faces that comes along with being able to better classify emotions one is showing. When using data such as faces that belong to people, it’s important to remember to respect and maintain privacy. Always make sure consent is given to use someone’s face. We don’t want our model to be used to classify anything that it does not have permission to, especially in private company use where we don’t know what results will be used for.

On the other side, we do believe that the improved models and interpretation could lead to development in virtual therapy applications, and education purposes for people with facial/emotion recognition disorders. Due to the recent rise in therapy and healthcare powered by AI, our method of increasing accuracy in classifying emotions can enhance patients’ experiences with these platforms.

6 Conclusion

Utilizing CLIP-Dissect, we identified which neurons in our vision model played key roles in predicting image classification. This process allowed us to pinpoint at which neurons images and concept labels matched up more often. Through automated labeling, we easily identified the neurons whose weights we wanted to modify.

Through measuring interpretability at each of these intermediate level neurons, we chose which weights to update to increase the accuracy of the model. We wanted to place more importance on these neurons that could predict the image class better. In our model, "good" neurons' weights were multiplied by a factor of 1.5 and "bad" neurons' weights were divided by a factor of 1.5. Through this strategy, we increased the model accuracy from 69.420% to 69.629%.

Our process also identified which concept labels yielded the most interpretable results. This uncovered a little about how our deep neural network made classifications, and what aspects of the face were used to determine the final classification. Knowing these concepts ultimately helps us better understand the decisions of neural network architectures.

References

Khaireddin, Yousif, and Zhuofa Chen. 2021. "Facial emotion recognition: State of the art performance on FER2013." *arXiv preprint arXiv:2105.03588*

Oikarinen, Tuomas, and Tsui-Wei Weng. 2022. "Clip-dissect: Automatic description of neuron representations in deep vision networks." *arXiv preprint arXiv:2204.10965*