Abstract

In this project, we choose SSD algorithm as training model for implementation. However, SSD algorithm has less accuracy in detecting and recognizing small objects, since these small objects don't have enough amount of feature in the top layer of the network. Besides, SSD algorithm don't utilize the feature in low-level layers, which is helpful with detecting small objects.

To ameliorate this dilemma, we can choose to enlarge the size of the input image, or implement Feature Pyramid Networks(FPN) algorithm to improve accuracy, because FPN produces a multi-scale feature representation where all levels are semantically strong, including the high-resolution levels. Consequently, the detection of small object can become more accurate, thus increase the successful rate of the hand gesture recognition.

We built our very own hand gesture detector by using the Tensorflow object detector and python. In order to reach the goal, first, we labeled images for object detection with LabelImg and generate the xml file. LabelImg will specify the detected object in the image by a rectangular bounding box, which is used for comparing the rate of overlapped area between ground-truth bounding box and predicted bounding box. Furthermore, we use Tensorflow model zoo to train custom dataset with a view to realizing real time hand gesture recognition.

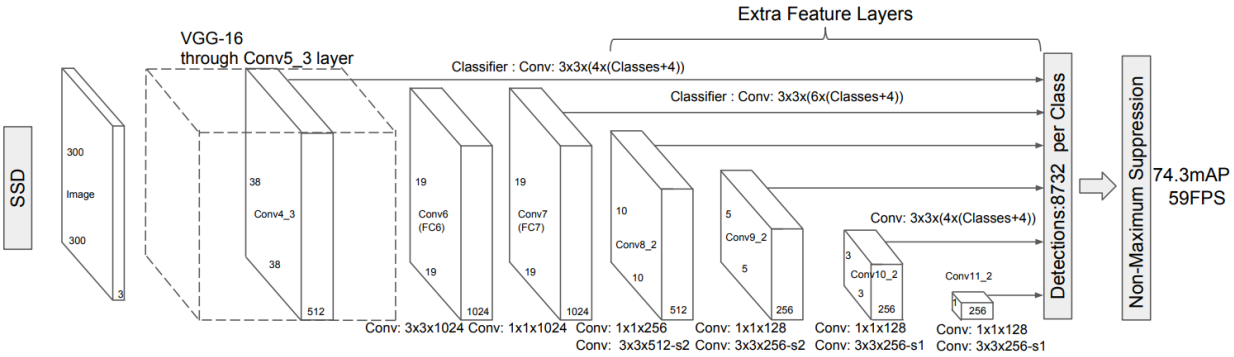
After model training, we can accomplish hand gesture recognition in real time precisely. However, we found that the background of the dataset and the difference between hand gestures would have an impact on the training outcome. At the same time, it is the capacity of dataset and the selection of the images that result in accuracy error of hand gesture recognition.

Introduction

在 Computer vision 的研究中有三個主要的主題，分別是 Image classification、 Object Localization 以及 Object Detection。而此次我們專題研究的方向針對 Object detection 進行更深入的探討。Image classification 是對影像中的物體進行分類，Object Localization 則是不僅對影像進行分類，還可以標出物體的位置，而 Object Detection 與 Object Localization 雖然都能夠標出物件所在位置，但 Object Localization 的輸入影像為單一物件，而 Object Detection 的輸入影像中則可能有零到多個物件，所以選擇 Object Detection 的方法才有助於我們實現即時影像中多個手勢的辨識。

當 Object Detection 演算法偵測到物件後，會用一個矩形的 bounding box 來標註偵測物件在影像中所在位置。這個 bounding box 有四個參數，分別是矩形的左下點與矩形的右上點，也就是 x, y, w, h，分別代表矩形框的中心座標以及其寬與高。而檢測目標若共有 c 個類別，則需要預測 c+1個 confidences，其中第一個 confidence 是不含目標或者屬於背景的類別。

而 SSD 全名為 Single Shot MultiBox Detector，Single Shot 說明 SSD 為 one-stage 的演算法，而 MultiBox 則代表 SSD 是多框預測。



資料來源：W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. SSD: Single shot multibox detector

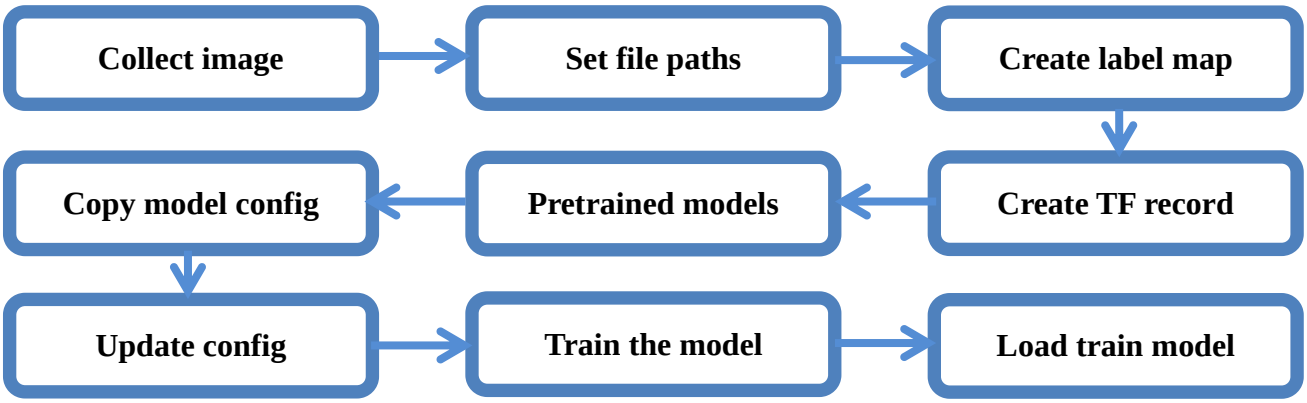
SSD 的主要概念是均勻的在圖片的不同位置進行密集抽樣，選取自己想要預測和偵測的物件範圍，抽樣時可以使用不同尺度和長寬的 bounding box，然後利用 CNN 提取 feature 後再進行 classification，相比需要 object proposals 的檢測模型，如 faster-RCNN，SSD 方法完全取消了 proposals generation、pixel resampling 或者 feature resampling 這些階段，使得 SSD 更容易去優化訓練，也更容易地將檢測模型融合進系統之中，所以能夠以更快的速度進行物件辨識。

SSD 採用的是 VGG16 作為基礎模型，然後在 VGG16的基礎上新增了 convolutional feature layer 來獲得更多的 feature map 以用於偵測與檢測。SSD 的演算法結構如下圖所示。

SSD 和 YOLO 一樣都是採用一個 CNN Network 來進行 detection，但是卻採用了 Multi-sca

le feature maps，而 SSD 核心架構的設計理念整理為以下三點：

(1) Multi-scale feature maps for detection
(2) Convolutional predictors for detection
(3) Default boxes and aspect ratios
以下是操作流程



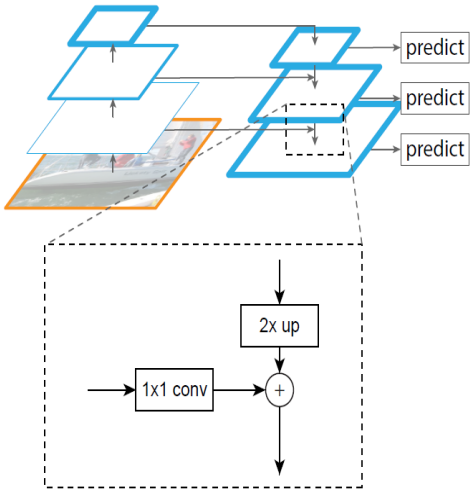
首先，我們利用 cv2來擷取我們需要用到的 images，作為 training 的素材。再來，設置需要用到的資料夾的路徑，像是 images、workspace 和 pre-trained models。接著，我們需要建立 label map，此時 label map 記錄了所有需要用來辨識的手勢，之後可以給 tensorflow object detection 的 library 使用。

我們還需要建立 TF record，這步驟非常重要，TF record 是以二進位的方式存取的一種簡易 format，並且用來存訓練要用的 data。接著就可以 從 tensorflow model zoo 下載 training models，而在這次的研究中，我們選擇的 model 是 ssd_mobilenet_v2_fpnlite_320x320_coco17_tpu-8，其中包含了 SSD 及 FPN 的演算法，此訓練 model 的辨識速度為22ms，而在 COCO dataset 上準確度的表現為22.2mAP，所以為了增加辨識速度，並且讓我們能實現即時手勢辨識，才選擇此 model。之後若要改變 training 的 model，可以修改 pipeline.config。

再來就可以開始利用上述 model 做 training，如下圖8，可以看到 loss rate 以及 learning rate。我們選擇 train 5000次，以達到可行的 loss rate。

訓練完便要 import 一些需要的 dependences，像是 label map utility、visualization utility from object detection libraries、model builder method，如此便可以利用 label map 以及將偵測範圍標示出來和從 config file 及 checkpoint 建立即時辨識的 model。
在訓練過程中，我們首先需要確定哪些 default boxes 對應 ground-truth detection，並且從不同位置、長寬比和尺寸的 default boxes 中選擇 ground-truth box，再將每個 ground-truth box 與具有最好的 Jaccard overlap 的 default boxes 相匹配。與 MultiBox 不同的是，我們將 default boxes 匹配到 Jaccard overlap 高於閾值的任何 ground-truth box，如此便可以讓 Network 為多個重疊的 default boxes 預測 high scores，它便不會只挑選具有最大 overlap 程度的一個 default box。

在進行預測時，我們選擇使用特徵金字塔網路 FPN 的演算法來增加對於小物件的準確度，因為 FPN 使用多尺度特徵融合，此作法會將上層 feature 進行取樣得到和下層 feature map 同樣大小的 feature map，然後兩個 feature map merged by addition 得到最終的 feature map，最後使用該 feature map 作為進一步處理的輸入特徵。下圖為 FPN 的 multi-scale feature representation。

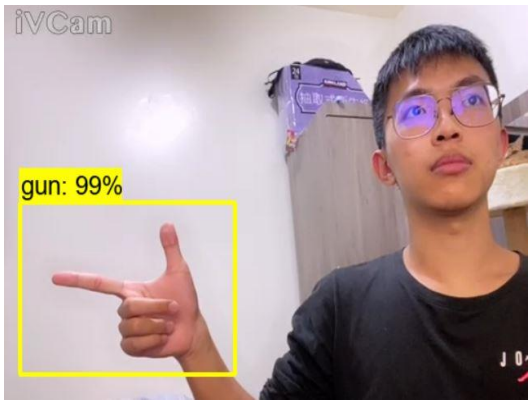


資料來源：Feature Pyramid Networks for Object Detection Tsung-Yi Lin1,2, Piotr Dollar´ 1 , Ross Girshick1 , Kaiming He1 , Bharath Hariharan1 , and Serge Belongie

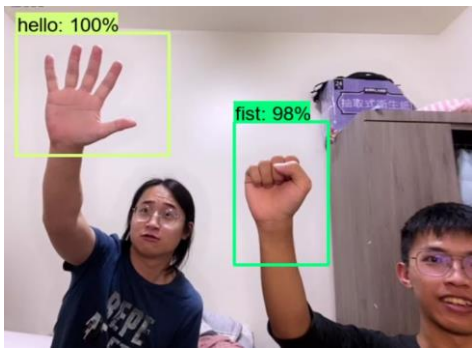
進行預測與偵測時，需先確定 predicted default box 的類別，也就是 confidence 最大者，並且過濾掉屬於背景的 predicted default box，以及 confidence 低於閾值的 default box；並對留下的 default box 進行 encoding，得到實際的位置參數，decoding 之後，根據 confidence 大小來進行降序排列，並保留最高的 k 個 default box，再進行 NMS 算法，過濾掉那些 overlap 程度比較大的 default boxes，最後剩餘的 default boxes 就是預測結果了。

由以上的過程，我們實現了即時手勢的辨識，如以下圖所示，可以看到我們可以辨識出所有自定義的手勢，並且辨識速度快速。

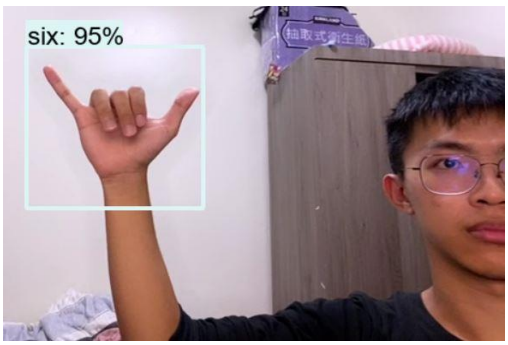
(a) seven



(b) gun



(c) hello、fist



(d) six

心得

在沒有修過機器學習的基礎下，我們從一開始的蒐集論文，到後來漸漸的熟悉了有關我們研究的這個領域，花了不少的時間，不僅研究了有關 CNN 的論文，也看了許多線上課程，都是為了能夠擁有更充分的知識來做專題研究。在操作的過程中，更是深刻的瞭解到理論與實作的差異，不管是從書籍或是論文，要將看到的及學到的知識應用在實作是有一定的難度，有時雖然看起來沒有那麼複雜，但等到真的開始研究後，才發現還有許多沒有考慮到的因素，或是還沒習得的知識要去補足。而做完此次的研究，讓我對於自身的研究能力更是有了很大的進步，對於挑選論文及閱讀論文的能力，我更能有效率的從中吸收我需要用到的觀念或是想法。非常感謝有這種機會能夠參與專題研究，不管是當作進入研究所前的跳板或是對於自身能力的提升，無疑是自我成長的很重要的一步。