

Dlhlp homework 4 - 1

學號：r08944008 系級：網媒所碩一 姓名：簡 義

學號：r07922104 系級：資工所碩二 姓名：林傳祐

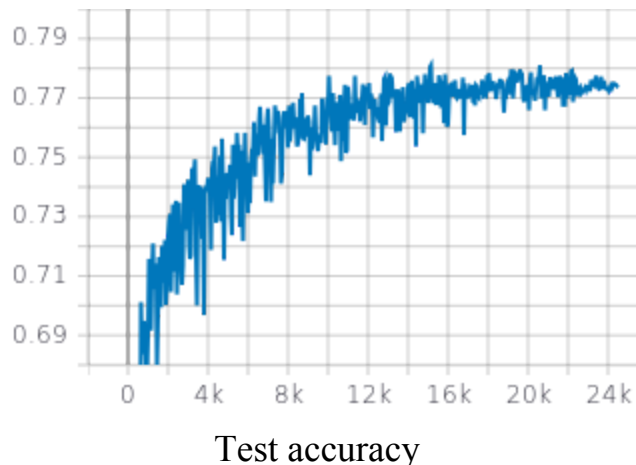
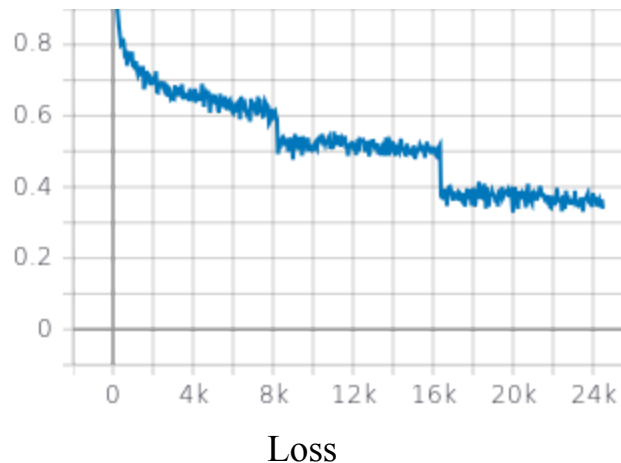
學號：r08944024 系級：網媒所碩一 姓名：陳品媛

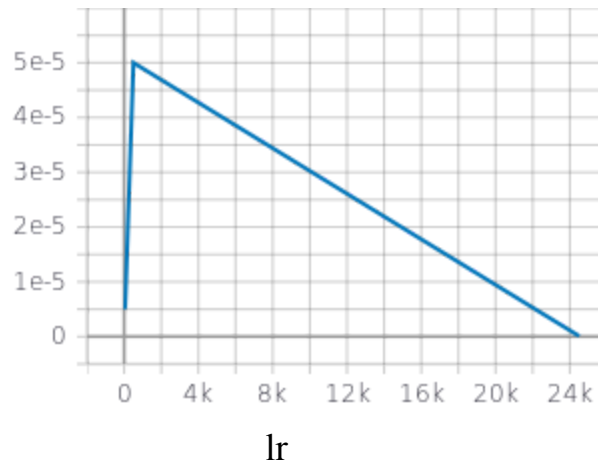
學號：r06922089 系級：資工所碩二 姓名：邱淳浩

Part 1:

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

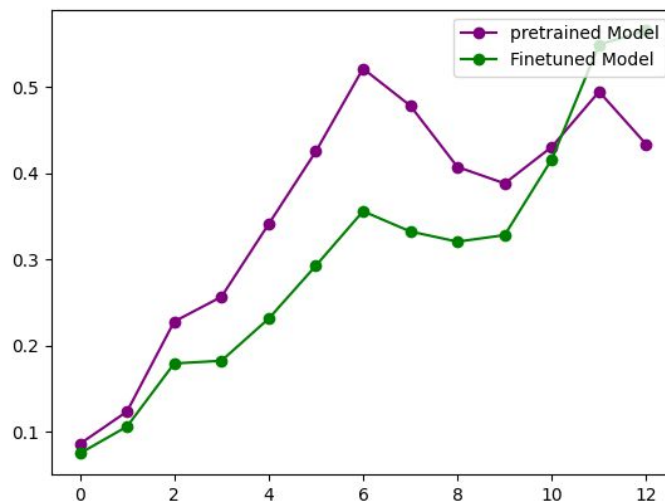
Ans: 在training中使用到了warm up的機制，讓learning rate先緩慢提升後再做decay，有一些實驗指出這樣對於模型的訓練會更穩定，並且比較容易收斂到較好的地方。Training過程中每過約8k個iteration，loss就會大幅的下降，可能表示model在優化的過程中剛好通過了plateau的部分。





2.(1%) 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，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

Ans:

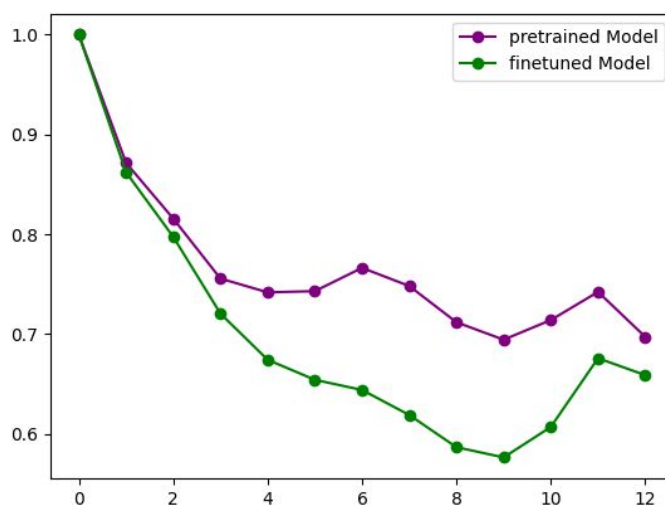


Anisotropy

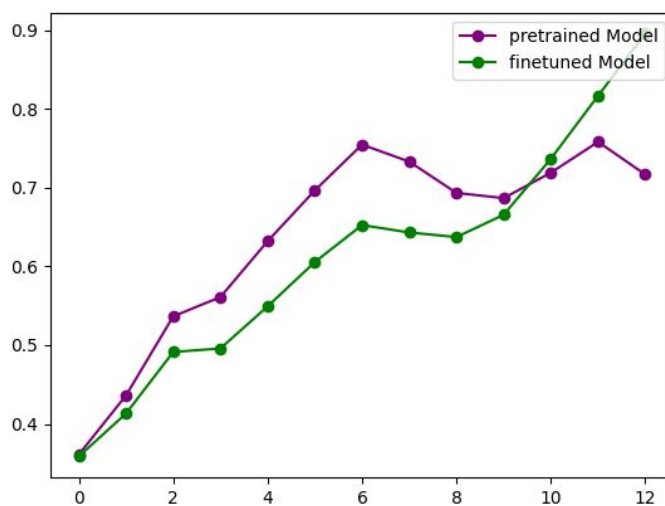
3.(1%) 請根據 self-similarity 跟 intra-sentence similarity 的定義 以及 助教提供的範例圖示，畫出兩張圖片，一張是 0 - 12 層 各層的 self-similarity 數值所連起來的線，一張是 0 - 12 層各層的 intra-sentence similarity 數值所連起來的線

，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

Ans:



self-similarity



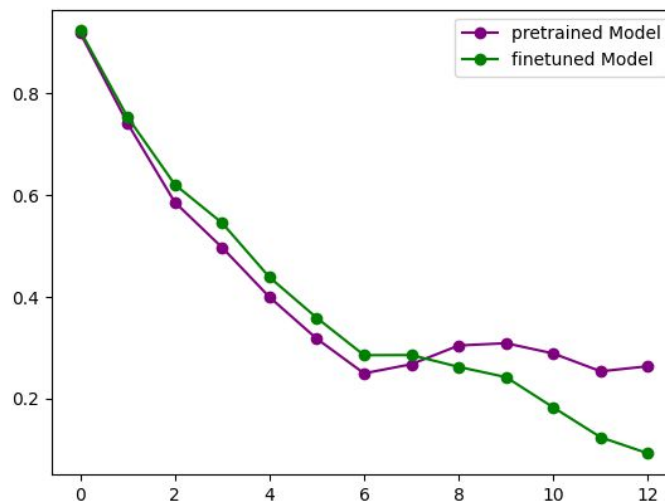
intra-sentence similarity

4.(2%) 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

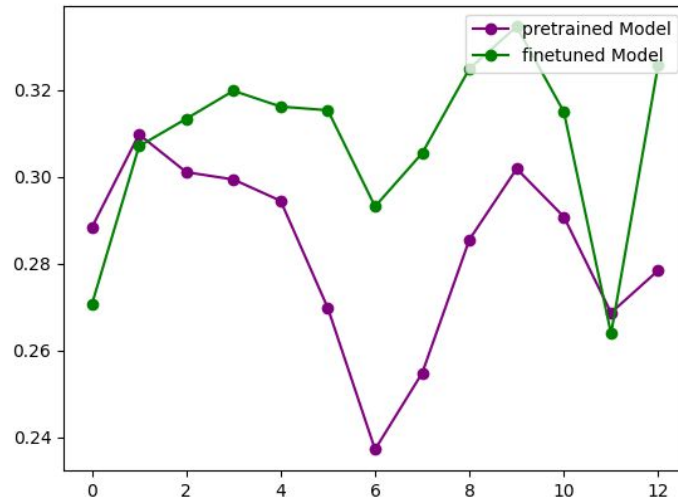
Ans:

在 self-similarity 的部分，減去 anisotropy 後值有明顯比較低，而 finetuned 後的 model 又比 pretrained 的更低，表示同樣 word 在不同的句子中，model 有成功 learn 到不同的表示。

在 intra-sentence similarity 的部分，還沒有減去 anisotropy 前，不論是 finetune 和 pretrained 的 intra-sentence similarity 隨著 model 層數越來越高，表示在同一個句子中，word 的表示都很相近。而在 adjust 過後，不同層間的 similarity 比較接近，而且 finetune 的值大多比 pretrain 的高，在同一個句子 word 有比較接近的表達，表示 finetune 後的 model 效果的確有比較好。



Anisotropy-adjusted-self-similarity

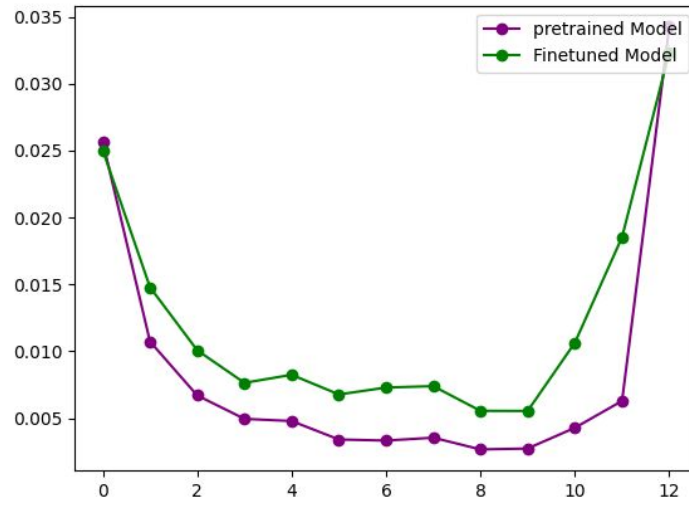


Anisotropy-adjusted-Intra-sentence-similarity

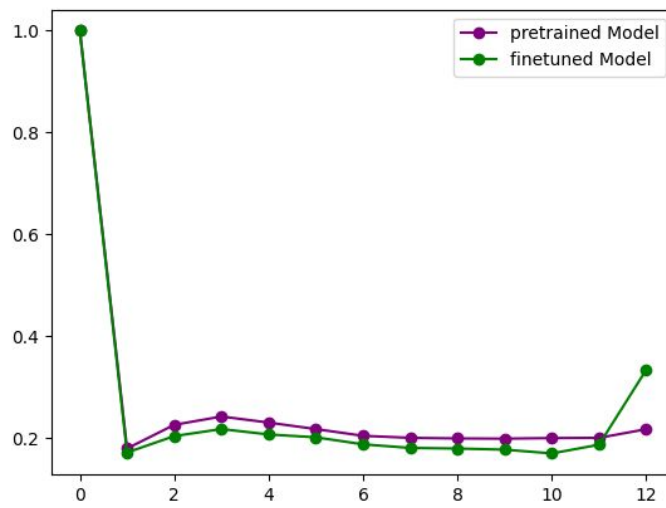
5. Bonus (3%): 請實做一種比較層與層之間 embedding 相似程度的方法，像是在該 [paper](#) 內還有第三種 metric MEV(Maximum Explainable Variance) 可以實作並如 2, 3, 4 題的要求把 MEV 的 anisotropy, MEV, adjust anisotropy MEV 在不同層的變化，畫出來 (finetune model 和 pretrained model 都要)，或是實做像是 [SVCCA](#), [PWCCA](#) 等衡量embedding 相似程度的方法，清楚標明你做的的方法，並簡單的說明一下你的實做細節

Ans:

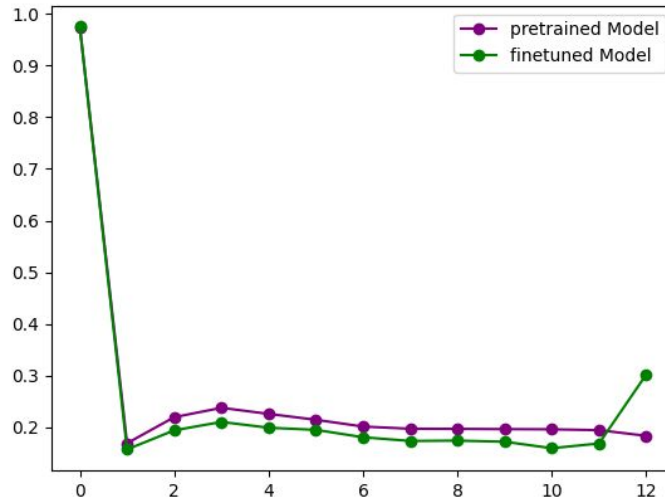
使用的方法為MEV，實作根據code的提示，MEV使用components=1的PCA，並計算最大singular value平方佔所有singular value平方和的比例。MEV 的 anisotropy則對sample的10000筆data計算components=100的TruncatedSVD，並一樣計算最大singular value平方佔所有singular value平方和的比例。



MEV_Anisotropy



MEV



Adjusted-Maximum-explainable-variance

Part 2:

1. (3%) Segment these sentences (有缺字該句不算分) :
 - (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) 輕關易道，通商寬衣。
 - (8) 人均國內生產總值接近八千萬美元。
 - (9) 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。

(10) 因為我那時候，扛兩百斤麥子，十里山路不換肩的。

Ans:

Segment result:

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. (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。

Ans:

無標點符號版本Segment result:

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
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

去除標點符號並不會對斷詞結果造成太多影響，大部分結果還是相同的，因此可能推測model有學到斷句的資訊。