學號：B07901169　系級： 電機三　姓名：楊宗桓

(1%) 請說明這次使用的 model 架構，包含各層維度及連接方式。

Input 圖片 1*48*48

第一層卷積層:

- Conv2d:
  - Input channel: 1, output channel:32, kernel_size: 5*5, padding: 2*2,stride:1
- Leakyrelu: 負的坡度設為 0.05
- MaxPool2d: 2*2

第二層卷積層:

- Conv2d:
  - Input channel: 32, output channel:64, kernel_size: 3*3, padding: 1*1,stride:1
- Leakyrelu: 負的坡度設為 0.05
- MaxPool2d: 2*2

第三層卷積層:

- Conv2d:
  - Input channel: 64, output channel:128, kernel_size: 3*3, padding: 1*1,stride:1
- Leakyrelu: 負的坡度設為 0.05
- Batchnorm2d(做標準化): 128 channels out
- MaxPool2d: 2*2

第四層卷積層:

- Conv2d:
  - Input channel: 128, output channel:128, kernel_size: 3*3, padding: 1*1,stride:1
- Leakyrelu: 負的坡度設為 0.05
- Batchnorm2d(做標準化): 128 channels out
- MaxPool2d: 2*2

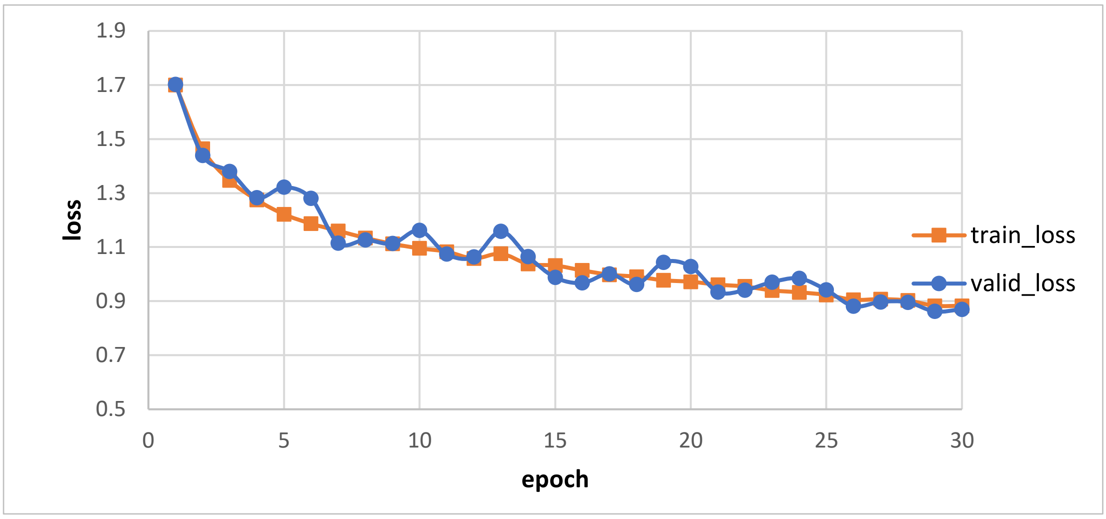
最後一層: (output)

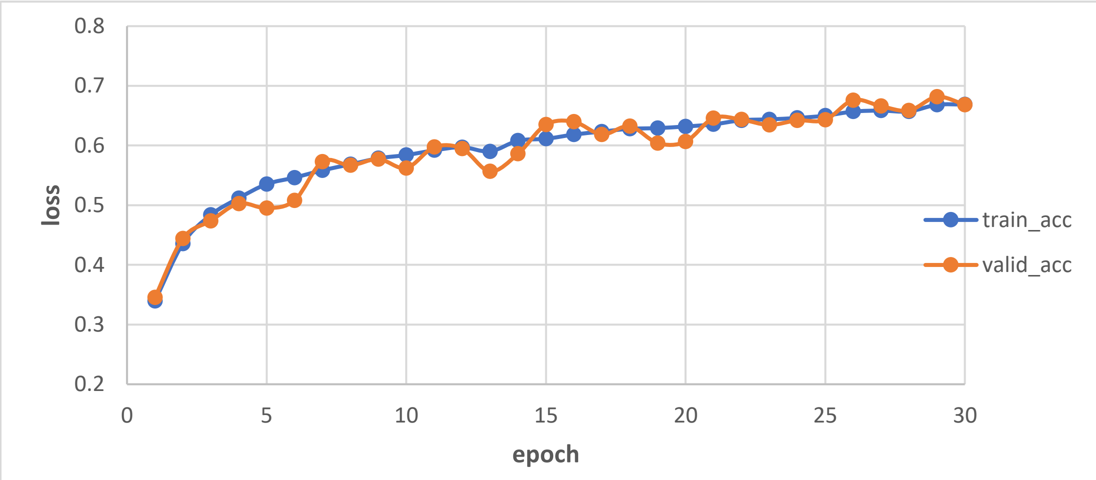
- Linear: input(3*3*128),output(256)
- Relu
- Batchnorm1d: 256 channels out
- Linear: input(256),output(7)

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

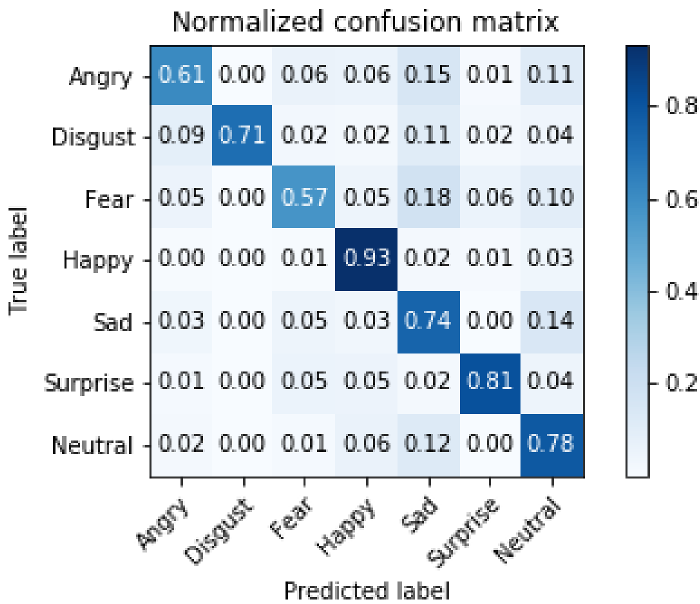
模型使用:

- Train_set: 前 25000 筆
- Validation_set: 剩下的
- 總 epoch: 30
- Learning rate: 0.0001
- Optimizer: Adam





(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。
(ref: https://en.wikipedia.org/wiki/Confusion_matrix)



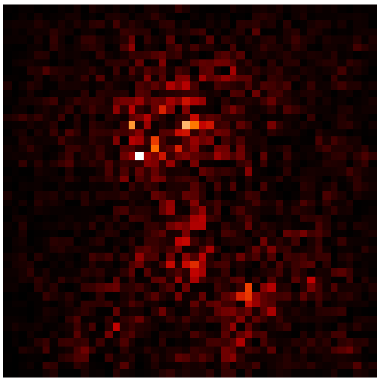
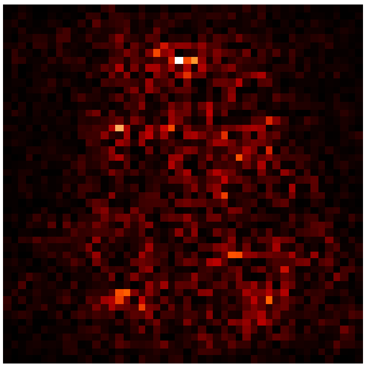
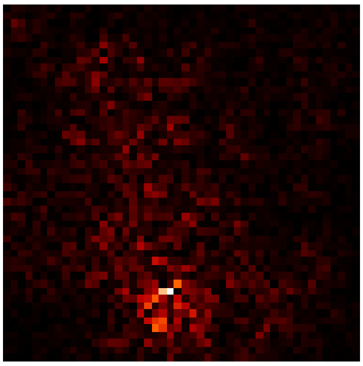
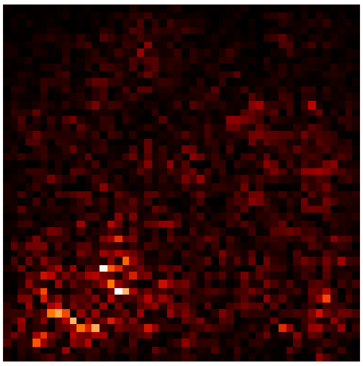
由 confusion_matrix 可看出，負面情緒互相被誤認的機會較大，可能是因為負面情緒的表情都很接近，對人來說可能就很難辨識，所以機器也會有較高機率便是錯。另外可以發現被誤認成中立還有難過的機率也較高，可能是因為兩者的表情都不明顯，所以比較可能被誤認。

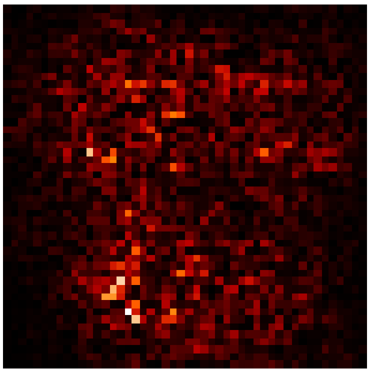
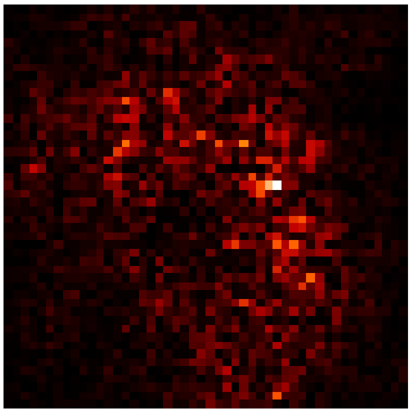
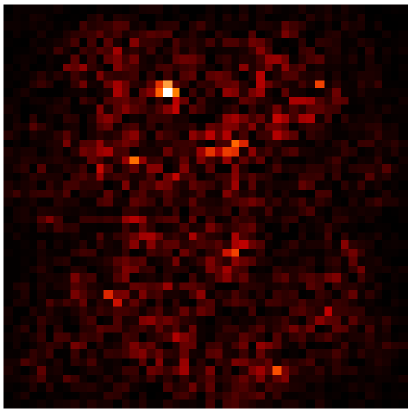
[關於第四及第五題]

可以使用簡單的 3-layer CNN model [64, 128, 512] 進行實作。

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

我挑出 7 個 class 的各一張，畫出其個別的 saliency map

生氣		
厭惡		
恐懼		
高興		

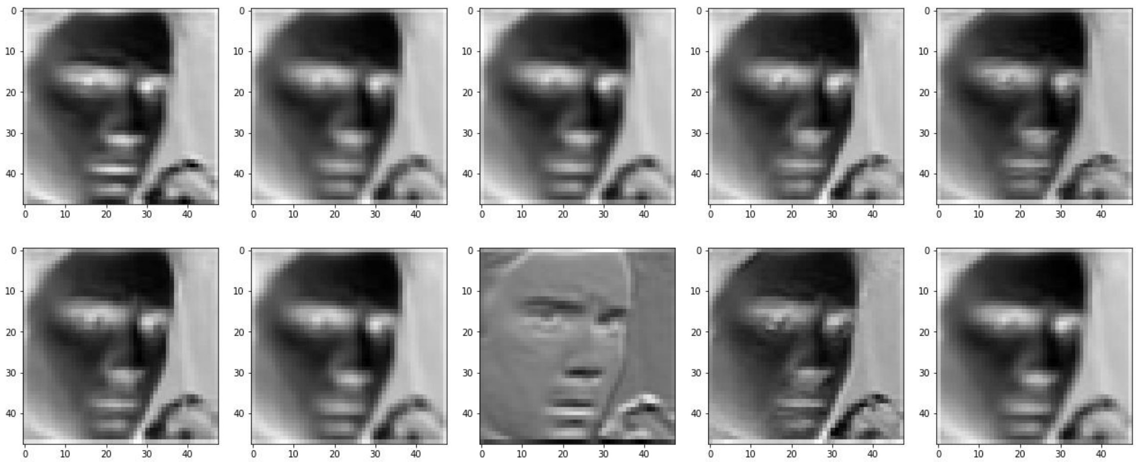
難過		
驚訝		
中立		

雖然不明顯，但兩相比對之後還是可以從 saliency map 發現，model 主要用來判別的部位(saliency map 中較亮的地方)是五官。

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

原圖:

從第四層卷積層隨機挑出來的 10 個 conv2d Filter



可以看到 filter extract 主要是被眼睛鼻子嘴巴所 activate

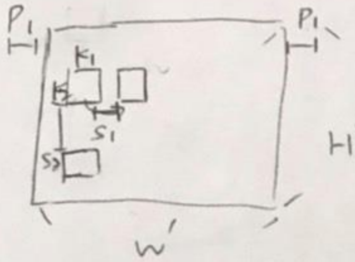
(3%)Refer to math problem

https://hackmd.io/@ASZWRvp7SjOEdYLqF3JYdg/HJMbtPOdD

HW 3.

1. Convolution.

W =



$$B' = B \quad \#$$

$$W' = \left[\frac{W+2P_1-K_1}{S_1} + 1\right] \#$$

$$H' = \left[\frac{H+2P_2-K_2}{S_2} + 1\right] \#$$

\# of channels = output_channels #

2. Batch Normalization.

To update γ & β from the optimization of loss,

we need to find gradients of loss, to optimize. (where ℓ is a function of "y" only)

$$\frac{\partial \ell(y)}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i} * \frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i} \cdot \frac{\partial}{\partial \hat{x}_i}(\gamma \hat{x}_i + \beta) = \frac{\partial \ell}{\partial y_i} \cdot \gamma \quad \#$$

$$\frac{\partial \ell}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma \cdot \frac{\partial \frac{x_i - u_B}{\sqrt{\sigma_B^2+\epsilon}}}{\partial \sigma_B^2}$$

$$= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma (x_i - u_B) \cdot \frac{-1}{2} (\sigma_B^2 + \epsilon)^{-\frac{3}{2}} \quad \#$$

$$\frac{\partial \ell}{\partial u_B} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i(u_B, \sigma_B^2)}{\partial u_B}$$

$$= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial u_B} + \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \sigma_B^2} * \frac{\partial \sigma_B^2}{\partial u_B}$$

$$= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma \frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \sum_{i=1}^{m} \frac{-2(x_i - u_B)}{m}$$

$$= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma \frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma (x_i - u_B) \frac{-1}{2} (\sigma_B^2+\epsilon)^{-\frac{3}{2}} \times \sum_{i=1}^{m} \frac{-2(x_i - u_B)}{m}$$

\#

2. 續

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial y_i}\cdot\frac{\partial y_i}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i(x_i,\mu_B,\sigma_B^2)}{\partial x_i}$$

$$= \frac{\partial l}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial l}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial \mu_B}\cdot\frac{\partial \mu_B}{\partial x_i} + \frac{\partial l}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial \sigma_B^2}\cdot\frac{\partial \sigma_B^2}{\partial x_i}$$

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

$$= \frac{\partial l}{\partial y_i}\cdot\gamma\frac{1}{\sqrt{\sigma_B^2+\epsilon}} + \underbrace{\frac{\partial l}{\partial \mu_B}}_{\text{代入前面求的}}\cdot\frac{1}{m} + \underbrace{\frac{\partial l}{\partial \sigma_B^2}}_{\text{代入前面所求的}}\cdot\frac{2(x_i-\mu_B)}{m}$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m}\frac{\partial l}{\partial y_i}\frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{m}\underbrace{\frac{\partial l}{\partial y_i}}_{\text{代入前面所求的}}\frac{\partial}{\partial \gamma}\hat{x}_i$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m}\frac{\partial l}{\partial y_i}\frac{\partial y_i}{\partial \beta} = \sum_{i=1}^{m}\frac{\partial l}{\partial y_i} \Leftarrow \text{代入前面所求}$$ #

以上可求得 $\frac{\partial l}{\partial \beta}$、$\frac{\partial l}{\partial \gamma}$ 之值(由其他 $\frac{\partial l}{\partial \hat{x}_i}$、$\frac{\partial l}{\partial \sigma_B^2}$ ……運算得到)

進而優化loss function #

3. Softmax and Cross Entropy

$$\frac{\partial L_t}{\partial z_t} = \frac{\partial L}{\partial y_t}\frac{\partial y_t}{\partial z_t} + \frac{\partial L}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial z_t}$$

note: cross entropy $L(y,\hat{y}) = -\sum_i y_i \log \hat{y}_i$

則 $L_t(y_t,\hat{y}_t) = -y_t\log\hat{y}_t$ 為 $y_t=1$ 的 cross entropy loss

否則 $L_t(y_t,\hat{y}_t)$ 應為 $-(1-y_t)\log(1-\hat{y}_t)$

$$\frac{\partial L}{\partial z_t} = \frac{\partial L}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial y_t} \quad (\because y_t \text{為定值}=1)$$

$$= \frac{\partial}{\partial \hat{y}_t}(-y_t\log\hat{y}_t)\cdot\frac{\partial}{\partial z_t}\left(\frac{e^{z_t}}{\sum_i e^{z_i}}\right)$$

$$= \frac{-y_t}{\hat{y}_t}\cdot\frac{\frac{\partial e^{z_t}}{\partial z_t}\cdot\sum_i e^{z_i} - e^{z_t}\cdot\frac{\partial\sum_i e^{z_i}}{\partial z_t}}{(\sum e^{z_i})^2}$$

$$= \frac{-y_t}{\hat{y}_t}\,\frac{e^{z_t}(\sum_i e^{z_t} - e^{z_t})}{(\sum e^{z_t})^2}$$

$$= \frac{-y_t}{\hat{y}_t}\left(\frac{e^{z_t}}{\sum e^{z_t}} - \left(\frac{e^{z_t}}{\sum e^{z_t}}\right)^2\right)$$

$$= \frac{-y_t}{\hat{y}_t}(\hat{y}_t - \hat{y}_t^2)$$

$$= -y_t + y_t\hat{y}_t$$

$$= \hat{y}_t - y_t \quad (\because y_t = 1)$$

#