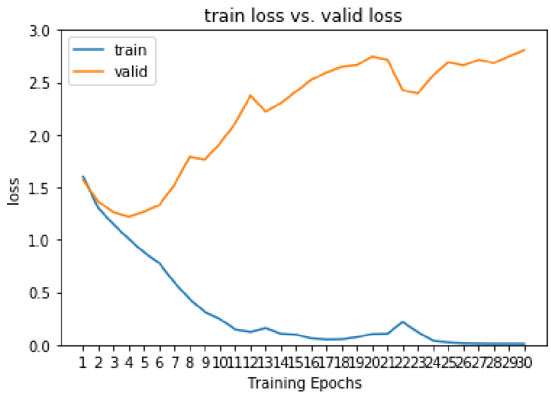
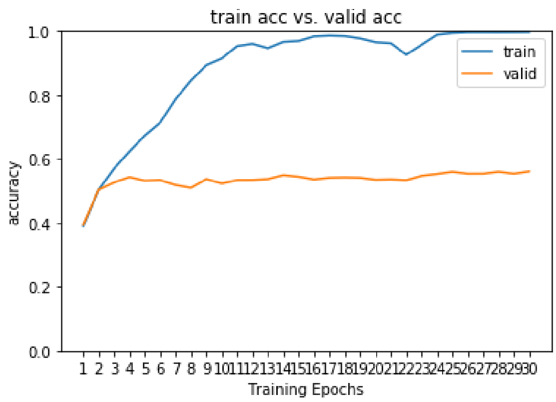
學號：R06522709　系級： 機械碩三　姓名：鄭呈毅

1. (1%) 請說明這次使用的 model 架構，包含各層維度及連接方式。
   我使用 TA 手把手 code。
   前面四層都是 convolutional layer+maxpooling
   Conv2d(input channels=1,output channels=32)//(48,48,32)
   Maxpool2d(2)//池化層 將圖片長寬縮為 1/2　(24,24,32)
   Conv2d(input channels=32,output channels=64)//(24,24,64)
   Maxpool2d(2)// (12,12,64)
   Conv2d(input channels=64,output channels=128)//(12,12,128)
   Maxpool2d(2)// (6,6,128)
   Conv2d(input channels=64,output channels=128)//(6,6,128)
   Maxpool2d(2)// (3,3,128)
   最後兩層 fully connected layer
   Linear(input channels=3*3*128, output layers=256)
   Relu()
   Linear(input channels=256, output layers=7)
   輸出 7 個 channel 對應到 7 個 label。

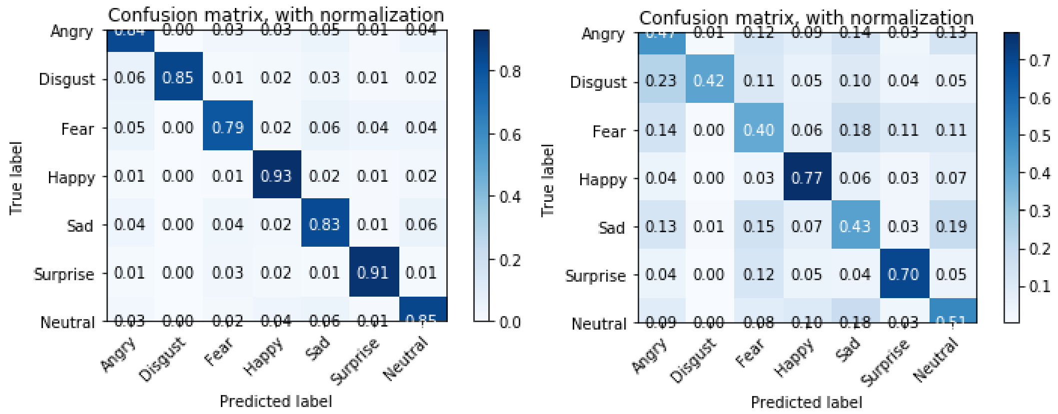
2. (1%) 請附上 model 的 training/validation history (loss and accuracy)。





3. (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)



左圖是 traning data 的 confusion matrix，右圖則是基於 validation data。

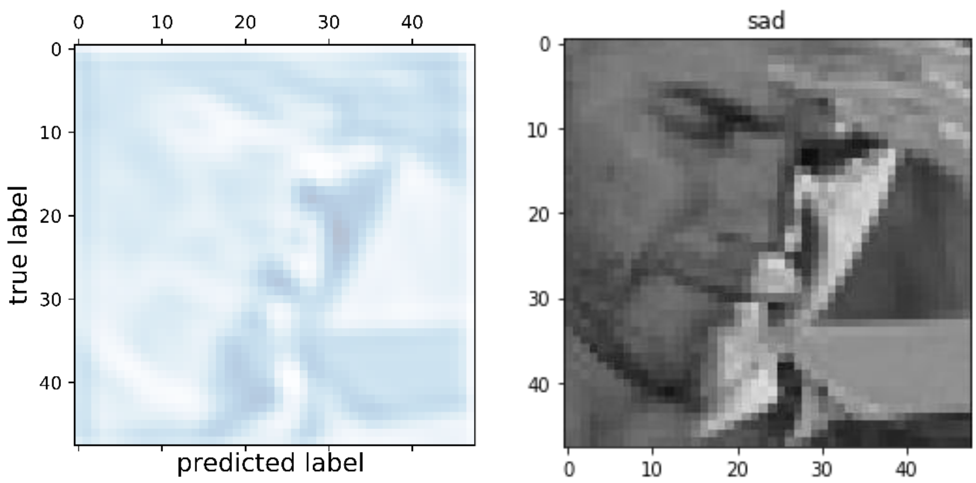
在兩張圖中，Happy 的準確度都特別高，從 raw data 中會發現該標籤的資料數量較多，加上 happy 是這七個標籤當中唯一表現正面能量的臉部表情，兩者都有可能是造成這個現象的原因之一。越接近的情感越容易被誤判:Disgust 被誤判的情況下，有很高的機率被誤判為 angry，這蠻合理的，當我們厭惡一個人，我們就容易對他生氣；再看看 sad 被誤判的情況下，有較高機率會被判為 neutral，當我們對一個人感到難過，他做得再多再好，你都覺得無感。

[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。

4. (1%) 畫出 CNN model 的 saliency map，並簡單討論其現象。

(ref: https://reurl.cc/Qpjg8b)



我是採用助教建議的[64,128,512]之 3-layer model。這是從第一層 convolution layer 所得到的 saliency map，可能由於是第一層的關係，故還是可以很明顯的看到輪廓，大致而言看起來像是原圖反白後的結果。

5. (1%) 畫出最後一層的 filters 最容易被哪些 feature activate。
   (ref: https://reurl.cc/ZnrgYg)
6. (3%)Refer to math problem

ML HW3 R06522709 鄭呈毅.

1. Convolution:

a batch of image data with shape ($\underset{\text{Batch size}}{B}$, $\overset{\text{width}}{W}$, $\underset{\text{height}}{H}$, Input_channels)

Conv2D(Input_channels, output_channels, kernal_size = (k_1, k_2), stride = (s_1, s_2), padding = (P_1, P_2))

Shape after convolution layer.

(B, $((W+2P_1)-k_1)/s_1+1$, $(H+2P_2-k_2)/s_2+1$, output_channels)

2. Batch Normalization:

parameters to be learned: γ, β

$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ // mini-batch mean.

$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2$ // mini-batch variance

$\hat{x}_i \leftarrow \frac{x_i-\mu_B}{\sigma_B^2+\epsilon}$ // normalize

$y_i \leftarrow \gamma\hat{x}_i+\beta \equiv BN_{\gamma,\beta}(x_i)$ // scale and shift.

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

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

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

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

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

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

3. Softmax and Cross Entropy.

$$\text{softmax}(z_t) = \frac{e^{z_t}}{\sum_i e^{z_i}}$$

$$\text{Cross\_entropy} = L(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i$$

$$\text{cross\_entropy} = L_t(y_t, \hat{y}_t) = -y_t \log \hat{y}_t \qquad (y_t = 1).$$

$$\hat{y}_t = \text{softmax}(z_t)$$

Derive that $\frac{\partial L_t}{\partial z_t} = \hat{y}_t - y_t$.

$$\frac{\partial \hat{y}_i}{\partial z_i} = \frac{\partial\left(\frac{e^{z_i}}{\Sigma_c}\right)}{\partial z_i} = \frac{e^{z_i}\Sigma_c - e^{z_i}e^{z_i}}{\Sigma_c^2} = \frac{e^{z_i}}{\Sigma_c}\,\frac{\Sigma_c - e^{z_i}}{\Sigma_c}$$

$$= \hat{y}_i(1 - \hat{y}_i)$$

$$\frac{\partial L_t}{\partial z_t} = \frac{\partial(-y_t \log \hat{y}_t)}{\partial z_t}$$

$$= -y_t \cdot \frac{\partial(\log \hat{y}_t)}{\partial z_t}$$

$$= -y_t \cdot \frac{1}{\cancel{\hat{y}_t}} \cdot \cancel{\hat{y}_t}(1 - \hat{y}_t)$$

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

Collaborator: R07543069 潘俊霖