Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 4, pp. 398-408, August 2022.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2022.48.4.398

장면 이미지 속 한글 문자를 종단 간 검출 및 인식 가능한 딥러닝 네트워크 모델

김정원 · 김성범†

고려대학교 산업경영공학과

Deep Learning Network-Based End-to-End Scene Text Spotter for Korean Characters

Jeong Won Kim · Seoung Bum Kim

Department of Industrial and Management Engineering, Korea University

Scene text spotting detects text boxes and recognizes the words from scene images in an end-to-end way. Existing studies proposed scene text spotters for English and showed their promising performance. However, studies for non-English text spotter are rarely conducted because of a lack of fine-quality labeled training datasets. In particular, Korean text spotting is considered to be harder than English because of a large number of characters in Korean. In this study, we propose an end-to-end scene text spotting network specialized in Korean texts that can read more than 2,300 characters. To overcome the lack-of-dataset problem, we propose using a transfer learning method. By pretraining both modules (detector and recognizer) of the network with multi-language datasets and fine-tuning only the recognizer with the Korean dataset, we could construct the robust scene text spotter. We expect that our work can offer useful learning guidelines to future scene text models for non-English language that does not have sufficient training datasets.

Keywords: Scene Text Spotting, Korean Text Spotting, Scene Text Detection, Scene Text Recognition, Transfer Learning

1. 서 론

장면 이미지 속 문자 판독(scene text spotting) 연구는 간판, 상품 등 여러 객체가 존재하는 이미지에서 특정 언어의 문자열을 읽어내는 컴퓨터 비전의 한 연구 분야다. 흔히 알려진 광학 문자 인식(optical character recognition, OCR) 기술의 일종이나, 그 중에서도 일반 이미지처럼 어떤 형태의 배경 또는 문자가 있든 해당 문자열을 읽어낼 수 있는 기술을 칭한다. 즉, 기존 OCR이 규격화된 용지에 인쇄된 문자만 읽는 기술이었다면, 장면 이미지 문자 판독은 거리 풍경이나 제품 사진 등 모든 이미지에 적용할 수 있다. 이같은 범용성 덕분에 장면 문자 판독 기술은 이미지 번역, 신분증 인식 등 다양한 서비스에서 쓰이고 있다.

다양한 상황에 적용 가능해야 하는 만큼 장면 문자 판독은 기존 OCR보다 어려운 과제로 분류된다(Long *et al.*, 2021). 그로 인해 기존에는 이미지에서 문자 영역을 찾는 문자 검출(scene text detection) 연구와 해당 문자열을 읽어내는 문자 인식(scene text recognition) 연구가 별개로 이뤄졌다. 하지만 딥러닝 네트워크 기반의 문자 검출 또는 인식 모델이 유의미한 성능을 기록하면서, 최근에는 두 작업을 모두 수행하는 종단 간 문자 판독(end-to-end scene text spotting) 연구로 합쳐지는 추세다.

This research was supported by BK21 FOUR.
† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145, 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-3290-4550, E-mail : sbkim1@korea.ac.kr
2022년 1월 6일 접수; 2022년 2월 10일 수정본 접수; 2022년 2월 11일 게재 확정.

Liu *et al.*(2018)가 주목할 만한 성능의 딥러닝 기반 문자 판독 네트워크를 제시한 이래 후행 연구들 대부분 영어 문자열의 판독을 목표로 진행되고 있다. 반면 한글을 비롯해 비영어권 언어를 위한 문자 판독 연구는 비교적 최근에야 이뤄지고 있다. 2017년 영문과 한글, 한자, 아랍어 등 6개 언어 문자열을 포함한 이미지 데이터셋 international conference on document analysis and recognition 2017 multi-lingual text(ICDAR 2017 MLT)가 배포된 것이 기점이다. 하지만 영문에 비하면 비영어 문자 판독 연구는 발전 속도가 더딘 편이고 특히 다국어 동시 판독이 아닌 개별 언어에 특화된 판독 방법론은 거의 연구되지 않았다.

비영어 문자 판독을 위해 해결해야 하는 주요 과제는 문자 인식 단계에서 영어보다 훨씬 많은 수의 문자를 분류해야 한다는 점이다. 예컨대 영어의 경우 대문자와 소문자를 합쳐 52자이나, 한자는 표준 문자(중화인민공화국 통용규범한자표)가 8,105자이고 한글은 완성형(KS X 1001) 기준 2,350자이다. 더불어 문자끼리 형태적으로 유사한 탓에 이를 인식하기 위해선 매우 복잡한 네트워크가 필요하다. 이처럼 복잡도 높은 딥러닝 네트워크를 학습하기 위해선 상대적으로 더 많은 수의 데이터셋이 필요하다. 하지만 한글의 경우 장면 이미지 문자 판독 모델의 학습용 데이터셋이 부족한 실정이다. 한글 판독 기술의 발달을 위해 한글 단어(word) 및 글자(character) 수준의 레이블을 가진 데이터셋이 일부 배포되기도 했으나 레이블이 불완전해 활용도가 낮은 편이다.

위와 같은 문제 상황에서 본 연구는 2,350자의 한글을 종단 간 검출 및 인식 가능한 딥러닝 기반 방법론을 제시한다. 본 연구에서는 한글 판독을 위한 데이터셋이 부족한 문제를 극복하기 위해 전이학습을 시도했다. 문자 영역 검출에 반드시 한글 이미지만을 사용할 필요는 없다는 점에 착안하여, 다국어 데이터셋으로 전체 네트워크를 사전학습한 뒤 한글 데이터셋으로 인식 모듈을 미세조정했다. 이를 통해 별도의 인공 데이터 생성 과정 없이 ICDAR 2017 MLT에 대해 유의미한 성능을 달성할 수 있었다. 또한 2,350가지 한글 문자에 대한 분류가 가능하도록 connectionist temporal classification(CTC; Graves *et al.*, 2006) 디코더 기반의 인식 모듈을 사용했다. 더불어 이미지 분할 기반의 문자 검출 알고리즘을 사용해 파이프라인을 단순화함으로써 종단간 방식의 학습이 가능하도록 했다. 본 연구의 주요 기여점을 정리하면 아래와 같다.

- 장면 이미지 내 한글과 영어 문자열을 종단간 검출 및 인식 가능한 딥러닝 기반 네트워크를 제안한다. 이는 한글 검출이나 인식 중 하나만 수행한 것이 아닌 두 과제를 종단간 방식으로 수행할 때의 성능을 최초로 제시하는 것이다.
- 한글 판독을 위한 학습용 데이터셋이 부족한 문제를 해결하는 데 있어 전이학습 방식이 효과적임을 제시한다. 모델 학습에 한글 이미지뿐 아니라 다양한 언어 이미지를 활용하는 식으로 학습 데이터 수를 늘림으로써 문자 검출과 판독 성능을 향상할 수 있음을 실험을 통해 확인했다.

본 논문은 다음과 같이 구성된다. 제2장에서는 최근 연구된 장면 문자 검출 및 인식, 판독 방법론들을 설명한다. 제3장에서는 본 연구에서 제안하는 한글 문자 판독 방법론을 소개한다. 제4장에서 데이터셋을 비롯한 실험 조건과 실험을 수행한 결과를 제시한다.

2. 관련 연구

2.1 장면 이미지 내 문자 검출 및 인식(Scene Text Detection and Recognition)

장면 문자 검출은 이미지 내 문자가 위치한 영역의 좌표를 예측하는 과제이다. 대부분의 방법론이 단어 단위 상자를 검출하는 것을 목표로 하며 드물게 줄 단위 검출을 수행하기도 한다(Cao *et al.*, 2020). 딥러닝 네트워크를 사용한 초기의 문자 검출 연구는 주로 기존 객체 검출(object detection) 모델을 변형한 구조를 사용했다(Liao *et al.*, 2017; Ma *et al.*, 2018). 하지만 이같은 방법론들은 수평이 아닌 다각도로 배열된 문자열을 탐지하는 데 한계를 드러냈다. 이 방법론들은 특정 형태의 후보 상자(anchor box)를 무수히 가정한 후 그중 문자가 존재할 가능성이 높은 상자를 추려 문자열을 탐지하는 영역 제안(region proposal) 방식의 모델을 사용했는데, 후보 상자의 형태를 미리 제한하다 보니 그외 형태의 문자열은 정교하게 탐지할 수 없었기 때문이다.

이후 다양한 형태의 문자열이 존재하는 장면 이미지에 맞게 끔 영역 제안 없이 픽셀 단위로 문자열 존재 여부를 바로 예측하는 이미지 분할 기반 네트워크들이 제안됐다(Zhou *et al.*, 2017; Deng *et al.*, 2018). 이 같은 방법론들은 후보 상자를 사용하지 않는다고 해서 앵커 미사용(anchor-free) 검출 모델이라고도 한다. 앵커 미사용 방법론은 영역 제안 방식보다 모델 구조가 단순하면서도 다각도의 단어를 잘 탐지해 점차 문자 검출 모델의 주류가 되고 있다. 특히 본 연구와 같이 검출 모델과 인식 모델을 합쳐 종단간 판독 모델을 설계해야 하는 경우 각 모델의 구조를 단순화하는 것이 필요하기 때문에 제안 방법론에서는 앵커 미사용 검출기를 적용했다.

한편 장면 문자 인식은 검출된 문자 영역을 입력 받아 해당 영역에 무슨 글자가 있는지 읽어내는 과제다. 대게 단어 단위로 인식하는 것이 목적인데, 단어는 개별 글자의 시퀀스이므로 시퀀스 내 한 글자씩 분류하는 과제와 같다. 따라서 문자 인식 모델은 공통적으로 합성곱 신경망을 통해 특징을 추출한 뒤 특징 지도로부터 단어를 디코딩할 때 순환 신경망(recurrent neural network, RNN)과 같이 단어 시퀀스를 처리할 수 있는 모델 구조를 사용한다. Shi *et al.*(2017)은 이런 방식을 사용한 대표적인 모델로, 순환 신경망과 CTC를 디코더로 사용해 주목받았다. CTC는 기존 음성 인식 모델 학습에 쓰인 알고리즘으로, 예측된 단어 시퀀스의 각 글자와 정답인 단어의 각 글자 간 명시적인 정렬 정보 없이도 시퀀스 학습을 가능하게 해 준다.

CTC 외에 인식 모델의 또 다른 축은 어텐션(attention) 구조(Bahdanau *et al.*, 2014)를 적용한 sequence-to-sequence(seq2seq; Sutskever *et al.*, 2014) 네트워크이다(Lee and Osindero, 2016; Shi *et al.*, 2016; Cheng *et al.*, 2017; Bai *et al.*, 2018; Shi *et al.*, 2018). seq2seq 기반 방법론은 CTC 기반 모델에 비해 언어 모델에 의존해 단어 자체를 학습하며 대규모 학습 데이터셋을 필요로 한다. 반면 CTC 기반 방법론은 이미지와 각 글자 간 연관성을 학습하기 때문에 비교적 적은 데이터셋으로 학습 가능할 뿐 아니라 한글, 한자와 같이 글자 수가 많은 언어에 적합하다(Long *et al.*, 2021). 본 연구는 한글 판독을 목표로 하기 때문에 seq2seq 네트워크가 아닌 합성곱 신경망과 CTC 디코더로만 이뤄진 인식기를 사용했다.

2.2 종단간 장면 이미지 문자 판독(end-to-end scene text spotting)

종단간 문자 판독은 제2.1절에서 언급한 문자 검출과 인식을 동시에 수행하는 것으로, 이미지를 입력 받아 문자 영역의 좌표와 해당 단어를 모두 출력하는 것이 목적이다. 초기 제안된 Liao *et al.*(2017)은 single shot multibox detector(SSD; Liu *et al.*, 2016) 기반의 문자 검출 모델과 Shi *et al.*(2017)의 인식 모델을 이어 붙여 두 단계로 판독을 수행했다. 그러나 이미지에서 검출된 영역을 잘라 내 인식하는 과정에서 오차 역전파가 제대로 이뤄지지 않아 성능 향상에 제한이 있었다. 따라서 이후 판독 연구에서는 어떻게 두 모델을 결합해 종단간 방식으로 제대로 학습할 것인지가 관건이었다(Long *et al.*, 2021).

Li *et al.*(2017), Busta *et al.*(2017), Liu *et al.*(2018), He *et al.*(2018)은 이미지가 아닌 특징 지도에서 검출 영역을 잘라 인식 모델을 통과시키는 방식으로 이 문제를 해결하고자 했다. Li *et al.*(2017)은 영역 제안 방식의 검출 모델과 어텐션 메커니즘을 적용한 long short-term memory(LSTM; Hochreiter and Schmidhuber, 1997) 기반의 인식 모델을 사용했다. Busta *et al.*(2017)과 Liu *et al.* (2018)은 각각 Zhou *et al.*(2017)와 Redmon and Farhadi(2017) 기반 검출 모델을 사용했으며, 특징 지도에서 검출 영역을 고정 크기로 분리해 Shi *et al.*(2017) 인식 모델에 입력하는 방식을 택했다. He *et al.*(2018)도 Zhou *et al.*(2017) 검출 모델을 사용했으나 인식 단계에서 Shi *et al.*(2016)의 어텐션 기반 디코더를 적용했다.

위 방법들은 기존 방법론에 비해 모델 결합도가 높아졌음에도 여전히 검출 영역을 분리하는 과정에서 학습 상 한계가 존재했다. 이에 이미지 분할을 통해 두 작업을 동시에 수행함으로써 종단간 학습을 가능하게 한 네트워크들이 제안됐다. Mask textspotter(Lyu *et al.*, 2018)는 mask R-CNN(He *et al.*, 2017)을 활용해 영어 알파벳 수만큼의 분할 마스크를 생성, 각 마스크가 이미지 내 해당 알파벳의 위치를 나타내도록 한 후 이를 합쳐 단어 영역을 예측했다. Xing *et al.*(2019)은 약지도 학습(weakly supervised learning)으로 단어 단위 상자뿐 아니라 글자 단위의 분할 마스크도 예측해 상자 좌표와 해당 영역 레이블을 한 단계에 출력했다. 이미지 분할 기반의 네트워크는 앞서 언급한 두 단계 네트워크에 비해 다양한 배열의 문자열을 보다 정확히 탐지해 주목 받았다. 그러나 한글과 같이 글자 수가 수천개일 경우 그만큼의 분할 마스크를 예측해야 하기 때문에 학습이 어렵다는 문제가 남는다.

2.3 비영어 언어에 대한 장면 이미지 문자 판독(non-English scene text spotting)

제2.2절에서 설명한 문자 판독 방법론은 모두 영어 문자열을 대상으로 하고 있다. 영어 외 언어의 문자 판독 연구가 진행된 것은 reading Chinese text in the wild(RCTW-17; Shi *et al.*, 2017), Chinese text in the wild(CTW; Yuan *et al.*, 2018), ICDAR 2017 MLT 등 관련 벤치마크 데이터셋이 공개된 이후부터다. 이중 영어, 중국어, 한글, 일본어, 벵골어, 아랍어 등 다국어 문자열의 동시 판독을 목표로 한 연구들이 주목 받는 추세다. 다국어 판독 네트워크의 경우 제2.2절의 영문 판독 네트워크와 구조적 차이는 크게 없으나 그중 인식 단계에서 훨씬 더 많은 문자를 읽을 수 있는 구조가 채택되고 있다. Busta *et al.*(2018)은 Zhou *et al.*(2017) 검출 모델과 CTC 기반 인식 모델을 결합했으며, 두 모델 중간에 어파인 변환(affine transformation)을 통해 검출 영역을 분리 및 교정했다. Baek *et al.*(2020)은 U-Net(Ronneberger *et al.*, 2015) 구조의 검출 모델과 어텐션 기반 인식 모델을 결합해 수평이 아닌 다양한 배열의 문자열을 판독했다. Huang *et al.*(2021)은 이미지 분할 기반의 영문 판독 모델인 Mask textspotter V3(Liao *et al.*, 2020)를 약 9,000자의 다국어 문자 판독이 가능하도록 개선했다. 특징적으로 공통

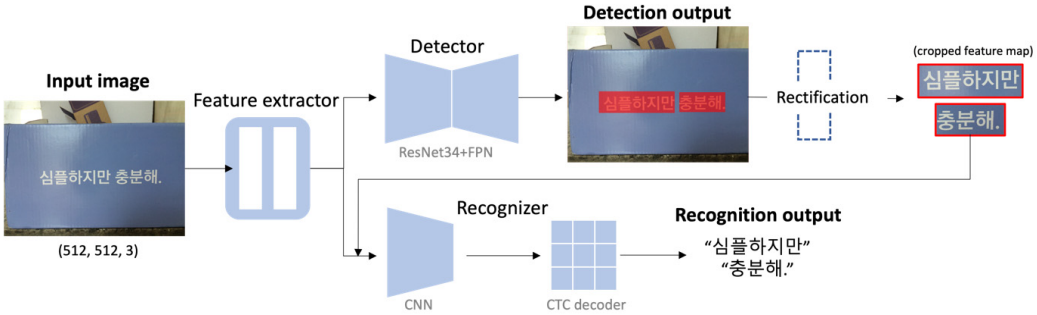

Figure 1. Structure of the Proposed end-to-end Scene Text Spotter

된 문자 검출 모듈을 거친 뒤, 각 영역이 어떤 언어인지 분류해 언어마다 별도의 인식 모듈에 입력되도록 설계했다. 인식 모듈은 각 언어에 특화된 학습이 필요하지만 검출 모듈은 언어와 무관하게 모든 데이터 정보를 이용할 수 있기 때문인데, 이는 본 연구에서 전이학습을 시도한 착안점과 유사하다. 한편 한글 검출과 인식을 수행한 연구로는 Hong and Kim(2018)이 있다. 해당 연구는 textboxes(Liao *et al.*, 2017) 검출 모델과 한글 자소 분해 및 병합 알고리즘, 합성곱 신경망 기반의 인식 모델을 연결해 1,000자의 한글을 검출 및 인식하는 모델을 제안했다. 하지만 이는 각 모델을 개별 학습한 뒤 이어 붙이는 형태이므로 종단간 문자 판독 모델로 보기는 어렵다.

본 연구에서는 앵커 미사용 검출기와 CTC 기반 인식기를 사용한 Busta *et al.*(2018)의 방법론을 기반으로 한글 판독 모델을 구축해 실험을 진행했다. 검출 단계에서는 앵커 미사용 검출기에서 많이 발생하는 1종 오류(false positive)를 줄이고자 어텐션 모듈을 추가했다. 인식 단계에서는 기존 방법론이 약 1,500자의 한글만 인식 가능하다는 점을 개선해 완성형 한글 2,350자를 모두 인식 가능하도록 모델을 학습했다. 또한 기존 비영어 판독 방법론이 모두 벤치마크 데이터셋과 더불어 대규모 합성 이미지 데이터셋을 생성해 모델 학습에 사용하는 데 반해, 본 연구는 벤치마크 데이터셋만으로 학습하면서도 전이학습을 통해 높은 판독 성능을 달성했다.

3. 제안 방법론

3.1 네트워크 구조

본 연구에서 사용한 모델은 합성곱 신경망으로 구성돼 있으며 구조는 <Figure 1>과 같다. 모델은 크게 특징 추출기, 문자 검출기, 문자 인식기로 나뉜다. 특징 추출기에서 출력된 특징 지도를 검출기와 인식기에 동시 입력함으로써, 특징 추출기가 오차 역전파 과정에서 문자 검출 및 인식에 필요한 정보를 모두 학습할 수 있도록 했다(Liu *et al.*, 2018). 문자 검출기는 feature pyramid network(FPN; Lin *et al.*, 2017) 구조를 사용했다. 이는 이미지 분할을 통해 문자열이 존재하는 픽셀을 직접 탐지하는 앵커 미사용 방식의 검출기로, 객체 탐지에 흔히 사용되는 영역 제안 방식의 네트워크보다 파이프라인을 단순화하면서도 다양한 배열의 문자열을 탐지할 수 있다. 또한 검출기에 특징 지도 상 문자열이 존재하는 영역을 주목하는 어텐션 모듈을 추가해 검출 성능을 높였다. 문자 인식기는 학습용 데이터셋이 한정적인 상황에서 2,350가지 한글에 대한 분류를 수행할 수 있도록 합성곱 신경망과 CTC 디코더로 구성했다.

3.1.1 문자 검출기(Text Detector)

문자 검출 단계에서는 특징 추출기가 출력한 특징 지도를 입력 받아 <Figure 2>에 나타낸 것과 같이 총 7개 채널의 이미지 분할 마스크를 예측한다. 채널 1은 해당 픽셀에 문자열이 존재할 확률값(text score) r_t을 나타낸다. 채널 2~5는 각 픽셀이 위치한 문자열 상자의 상・하・좌・우변까지의 최단 거리(side-distance)를 나타낸다. 채널 6~7은 각 픽셀을 포함한 문자열이 기울어진 각도(orientation angle) r_θ에 대해 $\sin(r_\theta)$, $\cos(r_\theta)$ 값을 나타낸다. 채널 1~5의 마스크를 합쳐 사각형의 문자열 상자를 탐지할 수 있으며, 채널 6~7을 통해 문자열 상자의 각도를 조정함으로써 더욱 정교하게 문자를 검출할 수 있는 구조다. 이렇게 산출된 분할 마스크에서 r_t>0.9인 영역을 대상으로 non-maximum suppression(NMS; Zhou *et al.*, 2017) 알고리즘을 적용해 문자열 테두리 상자를 산출한다.

문자 검출을 위한 신경망은 ResNet-34 구조의 FPN을 사용했다. <Figure 2>에서 특징 지도(feature map)를 추출한 뒤 채널 어텐션(channel attention) 모듈을 거치기 직전까지의 과정이 여

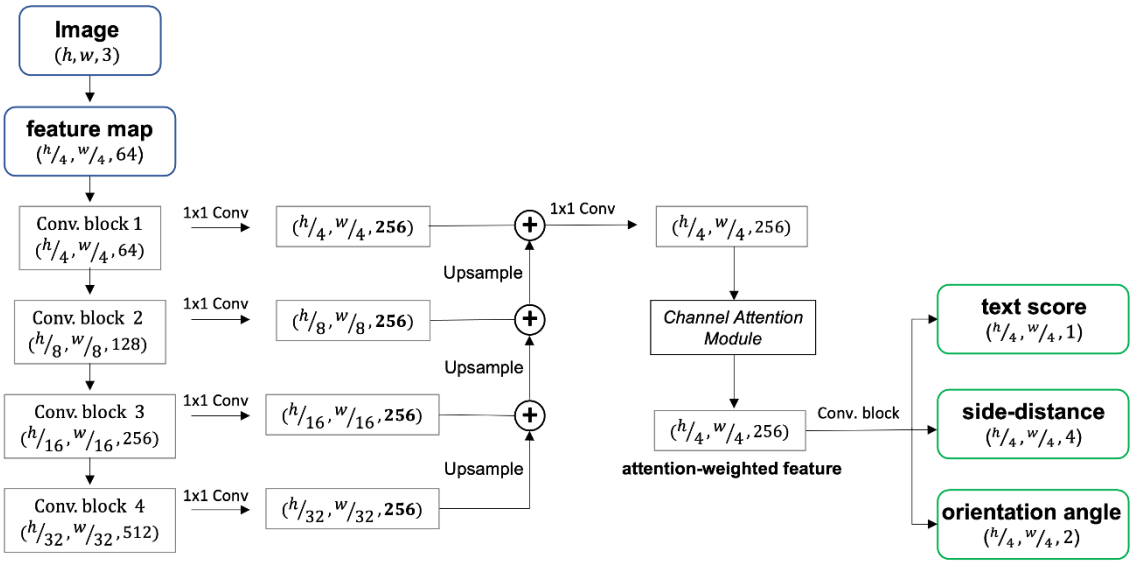

Figure 2. Structure of the Text Detector

기에 해당된다. 이 신경망은 특징 지도를 입력 이미지의 $\frac{1}{4^2}$~$\frac{1}{32^2}$ 크기까지 4단계에 걸쳐 압축한 뒤 이중선형 보간법 및 합성곱 연산으로 특징 지도를 복원한다. 복원은 원본 이미지의 위치 정보가 살아있는 하위 레벨의 특징 지도와, 특징을 압축적으로 담고 있는 상위 레벨의 특징 지도를 결합하는 방식으로 진행된다. 이를 통해 객체의 위치 정보를 최대한 보존하면서도 핵심적인 특징을 반영한 특징 지도를 산출할 수 있다. 복원된 특징 지도는 커널 크기 1의 합성곱 연산으로 융합한 뒤 채널 어텐션 모듈(Fu *et al.*, 2019; Cao *et al.*, 2020)에 입력된다. 채널 어텐션 모듈은 특징 지도의 채널 간 영향도를 계산해 각 채널에서 중요한 영역, 즉 문자열이 위치한 영역을 강조하고, 반대로 배경과 같이 중요도가 낮은 영역은 약화하는 역할을 한다. 구체적인 작동 방식은 아래 3.1.2에서 다룬다. 채널 어텐션 모듈을 거친 특징 지도는 다시 합성곱 연산을 거쳐 채널 1~7의 분할 마스크로 변환된다. 활성화 함수로 시그모이드 함수를 적용하여 분할 마스크 상 값을 0~1 사이의 확률값으로 변환한다.

3.1.2 채널 어텐션 모듈(Channel Attention Module)

채널 어텐션 모듈은 Fu *et al.*(2019)에서 일반적인 이미지 분할 모델의 정확도를 높이기 위한 메커니즘으로 소개됐다. Cao *et al.*(2020)은 이 모듈이 장면 이미지 문자 검출의 성능 향상에도 효과적임을 증명했다. 이미지 분할 기반의 문자 검출 시 어텐션 모듈을 도입하는 이유는 채널별 중요 영역에 더 높은 가중치를 부여함으로써 분할 마스크의 예측 정확도를 높이기 위해서다. 어텐션 모듈에 입력되는 특징 지도는 원본 이미지의 $\frac{1}{4^2}$ 크기, 256개 채널이며, 이후 합성곱 연산을 거쳐 7개 채널의 분할 마스크를 예측한다. 채널을 축소하기에 앞서 각 채널에서 중요한 영역이 강조되도록 특징 지도의 표현력을 높이는 것이 어텐션 모듈의 역할이다(Cao *et al.*, 2020). 본 제안 방법론에서 사용한 채널 어텐션 모듈은 합성곱 연산 없이 특징 지도 간 계산으로 이뤄져 있어 학습 파라미터 수를 늘리지 않는 장점이 있다. <Figure 3>에서 보듯이 입력 특징 지도 $A \in R^{C\times H\times W}$로부터 별도의 학습 없이 내부 연산을 통해 채널 어텐션 지도 $X \in R^{C\times C}$를 산출한다. x_{ji}는 j번째 채널에 대한 i번째 채널의 영향도를 나타낸다(식 (1)).

$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C}\exp(A_i \cdot A_j)} \tag{1}$$

X와 $A_{reshape} \in R^{C\times(H\times W)}$의 행렬 곱셈을 통해 A의 각 채널에 가중치를 적용한 $A_{weighted} \in R^{C\times H\times W}$를 구한다. 이후 식 (2)와 같이 파라미터 β를 적용한 $(\beta \cdot A_{weighted} + A) \in R^{C\times H\times W}$의 새로운 특징 지도($E$)를 출력한다. E가 최종적으로 어텐션에 의한 가중치가 적용된 특징 지도이다. 본 연구에서 H와 W는 각 원본 이미지 높이, 가로의 $\frac{1}{4}(h/4,\ w/4)$이며, C는 256이다.

$$E_j = \beta\sum_{i=1}^{C}(x_{ji}A_i) + A_j \tag{2}$$

어텐션 모듈은 특징 지도의 채널을 축소하는 과정에 위치해야 의도된 역할을 수행할 수 있다. 다만 문자 검출의 최종 단계인 7개 채널로 축소하기 바로 직전에 모듈을 위치시키는 것보다, 어텐션 모듈 이후 합성곱 연산을 일부 거친 다음 최종 예측을 수행해야 모듈이 효과적으로 작동함을 확인했다. 그렇지 않으면 일부 실험에서 긍정(positive) 예측이 과도하게 줄어 오히려 2종 오류(false negative)가 늘거나 이후 문자 인식 단계에서 정 확도가 떨어지는 등 불안정하게 학습되는 결과를 보였다. 이는 해당 모듈이 단순 연산을 통해 가중치를 부여하는 방식이기 때문인 것으로 해석된다. 따라서 본 연구에서는 어텐션 모듈 이후 2번의 합성곱 층을 추가함으로써 검출 성능을 향상하면서도 문자 검출기를 안정적으로 학습하고자 했다.

3.1.3 문자 인식기(Text Recognizer)

문자 인식 단계에서는 앞서 검출된 문자 영역들을 별도 분리한 뒤 각각 인식하여 단어를 출력한다. 문자 인식기는 글자가 왜곡 또는 회전돼 있을 경우 매우 취약하므로 문자 영역을 분리 시 교정 과정이 필요하다(Busta *et al.*, 2018). 본 제안 방법론에서는 Jaderberg *et al.*(2015)에서 사용된 어파인 변환을 사용해 교정을 진행했다. 교정된 영역은 원본 이미지 $\frac{1}{4^2}$ 크기의

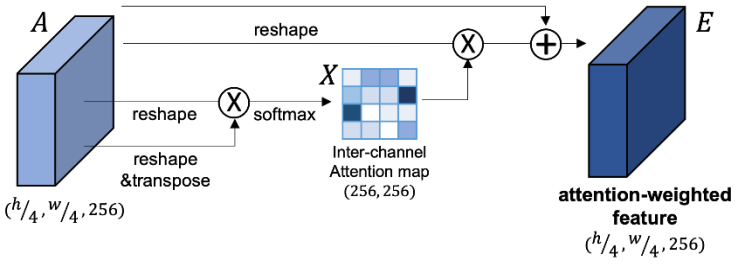

Figure 3. Structure of the Channel Attention Module(CAM)

Table 1. Layers of the Text Recognizer

Block	Layer	Size / Stride	Output Feature
1	Conv2D, Instance Norm, LReLU	3×3	*h*/4, *w*/4, 128
	Conv2D, LReLU, Conv2D, LReLU	3×3	*h*/4, *w*/4, 128
	Maxpool	2×1/2×1	*h*/8, *w*/4, 128
2	Conv2D, Instance Norm, LReLU	3×3	*h*/8, *w*/4, 256
3	Conv2D, LReLU, Conv2D, LReLU	3×3	*h*/8, *w*/4, 256
	Conv2D, LReLU, Conv2D, LReLU	3×3	*h*/8, *w*/4, 256
	Maxpool	2×1/2×1	*h*/16, *w*/4, 256
4	Conv2D, Instance Norm, LReLU	2×3	1, *w*/4, 256
	Dropout, Conv2D	1×1	1, *w*/4, |*C*|
	LogSoftmax		|*C*|, *w*/4

특징 지도 형태로 인식기에 입력된다. 인식기는 <Table 1>과 같이 합성곱 신경망과 CTC 디코더로 구성돼 있다. 합성곱 연산 시 세로 2, 가로 1 커널의 최대 풀링(max pooling)을 적용하여 가로 폭은 그대로 유지한 채 높이만 점차 줄인다. 합성곱 연산 후 로그소프트맥스 함수를 적용하면 $Z \in R^{|C| \times w/4}$의 시퀀스가 산출된다. $z_i (i = 1, 2, ..., w/4)$는 특징 지도의 왼편에서부터 i번째 위치의 수용 영역에 어떤 글자가 존재하는지에 대한 확률을 나타낸다. 여기서 $|C|$는 인식하고자 하는 글자 종류 수로 본 연구에서는 한글 2,350자와 영문, 숫자, 기호 등을 포함해 2,500으로 설정했다.

Z는 손실 및 기울기 계산을 수행하는 CTC 디코더에 입력돼 문자열 형태로 변환된다. CTC는 시퀀스가 길어지면 연산량이 기하급수적으로 늘어난다는 한계가 있으나, 한글 단어는 영어에 비해 길이가 짧기 때문에 CTC를 사용하기에 유리하다. CTC 디코더에 입력 전 RNN 신경망을 통과하도록 구성한 문자 인식 모델도 많지만 본 연구는 학습 데이터 수의 제약 아래 이뤄졌기 때문에 RNN 신경망을 사용하지 않는 것이 오히려 높은 인식 성능을 보여 생략했다.

인식기는 검출기와 함께 종단간으로 학습 가능하다. 하지만 검출기가 충분히 학습되지 않은 초반에는 검출 영역을 바로 인식기에 입력할 경우 인식기를 제대로 학습시킬 수 없다. 따라서 검출 손실이 2 미만으로 줄어들기 이전까지는 레이블 정보를 활용해 정답인 문자 영역을 분리하여 인식기를 학습시킨다. 이후 검출 손실이 2 미만일 때 검출된 영역 중 실제 문자 영역과 중첩된 부분이 90% 이상일 경우 분리하여 인식기에 입력한다.

3.2 손실 함수

총 손실은 식 (3)과 같이 검출 단계와 인식 단계 손실을 합하여 계산된다.

$$L_{final} = L_{\text{det}} + \lambda \cdot L_{rec} \tag{3}$$

검출 손실 L_{det}은 제3.1.1절에서 설명한 채널 1~7에 대해 식 (4)와 같이 산출한다. 먼저 문자열이 존재할 확률 r_t을 나타낸 분할 마스크는 Milletari *et al.*(2016)의 dice 손실(L_{dice})을 사용한다. 장면 이미지는 대부분의 픽셀이 배경이고 문자 영역은 극히 소수인데 이로 인해 발생하는 클래스 불균형 문제를 dice 손실로 개선해 준다.

$$L_{det} = \alpha \cdot L_{dice} + \beta \cdot L_{IoU} + L_{MSE} \tag{4}$$

이때 L_{dice}은 예측한 분할 마스크와 정답 분할 마스크의 모든 픽셀에 대해 계산한다. 반면 각 픽셀과 문자열 상자의 상・하・좌・우변 간 최단 거리를 나타낸 분할 마스크, 각 픽셀의 문자열 각도 r_θ를 나타낸 분할 마스크는 정답 상 실제 문자가 존재하는 영역에 대해서만 손실을 계산한다. 거리 분할 마스크에는 Zhou *et al.*(2017)의 intersection-over-union(IoU) 손실(L_{IoU})을 적용하고, 각도 분할 마스크에 대해선 mean-squared-error(MSE) 손실(L_{MSE})을 사용한다. 마지막으로, 인식 손실 L_{rec}은 3.1.3.에서 소개한 CTC 손실을 사용한다. 본 연구에서 $\alpha = \beta = 1$, $\lambda = 0.5$로 설정했다.

3.3 학습 방법

본 연구에서는 한글 판독 모델을 학습시킬 만한 데이터셋이 충분하지 않다는 한계를 극복하고자 전이 학습을 시도했다. 이는 문자 판독을 위한 두 단계 중 검출 성능을 올리기 위한 시도이다. 장면 문자 판독에서는 문자 검출이 인식에 선행되는 작업이기 때문에 검출 성능을 높이는 것이 매우 중요하다. 그런데 인간이 문자를 인지하는 방식에 비춰보면 단순히 이미지에서 문자를 구분(검출)해내는 데 있어서는 언어의 도메인적 정보가 거의 필요하지 않음을 알 수 있다. 즉, 인간은 전혀 알지 못하는 언어라 하더라도 글자를 보는 순간 그것이 글자임을 인지할 수 있다. 만약 딥러닝 모델 학습도 이와 유사하게 이뤄진다면 검출 모델을 반드시 타깃 언어 이미지로만 학습할

필요가 사라진다(Huang *et al.*, 2021). 따라서 적은 한글 이미지 데이터셋만으로 모델을 학습하는 것보다 다른 언어 레이블의 데이터셋도 추가로 활용함으로써 검출 성능을 향상하고자 한 것이다.

제3.1절에서 설명한 대로 제안 모델에서는 이미지 입력 시 특징 추출을 거쳐 문자 검출기와 인식기를 통과한다. 우선 사전학습(pre-training) 단계에는 모델의 모든 층위를 다국어 데이터셋으로 학습한다. 이후 미세조정(fine-tuning) 단계에서 문자 인식기 부분만 한글 및 영어 데이터셋으로 재학습한다. 이를 통해 한글 및 영어를 잘 탐지하면서도 불필요한 언어를 인식하지 않는 판독 모델을 만들고자 했다. 인식기를 이루는 여러 합성곱 신경망 층위 중 어느 층위까지 재학습할 것인지는 4.4.에서 다룬다. 가설 단계에서는 한글 이외 언어의 이미지를 사용할 때 1종 오류가 증가할 수 있다는 우려점도 있었지만 실험 결과 1종 오류가 크게 늘지 않았음을 확인했다. 전이 학습을 위한 데이터셋 구성은 4.1.에서 세부적으로 설명한다.

4. 실 험

4.1 데이터셋

ICDAR(https://rrc.cvc.uab.es/)은 장면 이미지의 문자 검출 및 인식, 판독을 위한 대표적인 공개 데이터셋이다. ICDAR 2015는 도로, 건물 등에서 촬영한 장면 이미지 1,500장(학습 1,000장, 평가 500장)으로 이뤄져 있으며 영어 및 숫자 문자열에 대한 레이블을 제공한다. 문자열을 정면에서 바라보거나 이미지 상 중앙에 위치시켜 촬영한 사진이 아닌(focused scene), 마치 행인이 거리를 걸을 때 주변을 인지하는 시점처럼 문자열이 여러 위치에 흩어져 있는 형태(incidental scene)의 사진이다. ICDAR 2017 MLT는 간판이나 표지판, 책 표지 등을 촬영한 장면 이미지 9,000장(학습 7,200장, 평가 1,800장)으로 구성돼 있다. ICDAR 2015와 달리 문자 영역을 정면에서 집중적으로 촬영했기 때문에 주로 수평 형태의 문자열을 담고 있다. 영어와 한글, 일본 한자, 중국 한자, 아랍어, 벵골어 등 6개 문자열과 숫자에 대한 레이블을 제공하며, 이중 한글을 주로 포함하고 있는 이미지는 학습용 850장, 평가용 200장이다. ICDAR 2019 MLT 역시 간판이나 표지판, 책 표지 등을 촬영한 장면 이미지 10,000장으로 구성된 데이터셋이다. ICDAR 2017 MLT와 마찬가지로 주로 수평 형태의 문자열 이미지이며, 영어와 한글, 일본 한자, 중국 한자, 아랍어, 벵골어, 힌디어 등 7개 문자열과 숫자에 대한 레이블이 존재한다. 10,000장 중 한글을 주로 포함하고 있는 이미지는 1,000장이다. 본 연구 실험에서는 ICDAR 2015, ICDAR 2017 MLT, ICDAR 2019 MLT 세 데이터셋의 학습 데이터를 이용하여 모델을 훈련하며, 성능 평가에는 ICDAR 2017 MLT의 평가용 한글 이미지 200장을 사용했다.

AIhub Text in the Wild(https://aihub.or.kr/aidata/133) 데이터는 서울 시내 간판, 도서, 상품 등 이미지 10만 장으로 구성된 한글 장면 이미지 데이터셋이다. 한글과 영어, 숫자 문자열에 대한 레이블이 포함돼 있으며 정면에서 촬영된 수평 형태의 문자열이 주를 이루고 있다. 다만 데이터 중 일부는 이미지에 문자열이 존재함에도 레이블이 누락돼 있어 문자 검출기 학습용으로는 부적합하다. 따라서 본 연구 실험에서는 레이블 정보를 기반으로 문자 영역을 분리해 최대 40,000장의 단어 이미지를 생성했으며, 이를 인식 단계에 바로 입력함으로써 인식기 학습용으로 사용했다.

제안 방법론은 전이학습을 통해 한글 문자 판독 성능을 향상하고자 했다. 이를 위해 세 가지의 학습 방식을 시도해 성능을 비교했다. 세 가지 학습 타입은 아래와 같다.

- KOR(Korean) : 한글 및 영문 이미지 데이터만으로 전체 네트워크를 지도학습
- ML(Multi-language) : 한글, 영문에 더해 다국어 이미지로 전체 네트워크를 지도학습
- ML-TL(Multi-language transfer learning) : ML 방식에서 최고 성능을 기록한 모델을 사전학습 모델로 삼아 한글 및 영문 이미지로 문자 인식 네트워크만 미세조정

이때 한글, 영문, 그외 언어 이미지에 해당하는 데이터셋은 아래와 같다. KOR, ML-TL 학습 방식에는 학습 데이터셋 A를, ML 방식에는 학습 데이터셋 B를 사용했다. 학습 데이터셋 A는 총 2,850장이며 B는 총 18,200장이다. 더불어 모든 타입에서 AIhub Text in the Wild 최대 40,000장을 인식기 학습에 사용했다.

- 학습 데이터셋 A : ICDAR 2015(영문), ICDAR 2017 MLT 중 한글, ICDAR 2019 MLT 중 한글
- 학습 데이터셋 B : 학습 데이터셋 A, ICDAR 2017 MLT 중 한글 외 모든 언어, ICDAR 2019 MLT 중 한글 외 모든 언어

4.2 학습 세부사항

제안 모델을 실험하면서 4.1.에서 서술한 각 학습 방식에서 최대 성능을 낼 수 있는 하이퍼파라미터를 탐색하여 적용했다. 특히 초기 학습율 설정에 따라 각 방식의 성능 차이가 커, 모델의 전 층위를 학습하는 KOR, ML 방식은 0.0001, 인식기만 미세조정하는 ML-TL 방식은 0.0005로 설정했다. 최적화 알고리즘은 Adam(Kingma and Ba, 2018)을 사용했다. 검증 데이터에 대한 손실 수렴 추이를 관찰해 KOR, ML 방식은 최대 500 에폭(epoch)을, ML-TL 방식은 최대 800 에폭을 학습했다. 배치 크기는 학습 데이터셋 A를 사용한 KOR, ML-TL 에서는 32로, 크기가 더 큰 학습 데이터셋 B를 사용한 ML 방식에는 40으로 두었다. 마지막으로, 인식기 학습 시 가로 폭이 모두 다른(세로는 40픽셀로 고정) 문자열 이미지를 미니배치 내 평균 폭으로 조정해 입력했다. 특히 CTC 디코더를 사용한 인식기

Table 2. Scene Text Detection and End-to-end Scene Text Spotting Results on ICDAR 2017 MLT(Korean)

	Detection			End-to-end spotting		
	Recall	Precision	F1 score	Recall	Precision	F1 score
KOR w/o attention	69.5	61.9	65.5	40.9	36.4	38.5
KOR	67.9	64.6	66.2	40.5	38.5	39.5
ML	69.5	**67.0**	68.2	46.2	44.5	45.3
ML-TL	**70.5**	66.0	**68.2**	**50.7**	**47.5**	**49.0**

특성 상 이미지 테두리에 공백을 채우는 식으로 크기를 조정할 경우 인식기가 제대로 학습되지 않음을 확인했으며, 대신 이미지를 가로 방향으로 늘이거나 줄이는 방식을 사용했다.

4.3 평가 방법

제안 모델의 성능을 평가하기 위해 두 가지 지표를 사용한다. 우선 문자 검출 성능은 검출한 영역과 실제 레이블 영역이 교차하는 비율인 IoU을 기반으로 한다. 일반적으로 IoU가 0.5를 넘을 경우 올바른 검출이 이뤄진 것(참 긍정)으로 간주한다(Nayef *et al.*, 2017). 모든 평가 이미지에 대해 추론을 시행한 후 재현율(recall), 정밀도(precision), F1 점수를 집계하여 성능을 파악한다. 종단간 문자 판독 성능은 문자 검출 결과와 인식 결과를 동시에 평가한다. 즉, 문자 검출 영역의 IoU가 0.5를 넘는 동시에, 문자 인식 결과가 실제 단어와 같은 경우 올바르게 판독한 것으로 분류한다.

이때 인식한 문자열과 실제 문자열이 동일한 지 판단하기 위해 편집 거리(edit distance)를 사용했다. 편집 거리는 문자열 간 유사도를 측정하는 주요 지표로 단어의 모든 글자가 동일할 경우 거리가 0이며 다른 글자가 많을수록 거리가 증가한다. 본 실험에서는 편집 거리가 0인 경우 올바른 인식이 이뤄진 것으로 간주한다.

본 연구는 장면 이미지 내 가로로 긴(horizontal) 배열의 한글 및 영문 단어를 정확히 탐지 및 인식하는 것으로 목적으로 하기 때문에, 정답인 단어 상자 중 가로 길이가 세로 길이보다 1.5배 이상인 상자를 대상으로 평가를 진행한다. 또한 단어 길이가 2자 이상인 경우를 대상으로 평가한다.

4.4 실험 결과 및 분석

ICDAR 2017 MLT의 테스트 데이터에 대한 제안 방법론의 문자 검출 및 종단간 판독 성능은 <Table 2>에서 확인 가능하다. 베이스라인은 제안 네트워크에서 3.1.2.의 채널 어텐션 모듈을 사용하지 않고 한글 데이터셋만으로 지도학습 했을 때의 성능이다(KOR w/o attention). 이 베이스라인 외에 한글 판독 성능을 제시한 선행 연구가 없기 때문에 정확한 비교는 어렵지만, ICDAR 2017 MLT의 공식 웹사이트에 공표된 대회 참가자들의 기록을 성능 지표로 참고할 수 있다. 기록 상 ICDAR 2017 MLT에 대해 문자 검출만 수행한 모델의 최고 성능은 한글의 경우 약 65% 수준으로 위 베이스라인과 유사하다. 다만 종단간 판독 성능은 ICDAR 2017 MLT에 대해 제시된 기록이 없으므로 유사 데이터셋인 ICDAR 2019 MLT의 공개 기록을 참고할 만하다. ICDAR 2019 MLT에 대해 영문을 포함한 다국어 판독 최고 성능은 약 48~52%로, 통상 영문 판독 성능이 그외 언어에 비해 월등히 높기 때문에 실제 해당 방법론들의 한글 판독 성능은 이보다 현저히 낮았을 것으로 추정된다. 따라서 본 연구는 한글 문자 검출 65% 이상, 판독 45% 이상일 경우 최신 기록들에 비해 준수한 성능으로 볼 수 있다고 판단하였다.

이 같은 기준에 따라 본 연구에서 제안한 전이학습 기반의 한글 판독 방법론(ML-TL)의 성능을 평가해보면, 베이스라인 대비 10.5%포인트 높은 판독 성능(49%)을 기록했을 뿐 아니라, 그 외 모든 학습 방식보다 검출 및 판독 성능이 뛰어난 것을 볼 수 있다. 결과를 세부적으로 살펴보면, 우선 한글 데이터셋만으로 지도학습 한 경우(KOR)보다 다국어 데이터셋을 추가해 지도학습(ML) 했을 때 문자 검출 성능이 2%포인트 향상됐다. 판독 성능도 크게 향상됐는데 이는 두 가지로 분석 가능하다. 첫째, 검출 단계에서 더 많은 문자 영역을 탐지했을 뿐 아니라 더 정확히 탐지하여 인식율이 높아졌을 가능성이 있다. 둘째, 제안 네트워크는 검출 모델과 인식 모델이 특징 추출기를 공유함으로써 상호 정보를 학습하도록 돼 있는데, 이런 구조로 인해 두 성능이 모두 개선됐을 수 있다.

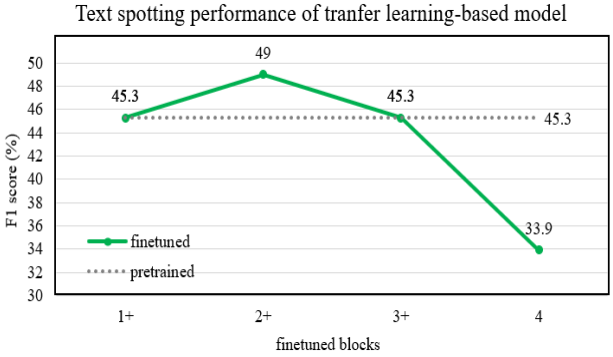

Figure 4. Text Spotting Performance of Transfer Learning-based Model(ML-TL)

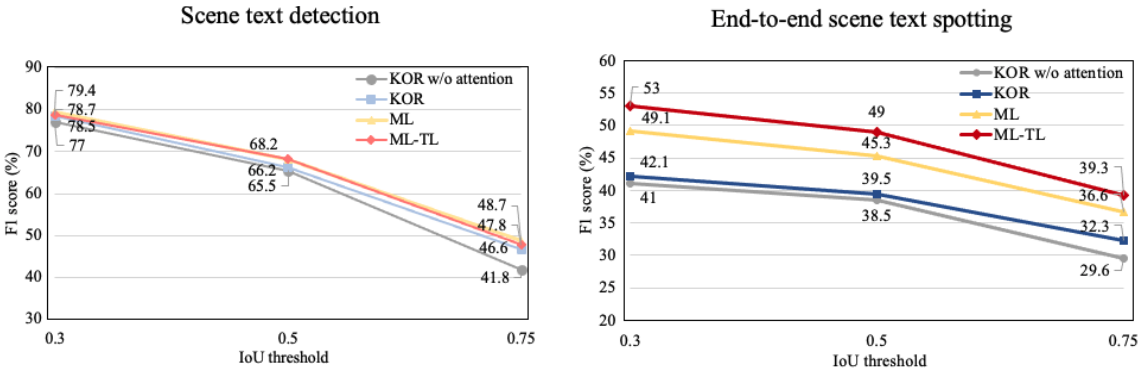

Figure 5. Scene Text Detection and End-to-end Scene Text Spotting Results on ICDAR 2017 MLT(Korean) with Different IoU Thresholds

Figure 6. Results on ICDAR 2017 MLT(Korean)

추가로 주목할 부분은 문자 검출 성능의 정밀도 지표이다. 한글이 아닌 데이터셋까지 학습에 활용 시 여러 문자의 특성을 학습해 문자가 아닌 부분을 문자 영역으로 탐지해 1종 오류가 늘어날 위험이 있다. 하지만 실제 실험에서는 정밀도 역시 2.4%포인트 상승한 것으로 나타나, 1종 오류가 크게 늘지 않은 것으로 확인됐다. 최종적으로, ML 타입에서 최고 성능을 기록한 모델의 문자 인식기를 한글 데이터셋으로 미세조정했을 시 (ML-TL) 문자 판독 성능이 4%포인트 가까이 높아졌다. ML 타입에서는 인식 단계에서 영문과 한글뿐 아니라 한자, 아랍어 등 너무 많은 글자를 분류해야 했지만, 이를 영문 및 한글만을 인식하도록 재학습함으로써 인식율을 높였다.

<Figure 4>는 미세조정 층위를 다르게 했을 때에 대한 평가 결과이다. 제안 모델의 문자 인식기는 <Table 1>에서 나타나듯이 여러 합성곱 블록으로 이뤄져 있다. 이중 미세조정 단계에서 몇 번째 블록부터 재학습하는지에 따라 판독 성능의 향상 폭이 달랐다. 실험 결과 block 2~4(2+)를 재학습했을 때 F1 score가 49.0%를 기록해 성능 향상 폭이 가장 컸다. 반대로 block 4만 재학습했을 때는 오히려 사전학습 모델(45.3%)보다 판독 성능(33.9%)이 크게 하락했다. 다만 이를 고정적인 결과로 보긴 어렵기 때문에 전이학습을 진행 시 상황에 맞는 최적 층위를 찾는 것이 중요할 것으로 보인다.

<Figure 5>는 제안 방법론에 따른 문자 판독 모델이 얼마나 정확하게 문자 영역을 검출했는지 알아보기 위해 IoU 임계값을 다르게 해 평가한 결과이다.

IoU 임계값은 문자 검출의 성공 여부를 정하는 기준으로 <Table 2>의 결과는 임계값이 0.5일 때이다. 임계값 0.3은 정답 영역과 30%만 겹쳐도 제대로 검출한 것으로 평가하는 반면, 0.75는 75%가 겹쳐야 맞게 탐지한 것으로 평가한다. 임계값 0.75에 대한 성능이 높을수록 가장 정답에 가깝게 검출한 것이다. <Figure 5>에서 보듯이 제안 방법론(ML-TL)의 문자 판독 성능은 모든 임계값에 대해 가장 우수하게 나타났다. 다만 ML-TL 그래프와 비교 방법론(KOR, ML) 그래프의 기울기를 비교했을 때, 임계값 0.5의 성능과 임계값 0.75에서의 성능 간 차이가 크게 다르지 않았다. 반면 채널 어텐션 모듈을 사용하지 않은 모델(KOR w/o attention)은 임계값 0.75에서 검출 성능과 판독 성능 모두 큰 폭으로 하락했다. 이는 어텐션 모듈의 사용으로 문자열 검출이 더욱 정확해졌음을 보여준다. 흥미로운 점은 모든 학습 타입에서 임계값 0.3일 때 문자 판독 성능이 상승한다는 점이다. 판독 성능은 정확한 검출과 인식을 모두 평가한 것이므로, 이는 문자 영역을 30% 이상만 검출해도 정답 영역의 단어를 상당 부분 인식할 수 있음을 뜻한다.

마지막으로 <Figure 6>은 평가 데이터셋인 ICDAR 2017 MLT 중 한글 이미지에 제안 모델을 적용한 결과이다. 두 글자 이상인 단어를 대상으로 평가 및 시각화했으며, 실제 길이가 1인 일부 글자를 제외한 대부분의 단어를 검출했음을 확인할 수 있다. 단어 상자의 왼편 상단에 녹색 점이 표시된 것은 검출과 인식이 모두 정확히 된 경우이며, 파란색 점은 검출이 제대로 됐으나 인식이 정답 단어와 다르게 된 경우이다. 붉은 점은 검출이 잘못된 경우다. 다만 오른쪽 상단 이미지에서 '인터넷', '모바일앱'의 경우 검출과 인식이 제대로 이뤄졌음에도 둘 다 틀린 것으로 평가돼 붉은 점으로 표시된 것을 볼 수 있다. 이는 정답 단어가 '인터넷/모바일앱'이기 때문인데, 사실상 문자 판독 목적에 어긋나지 않은 결과임에도 엄격한 평가 방법을 사용해 오류로 분류된 경우다.

5. 결 론

본 연구는 다양한 장면 이미지에서 한글 및 영어를 종단간 방식으로 읽어낼 수 있는 딥러닝 기반 문자 판독 네트워크를 제안했다. 네트워크는 2,350자의 한글을 분류할 수 있으면서도 문자 검출기와 인식기 구조를 간소화하여 종단간 학습이 가능하도록 설계했다. 한글 판독 모델 학습에 필요한 양질의 데이터셋이 부족한 상황에서 전이학습을 통해 50%에 가까운 판독 성능을 기록했으며, 이는 한글뿐 아니라 다국어 데이터셋을 모델 학습에 사용함으로써 데이터셋 부족 문제를 어느 정도 보완한 결과이다. 향후 한글과 같이 소수 언어에 특화된 문자 판독 모델을 구축하고자 할 때 전이학습이 효과적인 학습 방식으로 사용되기를 기대한다. 또한 이번 연구에서는 한글 판독 성능을 처음 측정한 연구인 만큼 가로 배열의 한글 문자열을 읽는 데 집중했다. 최근 장면 문자 판독 연구가 보다 다양한 배열의 문자를 읽어내는 방향으로 발전되고 있으므로, 한글 판독 역시 수직이나 대각선 형태로 배열된 단어를 읽을 수 있는 방법론을 추가 연구할 필요가 있다.

참고문헌

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019), Character region awareness for text detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9365-9374.

Baek, Y., Shin, S., Baek, J., Park, S., Lee, J., Nam, D., and Lee, H. (2020), Character Region Attention for Text Spotting, *ECCV 2020*, 504-521.

Bahdanau, D., Cho, K., and Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, ICLR 2015.

Bai, F., Cheng, Z., Niu, Y., Pu, S., and Zhou, S. (2018), Edit probability for scene text recognition, *CVPR 2018*, 1508-1516.

Busta, M., Neumann, L., and Matas, J. (2017), Deep TextSpotter: An end-to-end trainable scene text localization and recognition framework, *Proceedings of the IEEE International Conference on Computer Vision*, 2204-2212.

Busta, M., Patel, Y., and Matas, J. (2018), E2E-MLT - an unconstrained end-to-end method for multi-language scene text, *Asian Conference on Computer Vision 2018*, 127-143.

Cao, D., Zhong, Y., Wang, L., He, Y., and Dang, J. (2020), Scene text detection in natural images: A review, *Symmetry 2020*, **12**(12), 1956.

Cao, Y., Ma, S., and Pan, H. (2020) FDTA: Fully convolutional scene text detection with text attention, *IEEE Access*, **8**, 155441-155449.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., and Zhou, S. (2017), Focusing attention: Towards accurate text recognition in natural images, *2017 IEEE International Conference on Computer Vision*, 5086-5094.

Deng, D., Liu, H., Li, X., and Cai, D. (2018), PixelLink: Detecting scene text via instance segmentation, *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1).

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019), Dual attention network for scene segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3146-3154.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006), Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *Proceedings of the 23rd international conference on Machine learning*, 369-676.

He, K., Gkioxari, G., Dollár, P., and Girshick. R. (2017), Mask R-CNN, *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.

He, T., Tian, Z., Huang, W., Shen C., Qiao, Y., and Sun, C. (2018), An end-to-end textspotter with explicit alignment and attention, *Proceedings of CVPR*, 5020-5029.

Hochreiter, S. and Schmidhuber, J. (1997), Long short-term memory, *Neural Comput.*, **9**(8), 1735-1780.

Hong, J. and Kim, D. (2018), Natural Scene Text Detection and Recognition Framework for 1,000 Korean Characters based on CNN, *The Korean Institute of Information Scientists and Engineers KCC*, **45**(1), 1015-1017.

Huang, J., Pang, G., Kovvuri, R., Toh, M., Liang, K., Krishnan, P., Yin, X., and Hassner, T. (2021), A Multiplexed Network for End-to-End, Multilingual OCR, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 4547-4557.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015), Spatial transformer networks, *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, **2**, 2017-2025.

Kingma, D. and Ba, J. (2014), Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Lee, C. Y. and Osindero, S. (2016), Recursive recurrent nets with attention modeling for ocr in the wild, *IEEE Conference on Computer Vision and Pattern Recognition*, 2231- 2239.

Li, H., Wang, P., and Shen, C. (2017), Towards end-to-end text spotting with convolutional recurrent neural networks, *2017 IEEE International Conference on Computer Vision (ICCV)*, 5248-5256.

Liao, M., Shi, B., Bai, X., Wang, X., and Liu, W. (2017), Textboxes: A fast text detector with a single deep neural network, *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**(1).

Liao, M., Pang, G., Huang, J., Hassner, Tal., and Bai, X. (2020), Mask TextSpotter v3: Segmentation proposal net- work for robust scene text spotting, *Proceedings of the European Conference on Computer Vision (ECCV)*, 706-722.

Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017), Feature pyramid networks for object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117-2125.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A.C. (2016), SSD: Single shot multibox detector, *Computer Vision – ECCV 2016*, 21-37.

Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., and Yan, J. (2018), FOTS: Fast oriented text spotting with a unified network, *IEEE Conference on Computer Vision and Pattern Recognition*, 5676-5685.

Liu, Y., Jin, L., Zhang, S., and Zhang, S. (2017), Detecting curve text in the wild: New dataset and new solution, arXiv preprint arXiv:1712. 02170.

Long, S., He, X., and Yao, C. (2021), Scene text detection and recognition: The deep learning era, *International Journal of Computer Vision*, **129**, 161-184.

Lyu, P., Liao, M., Yao, C., Wu, W., and Bai, X. (2019), Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes, In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 532-548.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., and Xue, X. (2017), Arbitrary-oriented scene text detection via rotation proposals, *IEEE Transactions on Multimedia*, **20**(11), 3111-3122.

Milletari, F., Navab, N., and Ahmadi, S. A. (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation, *2016 Fourth International Conference on 3D Vision (3DV)*, 565-571.

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M. M., Burie, J., Liu, C., and Ogier, J. (2017), ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, **1**, 1454-1459.

Redmon, J. and Farhadi, A. (2017), YOLO9000: Better, faster, stronger, arXiv preprint.

Ronneberger, O., Fischer, P., and Brox. T. (2015), U-net: Convolutional networks for biomedical image segmentation, *MIC- CAI*, 234-241.

Shi, B., Bai, X., and Yao, C. (2017b), An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(11), 2298-2304.

Shi, B., Wang, X., Lyu, P., Yao, C., and Bai, X. (2016), Robust scene text recognition with automatic rectification, *IEEE Conference on Computer Vision and Pattern Recognition*, 4168-4176.

Shi, B., Yang, M., Wang, X., Lyu, P., Bai, X., and Yao C. (2018), ASTER: An attentional scene text recognizer with flexible rectification, *IEEE transactions on Pattern Analysis and Machine Intelligence*, **31**(11), 855-868.

Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., and Bai, X. (2017), ICDAR2017 competition on reading chinese text in the wild(RCTW-17), *Proceedings of ICDAR*, 1429-1434.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014), Sequence to sequence learning with neural networks, *In Advances in Neural Information Processing Systems*, 3104-3112.

Xing, L., Tian, Z., Huang, W., and Scott, M. R. (2019), Convolutional character networks, *Proceedings of the IEEE International Conference on Computer Vision*, 9126-9136.

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., and Liang, J. (2017), EAST: An efficient and accurate scene text detector, *CVPR 2017*, 2642-2651.

저자소개

김정원 : 고려대학교 경제학과에서 2015년 학사 학위를 취득하고 동 대학 산업경영공학과에서 석사과정에 재학 중이다. 연구분야는 인공지능, 머신러닝, 컴퓨터 비전이다.

김성범 : 고려대학교 산업경영공학부 교수로 2009년부터 재직하고 있으며, 인공지능공학연구소 소장 및 기업산학연협력센터 센터장을 역임했다. 미국 University of Texas at Arlington 산업공학과에서 교수를 역임하였으며, 한양대학교 산업공학과에서 학사 학위를 미국 Georgia Institute of Technology에서 산업공학 석사 및 박사학위를 취득하였다. 인공지능, 머신러닝, 최적화 방법론을 개발하고 이를 다양한 공학, 자연과학, 사회과학 분야에 응용하는 연구를 수행하고 있다.