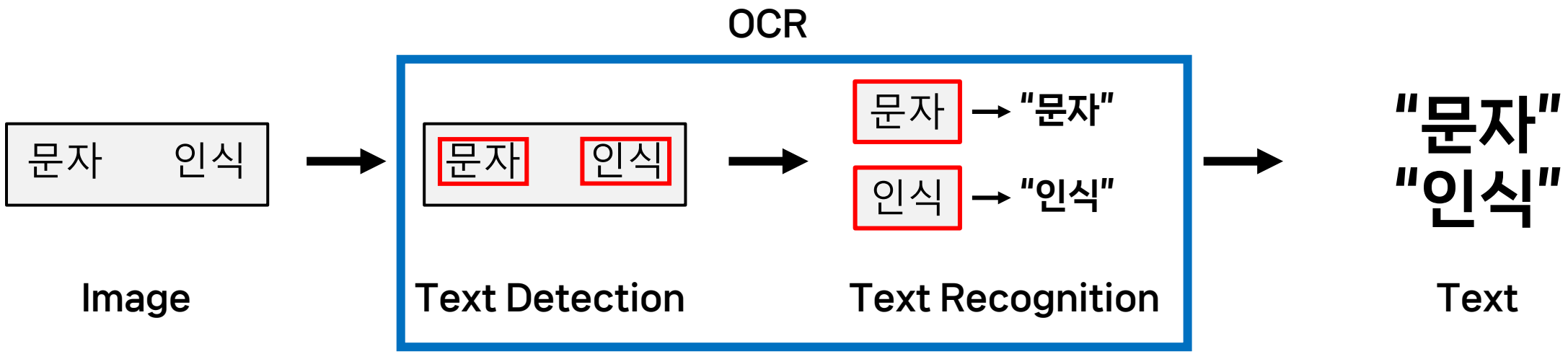
# TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models

# 목차

- Introduction
- TrOCR
- Experiments

# Introduction

- Optical Character Recognition(OCR)



# Introduction

- 기존 Text Recognition 기법
  - Encoder: CNN Backbone +  Self-Attention

    CNN Backbone : 입력 이미지를 feature map으로 변형시켜주는 부분
  - Decoder: Connectionist Temporal Classification(CTC) 사용

# Introduction

- Proposed Text Recognition model(TrOCR)
  - End-to-end Transformer based OCR model
    - Encoder - Pre-trained image transformer(ViT-style)
    - Decoder - Pre-trained text transformer(BERT-style)
      → 외부 언어 모델 필요 없이 이미지 이해 및
      언어 모델링을 위한 대규모 unlabeled data 활용
  - Backbone을 위한 Convolution network 사용 X
    → 유지보수 쉬움
  - SOTA
    - 인쇄체, 필기체, 길가 속 간판 등과 같은 글자 인식

# TrOCR

- TrOCR Architecture

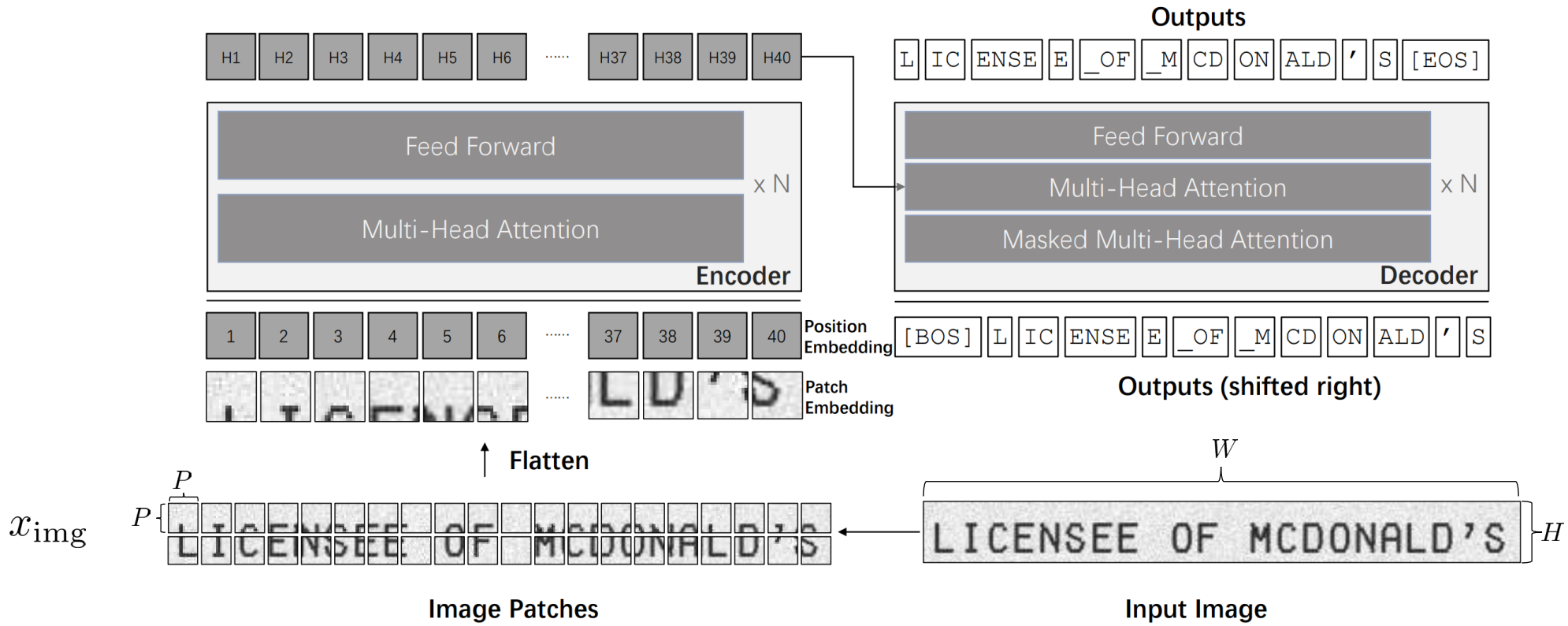



Figure 1: The architecture of TrOCR, where an encoder-decoder model is designed with a pre-trained image Transformer as the encoder and a pre-trained text Transformer as the decoder.

# TrOCR

- Model Initialization
  - Encoder Initialization
    - DeiT model
    - BEiT model
  - Decoder Initialization
    - RoBERTa model
    - MiniLM model

# TrOCR

- Task Pipeline
  - 이미지 입력 시 시각적 특징 추출 후 Wordpiece token 예측
  - 추론 시 Decoder
    - [BOS] token 부터 반복적으로 출력 예측
    - 새로 생성된 출력을 다음 입력으로 사용

# TrOCR

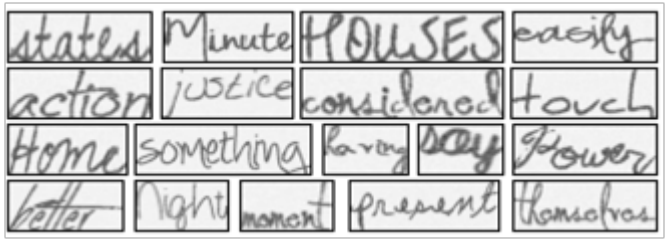
- Pre-training
  - 대규모(수억 개)의 인쇄된 Textline image 데이터
- Fine-tuning
  - Downstream text recognition tasks
  - Based on Byte Pair Encoding(BPE)
  - SentencePiece

# TrOCR

- Data Augmentation
  - Inversion
  - Curving
  - Blur
  - Noise
  - Distortion
  - Rotation

# Experiments

- Data
  - Pre-training Dataset
    - Handwritten Dataset
      - TRDG
      - IIIT-HWS dataset



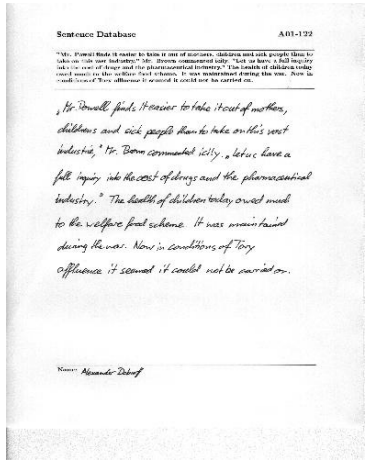


    - Text Recognition
      - Commercial OCR engines
      - MJSynth(MJ)
      - SynthText(ST)

# Experiments

- Evaluation Metrics
  - SROIE dataset
    - Precision, Recall, F1 Score


  - IAM dataset
    - Character Error Rate(CER)




  - Scene text dataset
    - Word Accuracy

# Experiments

- Results
  - Architecture Comparison

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

Table 1: Ablation study on the SROIE dataset, where all the models are trained using the SROIE dataset only.

# Experiments

- Results
  - SROIE Task2

| 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 |
| $\text{TrOCR}_{\text{SMALL}}$ | 95.89 | 95.74 | 95.82 |
| $\text{TrOCR}_{\text{BASE}}$ | 96.37 | 96.31 | 96.34 |
| $\text{TrOCR}_{\text{LARGE}}$ | 96.59 | 96.57 | **96.58** |

Table 3: Evaluation results (word-level Precision, Recall, F1) on the SROIE dataset, where the baselines come from the SROIE leaderboard (https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=2).

$\text{TrOCR}_{\text{SMALL}}$: $\text{DeiT}_{\text{SMALL}}$ + MiniLM
$\text{TrOCR}_{\text{BASE}}$ : $\text{BEiT}_{\text{BASE}}$ + $\text{RoBERTa}_{\text{LARGE}}$
$\text{TrOCR}_{\text{LARGE}}$: $\text{BEiT}_{\text{LARGE}}$ + $\text{RoBERTa}_{\text{LARGE}}$

# Experiments

- Results
  - IAM Handwriting Database

| Model | Architecture | Training Data | External LM | CER |
|---|---|---|---|---|
| (Bluche and Messina 2017) | GCRNN / CTC | Synthetic + IAM | Yes | 3.2 |
| (Michael et al. 2019) | LSTM/LSTM w/Attn | IAM | No | 4.87 |
| (Wang et al. 2020a) | FCN / GRU | IAM | No | 6.4 |
| (Kang et al. 2020) | Transformer w/ CNN | Synthetic + IAM | No | 4.67 |
| (Diaz et al. 2021) | S-Attn / CTC | Internal + IAM | No | 3.53 |
| (Diaz et al. 2021) | S-Attn / CTC | Internal + IAM | Yes | 2.75 |
| (Diaz et al. 2021) | Transformer w/ CNN | Internal + IAM | No | 2.96 |
| $\text{TrOCR}_{\text{SMALL}}$ | Transformer | Synthetic + IAM | No | 4.22 |
| $\text{TrOCR}_{\text{BASE}}$ | Transformer | Synthetic + IAM | No | 3.42 |
| $\text{TrOCR}_{\text{LARGE}}$ | Transformer | Synthetic + IAM | No | 2.89 |

Table 4: Evaluation results (CER) on the IAM Handwriting dataset.

# Experiments

- Results
  - Scene Text Datasets

| 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$_{\text{BASE}}$ (Syn) | 90.1 | 91.0 | 97.3 | 96.3 | 81.1 | 75.0 | 90.7 | 86.8 |
| TrOCR$_{\text{LARGE}}$ (Syn) | 91.0 | 93.2 | 98.3 | 97.0 | 84.0 | 78.0 | 91.0 | 89.6 |
| TrOCR$_{\text{BASE}}$ (Syn+Benchmark) | 93.4 | 95.2 | 98.4 | **97.4** | 86.9 | 81.2 | 92.1 | 90.6 |
| TrOCR$_{\text{LARGE}}$ (Syn+Benchmark) | 94.1 | **96.1** | **98.4** | 97.3 | 88.1 | 84.1 | **93.0** | **95.1** |

Table 6: Word accuracy on the six benchmark datasets (36-char), where "Syn" indicates the model using synthetic data only and "Syn+Benchmark" indicates the model using synthetic data and benchmark datasets.

# Experiments

- Results
  - Inference Speed

| Model | Parameters | Total Sentences | Total Tokens | Time | Speed #Sentences | Speed #Tokens |
|---|---|---|---|---|---|---|
| $\text{TrOCR}_{\text{SMALL}}$ | 62M | 2,915 | 31,081 | 348.4s | 8.37 sentences/s | 89.22 tokens/s |
| $\text{TrOCR}_{\text{BASE}}$ | 334M | 2,915 | 31,959 | 633.7s | 4.60 sentences/s | 50.43 tokens/s |
| $\text{TrOCR}_{\text{LARGE}}$ | 558M | 2,915 | 31,966 | 666.8s | 4.37 sentences/s | 47.94 tokens/s |

Table 5: Inference time on the IAM Handwriting dataset.

# Link

- https://github.com/microsoft/unilm/tree/master/trocr