【2019 佛教藏經會議專稿】

人工智慧視角下的佛教大藏經

釋賢超

北京龍泉寺藏經辦公室主任

一、導言

飛速發展的人工智慧，啟發我們從嶄新的角度看待佛教大藏經，那就是：將佛教大藏經轉化為大數據資源，促進人工智慧在佛教研究領域的落地。我們的眼光不應局限於「如何做好一部大藏經」，而應投向「如何利用大藏經造福人類」。大藏經不僅是人工智慧的應用對象，也是人工智慧的推進燃料。

深度學習是人工智慧的重要分支——機器學習的一種實現方式，也是人工智慧最熱門的技術流派。「端到端」訓練和「表徵學習」是深度學習的主要特徵。表徵學習又主要分為有監督學習和無監督學習。有監督學習是通過學習標記數據來獲得特徵，是目前人工智慧發展最成熟的方向。

本文從圖像和自然語言處理兩個方向介紹深度學習在佛教大藏經上的研究進展，特別是光學字符識別（Optical Character Recognition，簡稱 OCR）、版面分析、文字檢測、自動標點，以及相關展望。

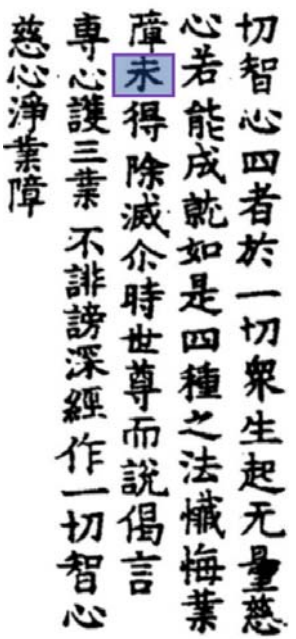
佛光學報 新七卷．第一期，2021 年 1 月，頁 223-245

二、OCR

（一）技術原理

傳統 OCR 針對的是印刷體文字的識別，基本原理是模板匹配。目前針對印刷體的 OCR 技術已經比較成熟，但是這種方法用於古籍 OCR 的效果並不理想。因為古籍字體的形狀、大小都存在一定的變化範圍，既不像印刷體那樣整齊劃一，也不像手寫體那樣自由發揮，其規範程度介於印刷體和手寫體之間。這種由人工寫就、相對規範的字體，可以稱為「手寫印刷體」。

「切分」是 OCR 的主要難點。傳統的 OCR 算法是先將文字區域切分為文字行，再把文字行切分為單字，然後進行單字識別。常見的算法有「過分割——動態規劃」和滑動窗口。前者的原理是將文本行過度切分成很多碎片，然後通過動態規劃合併碎片。後者則是通過窗口平移進行字符匹配。這些方法的識別效果都存在切分誤差積累的問題，上下文信息也沒有被充分利用起來。例如，下圖中畫框的文字，如果僅從字形判斷，應是「未」。但是根據上下文判斷，則應是「永」。



圖一　「永」或「未」。圖像取自《高麗藏・合部金光明經》

基於深度學習的 OCR 不依賴於切分，不必確定漢字的位置即可進行識別。常用方法分為三個階段：提取

特徵序列、預測特徵序列的標簽分佈、將標簽分佈轉錄為識別結果，簡稱為「提取——預測——轉錄」。

第一，提取階段，使用的是「卷積神經網絡」（Convolutional Neural Network，簡稱 CNN）。CNN 作為一種高效的特徵提取器，廣泛應用於圖像領域。CNN 的基本結構是「卷積層——激活層——池化層」。這是受到了視覺皮層細胞的啟發，可以對圖像中不同層次的概念進行分層提取。（1）卷積層（convolutional layer）的作用是對輸入產生響應，也就是將輸入圖像跟卷積核（kernel）進行卷積運算，獲得特徵響應圖。每一種卷積核代表一種特徵，每個特徵響應圖可以稱為一個通道（channel）。每個卷積層通常具有幾十、幾百的通道，每個通道都代表一種特徵。（2）激活層（activation layer）的作用對特徵響應圖進行非線性變換，目的是引入非線性，以增強擬合複雜解空間的能力，常見的激活函數是 ReLU，它的作用是忽略小於零的相應，保留大於零的相應。（3）池化層（pooling layer）的作用在於保留主要特徵，減少參數量和計算量，並且保持對微細位移的不變性。

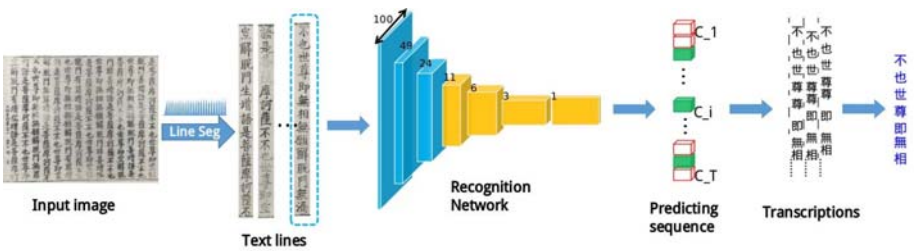
第二，預測階段，使用的是「循環神經網絡」（Recurrent Neural Network，簡稱 RNN）。這裏借鑒了語音識別技術，不同的是將輸入的聲學特徵序列替換成了圖像特徵序列。文本圖像相當於語音的視覺化。擅長處理序列數據的 RNN 可以捕捉特徵序列中的上下文信息，輸出的是特徵序列每個位置上所有可能字符（取決於字典的規模）的出現概率。這裏常用的 RNN 是雙向長短時記憶（Bi-directional Long Short-Term Memory Network，簡稱 Bi-LSTM）。

第三，轉錄階段。特徵序列的長度相比於輸出文本的長度，通常是過度冗餘的。所以需要把冗餘重複的部分去掉，整合為最終的識別結果。常用方法是採用 CTC（Connectionist Temporal Classification）或者

注意力（attention）。CTC 的處理過程，首先是合併重複字符，其次去掉占位符，剩餘字符便是識別結果。

深度學習 OCR 應用古籍字體識別的優勢在於：不受古籍中常見的筆劃交叉（即文字之間不存在連通的空隙）、字形交錯（文字之間的空隙不是水平直線，可能是斜線、折線或不規則曲線）等現象的影響，不存在切分誤差積累的問題，能夠充分利用上下文信息。

（二）基於 CNN+LSTM+CTC 的古籍 OCR



圖二　CNN+LSTM+CTC 算法[1]

華南理工大學的研究團隊基於 CNN+LSTM+CTC 開發了一種針對古籍字體的 OCR 算法，可參見上圖二所示。數據集是從《高麗藏》中選取的七萬張整版圖像，切分為 168 萬條文本行。數據集按照 4：1 的比例拆分為訓練集和測試集。

將整版圖像切分為文字行，採用的是投影算法。根據不完全統計，大約 10%的文本行圖像和標簽是沒有對齊的。為了減少由此帶來的訓練

[1] Yang Hailin, Lianwen Jin, and Jifeng Sun, “Recognition of Chinese Text in Historical Documents with Page-level Annotations,” in 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), New York: IEEE, 2018, pp. 199-204.

誤差，這裏使用了自適應梯度門（Adaptive Gradient Gate，簡稱 AGG）。原理是在訓練過程中，對於帶來高損失值的樣本將其傳遞梯度降至為零。使用 AGG 之後，錯誤率從 1.97%降低到 1.27%，降幅達到 35%，對於提升識別準確率有明顯效果。

除了常規的 CNN+LSTM+CTC 架構，論文還使用了一種簡化架構 CNN+CTC，去掉了中間的 LSTM，其識別準確率並沒有降低。原因可能是當特徵序列中的每個時間步長所對應的感受野大於圖像中單個字符的平均高度的時候，特徵序列中可能就已經包含了字符周圍的上下文信息，所以可以取代 LSTM 而實現類似的效果。如果修改結構來減少感受野的尺寸，使特徵序列不含有字符周圍的上下文信息，那麼理論上將降低識別的性能。這一點也在實驗中得到了印證。在 CNN+CTC 架構中，減少感受野的尺寸後，錯誤率從 1.29%上升到了 3.84%，增長了近二倍。

表一　不同架構的 OCR 效果[2]

TABLE II
THE RESULTS OF DIFFERENT ARCHITECTURE

Format	Architecture	Accuracy Rate (%)
Text Line	CNN+CTC	98.71 ± 0.03
Text Line	CNN+LSTM+CTC	98.60 ± 0.15
Single Character	CNN	97.36
Text Line	CNN+CTC_smallRF	96.16 ± 0.05

[2] Yang Hailin, Lianwen Jin, and Jifeng Sun, “Recognition of Chinese Text in Historical Documents with Page-level Annotations,” pp. 199-204.

三、文字檢測

（一）電子化校對

OCR 雖然可以解決絕大部分文字錄入工作，但是識別結果仍然需要人工來做逐字逐句的校對。最終的文本質量與校對工作的水平有直接關係。

《大正藏》、《卍續藏》早已實現了全文電子化，也便於進行圖文對照。但是這幾部大藏經是現代製作完成的，其中存在很多偏離底本的訛誤，需要認真核對最初底本才可能糾正。例如，CBETA 將《高麗藏》與《大正藏》的對應內容進行全文對比，發現《大正藏》新產生的文字訛誤還是相當多的。當然，《大正藏》也有未被以往大藏經收錄的經文。排查這些經文的訛誤，需要逐一尋找其對應的底本。如果完全依靠人工校對，無疑是成本巨大而且效率低下的。

電子化校對，就是利用電腦技術來提高校對的效率和準確率，有逐字校對和聚類校對兩種方式。逐字校對又被稱為橫校，就是按照自然順序進行文字與圖像之間的逐個比對。這是一種接近人工校對的方式。聚類校對又被稱為縱校，是將同種文字的字形圖像聚合在一起，在圖像之間對比。兩者相互補充，可以提高校對的準確性。逐字校對會受上下文、語言習慣等因素影響而產生失誤。聚類校對則可以排除上下文和語言習慣的干擾，聚焦在字形本身，彌補上述漏洞。

聚類校對需要藉助能夠檢測出文字邊界並識別文字種類的算法，這在計算機視覺中屬於檢測任務，簡單的說就是解決「在哪裡」和「是什麼」的問題。

現代印刷體的版面相對整齊，檢測文字邊界並不困難。古籍的情況更為複雜：（1）古籍文字大多排布緊密，單張圖像的檢測對象能有數百

之多，與物體檢測或場景字符檢測中排布鬆散、對象稀疏的情況有很大差別。（2）文字檢測對精度的要求更為苛刻。在一般目標檢測任務中，面積交並比（Intersection over Union，簡稱 IoU）達到 0.5 即可。但是對於古籍來說，IoU 偏小會令檢測結果發生部件丟失的情況，導致文字認讀失誤。根據經驗，IoU 達到 0.8 才是比較理想的。

按照自上到下的思路，古籍文字檢測分為三個環節：一、版面分析，從整篇圖像中提取文字區域；二、行分割，將文字區域分割為文本行的組合，投影法是常用方法；三、字分割，將文字行分割為單個字圖的組合，結合單字識別可以取得更好的分割效果。此外，按照自下而上的思路，也可以第一步就從整篇圖片中直接檢測單個文字，再根據單字邊界框的相對位置還原文字列和文字區域。

下文分別從文字檢測數據集、版面分析以及識別引導、弱監督學習、強化學習等三種文字檢測方法，介紹相關工作。

（二）文字檢測數據集

針對佛教大藏經的文字檢測數據集，有以下兩種：

1.《高麗藏》中文數據集（Tripitaka Koreana in Han，簡稱 TKH）。[3]《高麗藏》的字形統一，版式整齊，適合作為古籍文字檢測與識別的基線，具有完整的版面標注、行標注和字標注。

2.多種藏經中文數據集（Multiple Tripitaka in Han，簡稱 MTH），包

[3] Yang Hailin, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun, “Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector,” *IEEE Access* 6 (2018): 30174-30183.

括 MTH500、MTH1000、MTH1200 三種子集。[4]圖像來自八種大藏經（契丹藏、趙城金藏、思溪藏、洪武南藏、永樂北藏、徑山藏、乾隆藏、中華大藏經），包含雙行夾註、圖形等複雜版式，具有完整的版面標注、行標注和字標注。

表二　針對佛教大藏經的文字檢測數據集

	TKH	MTH500	MTH1000	MTH1200
頁面	1000	500	1000	1200
行	23471	17178	27559	21416
文字實例	323491	197886	420548	337613
文字種類	1471	3664	5341	5292
雙行夾註比例	0%	？	9.0%	27%

（三）基於級聯網絡的版面分析

版面分析的目的是確定文字區域的邊界，為下一步的行分割和字分割做準備。傳統的版面分析算法是通過形態學變換，提取投影信息，然後根據事先設定的閾值，選出圖片中的高峰區域，即為文本邊界的位置。這種方法依賴於手工設計特徵，只對特定的版面樣式有效，泛化能力（generalization ability）不佳，容易受到背景、汙漬等干擾。

Cascade R-CNN 是一種高精度目標檢測模型，適合古籍文字檢測這種對 IoU 要求較高的目標檢測任務。馬偉洪採用 Cascade R-CNN 作為基

[4] MTH500，參見 Yang *et al.*, “Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector,” pp. 30174-30183、MTH1000 與 MTH1200，參見馬偉洪，〈基於深度學習的文檔圖像版面分析與文字識別〉，廣州：華南理工大學電子與信息學院本科畢業設計，2019 年；伍思航，〈基於強化學習的古籍文字精準檢測〉，廣州：華南理工大學電子與信息學院本科畢業設計，2019 年。

礎網絡框架，以 Resnet-50 作為特徵提取層，並用單階段檢測網絡 YOLOv3 和兩階段檢測網絡 Mask R-CNN 作為對比組。[5]數據集採用 TKH、MTH1000、MTH1200 的總集。

結果顯示：當 IoU=0.8 和 0.9 的時候，Cascade R-CNN 的各項性能指標明顯優於 YOLO-v3 和 Mask R-CNN，具有較好的泛化能力。

表三　版面分析的性能指標對比[6]

	IoU=0.8			IoU=0.9		
模型	準確率	召回率	F 值	準確率	召回率	F 值
YOLOv3	84.47%	84.11%	84.47%	47.61%	46.91%	47.26%
Mask R-CNN	98.64%	98.95%	98.80%	96.25%	96.55%	96.40%
Cascade R-CNN	99.26%	99.05%	99.16%	98.21%	98.01%	98.11%

（四）識別引導檢測

單純根據圖像特徵進行文字分割，容易在上下結構的漢字產生錯誤。將識別與檢測相結合，可以減少錯誤發生。Yang Hailin 等人提出了識別引導檢測（Recognition Guided Detector，簡稱 RGD）的思路：對當前文本行進行整列識別，借助識別結果確認每個文字的具體位置。[7]這種方法能夠實現密集而精確的文字檢測。

識別引導檢測的結構是由兩個共同訓練的卷積神經網絡組成：一個是識別引導候選網絡（Recognition Guided Proposal Network，簡稱 RGPN），另一個是檢測網絡。這兩個網絡共享前三層的參數，可以實現

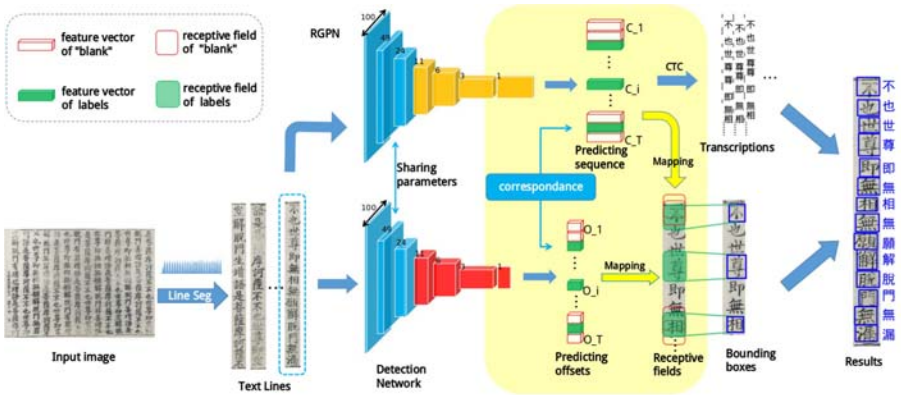
5　參見馬偉洪，〈基於深度學習的文檔圖像版面分析與文字識別〉，頁 27。

6　馬偉洪，〈基於深度學習的文檔圖像版面分析與文字識別〉，頁 27-28。

7　Yang *et al*., “Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector,” pp. 30174-30183.

最佳的整體性能。

完整的檢測過程由三個部分組成：文本行分割、生成候選文字和文字檢測。首先通過垂直投影方法從輸入圖像分割出文本行。然後由 RGPN 生成上下文。最後，通過 RGD 參照 RGPN 提供的上下文來取得精確的文字邊界。



圖三　識別引導檢測的基本流程[8]

實驗發現，當以整圖作為輸入並且在高 IoU 的條件下（等於 0.7、0.8）的情況下，RGD 在 TKH 上勝過了其他圖像檢測方法。而當以文本行作為輸入時，其他檢測方法效果更好。這說明 RGD 更適合於密集文字區域的檢測。由於文本行中的檢測對象更為稀疏，因此 RGD 的優勢並不明顯。

8 Yang *et al.*, "Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector," pp. 30174-30183.

表四　RGD 與其他方法在 TKH 的對比。[9]上半部以整圖為輸入，下半部以文本行為輸入

	Param	IoU:0.5			IoU:0.6			IoU:0.7			IoU:0.8		
		P	R	F	P	R	F	P	R	F	P	R	F
RFCN [1]	70.52M	99.69	90.96	94.98	99.34	90.37	94.64	97.32	88.54	92.72	78.22	71.15	74.52
Faster-RCNN [23]	130.07M	90.70	99.79	95.03	99.53	90.46	94.78	97.89	88.97	93.22	82.16	74.67	78.24
YOLO [22]	232.19M	-	-	-	-	-	-	-	-	-	-	-	-
SSD [15]	87.30M	99.92	66.03	75.23	99.78	60.22	75.10	98.54	59.56	74.24	86.60	52.31	65.23
TextBoxes [13]	90.64M	99.92	57.27	72.81	99.77	57.18	72.70	98.51	56.46	71.78	84.77	48.58	61.77
DMP-Nets [16]	178.56M	99.43	89.54	94.23	98.63	88.82	93.47	95.29	85.82	90.31	71.59	64.48	67.85
FEN [33]	176.85M	99.34	97.57	**98.44**	98.18	95.86	97.00	89.66	86.74	88.18	62.28	59.41	60.81
RGD[ours]	**9.29M**	98.32	97.52	97.92	97.23	96.50	96.86	94.93	94.27	94.60	84.08	83.56	83.82
RGD-VGG16[ours]	64.02M	98.58	96.96	97.76	97.64	96.39	**97.01**	95.40	94.56	**94.98**	85.97	85.72	**85.85**
RFCN-Line	70.52M	99.54	99.58	**99.56**	98.88	98.92	**98.90**	95.92	95.49	95.70	83.22	83.19	83.21
Faster-RCNN-Line	130.07M	99.44	99.33	99.38	98.35	98.46	98.40	94.11	94.32	94.22	79.73	79.74	79.74
YOLO-Line	232.19M	92.28	96.09	94.15	91.19	95.22	93.16	83.26	86.94	85.06	56.73	59.24	57.95
SSD-Line	87.30M	99.56	96.44	97.98	98.54	95.81	97.16	94.65	92.64	93.63	79.07	78.00	78.53
TextBoxes-Line	90.64M	98.49	98.49	98.66	97.89	98.23	98.06	95.75	96.08	95.91	86.82	87.12	**86.97**
DMP-Nets-Line	178.56M	99.56	99.46	99.51	98.98	98.67	98.82	96.37	96.06	**96.22**	81.19	80.93	81.06
FEN-Line	176.85M	99.62	99.46	99.54	99.01	98.65	98.83	96.52	94.91	95.71	77.15	74.48	75.79

在 MTH500 上，所有檢測方法在以整圖作為輸入的情況下，性能都出現急劇下降，故以文本行作為輸入與其他檢測算法進行比較。當 RGD 檢測網絡的骨幹部分採用 VGG-16 且高 IoU（大於 0.6）的條件下，檢測性能明顯優於其他方法。

表五　RGD 與其他方法在 MTH500 的對比。[10]

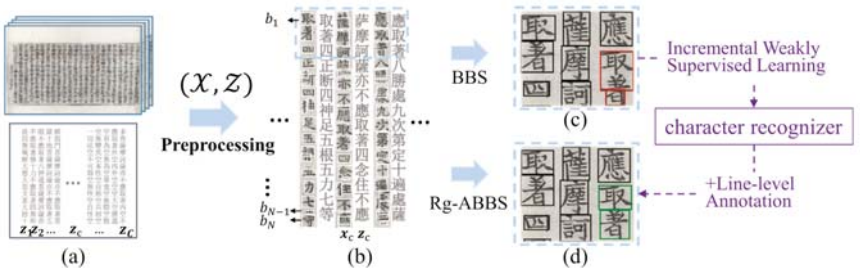
	IoU:0.5			IoU:0.6			IoU:0.7			IoU:0.8		
	P	R	F	P	R	F	P	R	F	P	R	F
RFCN-Line	96.30	97.68	**96.98**	94.36	95.72	95.04	84.50	85.72	85.11	46.94	47.61	47.27
Faster-RCNN-Line	96.15	97.44	96.79	93.26	94.52	93.89	81.60	82.70	82.15	43.97	44.57	44.27
YOLO-Line	92.09	88.51	90.26	88.96	85.50	87.20	77.14	74.14	75.61	42.16	40.52	41.33
SSD-Line	98.85	90.41	94.44	97.48	89.17	93.14	89.81	82.14	85.81	58.39	53.40	55.78
TextBoxes-Line	81.75	93.88	87.38	80.11	91.99	85.64	75.87	87.13	81.11	58.33	66.98	62.36
DMP-Nets-Line	96.45	95.91	96.18	94.65	94.12	94.38	85.13	84.66	84.89	46.05	45.79	45.92
FEN-Line	96.37	91.40	93.82	95.17	88.98	91.97	83.91	77.23	80.44	42.52	38.00	40.14
RGD[ours]	97.35	95.97	96.65	95.34	94.00	94.67	88.81	87.55	88.17	61.98	61.10	61.54
RGD-VGG16[ours]	97.71	95.86	96.78	96.44	94.61	**95.52**	92.17	90.42	**91.29**	73.72	72.31	**73.01**

9　Yang *et al*., “Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector,” pp. 30174-30183.

10　Yang *et al*., “Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector,” pp. 30174-30183.

（五）弱監督的文字檢測

基於深度學習的文字檢測方法，需要大規模的單字標注數據集。但是很多時候這種條件並不具備。在這種弱監督的條件下，Xie Zecheng 等人提出了實現文字精確分割的四步流程。[11]



圖四　基於弱監督學習的精確文字分割[12]

第一步利用垂直投影法，將頁面切分為文本行。

第二步進行邊界框分割（Boundary Box Segmentation，簡稱 BBS），給出近似的分割結果。

第三步，採用增量弱監督學習，訓練高性能字符識別器。這是一個循環迭代的過程。此處基於兩點考慮：（1）在分類正確的情況下，通常是第一選項（P_f）佔有絕大部分概率，其他選項幾乎可以忽略不計；（2）在分類錯誤的情況下，第二選項的概率（P_s）通常會顯著大於之後所有

[11] Xie Zecheng, Yaoxiong Huang, Lianwen Jin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao, and Xiaode Zhang, “Weakly Supervised Precise Segmentation for Historical Document Images,” *Neurocomputing* 350 (2019): 271-281.

[12] Xie Zecheng, Yaoxiong Huang, Lianwen Jin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao, and Xiaode Zhang, “Weakly Supervised Precise Segmentation for Historical Document Images,” *Neurocomputing* 350 (2019): 271-281.

選項。此處設計了一個判斷門，根據 P_f、P_s 的相對大小，在每次迭代過程中賦予字符樣本以「確定」或「不確定」的標簽。確定的樣本將用於更新樣本池。迭代過程終止於樣本池不再更改或者重新標記的樣本數量在容差範圍內。

第四步，對關注區域（即容易出現錯誤分割的區域）執行識別引導注意力邊界框分割（Recognition-guided Attention Boundary Box Segmentation，簡稱 Rg-ABBS）。Rg-ABBS 在兼顧了 Rg-BBS（Recognition-guided Boundary Box Segmentation）和 BBS 各自的優點，在保證足夠的分割性能同時，具有更少的時間消耗。

結果表明，在精確分割的要求（IoU=0.85）下， Rg-ABBS 在 TKH 數據集上超過其他檢測方法，取得了當前最優結果。

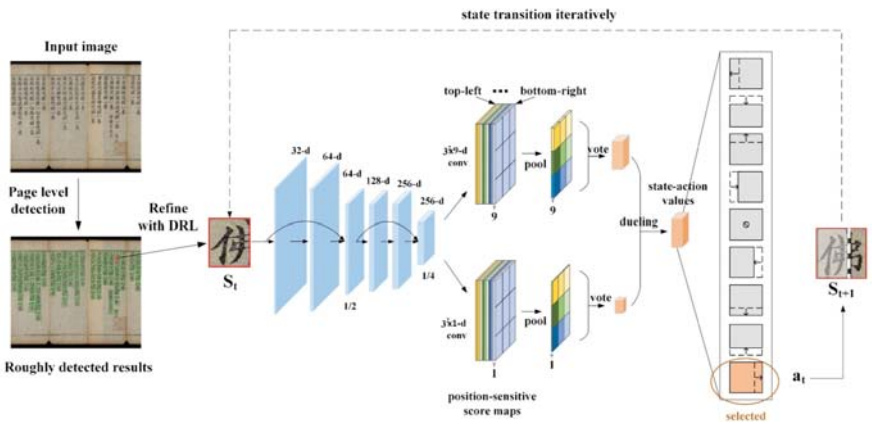
表六　Rg-ABBS 在 TKH 的表現[13]

Method	IoU 0.7			IoU 0.75			IoU 0.8			IoU 0.85		
	R	P	F	R	P	F	R	P	F	R	P	F
Projection [19]	32.61	34.34	33.45	22.23	23.16	22.69	15.38	15.79	15.58	12.16	12.41	12.28
Grouping [30]	26.60	21.07	23.51	15.73	10.93	12.90	15.02	10.81	10.87	14.08	9.97	11.67
CCS [31]	75.51	76.42	75.96	70.63	71.48	71.05	62.21	62.96	62.58	45.37	45.92	45.64
SS [60]	27.41	21.07	23.83	24.18	20.87	22.40	20.35	16.73	18.36	15.21	11.43	13.05
ACF [42]	25.48	26.17	25.81	23.41	24.35	23.87	21.48	22.27	21.87	20.53	21.42	20.97
R-fcn [43]	88.54	97.32	92.72	83.85	92.17	87.81	71.15	78.22	74.52	47.57	52.29	49.82
SSD [37]	59.56	**98.54**	74.24	57.46	**95.19**	71.66	52.31	**86.60**	65.23	42.69	70.56	53.20
YOLOv2 [36]	**93.92**	97.11	**95.49**	**90.42**	93.50	**91.93**	81.32	84.09	82.68	60.02	62.06	61.02
TextBox [52]	56.46	98.51	71.78	53.72	91.02	67.56	48.58	84.77	61.77	42.29	79.26	55.15
MNC [34]	90.70	91.22	90.96	86.17	84.66	85.41	76.29	76.73	76.51	56.14	56.45	56.29
FCIS [33]	74.89	75.78	75.33	55.17	55.82	55.49	30.82	31.19	31.01	21.94	22.08	22.01
BBS (ours)	88.24	85.76	86.98	84.21	82.07	83.13	81.21	80.18	80.69	74.21	73.41	73.81
Rg-ABBS (ours)	92.63	90.56	91.58	90.15	88.13	89.13	**86.54**	84.61	**85.56**	**78.80**	**77.04**	**77.91**

[13] Xie Zecheng, Yaoxiong Huang, Lianwen Jin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao, and Xiaode Zhang, “Weakly Supervised Precise Segmentation for Historical Document Images,” *Neurocomputing* 350 (2019): 271-281.

（六）基於深度強化學習的文字檢測

伍思航最近提出了一種基於深度強化學習（Deep Reinforcement Learning，簡稱 DRL）的古籍文字檢測算法。[14]其思想是：首先使用篇幅級檢測器對古籍文獻進行文字粗檢測，再利用強化學習算法對粗檢測結果進行精調，得到最終檢測結果。



圖五　基於強化學習的古籍文字檢測算法[15]

上圖左側是粗檢測，使用單階段檢測算法 YOLOv3 進行整篇圖片的檢測。右側是精調階段，基於值函數的「深度 Q 學習網絡」（Deep Q-Learning Networks），採用一種帶有位置敏感池化的「全卷積神經網絡」（Fully Convolutional Network with Position Sensitive RoI Pooling，簡稱 FCPN）來提取特徵信息並映射到動作價值函數。

主幹網絡包含兩個殘差模塊，每個模塊包含三個卷積層，末尾一個二倍下採樣池化層。主幹末尾展開兩個分支，分別計算當前的狀態價值

[14] 伍思航，〈基於強化學習的古籍文字精準檢測〉，頁 10。
[15] 伍思航，〈基於強化學習的古籍文字精準檢測〉，頁 10。

和動作優勢價值。最後使用競爭模塊將兩者價值融合後輸出各個動作的價值，選取價值最大的動作來調整文字邊界框，調整後的狀態繼續下一輪迭代。這裡設計了八個分別對應於在四個方向擴張或收縮邊界的動作和一個停止動作，動作調整步長為 2 像素。

在強化學習中，智能體需要從環境中獲得獎勵或懲罰的反饋信息進行學習。根據文字檢測的特點，這裡提出了一種新的「密集獎勵函數」（Dense Reward Function，簡稱 DRF），由當前狀態變量和下一個狀態變量共同決定獎勵信號，加快了調整過程。

圖六　強化學習調整過程的可視化。[16]

[16] 伍思航，〈基於強化學習的古籍文字精準檢測〉，頁 23。

在 TKH 和（MTH500+MTH1200）兩個數據集上分別進行訓練和測試，在高 IoU 和低 IoU 兩種條件下，本方法（YOLOv3+DRL）都超越了 RGD 以及其他方法，實現了當前最優結果（state-of-the-art）。

表七　強化學習在 TKH 的表現[17]

	IoU:0.5	IoU:0.6	IoU:0.7	IoU:0.8
SSD[10]	75.23	75.10	74.24	65.23
TextBoxes[38]	72.81	72.70	71.78	61.77
DMP-Nets[39]	94.23	93.47	90.31	67.85
RFCN[6]	94.98	94.64	92.72	74.52
Faster-RCNN[40]	95.03	94.78	93.22	78.24
FEN[41]	98.44	97.00	88.18	60.81
Xie-Line[42]	-	-	91.58	85.56
Gao-Line[19]	-	-	92.62	80.85
RGD-VGG16-Line[20]	97.76	97.01	94.98	85.85
YOLOv3[9]	99.40	98.64	96.11	82.89
YOLOv3 with DRL	**99.42**	**98.68**	**96.35**	**87.05**

表八　強化學習在 MTH 的表現[18]

	IoU:0.5	IoU:0.6	IoU:0.7	IoU:0.8
TextBoxes-Line[38]	87.38	85.64	81.11	62.36
FEN-Line[41]	93.82	91.97	80.44	40.14
SSD-Line[10]	94.44	93.14	85.81	55.78
RFCN-Line[6]	96.98	95.04	85.11	47.27
Faster-RCNN-Line[40]	96.79	93.89	82.15	44.27
DMP-Nets-Line[39]	96.18	94.38	84.89	45.92
RGD-VGG16-Line[20]	96.78	95.52	91.29	73.01
YOLOv3[9]	97.20	95.95	91.02	67.28
YOLOv3 with DRL	**97.30**	**96.23**	**92.57**	**73.83**

17　伍思航，〈基於強化學習的古籍文字精準檢測〉，頁 22。
18　伍思航，〈基於強化學習的古籍文字精準檢測〉，頁 22。

四、自動標點

廣義上的古文標點包括斷句、標點兩類。一般來說，斷句較為適合於學術讀者，注重的是文本嚴謹性。標點較為適合於大眾讀者，注重的是閱讀流暢性。

在自動斷句方面，有前後文 n-gram、[19]人工設計的模式識別庫、[20]條件隨機場（Conditional Random Field，簡稱 CRF）、[21]神經網絡語言模型（Neural Network Language Model，簡稱 NNLM）、[22]基於門控循環單元（Gate Recurrent Unit，簡稱 GRU）的雙向循環神經網絡（bi-directional recurrent neural network）、[23]帶有筆劃嵌入的雙向長短時記憶（Bi-LSTM with radical embedding）、[24]「BERT+微調」、[25] 「BERT+CRF」[26]等方法。

---

19 陳天瑩、陳蓉、潘璐璐、李紅軍、于中華，〈基於前後文 n-gram 模型的古漢語句子切分〉，《計算機工程》33：3，2007 年，頁 192-193。

20 黃建年，侯漢清，〈農業古籍斷句標點模式研究〉，《中文信息學報》22：4，2008 年，頁 31-38。

21 參見王川、張小紅、韓采華，〈古漢語句子切分與句讀標記方法研究〉，《河南大學學報（自然科學版）》39：5，2009 年，頁 525-529；張合、王曉東、楊建宇、周衛東，〈一種基於層疊 CRF 的古文斷句與句讀標記方法〉，《計算機應用研究》26：9，2009 年，頁 3326-3329；張開旭、夏雲慶、宇航，〈基於條件隨機場的古漢語自動斷句與標點方法〉，《清華大學學報（自然科學版）》10，2009 年，頁 1733-1736；Huang Hen-Hsen, Chuen-Tsai Sun, and Hsin-Hsi Chen. “Classical Chinese sentence segmentation.” In *CIPS-SIGHAN Joint Conference on Chinese Language Processing* 2010。

22 Wang Boli, Xiaodong Shi, Zhixing Tan, Yidong Chen, and Weili Wang. “A sentence segmentation method for ancient Chinese texts based on NNLM.” In *Workshop on Chinese Lexical Semantics*, Cham: Springer, 2016, pp. 387-396.

23 王博立、史曉東、蘇勁松，〈一種基於循環神經網絡的古文斷句方法〉，《北京大學學報（自然科學版）》53：2，2017 年，頁 255-261。

24 Xu Han, Wang Hongsu, Zhang Sanqian, Fu Qunchao, and Liu Jun, “Sentence segmentation for classical Chinese based on LSTM with radical embedding,” *The Journal of China Universities of Posts and Telecommunications* 26, 2 (2019): 1-8.

綜合各種文獻，前後文 n-gram 的 F1 值為 63.78%；人工設計的模式識別庫的準確率為 48%；CRF 的 F1 值介於 67.90 至 83.34%之間；NNML 的 F1 值為 81.13%；帶有 GRU 的 Bi-RNN 的 F1 值為 75.51%；帶有筆劃嵌入的 Bi-LSTM 的 F1 值為 70.5 至 81.3%之間；「BERT+微調」的 F1 值為 91.67%；「BERT+CRF」的 F1 值為 92.03%。

在自動標點方面，有人工設計的模式識別庫、[27]CRF、[28]殘差雙向長短時記憶（Residual Bi-directional Long Short-Term Memory，簡稱 Residual Bi-LSTM）、[29]「BERT+微調」[30]等方法。綜合各種文獻，人工設計的模式識別庫的準確率是 36%；CRF 的 F1 值處於 48.96%至 67.67%之間；Residual Bi-LSTM 的平均準確率為 79.92%；「BERT+微調」的 F1 值為 70.40%。

基於 Residual Bi-LSTM 的自動標點由北京市海淀區龍泉寺藏經辦公室與彩雲科技合作開發，部署於「古籍酷」網站，能夠標註七種現代標點（句號、逗號、問號、感歎號、冒號、分號、頓號）並給出每種標

25 俞敬松、魏一、張永偉，〈基於 BERT 的古文斷句研究與應用〉，《中文信息學報》33：11，2019 年，頁 57-63。

26 Hu Renfen, Shen Li, and Yuchen Zhu. “Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on Deep Language Models.” In *18th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 2019.

27 黃建年等，〈農業古籍斷句標點模式研究〉，頁 31-38。

28 詳見王川等，〈古漢語句子切分與句讀標記方法研究〉，頁 525-529；張合等，〈一種基於層疊 CRF 的古文斷句與句讀標記方法〉，頁 3326-3329；張開旭等，〈基於條件隨機場的古漢語自動斷句與標點方法〉，頁 40。

29 釋賢超、方愷齊、釋賢迥、釋賢菊、釋賢碉、釋賢繼、釋賢大、釋賢奉、宋延淳，〈一種自動標點的方法與實現〉，《數位典藏與數位人文》3，2019 年，頁 1-19。

30 俞敬松等，〈基於 BERT 的古文斷句研究與應用〉，頁 57-63。

點的出現概率。[31] CBETA 在新式標點專案中採用了本技術，對於標點效率和規範性有所幫助。

北京市海淀區龍泉寺藏經辦公室與華南理工大學深度學習與視覺計算實驗室合作，基於 Transformer 開發了新一代自動標點，實現了91.28%的平均準確率。[32]

為了檢驗自動標點與人類標點之間的實際差距，先是在二十萬五千字的古文樣本上，將從網絡搜集而來的兩種人工標點方案進行相互對比，得到的 F1 值為 87.0%。這個數值反映了人工標點方案之間的合理差異度，可以視作自動標點性能指標的經驗上限。如果自動標點的 F1 值超出經驗上限過多，意味著可能存在過擬合的情況，不利於陌生文本上的泛化能力。然後用基於 Transformer 的自動標點的標點結果對比前面兩種人工標點方案，F1 值分別為 84.8%和 86.4%，相當接近人工標點方案的 87%。以基於 Residual Bi-LSTM 的自動標點作為對照，在兩種人工標點方案上的 F1 值分別為 60.3%和 60.9%，遠低於基於 Transformer 的自動標點。通過其他測試也表明，基於 Transformer 的自動標點實際標點水平已經接近人類專業水平，在陌生文本上也具有很好的泛化能力。

五、前景展望

（一）BERT

2018 年出現的以 BERT（Bidirectional Encoder Representations from

31 古籍酷，http://www.gj.cool, 2020/1/28。

32 基於深度學習的中國古籍文檔智能理解項目演示系统，http://115.29.208.50:9000/punctuation, 2020/1/28。

Transformers）、[33]GPT（Generative Pre-Training）[34]為代表的預訓練技術，在自然語言處理領域引起了一場深刻的革命，令自然語言處理成為人工智慧最活躍的領域。以往成為主要瓶頸的標註數據稀疏問題，在將無標注數據引入預訓練之後，得到了緩解。此外，BERT 對於訓練數據的巨大需求，也要求在解決有關佛教古漢語問題的時候，不應僅限於佛教領域，而是應當全面利用所有古漢語資料。

BERT 可以捕捉到細粒度的語義信息，能夠根據某一語境向量的上下文環境，找到在語義上最近鄰的其他語境向量，實現語義聚類。[35]所以，人工智慧在原則上可以輔助字詞典的編纂，幫助學習者理解和掌握古漢語這門語言。

（二）MASS

雖然 BERT、GPT 等預訓練方法在情感分類、自然語言推理、命名實體識別、SQuAD 等語言理解任務上取得了最佳表現，但是從語言生成任務的角度來看，其或只具有編碼器，亦或是只具有解碼器，未必適合直接應用於限定性的語言生成任務，而這恰恰是 MASS 的優勢。

---

33 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805* (2018).

34 Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving language understanding by generative pre-training.” *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. Pdf* (2018).

35 Hu Renfen, Shen Li, and Yuchen Zhu. “Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on Deep Language Models.” In *18th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 2019.

MASS 是一種針對低資源場景的訓練方法，全稱是 Masked Sequence to Sequence Pre-training。[36] MASS 具有完整的編碼器——解碼器結構，適用於以生成特定自然語言句子為目的的語言生成任務。

MASS 在低資源（low resource）條件下六種語言對之間的神經網絡翻譯（NMT）和文本摘要等任務上，都大幅超過了基線模型。MASS 還在英語／法語和英語／德語這兩種無監督（即不使用任何對齊語料）神經網絡翻譯任務中超過之前所有的無監督訓練方法，取得了當前最佳結果。

關於數據集的準備方法，以翻譯任務為例，在預訓練階段，MASS 只需要兩種語言各自的單語種語料即可，兩者之間不需要進行對齊。在預訓練完成後，才需要少量的對齊語料進行微調。

利用 MASS，可以將中文「現代文／古文」作為一種語言對，各自準備充分的預訓練語料。再基於大藏經構建「現代文／古文」的標注數據，訓練翻譯模型，從而實現兩種語言的相互翻譯。

[36] Song Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. “Mass: Masked sequence to sequence pre-training for language generation.” *arXiv preprint arXiv:1905.02450* (2019).

引用書目

王川、張小紅、韓采華，2009，〈古漢語句子切分與句讀標記方法研究〉，《河南大學學報（自然科學版）》39：5，頁 525-529。

王博立、史曉東、蘇勁松，2017，〈一種基於循環神經網絡的古文斷句方法〉，《北京大學學報（自然科學版）》53：2，頁 255-261。

伍思航，2019，〈基於強化學習的古籍文字精準檢測〉，廣州：華南理工大學電子與信息學院本科畢業設計。

俞敬松、魏一、張永偉，2019，〈基於 BERT 的古文斷句研究與應用〉，《中文信息學報》33：11，頁 57-63。

馬偉洪，2019，〈基於深度學習的文檔圖像版面分析與文字識別〉，廣州：華南理工大學電子與信息學院本科畢業設計。

張合、王曉東、楊建宇、周衛東，2009，〈一種基於層疊 CRF 的古文斷句與句讀標記方法〉，《計算機應用研究》26：9，頁 3326-3329。

張開旭、夏雲慶、宇航，2009，〈基於條件隨機場的古漢語自動斷句與標點方法〉，《清華大學學報（自然科學版）》10，頁 1733-1736。

陳天瑩、陳蓉、潘璐璐、李紅軍、于中華，2007，〈基於前後文 n-gram 模型的古漢語句子切分〉，《計算機工程》33：3，頁 192-193。

黃建年、侯漢清，2008，〈農業古籍斷句標點模式研究〉，《中文信息學報》22：4，頁 31-38。

釋賢超、方愷齊、釋賢迥、釋賢菊、釋賢碉、釋賢繼、釋賢大、釋賢奉、宋延淳，2019，〈一種自動標點的方法與實現〉，《數位典藏與數位人文》3，頁 1-19。

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In arXiv preprint arXiv:1810.04805.

Hu, Renfen, Shen Li, and Yuchen Zhu. 2019. “Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on Deep Language Models.” In *18th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*.

Huang, Hen-Hsen, Chuen-Tsai Sun, and Hsin-Hsi Chen. 2010. “Classical Chinese sentence segmentation.” In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving language understanding by generative pre-training." *URL https:// s3-us-west-2.amazonaws. com/openai-assets/researchcovers/languageunsupervised/ language understanding paper. pdf.*

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. "Mass: Masked Sequence to Sequence Pre-training for Language Generation." In arXiv preprint arXiv:1905.02450v5 [cs.CL].

Wang, Boli, Xiaodong Shi, Zhixing Tan, Yidong Chen, and Weili Wang. 2016. "A Sentence Segmentation Method for Ancient Chinese Texts Based on NNLM." In *Workshop on Chinese Lexical Semantics*, Cham: Springer, pp. 387-396.

Xie, Zecheng, Yaoxiong Huang, Lianwen Jin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao, and Xiaode Zhang. 2019. "Weakly Supervised Precise Segmentation for Historical Document Images." *Neurocomputing* 350: 271-281.

Xu, Han, Wang Hongsu, Zhang Sanqian, Fu Qunchao, and Liu Jun. 2019. "Sentence Segmentation for Classical Chinese Based on LSTM with Radical Embedding." *The Journal of China Universities of Posts and Telecommunications* 26, 2: 1-8.

Yang, Hailin, Lianwen Jin, and Jifeng Sun. 2018. "Recognition of Chinese Text in Historical Documents with Page-level Annotations." In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), New York: IEEE, pp. 199-204.

Yang, Hailin, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun. 2018. "Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector." *IEEE Access* 6: 30174-30183.

網路資源

古籍酷，http://www.gj.cool, 2020/1/28。

基於深度學習的中國古籍文檔智能理解項目演示系统，
http://115.29.208.50:9000/ punctuation, 2020/1/28。