# Homework 5 – Convolution Accelerator

Handout: 2024/12/03

Due: 3 weeks later

1. (VGG-16) Fig. 1 shows the VGG-16 Convolutional Neural Network (CNN) model with a total of 16 layers, including 13 convolutional layers and 3 fully connected layers. The final softmax generates the probability of every class. Table 1 shows the information of the first two layers.

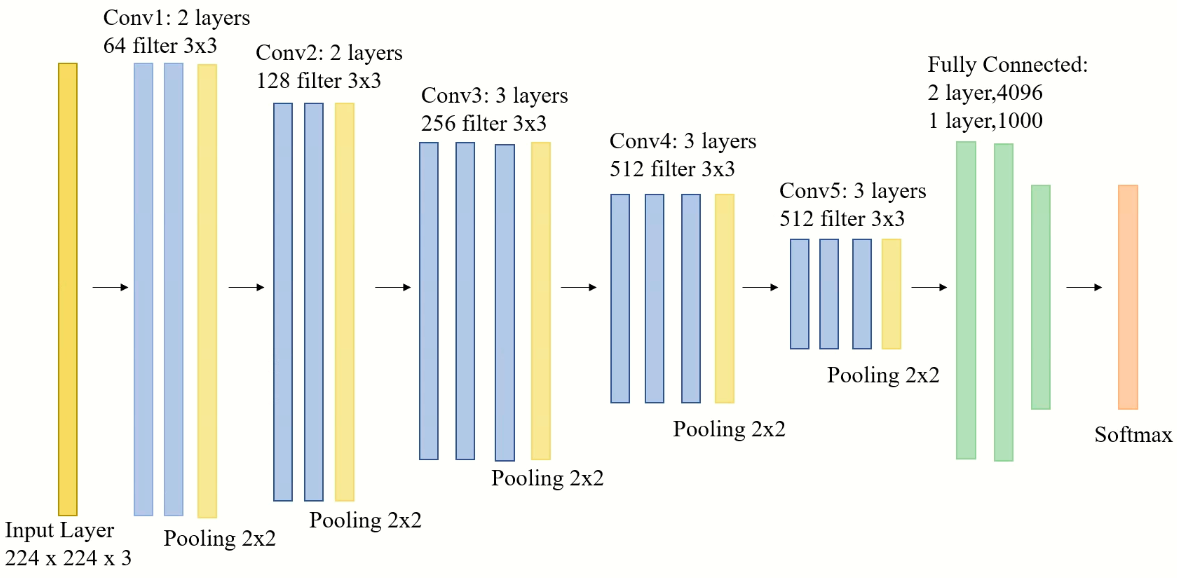


Fig. 1. VGG-16 CNN model.

Table 1. Information for the first two layers in VGG-16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # of input channels | # of output channels | Input feature map size (before padding) | Filter Kernel size | Stride |
| 3 | 64 | 224x224 | 3x3 | 1 |
| 64 | 64 | 224x224 | 3x3 | 1 |

1. (DNN Architecture) Design a deep neural network (DNN) hardware accelerator similar to that shown in Fig. 2 to speed up the convolution operations in the first two VGG-16 layers. You can determine the ICP and OCP for your own design. Calculate the total number of cycles and amount of memory accesses for exexecuting the first layer and the second layer. Note that the execution cycles and memory access depend on ICP and OCP

一張含有 圖表, 螢幕擷取畫面, 文字, 正方形 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 圖表, Rectangle 的圖片

自動產生的描述

1. (b)

Fig. 2. (a) Overall architecture of the DNN accelerator (ICP=4, KWP=9, OCP=4). (b) Architecture of PE.

1. (Padding) First, pad the input images as shown in Fig. 3 to obtain the output feature maps of the same size after 3x3 filter kernel convolution. Note that after padding, the featuremap size is 226x226.

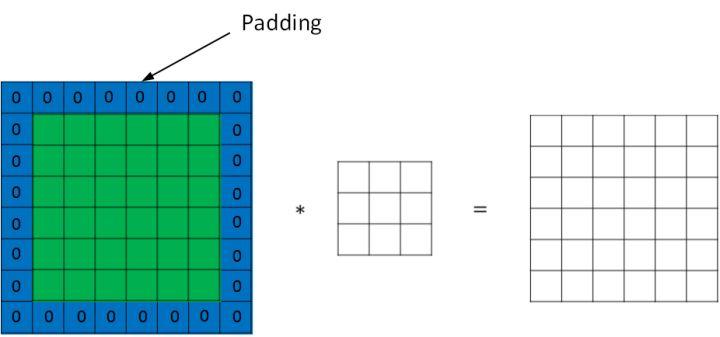


Fig. 3. Padding.

1. (Line Buffer) Use the line buffer shown in Fig. 4 to reduce repeated accesses of the same data.

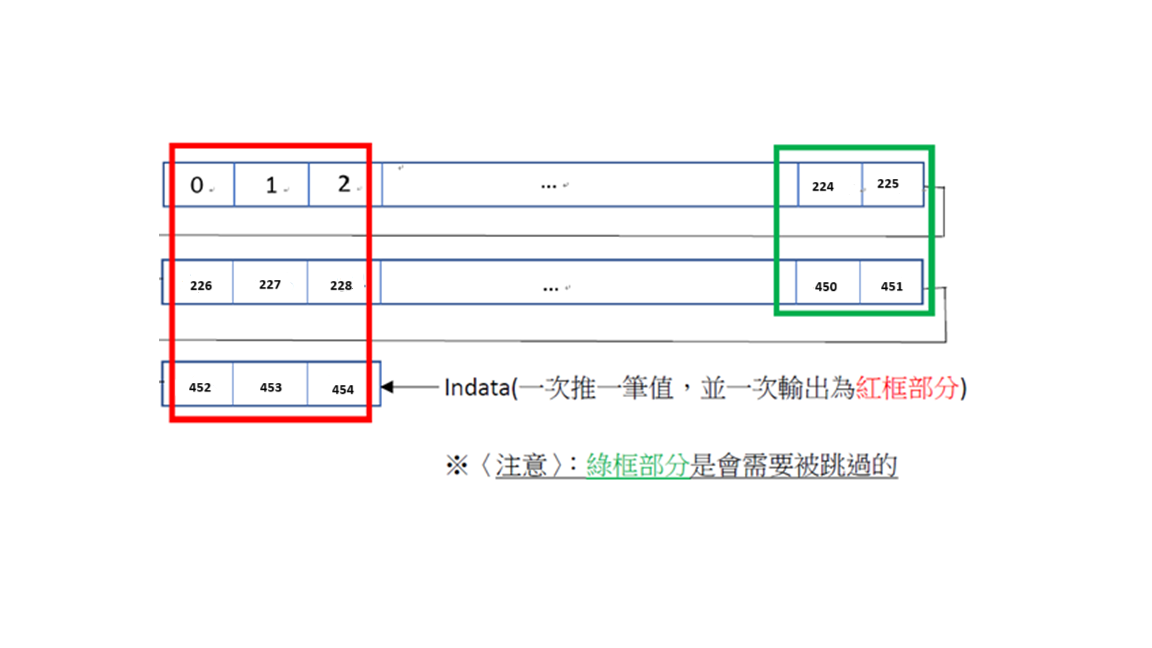
****

Fig. 4. Line buffer and the operation.

1. (ReLU) Non-linear activation function Rectified Linear Unit (ReLU) shown in Fig. 5 is used after convolution.

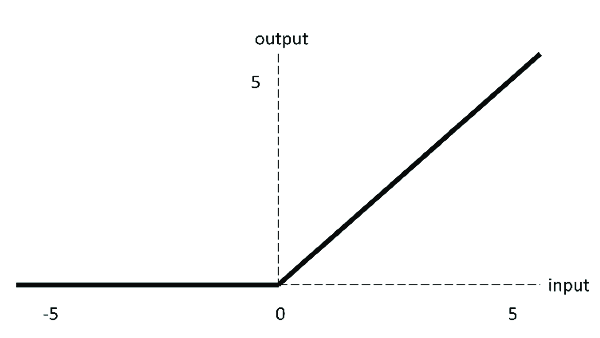
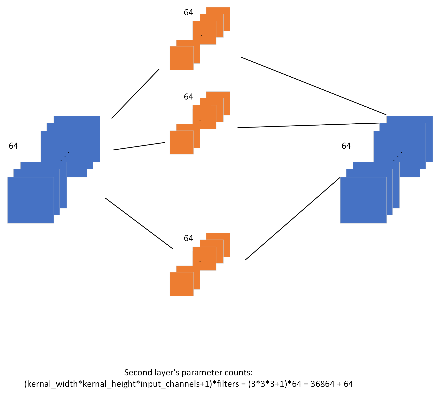
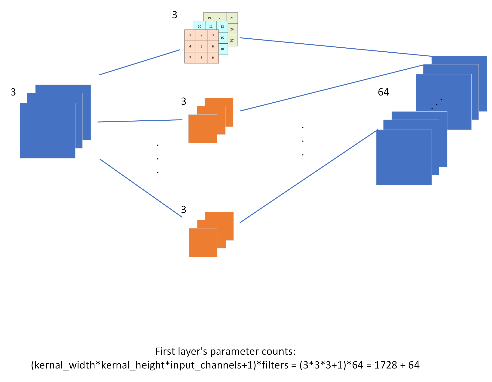


Fig. 5. ReLU.

1. (Provided Data) The following data (.txt), including initial input image data, the weights, and biases of the first two layers, are provided. The data structure is shown in Fig. 6.
2. input image(cat224.bmp)
3. filter weights (conv1\_kernel\_hex.txt, conv2\_kernel\_hex.txt)
4. biases (conv1\_bias\_hex.txt, conv2\_bias\_hex,txt)

一張含有 文字, 螢幕擷取畫面, 正方形 的圖片

自動產生的描述

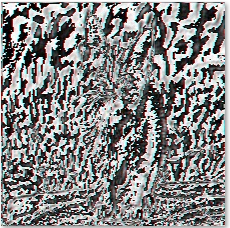
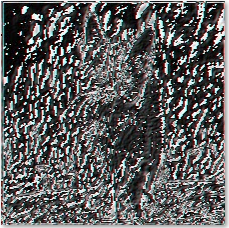
1. (b) (c)

Fig. 6. (a)order of the provided weights and biases in the text files.

(b)First layer of VGG-16. (c)Second layer of VGG-16.

1. (Featuremaps) Fig. 7 shows the initial input image and the first output featuremaps in the 1st and 2nd layers. Fig. 8 and Fig. 9 respectively show all the output featuremaps in the 1st and 2nd layers.

一張含有 室外, 草, 哺乳類, 坐 的圖片

自動產生的描述  

1. (b) (c)

Fig. 7. (a) Input image. (b) 1st output featuremap of the 1st layer. (c) 1st output featuremap of the 2nd layer.



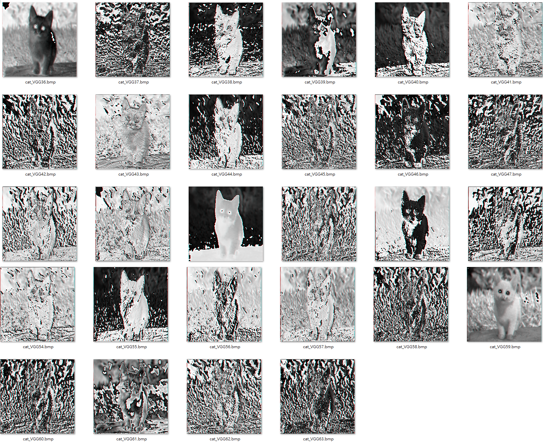
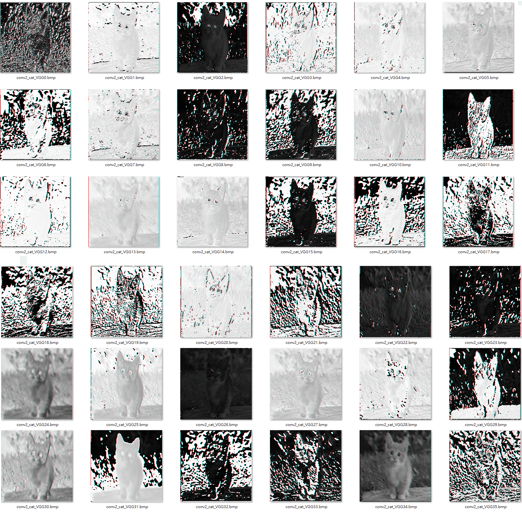


Fig. 8. Output featuremaps of the 1st layer.



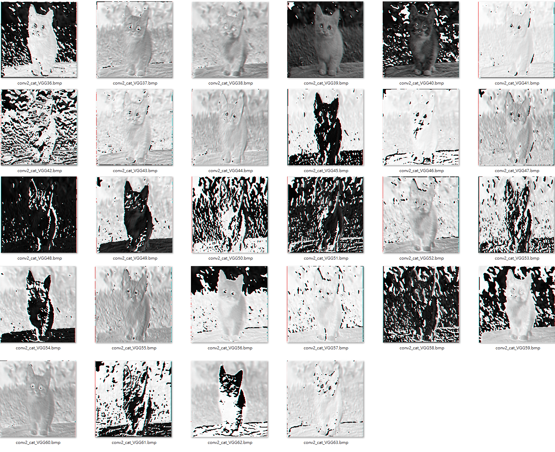


Fig. 9. Output featuremaps of the 2nd layer.

1. Report Requirement

檔案類須含有:

1. Testbench (20%)
2. Verilog RTL code & Gate-level code

-LineBuffer (5%)

-PE (5%)

-AdderTree (5%)

-ReLU (5%)

1. Image  
   -Conv1 image \*64(15%)

-Conv2 image \*64(15%)

word報告類須含有:

1. 硬體架構圖解釋(10%)
2. Area資訊和critical path資訊，不需要做optimization，合成出來即可(10%)
3. 心得(10%)

以上打包成HDL\_HW5\_MXXXXXXXXX.zip壓縮檔並繳交