Scene Text Recognition with Permuted Autoregressive Sequence Models (PARSeq)

Scene Text Recognition with Permuted Autoregressive Sequence Models

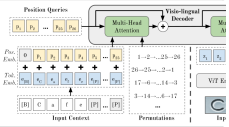
Context-aware STR methods typically use internal autoregressive (AR) language models (LM). Inherent limitations of AR models motivated two-stage methods which employ an external LM. The...

https://arxiv.org/abs/2207.06966

Papers with Code - Scene Text Recognition with Permuted Autoregressive Sequence Models

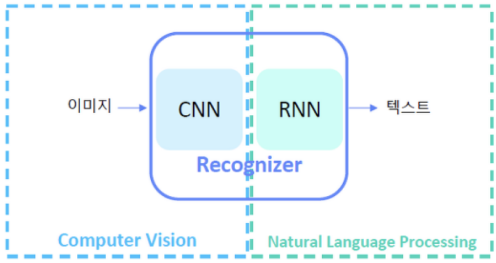
SOTA for Scene Text Recognition on ICDAR2013 (Accuracy metric)

https://paperswithcode.com/paper/scene-text-recognition-with-permuted

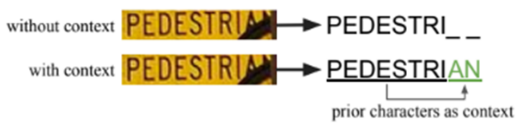


official github

1. Introduction
    a. Context-aware STR
        i. Scene Text Recognition(STR) - natural scene에서 det(감지) + rec(인식)를 통해 text reading



        ii. natural scene의 문제점으론 비정형인 텍스트의 노이즈, 변형으로 인해 acc가 낮은 문제 존재
        (이전 논문들은 context-free str 방식으로 개별 문자 단위로 인식 하였으며, 각 토큰간의 연관성을 고려하지 않아 노이즈성에 약하고 다양한 문맥이 sample로 들어오게 되면 성능 저하가 발생하는 이슈가 있었다.)
        iii. 이를 극복 하기 위해, STR + language semantics 개념(사전 언어 모델 개념,LM)이 추가
        iv. context-aware STR 방식은 문자들 간의 문맥 또는 연관성을 파악하고, 각 토큰을 분리하여 인식 하는 개념이 아닌 이전 토큰들을 활용하여 다음 토큰을 예측 하는 방식이다. (입력 시퀀스의 랜덤 순열을 문맥으로 사용한다.) → **문맥으로써 이전 문자를 예측**



    b. 이전 논문(ABINet) 같은 경우 외부 LM을 사용하였으며, 초기 모델 예측과 외부 LM의 예측을 비교하여 image feature, seq feature 간의 결합 과정을 통해 더 높은 확률을 가진 값을 결과로 출력

(**Vision - context free, Lanuage - externel LM**)

(a) ABINet (b) Unified STR model (Ours)

좌 - ABINet, 우 - PARSeq

2. Related Work

a. ABINet

i. IARLM 방법론

- Internal Autoregressive Language Model 뜻으로 내부적으로 언어 지식을 자기 회귀하는 구조이다.
- 단방향성 특징을 가지고 있으며, PAST to Future(무조건 전 - 후)로만 학습 / 예측이 가능하다.
- 주 문제로는 단방향에 치우진 학습과 예측, 토큰에 대한 예측은 학습에 사용되는 방향으로만 결과를 얻을 수 있다. (**학습 방향에 의존적인 예측을 할 수 밖에 없다는 문제 존재**)

ii. ABINet 방법론

- IARLM의 방향 의존적인 문제점을 극복하기 위해 등장
- 내부가 아닌 외부 LM 모델을 통해 context free model이 예측한 결과값과 비교하여 spelling을 체크한다.
  (vision - context free, language - externel LM)
- context free 개념은 문맥에 상관 없이 독립적으로 작동하는 모델을 의미
  (이전 문장 또는 내용을 고려하지 않고, 각각의 단어 또는 문장을 독립적으로 처리하는 모델)
- 앙상블 모델 개념이며, Fusion layer에서 vision + language의 결과값을 결합하여 조금 더 정확하고 잘못된 예측에 대해 다시 한 번 체크 할 수 있어, 성능이 향상 된다.
- LM의 파라미터 효율성 부족이라는(많은 양의 파라미터 개수 활용 불가) 단점이 존재한다
  (파라미터 효율적 활용 불가 쪽은 분리된 학습? 토큰마다 학습? 예측을 해서 파라미터의 수를 다 활용 못하는 건지..?)

b. Permuted Autoregressive Sequence Models (PARSeq)

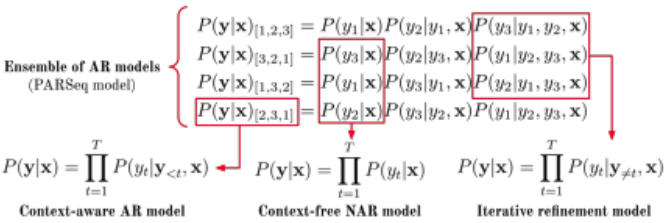
i. PLM (the cat is sitting)

1. Permutation Language Modeling은 transformer 구조에 적합하여 학습에 사용되고, Autoregressive Language Modeling (ARLM)에서 파생된 개념으로 기본적으로 입력 시퀀스의 순서를 임의로 섞어서 모델을 훈련
   → 각 토큰 간의 의존성을 낮추고 예측 확률이 상승 할 수 있게 된다. (scene text의 다양한 노이즈 및 방향성을 학습하는 데 있어 획기적인 방법론이다.)
   → 여기서 모델은 순열(입력 시퀀스를 섞는)된 입력에서 원래 시퀀스를 복원하는 방식으로 훈련
2. PLM 방식을 활용하여 Transformer model을 train 하고, Attention Mask를 활용하여 각 시퀀스의 순서를 관리하며, 이를 통해 주어진 입력 문맥의 임의의 부분 집합에 대해 출력 위치를 추론하는 능력을 모델에 부여
3. attention mask는 각 토큰들 간의 상호 작용, 연관성을 결정 짓게 해주는 기능을 가지며, 순열된 입력을 원래의 순서대로 복원하는 역할을 수행한다. (이를 통해 모델이 더욱 robust text 인식 능력을 갖춤)

ii. PARSq

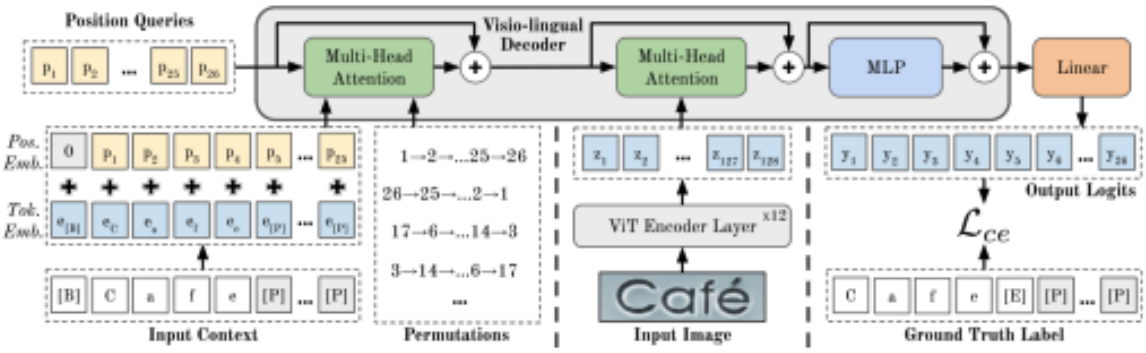
1. 토큰의 순서가 아닌 입력 시퀀스의 순서를 랜덤하게 섞는 개념으로 적용 했으며, 이미지의 일부 영역에서만 텍스트를 인식하는 데 있어,
   필요한 정보가 다른 영역에 흩어져 있을 경우에도 높은 acc를 가질 수 있게 된다. (robost, uniformly)

2. 아래 입력 시퀀스를 섞는 경우는 입력 x가 3개 주어 졌을 경우이며 총 6가지의 경우의 수가 발생 하는데, 그 중 4가지의 랜덤 토큰을
고른 것이며 가장 좌측의 4가지의 경우의 수다.



Unify models using AR ensemble

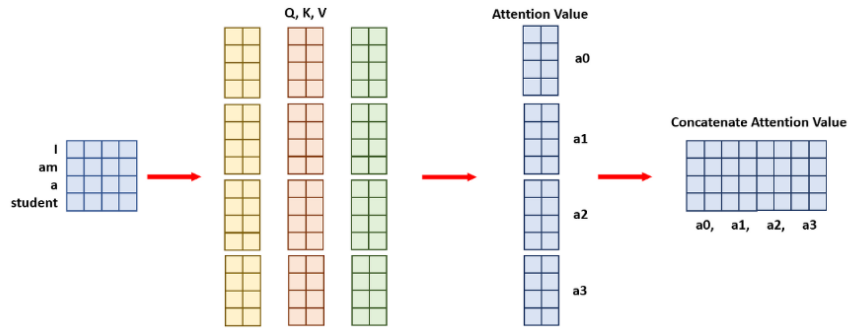
3. Learning PARSeq with Permutation Language Modeling



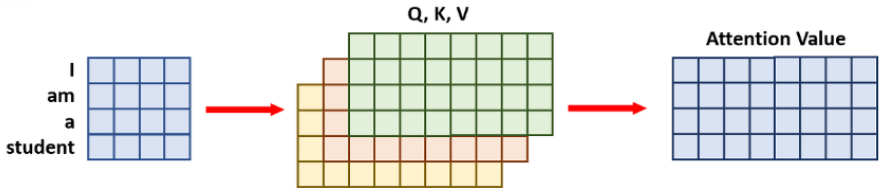
1. ViT Encoder
    a. ViT Encoder는 Transformer 구조에서 이미지 처리 시 주로 사용 되는 모델 (PARSeq Encoder는 12개의 ViT Layer들이 존재)
    b. 동작 순서
        i. **Patch Embedding을 통해 고정된 size의 patch로 나누고, 각 patch를 단일 vector로 flatting (16x16, 32x32 pixel patch)**
        ii. flatting되어 있는 vector를 초기 임베딩 레이어를 통해 패치 임베딩으로 변환시켜 줌
        iii. 1,2번의 동작 이후, 얻어진 연속된 토큰을 Transformer의 입력으로 사용
    c. self-attention(각 토큰 마다의 가중치를 할당하여 연관성을 지니게 함)을 위해 1개의 MHA Module이 존재 (q=k=v) - query, key, value
        i. **Self-Attention에서 q=k=v는 입력 토큰을 다른 특성을 가진 공간으로 매핑 하기 위해 쓰이며, 세 개의 서로 다른 학습 가능한 가중치 행렬을 통해 연산하게 됨**
        ii. Query, Key, Value 행렬 값을 head 수만큼 분할 한 후, Attention value값들을 도출하며, 해당 value를 통해 concatenate 하여 result data를 출력

$$\mathbf{z} = Enc(\mathbf{x}) \in \mathbb{R}^{\frac{WH}{p_w p_h} \times d_{model}}$$

2. Visio-lingual Decoder
    a. ViT Encoder에서 생성된 토큰 vector를 입력으로 받은 후, MHA Layer를 통과 시킴



multi head attention(MHA) 메커니즘



일반적인 attention 메커니즘

b. Encoder에서 받은 input들을 하나로 병합

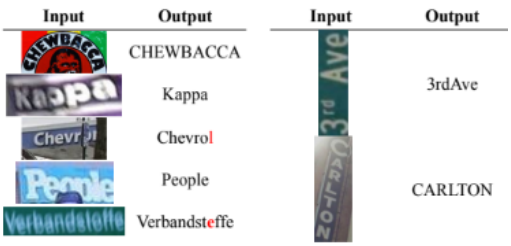
c. input data size만큼의 context를 생성한 것이 예측한 output 토큰이다.
→ 입력 텍스트를 조건부 확률 과정을 통해 이미지 토큰을 처리하고, 주어진 입력에 대해 가장 가능성(확률)이 높은 출력 문자열을 생성

4. Results and Analysis

a. SOTA in STR benchmarks

	Method	Conf.	Train data	36-char word acc. 7,248 samples	36-char word acc. 7,672 samples
Previous Work	SRN	CVPR'20	MJ,ST	90.4	–
	TextScanner	AAAI'20	MJ,ST+	–	91.0
	Bhunia *et al.*	ICCV'21	MJ,ST	–	90.9
	VisionLAN	ICCV'21	MJ,ST	91.2	–
	PREN2D	CVPR'21	MJ,ST	91.5	–
	ABINet	CVPR'21	MJ,ST+	92.7	–
Ours	$PARSeq_N$		MJ,ST	92.0±0.2	90.7±0.2
	$PARSeq_A$		MJ,ST	**93.2**±0.2	**91.9**±0.2
	$PARSeq_N$		*real*	95.7±0.1	95.2±0.1
	$PARSeq_A$		*real*	**96.4**±0.0	**96.0**±0.0

b. Robust vs occlusion and arbitrary orientation



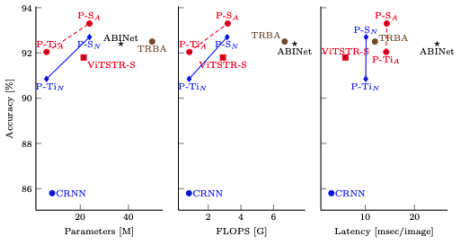
c. challenging datasets && acc, flops, latency Performance (PARSeq is flexible, accurate, and efficient)

Table 4. Mean word accuracy on the benchmark vs evaluation charset size

Method	Train data	36-char	62-char	94-char
CRNN	S	83.2±0.2	56.5±0.3	54.8±0.2
ViTSTR-S	S	88.6±0.0	69.5±1.0	67.7±1.0
TRBA	S	90.6±0.1	71.9±0.9	69.9±0.8
ABINet	S	89.8±0.2	68.5±1.1	66.4±1.0
PARSeq$_N$	S	90.7±0.2	72.5±1.1	70.5±1.1
PARSeq$_A$	S	**91.9**±0.2	**75.5**±0.6	**73.0**±0.7
CRNN	R	88.5±0.1	87.2±0.1	85.8±0.1
ViTSTR-S	R	94.3±0.1	92.8±0.1	91.8±0.1
TRBA	R	95.2±0.2	93.7±0.1	92.5±0.1
ABINet	R	95.2±0.1	93.7±0.1	92.4±0.1
PARSeq$_N$	R	95.2±0.1	93.7±0.1	92.7±0.1
PARSeq$_A$	R	**96.0**±0.0	**94.6**±0.0	**93.3**±0.1

Table 5. 36-char word accuracy on larger and more challenging datasets

Method	Train data	Test datasets and # of samples: ArT 35,149	COCO 9,825	Uber 80,551	Total 125,525
CRNN	S	57.3±0.1	49.3±0.6	33.1±0.3	41.1±0.3
ViTSTR-S	S	66.1±0.1	56.4±0.5	37.6±0.3	47.0±0.2
TRBA	S	68.2±0.1	61.4±0.4	38.0±0.3	48.3±0.2
ABINet	S	65.4±0.4	57.1±0.8	34.9±0.3	45.2±0.3
PARSeq$_N$	S	69.1±0.2	60.2±0.8	39.9±0.5	49.7±0.3
PARSeq$_A$	S	**70.7**±0.1	**64.0**±0.9	**42.0**±0.5	**51.8**±0.4
CRNN	R	66.8±0.2	62.2±0.3	51.0±0.2	56.3±0.2
ViTSTR-S	R	81.1±0.1	74.1±0.4	78.2±0.1	78.7±0.1
TRBA	R	82.5±0.2	77.5±0.2	81.2±0.3	81.3±0.2
ABINet	R	81.2±0.1	76.4±0.1	71.5±0.7	74.6±0.4
PARSeq$_N$	R	83.0±0.2	77.0±0.2	82.4±0.3	82.1±0.2
PARSeq$_A$	R	**84.5**±0.1	**79.8**±0.1	**84.5**±0.1	**84.1**±0.0



5. Summary

a. scene text data 내에서 다양한 방향 및 순서의 Text를 인식에 대해 새로운 방법론을 제시한 STR 논문 (OCR) → PARSeq, SOTA

b. 기존 CNN + RNN의 결합 방식의 문제점인 고정 적인 문자의 순서 대신, Feature의 Pixel마다 독립적인 seq(Context-aware STR)로 처리함으로써, 다양한 방향을 가진 글자 Sample에도 높은 인식률을 가진다.

c. 인식률은 이전 논문보다 향상된 것은 증명 됐으나, seq 순서 optimization에 대한 기준이 없다. (text seq opt 같은 경우, 강화 학습 기반으로 주어진 환경에 맞게 seq를 최적화 하는 것도 연구 해볼만한 가치가 있는 것 같습니다.)

1	**PARSeq**	98.4±0.2	✓	Scene Text Recognition with Permuted Autoregressive Sequence Models			2022
2	**MATRN**	97.9	×	Multi-modal Text Recognition Networks: Interactive Enhancements between Visual and Semantic Features			2021
3	**S-GTR**	97.8	✓	Visual Semantics Allow for Textual Reasoning Better in Scene Text Recognition			2021
4	**DPAN**	97.7	×	Look Back Again: Dual Parallel Attention Network for Accurate and Robust Scene Text Recognition			2021
5	**CDistNet** (Ours)	97.67	×	CDistNet: Perceiving Multi-Domain Character Distance for Robust Text Recognition			2021
6	**SVTR-L** (Large)	97.2	×	SVTR: Scene Text Recognition with a Single Visual Model			2022
7	**SVTR-B** (Base)	97.1	×	SVTR: Scene Text Recognition with a Single Visual Model			2022
8	**Yet Another Text Recognizer**	96.8	×	Why You Should Try the Real Data for the Scene Text Recognition			2021
9	**SVTR-T** (Tiny)	96.3	×	SVTR: Scene Text Recognition with a Single Visual Model			2022
10	**SVTR-S** (Small)	95.7	×	SVTR: Scene Text Recognition with a Single Visual Model			2022