學號：R10922111 系級： 資工所碩一 姓名：黃冠瑋

model checkpoint 連結：https://drive.google.com/file/d/1TmwdxoTSONjj5HXOaKYnXWyfb8trzWGo/view?usp=sharing

1. (1%) 請以block diagram或是文字的方式說明這次表現最好的model使用哪些layer module(如 Conv/Linear 和各類 normalization layer) 及連接方式(如一般forward 或是使用 skip/residual connection)，並概念性逐項說明選用該 layer module 的理由。

   conv --> normalize --> leakyReLU --> maxPool
   因為圖片有局部的特性，也就是假設要在圖中尋找鳥的嘴巴，不一定要觀察整張圖，只看一小部分也能辨認出那是一個鳥嘴，因此選擇conv達到這個目的。
   normalize 則是要標準化不同pattern，已加快訓練
   leakyReLU 相較於linear可以訓練得更快，而選擇leaky是因為它可以防止某些神經元一直沒有使用到
   maxPool 則是因為大部分的情況下，將一張圖縮小並不會影響到我們辨識圖中物體。

2. (1%) 嘗試使用 augmentation/early-stopping/ensemble 三種訓練 trick 中的兩種，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)。

   early-stopping: 記錄下valid資料精準度的最大值，當連續五次都沒有超過它就停止繼續訓練。
   我的訓練次數設定為60，有使用early-stopping的模型最後的valid acc為0.6456，而沒有使用的最後則為0.5849
   augmentation: 載入訓練資料時加上transforms，我選擇的有水平翻轉、左右旋轉15度
   使用augmentation的模型在train data的精準度只能達到0.84多，但在valid data的精準度能達到0.6456。而沒有使用augmentation的模型在train data的精準度能夠高達0.97，但在 valid data最高只能到0.58。

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)

	0	1	2	3	4	5	6
0	0.54	0.13	0.12	0.03	0.124	0.023	0.099
1	0.031	0.826	0.02	0.003	0.014	0.003	0.001

2	0.122	0.043	0.48	0.025	0.166	0.078	0.096
3	0.045	0	0.046	0.832	0.034	0.049	0.072
4	0.136	0	0.157	0.024	0.497	0.011	0.141
5	0.03	0	0.123	0.016	0.02	0.82	0.037
6	0.09	0	0.048	0.07	0.145	0.014	0.554

最常把恐懼、難過、中立這三者搞錯，而厭惡、高興、驚訝搞錯的比例則非常低，應該是因為像是高興很明顯會有嘴巴笑的樣子、驚訝則會有眼睛睜大，而恐懼、難過的特徵較相似。

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？針對上述內容，是否存在更好的方式來提升表現？例如設置不同條件來選擇預測結果/變更訓練資料抽樣的方式，或是直接回答「否」(但需要給出支持你論點的論述)
   我覺得當機器訓練出來時，當機率彼此都很接近時在直接選擇機率最大的類別會是比較好的結果，舉例來說，當某張圖對於恐懼、難過、中立的機率都很高且相近時，就從這三類中挑圖片總數量最多的那個做為結果。

5. (3%)Refer to math problem
   https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/SJy_akYUK

1. Each image's size from (W, H) to $(W+2p_1, H+2p_2)$

$$\begin{cases} k_1 + (x-1)s_1 \le W + 2p_1 \\ k_2 + (y-1)s_2 \le H + 2p_2 \end{cases} \Rightarrow \begin{aligned} x &\le \frac{W+2p_1-k_1}{s_1} + 1 \\ y &\le \frac{H+2p_2-k_2}{s_2} + 1 \end{aligned}$$

So, after convolution layer, it's shape will become $(B, \frac{W+2p_1-k_1}{s_1}+1, \frac{H+2p_2-k_2}{s_2}+1, \text{output\_channel})$

2. To do the optimization process of loss, let loss function $\ell(y_1, \dots, y_m)$, and $y_i = \gamma \hat{x}_i + \beta$

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

$$\frac{\partial \ell}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_B^2}, \quad \text{and} \quad \frac{\partial \hat{x}_i}{\partial \sigma_B^2} = \frac{\partial (x_i - \mu_B)(\sigma_B^2 + \varepsilon)^{-\frac{1}{2}}}{\partial \sigma_B^2} = -\frac{1}{2}(x_i - \mu_B)(\sigma_B^2 + \varepsilon)^{-\frac{3}{2}}$$

(b) ,d $\therefore \frac{\partial \ell}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \gamma \cdot -\frac{1}{2}(x_i - \mu_B)(\sigma_B^2 + \varepsilon)^{-\frac{3}{2}}$

$$\frac{\partial \ell}{\partial \mu_B} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_B} + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \frac{\partial \sigma_B^2}{\partial \mu}, \quad \text{and} \quad \frac{\partial \hat{x}_i}{\partial \mu_B} = \frac{-1}{\sqrt{\sigma_B^2 + \varepsilon}}, \quad \frac{\partial \sigma_B^2}{\partial \mu} = \frac{1}{m} \sum_{i=1}^{m} 2(x_i - \mu_B)(-1)$$

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

(a) $\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial \ell}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \frac{\partial \sigma_B^2}{\partial x_i}$

$\therefore \frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \varepsilon}}, \quad \frac{\partial \mu}{\partial x_i} = \frac{1}{m}, \quad \frac{\partial \sigma_B^2}{\partial x_i} = \frac{2(x_i - \mu_B)}{m}$

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

$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \hat{x}_i, \quad \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i}$

3.

$$\frac{\partial L}{\partial z_t} = -\frac{\partial}{\partial z_t}\sum y_i \log \hat{y}_i = -\sum y_i \frac{\partial}{\partial z_t}\log \hat{y}_i$$

Since $\frac{\partial \log \hat{y}_i}{\partial z_t} = \frac{\partial\left(\log e^{z_i} - \log \sum_j e^{z_j}\right)}{\partial z_t} = \frac{\partial z_i}{\partial z_t} - \frac{e^{z_t}}{\sum_j e^{z_j}} = \frac{\partial z_i}{\partial z_t} - \hat{y}_t$

so, $\frac{\partial L}{\partial z_t} = -\sum y_i\left(\frac{\partial z_i}{\partial z_t} - \hat{y}_t\right) = -\sum y_i \frac{\partial z_i}{\partial z_t} + \sum y_i \cdot \hat{y}_t$

$= -y_i + \hat{y}_t$ $\quad \because \sum y_i = 1$

4.

(a) $m^t = \beta_1 m^{t-1} + (1-\beta_1) g^t$

$= \beta_1^2 m^{t-2} + \beta_1(1-\beta_1) g^{t-1} + (1-\beta_1) g^t = \cdots = \beta_1^t \cdot m^0 + \beta_1^{t-1}\cdot(1-\beta_1) g^1 \cdots + (1-\beta_1) g^t$

$\therefore m^t = (1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g^i$

$v^t = \beta_2 v^{t-1} + (1-\beta_2)(g^t)^2$

$= \beta_2(\beta_2 v^{t-2} + (1-\beta_2)(g^{t-1})^2) + (1-\beta_2)(g^t)^2$

$= \beta_2^2 v^{t-2} + \beta_2(1-\beta_2)(g^{t-1})^2 + (1-\beta_2)(g^t)^2$

$= \cdots$

$= (\beta_2)^t v^0 + (\beta_2)^{t-1}(1-\beta_2)(g^1)^2 + \cdots + (1-\beta_2)(g^t)^2$

$= (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i}(g^i)^2$

$= \beta_1^2 (\beta_1 m^{t-3} + (1-\beta_1) g^{t-2})$

$\beta_1^3 m^{t-3} + \beta_1^2 (1-\beta_1) g^{t-2} + \dots$

$0^0 = 1$

(b) Adam: $w^t = w^{t-1} - \dfrac{\eta_0 \cdot \frac{1}{\sqrt{t}}}{\sqrt{\frac{v^t}{1-\beta_2}}} \cdot \dfrac{m^t}{1-\beta_1}$

by (a) $= w^{t-1} - \dfrac{\frac{\eta_0}{\sqrt{t}}}{\sqrt{\sum \beta_2^{t-i} (g^i)^2}} \cdot \sum_{i=1}^{t} \beta_1^{t-i} g^i$

$= w^{t-1} - \dfrac{\frac{\eta_0}{\sqrt{t}}}{\sqrt{\sum (g^i)^2}} \cdot g^t$

Adagrad: $w^t = w^{t-1} - \dfrac{\eta}{\sqrt{\sum_{i=0}^{t} (g^i)^2}} g^t$