Introduction to
Scene Text Detection and Recognition

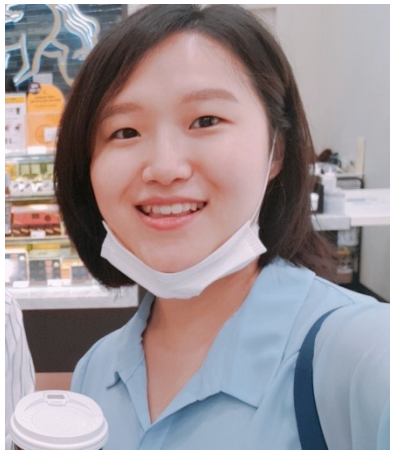
2021. 04. 23

발표자 : 김정원

발표자 소개

❖ **이름: 김정원**

- 고려대 산업경영공학과 석사과정(2020.9~)
- 데이터마이닝 및 품질애널리틱스 연구실(김성범 교수님)



❖ **관심 연구 분야**

- Machine learning & Deep learning
- Object detection & segmentation
- Scene text detection

❖ **연락처**

- jwnkim@korea.ac.kr

목차

Introduction

❖ What is Scene text detection and recognition?

- 일상적인 풍경 이미지에서 **글자가 있는 영역을 탐지**하고 이를 **컴퓨터 문자로 변환**하는 문제
- 이미지 인식 연구의 응용 분야로, Object detection·Semantic segmentation·Sequential image classification 등 알고리즘이 복합적으로 사용됨
- 이미지 번역, 차량 번호판·이정표·명함 인식, 이미지 검색 등에 실생활에서 다양하게 활용되고 있음

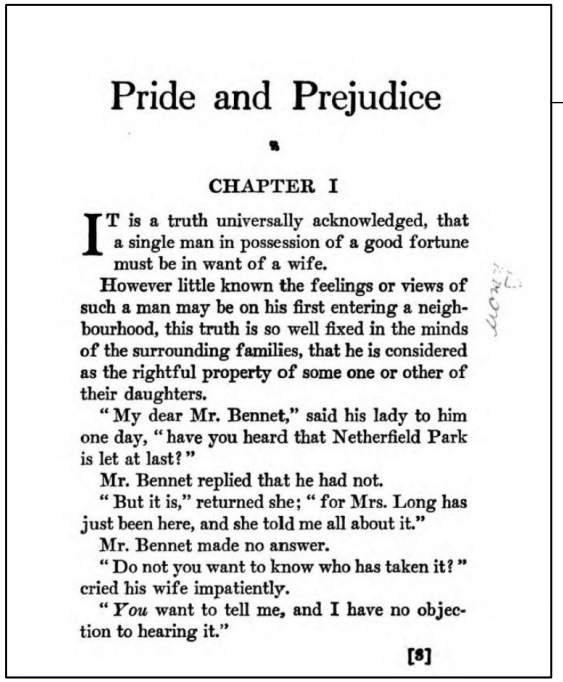
'Lyon',
'STARBUCKS',
'COFFEE'

'35리 6110'

Introduction

❖ What is different from traditional OCR?

- 고전적인 Optical Character Recognition은 종이 문서 등에 인쇄된 문자를 읽어내는 문제
- OCR에 비해 복잡도가 높아 더욱 정교한 모델이 필요함



Pride and Prejudice

CHAPTER I

IT is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife.

However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do not you want to know who has taken it?" cried his wife impatiently.

"*You* want to tell me, and I have no objection to hearing it."

[3]

OCR		Scene text
단순함	배경	복잡함
규칙적	글씨체	다양함
수평	글자 배열	각도 및 구도가 다양함
단조로움	색	다채로움

Challenge!

Introduction

❖ Challenge of Scene text detection and recognition

- Scene text detection and recognition을 어렵게 하는 주요 문제들
  - ✓ **비스듬 하거나 회전된 단어**는 어떻게 탐지할까? (Arbitrary-oriented/Multi-oriented text)
  - ✓ **서로 겹쳐져 있는 문자**는 어떻게 판별할까? (Occlusion)
  - ✓ **곡선형으로 나열된 단어**는 어떻게 할까? (Curved text)
- 특히 글자 배열 문제를 해결하게끔 다양한 특성을 가진 오픈 데이터셋이 존재함

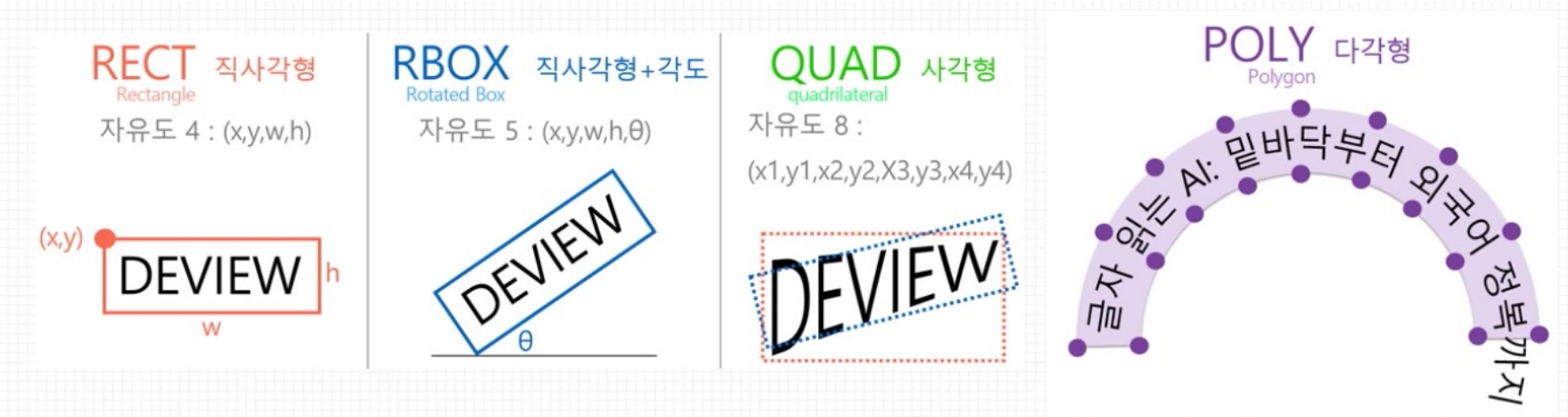
< Arbitrary-oriented/Multi-oriented text >

< Curved text >

Introduction

❖ Challenge of Scene text detection and recognition

- Detection format 예시
  - ✓ 직사각형
  - ✓ 회전된 직사각형
  - ✓ 사각형
  - ✓ 다각형

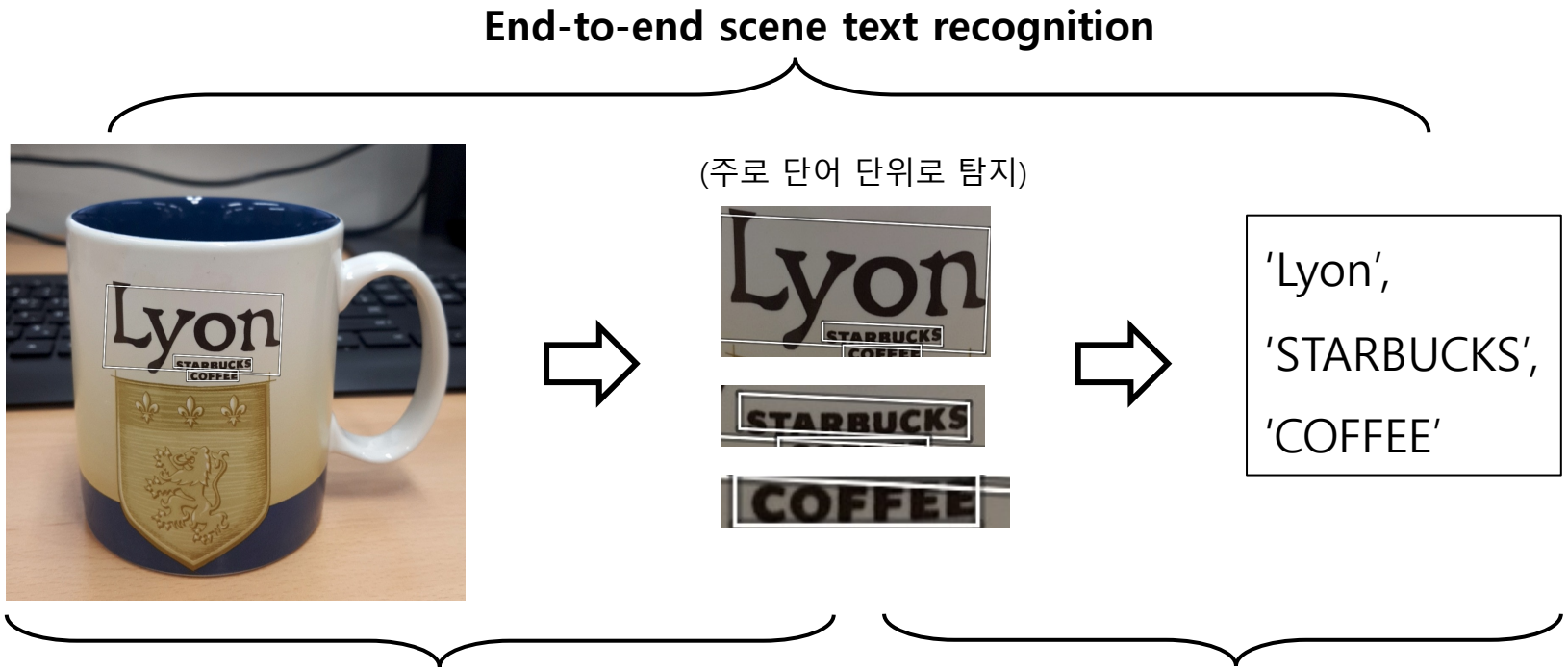


  - ✓ 그외 Segmentation 방식도 활발히 사용되고 있음

Introduction

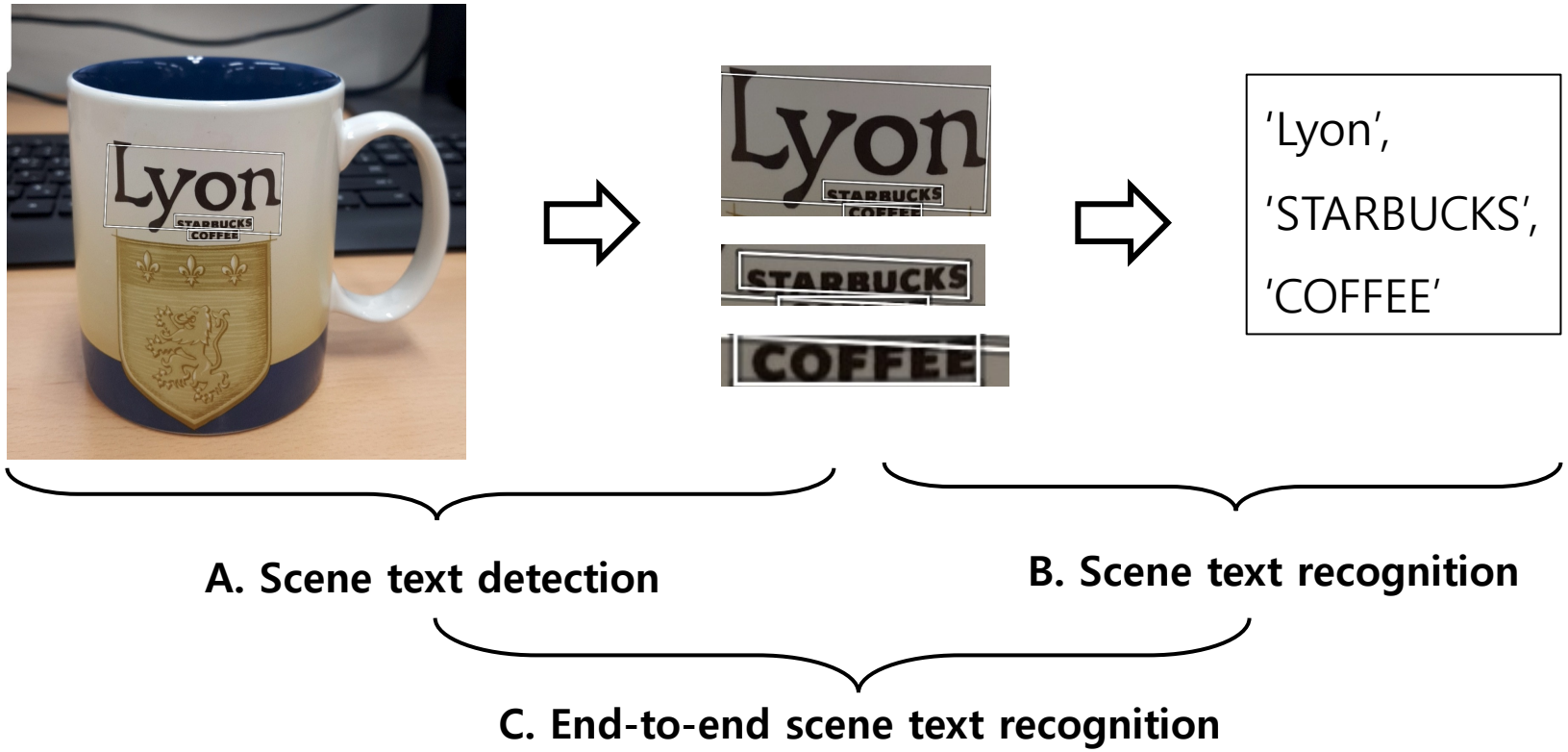
❖ **Scene text detection** vs **recognition** vs **End-to-end text recognition**

- 이미지에서 글자가 위치한 '영역'을 탐지하는 Detection
- 영역별로 잘라 영역 내 글자를 읽어내는 Recognition
- Detection과 Recognition을 한꺼번에 수행하는 End-to-end scene text recognition(text spotting이라고도 함)



Introduction

❖ **Scene text detection** vs **recognition** vs **End-to-end text recognition**



A. EAST: An Efficient and Accurate Scene Text Detector(Zhou. et al., 2017)

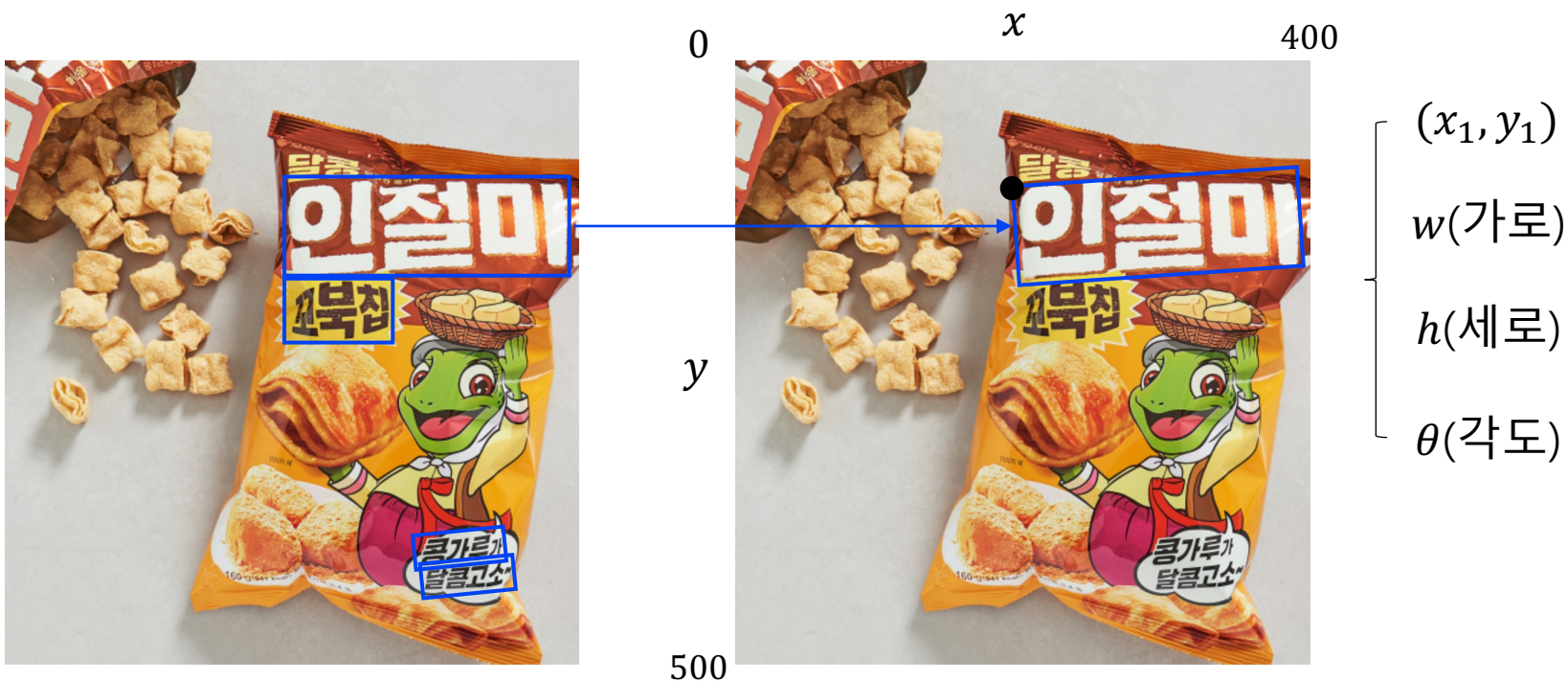
B. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition(Shi. et al., 2015)

C. FOTS: Fast Oriented Text Spotting with a Unified Network(Liu. et al., 2018)

Scene text detection

❖ Basic of scene text detection

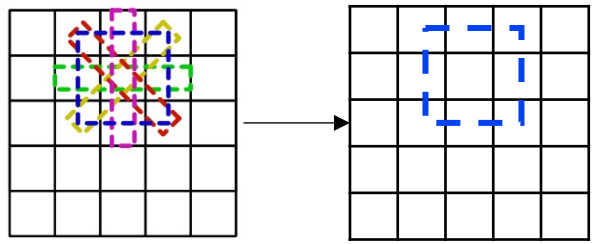
- 글자가 위치한 **bounding box의 좌표**를 최대한 정확히 맞추는 것이 목표이기 때문에 Regression 문제로 접근
- 글자 영역을 Region proposal 또는 Region of Interest(ROI)라고도 부름



Scene text detection

❖ Basic of scene text detection

- 주로 CNN으로 이미지의 특징을 추출한 뒤, 디코더를 통해 단어 영역을 생성함
- 단어 영역이 될 만한 여러 후보(Anchor box)를 만든 뒤 ROI를 추려내는, 고전적인 Object detection 방식도 사용됨



- 단어 정합성을 높이기 위해, 글자(character)와 글자 사이 여백(link)을 각각 탐지해 하나의 단어 영역으로 합치는 방식의 알고리즘도 다수 제안됨(Shi et al.(2017a), Baek et al.(2019b))

Scene text detection

❖ **EAST: An Efficient and Accurate Scene Text Detector(Zhou. et al., 2017)**

- 2021년 4월 현재 756회 인용
- 기존 Text detection 모델들이 3~5차례 Convolution 블록을 거치게 한 것과 달리 하나의 Convolution 블록으로 줄여 연산 시간을 대폭 단축함
- 이미지 분할을 위해 고안된 Fully Convolutional Network(FCN) 알고리즘을 활용해, 단어가 포함된 Rotated rectangle 또는 Quadrilateral box를 예측함

EAST: An Efficient and Accurate Scene Text Detector

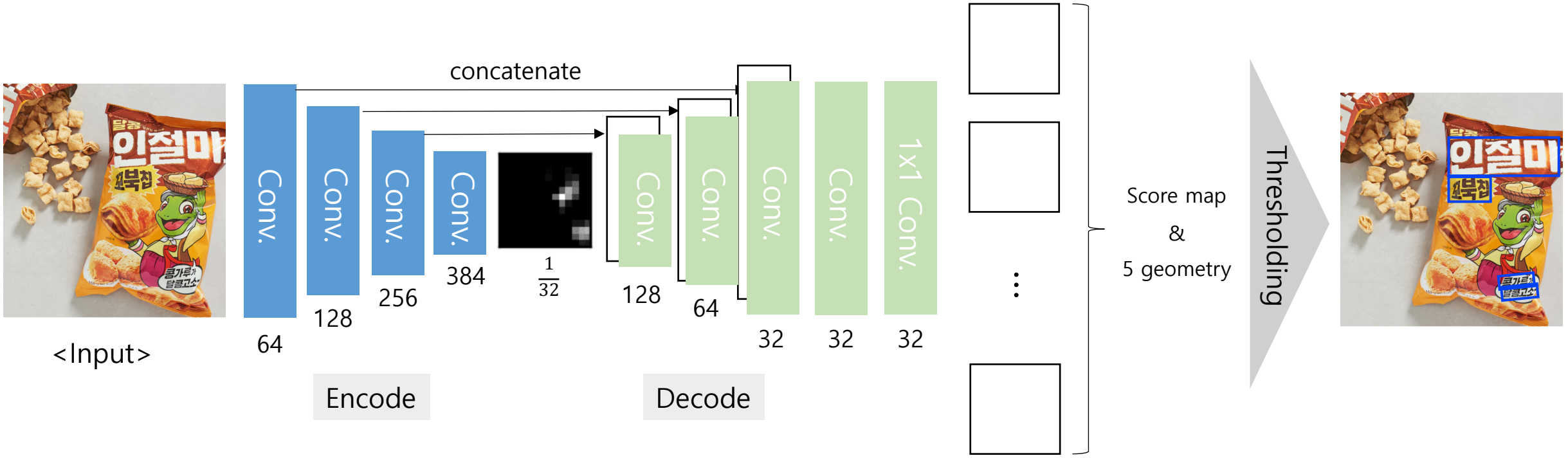
Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang

Megvii Technology Inc., Beijing, China
{zxy, yaocong, wenhe, wangyuzhi, zsc, hwr, liangjiajun}@megvii.com

Scene text detection

❖ Structure of EAST model

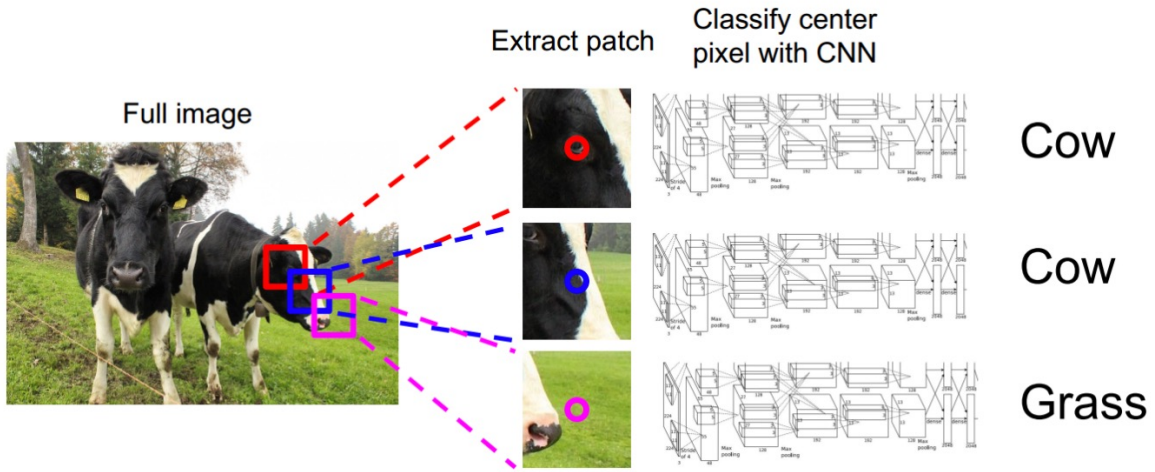
- Input : 가로 512, 세로 512의 RGB 이미지
- Output : Rotated rectangle bounding box의 5개 정보(x1, y1, w, h, 각도)
- U자 모양의 FCN 구조를 사용해 더욱 정확한 Localization을 하고자 함

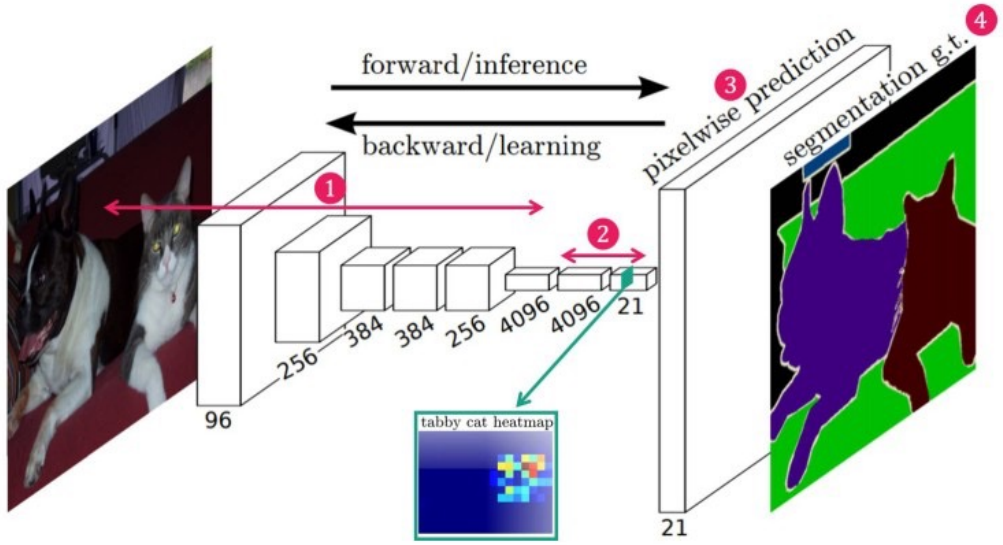


Scene text detection

❖ Fully Convolutional Network(FCN)

- Semantic Segmentation을 위해선 이미지의 각 픽셀이 어떤 Class에 속하는지 분류해야 함.
- 모든 픽셀을 분류 모델에 투입해 구분할 수도 있지만 매우 비효율적임.
- **FCN**은 기존 CNN에서 마지막 Fully Connected layer를 거치기 전 **위치 정보가 보존된 feature map을 활용해 Segmentation map으로 복원**하는 모델. 특징을 압축할 때 과정과 정반대로 Unpooling, Transposed convolution 등을 사용.

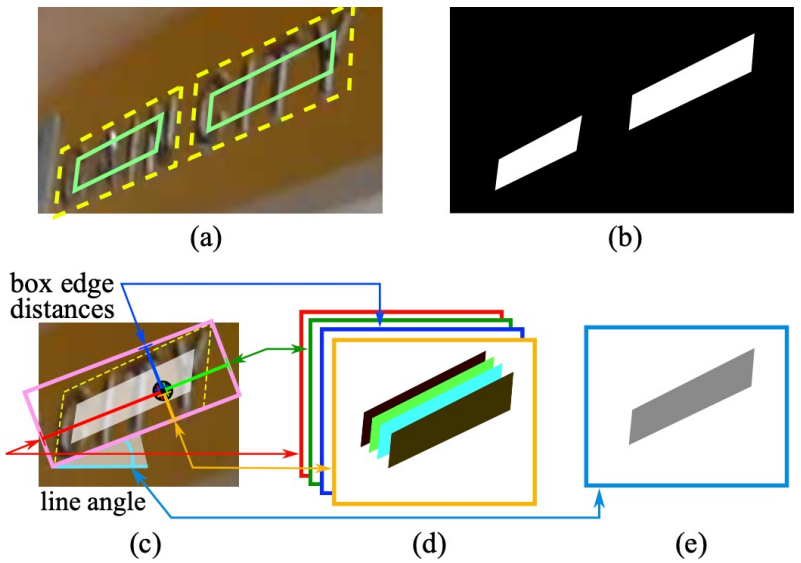




Scene text detection

❖ Output of EAST model

- Score map을 바탕으로 5개 정보가 출력
  - ✓ (b) Score map : 각 픽셀이 단어 영역 내에 있을 확률
  - ✓ (d) 단어 box 추정 후 각 픽셀과 box 4개 변 사이의 거리 정보
  - ✓ (e) 단어 box가 회전된 각도



Thresholding

$$L = L_s + \lambda_g L_g$$

예측 label

실제 label

Scene text detection

❖ Experiment result

- ICDAR2015 기준 F-score 80.72%의 성능 보임
- 2021년 현재 SOTA급 모델(90% 전후의 성능)보다 성능은 떨어지지만 이후 End-to-end 방식 모델에도 사용되는 등 Scene text detection 분야의 중요한 방향성을 제시함

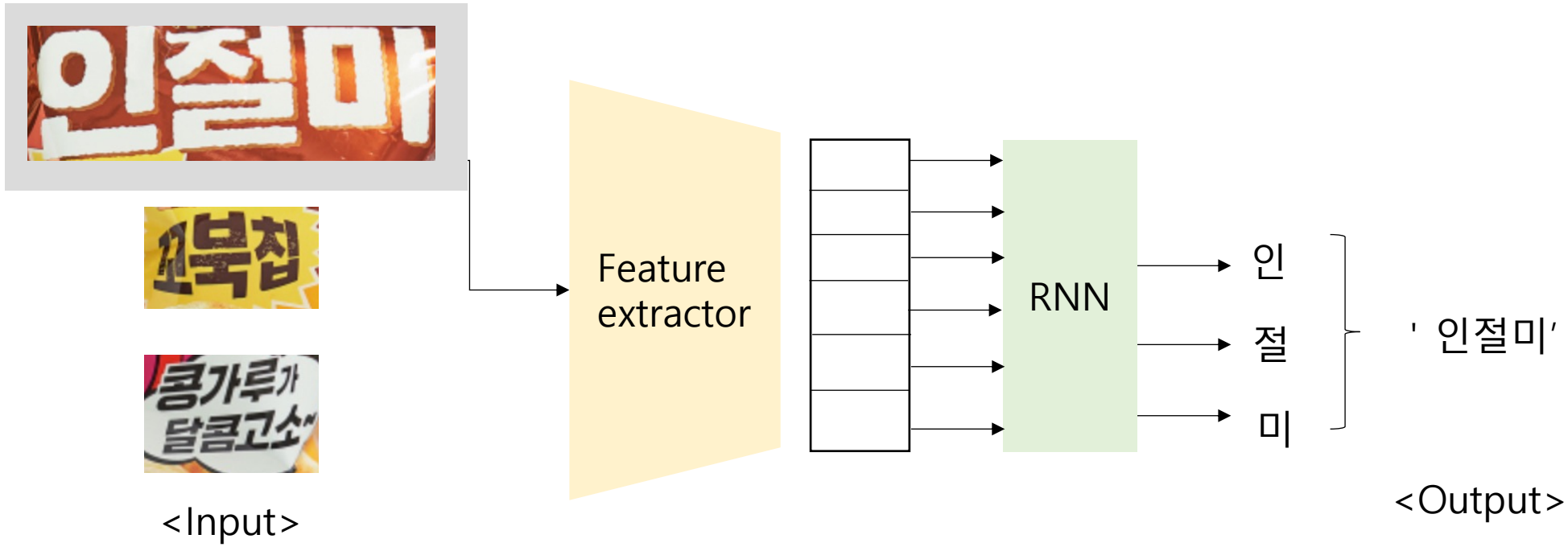
Algorithm	Recall	Precision	F-score
Ours + PVANET2x RBOX MS*	**0.7833**	0.8327	**0.8072**
Ours + PVANET2x RBOX	0.7347	**0.8357**	**0.7820**
Ours + PVANET2x QUAD	0.7419	0.8018	0.7707
Ours + VGG16 RBOX	0.7275	0.8046	0.7641
Ours + PVANET RBOX	0.7135	0.8063	0.7571
Ours + PVANET QUAD	0.6856	0.8119	0.7434
Ours + VGG16 QUAD	0.6895	0.7987	0.7401
Yao *et al.* [41]	0.5869	0.7226	0.6477
Tian *et al.* [34]	0.5156	0.7422	0.6085
Zhang *et al.* [48]	0.4309	0.7081	0.5358

<Table 3 in the paper>

Scene text recognition

❖ Basic of Scene text recognition

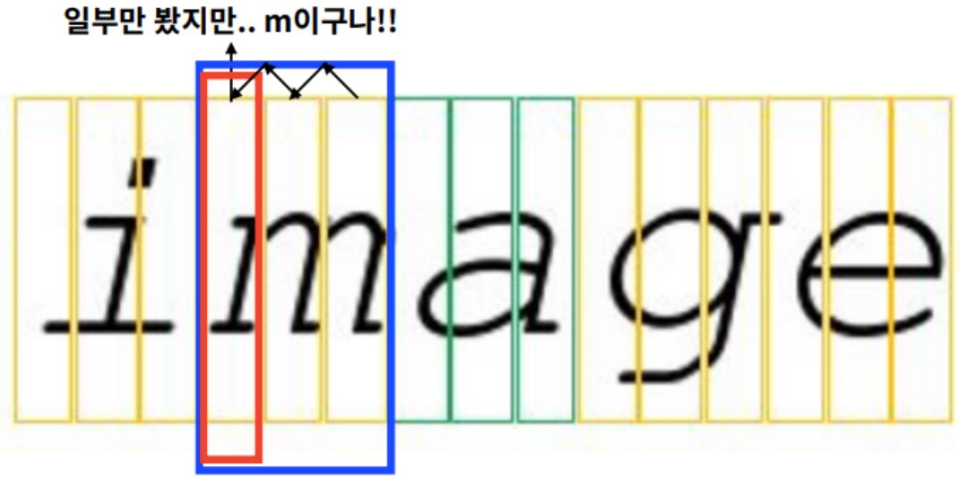
- 각 단어 영역이 어떤 문자인지 찾는 Classification 문제
- 단어 영역에 해당하는 이미지로부터 특징 추출 후 sequential하게 만들어 각 글자의 조합(단어)을 찾아가는 방식



Scene text recognition

❖ Why RNN?

- 이미지마다 글자의 크기, 배치가 다르기 때문에 글자 단위로 정확하게 나누는 것은 어려움
- feature map의 부분적 정보만을 이용해 글자를 예측하려면 앞뒤의 다른 정보를 종합적으로 고려해야 함
  - ✓ (예시) m의 일부는 i일 수도 l일 수도 있지만 시퀀스 내 다른 정보를 고려해 m으로 판단 가능
- 최근에는 Attention model을 활용한 text recognition 모델도 주목 받고 있음

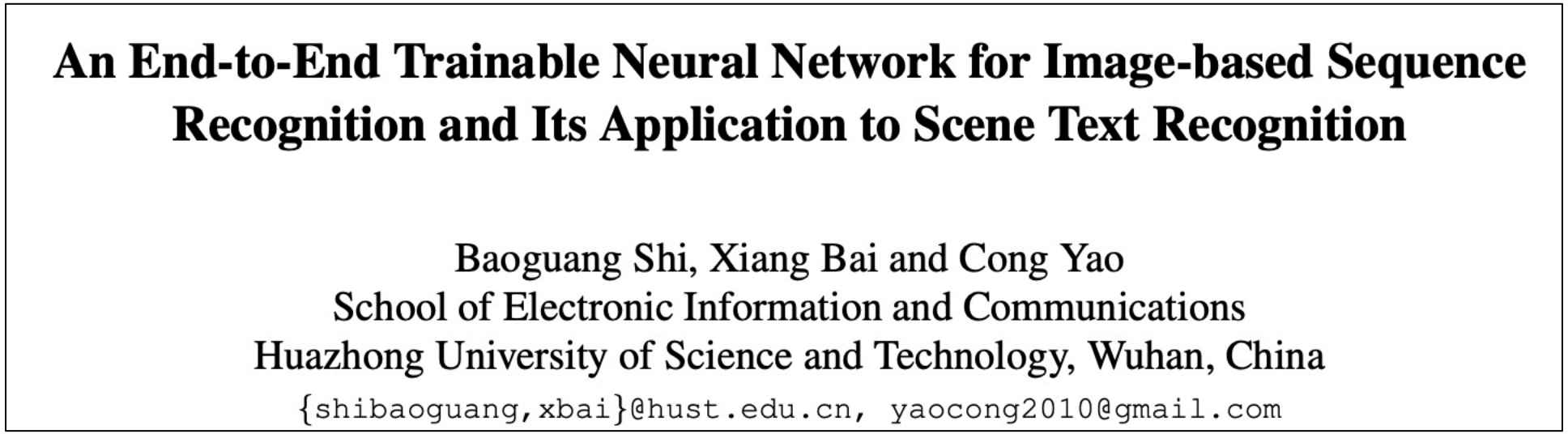


Source: https://brunch.co.kr/@kakao-it/304

Scene text recognition

❖ **An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition(Shi. et al., 2015)**

- 2021년 4월 현재 1,175회 인용
- CNN과 RNN을 결합해 Scene text recognition 문제를 해결한 초기 모델(CRNN)



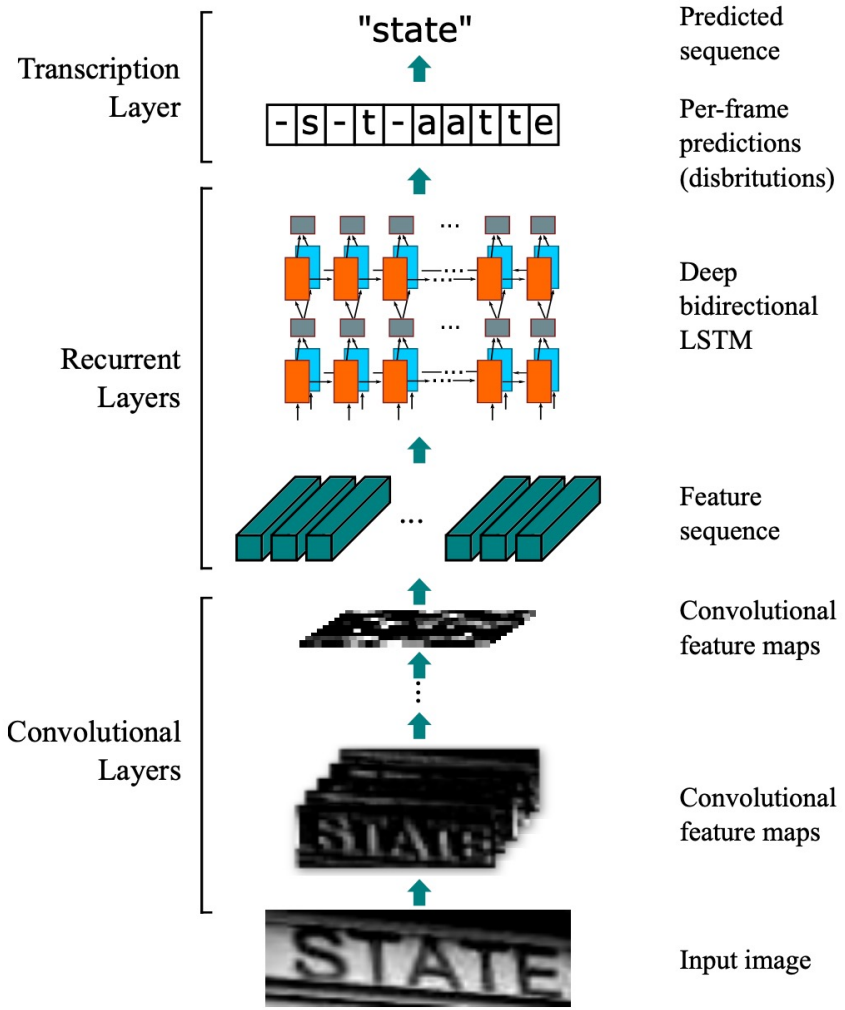
An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

Baoguang Shi, Xiang Bai and Cong Yao
School of Electronic Information and Communications
Huazhong University of Science and Technology, Wuhan, China
{shibaoguang,xbai}@hust.edu.cn, yaocong2010@gmail.com

Scene text recognition

❖ Structure of CRNN model

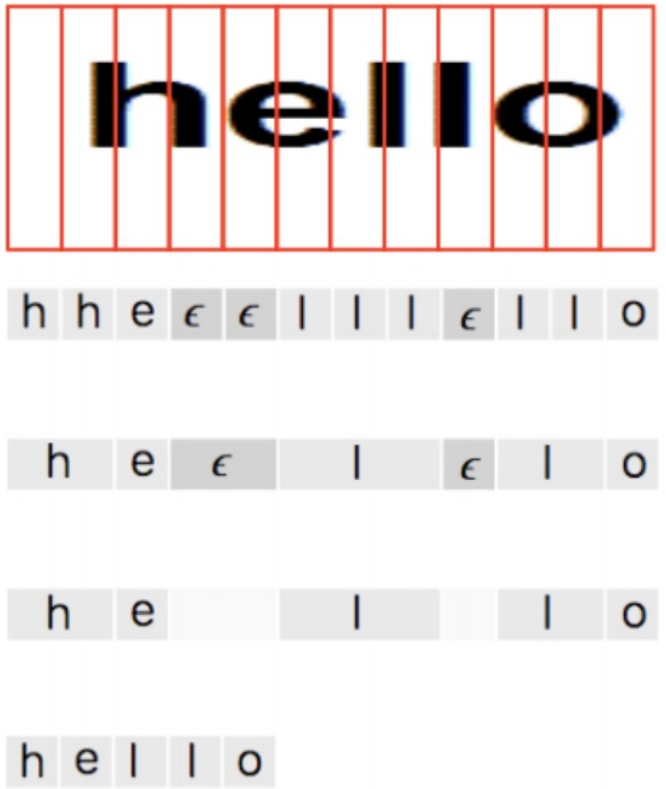
- **(Step 1)** Convolution 연산을 통해 특징 추출
- **(Step 2)** 추출된 feature map을 열 벡터 단위로 나눠 시퀀스 형태로 변환
- **(Step 3)** 길이가 긴 시퀀스 입력을 gradient vanishing 문제 없이 처리할 수 있는 bi-LSTM 모델을 이용해, 각 벡터에 대한 **글자 예측값** 출력
- **(Step 5)** CTC 알고리즘을 통해 중복 문자와 공백 등을 제거한 후 최종 **단어 예측값** 출력



Scene text recognition

❖ Connectionist Temporal Classification(CTC)

- 음성인식 모델에서 사용되는 알고리즘으로, 시퀀스 분류 모델 맨 마지막에 손실 및 그래디언트 계산 레이어 역할을 함
- Q1. '**hhe--lll-llo**'는 '**hellllo**'인가? '**hhello**'인가? '**hello**인가?'
- Q2. '**hello**'라면 몇번째 h, l을 사용할 것인가?
- 어떤 출력값을 선택/제거할 지에 대한 정보가 없는 상황에서, 정답 레이블을 만드는 모든 경우의 수에 대해 분류 확률을 최대화(Maximum Likelihood)하도록 모델을 학습함



Source: https://brunch.co.kr/@kakao-it/304

End-to-end Scene text recognition

❖ **FOTS: Fast Oriented Text Spotting with a Unified Network(Liu. et al., 2018)**

- 2021년 4월 현재 265회 인용
- EAST와 CRNN 모델을 하나로 합친 모델. 다만 각 모델을 단순히 이어 붙인 것이 아닌 한번의 특징 추출로 detection과 recognition을 수행함으로써, 연산 시간을 크게 줄임

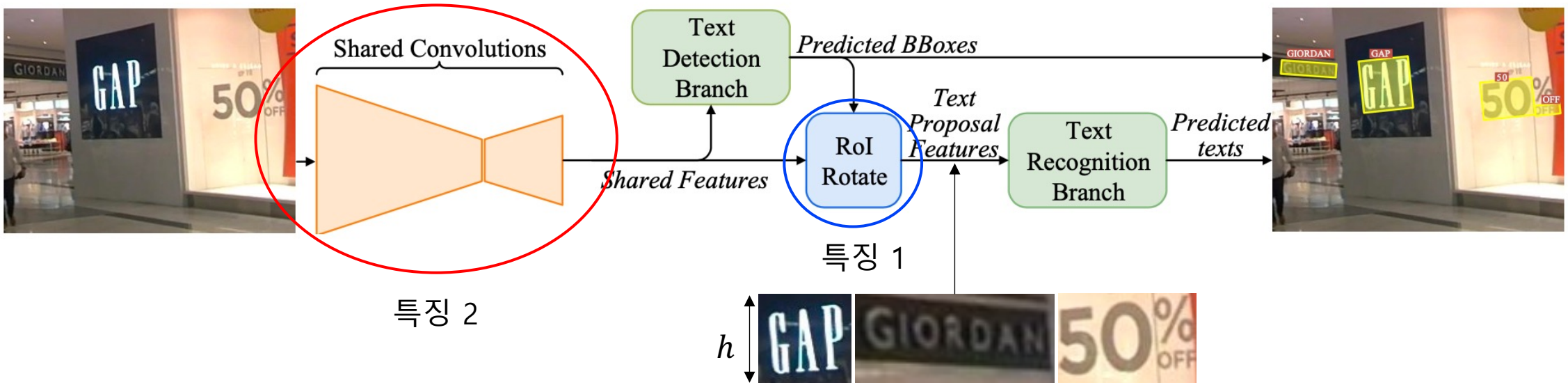
FOTS: Fast Oriented Text Spotting with a Unified Network

Xuebo Liu[1], Ding Liang[1], Shi Yan[1], Dagui Chen[1], Yu Qiao[2], and Junjie Yan[1]
[1]SenseTime Group Ltd.
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{liuxuebo,liangding,yanshi,chendagui,yanjunjie}@sensetime.com, {yu.qiao}@siat.ac.cn

End-to-end Scene text recognition

❖ Structure of FOTS model

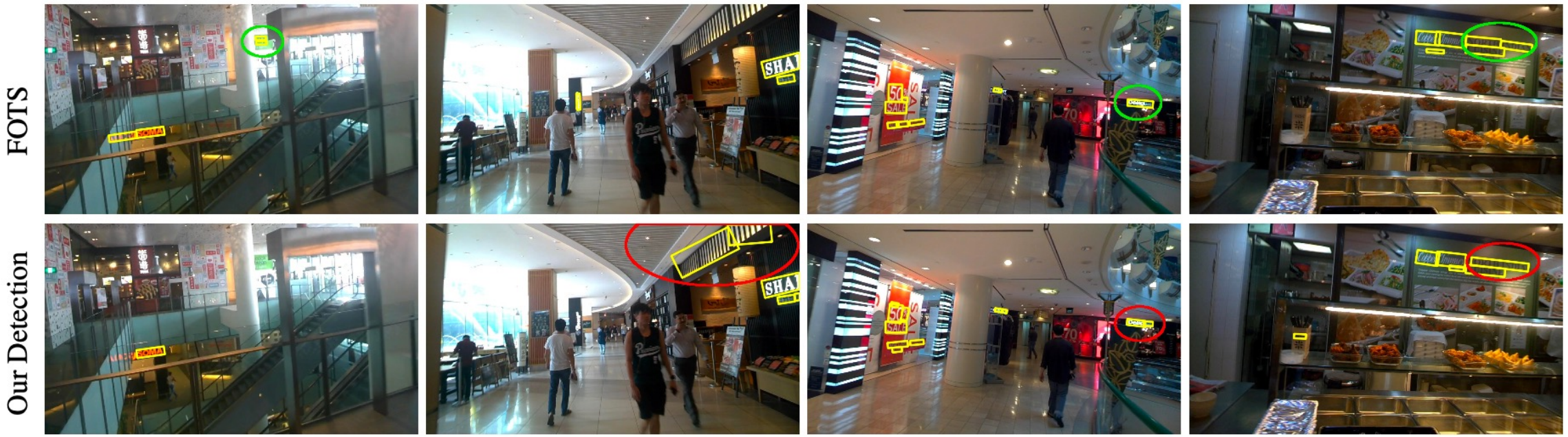
- Text detection 단과 Text recognition 단은 각 EAST, CRNN 모델과 동일함
- **ROI Rotate** 블록이 두 작업의 연결고리로, Text detection 단에서 출력된 단어 영역을 **수평으로 회전**시키고 같은 높이로 변환하여 Text recognition 단에 입력시켜 줌



End-to-end Scene text recognition

❖ Shared Convolutions

- Detection과 recognition을 End-to-end 방식으로 수행할 경우 각 작업에 쓰이는 정보를 교차로 활용할 수 있어 따로 할 때보다 성능이 올라감
- (예시) FOTS 모델의 실험 결과(위)와 FOTS에서 Text recognition 단을 제외하고 detection만 수행한 결과(아래)



End-to-end Scene text recognition

❖ **Experiment results**

- ICDAR2015에 대한 End-to-end 결과 기준 65(Generic)~87(Strong lexicon)%의 F-measure 성능을 보임
- Detection 네트워크로 EAST 모델을 사용했음에도 9%p 가까운 detection 성능 향상을 보임

Method	Detection			Method	End-to-End			Word Spotting		
	P	R	F		S	W	G	S	W	G
SegLink [43]	74.74	76.50	75.61	Baseline OpenCV3.0+Tesseract [26]	13.84	12.01	8.01	14.65	12.63	8.43
SSTD [13]	80.23	73.86	76.91	Deep2Text-MO [51, 50, 20]	16.77	16.77	16.77	17.58	17.58	17.58
WordSup [17]	79.33	77.03	78.16	Beam search CUNI+S [26]	22.14	19.80	17.46	23.37	21.07	18.38
RRPN [39]	83.52	77.13	80.20	NJU Text (Version3) [26]	32.63	-	-	34.10	-	-
EAST [53]	83.27	78.33	80.72	StradVision_v1 [26]	33.21	-	-	34.65	-	-
NLPR-CASIA [15]	82	80	81	Stradvision-2 [26]	43.70	-	-	45.87	-	-
R^2CNN [25]	85.62	79.68	82.54	TextProposals+DictNet [7, 19]	53.30	49.61	47.18	56.00	52.26	49.73
CCFLAB_FTSN [4]	88.65	80.07	84.14	HUST_MCLAB [43, 44]	67.86	-	-	70.57	-	-
Our Detection	88.84	82.04	85.31	Our Two-Stage	77.11	74.54	58.36	80.38	77.66	58.19
FOTS	91.0	85.17	87.99	FOTS	81.09	75.90	60.80	84.68	79.32	63.29
FOTS RT	85.95	79.83	82.78	FOTS RT	73.45	66.31	51.40	76.74	69.23	53.50
FOTS MS	**91.85**	**87.92**	**89.84**	FOTS MS	**83.55**	**79.11**	**65.33**	**87.01**	**82.39**	**67.97**

Conclusion

- Scene text detection and recognition은 Object detection, Semantic segmentation 등 이미지 인식 관련 다양한 문제를 접목한 연구 분야임
- Scene text detection은 이미지 내 여러 형태로 배열된 문자를 탐지할 수 있게끔 기존 직사각형 모양의 bounding box뿐 아니라 회전 사각형, 다각형 등의 box 모양의 탐지나 Segmentation을 수행하는 방식으로 발전 중임
- Scene text recognition은 CNN과 RNN을 결합한 모델이 기본 구조로 자리 잡았으며, 특히 RNN 계열의 최신 방법론을 적용하는 방향으로 발전되고 있음
- 향후 연산 시간을 줄일 수 있는 End-to-end Scene text recognition 모델이 더욱 주목 받을 것으로 예상됨

감사합니다.

Reference

[1] Long, S., He X., Yao, C. (2018). Scene Text Detection and Recognition: The Deep Learning Era.

[2] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J. (2017) EAST: An Efficient and Accurate Scene Text Detector. IEEE Conference on Computer Vision and Pattern Recognition.

[3] Shi, B., Bai, X., Yao, C. (2015). 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.

[4] Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan. J. (2018) FOTS: Fast Oriented Text Spotting with a Unified Network. IEEE Conference on Computer Vision and Pattern Recognition.