Test Time Augmentation을 이용한 문자 인식의 성능 향상

양희린, 소정민
서강대학교
hryang17@u.sogang.ac.kr, jso1@sogang.ac.kr

Improving Accuracy of Scene Text Recognition using Test Time Augmentation

Heerin Yang, Jungmin So
Sogang Univ.

요 약

최근 STR(Scene Text Recognition)에 관한 다양한 모델이 등장하였으며, 많은 논문에서 특정 문제 해결을 위해 모델을 추가하여 성능을 향상시킨다. 본 논문에서는 STR에서 높은 성능을 보이는 모델을 바탕으로 예측을 잘못하는 텍스트 이미지를 분석하고 이 중에서 잘린 문자를 가지고 있는 텍스트 이미지를 고려하여 Test Time Augmentation을 이용하여 STR의 성능을 향상시킬 수 있음을 보였다. 또한 최빈값, 신뢰도 등 다양한 투표 방식을 이용하여 결과에 어떤 영향을 미치는지 분석하였다.

Ⅰ. 서 론

실생활에서 한 장면의 텍스트를 읽는 STR은 산업 분야에서 중요한 기술이 되어 왔다. 광학 문자 인식(OCR, Optical Character Recognition) 시스템은 배경이 깨끗한 이미지에 대해 성공적으로 적용되고 있지만, 대부분의 기존 OCR은 다양한 텍스트의 배경과 캡처하는 순간의 불완전한 조건으로 인해 STR에서 효과적이지 못하였다. 이후 많은 논문에서 이러한 문제를 해결하기 위해 각 단계에서 특정 문제를 해결하는 Deep Neural Network(DNN)인 multi-stage pipelines를 제안하였다.[1]

다양한 모델의 공정한 비교를 위해 [1]은 STR의 데이터 세트와 framework를 정리하였다. 많은 논문에서 서로 다른 데이터 세트를 사용하여 모델을 비교하기가 쉽지 않은데, 이를 해결하기 위해 STR에서 자주 사용하는 데이터 세트에 대해 모델을 실험할 수 있도록 하였다. 또한, 통합된 four-stage STR framework를 소개하고 모든 조합을 동일한 조건 아래 정확도, 속도, 메모리에 대해 쉽게 실험할 수 있도록 하였다.

통합된 four-stage STR framework는 기능에 따라 Transformation, Feature Extraction, Sequence Modeling, Prediction 총 네 단계로 이루어진다. 먼저, Transformation은 텍스트 이미지를 정규화 하는 단계이다. RARE[2]에서는 TPS(Thin-Plate Spline)를 이용하여 왜곡된 이미지를 일반적인 형태의 이미지로 조정하여 인식 성능을 향상시켰다. Feature Extraction 단계에서는 VGG, RCNN, ResNet 등을 이용하여 텍스트 이미지의 특징을 추출한다. 이 중에서 ResNet은 가장 높은 정확도를 보이지만, 대신 많은 메모리가 필요하다. Sequence Modeling 단계는 양방향을 고려하는 LSTM인 BiLSTM을 사용한다. CRNN[3], RARE[2], FAN[4]에서는 BiLSTM을 사용하여 더 나은 sequence를 만들고자 한 반면에, Rosetta[5]에서는 계산 복잡도와 메모리 소비를 줄이기 위해서 이 단계를 제외하였다. 마지막으로 Prediction은 추출한 특징을 바탕으로 텍스트를 예측하는 단계로, CTC[3]와 Attention[4]을 사용한다. CTC(Connectionist Temporal Classification)는 추출한 특징을 연속적으로 읽으며 문자를 예측하는 모델이다. Attention은 현재 STR에서 가장 높은 정확도를 보이는 모델로서, 문자 각각에 주의를 주어 예측하는 모델이다. 일반적으로 attention-based decoder를 사용하여 문자를 예측하는데, FAN[4]에서는 기존의 AN(Attention Network)에 attention을 집중시키는(focus) FN(Focus Network)을 함께 학습하여 Attention의 성능을 높이는 모델을 제안하였다.

Ⅱ. 본 론

2.1. 기존 STR 모델 분석

TPS-ResNet-BiLSTM-Attn과 TPS-ResNet-BiLSTM-CTC 모델은 [1]에서 소개한 가장 높은 정확도를 보이는 조합으로, 두 모델을 IC13 1015, IC15 2077 데이터 세트에 적용해보고 잘못 예측한 텍스트 이미지를 분석하였다.

IC13 1015는 수평으로 배치된 Regular Dataset로, CTC와 Attn 모두 90%가 넘는 높은 정확도를 보인다. 1,015개의 텍스트 이미지 중에서 둘 다 틀리는 경우가 전체의 약 6.1%로 62개에 해당한다. 이미지 뚜렷하지 못하거나 폰트가 복잡하여 판단을 잘하지 못하는 경우가 대부분이었으며, 1, I, i 와 같이 비슷한 형태의 문자를 구분하지 못하여 틀리는 경우도 많았다. [그림1]과 같이 word box가 label에 비해 크게 설정되어 label이 아닌 문자를 인식하여 틀리는 경우도 있었다. 예를 들어, to를 tol, out을 loutj, price를 pricese로 CTC와 Attention 모두 잘못 예측하였다. CTC와 Attn에 따라 아래 그림 중 두세 개는 맞추었으나 대체로 맞추지 못하였다.

[그림1] IC13 1015의 잘린 문자가 있는 텍스트 이미지

IC15 2077은 회전되거나 왜곡된 텍스트로 이루어진 보다 어려운 Irregular Dataset로, Regular Dataset인 IC13에 비해 흐릿한 이미지, 왼쪽으로 90도 회전된 이미지, 훼손된 이미지 등으로 인해 정확도가 매우 낮았다.

투표 방법	모델	Regular Datasets						Irregular Datasets				평균 정확도	시간 (ms)
		IIIT	SVT	IC03		IC13		IC15		SP	CT		
		3000	647	860	867	857	1015	1811	2077	645	288		
None	CTC	86.267	86.090	94.651	94.348	91.482	90.542	75.814	72.841	76.434	72.474	83.570	0.428
	Attn	87.367	87.326	**95.116**	94.694	92.999	92.217	78.244	75.390	80.155	74.216	85.249	1.410
최빈값	CTC	87.933	87.017	94.419	93.887	92.649	91.429	77.305	74.558	78.760	72.125	84.770	0.966
	Attn	88.100	**88.563**	**95.116**	**95.040**	**94.049**	93.202	**79.072**	**76.223**	**80.310**	**74.913**	**85.971**	8.665
신뢰도	CTC	83.733	79.444	86.279	86.044	84.364	84.532	73.937	71.332	68.837	66.551	79.524	0.972
	Attn	87.500	86.708	93.023	93.195	92.182	91.724	77.305	74.662	76.899	72.474	84.409	8.699
최빈값+ 신뢰도	CTC	84.667	81.298	88.605	86.814	86.700	86.700	74.655	72.008	73.333	68.293	80.782	1.005
	Attn	**88.167**	87.944	95.000	94.579	93.932	**93.300**	78.410	75.598	79.225	73.868	85.627	5.191

[표1] 투표 방법에 따른 정확도

2.2 제안 방법

Test Time Augmentation(TTA)는 학습 데이터가 아닌 평가 데이터 세트에 대해 augmentation을 하여 평가한 후, 예측한 여러 결과를 바탕으로 하나의 결과로 만드는 것을 말한다. 본 논문에서는 잘린 문자로 인하여 생기는 문제를 해결하고자 TTA를 적용하였다.

평가 시에 이미지를 이동(translate)하거나 크기를 변경(scale)하여 augmentation을 한 후 빈도수와 신뢰도를 바탕으로 결과를 선택하였다. 먼저 이미지 이동과 크기를 변경하여 원본으로부터 8개의 이미지를 만들었다. 이동은 잘린 문자를 제거하기 위해 좌우로 움직였고, 크기는 문자 크기나 간격이 일반적이지 않아 생기는 문제를 해결할 가능성이 있어 확대 및 축소하였다. TTA에 사용한 이미지 변형 조합은 [표2]와 같다.

#		0	1	2	3	4	5	6	7	8
이동	L/R	-	L	R	L	R	L	R	L	R
	pixel	-	5		2		5			
크기		1.0					0.75		1.25	

[표2] TTA의 이미지 변형 조합 (L:left, R:right)

이렇게 9가지 데이터에 대하여 최빈값 또는 신뢰도를 바탕으로 투표하여 하나를 선택한다. 최빈값은 가장 많이 예측한 결과를 선택하였으며, 만약 모두 다른 예측을 하였다면 원본 이미지의 결과를 선택하였다. 신뢰도는 각각 예측한 결과에 대한 신뢰도로, 가장 높은 값을 가진 예측을 선택하였다. 또한 신뢰도만으로 투표하는 것은 다수의 예측을 무시할 가능성이 있어 신뢰도와 최빈값을 조합하여 실험하였다. 9개 신뢰도의 평균이 40~70%에 속하면 해당 예측을 선택하고, 그렇지 않다면 최빈값을 선택하였다.

2.3 결과 분석

[표1]에 다양한 평가 데이터 세트(IIIT, SVT, IC03, IC13, IC15, SP, CT)에 대해 TTA를 적용한 결과를 나타내었다. 먼저 최빈값으로 투표한 경우, CTC와 Attn 모두 [1]에서의 정확도보다 높아졌다. CTC는 평균 정확도에서 1.2%가 올랐지만, IC03 860, IC03 867, CT의 경우에는 정확도가 낮아졌다. Attn은 약 0.72% 올랐으며, 정확도가 그대로인 IC03 860을 제외하고는 모두 정확도가 올랐다. IC13의 경우에는 CTC와 Attn 모두 0.8% 이상 정확도가 높아졌다. 신뢰도로 투표한 경우에는 CTC와 Attn 모두 정확도가 낮아졌다. IIIT 3000이 Attn에서 대략 0.13% 올라간 것을 제외하고는 모두 정확도가 낮아졌다.

최빈값과 신뢰도를 조합한 경우에는 신뢰도만으로 투표하는 경우보다 정확도가 높아졌지만, 최빈값으로 투표하는 경우보다는 정확도가 높아지지 않았다. CTC는 기존보다 평균 정확도가 약 2.79% 정도 낮아졌다. Attn의 경우에는 평균 정확도가 0.38% 정도 높아졌으며, 특히 IC13 1015에서는 가장 높은 정확도를 보였다.

위에서 언급한 IC13 1015의 잘린 문자가 있는 경우를 확인하기 위해 [그림1]의 10개 데이터를 확인하였다. 투표 방법으로 신뢰도를 사용한 경우 비록 전체적인 정확도는 낮았지만, 10개의 모든 데이터에 대해서 맞추었다. 하지만 높은 정확도를 보인 최빈값으로 투표한 경우에는 몇 개의 데이터를 맞추지 못하였다.

Ⅲ. 결 론

본 논문에서는 TTA를 이용하여 기존 STR 모델의 정확도를 높이고자 하였다. 이동과 크기를 이용하여 이미지를 변형하였고, 여러 예측 중에서 하나를 선택하는 투표 방법을 실험하였다. 최빈값으로 투표하는 것이 가장 높은 정확도를 보여주었지만 잘린 문자 텍스트 이미지에 대해서는 오히려 신뢰도를 사용하였을 때 정확도가 높아졌다. 이 실험을 통해 TTA가 STR의 정확도를 높일 수 있다는 것을 알 수 있었다. 신뢰도와 최빈값의 조합을 단순히 이분법적으로 사용하여 조합하였지만, 둘의 특징을 이용하여 투표에 반영한다면 더 좋은 결과를 얻을 수도 있을 것이다. 또한 하나의 예측을 선택하는 것이 아니라 여러 예측을 조합하여 최종 결과를 예측하는 방법도 고려해볼 만하다. TTA를 이용하지 않고 Prediction 단계에서 CTC와 Attn의 서로 다른 특징을 이용하여 조합하는 방법 또한 STR 성능을 높일 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 한국연구재단 중견연구자지원사업(NRF-2019R1A2C1005881)과 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업(2015-0-00910)의 지원을 받아 수행하였음.

참 고 문 헌

[1] JeonHun Baek, “What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis”, 2019.

[2] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In CVPR, 2016.

[3] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. In TPAMI, volume 39, IEEE, 2017.

[4] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In ICCV, 2017.

[5] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In KDD, 2018.