



TrOCR

Transformer-based Optical Character Recognition with Pre-trained Models


(<https://arxiv.org/abs/2109.10282>)

한성대학교 1971336 김태민



A cluster of overlapping triangles in shades of blue, green, and red.


목차

- Abstract
 - Introduction
 - Model Architecture
 - Encoder Initialization & Decoder Initialization
 - Experiments
 - Competition
- 
- A cluster of overlapping triangles in shades of light gray.



TrOCR

Abstract

- 일반적인 OCR모델은 이미지 이해를 위한 CNN과 문자 수준의 텍스트 이해를 위한 RNN으로 정의 및 일반적으로 후처리를 위해 추가적인 언어 모델이 필요합니다. 이를 단순히 Transformer기반 이미지, 텍스트 모델을 사용하여 간단하지만 효과적인 모델을 구성하였다.
 - Transformer의 Pre-Training 모델이 CNN을 대체할수 있게 됨으로 기존 CNN 기반의 Backbone 모델을 사용하지 않고 Transformer모델 기반으로 모델을 구성
- 

TrOCR

Introduction

- 일반적으로 OCR 시스템은 크게 두가지 모듈로 구성됩니다. 텍스트 감지, 텍스트 인식 모듈입니다.
- 텍스트 감지는 일반적으로 YoLov5 및 DBNet과 같은 기존의 Object detection 모델을 적용할 수 있지만
- 한편 텍스트 인식은 텍스트 이미지에 대한 콘텐츠를 이해하고 이를 시각적 신호에서 자연어 토큰으로 바꾸는 것을 목표로 합니다.
- 이를 위해 일반적으로 CNN기반 인코더와 RNN기반 디코더를 활용하는 인코더-디코더 문제로 구성됩니다.
- 본 논문에서는 텍스트 인식에 초점을 맞추고 텍스트 감지는 향후 작업으로 남겨둡니다.

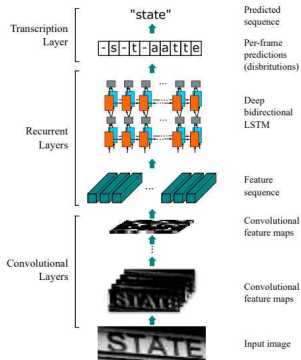
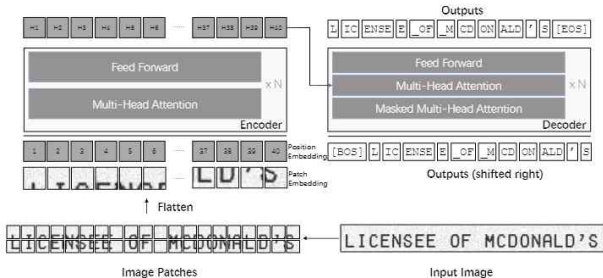


Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

TrOCR

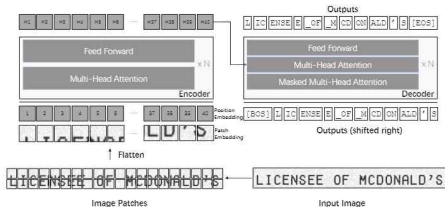
Introduction

- 1.) 결과적으로 CNN을 사용하지 않았으며 이점으로 image-specific inductive biases를 도입하지 않았습니다. 이를 통해 모델을 구현하고 유지하기 매우 간단합니다.
- 2.) 또한 간단하게 다국어 버전으로 확장을 할 수 있습니다.



TrOCR

Model Architecture



- TrOCR은 이미지 기반의 Transformer과 언어 모델링을 위한 Transformer로 인코더 디코더 구조로 구성됩니다.
- 인코더는 DeiT와 BEiT를 사용합니다.
- 디코더로는 RoBERTa과 MiniLM를 사용합니다.
- 실제로 Transformer모델에 인코더-디코더 구조는 없고 위 텍스트 모델 2개 모두 인코더 일뿐이므로 구조가 다릅니다. 이를 해결하기 위해 맵핑되는 부분은 수동으로 설정하여
- 위 두 모델로 초기화 후 부재 매개변수는 랜덤으로 초기화를 진행합니다.

TrOCR

Experiments

Encoder	Decoder	Precision	Recall	F1
DeiT _{BASE}	RoBERTa _{BASE}	69.28	69.06	69.17
BEiT _{BASE}	RoBERTa _{BASE}	76.45	76.18	76.31
ResNet50	RoBERTa _{BASE}	66.74	67.29	67.02
DeiT _{BASE}	RoBERTa _{LARGE}	77.03	76.53	76.78
BEiT _{BASE}	RoBERTa _{LARGE}	79.67	79.06	79.36
ResNet50	RoBERTa _{LARGE}	72.54	71.13	71.83

Model	Recall	Precision	F1
CRNN	28.71	48.58	36.09
Tesseract OCR	57.50	51.93	54.57
H&H Lab	96.35	96.52	96.43
MSOLab	94.77	94.88	94.82
CLOVA OCR	94.3	94.88	94.59
TrOCR _{SMALL}	95.89	95.74	95.82
TrOCR _{BASE}	96.37	96.31	96.34
TrOCR _{LARGE}	96.59	96.57	96.58

- 실제로 여러 모델을 인코더 디코더 구조로 설계하여 테스트한 결과 BEiT와 RoBERTa 조합이 높은 성능을 이끌어 냈으며
- 네이버에서 개발한 CLOVA OCR 보다 높은 성능을 달성 하였다.

TrOCR

Experiments

Model	Test datasets and # of samples							
	IIIT5k	SVT	IC13		IC15		SVTP	CUTE
	3,000	647	857	1,015	1,811	2,077	645	288
PlugNet (Mou et al. 2020)	94.4	92.3	–	95.0	–	82.2	84.3	85.0
SRN (Yu et al. 2020)	94.8	91.5	95.5	–	82.7	–	85.1	87.8
RobustScanner (Yue et al. 2020)	95.4	89.3	–	94.1	–	79.2	82.9	92.4
TextScanner (Wan et al. 2020)	95.7	92.7	–	94.9	–	83.5	84.8	91.6
AutoSTR (Zhang et al. 2020a)	94.7	90.9	–	94.2	81.8	–	81.7	–
RCEED (Cui et al. 2021)	94.9	91.8	–	–	–	82.2	83.6	91.7
PREN2D (Yan et al. 2021)	95.6	94.0	96.4	–	83.0	–	87.6	91.7
VisionLAN (Wang et al. 2021)	95.8	91.7	95.7	–	83.7	–	86.0	88.5
Bhunia (Bhunia et al. 2021b)	95.2	92.2	–	95.5	–	84.0	85.7	89.7
CVAE-Feed. ¹ (Bhunia et al. 2021a)	95.2	–	–	95.7	–	84.6	88.9	89.7
STN-CSTR (Cai, Sun, and Xiong 2021)	94.2	92.3	96.3	94.1	86.1	82.0	86.2	–
ViTSTR-B (Atienza 2021)	88.4	87.7	93.2	92.4	78.5	72.6	81.8	81.3
CRNN (Shi, Bai, and Yao 2016)	84.3	78.9	–	88.8	–	61.5	64.8	61.3
TRBA (Baek, Matsui, and Aizawa 2021)	92.1	88.9	–	93.1	–	74.7	79.5	78.2
ABINet (Fang et al. 2021)	96.2	93.5	97.4	–	86.0	–	89.3	89.2
Diaz (Diaz et al. 2021)	96.8	94.6	96.0	–	80.4	–	–	–
PARSeq _A (Bautista and Atienza 2022)	97.0	93.6	97.0	96.2	86.5	82.9	88.9	92.2
MaskOCR (ViT-B) (Lyu et al. 2022)	95.8	94.7	98.1	–	87.3	–	89.9	89.2
MaskOCR (ViT-L) (Lyu et al. 2022)	96.5	94.1	97.8	–	88.7	–	90.2	92.7
TrOCR _{BASE} (Syn)	90.1	91.0	97.3	96.3	81.1	75.0	90.7	86.8
TrOCR _{LARGE} (Syn)	91.0	93.2	98.3	97.0	84.0	78.0	91.0	89.6
TrOCR _{BASE} (Syn+Benchmark)	93.4	95.2	98.4	97.4	86.9	81.2	92.1	90.6
TrOCR _{LARGE} (Syn+Benchmark)	94.1	96.1	98.4	97.3	88.1	84.1	93.0	95.1

TrOCR

Competition

#	팀	팀 멤버	점수	재출수	종료일
7	한성대학교_김태민과 아이들		0.94693	35	한 시간 전
1	물리솔거기보병중야요원도리당삼여요		0.96298	34	2일 전
2	윤한섭사회법자들		0.96271	48	3시간 전
3	어나부기		0.95777	31	20시간 전
4	전두재씨		0.95479	39	5시간 전
5	20년대3아이브시도들		0.94967	24	6시간 전
6	mrncd		0.94828	20	9일 전
7	한성대학교_김태민과 아이들		0.94693	35	한 시간 전
8	최정영		0.94571	54	하루 전
9	마요우		0.94072	35	10시간 전
10	Derekim		0.93024	23	3일 전
11	AP		0.92934	21	2일 전
12	minibook		0.92705	6	5일 전

대회에서도 각종 데이터 전처리
통해 최종순위는 예선 9위를 달성
했으며 다른 다국어에서도 잘 동
작한다는것을 확인했습니다.

