Machine Cashier for Automatic Bill Calculation in Bakeries

Roy Chan
Stanford University
roychan@stanford.edu

Abstract

A convolutional neural network was developed to compute the final price of bakery products chosen by customers. Images of randomly oriented pastries were fed into a region proposal algorithm, which relayed selected regions into the convolutional network for item classification. Upon classification, the total price of the pastries in each image was computed. This proof of concept showcases a futuristic implementation of bill computation at retail stores, where there is no reliance on barcodes, RFID tags and mundane human labor.

1. Introduction

In the future of retail, there is likely to be less reliance on human labor for mundane tasks such as money collection and bill computation. At bakery stores, bill computation is mostly done manually by a cashier either by mental calculation or by selecting the item to be purchased from a touch screen. The human vision system is used for object detection and classification in the process. We envision a machine vision system that can perform the same task, detecting the number of pastries to be bought by the customer, classifying them and computing the final price to be paid. In our system, we applied transfer learning by utilizing pretrained weights from the YOLO convolutional neural network and fine tuned additional fully connected neural network layers. The bakery product dataset used to fine tune the network was 4 pastry classes from ImageNet. As this dataset did not come with bounding boxes, two region proposal algorithms were tested to provide regions of interest for the convolutional neural network. Upon object classification, only high probability regions were kept, and non maximal suppression was used to remove overlapping regions. A lookup table with the pastry prices could then be referenced for the autonomous display of the final bill.

2. Related Work

There have been numerous instances of machine vision applied to bakery products. Notable papers related to our work include the use of a support vector machine and color thresholding to inspect biscuits on a conveyor belt [1] and the use of discriminant analysis for the classification of muffins [2]. Various region proposal algorithms have been developed, such as Bing [3], Edgeboxes [4], and Selective Search [5], all of which can be adapted to work with our machine cashier. There are also many pretrained convolutional neural network architectures such as AlexNet [6], VGG [7] and ResNet [8] that can be used for the image classification part of our machine cashier. Additionally, convolutional neural networks can be made to do simultaneuous object detection and classification, such as is the case with YOLO [9] and Faster-RCNN [10].

3. Methods

An image containing bakery pastries is fed into a region of interest proposal algorithm. Two such algorithms were tested, Selective Search, and a custom adaptive threshold based algorithm. For Selective Search, the image was scaled down in size in order to achieve fast detection times of ~50ms for real time image detection. However, as Selective Search did not always propose all regions of interest, a custom adaptive threshold algorithm was used to pick out bright regions against a dark background. The cropped regions proposed by the ROI algorithm was passed into the convolutional neural net, with an architecture of 6 consecutive convolutional and pooling layers, followed by 3 convolutional layers, and 3 fully connected layers. Softmax regression was used to compute the loss during training. The final 3 fully connected layers were fine tuned on the bakery dataset for 2 days, achieving a training accuracy of 50%. Validation accuracy at the end of the training phase was also 50%. For price computation, only high probability regions were considered, overlapping regions were removed with non maximal suppression, and a lookup table containing the price of each pastry was referenced. The implementation was in

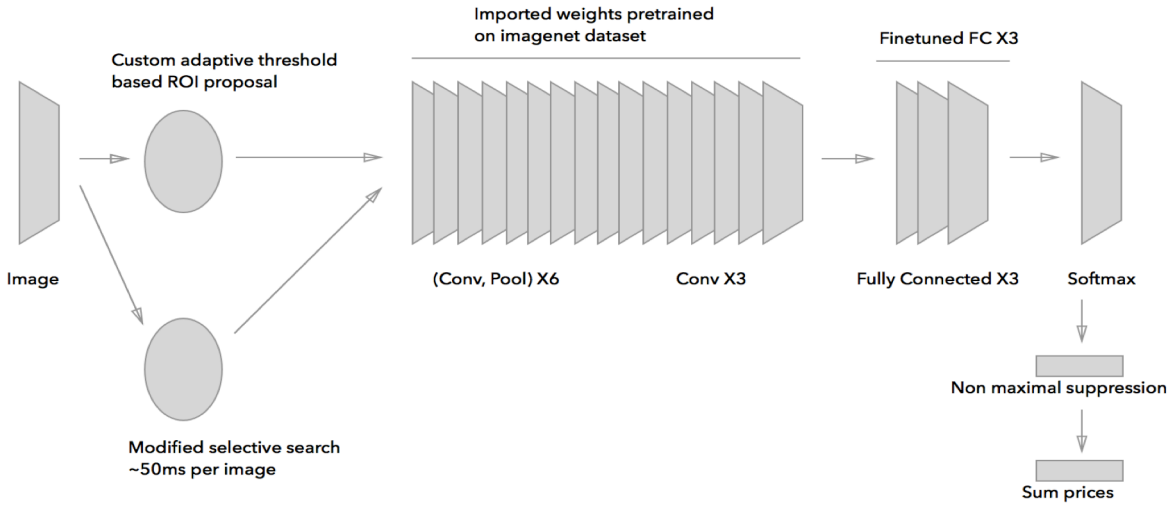


Figure 1: Neural network architecture of machine cashier. One of two region proposal algorithms feed proposed regions (cropped to 448x448) into an 18 layer convolutional network (15 layers from pretrained YOLO, and 3 finetuned fully connected layers). A Softmax function computes the loss and non maximal suppression (NMS) removes overlapping proposals.

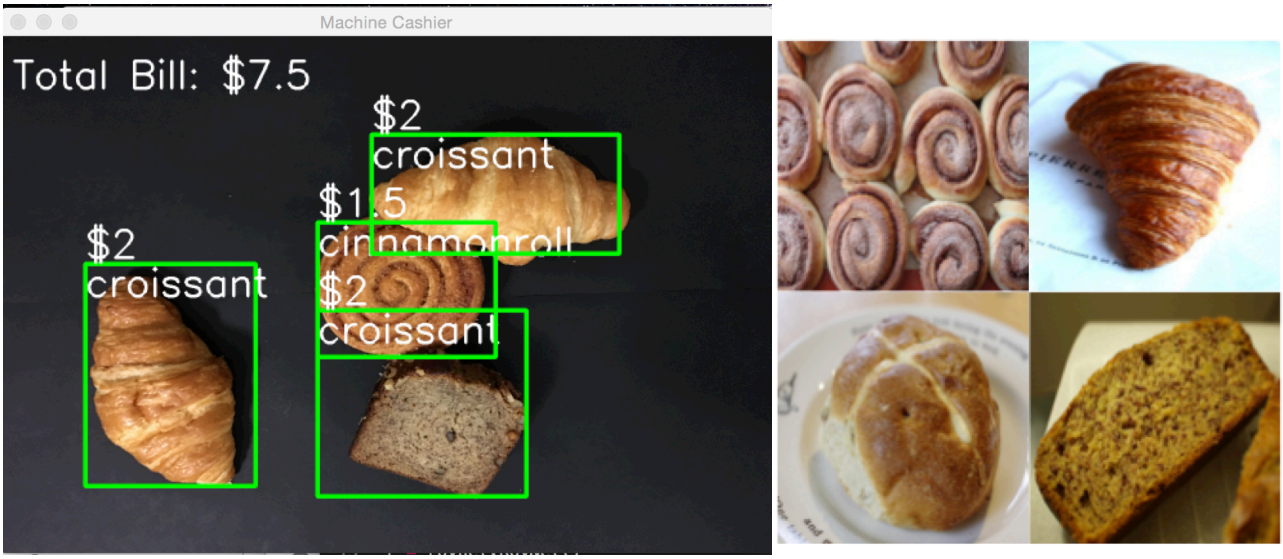


Image 1 (Left): Bill computation in action by machine cashier on Starbucks pastries.
Image 2 (Right): Sample ImageNet images of the 4 classes of pastries used for fine tuning the neural network.

python and the framework for the neural network was TensorFlow.

4. Dataset and Features

The pretrained weights from YOLO convolutional neural network were pretrained on ImageNet data. For the bakery dataset, 4 classes of pastries were obtained form ImageNet, croissant, banana bread, cinnamon roll, and hot cross buns. There were 1000 - 1500 images for each class. These classes were chosen among the bakery produce classes in ImageNet because they offered more or less consistent pastry appearance from image to image. Other available classes such as 'cake' and 'bread' were dropped there was too much variation in product appearance from image to image.

5. Discussion and Results

The original YOLO had region proposal built into it's convolutional neural network, and our intended goal was to utilize the internal regional proposal for fast, real time

object detection and recognition. However, due to the unavailability of a bakery product dataset with bounding box labels, we had to compromise by using an external region proposal algorithm such as selective search or our custom adaptive thresholding region proposer. The external region proposals resulted in additional computational time ranging from 15 seconds on unmodified selective search to 0.05ms on our adaptive thresholding region proposer. By shrinking the image that was fed into selective search to 100x100 pixels, we managed to speed up the computational time per image to 50ms, but at the expense of proposal quality. However, these algorithms had difficulty when pastry were overlapped or in very close proximity to each other, and were not always able to segregate all pastries into their respective bounding boxes. When the background of the pastry test image was complex, or of the same color as the pastries, very poor region proposals arose. Therefore, a tray with a black background was used to hold the pastries in the validation images in order to create images with high contrast between the objects and background. Having the black background helped significantly, as the object proposals were more accurate.

Fine tuning of the final 3 fully connected layers of the convolutional network was done over 2 days until a plateau was reached for training and validation accuracies. Adam was used at the initial stage of fine tuning, but swapped with standard gradient descent when faced with difficulty in obtaining the right hyperparameters for convergence. The learning rate on standard gradient descent was gradually decreased from 0.01 to 0.00001 by powers of 10 so as to achieve a lower loss value. One reason for the low training and validation accuracies of ~50% could be the large amount of variation between appearance of the pastries in the dataset. For example, looking at banana bread, they all look very different from image to image, with some being sliced open, and some only showing the outer brown crust. Unlike banana bread, hot cross buns had a more consistent look, with most being square and having a cross pattern over the top. To improve the training and validation accuracies, a plausibly good method would be to procure photos of multiple angles of the exact pastry from the bakery it was made in, and to only detect that specific pastry. Also, all the pastries were brown, and the YOLO network might be too small (or be of the wrong architecture) to finely classify the nuances of different brown objects with almost similar features. In this case, a more powerful network such as ResNet might give better results.

Non maximal suppression was used to remove the overlapping regions, but this algorithm often grouped together separate pastries as a single object, thereby resulting in the wrong bill price. Non maximal suppression is an inelegant hack meant to get rid of overlapping bounding boxes, and truly elegant future solutions to object detection should not rely on it.

Additionally, as the yet to be optimized Tensorflow beta release (version r0.7) was used as the framework for the neural network, each forward pass took 300ms, and a single image with 5 object proposals took up to 1.5s. This excludes the capability of the current implementation of machine cashier for real time object detection and recognition. OpenCV was used to capture frame by frame images from the webcam of a macbook for the processing by the machine cashier. It was found that frame to frame images often had low reproducibility if a frame was just shifted slightly. Adaptive thresholding gave significantly better frame to frame reproducibility than selective search as a region proposer.

Finally, to test the capability of this machine cashier, several pastries were bought from Starbucks and placed at random positions on the black tray. Cinnamon rolls and croissants were classified with ease, likely due to their distinctive surface pattern, but not banana bread. The biggest issue with the machine cashier came from region proposals, as obviously positioned pastries were often missing in the proposed regions. Selective search, even though state of the art, is a very weak algorithm for region proposal when compared to human level capability (it certainly will get the bakery shop boss angry by missing out pastries in the bill computation). Therefore, there is high certainty future working implementations of machine cashers will not depend on selective search.

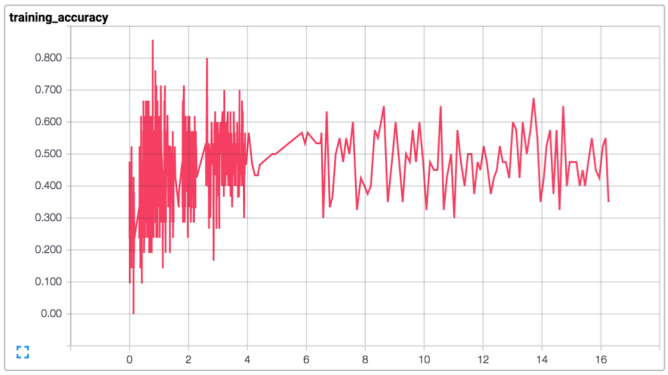
6. Conclusion

The machine cashier presented here is a proof of concept to show how it might be possible for retail product prices to be computed not through barcode scanning, RFID, or manual price summing by humans, but though the swift eyes of a machine vision algorithm. Before such machine cashiers become commercially viable, object detection and classification accuracy has to be improved.

References

[1] S. Nashat, A. Abdullah, S. Aramvith, M.Z. Abdullah, Support vector machine approach to real-time inspection of biscuits on moving conveyor belt, Computers and Electronics in Agriculture, Volume 75, Issue 1, January 2011.

[2] Mohd Zaid Abdullah, Sabina Abdul Azizi, and Abdul Manan Dos Mohamed, Quality Inspection of Bakery Products Using A Color-based Machine Vision System. Journal of Food Quality, Volume 23, Issue 1, pages 39–50, March 2000.



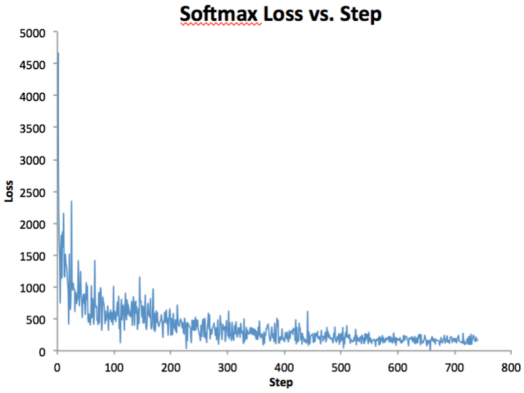
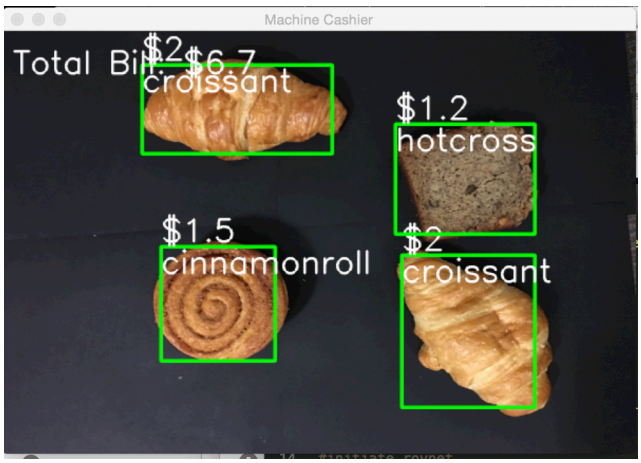


Figure 3 (left): Training accuracy plot versus time (screenshot from tensorboard API).
Figure 4 (right): Softmax loss versus training step.



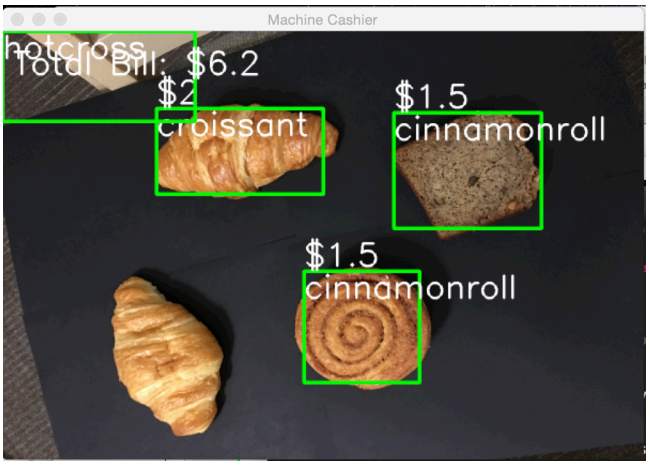


Image 3: Successful object detection but misclassifcation of banana bread as hotcross buns
Image 4: Misclassification of background object as hotcross bun.

Image 5 (left): Clustered pastries in box with brown background caused Selective Search to work poorly. Hence the introduction of a black background to improve contrast.
Image 6 (right): Adaptive thresholding of pastries against a black background allows brighter pastry regions to be selected.

[3] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, Philip Torr,, BING: Binarized Normed Gradients for Objectness Estimation at 300fps.
[4] C. Lawrence Zitnick and Piotr Dolĺar, Edge Boxes: Locating Object Proposals from Edges.
[5] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders,Selective Search for Object Recognition, IJCV.
[6] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks.
[7] Karen Simonyan and Andrew Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition
[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition.
[9] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You Only Look Once: Unified, Real-Time Object Detection.
[10] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.