深度學習
Deep Learning

可解釋性人工智慧
Explainable AI

Instructor: 林英嘉 (Ying-Jia Lin)
2025/05/19

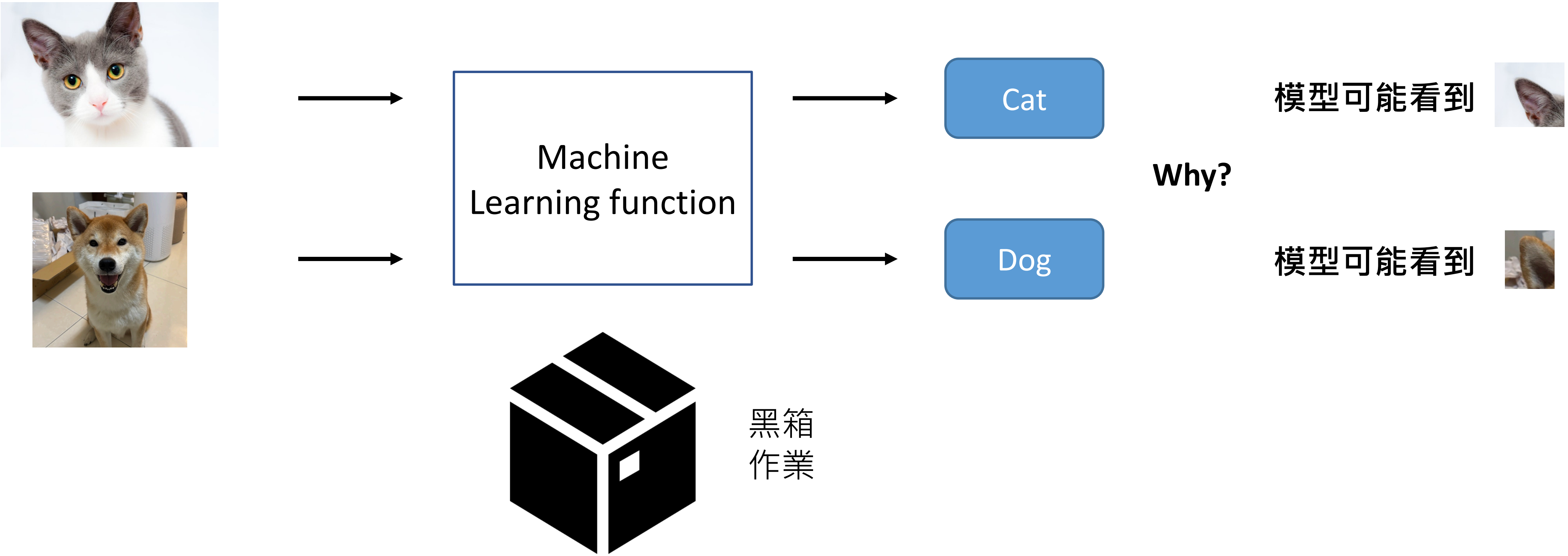
Course GitHub

Slido # DL_0519

Outline

- Introduction
- Class-activation Map (CAM)
- Grad-CAM
- Code

Prediction of a Machine Learning Model



Machine Learning function

Cat

Dog

Why?

模型可能看到

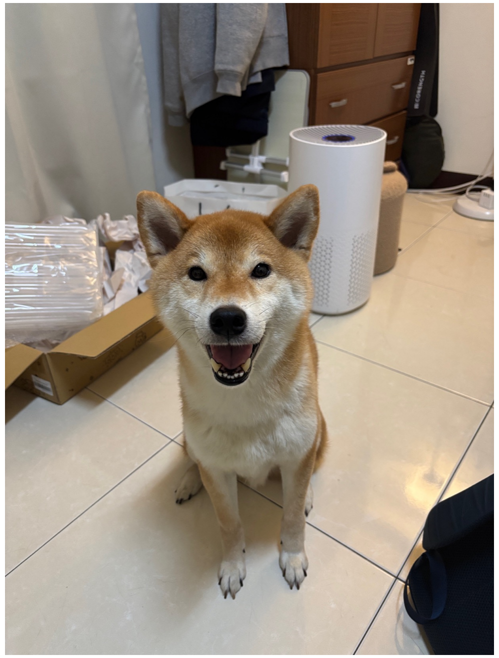
模型可能看到

黑箱
作業

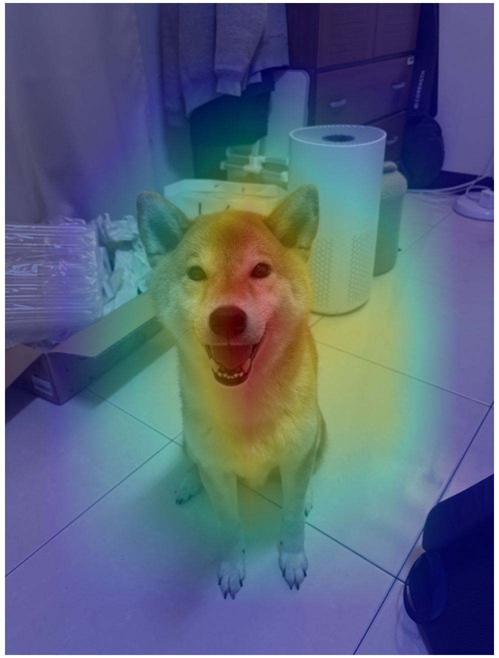
Example: Classification

- Test model: resnet18 pre-trained on ImageNet-1K
- Method: Class activation map (CAM)

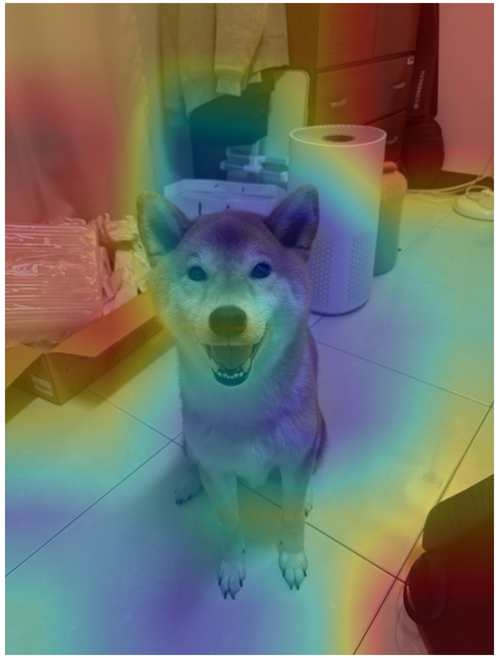
Input Image



Pembroke Welsh Corgi



window shade



Visualizing Feature Maps in a CNN

- 151st channel on the conv5 layer of a deep neural network trained on ImageNet

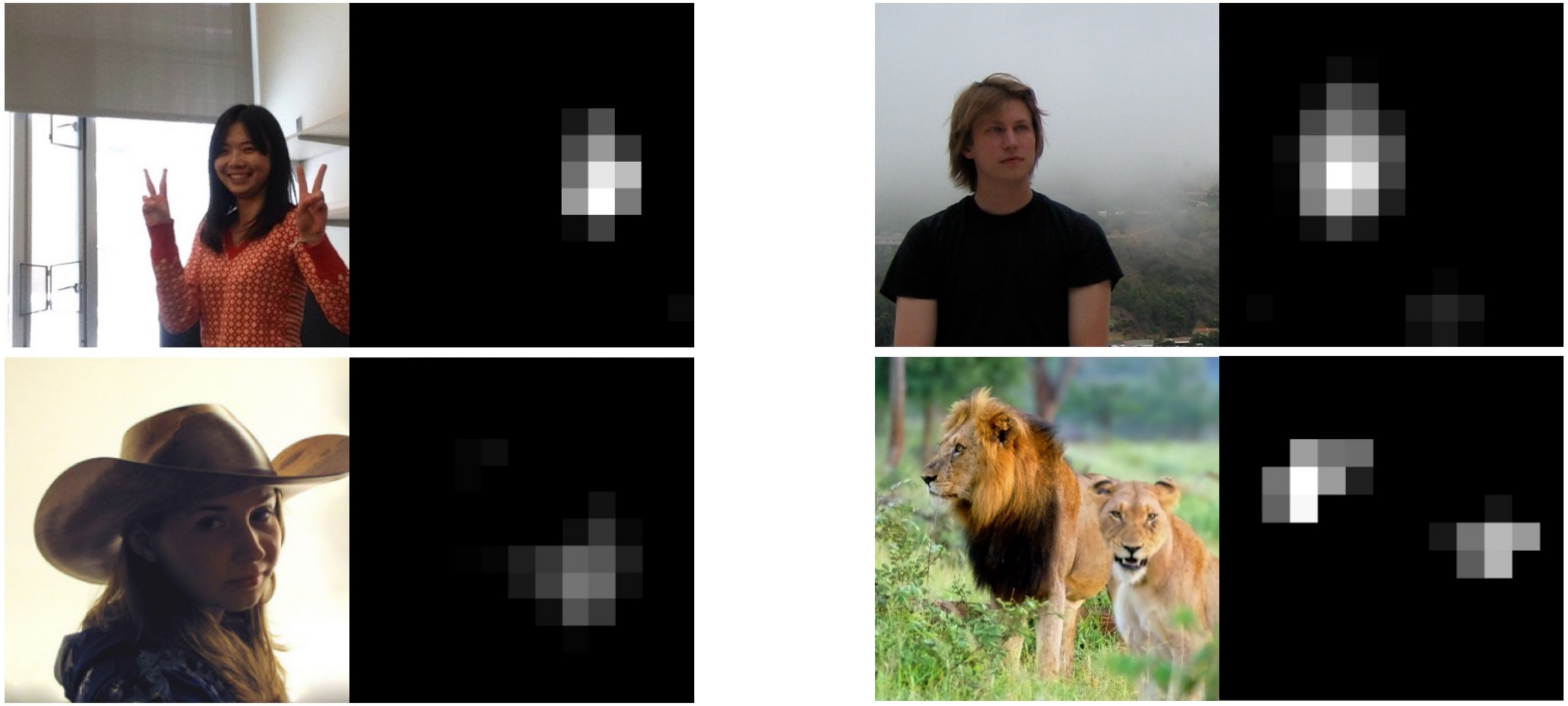


Figure source: Yosinski, Jason, et al. “Understanding neural networks through deep visualization.” 2015 ICML Deep Learning Workshop.

Example with SHAP

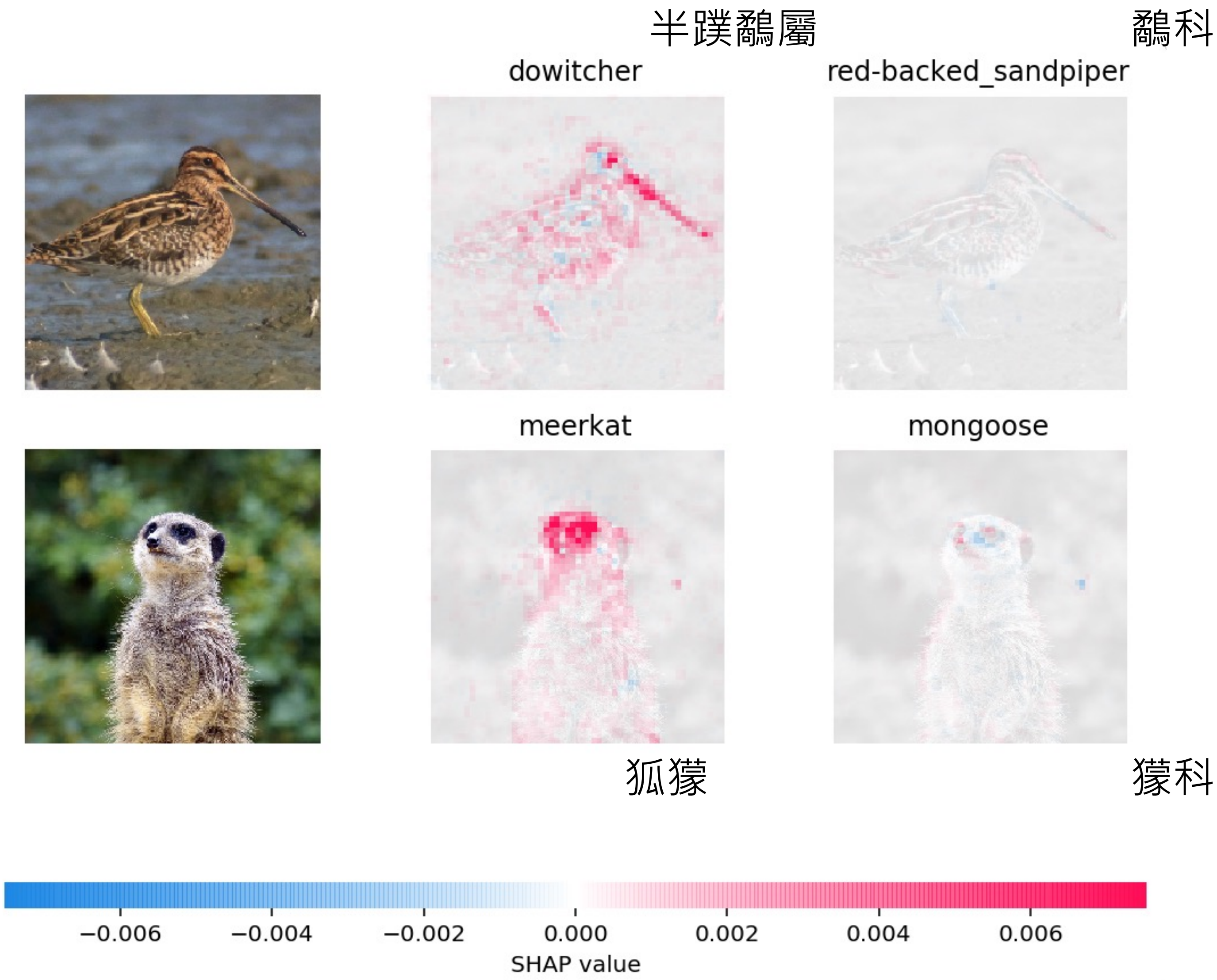
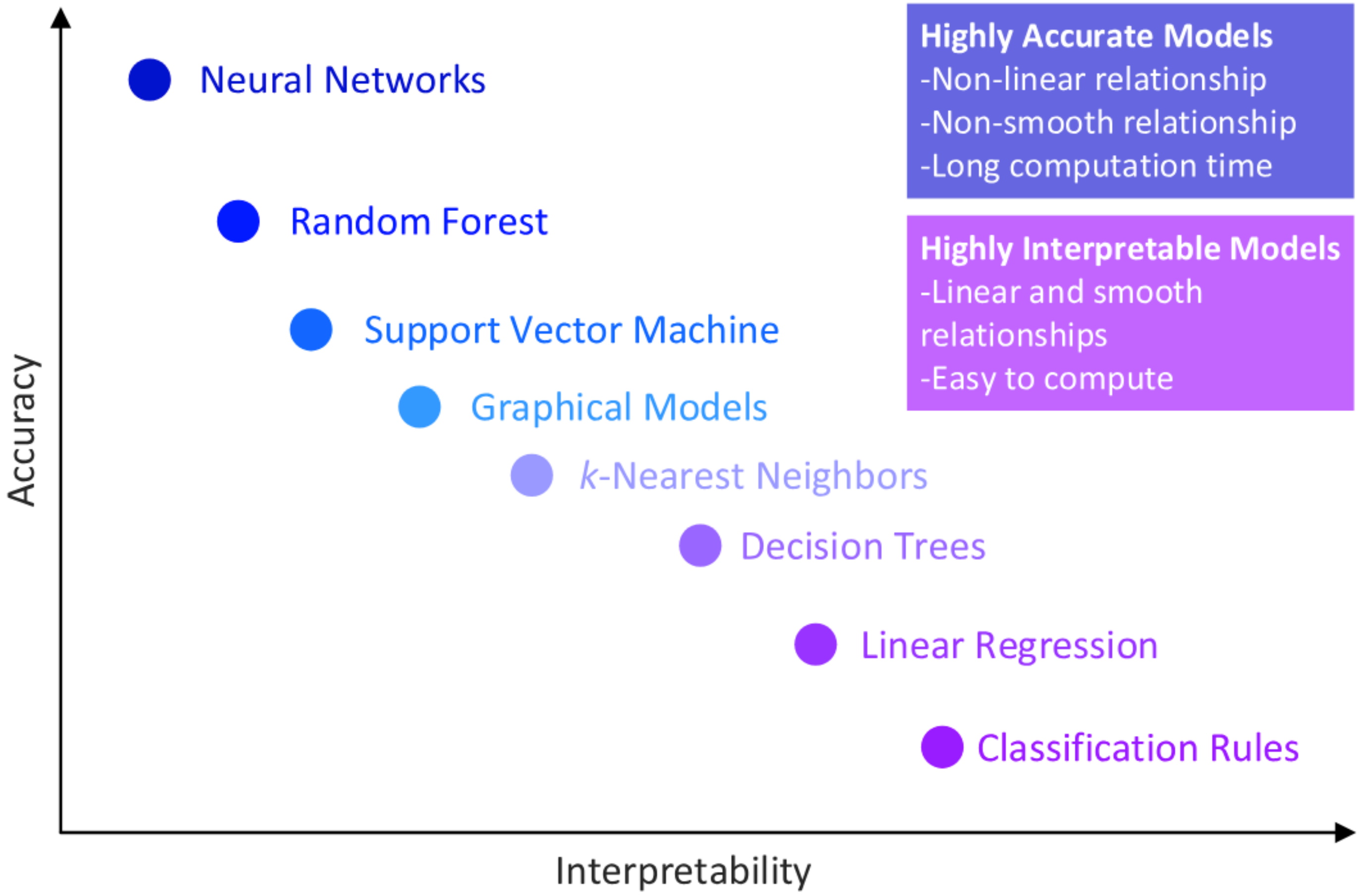


Figure source:
https://github.com/shap/shap

模型的可解釋性與效能



Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time

Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute

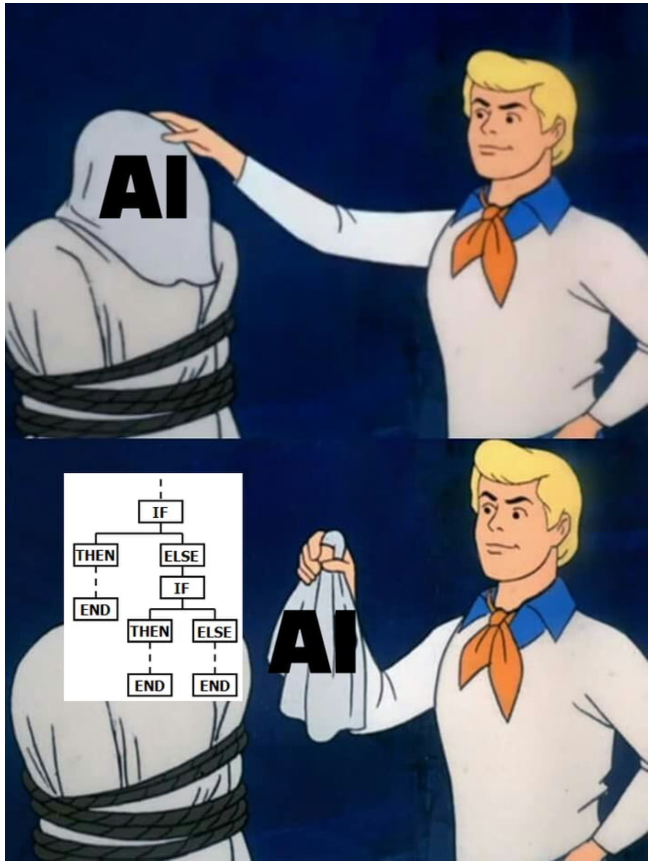


Figure source: Morocho-Cayamcela, Manuel Eugenio, Haeyoung Lee, and Wansu Lim. "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions." IEEE access 7 (2019): 137184-137206.

Figure source:
https://9gag.com/gag/aOYA1mE?ref=pn.mw

Why is Explainable AI important?

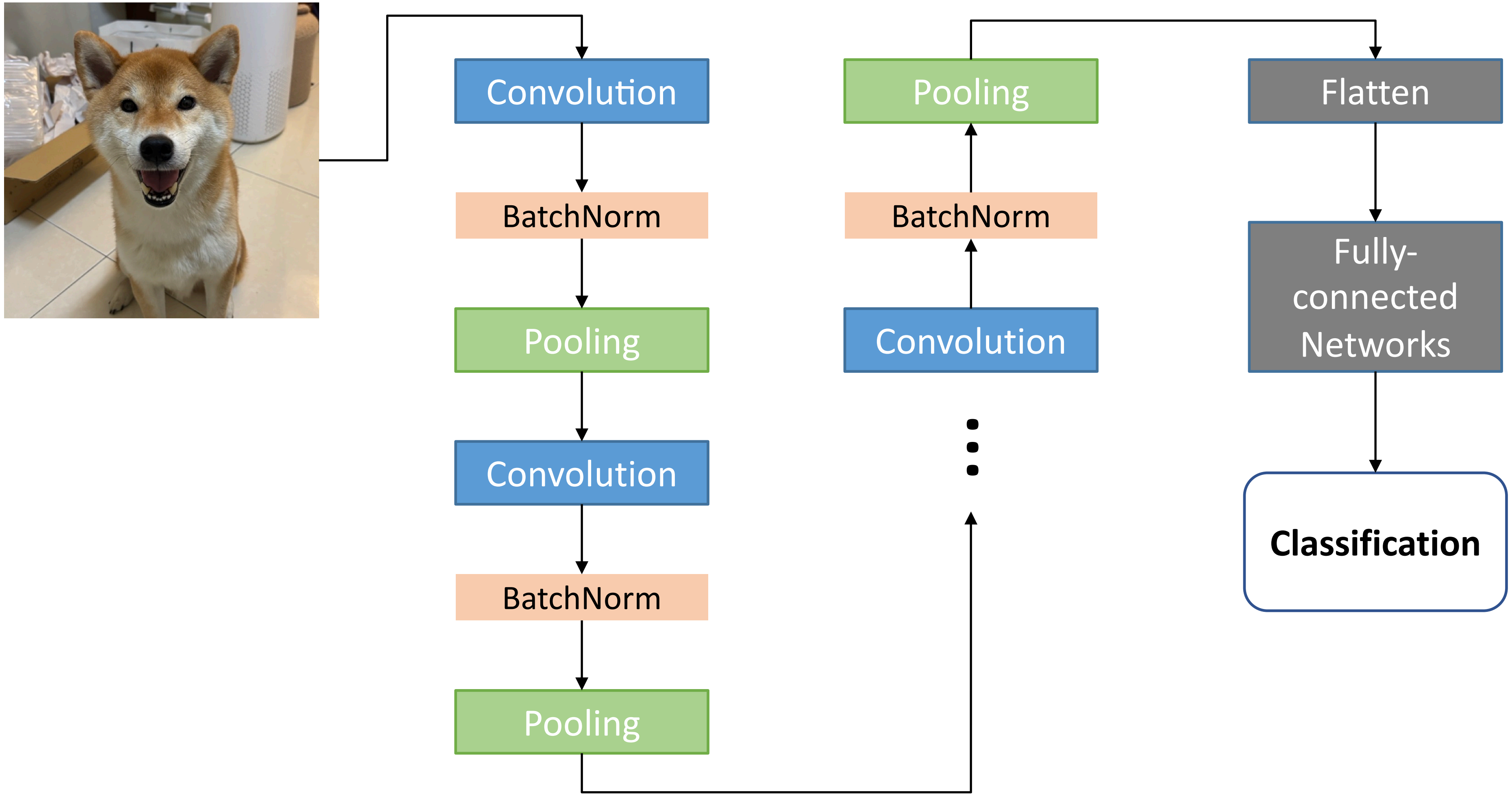
- 確認機器學習模型的判斷合理
  - 建立信任 (使用者 / 政府)
- 改進機器學習模型
  - 從模型輸出找出改進的策略

https://www.ibm.com/think/topics/explainable-ai

Convolutional Neural Networks (Recap)

[Recap] The whole Process of a CNN

CNN: Convolutional Neural Networks



[Recap] Convolutions (stride = 1)

1	1	1	1	0	0
0	1	1	0	1	0
0	0	1	1	0	0
0	0	1	1	1	0
0	0	0	1	1	0
0	0	0	0	0	0

Stride = 1

1	0	0
0	1	0
0	0	-1

Filter

Element-wise multiplication

1

[Recap] Convolutions (stride = 1)

1	1	1	1	0	0
0	1	1	0	1	0
0	0	1	1	0	0
0	0	1	1	1	0
0	0	0	1	1	0
0	0	0	0	0	0

Stride = 1

1	**0**	**0**
0	**1**	**0**
0	**0**	**-1**

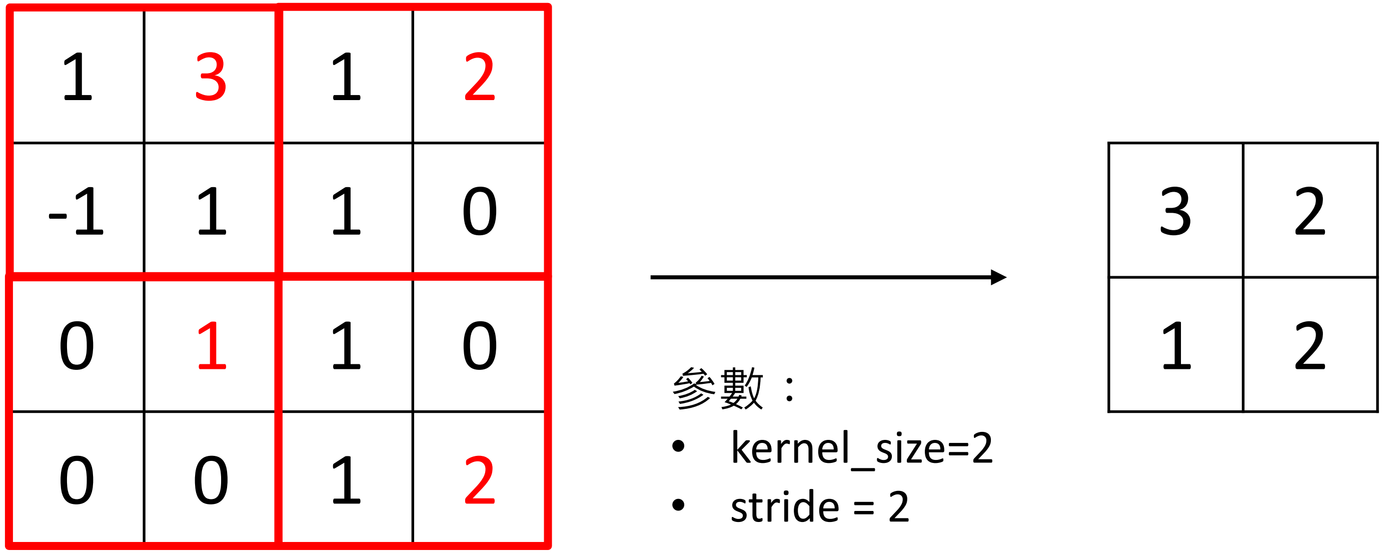
Filter

Element-wise multiplication

1	1	1	2
-1	1	1	0
0	0	1	2
0	0	1	2

feature map

[Recep] 2x2 Pooling (example of Max Pooling)



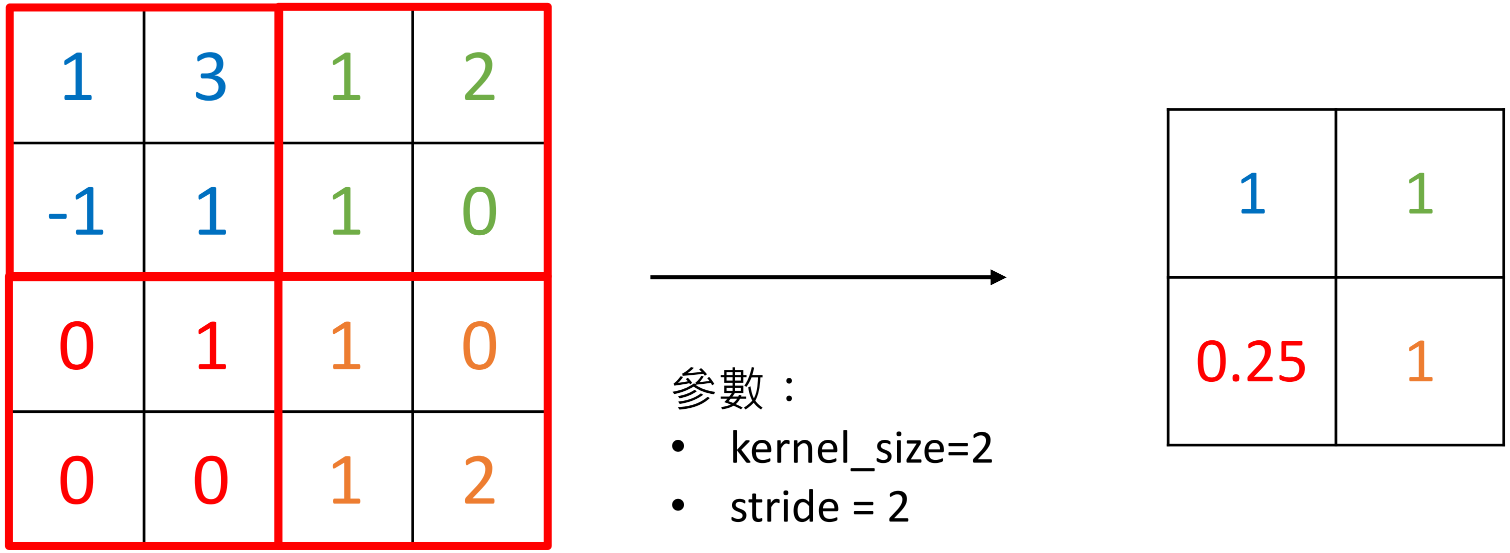
1	3	1	2
-1	1	1	0
0	1	1	0
0	0	1	2

3	2
1	2

參數：

- kernel_size=2
- stride = 2

[Recep] 2x2 Pooling (example of Average Pooling)



1	3	1	2
-1	1	1	0
0	1	1	0
0	0	1	2

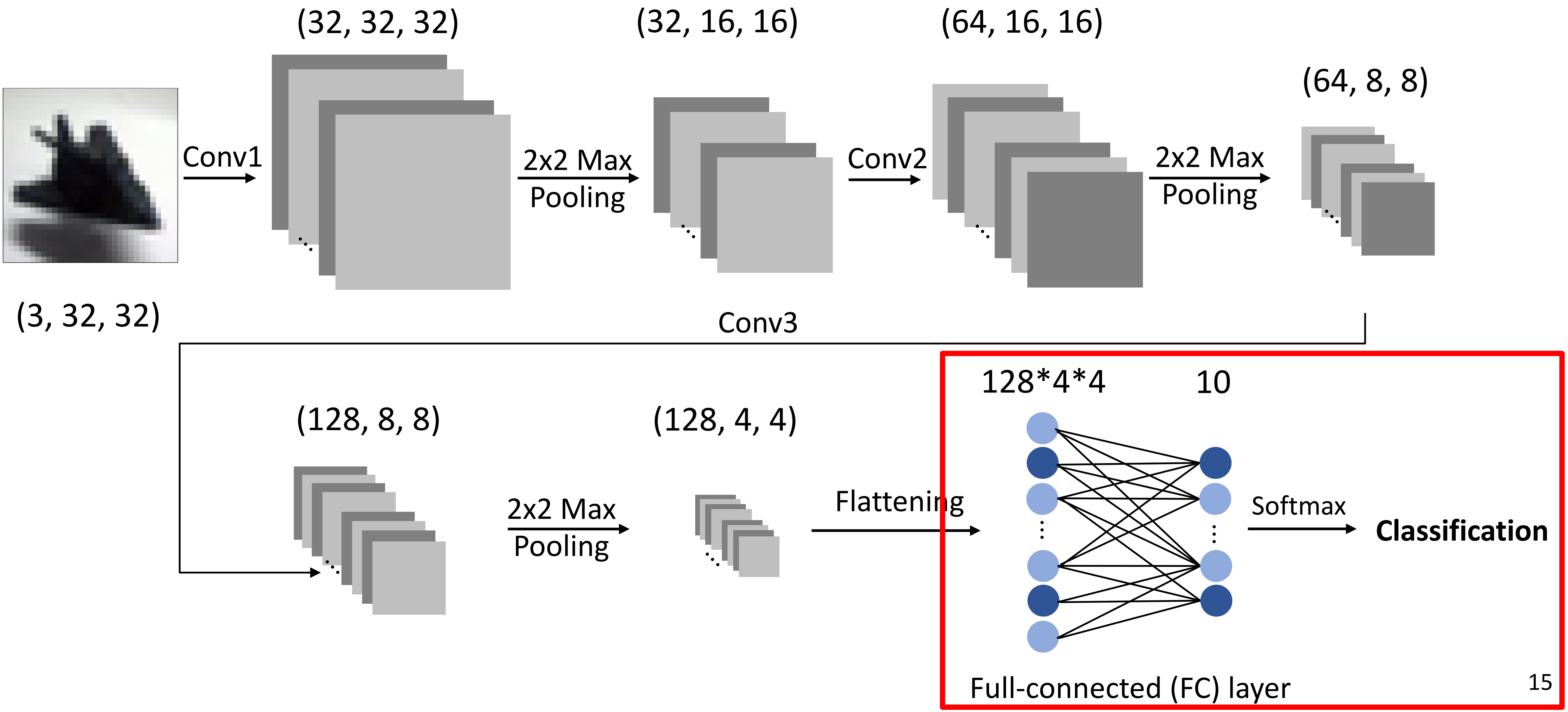
參數：
- kernel_size=2
- stride = 2

1	1
0.25	1

[Recap] Convolutional Neural Networks (CNN)

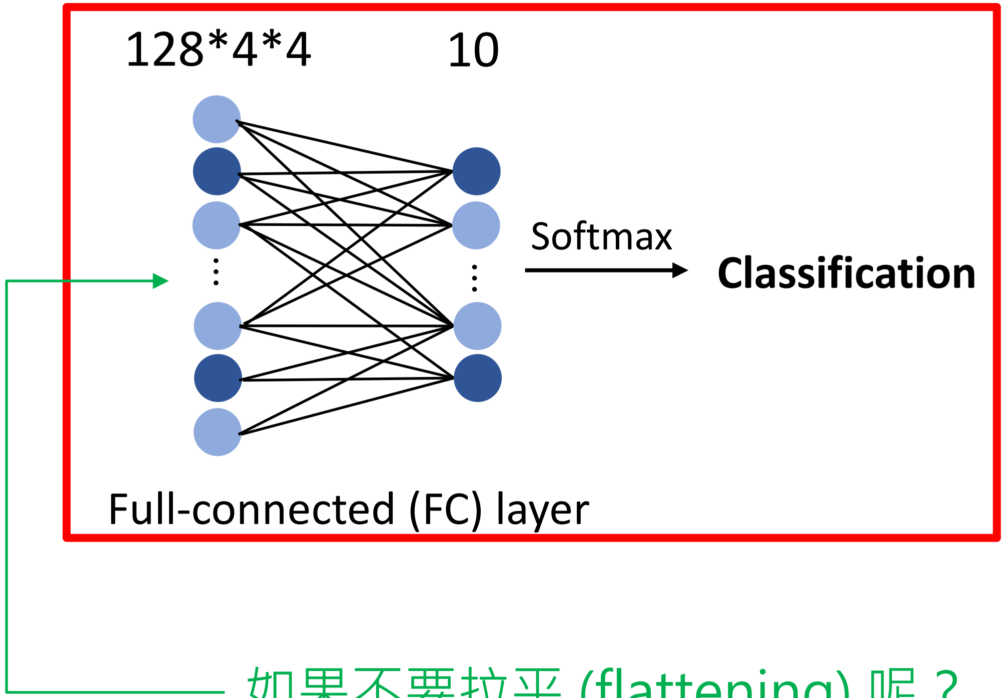
*維度意義：(C, H, W)

CGUDL_2025_Spring/code/pytorch_mnist.ipynb



[Recap] FC layer 參數量

FC: fully-connected



RGB images	參數量比較 (不算 bias 數)
FC layer	128* 4 * 4 * 10 = 20480

如果不要拉平 (flattening) 呢？

Global Average Pooling (GAP)

Feature Map

1	3	1	2
-1	1	1	0
0	1	1	0
0	0	1	2

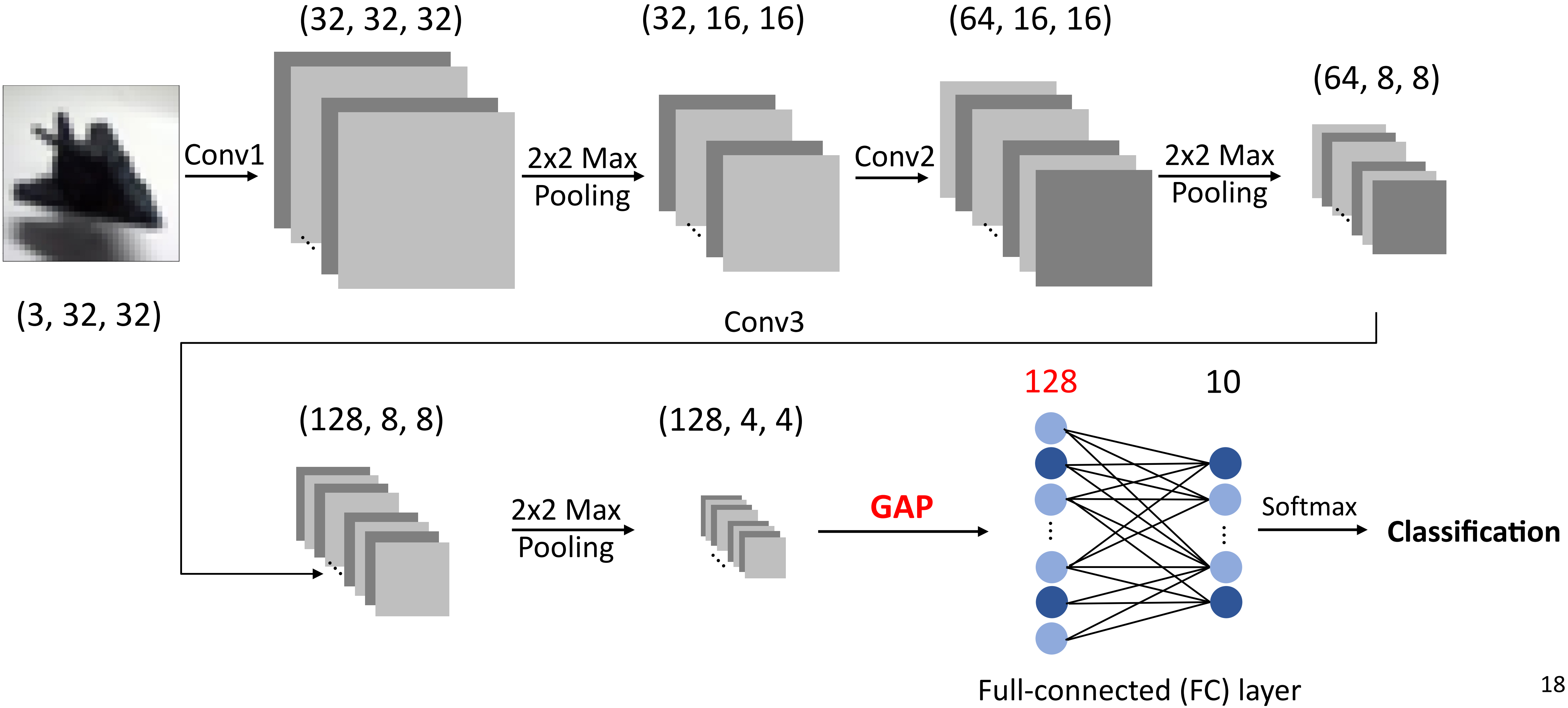
$\xrightarrow[\text{(13/16)}]{\text{全局數值取平均}}$ 0.8125

一張 feature map
經過GAP後變成一個數值

Lin, Min, Qiang Chen, and Shuicheng Yan. “Network in network.” ICLR 2014.

CNN with Global Average Pooling (GAP)

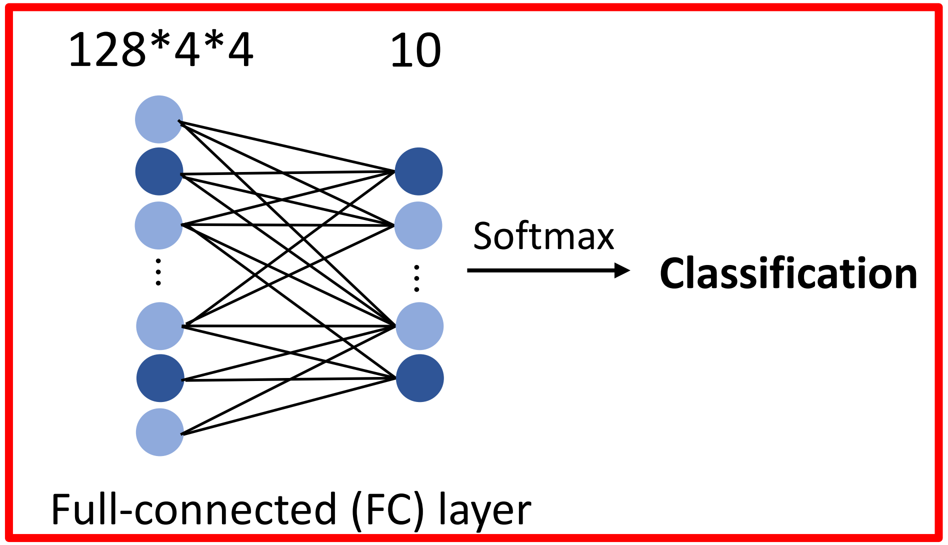
*維度意義：(C, H, W)



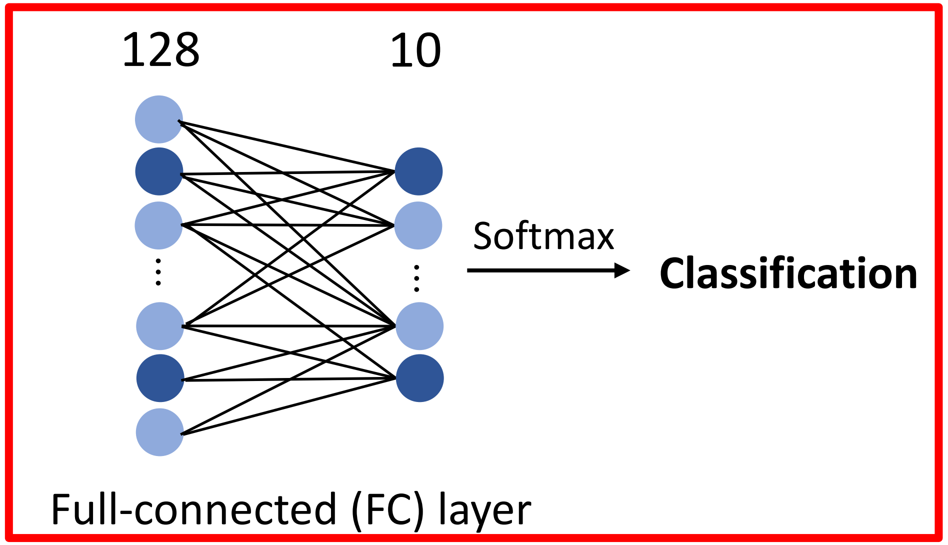
加入 GAP 後參數量下降

FC: fully-connected

Original version



GAP version



RGB images	參數量比較 (不算 bias 數)
Original	128* 4 * 4 * 10 = 20480
GAP	128 * 10 = 1280

加入 GAP 後 Testing Error 下降

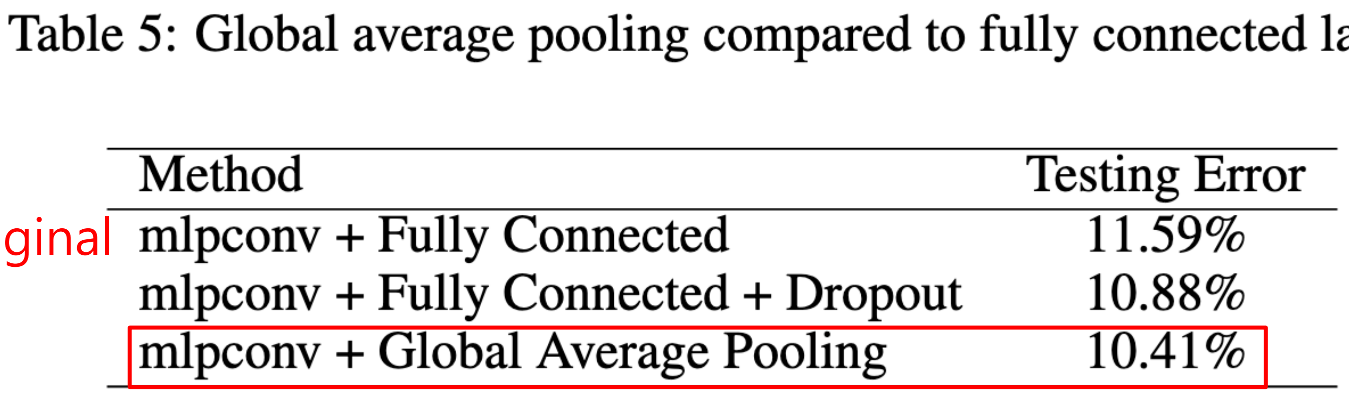


Table 5: Global average pooling compared to fully connected layer.

	Method	Testing Error
Original	mlpconv + Fully Connected	11.59%
	mlpconv + Fully Connected + Dropout	10.88%
	mlpconv + Global Average Pooling	10.41%

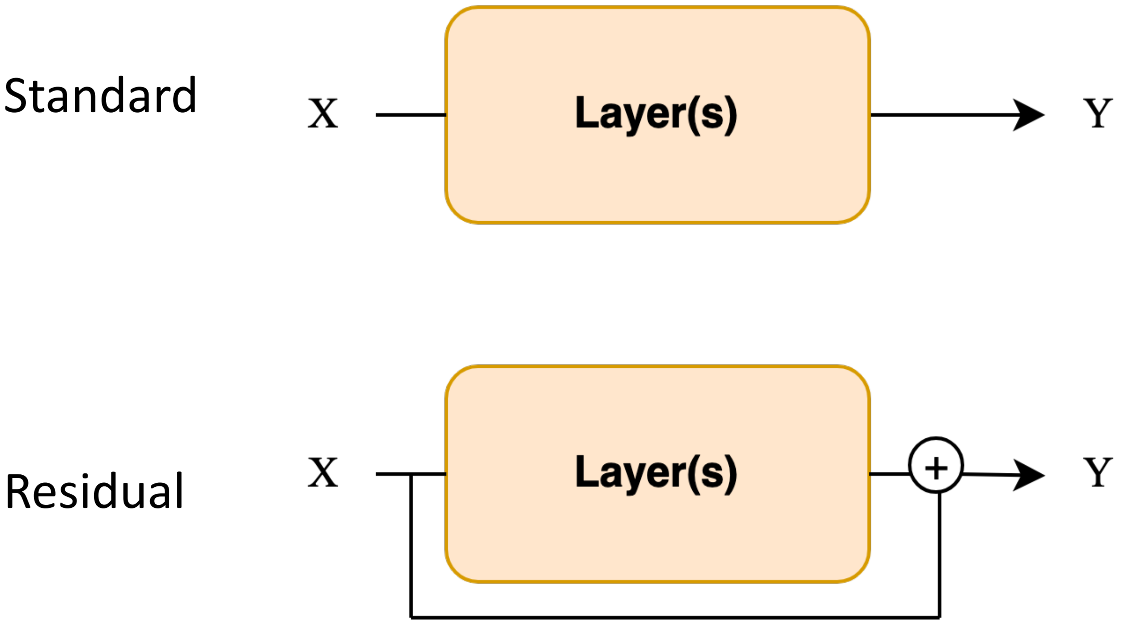
*Dropout 也是減少FC layer中node連接數量的方法

Lin, Min, Qiang Chen, and Shuicheng Yan. “Network in network.” ICLR 2014.

ResNet 架構

https://arxiv.org/abs/1512.03385

(Recap)



ResNet 的最後也是 GAP + FC

He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (2016).

Class-activation Map (CAM)

[Recap] Convolutional Neural Networks (CNN)

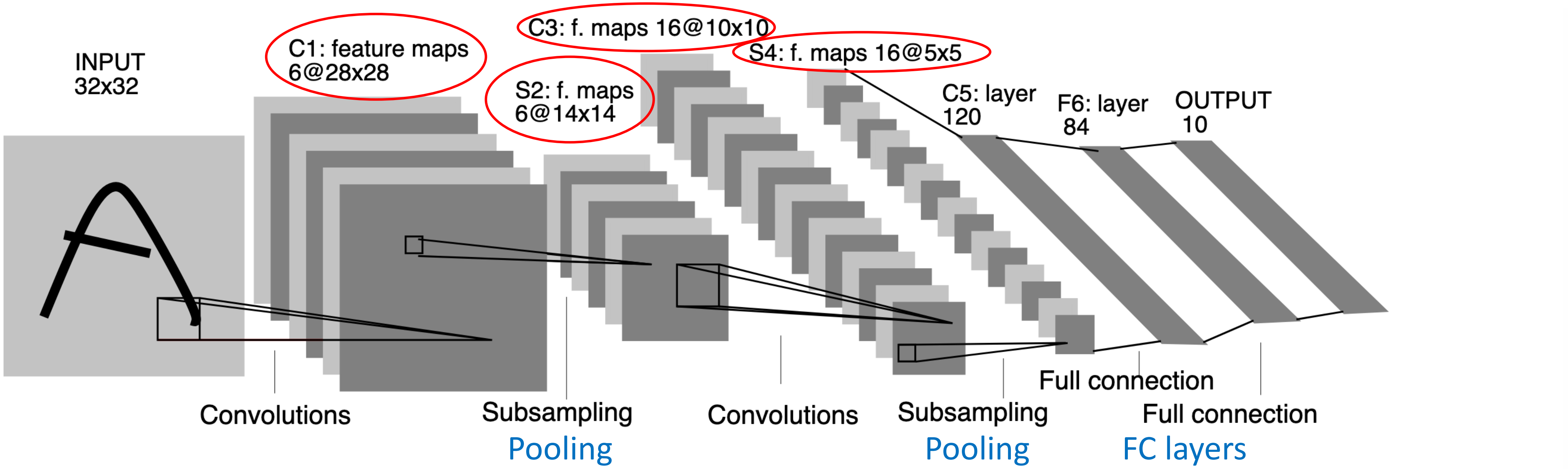
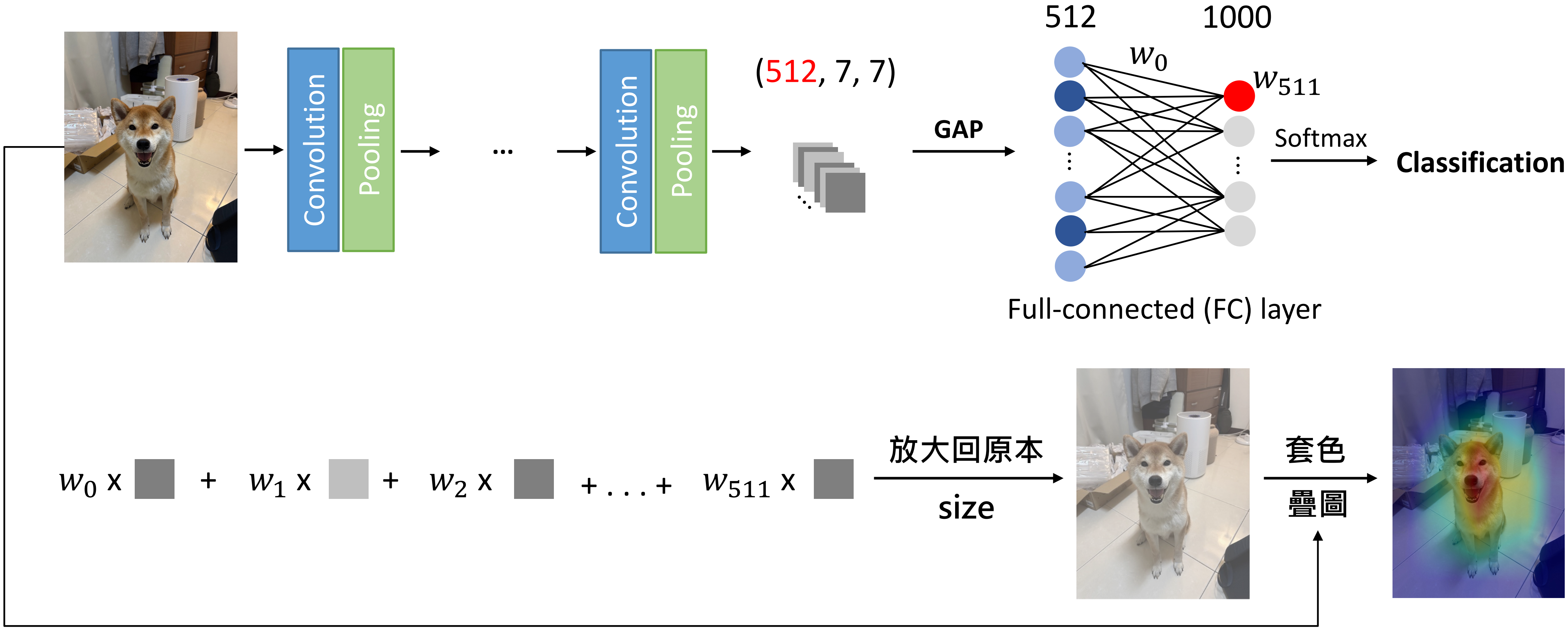


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Class-activation Map (CAM)

*假設使用 ImageNet pre-trained model



Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Why Global “Average” Pooling?

How about Global “**Summation**” Pooling?

How about Global “**Max**” Pooling (GMP)?

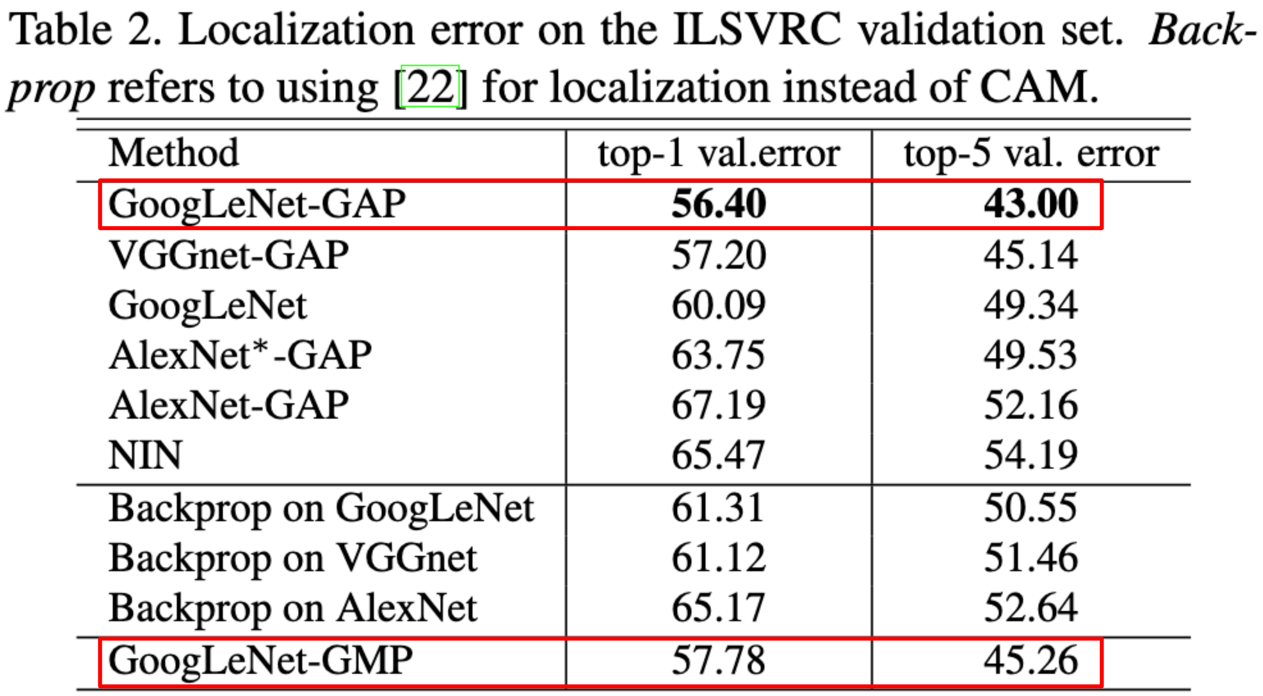
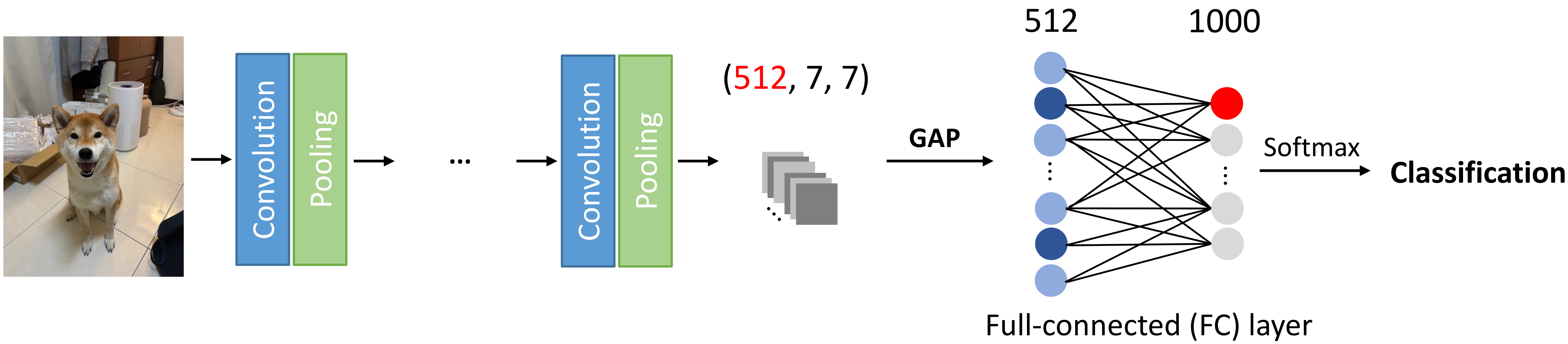


Table 2. Localization error on the ILSVRC validation set. *Backprop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

CAM 的問題

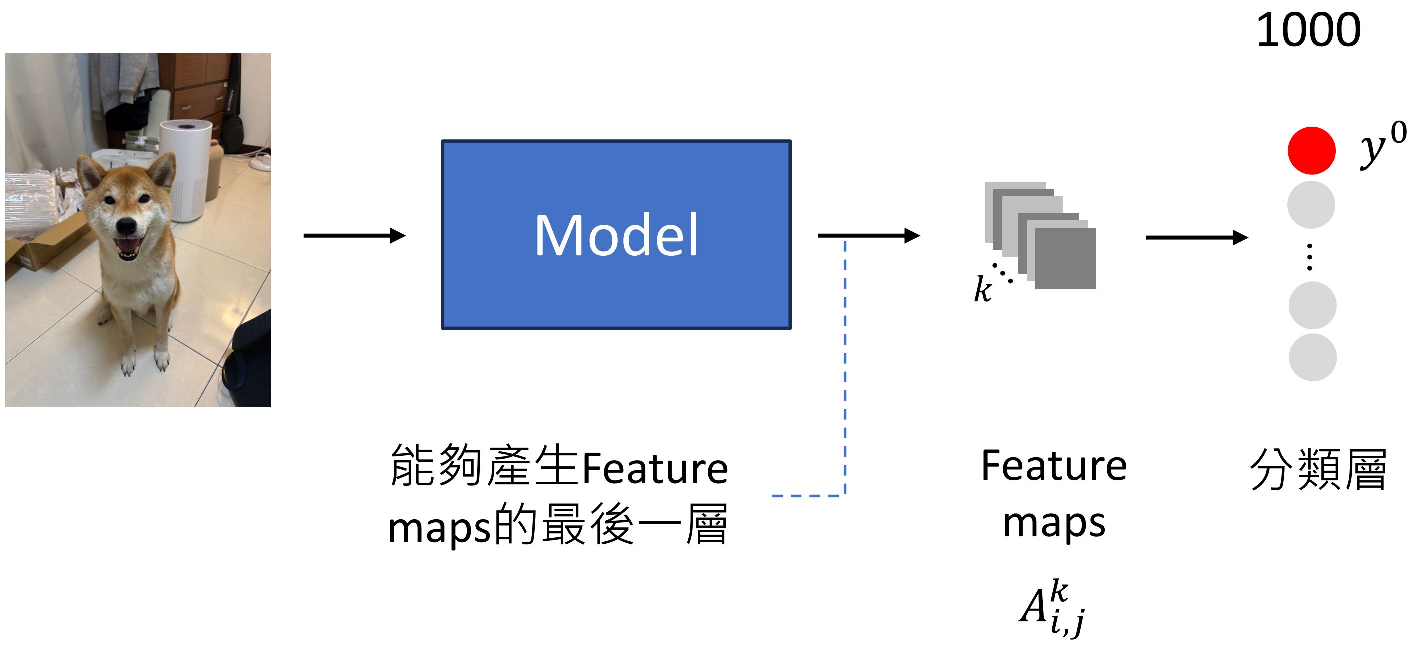


不是每個模型最後面都是 GAP + FC
(E.g., VGG-16 的最後是 3 層 FC、ViT 的最後只有 FC)

Grad-CAM (1)

Grad: gradients

i, j: feature map 中x軸與y軸位置
A: feature map
k: feature map 的數目
y^c: 對應到 class c 的 label ID



1000

y^0

Model

k

能夠產生Feature maps的最後一層

Feature maps

$A^k_{i,j}$

分類層

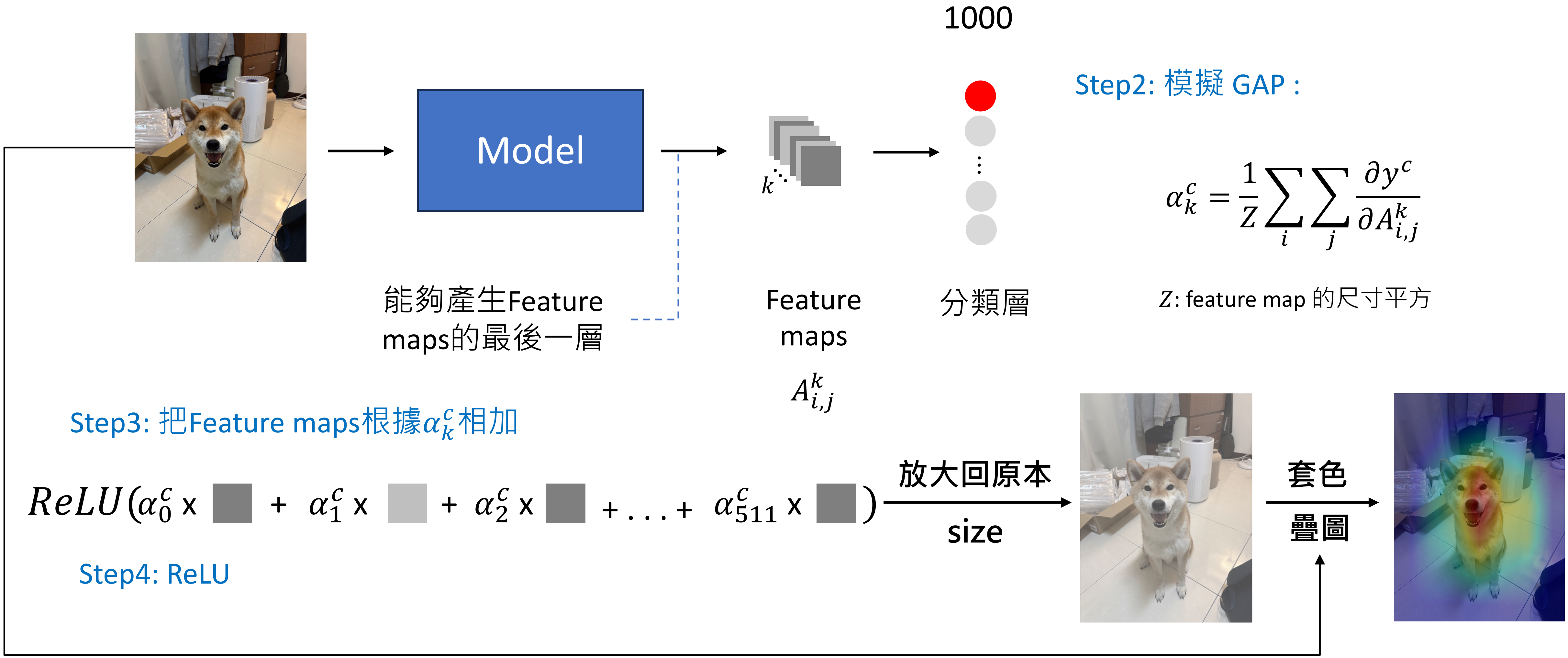
Step1: 計算特定 class (假設是y^0)

對於任一張 feature map 中每個

位置 ($A^k_{i,j}$) 的梯度 $\frac{\partial y^c}{\partial A^k_{i,j}}$

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

Grad-CAM (2)



1000

Model

能夠產生Feature maps的最後一層

Feature maps

$A_{i,j}^{k}$

分類層

Step2: 模擬 GAP :

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$$

Z: feature map 的尺寸平方

Step3: 把Feature maps根據α_k^c相加

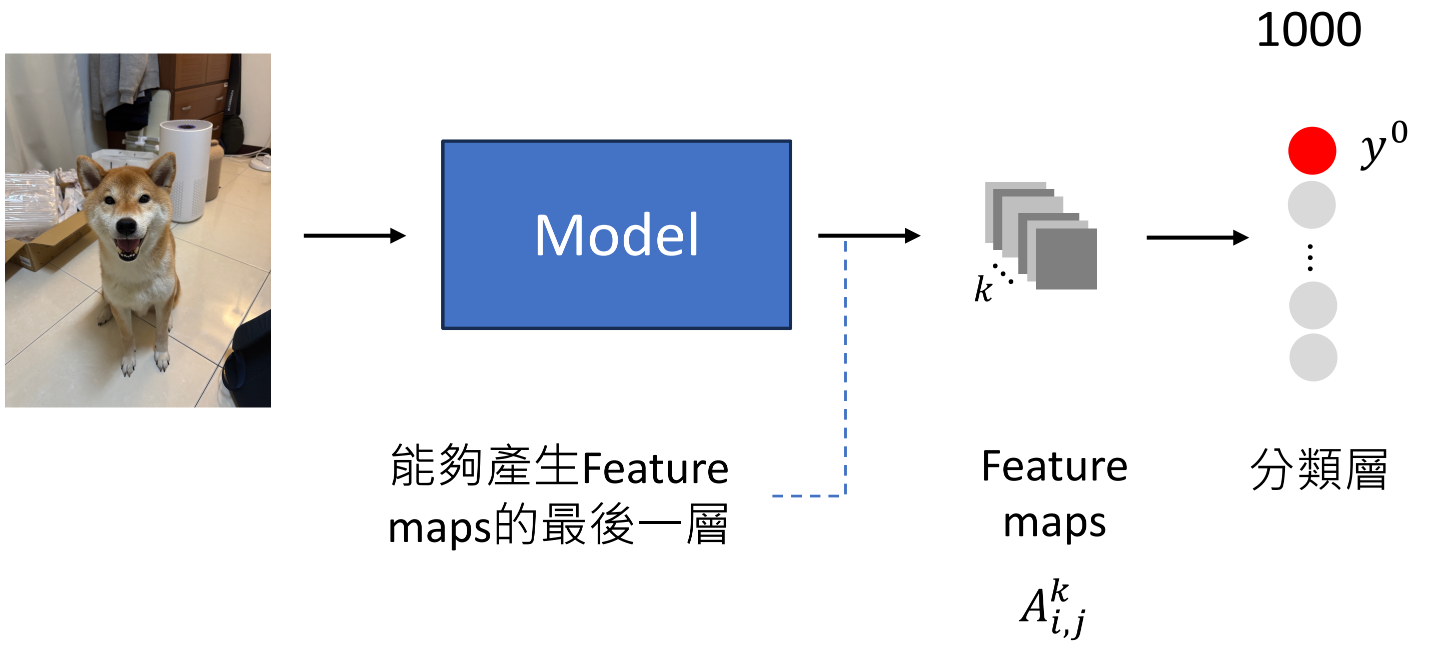
$$ReLU(\alpha_0^c \text{ x } ■ + \alpha_1^c \text{ x } ■ + \alpha_2^c \text{ x } ■ + \ldots + \alpha_{511}^c \text{ x } ■)$$

Step4: ReLU

Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

為什麼要使用梯度？

i, j: feature map 中x軸與y軸位置
A: feature map
k: feature map 的數目
y^c: 對應到 class c 的 label ID



Step1: 計算特定 class (假設是y^0)

對於任一張 feature map 中每個

位置 ($A_{i,j}^k$) 的梯度 $\frac{\partial y^c}{\partial A_{i,j}^k}$

- $\frac{\partial y^c}{\partial A_{i,j}^k}$ 代表 feature map 中任意位置的數值 ($A_{i,j}^k$) 對 y^c 的影響
- 如果一 feature map 有位置 $A_{i,j}^k$ 對 y^c 的影響很大 ($\frac{\partial y^c}{\partial A_{i,j}^k}$ 很大)， α_k^c 也會跟著被放大，代表該 feature map 可能對 y^c 特別重要

Automatic Evaluations

- Annotated datasets
  - ILSVRC (ImageNet Large Scale Visual Recognition Challenge) **Localization**
- Human Evaluation

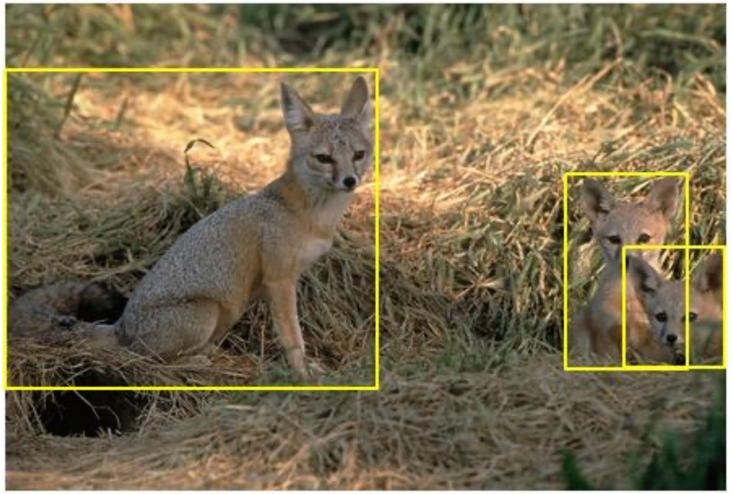
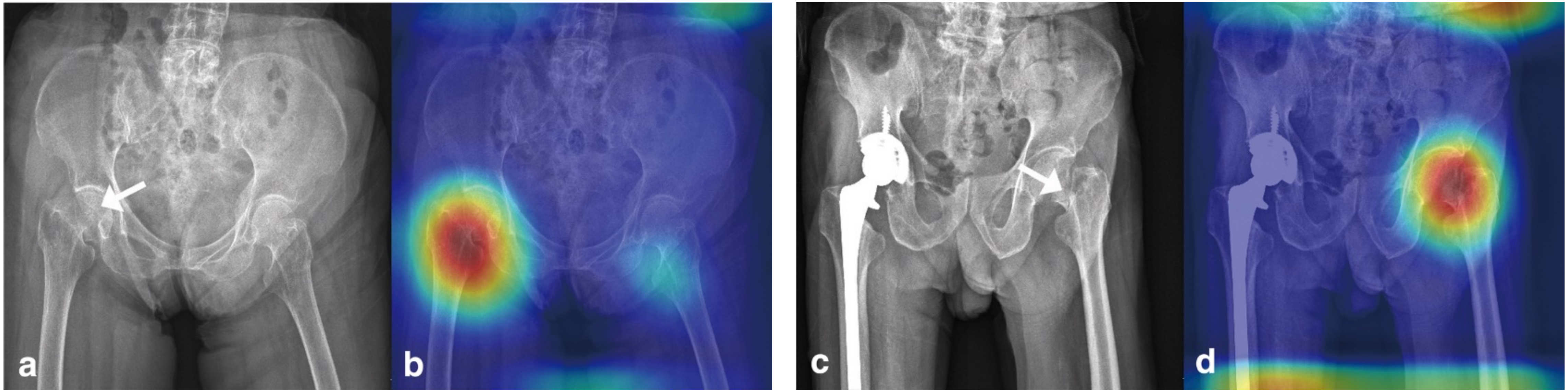


Figure source: https://www.kaggle.com/c/imagenet-object-localization-challenge/overview/description

Example: Medical Image Classification

- Pelvic X-ray fracture classification with Grad-CAM

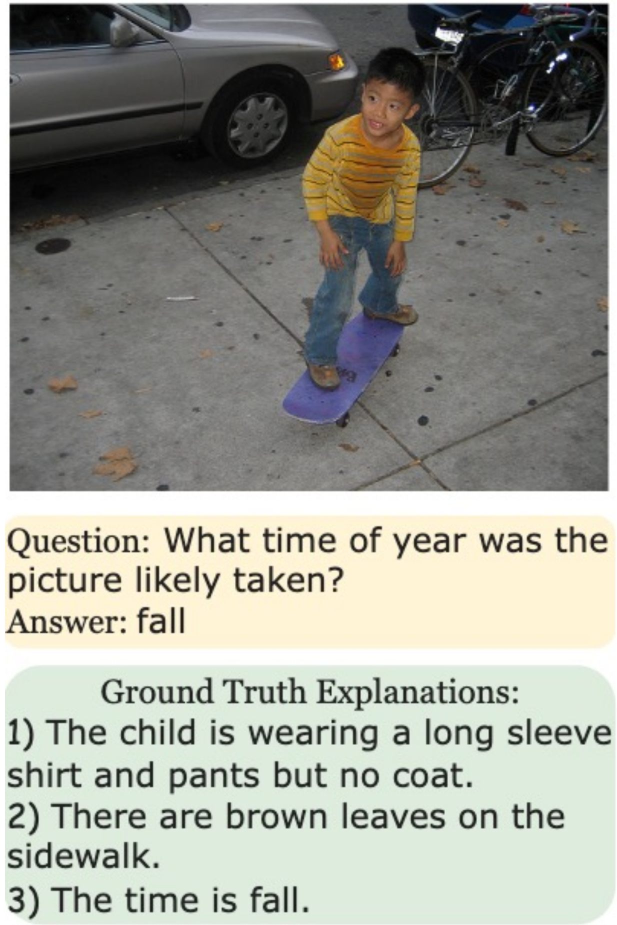


Cheng, Chi-Tung, et al. "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs." *European radiology* 29.10 (2019): 5469-5477.

Example: VQA (Visual Question Answering)

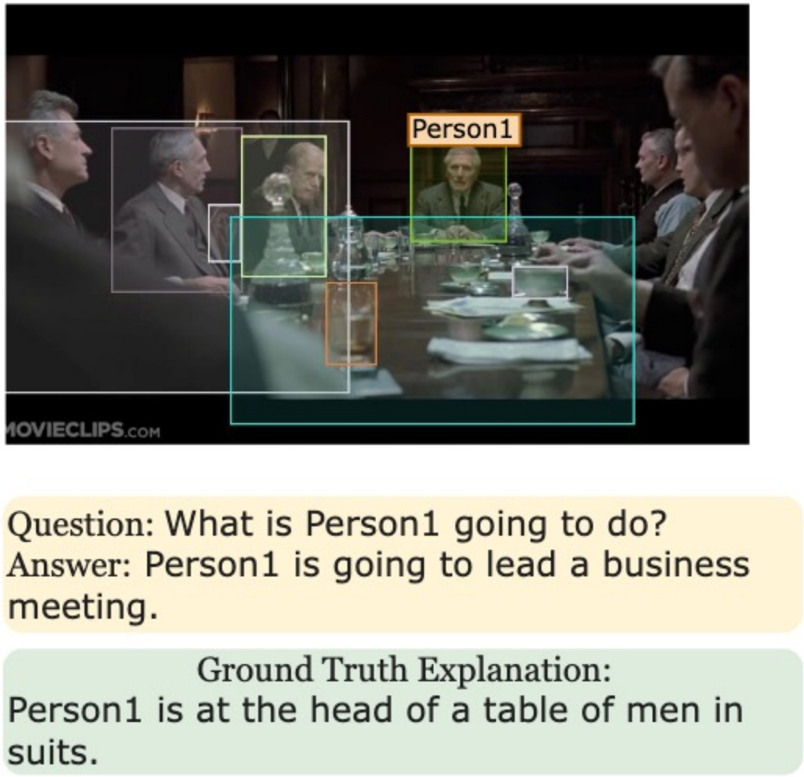
- VQA with explanations

Whitehouse, Chenxi, Tillman Weyde, and Pranava Swaroop Madhyastha. "Towards a Unified Model for Generating Answers and Explanations in Visual Question Answering." Findings of EACL 2023.



Question: What time of year was the picture likely taken?
Answer: fall

Ground Truth Explanations:
1) The child is wearing a long sleeve shirt and pants but no coat.
2) There are brown leaves on the sidewalk.
3) The time is fall.



Question: What is Person1 going to do?
Answer: Person1 is going to lead a business meeting.

Ground Truth Explanation:
Person1 is at the head of a table of men in suits.

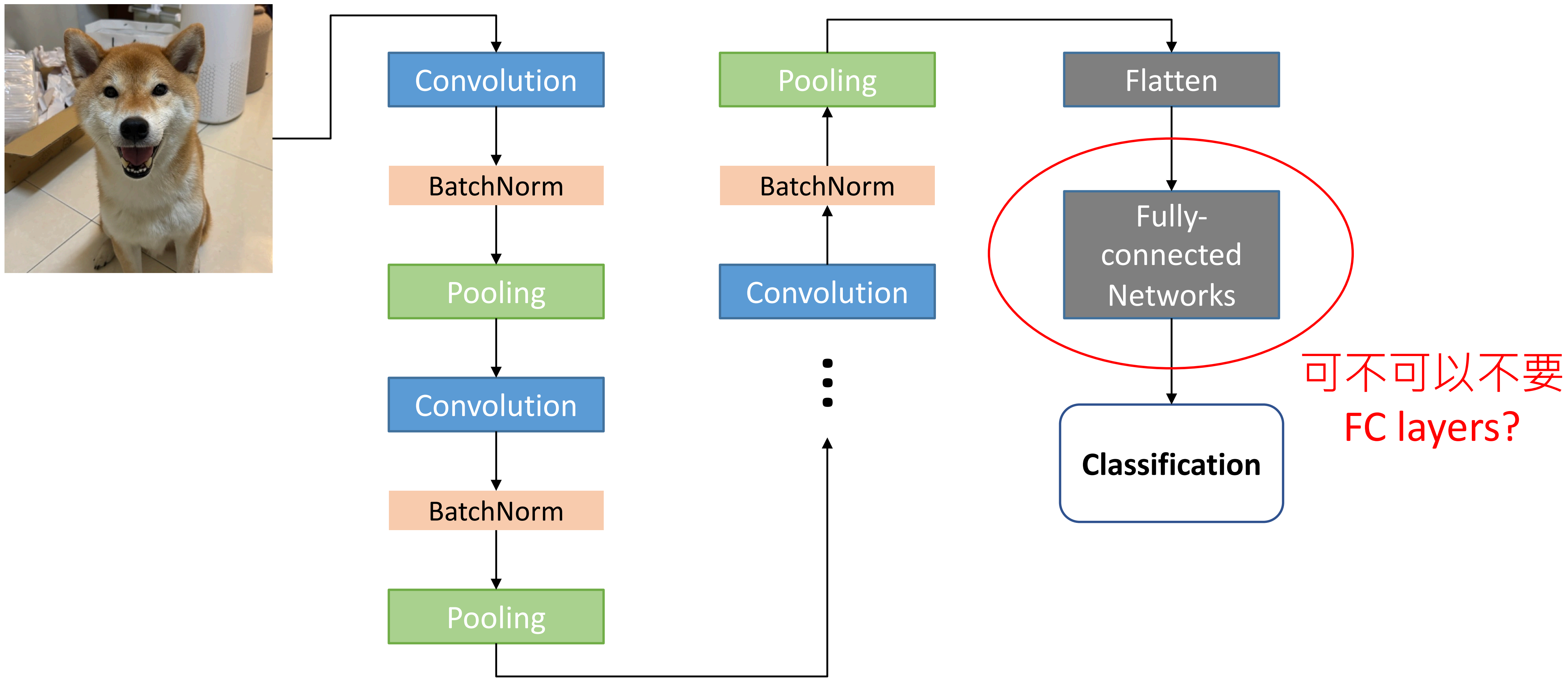
重要論文

- Network In Network: https://arxiv.org/abs/1312.4400
- Class-activation Map (CAM): https://arxiv.org/abs/1512.04150
- Grad-CAM: https://arxiv.org/abs/1610.02391
- SHAP: https://arxiv.org/abs/1705.07874
- Guided back-propagation: https://arxiv.org/pdf/1412.6806

延伸主題

可不可以完全不要FC LAYERS?

FC layers 會大量增加參數



可不可以不要 FC layers?

完全沒有採用 FC layers 的模型 (論文)

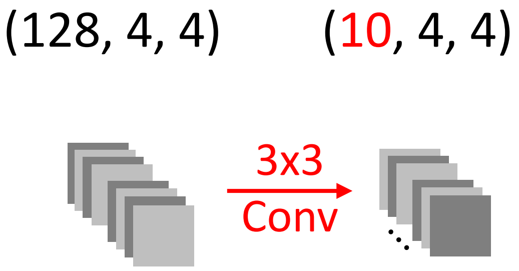
- Network In Network: https://arxiv.org/abs/1312.4400
- SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size (ICLR 2017): https://openreview.net/forum?id=S1xh5sYgx

Channels 數目調整

*維度意義：(C, H, W)

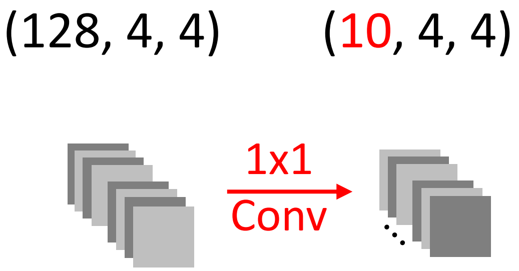
PyTorch 寫法

Filters 參數 (weights) 數量



```
torch.nn.Conv2d(
  in_channels=128,
  out_channels=10,
  kernel_size=3,
  padding=1,
)
```

128*10*3*3 = 11,520



1 x 1 convolution

```
torch.nn.Conv2d(
  in_channels=128,
  out_channels=10,
  kernel_size=1,
  padding=0,
)
```

128*10*1*1 = 1,280

Reference:
https://docs.pytorch.org/docs/stable/generated/torch.nn.Conv2d.html

1 x 1 Convolution

Stride = 1

1	1	1	1	0	0
0	1	1	0	1	0
0	0	1	1	0	0
0	0	1	1	1	0
0	0	0	1	1	0
0	0	0	0	0	0

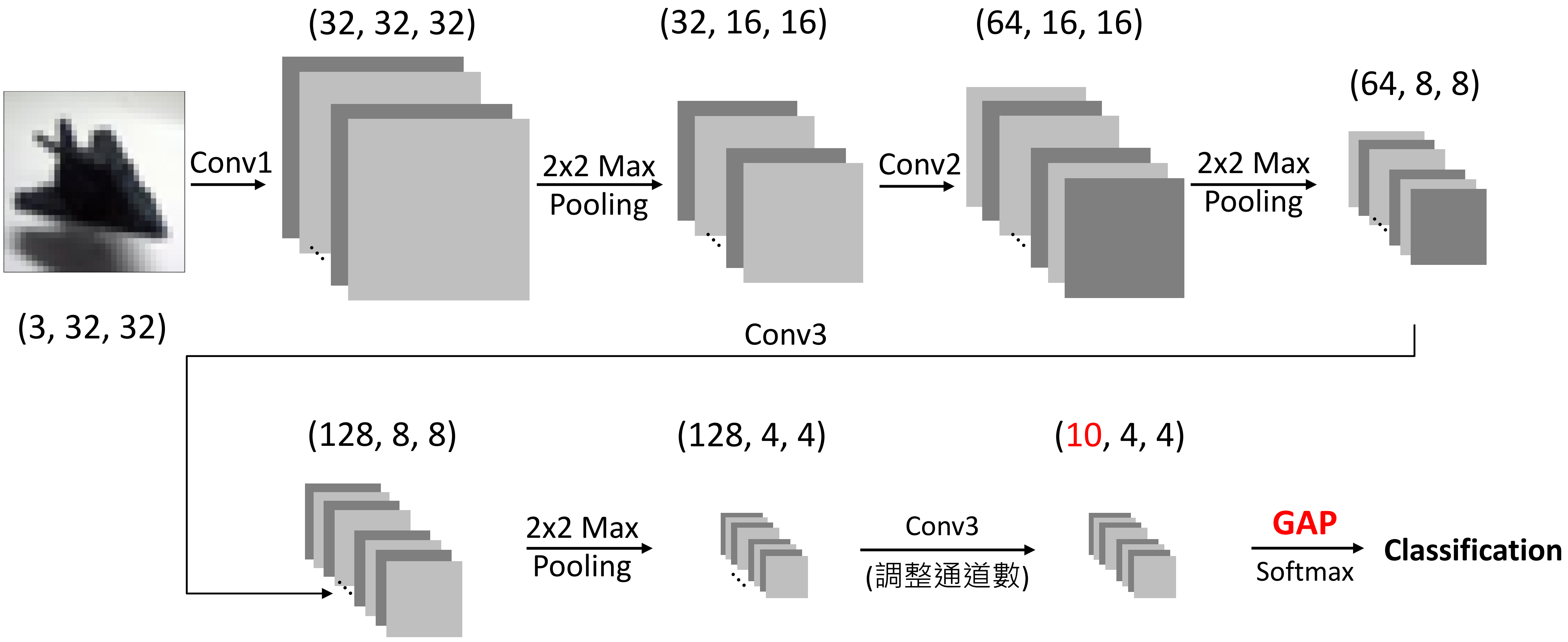
1

Filter
(假設數值為1)

1 1 1 1 0 0

CNN with Global Average Pooling (GAP)

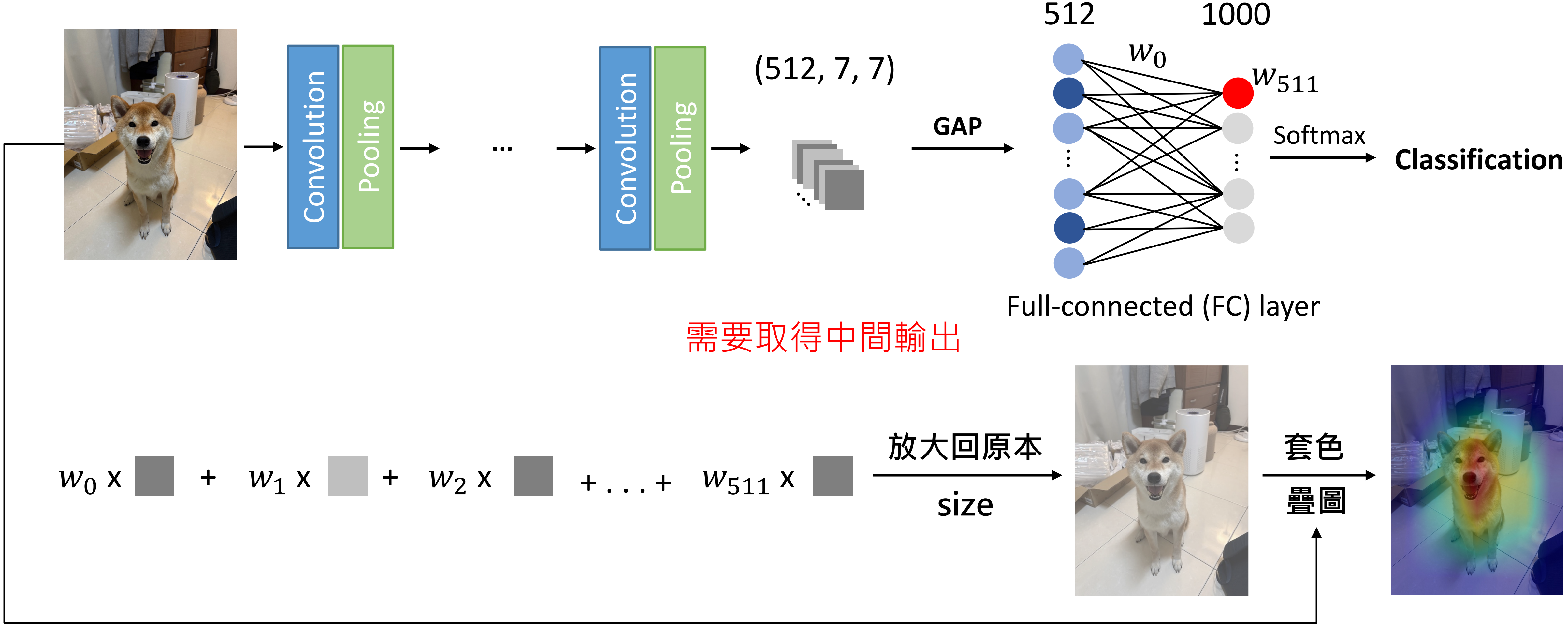
*維度意義：(C, H, W)



PyTorch

Class-activation Map (CAM)

*假設使用 ImageNet pre-trained model



Convolution → Pooling → … → Convolution → Pooling → (512, 7, 7) $\xrightarrow{\text{GAP}}$

512 1000

w_0 w_{511}

Softmax → Classification

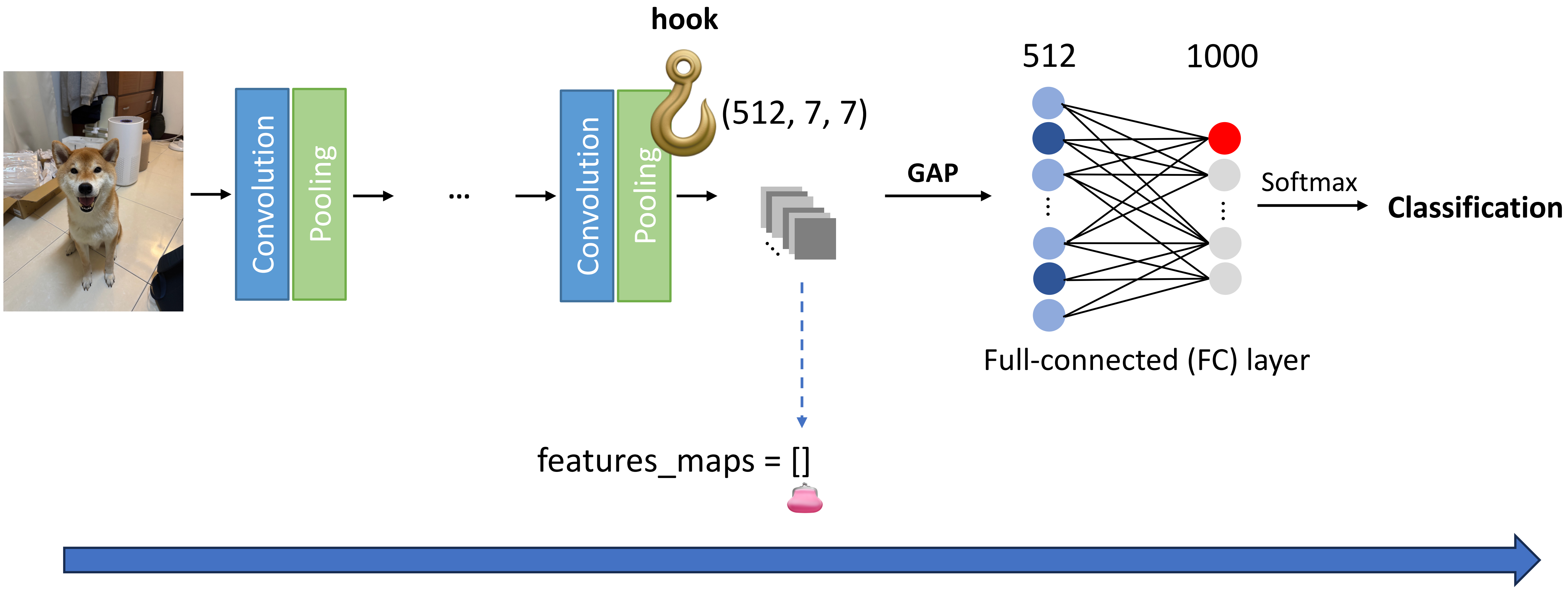
Full-connected (FC) layer

需要取得中間輸出

w_0 x ■ + w_1 x ■ + w_2 x ■ + . . . + w_{511} x ■ $\xrightarrow[\text{size}]{\text{放大回原本}}$ $\xrightarrow[\text{疊圖}]{\text{套色}}$

Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

register_forward_hook



https://docs.pytorch.org/docs/stable/generated/torch.nn.Module.html#torch.nn.Module.register_forward_hook

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 林君襄

✉ becky890926@gmail.com