CSC411: Assignment #2 Bonus

Due on Friday, February 26, 2018

Lukas Zhornyak

February 25, 2018

1 Environment

This submission was created with Python 3.6.4 with numpy 1.14.1, scipy 1.0.0, scikit-image 0.13.1, and matplotlib 2.1.2, pytorch 0.3.0, torchvision 0.2.0, opencv 3.3.1, as well as all associated dependencies.

The various visualizations shown in section 3 are produced with the help of the Convolutional Neural Network Visualizations repository created by Utku Ozbulak[1], use with permission under the MIT license. A slightly modified version of the code is included with this submission. Modifications were made to facilitate inclusion as a package in this submission and to facilitate use of a custom neural network model.

All images used are 277 by 277 pixels[2] from the same database used in Assignment 2.

[1] Available from `https://github.com/utkuozbulak/pytorch-cnn-visualizations`

[2] Odd size is due to misreading the typical size for AlexNet. I would expect all results to be extensible to different image sizes however.

2 Warm-up

The weights of the first layer of AlexNet are shown in fig. 1. Each filter displays some sort of simple colour pattern, most commonly either stripes, boundaries, or single dots. The more complicated patterns also tend to be nearly black and white, suggesting a separation in the detections of colour boundaries and patterns. In either case, the majority of the patterns seem to be focused on some variety of edge detection, as might be expected from the first layer of a convolutional neural network (CNN).

Figure 2 shows the regions in the available dataset the most activate their corresponding filter. A brief examination of the images reinforces the idea that the filter in this first layer are focused on edge detection since the majority of the images in fig. 2 show some sort of distinct boundary and/or line. Figure 3 provides context to these crops. Some notable repeats include glasses, mouths, and eyes, and the letter "W". It also seems that many of the crops come from a smaller subset of images, suggesting that there is an overlap in what the activates the different neurons.

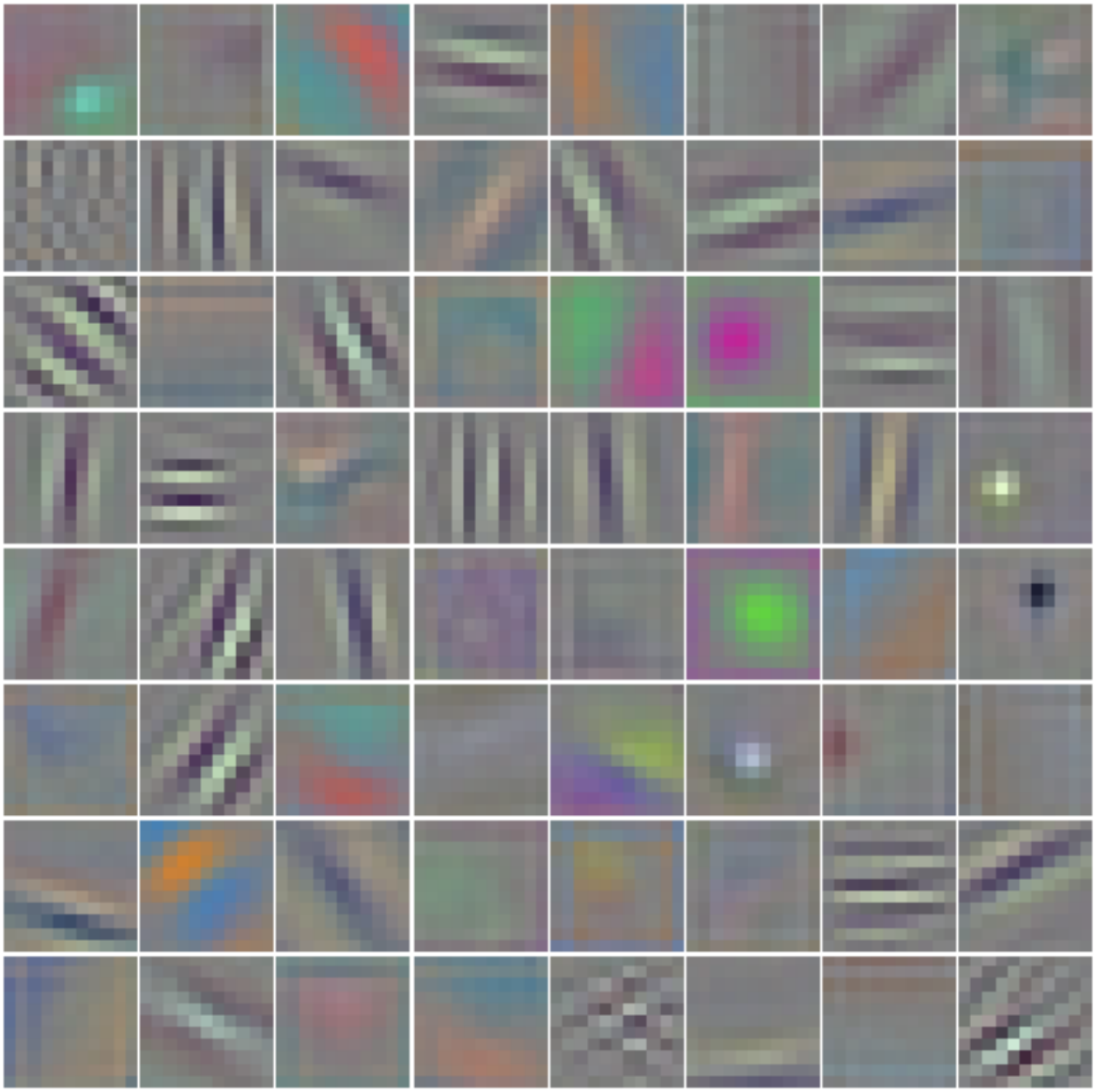


Figure 1: Visualization of the weights of the sixty-four $11 \times 11 \times 3$ filters in the first layer of the pre-trained AlexNet, visualized in colour based on which colour channel the weights access.

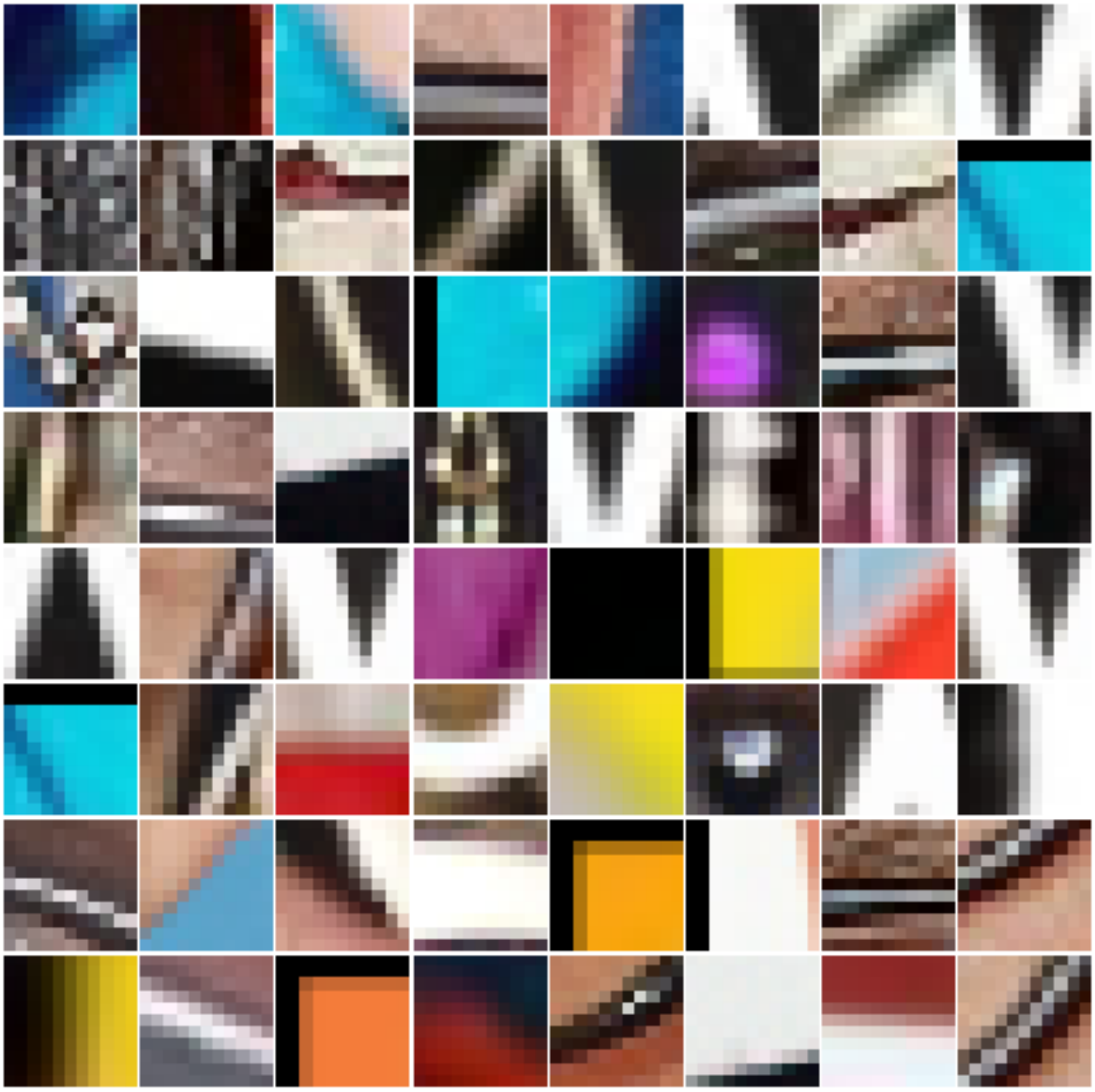


Figure 2: Crop of the regions in the dataset which most activate the filters shown in fig. 1.

Figure 3: Expanded view of the regions shown in fig. 2 to shown context of crop.

3 Visualizing the Deep Network

In creating many of these visualization, a base image is needed. For this reason, the image associated with each actor that had the smallest loss against the expected label was selected. This can be interpreted as the image that the CNN is most confident is an image of the person. These images (fig. 4) show clear, nearly head-on views of the person without any obstructing features. The case may be made to use the image the had the largest one-hot value for that label, as this in some ways represents what the CNN views as the "most" like the person.

The simplest way to visualize what a CNN views as the most important features is by simply visualizing the gradients produces from backpropagation. This is shown in the middle column of fig. 5. It is difficult to make out much from just these gradient, but if the results are normalized we get a saliency map (right column of fig. 5) that shows much more important features. The eyes and nose can be made out in several of these images, suggesting their importance in classification.

An improvement to backpropagation can be made by ignoring the negative gradients. In this way, only those features which given a positive contribution to the classification will be shown – those features which given the impression of that person. This is shown, along with the saliency map, in fig. 6. Again, the importance of the eyes and nose are shown, but some additional features are also expressed. Particularly, the mouth and the contour of the face (e.g. the jawline) also seem to be important for the classification of face.

An even better visualization can be obtained with a technique known as Gradient-weighted Class Activation Mapping (Grad-CAM). Simply, it "uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept." A more detailed description can be found in [1]. Applying it to the output of the first convolutional layer (left column of fig. 7), the observations made in section 2 are reinforced as the edges of the images are the areas that shows the greatest heat on the heatmap.

Applying GradCAM to later layers, larger features become visible. The middle column of fig. 7 shows the heatmap of the output of the second convolutional layer. Already, the eyes, nose and mouth are highlighted clearly, though there is still a lot of noise in the output. By the output of the CNN however (shown in the right column of fig. 7), much of the noise is gone and clear regions of importance are shown, concentrated around the major facial features.

All the visualizations produced so far have relied on using a base image. By choosing a target class and generating a random image, then performing a gradient descent of the image to minimize loss, a idea of the neural networks conception of each person can be obtained. This is shown in fig. 8. Vaguely eye-like swirls where the eyes generally are, circles around where the mouth would, and lines down the nose all serve to provide further evidence towards the importance of these features towards classification. The evolution of these features from the initial random state for Angie Harmon can be found at `https://drive.google.com/open?id=17TaAyBZ6JvGKEvvBlIhP7RKIWS0ud4DR`.

Throughout all of the visualizations presented, a common trend has emerged. It seems that the eyes and the nose – and to a lesser extent the mouth and face shape – serve as the primary features that the neural network uses to classify features. This would make sense; these features are the major landmarks that form a face, and so differences in them would be an easy way to identify a person.

(a) Lorraine Bracco

(b) Peri Gilpin

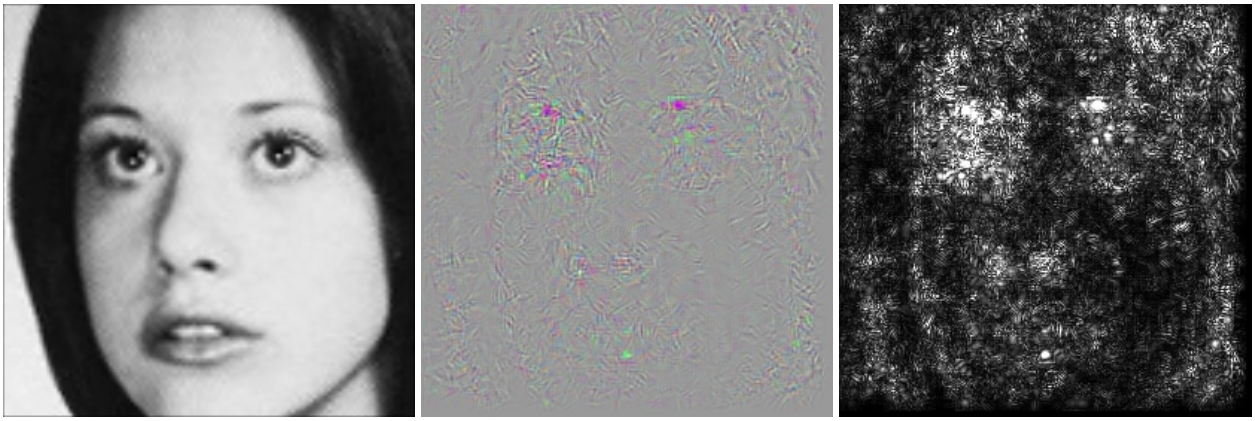
(c) Angie Harmon

(d) Alec Baldwin

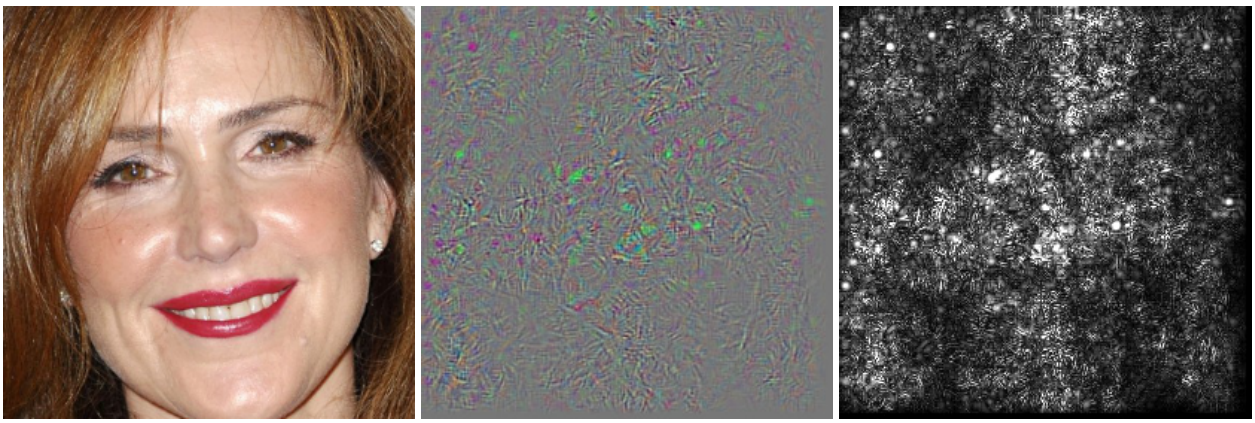
(e) Bill Hader

(f) Steve Carell

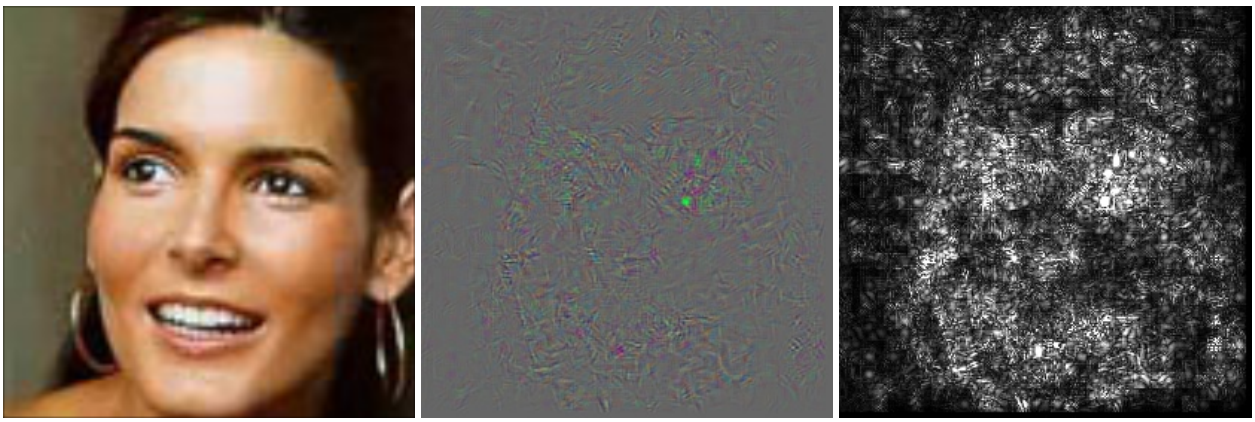
Figure 4: Prototypical images associated with each actor, representing the least loss of any of the images in the dataset.



(a) Lorraine Bracco

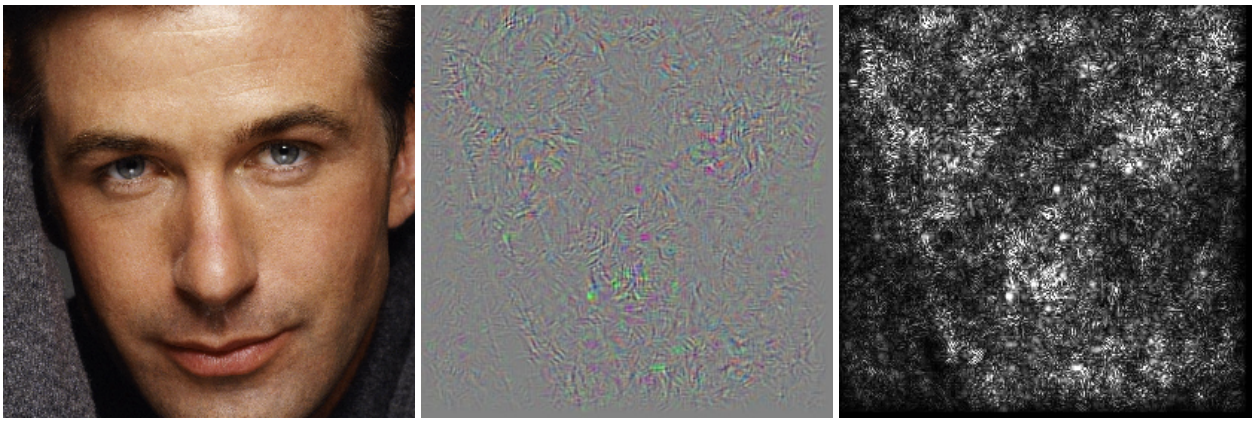


(b) Peri Gilpin

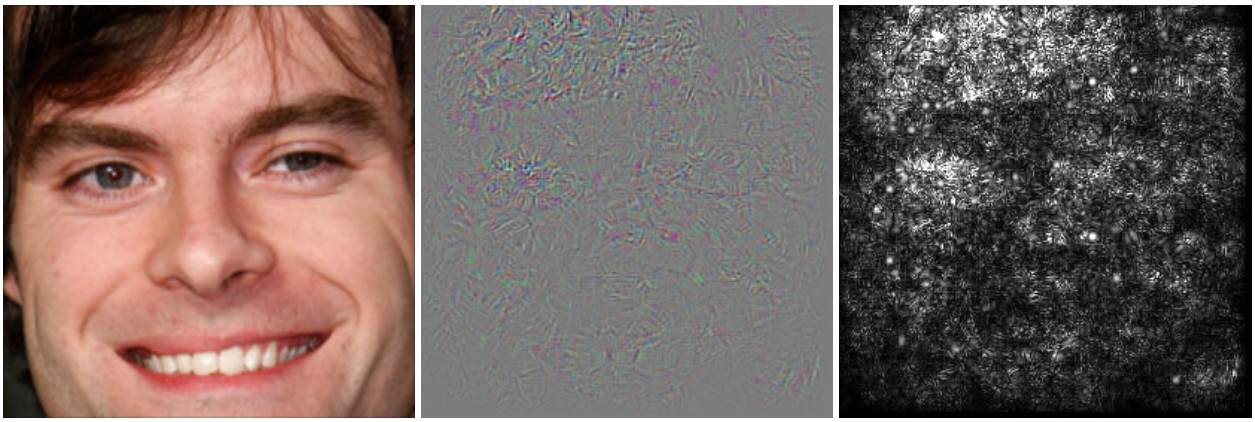


(c) Angie Harmon

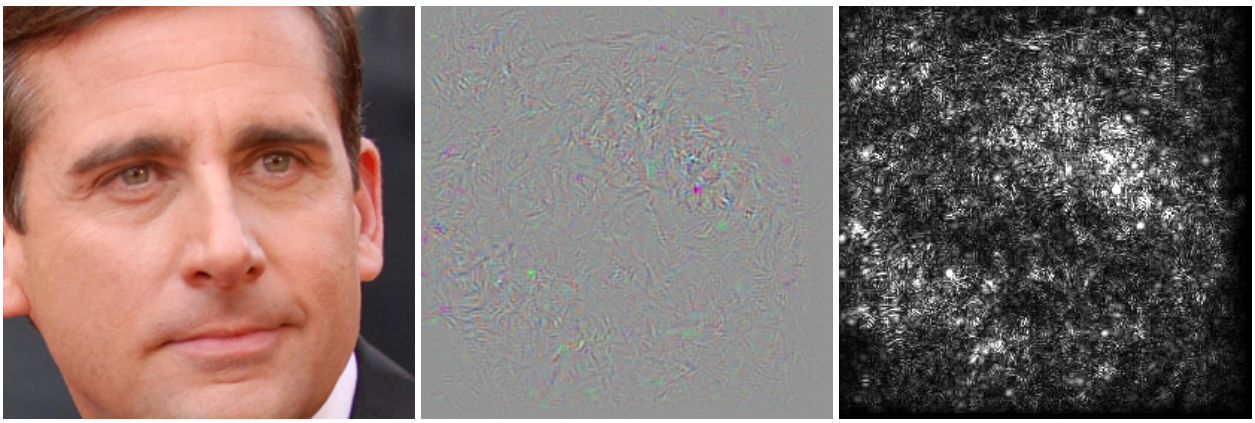
Figure 5: Original image (left), gradients of the image produced via backpropagation (middle), and associated saliency map (right).



(d) Alec Baldwin

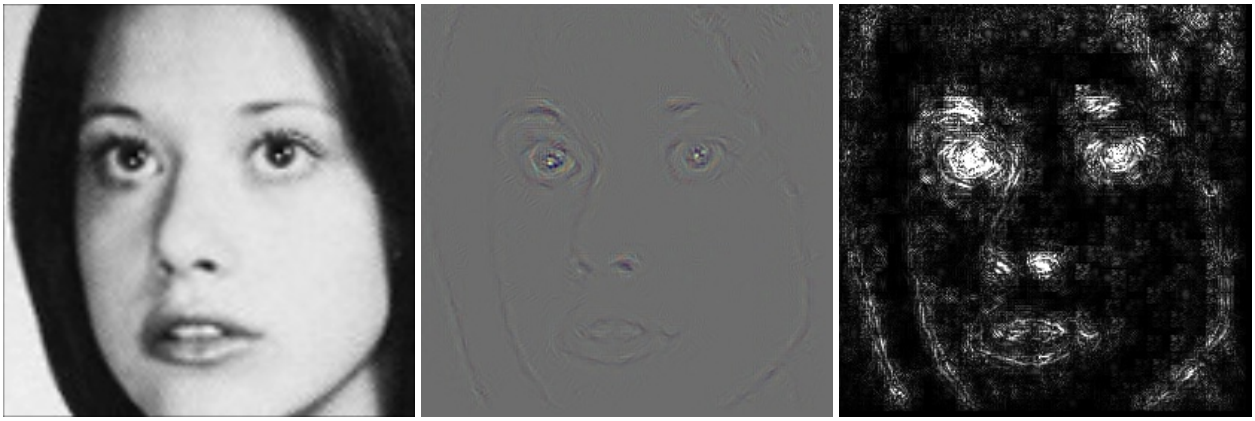


(e) Bill Hader



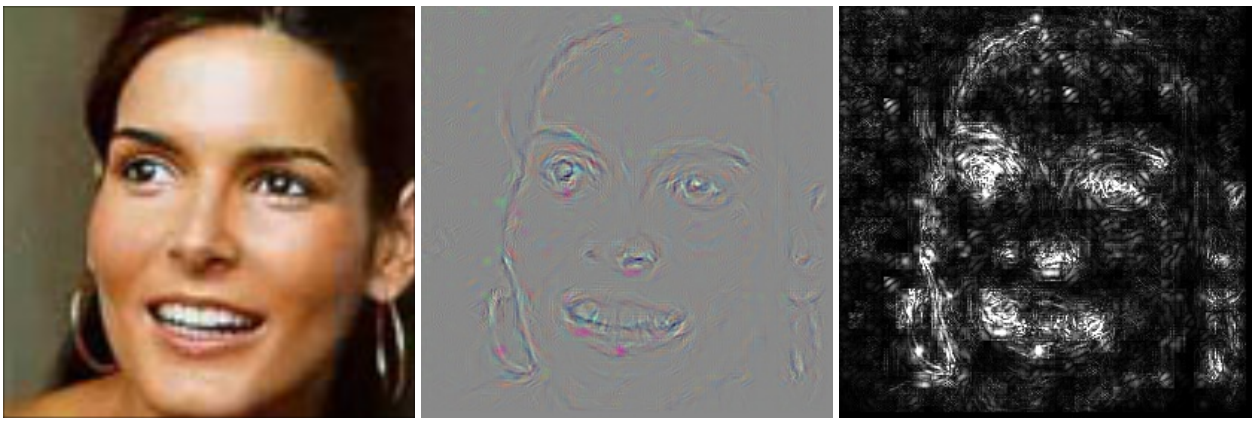
(f) Steve Carell

Figure 5: (Cont.) Original image (left), gradients of the image produced via backpropagation (middle), and associated saliency map (right).



(a) Lorraine Bracco

(b) Peri Gilpin



(c) Angie Harmon

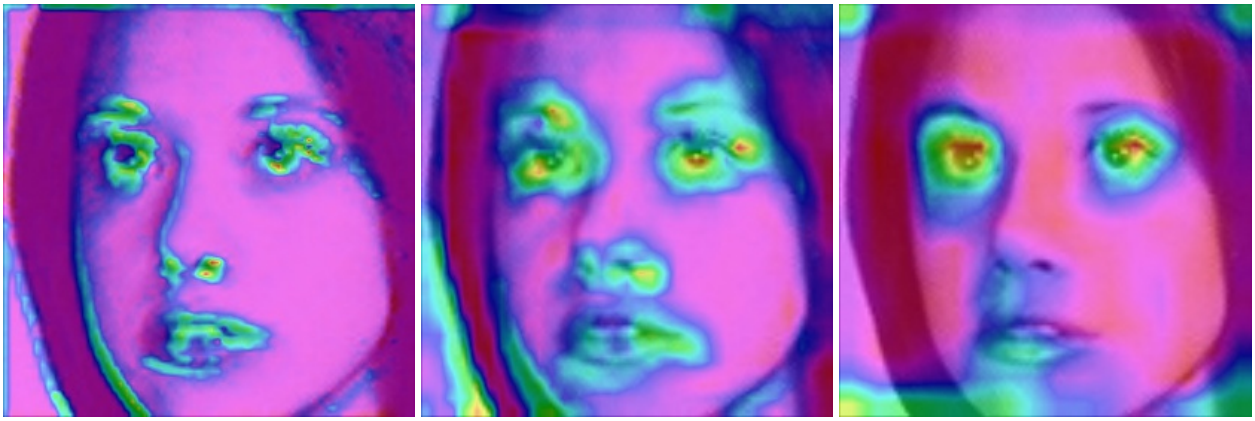
Figure 6: Original image (left), gradients of the image produced via guided backpropagation (middle), and associated saliency map (right).

(d) Alec Baldwin

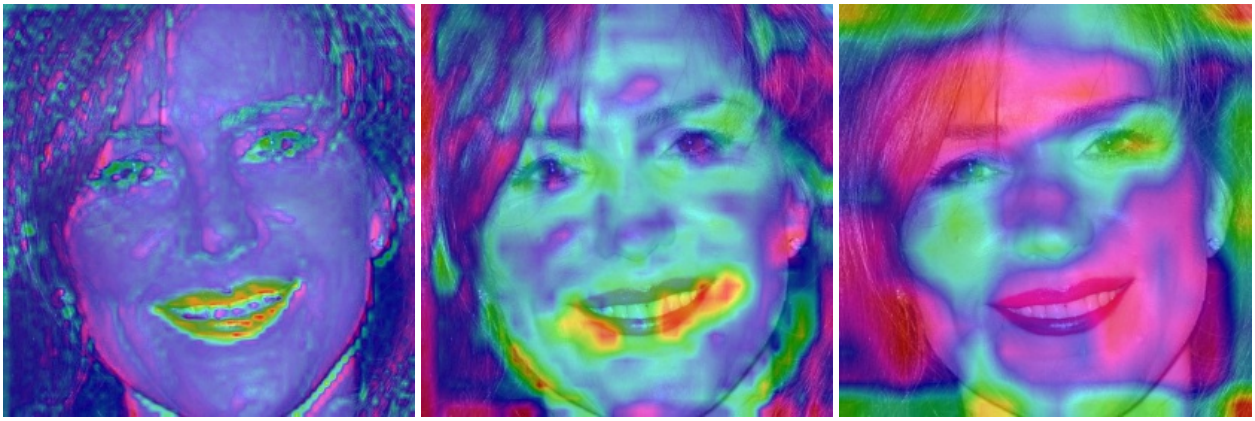
(e) Bill Hader

(f) Steve Carell

Figure 6: (Cont.) Original image (left), gradients of the image produced via guided backpropagation (middle), and associated saliency map (right).



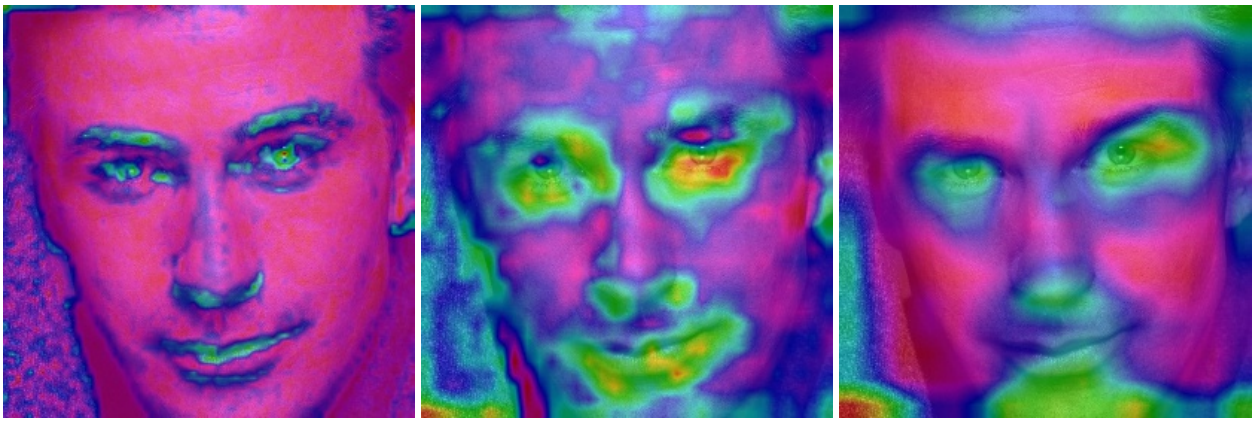
(a) Lorraine Bracco



(b) Peri Gilpin

(c) Angie Harmon

Figure 7: GradCAM after first convulutional layer (left), GradCAM after second convulutional layer (middle), and GradCAM at layer before fully connected layer (right).

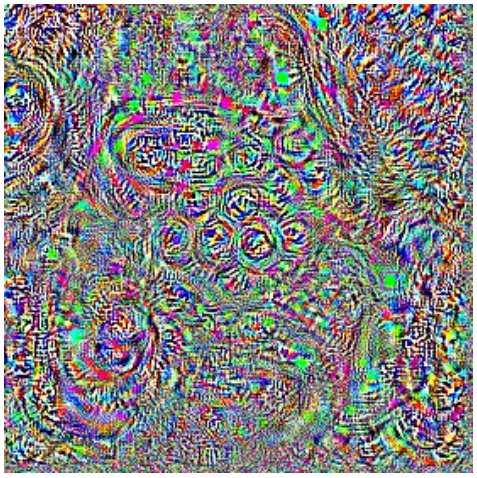


(d) Alec Baldwin

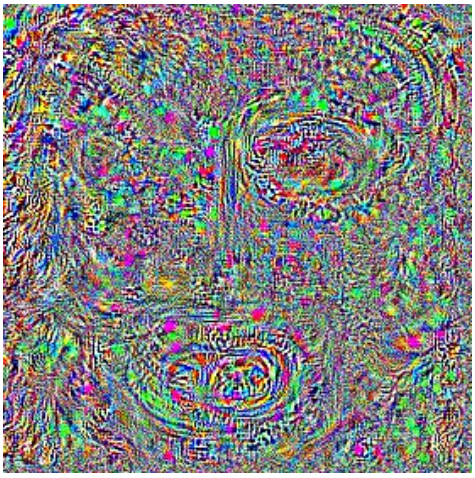
(e) Bill Hader

(f) Steve Carell

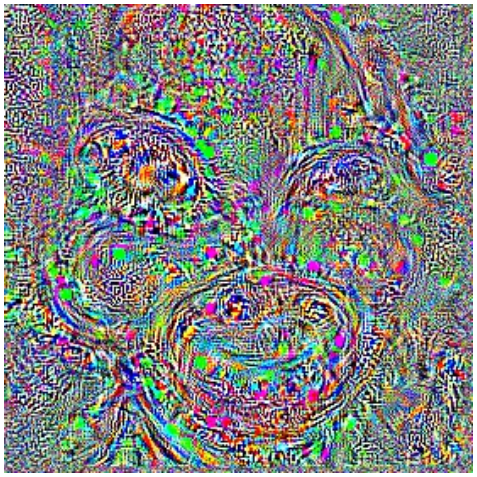
Figure 7: (Cont.) GradCAM after first convulutional layer (left), GradCAM after second convulutional layer (middle), and GradCAM at layer before fully connected layer (right).



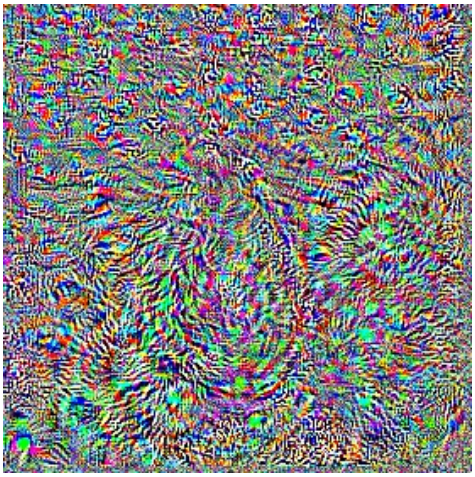
(a) Lorraine Bracco



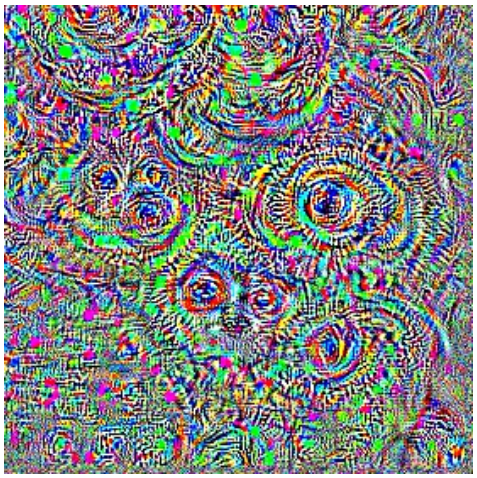
(b) Peri Gilpin



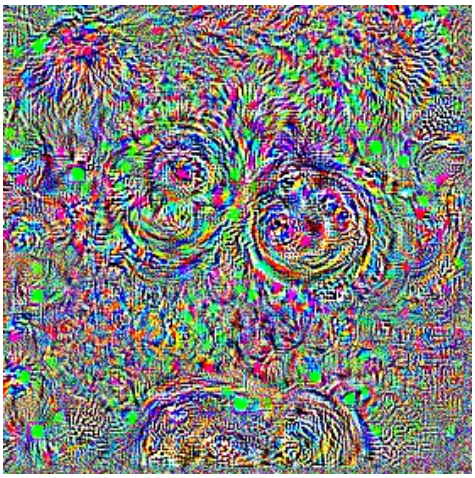
(c) Angie Harmon



(d) Alec Baldwin



(e) Bill Hader



(f) Steve Carell

Figure 8: Results of class specific image generation after 150 iterations with no regularization.

References

[1] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391.