Midterm project

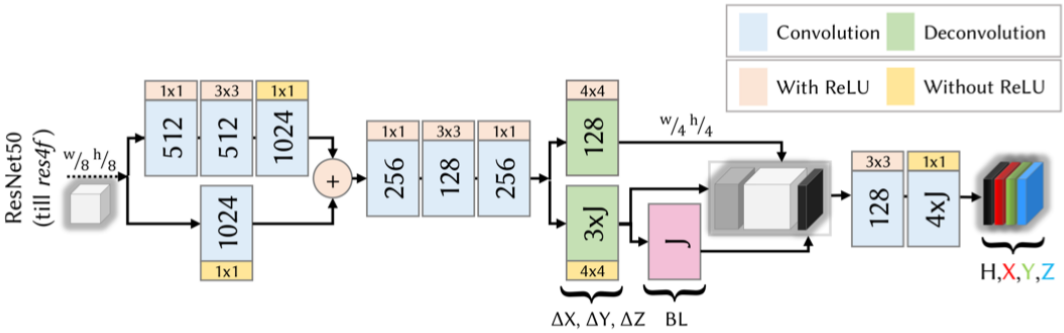
STuser19 賴昱凱

一、模型架構

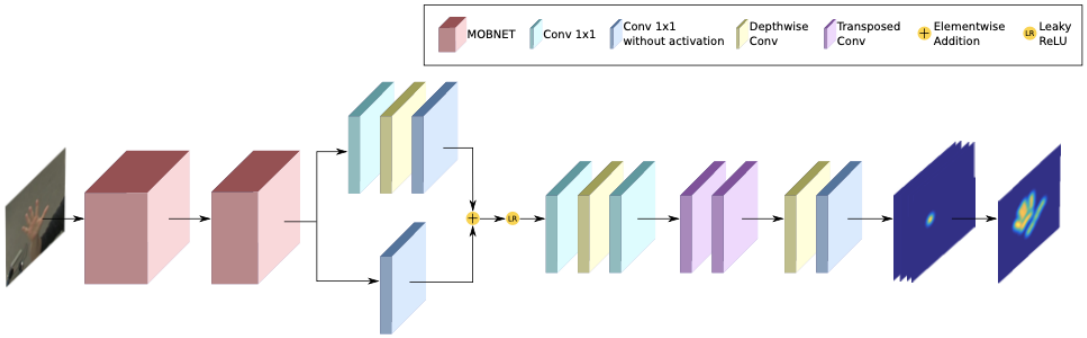
1. 模型架構概述

本期中專題我參考“Accurate Hand Keypoint Localization on Mobile Devices”（後簡稱甲論文）中的模型架構，不過該篇論文也是基於“VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”（後簡稱乙論文）之模型，並以創建可於行動裝置上執行之模型為目的，大幅減少乙論文計算量及參數量所創照之模型架構。而我基於甲論文之模型做更進一步的縮減規模，同時也參考乙論文之 transposed convolution 架構（於甲論文中無詳細說明），做出在模型大小及表現上有優異平衡的模型。

乙論文之模型架構於 ResNet50 後接上一連串的 convolution layers 做後續處理（圖一），但其模型規模龐大，無法於行動裝置做即時影像處理，因此甲論文將原 Resnet50 利用 MobileNetV2 代替，並將 mobilenetv2 的第十四層 stride 由（2, 2）改為（1, 1），使特徵輸出由原先的 1080*7*7 改為 1080*14*14，以避免過小的資料量造成特徵得提取困難，同時為了在後續可以繼續對特徵做處理，甲論文不使用 mobilenetv2 之 linear layers，僅使用其 feature maps 的輸出 1080 * 14 * 14，並將乙論文原先後續處理的 convolution layers 更換使用 depth-wise convolution 進一步減少參數及計算量（圖二）。



圖一、”VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”模型架構圖



圖二、”Accurate Hand Keypoint Localization on Mobile Devices" 模型架構圖

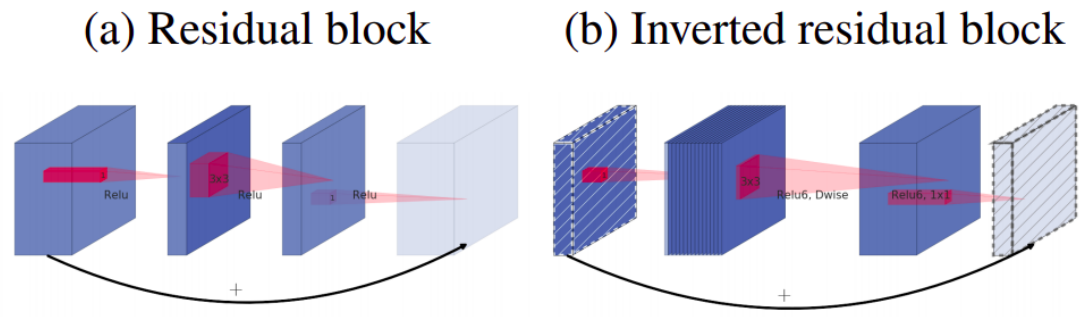
2. MobileNet

i. MobileNetV2

MobileNetV2 之中最重要的概念就是 inverted residual block。在大部分常見 Efficient CNN 中，都會使用 Depthwise Separable Convolution 來減少參數量，包括乙論文所使用的 ResNet50。然而在甲論文中所使用的 MobileNetV2 中，其利用了 inverted residual blocks 技巧來達成更少的參數及計算量。

簡單來說，residual blocks 是藉由減少計算時的 channel 數量來達到計算量及參數量的減少，因此利用 1x1 的 convolution layers 在兩端做 channel 數量的調整，兩端連接的是 channel 數較多的 expansion layers 做資料傳遞。然而 inverted residual blocks 在做完 convolution layers 後的資料傳遞是使用 linear bottleneck layers，是將 channel 數減少而不是增加，且因為“manifold of interest”在降至低維度時不會有大多的資料損失，只會在經過 non-linear function 時才會，因此只要不使用 relu 等 activation function 就可以完整保留特徵資訊同時減少參數量及計算量，convolution layers 前的 inear bottleneck layers 利用 1x1 convolution 後也是增加 channel 而非減少。

由於前面提到 linear bottleneck layers 並不會遺失資訊，因此在 MobileNetV2 中是對非 expansion layers 做 shortcut，與一般 residual blocks 相反，因此稱為 inverted residual blocks。

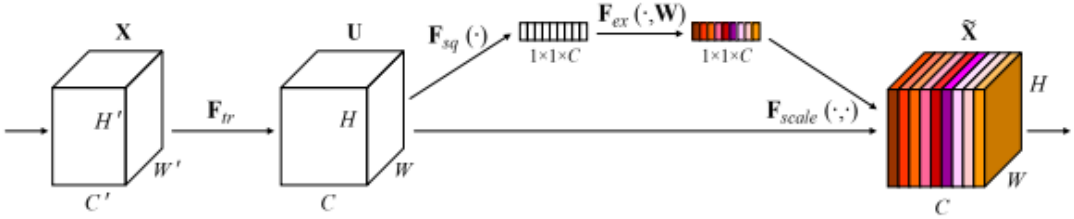


圖三、Residual / Inverted residual block 架構圖

ii. MobileNetV3

而在 MobileNetV3 中，使用了 SENet (Squeeze and Excitation) 架構，透過 Global Average Pooling 計算每個 feature map 的權重，用來強化重要的 feature map 並減弱不重要的 feature map 。且在 SE block 中使用 Hard-sigmoid 代替 sigmoid 以實現更高效率的計算。

SE block 首先先利用 Squeeze function（Fsq）對每個 channel（C）分別作全域平均（Global Average Pooling），得到一個 1*1*C 的資料，再利用 Excitation function（Fex）做兩層 fully connection layers，分別使用 Relu 及 Hard-sigmoid 做 activation function，此時的輸出即為該 channel 的權重，並與對應的 channel 相乘即可達到強化重要的 feature map 並減弱不重要的 feature map 的效果。



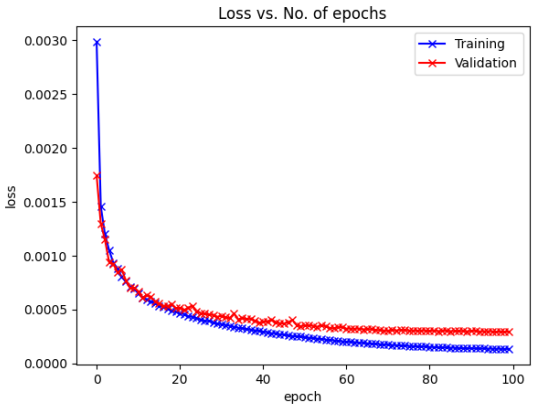
圖四、Squeeze and Excitation block 架構圖

3. 專題使用之模型架構

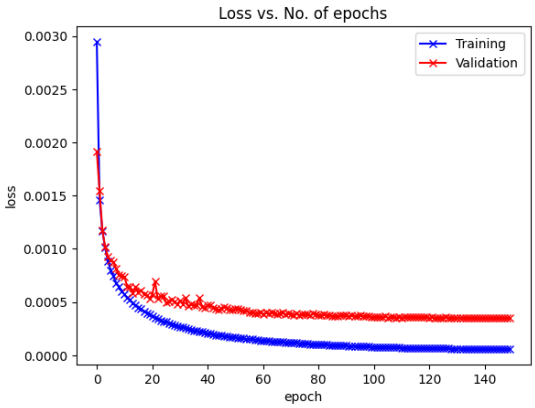
在研讀甲論文後，因為專題評分方式會將模型規模納入考慮，因此我決定以甲論文之模型為雛形，創建一個計算量及參數量皆小的模型。在多次實驗後，我得到以下結論。

i. 雖然模型 channel 數量與準確度有正相關，但考慮到 ranking formula 後，將 channel 數量大幅減少有利於分數計算，且損失的準確度並無影響 Kaggle grade 太多。

ii. 使用 MobileNetV2 的準確度比 MobileNetV3_small 高，但在同樣的前後期處理上，MobileNetV3_small 的規模（計算量*參數量）為 MobileNetV2 的 0.27 倍，MobileNetV2 需在準確度上有約 1.3 pixel difference 的領先才有優勢，實驗後發現 V2 並無如此優勢，因此使用 V3 在最終模型中，實驗結果如下(非最終結果)。



圖五、使用 MobileNetV2 之 loss 曲線



圖六、使用 MobileNetV3 之 loss 曲線

FLOPS: 1.031351216 G

Params: 4.152448 M

avg pixeldiff: 5.108583596827928

avg loss: 0.0007516901241615415

ranking score: 708.5009779051727

FLOPS: 0.474115488 G

Params: 2.585248 M

avg pixeldiff: 5.965919247358571

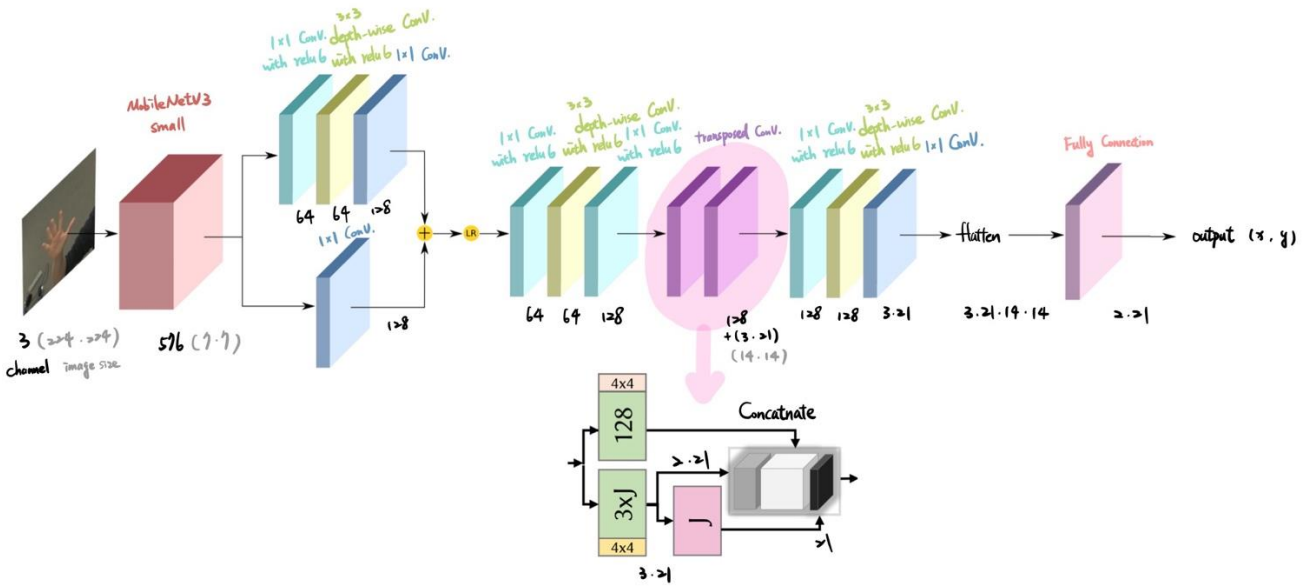
avg loss: 0.0011419530492275953

ranking score: 477.91665120700634

最終模型如下：

FLOPS: 0.302601376 G

Params: 2.007968 M.

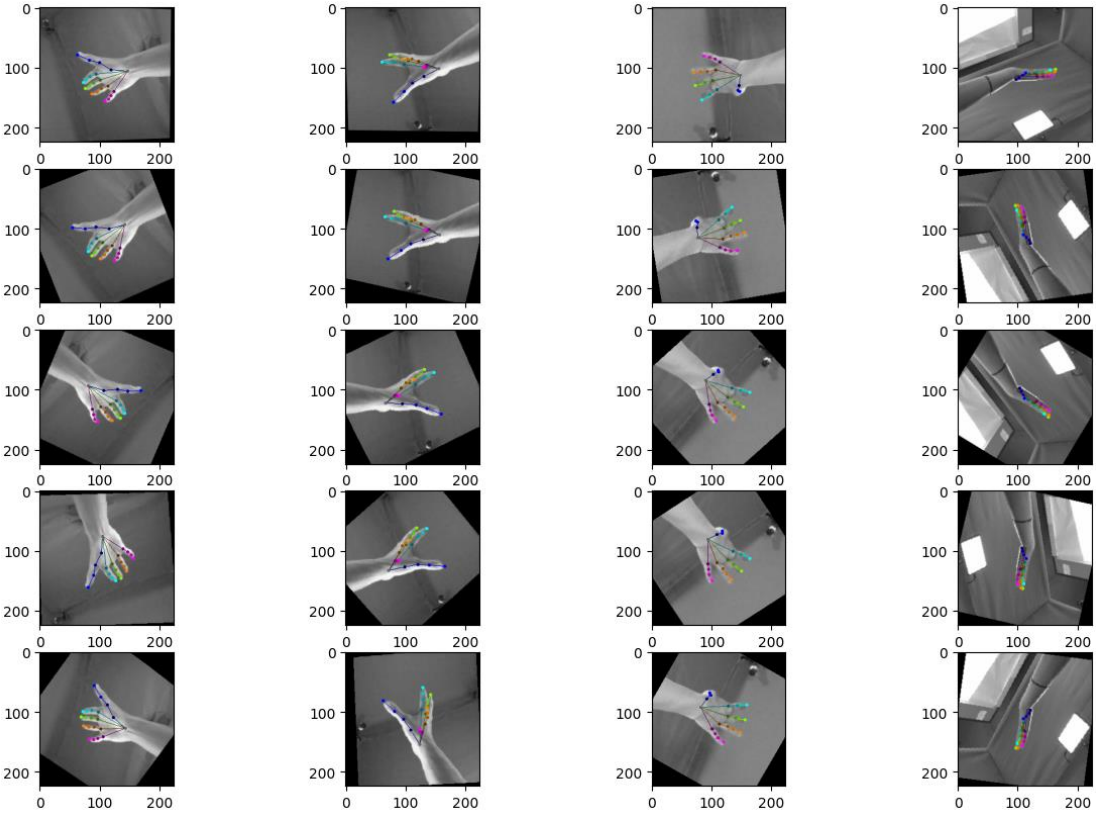


圖七、專題模型架構圖

二、Training skill

1. 資料前處理

訓練資料數量與準確率有正相關，我將原先 30000 筆的 training data 依靠旋轉、鏡像及縮放增加資料量。但實驗後發現縮放對結果為負面影響，因此只使用了隨機旋轉以及隨機水平鏡像處理，同時也發現過多使用前處理新增的資料也會對結果有負面影響，因此最終使用 7 倍的訓練資料。



圖八、資料前處理結果示意圖

2. Optimizer

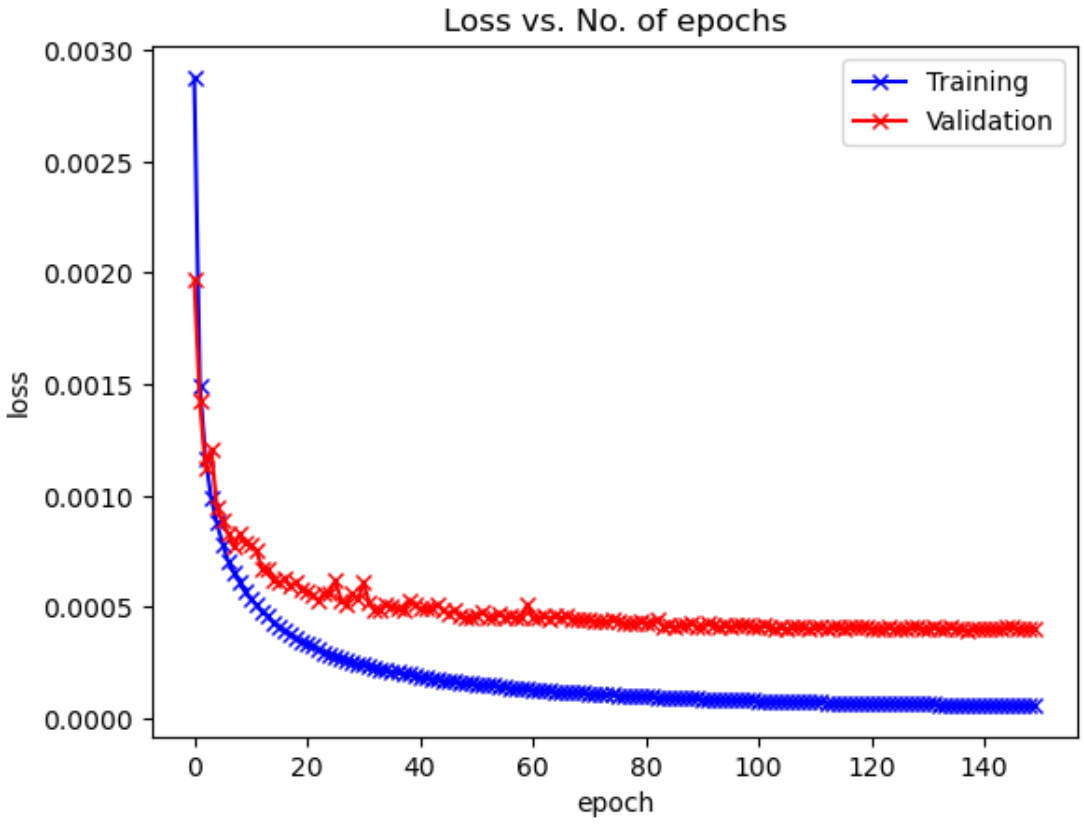
於甲論文中所使用的 Optimizer 是 Adadelta，然而我在嘗試過後表現並沒有比 Adam 好，因此最終我使用的是 Adam，且參數皆為預設。

3. Scheduler

我一開始使用 warn up 的技巧來訓練模型，但後來發現並沒有太大的幫助，因此之後使用“CosineAnnealingLR”，由初始 learning rate：0.0005 至最小的 0.00005，T_max 為（config["rot_num"] * (len(train_dl.dataset) / config["batch_size"]) * config["n_epochs"]），就是半個完整的 cos 波形。

三、訓練結果

1. Loss

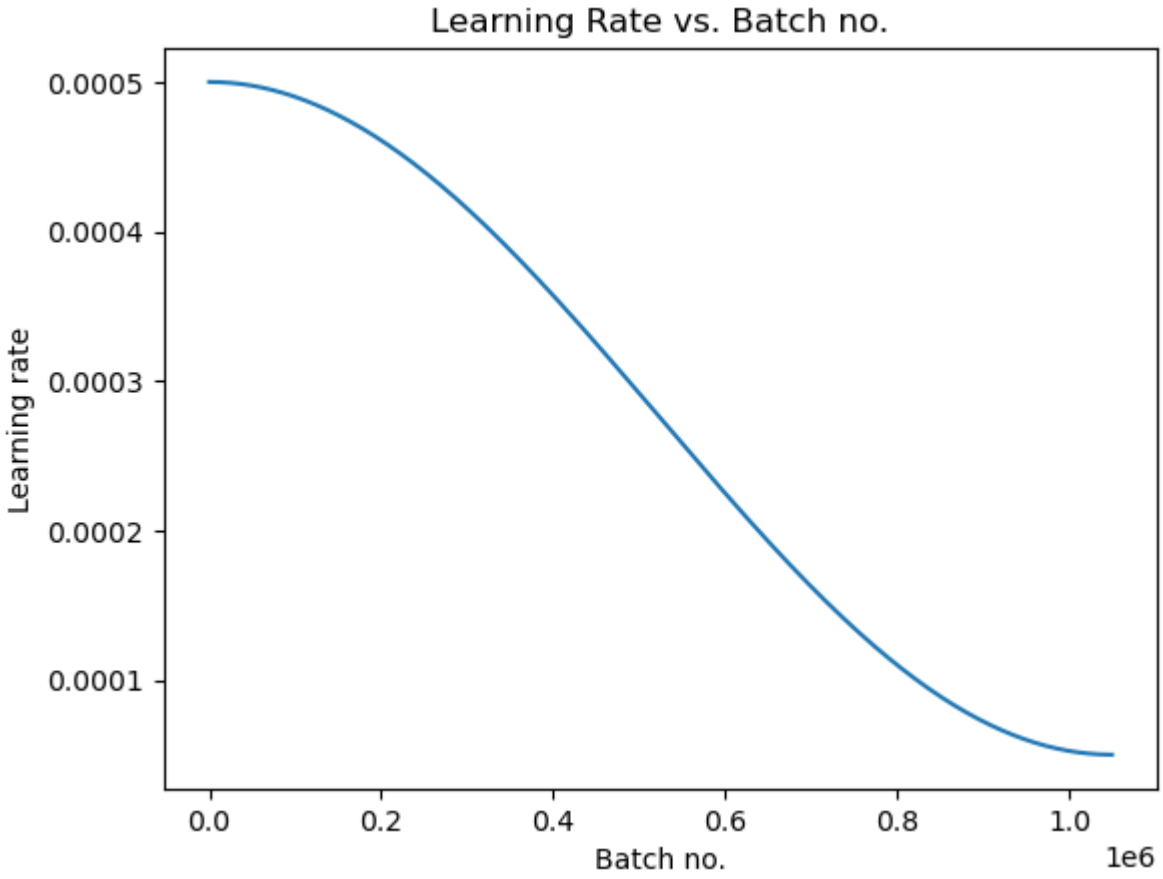


圖九、loss vs. epochs

可以明顯的發現 training loss 與 validation loss 是有相同的變化趨勢，然而 training loss 皆會穩定地低於 validation loss，這是因為 model 的參數調整是基於 training data 與該 model 的差距做 gradient descent，

然而 validation data 僅用來判斷這次 model 調整的好壞，並不會直接與 model 計算有關。因此對於 model 來說，計算上本身就是為了迎合 training data，有這樣的結果也十分合理。基於同個理由，testing loss 也會比 validation loss 更高，因為 model 是選擇使用對於 validation data 有最好表現的參數，與 testing data 毫無相關。

2. Scheduler



圖十、learning rate vs.batch number

3. Result

FLOPS: 0.302601376 G

Params: 2.007968 M.

avg pixeldiff: 6.152803876982536

avg loss: 0.00122355786152184

ranking score: 285.59885234255654

四、Reference

1. Filippos Gouidis et al. "Accurate Hand Keypoint Localization on Mobile Devices." Proceedings of the 16th International Conference on Machine Vision Applications (MVA), National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

2. Dushyant Mehta et al. "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera." ACM Transactions on Graphics (TOG), SIGGRAPH 2017.

3. 速讀論文 Accurate Hand Keypoint Localization on Mobile Devices https://allen108108.github.io/blog/2020/01/11/%5B%E8%AB%96%E6%96%87%5D%20%E9%80%9F%E8%AE%80%E8%AB%96%E6%96%87%20Accurate%20Hand%20Keypoint%20Localization%20on%20Mobile%20Devices/

4. 手部關鍵點 (Hand Keypoints) 的預測及其應用 https://allen108108.github.io/blog/2020/06/06/%E6%89%8B%E9%83%A8%E9%97%9C%E9%8D%B5%E9%BB%9E%20(Hand%20Keypoints)%20%E7%9A%84%E9%A0%90%E6%B8%AC%E5%8F%8A%E5%85%B6%E6%87%89%E7%94%A8/

5. Efficient CNN 介紹(二)：MobilenetV2 https://medium.com/ai-academy-taiwan/efficient-cnn-%E4%BB%8B%E7%B4%B9-%E4%BA%8C-mobilenetv2-7809721f0bc8

6. [論文筆記] MobileNet 演變史-從 MobileNetV1 到 MobileNetV3 https://chihangchen.medium.com/%E8%AB%96%E6%96%87%E7%AD%86%E8%A8%98-mobilenetv3%E6%BC%94%E8%AE%8A%E5%8F%B2-f5de728725bc