Deep learning for human intention anticipation
深度學習在事件預測的應用

組別：B21　　指導教授： 孫民　　組員姓名：周柏睿、石孟立

摘要

結合深度學習與嵌入式相機，做到日常生活中的物品分類。相較於以往大多使用胸前相機或頭部相機，我們將攝影機穿在手上，以手部的視角拍攝物體。如此一來，物體往往就在畫面的正中央，我們不用對物體做偵測。同時，也不會有手上拿著東西，但頭部相機拍不到的狀況。

實驗介紹

一、資料庫 Daily Object 50 dataset

因為是從手部拍攝，視角也跟以往的照片非常不同。因此，同樣的分類器是否適合我們的資料庫會是一個問題。為了避免這樣的問題，我們去蒐集了一份資料庫，其中包含了 50 種會出現在日常生活中的物品，資料庫中的照片都是以手部視角拍攝的日常生活物品。

每一類物品平均有 7000 張圖片、18 個不同的物品。舉例來說，資料庫中有 18 隻不同的水瓶，7000 張關於水瓶的圖片，那為什麼需要這麼多的資料量？因為從監督式學習(Supervised Learning)的角度來看，資料量愈多，model 的效果愈好。

二、架構

VGG-16

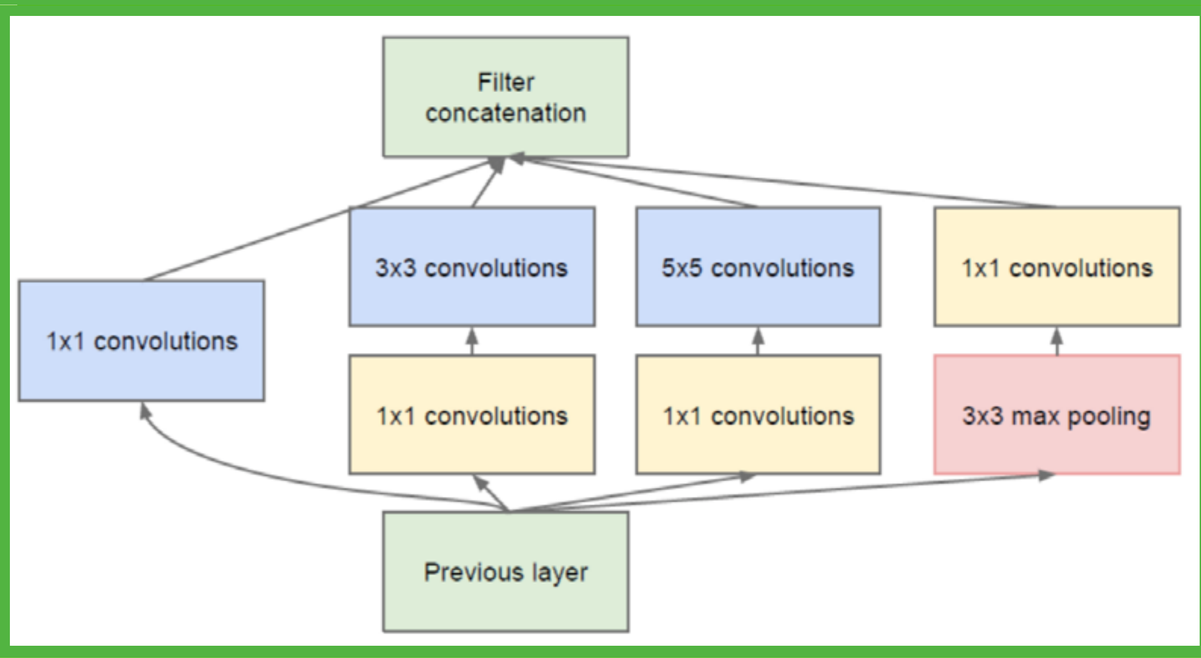
Very Deep Convolutional Networks for Large-Scale Visual Recognition(VGG)，是由英國牛津大學的團隊改良 AlexNet 後產生的架構。其核心概念就是越深的網絡，越能處里大量的資訊。所以，有別于 AlexNet 不到十個捲積層的結構，VGG 的設計有十六個捲積層。而他們也因此獲得了 IMAGE NET 2014 的冠軍。

GoogleNet

GoogleNet 的文章中認為像是 VGG 這種只增加寬度與深度的網絡是不夠的。除了大量的參數會消耗記憶體之外，也會有非線性度不足的問題。因此，作者用 Inception Block 來組成 GoogleNet。Inception Block 具有相對稀疏，但是更高的非線性度等特性。與 VGG 比較，GoogleNet 在消耗較少記憶體下，有更高準確率。

Inception Block

為了減少參數數量，Inception Block 會先透過 1x1 的 Convolution，將輸入資料的 Channel 數量縮減，之後再去做 3x3、5x5 的 Convolution。最後，將不同的結果連接並輸出。如此一來，不但減少參數數量，也能增加非線性度。結構如右圖



實驗結果

	訓練時的準確率	測試在相同場景	測試在不同場景	測試時記憶體需求
VGG-16	98.50%	77.51%	34.91%	2G
	訓練時的準確率	測試在相同場景	測試在不同場景	測試時記憶體需求
GoogleNet	98.60%	80.26%	36.24%	1.2G

結論

相較於 VGG-16，GoogleNet 在準確率上有更好的表現。不論是否在同一個場景，其準確率都比較高。同時，GoogleNet 消耗的記憶體也比 VGG-16 少，非常適合穿戴式像機這樣的嵌入式系統。GoogleNet 能在硬體有限的情況下達到可接受的成果。

未來發展

可以發現當我們測試資料是在不同場景取得時，準確率會下降非常多。我們認為這與，光影、使用者拿東西的姿勢有關。因此，如何解決這些問題，是這個專題未來的研究方向。