

DLHLP HW4-1 Report

組長 Github ID: openopentw

組員 (姓名 + 學號): 鄭淵仁 R08922067

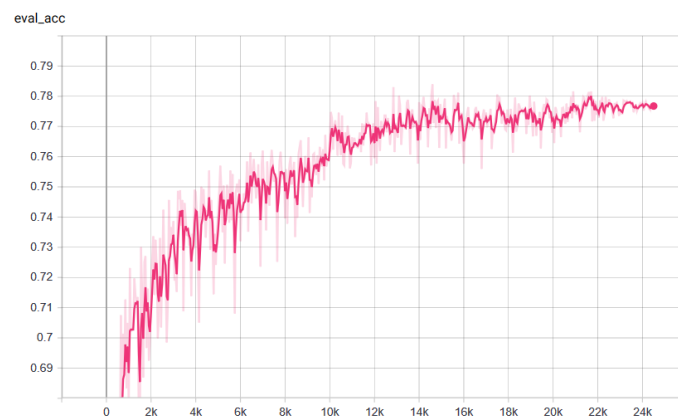
Part 1

1 請將 **training** 的 **loss curve** 跟 **test accuracy** 截圖放在下面，並簡單交待一下你的 **training** 過程

1.(a) **loss curve**



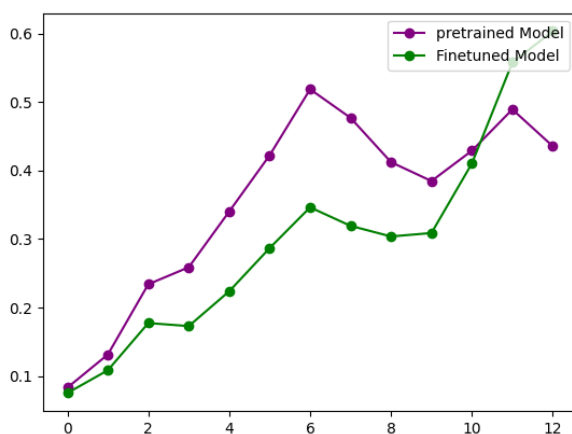
1.(b) **test accuracy**



1.(c) training 過程

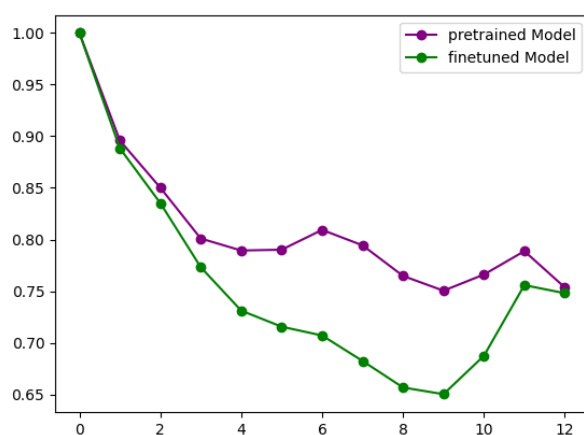
根據助教給的 Github Repository，先安裝好 Nvidia-Apex 等套件，再跑 `run_xnli.sh`，最後從 tensorboard 把圖片截圖下來。

- 2 請根據 Anisotropy 的定義以及助教提供的範例圖示，畫出 0 - 12 層各層的 Anisotropy 數值所連起來的線，每張圖片需要包含 pretrained-model 的版本跟 fine-tune 的版本，pre-trained model 的版本請用紫色線畫出來，finetune model 的版本請用綠色線畫出來

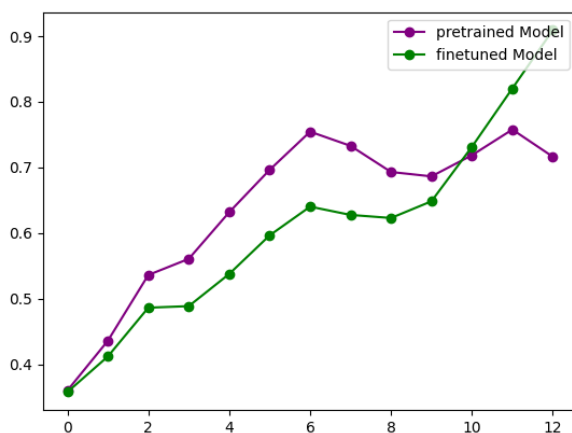


- 3 請根據 self-similarity 跟 intra-sentence similarity 的定義以及助教提供的範例圖示，畫出兩張圖片，一張是 0 - 12 層各層的 self-similarity 數值所連起來的線，一張是 0 - 12 層各層的 intra-sentence similarity 數值所連起來的線，每張圖片，都會有兩條線，一條線是 pretrained model 的版本，另一條線是 finetune model 的版本，pretrained model 的版本請用紫色線畫出來，finetune model 的版本請用綠色線畫出來

3.(a) self-similarity

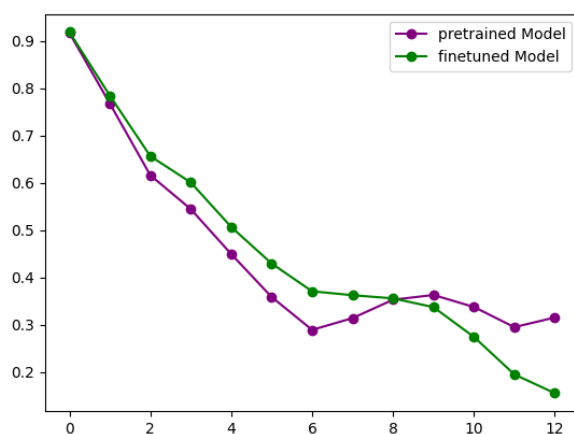


3.(b) intra-sentence similarity

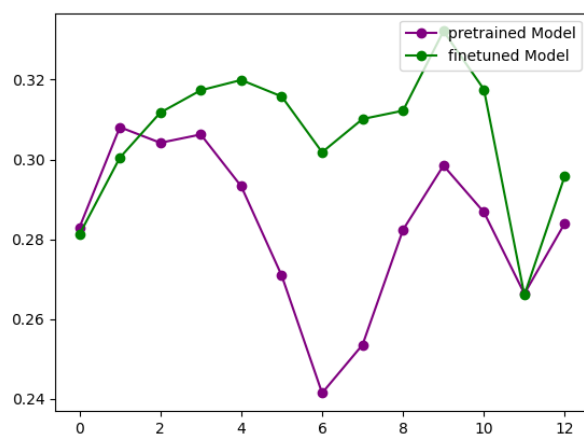


- 4 請把 self-similarity 以及 intra-sentence similarity 有減去各層的 anisotropy 的圖 (adjust version) 畫在本題 (配色如同第 2 題並標示清楚)，並且比較一下兩張圖和前一題做出來的兩張圖的差異，試著解釋一下 finetune 前後的變化

4.(a) self-similarity 減去 anisotropy



4.(b) intra-sentence similarity 減去 anisotropy



Part 2

1 Segment these sentences

- (1) 我，一直，親自，指揮，親自，部署，我，相信，只要，我們，堅定，信心，同舟共濟，科學，防治，精準，施策，我們，一定，會，戰勝，這，一，次，疫情。
- (2) 這，個，聲明，讓，我，再次，想起，了，安徒生，的，童話，《皇帝，的，新裝》，。
- (3) 希望，他們，能夠，聽，一，聽，這，個，忠告，不，要，再，信口雌黃，地，抹黑，居心叵測，地，挑撥，煞有介事，地，恫嚇。
- (4) 有關，部門，當然，就，是，有關，的，部門，了，無關，的，就，不，能，稱為，有關，部門，所以，我，建議，你，還是，要，向，他們，詢問。
- (5) 不，要，搞，奇奇怪怪，的，建築。
- (6) 現在，提請，表決，同意，的，代表，請，舉手，請，放下，不，同意，的，請，舉手，沒有，棄權，的，請，舉手，沒有，通過！
- (7) 人均，國內，生產，總值，接近，八千萬，美元。
- (8) 我，青年，時代，就，對，法國，文化，抱有，濃厚，興趣，法國，的，歷史，哲學，文學，藝術，深深，吸引，著，我，讀，法國，近現代史，特別是，法國，大，革命史，的，書籍，讓，我，豐富，了，對，人類，社會，政治，演進，規律，的，思考，讀，孟德斯鳩，伏爾泰，盧梭，狄德羅，聖西門，傅立葉，薩特，等，人，的，著作，讓，我，加深，了，對，思想，進步，對，人類，社會，進步，作用，的，認識，讀，蒙田，拉封丹，莫里哀，司湯達，巴爾扎克，雨果，大，仲馬，喬治·桑，福樓拜，小，仲馬，莫泊桑，羅曼·羅蘭，等，人，的，著作，讓，我，增加，了，對，人類，生活，中，悲歡離合，的，感觸，冉阿讓，卡西莫多，羊脂球，等，藝術，形象，至今，仍，栩栩如生，地，存在，於，我，的，腦海，之中，欣賞，米勒，馬奈，德加，塞尚，莫內，羅丹，等，人，的，藝術，作品，以及，趙無極，中，西，合璧，的，畫作，讓，我，提升，了，自己，的，藝術，鑑賞，能力，還有，讀，凡爾納，的，科幻，小說，讓，我，的，頭腦，充滿，了，無盡，的，想像。
- (9) 輕關，易道，通商，寬衣。
- (10) 因為，我，那，時候，扛，兩百，斤，麥子，十，里，山路，不，換肩，的。

2 從助教給的例子中，我們會發現機器遇到標點符號時必定預測其 label 為 S。如果去除標點符號，是否會對句子的 segmentation 造成影響呢？請對上述 10 句的無標點符號版本進行 segmentation，並敘述你的觀察。

- (1) 我，一直，親自，指揮，親自，部署，我，相信，只要，我們，堅定，信心，同舟共濟，科學，防治，精準，施策，我們，一定，會，戰勝，這，一，次，疫情
- (2) 這，個，聲明，讓，我，再次，想起，了，安徒生，的，童話，皇帝，的，新裝

- (3) 希望, 他們, 能夠, 聽, 一, 聽, 這, 個, 忠告, 不, 要, 再, 信口雌黃, 地, 抹黑, 居心叵測, 地, 挑撥, 煞有介事, 地, 恫嚇
- (4) 有關, 部門, 當然, 就, 是, 有關, 的, 部門, 了, 無關, 的, 就, 不, 能, 稱為, 有關, 部門, 所以, 我, 建議, 你, 還是, 要, 向, 他們, 詢問
- (5) 不, 要, 搞, 奇奇怪怪, 的, 建築
- (6) 現在, 提請, 表決, 同意, 的, 代表, 請, 舉手, 請, 放下, 不, 同意, 的, 請, 舉手, 沒有, 棄權, 的, 請, 舉手, 沒有, 通過
- (7) 人均, 國內, 生產, 總值, 接近, 八千萬, 美元
- (8) 我, 青年, 時代, 就, 對, 法國, 文化, 抱有, 濃厚, 興趣, 法國, 的, 歷史, 哲學, 文學, 藝術, 深深, 吸引, 著, 我, 讀, 法國, 近現代史, 特別是, 法國, 大, 革命史, 的, 書籍, 讓, 我, 豐富, 了, 對, 人類, 社會, 政治, 演進, 規律, 的, 思考, 讀, 孟德斯鳩, 伏爾泰盧梭, 狄德羅, 聖西門, 傅立葉薩特, 等, 人, 的, 著作, 讓, 我, 加深, 了, 對, 思想, 進步, 對, 人類, 社會, 進步, 作用, 的, 認識, 讀, 蒙田拉封丹, 莫里哀司湯達巴爾扎克, 雨果, 大, 仲馬, 喬治桑福樓拜, 小, 仲馬, 莫泊桑, 羅曼羅蘭, 等, 人, 的, 著作, 讓, 我, 增加, 了, 對, 人類, 生活, 中, 悲歡離合, 的, 感觸, 冉阿讓卡西莫多, 羊脂球, 等, 藝術, 形象, 至今, 仍, 栩栩如生, 地, 存在, 於, 我, 的, 腦海, 之中, 欣賞, 米勒馬奈德加塞尚莫內羅丹, 等, 人, 的, 藝術, 作品, 以及, 趙無極, 中西合璧, 的, 畫作, 讓, 我, 提升, 了, 自己, 的, 藝術, 鑑賞, 能力, 還有, 讀, 凡爾納, 的, 科幻, 小說, 讓, 我, 的, 頭腦, 充滿, 了, 無盡, 的, 想像
- (9) 輕, 關, 易道, 通商, 寬衣
- (10) 因為, 我, 那, 時候, 扛, 兩百, 斤, 麥子, 十, 里, 山路, 不, 換肩, 的

2.(a) 我的觀察

如果去除標點符號, 是會對句子的 segmentation 造成影響的。上述 10 句中, (8) 和 (9) 的 segmentation 就不一樣, 除此之外都相同。

其中 (8) 有許多人名, 例如「伏爾泰盧梭」或是「傅立葉薩特」等等, 在沒有標點符號時會被接在一起, 但其實是兩個人名。我想可能是因為這些人名的翻譯在訓練的語料庫中沒有很常出現, 所以不會被斷得很好。

另外, (9) 的「輕關」在沒有標點符號時會被斷開, 我想可能是因為語料庫中大多是現代的中文, 因此在古文方面的斷詞會不穩定。