學號：T08902109 系級：資工三　姓名：賈成銪

(1%) 請說明這次使用的 model 架構，包含各層維度及連接方式。

Model 架構：整體采用 Resnet18 的 pre-train model,在最後一層增加了 dropout。具體架構如下(BatchNorm 和 Relu 未影響 output_size，故沒有顯示)：

Layer1:

Input --> 48*48*3

Conv2d --> 24*24*64

MaxPool2d --> 12*12*64

Layer2:

Conv2d --> 12*12*64

Conv2d --> 12*12*64

Conv2d --> 12*12*64

Layer3:

Conv2d　--> 6*6*128

Conv2d　--> 6*6*128

Conv2d　--> 6*6*128

Conv2d　--> 6*6*128

Layer4:

Conv2d 3*3*256

Conv2d 3*3*256

Layer5:

Conv2d: 2*2*512

Conv2d: 2*2*512

Conv2d: 2*2*512

Layer6:

AdaptiveAvgPool2d --> 1*1*512

)

Layer7:(fn)

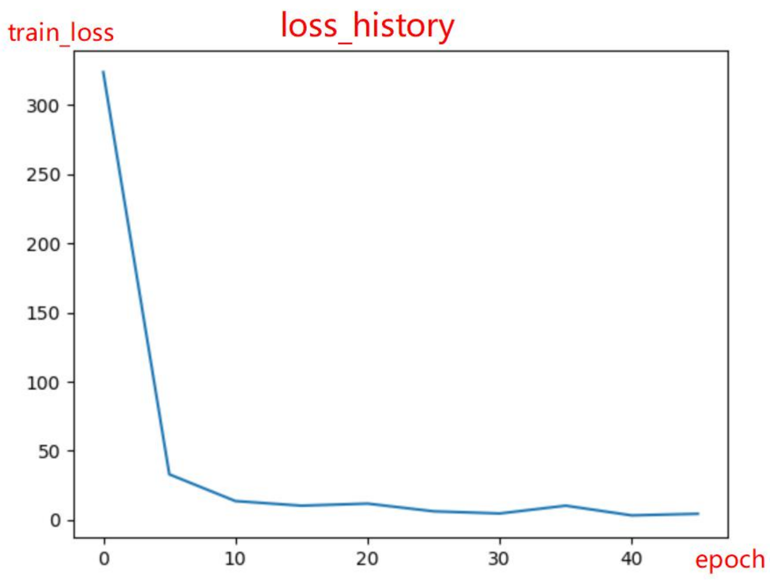
Dropout(p=0.4)

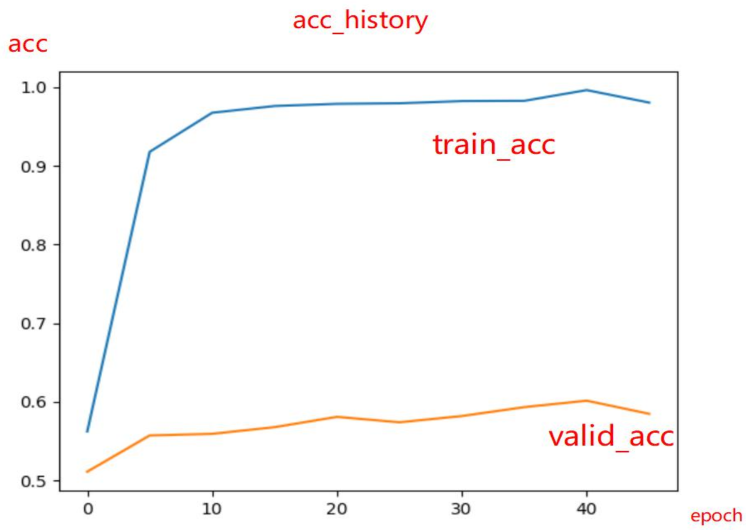
Linear -->　1*1*128

Dropout(p=0.4)

Linear --> 1*1*7

(1%) 請附上 model 的 training/validation history (loss and accuracy)。





(1%) 畫出 confusion matrix 分析哪些類別的圖片容易使 model 搞混，並簡單說明。

confusion matrix：

	0	1	2	3	4	5	6
0	270	6	51	41	75	12	42
1	7	30	4	1	1	0	0
2	67	3	268	28	82	31	48
3	56	0	28	822	54	19	66
4	74	7	90	40	320	8	115
5	12	1	38	24	14	348	9
6	65	3	59	62	114	22	351

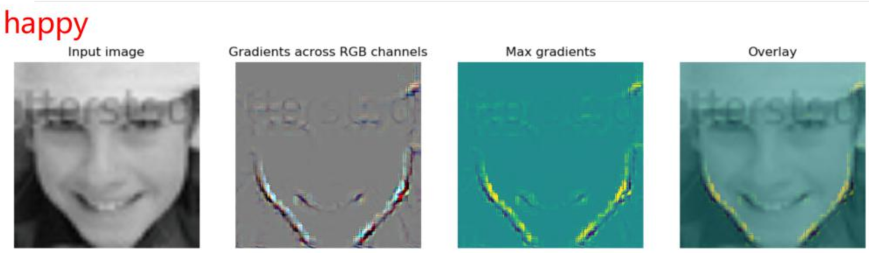
可以看恐懼和難過（2,4），中立和難過（4,6）容易使 model 搞混。

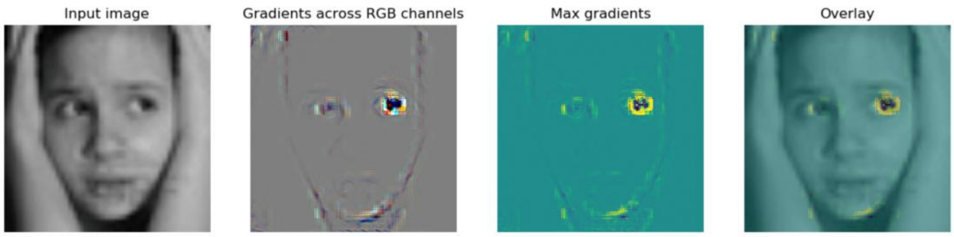
說明：恐懼（2）的正確識別率為 49.8% ，而誤認為難過（4）的幾率為 16.7%

難過（4）的正確識別率為 48.5%　而誤認為中立（6）的幾率為 17.27%

中立（6）的正確識別率為 55.1%　而誤認為難過（4）的幾率為 18.1%

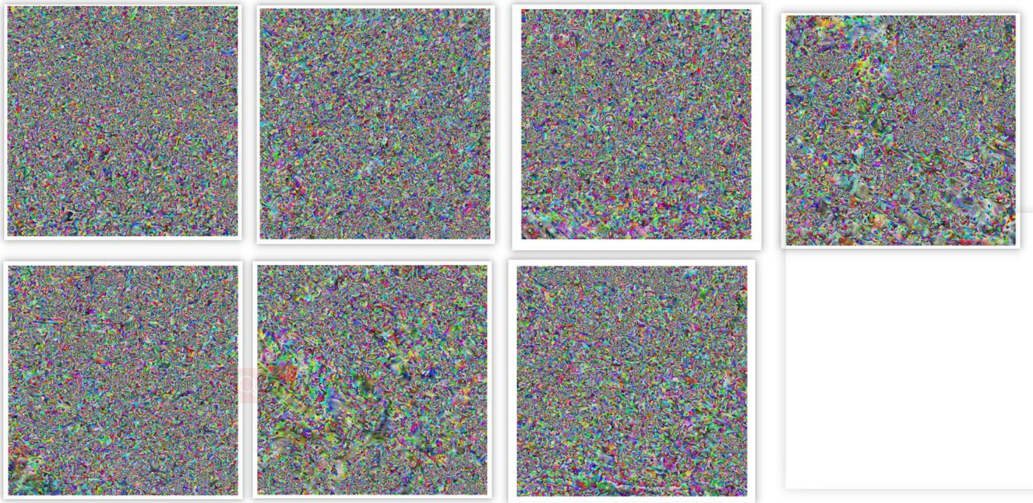
(1%) 畫出 CNN model 的 saliency map，並簡單討論其現象。





從上述例子可以看出，梯度下降的部分主要集中在眼睛，嘴巴和臉部輪廓，可以知道 CNN 的 model 主要是通過識別這些地方的特征來實現分類。

(1%) 畫出最後一層的 filters 最容易被哪些 feature activate。



filters visualization（0-6）

(3%)Refer to math problem
Nextpage：

1. $W' = \left\lfloor \frac{W + 2S_1 - k_1}{S_1} \right\rfloor + 1$ $\quad B' = B$

$H' = \left\lfloor \frac{H + 2S_2 - k_2}{S_2} \right\rfloor + 1$ $\quad$ new_input_channels = output_channels

shape = (B, W', H', output_channels)

2. $\partial y_i = \gamma \, \partial \hat{x}_i$

① $\frac{\partial l}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i / \gamma} = \frac{\partial l}{\partial y_i} \cdot \gamma$

② $\frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} (x_i - \mu_B) \cdot \frac{-1}{2} (\sigma_B^2 + \varepsilon)^{-3/2}$

③ $\frac{\partial l}{\partial \mu_B} = \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_B^2 + \varepsilon}} \right) + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{\sum_{i=1}^{m} -2(x_i - \mu_B)}{m}$

④ $\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \frac{1}{\sqrt{\sigma_B^2 + \varepsilon}} + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B} \cdot \frac{1}{m}$

⑤ $\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \cdot \hat{x}_i$

$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}$

3. $\frac{\partial \hat{y}_t}{\partial z_t} = \frac{e^{z_t}}{(\sum_i e^{z_i})^2} = \begin{cases} (1 - \hat{y}_t)\,\hat{y}_t, & k = t \\ -y_t \hat{y}_k, & k \neq t \end{cases}$ ①

$\partial l \quad \partial(-\sum_k y_k \log \hat{y}_k) \quad , \quad d\hat{y}_k$

$$\frac{}{\partial z_t} = \frac{}{\partial z_t} = -\sum_k y_k \frac{1}{\hat{y}_k} \frac{}{\partial z_t}$$

$$= -\left(y_t \frac{1}{\hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t}\right) - \left(\sum_{k \neq t} y_k \frac{1}{\hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_t}\right) \quad ②$$

① and ②: $$\frac{\partial L}{\partial z_t} = -\left(y_t \frac{1}{\hat{y}_t} (1-\hat{y}_t)\hat{y}_t\right) - \left(\sum_{k \neq t} y_k \frac{1}{\hat{y}_k} \cdot (-\hat{y}_t \cdot \hat{y}_k)\right)$$

$$= -y_t(1-\hat{y}_t) - \sum_{k \neq t} y_k \hat{y}_t$$

$$= -y_t + y_t\hat{y}_t - \sum_{k \neq t} y_k \hat{y}_t$$

$$= -y_t + \hat{y}_t\left(\sum_k y_k\right)$$

$\sum_k y_k = 1 \rightarrow$ $$\frac{\partial L}{\partial z_t} = \hat{y}_t - y_t$$