Rafael Calleja and Suvansh Sanjeev
Professor Avideh Zakhor
EE 225B Spring 2018
11 May 2018

Final Project Report

Overview and Design

Our project was to use a saliency algorithm and a convolutional neural network to accomplish the task of producing a real-time classification and bounding box algorithm through the computer webcam in a Jupyter Notebook. We wrote the convolutional neural network using the Keras framework, using the following architecture, starting with 64x64x3 images:

- Convolutional layer with 32 3x3 filters and a ReLU activation
- Max pooling layer with 2x2 pool size
- Dropout layer with p=0.25 chance of dropping
- Convolutional layer with 32 3x3 filters and a ReLU activation
- Max pooling layer with 2x2 pool size
- Dropout layer with p=0.25 chance of dropping
- Flatten layer in order to feed into the fully connected layer
- Fully connected layer with 64 outputs and a ReLU activation
- Dropout layer with p=0.5 chance of dropping
- Fully connected layer with 5 outputs and a softmax activation for the final classification

For the saliency algorithm we followed a general, low-cost procedure to maximize utility while reducing overhead processing. This allowed our classifier to more easily pick out desired objects when it came into its field of view while still retaining its real-time feel. The saliency algorithm took in as input a frame from the webcam's stream:

- Mean Shift using OpenCV's multiscale pyramidal mean shift
- Back projection using the hue and saturation histograms of the entire image
- Smoothing the back projection again using mean shift
- Thresholding to estimate the salient areas
- Find the largest bounded region of saliency
- Mask original image with this created saliency region

We classified the following objects: exit sign, security camera, sprinkler, power outlet, red fire alarm.

Design Process Details

We initially selected the following five objects for classification in Moffitt Library on the 4th floor: exit sign, security camera, WiFi router, power outlet, and white fire alarm. We trained the CNN for an hour on a dataset of 36 training images, with 18 test images to evaluate performance. This yielded particularly strong performance on the exit sign, perhaps due to its distinguishing vibrant, glowing green hues. Performance was poor on the WiFi router, power outlet, and white fire alarm, however, which we speculate is due to the all-white nature of the images. We anticipated that the different shapes would be sufficient to differentiate them, but perhaps our relatively shallow network lacked the necessary capacity to adequately represent these higher-level features.

We then refined the list of objects, taking care to ensure diverse colors in addition to shapes. We decided to use the security camera, exit signs, red fire alarms, power outlets, and ceiling sprinklers. We were concerned that the security camera and ceiling could look too similar, but

decided that the difference in size as well as the stronger black region of the security camera would be adequate for differentiation. Additionally, we were concerned that the previous model was overfitting to the training data, since there were relatively few images, and since the validation error was increasing towards the end of training. To alleviate this concern, we took the following actions:

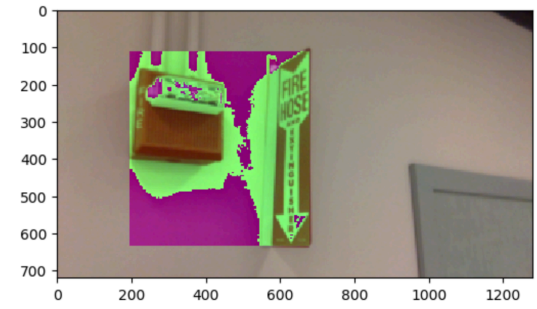
1. We expanded our dataset to a total of 120 images (80 train, 40 test).
2. We added the three dropout layers described in the Overview and Design section, which serve as regularization at the level of the model[1]
3. We decreased the number of training epochs to two, at which point the lowest validation error was observed in previous training runs

Results

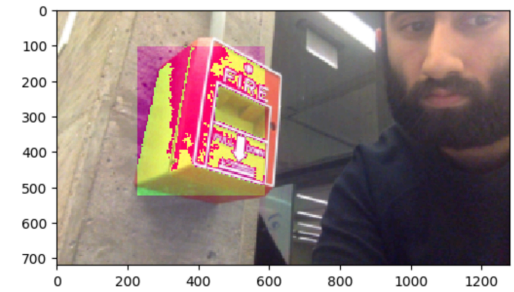
At this point, we had our final model. Our efforts to prevent overfitting seemed to make a noticeable difference, as seen below. The framerate of the live classification is around 0.4 Hz. The video is very noticeably choppy, but sufficiently smooth as to be easily recognizable as real-time. Classification screenshots are taken from live, real-time usage of the software, as opposed to still images fed to the CNN:

```
Epoch 1/2
 - 2088s - loss: 0.7692 - acc: 0.6942 - val_loss: 0.2845 - val_acc: 0.9750
Epoch 2/2
 - 2063s - loss: 0.2792 - acc: 0.9132 - val_loss: 0.1048 - val_acc: 0.9750
```

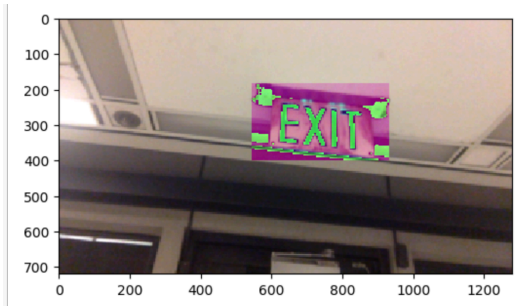
Successes



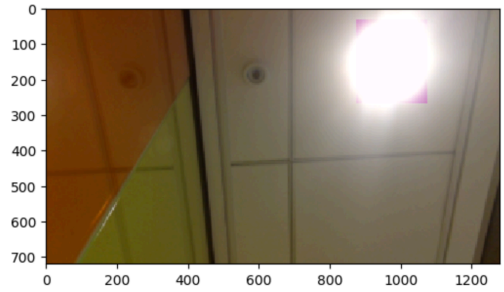
fire alarm



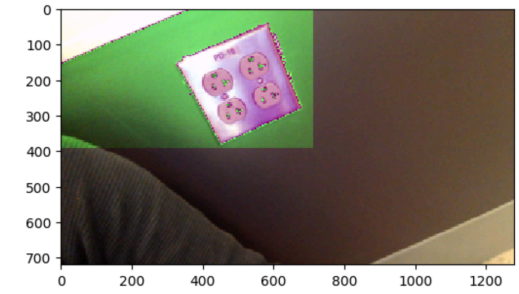
fire alarm



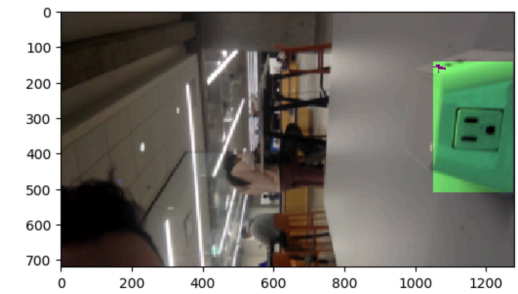
exit sign



sprinkler



power outlet

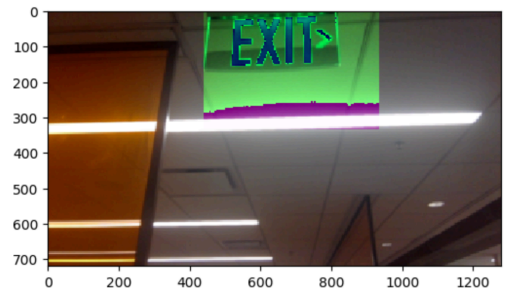


power outlet

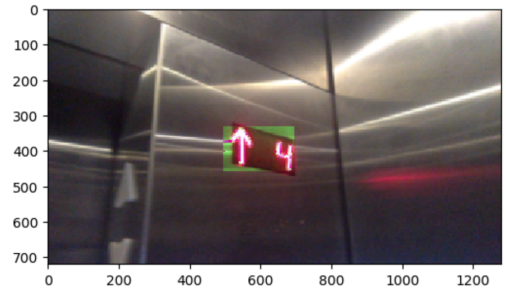
Particularly impressive generalization on the power outlets in the last two images serves as evidence of the reduction in overfitting. The CNN was trained only on standard two-outlet images, such as the following:

Failures

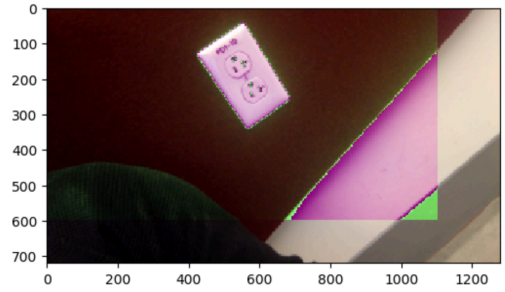
However, even in our final iteration of the CNN, there were still classification missteps, particularly if the object of interest was obscured or the background was cluttered. We also gave the CNN some inputs that belonged to none of the learned classes, expecting a certain classification, such as the elevator sign below, which was predictably classified as a fire alarm due to the striking red hues. The exit sign and power outlet images may have suffered from a similar fate due to the orange glass and red velvet backing, respectively, which would have yielded a high activation on a red detection filter, which the CNN may have learned for classification of the fire alarms.



fire alarm



fire alarm



fire alarm

This is a sample image of the fire alarm from the training set, for comparison:

Improvements

The framerate, as noted above, is rather low, and could be increased either through conversion of images to grayscale, in effect decreasing the input size by two-thirds, or by decreasing the input image resolution to the same effect. Additionally, we selected a square-shaped region in

the middle of the webcam frame to classify, which means classification fails on objects at the edge of the webcam view. The fix to this would be to involve the saliency algorithm not only in drawing the bounding box, but also in selecting the region of the webcam view to pass to the CNN. The reason we chose not to implement this is due to instability of the saliency algorithm. While it empirically usually does a reasonable job of highlighting the object of interest, it is not accurate enough to rely on for classification, and the salient region selected can move around rapidly with slight movements of the webcam.

References

1. Srivastava, Nitish, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. 2014.
2. Gil, Jacob. "Simple Image Saliency Detection from Histogram Backprojection." *Jacob's Computer Vision and Machine Learning Blog*, 24 Apr. 2015, jacobgil.github.io/computervision/saliency-from-backproj.
3. Squirrel. "Webcam Based Image Processing in IPython Notebooks – Squirrel – Medium."*Medium*, Augmenting Humanity, 1 Nov. 2016, medium.com/@neotheicebird/webcam-based-image-processing-in-ipython-notebooks-47c75a022514.
4. Keras Documentation. https://keras.io/getting-started/sequential-model-guide/