漢文學論集, 근역한문학회
Journal of Korean Literature in Chinese
2023, 第六十五輯, pp.211~236
http://dx.doi.org/10.17260/jklc.2023.65..211

인공지능 활용 초서 OCR 개발 과정과 과제*

허철** · 조성덕*** · 최동빈****

「차 례」

[국문초록]

고전적 디지털화와 관련된 비용은 지속적으로 증가하고 있으며, 속도 또한 제한적이다. 이를 극복하기 위한 대안으로 오래 전부터 자동화 입력 관련 기술이 논의되었으나 기술적 한계가 있었다. 최근 이러한 어려움과 효율성 제고를 인공지능으로 극복하려는 다수의 노력이 진행되고 있다. 관련 기술 개발의 중요성은 경제적 효율화를 이루는 것뿐 아니라, 한자문화권에서의 한국만의 고유한 원천 기술 확보를 통해 국가 경쟁력이라는 의미도 함께 지니고 있다.

본고는 이러한 현실의 추세에 맞추어 인공지능 활용 초서OCR 개발 과정 경험과 발견 문제점을 공유하고자 한다. 이 연구가 향후 보다 효율적인 초서 OCR 관련 기술 개발의 계기가 될 것으로 기대한다.

주제어: 한문, 한자, OCR 문자인식 초서

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A6A3A02043693)
** 주저자, 경성대학교 한국한자연구소 HK교수 / heochul@gmail.com
*** 교신저자, 경성대학교 한국한자연구소 HK연구교수 / isanjo@nate.com
**** 공동저자, 단국대학교 컴퓨터학과 박사수료 / dbchoi85@gmail.com

1. 서론

오랜 역사 동안 한자로 구축된 다양한 문헌을 컴퓨터로 이용할 수 있는 디지털 데이터로 전환하는 과정은 다른 어떤 문자 기록보다도 많은 어려움이 있다.[1] 특히 대다수의 필사본 고전적의 경우 다른 자형보다 초서가 대부분을 차지하고 있다. 고전적 전산화에서 초서 전적의 디지털화는 불가결하다고 할 수 있으나, 초서 필사본의 경우 전문가에 의한 디지털화나 자동 입력을 위한 도구 개발 등은 불가능한 것으로 인식되었다.

익히 알려진 것처럼 일반적인 한자에 익숙한 입력자 대부분도 초서 자형의 서사 습관에는 익숙하지 않은 경우가 대부분이다. 이는 초서 자형의 독특함에 그 이유가 있다. 초서 자형은 독립적이지 않고 연결되어 있으며, 자형의 크기 또한 서사자의 인식에 따라 대소와 편폭이 결정되며, 필획과 부건의 위치 또한 가변적이다. 하나의 초서 자형은 경우에 따라 몇 개의 자종으로 해서화하한다. 이와 같은 초서의 특성은 해서를 중심으로 한자를 이용하는 대부분의 사람들에게 매우 어려운 숙제로 인식되고 있다.[2]

따라서 초서를 직접 읽고 입력할 수 있는 전문 인력은 늘 부족한 현실이다. 초서 자형에 익숙한 전문가의 수가 적으니 비용은 증가하고 속도는 제한적이다. 이

---

1) 간략하게 서술하면 아래와 같다. 첫째, 한자는 그 수량이 너무 많다. 물론 다량의 문헌에서 가장 많이 사용되는 한자의 자종만을 고려한다면 몇 천자 이하로 제시할 수도 있으나, 문헌 기록의 데이터 과정에서는 단지 1회 출현하였더라고 이를 무시할 수 없다. 현재의 유니코드 개발 이상, 즉 10여 만 자종 이상의 한자를 디지털 변환할 수 있어야 한다. 둘째, 한자는 1개 자종에 대응하여 다양한 자체가 존재한다. 역사적으로 甲骨文부터 草書에 이르기까지 최소한 1개의 대응 자체(字體)부터 많게는 수십 개의 대응 자체가 존재한다. 자종이 8만 개라면 대응 자체는 무려 40만 개이며, 자체 사이의 자체 변화는 규칙적이기도 하지만 규칙적이지 않다. 무규칙성에서 규칙성을 찾아내려는 노력은 그간 한자 학계의 오랜 숙제이기도 할 정도이다. 셋째, 한자 사용상에 다양한 異體字와 通假字 관계가 존재한다. 기존의 이체자와 통가자뿐 아니라 때로는 당시 서사자의 습관이나 관념으로 인해 새로운 자체를 생성하기도 한다. 이렇게 만들어진 자체는 역대 어떤 자전이나 사전에도 존재하지 않는 독특한 것이다. 넷째, 필사, 판각, 인쇄 과정에서 일부 자형이 변화되는 현상이 존재한다. 다섯째, 문헌마다 조판 형태 즉 레이아웃이 너무 다양하다. 글자 크기의 대자와 소자는 물론이고, 조판 과정에서도 상하 좌우 2단 뿐 아니라 3단도 존재하며, 중간에 그림이나 표가가 포함하기도 한다.

2) 대부분의 초서 문헌의 디지털 데이터화 과정에는 반드시 "탈초" 공정이 포함되어 있다. 탈초란 초서 전문가가 해당 자형을 판별하여 해서체로 치환하는 과정을 말한다. 초서를 해서로 치환한 후에야 디지털 데이터를 구축할 수 있는 셈이다. 곧 탈초 공정의 정확도 여부가 해당 문헌 데이터의 품질과 직접적 연관이 된다.

를 극복하는 대안으로 한자문화권 국가 대부분에서 오랫동안 컴퓨터를 이용한 자동 문자 인식 기술(OCR)이 여러 차례 논의되고 시행되었으나 초서의 고유한 특성을 극복하여 만족할 만한 결과를 얻지는 못하였다. 그러나 최근 다양한 인공지능을 활용한 이미지 인식 기술이[3] 발달하면서 탈초 공정에도 새로운 전기가 마련되고 있다. 중화권뿐 아니라 국내에서도 고문헌 소재 한자나 초서 OCR과 관련하여 다수의 연구가 진행 중이며 그 결과로 인식률 몇 %라는 제호로 지상에 소개되고 있으나[4] 구체적인 실험 결과와 그 과정에 대해서는 상세히 소개되지 않고 있다.

본고는 우리나라와 중국의 초서 이미지 데이터를 수집하고 전문 정제 과정을 거친 후, 한자 OCR에 활용할 수 있는 다수의 알고리즘을 투입하여 실험 후 최종 결과가 도출되는 일련의 과정을 소개하고자 한다. 이를 통해 실험 과정 속에서 발견된 성과와 문제점을 공유하여, 본 연구의 경험이 향후 관련 기술 개발은 물론 관련 연구자들의 이해에 도움이 되기를 희망한다.

---

3) 최근 10년 사이 우리 사회 각 분야에서 다양한 형태의 인공지능 활용 연구와 서비스가 제공되고 있다. 번역, 음악, 미술, 내용 요약, 쓰기, 그리고 최근의 챗GPT에 이르기까지 특정 데이터를 학습하여 학습된 결과물을 제시하는 모든 분야에서 인공지능은 이미 인간 숙련자의 능력을 넘어서기 시작하였다. 이 과정에서 인공지능이 가지는 기술적•윤리적 한계도 분명하다. 인공지능의 결과물에 대한 우리의 윤리적 판단이나 추론, 인식의 과정에 대한 명확한 인식뿐 아니라 결과물의 정확도와 효율성에 대한 명확한 관찰과 취사선택도 필요하다. 지금의 한계에도 불구하고 우리가 인공지능이라는 새로운 기술 방식을 선호하고 개발하여 사용하는 근본적인 이유는 인간 스스로가 "편안함"과 "효율"을 추구하기 때문이다. 반복적이고 복잡한 여러 작업을 인간이 아닌 기계가 쉽고 빠르게 대신할 수 있다는 인식과 기계로 인간의 노동을 줄이거나 대신하자는 지향은 먼 옛날 인간이 "도구"를 개발하고 사용하면서부터 시작되었다. 글자를 대량으로 인쇄하기 위해 활자와 판각 등 인쇄술이 발달하였고, 먼 거리를 효율적으로 이동하기 위해 다양한 이동 수단이 발달하였다. 이는 고전적 한문 문헌의 디지털 영역에서도 동일하다.

4) 어떤 기사에서도 가장 중요한 개별 자종과 자체에 따른 인식률을 정확한 정보로 제공하지 않고 있다. 실험 대상이 목판인지, 활자인지, 필사인지 혹은 해서인지, 예서인지, 행초서인지, 그리고 어떤 한자 자종에서 인식률이 얼마인지 등에 대한 정확한 정보를 제공하지 않는다. 대개의 경우에 데이터셋을 자체 구축 혹은 기존 데이터를 이용하여 한자인식 알고리즘을 개발한 결과 몇%의 인식 결과가 나왔다는 결과만 신문 보도나 외국계 잡지에 소개하는 정도이다. 미래 연구자를 위해 보다 구체적으로 어떤 데이터 셋과 알고리즘을 활용하여 무엇을 대상으로 실험한 결과 어떤 자종에서 어떤 결과가 나오지는 소개하여 차기 연구에 도움이 되어야 한다.

2. 한자 관련 OCR 개발 현황

(1) OCR의 기본 이해

본고에서 논의하는 OCR[5]이라는 용어를 이해하기 위해서는 아날로그 판본을 컴퓨터가 이해할 수 있는 가독형 언어로 변환하는 두 가지 방식을 이해할 필요가 있다. 일반적으로 특정 아날로그 대상물을 컴퓨터가 이해하는 가독형 언어로 변환할 때 두 가지 방식을 사용한다. 첫째 방식은 디지털 카메라 혹은 스캐너를 통해 사진을 찍거나 스캔하여 디지털 데이터로 변환하는 방식이다. 예를 들어, 대상물을 사진으로 찍거나 스캔하는 등의 방법으로 디지털 데이터로 변환하는 방법이다. 사용자는 컴퓨터로 입력할 이미지 데이터의 전체 크기를 설정하여 픽셀을 결정한다. 입력된 데이터는 실제로는 각 픽셀의 위치 정보인 좌표와 그 속에 포함되는 색의 정보 등은 디짓으로 재구성되어 디지털 데이터로 저장된다. 두 번째 방식은 한 단계를 더 거쳐 문자 인식을 하는 방법이다. 문자인식이란 외부로부터 입력된 이미지와 컴퓨터 안에 이미 정해진 문자 관련 이미지 정보가 일치 혹은 유사할 경우 이를 이미지의 정보의 디짓이 아닌 해당 문자코드 정보로 변환함을 말한다. 따라서 이미지의 정보 자체를 기록하거나 활용하지 않고, 이미 컴퓨터에 비교 대상이 되는 문자의 이미지와 해당되는 문자코드의 정보만이 존재한다.

두 방식은 목적에 따라 다른 결과를 도출한다. 첫 번째 방식은 이미지를 재구성하는데 필요한 방식이며, 두 번째 방식은 이미지는 처리 과정의 중간 과정일 뿐 최종 결과물은 문자코드로 남는다. 문자코드로 치환되기 때문에 이 데이터는 다양한 언어 자료로 활용이 가능하다. 이 방식은 수요자가 이미지를 시각적으로 보고 자형을 파악하여 해당 문자를 손으로 입력하여 디지털화하는 방식과 비교할 때 속도 면에서 큰 차이가 있다. 문제는 속도는 확보할 수 있으나 컴퓨터가 실행한 문자화의 정확도이다. OCR의 평가 기준은 속도가 아니라 인식률 곧 정확도이다.

---

5) OCR(optical character recognition)은 이미지 상에서 글자가 있는 부분과 없는 부분을 확인하기 위해 빛을 투사하고 반사된 빛을 수집하여 이용한다. 곧 색의 다름을 이용하는데, 현재 논의하는 인공지능 문자 인식기도 원리상으로는 같다.

따라서 정확도를 상승시키기 위해 알고리즘 상에 사람이 서사한 혹은 인쇄한 자형과 같은 비교 데이터 즉 다수의 스크립트가 존재할 때 이를 모두 포함하여 비교할 대상을 증가시켜 놓거나, 디짓으로 변화된 이미지와 이미 입력된 이미지의 매칭률을 높이기 위해 디짓화된 이미지를 가공하는 과정 등을 사용한다. 예를 들어 글자 주변에 노이즈가 많아 글자와 노이즈의 판별이 어려운 경우 글자만 부각시키는 기술을 적용한다거나 글자의 장평이 이미 설계한 장평과 다를 경우 자동으로 조정하는 경우이다.

이런 이유로 라틴문자나 아라비아 숫자와 같이 대상 문자의 형태가 단순할수록 OCR는 좋은 효능을 발휘한다.[6] 반면 획과 점 하나로도 구분되어 10만 종이 넘는 모양을 가진 한자의 경우는 OCR 개발이 원천적으로 어려운 한계를 지니고 있다.[7] 특히 기존의 원형정합(template matching)이나 통계(statistical), 구조 분석(structural analysis) 등의 방법으로는 초서는 말할 것도 없고 해서 자형 조차도 OCR개발의 한계가 명확하였다.[8] 그러나 최근 인공신경망을 이용한 다양한 이미지 처리 알고리즘은 한자 대상 OCR 개발의 가능성과 정확도를 높였다. 중국에서 개발되어 상용화되고 있는 OCR의 경우 해서 자형을 대상으로한 실험에서 매우 높은 인식률을 보이고 있으며, 초서 자형 또한 인식률이 높이 상승하고 있다. 그렇다면 인공지능을 활용한 OCR은 어떤 원리로 작동하는 것인지 살펴볼 필요가 있다.

---

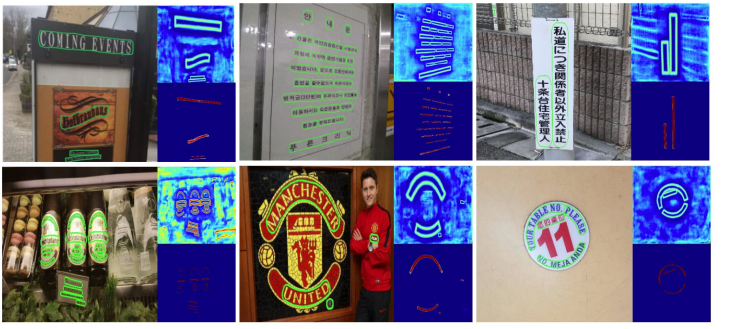
6) 한글도 비교적 단순한 문자이기에 개발이 용이하다고 생각하기 쉬우나 이는 오해이다. 한글은 자음과 모음을 정방형에 모아놓는 집합의 형태이기 때문에 24개의 자모를 조합할 경우 수백 개의 자종이 존재하기 때문에 라틴문자나 숫자에 비해 난이도가 높다고 알려져 있다.

7) 우리에게 전래된 많은 문헌들은 판형의 형태가 다양할 뿐 아니라, 개별 시대마다 자형과 자체의 특징이 다르며, 대상이 되는 문헌의 품질 상태도 각기 다양할 뿐 아니라, 서론에서 밝혔듯 너무 많은 한자가 있다는 것은 인식률에도 큰 영향을 미치게 된다.

8) 원형정합은 입력된 문자 패턴과 원형패턴과 비교하여 가장 유사한 형태를 찾아내는 방법이다. 다만 입력된 문자 패턴이 일정할 수 없고, 원형 패턴을 그만큼 증가시킬 수 없다는 한계를 지니고 있다. 이후 등장한 통계기반은 인식 대상에서 특정 벡터를 추출하여 확률 분포를 구성하고 특정 벡터 공간을 이용하여 매칭하는 방법이다. 이 또한 원형정합에서의 데이터 한계를 극복하기 어려운 문제를 여전히 지니고 있다. 이후 등장한 구조분석법은 해당 언어에서의 문자가 사용될 때 문자 구성 원리를 기초하고 필획과 위치 배열 요소 등과 기본원칙과 연관성을 추출하여 이용하는 방법이다. 이 또한 개별 문자의 특성을 미리 설계해야 할 뿐 아니라 복합적으로 다종 문자가 포함될 경우 인식률이 저하되는 문제점을 극복할 수 없었다.

2) 인공지능 활용 OCR 연구 현황

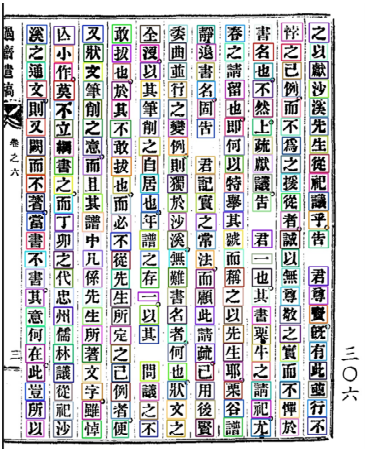
문자 OCR 개발을 촉진시킨 중요한 사건은 하나의 이미지로부터 여러 객체를 탐지할 수 있는 인공지능 이용 모델의 개발이었다. 〈그림1〉의 Scene text detection은[9] 근에 표지판 인식 등 자율주행 관련 기술 개발에서 활용하는 기술로, 해당 이미지 중 어느 부분에 글자가 있는지를 검색 후 그 글자가 무슨 뜻인지 파악하는 알고리즘의 한 예이다.

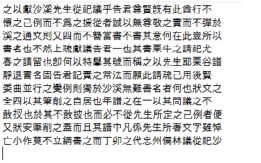


〈그림1〉 인공지능 모델을 통한 글자 위치 추적

이와 같은 객체탐지 모델 중 하나인 Mask-RCNN을 이용하여 한자를 학습하면 아래 [그림2]와 같은 OCR 기능이 구현된다.

9) Liao M, Wan Z, Yao C, Chen K, Bai X (2019) 참조.





〈그림2〉 객체 탐지 모델을 이용한 한자 OCR

문제는 기술의 측면에서 볼 때 객체 탐지 모델은 다양한 종류가 개발되었을 뿐 아니라 모델에 따라 개별 대상물에 따른 성능 또한 다양한 것으로 알려져 있다는 점이다. 곧 하나의 모델이 모든 것에 동일한 성능을 보이지 않기 때문에 대상물에 적합한 모델을 선택하는 과정이 요구된다. 모델에 따른 성능의 차이는 〈표1〉과 [그림3]을 통해 알 수 있다.

Methods	Pretraining	Scene	Web	Document	Handwriting	Avg	#Params
CRNN [42]	×	53.4	54.5	97.5	46.4	67.0	-
ASTER [43]	×	54.5	52.3	93.1	38.9	64.7	-
MORAN [33]	×	51.8	49.9	95.8	39.7	64.3	-
SAR [26]	×	62.5	54.3	93.8	31.4	67.3	-
SRN [50]	×	60.1	52.3	96.7	18.0	65.0	-
SEED [39]	×	49.6	46.3	93.7	32.1	61.2	-
TransOCR[4]	×	63.3	62.3	96.9	53.4	72.8	84M
MaskOCR (ours, ViT-S)	√	71.4	72.5	98.8	55.6	78.1	36M
MaskOCR (ours, ViT-B)	√	73.9	74.8	99.3	63.7	80.8	100M
MaskOCR (ours, ViT-L)	√	**76.2**	**76.8**	**99.4**	**67.9**	**82.6**	318M

〈표1〉 인공지능 모델별 성능 비교

〈표1〉의 Pretraining의 항목은 ImageNet이라는 대규모 이미지 학습데이터를 사용하여 사전 학습 진행 여부를 나타내며, Scene은 도로나 거리와 같은 사진에서의 글자 인식, Web은 웹 브라우저 환경에서 글자 인식, Document는 문서 내

글자, Handwriting은 필기체에 대한 글자 인식 성능을 말한다.[10] 앞의 methods는 모델명을 말하며, Avg는 평균적인 인식률을 말한다. 객체 탐지 모델들은 CNN(Convolution Neural Network)이라는 이미지 처리 모듈을 기반으로 여러 다른 모듈을 연결해 놓았기에, 이때 사용된 CNN을 backbone이라고 부른다. [그림3]은 2022년 Pengyuan Lye et al[11]에 수록된 성능비교표로 여러 Scene text detection 모델들의 성능을 검토한 것이다. 일반적으로 CNN 중 VGG(Visual Geometry Group)라는 모델을 많이 사용한다. VGG[12]는 CNN을 구성하는 주요 구성물인 컨볼루션 레이어를 허용 가능한 만큼 최대한 많이 쌓는 모델이다. 레이어의 숫자에 따라 16개의 컨볼루션 레이어를 지닌 VGG16과 19개의 컨볼루션 레이어를 지닌 VGG19로 구분한다. 본 CNN 알고리즘의 경우 특성상 컨볼루션 레이어를 통과할 때마다 일정량의 정보 손실이 발생한다. 순차적으로 레이어를 통과하다보니 초기 정보가 다음에서 다음 레이어로 옮겨지면서 정보 손실이 발생하게 되는 셈이다. 이처럼 여러 컨볼루션 레이어를 연결한 VGG는 19개를 초과할 경우 정보 손실이 극대화되어 성능이 하락하는 문제가 발생하는 것으로 알려졌다. 이런 단점을 극복하기 위해 개발된 ResNet[13]계열은 〈그림3〉[14]처럼 컨볼루션 레이어 사이에 우회로를 설치하여 상단 컨볼루션 레이어의 정보가 하단 컨볼루션 레이어를 거치지 않고 더 아래의 레이어로 정보를 전달하는 방식을 선택하는 모델이다. 우회로가 있기 때문에 순차적이 아닌 선택적 레이어가 가능해지고, VGG계열보다 더 많은 컨볼루션 레이어를 쌓을 수 있게 되었다. 최근 여러 객체 탐지모델의 backbone으로 사용되는 ResNet152V2가 대표적이다. 이 모델은 숫자 152에서 알 수 있듯이 152개의 컨볼루션 레이어를 쌓은 것으로, 현재 V2 즉 2번째 개선된 모델까지 개발되었다.

한자 관련 OCR의 경우 앞서 언급된 Scene text detection의 기법과 유사한 알고리즘이 사용되고 있다. 탐지된 객체, 즉 글자를 개별 글자로 남길 것이냐 아니면 합쳐서 하나의 글자로써 인식할 것이냐의 차이가 존재한다. 이는 앞서 제시한

10) Dong, H.; Cheng, Z.; et al.(2022)참조.
11) Pengyuan Lyu etc.(2022) 참조.
12) Simonyan, K. & Zisserman(2014).참조.
13) He K, Zhang X(2016) 참조.
14) He K, Zhang X,(2016)참조.

성능평가 모델들은 연속된 글자들을 인식하여 하나의 글자 혹은 문장으로 평가하는가에 따라 달라진다.

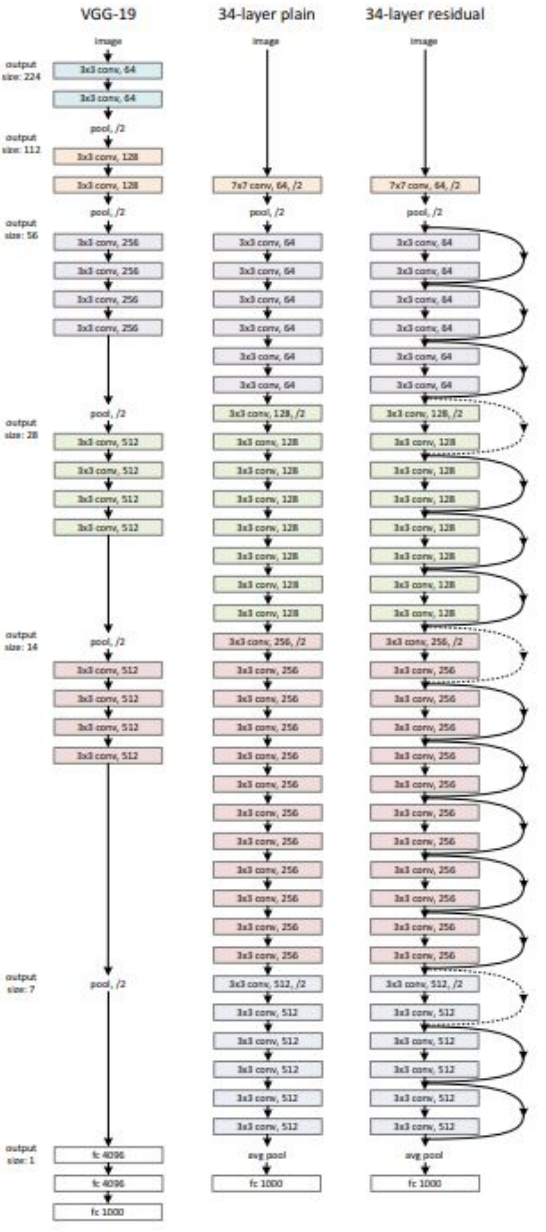
이처럼 동일 대상도 모델 선정에 따라 각기 다른 결과를 보이기 때문에 각 모델 중 본 연구의 실험에 가장 적합한 모델을 찾아내는 과정은 필수적이다.

한편, 본 실험에서 중요한 것은 학습데이터와 테스트데이터 등을 포함한 데이터셋이다. 인공지능 학습 모델의 경우 여러 연구를 통해서 목표로 하는 이미지와 관련 없는 이미지까지 포함하여 대규모의 학습데이터를 미리 학습시키면 목표 이미지만 학습하는 경우와 비교할 때 오히려 학습 속도가 빠르며 성능 또한 높게 나타난다는 결과가 있다.[15] 결국 대량의 데이터가 필요했고, 이 데이터는 학습뿐 아니라 테스트와 검증을 위해서도 사용된다. 주지하듯 인공지능 학습은 Trainning 데이터를 이용하여 모델에 따라 학습하게 한 후 결과를 검증하기 위해 Test 데이터를 투입한다. 이때 인공지능 모델은 학습을 통해 Test 데이터가 정답값에 가장 가깝게 구현됨에 목표를 둔다. 다만 Test 데이터만을 검증하기 위한 목적으로 학습이 일어나기 때문에, Test 데이터만 잘 맞추는 overfitting 현상이 발생할 수 있는데, 이를 감지하기 위해 최종 학습 후 별도로 준비한 Valid 데이터로 성능 평가를 진행하는 과정을 포함하는 것이 일반적이다..[16]

곧 데이터셋은 학습과 테스트, 검증을 위한 기본적인 데이터인 셈이고, 이 데이터는 대량과 정확함이 우선되어야 한다. 따라서 본 연구에서는 객체를 탐지하는 수많은 모델 중 초서 자형 인식에 가장 적합한 모델을 선별하는 실험을 위해 실험에 투입할 정제 데이터셋을 준비하였다.

15) 〈표1〉에서도 pretraining의 유무가 성능에 영향을 미친 것을 통해서 확인할 수 있다.

16) Train은 학습에 사용된 글자의 수량, Test는 학습 결과를 확인할 때 사용되는 글자의 수량, Valid는 학습이 완료된 후 최종 성능 평가에 사용되는 글자의 수량을 말한다.



〈그림3〉 VGG와 ResNet의 비교도

3. 초서 OCR 개발 실험 과정

언급하였듯 인공지능 활용 기술 개발에는 데이터셋과 모델의 두 축이 필요하다. 본 연구에서도 이 두 축을 중심으로 연구를 진행하였다.

1) 학습데이터 정제

데이터셋을 구축하기 위해서는 무엇보다 목적에 가장 적합한 데이터를 수집하고 정제하여 최대한 정확도가 높은 데이터를 구축해야 한다. 중국의 경우 다양한 문헌과 조판 형태, 고문자와 현대 한자까지 다양한 고문헌 처리 관련 데이터셋이 개발되어 연구에 활용되고 이를 통해 상업화까지 진행되고 있다.[17] 반면 한국의 경우 이용이 가능한 데이터셋이 제한적이다. 한국학중앙연구원에서 개발한 한국고문서 서체 용례사전[18]과 AI HUB에서 제공하는 고서한자인식OCR 데이터[19] 등 2종이 이용 가능하였다. 본 연구에서는 모델에 투입할 데이터셋을 준비하기 위해 한국학중앙연구의 "한국학자료포털"을 통해서 초서 85,285자, 행서 184,800자를 크롤링하였다. 크롤링 결과 수집된 최종 결과는 유니코드 4,537종에 해당하는 총 287,095개의 이미지였다.

그러나 자종에 따라 충분한 양의 데이터가 확보되지 않아 수집 결과를 바탕으로 충분히 수집되지 않은 한자 자종 2,227종을 중국의 "書法大師" 사이트를 통해서 추가로 수집하였다. 추가로 수집된 한자는 1,567종 17,010개 이미지였다.

본 데이터를 사용하기 이전에 데이터의 정확성 유무에 대한 정합성 검토가 필요하였다. 데이터의 정합성에 따라 학습과 도출 결과물이 달라지기 때문이다.

---

17) 대표적인 데이터셋은 다음과 같다.
Tripitaka Koreana in Han (TKH) Dataset, Multiple Tripitaka in Han (MTH) Dataset, MTH v2 Dataset (Extension Version), 古籍单字数据集：CASIA-AHCDB-中科院数据集, 中文古籍族谱数据集：HDRC-Chinese, 甲骨文数据集Oracle-20K, Oracle AYNU, OBC306, 外文古籍数据集：IAM-HistDB, 公开数据集：DIVA-HisDB精细标注的中世纪手稿, 公开数据集：HJDataset 复杂版面的日语文档, 公开数据集：READ-BAD欧洲历史档案文件, 公开数据集：REID2019早期的印度印刷文档, 公开数据集：MapSeg2021历史地图文档

18) https://kostma.aks.ac.kr/segment/segmentList.aspx

19) https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=603

검토 중 수집한 기존 데이터가 많은 오류를 내포하고 있음을 확인하여, 기존 데이터의 문제를 해결하는 과정이 필요했다. 오류가 있는 데이터를 학습할 경우 당연히 학습 알고리즘에도 오염이 될 뿐 아니라 실험 자체를 신뢰하기 어렵기 때문이었다. 검토 과정 중 발견된 오류의 유형은 (1) 자형을 오독한 경우, (2) 앞뒤 글자의 순서가 바뀐 경우, (3) 단순 자형오류인 경우, (4) 한국식 한자의 아래 부분이 누락된 경우, (5) 동일음가 한자를 잘못 분류한 경우, (6) 유사자형 한자를 잘못 분류한 경우, (7) 이미지의 각도가 잘못된 경우, (8) 불량한 이미지가 포함된 경우, (9) 여러 개의 글자가 포함된 경우 등의 9종류로 구분할 수 있었다. 각각의 예를 보면 다음과 같다.

(1) 자형을 오독한 경우

繡를 稱으로(稱 ← ), 粉을 粘으로(粘 ← ), 問을 古로(古 ← ), 權을 栰로(栰 ← ), 祝을 祀로 분류한 경우(祀 ← ) 등이 확인되었다.

(2) 앞뒤 글자의 순서가 바뀐 경우

上下에서 下를 上으로, 不有~에서 有를 不로, ~學錄에서 錄을 學으로, 疾病에서 疾을 病으로, 四年에서 四를 年으로, 捧上에서 上을 捧으로, 壬癸庚申에서 庚을 癸로, 行下에서 行을 下로 분류한 경우 등이 확인되었다.

(3) 단순 자형 오류인 경우

衛를 尹으로, 之를 汲으로, 大를 籃으로, 保를 證으로, 一을 石으로, 叱을 '臥'로 분류한 경우 등이 확인되었다.

(4) 한국식 한자의 아래 부분이 누락된 경우

廛을 庫로(), 𡈽을 末로(), 岔을 分으로 분류한 경우() 등이 확인되었다.

(5) 동일음가 한자를 잘못 분류한 경우

相()을 上으로, 貳()를 二로,  (一)을 日로, 斫()을 作으로, 

(脩)를 修로,  (庭)을 定으로,  (傳)을 專으로,  (小)를 少로,  (天)을 川으로,  (嗟)를 差로,  (夫)를 府로,  (貼)을 帖으로,  (遑)을 徨으로,  (此)를 徨으로,  (呈)을 正으로,  (嬪)을 殯으로,  (之)를 池로,  (忱)을 沈으로,  을 洞으로,  (談)을 淡으로,  (宗)을 種으로,  (拾)을 習으로,  (是)를 詩로,  (訂)을 證으로,  (驚)을 警으로,  (臣)을 辛으로,  을 轉으로 ,  (告)를 高로,  (渝)를 喩로,  (修)를 壽로,  (果)를 科로,  (藏)을 臧으로,  (不)를 部로,  (陸)을 六으로,  (錢)을 前으로,  (則)을 卽으로,  (聞)을 問으로,  (監)을 感으로,  (乘)을 承으로,  (正)을 政으로,  (始)를 是로,  (腹)을 服으로,  (祖)를 朝로,  (復)을 福으로,  (棄)를 其로, (倒)를 到로, (製)를 制로, (畫)을 劃으로, (男)을 南으로, (園)을 圓으로, (棄)를 其로, (玆)를 慈로, (姑)를 故로, (雖)를 數로, 斫을 昨으로, (母)를 眸로, (婢)를 碑로, (巡)을 順으로, (婢)를 批로 분류한 경우 등이다.

(6) 유사 자형 한자를 잘못 분류한 경우

住를 任으로, 列을 例로, 伐을 代로, (宏)을 宏으로, (宣)을 宜으로, (字)를 子로, (宮,宮)을 官으로, (欲)을 就, (竹)을 幷으로, (從)을 徒로, (洗)을 決로, (沈)을 沐으로, (步)를 涉으로 , (瑞)를 端으로, (綻)을 綜으로, (美)를 義로, (意)를 言으로, (誠)을 誠으로 (論)를 論으로, (賊)을 賤으로, (蹈)를 踏으로, (幸)을 辛으로, (麃)를 鹿으로, (兀)을 几으로, (靳)을 勒으로, (勒)을 靳으로, (勒)을 勤으로, (乃)를 及으로, (反)을 及으로, (古)를 右로, (晳)을 哲로, (回)을 回로, (姝)를 妹로, (扶)를 拔로, (愴)을 拾으로, (採)를 探으로, (欺)를 斯로, (雜)을 新으로, (拖)를 施로, (葉)를 柴로, (寀)을 棠으로, (互)를 瓦로, (盃[鬱])을 盃로, (婢)를 秘로, (祇)를 秖로, (苦)를 若으로, (歲)를 若으로, (壓)을 釐로, (雙)을 隻으로, (雄)을 雜으로, (雖)를 難으로, (焉)을 馬로, (依)를 倚로, (措)를 借로, (旣)를 卽으로, (罔)을 同으로, (恩)을 思로, (捐)을 損으로, (累)를 果로,

 (之)를 枝로,  (玉)을 王으로,  (瑋)를 瑞로,  (其)를 甚으로,  (姓)을 生으로,  (服)을 眼으로,  (杜)를 社로,  (關)을 開로,  (叵)를 區로,  (郎)을 即으로,  (名)을 各으로,  (基)를 墓로, (懍)을 慄로, (愛)를 慶으로, (久)를 文으로,  (村)을 材로,  (材)를 林으로,  (畜)을 畜으로,  (祭)를 癸로,  (敬)을 的으로,  (虛)를 處로,  (皿)을 血로,  (衰)를 襄으로,  (達)을 違로,  (遺)를 遣으로,  (陛)를 陸으로,  (聘)을 聘으로 잘못 분류한 경우 등이다.

(7) 이미지 각도가 잘못된 경우

上 ←  , 不 ←  , 濟 ←  , 第 ← , 等 ←  , 己 ←  , 三 ←  , 宣 ← , 泰 ← , 少 ←  , 人 ←  , 翼 ←  , 許 ←  , 講 ←  , 謝 ← , 謹 ← , 輿 ← , 友 ← , 患 ← , 惟 ← , 拜 ← , 方 ← , 日 ← , 暫 ← , 書 ← , 直 ← , 矣 ← , 華 ← , 金 ← , 鎭 ← , 韓 ← , 八 ← , 前 ← , 好 ← , 敎 ← , 李 ← , 百 ← , 而 ← , 落 ← , 行 ← , 見 ← , 候 ← , 兄 ← , 先 ← , 入 ← , 廿 ← , 圖 ← , 其 ← , 慈 ← , 憂 ← , 撓 ← , 朴 ← , 東 ← , 照 ← , 爲 ← , 者 ← , 耳 ← , 送 ← , 適 ← , 餘 ← .

(8) 불량한 이미지가 포함된 경우

① 글자 훼손

不 ← , 作 ← , 着 ← , 祈 ← , 考 ← , 者 ← , 衙 ← , 錄 ← , 處 ← , 王 ← , 非 ← , 遣 ← , 褒 ← , 相 ← , 是 ← , 敬 ← , 推 ← , 志 ← , 庚 ← , 今 ← , 今 ← , 下 ← , 三 ← , 題 ← , 郡 ← , 道 ← , 白 ← , 漢 ← , 月 ← , 德 ← , 千 ← , 全 ← , 金 ← , 等 ← , 消 ← , 婢 ← , 妻 ← , 咸 ← , 目 ← .

② 글자 흐림

作  , 的 ←, 皆 ←  , 着 ←  , 蔡 ←  , 行 ←  , 是 ←  , 指 ←  , 彩 ←  , 敬 ← .

(9) 여러 개의 글자가 포함된 경우

① 상하나 좌우의 글자 포함된 경우

不 ←  , 升 ←  , 月 ←  , 令 ←  , 香 ←  , 行 ←  , 聃 ←  , 相 ←  , 病 ←  , 父 ←  , 是 ← , 指 ←  , 悌 ← , 志 ← , 尸 ← , 察 ←  , 宿 ←  , 六 ←  , 金 ←  , 乙 ←  , 三 ←  , 鴈 ←  , 鎭 ←  , 量 ←  , 知 ←  , 反 ←  , 申 ←  , 日 ←  , 尹 ←  , 子 ←  , 姓 ← , 夫 ← , 太 ← , 右 ← , 內 ← , 先 ← , 全 ← , 一 ← , 者 ← , 折 ← , 名 ← , 十 ← , 北 ← , 八 ← , 班 ← , 玉 ← .

② 小字와 大字가 쌍행으로 포함된 경우

不 ← , 行 ← , 烈 ← , 極 ← , 史 ← , 題 ← , 掘 ← , 面 ← , 邦 ← , 百 ← , 日 ← , 廿 ← , 使 ← , 一 ← , 聞 ← , 禁 ← , 石 ← , 此 ← , 尙 ← , 尺 ← , 少 ← , 尤 ← , 吏 ← , 公 ← , 八 ← , 任 ← , 館 ← .

검토 결과 위에 제시한 오류 유형 중 '(5) 동일음가 한자를 잘못 분류한 경우'와 '(6) 유사자형 한자를 잘못 분류한 경우'에서 가장 많은 오류가 발견되었다.

이외에도 한국학중앙연구원 "한국학자료포털" 행초자료에는 浸 ←  , 秩 ←  , 生 ← , 耿 ← , 甲 ←  , 禁 ← , 邑 ←  등 첩자 표시가 포함된 경우와 息 ←  (自), 斗 ←  , 褒 ←  등처럼 글자와 윗부분이나 아랫부분이 잘린 사례도 발견되었다.

위에서 언급한 예를 모두 수정하여 정제된 데이터를 실험에 사용하였다.

3) 실험 과정 및 결과

본 실험의 목적은 초서 이미지를 인식하여 탈초를 얼마나 정확하게 하는가를 알아보는데 있으므로, 초서 이미지 인식에 최적화된 모델을 찾는 과정은 필수적이다.[20]

일반적으로 CNN에 사용된 컨볼루션 레이어의 수가 많을수록 성능도 좋아지나, 그만큼 계산량이 많아지기 때문에 높은 사양의 컴퓨터와 많은 학습 시간이 소요된다. 따라서 VGG계열과 ResNet계열 중 학습데이터 및 목적에 맞는 모델을 파악하고자 수집된 학습데이터를 이용하여 VGG16, VGG19, ResNet50, ResNet152V2 모델들을 적용하여 평가하였다. 적은 컨볼루션 레이어를 사용한 모델로도 충분한 성능을 보인다면, 굳이 더 많은 컨볼루션 레이어를 사용한 모델을 이용하여 많은 시간과 자원을 사용할 필요가 없기 때문이었다.

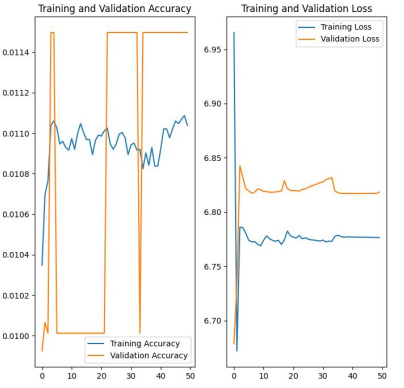
평가 및 학습에 사용된 학습데이터는 초서로 된 유니코드 4,537종이었다. 이중 이미지가 1개만 존재하는 유니코드 382종 및 이미지 10개 미만인 유니코드 2,044종을 대상으로 데이터 증폭을 실시하여 최종적으로 실험에 투입한 데이터는 총 286,153개의 이미지였다. 학습에는 전체 이미지 286,153개중 80%인 228,923개를 성능 평가에 사용하고 나머지 20%인 57,230개는 학습에 사용하였다.

각 모델별 성능 실험 결과는 다음과 같다.

VGG16에 대한 학습 및 평가 결과는 〈그림4〉이다. 파란색 선은 학습 성능을

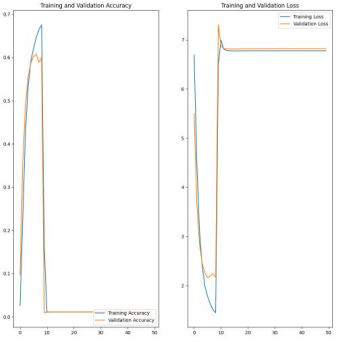
20) 본 실험의 목적은 기계의 탈초 인식률을 알아보는 것이다. 따라서 초서가 포함된 문서 전체에서 한자를 인식하고 구분하는 세그먼트 모델 개발 실험은 차후 과제로 남기고 단일 자종에 단일 이미지만을 투입하였다. 실험 모델 또한 객체 탐지와 같은 복잡한 모델이 아닌 CNN 중 여러 분야에 적용되는 VGG와 ResNet을 투입하였다.

노란색 선은 평가 성능을 나타낸다. x축은 학습량을 나타내는 epoch 값이며 y축은 정확도를 나타낸다. VGG16은 학습할 때 최대 정확도가 1%를 넘기지 못했으며, 이는 성능 평가에도 동일하였다. 다음 〈그림4〉에서도 결과를 확인할 수 있다.



〈그림4〉 VGG16 학습 및 성능 평가〉

한편 레이어를 19층으로 쌓는 VGG19에 대한 학습 및 평가는 〈그림5〉와 같다. 학습할 때 최대 정확도는 68%로 VGG16보다는 높은 정확도를 보이며 평가에도 최대 정확도 60%를 달성하였다. 그러나 학습이 지속됨에 따라 성능이 오히려 하락하는 모습을 보이고 있다.

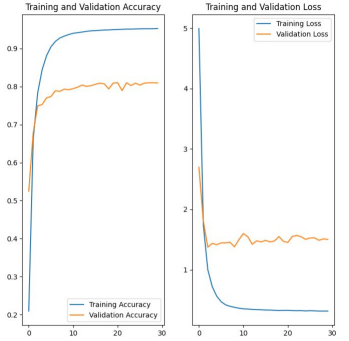


〈그림5〉 VGG19 학습 및 성능평가

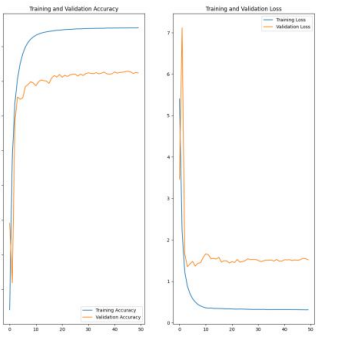
위와 같은 결과를 보인 것은 VGG계열은 단순히 CNN레이어를 중첩으로 쌓아

만든 모델이기 때문에 CNN레이어를 거칠수록 일어나는 정보축약 그리고 그로 인한 정보 손실 때문으로 추정된다.

ResNet계열을 실험한 결과는 〈그림6〉, 〈그림7〉과 같다. ResNet50, ResNet152V2 모두 학습 정확도는 95%이상 달성하였으나 평가시 정확도만 놓고 비교해보면 ResNet152V2 모델이 82.3%로 더 높게 나타났다. ResNet계열의 특성인 우회로의 존재로 정보 손실 없이 학습된 결과가 결국 더 높은 성능을 보여주는 것으로 나타난 것으로 파악할 수 있다.

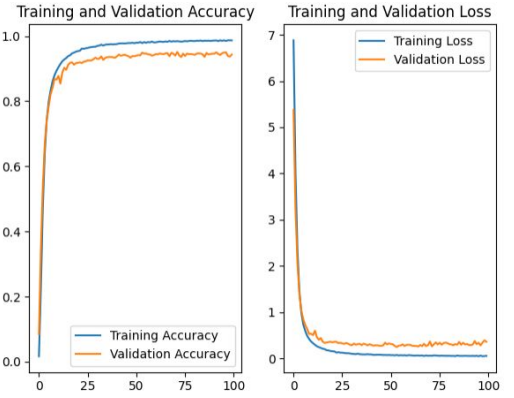


〈그림6〉 ResNet50 학습 및 성능 평가

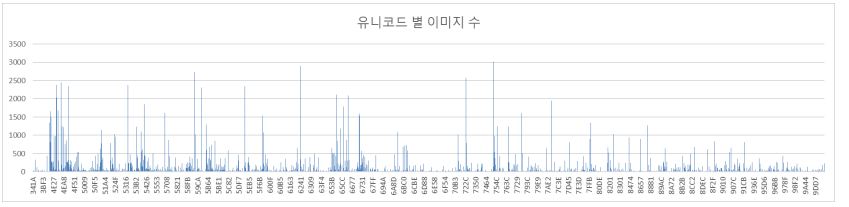


〈그림7〉 ResNet152V2 학습 및 성능평가

만약 VGG계열이 과도한 CNN레이어의 사용으로 인해 정보 손실이 일어나고 이로 인해 성능이 나타나지 않는다고 생각하면, CNN레이어를 줄일 경우 제대로 된 성능이 나타날 수 있을 가능성이 있었다. 따라서 CNN레이어 3개만을 사용한 Simple CNN model을 구성하여 추가 실험을 진행하였다. 그 결과 〈그림8〉와 같이 테스트 정확도 97%를 달성하였다. 단순 결과로는 성능이 높았으나, 평가데이터를 좀 더 분석한 결과 유니코드 1,672개 중 882개의 validation 정확도는 0%, 1672개 중 19개의 validation 정확도는 100%로 학습데이터 중 많은 데이터가 있는 한자 자종만의 인식률이 높았다. 이는 모델 자체가 대상 한자 자종을 골라서 학습한 결과로 보인다. 실험에 투입한 한자 자종이 자종당 동일한 이미지 숫자가 아닌 것이 문제였다. 이는 〈그림9〉에서 확인할 수 있다.



〈그림8〉 Simple CNN model 학습 및 성능 평가〉



〈그림9〉 학습데이터의 분포

한편 본 실험에 적용한 학습데이터는 유니코드 4,536종으로 그중 3,015개의 이미지를 지닌 유니코드 “751F”가 있는가 하면, 이미지가 1개뿐인 유니코드 또한 412종이나 존재했다. 유니코드 별로 학습데이터가 고르게 분포되어 있지 않기에 Simple CNN의 문제가 발생할 수 있다고 추정하였다. 이 문제를 보완하기 위해서는 자종 당 전체 이미지수를 유니코드 수로 나눈 평균 56.4 즉, 자종 당 최소 57개의 이미지가 있어야 좀 더 제대로 된 학습데이터의 역할을 수행 가능할 것으로 판단되었다.

이러한 추정을 토대로 이미지 수가 100개 보다 많은 자종은 샘플링을 통해서 이미지 개수를 줄였으며, 이미지 개수가 100개 보다 적은 자종은 데이터 증폭을 통해서 새롭게 이미지를 생성하였다. 데이터 증폭은 단 1개의 이미지만 존재하는 자종의 경우 이미지의 크기 조절, 노이즈 생성, 영역 확대 등을 통해서 〈그림10〉과 같이 이미지를 생성하여 해당 유니코드의 데이터 개수를 100개로 증가하였다.

〈그림10〉 증폭 생성된 이미지의 예

이처럼 데이터 증폭 및 샘플링을 통해서 각 자종별 이미지 수를 100개로 맞추었으며, 모델 선정 실험을 통해서 가장 높은 성능을 보인 ResNet152V2를 기준 모델로 새롭게 정제된 학습데이터를 학습하여 평가하였다.

그 결과는 〈그림11〉과 같다. 이렇게 정제된 학습데이터를 학습하였을 때 학습 시 최고 정확도는 99.48%를 달성하였으나, 평가 결과는 최고 90%에서 마지막 35%까지 변동폭이 매우 크게 나타났다.

〈그림11〉 정제된 학습데이터로 학습한 ResNet152V2

다음은 이러한 결과의 원인이 모델에 있는지 혹은 학습데이터의 특성에 인한 것인지 확인하기 위해서 평가데이터를 좀 더 분석을 진행해 보았다.

상위		하위	
name	f1score	name	f1score
5F55	1	9D42	0
633A	0.974359	9D51	0
6790	0.974359	9DF9	0
906D	0.95	9E92	0
34B1	0.947368	9ECC	0
39FE	0.947368	9EDB	0
6476	0.947368	9EFE	0
7235	0.947368	9F20	0
73EA	0.947368	9F52	0

〈표 2〉 인식률 상위 10종 및 하위 10종

〈표2〉와 같이 ResNet152V2의 한자 자종별 인식률은 확연한 차이를 나타내고 있다. 〈표2〉에서 사용된 f1 score(실현률)는 아래의 수식을 사용하였다.

$$F1score = 2 \times \frac{정밀도 \times 재현율}{정밀도 + 재현율}$$

정밀도는 모델이 예측한 정답 중 실제로 정답인 비율이며, 재현율은 실제 정답 중에서 모델이 정답으로 예측한 비율이다.

F1score가 1% 미만인 한자 자종은 1015종, 반대로 f1score 70% 이상은 620종으로 나타났다. 데이터 증폭 및 샘플링으로 인한 학습데이터가 오염의 가능성을 보기 위해 가장 많은 이미지가 있는 유니코드 10종과 하위 10종을 비교하면 아래 〈표3〉과 같다.

상위			하위		
유니코드	f1score	샘플링 전 이미지 수	유니코드	f1score	증폭 전 이미지 수
751F	0.26506	3015	9AE5	0.823529	1
6240	0.14359	2894	9B29	0	1
5974	0.136126	2724	9B31	0.888889	1
7232	0.103448	2575	9B41	0.333333	1
4E8B	0.062338	2451	9E81	0.095238	1
5341	0.367816	2379	9E92	0	1
4E4E	0.082353	2375	9EDE	0.787879	1
5E74	0.103806	2340	9EEF	0.461538	1
4EE5	0.395062	2338	9F0B	0.26087	1
5A62	0.141593	2307	9F6E	0.62069	1

〈표 3〉 원본 이미지 수에 따른 결과 비교

종합적으로 볼 때, 본 실험에 투입된 모델 중 초서 인식기에 가장 적합한 것은 ResNet152V2로 확인되었다. 또한 실험 결과 샘플링이 된 자종이 비교적 낮은 성능을 보이고, 증폭의 대상이 된 유니코드는 비교적 높은 성능을 보이는 것으로 나타났다. 그 원인은 이미지가 많아 다양한 글씨의 형태가 있는 자종을 샘플링하여 충분한 학습데이터를 확보하지 못한 경우 평가 정확도가 떨어진 것으로 추측할 수 있었다.

또한 학습 및 평가 그래프인 〈그림11〉를 보면 평가 정확도가 최대 90%이상 올라간 모습이 보이나 그 이후 지속적으로 하락하고 있음을 확인할 수 있다. 학습 정확도는 지속적으로 상승하는데 반해 평가 정확도의 변동이 학습 곡선과 다를 때는 보통 오버피팅이라 불리는 학습데이터 의존성 문제가 나타난 경우가 대부분이다. 오버피팅은 앞서 언급했듯 학습데이터와 테스트데이터만을 학습하여 실제 평가데이터인 validation 데이터에 대한 정답률이 떨어지는 현상을 말한다. 결국 학습데이터 증폭 및 샘플링을 통한 오염, 그리고 모델의 오버피팅 문제로 인해서 성능저하가 일어난 것으로 보인다.

본 실험을 통해 ResNet152V2가 가장 좋은 성능을 가진 모델로 보이나, 실험 결과를 좌우하는 것은 결국 데이터의 품질과 데이터의 균질함이었다. 양과 질에

있어 균형화된 데이터셋 확보의 중요성을 다시 확인할 수 있었다.

4. 결론

초서 관련 OCR의 개발은 효율적인 고문헌 데이터 구축의 기초 기술이다. 이런 인식은 우리나라뿐 아니라 중국에서도 동일하다. 중국은 오랫동안 이 문제를 연구하면서 개발 데이터셋을 확보하는 한편, 관련 알고리즘 개발 또한 노력을 경주하고 있다.

본 연구를 통해 우리나라의 한자 자종 관련 기초 데이터셋에 대한 오류 수정의 필요성을 충분히 살펴보았고, 기초 실험을 진행하면서 그 과정과 결과를 공유하였다. 본 연구를 통해 제시한 실험 방법과 결과가 얼마나 좋은 평가를 받게 될지 알 수 없다. 다만 본고와 같이 관련 연구 과정과 결과를 공유함으로써 향후 관련 연구에 대한 관심과 집중, 그리고 보다 정확한 연구 결과에 대한 정보 제시와 공유를 통해 인문학의 발전을 위한 진일보한 연구가 진행되기를 바란다.

▌참고문헌

1. 전자문헌

https://kostma.aks.ac.kr/segment/segmentList.aspx

https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=603

2. 논문

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 770-778.

Liao M, Wan Z, Yao C, Chen K, Bai X (2019) Real-time scene text detection with differentiable binarization. *In: arXiv:1911.08947*

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr(2022): Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*.

Simonyan, K. & Zisserman,(2014). A. Very deep convolutional networks for large-scale image recognition. *Preprint at https://arxiv.org/abs/1409.1556*.

The process and task of developing ChoSeo(草書) OCR Detection by artificial intelligence algorithm*

Heo, Chul** · Cho, SungDuk*** · Choi, Dong Bin****

In recent years, the reason for developing Chinese character OCR in Korea, including the Chinese region, is a problem of economic efficiency due to the lack of professional manpower beyond Chinese character-related skills. Costs increase and speed is limited. OCR-related technologies have been discussed as an alternative to overcoming this, but the difficulties revealed above have become limitations of technology and are in the process of overcoming them with artificial intelligence.

In line with this trend of reality, this paper intends to directly implement the process of developing Chaucer OCR using artificial intelligence and share the problems found in the process. This study became an opportunity to pay attention to the development of production tools through the development of more efficient Chaucer OCR-related technologies in the future, as well as to recognize that it is one of the strengthening of national competitiveness by securing original technologies in the Chinese character culture.

Keywords: literrary Sinitic works, Chinese characters, OCR character recognition Chose

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A6A3A02043693)

** HK Professor, KyungSung University, Korea / heochul@gmail.com
*** HK Research Professor, KyungSung University, Korea, / isanjo@nate.com
**** ABD, Dankook University / dbchoi85@gmail.com

본 논문은 2023년 5월 28일 투고되어 2023년 6월 12일 심사를 완료하여 2023년 6월 12일에 게재를 확정하였음