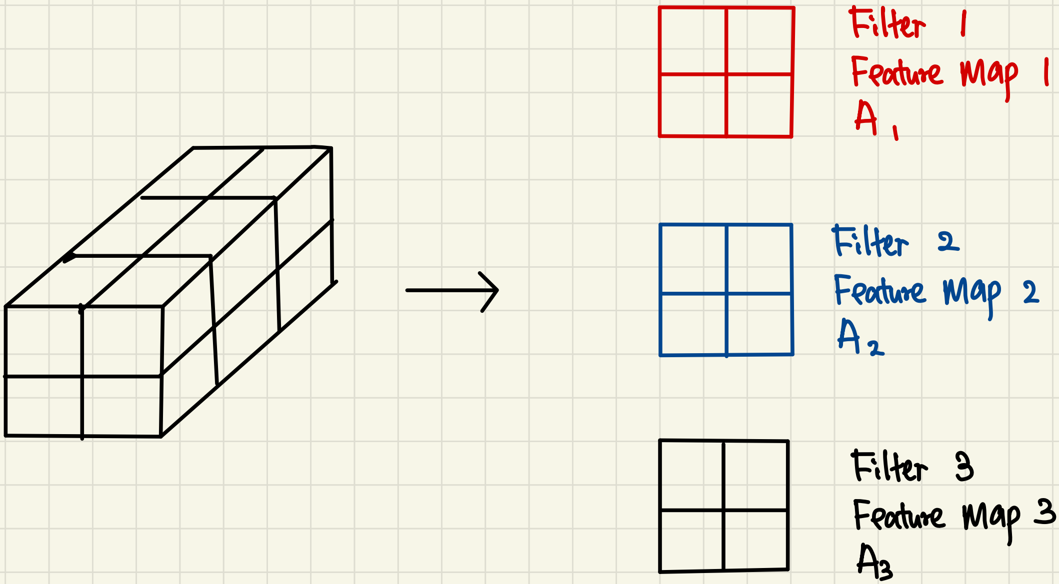
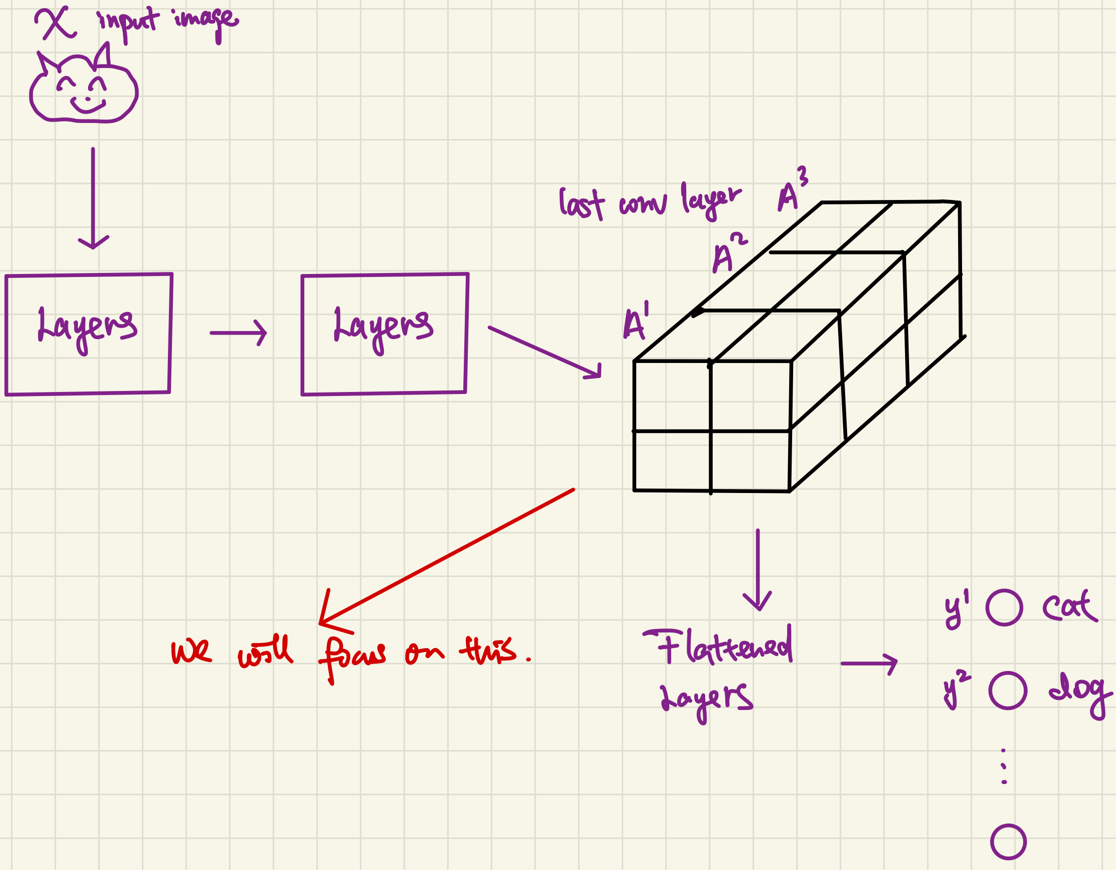
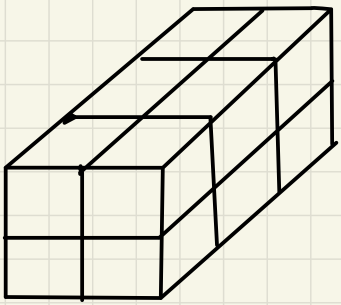

Filter 1
Feature Map 1
A_1
Filter 2
Feature Map 2
A_2
Filter 3
Feature Map 3
A_3

Typical CNN Architecture



We will focus on this.

Feature Maps. (Last Conv Layer)



$$\begin{bmatrix} 1 & -1 \\ 2 & 0.1 \end{bmatrix}$$

Filter 1
Feature Map 1
A_1

$$\begin{bmatrix} 1.5 & 0.1 \\ 2.5 & 0.5 \end{bmatrix}$$

Filter 2
Feature Map 2
A_2

$$\begin{bmatrix} -2 & 3 \\ 0 & 2 \end{bmatrix}$$

Filter 3
Feature Map 3
A_3

A^1, A^2, A^3 are feature maps of a CAT.
By design, values in A^3 are "more different" than those in A^1 & A^2. This is to highlight that A^3 is the "background kernel" and it is quite irrelevant to whether a cat is a cat.
We will certainly hope our Grad-CAM does not focus on this.
Our α_3^c should reflect this later.

If input image is an elephant on grass!

For eg. this feature map A^3 is designed to focus on the grasses of the elephant image & thus is not "important".

Gradient MAP $\frac{dy^c}{dA^k}$

Differentiate y^c wrt A^k_{ij}

$$\left\{ \frac{dy^c}{dA_1}, \frac{dy^c}{dA_2}, \frac{dy^c}{dA_3} \right\}$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 2 \end{bmatrix} \quad \frac{dy^c}{dA_1}$$

$$\begin{bmatrix} 3 & 5 \\ 6 & 3 \end{bmatrix} \quad \frac{dy^c}{dA_2}$$

$$\begin{bmatrix} 0.1 & -0.1 \\ 0.2 & 0.2 \end{bmatrix} \quad \frac{dy^c}{dA_3}$$

Computing $\frac{dy^c}{dA^k}$ helps us understand how feature map affects the class of interest. In other words, we see that $\frac{dy^c}{dA^1} = \begin{bmatrix} 2 & 3 \\ 4 & 2 \end{bmatrix}$ has these values. Then we can intuitively understand the value 2 means a unit change in pixel $A^{11} = 1$ will cost 2 unit change to y^c.

We perform GAP to get hold of the rate of change of individual feature map A^k wrt y^c.

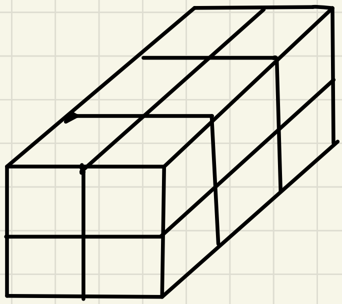
GAP

$$\xrightarrow{GAP} \left\{ \frac{2+3+4+2}{4} = 2.75 \right\} \rightarrow \alpha_1^c$$

$$\xrightarrow{GAP} \left\{ \frac{3+5+6+3}{4} = 4.25 \right\} \quad \alpha_2^c$$

$$\xrightarrow{GAP} \left\{ \frac{0.1-0.1+0.2+0.2}{4} = 0.1 \right\} \quad \alpha_3^c$$

$$\left\{ \alpha_1^c \quad \alpha_2^c \quad \alpha_3^c \right\}$$



Feature Maps $\{A_1, A_2, A_3\} \longrightarrow$

1	-1
2	0.1

Filter 1
Feature Map 1
A_1

1.5	0.1
2.5	0.5

Filter 2
Feature Map 2
A_2

-2	3
0	2

Filter 3
Feature Map 3
A_3

Differentiate y^c wrt A^k_{ij}

$$\left\{ \frac{dy^c}{dA_1}, \frac{dy^c}{dA_2}, \frac{dy^c}{dA_3} \right\}$$

2	3
4	2

$\frac{dy^c}{dA_1}$

3	5
6	3

$\frac{dy^c}{dA_2}$

0.1	-0.1
0.2	0.2

$\frac{dy^c}{dA_3}$

GAP

$\xrightarrow{GAP} \left\{ \frac{2+3+4+2}{4} = 2.75 \right\}$

α_1^c

α_2^c

$\xrightarrow{GAP} \left\{ \frac{3+5+6+3}{4} = 4.25 \right\}$

α_3^c

$\xrightarrow{GAP} \left\{ \frac{0.1-0.1+0.2+0.2}{4} = 0.1 \right\}$

$$\{ \alpha_1^c \quad \alpha_2^c \quad \alpha_3^c \}$$

Weighted Localization MAP.

$$\begin{bmatrix}1&-1\\2&0.1\end{bmatrix} \times \underset{2.75}{\alpha_1^c} \qquad \begin{bmatrix}2.75&-2.75\\5.5&0.275\end{bmatrix}$$

$$\begin{bmatrix}1.5&0.1\\2.5&0.5\end{bmatrix} \times \underset{4.25}{\alpha_2^c} \qquad \begin{bmatrix}6.375&0.425\\10.625&2.125\end{bmatrix}$$

$$\begin{bmatrix}-2&3\\0&2\end{bmatrix} \times \underset{0.1}{\alpha_3^c} = \begin{bmatrix}-0.2&0.3\\0&0.2\end{bmatrix}$$

Weighted sum.

$$\begin{bmatrix}8.925&-2.025\\16.125&2.6\end{bmatrix}$$

Weighted Sum of All Feature maps

$$L = \alpha_1^c A_1 + \alpha_2^c A_2 + \alpha_3^c A_3 = \begin{bmatrix}8.925&-2.025\\16.125&2.6\end{bmatrix}$$

Notice that $\alpha_3^c A_3$ has very small values and hence contribute lesser to how our CNN looks at y^c.

Lasty apply ReLU to L: $\underbrace{ReLU(L)}_{L\text{-grad-cam}} = \begin{bmatrix}8.925&0\\16.125&2.6\end{bmatrix}$

This has an intuitive meaning, negative values may indicate regions not related to y^c.

We overlay $L_{gradcam}$ to the original Image.