Understanding Emotion Recognition in Neural Networks

Ivan Montero, Terrell Strong, and Caleb Kierum

Department of Computer Science and Engineering
University of Washington

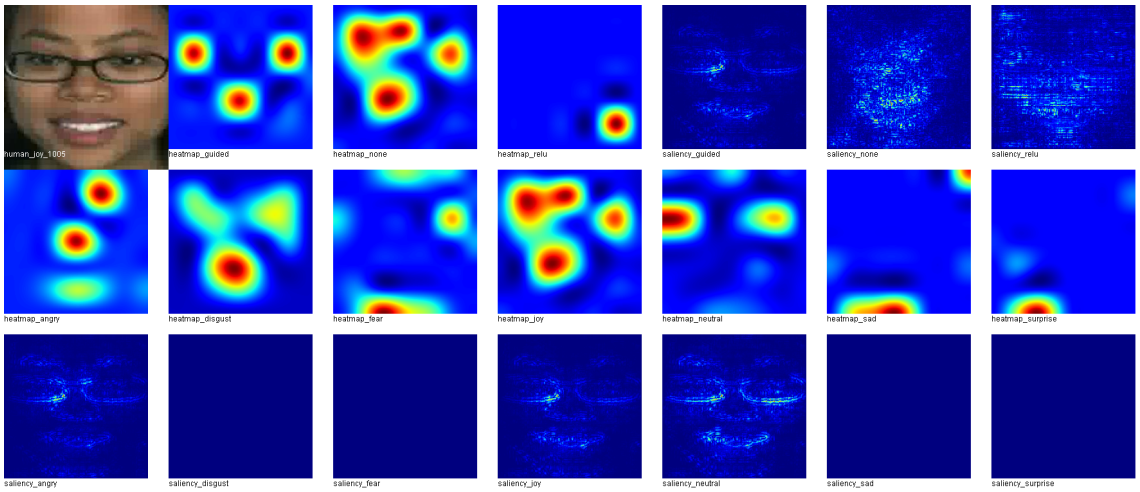

Figure 1: *A collection of activation maps and saliency maps associated with the face in the upper-left. The top row shows the maps associated with the predicted emotion, the middle row shows activation maps for individual emotions, and the bottom row shows saliency maps for individual emotions. The emotions from left to right are anger, disgust, fear, joy, neutral, sadness, and surprise.*

Abstract

We apply neural network visualization methods to networks used to identify emotions in human faces in order to gain a better understanding of how the network interprets facial features as a particular cardinal emotion.

Techniques such as saliency and class activation maps allow us to determine that networks do indeed seem to focus on the eyes and mouth to determine emotions. Furthermore some emotions such as joy, disgust, and neutral seem to rely on the mouth more than the eyes whereas surprise seems to focus mostly on the eyes. This explains issues like the frequent miscategorization of joy as neutral.

1 Introduction

Neural networks have been applied to many problems recently with significant successes, yet the inner workings of neural networks are often poorly understood. By using a neural network that recognizes emotions in human faces and analyzing the results of various visualization techniques, we aim to get a better understanding of how neural networks interpret features of an image in order to make its predictions.

Although our research is specifically related to human emotions, the idea of visualizing what a neural network has learned fits under the larger umbrella of explainable neural networks which has a growing importance as artificial intelligence starts to proliferate into more areas of our everyday lives.

This project, in addition to understanding what areas of the face dictate an emotion, was a learning experience for us. Having no prior experience with machine learning, we progressed through this project to understand the fundamental concepts relating to machine learning and how neural networks work in order to understand how to go about visualizing its predictions.

2 Related Work

Nguyen et al. [2014] address the problem of problem of visualizing deep neural networks by using gradient ascent to produce images that maximize recognition of a class and by creating saliency maps of images to visualize what pixels in the image are most important for determining its classification. The images produced with gradient ascent were not human-recognizable, but often contain some distinguishable features that could be associated with the optimized class. The saliency maps described in the paper were significant because they visibly highlighted the parts of the image that the neural network depended on the most in order to classify it. In our study, we produce saliency maps from human faces in order to visualize what features of the image are used by the network to identify cardinal emotions.

Simonyan et al. [2013] address multiple methods that can be used to produce fooling images that are not recognizable by humans, but are interpreted by neural networks as belonging to a specific class with high certainty. One way that was used to create fooling images was evolutionary algorithms on a classified image in order to maximize the networks recognition of the image as a member of the given class. Although the image produced using this method were rarely human-recognizable, the image often provided valuable insight about the features that a network uses to discriminate one class of object from another. Although we dont apply this method of visualizing classes in our study, this method could provide valuable insights in future studies.

Mahendran and Vedaldi [2014] address the lack of understanding in how neural networks interpret and classify images by implement-

ing a method of reconstructing an image from its representation throughout the layers of the network. Their method of visualizing the network is interesting because it can be applied to individual layers of the network to understand how an image is manipulated and interpreted at each step.

Dosovitskiy and Brox [2015] address the problem of visualizing neural networks by implementing a method of reversing a network so that it estimates an input image based on the feature representation of an image from the original network. This is similar to the work of Mahendran and Vedaldi [2014] that we also refer to since both papers aim to visualize the features used by the network as it classifies an image, but this is different because it works without the use of a image prior.

3 Methodology

For our experiments we mostly exclusively used the emotions in the wild dataset and a neural network trained on the data set, which we successfully converted from the Caffe model zoo to Keras, for our visualizations. The generated visualizations are in one of three formats: activation maximization, saliency, and class activation maps.

Activation maximization visualizations most often look similar to the input image just more blurred and with textures shown in some areas. Activation maximization does gradient descent, a process similar to training an actual neural network, on the provided input image. This is the technique utilized in Googles Deep Dream algorithm [Tyka et al. 2015]. Gradient descent gradually changes each pixel in the input, over many iterations, according to the gradient provided by:

$$\frac{\partial ActivationMaximizationLoss}{\partial input}$$

Where the input is the specific pixel and the ActivationMaximizationLoss is a loss function that maximizes the output for a specific layer or classification. This algorithm, after each iteration, produces an image which the neural network has a greater confidence value in being the specified class being maximized for. This method can also be applied to any convolutional layer in the network, and to specific filters.

Saliency maps display the gradient of output categorization with respect to the input image as a dotted blue image with lighter regions showing the portions of the image that had a larger effect on the images categorization. The noisy, dotty appearance of its output is due to the algorithm doing back-propagation, similar to training a neural network, and displays the magnitude of gradient of each pixel, determined by:

$$\frac{\partial output}{\partial input}$$

Where the input is the specific pixel and the output is the specified class. In effect, this method shows which pixels in the input image have the greatest influence on the network's resulting interpretation. Beyond this there are three sub-categorizations of the technique that produce distinctly different saliency maps. With no modification (none) to the back propagation, we get images that show areas of general interest on the image. These results are generally noisy, as they communicate all the information carried through an unmodified gradient. However by modifying the back propagation step to only propagate positive gradient information (relu) we get a saliency map communicating the increase in output. This is accomplished by utilizing ReLU (Rectified Linear Unit) activation at every layer, which clips the output in the [0, 1] range. This method, through practice, produces nosier values than no modification at all, due to the fact that only positive gradients are always communicated through the backpropagation, and the absence of negative gradients means the resulting gradients only increase after each successive layer. This method communicates increase in output. Finally by using a guided technique that only propagates the positive gradients of positive activations we get very clear shapes from the saliency maps showing sections of the images that are directly related to the specified classification. This is because the information thats communicated in the back-propagation is directly related to the influence on the output–both positive gradients and positive activations narrows down the areas of interest in the input image. Class activation maps use a similar technique to generate heatmap-like maps from the input image. This method utilizes the output of the penultimate–the layer before the dense, fully-connected layers–convolutional layer, retaining spatial information that is completely lost in dense layers.

$$\frac{\partial penultimate\_output}{\partial input}$$

This methods allow us to get an idea of what groups of pixels in the input image have the greatest influence on the networks interpretation, rather than the influence of specific pixels.

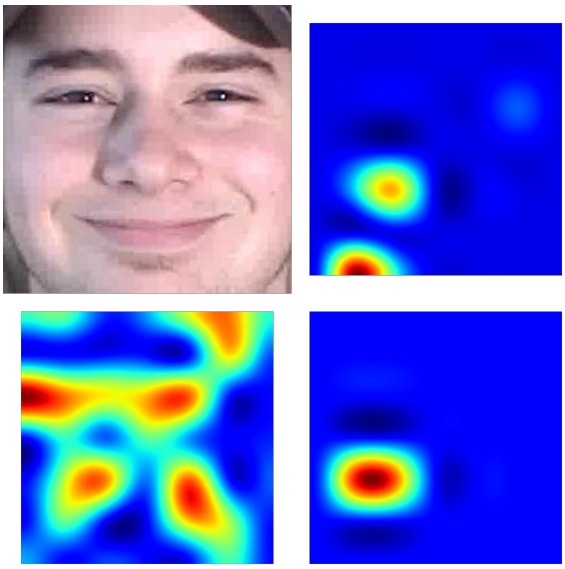

Figure 2: *A set of class activation maps produced from the photo in the upper-left. The upper-right image is a guided saliency map, the lower-left is an unguided saliency map, and the lower-right is a relu saliency map.*

4 Results

Activation Maximization

We used an edited image of a female face without eyes or a mouth in order to test the results of activation maximization. The motivation behind using an edited image was to remove features that were often used by the network to determine emotions so that maximizing an emotion would generate features that were not represented in the original image.

Maximizing this image for fear produced an image that seemed to be blurred, offset, and have lower contrast when compared to the

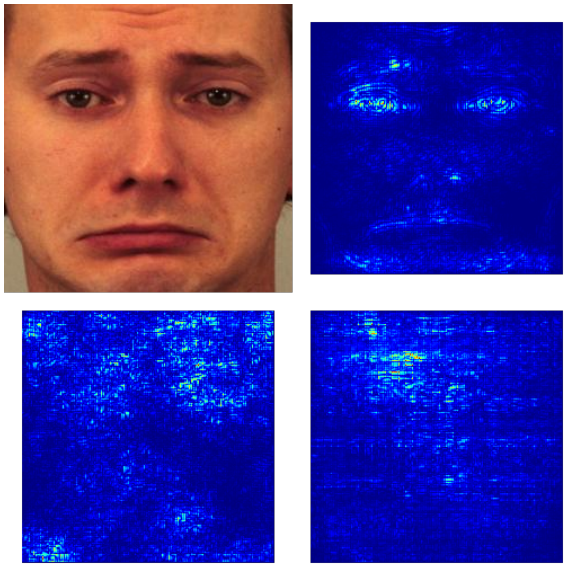

Figure 3: *A set of saliency maps produced from the photo in the upper-left. The upper-right image is a guided activation map, the lower-left is an unguided activation map, and the lower-right is a relu activation map.*

original image. In the center of the generated image, there were faint lines and shapes that were barely distinguishable from the rest of the image. In order to make the shapes easier to see, we tried applying several different image filters to the image.

As we tried interpreting the features of the generated image, we also tried applying an offset to the generated image so that the image more closely resembled the shape of the original face.

Saliency Maps

The guided saliency maps often included the finest level of detail when considering other methods we used since they would often produce shapes that were identifiable as facial features such as the mouth and eyes. In some cases, the guided saliency maps also highlighted the outline of the face, the brow of the face, and areas around the nose where the skin creased due to the mouth shape.

This pattern of highlighting continued even during animations. In the animation of the saliency maps of a image sequence where an actor changed their emotion from neutral to happy you could clearly see the saliency map following the creases formed near the corners of the mouth indicating that this may have been a significant factor in the network's categorization.

Although the unguided and relu saliency maps did not produce shapes with as much detail as the guided saliency maps, they provided useful information about the more general areas that the network used to interpret emotions. Notably, the relu saliency maps highlighted areas in the middle of the face more brightly than areas towards the edge of the face or entirely off of the face. This implies that the network was interpreting the face in order to determine emotion rather than trying to interpret other details that are not consistent across the faces and images we used.

The unguided saliency maps often provided information similar to the relu saliency maps, but would sometimes highlight features of the face more brightly and distinctly than the relu saliency maps. The unguided map would sometimes highlight areas of the image

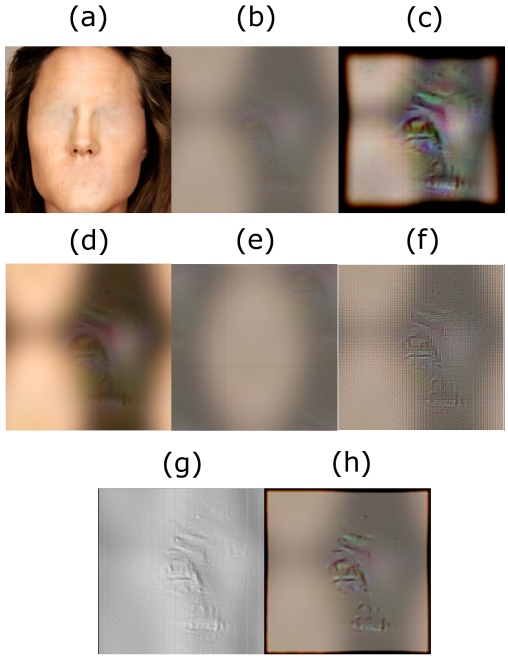

Figure 4: *The image we used for our activation maximization. (b) The image produced after optimizing for fear. There are very faint shapes in the center of the image that were produced in the process of maximizing. The remaining images are filters applied to b in order to make the new features more visible. (c) This is the result of applying an unsharp mask to b. (d) This is the result of increasing the contrast of b. (e) This is the result of offsetting b in order to more closely match the shape of the face in a. (f) This is the result of increasing the sharpness of b. (g) This is the result of applying an emboss filter to b. (h) This is the result of applying an unsharp filter that is weaker than 'c' to 'b'.*

where hair occluded the face or the background was visible. Since these highlights were often exclusive to the unguided saliency map and did not appear in the guided or relu saliency map in most cases, these highlights could be interpreted as being negative activations that push the network's focus away from the edges of the image towards the center where the features are more useful.

Activation Maps

The activation maps highlighted features similar to the saliency maps for the most part, although the highlights were not always focused in the middle of the image.

The unguided activation map often had more highlights than the guided and relu activation maps and would focus on features in a way similar to the saliency maps. Areas such as the eyes and mouth were highlighted in the activation maps fairly consistently, although it is notable that the bottom edge of the image was also highlighted in many of the unguided activation maps. It is possible that this highlighted edge shows that the network sometimes uses the area beneath the mouth and around the chin as a facial feature to predict emotions.

The guided and relu activation maps normally had small activations focused around a single point in the image. Although the highlighted point would often be on or near important features such as the eyes or mouth, the highlight sometimes appeared at the edge or corner of the image instead.

5 Future Work

A lot of our results could have been improved, due to the time constraints we were given. One aspect we could have changed in our exploration would be using a model we specifically trained, rather than one trained by someone else. This would have improved the coherence of our results, since the network we utilized was trained on a data set that included faces expressing emotions in various orientations, thus producing results that were somewhat hard to interpret in our activation maximization methods. Additionally, the data set was the Emotions in The Wild data set, which were images of people expressing emotion with the surrounding environment in view. This inclusion of the environment in the data set may have led the network to utilize color to determine emotion, which can be seen when "fear" was maximized in our activation maximization method since the network tended to blur the image with a darker, greenish tint. We would train the network on cropped images of emotions, further narrowing the scope of what the network focuses on. Additionally, we would gain better certainties and clearer results since the network would be more specialized for our focus.

Another aspect we could have changed is the images we utilized to produce the results. The images we utilized have never been seen by the network, and are in a different orientation than that which the network has been trained with. This is evident, the network's predictions were fairly off when identifying the emotion conveyed in the image. If we utilized images which the network was trained on, it'd produce a greater certainty value and corresponding visualization results on those images.

It could also be insightful for animators and artists to apply this technique to artificially created faces and emotions with a network trained on human tagged data. Running these visualizations with such a network and data would potentially give insights on statistically what common features from these artificial faces help it be categorized as a specific emotion.

6 Conclusion

Overall each of our visualization techniques did clearly show that this neural network correctly focused on the eyes and the mouth to determine their categorization which is refreshing considering how many times neural networks have been proven to categorize based on surroundings rather than objects.

Although the results of our study show that the network we used interprets facial features such as the mouth, eyes, and brow when predicting the cardinal emotions, we are not sure if this conclusion can be generalized to different neural networks that use other architectures. Some of our results may have been clearer if we had used a network that was not focused on recognizing images "in the wild" where different face orientations, lighting, and face locations within the frame were intended to be part of the challenge of training the network. Because of this, future studies would investigate if the architecture of a network influences the features visualized.

References

Mike Tyka Alexander Mordvintsev, Christopher Olah. 2015a. DeepDream - a code example for visualizing Neural Networks. https://research.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html. (July 2015).

Mike Tyka Alexander Mordvintsev, Christopher Olah. 2015b. Inceptionism: Going Deeper into Neural Networks. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. (June 2015).

Mike Tyka Alexander Mordvintsev, Christopher Olah and contributors. 2015. deepdream. https://github.com/google/deepdream. (2015).

Alexey Dosovitskiy and Thomas Brox. 2015. Inverting Convolutional Networks with Convolutional Networks. *CoRR* abs/1506.02753 (2015). arXiv:1506.02753 http://arxiv.org/abs/1506.02753

Raghavendra Kotikalapudi and contributors. 2017. keras-vis. https://github.com/raghakot/keras-vis. (2017).

Aravindh Mahendran and Andrea Vedaldi. 2014. Understanding Deep Image Representations by Inverting Them. *CoRR* abs/1412.0035 (2014). arXiv:1412.0035 http://arxiv.org/abs/1412.0035

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *CoRR* abs/1412.1897 (2014). arXiv:1412.1897 http://arxiv.org/abs/1412.1897

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* abs/1610.02391 (2016). arXiv:1610.02391 http://arxiv.org/abs/1610.02391

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* abs/1312.6034 (2013). arXiv:1312.6034 http://arxiv.org/abs/1312.6034

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2014. Striving for Simplicity: The All Convolutional Net. *CoRR* abs/1412.6806 (2014). arXiv:1412.6806 http://arxiv.org/abs/1412.6806

Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2901 (2013). arXiv:1311.2901 http://arxiv.org/abs/1311.2901