Evaluating CNN Layers Given Emotional Stimuli for FFA Mapping

Mihir Kadiwala (kadi@gatech.edu)

Atlanta, GA

Abstract

Logic and reasoning are integral parts of human cognition, but so is emotion. Convolutional neural networks have their foundation in cognitive science and neuroscience. Particularly, the different regions of the brain that serve different purposes serve as the inspiration for different layers in a CNN. It has been noted that the progression of layers in a convolutional neural network map to the ventral visual stream. In human brains, we have specific cortical regions that can detect individual faces and the emotion that they present. I hope to use a list of emotionally relevant stimuli to pass into a trained ResNet50 CNN model and evaluate how well this model represents emotions using the facial imagery shown in images. I will also look at the different layers to see if the model can discern emotions inherent within its layers of representation. Since humans process faces at the fusiform face area in the brain, I will be most interested in finding the ResNet50 equivalent of this area and evaluate if the different types of stimuli are represented differently for this layer.

Understanding The Human Ventral Visual Stream

The basis of this paper explores how select models for computing produce similar patterns shown in human neural pathways. First, it is important to understand the pattern of activation that occurs in human brains for any visual stimuli. This understanding will help us understand what a machine representation should look like.

The ventral visual stream is also known as the “what” pathway in humans. It is responsible for processing visual input from the eyes. Information is first collected from the ganglion cells in the retina. Then, the information is sent to the primary visual cortex, which is responsible for initial processing. This area detects edges, lines, and other basic stimuli. Then, the secondary visual cortex integrates information from the primary visual cortex and is able to detect more complex patterns like texture and contours. The tertiary visual cortex is the first region of the visual stream that can detect motion based on information from the other two regions. To continue, the brain continues to different regions for tasks such as motion processing, spatial processing, and object recognition.

Object recognition happens in the Inferotemporal cortex (IT). This is where we begin to understand what we are receiving in visual input with a higher degree of understanding. Within the inferotemporal cortex lies the fusiform face area (FFA). The FFA is not the only cortical region that is able to process visual information at a high level. There are other regions that can also process places, and for other senses like hearing and touch, there exists higher level processing for locality meaning too. In this project, we are mostly concerned with the fusiform face area or the inferotemporal cortex since it seems to be the primary region in the brain that is responsible for processing stimuli of faces, which is the same stimuli we will be feeding machine models.

The fusiform face area has been studied to hold specific activation for individual faces (Quiroga et. al 2023). Their study was able to measure fMRI activation in the FFA for different kinds of stimuli such as known and unknown faces and places. In the past, it has been measured that faces tend to invoke more activation compared to other stimuli. Their research showed that there is also specific activation for specific people. Thus, much of the emotional processing in the visual ventral stream is most prevalent in the inferotemporal cortex and the fusiform face area. It is important to note that while activation in the FFA is similar for most facial stimuli, it is still unique for each stimulus. The FFA, for the most part, can distinguish between a friend's face and a parent's face. It knows the difference between a new face and known face, but more importantly it also holds some semantic meaning of the emotional expression that any given face carries. This suggests that the FFA has neural firing that is unique not only for the person, but it also has differential activation for discrete emotions for the same person.

Mapping CNN Models to the Visual Ventral Stream

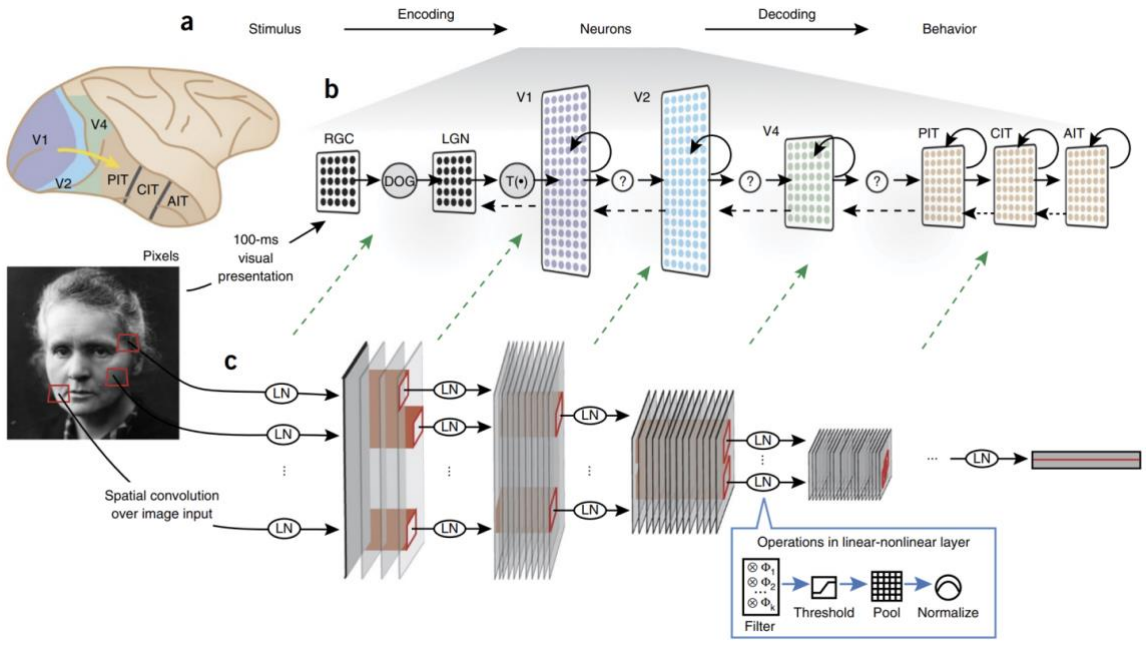

Figure 1— Figure taken from Yamins et al. (2016) showing CNN similarity to visual ventral stream.

Convolutional Neural Networks (CNN) take their inspiration from the visual stream, and as such, their architecture tends to resemble architecture of human visual processing. Figure one shows the similarity that CNNs have to brain processing. Given the similarity, it is no surprise that CNNs have been used to model human cognition in the past and even tested for human similarity in cortical responses.

Kragel et. al., in their research, take on the similarities between computer models and the visual ventral stream. They found that throughout this whole process exists some resemblance of emotional understanding through the CNN (2019). They created their own convolution network and called it EmoNet. They were able to correctly predict specific emotional responses given a wide range of visual stimuli. For example, the model was given an image taken from a large height, which it categorized as fear evoking, and another example of a baby animal, which it categorized as adoring. They looked at a specific layer of EmoNet, which they called the emotional category layer and compared it to human brain activations for different emotions. They found that their model held some similarities to fMRI data taken from the human brain when posed with some of the same stimuli. they did this by comparing specific cortical regions like the occipital lobe to the clustering found in their model.

While we can certainly take the human brain and use it as inspiration for machine learning models, we can also try doing the opposite and take a machine learning model to help us better understand our own cognitive architecture. Kriegeskorte et. al. successfully showed that CNN architectures can be used to model brain information processing. They posed the issue of having to take a specific snapshot of person's brain at a given moment to measure activations. At that point, the processing of a stimuli could have moved onto a different region and tamper with any conclusions we might draw from the image. It can be difficult to measure activations of

small cortical regions in the brain for specific stimuli because it is difficult to measure the exact time frame that the region receives and outputs information. Kriegeskorte et. al. decided to exploit the similarity of layers in CNNs and the ventral visual system to model human brain activity for intermediate and preliminary layers of this system. They were able to draw conclusions on how the biological system works by using a CNN as a framework. Although their findings can't be fully verified by medical professions, their research shows the power of CNNs since they can give us a logical mapping between connecting regions in the brain.

There is a clear connection between the biological framework that exists in human and primate brains and the architecture and perceptron firings of specific layers of the CNN. These models are not necessarily equipped to complex emotional stimuli since they are simply large chains of mathematical representations, but they are still able to classify emotional stimuli. Modi et. al. trained a basic CNN model on emotional stimuli using a supervised learning technique and found that the model was able to correctly predict the emotional classification by reading an image of a person's face (2021). Modi et. al. is able to build a model that could predict emotions with up to 82.5% accuracy, which was higher than a Transfer Learning Model. This is the model that uses the end of a task to help it begin another task. The Transfer Learning Model has been commended with its ability to complete computations and tasks on items that were not originally in its training. However, when measured against the CNN for the task of classifying a person's face from neutral, happy, or sad, the CNN did much better, since the highest accuracy for the Transfer Learning model was only 73.5%.

The recent research has proven the effectiveness of CNNs and have shown that they hold promising results to belter model human cognition since they excel at tasks that are not purely computational, including classifying emotional stimuli. Perhaps, the layering of networks in a

model like the brain and a CNN gives it the ability to break down stimuli into its most fundamental parts and then build up using a bottom-up approach, and somewhere inherent within the structure allows it process nontrivial concepts like emotions. I am not looking to evaluate a specific model on emotional faces, since that has already been done. Instead, I hope to gauge a better understanding of how the architectural connections between the human visual system and CNNs emerge throughout the layers of a given CNN model. Particularly, I am interested in seeing evidence of the visual ventral stream in a CNN. Since the overall analogical mappings have been shown to have validity, I am interested in seeing the makeup of specific layers that would most closely map to the FFA in the human brain to gauge if that layer is also able to resemble something similar to how a FFA or the IT processes information.

The possibility of a layer that resembles information captured by the FFA would be made possible by the overall similarity to the human visual ventral stream, so if the layer most closely related to the FFA shows great similarity, then I should also expect to see earlier layers that capture more fundamental information like lines, edges, contours, and important features. This experiment tests if the architectural design of CNNs can mirror human processing by paying attention to activation in specific layers. It is interesting because I am measuring if CNNs have a unique capability to hold representations of emotions simply because they have layered representation of stimuli.

Data and Model Selection

I chose to use a dataset that had been used in previous research since I did not want to introduce new stimuli that can cause variability and hadn't been tested and worked on before. I decided to use the FACES database, as characterized by Grühn and Sharifian (2016). This dataset has discrete labels for five basic emotions: anger, disgust, fear, happiness, and sorrow. It also has

sixth category for neutral faces. The database gave me seventy-two total images. There are six individuals that pose slightly differently twice for each category. This gives me twelve total images for each emotional category. I have an example of one individual's faces in figure 2.

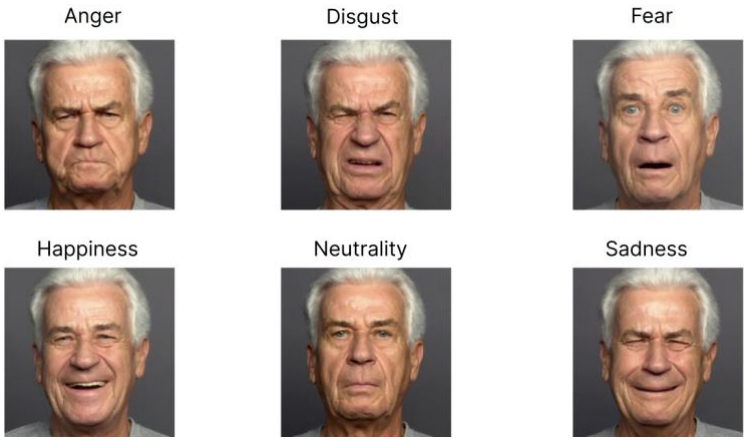

Figure 2— Example of six basic emotions from the FACES database

Next, I needed to choose a model of a CNN that I could plug stimuli into. I came to understand that multiple CNN models had great success at mirroring sensory cortex (Yamins and DiCarlo 2016). Yamins and DiCarlo used CNNs and optimized them for specific tasks, and they called these goal driven networks. Then, they tested these models to see if they may hold some similarities in firings to the processing that happens in the first few regions of the visual stream. They found that the top hidden layers in these models were able to predict neural firings in the Inferotemporal cortex. They tested this for specific models and found this to be true for most of them. In particular, they were able to show specific neural activations that were similar for the seventh layer of AlexNet and the IT of a monkey.

I would have liked to explore this seventh layer of AlexNet since it has already been shown to hold similar firings in a different context, but I wanted to test another model to see if it may perform similarly. Additionally, I faced implementation issues with AlexNet. I decided to test my hypothesis using the popular ResNet50 model simply because it has grown in popularity

and usage in recent years. I chose to use a pretrained model from TensorFlow since I was looking to see how it was able to capture new stimuli given its architectural makeup and I wasn't necessarily interested in probing how well it learned stimuli. However, this meant that I would not have a goal-driven model that Yamins and DiCarlo were able to create. The difference turns out to make for a great experiment since I am hoping to replicate the findings that they found at the top hidden layers of a CNN, but without personally training the network and by feeding it simple facial imagery meant to represent discrete emotional expressions.

In order to accurately measure how ResNet50 processed images, I partitioned the dataset into the six different categories based on emotional classification of faces and passed each set into the model. Then, I measured the activation for each image at a specific layer and measured the average activation for the layer for each emotion. I looked at a few layers since I did not immediately know which layer would correspond to the FFA, but I had a good idea that it would be near the end of the network.

Results

First, I looked at earlier layers of the model to see if I could pick out any semantic meaning or evidence of earlier pathways in the visual ventral stream. Figure 3 below has one of these layers. You can see that you can pick out the contour of a person's shape and also notice that the most important features for identifying emotional expression have higher activations. The eyes and mouth are clearly represented within the layers. I believe this is similar to what we might find in the second or tertiary cortical regions of the visual ventral stream since it has knowledge of edges, lines, and contours, but it is also paying attention to the specific parts that tend to move the most and provide the most insight regarding a person's emotional state.

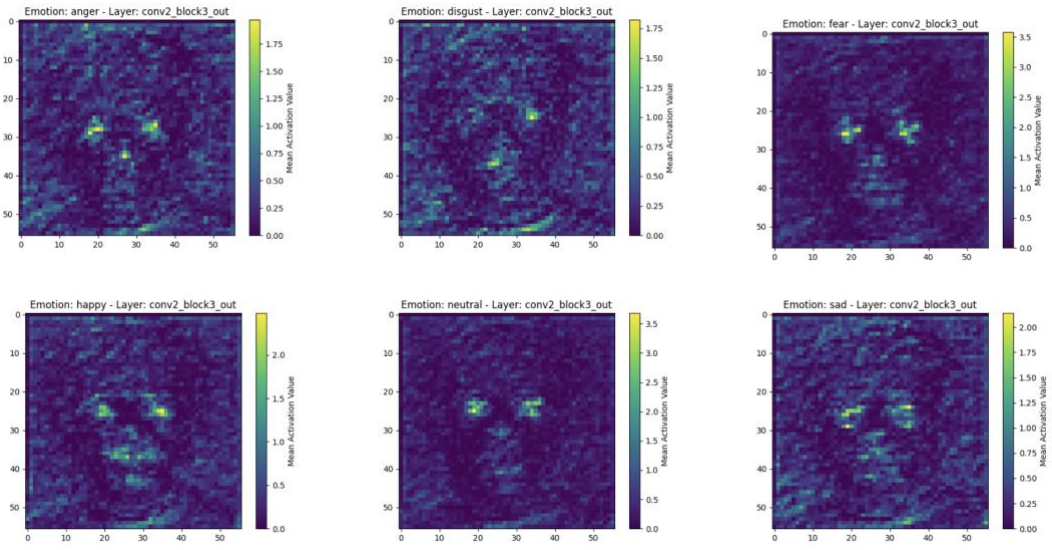

Figure 3— ResNet average activation for emotional stimuli at the intermediate 'conv2_block3_out' layer.

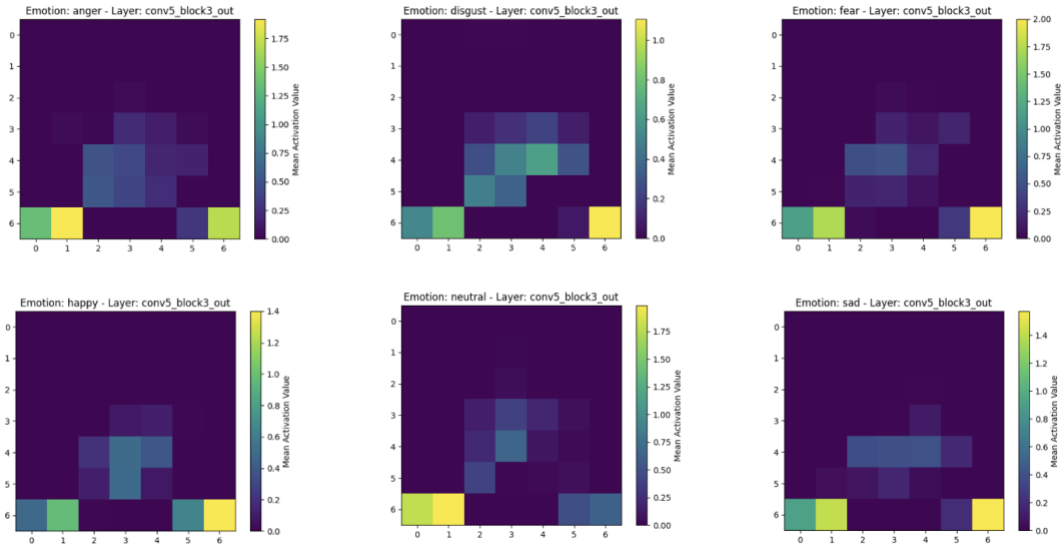

Figure 4— ResNet average activation for emotional stimuli at the later hidden layer 'conv5_block3_out.'

Next, I tried several different layers before settling on the last hidden output layer in the model to represent the CNN. This is consistent with Yamins and DiCarlo's findings. Figure 4 shows the average activation at this layer. I believe the activations provide evidence for a FFA or IT equivalent in ResNet50. Each emotional state has a

similar structure to it with slight variations for each emotion. The bottom row has the most activation ordered by bottom right with the most and then the bottom left. For neutral expressions, the bottom right is the least activated in the bottom right.

Discussion

The layer activation has a common structure in the middle of the unit. Disgust has the highest activations probably because the facial imagery for disgust is likely the most different than other facial expressions. Each midsection is lightly activated but unique, and each bottom row has the same activation symmetry to varying degrees of activations. These patterns provide evidence to support the claim that CNN layers are able to accurately replicate the ventral visual stream present in biological beings. A similar pattern of neural firing is what we would expect to see from human or primate fMRI data.

The evidence is further compelling because we have an earlier layer that would easily map to early layers of the visual process. Additionally, this evidence is consistent with previous research that suggested that the top hidden layers is where we would see layers correlated to the IT.

While this research is exciting, there definitely needs to be more investigation into these layers to fully evaluate its mapping to cortical regions in the brain. A goal driven CNN model would be the next step, since it should give us a more robust result since its parameters would be optimized for emotional classification. Additionally, a larger dataset with more variability that includes race should be examined. My dataset only consisted of seventy-two images of white Americans.

To further test this hypothesis, it would be beneficial to test other CNNs, but I am confident that they would have similar results. If we are to build on this analogical mapping, it

might be beneficial to see if something like the distance effect we see for numbers and words also exists for faces and emotional expressions. If we are more likely to label an expression using a logarithmic scale depending on how far it deviates from neutral, we may find such effects inherent within the representations for emotions when given facial stimuli.

References

Baghbani, F., Akbarzadeh-T, M.-R., Naghibi-Sistani, M.-B., & Akbarzadeh, A. (2020). Emotional neural networks with universal approximation property for Stable Direct Adaptive Nonlinear Control Systems. *Engineering Applications of Artificial Intelligence*, *89*, 103447. https://doi.org/10.1016/j.engappai.2019.103447

Grühn, D., & Sharifian, N. (2016). Lists of emotional stimuli. *Emotion Measurement*, 145–164. https://doi.org/10.1016/b978-0-08-100508-8.00007-2

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and Artificial Intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science Advances*, *5*(7). https://doi.org/10.1126/sciadv.aaw4358

Kriegeskorte, N. (2015). *Deep Neural Networks: A New Framework for Modelling Biological Vision and Brain Information Processing*. https://doi.org/10.1101/029876

Lotfi, E., & Akbarzadeh-T., M.-R. (2014). Practical Emotional Neural Networks. *Neural Networks*, *59*, 61–72. https://doi.org/10.1016/j.neunet.2014.06.012

Modi, S., & Bohara, M. H. (2021). Facial emotion recognition using convolution neural network. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iciccs51141.2021.9432156

Rust, N. C., & Jannuzi, B. G. (2022). Identifying objects and remembering images: Insights From Deep Neural Networks. *Current Directions in Psychological Science*, *31*(4), 316–323. https://doi.org/10.1177/09637214221083663

Quian Quiroga, R., Boscaglia, M., Jonas, J., Rey, H. G., Yan, X., Maillard, L., Colnat-Coulbois, S., Koessler, L., & Rossion, B. (2023). Single neuron responses underlying face recognition in the human midfusiform face-selective cortex. *Nature communications*, *14*(1), 5661. https://doi.org/10.1038/s41467-023-41323-5

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Zhou, Z. (2021). Emotional thinking as the foundation of Consciousness in Artificial Intelligence. *Cultures of Science*, *4*(3), 112–123. https://doi.org/10.1177/20966083211052651