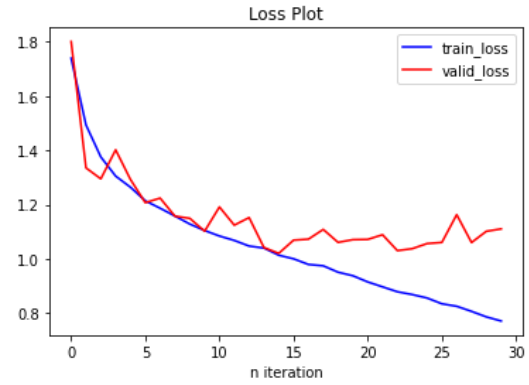
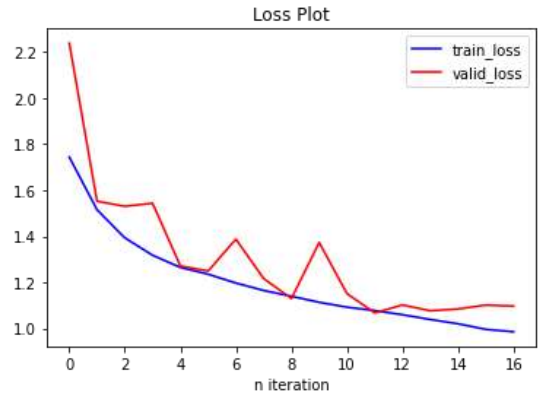
學號：R10921019　系級： 電機所碩二　姓名：鄭皓方

1. (1%) 實作early-stopping, 繪製training, validation loss 的 learning curve, 比較實作前後的差異, 並說明early-stopping的運作機制

我使用的early stopping機制十分簡單，就是如果連續五次沒有產生更高的分數，trainning就會停下來，如下方程式我用break_count的參數計算沒有更新的次數，如果創新高就會更新，下圖為有沒有early stopping 的loss情形，可以發現左圖的validation的error緩緩升高，右圖有及時停下來就可以適當避免overfitting。

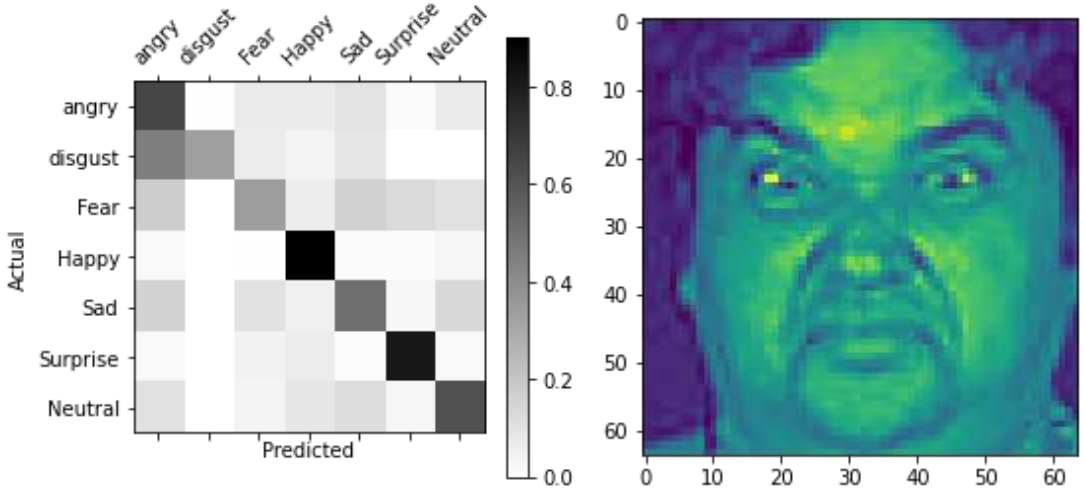
	
沒有early stop	有 early stop

2. (1%) 嘗試使用 augmentation, 說明實作細節並比較有無該 trick 對結果表現的影響(validation 或是 testing 擇一即可), 且需說明為何使用這些augmentation的原因。
(ref: https://pytorch.org/vision/stable/transforms.html)

我使用torchvision 的 transform來實作 augenmetation，用transform.compose可以把很多功能合併一起使用，我使用了兩種方法，一種是旋轉，範圍是15度，我發現旋轉的範圍不能太大，不然效果會變差，我猜想可能是因為model會誤認成其他表情；第二種方法是水平翻轉，我選擇翻轉的機率是0.5，由於臉是水平對稱的，水平翻轉的話我認為可以增加測資的多樣性。

使用方法	Testing Public Result
沒有augenmentation	0.59800
有rotation	0.61466
有horizental flip	0.62400
rotation + horizental flip	0.63733

3. (1%) 畫出 confusion matrix 分析哪些類別的圖片容易使 model 搞混，找出模型出錯的例子，並分析可能的原因。
(ref: https://en.wikipedia.org/wiki/Confusion_matrix)
從左圖的confusion matrix 可以發現，model很容易將disgust判斷成angry，我覺得最直接的原因就是disgust的數量太少，如第四題的表格，disgust是angry的9分之一，至於為甚麼是認成angry而不是其他的表情，我舉了一個例子在右圖。我猜想應該是生氣和覺得噁心的時候，眉毛都會呈現倒八字形，所以model把這張圖片誤認為angry。



，

4. (1%) 請統計訓練資料中不同類別的數量比例，並說明：
對 testing 或是 validation 來說，不針對特定類別，直接選擇機率最大的類別會是最好的結果嗎？
(ref: https://arxiv.org/pdf/1608.06048.pdf, or hints: imbalanced classification)

表情	數量
angry	3139
disgust	358
fear	3296
happy	5762
sad	3785
surprise	2515
neutral	3925

我認為在測試資料中猜最多的種類不會準確，這個分兩個面向來討論，第一個是正確率，以我們這題當例子，假設testing data分布跟 training data類似，全猜happy，正確率也只有0.25，離我訓練出來的0.6還有一段距離，如果資料分布不同，結果還會更糟。第二個面向是在imblance data中，有時候數量比較少的答案是我們需要的，例如說判斷有沒有確診(有確診比沒確診的人少很多)，如果全部猜沒有確診，雖然acc很高，但是

一點用處都沒有，因此我認為即使面對imblance data也不可以全部猜最多的。比較好的方式我覺得可以對較少的測資作多一點的augenmentation並增加該類的總量，或者是減少一些數量很多種類的data。

5. (4%)Refer to math problem
https://hackmd.io/@lH2AB7kCSAS3NPw2FffsGg/r1otQp7Gi?fbclid=IwAR0cs5CajVy_zhDmHEDgze2V1_Jlxp95N45BF6hg1l6CgG-6IViYGAIGReE

Problem 1.

Define the Liklyhood function L as

$$L = P(x_1)P(x_2)P(x_3)\cdots P(x_i)$$

I replace the small k for i to prevent confused

due to the independency of each point

Define each point's classification as $C_{x_1}, C_{x_2}, C_{x_3} \cdots C_{x_N}$

$$\Rightarrow L = P(C_{x_1})P(C_{x_1}|C_k)\cdot P(C_{x_2})P(C_{x_2}|C_k)\cdots P(C_{x_N})P(C_{x_N}|C_k)$$

$$= \prod_{i=1}^{N} P(C_{x_i})\,P(C_{x_i}|C_k)$$

take ln

$$\Rightarrow \ln L = \sum_{i=1}^{N} \ln P(C_{x_i}) + \sum_{i=1}^{N} \ln P(C_{x_i}|C_k)$$

$$\Rightarrow \sum_{i=1}^{N} \ln P(C_{x_i}) = N_1 \ln P(C_1) + N_2 \ln P(C_2) \cdots N_K \ln P(C_k)$$

$$= \sum_{i=1}^{K} N_i \ln P(C_k) = \sum_{i=1}^{K} N_i \ln \pi_i$$

where $\sum_{i=1}^{K} \pi_i = 1$, the Lagragian of $\ln L$

will become $\ln L = \sum_{i=1}^{K} N_i \ln \pi_i - \lambda\left(\sum_{k=1}^{K} \pi_{(k)} - 1\right)$

$$\frac{\partial \ln L}{\partial \pi_i} = \sum_{i=1}^{K}\left(\frac{N_i}{\pi_i} - \lambda\right) \quad \pi_i = \frac{N_i}{\lambda} \Rightarrow \lambda = N$$

$$\frac{\partial \ln L}{\partial \lambda} = 1 - \sum_{i=1}^{K} \pi_i = 1 - \sum_{i=1}^{K} \frac{N_i}{\lambda} \quad \pi_i = \frac{N_i}{N}$$

#

Problem 2.

a. $\frac{\partial w^T A w}{\partial w} = \nabla_w (w^T A w) = \nabla_w f(w)$

$f(w+\Delta w) = (w+\Delta w)^T A (w+\Delta w) = w^T A w + \Delta w^T A w + w^T A \Delta w + \Delta w^T A \Delta w$

$= f(w) + \Delta w^T A w + \Delta w^T A^T w + \Delta w^T A \Delta w$

$\frac{f(w+\Delta w) - f(w)}{\Delta w} = A w + A^T w + A \Delta w \quad \Delta w \to 0$

$\nabla_w f = A w + A^T w$, if A is symmetric, $\nabla_w f = 2Aw$ ※

b. $(AB)_{11} = \sum_{i=1}^{m} a_{1i} b_{i1}$

$(AB)_{22} = \sum_{i=1}^{m} a_{2i} b_{i2}$

$\vdots$

$(AB)_{mm} = \sum_{i=1}^{m} a_{mi} b_{im}$

$\Rightarrow tr(AB) = \sum_{j=1}^{m} \sum_{i=1}^{m} a_{ij} b_{ji}$

故 $\frac{\partial tr(AB)}{\partial a_{ij}} = b_{ji}$ ※

c. $|\Sigma| = \sum_{j=1}^{n} (-1)^{i+j} \sigma_{ij} |\Sigma_{ij}| = \sum_{i=1}^{n} (-1)^{i+j} \sigma_{ij} |\Sigma_{ij}|$

$\frac{\partial \log|\Sigma|}{\partial \sigma_{ij}} = \frac{(-1)^{i+j} |\Sigma_{ij}|}{|\Sigma|}$ from cramer rule

$\frac{\partial \log|\Sigma|}{\partial \sigma_{ij}} = e_j^T \Sigma^{-1} e_i$ ※

Problem 3.

a. $\ln L = \sum_{k=1}^{K} N_k \ln \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik} \ln P(x_i|C_k)$

from Gaussion distribution → define as B

$$P(x_i|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$$\Rightarrow \ln P(x_i|C_k) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + B$$

$$\frac{\partial \ln L}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\sum_{i}^{N}\sum_{k}^{K} t_{ik} \ln P(x_i|C_k) + B$$

$$= \sum_{i=1}^{N} t_{ik} \frac{\partial}{\partial \mu_k} \ln P(x_i|C_k)$$

$$= \sum_{i=1}^{N} t_{ik} \frac{\partial}{\partial \mu_k}\left[-\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right]$$

$$= \sum_{i=1}^{N} t_{ik}\left\{-\Sigma^{-1}(x_i-\mu_k)\right\}$$

$$= -\Sigma^{-1}\sum_{i=1}^{N} t_{ik}x_i + \Sigma^{-1}N_k\mu_k = 0$$

$$\mu_k = \frac{\sum_{i=1}^{N} t_{ik}x_i}{N_k}$$ ※

b. $\frac{\partial \ln L}{\partial \Sigma^{-1}} = \frac{\partial}{\partial \Sigma^{-1}}\sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}\ln P(x_i|C_k)$

$$= \frac{\partial}{\partial \Sigma^{-1}}\sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}\left(-\tfrac{1}{2}(\mu_k-x_i)^T\Sigma^{-1}(\mu_k-x_i) - \tfrac{1}{2}\ln|\Sigma| - \tfrac{D}{2}\ln 2\pi\right)$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}\left(\underbrace{-\tfrac{1}{2}(\mu_k-x_i)(\mu_k-x_i)^T}_{1} + \tfrac{1}{2}\Sigma\right) \Rightarrow \text{transfer 1 with } \Sigma_k$$

$$= -\frac{1}{2}\sum_{k=1}^{K} N_k\Sigma_k + \frac{1}{2}N\Sigma = 0$$

$$N\Sigma = \sum_{k=1}^{K} N_k\Sigma_k \qquad \Sigma = \sum_{k=1}^{K}\frac{N_k}{N}\Sigma_k$$ ※

Problem 4.

b. $L(C^t, \mu^t) = \sum_{q=1}^{k} \sum_{i=C^t(i)=q} \|x_i - \mu_q^t\|_2^2$

The step 1 will minimize the C^t as

$$C^{t+1}(i) = \arg\min_{j=1,\dots,k} \|x_i - \mu_j\|_2^2$$

thus, $\sum_{q=1}^{k} \sum_{i=C^t(i)=q} \|x_i - \mu_q^t\|_2^2 \geq \sum_{q=1}^{k} \sum_{i=C^{t+1}(i)=q} \|x_i - \mu_q^t\|_2^2$

$$L(C^t, \mu^t) \geq L(C^{t+1}, \mu^t)$$

c. Accordding to the result of a

$$\sum_{i=C^{t+1}(i)=q} \|x_i - \mu_q^{t+1}\|_2^2 \leq \sum_{i=C^{t+1}(i)=q} \|x_i - x\| \quad x \in R$$

where $\mu_q^{t+1} = \frac{\sum_{i=C(i)=j} x_i}{|i = C^{t+1}(i) = j|} = \overline{x_i}$

thus, $\sum_{q=1}^{k} \sum_{i=C^{t+1}(i)=q} \|x_i - \mu_q^t\| \geq \sum_{q=1}^{k} \sum_{i=C^{t+1}(i)=q} \|x_i - \mu^{t+1}\|$

$$L(C^{t+1}, \mu^t) \geq L(C^{t+1}, \mu^{t+1})$$

d. $L(C^t, \mu^t) \geq L(C^{t+1}, \mu^t) \geq L(C^{t+1}, \mu^{t+1}) \geq 0$

sequence $\{l_t\}$ is monotone decreasing and bounded,

thus, according to the monotone theorm, $\{l_t\}$

converges.

e. since the number of $C(i)$ is limited (there's only k^n combinations), and the loss is keep decreasing, once $\mu^{t+1} = \mu^t$, then $L(C^{t+1}, \mu^{t+1}) = L(C^t, \mu^t)$, and the K-means clustering would stop in finite steps.