中華大學生物資訊學系系統開發專題報告

手寫繁體中文字影像辨識與其訓練策略比較

Image Recognition of Handwritten Mandarin Chinese and Comparison of Model Training Strategies

專題組員:李翰威、傅聖修、林哲緯

專題編號：107011

指導老師:董其樺老師、張慧玫老師

一、摘要

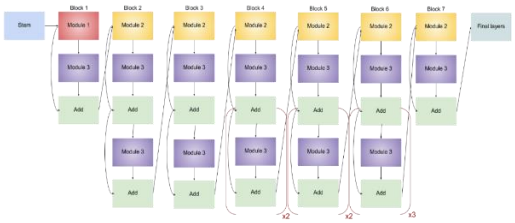
本專題為組員競賽作品之延伸，該作品為一卷積神經網路模型，用於分類主辦方指定的800個中文字之手寫中文圖像，將針對不同優化器與不同訓練資料集對於模型準確度的影響，以及模型對於主辦方所使用之資料集與其他來源之手寫中文圖片的準確度做比較，以此分別評估模型對於比賽需求的適用性與模型實際泛用性。

關鍵詞:手寫中文辨識、卷積神經網路、資料增強、參數優化器

二、簡介

近年來影像辨識技術快速發展，然而在這個資訊爆炸的時代，為了更有效率的創造更大的利益與競爭價值，在有限的資源下有效的分析資訊並轉化成知識與剃除雜訊是越來越迫切的需求，對手寫字的辨識也是在目前需求下積極發展的領域。

從2014年Network in Network[1]與2015年 ResNet[2]等強調減少參數、收縮尺度、簡化與合併架構的模型問世，開啟了最佳化資源分配與特徵擷取以優化模形效益的新思維，當前佔據 ImageNet 榜首的 EfficientNet[3]系列，便是這個領域的佼佼者，其承襲與 MnasNet[4]用來優化延遲時間(latency)相同的搜索空間(search space)，來尋找最佳運算量 (FLOPS)，藉此找出模形架構的最佳比例，而本專題所使用的模型基線架構便是該系列的 EfficientNet-B1(圖一)，其有效利用運算資源的效果，能在不犧牲準確度的情況下大福縮減訓練與計算時間，由於主辦方要求每筆資料要在一秒內回傳結果，因此成為首選。



圖一：EfficientNet-B1架構

本專題為組員於<<玉山人工智慧公開挑戰賽 2021 夏季賽>>[5] 之作品的延伸，原參賽之作品直接使用主辦方提供的原始資料集，並未實際對資料集做整理，因而決定探討經過整理之資料集是否能提升模型表現，並同時探索更為適合的優化器以產出最佳模型。

原作品 Adam_old 模形之設置：將資料集餵入 EfficientNet-B1，並使用 Adam 優化器(Optimizer)優化參數，學習率(Learning Rate)設為0.00005，損失函數(Loss Function)為交叉熵(Cross-Entropy)，訓練30代(epoch)，批次大小(batch size)為16，將圖檔 reshape 成240X240，訓練(train)與驗證(validation)資料比為3：1，資料餵入時隨機旋轉20度、隨機剪裁(shear)20度、隨機縮放0.1倍，產出 Adam_old 模型，剩餘細節將於後續段落補述。

三、專題進行方式

(一)、資料集來源與處裡

主要資料由比賽主辦方玉山銀行所提供的約68000張大小不一的中文手寫照片，可分類成800個不同的中文字，經過編輯與處裡，衍生為以下不同的資料集。

1、Old：未經任何處裡的原始資

料集。

2、Best：依照圖檔名稱分類成800個不同的字，如：[10003_琪.jpg]，再以人工清理資料，將無法辨識、包含多於一個完整的字、損失一半以上字體的圖片捨棄，而可辨識、僅包含一個完整字但標籤錯誤的字則重新標籤並分類，最後每個字各移出一張圖片，形成共800張圖片的獨立測試資料集 splited_800，Best訓練資料集。

(二)、第一次模型訓練與測試：

將 Best 訓練資料集套入與 Adam_old 相似的配置，訓練30代，批次大小改為30，產出的 Adam_best 模型之訓練驗證準確度高達0.9577，使用 splited_800做獨立測試(independent test)準確度則有0.98，然而 splited_800對於 Adam_Old 來說並非獨立測試，因此無法比較。

(三)、主辦方的額外圖片 S2:

回頭檢視比賽過程，發現一份意外擷取的賽前測試資料集 S2，總共有七千張無命名與標籤，且藍紅 Channel 相反而無法直接使用的圖片。

(四)、S2轉色與重新標籤：

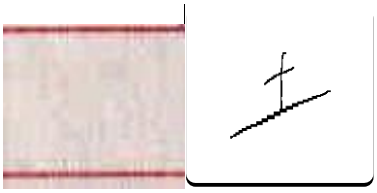
將圖片的藍紅 Channel 對調之後，丟入 Adam_best 預測以分類這些圖片，再人工清理分類結果。

(五)、資料回填 生成 S2fix:

為了填補資料分布，將 splited_800放回 best 資料集，並將132張 S2資料用於填補 Best 中數量較少的字，形成新的訓練資料集 S2fix。

(六)、合成圖片：

使用合成圖片的方式來擴充並且填平資料分布，使用開源手寫中文字圖片[6]，然而其空白的背景**(圖二右)**與原始圖片之背景非常不符，可能影響模型表現，於是用 Old 資料集中只有背景之圖片**(圖二左)**與開源手寫中文字合成新圖片，以填充資料分布。



圖二：空白背景(左)開源手寫字圖片(右)

(七)、圖成片合成的方法[7]：

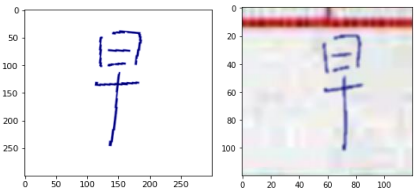
合成圖片分成兩個部分，開源字體的截取與調整、合成與模糊。

(八)、開源字體的截取與調整：

截取目標800字，字體原先為黑色的**(如圖二右)**，而原始資料當中字體大多數都是藍色，因此先把字體調整成藍色使資料更貼近需求，並且微幅的隨意旋轉或位移避免過度擬合**(圖三左)**。

(九)、背景調整與合成：

將背景圖與藍色字體圖片轉化為矩陣後重疊，並選取相同像素上數值較小者合併成新圖片，再透過高斯模糊(Gauss blur)篩去雜訊與合成邊際，避免模型在訓練時將「是否是合成圖」當成重要特徵**(圖三右)**。



圖三：調色手寫圖片(左)，最終合成圖片(右)

(十)、回填資料 生成 blueFB：

將合成好的圖片人工清理後回填到 S2fix 資料集以填補資料分布，成為新的 blueFB 訓練資料集，是資料分布最平均的。

圖四：不同訓練資料集的形成

(十一)、模型參數與權重：

載入 EfficientNet 作者事先於 ImageNet 訓練後的參數與權重，再將輸出層(Top-Layer)設置為分成800類的全連接層，再開始餵入資料以訓練模型的參數權重。

(十二)、Ranger optimizer：

Ranger[8]優化器為 RAdam[9]以及 LookAhead[10]兩個優化器的結合，利用 RAdam 調整 learning rate 並且利用 LookAhead 來調整模型收斂的方向，避免模型收斂至區域最小值(local minimum)，以及改善梯度消失(gradient vanishing)的問題同時加速模型的訓練，本專題使用建議超參數。

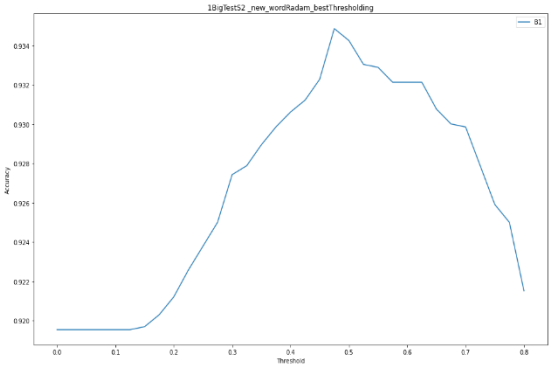
(十三)、獨立測試資料集：

BigTestS2：由 S2資料集組成的測試資料集，共6587張照片，其中405張為不屬於800個目標字內並標記成 isnull 的生字圖片。

blueFB_test：由填補資料集後剩下的合成圖片組成，並無生字。

(十四)、最佳閾值(thresh-old)：

測試模型認定生字的能力，將模型預測結果之機率低於閾值者，其預測結果改為 isnull，再透過改動閾值**(圖五)**以找到使每個模型準確度最高的閾值(表一)。



圖五：Adam_best 的閾值-準確度曲線

表一

N:Modl name, A:Accuracy, T:Threshpld

N	Adam_old	Adam_best	Adam_blueFB	Adam_S2fix
A	0.89874	0.934872	0.909974	0.873235
T	0.4	0.475	0.5	0.4
N	Ranger_old	Ranger_best	Ranger_blueFB	Ranger_S2fix
A	0.888872	0.924397	0.920753	0.912707
T	0.325	0.475	0.45	0.425

四、主要成果

(一)、最佳模型：

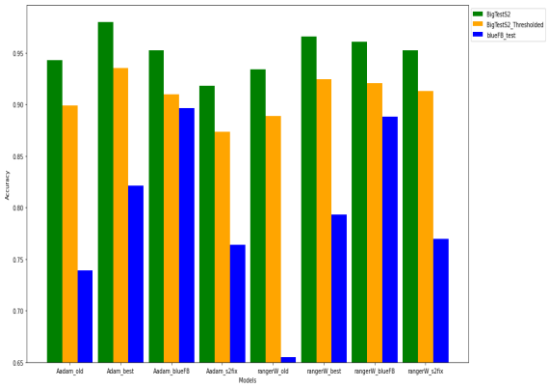
Adam_best，學習率為0.00005，於第26代降為1.5811387e-05，BigTestS2 不含生字與閾值獨立測試準確度 0.9796，含生字並套上閾值，準確度 0.9348。

(二)、最佳訓練資料集：

在兩種優化器中，表現最好的皆為 best 資料集所訓練的模型。

(三)、優化器比較：

針對主辦方資料集，Adam 能夠產出最好的模型，而 Ranger 產出的模型差異性最小，甚至顯著提升使用 Adam 表現較差的資料集所產出之模型表現，而透過 blueFB_test 表現出的泛化能力上，Adam 明顯能夠保持較好的表現。



圖六：BigTstS2準確度(綠)、BigTstS2閾值準確度(橘)、blueFB 準確度(藍)

(四)、BigTestS2 錯字分布：

依照不同訓練資料集或優化器分組觀察共同分類錯誤的圖片，分成 old(308 張)、best(179 張)、blueFB(211 張)、S2fix(276 張)、Adam(63張)、Ranger(87張)與全模型共同測錯總共7組。

看得出來不同組別間確實有明顯

的錯誤分布區別，也具有相似的趨勢**(圖七)**，圖七中的橫軸為所有被分辨錯誤的圖片，依照類別分類後排序，屬於同一個字的圖片會在同一個區間內，最左側的類別為 isnull，顯示所有模形對於生字的辨識能力並不高，且高度集中在相同圖片上，其中 Rager 組(綠)的錯誤比 Adam 組(青)更集中於相同的圖片上，best 組與 blueFB 的錯誤分部有較明顯的差異，代表兩者善於辨識的特徵較為不同，old 組(紅)明顯有特別能辨識的字與特別無法辨識的字，S2fix 組(黃)是共同錯誤分布最分散的，表示其對每個字的辨識能力差異性較小。

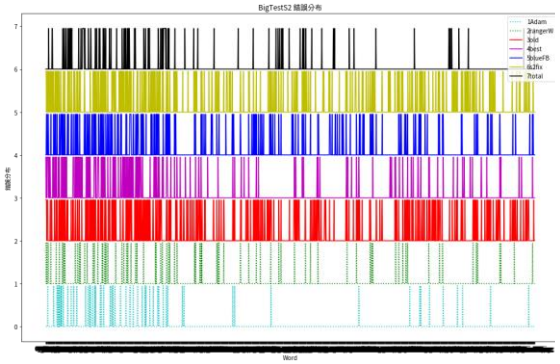
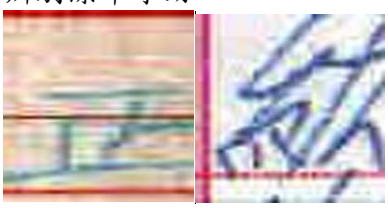


圖 七：每組共同測錯圖片在所有測錯圖片之分布，最頂層(黑)為全體模形共同測錯。

實際細看全體模形共同測錯之圖片，發現所有模形對於淡化、字體比例傾斜、筆畫粗細不一的圖片辨識度較差**(圖八)**，best 訓練資料集對於淡化、比例傾斜、筆畫粗細不一較為敏感，blueFB 資料集則是較無法判別潦草的字跡，Adam 組也容易在淡化、比例傾斜、筆畫粗細不一的圖片上出錯，Ranger 組也比較無法辨別潦草字跡。



圖八：共同錯誤之圖片

五、評估與展望

回顧比賽結果，當初比賽時的第一名準確度高達0.99，本作品還有許多進步空間，未來希望調整優化器超參數，以找到更適合此問題的超參數配置，並釐清目前模形表現欠佳是否與模形收斂至區域最小值有關，以及使用更好的合成圖片，例如使用對抗生成網路訓練強大的圖片合成模型以產出最逼真的手寫圖片，並且探討如果加入合成的印刷字體會有何影響。

本專題遺留下最大的困惑是即便 best 資料集與 S2fix 差別甚小，卻不知道為何模形表現差異如此之大。

六、結論

對於玉山提供之目標資料集而言，我們發現影響最大的是資料集的差異，尤其資料集的乾淨程度最為重要，其次是訓練資料之分布狀態與訓練資料對需求的貼近程度，而優化器的選擇與超參數的調校則會進一步影響模形如何學習資料特徵，進而影響泛化性與適用性。

七、參考文獻

[1] Min Lin, Qiang Chen, Shuicheng Yan, Network In Network, arXiv:1312.4400, (2014).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385(2015)

[3] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In ICML, pages 6105-6114, 2019.

[4] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. MnasNet: Platform-aware neural architecture search for mobile. CVPR, 2019.
[5] 玉山人工智慧公開挑戰賽 2021 夏季賽：https://tbrain.trendmicro.com.tw/Competitions/Details/14
[6] The dataset is AI . FREE Team development from [STUST EECS_Chinese MNIST(總集)]. If used, modified, or shared, please cite the source and the mesage.
[7] T-brain 2021 夏季賽，中文手寫影像辨識：Top 4% Solution：https://github.com/KuanHaoHuang/tbrain-esun-handwriting-recognition
[8] Wright,L. and Demeure,N. (2021) Ranger21: a synergistic deep learning optimizer.arXiv:2106.13731 [cs].
[9] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
[10] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In Advances in Neural Information Processing Systems, pages 9597-9608, 2019