DeepLearning Lab1: Dataset and DataLoader

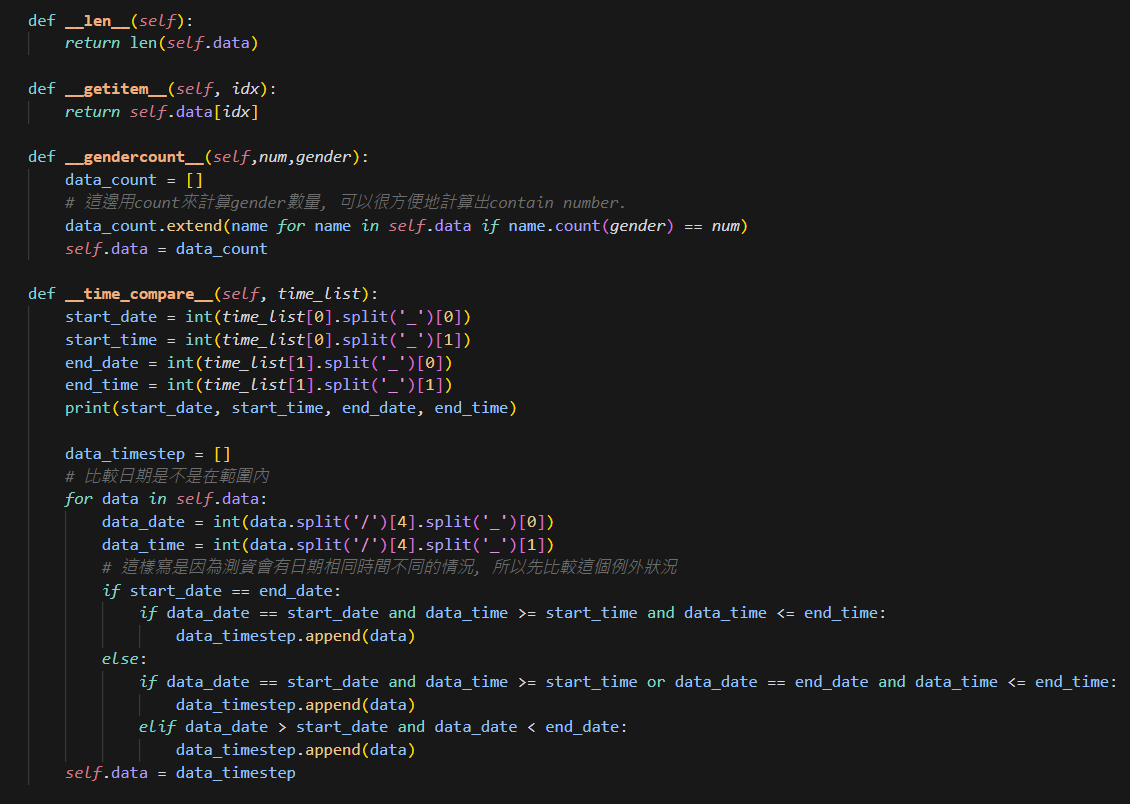
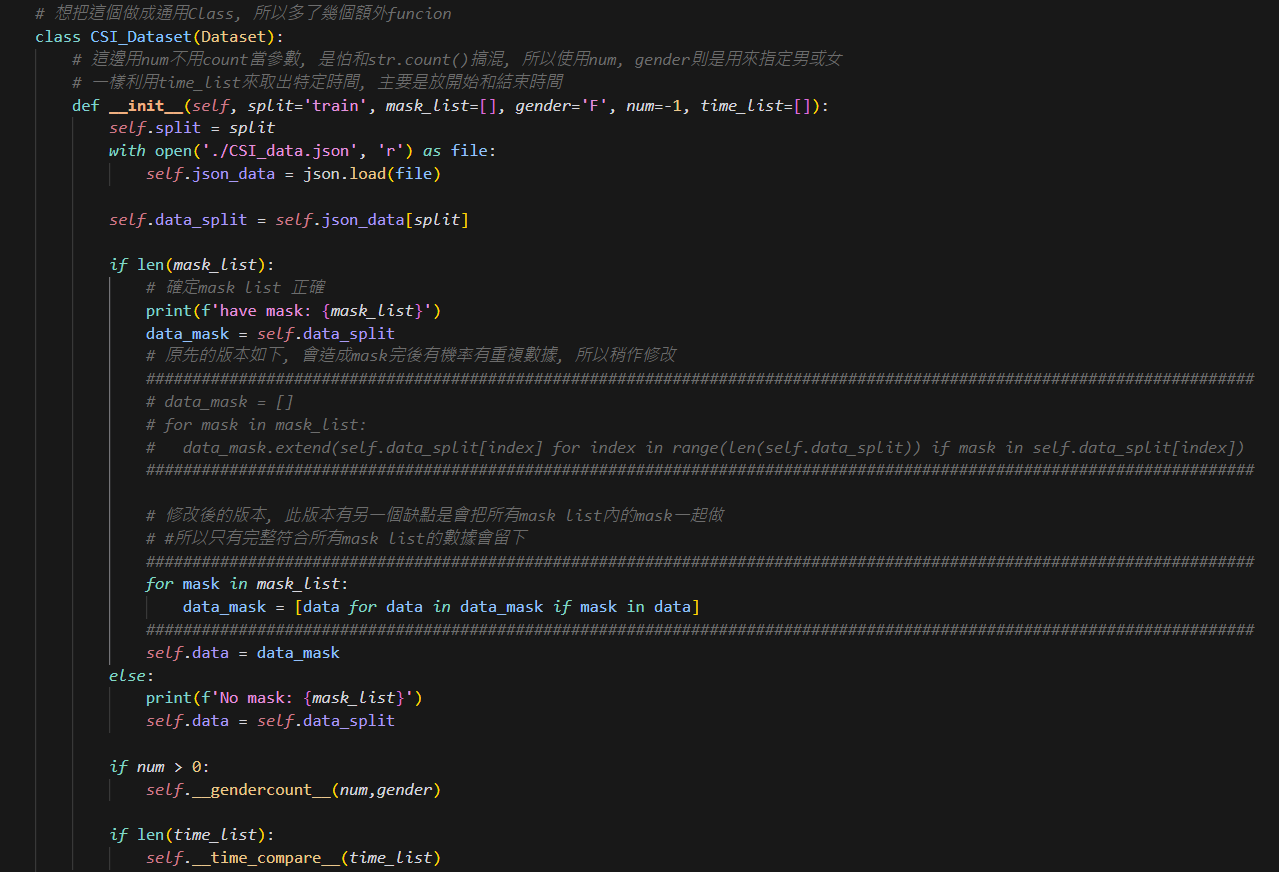
學號:31383404 姓名:周彥宏

[Github link](https://github.com/Qmooo/NYCU-DeepLearning/tree/main/HW1)

這次的Lab實作中,主要目的是想最大程度減小rewrite Class CSI\_DataSet的次數, 讓我在面對每個requirement都可以直接用已經寫好的**通用**Class去做。

Step 1. 觀察CSI\_data.json

1. Dataset中分成三個不同的資料集, 分別是”train”, “val”, “test”.
2. 主要的format為: “CLASS\_NAME/npy/THE\_GENDER\_AND\_COUNT/POSITION/TIME/FILE\_NAME”
3. 在”val” and “test”中, Gender and count這欄只會出現**單一男性**, 或是**單一女性**, 這在requirement 2,3,5都很重要.
4. 只有在”train”內, 會出現”Env(i)”的Class name, 其餘都是”val\_set” or “test\_set”.
5. Time中的format為 “date”\_”time”, eg. 240509\_111208(2024/05/09 11:12:08), 這種format會讓我們在做requirement 4,5 的時候很好判斷.
6. 承5. 如果利用判斷數字”240509” or “111208”, 會出現一種錯判是數字有機率會出現在後面的filename, 在觀察時沒有觀察出一個特定的規律, 所以認定是random number所組成的filename.

Step 2. Rewrite Class

在設計Class CSI\_Dataset時, 有先去查閱一下torch dataset and dataloader運作, 發現基本上Dataset都要在這個Class先處理完,就直接拿來做訓練, 所以我將所有data clear的function都放在這個Class內去做, 並讓它變成通用的.

以下會分幾個重點來說: (**BTW 我怕太多PDF放不下, 所以code都有寫註解**)

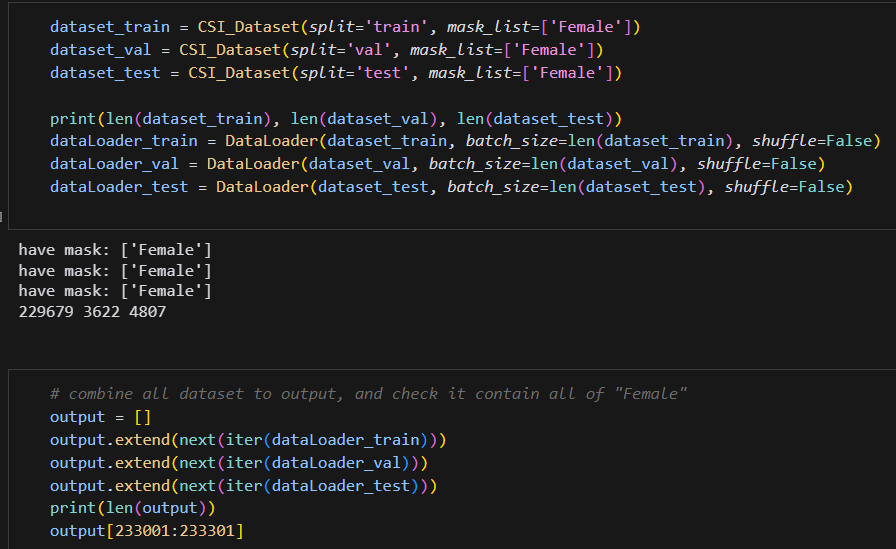
1. 在做mask時, 原先想法是每個符合mask list就利用extend丟進去mask\_data, 但是會造成data會有**重複情況**發生.eg. mask\_list = [‘F1’, ‘F2’, ‘F3’]
2. 承1. 之後修改成只有完整符合所有mask\_list的format才算.
3. 接著是gendercount, 就是簡單去判斷string內, 所含的字母”F” or “M”個數, 就可以拿出我們需要的資料. 那這次資料集不會在其他地方出現”F” or “M”, 所以就沒有做很嚴謹的處理.
4. Time compare一樣是利用time\_list傳入起始和結束時間, 然後判斷時間有沒有在範圍內, 這是可以解決我們剛剛講到date or time number會出現在filename中的問題.
5. 承4. 我利用spilt方法把format分段, 取出我想要的date and time. 因為requirement中, 會出現**開始結束的日期是同一天**, 所以我有稍微修改一下判斷式, 先判斷這個例外狀況. 接著判斷**相同日期不同時間**, 再來才是**超過開始日期但是還未到結束時間**, 這種只要判斷date, 不用判斷time.

Step 3. Thinking Requirement

分別對每個Requirement做不同的Constructor 傳入參數的設定.(相關function介紹都在step 2).

**Requirement 1.** 因為只需要Class name, 所以只需要傳入mask\_list = [‘Env3’], 且這個class只有在”train”才有, 所以spilt = “train”.

**Requirement 2.** 因為要女生, 先利用mask\_list = [‘F’]把單一男性的拿掉, 在利用gender = ‘F’, num = 2,來拿到需要的data. 另外”val” and “test”只有單一性別, 所以只看”train”.

**Requirement 3.** 同理於Requirement 2.但不一樣的是只有單一女性, 所以mask\_list = [“Female”], 然後需要看”val” and “test”, 所以我採取分別將”train” “val” “test” 利用spilt取出後, 在透過output合併.

**Requirement 4**. 時間部分, “val” and “test”只有240509的資料, 所以只看”train”,time\_list = [“240506\_181307”,”240507\_232434”].

**Requirement 5.** 同理1 and 2, 只看”train”然後只有一個男性, 所以只要看”Male(i)”, 利用上述修改過的mask特性, 直接設定mask\_list = [“Env3”, “5\_posi”,”Male”], 由於這邊設定Male當mask, 所以不會有其他女性在內,因此不用設定gender and num. time\_list = ['240508\_090000', '240508\_110000'].