영상내 문자 인식 기술 동향

Research Trends for Scene Text Recognition

JIN DONG(kimdoubled@gmail.com)　　컴퓨터공학과 박사과정

Scene text recognition (STR) enables computers to read text in natural scenes such as sign boards, road signs and object labels. Reading text in natural scenes has been an important task in industrial application. It's helps machines perform informed decisions such as which direction to go, what object to pick and what is the next step of action. Traditional text recognition uses a manually designed feature extractor to extract features from a given image, and the machine learning algorithm is then used to classify the extracted features into one of the pre-defined categories, where each category represents a character or word. In recent years, with the development of deep learning, numerous methods have shown promising performance in scene text recognition. In this article, we provide a brief descriptive summary of several recent deep-learning methods for scene text recognition. As part of the improvement in character recognition performance can be attributed to the large number of character recognition datasets in natural scenes, this article also introduces the dataset commonly used in the research. We also compare the performance of these method and present a research guide of the object detection field.

Ⅰ. 서론

문자의 역사는 수천 년 전으로 거슬러 올라갈 수 있다. 문자로 전달되는 풍부하고 정확한 의미 정보는 다양한 비전 기반 애플리케이션에서 중요한 역할을 한다. Natural scenes에서 텍스트를 읽는 작업은 일반적으로 text detection과 text recognition 두 단계를 필요로 한다. Text detection은 텍스트가 존재하는 영역의 bounding box의 위치를 결정한다. 텍스트가 존재하는 영역이 알려지면 text recognition 작업은 해당 영역의 text의 symbol를 인식해 낸다. 하나의 모델로 두 가지 작업을 한 번에 수행할 수 있으면 좋겠지만 두 작업이 동시에 가능한 end-to-end 모델은 아직 연구의 초기 단계라 할 수 있어 만족스러운 성능을 보여주지 못하고 있다. 본 동향 분석에서는 두 task 중에 text recognition 부분에 초점을 두고 있다.

Scene Text Recognition(STR)은 컴퓨터비전 및 패턴 인식 분야의 중요한 연구 주제로 활발한 연구가 진행되고 있다. Scene Text Recognition은 광학 문자 인식(Optical Character Recognition, OCR)에서 카메라 등을

통해 실제 환경에서 촬영한 scene text image를 대상으로 문자 인식을 진행하는 task이다. OCR는 일반적으로 정형화된 문서에 인쇄되어 있는 문자를 읽어내는 작업을 의미한다. 반면 STR은 일상적인, 보통은 카메라 센터로 촬영한 이미지 내의 문자를 읽어내는 것을 목표로 하고 있다. 그렇기 때문에 STR의 대상이 되는 이미지의 배경이 훨씬 복잡하고 글꼴이나 배열이 다양해 OCR보다 어려운 문제라고 할 수 있다. 하지만 이 둘을 명확히 구분하지 않고 통틀어 OCR라고 부리기도 한다. 앞서 언급했듯이 STR의 목적은 간판, 제품의 라벨, 도로 표지판과 같은 다양한 비정형 환경에서 나타나는 문자를 인식하는 것이다. 그렇기 때문에 입력된 문자는 글꼴, 방향, 모양, 크기, 색상, 질감, 조명 등이 모두 다양한 형태로 되어있다. STR의 성능은 촬영할 때 카메라 센서의 방향, 위치 및 이미지 흐림, 노이즈, 기하학적 왜곡에 영향을 받는다. 그 뿐만 아니라 눈부심, 그림자, 비, 눈 및 서리와 같은 날씨 상태도 STR의 성능에 영향을 줄 수 있다. 이러한 이유 때문에 STR는 풀기 어려운 문제로 알려져 있다.

이를 해결하기 위해 연구자들은 기계학습 및 딥러닝 등 기술을 이용해 STR 분야의 발전을 도모해 왔다. 이번 동향 분석에서는 먼저 기술 발전에 중요한 역할을 하고 있는 데이터셋에 대해 간단히 소개하고 기계학습을 이용한 STR 연구를 간단하게 소개한다. 그리고 최근 딥러닝 기술을 이용한 STR 연구들에 대해 소개한다.

Ⅱ. 데이터셋

2.1 Synthetic dataset

STR에 사용되는 데이터셋은 크게 synthetic datasets와 real-world datasets으로 나눌 수 있다. STR 모델 훈련에 필요한 scene text image를 labeling하는 작업은 많은 인적, 시간적 비용이 필요하기 때문에 잘 labeling 된 큰 사이즈의 scene text image datasets를 확보하기는 쉽지 않다. 그래서 기존 연구에서는 real data를 사용하는 대신 코드를 사용해 합성한 synthetic data를 사용하고 있다. STR 모델 훈련에 사용되는 synthetic datasets로는 MJSynth[1]와 SynthText[2]가 있다.

MJSynth(MJ)[1] dataset은 STR 모델의 훈련을 위해 제안된 synthetic dataset으로 총 890만 장의 합성된 text images로 구성되어 있다. MJSynth 이미지는 아래와 같은 과정을 통해 생성된다. 1) 글꼴 렌더링, 2) 테두리 및 그림자 렌더링, 3) 배경 색칠하기, 4) 글꼴, 테두리, 배경을 결합하기, 5) 이미지에 대해 labeling 진행, 6) 노이즈 추가. Figure 1a에서는 MJSynth dataset의 예시를 보여주고 있다.

SynthText(ST)[2] dataset는 합성을 통해 생성한 데이터셋으로 처음에는 text detection을 위해 만들어졌다. 그렇기 때문에 SynthText dataset의 이미지는 Figure 1b와 같이 한 장의 그림에 text가 놓여 있고 그 주위에 box가 쳐져 있는 형태로 되어있다. Text detection을 위해 생성된 SynthText dataset은 box 부분을 cropping해 STR 모델의 훈련에 사용할 수 있다. SynthText dataset에서 약 550만 장의 cropping 된 text image를 얻을 수 있다.

기존 연구에서는 STR 모델을 훈련할 때 다양한 방법으로 앞서 소개한 두 개의 synthetic dataset을 활용했다. 예를 들어 연구 [3-5]에서는 MJ dataset만 사용해 훈련을 진행하고 연구 [6]에서는 MJ dataset과 ST dataset을 함께 사용해 STR 모델을 훈련했으며 연구 [7]에서는 이들을 각각 50%씩만 활용했다. MJ 데이터셋과 ST 데이터셋의 예시는 그림 1과

같다.

(a) MJSynth word boxes (b) SynthText scene image

(그림 1) (a) MJ dataset (b) ST dataset

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

2.2 Real-world datasets

실제 환경에서 수집하고 사람이 수작업으로 labeling한 real-world datasets는 evaluation 단계에서 사용된다. 기존 연구에는 다양한 real-world datasets이 STR model의 성능 검증을 위해 사용되었는데 이들은 regular datasets와 irregular datasets로 나눌 수 있다.

Regular datasets는 균일한 간격에 가로로 배치된 규칙적인 문자가 포함되어 있는 text images로 구성됐는데 대표적으로 아래와 같은 데이터셋이 있다.

IIIT5K-Words(IIIT)[8]: Google image search를 통해 수집한 데이터셋으로 "billboards", "signboard", "house numbers"등 단어로 검색했을 때 나타나는 이미지로 구성된 데이터셋이다. 해당 데이터셋은 2000장의 훈련용 이미지와 3000장의 테스트용 이미지고 구성되었다.

Street View Text(SVT)[9]: Google Street View를 통해 수집한 야외 거리 이미지로 노이즈가 있거나 해상도가 낮은 이미지들을 포함하고 있다. SVT는 257장의 훈련 이미지와 647장의 테스트 이미지로 구성되었다.

ICDAR2003(IC03)[10]: 카메라를 통해 촬영한 scene text image를 인식하기 위한 ICDAR 2003 Robust Reading competition을 위해 공개된 데이터셋이다. 1156장의 훈련용 이미지와 1110장의 테스트용 이미지가 포함되어 있다.

ICDAR2013(IC13)[11]: ICDAR 2013 Robust Reading competition을 위해 제안된 데이터셋으로 IC03 dataset의 일부 이미지를 차용하고 있다. IC13 데이터셋은 848장의 훈련용 이미지와 1095장의 테스트용 이미지로 구성되어 있다.

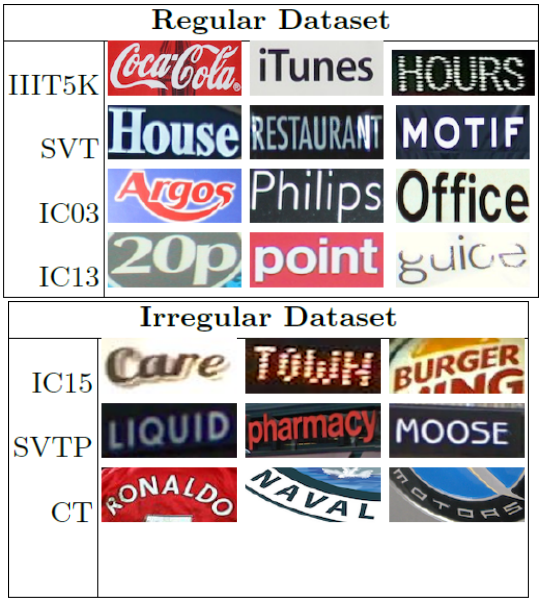
Irregular datasets에는 일반적으로 곡선 및 회전 등으로 인해 왜곡된 텍스트가 포함되어 STR 모델이 인식하기 어려운 이미지들로 구성된 이미지가 포함되어 있다. 대표적인 irregular datasets는 아래와 같은 것들이 있다.

ICDAR2015(IC15)[12]: ICDAR 2015 Robust Reading competitions을 위해 공개된 데이터셋으로 4468장의 훈련용 이미지와 2077장의 테스트용 이미지로 구성된 데이터셋이다. IC15 데이터셋은 Google Glasses를 통해 수집한 이미지로 구성되었는데 사람의 움직임의 영향을 받아 노이즈가 있거나, 왜곡되거나 해상도가 낮은 이미지가 포함되어 있다.

SVT Perspective(SP)[13]: Google Street View를 통해 수집된 645장의 테스트용 이미지로 구성된 데이터셋이다. 그중 일부 이미지는 정면에서 촬영한 것이 아닌 관계로 perspective projection을 가지고 있다.

CUTE80(CT)[14]: Natural scenes에서 수집한 288장의 테스트용 이미지로 구성된 데이터셋이다. 데이터셋 중 대부분은 이미지 내의 텍스트는 왜곡된 형태를 가지고 있다.

그림 2에서는 regular 및 irregular datasets의 sample를 보여주고 있다.



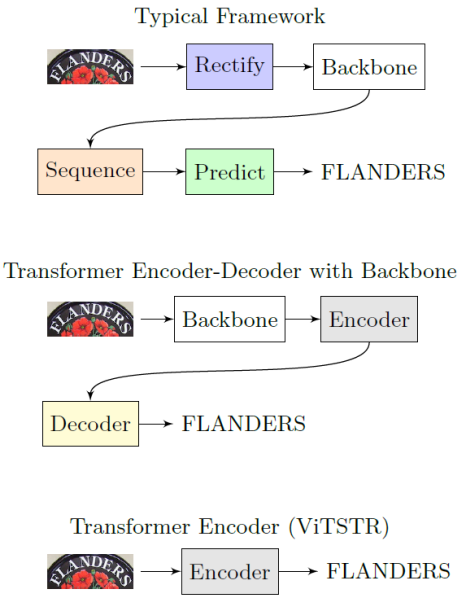
(그림 2) Regular and Irregular dataset

[출처] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. *arXiv preprint arXiv:2105.08582*.

Ⅲ. STR 모델

STR 모델은 이미지에 포함되어 있는 문자를 올바른 순서로 식별하는 작업을 한다. 일반적으로 예측하려는 객체의 범주가 몇 개뿐인 object recognition과 달리 STR가 예측하는 이미지에는 여러 개의 다른 문자가 포함될 수 있다. 그렇기 때문에 STR 모델은 알파벳 및 숫자를 포함한 다양한 문자를 식별할 수 있어야 한다. 그래서 STR 문제는 object recognition 보다 어려운 문제라고 할 수 있다. 연구 초기에는 STR 문제를 해결하기 위해 연구[15] 및 연구[16]와 같이 hand-crafted features와 기계학습 방법을 사용했다. 하지만 이러한 방법은 만족할 만한 성능을 달성하기 못했다. 최근에는 딥러닝 기반 방법이 STR 분야를 극적으로 발전시켰다.

기존 연구에서 제안한 STR 모델은 모두 그림 3에서 보여주고 있는 3가지 종류의 framework 중 하나로 분류될 수 있다. 그중 대부분 모델의 구조는 그림 3의 Typical Framework로 분류될 수 있다. Typical Framework는 Rectification - Feature Extraction(Backbone) - Sequence Modelling - Prediction stage로 구성된 framework이다.



(그림 3) STR 모델 Framework

[출처] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. *arXiv preprint arXiv:2105.08582*.

Rectification stage는 transformation stage라고도 하는데 입력된 이미지의 왜곡을 제거해 입력 이미지 내의 문자가 수평으로 배열되도록 정규화를 진행하는 역할을 한다. 이는 Feature Extraction(Backbone) 모듈이 안정적으로 invariant features를 추출할 수 있도록 한다. 이때 사용되는 기술로는 Thin-Plate-Spline(TPS)[17], Spatial Transformer Network(STN)[18] 등이 있다.

Feature extraction stage에서는 입력된 이미지에서 글자 정보와 관련된 feature를 추출하는 역할을 한다. 기존 연구에서는 VGG[19], ResNet[20], RCNN[5] 등 모델을 특징 추출기로 사용했다.

Sequence modeling stage에서는 문자열에 포함된 contextual information을 파악한다. 이

는 입력 feature를 통해 문자열을 예측하는 prediction stage에서 더 robust한 결과를 출력할 수 있도록 도와주는 역할을 한다. Sequence modeling stage에서는 contextual information을 모델링하기 위해 일반적으로 BiLSTM을 사용한다.

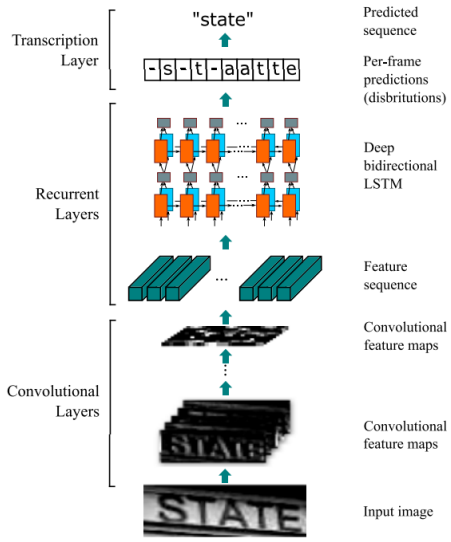
Prediction stage에서는 앞 단계에서 추출한 features를 사용해 이미지에 있는 문자 시퀀스를 예측한다. 기존 연구에서는 Connectionist temporal classification(CTC)[21] 혹은 attention-based sequence prediction(Attn [4][22] 방법을 사용하고 있다. 기존 제안된 대부분 STR 방법은 모두 Rectification-Feature Extraction(Backbone)-Sequence Modelling-Prediction framework에 적용할 수 있다. 다만 특정 stage를 생략했거나 stage 내에서 서로 다른 방법론을 적용했다는 차이점이다. 아래에서는 이와 같은 프레임워크의 STR 모델을 제안한 문자 인식을 진행한 모델을 소개한다.

3.1 CRNN

CRNN[3]은 rectification stage를 거치지 않고 feature extraction stage에서 VGG 모델을 사용하고 sequence modeling stage에서 BiLSTM을 사용하고 prediction 단계에서 CTC를 사용한 모델이다. CRNN 모델은 최초로 딥러닝을 이용해 STR 문제에 대해 연구를 진행한 end-to-end 모델이며 그 구조는 그림 와 같다.

3.2 R2AM

R2AM[5]은 오직 feature extraction stage와 prediction stage로 구성된 모델이다. Feature extraction stage에서는 RCNN을 사용했으며 prediction stage에서는 attention 기반의 방법을 사용했다. 해당 연구에서는 recursive CNNs를 이용해 기존보다 우수한 STR 성능을 기록했다.



(그림 4) CRNN 모델 구조

[출처] Shi, B., Bai, X., & 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.

3.3 GRCNN

GRCNN[23]은 R2AM에서 사용하는 recurrent convolution neural network 대신 새롭게 제안한 gated recurrent convolution neural network를 사용해 기존 recurrent convolutional neural network를 사용한 R2AM 모델보다 IIIT-5K, SVT 등 데이터셋에서 우수한 성능을 기록했다. GRCNN은 rectification stage를 거치지 않았으며 sequence modeling stage에서는 BiLSTM을, prediction stage에서는 attention-based 방법을 사용했다.

3.4 Rosetta

Rosetta[24]는 STR의 속도 계선을 위해 rectification state와 sequence modeling stage를 제외하고 feature extraction stage와 prediction stage로만 구성된 STR 모델이다. feature extraction stage와 prediction stage는

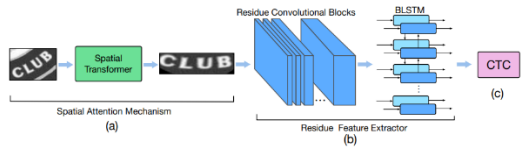
각각 ResNet과 CTC를 사용했다.

3.5 RARE

RARE[4]는 rectification stage에서 TPS 모델을 이용해 이미지 내 문자에 대해 rectification을 진행했다. Feature extraction stage에서는 VGG를 사용하고 있으며 BiLSTM을 이용해 sequence modeling을 진행하였다. Prediction stage에서는 attention-based 방법을 사용했다. RARE는 처음으로 rectification stage를 end-to-end STR 모델에 도입한 모델로 기존 모델보다 좋은 성능을 기록했으며 그 후로 나온 모델들은 rectification stage를 적극적으로 활용하기 시작했다.

3.6 STAR-Net

SpaTial Attention Residue Network(STAR-Net)[25]은 rectification, feature extraction, sequence modeling, prediction stage에서 각각 TPS 모델, ResNet, BiLSTM, CTC를 사용했으며 그 구조는 그림 x와 같다. STAR-Net은 spatial attention mechanism을 제안해 text image의 distortions을 제거했다. 이를 통해 기존 모델보다 우수한 성능을 기록할 수 있었다.



(그림 5) STAR-Net 구조

[출처] Liu, W., Chen, C., Wong, K. Y. K., Su, Z., & Han, J. (2016, September). STAR-Net: a spatial attention residue network for scene text recognition. In *BMVC* (Vol. 2, p. 7).

3.7 TRBA

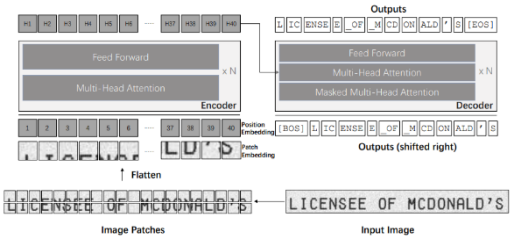
TRBA 모델은 연구[6]에서 제안한 모델로 각 stage에서 사용할 수 있는 방법들을 조합하여 grid search하는 실험을 통해 찾아낸 성능이 제일 좋은 모델이다. TRBA 모델은 rectification, feature extraction, sequence modeling, prediction stage에서 각각 TPS, ResNet, BiLSTM, Attention based 방법을 사용한 모델이다. 해당 연구에서는 각 stage에 사용되는 방법을 어떻게 조합했을 때 성능이 가장 높은지를 찾아냈을 뿐만 아니라 각 stage가 모델 성능에 주는 영향도 분석하였다.

IV. 기타 구조의 STR 모델

자연어 처리(natural language processing, NLP)에 널리 사용되고 있는 Transformers는 parallel self-attention 및 prediction을 통해 sequence modeling과 prediction을 가능하게 했다. 이는 빠르고 효율적은 모델의 기초가 되었다. 현제까지 제안된 transformer를 이용한 STR 모델은 그림 3의 transformer encoder-decoder with backbone 구조와 transformer encoder 구조 두 가지가 있다.

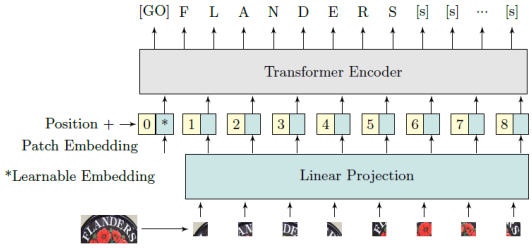
4.1 TrOCR

TrOCR[26] 모델은 transformer encoder-decoder 구조를 가진 STR 모델로 pre-trained image transformer 및 text transformer 모델로 구성된 모델이다. TrOCR는 그림 6과 같이 사전 훈련된 image transformer를 encoder로 사용하고 사전 훈련된 text transformer를 decoder로 사용해 STR 모델을 구성하고 있다.



(그림 6) TrOCR 모델 구조

[출처] Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., ... & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *arXiv preprint arXiv:2109.10282*.



(그림 7) ViTSTR 모델 구조

[출처] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. *arXiv preprint arXiv:2105.08582*.

4.2 ViTSTR

ViTSTR[7]는 그림 3의 transformer encoder 구조와 같이 기존 transformer를 이용한 프레임워크에서 backbone과 transformer decoder를 제거하고 오직 transformer encoder로 구성된 simple한 STR 모델이다. ViTSTR 모델의 구조는 그림 7과 같은데 해당 모델은 먼저 이미지를 여러 개의 patches로 나눈다. 나누어진 2D patches는 Linear Projection을 통해 1D vector로 된 Patch Embedding으로 변환한다. Patch embedding은 positional embedding과 합한 후 transformer encoder에 입력된다. Transformer encoder는 입력에 근거해 예측된 character sequence를 출력한다.

V. STR 모델 성능 비교

앞서 언급한 모델의 각 benchmark dataset에 대한 성능은 표 x와 같다. 표를 통해 STR 모델은 초기 딥러닝을 STR task에 적용한 CRNN부터 점점 성능이 향상되고 있다는 것을 알 수 있으며 그중 TRBA 모델과 ViTSTR 모델의 성능이 각 데이터셋에서 SOTA를 차지하고 있다는 것을 알 수 있다. 그리고 그림 x을 통해 STR 모델의 성능과 파라미터 사이즈 및 inference 속도 사이의 관계를 파악할 수 있다. 그림을 통해 ViTSTR-Small 모델이 성능과 파라미터 사이즈, 그리고 inference 속도 면에서 균형을 이룬 모델이라는 것을 알 수 있다

Table 1 모델 성능 비교

Model	IIIT 3000	SVT 647	IC03 860	IC03 867	IC13 857	IC13 1015	IC15 1811	IC15 2077	SVTP 645	CT 288	Acc %	Std
CRNN [30]	81.8	80.1	91.7	91.5	89.4	88.4	65.3	60.4	65.9	61.5	76.7	0.3
R2AM [17]	83.1	80.9	91.6	91.2	90.1	88.1	68.5	63.3	70.4	64.6	78.4	0.9
GCRNN [36]	82.9	81.1	92.7	92.3	90.0	88.4	68.1	62.9	68.5	65.5	78.3	0.1
Rosetta [4]	82.5	82.8	92.6	91.8	90.3	88.7	68.1	62.9	70.3	65.5	78.4	0.4
RARE [31]	86.0	85.4	93.5	93.4	92.3	91.0	73.9	68.3	75.4	71.0	82.1	0.3
STAR-Net [21]	85.2	84.7	93.4	93.0	91.2	90.5	74.5	68.7	74.7	69.2	81.8	0.1
TRBA [1]	87.8	87.6	94.5	94.2	**93.4**	92.1	77.4	71.7	78.1	75.2	84.3	0.1
ViTSTR-Tiny	83.7	83.2	92.8	92.5	90.8	89.3	72.0	66.4	74.5	65.0	80.3	0.2
ViTSTR-Tiny+Aug	85.1	85.0	93.4	93.2	90.9	89.7	74.7	68.9	78.3	74.2	82.1	0.1
ViTSTR-Small	85.6	85.3	93.9	93.6	91.7	90.6	75.3	69.5	78.1	71.3	82.6	0.3
ViTSTR-Small+Aug	86.6	87.3	94.2	94.2	92.1	91.2	77.9	71.7	81.4	77.9	84.2	0.1
ViTSTR-Base	86.9	87.2	93.8	93.4	92.1	91.3	76.8	71.1	80.0	74.7	83.7	0.1
ViTSTR-Base+Aug	**88.4**	**87.7**	**94.7**	**94.3**	93.2	**92.4**	**78.5**	**72.6**	**81.8**	**81.3**	**85.2**	0.1

[출처] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. *arXiv preprint arXiv:2105.08582*.

VI. 결론

딥러닝 기술이 STR task에 적용된 후부터 STR 성능은 장족의 발전을 가져왔다. 그동안 rectification-feature extraction-sequence modeling-prediction framework을 메인으로 각 stage에서 다양한 기법과 모델들이 개발되었다. 하지만 STR에는 여전히 해결해야 할 문제가 존재한다. STR 모델은 글꼴이 많이 왜곡되거나 강한 빛이 있는 이미지에 대한 인식에 어려움을 겪고 있다. 그리고 vertical text에 대한 인식 역시 해결해야 할 과제로 남아 있다. 다양한 환경의 글자들을 robust하게 인식할 수 있는 STR 모델에 대한 지속적인 연구가 진행되어야 한다.

References

[1] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227.

[2] Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2315-2324).

[3] Shi, B., Bai, X., & 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.

[4] Shi, B., Wang, X., Lyu, P., Yao, C., & Bai, X. (2016). Robust scene text recognition with automatic rectification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4168-4176).

[5] Lee, C. Y., & Osindero, S. (2016). Recursive recurrent nets with attention modeling for ocr in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2231-2239).

[6] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., ... & Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4715-4723).

[7] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. arXiv preprint arXiv:2105.08582.

[8] Mishra, A., Alahari, K., & Jawahar, C. V. (2012, September). Scene text recognition using higher order language priors. In BMVC-British Machine Vision Conference. BMVA.

[9] Wang, K., Babenko, B., & Belongie, S. (2011, November). End-to-end scene text recognition. In 2011 International Conference on Computer Vision (pp. 1457-1464). IEEE.

[10] Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., ... & Lin, X. (2005). ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR), 7(2-3), 105-122.

[11] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., ... & De Las

Heras, L. P. (2013, August). ICDAR 2013 robust reading competition. In 2013 12th International Conference on Document Analysis and Recognition (pp. 1484-1493). IEEE.

[12] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., ... & Valveny, E. (2015, August). ICDAR 2015 competition on robust reading. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 1156-1160). IEEE.

[13] Phan, T. Q., Shivakumara, P., Tian, S., & Tan, C. L. (2013). Recognizing text with perspective distortion in natural scenes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 569-576).

[14] Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. Expert Systems with Applications, 41(18), 8027-8048.

[15] Neumann, L., & Matas, J. (2012, June). Real-time scene text localization and recognition. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3538-3545). IEEE.

[16] Yao, C., Bai, X., & Liu, W. (2014). A unified framework for multioriented text detection and recognition. IEEE Transactions on Image Processing, 23(11), 4737-4749.

[17] Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence, 11(6), 567-585.

[18] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. Advances in neural information processing systems, 28, 2017-2025.

[19] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[21] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

[22] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017). Focusing attention: Towards accurate text recognition in natural images. In Proceedings of the IEEE international conference on computer vision (pp. 5076-5084).

[23] Wang, J., & Hu, X. (2017, December). Gated recurrent convolution neural network for ocr. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 334-343).

[24] Borisyuk, F., Gordo, A., & Sivakumar, V. (2018, July). Rosetta: Large scale system for text detection and recognition in images. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge

Discovery & Data Mining (pp. 71-79).

[25] Liu, W., Chen, C., Wong, K. Y. K., Su, Z., & Han, J. (2016, September). STAR-Net: a spatial attention residue network for scene text recognition. In BMVC (Vol. 2, p. 7).

[26] Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., ... & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv preprint arXiv:2109.10282.