DLHLP HW4-1 Report

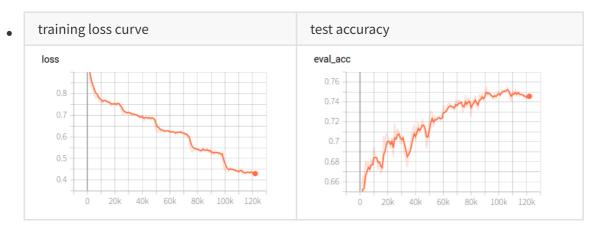
組長 Githud ID: lou-tun-chieh

學號:b05902111系級:資工四 姓名:婁敦傑

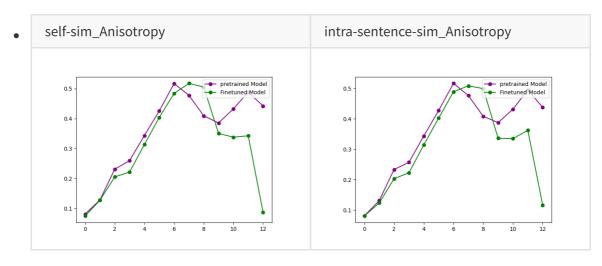
學號:b05902010系級: 資工四 姓名:張頌平

Part1

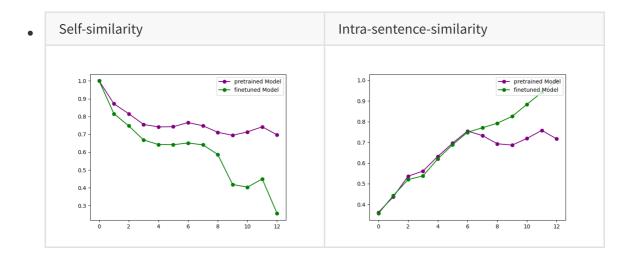
1. 請將training 的 loss curve 跟 test accuracy 截圖 放在下面,並簡單交待一下你的training 過程



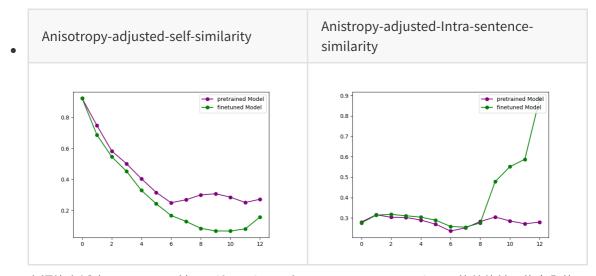
- 我們使用 bert-base chinese pretrained model,再多做 5 個 epoch 的 finetune,就能達到 75 % 左右的正確率,參數上有調整 batch size = 8
- 2. (1%) 請根據 Anisotropy 的定義 以及助教提供的範例圖示,畫出 0 12 層各層的 Anisotropy 數值 所連起來的線 ,每張圖片需要包含pretrained-model的版本跟fine-tune的版本,pretrained model的版本請用紫色線畫出來,finetune model的版本請用綠色線畫出來)



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 intrasentence similarity 數值所連起來的線.



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



- 在還沒有減去 anisotropy 前,self-similarity 和 intra-sentence similarity 的數值並不能完全的反應 context-specific,就如投影片提到的 Self-Similarity(w) = 0.95 is relatively high if all embeddings are isotropic but relatively low if they are anisotropic。未減去 anisotropy下, self-similarity 在第十二層才調到 0.3 之下,但在減去 anisotropy 後,finetune 的 self-similarity 在第六層後幾乎都在 0.2 下,代表其實在第六層就已經有學到同個字在不同句子間的不同 embedding;而對於 intra-sentence similarity來說,在還沒減去 anisotropy時,第十和十一層 的intra-sentence similarity皆超過0.8,和第十二層的結果相去不遠,但在減去 anisotropy 後, 我們可以發現第十二層的表現是遠勝於前面所提到的第十和十一層。從這兩張圖前後的差異,我們 可以知道透過減去各層的 anisotropy,可以幫助我們更全面地探討model是否有真的學習到 context-specific的資訊,以免因為embedding所在空間較小而導致出現的相似度和實際的效果有 所落差。
- finetune 前後可以明顯看出 finetune 後的結果都更加地好,尤其是在後面幾層 layer, self-similarity 越低和 intra-sentence similarity 越高,此外, pretrained model 的 intra-sentence similarity 的數值都是偏低的,這是因為 pretraind model 可能沒有遇過類似我們測試的句子,但經過 finetune 後 intra-sentence similarity 就會有顯著的提高。

Part2

1. (3%) Segment these sentences (有缺字該句不算分):

這,個,聲明,讓,我,再次,想起,了,安徒生,的,童話,《,皇帝,的,新裝,》,。

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,。

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,。 不,要,搞,奇奇怪怪,的,建築,。

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,!

人均,國內,生產,總值,接近,八千萬,美元,。

輕關,易道,,,通商,寬衣,。

因為,我,那,時候,,,扛,兩百,斤,麥子,,,十里山路,不,換局,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。

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 這,個,聲明,讓,我,再次,想起,了,安徒生,的,童話,皇帝,的,新裝

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 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

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

● 人均,國內,生產,總值,接近,八千萬,美元

不,要,搞,奇奇怪怪,的,建築

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,伏爾泰區梭狄德羅,聖西門傳立葉薩特,等,人,的著作。讓,我,加深了了對,思想,維步,對,人類,社會,推步,作用,的。認識,讀,竟田拉封丹,莫里哀司,湯達巴爾扎克,兩果,大,中馬、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,悲歌離合,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,趙無像,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

輕關,易道,通商,寬衣

因為,我,那,時候,扛,兩百,斤,麥子,十里山路,不,換局,的

從將標點符號去除的這張圖和前一張圖比較可以發現,在大部分的例子中,有無將標點符號去除對於斷詞是沒有影響的,這也證明了BERT的強大,不過,對於專有名詞(也可以說是較少被看過的詞語),像是第九行範例中的英譯人名,大部分就都被串在一起,或是被機器用一般句子的語法來切斷(例如:大仲馬->大,仲馬),因此,加上標點符號仍然是有助於我們對於切詞這個任務表現得更好!