Machine Learning Fall 2020 ——— Homework 3

學號：B07902037 系級： 資工三 姓名：蔡沛勳

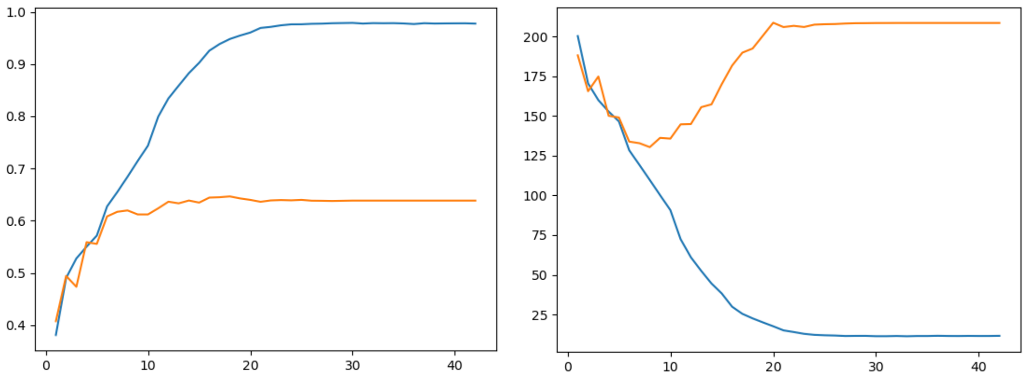
1. (1%) 請說明這次使用的 model 架構，包含各層維度及連接方式。

我的 model 使用了三層 CNN，具體結構如下：

```
self.cnn = nn.Sequential(
# Convolution 1 (1 * 48 * 48) -> (64 * 23 * 23)
        nn.Conv2d(in_channels = 1, out_channels = 64, kernel_size = 3, stride = 1, ),
        nn.BatchNorm2d(num_features = 64),
        nn.LeakyReLU(negative_slope = 0.1),
        nn.MaxPool2d(kernel_size = 2, stride = 2),

# Convolution 2 (64 * 23 * 23) -> (128 * 9 * 9)
        nn.Conv2d(in_channels = 64, out_channels = 128, kernel_size = 6, stride = 1, ),
        nn.BatchNorm2d(num_features = 128),
        nn.LeakyReLU(negative_slope = 0.1),
        nn.MaxPool2d(kernel_size = 2, stride = 2),

# Convolution 3 (128 * 9 * 9) -> (256 * 3 * 3)
        nn.Conv2d(in_channels = 128, out_channels = 256, kernel_size = 4, stride = 1, ),
        nn.BatchNorm2d(num_features = 256),
        nn.LeakyReLU(negative_slope = 0.1),
        nn.MaxPool2d(kernel_size = 2, stride = 2),
)

#Fully connection (256 * 3 * 3) -> (1024) -> (128) -> (7)
self.fc = nn.Sequential(
        nn.Linear(in_features = 256 * 3 * 3, out_features = 1024),
        nn.LeakyReLU(negative_slope = 0.1),
        nn.Dropout(p = 0.5),
        nn.Linear(in_features = 1024, out_features = 128),
        nn.LeakyReLU(negative_slope = 0.1),
        nn.Dropout(p = 0.5),
        nn.Linear(in_features = 128, out_features = 7)
)
```

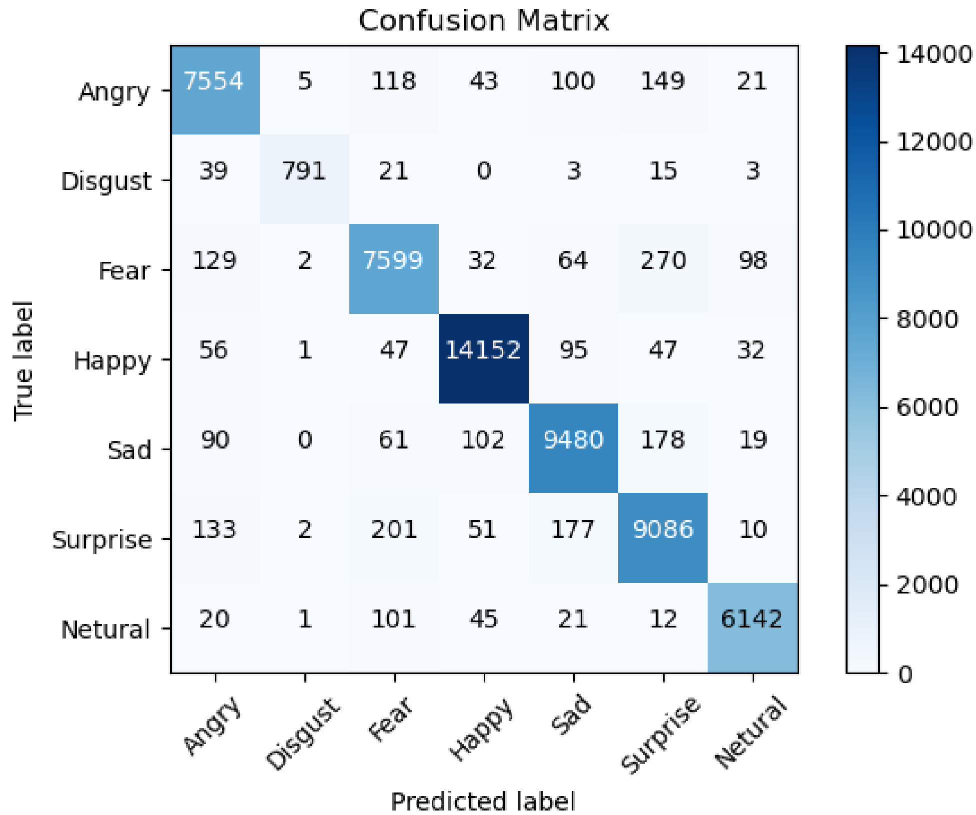
2. (1%) 請附上 model 的 training/validation history (loss and accuracy)。

我先將所有資料做左右翻轉得到兩倍資料後，再對全部資料的 90%拿來做 training，10%做 validation，在 batch size 為 128 時做 40 次 epoch 所得到結果如下。其中左圖為 training/validation 的 accuracy history，右圖為 training/validation 的 loss history。 (藍線為 training，橘線為 validation)



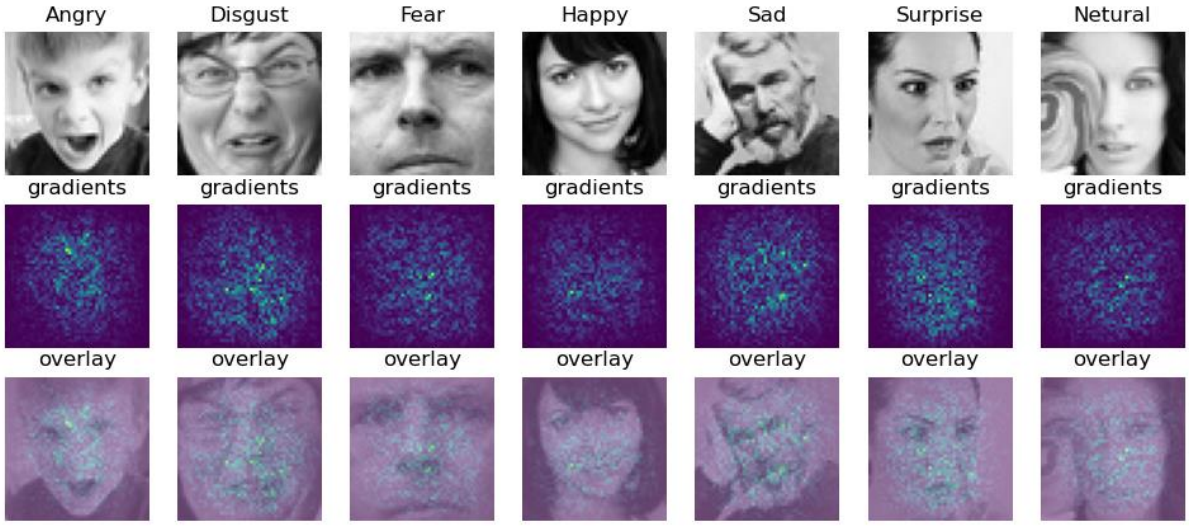
3. (1%) 畫出 confusion matrix 分析哪些類別的圖片容易使 model 搞混，並簡單說明。

下圖為原先資料加上左右翻轉的資料合計 57418 筆資料產生的 confusion matrix。



可以觀察出 Fear 跟 Surprise 彼此最容易使 model 搞混。推測原因為人類對這兩種情緒的反應較為相似(ex. 睜大眼睛、嘴巴)，導致 CNN 產生的結果也較為接近。

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



可以看出 gradient 再眼睛、鼻子及嘴巴等處數值較大，代表 model 主要是利用人的五官來辨識照片中人物的情緒。

5. (1%) 畫出最後一層的 filters 最容易被哪些 feature activate。

以下為我的 model 中最後一層的 256 筆 4 * 4 的 filters。

6. (3%) Refer to math problem

1.

設變化後大小變為 $\left(B^*, W^*, H^*, input\_channels^\wedge *\right)$。

$$B^* = B$$

$$W^* = \frac{W + 2 * p_1 - k_1}{s_1} + 1$$

$$H^* = \frac{W + 2 * p_2 - k_2}{s_2} + 1$$

$$input\_channels^* = output\_channels$$

2.

由題目得到

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$$

$$\widehat{x_\iota} = (x_i - \mu_B)\left(\sigma_B^2 + \epsilon\right)^{-\frac{1}{2}}$$

$$y_i = \gamma\widehat{x_\iota} + \beta$$

可推得

$\frac{\partial l}{\partial \widehat{x_\iota}}$：

$$\frac{\partial l}{\partial \widehat{x_\iota}} = \frac{\partial l}{\partial y_i}\frac{\partial y_i}{\partial \widehat{x_\iota}} = \frac{\partial l}{\partial y_i}\gamma$$

$\frac{\partial l}{\partial \sigma_B^2}$：

$$\frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m}\frac{\partial l}{\partial \widehat{x_\iota}}\frac{\partial \widehat{x_\iota}}{\partial \sigma_B^2} = -\sum_{i=1}^{m}\frac{\partial l}{\partial \widehat{x_\iota}}\frac{1}{2}(x_i - \mu_B)\left(\sigma_B^2 + \epsilon\right)^{-\frac{3}{2}}$$

$\frac{\partial l}{\partial \mu_B}$:

$$\frac{\partial l}{\partial \mu_B} = \sum_{i=1}^{m} \left(\frac{\partial l}{\partial \widehat{x_i}} \frac{\partial \widehat{x_i}}{\partial \mu_B} \right) + \frac{\partial l}{\partial \sigma_B^2} \frac{\partial \sigma_B^2}{\partial \mu_B}$$

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

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

$$= -\sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} + \frac{\partial l}{\partial \sigma_B^2} \frac{-2}{m} \left(\sum_{i=1}^{m} x_i - \sum_{i=1}^{m} x_i \right)$$

$$= -\sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} + 0 = -\sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}}$$

$\frac{\partial l}{\partial x_i}$:

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \widehat{x_i}} \frac{\partial \widehat{x_i}}{\partial x_i} + \frac{\partial l}{\partial \mu_B} \frac{\partial \mu_B}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2} \frac{\partial \sigma_B^2}{\partial x_i}$$

$$= \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} + \frac{\partial l}{\partial \mu_B} \frac{1}{m} + \frac{\partial l}{\partial \sigma_B^2} \frac{2}{m} (x_i - \mu_B)$$

$$= \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} - \frac{1}{m} \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}}$$

$$- \frac{1}{m} (x_i - \mu_B) \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} \left(x_j - \mu_B \right) \left(\sigma_B^2 + \epsilon \right)^{-\frac{3}{2}}$$

$$= \frac{\partial l}{\partial \widehat{x_i}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} - \frac{1}{m} \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}} - \widehat{x_i} \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} \frac{\widehat{x_j}}{m} \left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}}$$

$$= \frac{\left(\sigma_B^2 + \epsilon \right)^{-\frac{1}{2}}}{m} \left(m \frac{\partial l}{\partial \widehat{x_i}} - \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} - \widehat{x_i} \sum_{j=1}^{m} \frac{\partial l}{\partial \widehat{x_j}} \widehat{x_j} \right)$$

$\frac{\partial l}{\partial \gamma}$:

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \widehat{x_\iota}$$

$\frac{\partial l}{\partial \beta}$:

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

3.

已知

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

$$cross\_entropy = L_t(y, \hat{y}) = -\sum_i y_i \log \hat{y}$$

$$cross\_entropy = L_t(y_t, \widehat{y_t}) = -y_t \log \widehat{y_t}$$

$$\widehat{y_t} = softmax(z_t)$$

在 $y_t = 1$ 時 $cross\_entropy = L_t(y_t, \widehat{y_t}) = -y_t \log \widehat{y_t}$，故

$$\begin{aligned}
\frac{\partial L_t}{\partial z_t} &= -\frac{\partial}{\partial z_t} y_t \log \frac{e^{z_t}}{\sum_i e^{z_i}} \\
&= -\frac{\partial}{\partial z_t} y_t \left(\log e^{z_t} - \log \sum_i e^{z_i} \right) \\
&= -y_t \left(1 - \frac{e^{z_t}}{\sum_i e^{z_i}} \right) \\
&= -y_t (1 - \widehat{y_t}) = \widehat{y_t} y_t - y_t = \widehat{y_t} - y_t
\end{aligned}$$

在 $y_t = 0$ 時 $cross\_entropy = L_t(y_t, \widehat{y_t}) = -(1 - y_t) \log(1 - \widehat{y_t})$，故

$$\begin{aligned}
\frac{\partial L_t}{\partial z_t} &= -\frac{\partial}{\partial z_t} (1 - y_t) \log \left(1 - \frac{e^{z_t}}{\sum_i e^{z_i}} \right) \\
&= -\frac{\partial}{\partial z_t} (1 - y_t) \left(\log \sum_{i, i \neq t} e^{z_i} - \log \sum_i e^{z_i} \right) \\
&= -(1 - y_t) \left(0 - \frac{e^{z_t}}{\sum_i e^{z_i}} \right) \\
&= (1 - y_t)\, \widehat{y_t} = \widehat{y_t} - y_t \widehat{y_t} = \widehat{y_t} - y_t
\end{aligned}$$