Day 46

深度學習與電腦視覺

學習馬拉松

陪跑專家：周俊川

CUPOY

輕量化模型簡介-MobileNet

重要知識點

CUPOY

- 認識輕量化模型的方法
- MobileNet 架構設計
  - Separable Convolution

為什麼需要輕量化模型？

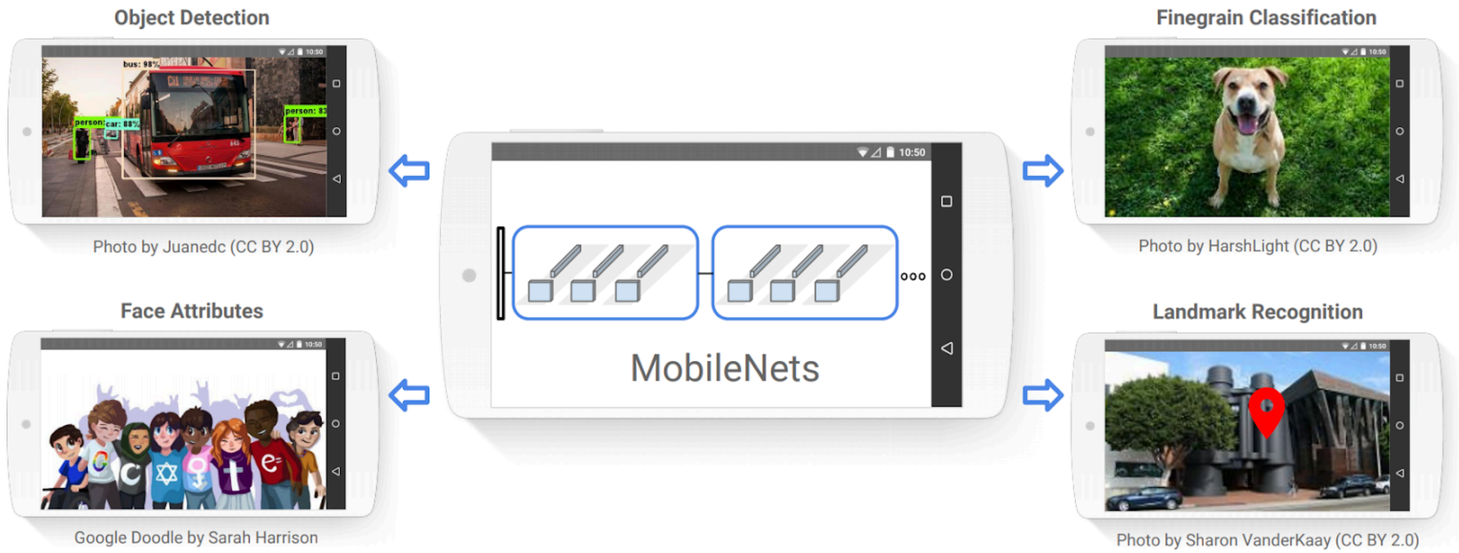
CUPOY

- 為了追求準確率，深度學習模型架構越來越深，越趨複雜，導致在真實的應用場景，如手機端和嵌入式設備，需要低內存以及追求速度效能的應用上難以被廣泛應用。

- 輕量化模型的一個重點在於提升模型推論速度同時保持不弱的準確率。

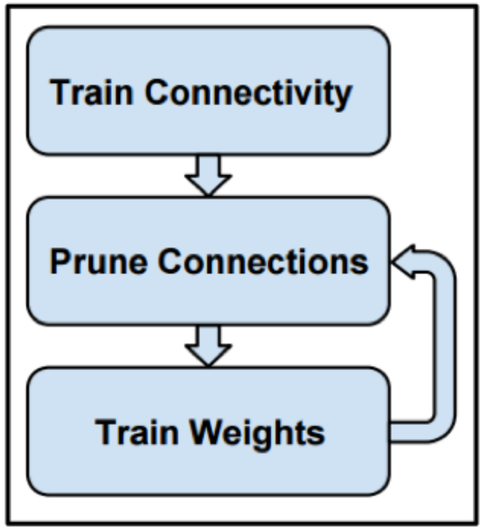
輕量化模型方式

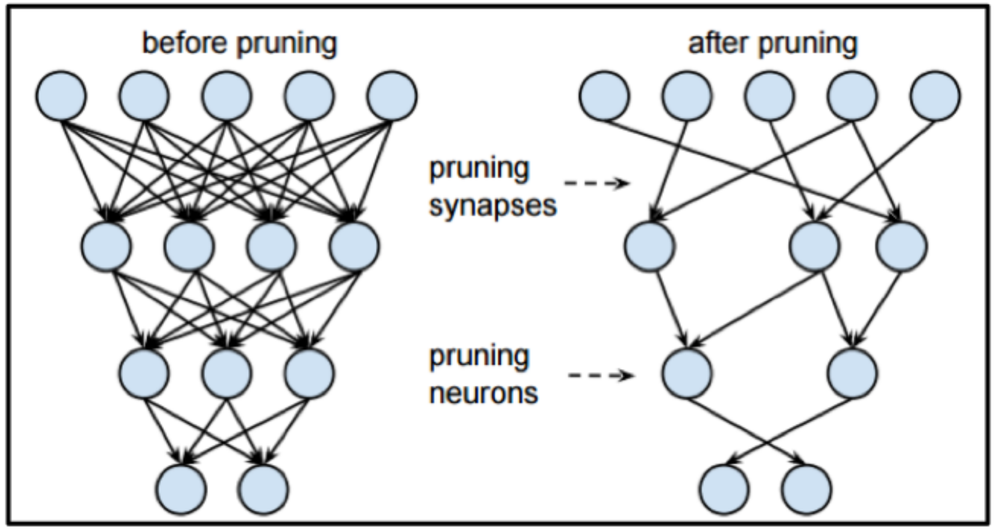
- Model prunning: 修剪模型
- Quantization: 降低附點數
- Architecture Design: 網路架構設計

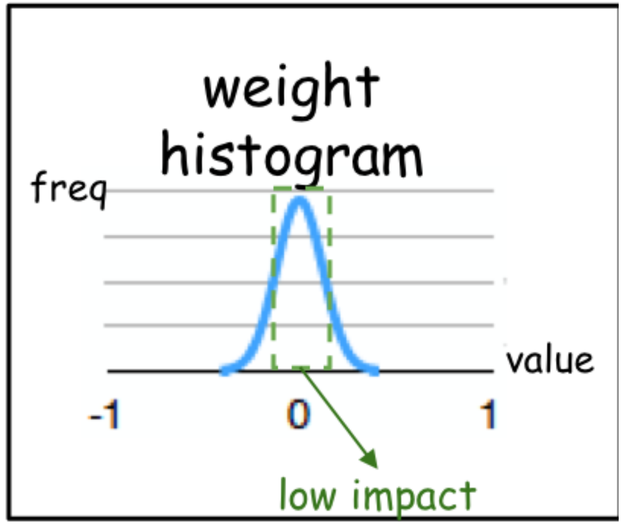


Model Prunning 修剪模型

模型絕大部分時候 weights 都接近 0, impact 較小，因此可以對模型做 prunning, 把不重要的 weights 砍掉，再繼續 train.



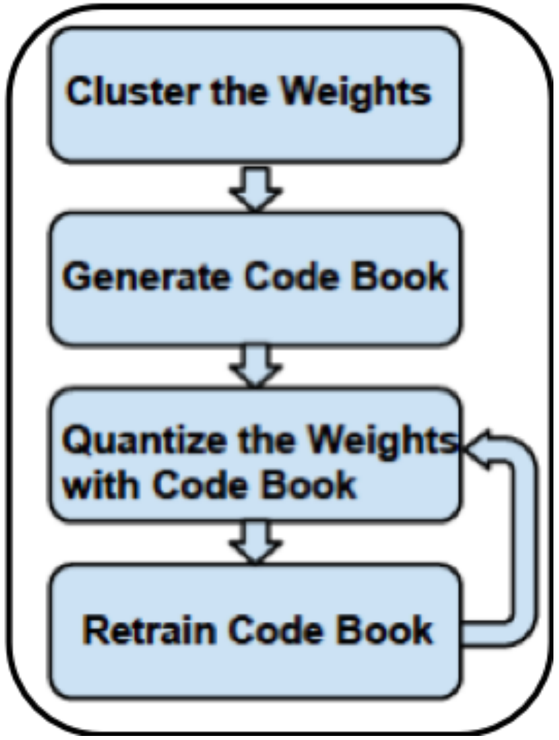


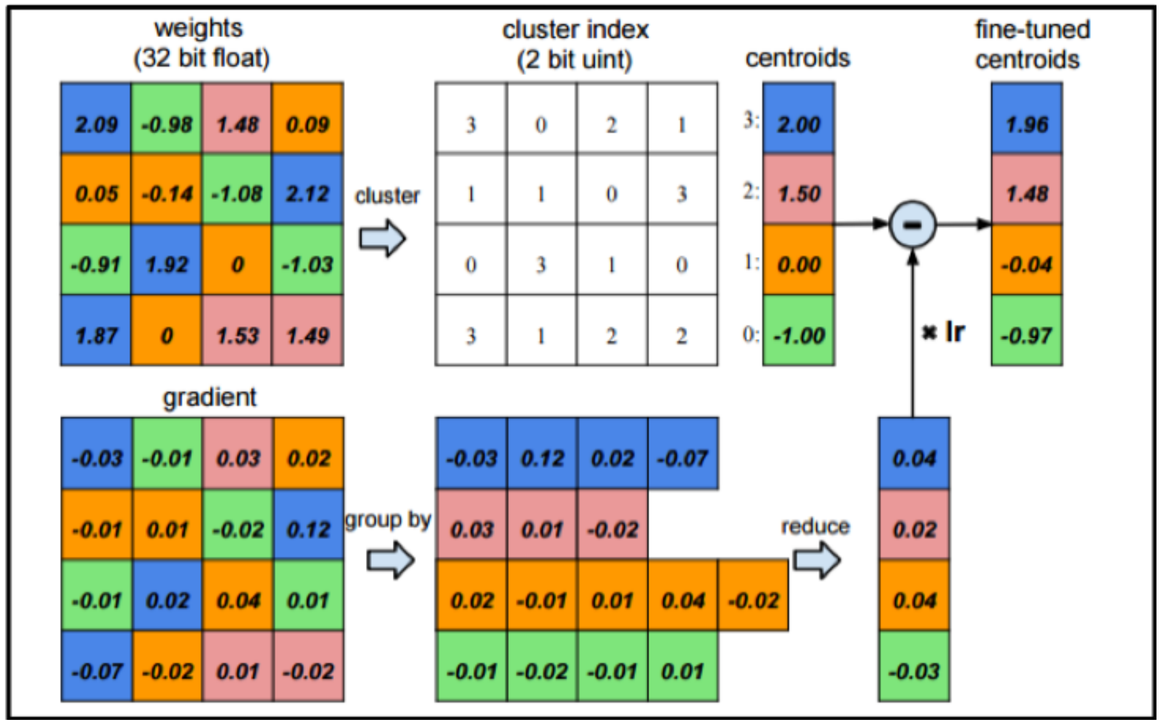


Quantization 降低附點數

CUPOY

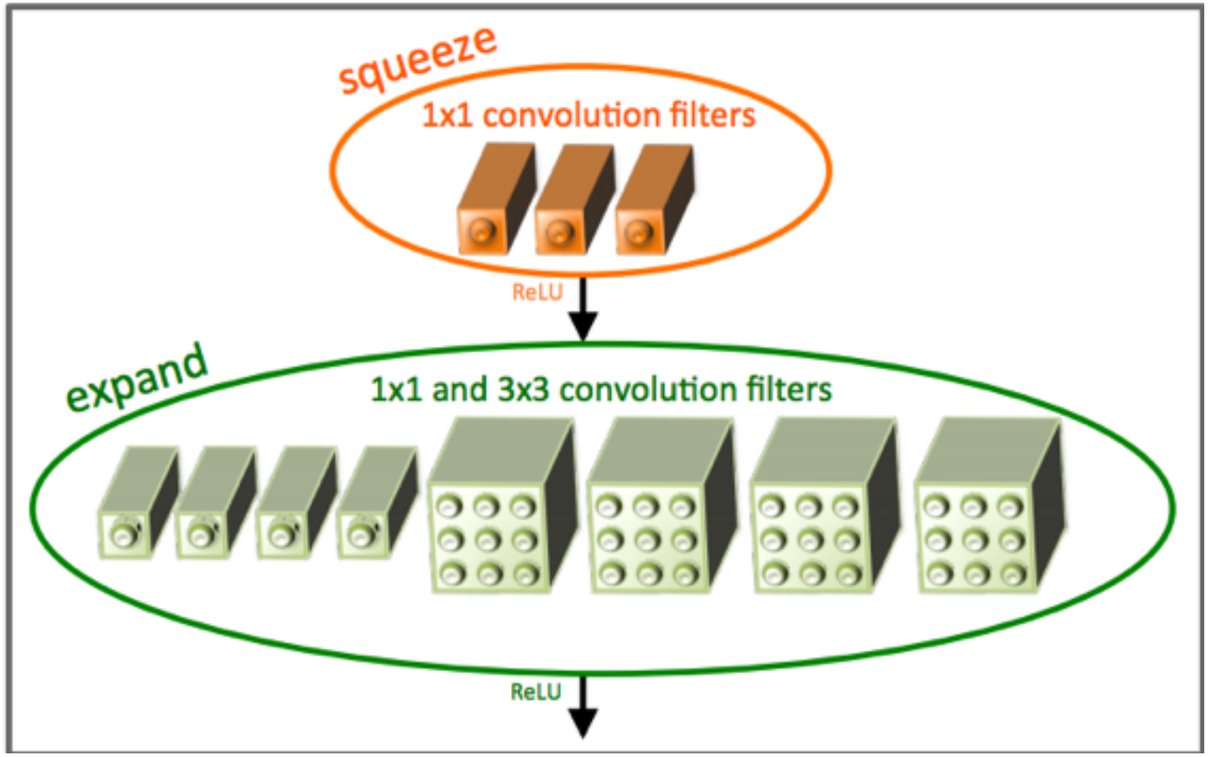
一般模型訓練預設使用 fp32/ fp64, 可以直接改用 fp16, 抑或是對於權重做分群，再以 fp16 / int8 取代之。





Architecture Design 網路架構設計

針對網路 / 計算架構本身做一些修改，目標是在相同輸入 / 輸出維度的情況下，能不能降低參數量及計算量。



MobileNet: 一種輕量化的網路架構

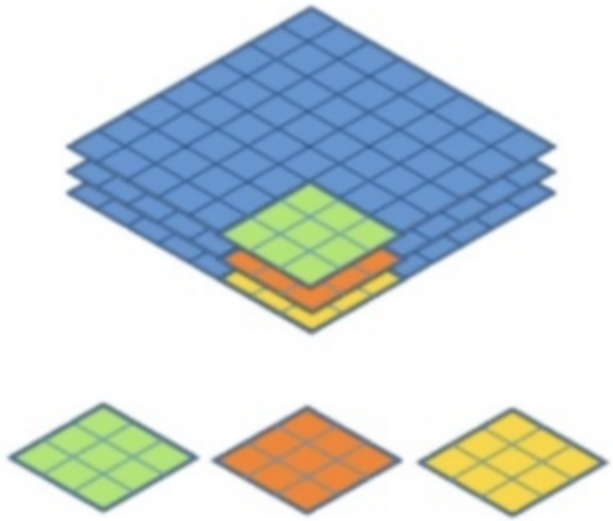
CUPOY

使用了 (separable convolution) depthwise convolution + pointwise convolution 構建的輕量級神經網路，並通過兩個超參數使得開發人員可以基於自己的應用和資源選擇合適模型。

Depthwise Convolution

CUPOY

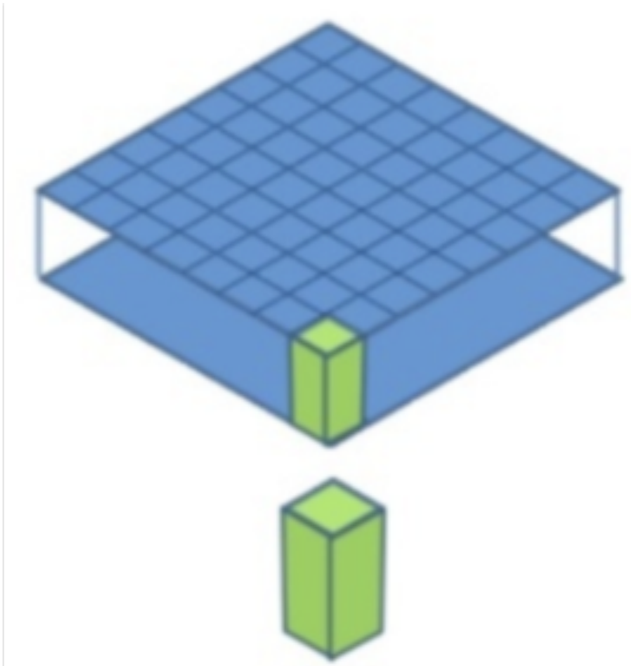
Depthwise convolution 是指每個 input channels 都會採用不一樣的 / 自己的 filter 來做 conv. computation (一般 convolution 是 filters 用在所有 input channels)



Pointwise Convolution

CUPOY

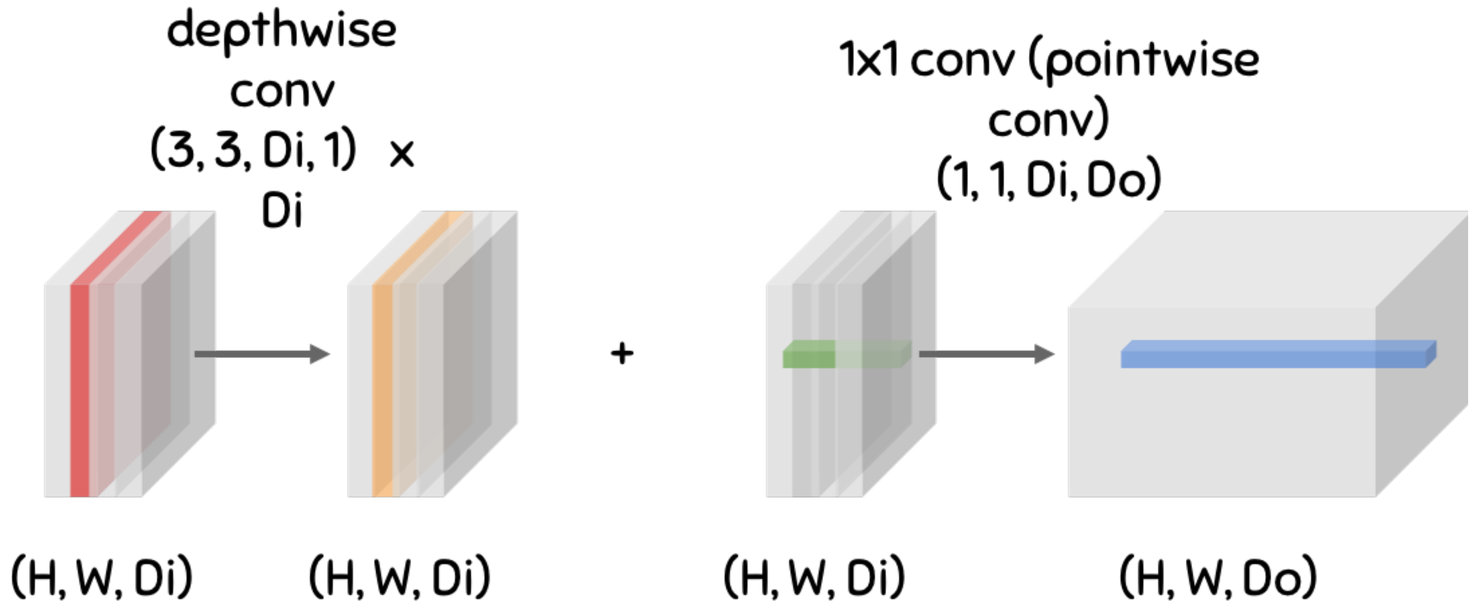
Pointwise convolution 就是 1x1 convolution.



Separable Convolution

CUPOY

Depthwise + pointwise , 這兩個 convolution 加在一起輸出與一般 convolution 相等，但參數量 / 計算量卻降低



 # All convs are followed by BN + Relu

Separable Convolution 計算量

CUPOY

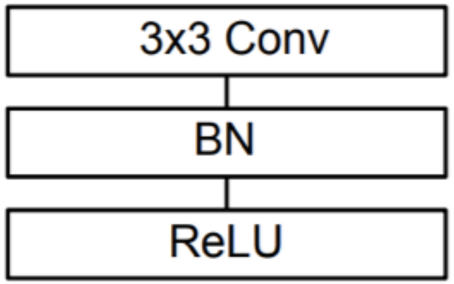
- 一般 convolution 計算方式為
  - filter size * d * D * H * W
- Depthwise convolution 為
  - filter size * d * H * W

Input : (h,w,d)

Output : (H.W.D)

Convolution types	Computation	Ratio
3x3 Conv	3*3*d*D*H*W	1
3x3 DW-conv	3*3*d*H*W	1/D
1x1 conv (PW-conv)	1*1*d*D*H*W	1/9
Separable Conv	9dHW + dDHW	1/D + 1/9

Mobilenet 架構



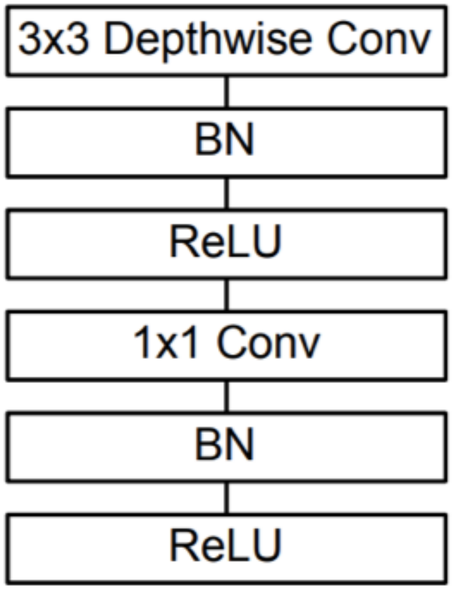


Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Mobilenet 與其他模型比較

CUPOY

Table 8. MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

Table 9. Smaller MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
0.50 MobileNet-160	60.2%	76	1.32
Squeezenet	57.5%	1700	1.25
AlexNet	57.2%	720	60

Mobilenet 超參數調整

CUPOY

- Width multiplier (α)
  - channel (D) 數的縮放因子，介於 0 - 1
  - - D *= **αD**
- Resolution multiplier **(β)**
  - input resolution (H, W) 的縮放因子，介於 0 - 1
  - - H *= **βH; W *= βW**

Mobilenet 超參數性能分析

CUPOY

Table 6. MobileNet Width Multiplier

Width Multiplier	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
0.75 MobileNet-224	68.4%	325	2.6
0.5 MobileNet-224	63.7%	149	1.3
0.25 MobileNet-224	50.6%	41	0.5

Table 7. MobileNet Resolution

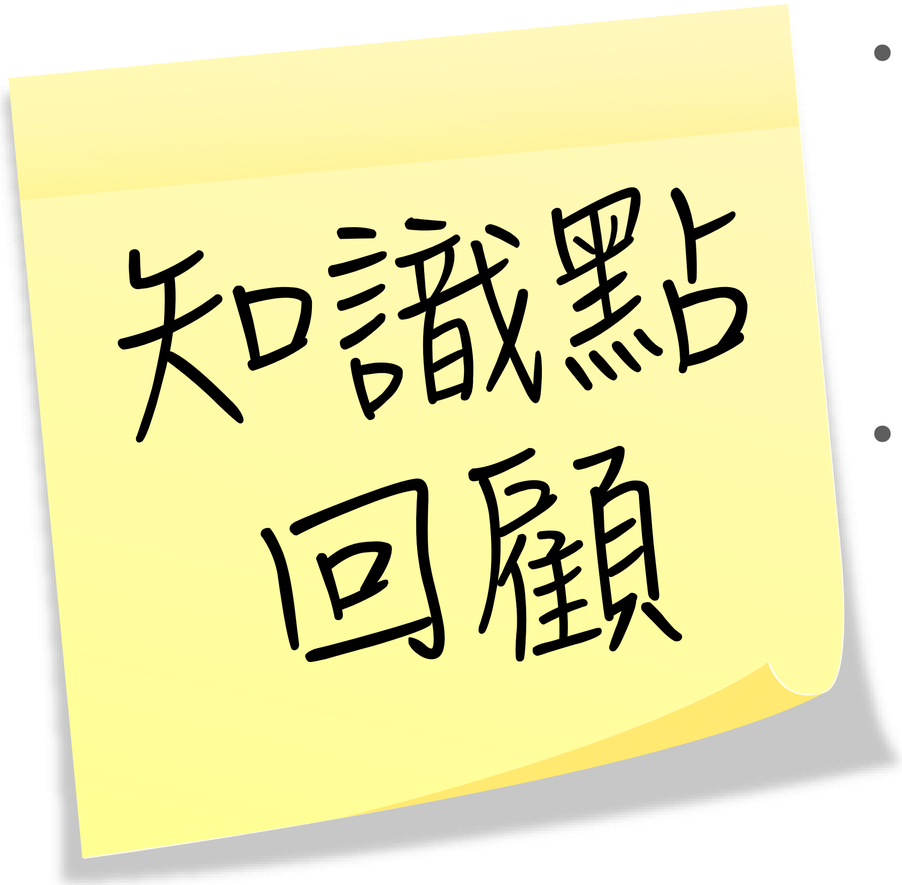
Resolution	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
1.0 MobileNet-192	69.1%	418	4.2
1.0 MobileNet-160	67.2%	290	4.2
1.0 MobileNet-128	64.4%	186	4.2

Separable Convolution 計算量 (Revisited)

CUPOY

Input: (h, w, d) ， Output (H, W, D)
Multiplier: width **α**, resolution **β**

Convolution types	Computation	Ratio
Regular 3x3 Conv	3*3*d*D*H*W	1
3x3 DW-conv	$3*3*\alpha d*\beta H*\beta W$	$\alpha\beta^2/D$
1x1 conv	$1*1*\alpha d*\alpha D*\beta H*\beta W$	$\alpha^2\beta^2/9$
Separable Conv	$9\alpha\beta^2dHW + \alpha^2\beta^2dDHW$	$\alpha\beta^2/D + \alpha^2\beta^2/9$



- 認識輕量化模型的方法
  - Model prunning: 修剪模型
  - Quantization: 降低附點數
  - Architecture Design: 網路架構設計
- MobileNet 架構設計
  - Separable Convolution
    - 分為 depthwise convolution 和 pointwise convolution (1x1 conv.) 達到降低計算量和參數量的目的
    - 有 width 和 resolution 超參數調整空間

參考資料

CUPOY

- 李宏毅 Youtube [Next Step for Machine Learning 系列 - Network Compression]
- https://arxiv.org/pdf/1704.04861.pdf

解題時間 Let’s Crack It

CUPOY

請跳出 PDF 至官網 Sample Code & 作業開始解題