影響 CNN 中文古籍 OCR 辨識率的因素探討

蘇冠宇[1]　吳昱賢[1]　王祥安[2]
[1]元智大學資訊工程學系　　[2]中央研究院數位文化中心
[1]{watermelo0326, ymmh123456789}@gmail.com　　[2]sawang@gate.sinica.edu.tw

摘要

為加速古籍數位化，本研究透過卷積神經網路辨識古籍影像中的文字。我們訓練多個不同的卷積神經網路模型，比較不同字體是否會影響模型的辨識能力，也比較訓練資料集的數量對模型辨識能力的影響，根據不同的實驗結果，了解影響模型辨識率的原因。希望未來能開發出一個針對古籍文字辨識的系統，以加速古籍數位化的工作。

關鍵詞： 卷積神經網路、文字辨識、深度學習

1. 前言

Google 開發的 AlphaGo[1]在圍棋人機大戰中擊敗了人類，使深度學習技術越來越被重視，其中由 LeCun 提出的 Convolutional Neural Network（卷積神經網路，以下簡稱 CNN）[2][3]在圖像辨識、語音檢測、人臉識別、手寫辨識等領域被廣泛採用。

為了能加快古籍數位化及降低人工成本，本研究利用 CNN 技術，訓練出針對古籍文字辨識的神經網路模型，並嘗試了解影響不同模型辨識率的原因，已發展出最佳的影像文字辨識方法。

至今已有許多學者將深度學習運用在影像文字辨識的領域，然而他們的資料集多是來自 MNIST[4]及 CASIA-HWDB[5]，並沒有針對古籍文字辨識的研究，因此我們希望能夠利用中央研究院歷史語言研究所漢籍電子文獻資料庫[6]所蒐集的古籍資料，訓練出能夠針對古籍文字進行辨識的神經網路模型，未來希望應用此成果，讓不同字體都能被更準確的辨識出來。

2. 文獻探討

與中文影像辨識相似的是數字影像辨識，常見的作法有 Back-Propagation Network[7]、MLP-SVM[8]、CNN-SVM[9]等，在2012年的 ImageNet 競賽當中，以 CNN 為核心的 AlexNet[10]摘下冠軍，而之後有關 CNN 的模型也如雨後春筍般出現，且都有超過人類肉眼極限的辨識率。數字與漢字不同的是，漢字的數量可能性高達上萬種且字體細節繁雜，相較於阿拉伯數字難度更高。

對於中文影像辨識的相關論文：Dan Cireşam 與 Ueli Meier 所發表的[11]，利用合併多層 DNN 網路，達到接近人類辨識能力的準確率。然而他們使用的資料集多為網路上的公開資料庫－CASIA-HWDB，並沒有針對古文字體所設計的辨識系統，使用古籍影像當作測試集與他們最大不同的是，我們的影像是直接從古籍擷取下來的，因此影像中有許多雜訊，而前者的資料是較少雜訊的，因此本實驗的結果更接近於真實應用的情況。

在觀察多種神經網路系統後，我們使用 CNN 深度學習模型，用來訓練古籍影像文字辨識系統，因為 CNN 透過卷積有效地減少傳入隱含層的參數數量，減少模型訓練時間，也解決模型參數過多的問題；此外，CNN 相較於 DNN，可以透過前面的卷積層，有效地限制參數數量避免陷入局部最佳解的狀況。

3. 研究方法

實驗流程如圖1，我們將於3.1節說明蒐集資料的方法，3.2節說明如何將資料切割成訓練、測試與交叉驗證集，3.3節說明如何設計與訓練模型，並於3.4節中說明如何調整訓練模型的參數。

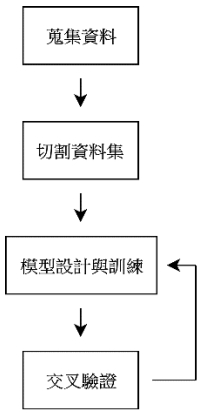

圖 1　實驗流程圖

3.1 蒐集圖片與文字資料集

為了實驗不同字體訓練出的模型是否對辨識率會造成影響，我們從漢籍電子文獻資料庫中，挑選兩本字體相異(一本印刷體與一本手寫字體)的古籍影像，分別為本草述鉤元及景岳全書，其字體如圖2所示。

我們從影像中切割出每個字元，並與文字檔做

交互參照，從中挑選常用字以及影像品質較佳的圖片，進而得到實驗的資料集，其中本草述鉤元共挑選了103,413字，景岳全書共72,783字。

圖2 左圖為本草述鉤元之字體，右圖為景岳全書之字體

3.2 切割資料集

為了避免訓練集與測試集使用相同的影像而失去實驗的公正性，因此我們將前述的影像檔案分成三部分，分別為：60%的訓練集，20%的測試集，以及剩餘20%的交叉驗證集；訓練集是用來訓練模型；測試集是最後用來驗證模型的正確率；交叉驗證集是模型評估指標，如果訓練集的準確率隨著訓練次數增加而提升，但交叉驗證集的準確率不變甚至於下降，此狀況說明了模型已經產生overfitting[12]，如果模型發生 overfitting 則會在測試集上的準確率與訓練集上的準確率差異過大，無法正確預測出真實答案此時應該要停止訓練；也就是說，交叉驗證集是可以避免訓練出的模型與特定的訓練集資料相依性太高，而無法適用於大部分的狀況[13]。

3.3 模型設計與訓練

我們參考 VGGNet 模型架構[14]，在此架構中他們使用最小的3*3 filter 以減少訓練參數並加快訓練速度，在通過多卷積層的特徵提取之後再用三個全連接層連接。在此研究中我們使用較大的5*5 filter，原因為原本 VGGNet 模型架構輸入圖片大小為28*28，而我們圖片大小為64*64，因此我們將一開始的 filter 放大兩倍，待於後面的卷積層才使用較小的3*3 filter 捕捉較多細節，並能有較快的訓練速度，而最後進行 softmax 分類時改成各自書本對應的字型數量。

為了避免 overfitting 的狀況發生，我們在網路當中加入 Dropout[15]，藉由隨機關閉隱藏層的節點權重，阻止對於某些特徵的過度依賴。當神經網路隨機丟棄一些特徵，神經網路必須學會如何從剩餘的特徵進行分類，而非透過特定的特徵進行分類。而[16]提到，在輸入層的 Dropout rate 設定為20%且隱藏層的 Dropout rate 設定為50%的實驗結果是最佳的，原本我們將輸入層的 Dropout rate 設定為30%，不過發現訓練效果並不如此設定，因此我們也採用了此設定。

3.4 交叉驗證

在訓練過程中，我們發現加入 Dropout 後，在每一次 epoch 中交叉驗證集的準確率會略高於訓練集的準確率，原因是在訓練模型時，加入 Dropout 會暫時不更新所關閉的節點權重，而在測試集中並不會關閉節點，即每一個節點皆會傳送訊息給下一層節點，因此交叉驗證集的準確率會略高於訓練時的準確率。而我們也在交叉驗證當中，挑選辨識率最高的模型，用以當作最終的辨識模型。

4. 實驗

為了評估 CNN(以下稱 CNN-OCR)是否適用於古籍文字辨識，我們同時比較了 Tesseract-OCR 及 ABBYY Cloud OCR 的辨識結果，其中 Tesseract-OCR[17]為 Google 所維護的開源文字辨識軟體，而 ABBYY[18]是 Adobe 公司的光學文字辨識軟體，我們於4-1節說明印刷體與手寫字體對於上述軟體與 CNN-OCR 比較的結果，4-2節探討不同字體所訓練出的模型對於辨識率的影響，4-3節探討增加資料集是否影響增加辨識率，4-4節探討混合不同字體來訓練模型是否有助提升於文字辨識率。

4.1 比較 CNN 與其他軟體的辨識率

印刷體測試集在 ABBYY、Tesseract-OCR 與 CNN-OCR 的辨識率分別為75.33%、74.73%及94.49%；在手寫字體測試集上的辨識率分別為54.23、68.92%及91.08%，實驗結果如圖3，從上述結果可以發現，不管在印刷體、手寫字體測試集上 CNN-OCR 的辨識率皆高於其他軟體約18%以上，由此可見 CNN-OCR 的方法相較於其他軟體更適用於古籍文字辨識。

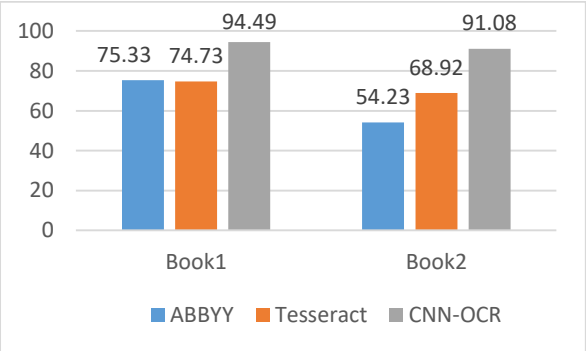

圖3 CNN-OCR 與其他軟體比較

4.2 不同模型的辨識率影響

為了探討不同字型所訓練出的模型，是否影響辨識率，我們將訓練出的印刷體、手寫字體模型，分別對印刷體、手寫字體測試集做交互測試，為了確保對於模型的公平性，我們挑出兩本書的共同字作為訓練集、測試集一共2058種類別，以便評估字體對於模型的影響。以下是我們的實驗結果，如圖

4，可以發現當模型與測試的字型不同時辨識率會明顯的下滑，因此可以得知資料集的字型，會大幅影響 CNN-OCR 的準確率。

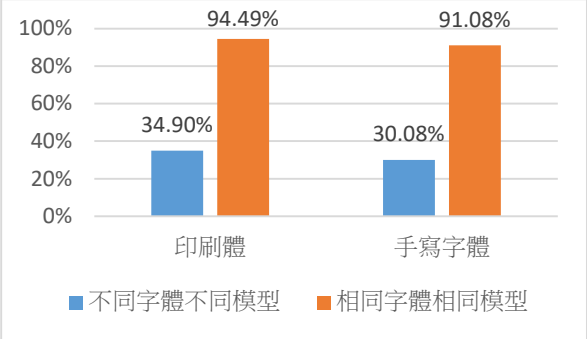

圖4 字體相同/相異對於模型影響

4.3 複製冷門字圖片增加訓練集

實驗中我們發現古籍中冷門字的圖片數量與熱門字的數量差異甚多，可能導致 CNN-OCR 將測試集中的冷門字預測偏向熱門字，因此我們複製冷門字圖片的類別數量為734以及976個類別，在總共2058個類別當中，複製將近一半類別訓練集的情況下來增加冷門字的圖片，使其與熱門字的數量相當來進行實驗。

在我們的實驗結果當中，如圖5，可以發現在印刷體上複製冷門字的效果並不顯著，而比較前後模型對於測試集的表現，雖然訓練出來的模型對於同一測試集的錯字有改善，但印刷體準確率卻只有提高約0.15%，而在手寫字體上也只有提高約1.5%。

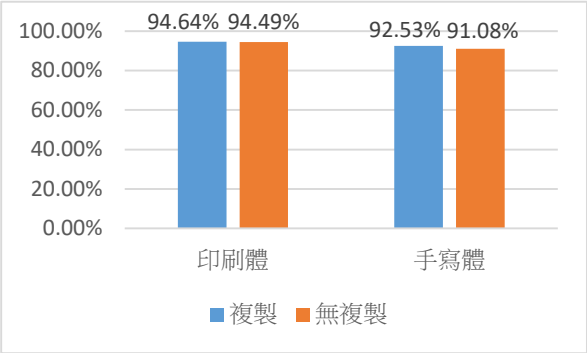

圖 5 複製與無複製測試結果

4.4 混合字體模型結果

為了瞭解訓練字體的多樣化，是否影響辨識能力，因此我們將兩種字體的訓練集混和後進行訓練，並與單一字體模型做比較，實驗結果如圖6，從中可以發現混合字體資料以得到的模型，相較於單一字體所訓練出的模型效果更好。

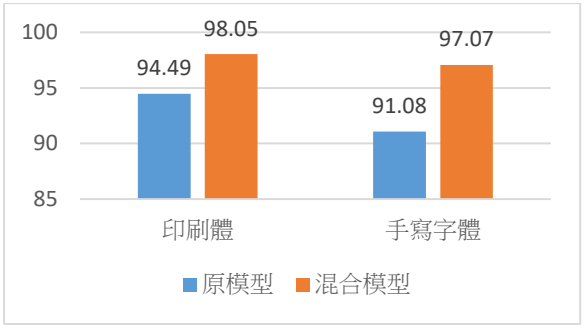

圖6 混合字體模型比較

5 結論與未來工作

從我們的實驗中可以看出，比起其他 OCR 軟體，使用 CNN 更適合於古籍文字辨識，雖然在4-2節中發現字體的不同會對於 CNN 辨識率造成影響，但我們於4-5節發現混和不同字體所訓練出的模型相較於針對單一字體所訓練的模型，準確率大為提升，此實驗也間接說明了訓練集圖片的多樣性影響大於字體對於模型的影響，越多樣性的圖片越可以適應大部分的狀況，且針對不同字體去選擇相對應字體的模型是一件相當麻煩的事情；總和上面的實驗結果，我們在混合模型實驗結果當中發現，真正影響 CNN 辨識率的原因，字體只佔了小部分的因素，真正影響 CNN 的是訓練集圖片的多寡以及多樣性。

未來我們將收集更多不同字體的資料集，繼續實驗字體混和或單一訓練的優缺點，並改善模型內部的參數設計，以利用更少的時間得到更高的準確率。

參考文獻

[1] Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., ... & Yang, L. (2016). Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, *3*(2), 113-120.

[2] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396-404).

[3] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

[4] LeCun, Yann, Corinna Cortes, and Christopher JC Burges. "MNIST handwritten digit database." *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* 2 (2010).

[5] CASIA Online and Offline Chinese Handwriting Databases http://www.nlpr.ia.ac.cn/databases/handwriting/Home.html

[6] 漢籍電子文獻資料庫， http://hanchi.ihp.sinica.edu.tw

[7] LeCun Y, Boser B E, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C]//Advances in neural information processing systems. 1990: 396-404.

[8] Bellili A, Gilloux M, Gallinari P. An MLP-SVM combination architecture for offline handwritten digit recognition[J]. International Journal on Document Analysis and Recognition, 2003, 5(4): 244-252.

[9] Niu X X, Suen C Y. A novel hybrid CNN–SVM classifier for

recognizing handwritten digits[J]. Pattern Recognition, 2012, 45(4): 1318-1325.

[10] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[11] Cireşan, Dan, and Ueli Meier. "Multi-column deep neural networks for offline handwritten Chinese character classification." Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, 2015.

[12] Hawkins, Douglas M. "The problem of overfitting." *Journal of chemical information and computer sciences* 44.1 (2004): 1-12.

[13] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[15] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012).” Improving neural networks by preventing co-adaptation of feature detectors.” *arXiv preprint arXiv:1207.0580*.

[16] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). “Dropout: a simple way to prevent neural networks from overfitting.” *Journal of machine learning research*, *15*(1), 1929-1958.

[17] Smith, Ray. "An overview of the Tesseract OCR engine." *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*. Vol. 2. IEEE, 2007.

[18] ABBYY Cloud SDK http://cn.ocrsdk.com/