Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

DLSU Innovation and Technology Fair
Ideas and Solutions for Nation-building

Single Shot Multi-Box Detector with Multi Task Convolutional Network for Carabao Mango Detection and Classification using Tensorflow

Ryan Joshua H. Liwag[1], Kevin Jeff T. Cepria[2], Anfernee S. Rapio[3],
Karlos Leo F. Castillo[4], Melvin K. Cabatuan[5], Edwin J. Calilung[6]
Gokongwei College of Engineering
Electronics and Communications Engineering Department
De La Salle University - Manila, Philippines
[1] *ryan_liwag@dlsu.edu.ph*
[2]*kevin_cepria@dlsu.edu.ph*
[3]*anfernee_rapio@dlsu.edu.ph*
[4]*karlos_castillo@dlsu.edu.ph*
[5]*melvin.cabatuan@dlsu.edu.ph*
[6]*edwin.calilung@dlsu.edu.ph*

Abstract: Carabao mangoes are among the most desired exports of the Philippines in the world due to its very sweet taste but before these mangoes are exported, they are first sorted so that those products with the desired characteristics would be the only ones sent abroad. To aid the Filipino mango farmers in increasing their profit by increasing their yield, a system would be necessary so that it would provide a fast and reliable system to classify mangoes based on their ripeness. As the input of the system would be images of mangoes, convolutional neural networks are the most appropriate deep learning system for this application. This study would be using Visual Geometry Group 16 deep convolutional model due to its lower overhead delay compared to other models. The algorithm was then retrained and refined using transfer learning so that it could be used to classify images of mangoes into three ripeness categories, unripe, partially ripe, and ripe, on a Raspberry Pi 3-controlled hardware system for portability and mobility. After obtaining an accuracy that is less than ideal for the industry, the neural network was adjusted by decreasing its learning rate and adding a dropout layer within the network. The new accuracy achieved by the system was at 98.32%. The neural network was also installed on a laptop for increased computing power and to show the versatility of the system with regards to hardware. This research therefore shows the potential of using technology in supporting the advancement of the agriculture industry in the Philippines.

Key Words: Computer Vision; Machine Learning; Mango Classification; Neural Networks; TensorFlow

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

DLSU Innovation and Technology Fair
Ideas and Solutions for Nation-building

1. INTRODUCTION

The carabao mango is one of Philippines' quality export product because of its sweetness. Before the fruits are exported, it must first be processed and sorted with regards to their size, quality and ripeness. This research wants to focus on the potential of current state of the art technology on image detection and classification, and find ways to leverage this technology to potentially help the Filipino farmers.

The initial hypothesis of the research is to build a cascading process of first detecting the carabao mango and then classifying it. The reasoning of opting to use detection is so that the system could specifically target and classify the mango in the image , and this would allow the system to be used in different types of background environment. There are multiple research papers that use computer vision, but they are held in controlled environments, with lighting and background being constants this limits the flexibility of integrating their techniques with sorting machinery (Arakery & Lakshmana, 2016) (Patil et al, 2016) (Nandi et al, 2014). For the detection task, the researcher opted to using one of the variations of the single shot multi box detector (SSD) model (Liu et al, 2016). SSD is a meta architecture used for detecting objects in images through the use of a single deep Convolutional neural network.

After the object has been detected, it's resulting fitted bounding box will be cropped. This allows us to have a image where the carabao mango is specifically focused and other background objects are removed. This image is then fed into a Convolutional Neural Network (CNN) (Krizhervsky et al, n.d) model that we have designed. The designed model uses a technique known as Multi-task Learning (Ruder, 2017), which allows it to attention focus features of ripeness and quality of the carabao mango. The classification model will have 3 classes of ripeness which are green, semi-ripe and ripe. It will also have 2 classes of quality, the mango either has defects or not.

2. RELATED WORK

2.1 Multi Task Learning

In Machine Learning (ML), the goal is to build a model that can generalize, its performance is graded on its ability to generalize the real world's unseen data. Typically a single or an ensemble of models are trained on a desired task, then later fine tuned until performance no longer increases (Lee et al, 2010). But by being laser focused on a single task, the model could ignore signals that are important and essential. By learning multiple individual features, the focus is on different aspects of the input, that allows the model to generalize even better on the original task. Rich Caruana (Lee et al, 2010) summarizes the objective of MTL (Multi Task Learning): "MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks".

With the success of Deep Convolutional Neural Networks (CNN) when it comes to classifying images which has led to the MTL approach to computer vision supplemented by deep CNN. They are called Deep Relationship Networks, below is a figure exemplifying the common usage of Deep relationship networks, where you have convolutional layers acting as a feature extractor and then it is fed to multiple task specific fully connected layers(Long et al, 2017).

2.2 Object Detection Using SSD

Object detection is helpful in cases where the object appears multiple times in an image, or if the object's exact size and location is needed. CNN's have also entered this field and has shown significant advantages over previous methods of object detections such as Haar and LBP classifier (Abdulnabi et al, 2016). It has been shown that traditional detection methods involved using block-wise orientation histogram feature on which has not achieved high accuracy in standard datasets such as PASCAL VOC. Deep CNN has become the state of the art with regards to object detection tasks. Currently these state of the art object detection models used region proposal algorithms to hypothesize object location, a several papers have been released studying and improving this architecture, focusing on improving run-time efficiency or accuracy which can be seen in the progression of YOLO (Redmon et al, 2016) and Faster R-CNN (Ren et al, 2016).

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

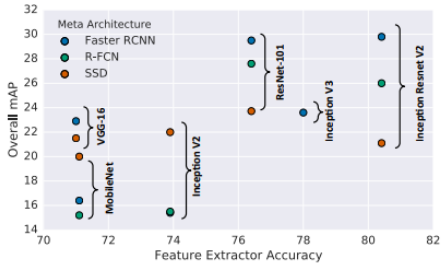


Fig. 2. mAP score for the most popular CNN architectures also grouped by their Meta Architecture (Huang et al, 2017).

The figure above shows a detailed list of different model and their meta architectures, their mAP score is based on COCO mAP (COCO Consortium, 2017), which is the score the model achieved on the COCO dataset. It shows that the SSD detection outperforms the R-FCN (Region-based Fully Convolutional Networks) meta architecture. Also the authors stated that SSD models with the Inception v2 and mobilenet feature extractors are the most accurate among the fastest models (Huang et al, 2017).

3. SYSTEM MODEL

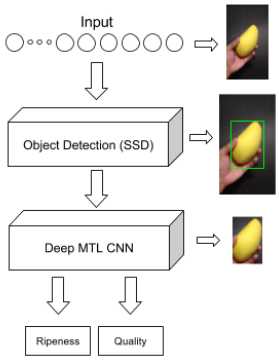


Fig. 3. System Pipeline Model

In designing the system model., it is necessary to explicitly target problems of past papers when it comes to detecting and grading mangoes is necessary. There was always a form of controlled environment present in their papers, where the mangoes are processed in a high contrast background. The problem with their approach is the viability of being used in real agricultural setups, it's harder for their grading process to be integrated into existing agricultural processes. That is why in our design we propose to use one of meta-architectures used in the state of the art CNN object detection, which is the SSD. This would allow the system to target only mango objects, additionally since it may detect multiple objects it can allow our system to classify multiple mango's in the same image.

The assumption is that the background of the object would not be variable, but there must be sufficient lighting to clearly identify its color. As lighting can affect the mangoes color, the test are accomplished with the relative lighting setup that clearly shows mango's color. Another assumption is that as more mangoes are detected our deep MTL CNN model needs to run through each mango detected. This process would naturally have an effect that as more mangoes are being detected, there will be a significant lag time increase in processing the image.

4. SOLUTION

The solution is to build a system that can detect multiple amount of mangoes despite the object's state and background. After detection we would feed it into a Model specifically trained for determining its class and ripeness. The first step to training the models, is gathering and building a dataset, to train it with. From this dataset we would apply augmentations to increase the datasets variance and fix any imbalances among classes.

4.1 DATA GATHERING

The first step before training a model, is to first build the dataset. The same dataset will be used to train both, object detection and classification model. The initial dataset is obtained from 270 mangoes, and 2,800 images were produced from it. Take into account that the mangoes were bought while unripe, so as they ripen more images are taken.

4.1 1 OBJECT DETECTION DATASET

For object detection, the dataset must first be annotated with bounding boxes. So, the research made use of the tool called lableimg (Github, 2017),

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

this was used to draw the bounding box on each mango object in the image

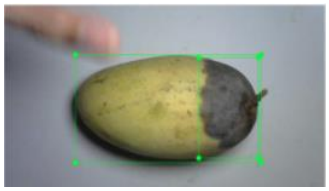


Fig. 4. Process of annotating Images using LabelImg (Github, 2017).

4.1 2 OBJECT CLASSIFICATION DATASET

The object classification dataset will be built, from the annotated dataset used in the object detection model. All the specific bounding boxes were cropped from each image, producing a dataset where the mango object is the only focus of the image. Then they are all segregated to their respective classes. For ripeness, the Green, semi-ripe and ripe are the chosen classes to tackle. Below shows a figure of different categories of mango ripeness on which dataset has. Dataset will also have 2 quality classes, good or bad mango. Since the research is limited by dataset to identify every type of defect, they generalized and labeled all defects present as bad, basis for defects are taken from the Philippine standard (Department of Trade and Industry Philippines, 2004).

Fig. 5. An unripe, partially ripe, and ripe mango sourced from local markets.

As the dataset contains relatively few images and since these images were obtained from videos, several images would display similar data. Being the case, the neural network simply "memorizes" the data rather than finding patterns appropriate to that classification; memorization leads to overfitting. Employment of data augmentation was done to supplement the lack of images in the dataset by modifying several parameters of the images such as rotation, width shifting, height shifting, zoom and channel shifting. Modifying the zoom condition and size parameters of the mangoes in the images by cropping and resizing the images avoids the distortion of the mango within the frame of the image but allows the creation of new images that contain unique data points for the neural network to learn.

As the dataset for the study is composed of a relatively few images, adding a linear classifier on the top layer of the model aids in the characterization of the images as tested by the VGG group and Caffe Model Zoo (Abdulnabi et al, 2016). The linear classifier was used by to reduce the original number of classes of the model into the number necessary in this research.

4.2 TRAINING OBJECT DETECTION (SSD-Mobile net)

With the dataset ready to be trained for the object detection task, we choose among different CNN meta-architectures and model, and decide which is most beneficial to the system. In this study we needed a model that has a high speed and high precision. Since intention is to train it on 1 class which is mangoes, its classification score is not as important. The model chosen is called the SSD mobile-net v1 COCO, this is essentially the mobile-net model pretrained on the COCO dataset (Lin et al, 2015). This model was chosen because despite having the lowest score among the COCO models, it has shown to have achieved the highest speed among them (TensorFlow Detection Model Zoo, 2017).

Table 1. SSD Mobile net Training Configurations

Model	SSD mobile-net v1 COCO
Training Dataset	2600
Evaluation Dataset	200

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

Image input resize	300 x 300
Batch sizes	24
Training Steps	27,000
Learning rate	Learning rate: 0.004 Decay steps: 1000 Decay rate: 0.95
Image Augmentations	1. Random Horizontal Flip 2. Random cropping with fixed aspect ratio

It took many iterations in configuring an optimal learning rate as having too high of a learning does not allow the model to converge to a minima, and having too low of a learning rate takes the much longer to train. Basic image augmentations are included to increase variety in training data.

4.3 DEEP MULTI TASK CNN

4.3.1 MODEL DESIGN

The dataset currently has 3 levels of ripeness and 2 levels of mango quality, Combining them would create 6 different distinct classes. The distribution of classes on total amount of images can be seen on Figure 6.

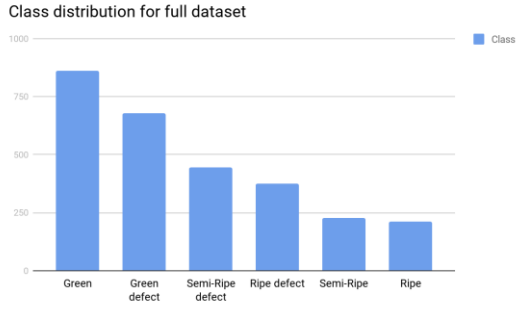


Fig. 6. Distribution of Classes and number of images per class

This dataset contains closely related classes and training a deep CNN model to classify 6 classes that are very much closely related can yield to unsatisfactory results. Previous iterations include the attempted training of the Mobile-Net model, resulting in below 90% accuracy across all 6 classes.

To solve this solution, It is necessary to design a Deep multi task learning Convolutional neural network model. This allows the model to specify some layers to train very specific tasks. This also allows the dataset to be used much more effectively on each task.

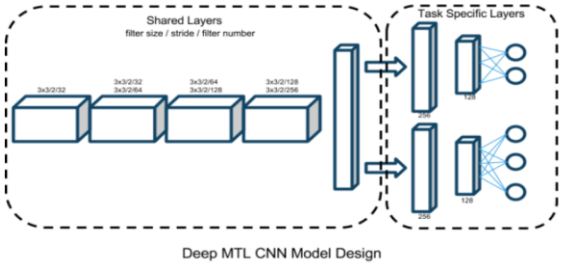


Figure 7. Deep MTL CNN Model Design

Fig. 8. The 2 graphs above shows the mango image distribution, if the dataset where to be trained through a MTL approach.

Designing the Deep Multi task learning model, the research took inspiration from the CNN design

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

pattern used in VGG16 architecture (Simonyan & Zisserman, 2015). The model was designed with a series of convolutional layers which are at the end flattened then distributed to the task specific layers. Each task specific layer is built as a regular multi layer perception network.

This design uses the concept of hard parameter sharing, where there are no soft weight sharing in between the task specific layers. Each task specific layer will optimize its weights to classify their target class which is ripeness or quality. Calculated softmax loss for each task specific layer is added together to form a combined loss which is where the model's optimizer will be looking to minimize.

4.3.2 TRAINING

Final training Configurations of the Deep MTL CNN model can be seen on the table below.

Table 2. Deep MTL CNN Model configurations

Input shape	1 x 50 x 50 x 3
Number of images per epoch	2700
Learning Rate	0.001
Batch Size	24
Optimizer	Adam Optimizer
Image augmentations	1. Shearing 2. Cropping 3. Tilting

After 17 epochs the model's has obtained an accuracy above 90% on both classification tasks. This training was not held over at the Google Cloud Platform but done locally using a GTX 960 GPU unit, training took 3 hours. Training on Google Cloud Platform would cost higher as there are fees on renting the cloud GPU service.

5. ANALYSIS

5.1 OBJECT DETECTION MODEL

The object detection model which is the SSD Mobile-net, was trained for 27,000 steps. A checkpoint would be saved every 1000 steps, this is to evaluate model performance in detection as it was being trained.

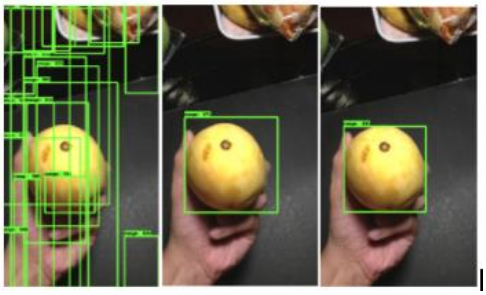
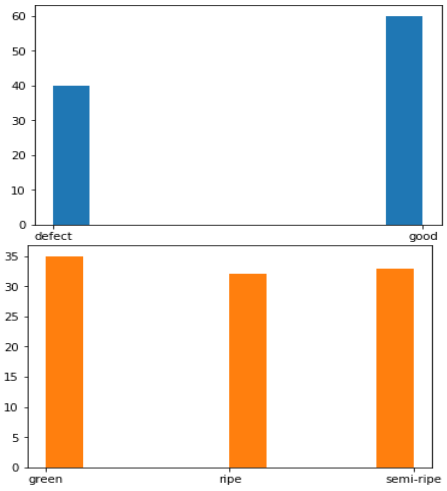


Fig. 9. (left to right) Models Evaluation Performance at epoch 0, epoch 18k, epoch 25k

5.1 OBJECT CLASSIFICATION EVALUATION

The testing data and evaluation data was withheld while evaluating the mango classification task. Evaluation is to data is tested while model is being trained, and had a final accuracy of 95% on both ripeness and quality.

To further test whether the research model is successful and if the same accuracy is reproducible, the system has been tested against a testing set of 100 images. Figure 7 shows that distribution of testing images among the different classes.



Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

Fig. 10. Class distribution for testing Data, where X values are classes and Y axis represent the number of images.

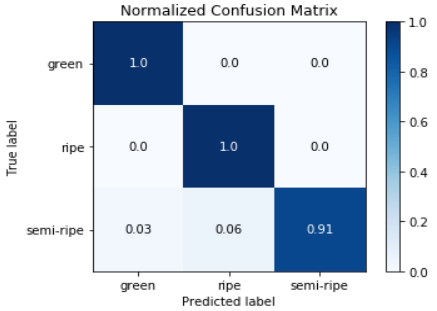


Fig. 11. Normalized Confusion matrix for Ripeness classification

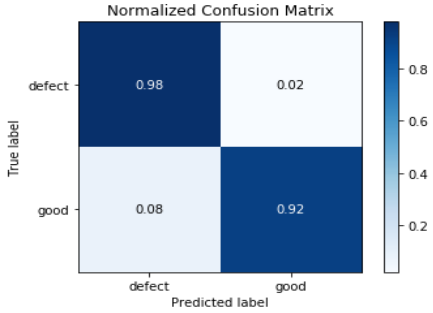


Fig. 12. Normalized Confusion matrix for Ripeness classification

Figure 11 and 12 above are the result of applying the classification model on the testing dataset. It can be seen that the model has been able to classify above 90% on each factor of sorting. With minimal misclassifications on semi-ripe and ripe, but overall the model has succeeded in accurately classifying ripeness and quality.

6. SIMULATION AND EXPERIMENTATION

The researchers measured the performance of each neural network model which are object detection model (SSD Mobile-net) and the Classification model (MTL CNN).

Firstly the execution time for the mobile-net object detection network was tested. It ran successfully for all 100 iterations and has resulted in a program execution time average of 0.13 ms. The classification model was also tested, it generated a average mean of 0.059 ms per classification. It can also be noted that when both test were run, it can be seen that both models had significant time lag at the start as compared to the mean run time. This is a time delay taken as spinning up a session in TensorFlow the first few images will experience a significant time delay.

To further test and experiment the model on a specific setting with accordance to the distance of the camera and mango, a hardware setup was arranged to test the mango classification objectively. The setup would consist of 2 rollers, and camera on top. The idea is to build the software with a GUI and load the system to a Raspberry Pi 3 (Rpi3) model B.

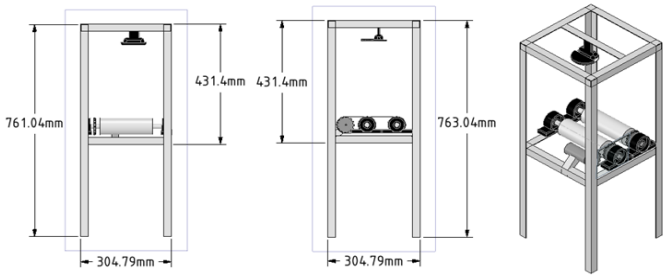


Fig. 13. Front view, Side view and Isometric Projection of the Frame with Rollers

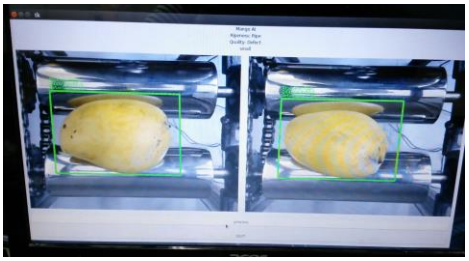


Fig. 14. GUI showing top and bottom side of the mango, the general and overall classification will be displayed at the top.

The researchers wanted a standardized way to emulate how the software could be integrated with hardware. In this system a mango can be loaded and the camera will take 2 images, top side of the mango and bottom side. The rollers will be responsible for automatically turning the mango. Rpi3 will be integrated with arduino and motor driver to control the rollers, while the camera is controlled by the Rpi3.

Presented at the 5th DLSU Innovation and Technology Fair 2017
De La Salle University, Manila, Philippines
November 28 & 29, 2017

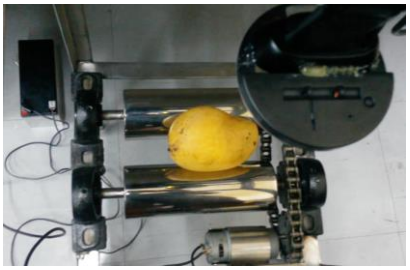


Fig. 15. Top view of the rollers with a mango in place

7. CONCLUSION

This research has looked into the possibility of using current state of the art techniques in object detection and classification, in aiding or improving the Philippine mango industry. This paper is also successful in exploring and implementing Multi-task learning to solve class problems of classifying different mango features. The group have managed to develop a system with a classification accuracy of 90% on the 2 factors of sorting. This system

8. REFERENCES

Arakeri, M., & Lakshmana. (2016). Computer Vision Based Fruit Grading System for Quality Evaluation of Tomato in Agriculture industry. *Procedia Computer Science*, *79*.

Patil, K., Kadam, S., Kale, S., Rachetti, Y., Jagdan, K., & Inamdar, K. (2016). Machine Vision Based Autonomous Fruit Inspection And Sorting. *International Research Journal Of Engineering And Technology (IRJET)*, *3*(7).

Nandi, C., Tudu, B., & Koley, C. (2014). Computer Vision Based Mango Fruit Grading System. *International Conference On Innovative Engineering Technologies*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & Berg, A. (2016). SSD: Single Shot MultiBox Detector.

Krizhevsky, A., Sutskever, I., & Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks.

Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks.

Lee, C., Jung, S., Kim, K., & Lee, G. (2010). Hybrid approach to robust dialog management using agenda and dialog examples. *Computer Speech & Language*, *24*.

Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016). Cross-Stitch Networks for Multi-task Learning. *Computer Vision And Pattern Recognition (CVPR)*.

Long, M., Cao, Z., Wang, J., & Yu, P. (2017). Learning Multiple Tasks with Multilinear Relationship Networks.

Abdulnabi, A., Wang, G., Lu, J., & Jia, K. (2016). Multi-task CNN Model for Attribute Prediction.

Ozhiganov, I. (2017). Convolutional Neural Networks vs. Cascade Classifiers for Object Detection.

Redmon, J., Divvala, S., Girshik, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

COCO Consortium. (2017). *COCO - Common Objects in Context*. *COCO*. Retrieved 26 November 2017, from http://cocodataset.org/#people

Huang, J., Rathod, V., Sun, C., Zhu, M., Korratikara, A., & Fathi, A. et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors.

Github. (2017). *labelimg*. *Github*. Retrieved 26 November 2017, from https://github.com/tzutalin/labelImg.

Department of Trade and Industry Philippines. (2004). Fresh Fruit - Mangoes - Specification. *Philippine National Standard*.

Lin, T., Maire, M., Belongie, S., BOurdev, L., Girshick, R., & Hays, J. et al. (2015). Microsoft COCO: Common Objects in Context.

TensorFlow Detection Model Zoo. (2017). *Github*. Retrieved 26 December 2017, from https://github.com/tensorflow/models/blob/master/research/object_de tection/g3doc/detection_model_zoo.md.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition