Occlusion Sensitivity Analysis of AlexNet using MATLAB Natural Images

Cody Costa
KLA Masters of Engineering Program
Department of Electrical Engineering, San Jose State University
Milpitas, CA
cody.costa@sjsu.edu

Abstract—**In this work, we investigate the robustness of convolutional neural networks to local image occlusion using test images from MatLab's stock image log. Following the methodology introduced by Zeiler and Fergus (2014), we apply sliding-window occlusion to benchmark images and analyze the change in classification confidence among various mask sizing. We present sensitivity heatmaps, accuracy curves, and region-of-interest analyses on both natural objects (*llama*) and synthetic objects (*peppers*).**

Keywords—foreground, separation, image processing, binarization, python

I. Introduction

Deep convolutional networks achieve strong performance on image classification benchmarks, but the basis of their predictions is often difficult to interpret. Occlusion sensitivity analysis provides us a straightforward and concrete diagnostic. Through masking portions of the input image and measuring the effect on classification scores, one can infer which regions are critical for the network's decision as well as the probability of its interpretation being correct.

This report replicates to some extent the visualization experiments in Visualizing and Understanding Convolutional Networks by Zeiler and Fergus, adapting the method to AlexNet and canonical images.

II. METHODOLOGY

A. Goal/Specification

Record AlexNet's performance categorizing both the llama and peppers image with various mask sizing and positioning, emphasizing critical areas and distinguishing features within the frame.

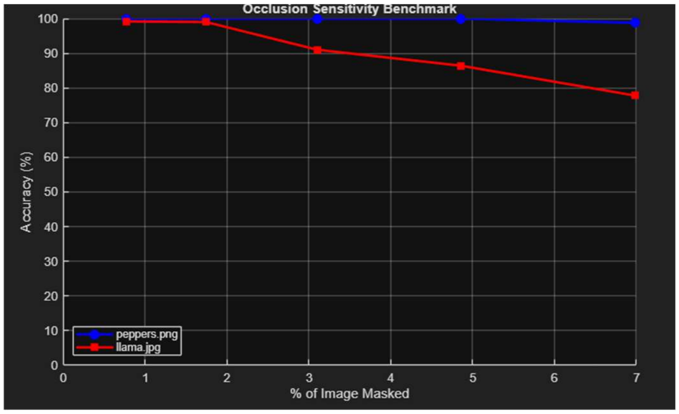
B. AlexNet Processing Procedure

Utilizing MATLAB's deep learning toolbox, AlexNet and its framework are readily available. Our process begins with loading said test images 'llama.jpg' and 'peppers.png', which come built into the IDE, and resizing them to 227 x 227 pixel matrices.

A variable pixel mask, or gray area, is then applied to the frame, moving discretely with a stride of 10 pixels at a time, and each new photo with its own unique mask positioning is fed into the network to approximate its class. We are able to record AlexNet's best initial guess and plot its accuracy over the occlusion cycles, allowing us to record certain ROI's (regions of interest) within the photograph. These ROI's become either the critical zones, or spots were certain deciding details are found that alert the CNN with high probability the photo matches the assumed classification, or the contrary; regions of confusion or ambiguity wherein no determinate guess can be calculated.

III. EXPERIMENTAL RESULTS

First, let's compare the size of the mask versus the occlusion sensitivity of the CNN. Using an increasing mask size of 20x20 up to 60x60 pixels, we can plot the performance of AlexNet:



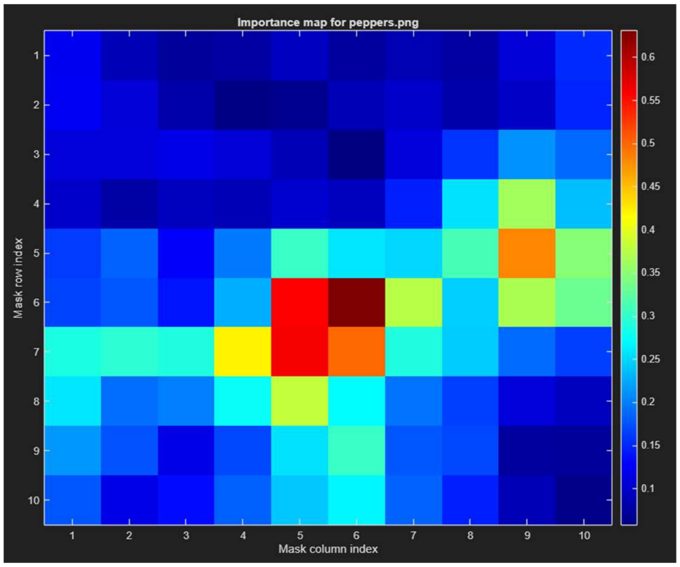
As more of the bell peppers photo is covered, AlexNet did not seem to flinch, with its guesses still scoring above 98% accurate in every case.

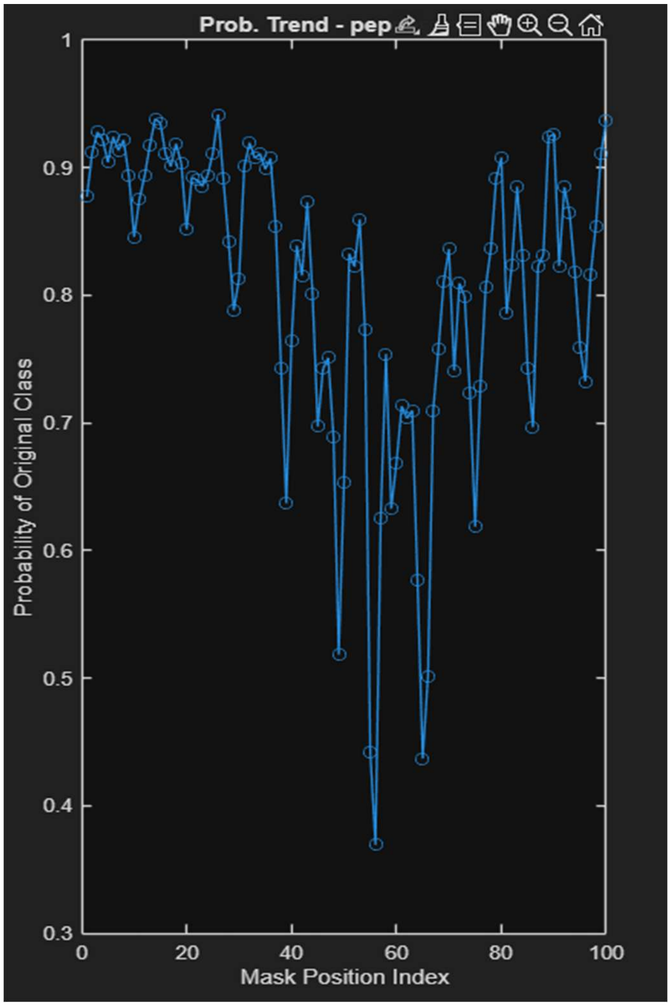
The variable coverage of the llama had worse outcomes. What started at 100% confidence eventually dwindled down with the mask size of 60x60 scoring just above 77%, still not bad.

Again, these results are an average of the total mask strides from start to finish for each coverage area.

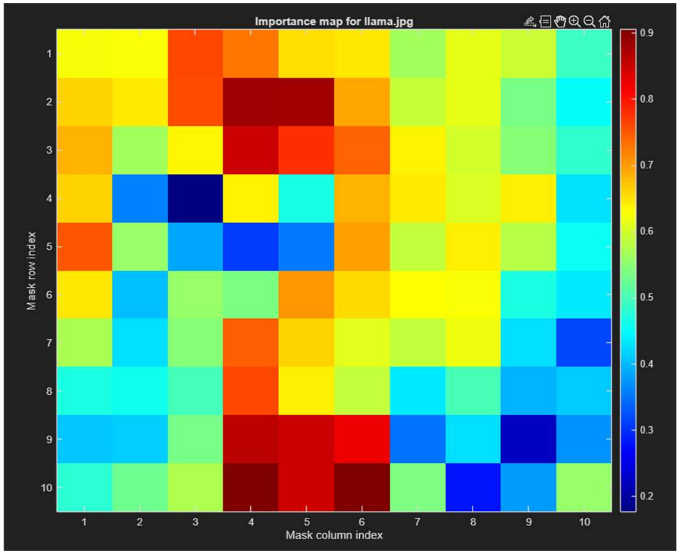
The mask motion and testing provides us with the following importance maps and probability trend as a function of mask location for the peppers and the llama. The mask index increments from left to right, top to bottom:

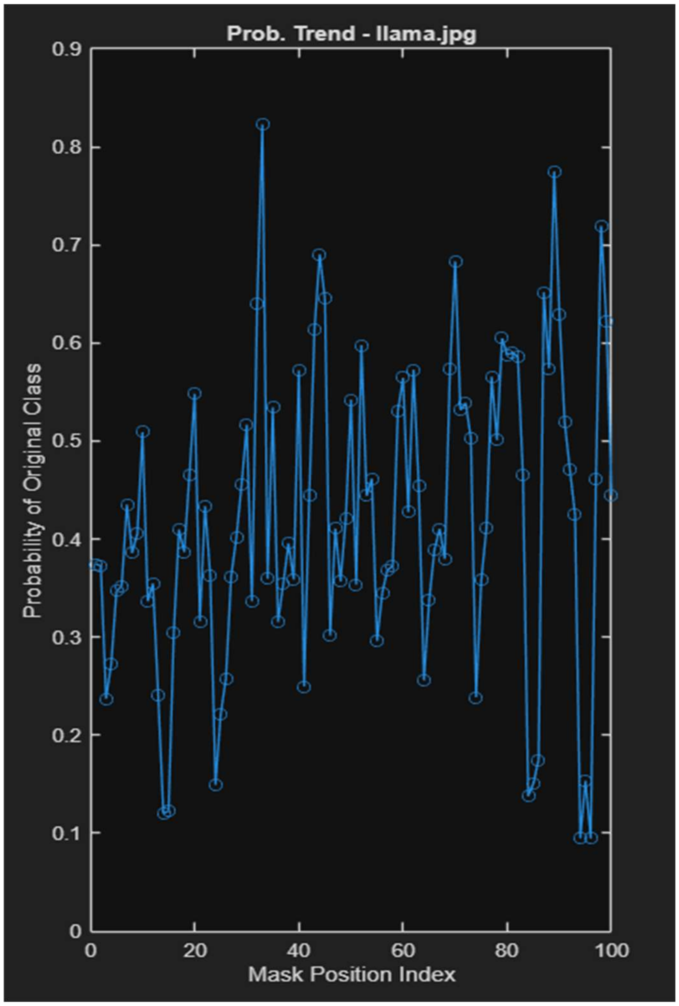
Peppers:





Llama:



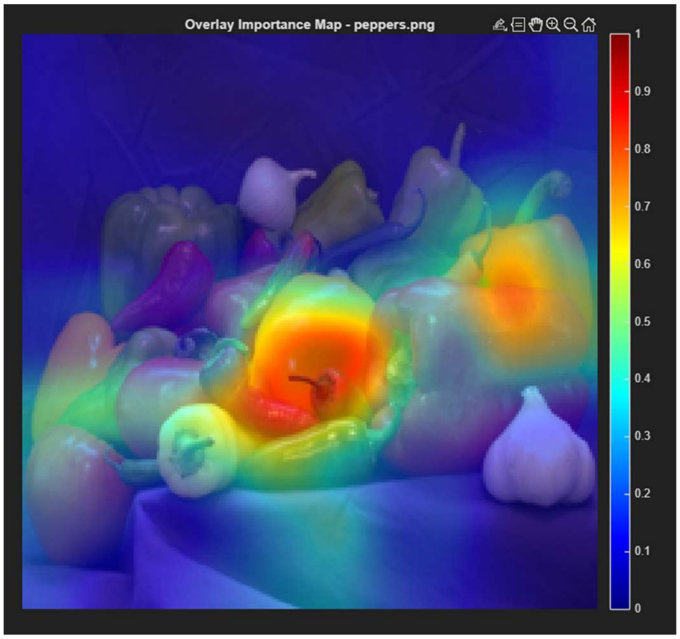


We see that the peppers image has a more predictable trend in the CNN’s probability to predict the correct object classification, with the stems of the fruits, particularly in the middle of the frame providing the strongest inference.

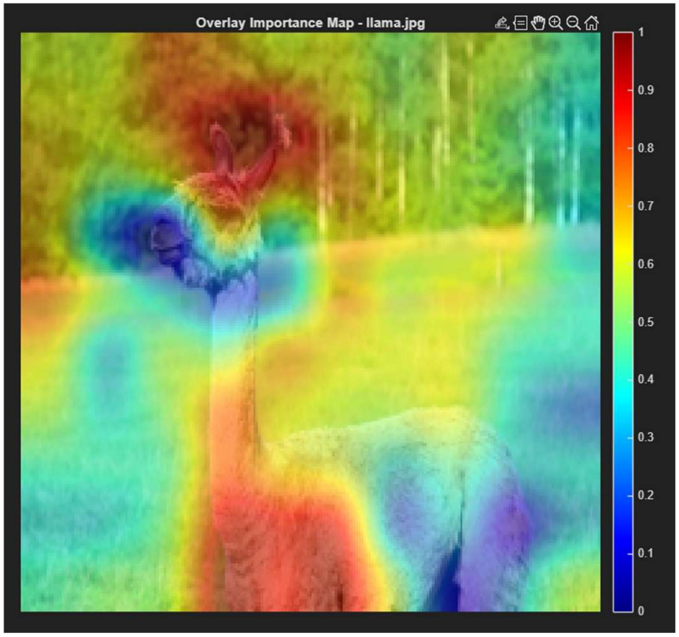
Contrarily, the llama showed a bit more random organization of the predicted outcome. The data does suggest that the llama's chest and lower neck, as well as the top of it's head near the ears, provided the strongest inference however.

Plotting and smoothing the importance mask over the base test image maps the ROI's previously discussed, where the heatmap displays the relative importance of each region:

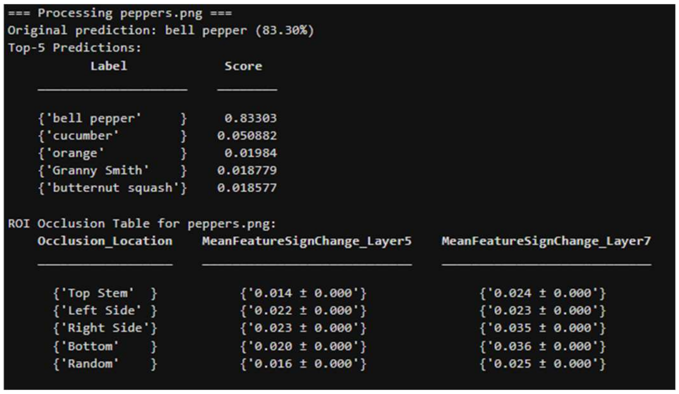
Peppers:



Llama:

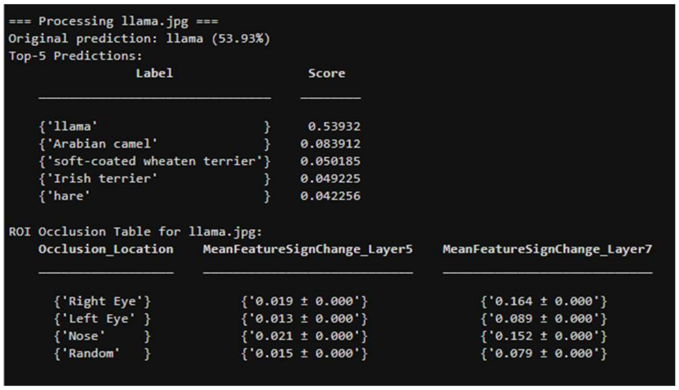


The results of occlusion tests showed that AlexNet had little to no issue predicting the classification of the bell peppers image, with an initial guess of 83.3% confidence. Regions such as the top stem, and left and right sides of the image (at about the horizontal midline of the image) provided the highest sign change during analysis:



```
=== Processing peppers.png ===
Original prediction: bell pepper (83.30%)
Top-5 Predictions:
           Label              Score
    ____________________    ________

    {'bell pepper'     }     0.83303
    {'cucumber'        }    0.050882
    {'orange'          }     0.01984
    {'Granny Smith'    }    0.018779
    {'butternut squash'}    0.018577

ROI Occlusion Table for peppers.png:
    Occlusion_Location    MeanFeatureSignChange_Layer5    MeanFeatureSignChange_Layer7
    __________________    ____________________________    ____________________________

      {'Top Stem'  }           {'0.014 ± 0.000'}               {'0.024 ± 0.000'}
      {'Left Side' }           {'0.022 ± 0.000'}               {'0.023 ± 0.000'}
      {'Right Side'}           {'0.023 ± 0.000'}               {'0.035 ± 0.000'}
      {'Bottom'    }           {'0.020 ± 0.000'}               {'0.036 ± 0.000'}
      {'Random'    }           {'0.016 ± 0.000'}               {'0.025 ± 0.000'}
```

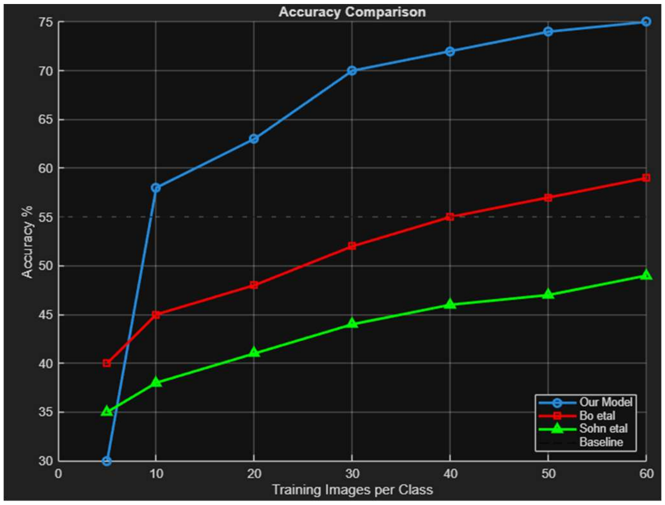
AlexNet tended to struggle with the llama more however. Its initial guesses were only 53.9% confident, which is understandable given the similarity between llamas and some of the close second guesses such as camels, and rabbits when comparing the ears. Regions that center around the animal's face were most heavily weighted:



```
=== Processing llama.jpg ===
Original prediction: llama (53.93%)
Top-5 Predictions:
                 Label                   Score
    _______________________________    ________

    {'llama'                      }     0.53932
    {'Arabian camel'              }    0.083912
    {'soft-coated wheaten terrier'}    0.050185
    {'Irish terrier'              }    0.049225
    {'hare'                       }    0.042256

ROI Occlusion Table for llama.jpg:
    Occlusion_Location    MeanFeatureSignChange_Layer5    MeanFeatureSignChange_Layer7
    __________________    ____________________________    ____________________________

      {'Right Eye'}           {'0.019 ± 0.000'}               {'0.164 ± 0.000'}
      {'Left Eye' }           {'0.013 ± 0.000'}               {'0.089 ± 0.000'}
      {'Nose'     }           {'0.021 ± 0.000'}               {'0.152 ± 0.000'}
      {'Random'   }           {'0.015 ± 0.000'}               {'0.079 ± 0.000'}
```

IV. CONCLUSIONS

The training model showed good promise upon review of the accuracy comparison. With 50 or more training images per class (50+ unique images per distinct image category) we saw the scores jump into the 75% or greater threshold, which is much better when comparing to the initial guess of only 54% roughly.



This finding shows that occlusion training is a viable method to tune one’s CNN to the next level as it forces the program to focus in on specific, crucial details to make an informed decision, which could provide healthier or more effective training data for others, should it be distributed.