스포츠 중계 영상 내 문자 인식 모델의 성능 분석 연구

이채은[1], 김민지[1], 오수민[1], 장현겸[1], 최은서[1], 신윤호[2], 이웅희[1]*

[1]한성대학교, [2]엘지유플러스

[1] {rjh2436, 1971273, ohsoomin, hyungyumjang, irenelove112, whlee}@hansung.ac.kr, [2] yoon731@gmail.com

A Study on Performance Analysis of Text Recognition Models in Sports Broadcast Videos

Chae-Eun Lee[1], Minji Kim[1], Soomin Oh[1], Hyungyum Jang[1], Eunsuh Choi[1], Yoonho Shin[2], Woonghee Lee[1]*

[1]Hansung University, [2]LG Uplus

요 약

본 논문은 스포츠 중계 화면에 적합한 OCR(Optical character recognition) 모델 선정을 목표로 다양한 OCR 모델의 성능을 비교 분석하였다. 네이버 Clova AI가 제안한 STR(Scene Text Recognition) 모델 중 상위 9개의 모델을 한국어 데이터셋을 사용하여 새롭게 학습을 진행하였으며, 한글의 특성에 맞춘 적합한 모델 조합을 찾기 위해 성능 분석을 진행하였다. 그 중, 성능이 뛰어난 상위 3개의 모델을 재선별하여 KBO(Korea Baseball Organization) 프로야구 중계 영상의 스코어보드 문자인식 수행 결과를 비교하였고, 최종적으로 가장 우수한 성능을 가지는 모델 조합을 선정하였다. 이러한 결과를 비교 분석하는 과정을 통해 스포츠 중계 화면과 같은 환경에서 적절한 OCR 모델을 선택할 수 있는 기준을 제시할 수 있을 것으로 기대한다.

Ⅰ. 서 론

OCR(Optical character recognition)은 이미지나 영상 속 문자를 컴퓨터로 인식하는 기술이다. 이러한 OCR 시스템은 크게 문자 영역 검출(Text Detection)과 검출된 문자를 인식하는 Text Recognition의 두 단계로 구성되며 전체적인 OCR 시스템의 프로세스는 그림 1과 같다[1]. 본 연구의 초점은 Text Recognition 단계에서 활용되는 STR(Scene Text Recognition) 모델에 맞춰져 있다. STR 모델은 특히 실제 환경에서 복잡한 배경 속 텍스트를 인식하는 데 적합하게 설계되었으며, 스포츠 중계 화면과 같이 동적이고 복잡한 상황에서의 문자 인식에 특히 유용하다[2]. 본 연구는 다양한 OCR 모델 중에서 스포츠 중계 화면에 적합한 최적의 모델을 선정하기 위해, STR 모델을 중점적으로 분석하고 한국어 데이터셋의 특성에 적합한 옵션을 찾는 데 중점을 둔다. 이러한 분석을 통해, 효과적인 OCR 모델 선정에 대한 중요한 기준을 제시할 것으로 기대한다.

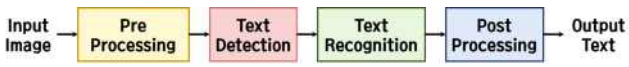

그림 1. OCR 시스템의 전체적인 프로세스 개요

Ⅱ. 실험 방법

본 연구에서는 네이버 Clova AI가 제안한 STR 모델들의 성능을 평가하는 방식으로 진행되었다[3].

STR 모델은 '변환(Transformation)-특징 추출(Feature Extraction)-시퀀스 모델링(Sequence Modeling)-예측(Prediction)'의 4단계로 구성되며, 그림 2와 같이 Text Recognition 단계를 수행하여 문자를 검출한다. 각 단계에서 사용 가능한 옵션들은 표 1과 같으며, 총 24가지 모델 조합이 가능하다. 본 연구에서는 [3]에 언급된 두 가지 기준을 사용하여 최상위 모델 9개를 선정했다. '정확도-학습시간' 기준에서는 학습시간이 짧고 정확도가 높은 모델을, '정확도-매개변수 수' 기준에서는 매개변수가 적으면서 정확도가 높은 모델을 선택했다. 'T모델'은 학습시간 기준 상위 모델로, T1이 가장 효율적이고 T5는 상대적으로 더 긴 학습시간이 필요하다. 'P모델'은 매개변수 수 기준 상위 모델로, P1이 가장 효율적이며 P5는 더 많은 매개변수를 가진다. 각 기준별로 상위 5개 모델을 선택했으며, 표 2와 같이 중복을 포함해 총 9개의 모델이 최종적으로 선정되었다.

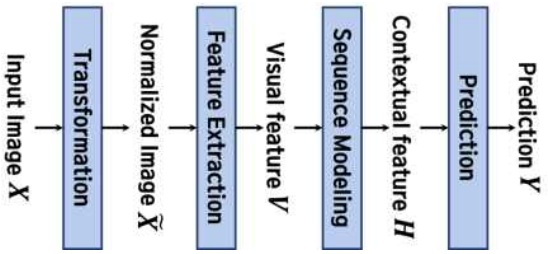

그림 2. STR 모델의 4단계 프로세스

또한 영어 데이터셋인 MJSynth(MJ)와 SynthText(ST)을 활용해 학습된 기존 모델들의 한글 인식 정확도를 높이기 위해 한국어 글자체 데이터셋을 사용하여 재학습하였다[4]. Intel Xeon Gold 5218 CPU, 128GB RAM, NVIDIA Tesla V100 PCIe 32GB GPU 2개를 포함한 기기를 사용하여 Ubuntu 22.12 운영체제에서 모델 학습을 수행하였다. 각 모델에 대해 3개의 Seed 값을 설정하고 평균값을 성능 기준으로 삼았다. 이후 테스트 데이터셋에서 성능이 가장 뛰어난 상위 3개 모델을 선정하여 2023 KBO(한국야구위원회) 프로야구 하이라이트 영상 중 선별된 장면에서 스코어보드의 문자인식을 수행하고 그 결과를 비교하였다.

Stage	Options
Transformation	None | TPS(Thin-Plate Spline)
Feature Extraction	VGG | RCNN | ResNet
Sequence Modeling	None | BiLSTM
Prediction	CTC Loss | Attention

표 1. STR 모델의 각 단계별 사용 가능한 옵션들

별칭	모델 구성
T1	None-VGG-None-CTC
T2	None-Resnet-None-CTC
T3	None-Resnet-BiLSTM-CTC
T4	TPS-ResNet-BiLSTM-CTC
T5/P5	TPS-ResNet-BiLSTM-Attn
P1	None-RCNN-None-CTC
P2	None-RCNN-None-Attn
P3	TPS-RCNN-None-Attn
P4	TPS-RCNN-BiLSTM-Attn

표 2. 9개의 STR 모델의 별칭 및 구성 요소

Ⅲ. 실험 결과

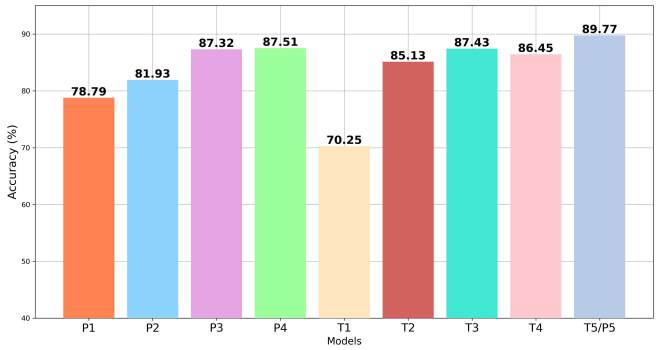

그림 3. 한국어 데이터셋을 학습한 모델들의 성능 비교

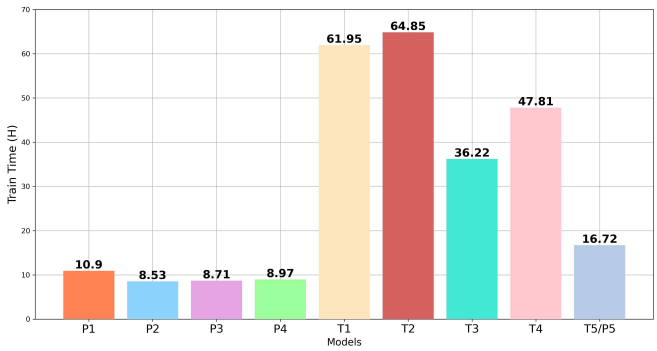

그림 4. 각 모델별 한국어 데이터셋 학습 소요 시간

한국어 데이터셋을 학습한 모델들의 성능을 비교한 결과는 그림 3에 나타나있다. T5/P5 모델이 가장 뛰어난 성능을 보였으며, P4와 T3모델도 우수한 성능을 나타냈다. 그림 4는 모델 학습에 소요된 시간을 비교한다. RCNN을 사용한 모델들은 ResNet이나 VGG를 사용한 모델들에 비해 학습 시간이 짧으나, 성능과 학습 시간의 관계 분석 결과 ResNet을 사용한 모델이 뛰어난 성능을 보이는 대신 학습 속도가 상대적으로 느린 것으로 확인되었다. 이는 성능과 학습 속도 사이에 trade-off 관계가 존재함을 의미한다. 추가적으로 4단계 중 각 단계에서 가능한 선택지에 따른 성능을 비교하였다.

1) 변환(Transformation)

A(모델 T5/P5와 T3), B(모델 P2와 P3), C(모델 T3과 T4)의 경우를 통해 TPS의 적용 여부에 대한 성능을 비교할 수 있다. A와 C 경우에 TPS를 적용하지 않았을 때 성능이 각각 2.34%와 0.98% 더 우수했으며, B 경우에는 TPS 적용 시 성능이 5.4% 향상되었다. 이러한 결과는 변환 단계의 모듈들이 다른 단계와의 상호작용에 영향을 받는다는 것을 보여준다. 특히, TPS로 변형된 이미지는 시퀀스 모델링 단계에서 시퀀스의 특성을 변경할 수 있어, TPS 적용 시 OCR 프로세스 내의 다른 모듈들과의 상호작용을 면밀히 고려해야 한다.

2) 특징 추출(Feature Extraction)

A(모델 T5/P5와 P4)와 B(모델 T2와 P1)의 경우를 통해 RCNN과 ResNet을 사용했을 때의 성능을 비교하면, A와 B 모두 ResNet을 사용할 때의 성능이 각각 2.25%와 6.33% 더 우수했다. 이는 ResNet의 잔차 연결(Residual Connections)이 OCR에서 다양한 텍스트 변형과 스타일을 더 잘 포착하는 데 도움이 될 수 있음을 시사한다.

3) 시퀀스 모델링(Sequence Modeling)

A(모델 P3과 P4)와 B(T2와 T3)의 경우를 통해 BiLSTM의 적용 여부에 대한 성능을 비교하면, A와 B 모두 BiLSTM을 사용했을 때 0.19%와 2.3% 더 향상된 성능을 보인다. 이는 BiLSTM의 양방향 처리 기능이 한국어의 초성 및 중성과 같은 음절 구조에 효과적으로 작용하며, 다양한 폰트에서 문자 형태가 변할 때 전체 단어를 이해하는 데 효율적임을 보여준다.

4) 예측(Prediction)

A(모델 T5/P5와 T4)와 B(P1과 P2)의 경우를 통해 Attention과 CTC Loss 중 어떤 메커니즘이 더 적합한지 평가했다. A와 B에서 각각 Attention 메커니즘이 3.31%와 3.13% 더 우수한 성능을 나타냈다. 이는 한국어가 조합성이 높은 언어이기 때문에, 데이터의 중요한 부분에 더 큰 가중치를 두어 복잡한 패턴이나 연결을 더 잘 해석하는 Attention이 더 유용하다고 볼 수 있다.

추가로, 2023 KBO 프로야구 하이라이트 영상 중 선별된 장면 내 스코어보드의 문자열 데이터 23개 중 T5/P5 모델은 3개를 맞추지 못했으며, P4 모델은 6개, T3 모델은 7개를 맞추지 못하였다. 최종적으로 T5/P5 모델이 스포츠 중계영상의 문자인식에 가장 적합한 모델임을 확인하였다.

Ⅳ. 결론

본 연구는 스포츠 중계 화면 내 문자 인식에 적합한 OCR 모델을 선정하기 위해 다양한 모델들을 비교 분석하였다. 4단계로 구성된 STR 모델들을 중점적으로 검토한 결과, 한국어 데이터셋을 학습한 모델 중 T5/P5 모델이 가장 우수한 성능을 보였다. T5/P5 모델의 특징 추출 단계에서 ResNet의 사용, 시퀀스 모델링 단계에서의 BiLSTM의 사용, 예측 단계에서 Attention 메커니즘의 사용은 한국어 인식에 효과적임을 입증했다.

향후 TPS 적용 유무에 대한 추가 분석을 수행할 예정이며, 한국어에 특화된 모듈들을 결합하여 새로운 문자 인식 모델의 설계 가능성을 탐색할 것이다. 이러한 연구 결과는 스포츠 중계와 같이 동적이고 복잡한 환경에서 효과적인 OCR 모델을 선택하는 데 중요한 기준을 제공할 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 한성대학교 학술연구비 지원과제임.

참 고 문 헌

[1] Ye, Qixiang, and David Doermann. "Text detection and recognition in imagery: A survey." IEEE transactions on pattern analysis and machine intelligence 37.7 (2014): 1480-1500.

[2]Shi, Baoguang, Xiang Bai, and Cong Yao. "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 (2016): 2298-2304.

[3] Baek, Jeonghun, et al. "What is wrong with scene text recognition model comparisons? dataset and model analysis."

[4] Korean Character Font Images Dataset, AI Hub, accessed on 2019.