

학습 데이터 추가 및 수정을 통한 이미지 속 글자 영역 검출 성능 개선 대회

7조 - 컴퓨터 구조

1. 프로젝트 개요

- 프로젝트 주제 : 글자 검출 대회
- 프로젝트 개요 및 목표

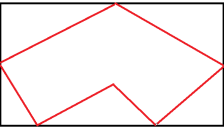
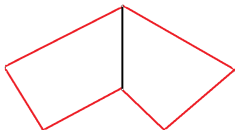
학습 데이터 추가 및 수정을 통한 이미지 속 글자 영역 검출 성능 개선 대회

- 목표 : 데이터의 검수가 얼마나 성능을 올려줄 지 확인
task에 맞추어서 데이터를 넣었을 때 성능이 얼마나 오르는지 확인
데이터 특성에 맞춘 optimizer, batch size, scheduler을 고민

2. 프로젝트 팀 구성 및 역할

- 김성민 : optimizer, scheduler, TTA, data - camp에서 준 것 넣기, 500~1000번 검수,
train에 wandb, data여러 종류 줄 수 있게 dataset 및 train수정, visualized tool
- 박지민 : input size 변경 실험, epoch 변경 실험
- 박진형 : visualize 및 검수해야할 파일 체크해주는 코드, aihub데이터 추가
- 심세령 : polygon cutting방법 고민, 0~500번 검수
- 윤하정 : 팀원들의 실험 케이스를 보고 조합

3. 프로젝트 수행 절차 및 방법

적용 내용	적용 이유	적용 결과(AVG F1)	아쉬웠던 점
Camp dataset 추가	데이터 양을 증가시켜 성능 향상 기대	0.4655 -> 0.5595	실수로 삼각형 형태로 자름
epoch test	valid가 없어서 에폭을 여러개 제출해서 실험해야만 했음	130epoch 0.5470 ->0.5725, 0.5755	valid가 필요했음
데이터 검수(0~500)	annotation에 문제가 있는 데이터를 제거하여 양질의 데이터를 제공하기 위함	0.5755 -> 0.5710	데이터가 줄었는데 동일 에폭을 제출
폴리곤을 외접 사각형으로 자름 	다양한 모양의 글자 영역 처리 시도함과 동시에 one-to-many를 줄여보고자 함	0.5710 -> 0.5695	글자 영역 이외의 여백이 많아져 성능 하락
adamW	동일 batch 내에서 lr을 조절함으로써 학습의 효율 향상	0.5695->0.5710	
폴리곤을 내부 사각형들로 자름 	글자 크기에 따라 잘라 여백을 줄여 성능향상 기대	0.5710 -> 0.5845	

cos annealing	기존 scheduler의 lr이 급격하게 변경되는 지점의 완화	0.5845 -> 0.5920	
Nadam	nag의 원리가 적용된 아담으로 가장 높은 성능 향상을 기대	0.5920 -> 0.4475	reference 확인이 제대로 되지 않아 lr 설정에 오류
aihub data 사용	데이터를 추가로 사용함으로써 성능 향상을 기대(기존 데이터 비해 너무 양이 많아 일부 데이터만 사용)	0.4635	기존 data의 annotation을 잘못 된 것을 넣음
TTA	짧은 대회기간에 대응하여 augmentation을 다양하게 적용해볼 수 있을 것을 기대함	0.1470	ensemble 방법에 대한 고민

4. 프로젝트 수행 결과

- 탐색적 분석 및 전처리 (학습데이터 소개)
 - 학습데이터: ICDAR17_Korean dataset + 직접 annotation 진행한 dataset
 - Input: Image data
 - Output: 글자 영역 bounding box
 - 전처리
 - 데이터 검수 및 제거
 - Polygon cutting(제거, 삼각형, 외접사각형, 사각형 여러개)
- 모델 개요 : EAST로 고정

5. 자체 평가 의견

- 잘한 점들
 - 실험을 치밀하게 해서 특정 기능이 영향을 주는 것을 명확하게 볼 수 있었음
 - 데이터에 영향을 주거나 받을 수 있는 것(optimizer)도 다양하게 실험해봄
 - annotation을 검수함
 - 직접 코드를 짜서 씀
 - bound box 종류 3가지 비교 해봄
- 시도 했으나 잘 되지 않았던 것들
 - AI hub데이터를 가져왔으나 성능이 오르지 않음 : camper data의 annotation이 잘못 되어 있었음
 - Nadam : learning rate에 대한 논문 자료를 리뷰하고 만들었으면 더 좋았을 듯함
- 아쉬웠던 점들
 - annotation을 우리가 원하는 형태로 만들었는지 점검하고 제출하면 좋았을 듯함
 - Validation set을 만들지 않음
 - 우리가 실험한 것에 대해 확신을 가지고 했으면 더 좋았을 듯함
 - 대회기간이 짧아서 많은것을 시도해보지 못하였음
 - augmentation을 거의 하지 않음(글자영역 crop 등)
 - 데이터 version을 만들지 않아서 추후에 데이터가 헷갈렸음
 - TTA의 앙상블 방법을 마치기 전에 프로젝트가 끝남