BoostCamp AI Tech – Project2 Wrap-up Report

CV-23

(김세연, 안지현, 김상유, 김태욱, 김윤서)

1. 프로젝트 개요

1.1 프로젝트 주제

영수증 이미지 데이터에서 글자를 검출하는 ocr 대회로 이미지에서 글자가 있는 위치를 예측하는 text detection만을 수행한다. 해당 대회는 data-centric 대회로 모델은 EAST로 고정하고 data만을 수정해 성능을 높이는 것을 목표로 한다.

1.2 대회 데이터셋

다국어(중국어, 일본어, 태국어, 베트남어) 영수증 이미지 train 400개(언어별 100개), test 120개(언어별 30개)로 구성되어 있다.

2. 프로젝트 팀 구성 및 역할

이름	역할
김세연	Data augmentation, 모델 실험 및 평가, cv2 데이터 로드 및 기본 증강 구현
안지현	Wandb 기반 데이터셋 버전 및 실험 관리, annotation 수정, 모델 실험
김상유	Val_data, 모델 실험 및 평가, 모델 앙상블, Data cleansing
김태욱	외부 데이터 추가, Data cleansing, 모델 실험
김윤서	Streamlit 결과 시각화, 외부 데이터 추가, Data cleansing, 모델 실험

3. 프로젝트 실험 방법 및 결과

3.1 Data augmentation

대회 데이터는 영수증을 사람이 찍은 데이터이기에 손의 떨림 등으로 인한 노이즈가 있는 데이터가 다수 존재한다. 따라서 MotionBlur, GaussianBlur, MultiplicativeNoise 등의 증강 기법을 적용하였다.

다음은 validation data를 통해 DetEval을 평가한 결과이다.

Metric	baseline	Motionblur	GaussianBlur	MultiplicativeNoise	all
Precision	0.1617	0.6512	0.2638	0.5993	0.5632
Recall	0.5442	0.5062	0.6931	0.4116	0.6350
F1-score	0.2493	0.5696	0.3821	0.4880	0.5970

Precision은 Motionblur가 가장 높고, Gaussianblur가 가장 낮다. recall은 Gaussianblur가 가장 높으며, MultiplicativeNoise이 가장 낮다. 또한, 각 증강 기법의 단점들을 보완하고, 일반화 성능을 높이기 위해 Motionblur, GaussianBlur, MultiplicativeNoise의 증강을 동시에 하는 all을 시도했으며, F-1 score는 style 증강 기법 중 가장 높은 점수를 보였다.

3.2 외부 데이터 추가

SROIE - 2019년 ICDAR에서 진행한 SROIE 대회 데이터셋으로 대부분 알파벳으로 구성되어있다. Base 데이터와 달리 구분선에는 bbox가 그려져 있지 않고 스캔한 영수증과 직사각형 bbox로만 이루어져있다는 특징이 있다. (train 약 950개)

CORD – Crowd-sourcing을 통해 11,000개 이상의 인도네시아 영수증 이미지를 수집 후, OCR을 위한 image와 bbox/text annotations. 구문 분석을 위한 multi-level semantic labels을 포함한다.

다음은 validation data를 통해 DetEval을 평가한 결과이다. (3.1과 다른 validation data을 사용)

Metric	base	Sroie+base	Sroie -> base	Cord+base	Cord -> base
Precision	0.581	0.7282	0.7643	0.8613	0.8076
Recall	0.765	0.6382	0.8052	0.7952	0.8015
F1-score	0.661	0.6802	0.7842	0.8269	0.8045

데이터의 개수를 증가시켜 성능 향상을 이끌었다. Base 데이터와 분포 차이가 있는 Sroie의 경우 pretrain 했을 때 좀 더 확실한 성능 향상 효과를 보였으며, base 데이터와 bbox 형태, style이 더 유사한 Cord가 학습 방법에 관계없이 Sroie보다 높은 성능을 보였다.

3.3 Data Cleansing

대회 데이터셋 (영수증 데이터셋)에 존재하는 BBOX 노이즈 제거 및 추가한 외부 데이터셋과의 일관성 유지를 위해 Data cleansing 수행하였다.

3.3.1 Annotation Tool

웹 환경에서 여러 팀원이 협업하여 word 별 태깅 및 points 수정을 할 수 있는 CVAT을 선택하였다.

3.3.2 Annotation Guideline

Try 1. 기존 데이터셋의 구분선과 같은 모델에 혼동을 줄 수 있는 특수 문자(, . ;)를 포함하는 BBOX는 annotation 파일에 transcription이 비어있는 경우 제거한다.

Try 2. 경진 대회의 평가 기준인 테스트 데이터셋의 경우 구분선과 같은 특수문자를 포함하는 BBOX를 탐지하도록 세팅되어 있으므로 학습에 추가로 사용하는 외부 데이터셋의 Annotation을 기존 데이터셋과 정답에 대해 일관성이 있도록 구분선을 포함하여 BBOX를 추가한다. 또한 외부 데이터셋에서 BBOX가 그려져 있지 않은 텍스트 영역에 BBOX를 추가하거나 올바르게 그려져 있지 않은 (텍스트 영역을 전부 포함하지 않거나 텍스트 영역에 비해 너무 큰) BBOX를 수정한다.

3.4 CV2

데이터 로드를 PIL에서 CV2로 변경하여 속도 개선 효과를 확인해보았다.

- pretrained : new_SROIE, fineturning : CORD + base

데이터 로드	precision	recall	f1	elapsed time
PIL	0.7833	0.8651	0.8222	평균 280초
CV2	0.8000	0.8746	0.8356	평균 240초

CV2로 데이터를 로드한 실험 F1 스코어 성능이 0.02 정도 높으며 학습시간은 40초 가량 빠르다.

- pretrained : new_ SROIE, fineturning : new_CORD + base

데이터 로드	precision	recall	f1	elapsed time
PIL	0.7939	0.8795	0.8345	평균 290초
CV2	0.7789	0.8831	0.8277	평균 230초

CV2로 데이터를 로드한 실험의 F1 스코어 성능이 0.007 정도 낮으며, 학습시간은 60초 가량 빠르다.

4. Ensemble 결과

앙상블은 test data의 public 점수를 기준으로 여러 조합을 시도하였다.

데이터	비고	precision	recall	f1
Cord+base	30epoch	0.8797	0.8005	0.8382
Sroie -> base		0.7644	0.8192	0.7908
Cord -> base		0.8172	0.8271	0.8221

앙상블1 결과 = precision = 0.9072 recall = 0.7761 f1-score = 0.8365

데이터	비고	precision	recall	f1
Cord+base	76epoch	0.8919	0.8417	0.8660
Cord+base	100epoch	0.8635	0.8607	0.8621
New_Sroie-> Cord+base		0.7861	0.8688	0.8254
New_sroie -> Cord+base	cv2	0.7956	0.8724	0.8322

앙상블2 결과 = precision = 0.8980 recall = 0.8526 f1-score = 0.8747

데이터	비고	precision	recall	f1
Cord+base	30epoch	0.8797	0.8005	0.8382
Cord+base	76epoch	0.8919	0.8417	0.8660
New_Sroie-> newcord+base	cv2	0.7917	0.8926	0.8391
앙상블1		0.9072	0.7761	0.8365
앙상블2		0.8980	0.8526	0.8747

최종 앙상블 = precision = 0.9176 recall = 0.8606 f1-score = 0.8882

(iou=0.4, vote=3, wbf)

5. 자체 평가 의견

이번 프로젝트에서는 Wandb에서 통일된 평가 지표를 사용하고 데이터셋 버전 관리를 위해 Wandb artifacts를 사용했으며, 실험 템플릿을 통해 체계적인 실험 기록이 이뤄졌다. 팀원 간 공유의 측면에서 이전보다 개선된 모습을 보였다.

하지만 데이터 개수를 증가시켰을 때 효과가 있다는 점을 확인했으나 2개의 외부 데이터셋만 사용해본 것과 합성 데이터를 사용해보지 못한 것, annotation 수정 방법을 다양하게 실험해보지 못한 것에서 아쉬움이 남았다. 다음 프로젝트에서는 실험의 수 자체를 늘려 프로젝트 결과 개선에 초점을 두고자 한다.