DMQA Open Seminar

Self/Semi-supervised Learning for Scene Text Recognition

2022. 12. 02.

김성수

Data Mining and Quality Analytics

Data Mining Quality Analytics

고려대학교 KOREA UNIVERSITY

발표자 소개

❖ **김성수 (Sungsu Kim)**

- 경희대학교 산업경영공학과 학부 졸업 (2022.02)
- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2022.03 ~ Present)

❖ **Research Interest**

- Self-supervised Learning
- Semi-supervised Learning
- Scene Text Recognition

❖ **Contact**

- 2022020650@korea.ac.kr

목차

Introduction

Introduction

What is Scene Text Recognition? – Overall Concept

❖ Scene Text Spotting

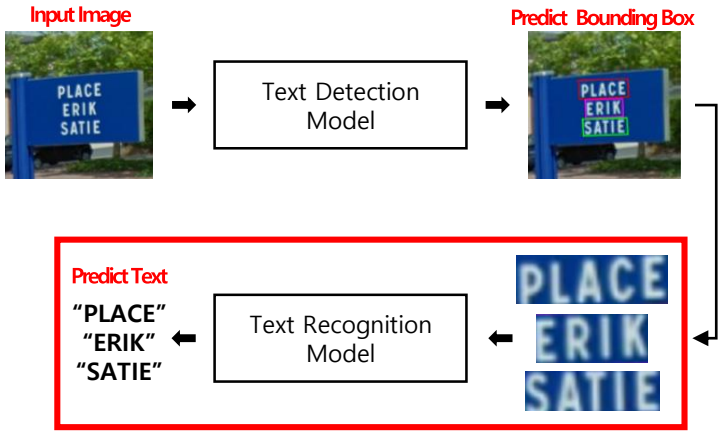
- 일상 이미지에서 글자를 검출 및 인식하는 분야
- 글자 검출모델(Scene Text Detection)과 글자 인식모델(Scene Text Recognition)로 구성

Scene Text Spotting

Detection

Recognition

Model Architecture



Introduction

What is Scene Text Recognition?: Overview

❖ Scene Text Recognition (STR)

- 글자가 존재하는 이미지 내 글자들을 인식
- 문자열이 포함된 이미지가 모델에 입력으로 들어가면 이미지에 존재하는 글자들을 출력

Overview of Scene Text Spotting

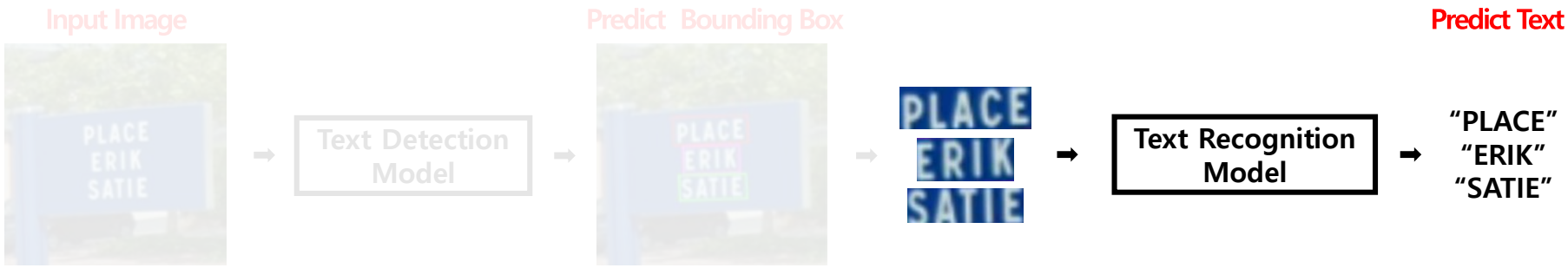
Introduction

What is Scene Text Recognition?: Overview

❖ Scene Text Recognition (STR)

- 글자가 존재하는 이미지에서 이미지 내 글자들을 인식
- 문자열이 포함된 이미지가 모델에 입력으로 들어가면 이미지에 존재하는 글자들을 출력

Overview of Scene Text Spotting



Introduction

What is Scene Text Recognition?: Overview

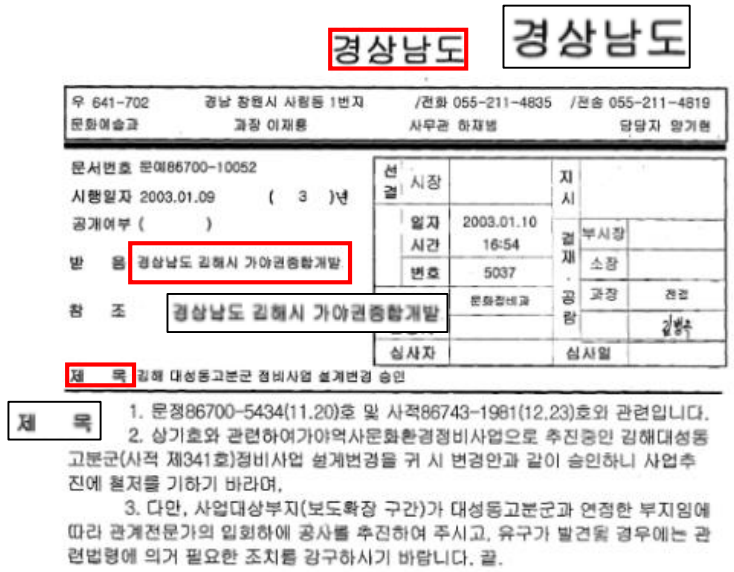
❖ Scene Text Recognition VS Optical Character Recognition (OCR)

- STR: 다양한 배경과 여러 형태의 글꼴이 존재하는 일상 이미지 내 글자를 인식하는 연구분야
- OCR: 규격화된 인쇄체 문자를 인식하는 연구분야

Scene Text Recognition

불규칙적인 이미지를 인식하기에 더욱 정교한 모델이 필요

Optical Character Recognition



Introduction

What is Scene Text Recognition?: Model Architecture

❖ Scene Text Recognition in Deep Learning: Input & Output

- 하나의 입력값에 대해 여러 개의 순차적인 출력값을 갖는 Sequence Prediction Task
  - ✓ 입력: 이미지 / 출력(Label): 글자(들)
  - ✓ Sequence Prediction Task: Scene Text Recognition, Image Captioning …
- 전체 문자열 단위로 학습하는 것이 아닌, 각 글자 단위로 학습

데이터의 전체적인 구조

입력(X)	출력(Y)
	수동골영양탕
	만두찐빵
	샤브미가
	한우리교회

Image Captioning

A man is using a mobile phone in a cafe

Introduction

What is Scene Text Recognition?: Model Architecture

❖ Scene Text Recognition in Deep Learning: Input & Output

- 하나의 입력값에 대해 여러 개의 순차적인 출력값을 갖는 Sequence Prediction Task
  - ✓ 입력: 이미지 / 출력(Label): 글자(들)
  - ✓ Sequence Prediction Task: Scene Text Recognition, Image Captioning …
- 전체 문자열 단위로 학습하는 것이 아닌, 각 글자 단위로 학습

순차적인 출력을 고려한 데이터의 구조

입력(X)	출력(Y)	순차적인 출력(Y)
	수동골영양탕	수+동+골+영+양+탕
	만두찐빵	만+두+찐+빵
	샤브미가	샤+브+미+가
	한우리교회	한+우+리+교+회

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Introduction

What is Scene Text Recognition?: Model Architecture

❖ Scene Text Recognition in Deep Learning: Input & Output

- 하나의 입력값에 대해 여러 개의 순차적인 출력값을 갖는 Sequence Prediction Task
  - ✓ 입력: 이미지 / 출력(Label): 글자(들)
  - ✓ Sequence Prediction Task: Scene Text Recognition, Image Captioning …
- 전체 문자열 단위로 학습하는 것이 아닌, 각 글자 단위로 학습

순차적인 출력을 고려한 데이터의 구조

입력(X)	출력(Y)	순차적인 출력(Y)
	수동골영양탕	수+동+골+영+양+탕
	만두찐빵	만+두+찐+빵
	샤브미가	샤+브+미+가
	한우리교회	한+우+리+교+회

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

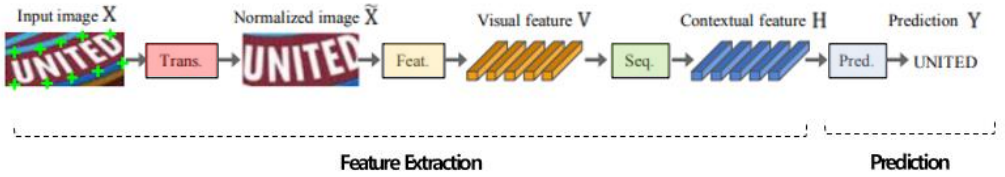
Introduction

What is Scene Text Recognition?: Model Architecture

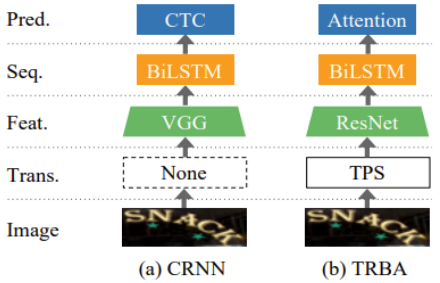
❖ Scene Text Recognition in Deep Learning: Architecture

- Scene Text Recognition은 크게 4단계로 구성
  - ① Transformation: 입력된 이미지를 올바른 형태로 정렬 (Affine, TPS)
  - ② Feature Extraction: 정렬된 이미지에서 Visual Feature를 추출 (ResNet, VGGNet)
  - ③ Sequence Modeling: Visual Feature를 Context Feature로 변환 (BiLSTM)
  - ④ Prediction: Context Feature를 통해 이미지 내 글자들을 예측 (CTC, Attention)

Overview of Scene Text Recognition Model

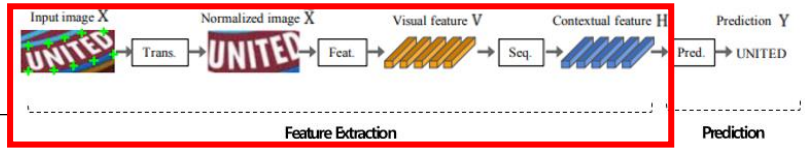


Examples for STR Model



* Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., ... & Lee, H. (2019), What is wrong with scene text recognition model comparisons? dataset and model analysis, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Introduction



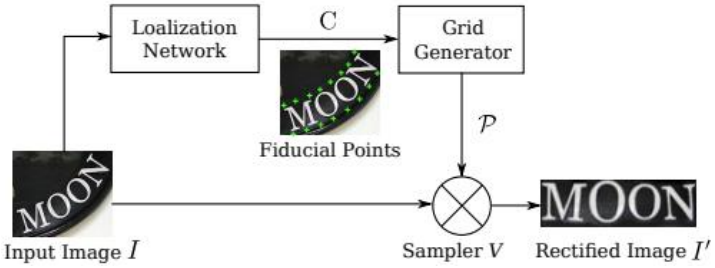
What is Scene Text Recognition?: Model Architecture

❖ Architecture (1): Encoder

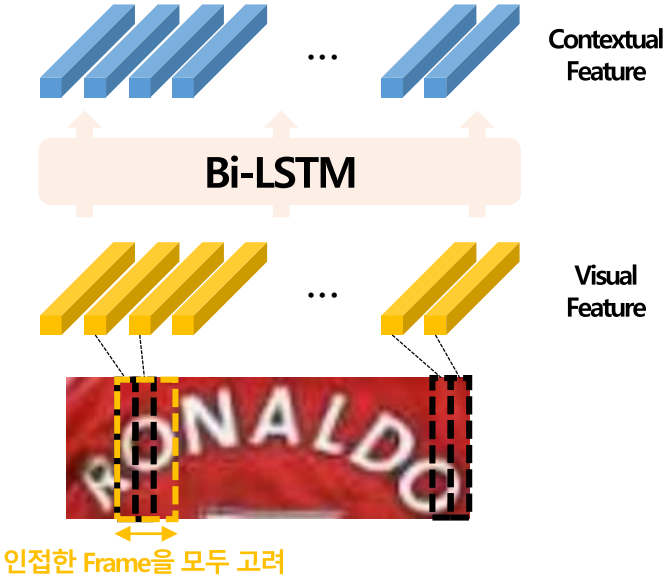
- Scene Text Recognition은 크게 4단계로 구성
  - ① Transformation: 입력된 이미지를 올바른 형태로 정렬 (Affine, TPS)
  - ② Feature Extraction: 입력된 이미지에서 Visual Feature를 추출 (ResNet, VGGNet)
  - ③ Sequence Modeling: Visual Feature를 Context Feature로 변환 (BiLSTM)
  - ④ Prediction: Context Feature를 통해 이미지 내 글자들을 예측 (CTC, Attention)

Encoder

Transformation

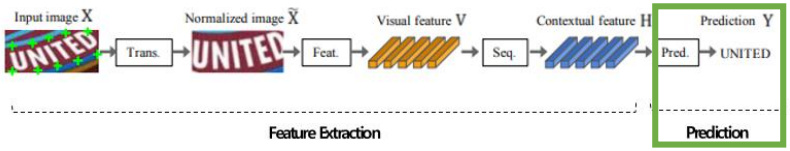


Feature Extraction & Sequence Modeling



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Introduction

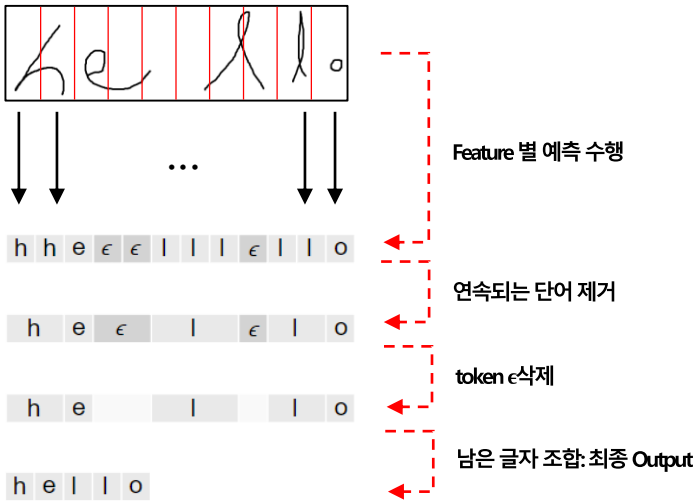


What is Scene Text Recognition?: Model Architecture

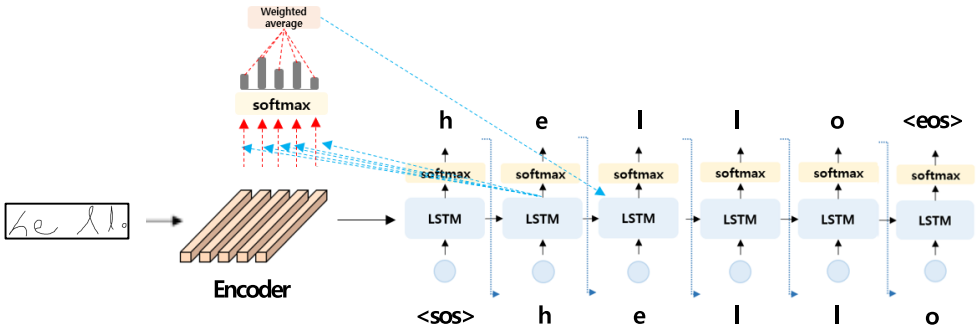
❖ Architecture (2): Decoder

- Scene Text Recognition은 크게 4단계로 구성
  - ① Transformation: 입력된 이미지를 올바른 형태로 정렬 (Affine, TPS)
  - ② Feature Extraction: 입력된 이미지에서 Visual Feature를 추출 (ResNet, VGGNet)
  - ③ Sequence Modeling: Visual Feature를 Context Feature로 변환 (BiLSTM)
  - ④ Prediction: Context Feature를 통해 이미지 내 글자들을 예측 (CTC, Attention) — **Decoder**

CTC 기반 Decoder



Attention 기반 Decoder



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

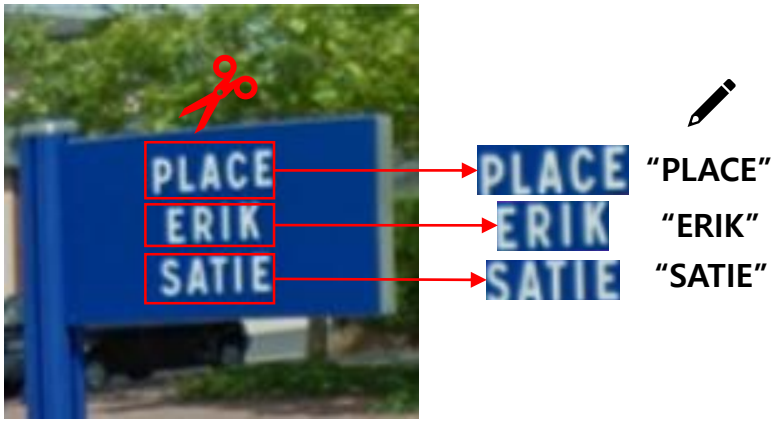
Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Limitation of STR: Insufficient Labeled Data

- Scene Text Recognition은 레이블링 비용이 크다는 한계가 존재
  - ✓ 이미지에서 글자가 있는 위치를 식별(Detection)한 후, 해당 글자가 무엇인지 표기(Labeling)하는 2단계 절차
- 각기 다른 언어에 대한 데이터가 필요하기에, 소수 언어의 경우 데이터 수집이 어려움

STR 데이터 레이블링 과정



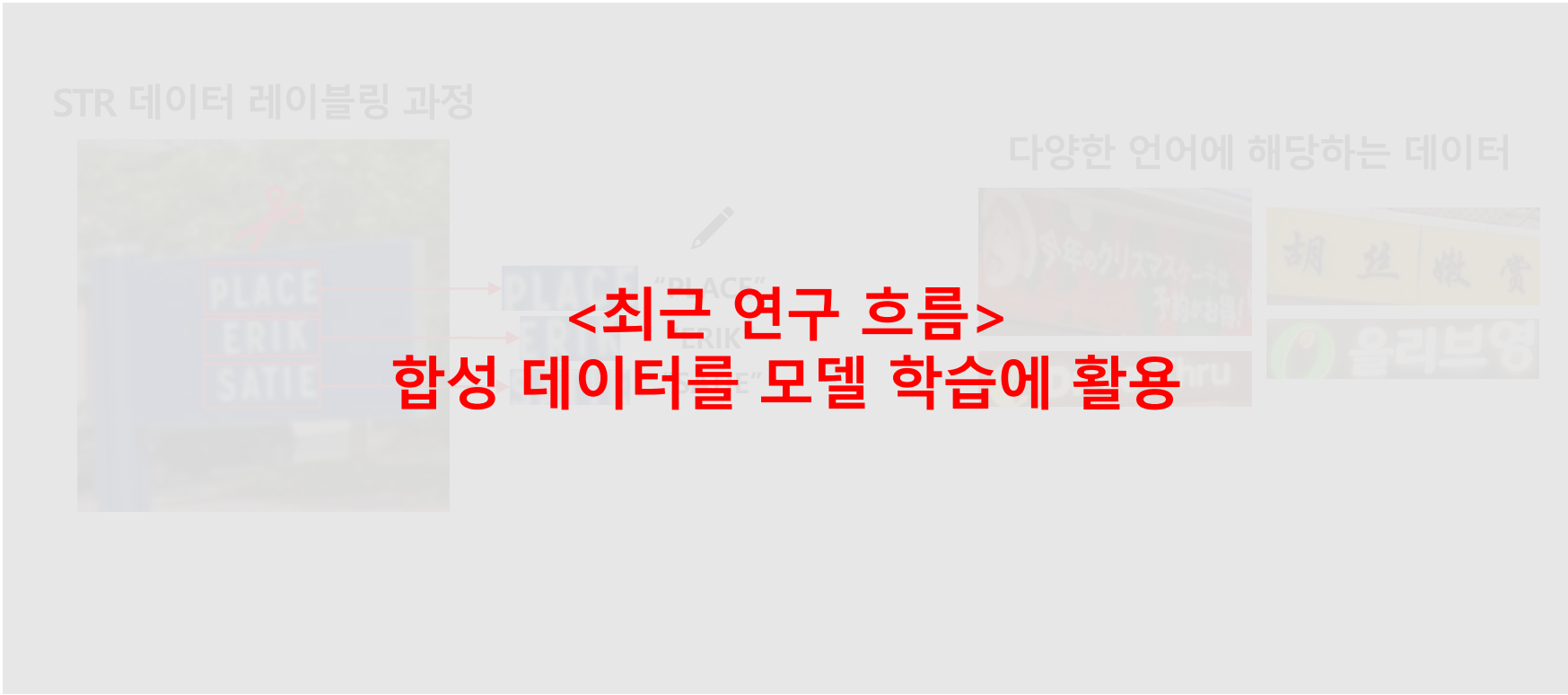
다양한 언어에 해당하는 데이터

Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Limitation of STR: Insufficient Labeled Data

- Scene Text Recognition은 레이블링 비용이 크다는 한계가 존재
  - ✓ 이미지에서 글자가 있는 위치를 식별(Detection)한 후, 해당 글자가 무엇인지 표기(Labeling)하는 2단계 절차
- 각기 다른 언어에 대한 데이터가 필요하기에, 소수 언어의 경우 데이터 수집이 어려움



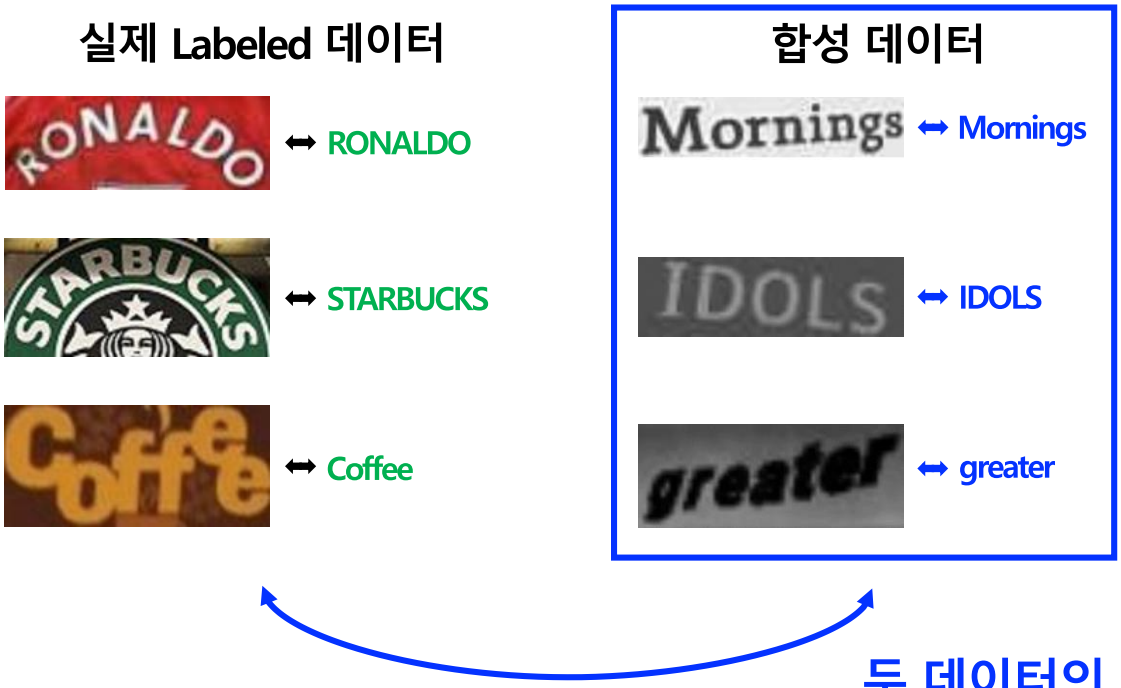
<최근 연구 흐름>
합성 데이터를 모델 학습에 활용

Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Synthetic Data

- 합성 데이터: 인간이 규칙에 기반하여 인위적으로 만들어 낸 데이터
  - ✓ 무한하게 만들 수 있지만, 실제 이미지와 비교했을 때 부자연스러움
- 평가 데이터와 분포가 다른 합성 데이터로 학습하기에, 일반화 성능이 저하될 수 있음

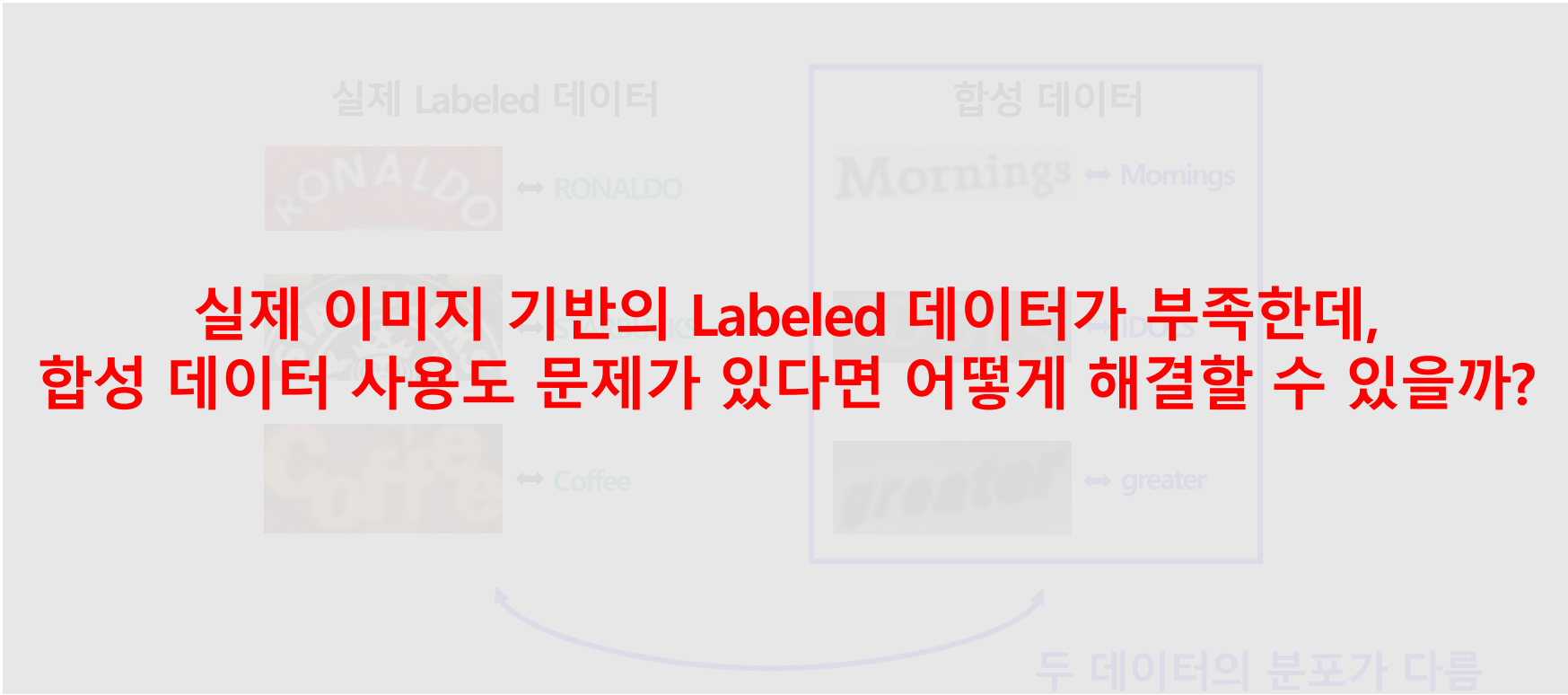


Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Synthetic Data

- 합성 데이터: 인간이 규칙에 기반하여 인위적으로 만들어 낸 데이터
  - ✓ 무한하게 만들 수 있지만, 실제 이미지와 비교했을 때 부자연스러움
- 평가 데이터와 분포가 다른 합성 데이터로 학습하기에, 일반화 성능이 저하될 수 있음



실제 이미지 기반의 Labeled 데이터가 부족한데,
합성 데이터 사용도 문제가 있다면 어떻게 해결할 수 있을까?

Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Unlabeled Data

- 데이터 레이블링 소요가 적기에, 데이터 수집비용이 적음
- 실제 이미지를 기반으로 하기에 일반화 성능 저하에 대한 우려도 낮음

실제 Labeled 데이터

↔ RONALDO

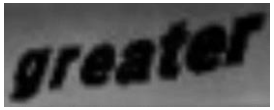
↔ STARBUCKS

↔ Coffee

합성 데이터

↔ Mornings

↔ IDOLS



↔ greater

실제 Unlabeled 데이터


↔ No label


↔ No label

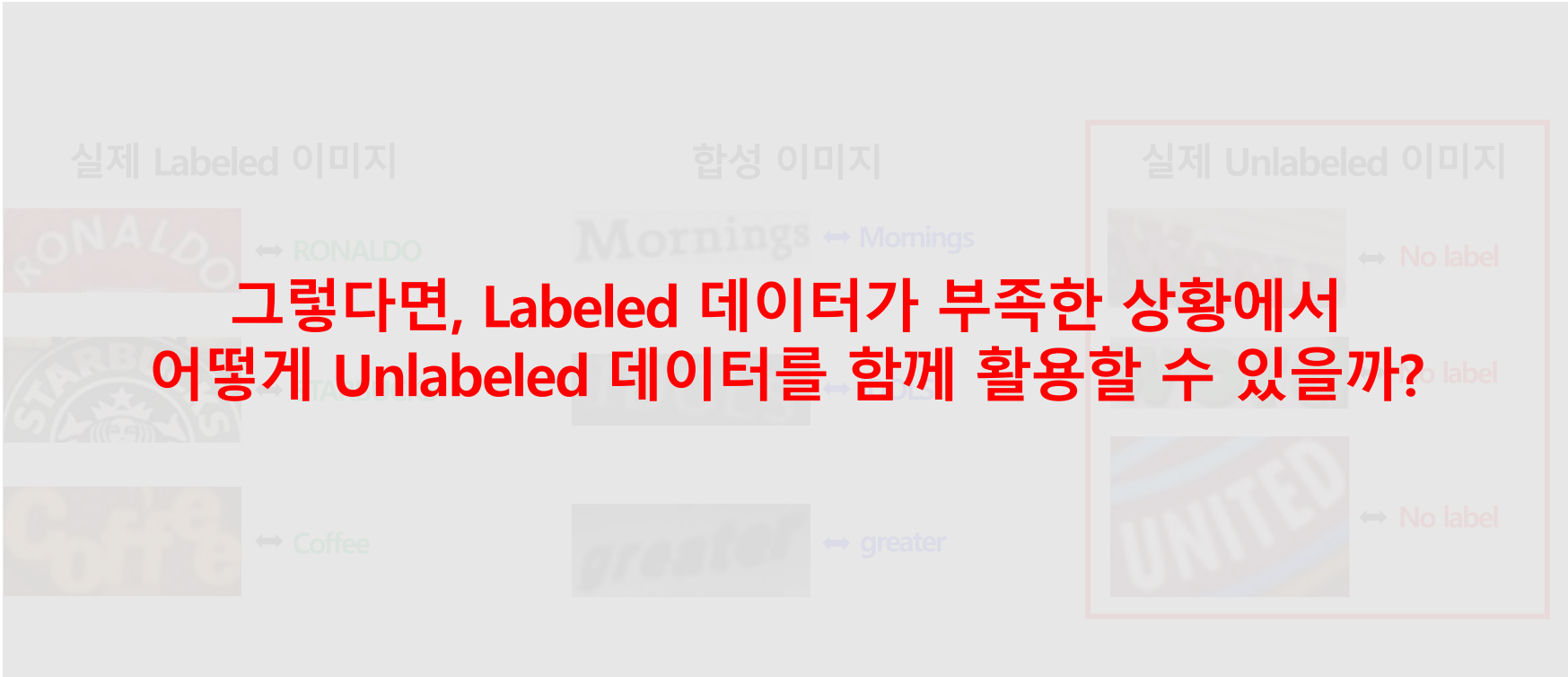

↔ No label

Introduction

What is Scene Text Recognition?: Limitation of Previous Studies

❖ Unlabeled Data

- 데이터 레이블링 소요가 적기에, 데이터 수집비용이 적음
- 실제 이미지를 기반으로 하기에 일반화 성능 저하에 대한 우려도 낮음



Introduction

What is Self/Semi-supervised Learning?: Overview

❖ Methodology using Unlabeled Data

- 소수의 Labeled 데이터만 존재할 때, Unlabeled 데이터를 함께 활용하여 Labeled 데이터가 부족한 한계를 극복
  - ✓ Self-supervised Learning
  - ✓ Semi-supervised Learning

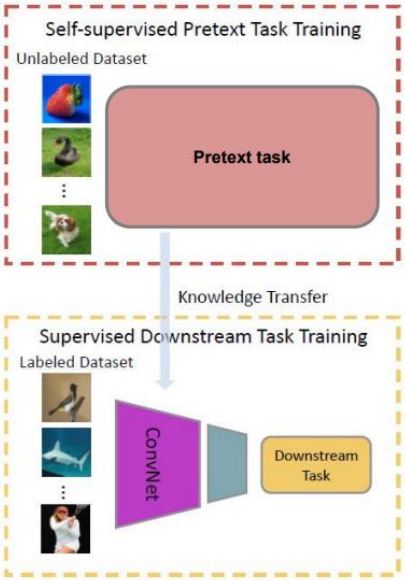
Introduction

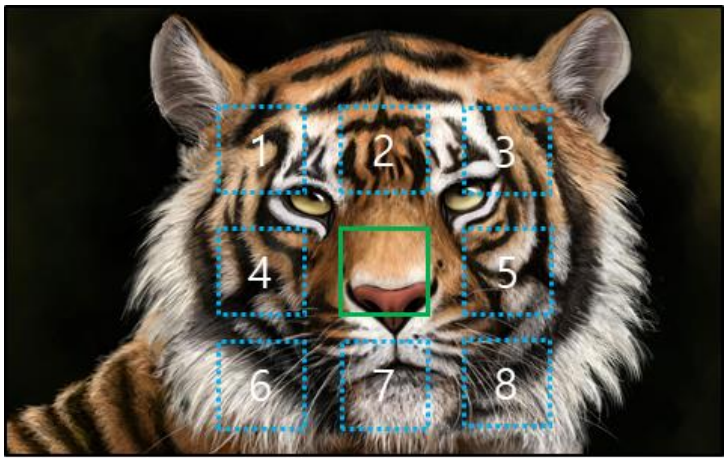
What is Self/Semi-supervised Learning?: Self-supervised Learning

❖ Self-supervised Learning

- 입력 데이터를 변형하여 기존의 입력 데이터에 Supervision(지도)을 주는 방식으로 데이터의 특징을 학습
  - ✓ Stage1: Unlabeled 데이터로 Pretraining
  - ✓ Stage2: Labeled 데이터로 Fine-tuning

대표적인 자기지도학습 Framework (Context Prediction, 2015)





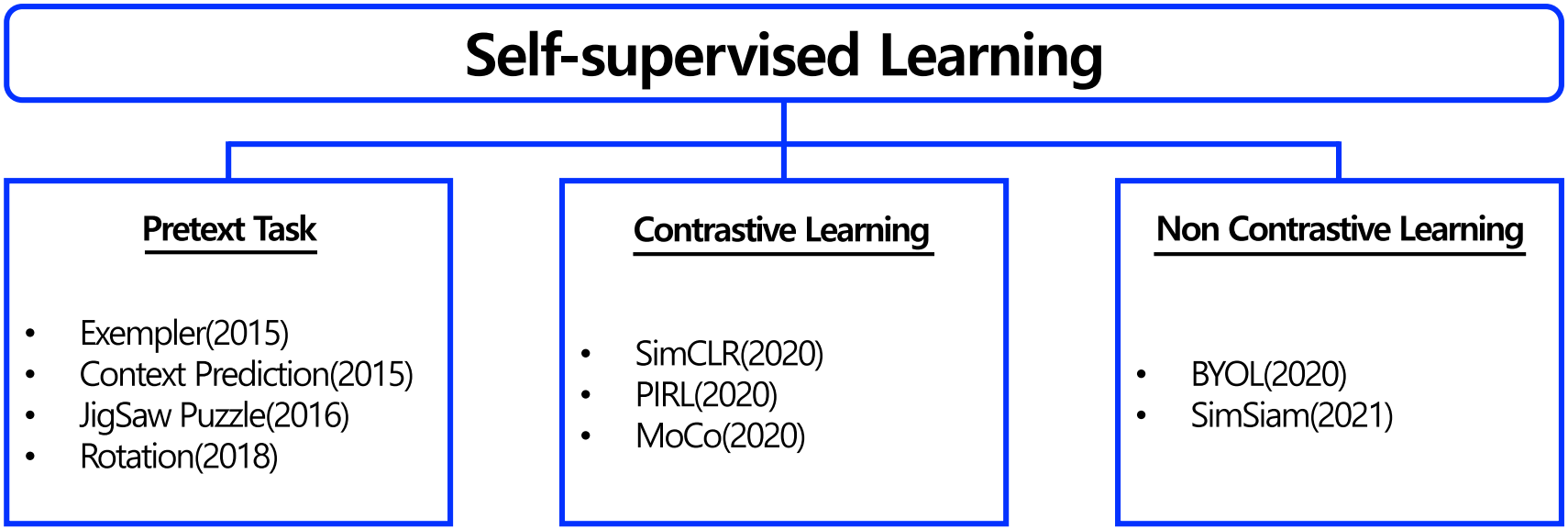
X = (,) → Y = 7

Introduction

What is Self/Semi-supervised Learning?: Self-supervised Learning

❖ Self-supervised Learning

- Pretext Task: 새로운 Self-supervision 문제를 정의하여 이를 예측함으로써 특징을 학습
- Contrastive Learning: 주어진 입력 데이터에 대해 Positive/Negative Pair를 정의한 후 데이터 간 관계를 통해 특징을 학습
- Non Contrastive Learning: Negative Sample을 정의하지 않고, Positive Pair만으로 학습

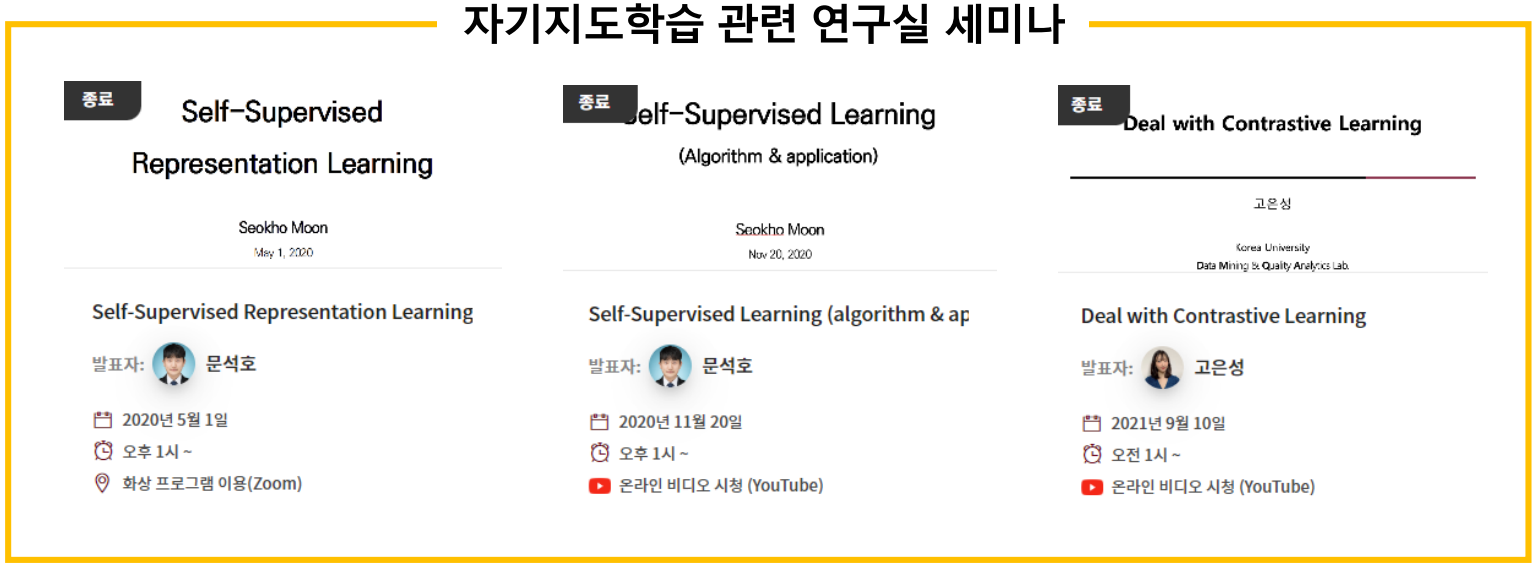


Introduction

What is Self/Semi-supervised Learning?: Self-supervised Learning

❖ Self-supervised Learning

- Pretext Task: 새로운 Self-supervision 문제를 정의하여 이를 예측함으로써 특징을 학습
- Contrastive Learning: 주어진 입력 데이터에 대해 Positive/Negative Pair를 정의한 후 데이터 간 관계를 통해 특징 학습
- Non Contrastive Learning: Negative Sample을 정의하지 않고, Positive Pair만으로 학습



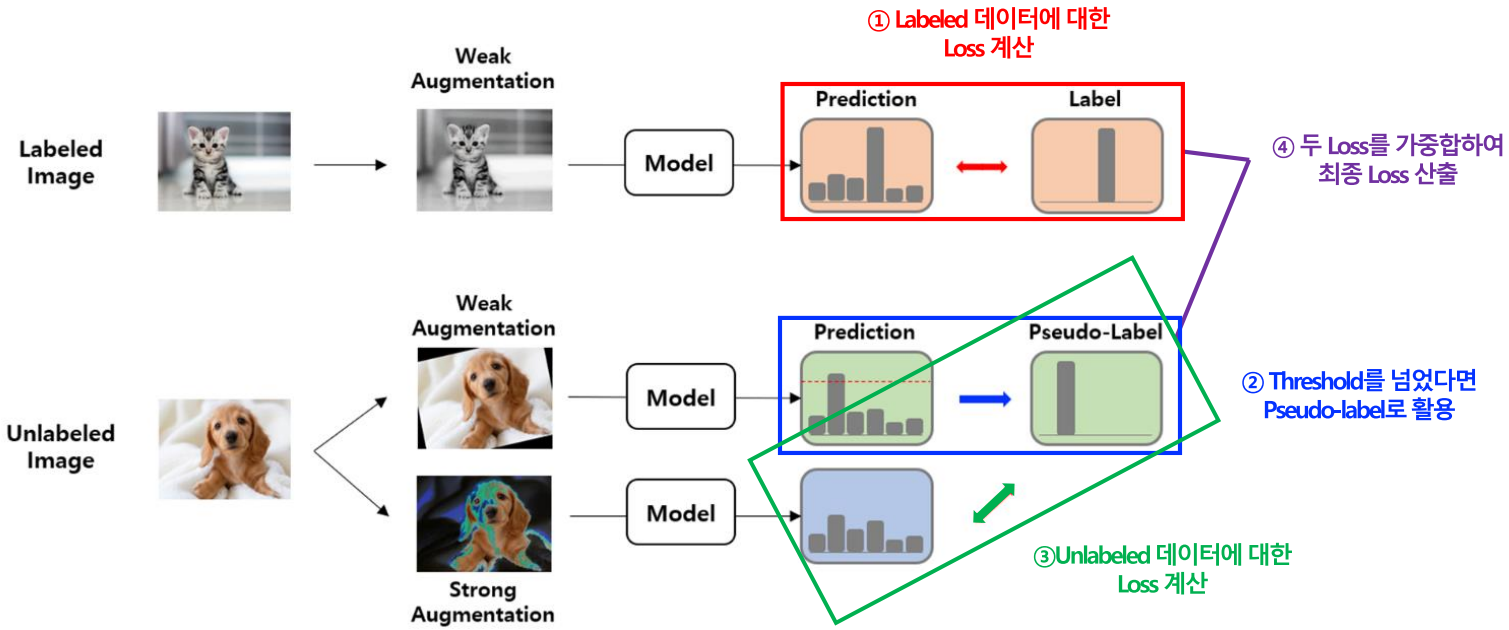
Introduction

What is Self/Semi-supervised Learning?: Semi-supervised Learning

❖ Semi-supervised Learning

- Label이 없는 데이터에 대해 모델의 예측결과로 Label을 임의로 만들어주어 학습
- Labeled 데이터와 Unlabeled 데이터를 함께 활용하여 한 번에 학습

대표적인 준지도학습 방법론 (FixMatch, 2020)

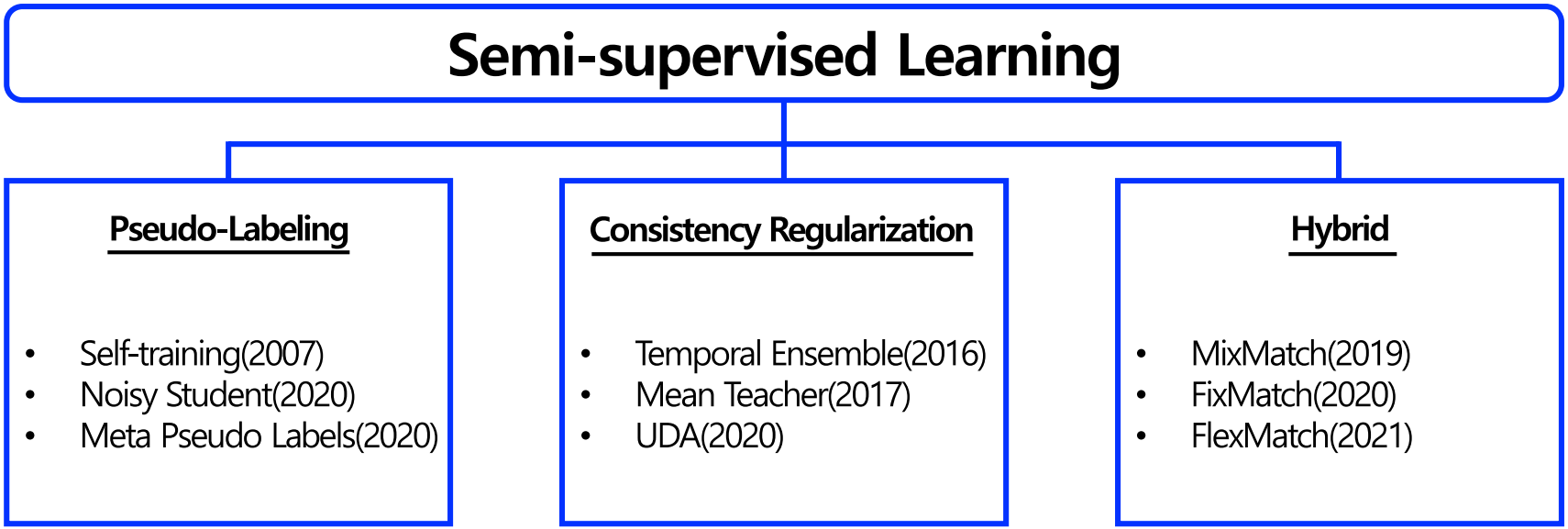


Introduction

What is Self/Semi-supervised Learning?: Semi-supervised Learning

❖ Semi-supervised Learning

- Pseudo-Labeling Method: Unlabeled 데이터를 예측결과를 활용하여 가짜로 레이블링 후, Labeled 데이터처럼 활용
- Consistency Regularization Method: 데이터 및 모델에 변형을 가해도 예측에 대한 일관성을 갖도록 학습
- Hybrid Method: 여러 준지도학습 알고리즘의 아이디어를 혼합하여 활용하여 학습



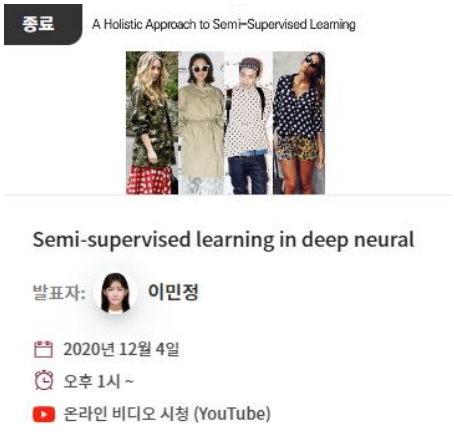
Introduction

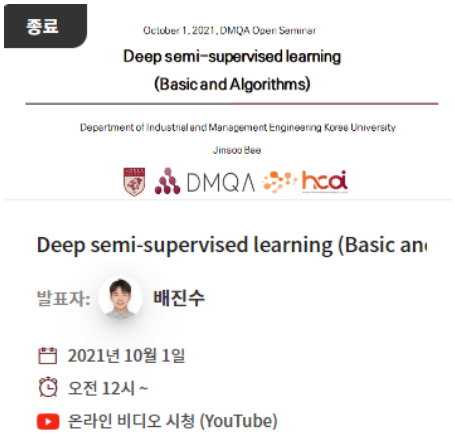
What is Self/Semi-supervised Learning?: Semi-supervised Learning

❖ Semi-supervised Learning

- Pseudo-Labeling Method: 예측결과를 활용하여 Unlabeled 데이터를 가짜로 레이블링 후, Labeled 데이터처럼 활용
- Consistency Regularization Method: 데이터 및 모델에 변형을 가해도 예측에 대한 일관성을 갖도록 학습
- Hybrid Method: 여러 준지도학습 알고리즘의 아이디어를 혼합하여 활용하는 방법론

준지도학습 관련 연구실 세미나





Algorithms

Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY

Algorithms

Overview

❖ Overview of Algorithms

- Labeled 데이터가 부족한 상황에서, Unlabeled 데이터를 함께 활용하여 모델을 학습
- Self/Semi-supervised Learning을 Scene Text Recognition에 적용

① Self-supervised Learning-based Scene Text Recognition

** Scene Text Recognition보다는, Hand written Text Recognition에 적용*

② Semi-supervised Learning-based Scene Text Recognition

③ Self&Semi-supervised Learning-based Scene Text Recognition

Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Self-supervised Learning-based Scene Text Recognition

- Text Recognition에 Self-supervised Learning의 Contrastive Learning을 적용
- SeqCLR: Sequence-to-Sequence Contrastive LeaRning

Sequence-to-Sequence Contrastive Learning for Text Recognition

Aviad Aberdam*
Technion
aaberdam@cs.technion.ac.il

Ron Litman*
AWS
litmanr@amazon.com

Shahar Tsiper
AWS
tsiper@amazon.com

Oron Anschel
AWS
oronans@amazon.com

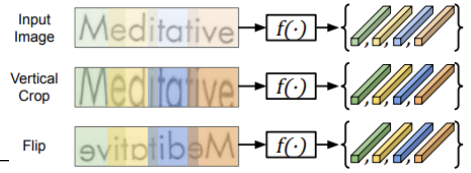
Ron Slossberg
Technion
ronslos@cs.technion.ac.il

Shai Mazor
AWS
smazor@amazon.com

R. Manmatha
AWS
manmatha@amazon.com

Pietro Perona
Caltech and AWS
peronapp@amazon.com

Algorithms

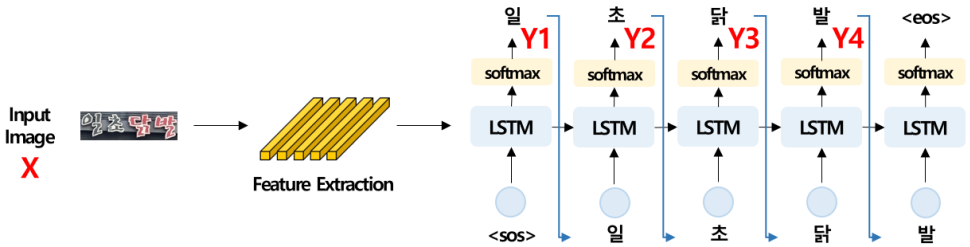


Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

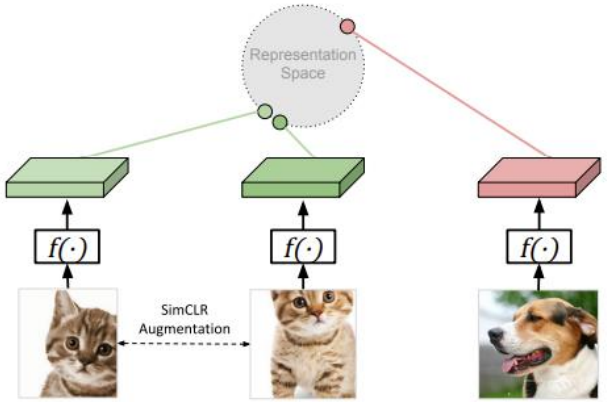
❖ Overview & Background

- 문자인식에 Unlabeled 데이터를 함께 활용할 수 있는 자기지도학습 Framework 제안
- 자기지도학습 중 Contrastive Learning을 활용
  - ✓ Contrastive Learning: Positive Pair는 유사하도록, Negative Pair는 상이하도록 학습
- 일반적인 자기지도학습(SimCLR 등)을 STR에 적용한다면, 아래와 같은 한계가 존재
  - ✓ 기존 RandAugment기반 Data Augmentation은 Sequence를 해칠 수 있음
  - ✓ 문자인식 모델이 지니는 Sequential한 특징을 반영하기 어려움

Text Recognition의 Sequential한 특징



Contrastive Learning (SimCLR)



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

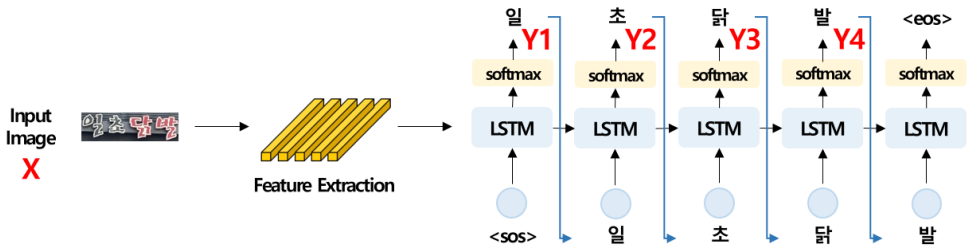
Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Text Recognition + Contrastive Learning?

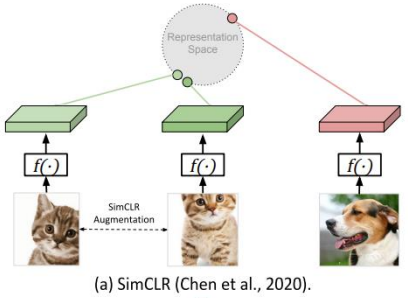
- STR은 1개의 입력값에 대해 여러 개의 출력값을 갖는 구조
- 출력값에 Sequence가 존재하는 데이터 형태

→ 일반적인 이미지 분류문제와 다르게 여러 개의 Sequential한 출력값을 반영해야 함

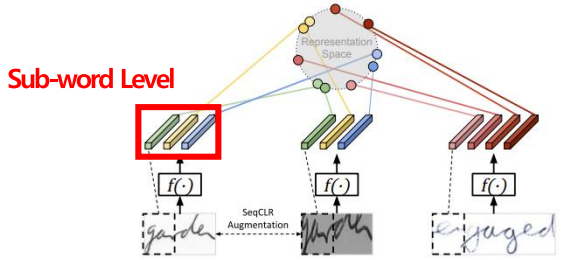
Text Recognition의 특징



Contrastive Learning



(a) SimCLR (Chen et al., 2020).



(b) Sequence-level Contrastive Learning (SeqCLR).

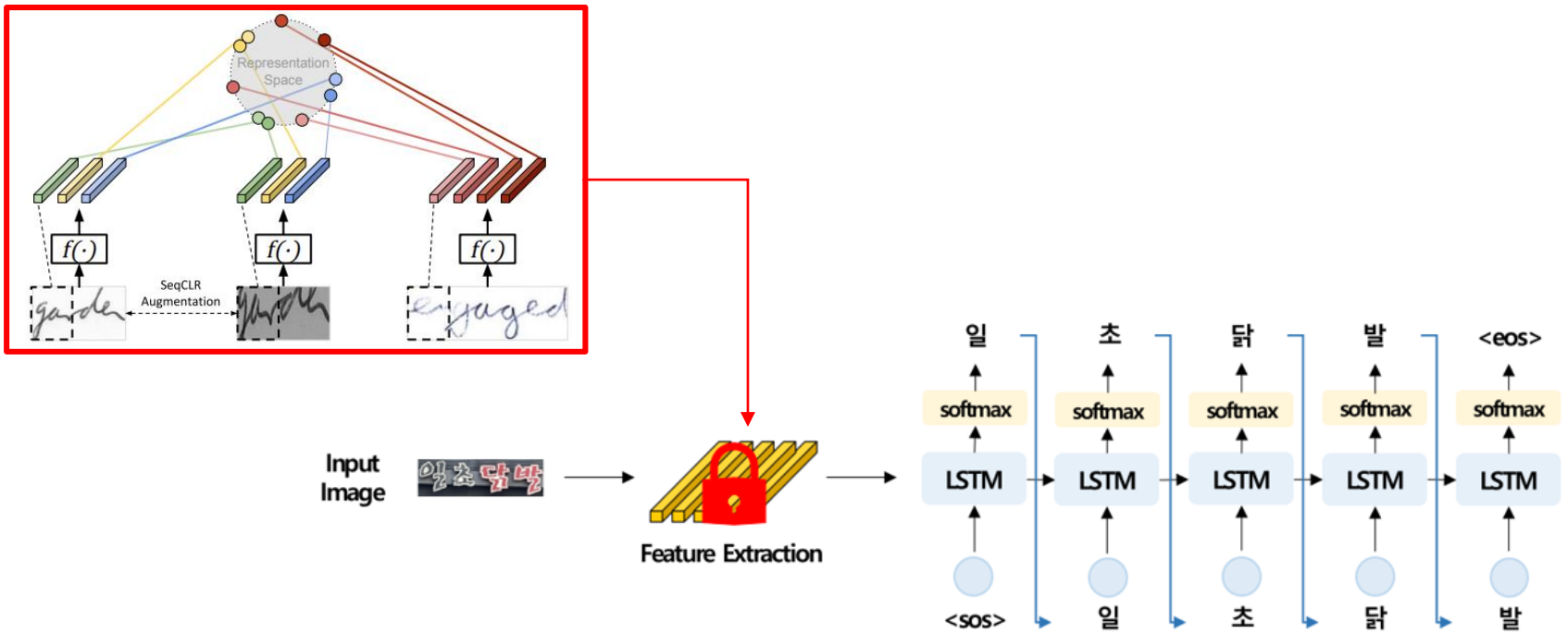
Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Overview of Model Architecture

- Phase (1): Pretraining (Contrastive Learning) – Unlabeled 데이터 활용
- Phase (2): Fine-tuning – Labeled 데이터 활용

Overview of Model Architecture

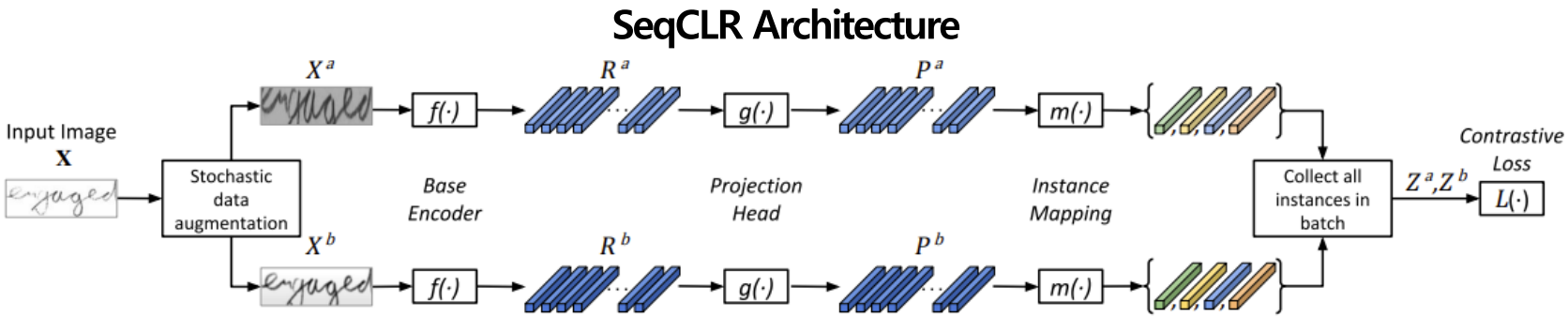


Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

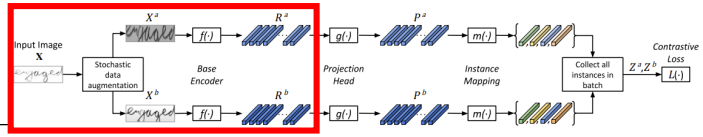
❖ Overview of Phase1: Contrastive Learning

- Data Augmentation
- Base Encoder
- Projection Head
- Instance Mapping Function
- Contrastive Loss



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

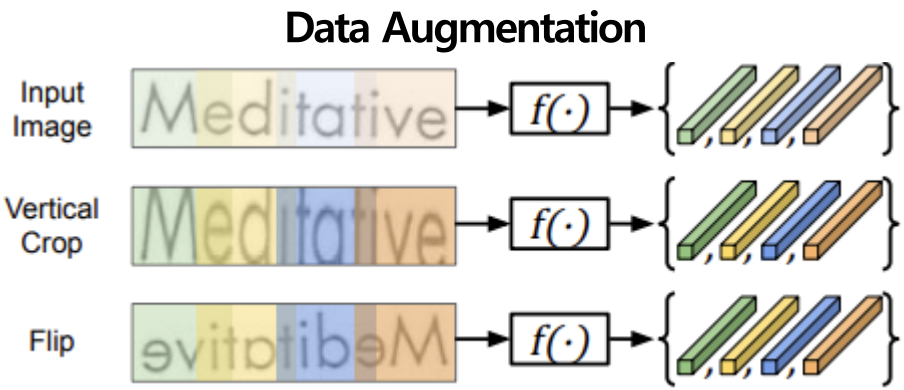


Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Model Architecture (1)

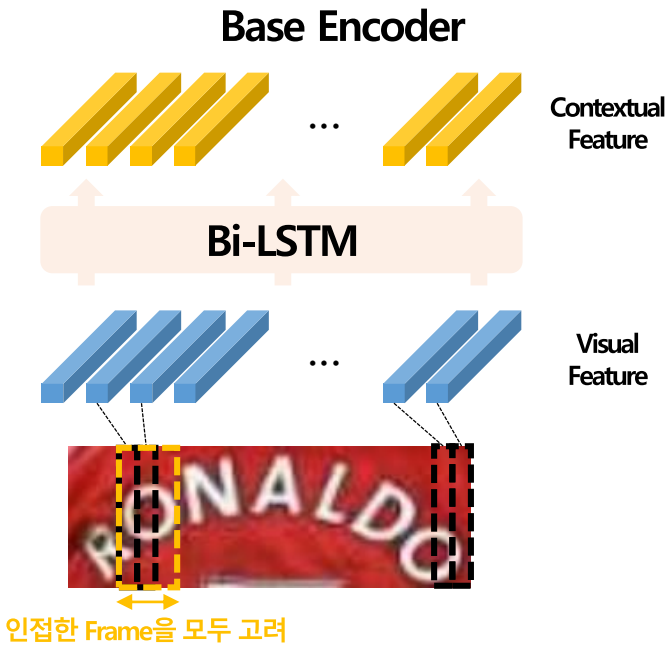
- 문자열의 순서를 훼손하지 않는 Data Augmentation을 2회 수행
- Base Encoder에서 Visual Feature 추출 후 Contextual Feature로 변환
  - ✓ CNN + Bi-LSTM

VGGNet, ResNet …



한계① 극복

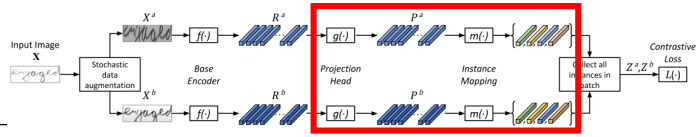
Sequence를 해치지 않는 Data Augmentation 전략 제안



인접한 Frame을 모두 고려

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

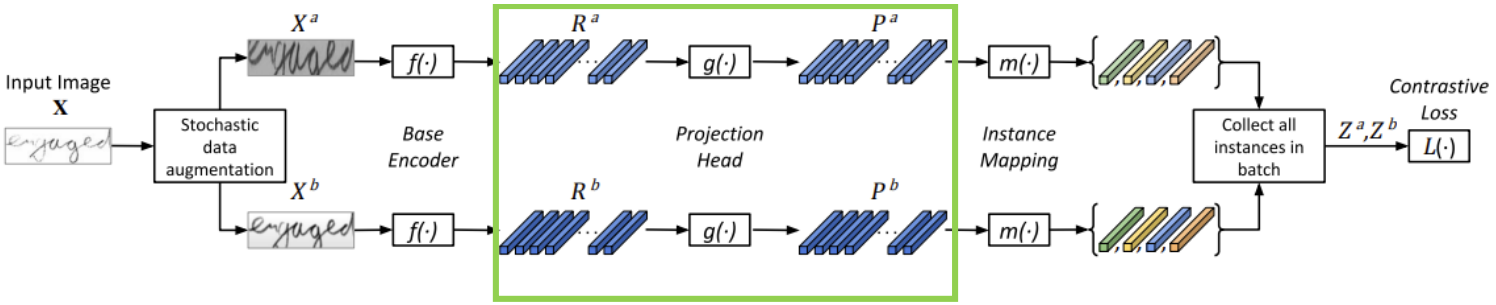


Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

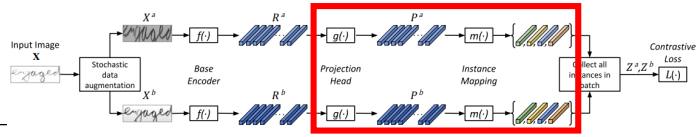
❖ Model Architecture (2)

- 각 Frame별로 Representation의 Quality를 향상 시키기 위해 Projection Head를 거침
  - ✓ None, MLP, Bi-LSTM
- Projection Head를 거친 각 Frame들을 Instance Mapping Function를 통해 Contrastive Loss를 연산하기 위한 Instance로 변환

SeqCLR의 개요



Algorithms

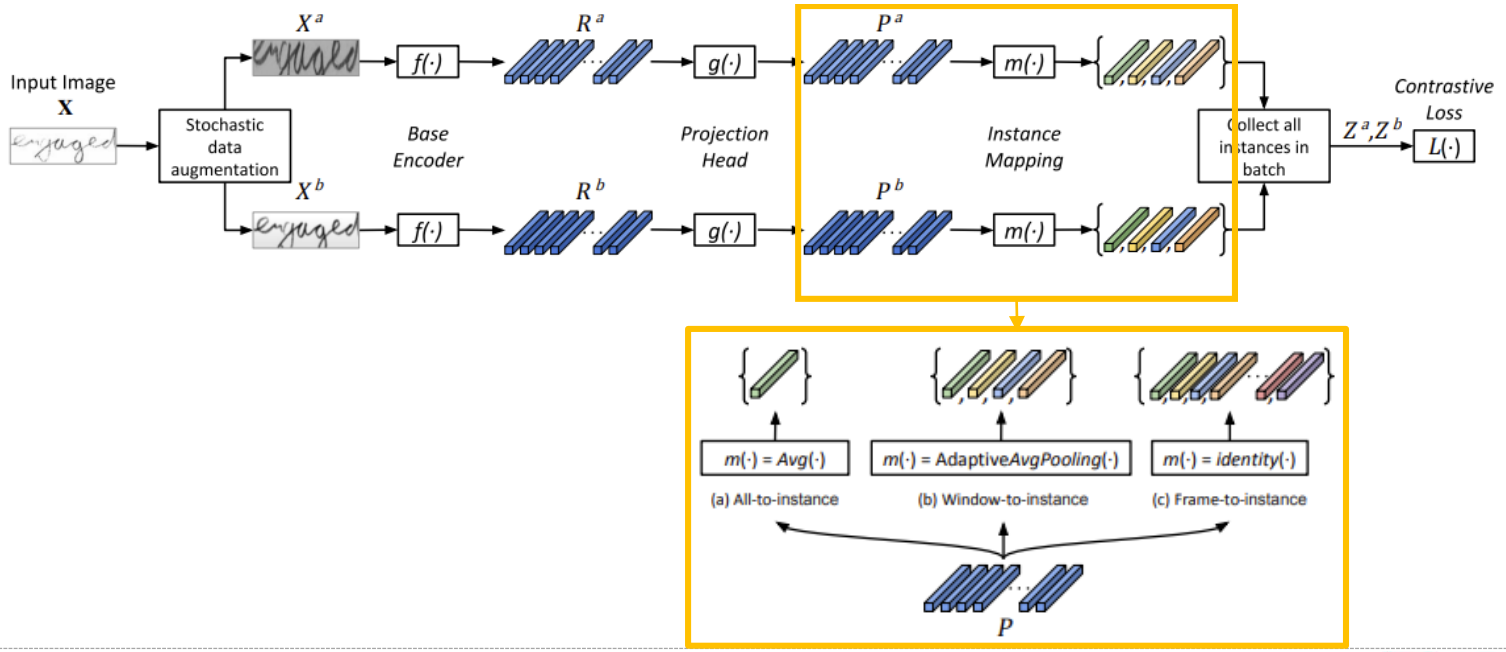


Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Model Architecture (2)

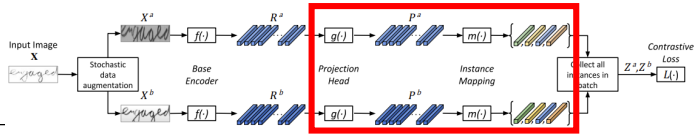
- 각 Frame별로 Representation의 Quality를 향상 시키기 위해 Projection Head를 거침
  - ✓ None, MLP, BiLSTM
- Projection Head를 거친 각 Frame들을 Instance Mapping Function를 통해 Contrastive Loss를 연산하기 위한 Instance로 변환

SeqCLR의 개요



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

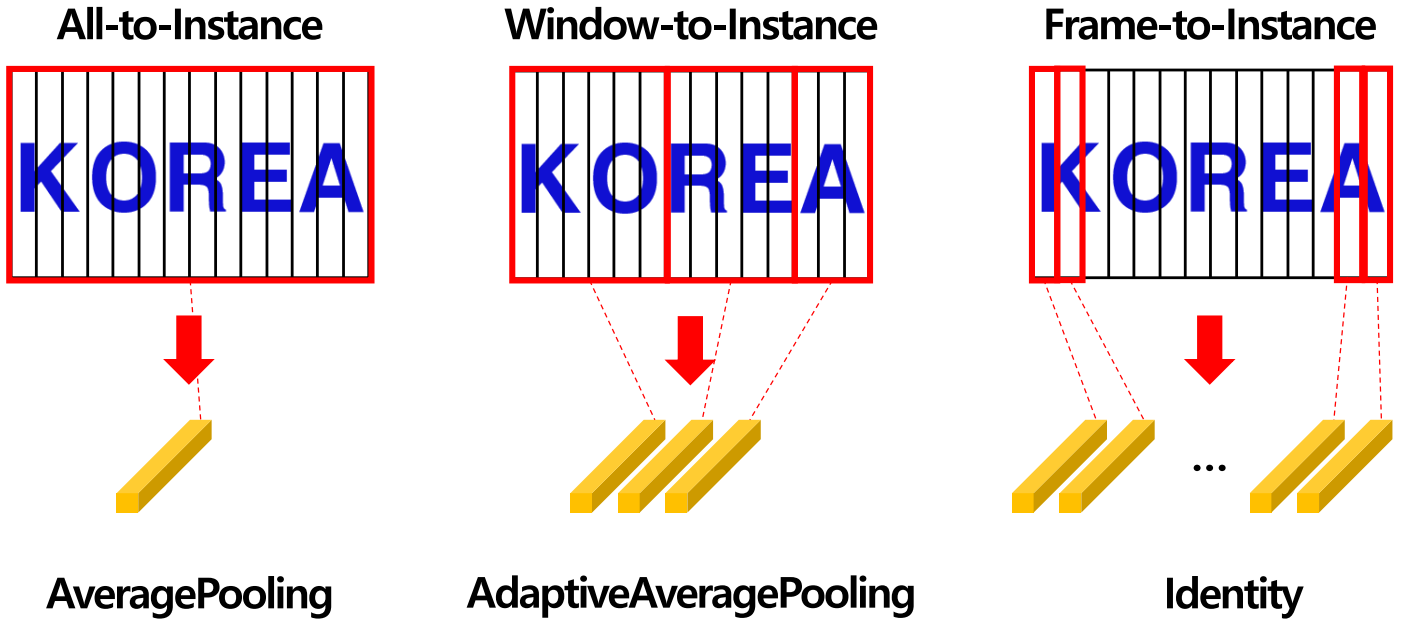
Algorithms



Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

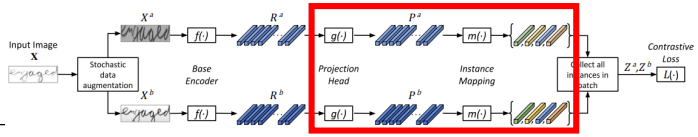
❖ Model Architecture (2)

- 각 Frame별로 Representation의 Quality를 향상 시키기 위해 Projection Head를 거침
  - ✓ None, MLP, BiLSTM
- Projection Head를 거친 각 Frame들을 Instance Mapping Function를 통해 Contrastive Loss를 연산하기 위한 Instance로 변환



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms



Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Model Architecture (2)

- 각 Frame별로 Representation의 Quality를 향상 시키기 위해 Projection Head를 거침
  - ✓ None, MLP, BiLSTM
- Projection Head를 거친 각 Frame들을 Instance Mapping Function를 통해 Contrastive Loss를 연산하기 위한 Instance로 변환

All-to-Instance | Window-to-Instance | Frame-to-Instance

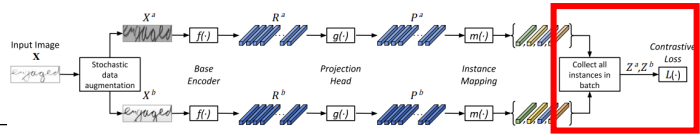
한계② 극복

Instance Mapping Function을 통해 문자 인식 모델이 지니는 Sequential 한 특징을 반영

AveragePooling | AdaptiveAveragePooling | Identity

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms



Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Model Architecture (3)

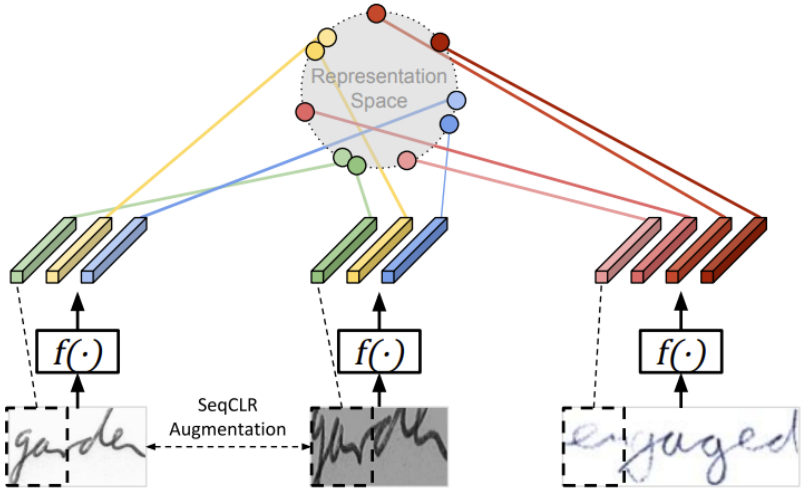
- Contrastive Learning: 유사한 것은 가까워지도록, 상이한 것은 멀어지도록 학습
- Instance Mapping Function을 거친 Sub-word들에 대하여 Contrastive Loss 산출

$$\mathcal{L}(\mathcal{Z}^a, \mathcal{Z}^b) = \sum_{r \in |\mathcal{Z}^a|} \ell_{\text{NCE}}\left(\mathbf{z}_r^a, \mathbf{z}_r^b; \mathcal{Z}^a \cup \mathcal{Z}^b\right) + \sum_{r \in |\mathcal{Z}^b|} \ell_{\text{NCE}}\left(\mathbf{z}_r^b, \mathbf{z}_r^a; \mathcal{Z}^a \cup \mathcal{Z}^b\right)$$

$$\ell_{\text{NCE}}(\mathbf{u}^a, \mathbf{u}^b; \mathcal{U}) = -\log \frac{\exp\left(\text{sim}(\mathbf{u}^a, \mathbf{u}^b)/\tau\right)}{\sum_{\mathbf{u} \in \mathcal{U} \setminus \mathbf{u}^a} \exp\left(\text{sim}(\mathbf{u}^a, \mathbf{u})/\tau\right)}$$

Cosine Similarity

Contrastive Learning



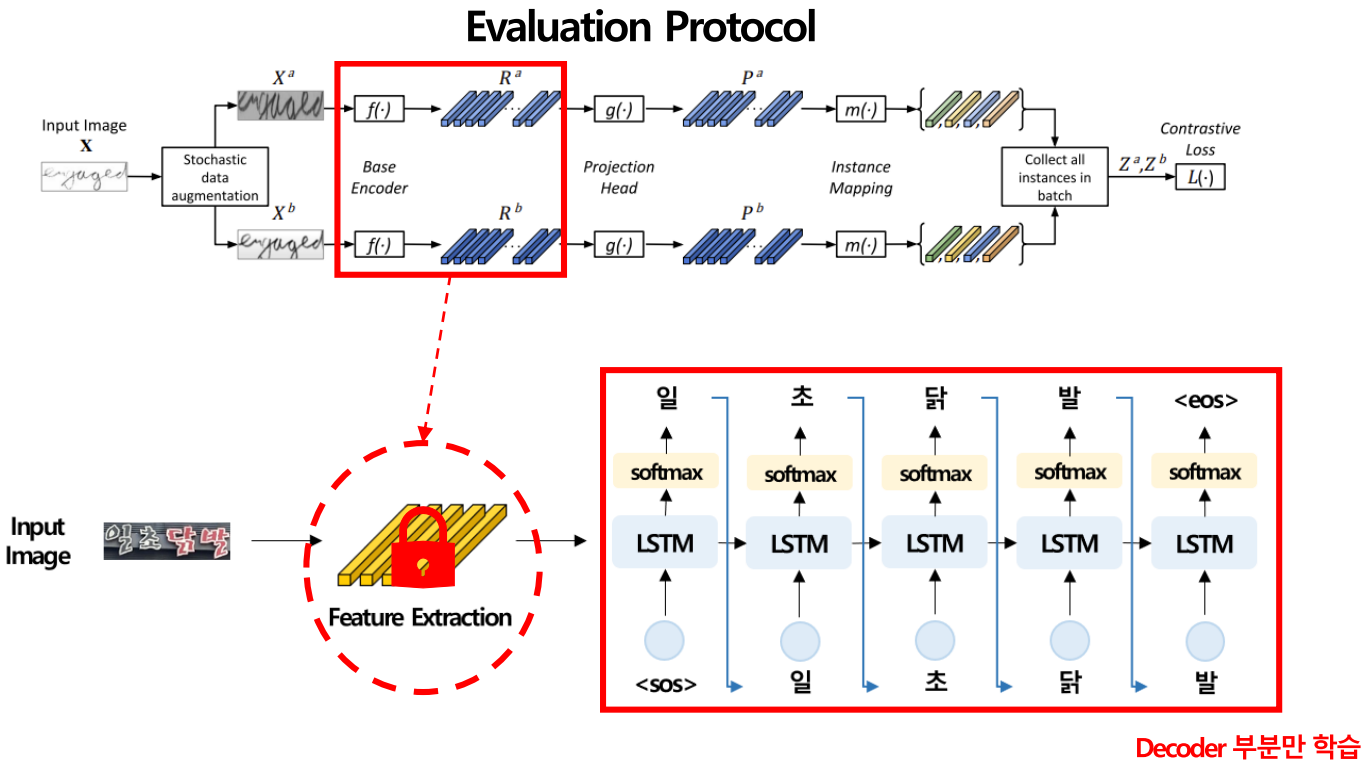
Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Fine-tuning Phase

- 일반적인 자기지도학습과 마찬가지로 Feature Extractor를 Freeze 후, Decoder 부분만 학습
- 이때, Base Encoder만 활용



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Experiment Result

- Sequence를 고려하여 Contrastive Learning을 수행했을 때, 문자 인식에서 좋은 성능을 보임
  - ✓ STR에서는 합성 이미지를 활용하여 풍부한 Representation을 학습하지 못하였기에, Supervised Learning과 유사한 성능
- Window-to-instance 방식이 대체로 우수한 성능을 보임
  - ✓ Frame별로 예측을 수행 후 후처리를 하는 CTC 알고리즘에서는 Frame-to-instance가 효과를 보이기도 함

실험결과

Method	Decoder	Handwritten Dataset									Scene-Text Dataset		
		IAM			RIMES			CVL			IIT5K	IC03	IC13
		Label fraction			Label fraction			Label fraction			Label fraction		
		5%	10%	100%	5%	10%	100%	5%	10%	100%	100%	100%	100%
Supervised Baseline	CTC	21.4	33.6	75.2	35.9	59.7	86.9	48.7	63.6	75.6	76.1	87.9	84.3
SimCLR [9]		15.4	21.8	65.0	36.5	52.9	84.5	52.1	62.0	74.1	69.1	83.4	79.4
SimCLR Contextual		20.4	27.8	63.7	48.6	55.6	84.4	51.8	62.3	74.1	64.5	81.7	78.1
SeqCLR All-to-instance		27.5	44.8	**76.7**	50.4	66.4	89.1	60.1	69.4	76.9	74.7	88.2	83.2
SeqCLR Frame-to-instance		**31.2**	**44.9**	75.1	**61.8**	**71.9**	**90.1**	**66.0**	**71.0**	**77.0**	69.8	84.2	81.8
SeqCLR Window-to-instance		26.2	42.1	**76.7**	56.6	62.5	89.6	61.2	69.7	76.9	**80.9**	**89.8**	**86.3**
Supervised Baseline	Attention	25.7	42.5	77.8	57.0	67.7	89.3	64.0	72.1	77.2	**83.8**	91.1	**88.1**
SimCLR [9]		22.7	32.2	70.7	49.9	60.9	87.8	59.0	65.6	75.7	77.8	88.8	84.9
SimCLR Contextual		24.6	32.9	70.2	51.9	63.0	87.3	59.7	66.2	75.2	72.2	87.0	82.3
SeqCLR All-to-instance		**40.3**	51.6	79.8	69.7	76.9	**92.5**	69.5	73.2	77.6	80.9	90.0	87.0
SeqCLR Frame-to-instance		37.2	48.5	78.2	68.8	75.9	92.3	69.7	73.4	77.5	76.3	90.2	85.8
SeqCLR Window-to-instance		38.1	**52.3**	**79.9**	**70.9**	**77.0**	92.4	**73.1**	**74.8**	**77.8**	82.9	**92.2**	87.9

Algorithms

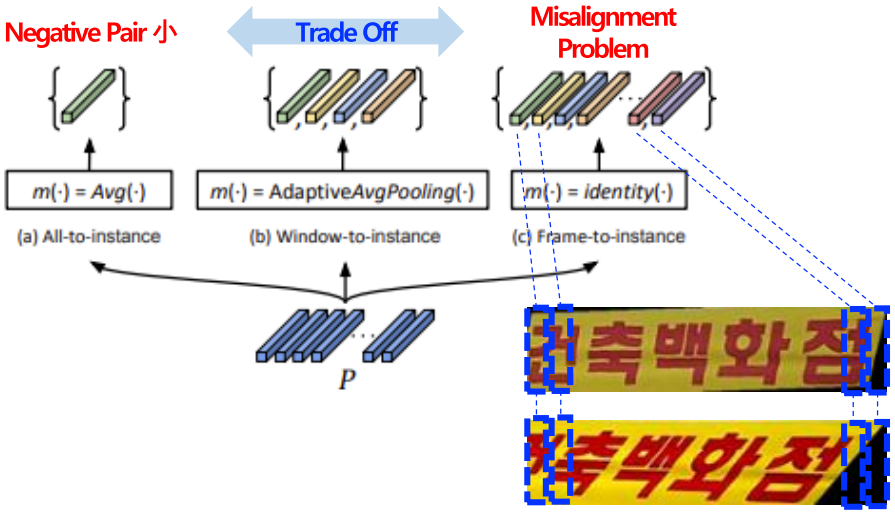
Method	Decoder	IAM			Handwritten Dataset RIMES Label fraction			CVL			Scene-Text Dataset IIIT5K	IC03	IC13
		5%	10%	100%	5%	10%	100%	5%	10%	100%	100%	100%	100%
Supervised Baseline	CTC	21.4	33.6	75.2	35.9	59.7	86.9	48.7	63.6	75.6	76.1	87.9	84.3
SimCLR [1]		15.4	21.8	65.0	36.5	52.9	84.5	52.1	62.0	74.1	69.1	83.4	79.4
SimCLR Contextual		20.4	27.8	63.7	48.6	55.6	84.4	51.8	62.3	74.1	64.5	81.7	78.1
SeqCLR All-to-instance		27.5	44.8	**76.7**	50.4	66.4	89.1	60.1	69.4	76.9	74.7	88.2	83.2
SeqCLR Frame-to-instance		**31.2**	**44.9**	75.1	**61.8**	**71.9**	**90.1**	**66.0**	**71.0**	**77.0**	69.8	84.2	81.8
SeqCLR Window-to-instance		26.2	42.1	**76.7**	56.6	62.5	89.6	61.2	69.7	76.9	**80.9**	**89.8**	**86.3**
Supervised Baseline	Attention	25.7	42.5	77.8	57.0	67.7	89.3	64.0	72.1	77.2	**83.8**	91.1	**88.1**
SimCLR [1]		22.7	32.2	70.7	49.9	60.9	87.8	59.0	65.6	75.7	77.8	88.8	84.9
SimCLR Contextual		24.6	32.9	70.2	51.9	63.0	87.3	59.7	66.2	75.2	72.2	87.0	82.3
SeqCLR All-to-instance		**40.3**	51.6	79.8	69.7	76.9	**92.5**	69.5	73.2	77.6	80.9	90.0	87.0
SeqCLR Frame-to-instance		37.2	48.5	78.2	68.8	75.9	92.3	69.7	73.4	77.5	76.3	90.2	85.8
SeqCLR Window-to-instance		38.1	**52.3**	**79.9**	**70.9**	**77.0**	92.4	**73.1**	**74.8**	**77.8**	82.9	**92.2**	87.9

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

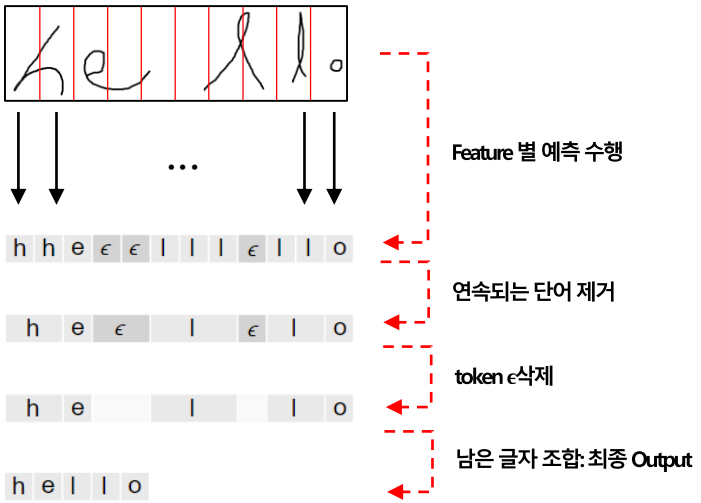
❖ Experiment Result

- Sequence를 고려하여 Contrastive Learning을 수행했을 때, 문자 인식에서 좋은 성능을 보임
  - ✓ STR에서는 합성 이미지를 활용하여 풍부한 Representation을 학습하지 못하였기에, Supervised Learning과 유사한 성능
- Window-to-instance 방식이 대체로 우수한 성능을 보임
  - ✓ Frame별로 예측을 수행 후 후처리를 하는 CTC 알고리즘에서는 Frame-to-instance가 효과를 보이기도 함

Instance Mapping Function



CTC 기반 Decoder



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Self-supervised Learning-based STR: Sequence-to-Sequence Contrastive Learning for Text Recognition (CVPR, 2021)

❖ Summary

- 문자인식의 Labeled 데이터가 부족한 상황에서 Unlabeled 데이터를 활용하는 자기지도학습 Framework 제안
  - ✓ Contrastive Learning을 활용하여 학습
- 모델구조
  - ✓ Data Augmentation
  - ✓ Base Encoder + Projection
  - ✓ Instant Mapping Function
- 기여점
  - ✓ 문자열의 순서를 유지하기 위해 Text Recognition에 적합한 데이터 증강 기법 제시
  - ✓ 여러 개의 출력 값을 Contrastive Learning에 활용할 수 있도록 Instance Mapping Function을 제안
    - ➢ 너무 많은 Instance에 Mapping할 경우, 데이터 증강 세기에 영향을 받아 Misalignment Pair 문제 발생
    - ➢ 너무 적은 Instance에 Mapping할 경우, Contrastive Learning에서 Negative Pair의 개수 감소

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Semi-supervised Learning-based Scene Text Recognition

- STR에 Semi-supervised Learning의 Consistency Regularization을 적용

Pushing the Performance Limit of Scene Text Recognizer without Human Annotation

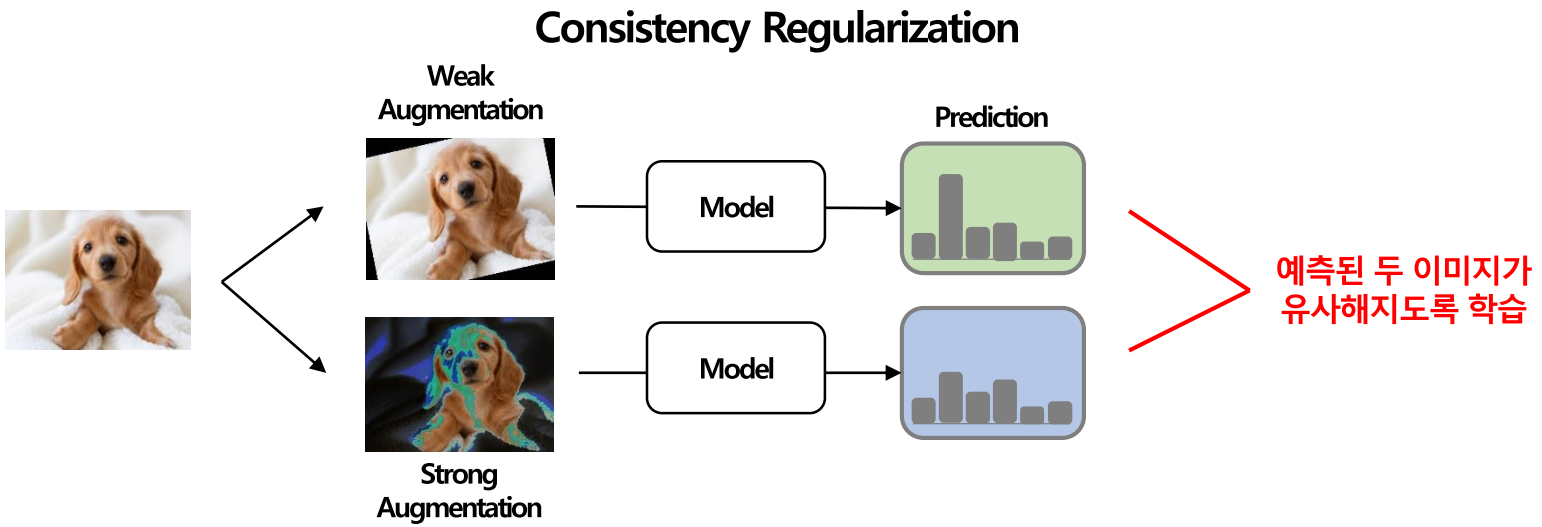
Caiyuan Zheng[1,2]*, Hui Li[3], Seon-Min Rhee[4], Seungju Han[4], Jae-Joon Han[4], Peng Wang[1,2]†
[1]School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, China,
[2]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, China,
[3]Samsung R&D Institute China Xi'an (SRCX),
[4]Samsung Advanced Institute of Technology (SAIT), South Korea
2020202704@mail.nwpu.edu.cn, {hui01.li, s.rhee, sj75.han, jae-joon.han}@samsung.com
peng.wang@nwpu.edu.cn

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Overview & Background

- STR에 합성 데이터와 실제 Unlabeled 데이터를 함께 활용하는 준지도학습 Framework 제안
  - ✓ Attention-based Decoder에 적합한 모델
- 준지도학습 중 Consistency Regularization을 활용
  - ✓ Consistency Regularization: 동일한 이미지에서 다르게 변형된 이미지를 입력으로 받더라도, 동일한 결과를 갖도록 학습
- 일반적인 준지도학습(FixMatch, UDA 등)을 STR에 적용한다면, 아래와 같은 한계가 존재
  - ① 합성 이미지와 실제 이미지 사이 데이터 분포가 달라서 학습이 붕괴
  - ② 글자 간 Misalignment 문제가 발생하여 동일하지 않은 글자에 대해 Consistency Regularization이 이루어져 학습에 방해



Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Overview of Model Architecture

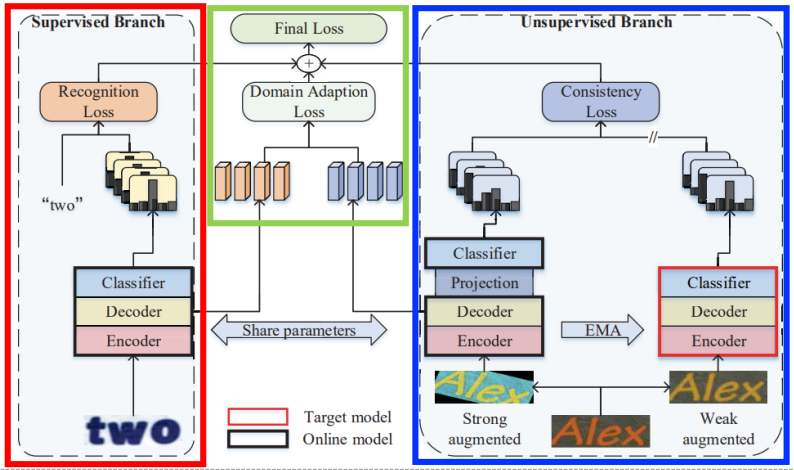
- Supervised Branch
- Unsupervised Branch

Projection Layer, Exponential Moving Average(EMA), Weight Decay, Stop Gradient

  - ✓ Online Model 및 Target Model로 구성된 Asymmetric한 구조 (Inspired from BYOL)
  - ✓ Character-level Consistency Regularization

- Domain Adaptation
  - ✓ Supervised Branch와 Unsupervised Branch의 Online 모델 간 Deep CORAL Loss (ECCV, 2016)활용
  - ✓ 합성 데이터와 실제 데이터의 도메인 차이를 최소화

Model Architecture



Algorithms

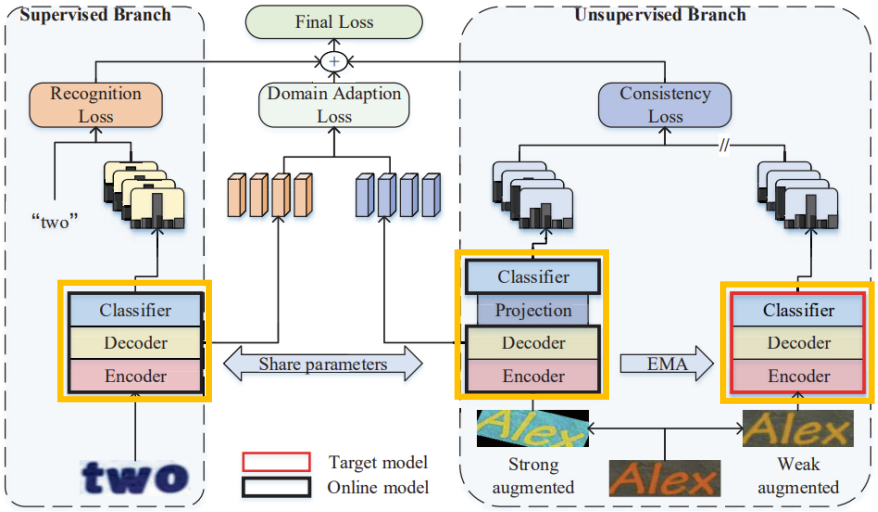
Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Overview of Model Architecture

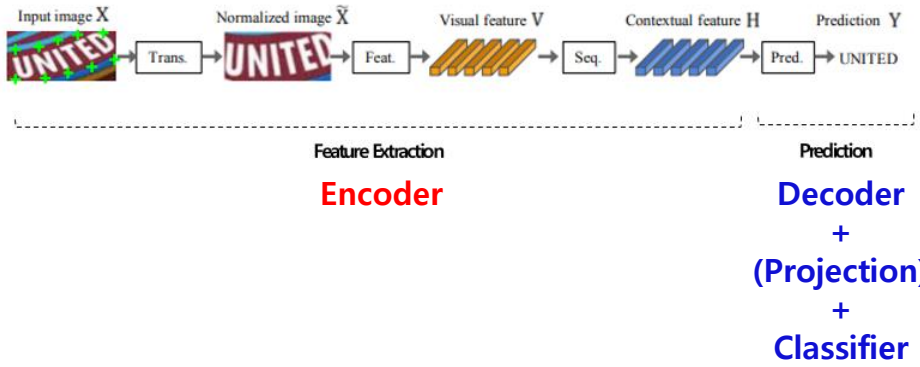
- Supervised Branch 및 Unsupervised Branch 내 모델의 기본구조는 아래와 같음
  - ① Encoder: 입력 이미지에서 Feature를 추출
  - ② Decoder: 이미지 단위 Feature에서 글자 단위 Feature를 생성
  - ③ Classifier: 글자 단위 Feature에서 각 글자들을 예측 (Linear Transformation + Softmax)

  * Projection: Unsupervised Branch의 Online Model에만 있는 구조

Model Architecture



Scene Text Recognition의 구조



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

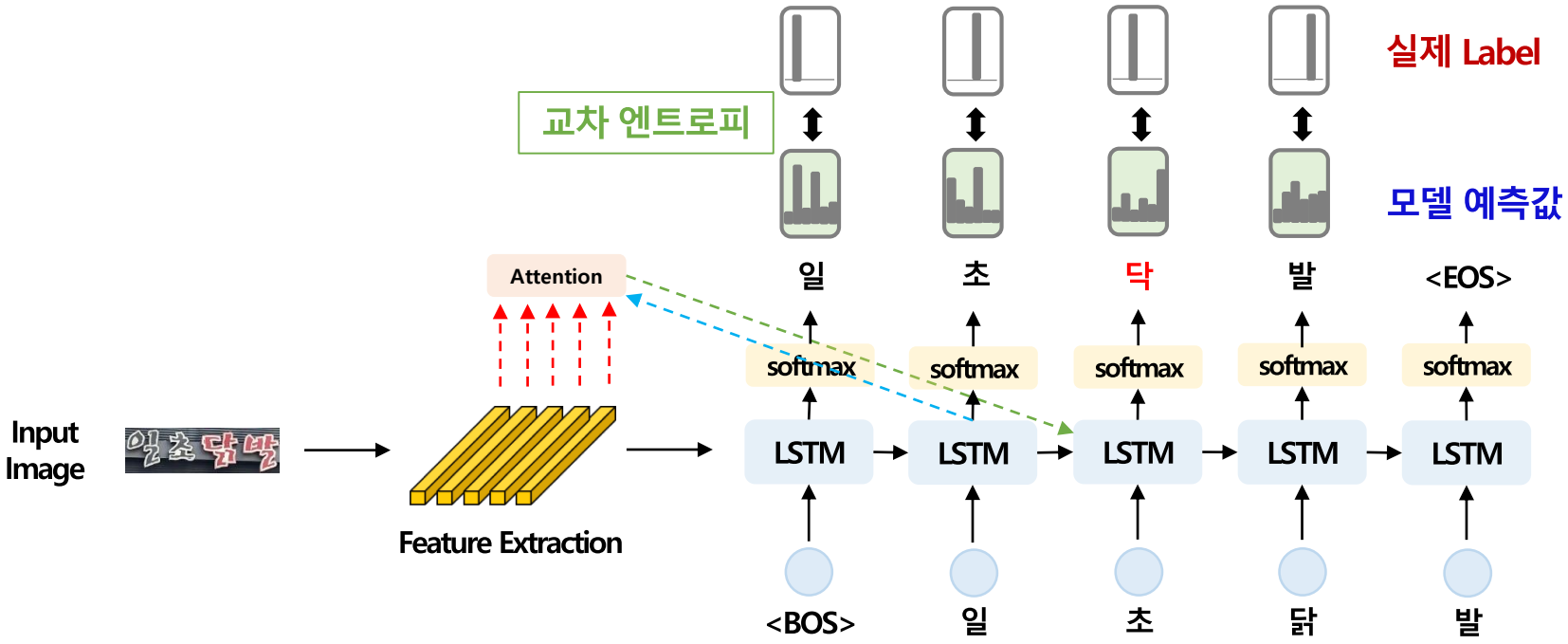
Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (1): Supervised Branch

- Labeled 데이터(합성 데이터)를 활용하여 학습
  - ✓ 목적함수: 교차 엔트로피

$$\mathcal{L}_{reg} = \frac{1}{T}\sum_{t=1}^{T} \log p_t^L(y_t^{gt} \mid \mathbf{X}^L)$$

- 레이블 글자들이 Decoder의 입력 글자로 활용된다.
- 학습한 Weight를 Unsupervised Branch의 Online Model과 공유



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

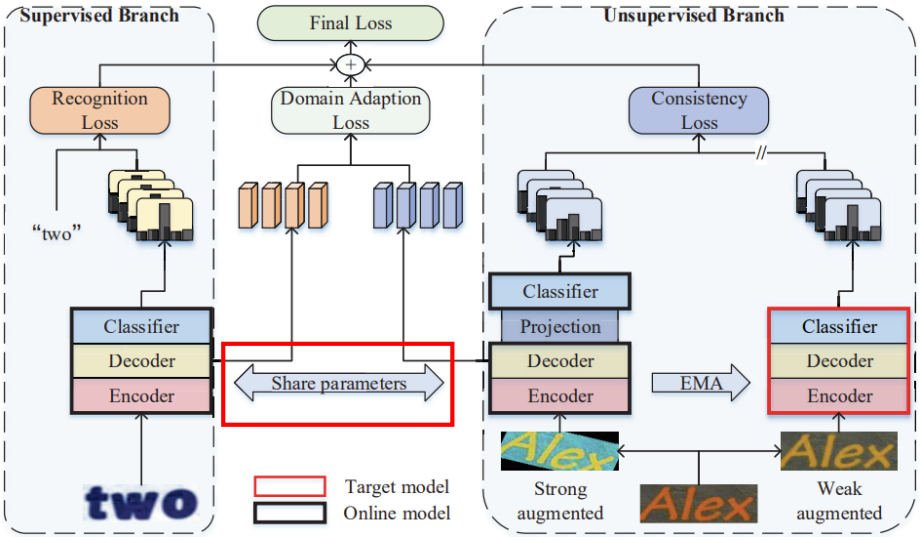
❖ Model Architecture (1): Supervised Branch

- Labeled 데이터(합성 데이터)를 활용하여 학습
  - ✓ 목적함수: 교차 엔트로피

$$\mathcal{L}_{reg} = \frac{1}{T}\sum_{t=1}^{T} \log p_t^L(y_t^{gt} \mid \mathbf{X}^L)$$

- 레이블 글자들이 Decoder의 입력 글자로 활용된다.
- 학습한 Weight를 Unsupervised Branch의 Online Model과 공유

Model Architecture



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch

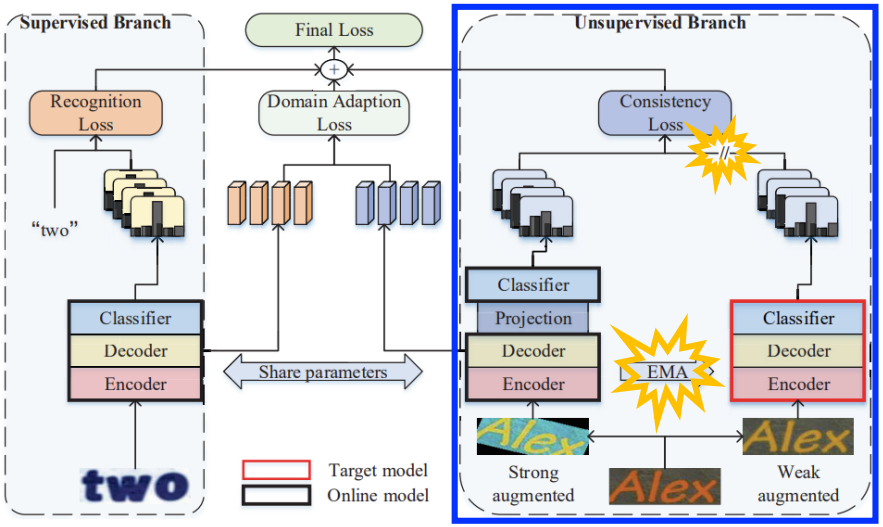
- Unlabeled 데이터를 활용하여 학습
- Online 모델과 Target Model은 Asymmetric
- 하나의 이미지를 두 번 증강 후 두 예측 값이 글자 단위로 유사해지도록 학습 (KL-Divergence)

Character-level Consistency Regularization

  - ✓ Online Model: Strong Augmentation 이미지가 입력 / Encoder + Decoder + Projection + Classifier / Weight Decay
  - ✓ Target Model: Weak Augmentation 이미지가 입력 / Encoder + Decoder + Classifier / Stop Gradient (EMA 활용)

$$\theta_t = \alpha\theta_t + (1-\alpha)\theta_o$$

Model Architecture



한계① 극복 포인트

학습의 안정성 개선 및 성능 향상

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

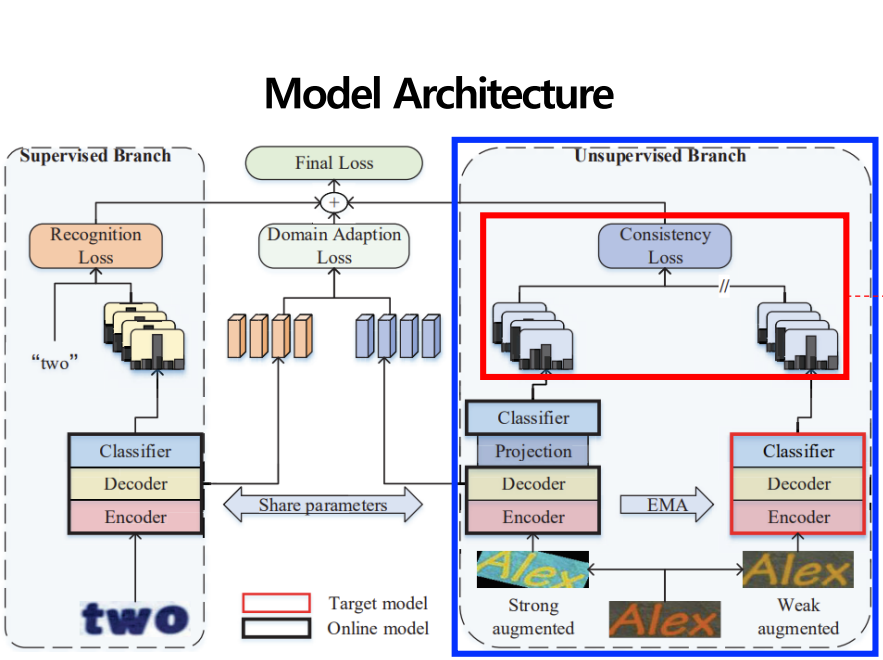
❖ Model Architecture (2): Unsupervised Branch

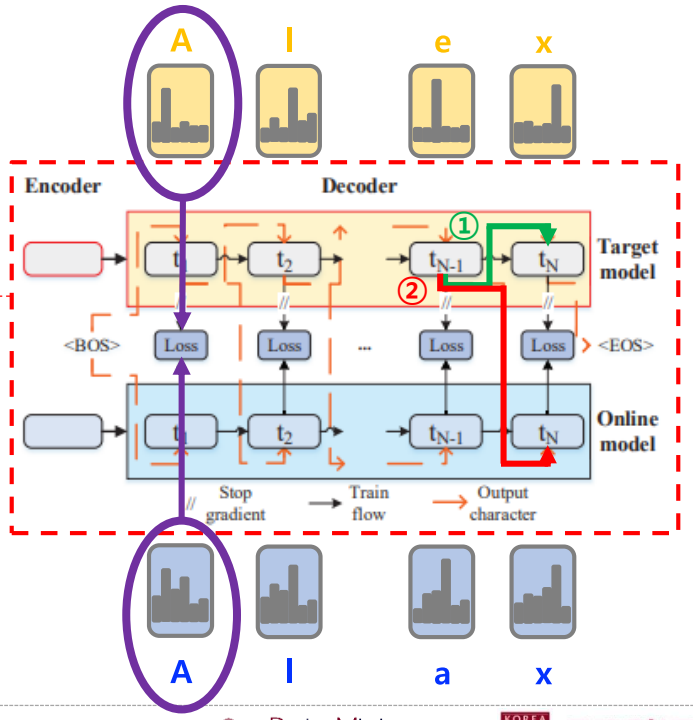
- Unlabeled 데이터를 활용하여 학습
- Online 모델과 Target Model은 Asymmetric
- 하나의 이미지를 두 번 증강 후 두 예측 값이 글자 단위로 유사해지도록 학습 (KL-Divergence)

Character-level Consistency Regularization

  - ✓ Online Model: Strong Augmentation 이미지가 입력 / Encoder + Decoder + Projection + Classifier / Weight Decay
  - ✓ Target Model: Weak Augmentation 이미지가 입력 / Encoder + Decoder + Classifier / Stop Gradient (EMA 활용)

Model Architecture



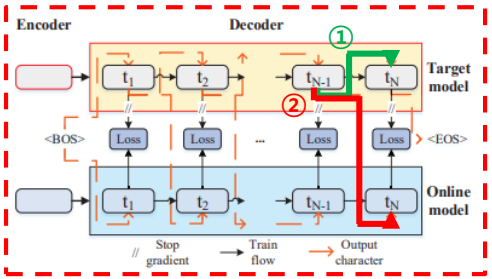


Algorithms

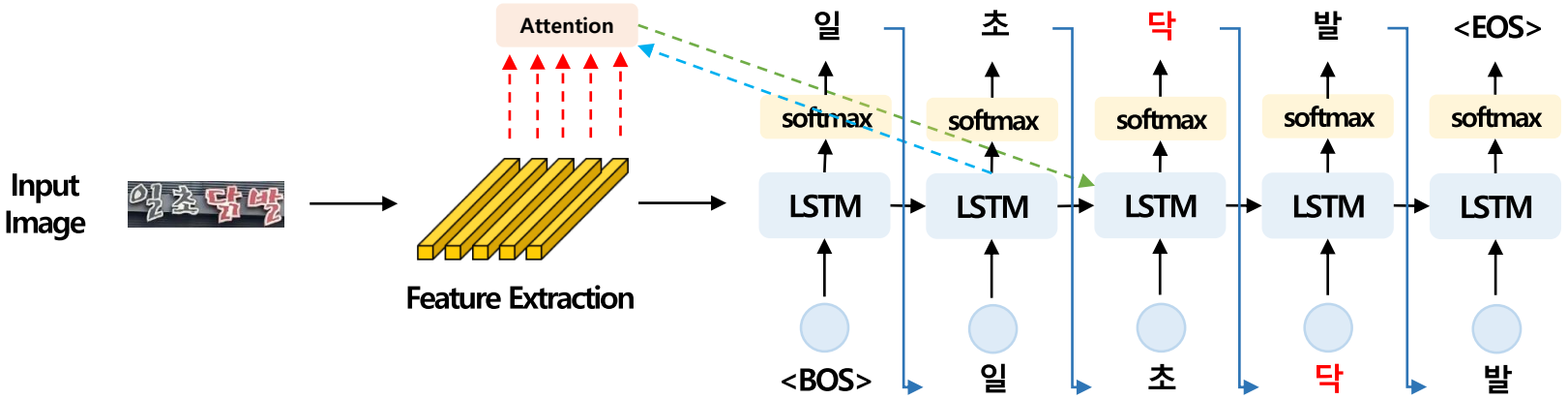
Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch - Decoder

- Decoder는 Autoregressive 하게 이전 시점 출력값 활용
- Target Model의 Output을 Online Model과 공유
  - ✓ Target Model: Autoregressive하게 이전 시점의 값을 활용
  - ✓ Online Model: Target Model의 이전 시점 값을 활용



Attention-based Decoder in Unsupervised Branch

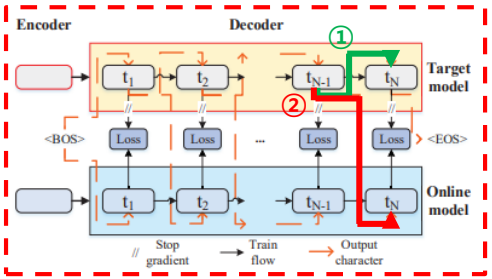


Algorithms

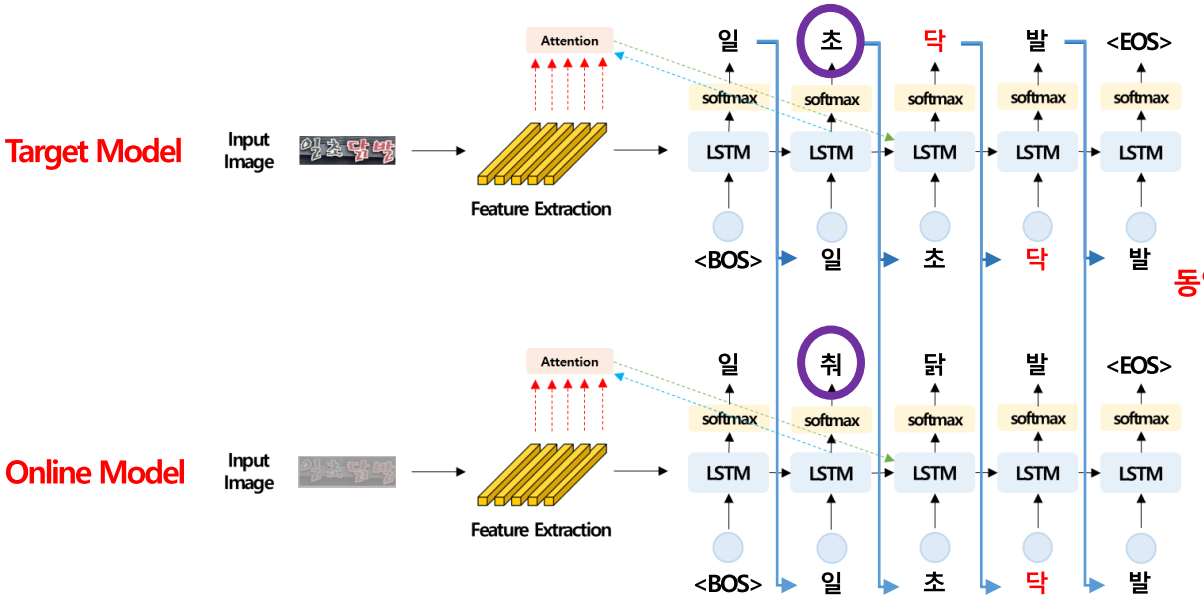
Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch - Decoder

- Decoder는 Autoregressive 하게 이전 시점 출력값 활용
- Target Model의 Output을 Online Model과 공유
  - ✓ Target Model: Autoregressive하게 이전 시점의 값을 활용
  - ✓ Online Model: Target Model의 이전 시점 값을 활용



Attention-based Decoder in Unsupervised Branch



한계② 극복 포인트
동일한 Context Information을 공유하여
글자 간 Misalignment를 최소화

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

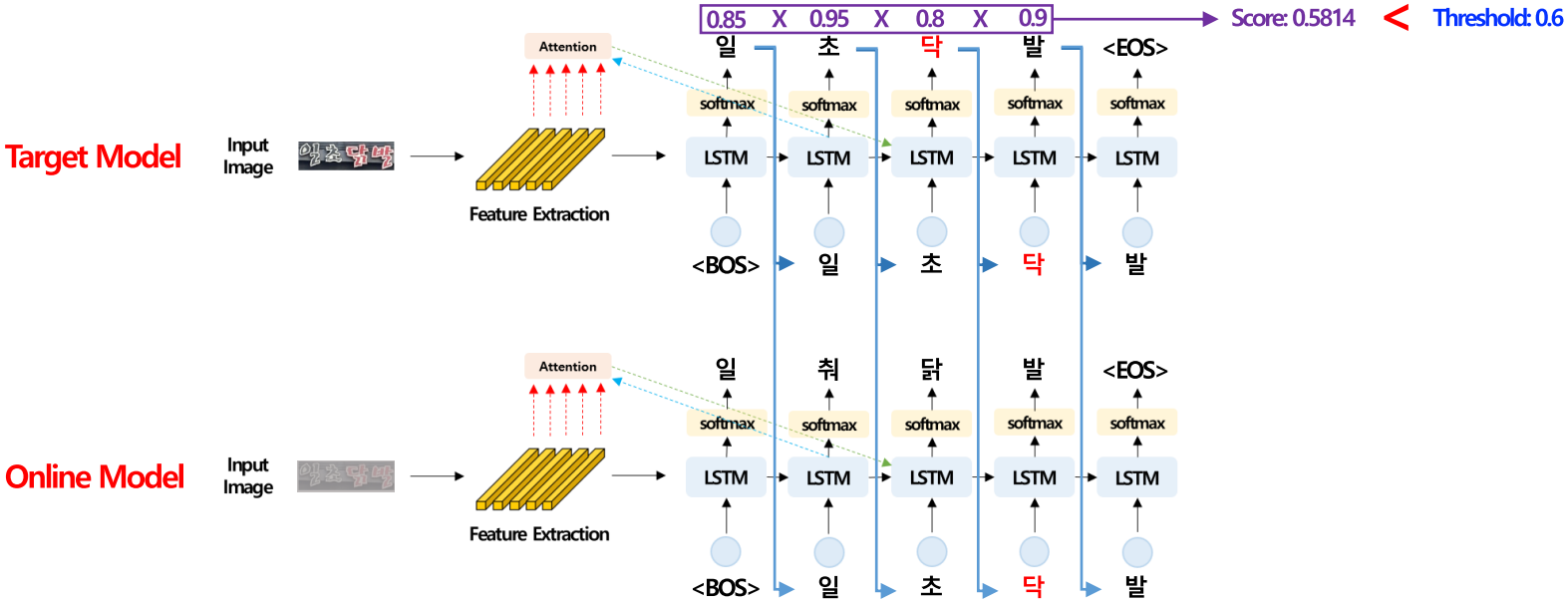
Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch - Decoder

- Target Model의 각 글자 별 예측 확률을 활용하여 Noisy 데이터를 필터링
  - ✓ 각 시점에서 예측된 글자들의 가장 높은 확률들을 곱하여 점수 산정
  - ✓ 산정한 점수와 Threshold와 비교했을 때, 점수가 더 높은 이미지만 학습에 활용

Attention-based Decoder in Unsupervised Branch



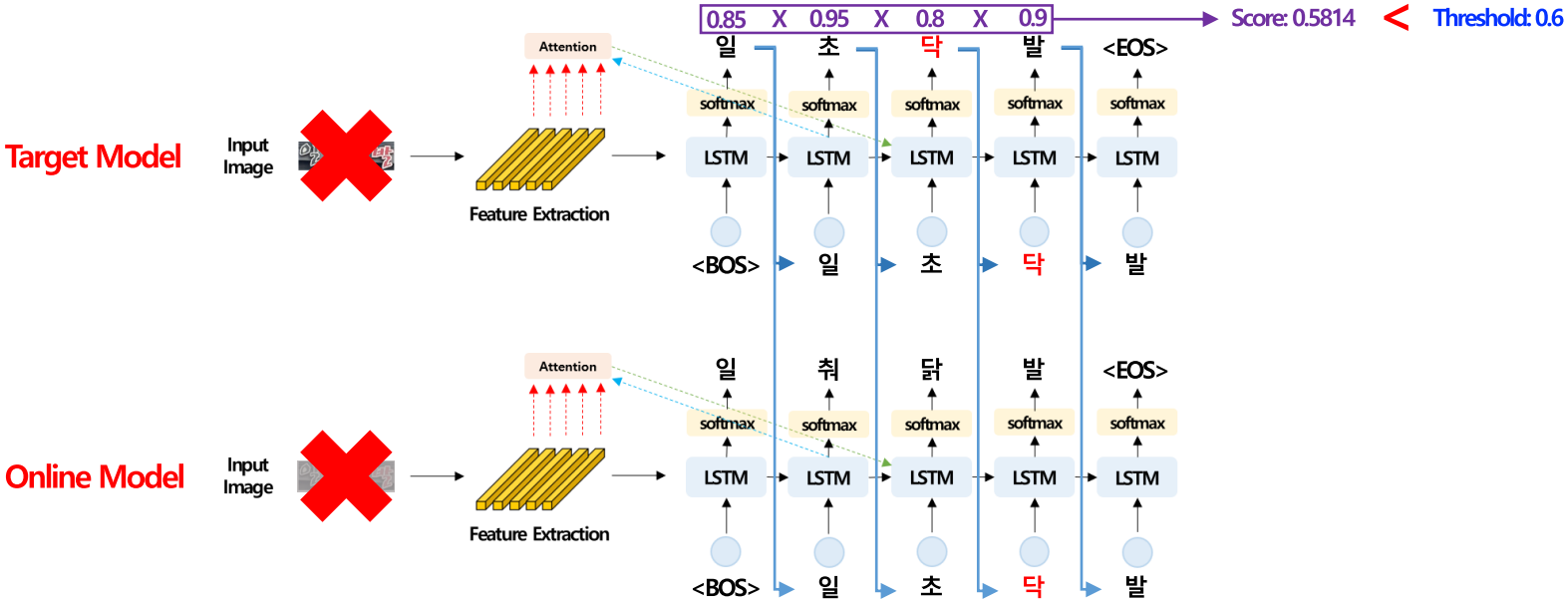
Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch - Decoder

- Target Model의 각 글자 별 예측 확률을 활용하여 Noisy 데이터를 필터링
  - ✓ 각 시점에서 예측된 글자들의 가장 높은 확률들을 곱하여 점수 산정
  - ✓ 산정한 점수와 Threshold와 비교했을 때, 점수가 더 높은 이미지만 학습에 활용

Attention-based Decoder in Unsupervised Branch



Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (2): Unsupervised Branch Loss

- 목적함수: KL-Divergence
  - ✓ Noisy 데이터를 필터링
  - ✓ Consistency Regularization

$$\mathcal{L}_{cons} = \mathbb{I}(\mathbf{s}^{U_w} > \beta_U)\frac{1}{T}\sum_{t=1}^{T} Dist(\mathbf{p}_t^{U_w}, \mathbf{p}_t^{U_s})$$

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (3): Domain Adaptation

- Domain Adaptation을 통해 합성 데이터와 실제 데이터 간 Domain Shift를 최소화
  - ✓ Supervised Branch와 Unsupervised Branch의 Target Model의 Vision Feature에서 각각 공분산 행렬을 구한 후, 이들의 차이를 통해 Domain Shift 최소화 (Deep CORAL Loss, ECCV, 2016)

Unsupervised Branch의 공분산 행렬

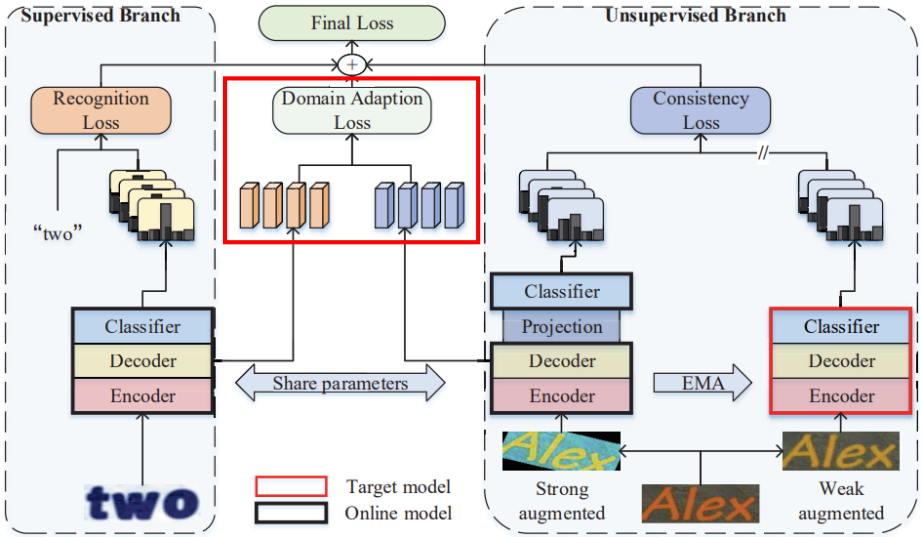
$$\mathcal{L}_{da} = \frac{1}{4d^2} \| (cov(\mathbf{H}^L) - cov(\mathbf{H}^{U_s}) \|_F^2$$

프로베니우스 놈

Feature Dimension

Supervised Branch의 공분산 행렬

Model Architecture



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture (4): Overall Loss

- Supervised Branch Loss + Unsupervised Branch Loss + Domain Adaptation Loss
  - ✓ λ_{cons}: Unsupervised Branch Loss에 대한 가중치
  - ✓ λ_{da}: Domain Adaptation Loss에 대한 가중치

$$\mathcal{L}_{overall} = \mathcal{L}_{reg} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{da}\mathcal{L}_{da}$$

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

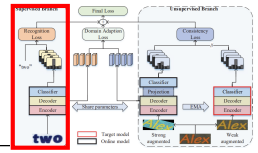
Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Model Architecture Summary

1. Supervised Branch Loss 산출
   - ① 예측값과 레이블을 활용하여 교차 엔트로피를 통해 산출
2. Unsupervised Branch Loss 산출
   - ① 데이터 증강 2회 후 Online 및 Target Model에 입력
   - ② Encoder ~ Classifier를 통과하여 각 Model에서 예측 수행
   - ③ Target Model의 예측 확률 값을 통해 점수를 산정하고, 임계값을 비교하여 Noisy 데이터 여부 판별
   - ④ Noisy 데이터가 아니라면, 글자 단위로 Consistency Loss 산출 (이때, Context Information을 공유)
3. Domain Adaptation Loss 산출
   - ① Supervised Branch와 Unsupervised Branch Target Model의 Visual Feature를 활용하여 Deep CORAL Loss 산출
4. Overall Loss = Supervised Branch Loss + Unsupervised Branch Loss + Domain Adaptation Loss
5. Weight Update
   - ① Supervised Branch Weight 업데이트
   - ② Unsupervised Branch (Online Model): Projection 가중치만 업데이트, 그 외 Encoder/Decoder/Classifier는 Supervised Branch 가중치를 공유
   - ③ Unsupervised Branch (Target Model): Online Model의 가중치를 EMA하여 업데이트

Algorithms



Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Experiment Result

- Inference 시에는 Supervised Branch만 활용
- 본 연구의 준지도학습 Framework를 적용했을 때, 기존 지도학습 모델보다 좋은 성능을 보임
  - ✓ 기존의 SOTA 지도학습 모델보다 좋은 성능

다른 모델과 비교 실험 결과

	Methods	Labeled Dataset	Unlabeled Dataset	Regular Text IC13	SVT	IIIT	Irregular Text IC15	SVTP	CUTE	Avg
SOTA Methods	Shi *et al.* [39] (CRNN)	MJ	-	-	80.8	78.2	-	-	-	-
	Luo *et al.* [28](MORAN)	SL	-	-	88.3	93.4	77.8	79.7	81.9	-
	Yang *et al.* [52](HGA)	SL	-	-	88.9	94.7	79.5	80.9	85.4	-
	Baek *et al.* [2](TRBA)	SL	-	-	87.5	87.9	-	79.2	74.0	-
	Liao *et al.* [24](Mask TextSpotter)	SL	-	95.3	91.8	93.9	77.3	82.2	87.8	88.3
	Wan *et al.* [46](TextScanner)	SL	-	92.9	90.1	93.9	79.4	84.3	83.3	88.5
	Wang *et al.* [49](DAN)	SL	-	93.9	89.2	94.3	74.5	80.0	84.4	87.2
	Yue *et al.* [55](RobustScanner)	SL	-	94.8	88.1	95.3	77.1	79.5	90.3	88.4
	Qiao *et al.* [34](SRN)	SL	-	95.5	91.5	94.8	82.7	85.1	87.8	90.4
	Zhang *et al.* [57](SPIN)	SL	-	-	90.9	95.2	82.8	84.3	83.2	-
	Mou *et al.* [31](PlugNet)	SL	-	-	92.3	94.4	-	84.3	84.3	-
	Qiao *et al.* [33](PIMNet)	SL	-	95.2	91.2	95.2	83.5	84.3	84.4	90.5
	Fang *et al.* [9](ABINet)	SL	-	97.4	93.5	96.2	86.0	89.3	89.2	92.7
	Gao *et al.* [12]	10% SL	90% SL	-	78.1	74.8	-	-	-	-
	Baek *et al.* [3](CRNN)	RL	Book32 et al.	-	84.3	89.8	-	74.6	82.3	-
	Baek *et al.* [3](TRBA)	RL	Book32 et al.	-	91.3	94.8	-	82.7	88.1	-
	Fang *et al.* [9](ABINet)	SL	Uber-Text	97.3	94.9	**96.8**	87.4	90.1	93.4	93.5
Ours	CRNN-pr	SL	-	91.0	82.2	90.2	71.6	70.7	81.3	82.8
	CRNN-cr	SL	RU	92.4	87.9	92.0	75.8	75.7	85.8	85.9
	MORAN-pr	SL	-	95.1	90.4	93.4	79.7	80.6	85.4	88.5
	MORAN-cr	SL	RU	96.5	93.0	94.1	82.6	82.9	88.5	90.2
	HGA-pr	SL	-	95.0	89.5	93.6	79.8	81.1	87.8	88.7
	HGA-cr	SL	RU	95.4	93.2	94.9	84.0	86.8	92.0	91.2
	TRBA-pr	SL	-	97.3	91.2	95.3	84.2	86.4	92.0	91.5
	TRBA-cr	10% SL	10% RU	97.3	94.7	96.2	87.0	89.6	**94.4**	93.2
	TRBA-cr	SL	RU	**98.3**	**96.3**	96.5	**89.3**	**93.3**	93.4	**94.5**

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Experiment Result

- 본 연구의 준지도학습 Framework가 준지도학습 선행연구보다도 더 좋은 성능을 보임
  - ✓ FixMatch나 UDA는 Labeled 데이터와 Unlabeled 데이터가 동일한 도메인(In-domain)일 때는 효과가 있었지만, 다른 도메인(Cross-domain) 일 때는 오히려 악영향을 미침
    → 기존 방법론들은 합성 데이터와 분포 차이를 극복하지 못한 것으로 추정

준지도학습 모델과 비교 실험 결과

	Labeled/ Unlabeled Data	Methods	IC13 IC15	SVT SVTP	IIIT CUTE	Avg
In-domain	RL_{20p}(55.7K)/ -	Sup	90.1 77.6	87.5 78.0	88.8 83.0	84.8
	RL_{20p}(55.7K)/ RL_{80p}(223K)	FixMatch	93.0 82.3	88.6 82.5	92.0 88.5	88.4
		UDA	92.5 80.7	88.6 80.9	91.4 88.5	87.4
		Ours	93.8 82.5	91.5 83.6	92.9 88.5	89.3
Cross-domain	SL_{sm}(1.45M)/ -	Sup	96.0 82.4	90.0 82.6	94.4 88.9	89.9
	SL_{sm}(1.45M)/ RU_{sm}(1.06M)	FixMatch	90.0 72.6	86.2 77.2	79.2 69.1	78.9
		UDA	94.2 75.7	85.3 79.5	90.0 82.3	85.3
		Ours	97.3 87.0	94.7 89.6	96.2 94.4	**93.2**

Method	IC13 IC15	SVT SVTP	IIIT CUTE	Avg
Pseudo Label (PL)	95.9 82.9	91.2 85.7	95.4 90.6	90.9
Noisy Student (NS)	96.3 85.5	94.4 86.7	96.1 94.1	92.4
Ours	97.3 87.0	94.7 89.6	96.2 94.4	**93.2**

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Experiment Result

- Asymmetric 구조에 영향을 미치는 4가지 요소에 대한 효과 검증
- 본 연구에서 제안하는 Character-level Consistency Regularization(CCR)의 효과 검증

Asymmetric 구조에 대한 실험 결과

Projection	WD	EMA	DA	IC13 IC15	SVT SVTP	IIIT CUTE	Avg
				94.2 80.5	91.5 84.0	88.7 84.0	87.0
✔				94.5 81.6	90.1 86.1	89.5 85.4	87.7
✔	✔			97.2 85.9	93.0 87.0	93.5 91.3	91.2
✔	✔	✔		96.7 86.7	94.6 89.3	95.9 92.7	92.8
✔	✔	✔	✔	97.3 87.0	94.7 89.6	96.2 94.4	**93.2**

CCR에 대한 실험 결과

Method	IC13 IC15	SVT SVTP	IIIT CUTE	Avg
SCR	96.6 84.9	93.0 85.9	96.4 93.1	92.2
CCR	97.3 87.0	94.7 89.6	96.2 94.4	**93.2**

Algorithms

Semi-supervised Learning-based STR: Pushing the Performance Limit of Scene Text Recognizer without Human Annotation (CVPR, 2022)

❖ Summary

- STR에서 합성 데이터와 실제 Unlabeled 데이터를 활용하는 준지도학습 Framework 제안
  - ✓ Consistency Regularization을 활용하여 학습
- 모델 구조
  - ✓ Supervised Branch
  - ✓ Unsupervised Branch: Asymmetric Structure, Character-level Consistency Regularization
  - ✓ Domain Adaptation: Deep CORAL Loss
- 기여점
  - ✓ Asymmetric Structure로 학습 안정화 및 성능 개선
  - ✓ Character-level Consistency Regularization에서 Contextual Information을 공유함으로써 글자 간 Misalignment 문제 해결
  - ✓ Deep CORAL Loss를 활용하여 합성 데이터에 의해 발생하는 Domain Shift 최소화

Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Self&Semi-supervised Learning-based Scene Text Recognition

- STR에 Self-supervised Learning과 Semi-supervised Learning을 모두 적용
- SemiMTR: Semi-supervised Learning-based Multimodal Text Recognition

Multimodal Semi-Supervised Learning for Text Recognition

Aviad Aberdam[1], Roy Ganz[2⋆], Shai Mazor[1], and Ron Litman[1]

[1] AWS AI Labs
[2] Technion, Israel
{aaberdam,smazor,litmanr}@amazon.com, ganz@cs.technion.ac.il

Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Overview & Background

- 최근 연구 흐름
  - ✓ STR은 학습을 위한 Labeled 데이터가 매우 부족한 실정
  - ✓ STR에 Unlabeled 데이터를 활용한 연구는 Vision Feature만 고려하여 수행되고 있음
- STR에 Self-supervised Learning과 Semi-supervised Learning을 모두 적용한 Framework 제안
  - ✓ Self-supervised Learning: Contrastive Learning
  - ✓ Semi-supervised Learning: Consistency Regularization
- Vision & Language를 모두 고려한 Multimodal기반의 모델을 제안
  - ✓ 1단계: Vision Model Pretraining
  - ✓ 2단계: Language Model Pretraining
  - ✓ 3단계: Vision & Language Model Fine-tuning & Fusion Model Training

Scene Text Recognition

Natural Language Processing **Computer Vision**

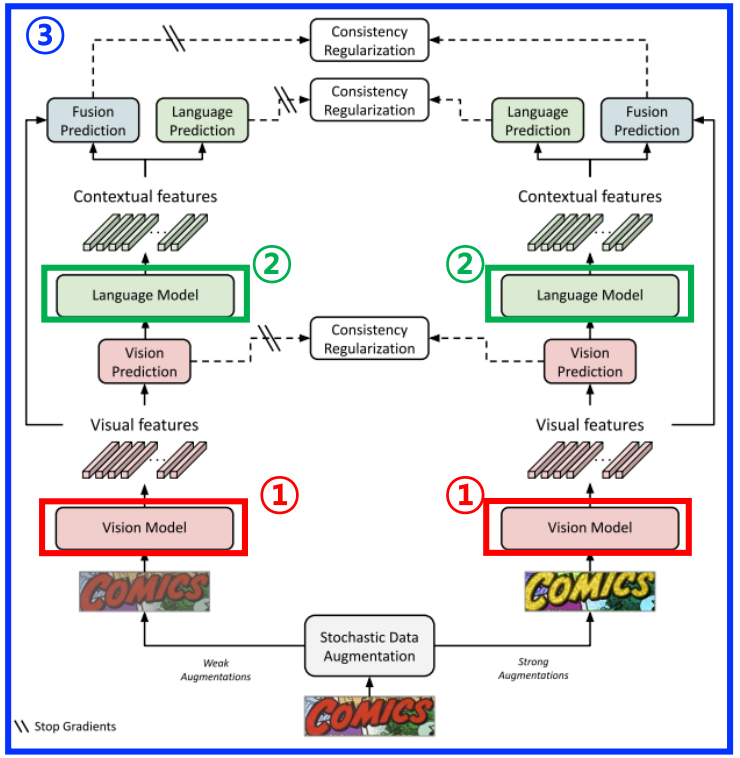
Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Overview of Model Architecture (SemiMTR)

- Vision Model Pretraining
  - ✓ Contrastive Learning에 Supervised Loss를 결합하여 활용
- Language Model Pretraining
  - ✓ Masked Language Model
- Fine-tuning & Fusion Model Training
  - ✓ 각 Modality별 Prediction
  - ✓ 각 Modality별 Consistency Regularization

Model Architecture



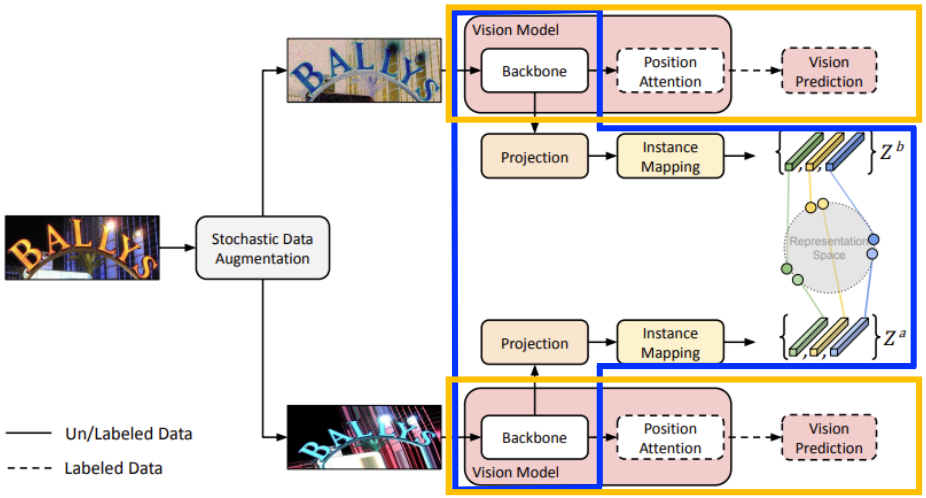
Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

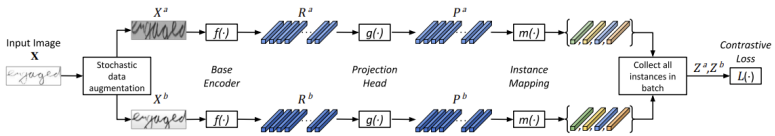
❖ Model Architecture (1) Vision Model Pretraining

- Contrastive Learning 및 Supervised Learning을 동시에 활용하여 사전학습
  - ✓ Contrastive Learning은 SeqCLR(CVPR, 2021)의 구조 (Unlabeled 데이터 및 Labeled 데이터 모두 활용)
  - ✓ Supervised Learning은 Backbone + Position Attention의 구조 (Labeled 데이터만 활용)
- 추후 Fusion Model에는 Visual Backbone만 활용

Vision Model Architecture



SeqCLR Architecture



Algorithms

$$\mathcal{L}_{\text{SeqCLR}}(\mathcal{Z}^a, \mathcal{Z}^b) = \sum_{0 \le i < NT} \ell_{\text{NCE}}\left(\mathbf{z}_i^a, \mathbf{z}_i^b; \mathcal{Z}^a \cup \mathcal{Z}^b\right) + \ell_{\text{NCE}}\left(\mathbf{z}_i^b, \mathbf{z}_i^a; \mathcal{Z}^a \cup \mathcal{Z}^b\right)$$

$$\ell_{\text{NCE}}(\mathbf{u}^a, \mathbf{u}^b; \mathcal{U}) = -\log \frac{\exp\left(\text{sim}(\mathbf{u}^a, \mathbf{u}^b)/\tau\right)}{\sum_{\mathbf{u} \in \mathcal{U} \setminus \mathbf{u}^a} \exp\left(\text{sim}(\mathbf{u}^a, \mathbf{u})/\tau\right)}$$

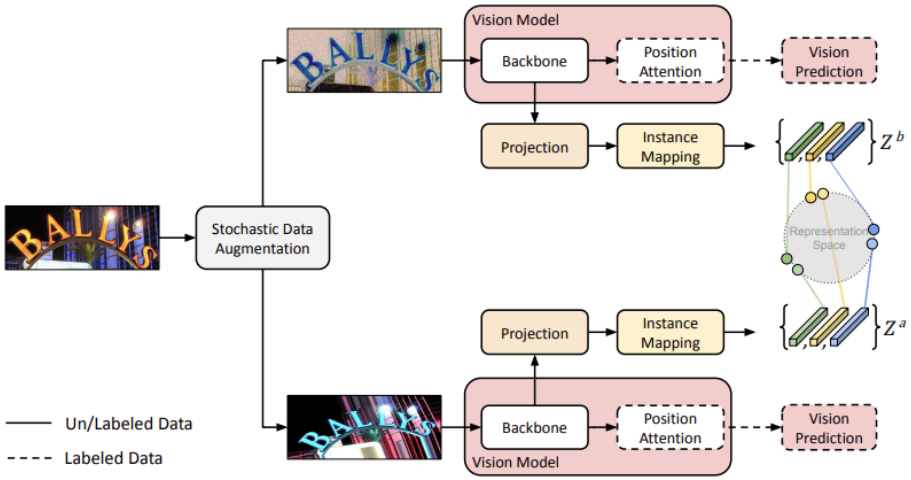
Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Model Architecture (1) Vision Model Pretraining

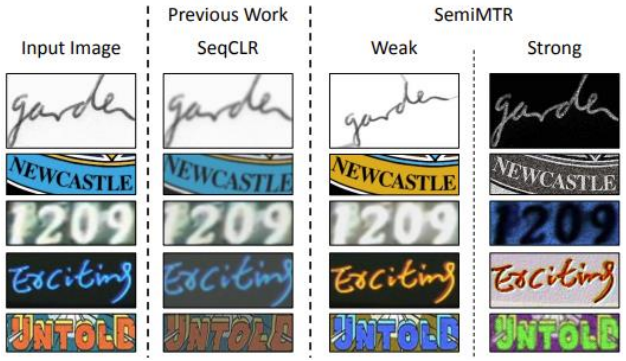
- 학습과정
  - 데이터 증강 2회: STR의 특성을 반영하여, 순서를 훼손하지 않는 선에서 SeqCLR보다 Stronger Augmentation (for Color)
  - Contrastive Learning: Visual Backbone – Projection – Instance Mapping – Contrastive Loss (SeqCLR) / NCE Loss
  - Supervised Learning: Visual Backbone – Position Attention – Vision Prediction – Supervised Loss / Cross Entropy

    * Visual Backbone: ResNet / Position Attention: Transformer / Instance Mapping: Window-to-Instance
- 목적함수: NCE Loss와 Cross Entropy를 가중합 $\mathcal{L} = \lambda_U \sum_{\mathcal{D}_U, \mathcal{D}_L} \mathcal{L}_{\text{SeqCLR}} + \lambda_L \sum_{\mathcal{D}_L} \mathcal{L}_{\text{ce}}$

Vision Model Architecture



Data Augmentation in SemiMTR



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

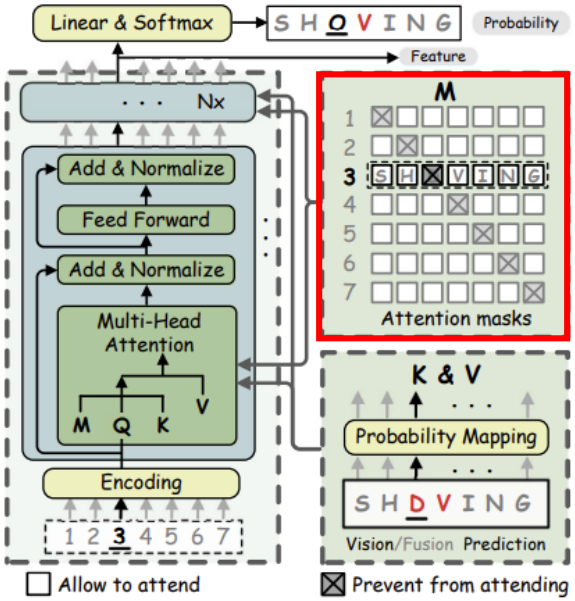
Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Model Architecture (2) Language Model Pretraining

- Masked Language Model(MLM)으로 사전학습 수행
  - ✓ MLM: 특정 Text Token을 가리고 가려진 부분의 Text Token을 맞추는 방식

    → 본 모델은 연산량이 많은 MLM 특성의 한계를 극복하기 위해 Token이 아닌 Attention Map에 Masking을 수행
  - ✓ Unlabeled data만 활용, Large Text Corpus를 활용하여 사전학습
- Vision Model과는 별개로 Text 데이터만 활용하여 학습

Language Model Architecture



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Model Architecture (3) Fine-tuning & Fusion Model Training

- 두 가지 방식을 동시에 활용하여 학습
  - ✓ 각 Modality(Vision, Language, Fusion)에 대해 Consistency Regularization으로 학습 (Unlabeled + Labeled 데이터)
  - ✓ 각 Modality(Vision, Language, Fusion) Prediction 값에 대한 Cross Entropy (Labeled data)
- 과정
  - ① 데이터 증강을 2회 수행: 이때, Vision Pretraining처럼 강력한 Color Augmentation 수행
  - ② Consistency Regularization Loss: 모든 Modality에 대해 Loss 산출, Threshold를 반영하여 Noisy한 시퀀스는 필터링

Threshold와 비교하여 Noisy 데이터는 필터링

$$\mathcal{L}_{\text{Consist}}(\mathbf{Y}^{\text{strong}}, \mathbf{Y}^{\text{weak}}) = \sum_{0 \le i < N^{\text{weak}}} \mathbf{1}(\max(\mathbf{y}_i^{\text{weak}}) > t)\,\ell(\mathbf{y}_i^{\text{strong}}, \mathbf{y}_i^{\text{weak}})$$

Consistency Loss 산출

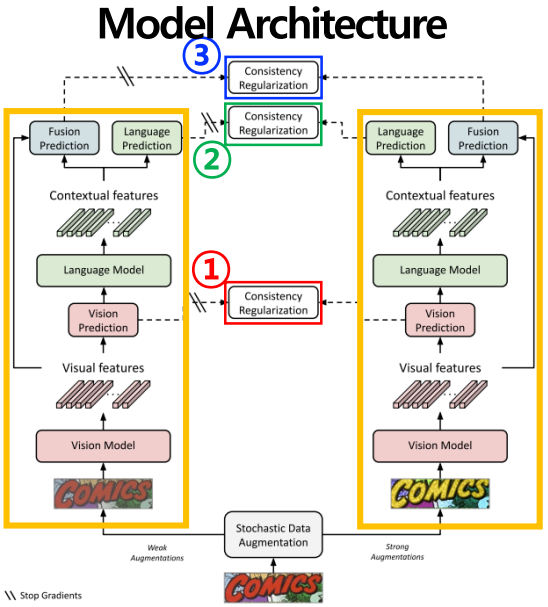
  - ③ Supervised Loss: Vision, Language, Fusion Prediction에 대해 Cross Entropy 도출

    * Stop Gradient 적용

  - ④ Overall Loss: Consistency Regularization Loss와 Supervised Loss의 가중 합

$$\sum_{d \in \text{decoders}} \lambda_{U_d} \sum_{\mathcal{D}_U, \mathcal{D}_L} \mathcal{L}_{\text{Consist}}(\mathbf{Y}_d^{\text{strong}}, \mathbf{Y}_d^{\text{weak}}) + \lambda_{L_d} \sum_{\mathcal{D}_L} \mathcal{L}_{\text{ce}}(\mathbf{Y}_d^{\text{strong}}) + \mathcal{L}_{\text{ce}}(\mathbf{Y}_d^{\text{weak}})$$

모든 Modality | Consistency Loss | Supervised Loss



Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Experiment Result

- Unlabeled 데이터를 함께 활용한 SemiMTR이 다른 모델들보다 좋은 성능을 보임
- Unlabeled 데이터를 Vision 및 Fusion에 모두 활용 시 성능 향상
- [실제 Labeled 데이터 + 실제 Unlabeled 데이터]에 합성 데이터까지 활용 시 성능 향상

→ 합성 데이터가 Labeled 데이터(합성 데이터의 1.7%)가 적은 한계를 어느정도 보완

다른 모델과 비교 실험 결과

Method	Labeled Data	Unlabeled Data	Common Benchmarks IIIT 3,000	SVT 647	IC13 1,015	IC15 2,077	SVTP 645	CUTE 288	Avg. 7,672	Non-Common Benchmarks COCO 9,835	RCTW 1,050	Uber 80,826	ArT 35,284	LSVT 4,257	MLT19 5,693	ReCTS 2,592	Avg. 139,537
PlugNet [28]	Synth	✗	94.4	92.3	95.0	82.2	84.3	85.0	89.8	-	-	-	-	-	-	-	-
RobustScanner [61]	Synth	✗	95.3	88.1	94.8	77.1	79.5	90.3	88.2	-	-	-	-	-	-	-	-
SCATTER [28]	Synth	✗	93.7	92.7	93.9	82.2	86.9	87.5	89.7	-	-	-	-	-	-	-	-
Plugnet [34]	Synth	✗	94.4	92.3	95.0	82.2	84.3	85.0	89.8	-	-	-	-	-	-	-	-
SRN [59]	Synth	✗	94.8	91.5	95.5	82.7	85.1	87.8	90.3	-	-	-	-	-	-	-	-
VisionLAN [56]	Synth	✗	95.8	91.7	95.7	83.7	86.0	88.5	91.1	-	-	-	-	-	-	-	-
TRBA [4]	Synth	✗	92.1	88.9	93.1	74.7	79.5	78.2	85.7	50.2	59.1	36.7	57.6	58.0	80.3	80.6	46.3
TRBA [5]	Real-L	✗	93.5	87.5	92.6	76.0	78.7	86.1	86.6	62.7	67.7	52.7	63.2	68.7	85.8	83.4	58.6
$TRBA_{PL}$ [5]	Real-L	Real-U	94.8	91.3	94.0	80.6	82.7	88.1	89.3	66.9	71.5	54.2	66.7	73.5	87.8	85.6	60.9
$TRBA_{PL}$ [5]	Real-L,Synth	Real-U	95.2	92.0	94.7	81.2	84.6	88.7	90.0	-	-	-	-	-	-	-	-
$TRBA_{PL}$ [5]	Real-L,Synth	Real-U	95.2	92.0	94.7	81.2	84.6	88.7	90.0	-	-	-	-	-	-	-	-
$ABINet^{Git}$ [13]	Synth	✗	96.4	93.2	95.1	82.1	89.0	89.2	91.2	63.1	59.7	39.6	68.3	59.5	85.0	86.7	52.0
ABINet* [13]	Real-L	✗	95.5	93.4	94.4	83.0	87.1	89.6	90.8	69.2	71.6	55.7	67.7	73.7	88.2	90.6	62.4
$ABINet_{PL}$* [13,5]	Real-L	Real-U	96.4	94.1	95.0	83.7	88.8	93.1	91.8	71.2	74.2	56.8	70.5	75.0	89.1	90.9	63.9
$ABINet_{est}$* [13]	Real-L	Real-U	96.5	**96.3**	95.7	83.7	89.1	92.0	92.1	71.7	73.8	56.8	70.1	75.7	89.3	91.6	63.9
SemiMTR-V	Real-L	Real-U	95.6	93.5	95.2	82.5	88.1	90.6	91.0	70.5	75.1	57.7	69.5	75.2	89.6	92.3	64.2
SemiMTR-F	Real-L	Real-U	96.5	95.4	96.5	84.2	89.6	90.6	92.3	70.9	74.9	57.7	70.3	75.5	89.3	91.5	64.4
SemiMTR	Real-L	Real-U	**96.7**	95.5	**96.6**	**83.8**	**90.5**	**93.8**	**92.4**	**72.0**	**75.8**	**58.5**	**70.8**	**77.1**	**90.3**	**92.5**	**65.2**
SemiMTR	Real-L,Synth	Real-U	**97.3**	**96.6**	**97.0**	**84.7**	**93.0**	**93.8**	**93.3**	**72.7**	**76.3**	58.4	**72.3**	**77.1**	90.2	**93.2**	**65.6**

Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

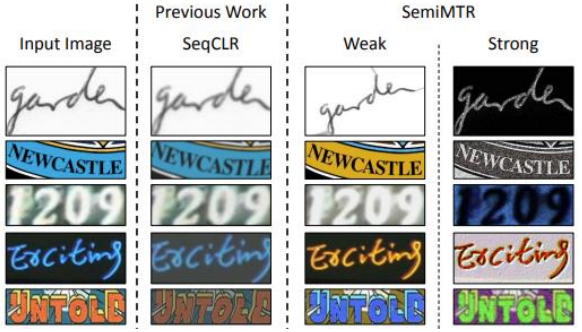
❖ Experiment Result

- Vision Model Pretrain 시 [Contrastive Loss + Supervised Loss]로 통합한 형태가 가장 효과적
  - ✓ 가장 좋은 성능
- STR에 적합하도록 Color Transformation 세기를 강하게 한 Augmentation 전략의 효과 입증

Vision Model의 Pretrain 방식과 증강기법에 대한 실험

Method	Augmentations	Vision Model	
		Common Benchmarks	Non-Common Benchmarks
Supervised Baseline [13]	ABINet [13]	85.3	57.9
Two-Stage	SeqCLR [1]	86.7	59.6
Unified	SeqCLR [1]	87.0	60.2
Unified	Ours	**88.1**	**60.3**

Data Augmentation in SemiMTR



Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Experiment Result

- Consistency Loss를 계산할 때 활용되는 5가지 요소 중 아래 요소들이 효과를 가짐
  - ✓ 목적함수: Cross Entropy
  - ✓ Stop Gradient 활용
  - ✓ One-hot Label 활용
  - ✓ Threshold를 통한 Noisy 데이터 필터링
  - ✓ 각 Modal Output에 대해 Consistency Loss 산출

Consistency Loss의 요소들에 대한 실험

Consistency Loss	Stop Gradients	Soft labels	Threshold	Common Benchmarks	Non-Common Benchmarks
CE	✓	✗	✗	91.8	**65.0**
CE	✗	✓	✗	91.8	64.1
CE	✓	✓	✗	91.8	64.5
KL Divergence	✗	✓	✗	**92.2**	64.5
CE	✓	✗	✓	**92.2**	**65.0**
CE	✗	✓	✓	92.0	**65.0**
CE	✓	✓	✓	92.0	64.7
KL Divergence	✗	✓	✓	92.1	64.8

Teacher	Student	Common Benchmarks	Non-Common Benchmarks
Vision	Vision	91.7	64.8
Vision	Language	91.9	64.5
Vision	Fusion	92.1	**65.2**
Vision	Vision, Language, Fusion	91.8	65.0
Fusion	Vision	92.0	64.7
Fusion	Language	91.7	64.6
Fusion	Fusion	92.2	65.1
Fusion	Vision, Language, Fusion	92.3	65.1
Vision, Language, Fusion	Vision, Language, Fusion	**92.4**	**65.2**

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Algorithms

Self&Semi-supervised Learning-based STR: Multimodal Semi-Supervised Learning for Text Recognition (arXiv, 2022)

❖ Summary

- STR에 Self/Semi-supervised Learning을 모두 활용하는 Framework 제안
  - ✓ Vision Model 뿐만 아니라 Language Model도 함께 활용
- 모델구조
  - ✓ Vision Model Pretraining
  - ✓ Language Model Pretraining
  - ✓ Fine-tuning & Fusion Model Training
- 기여점
  - ✓ Labeled 데이터가 부족한 한계를 Unlabeled 데이터를 활용하는 Self/Semi-supervised Learning을 모두 적용하여 극복
  - ✓ Vision Feature 뿐만 아니라 Language Feature까지 고려함으로써 기존보다 많은 정보를 활용하여 성능 향상

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

Conclusions

Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY

Conclusions

❖ Conclusions

- Scene Text Recognition
  - ✓ Scene Text Spotting = Scene Text Detection + **Scene Text Recognition**
  - ✓ 딥러닝 기반의 Scene Text Recognition에서 Input과 Output의 형태 및 모델구조 소개
  - ✓ 최근 문제상황: Insufficient Labeled Data
- Self/Semi-supervised Learning
  - ✓ Labeled 데이터가 부족한 상황에서 Unlabeled 데이터를 함께 활용할 수 있는 방법론
- Self/Semi-supervised Learning for Scene Text Recognition 사례 3가지 소개
  - ① Self-supervised Learning-based STR: STR에 Contrastive Learning을 적용한 방법론 소개
  - ② Semi-supervised Learning-based STR: STR에 Consistency Regularization을 적용한 방법론 소개
  - ③ Self&Semi-supervised Learning-based STR: STR에 Contrastive Learning + Consistency Regularization + (Multimodal Learning)을 활용

References

References

1. Aberdam, A., Ganz, R., Mazor, S., & Litman, R. (2022). Multimodal Semi-Supervised Learning for Text Recognition. arXiv preprint arXiv:2205.03873.
2. Aberdam, A., Litman, R., Tsiper, S., Anschel, O., Slossberg, R., Mazor, S., ... & Perona, P. (2021), Sequence-to-sequence contrastive learning for text recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15302-15312.
3. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., ... & Lee, H. (2019), What is wrong with scene text recognition model comparisons? dataset and model analysis, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4715-4723.
4. Chen, X., Jin, L., Zhu, Y., Luo, C., & Wang, T. (2021). Text recognition in the wild: A survey. ACM Computing Surveys (CSUR), 54(2), 1-35.
5. Zheng, C., Li, H., Rhee, S. M., Han, S., Han, J. J., & Wang, P. (2022), Pushing the Performance Limit of Scene Text Recognizer without Human Annotation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14116-14125.

감사합니다.

Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY