Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 3, pp. 258-266, June 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.3.258


# 장면 이미지 속 한글 문자 인식을 위한 글자 단위 일관성 정규화 기반의 준지도학습 모델

김성수 · 김성범†

고려대학교 산업경영공학과

# Korean Scene Text Recognition Using Semi-Supervised Learning with Character-Level Consistency Regularization

Sungsu Kim · Seoung Bum Kim

Department of Industrial and Management Engineering, Korea University

Scene text recognition is a task that recognizes characters in scene images. Existing studies have been actively conducted based on English but little has been done on Korean. Because Korean does not have enough labeled data and has a large number of characters compared to English, it is more difficult to train Korean scene text recognition model. In addition, most of the previous studies trained the model using synthetic images rather than real images because of insufficient labeled data. However, using synthetic images can reduce generalization performance because of domain gap between real and synthetic images. In this study, we propose a Korean scene text recognition model using semi-supervised learning that overcomes the insufficient labeled data and domain gap. By using text alignment and consistency regularization specialized for Korean scene text recognition, we can obtain better performance than the existing supervised and semi-supervised scene text recognition models for three evaluation datasets. To the best of our knowledge, this is the first study that attempts semi-supervised learning for Korean scene text recognition.

***Keywords:*** Consistency Regularization, Korean Text Recognition, Scene Text Recognition, Semi-Supervised Learning, Text Alignment

## 1. 서 론

장면 이미지 속 문자 인식(scene text recognition)은 이미지 내 어떤 문자가 있는지 예측하는 과제이다. 이는 하얀색 배경과 검정색 글씨처럼 규격화된 인쇄체 글씨를 인식하는 광학 문자 인식(optical character recognition, OCR)과 유사한 특징을 갖는다. 하지만 장면 이미지 속 문자 인식은 여러 형태의 글꼴과 다양한 배경을 지닌 간판이나 책 표지와 같은 이미지들을 인식하는 과제로 OCR보다 난이도가 높은 연구분야이다. 최근 이러한 장면 이미지 속 문자 인식은 딥러닝을 활용하여 우수한 성능을 보여주고 있다(Chen *et al.*, 2021).

장면 이미지 속 문자 인식에 대한 선행연구는 영어를 위주로 활발하게 진행되고 있다(Shi *et al.*, 2016; Baek *et al.*, 2019; Aberdam *et al.*, 2021). 하지만 한글을 포함한 그 외 소수언어에 대해서는 연구 사례가 저조한 실정이다. 이는 전 세계에서 공용으로 활용되는 영어에 비해 데이터 수가 적은 것이 주된 이유이다. 또한 영어의 경우 분류해야 하는 글자가 대/소문자 각각 26개씩 총 52개가 존재하지만, 한글의 경우 한국 산업 규격으로 지정된 한국어 문자 집합인 완성형(KS×1001) 기준 2,350개의 글자를 분류해야만 한다. 더불어 한글은 자음과 모

This research was supported by the BK21 FOUR funded by the Ministry of Education of Korea and National Research Foundation of Korea.
† 연락저자 : 김성범 교수, 02841, 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2022년 12월 24일 접수; 2023년 1월 17일 수정본 접수; 2023년 1월 19일 게재 확정.

음으로 구성된 복합적인 언어로 '가'와 '갸'처럼 상대적으로 유사한 글자들이 다수 존재하기에 정교한 기술을 필요로 한다. 이에 따라 영어보다 복잡한 한글은 장면 이미지 속 문자 인식에 어려움이 존재하여 다양한 연구사례가 존재하지 않는다.

활발히 연구가 진행되고 있는 영어에 대한 장면 이미지 속 문자 인식에 관한 선행연구들은 합성 이미지를 활용한 지도학습(supervised learning)을 기반으로 수행되었다(Shi *et al.*, 2018; Baek *et al.*, 2019). 합성 이미지는 일반적인 이미지에 글자를 삽입하여 만든 인위적인 이미지로, 데이터 수집 및 레이블링 비용이 적다는 장점이 있다. 반면 실제 이미지의 경우 레이블링을 위해 원본 이미지에서 문자열의 위치를 찾아 해당 부분만 잘라내고 해당 문자가 무엇인지 표기해야 하기에 시간적 비용이 크다는 한계가 존재한다. 이처럼 학습할 만큼 충분한 실제 이미지 기반의 레이블을 갖는 데이터(labeled data)를 수집하는 것은 많은 비용이 소모된다. 따라서 선행연구들은 수집하기 용이한 합성 이미지를 모델 학습에 활용하여 연구를 수행하였다. 하지만 이러한 합성 이미지로 학습할 경우 실제 이미지와 합성 이미지 사이에 도메인 차이(domain gap)가 나타나기에, 모델이 충분히 학습했더라도 일반화 성능이 떨어질 수 있다는 한계가 존재한다.

이와 같은 데이터의 부족 그리고 도메인 차이를 완화하기 위해서 합성 이미지와 실제 이미지 기반의 레이블이 없는 데이터(unlabeled data)를 함께 활용한 연구들도 수행되고 있다(Aberdam *et al.*, 2021; Zheng *et al.*, 2022). 하지만 합성 이미지가 포함되어 있기 때문에 도메인 차이는 완전히 극복하기 어렵다는 한계점이 존재한다. 따라서 최근 장면 이미지 속 문자 인식에 대한 연구는 이러한 도메인 차이를 극복하기 위해서, 합성 데이터가 아닌 소수의 labeled 데이터와 다수의 unlabeled 데이터를 활용한 연구가 수행되고 있다(Baek *et al.*, 2021). Unlabeled 데이터는 labeled 데이터에 비해 수집 비용이 낮고, 합성 이미지가 아니기에 도메인 차이가 존재하지 않는 장점이 있다. 그러나 unlabeled 데이터를 활용하여 장면 이미지 속 문자 인식을 수행한 사례는 영어에만 한정되어 있으며, 장면 이미지 속 한글 문자 인식에 대한 연구사례는 전무한 실정이다.

이에 따라 본 연구는 labeled 데이터와 unlabeled 데이터를 동시에 사용하는 준지도학습(semi-supervised learning) 기반 장면 이미지 속 한글 문자 인식 모델을 제안하고자 한다. 이때, unlabeled data를 활용함으로써 데이터가 부족한 한계를 극복하였으며, 합성 데이터를 사용하지 않음으로써 도메인 차이를 극복하였다. 또한 자음과 모음으로 구성되어 글자 자체에 정보가 많은 한글의 특징을 고려하여 글자 단위로 모델링을 수행하였다. 본 연구에서 제안하는 모델은 이미지 분류 문제에서 활용된 준지도학습 모델인 Sohn *et al.*(2020)에서 고안되었다. Sohn *et al.*(2020)의 준지도학습 모델은 해당 분야에서 우수한 성능을 달성했고, 비교적 간단한 구조를 갖고 있어 학습 및 응용이 용이하다. 본 연구에서 제안하는 모델은 Sohn *et al.*(2020)에서 활용되는 이미지 단위 일관성 정규화가 아닌, 글자 자체에 정보가 많은 한글의 특성을 반영한 글자 정렬과 글자 단위 일관성 정규화 기법을 새롭게 제안한다.

장면 이미지 속 문자 인식 모델에서 글자 단위로 일관성 정규화를 통해 학습하는 것은 이전 Zheng *et al.*(2022)에서 영어 문자에 대해 최초로 제안되었다. 하지만 해당 방법론은 몇 가지 한계점이 존재한다. 첫째, 글자들이 예측된 각 확률을 모두 곱하여 임계값(threshold)과 비교할 점수(score)를 정의한다. 이는 0과 1사이의 확률들을 곱하여 점수를 계산하기 때문에, 문자열의 길이가 길어질수록 각 글자에 대한 예측 확률이 높음에도 불구하고 해당 이미지의 점수는 낮아질 수 있다는 한계가 존재한다. 예를 들어 '농수산물도매시장'이라는 문자열에 대하여 각 글자를 90%의 확률로 우수하게 예측했음에도 불구하고, 해당 이미지의 점수는 $0.9^8$(≒ 0.4305)로 낮게 산출될 수 있다. 둘째, 글자 단위 일관성 정규화를 기반으로 학습하지만 임계값과 비교하는 점수를 전체 이미지 단위에 적용한다. 이는 이미지 내 학습에 도움이 될 수 있는 글자들까지 학습에 활용하지 않는다는 한계가 존재한다. 예를 들어 '항정살'이라는 문자열을 각각 '0.95', '0.8', '0.01'로 예측했다면, 비교적 모호하게 예측된 '살'이라는 글자의 영향으로 점수가 '0.95×0.8×0.01 = 0.0076'로 낮게 산정되어 해당 이미지가 학습에 활용되지 않을 수 있다. 이때, 각 이미지 단위로 임계값과 비교하기 때문에 우수하게 예측된 '항'과 '정'은 학습에 도움이 될 수 있지만 '살'의 영향으로 이미지 내 모든 글자들이 학습에 활용되지 않는다. 이에 따라 본 연구는 위 두 가지 한계를 극복하기 위하여 이미지가 아닌 각 글자 단위로 점수를 산정하고, 각 점수를 전체 이미지가 아닌 글자 단위로 비교함으로써 우수한 성능을 달성할 수 있었다. 본 연구의 주요 기여점은 아래와 같다.

- 본 연구는 labeled 데이터가 부족한 상황에서 효과적인 준지도학습을 장면 이미지 속 한글 문자 인식에 활용한 모델을 제안하며, 이는 우리가 파악하고 있는 한 장면 이미지 속 한글 문자 인식 분야에서 최초의 시도이다.
- 글자 자체에 정보량이 많은 한글의 특징을 고려한 글자 정렬 및 글자 단위 일관성 정규화 기반 장면 이미지 속 문자 인식 모델을 새롭게 제안하여 기존 지도학습 및 준지도학습 모델 대비 성능 향상을 이루었다.

본 논문은 다음과 같이 구성된다. 제2장에서는 최근 연구된 장면 이미지 속 문자 인식 및 준지도학습을 활용한 장면 이미지 속 문자 인식 모델들을 소개한다. 제3장에서는 본 연구에서 제안하는 준지도학습을 활용한 장면 이미지 속 한글 문자 인식 모델을 설명한다. 제4장에서는 본 연구에서 활용한 데이터셋과 실험조건 및 실험결과를 제시한다. 마지막으로 제5장에서는 결론과 함께 추후 연구방향을 제시한다.

## 2. 관련연구

### 2.1 장면 이미지 속 문자 인식

장면 이미지 속 문자 판독(scene text spotting)은 이미지 내에서 문자의 위치를 찾고 어떤 문자인지 인식하는 연구분야이다. 이러한 장면 이미지 속 문자 판독은 크게 장면 이미지 속 문자 탐지(scene text detection)와 장면 이미지 속 문자 인식으로 구분된다. 장면 이미지 속 문자 탐지는 원본 이미지에서 글자가 있는 위치를 식별하는 과제이며, 장면 이미지 속 문자 인식은 식별된 위치에 존재하는 문자들을 예측하는 과제이다. 이때, 장면 이미지 속 문자 인식 모델은 하나의 이미지가 입력으로 들어오면 여러 개의 글자를 순차적으로 반환하는 구조로, 1개의 입력 데이터에 대해 여러 개의 출력 값을 갖는다. 이는 일반적인 분류 문제와는 달리 각 글자 단위로 예측한다는 특징을 가지며, 각 글자를 시퀀스(sequence)로 보았을 때 시퀀스 분류 과제로 볼 수 있다. 최근 연구들은 장면 이미지 속 문자 인식을 위해서 합성곱 신경망(convolution neural network, CNN)을 이용하여 특징을 추출한 후, 디코딩 과정에서 글자 시퀀스들을 예측한다(Shi *et al.*, 2016; Baek *et al.*, 2019). 이때, 디코딩 과정에서 일반적으로 2가지 방법론이 활용된다. 첫번째는 connectionist temporal classification(CTC; Graves *et al.*, 2006)으로 디코딩 시 각 시퀀스별로 예측한 후 공백 토큰(blank token)을 활용하여 중복되는 단어들을 제거하는 등의 후처리를 수행하는 특징을 갖는다. CTC를 사용한 대표적인 연구로는 Shi *et al.*(2016)와 Busta *et al.*(2017)이 존재한다. 두번째는 어텐션(attention; Bahdanau *et al.*, 2014) 메커니즘이다. 이는 주로 sequence-to-sequence(seq2seq; Sutskever *et al.*, 2014)모델에 어텐션 메커니즘을 적용한 디코더 구조를 활용하며 각 시퀀스 별 글자와 이미지간 관계를 학습한 후 글자들을 순차적으로 디코딩한다. 이때 <sos>토큰을 입력 받아 디코딩 시작 여부를 판단하고, 문자열이 모두 디코딩이 되었다면 <eos> 토큰을 출력하여 디코딩 완료 여부를 판단한다. Shi *et al.*(2018)와 <Figure 1>에 명시된 Baek *et al.*(2019)이 대표적으로 어텐션 메커니즘을 활용한 구조이며, 어텐션 메커니즘을 활용함으로써 CTC보다 우수한 성능을 달성할 수 있었다.

### 2.2 준지도학습을 활용한 장면 이미지 속 문자 인식

준지도학습은 labeled 데이터가 부족할 때 unlabeled 데이터를 함께 학습에 활용하는 방법론으로, 자가학습(self-training; Scudder, 1965)과 일관성 정규화를 위주로 활발하게 연구가 진행되고 있다. 자가학습은 labeled 데이터로 모델을 학습하고 해당 모델로 unlabeled data를 예측한 후, 신뢰도가 높은 데이터들을 선별하여 기존의 학습데이터와 합쳐서 모델을 반복적으로 학습한다. 이러한 자가학습 기반의 모델은 Lee(2013)가 대표적이다. 또한 일관성 정규화는 원본 이미지와 원본 이미지에 변형을 가하거나 다른 모델에 들어갔을 때 예측 결과들이 일관성을 갖도록 학습한다. Tarvainen *et al.*(2017), Xie *et al.*(2020), Sohn *et al.*(2020)은 일관성 정규화를 기반으로 학습하는 준지도학습 모델로, 자가학습보다 우수한 성능을 보였다.

더불어 장면 이미지 속 문자 인식에 unlabeled 데이터를 활용하기 위해 준지도학습을 적용하여 우수한 성능을 보인 연구들도 존재한다. Baek *et al.*(2021)은 준지도학습 중 Lee(2013)와 Tarvainen *et al.*(2017)를 장면 이미지 속 문자 인식에 적용하여 지도학습보다 우수한 성능을 얻었다. 또한 Zheng *et al.*(2022)은 합성 이미지를 활용했지만 준지도학습의 일관성 정규화를 기반으로 학습했다는 특징을 가지며, 글자 단위 일관성 정규화를 장면 이미지 속 문자 인식에 최초로 적용한 연구이다. 이는 한 이미지를 두 번 증강(augmentation)한 후 약하게 증강(weak augmentation)된 이미지에 대해 점수를 산출한다. 이때, 점수는 각 단어로 예측될 확률을 모두 곱하여 정의한다. 만약 점수가 임계값보다 더 크다면, 약하게 증강된 이미지와 강하게 증강(strong augmentation)된 이미지에 대해 일관성을 갖도록 학습하고, 그렇지 않으면 학습에 활용하지 않는다. 이때, 증강된 두 이미지를 유사하게 정렬하기 위하여 디코딩 시 약하게 증강된 이미지의 각 시퀀스 별 출력값을 공유하는 특징을 갖는다.

## 3. 제안방법론

본 연구에서는 장면 이미지 속 한글 문자 인식을 위한 글자 단위 일관성 정규화(Character-Level Consistency Regularization, ChaLCoR)기반의 준지도학습 모델을 제안하며, 이에 대한 전체적인 구조를 <Figure 2>에 도식화하였다. ChaLCoR은 글자 자체에 정보량이 많은 한글의 특징을 반영하여 문자열 내 글자들의

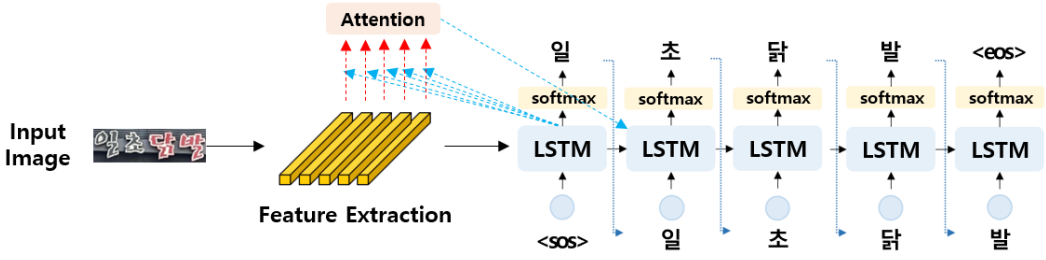



**Figure 1.** Architecture of Attention-based Scene Text Recognition Model

정렬 여부를 고려하고 글자 단위로 임계값과 비교하는 글자 단위 일관성 정규화를 적용한다는 특징을 갖는다. ChaLCoR의 학습 과정은 지도학습과 비지도학습으로 구분되며, 공통적으로 RandAugment(Cubuk *et al.*, 2020)로 데이터를 증강한 후 학습한다. 그러나 장면 이미지 속 문자 인식 모델의 특성상 RandAugment 내 Mixup(Zhang *et al.*, 2017)과 Cutout(DeVries *et al.*, 2017) 증강 기법은 글자 고유의 본질을 해칠 수 있다고 판단하여 데이터 증강에 활용하지 않았다. 먼저 ChaLCoR의 지도학습은 labeled 데이터를 학습에 활용하고, 글자가 디코딩 될 때마다 시퀀스별로 예측된 글자와 실제 레이블을 비교하여 학습을 수행한다. 비지도학습은 unlabeled 데이터를 학습에 활용하며, 약하게 증강된 이미지와 강하게 증강된 이미지 사이에 일관성을 갖도록 학습한다. 이때, 기존 Sohn *et al.*(2020)처럼 이미지 단위로 학습하는 것이 아닌 장면 이미지 속 문자 인식 모델과 한글의 특성을 고려하여 글자 단위로 학습한다는 특징을 가진다.

지도학습은 labeled 데이터를 약하게 증강한 후 모델에 통과시켜 학습을 수행한다. 본 연구에서는 이러한 labeled 데이터를 입력받는 장면 이미지 속 문자 인식 모델로 Baek *et al.*(2019)를 활용하였다. Baek *et al.*(2019)는 TPS변환(Jaderberg *et al.*, 2015), ResNet(He *et al.*, 2016), BiLSTM(Schuster *et al.*, 1997)기반의 인코더와 어텐션 기반의 디코더로 구성된 장면 이미지 속 문자 인식 모델로, 각 시퀀스 별 글자와 이미지간 관계를 학습하며 글자들을 순차적으로 디코딩한다는 특징을 갖는다. 이때, ChaLCoR은 이미지 단위가 아닌 글자 단위로 학습을 수행하기 때문에 디코딩 과정에서 예측된 각 글자 시퀀스의 확률에 대하여 교차 엔트로피(cross entropy)를 통해 지도학습에 대한 손실함수($L_{labeled}$)를 산출한다. 본 연구에서는 각 $B$개의 데이터에 대하여 최대 시퀀스 길이가 $T$일 때, labeled 데이터를 $X=\{(X_b,\ P_b):b\in(1,\ ...,\ B)\}$ 로 정의하였다. 장면 이미지 속 문자 인식은 시퀀스 분류 과제이므로 각 시퀀스를 고려하여 레이블 값을 $P_b=\{p_{b,t}:t\in(1,\ ...,\ T)\}$ 의 형태로 세분화하였다. 또한 두 분포 $p$와 $q$가 주어졌을 때 두 분포에 대한 교차 엔트로피는 $H(p,\ q)$로, 약한 데이터 증강과 강한 데이터 증강은 각각 $\alpha(\bullet)$ 및 $A(\bullet)$로 표기하였다. 추가적으로, 주어진 이미지 $x$에 대하여 산출된 각 시퀀스들의 확률분포는 $p_m(y\mid x)$로 정의하였다. 결론적으로 지도학습에 대한 손실함수는 아래 식 (1)과 같으며 <sos> 토큰과 <eos> 토큰은 제외하고 계산하였다.

$$L_{labeled}=\frac{1}{B}\frac{1}{T}\sum_{b=1}^{B}\sum_{t=1}^{T}H(p_{b,t},\ p_m(y|\alpha(x_b))) \tag{1}$$

비지도 학습은 unlabeled 데이터를 활용하며, 크게 4단계로 구성된다. 첫 단계에서는 unlabeled 데이터를 약하게 1번, 강하게 1번 증강한 후 증강된 이미지들을 모델에 각각 통과시킨다. 이때, unlabeled 데이터를 입력받는 장면 이미지 속 문자 인식 모델은 지도학습과 동일하게 Baek *et al.*(2019)을 활용하였다. 두 번째 단계에서는 증강된 두 데이터의 예측 값을 통해 글자 정렬 여부를 확인하는 과정을 거친다. 이러한 과정은 증강된 두 이미지가 각각 $L$과 $L'$의 서로 다른 길이를 가진 문자열로 예측될 경우 각 예측된 글자 시퀀스가 $|L-L'|$만큼 밀릴 수 있다는 것에서 기인하였다. 각 글자단위로 학습하는 것이 효과를 가지기 위해서는 동일하게 예측된 결과에 대하여 일관성을 갖도록 학습을 해야 하지만, 서로 다른 글자에 대해 일관성을 갖도록 학습한다면 오히려 혼란을 줄 수 있다. 이에 대한 예시는 <Figure 3>과 같다.

<Figure 3(a)>처럼 올바르게 정렬된 두 값에 대하여 학습해야 하지만, <Figure 3(b)>처럼 글자가 밀릴 경우 서로 다른 값

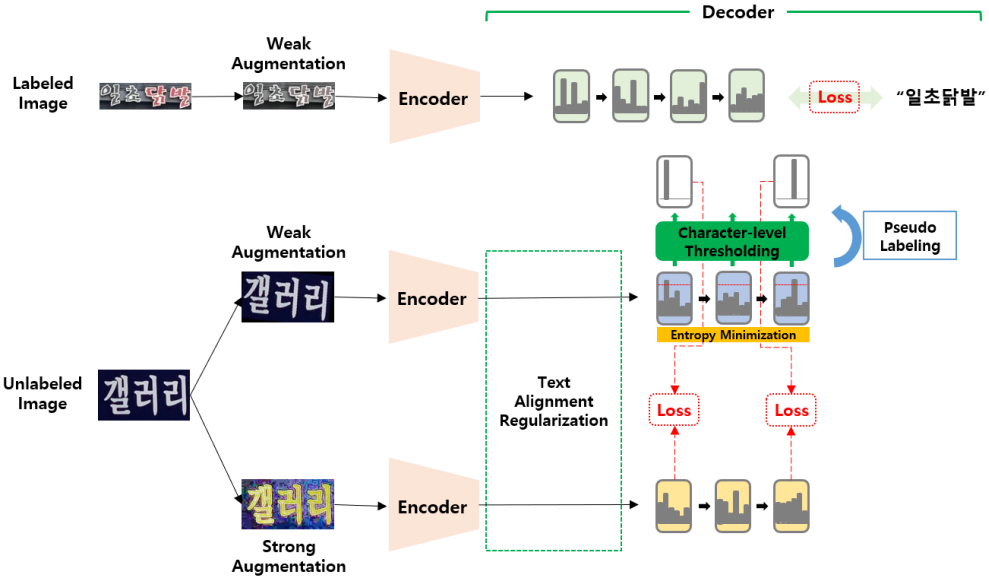
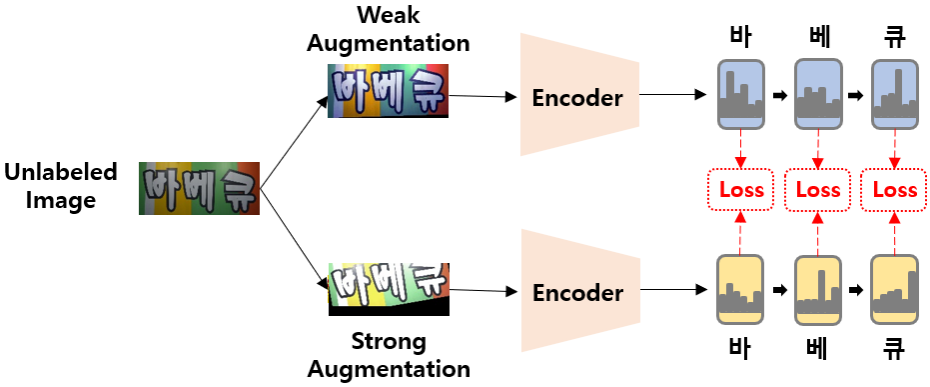
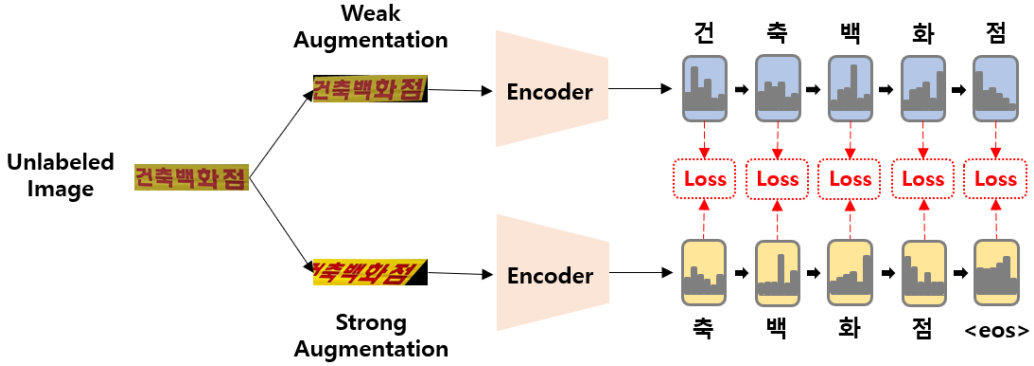



**Figure 2.** Overview of the Proposed Scene Text Recognition Model




(a) well aligned example




(b) poorly aligned example.

**Figure 3.** Example for Text Alignment

에 대해 일관성을 갖도록 학습하는 오류가 생길 수 있다. 이에 따라 본 연구는 증강된 두 이미지에 대해 문자열 길이를 다르게 예측한 이미지는 학습에 방해가 될 수 있다고 판단하여 학습에서 제외하였다. 즉, 증강된 두 이미지의 예측된 문자열 길이가 동일한 이미지만 선별하여 학습에 반영함으로써 발생 가능한 노이즈를 최소화하였다. 세 번째 단계에서는 글자 단위로 임계값과 비교하는 과정을 수행한다. 기존 Sohn *et al.*(2020)은 단순히 이미지 분류에 적용하였기 때문에 예측 결과에 시퀀스가 존재하지 않았지만, 장면 이미지 속 문자 인식은 예측 결과인 글자가 시퀀스로 반환되기 때문에 이를 고려해야만 한다. 기존 Sohn *et al.*(2020)이 전체 이미지 단위로 임계값과 비교하는 과정을 수행했다면 본 연구는 각 예측된 시퀀스 단위로 임계값과 비교하였다. 예를 들어, 약하게 증강된 이미지의 시퀀스 예측 결과가 각각 '갤', '러', '리'이고 각 시퀀스로 예측한 확률이 '0.95', '0.8', '0.92'이며 임계값을 0.9라고 가정했을 때, '갤'과 '리'만 학습에 활용하였다. 이를 통해 본 연구는 예측이 모호한 글자들은 학습에 반영하지 않고, 학습에 도움이 될 수 있는 글자들만 선별함으로써 노이즈 시퀀스의 영향을 최소화하고자 하였다. 이와 같이 글자 단위 학습 시 비교하는 임계값을 본 연구에서는 $\beta_u$로 정의하였다.

추가적으로 본 연구에서는 예측된 글자의 확률을 보정하기 위하여 약하게 증강된 이미지에서 각 시퀀스 별 확률분포에 대해 엔트로피 최소화(entropy minimization; Grandvalet *et al.*, 2004)를 적용하였다. 한글의 경우 글자가 서로 유사하고 분류해야 할 글자 수가 2,350개로 많기에 예측확률이 모호할 수 있지만, 엔트로피 최소화를 통해 높은 확률을 강조함으로써 글자의 모호함을 해소하고자 하였다. 본 연구에서는 unlabeled 데이터를 $U=\{U_b : b\in(1, ..., B)\}$ 로 정의하였고, 이들 중 약하게 증강된 이미지의 예측 값에 대한 확률분포는 $q_b = p_m(y \mid \alpha(U_b))$로 정의하였다. 아울러 각 시퀀스를 고려하여 $q_b = \{q_{b,t} : t\in(1, ..., T)\}$ 로 세분화하였다. 또한 전체 글자 개수가 $N$일 때, 각 시퀀스의 확률분포 $q_{b,t}$를 산출하는 소프트맥스(softmax) 함수를 거치기 전 시퀀스 벡터를 $z_{b,t} = \{z_{b,t,C} : C\in(1, ..., N)\}$ 로 표기하였다. 이때, 엔트로피 최소화 상수가 $\tau$일 때, $b$번째 데이터에서 $t$번째 시퀀스에 해당하는 $C$번째 글자의 예측 확률($q_{b,t}(C)$)에 대한 엔트로피 최소화 수식은 아래 식 (2)와 같다.

$$q_{b,t}(C) = \frac{exp(z_{b,t,C}/\tau)}{\sum_{k=1}^{N} \exp(z_{b,t,k}/\tau)} \tag{2}$$

네 번째 단계로 임계값을 넘은 시퀀스들은 약하게 증강된 이미지의 시퀀스 예측 결과를 원-핫 벡터 형태의 임의 레이블로 바꾸어 강하게 증강된 이미지의 시퀀스 예측 결과와 일관성

정규화를 활용하여 교차 엔트로피로 학습한다. 이때, 약하게 증강된 이미지를 통해 생성한 임의 레이블과 각 시퀀스를 고려한 표기를 각각 $\hat{q}_b = argmax(q_b)$, $\hat{q}_b = \{\hat{q}_{b,t} : t \in (1, ..., T)\}$ 로 정의했을 때, 각 이미지 내 개별적인 글자 시퀀스의 학습여부($I_{b,t}$)는 아래 식 (3)과 같으며, 비지도학습에 대한 손실함수($L_{unlabeled}$)는 아래 식 (4)와 같다.

$$I_{b,t} = len(p_m(y \mid A(x_b))) == len(p_m(y \mid \alpha(x_b))) \cdot l(max(q_{b,t} \geq \beta_u)) \quad (3)$$

$$L_{unlabeled} = \frac{1}{B}\frac{1}{T}\sum_{b=1}^{B}\sum_{t=1}^{T} I_{b,t} \cdot H(\hat{q}_{b,t}, p_m(y|A(x_b))) \quad (4)$$

최종적인 손실함수($L_{overall}$)는 지도학습에서 산출된 손실함수와 비지도학습에서 산출된 손실함수를 가중 합하여 정의한다. 이때, $\lambda_u$는 비지도학습에서 산출된 손실함수에 대한 가중치로 비지도학습의 손실함수를 전체 손실함수에 얼마나 반영할지 결정한다. 전체 손실함수 식은 아래 식 (5)와 같다.

$$L_{overall} = L_{labeled} + \lambda_u L_{unlabeled} \quad (5)$$

## 4. 실험결과

### 4.1 데이터셋

본 연구에서 사용한 데이터셋은 International Conference on Document Analysis and Recognition(ICDAR) 2017의 장면 이미지 속 문자 인식 데이터셋(https://rrc.cvc.uab.es/), AI Hub (www.aihub.or.kr)의 야외 실제 촬영 이미지 중 간판(sign) 이미지 데이터셋과 책 표지(book cover) 이미지 데이터셋이다. 학습, 검증 및 평가에 세 개의 데이터셋을 모두 사용하였고, 각 문자열 이미지에서 한글만 존재하는 이미지들을 활용하여 실험을 수행하였다. 또한 가로형 문자열과 세로형 문자열은 이미지의 특성상 차이가 있다고 판단하여, 가로의 길이가 세로보다 긴 이미지들만 활용하였다.

ICDAR 2017 데이터셋은 도로/건물 등에서 촬영한 여러 국가의 문자열이 포함된 이미지 데이터셋으로, 학습 데이터와 평가 데이터가 각각 68,623개, 16,255개씩 존재한다. 이때, 가로의 길이가 세로보다 긴 한글 이미지 데이터만 선별하여 학습 및 평가 데이터 각각 3,868개, 875개를 실험에 활용하였다. 추가적으로 준지도학습 모델링을 위해 학습 데이터는 labeled 데이터, unlabeled 데이터, 검증 데이터로 각각 1,934개, 1,547개, 387개씩 활용하였다. 또한 AI Hub의 간판 이미지 데이터셋은 국내 전국의 간판, 안내판, 광고물, 현수막 등을 촬영한 이미지 데이터셋이다. 이는 장면 이미지 속 문자 인식에 적합한 데이터셋이 아닌 장면 이미지 속 문자 탐지(detection) 및 장면 이미지 속 문자 인식을 모두 고려한 데이터셋으로 이미지 내에 1개 이상의 문자열이 존재한다. 이를 장면 이미지 속 문자 인식 모델의 학습에 적합한 데이터셋으로 구축하기 위해 문자열이 있는 위치만 잘라낸(crop) 후 한글 문자 및 가로형 문자열만 선별하여 학습 데이터와 평가 데이터 각각 474,800개, 58,093개를 실험에 활용하였다. 추가적으로, ICDAR 2017 데이터셋과 같이 준지도학습 학습환경을 위해서 학습 데이터를 labeled 데이터, unlabeled 데이터, 검증 데이터로 각각 233,210개, 175,318개, 66,272개씩 활용하였다. 이때, 각 데이터는 간판 이미지의 소분류 8개에 대한 데이터셋의 비율을 고려하여 분할하였다. 마지막으로 AI Hub의 책 표지 이미지 데이터셋은 실제 책 표지를 촬영한 데이터셋으로, AI Hub의 간판 이미지 데이터셋과 동일한 구조로 되어있다. 이에 따라 문자열이 있는 위치만 잘라내고 가로형 한글 문자열만 선별하여 학습 데이터와 평가 데이터 각각 45,403개, 5,882개를 실험에 활용하였다. 여기서도 준지도학습 환경을 위해 학습 데이터를 labeled 데이터, unlabeled 데이터, 검증 데이터로 각각 16,001개, 17,835개, 11,567개씩 활용하였다.

이처럼 학습, 검증 및 평가의 목적으로 분리된 세 개의 데이터셋에 대하여 본 연구는 학습 데이터로 모델을 충분히 학습한 후, 검증 데이터로 최적의 하이퍼파라미터를 탐색하였고, 선정된 최적모델에 대해 평가 데이터로 예측성능을 측정하였다. 또한 본 제안 방법론이 labeled 데이터가 적은 상황에서도 효과적인지 입증하기 위하여 labeled 데이터의 비율을 10%(19,470개), 15%(29,205개), 25%(48,675개), 100%(194,700개)로 나누어 추가적으로 실험하였다. 이때, labeled 데이터들의 비율만 조정하고, unlabeled 데이터나 검증 데이터는 수정하지 않고 적용하였다. 데이터셋의 세부 정보는 <Table 1>과 같다.

### 4.2 평가지표 및 학습 세부사항

본 연구는 선행연구에서 주요 지표로 활용되는 이미지 단위

**Table 1.** Summary of the Training and Evaluation Datasets

| Dataset | Training | | | Evaluation |
|---|---|---|---|---|
| | Unlabeled | Labeled | Validation | Test |
| ICDAR 2017 | 1,934 | 1,547 | 387 | 875 |
| AI Hub Sign | 16,001 | 17,835 | 11,567 | 5,882 |
| AI Hub Book Cover | 233,210 | 175,318 | 66,272 | 58,093 |
| Total | 251,145 | 194,700 | 78,226 | 64,850 |

정확도(accuracy)를 평가척도로 활용하였다. 예를 들어 '갤러리'라는 이미지가 모델 예측 결과 '갤러리'로 정확하게 인식하였다면 정답으로 여기지만, '갤리리'나 '갤러' 등 한 글자라도 인식하지 못하였을 경우 오답으로 간주하였다. 또한 제안 모델인 ChaLCoR의 우수성을 입증하기 위하여 다음과 같이 실험환경을 구성하였다. 모든 모델은 100에폭(epoch)을 학습하였으며, 최적화 알고리즘은 Adadelta(Zeiler, 2012)를 활용하였고, 학습률(learning rate)은 1, 그리고 배치 사이즈(batch size)는 128을 사용하였다. 추가적으로, ChaLCoR의 준지도학습 모델에서 활용되는 최대 시퀀스의 길이는 25, 임계값은 0.9, 비지도학습의 손실함수 가중치인 $\lambda_u$는 1로 적용하였다. 또한 엔트로피 최소화에 활용되는 $\tau$는 각각 레이블을 갖는 데이터가 10%, 15%, 25%, 100% 일 때, 0.8, 0.6, 0.4, 0.8로 다르게 적용하였다. 마지막으로 학습할 글자의 총 개수는 완성형(KS×1001) 한글 기준 2,350개를 활용하였으며, 동등한 평가를 위해 다른 비교 모델들도 동일한 실험환경을 구성하였다.

### 4.3 실험 결과 및 분석

Labeled 데이터의 비율에 따른 ChaLCoR과 비교 방법론들의 성능은 <Table 2>와 같으며, 각 데이터셋에 따라 가장 높은 성능에 굵은 글씨로 표기하였다. 제안 방법론인 ChaLCoR과 비교대조군의 준지도학습 방법론들 내 장면 이미지 속 문자 인식 모델은 지도학습에서 비교 모델로 활용한 TRBA(Baek *et al.*, 2019)를 활용하였다. 준지도학습 방법론에 나타난 pseudo labeling과 mean teacher는 Baek *et al.*(2021)를 재현한 결과이며, UDA 및 FixMatch는 기존 Xie *et al.*(2020)와 Sohn *et al.*(2020)의 방법론을 참고하여 재구성한 결과이다. 본 연구에서 제안하는 ChaLCoR은 지도학습 그리고 다른 준지도학습 모델들과 비교했을 때, 가장 우수한 성능을 달성하였다. Mean teacher기반의 모델을 제외한 다른 준지도학습 모델들은 지도학습에 비해 우월한 성능을 도출하지 못하였다. 특히, labeled 데이터의 비율이 10%로 가장 적을 때 지도학습보다 낮게 성능이 도출되었다. 그러나 ChaLCoR은 지도학습보다 전반적으로 우수한 성능을 보여주었다. 또한 준지도학습 모델 중 mean teacher기반의 모델이 지도학습보다 성능이 좋았지만, ChaLCoR은 mean teacher기반의 모델보다 전반적으로 우수한 성능을 보여주었다. 이는 제안 방법론인 ChaLCoR이 기존의 준지도학습 모델보다 unlabeled 데이터를 효과적으로 활용함으로써 labeled 데이터가 부족한 한계를 극복할 수 있음을 시사한다.

### 4.4 구성 요소별 성능 기여도 평가

첫 번째 실험은 ChaLCoR이 제안하는 글자 정렬 및 글자 단위 일관성 정규화의 효과를 입증하려는 목적으로 수행되었다. 추가적으로, 글자 단위로 학습하는 것은 앞서 언급한 것처럼 Zheng *et al.*(2022)에서도 찾아볼 수 있었는데, ChaLCoR이 Zheng *et al.*(2022)의 글자 단위 일관성 정규화보다 더 효과적인지 실험을 통해 증명하고자 하였다. 실험결과는 <Table 3>과 같으며, 각 데이터셋 별 가장 우수한 성능과 두 번째로 우수한 성능에 각각 굵은 글씨와 밑줄로 표기하였다. 또한 <Table 3>에는 선행연구(FixMatch 및 Zheng *et al.*(2022))에 대한 실험

**Table 2.** Comparison of Accuracy among the Supervised, Semi-supervised, and Proposed Models for Korean Text Recognition with Different Percentages of Labeled Data Used (10%, 15%, 25%, and 100%). Bold Indicates the Best Accuracy among Comparative Models.

| Type | Model | ICDAR 2017 | | | | AI Hub Book Cover | | | | AI Hub Sign | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% |
| Supervised Learning | TRBA | 67.8 | 71.8 | 76.0 | 80.6 | 89.0 | 90.9 | 93.2 | 95.3 | 72.4 | 82.7 | 88.9 | 96.1 |
| Semi-supervised Learning | Pseudo labeling | 64.7 | 72.3 | 75.7 | 81.1 | 86.6 | 90.3 | 92.8 | 95.3 | 69.6 | 81.5 | 88.3 | 95.8 |
| | Mean teacher | 68.7 | 74.2 | 77.4 | 82.2 | **89.5** | 92.2 | 93.5 | 96.0 | 76.2 | 85.8 | 90.9 | 96.6 |
| | UDA | 64.2 | 73.1 | 76.6 | 81.6 | 86.4 | 90.7 | 92.5 | 95.8 | 72.0 | 84.0 | 89.9 | 96.3 |
| | FixMatch | 65.6 | 73.8 | 77.0 | 81.6 | 86.4 | 88.9 | 93.2 | 95.8 | 73.1 | 83.8 | 90.9 | 97.1 |
| Proposed | ChaLCoR | **71.4** | **76.2** | **77.9** | **82.4** | 88.4 | **92.3** | **94.2** | **96.4** | **78.4** | **89.4** | **93.2** | **97.2** |

**Table 3.** Effect of Text Alignment and Character-level Consistency Regularization in ChaLCoR

| Model | ICDAR 2017 | | | | AI Hub Book Cover | | | | AI Hub Sign | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% |
| FixMatch | 65.6 | 73.8 | 77.0 | 81.6 | 86.4 | 88.9 | 93.2 | 95.8 | 73.1 | 83.8 | 90.9 | 97.1 |
| Zheng *et al.*(2022) | 59.4 | 75.1 | 77.9 | **82.4** | 80.7 | 91.2 | 93.1 | 95.6 | 65.1 | 88.1 | 92.5 | 97.0 |
| ChaLCoR(w/o Text Alignment) | 69.0 | 76.0 | **78.4** | 82.3 | 88.3 | 92.1 | **94.2** | 95.9 | 77.7 | 88.6 | 93.0 | 96.9 |
| ChaLCoR | **71.4** | **76.2** | 77.9 | **82.4** | **88.4** | **92.3** | **94.2** | **96.4** | **78.4** | **89.4** | **93.2** | **97.2** |

**Table 4.** Effect of Entropy Minimization in ChaLCoR

| Model | ICDAR 2017 | | | | AI Hub Book Cover | | | | AI Hub Sign | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% |
| ChaLCoR (w/o entropy minimization) | 68.8 | 75.3 | 77.6 | **82.4** | 88.3 | 91.4 | 93.8 | **96.4** | **78.5** | 87.8 | 92.7 | **97.2** |
| ChaLCoR | **71.4** | **76.2** | **77.9** | **82.4** | **88.4** | **92.3** | **94.2** | **96.4** | 78.4 | **89.4** | **93.2** | **97.2** |

**Table 5.** Effect on One-hot Labeling in ChaLCoR

| Model | ICDAR 2017 | | | | AI Hub Book Cover | | | | AI Hub Sign | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% | 10% | 15% | 25% | 100% |
| ChaLCoR (w/o one-hot label) | 68.1 | 75.3 | 77.6 | **82.4** | 88.1 | 91.6 | 93.9 | 96.3 | 77.5 | 88.1 | 92.7 | **97.3** |
| ChaLCoR | **71.4** | **76.2** | **77.9** | **82.4** | **88.4** | **92.3** | **94.2** | **96.4** | **78.4** | **89.4** | **93.2** | 97.2 |

결과, ChaLCoR의 글자 단위 일관성 정규화만 적용한 실험 결과, ChaLCoR의 글자 정렬 및 글자 단위 일관성 정규화를 모두 포함한 실험결과를 명시하였다. 실험결과들을 비교해 보았을 때, ChaLCoR의 글자 단위 일관성 정규화만 적용했을 때도 선행연구들에 비해 우수한 성능을 보였다. 또한 글자 정렬을 추가함으로써 ChaLCoR의 성능이 글자 단위 일관성 정규화만 적용했을 때보다 개선된 것을 확인할 수 있었다. 이러한 실험결과는 본 연구에서 제안하는 글자 정렬과 글자 단위 일관성 정규화가 장면 이미지 속 문자 인식에 효과적이라는 것을 증명한다. 이는 글자 정렬을 통해 학습에 도움이 되는 이미지를 선별할 뿐만 아니라, 글자 단위 일관성 정규화를 통해 학습에 도움이 되는 글자들을 선별함으로써 발생가능한 노이즈를 최소화한 것으로 사료된다.

두 번째 실험은 ChaLCoR에서 엔트로피 최소화의 효과를 입증하기 위해 수행하였다. 엔트로피 최소화를 사용하지 않았을 때는 엔트로피 최소화 상수 $\tau$를 1로 설정하여 일반적인 소프트맥스 함수 연산을 수행했고, 엔트로피 최소화를 사용했을 때는 1보다 작은 $\tau$값을 통해 일반적인 소프트맥스 함수의 출력값에서 높은 확률 값을 강조하였다. 실험 결과는 <Table 4>와 같으며, 엔트로피 최소화를 했을 때가 그렇지 않았을 때보다 더 우수한 성능을 보여주었다. 이는 높은 확률을 강조함으로써 유사한 글자가 많은 한글에서 발생하는 글자의 모호함을 해소하였다고 볼 수 있다. 따라서 이러한 실험결과는 엔트로피 최소화가 ChaLCoR에서 효과적이라는 것에 대한 타당성을 보여준다.

세 번째 실험은 ChaLCoR에서 글자 정렬 조건과 글자 단위의 임계값 비교 조건을 모두 만족했을 때, 원-핫 레이블로 바꿔주는 것에 대한 효과를 증명하고자 하였다. 원-핫 레이블은 기존과 마찬가지로 교차 엔트로피를 기반으로 학습하였고, 원-핫 레이블이 아닌 경우는 평균 제곱 오차(mean squared error, MSE)를 활용하여 학습하였다. 실험결과는 <Table 5>와 같으며, AI Hub 간판 이미지 데이터셋에서 레이블을 갖는 데이터를 100% 활용하여 학습한 경우를 제외하고, 모든 경우에서 원-핫 레이블이 우수한 성능을 보여주었다. 이를 통해 원-핫 레이블로 변환하여 학습하는 것이 ChaLCoR에서 효과가 있음을 실험을 통해 입증할 수 있었다.

## 5. 결론

본 연구는 딥러닝을 활용한 장면 이미지 속 한글 문자 인식 모델인 ChaLCoR을 제안했다. 한글 특성상 유사한 글자가 많고 학습해야 할 글자수가 많기에, 장면 이미지 속 한글 문자 인식 모델은 영어보다 학습에 어려움이 존재한다. 그러나 본 연구에서 제안하는 ChaLCoR은 글자 자체에 정보가 많은 한글의 특성을 고려한 글자 정렬 및 글자 단위 일관성 정규화 기반의 준지도학습을 활용하여 이를 극복할 수 있었다. 실험 결과, ChaLCoR은 기존 지도학습 및 준지도학습 모델보다 전반적으로 우수한 성능을 보여주었다. 아울러 구성 요소별 성능 기여도 평가를 통해 ChaLCoR에서 제안하는 글자 정렬 및 글자 단위 일관성 정규화가 학습에 도움이 되는 글자들을 이중으로 선별함으로써 장면 이미지 속 한글 문자 인식 모델에 효과가 있다는 것을 증명하였다. 또한 본 연구는 장면 이미지 속 한글 문자 인식 모델에 준지도학습을 최초로 적용했다는 의의를 갖는다. 향후 한글이나 한자와 같이 글자에 정보량이 많은 언어에 특화된 장면 이미지 속 문자 인식 모델을 구축하고자 할 때, ChaLCoR은 데이터가 부족한 상황을 극복할 수 있을 것으로 본다. 이처럼 적은 수의 데이터만으로 학습이 가능한 ChaLCoR은 데이터 레이블링의 소요를 줄일 수 있다는 장점을 갖는다. 이러한 장점들은 국내 자율주행의 노상 표지판 인식이나 간판인식 등 실제 현업에서 유용하게 활용될 수 있다. 현업에서는 특정 도메인 데이터가 부족하여 사전 학습 모델을 사용할 때, 도메인 차이가 생길 수 있다는 한계를 갖는다. 하지만 ChaLCoR은 적은 수의 데이터만으로 학습이 가능하기에 도메인 차이가 없는 모델을 구성할 수 있다. 최근 일관성 정규화 기반의 준지도학습은 동적 임계값(adaptive threshold) 적용을 통해 각 범주 별 학습 정도를 고려하여 발전되고 있다. 따라서, 향후 각 한글의 글자간 학습 정도를 고려하여 학습할 수 있다면 더 우수한 예측모델을 수립할 수 있을 것으로 기대된다.

## 참고문헌


Aberdam, A., Litman, R., Tsiper, S., Anschel, O., Slossberg, R., Mazor, S., Manmatha, R., and Perona, P. (2021), Sequence-to-sequence contrastive learning for text recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15302-15312.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., and Lee, H. (2019), What is wrong with scene text recognition model comparisons? Dataset and model analysis, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4715-4723.

Baek, J., Matsui, Y., and Aizawa, K. (2021), What if we only use real datasets for scene text recognition? Toward scene text recognition with fewer labels, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3113-3122.

Bahdanau, D., Cho, K., and Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

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 (ICCV)*, 2204-2212.

Chen, X., Jin, L., Zhu, Y., Luo, C., and Wang, T. (2021), Text recognition in the wild: A survey, *ACM Computing Surveys (CSUR)*, 54(2), 1-35.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020), Randaugment: Practical automated data augmentation with a reduced search space, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 702-703.

DeVries, T., and Taylor, G. W. (2017), Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552.

Grandvalet, Y. and Bengio, Y. (2004), Semi-supervised learning by entropy minimization, *Advances in Neural Information Processing Systems*, **17**, 529-536.

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*, 369-376.

He, K., Zhang, X., Ren, S., and Sun, J. (2016), Deep residual learning for image recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

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

Lee, D. H. (2013), *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*, ICML Workshop on challenges in Representation Learning.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016), Regularization with stochastic transformations and perturbations for deep semi-supervised learning, *Advances in Neural Information Processing Systems*, **29**, 1163-1171.

Schuster, M., and Paliwal, K. K. (1997), Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing*, **45**(11), 2673-2681.

Scudder, H. (1965), Probability of error of some adaptive pattern-recognition machines, *IEEE Transactions on Information Theory*, **11**(3), 363-371.

Shi, B., Bai, X., and Yao, C. (2016), 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., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2018), Aster: An attentional scene text recognizer with flexible rectification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(9), 2035-2048.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C. L. (2020), Fixmatch: Simplifying semi-supervised learning with consistency and confidence, *Advances in Neural Information Processing Systems*, **33**, 596-608.

Sung, S. H., Lee, K. B., and Park, S. H. (2020), Research on Korea Text Recognition in Images Using Deep Learning, *Journal of the Korea Convergence Society*, **11**(6), 1-6.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014), Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, **27**, 3104-3112.

Tarvainen, A. and Valpola, H. (2017), Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *Advances in Neural Information Processing Systems*, **30**, 1195-1204.

Usmankhujaev, S., Lee, S., and Kwon, J. (2019), Korean license plate recognition system using combined neural network, *International Symposium on Distributed Computing and Artificial Intelligence*, 10-17.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020), Unsupervised data augmentation for consistency training, *Advances in Neural Information Processing Systems*, **33**, 6256-6268.

Zeiler, M. D. (2012), Adadelta: An adaptive learning rate method, arXiv preprint arXiv:1212.5701.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017), mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412.

Zheng, C., Li, H., Rhee, S. M., Han, S., Han, J. J., and Wang, P. (2022), Pushing the Performance Limit of Scene Text Recognizer without Human Annotation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14116-14125.


## 저자소개


**김성수:** 경희대학교 산업경영공학과에서 2022년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석사과정에 재학 중이다. 연구 분야는 Semi-Supervised Learning, Scene Text Recognition이다.

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