

# Light-Weight Facial Landmark Prediction Challenge

B07703078 艾芯、B08902092 陳翰雯、R09945067 王馨

## Introduction

Face landmark 是許多面部影像處理的重要步驟，例如虛擬面部重演、表情追蹤、生物特徵識別、駕駛員狀態追蹤等。多年來，儘管面臨各種挑戰，如大動作、面部遮擋、不尋常的情緒等，人們仍然實現了許多高性能的方法和想法。然而，隨著科技的進步，新型的移動設備和其他 IoT (物聯網) 的開發越來越需要快速推理和低成本模型。

在本次專題中，我們將提出一個以 15 MB 為上限，但在測試數據集中可以顯示出高性能結果的模型。我們所使用的模型為眾所周知的輕量級模型 MobileNet，並進一步提高其性能。

## Method

在這次實驗中，我們使用 MobileNet 來實現 face landmark，並進一步改良模型以提高其性能。我們以簡單 CNN 作為基礎模型，並與 MobileNet、MobileNetV2 和 MobileNetV3 進行比較。以下我們將簡述模型特徵與實驗方法。

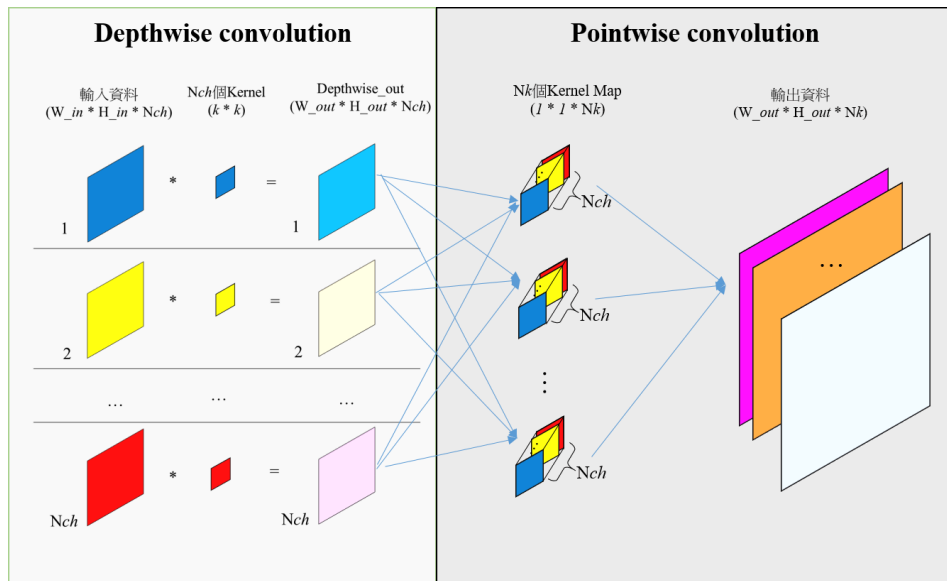
### 一、模型簡介

#### 1. MobileNet (V1)

傳統 CNN 為了提升準確度，將神經網路增加越來越多層，導致運算量越趨複雜且運算速度慢。因此 MobileNetV1 便是為解決此問題，於 2017 年 Google 所提出的一個模型架構。透過 Depthwise Separable Convolution，大幅減少傳統 CNN 的運算量，其使用的方法便是將傳統的 CNN 拆解成以下兩個步驟：

- (1) Depthwise convolution：對於 Input 中的每層 channel 各自單獨做 convolution。
- (2) 1x1 convolution (pointwise convolution)：將經過 Depthwise convolution 後的 output feature map 轉換成為我們所希望的 output 大小。

關於 MobileNet 的結構，可參照下圖。



## 2. MobileNetV2

MobileNetV2 是在 V1 基礎上的改良版，提出一個全新的模型架構：The inverted residual with linear bottleneck。

- (1) **Linear Bottleneck**：研究發現，Depthwise Separable Convolution 中有大量卷積核為 0，即有很多卷積核沒有參與實際計算。其原因在於 ReLU，當 Feature Map 經過 ReLU 激活後，所有值都會大於等於零，造成大量訊息的流失。因此有別於 V1 做 Depthwise separable convolution，MobileNetV2 先透過 Pointwise convolution 擴張 Feature Map 深度。也就是將 pointwise convolution 中的 ReLU 都換成線性函數，以避免非線性函數使重要資訊流失。
- (2) **Inverted residuals**：MobileNetV2 中的神經網路，借鑑了 ResNet 的殘差結構，在 V1 網絡結構基礎上加入了跳躍連接。由於 Depthwise 本身沒有改變通道的作用，為了能讓深度卷積能在高維上運作，V2 提出在深度卷積之前加一個擴充通道的卷積操作。也就是對 bottleneck 使用 shortcuts，而並非對 expansion layer 使用 shortcuts 連結（因和過去連接 shortcut 的方法相反，稱為 Inverted residuals）。

### 3. MobileNetV3

MobileNetV3 不僅有 V1 的 Depthwise separable convolution、V2 的跨接與先放大再壓縮觀念，另外加入了 Squeeze-and-Excitation Networks，並以 NAS 神經架構搜索確定網絡結構。

- (1) Activated function: 其將部分的 ReLU 使用 H-swish 取代，Sigmoid 則使用 H-sigmoid 取代。H-swish 是參考 swish 函數設計，主要是由於 swish 函數運算較慢，而 H-swish 能提高準度。

$$\text{h-swish}[x] = x \frac{\text{ReLU6}(x + 3)}{6}$$

- (2) Squeeze-and-Excitation 結構 (SENet)：MobileNetV3 在 V2 的基礎上引入 SENet 結構。SENet 的核心思想在於通過網絡去學習特徵權重，使得有效的特徵圖權重大，無效或效果小的特徵圖權重小的方式訓練模型達到更好的結果。
- (3) 改良尾部結構：在 MobileNetV2 中，在 Avg Pooling 之前，存在一個 1\*1 的卷積層，目的是提高 feature map 的維度，更有利於結構的預測，但也會帶來一定的計算量。因此 MobileNetV3 改良了尾部網路架構，以減少運算量，並維持原準確度。

## 二、實驗方法

我們使用各個模型來進行 face landmark 預測的訓練，其中模型種類包含基礎 CNN、MobileNet、MobileNetV2 和 MobileNetV3。各模型大小均控制在 13MB 至 15MB 之內，且訓練過程皆使用 MSE (mean square error) 來計算 loss。下表為本次實驗所控制的各參數。

表：訓練模型參數表

Parameters	Value
epoch number	20
learning rate	0.001
learning rate schedule	10, 15
learning rate gamma	0.1
momentum	0.9
weight decay	0.0005

訓練完成後，我們使用所訓練出的模型對 validation set 的圖片進行 face landmark 的預測，並使用 normalized mean error (NME) 比較其結果的成效。





## Results

關於本次實驗成果，我們使用 NME 值來比較各模型的預測成效，如下表。

Model	NME (%)
Base CNN	7.89
MobileNet	6.38
MobileNetV2	5.82
MobileNetV3	4.56

由上表我們可以發現，各模型的預測成效為 MobileNetV3 > MobileNetV2 > MobileNet > Base CNN。由此可知，V1 的 Depthwise separable convolution、V2 的跨接與先放大再壓縮觀念和 SENet 的結構皆有助於訓練的成效。

各模型預測圖片的成效如下表。

Base CNN	MobileNet
	
MobileNetV2	MobileNetV3
	

由圖片結果可看出，從 Base CNN 到 MobileNetV3，模型預測有越來越準確的趨勢。

## Conclusion

本次專題中，我們使用 MobileNet 來進行 face landmark 的預測，其中以 MobileNetV3 的訓練成效最好，其 NME 為 4.56。

## Reference

1. <https://arxiv.org/pdf/1704.04861.pdf>
2. <https://arxiv.org/pdf/1602.07360.pdf>
3. <https://arxiv.org/pdf/1709.01507.pdf>
4. <https://cinnamonaitaiwan.medium.com/cnn%E6%A8%A1%E5%9E%8B-resnet-mobilenet-densenet-shufflenet-efficientnet-5eba5c8df7e4>
5. [https://github.com/lzx1413/pytorch\\_face\\_landmark](https://github.com/lzx1413/pytorch_face_landmark)
6. <https://github.com/d-li14/mobilenetv2.pytorch>
7. <https://github.com/d-li14/mobilenetv3.pytorch>
8. <https://github.com/leaderj1001/MobileNetV3-Pytorch>