# 한국어 손글씨 인식을 위한 TrOCR 모델 미세조정

박시은, 안진현*
명지대학교

plscew@mju.ac.kr, *wlsgus3396@mju.ac.kr

# A Study on Fine-Tuning TrOCR for Korean Handwritten Text Recognition

Sieun Park, Jin-hyun Ahn*

Myongji Univ.

## 요 약

본 연구는 한국어 손글씨 문장 인식 성능 향상을 위해 Transformer 기반 OCR 모델인 TrOCR 을 한국어 손글씨 데이터에 맞추어 파인튜닝하였다. 제안 모델은 Vision Encoder-Decoder 구조를 활용하여 손글씨 이미지로부터 문장 텍스트를 생성하며, 영어 중심 사전학습 모델을 한국어 손글씨 환경에 효과적으로 전이하는 것을 목표로 한다. 이를 위해 학습 안정성과 수렴을 고려한 2 단계 학습 전략을 적용하였다. 1 단계에서는 인코더를 고정한 채 디코더를 중심으로 학습하여 한국어 문자 분포에 적응시켰으며, 2 단계에서는 인코더의 후반부 레이어와 정규화 계층을 해제하여 미세조정을 수행하였다. 성능 평가는 Character Error Rate(CER)를 지표로 수행하였고, 실험 결과 CER 는 1.27 에서 0.14 로 감소하여 한국어 손글씨 문장 인식 성능 향상을 확인하였다.

## Ⅰ. 서론

본 연구는 캡스톤 디자인 프로젝트 IVO 의 기능 중 하나인 한국어 손글씨 인식 모듈을 구현하기 위해 수행되었다. IVO 프로젝트는 스마트워치 기반 제스처 입력과 OCR 을 결합하여 발표·강의 상황에서 슬라이드 제어를 지원하고, 필기/판서 내용을 텍스트로 변환해 화면에 제공하는 발표·강의 보조 시스템이다.[1] 이러한 환경에서 손글씨 인식 성능은 시스템의 실사용성과 사용자 경험에 직접적인 영향을 주므로, 한국어 손글씨 문장 인식 성능 확보가 필요하다.

이를 위해 본 연구에서는 Transformer 기반 OCR 모델인 TrOCR 을 기반 모델로 선택하고, 한국어 손글씨 데이터에 맞추어 파인튜닝하여 문장 단위 인식 성능을 개선하고자 한다.

## Ⅱ. 관련 연구

손글씨 인식은 이미지로부터 텍스트를 복원하는 문제로, 최근에는 입력 이미지를 특징으로 변환한 뒤 텍스트 시퀀스를 생성하는 인코더-디코더 기반 접근이 널리 사용된다. 이 방식은 글자 형태가 불규칙하거나 문장 길이가 긴 환경에서도 문맥 정보를 활용해 안정적으로 문자열을 생성할 수 있다.

최근 손글씨 텍스트 인식 분야에서는 Transformer 기반 인식기가 활발히 연구되고 있다. ViTSTR 은 ViT 인코더를 활용해 전역 문맥 특징을 추출하고 이를 기반으로 단어/문장 수준 텍스트를 예측하는 방식으로 성능을 얻는다.[2] PARSeq 은 permuted autoregressive 학습을 통해 디코딩 순서 의존성을 완화하여 다양한 길이의 텍스트에 대해 강건한 성능을 보인다.[3] ABINet 은 시각 인식 결과를 언어 모델로 반복 보정하여 철자 및 문맥 오류를 줄이는 방향으로 성능을 향상시킨다.[4] 다만 이러한 모델들은 사전학습 데이터 분포와 목표 도메인 간 차이가 클 경우 성능 저하가 발생할 수 있어, 실제 적용 환경에 맞춘 전이학습 및 파인튜닝 전략이 중요하다.

TrOCR 은 Transformer 기반 end-to-end OCR 모델로서 사전학습 모델을 다양한 텍스트 인식 환경에 전이하는 전략을 제시하였다. 본 연구는 TrOCR 의 전이학습 특성을 활용하여 한국어 손글씨 문장 인식 성능을 향상시키는 데 초점을 둔다.[5]

## Ⅲ. 실험

본 연구는 Transformer 기반 OCR 모델인 TrOCR 을 사용하여 한국어 손글씨 문장 이미지를 텍스트로 변환하는 문제를 다룬다. 초기 가중치는 영어 손글씨에 대해 사전 학습된 microsoft/trocr-base-handwritten 체크포인트를 사용하였고, 이를 한국어 손글씨 데이터에 맞추어 단계적으로 파인튜닝하였다.

데이터는 AI Hub 의 '한국어 손글씨 이미지' 데이터셋을 기반으로 구성하였다. 파일명 기반 매칭을 통해 이미지-라벨 쌍을 구성한 뒤 학습/검증 세트로 분리하여 학습 및 평가에 활용하고 입력 이미지는 PIL 을 통해 로드한 뒤 RGB 로 변환하여 사용하였다.

한국어 손글씨 문장 인식 성능을 안정적으로 향상시키기 위해 총 2 단계 파인튜닝 전략을 사용하였다. 이는 영어 중심으로 사전 학습된 TrOCR 의 표현을 무리하게 한

번에 바꾸기보다, 단계적으로 한국어 데이터 분포에 적응시키는 방식이다.

Stage 1 에서는 비전 인코더의 파라미터를 고정하고, 디코더를 중심으로 학습을 수행하였다. 이를 통해 사전학습 모델이 보유한 기본적인 손글씨 인식 표현은 유지하면서, 한국어 문장 생성에 필요한 문자 분포 및 디코딩 패턴을 우선적으로 적응시키는 것을 목표로 하였다.

Stage 2 에서는 Stage 1 체크포인트를 초기값으로 사용하되, 인코더 전체를 학습하지 않고 마지막 2 개 레이어와 LayerNorm 만 선택적으로 언프리징하여 추가 파인튜닝을 수행하였다. 또한 낮은 학습률(=1e-5)을 적용해 과도한 파라미터 변형과 과적합을 방지하고, 한국어 손글씨의 시각적 분포에 부분적으로 적응하도록 유도하였다.

## Ⅳ. 결과

성능 평가는 CER 을 사용하였다. CER 은 예측 문자열과 정답 문자열 사이의 편집거리를 기반으로 계산되며 치환(Substitution), 삭제(Deletion), 삽입(Insertion) 오류의 합을 정답 문장의 문자 수로 정규화한 값으로 정의된다.

$$CER = \frac{S + D + I}{N}$$

평가의 일관성을 위해 문자열은 유니코드 정규화(NFKC) 및 공백 정리를 수행한 후 CER 을 산출하였다. 동일한 검증 설정에서 사전학습 모델과 단계별 파인 튜닝의 성능을 비교한 결과는 다음과 같다.

| Model | Fine-tuning strategy | Generate CER |
|---|---|---|
| Baseline | Pretrained TrOCR | 1.2765 |
| Stage 1 | Encoder frozen | 0.2456 |
| Stage 2 | Partial unfreezing | 0.1407 |

Stage 1 적용만으로도 CER 가 크게 감소하여 한국어 문자 분포 및 문장 생성 방식에 대한 적응 효과를 확인하였다. 이후 Stage 2 에서 인코더의 일부 레이어를 선택적으로 unfreezing 하여 추가 학습을 수행한 결과 CER 가 0.1407 까지 추가로 감소하여, 한국어 손글씨의 시각적 분포에 대한 적응이 성능 향상에 기여했음을 확인하였다.

## Ⅴ. 결론

제안한 방법은 TrOCR 기반 Vision Encoder-Decoder 를 한국어 손글씨 문장 데이터에 맞춰 파인튜닝하여 문장 인식 성능을 개선하였다.

학습 안정성과 도메인 적응을 위해 인코더 고정 기반 Stage 1 과 인코더 부분 언프리징 기반 Stage 2 로 구성된 2 단계 파인튜닝 전략을 적용하였다. 검증 샘플 300 개 기준 generate CER 은 baseline 1.2765 에서 Stage 1 0.2456, Stage 2 0.1407 로 감소하여 한국어 손글씨 문장 인식 성능 향상을 확인하였다.

향후에는 더 큰 규모의 데이터로 추가 검증을 수행하고, 다양한 필기체 및 촬영 조건에 대한 일반화 성능을 평가할 예정이다. 또한 언어모델 결합 기반 후처리(교정)를 적용하여 인식 오류를 줄이고, 영문과 한국어가 혼합된 문장에서도 안정적으로 인식할 수 있도록 성능을 향상시킬 계획이다.


## ACKNOWLEDGMENT

본 과제(결과물)는 2025 년도 교육부 및 경기도의 재원으로 경기 RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.

(2025-RISE-09-A15)

This research was supported by the Regional Innovation System & Education RISE) program through the Gyeonggi RISE Center, funded by the Ministry of Education MOE) and the Gyeonggi-do, Republic of Korea.

(2025-RISE-09-A15)


## 참 고 문 헌


[1] blueion0612, IVO (Smartwatch-Based Presentation & Lecture Assistance System), GitHub repository, https://github.com/blueion0612/IVO

[2] Atienza, R., “Vision Transformer for Fast and Efficient Scene Text Recognition,” arXiv:2105.08582, 2021.

[3] Bautista, D., and Atienza, R., “Scene Text Recognition with Permuted Autoregressive Sequence Models,” arXiv:2207.06966, 2022.

[4] Fang, S., Xie, H., Wang, Y., Mao, Z., and Zhang, Y., “Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition,” arXiv:2103.06495, 2021.

[5] Li, M., et al., “TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models” arXiv:2109.10282, 2021.