

# A Comparison of Embedded Deep Learning Methods for Person Detection

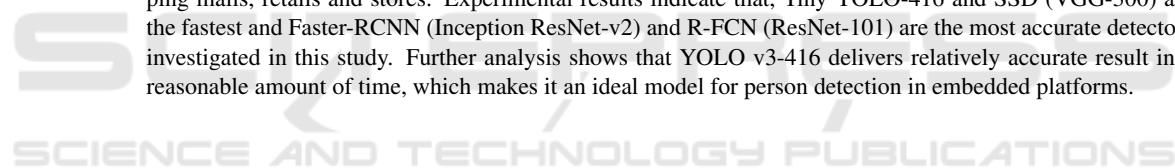
Chloe Eunhyang Kim<sup>1</sup>, Mahdi Maktab Dar Oghaz<sup>2</sup>, Jiri Fajtl<sup>2</sup>, Vasileios Argyriou<sup>2</sup>  
and Paolo Remagnino<sup>2</sup>

<sup>1</sup>VCA Technology Ltd, Surrey, U.K.

<sup>2</sup>Kingston University, London, U.K.

**Keywords:** Embedded Systems, Deep Learning, Object Detection, Convolutional Neural Network, Person Detection, YOLO, SSD, RCNN, R-FCN.

**Abstract:** Recent advancements in parallel computing, GPU technology and deep learning provide a new platform for complex image processing tasks such as person detection to flourish. Person detection is fundamental preliminary operation for several high level computer vision tasks. One industry that can significantly benefit from person detection is retail. In recent years, various studies attempt to find an optimal solution for person detection using neural networks and deep learning. This study conducts a comparison among the state of the art deep learning base object detector with the focus on person detection performance in indoor environments. Performance of various implementations of YOLO, SSD, RCNN, R-FCN and SqueezeDet have been assessed using our in-house proprietary dataset which consists of over 10 thousands indoor images captured form shopping malls, retails and stores. Experimental results indicate that, Tiny YOLO-416 and SSD (VGG-300) are the fastest and Faster-RCNN (Inception ResNet-v2) and R-FCN (ResNet-101) are the most accurate detectors investigated in this study. Further analysis shows that YOLO v3-416 delivers relatively accurate result in a reasonable amount of time, which makes it an ideal model for person detection in embedded platforms.



## 1 INTRODUCTION

The rise of industry 4.0, IoT and embedded systems pushes various industries toward data driven solutions to stay relevant and competitive. In the retail industry, customer behavior analytic is one of the key elements of data driven marketing. Metrics such as customer's age, gender, shopping habits and moving patterns allow retailers to understand who their customers are, what they do and what they are looking for. These metrics also enables retailers to push customized and personalized marketing schemes to their customers across various stages of the customer life-cycle. Additionally, with the help of predictive models, retailers are now enable to predict what their customers are likely to do in the future and gain edge over their competitors. In recent years, there has been an increasing interest in the analysis of in-store customer behavior. Retailers are looking for insights on in-store customer's journey; Where do they go? What products do they browse? and most importantly, which products do they purchase (Ghosh et al., 2017) (Majeed and Rupasinghe, 2017) (Balaji and Roy, 2017)?

Over the last decade, several tracking approaches such as sensor based, optical based and radio based have been proposed. However, The majority of them are not efficient and reliable enough, or they expect some form of interaction with customers which might compromise their shopping experience (Jia et al., 2016)(Foxlin et al., 2014). Analysis of in-store customer behavior through optical video signal recorded by security cameras has clear advantage over other approaches as it utilizes the existing surveillance infrastructure and operates seamlessly with no interaction and interference with customers (Ohata et al., 2014)(Zuo et al., 2016). Despite the clear advantage of this approach, analysis of video signal requires complex and computationally expensive models, which up until recent years, was impractical in the real world. Recent advancements in parallel computing and GPU technology diminished this computational barrier and allowed complex models such as deep learning to flourish (Nickolls and Dally, 2010).

Aside from hardware limitations, classic computer vision and machine learning techniques had hard time to model these complex patterns, however the

rise of data driven approaches such as deep learning, simplified these tasks, eliminating the need for domain expertise and hard-core feature extraction. A reliable yet computationally reasonable person detection model is fundamental requirement for in-store customer behavior analysis. Numerous studies focused on person detection using deep neural network models. However, none of which particularly focused on the person detection in indoor retail environments. Despite the similarity of these topics, there are a number of unique challenges, such as lighting condition, camera angles, clutter and queues in retail environments, which questions the adaptability of the existing person detection solutions for retail environments.

In this regard, this research is mainly focused on person detection as a preliminary step for in-store customer behavior modeling. We are particularly interested in evaluation and comparison of deep neural network (DNN) person detection models in cost-effective, end-to-end embedded platforms such as the Jetson TX2 and Movidius. State of the art deep learning models use general purpose datasets such as PASCAL VOC or MS COCO to train and evaluate. Despite their similarities, these dataset cannot be true representative of the retail and store environments. In data driven techniques such as deep learning, this adaptability issues are more pronounced than ever before (LeCun et al., 2015). To address these issues, this research investigates the performance of state of the art DNN models including variations of YOLO, SSD, RCNN, R-FCN and SqueezeDet in person detection using an in-house proprietary image dataset were captured by conventional security cameras in retail and stores environments.

These images were manually annotated to form the ground truth for training and evaluation of the deep models. Having deep models trained by the similar type of images that could be found in target environment, can significantly improve the accuracy of the models. However, preparation of a very large annotated dataset is a big challenge. This research employs *average precision* metric at various *intersection over union* (IoU) as the figure of merit to compare model performance. As processing speed is a key factor in embedded systems, this research also conducts a comprehensive comparison among the aforementioned DNN techniques to find the most cost-effective approach for person detection in embedded systems.

The major contributions of this study can be summarized as: first, integration and optimization of the state of the art person detection algorithm into embedded platforms; second, an end-to-end comparative study among the existing person detection models in

terms of accuracy and performance and finally, a proprietary dataset, which can be used in indoor human and analysis studies.

The paper is organized as follow. Section 2 briefly describes the state of art object detection models used in this research. Section 3 presents the overall framework, data acquisition process as well as experimental setup of the research. Section 4 describes the experimental results and discussions and finally, sections 5 concludes the research.

## 2 CNN BASED OBJECT DETECTION

Various DNN based object detector have been proposed in the last few years. This research investigates the performance of state of the art DNN models including variations of YOLO, SSD, RCNN, R-FCN and SqueezeDet in person detection. The models have been trained using an in-house proprietary image dataset were captured by conventional security cameras in retail and stores environments. The following sections describes aforementioned DNN models in more details.

### 2.1 RCNN Variants

The region-based convolutional neural network (RCNN) solution for object detection is quite straightforward. This technique uses selective search to extract just 2000 regions (region proposal) from the image and then, instead of trying to classify a huge number of regions throughout the image only these 2000 region will be investigated. Selective search initially generates candidate regions, then uses a greedy algorithm to recursively combine similar regions into larger ones. Finally, it uses the generated regions to produce the final candidate region proposals. The region proposals will be passed to a conventional neural network (CNN) for classification. Despite RCNN has lots of advantages over the conventional DNN object detector (Girshick et al., 2016), this technique is still quite slow for any real-time application. Furthermore, a predefined threshold of 2000 region proposal cannot be suitable for any given input image.

To address these limitations, other variants of RCNN have been introduced (Ren et al., 2015). Faster RCNN is one popular variant of RCNN which mainly devised to speed up RCNN. This algorithm eliminates the selective search algorithm used in the conventional RCNN and allows the network learn the region proposals. The mechanism is very similar to fast RCNN where an image is provided as input to a

CNN to generate a feature map but, instead of using a selective search algorithm on the feature map to identify the region proposals, a separate network is used to predict region proposals. The predicted region proposals are then reshaped using a region of interest (RoI) pooling layer and used to classify the image input within the proposed region (Ren et al., 2015). To train the Region Proposal Network, a binary class label has been assigned to each anchor (1: being object and 0: not object). Any with IoU over 0.7 determines object presence and anything below 0.3 indicates no object exists. With these assumptions, we minimize an objective function following the multi-task loss in Fast R-CNN which is defined as:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i N_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i P_i^* L_{reg}(t_i, t_i^*) \quad (1)$$

where  $i$  is the index of anchor in the batch,  $p_i$  is its predicted probability of being an object;  $p_i^*$  is the ground truth probability of the anchor (1: represents object, 0: represents non-object);  $t_i$  is a vector which denotes the bounding box coordinates;  $t_i^*$  is ground truth bounding box coordinates;  $L_{cls}$  is classification log loss and  $L_{reg}$  is regression loss. We have also deployed the Faster RCNN model using the Google inception framework which is expected to be less computational intensive.

## 2.2 R-FCN Variants

In contrast to the RCNN model which applies a costly per-region subnetwork hundreds of times, region based fully convolutional network (R-FCN) is an accurate and efficient object detector that spreads the computation across the entire image. A position-sensitive score map is used to find a trade-off between translation-invariance in image classification and translation-variance in object detection. A position-sensitive score defined as following:

$$r(i, j | \theta) = \sum_{(x,y) \in bin(i,j)} z_{i,j,c}(x+x_0, y+y_0 | \theta) / n \quad (2)$$

where  $r_c(i, j)$  is the pooled response in the  $(i, j)^{th}$  bin in the  $c^{th}$  category;  $z_{i,j,c}$  is one score map out of the  $k^2(C+1)$  score map;  $n$  in the number of the pixels in the bin;  $(x_0, y_0)$  represents the top left corner of the region of interest and  $\theta$  denotes network learning parameters. The loss function defined on each region of interest which calculated by summation of the cross entropy loss and box regression loss as following:

$$L(s, t_{x,y,w,h}) = L_{cls}(s_{c^*}) + \lambda[c^* > 0] L_{reg}(t, t^*) \quad (3)$$

where  $c^*$  is the region of interest ground truth label;  $L_{cls}(s_{c^*})$  is cross entropy loss for classification;  $t^*$  represents the ground truth box and  $L_{reg}$  is the bounding box regression loss. Aside from the original R-FCN, this study also investigates the R-FCN model with the Google inception framework (Dai et al., 2016).

## 2.3 YOLO Variants

You only look once (YOLO) is another state of the art object detection algorithm which mainly targets real time applications. it looks at the whole image at test time and its predictions are informed by global context in the image. It also makes predictions with a single network evaluation unlike models such RCNN, which require thousands for a single image. YOLO divides the input image into an  $S \times S$  grid. If the center of an object falls into a grid cell, that cell is responsible for detecting that object. Each grid cell predicts five bounding boxes as well as confidence score for those boxes. The score reflects how confident the model is about the presence of an object in the box. For each bounding box, the cell also predicts a class. It gives a probability distribution score over all the possible classes designate the object class. Combination of the confidence score for the bounding box and the class prediction, indicates the probability that this bounding box contains a specific type of object. The loss function is defined as:

$$\begin{aligned} \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \\ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] + \\ \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \\ \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \quad (4)$$

where  $1_i^{obj}$  indicates if object appears in cell  $i$  and  $1_{ij}^{obj}$  denotes the  $j^{th}$  bounding box predictor in cell  $i$  responsible for that prediction;  $x, y, w, h$  and  $C$  denote the coordinates represent the center of the box relative to the bounds of the grid cell, the width and height are predicted relative to the whole image and finally  $C$  denotes the confidence prediction represents

the IoU between the predicted box and any ground truth box. This study also investigates the other variants of YOLO including YOLO-v2 as well as Tiny YOLO models performance for person detection in retail environments (Redmon et al., 2016)(Redmon and Farhadi, 2017).

## 2.4 SSD Variants

Single shot multi-box detector (SSD) is one of the best object detector in terms of speed and accuracy. The SSD object detector comprises two main steps including *feature maps extraction*, and *convolution filters application* to detect objects. A predefined bounding box (prior) is matched to the ground truth objects based on IoU ratio. Each element of the feature map has a number of default boxes associated with it. Any default box with an IoU of 0.5 or greater with a ground truth box is considered a match. For each box, the SSD network computes two critical components including *confidence loss* which measures how confident the network is at the presence of an object in the computed bounding box using categorical cross-entropy and *location loss* which computes how far away the networks predicted bounding boxes are from the ground truth ones based on the training data (Huang et al., 2017)(Liu et al., 2016). The overall loss function is defined as following:

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (5)$$

where  $N$  is the number of matched default boxes. Other variants of the standard SSD with 300 and 512 inputs as well as MobileNet and Inception models has been implemented and tested in this research (Howard et al., 2017)(Szegedy et al., 2015).

## 2.5 SqueezeDet

SqueezeDet is a real-time object detector used for autonomous driving systems. This model claims high accuracy as well as reasonable response latency, which are crucial for autonomous driving systems. Inspired by the YOLO, this model uses fully convolutional layers not only to extract feature maps, but also to compute the bounding boxes and predict the object classes. The detection pipeline of SqueezeDet only contains a single forward pass over the network, making it extremely fast (Wu et al., 2017). SqueezeDet can be trained end-to-end, similarly to the YOLO and it shares similar loss function with YOLO object detection.

## 3 RESEARCH FRAMEWORK

Similar to any other machine learning task, this research employs training/testing and validation strategy to create the prediction models. All CNN models were trained and tested using our proprietary dataset. Predictions were compared against ground truth by means of cross entropy loss function to back propagate and optimize network weights, biases and other network parameters. Finally, the trained models were tested against an unseen validation set to identify the models performance in real life. Figure 1 shows overall experimental framework.

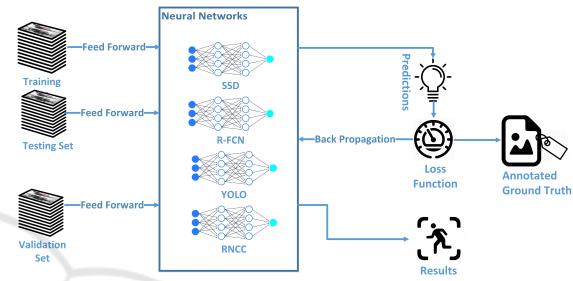


Figure 1: Overall experimental framework.

### 3.1 Data Acquisition

We have prepared a relatively large dataset comprising total number of 10,972 image were mostly captured from CCTV cameras placed in department stores, shopping malls and retails. Majority of the images were captured in indoor environments under various conditions such as distance, lighting, angle, and camera type. Given the fact that each camera has its own color depth and temperature, field of view and resolution, all images passed through a preprocessing operation which ensures consistency across entire input data. Figure 2 shows some examples of our dataset.



Figure 2: An example of tilt (left) and top-down (right) frame in dataset.

In order to ease and speed up the annotation process, we have employed a semi-automatic annotation mechanism which uses a Faster RCNN inception model to generate the initial annotations for each given

Table 1: Average precision at IoU 0.95 and 0.50.

#	Model	Framework	AP [IoU=0.95]	AP [IoU=0.50]
1	Faster RCNN (ResNet-101)	Tensorflow	<b>0.245</b>	<b>0.476</b>
2	YOLOv3-416	Darknet	0.143	0.367
3	Faster RCNN (Inception ResNet-v2)	Tensorflow	<b>0.317</b>	<b>0.557</b>
4	YOLOv2-608	Darknet	0.198	<b>0.463</b>
5	Tiny YOLO-416	Darknet	0.035	0.116
6	SSD (MobileNet v1)	Tensorflow	0.094	0.233
7	SSD (VGG-300)	Tensorflow	0.148	0.307
8	SSD (VGG-500)	Tensorflow	0.183	0.403
9	R-FCN (ResNet-101)	Tensorflow	<b>0.246</b>	<b>0.486</b>
10	Tiny YOLO-608	Darknet	0.06	0.185
11	SSD (Inception ResNet-v2)	Tensorflow	0.116	0.267
12	SqueezeDet	Tensorflow	0.003	0.012
13	R-FCN	Tensorflow	0.124	0.319

input image. The detection results were manually investigated and fine tuned to insure the reliability and integrity of the ground truth. Moreover, images with no person presence have been removed from the dataset. Finally, a random sampling process performed over entire images. The final dataset consists of total number of 10,972 images no background overlap, divided into training set (5,790 images), testing set (2,152 images) and validation set (3,030 images).

### 3.2 Experimental Setup

To measure and compare the average precision (AP) and IoU of the deep models, we have used a workstation powered by 16 GB of internal memory and Nvidia GTX 1080ti graphics accelerator. To measure and compare the time complexity metrics, we have utilized two common embedded platforms including the Nvidia Jetson TX2 as well as Movidius to run the experiments.

## 4 EXPERIMENTAL RESULTS AND DISCUSSIONS

We investigated 13 different object detector deep models including variants of YOLO, SSD, RCNN, RFCN and SqueezeDet. To measure the accuracy of these models, we have used AP at two different IoU ratios, including 0.5 which denotes a fair detection and 0.95 which indicates a very accurate detection. Table 2 summarizes the AP across various object detectors. It can be observed that, when IoU is 0.95, Faster RCNN (Inception ResNet-v2) with average precision of 0.317 outperforms other object detector in this research. Faster RCNN (ResNet-101) alongside R-FCN (ResNet-101) with respective AP of 0.245 and 0.246 are among the best performers in this category.

On the other hand, SqueezeDet and Tiny YOLO-608 with respective AP of 0.003 and 0.06 performed poorly in this category. Results with IoU = 0.50 show a very similar trend. Once again, Faster RCNN (In-

Table 2: Total latency of inference in both CPU and GPU modes.

#	Model	CPU Latency (S)	GPU Latency (S)
1	Faster RCNN (ResNet-101)	3.271	0.232
2	YOLOv3-416	5.183	0.017
3	Faster RCNN (Inception ResNet-v2)	10.538	0.478
4	YOLOv2-608	11.303	0.035
5	Tiny YOLO-416	1.018	0.011
6	SSD (MobileNet v1)	<b>0.081</b>	0.03
7	SSD (VGG-300)	0.361	<b>0.015</b>
8	SSD (VGG-500)	0.968	<b>0.026</b>
9	R-FCN (ResNet-101)	1.69	0.131
10	Tiny YOLO-608	2.144	<b>0.025</b>
11	SSD (Inception ResNet-v2)	<b>0.109</b>	0.04
12	SqueezeDet	<b>0.14</b>	<b>0.027</b>
13	R-FCN	3.034	0.084

ception ResNet-v2) with AP 0.557 outperformed other detector. R-FCN (ResNet-101), Faster RCNN (ResNet-101) and YOLOv2-608 with average precision of 0.486, 0.476 and 0.463 respectively, are showing superior performance. In contrast, SqueezeDet and Tiny YOLO-416 with respective AP of 0.012 and 0.116 generate poor results. Results also indicates, that, in terms of robustness and resiliency of the detector against increase in IoU, all models perform roughly equally and there is no significant variance. Another noteworthy observation in this experiment is the superiority of the Faster RCNN over other detectors that could be influenced biased by the approach used to prepared the ground truth. As we mentioned earlier in section 3.1, the dataset annotation initialized with the help of Faster RCNN inception model detector. Despite the significant manual adjustments and fine-tuning in annotation, we believe it introduces some level of bias to the results.

The time complexity of detectors were evaluated with measurement of execution latencies in two different approaches. In the first approach total latency of inference of a single test image has been measured in both CPU and GPU modes. In the second approach throughput of continuous inference with repeating camera capture. Table 3 shows the total latency of inference of a single test image on both CPU and GPU. Apparently, GPU is considerably faster than a CPU in matrix arithmetics such as convolution due to their high bandwidth and parallel computing capabilities, but it is always interesting to learn this advantage objectively. According to the results shown in table 3, in CPU mode, SqueezeDet, SSD (Inception ResNet-v2) and SSD (MobileNet-v1) are the fastest deep models in this study.

These models benefit relatively simpler deep network with fewer arithmetic operations. This significantly reduced their computational overhead and increased their performance. However, considering the AP result in table 2, it can be inferred that this performance gains, came with an expensive cost of accuracy and precision. Results in GPU mode shows a

very similar trend however due to high bandwidth and throughput of GPU, the variance in results are significantly lower. According to Table 3, in GPU mode, SSD (VGG-300), Tiny YOLO-608, and SqueezeDet are among the fastest models in our experiments. Aside from CPU and GPU latency, we also measured the throughput of continuous inference with repeating image feed. Due to several factors in the experimental setup and model architecture throughput of continuous inference might not be necessarily correlated with the CPU and GPU latency. Figure 3 shows, Tiny YOLO-416 followed by SSD (VGG-300) with over 80 and 60 FPS respectively have the overall highest throughput among the models investigated in this study. On the other hand, Faster RCNN (Inception ResNet-v2) and Faster RCNN (ResNet-101) are slowest in this regard. In order to deploy the deep models in embedded platforms, Caffe or Tensorflow models should be optimized and restructured using Movidius SDK or TensorRT. This enables the CNN model to utilize the target height/width effectively.

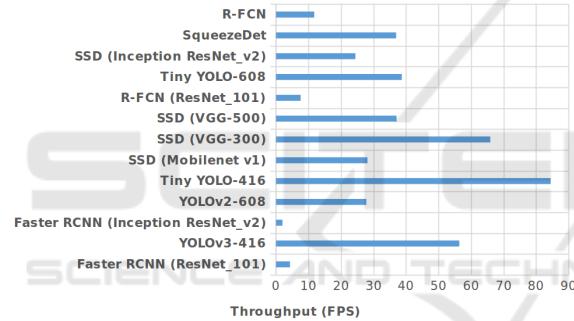


Figure 3: Throughput of continuous inference across various models.

However, the supported layers by Movidius SDK or TensorRT are relatively basic and limited and complex models such as ResNet cannot be truly deployed in these platforms. As an example, leaky rectified linear unit activation function in inception models is not supported by the Jetson platform and cannot be fully replicated. Table 4 summarizes the throughput of continuous inference across various deep models in embedded platforms. It can be observed that the Nvidia Jetson performed significantly better than the Movidius across all different models. Furthermore, TensorRT outperformed Caffe by a relatively large margin. However, in terms of features and functionality, Caffe allows to reproduce more complex networks.

Finding the right deep model for embedded platform is not about accuracy neither performance but is about finding the right tradeoff between accuracy and performance, which satisfies the requirements. Deep models such as Tiny-YOLO can be extremely fast. However, their accuracy is questionable. Fi-

Table 3: Throughput of continuous inference across various models using embedded platform including Movidius and Jetson.

#	Model	Framework	Movidius	Jetson	TensorRT	
			(FP16) Throughput	(FP32) Throughput	(FP16) Throughput	(FP32) Throughput
1	AgeNet	Caffe	18	56	192	127
2	AlexNet	Caffe	10	37	65	54
3	GenderNet	Caffe	18	62	198	119
4	GoogleNet	Caffe	9	19	120	73
5	SqueezeNet	Caffe	17	37	166	124
6	TinyYolo	Caffe	7	19	-NA-	-NA-
7	Inception v1	Tensorflow	10	-NA-	-NA-	-NA-
8	Inception v2	Tensorflow	7	-NA-	-NA-	-NA-
9	Inception v3	Tensorflow	3	-NA-	-NA-	-NA-
10	MobileNet	Tensorflow	19	-NA-	-NA-	-NA-

Figure 4 plots the deep models Average Precision across their throughput. The closer to the top right corner of the plot, the better the overall performance of the model. Figure 4 shows among the various models that we investigated in this research, YOLO v3-416 and SSD (VGG-500) are the best tradeoff between Average precision and throughput.

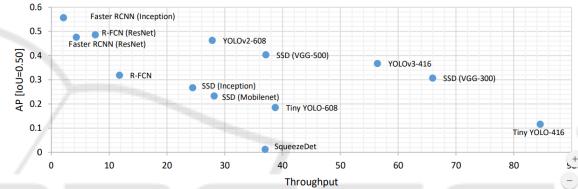


Figure 4: Average Precision [IoU=0.5] across throughput.

## 5 CONCLUSION

Person detection is essential step in analysis of the in-store customer behavior and modeling. This study focused on the use of DNN based object detection models for person detection in indoor retail environments using embedded platforms such as the Nvidia Jetson TX2 and the Movidius. Several DNN models including variations of YOLO, SSD, RCNN, R-FCN and SqueezeDet have been analyzed over our proprietary dataset that consists of over 10 thousands images in terms of both time complexity and average precision. Experiments results shows that Tiny YOLO-416 and SSD (VGG-300) are among the fastest models and Faster RCNN (Inception ResNet-v2) and R-FCN (ResNet-101) are the most accurate ones. However, neither of these models nail the tradeoff between speed and accuracy. Further analysis indicates that YOLO v3-416 delivers relatively accurate result in reasonable amount of time, which makes it a desirable model for person detection in embedded platforms.

## ACKNOWLEDGEMENTS

We thank our colleagues from VCA Technology who provided data and expertise that greatly assisted the research. This work is co-funded by the EU-H2020 within the MONICA project under grant agreement number 732350. The Titan X Pascal used for this research was donated by NVIDIA.

## REFERENCES

- Balaji, M. and Roy, S. K. (2017). Value co-creation with internet of things technology in the retail industry. *Journal of Marketing Management*, 33(1-2):7–31.
- Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387.
- Foxlin, E., Wormell, D., Browne, T. C., and Donfrancesco, M. (2014). Motion tracking system and method using camera and non-camera sensors. US Patent 8,696,458.
- Ghosh, R., Jain, J., and Dekhil, M. E. (2017). Acquiring customer insight in a retail environment. US Patent 9,760,896.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, volume 4.
- Jia, B., Pham, K. D., Blasch, E., Shen, D., Wang, Z., and Chen, G. (2016). Cooperative space object tracking using space-based optical sensors via consensus-based filters. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1908–1936.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.
- Majeed, A. A. and Rupasinghe, T. D. (2017). Internet of things (iot) embedded future supply chains for industry 4.0: an assessment from an erp-based fashion apparel and footwear industry. *International Journal of Supply Chain Management*, 6(1):25–40.
- Nickolls, J. and Dally, W. J. (2010). The gpu computing era. *IEEE micro*, 30(2).
- Ohata, Y., Ohno, A., Yamasaki, T., and Tokiwa, K.-i. (2014). An analysis of the effects of customers migratory behavior in the inner areas of the sales floor in a retail store on their purchase. *Procedia Computer Science*, 35:1505–1512.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. *arXiv preprint*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Wu, B., Iandola, F. N., Jin, P. H., and Keutzer, K. (2017). Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *CVPR Workshops*, pages 446–454.
- Zuo, Y., Yada, K., and Ali, A. S. (2016). Prediction of consumer purchasing in a grocery store using machine learning techniques. In *Computer Science and Engineering (APWC on CSE), 2016 3rd Asia-Pacific World Congress on*, pages 18–25. IEEE.