Donut: Document Understanding Transformer without OCR

ECCV 2022

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Senghyun Park

Index

Introduction	Optical Character Recognition (OCR) Visual Document Understanding (VDU) 기존의 OCR-based 방법의 문제점
Method	Document Understanding Transformer Pre-Training Application
Experiments	Document Classification/Parsing/VQA
Conclusion	

Introduction

Optical Character Recognition (OCR)

> 사람이 쓰거나 기계로 인쇄한 문자의 이미지를 이미지 스캐너로 획득하여 기계가 읽을 수 있는 문자로 변환하는 것

- Text Detection + Text Recognition

Visual Document Understanding (VDU)

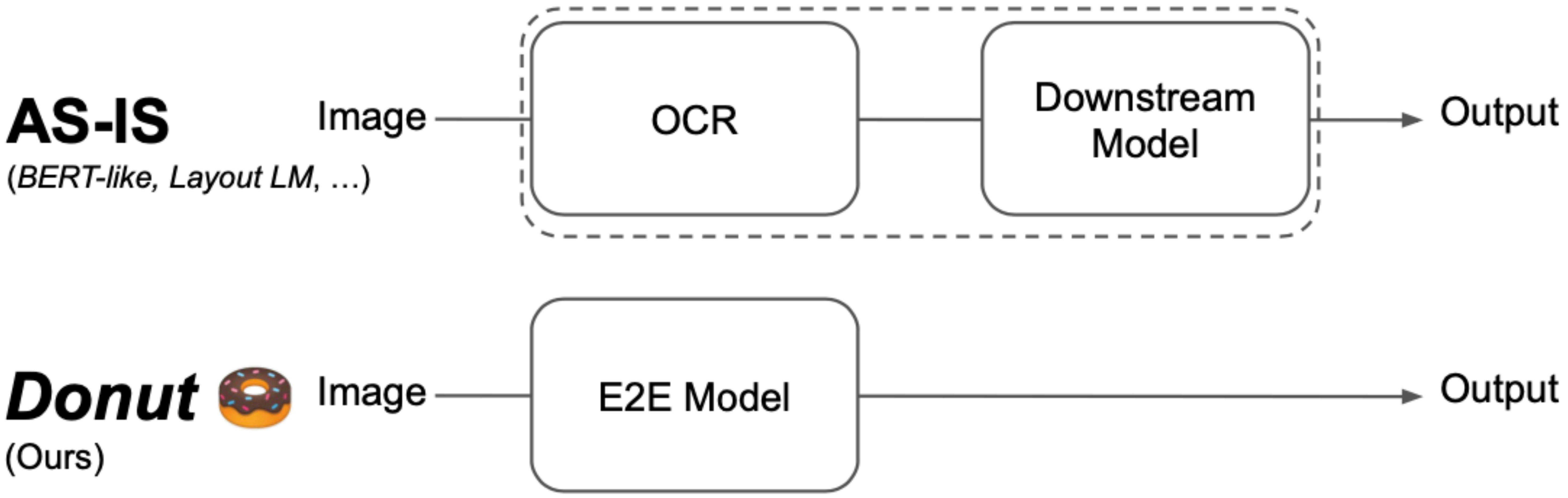
> 다양한 Formats, Layouts, Contents를 가진 Document Image를 이해하는 것이 목적
>
> e.g. Document Classification, Parsing, Visual Question Answering(VQA)

- Invoices, Receipts, Business Card와 같은 Semi-Structured Document를 주로 활용
- Semi-Structured Document는 Digital Files, Scanned Images, Photographs로 존재함

Introduction

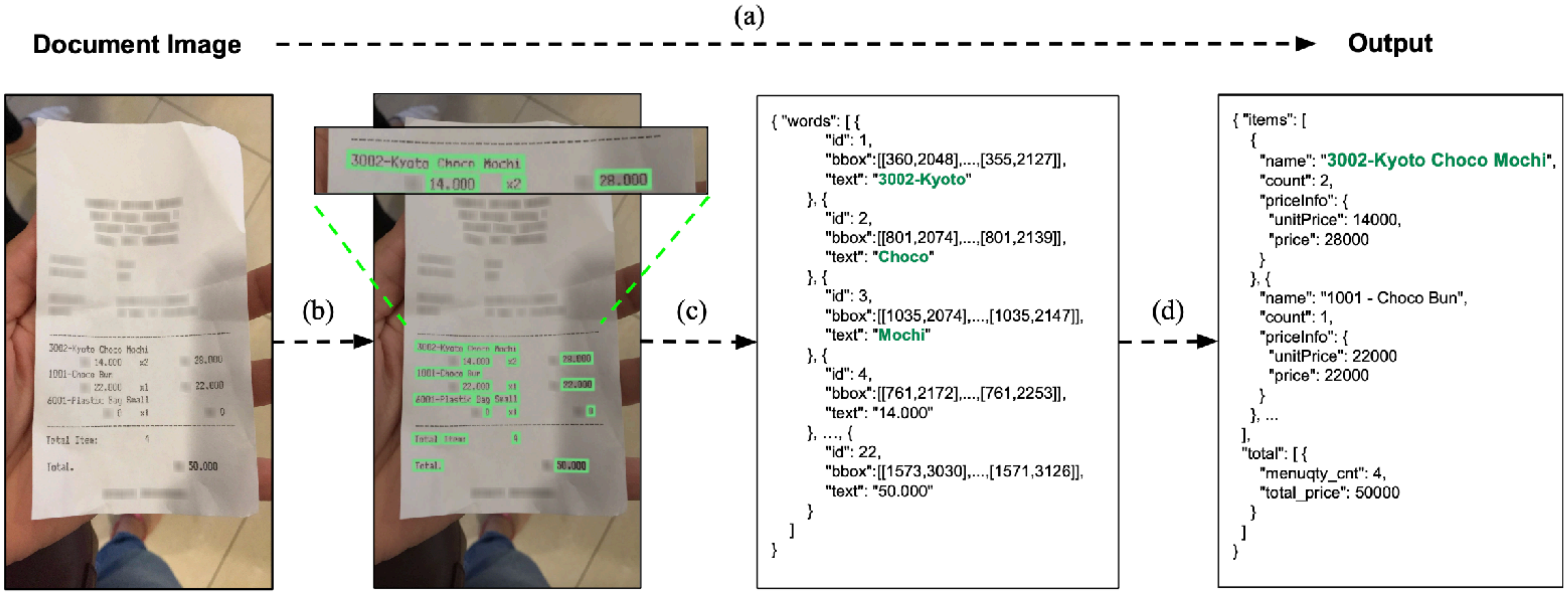
기존의 OCR-based 방법의 문제점

- Expensive Computational Costs
- Performance Degradation due to the OCR Error Propagation



Introduction

기존의 OCR-based 방법의 문제점



Introduction

기존의 OCR-based 방법의 문제점

- Expensive Computational Costs

✓ Own OCR Model의 학습을 위해서는 추가적인 Supervision과 Large-Scale Dataset이 필요함

✓ 최근의 Model은 학습을 위해 비싸고, 유지 비용을 증가시키는 GPU를 필요로 함

✓ 기존의 OCR Engine을 활용하여 비용을 줄일 수 있지만, Target Domain의 성능은 보장하지 못함

Introduction

기존의 OCR-based 방법의 문제점

- Performance Degradation due to the OCR Error Propagation

  ✓ OCR Model에서 발생하는 Error는 Subsequent Processes에 부정적인 영향을 미침

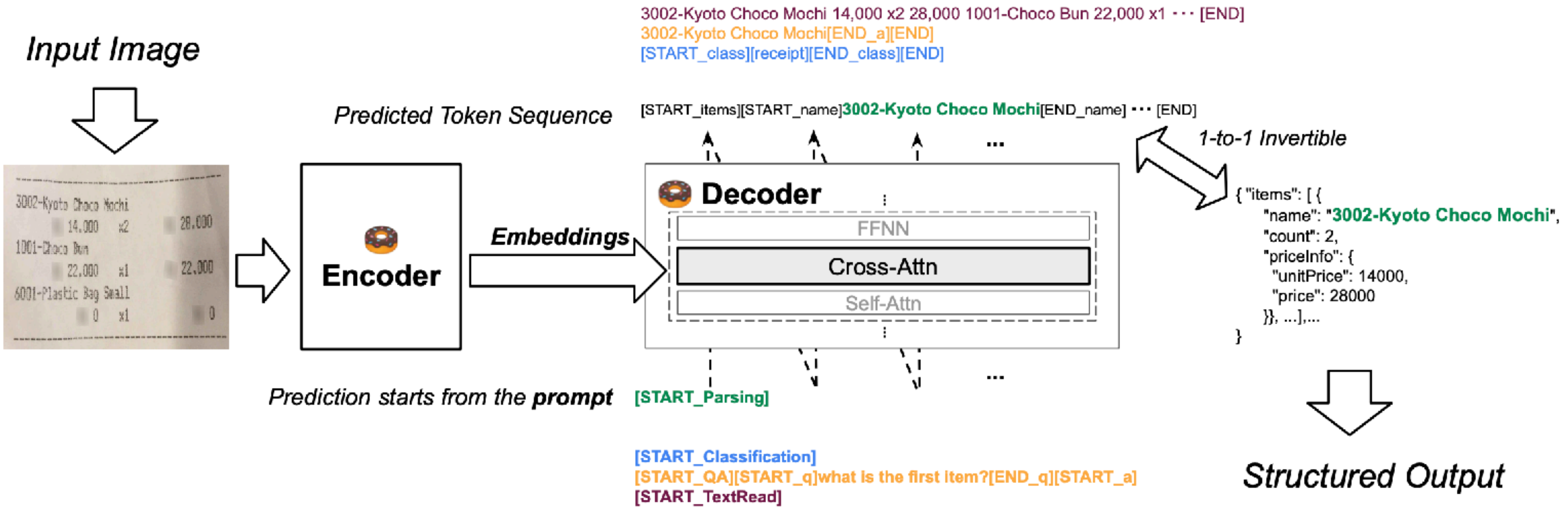
  ✓ 이는 복잡하고 큰 Character Sets를 가진 언어(e.g. Korean, Japanese)에서는 심각함

  ✓ 별도의 Post-OCR Correction Module을 활용할 수는 있지만, 전체 시스템의 크기와 유지 보수 비용 문제 때문에 실질적으로 좋은 방법이 아님

Introduction

Donut: A Simple OCR-free Transformer Architecture (End-to-End Manner)

SynthDoG: Synthetic Document Generator



Method

Document Understanding Transformer

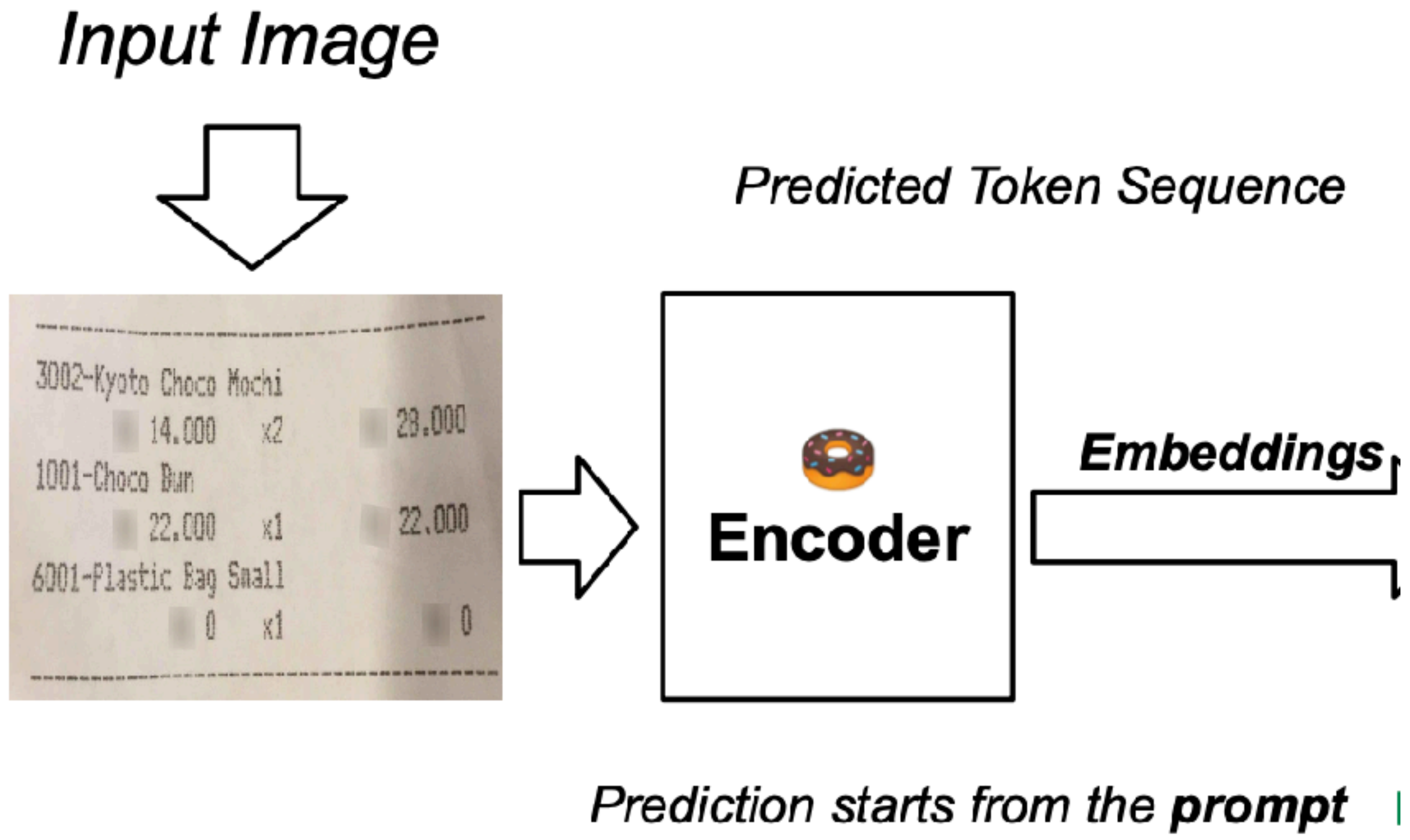
Encoder: Swin Transformer

Input_Image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

Encoder_I/O $\{\mathbf{z}_i | \mathbf{z}_i \in \mathbb{R}^d, 1 \leq i \leq n\}$,

n : Feature Map Size or Number of Image Patches

d : Dimension of the Latent Vectors



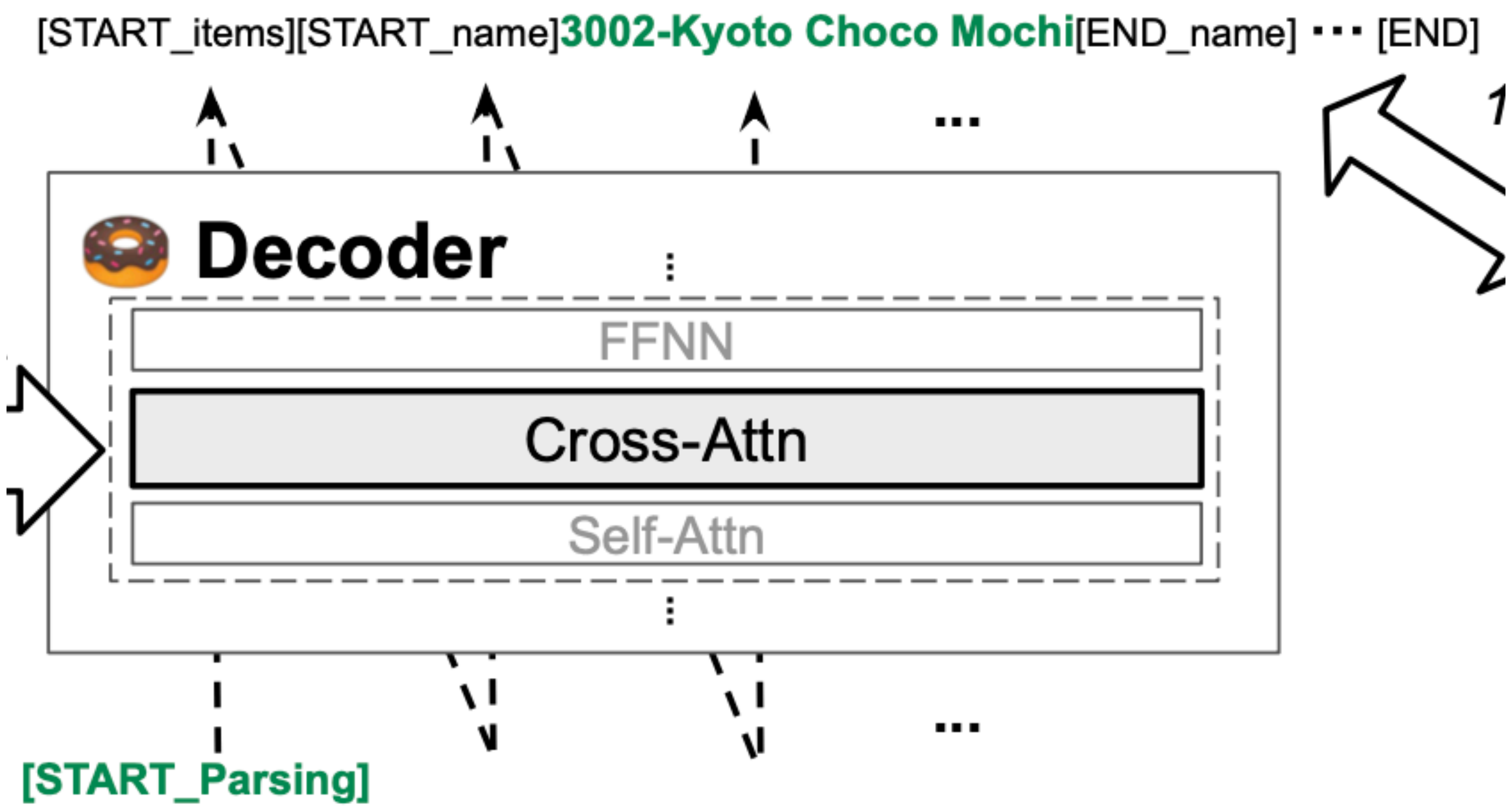
Method

Document Understanding Transformer

Decoder: Multilingual BART Decoder (use the first 4 layers)

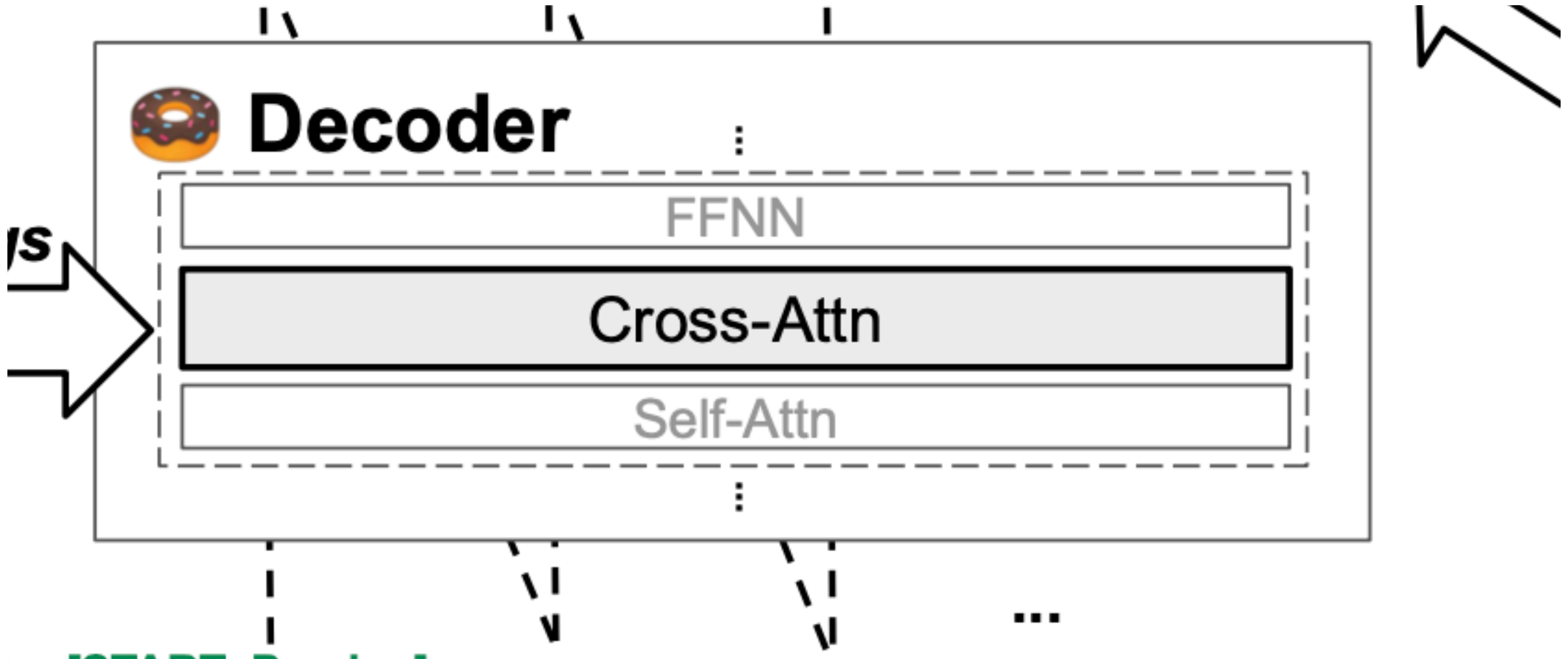
Decoder_Output $(\mathbf{y}_i)_1^m,\ \mathbf{y}_i \in \mathbb{R}^v$

i : One – hot Vector for the Token
v : Size of Token Vocabulary
m : HyperParameter



Method

Document Understanding Transformer



Model Input

[START_Parsing]

[START_Classification]
[START_QA][START_q]what is the first item?[END_q][START_a]
[START_TextRead]

Train Phase: Teacher Forcing Manner

Test Phase: Prompt/Special Tokens for Each Downstream Task (Like GPT-3)

Method

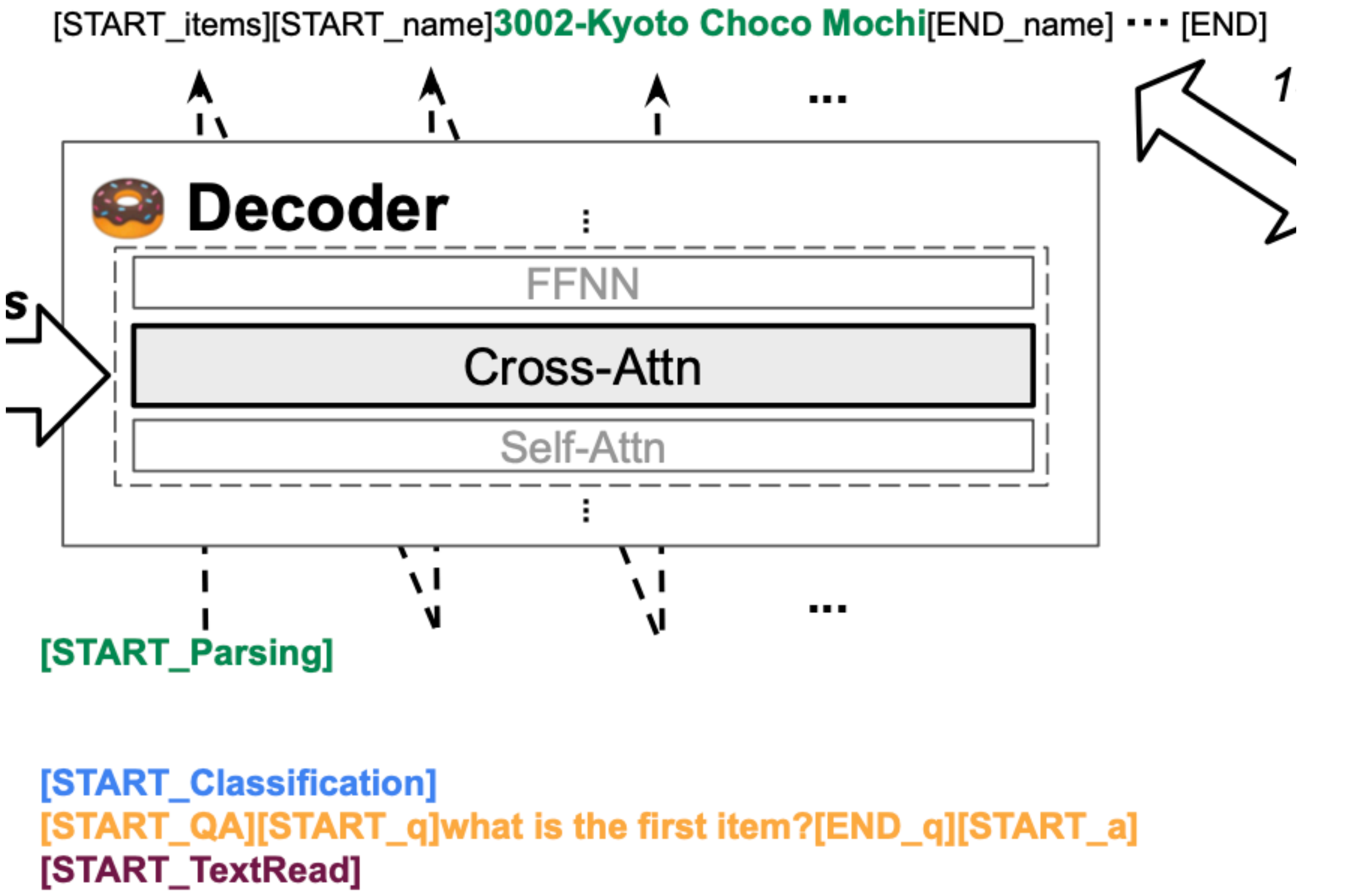
Document Understanding Transformer

<u>Output Conversion</u>

JSON Format

Special Token: [START_*] and [END_*]

[START_name]은 있는데, [END_*]이 없으면 Regular Expression을 통해
name field는 손실되었다고 가정



Method

Pre-Training - How to Read

<u>Synthetic Document Generator (SynthDog)</u>

1.2M Synthetic Document Images 생성

Wikipedia(English, Korean, Japanese)에서 추출된 Corpus 활용

각 언어마다 400K Images 생성

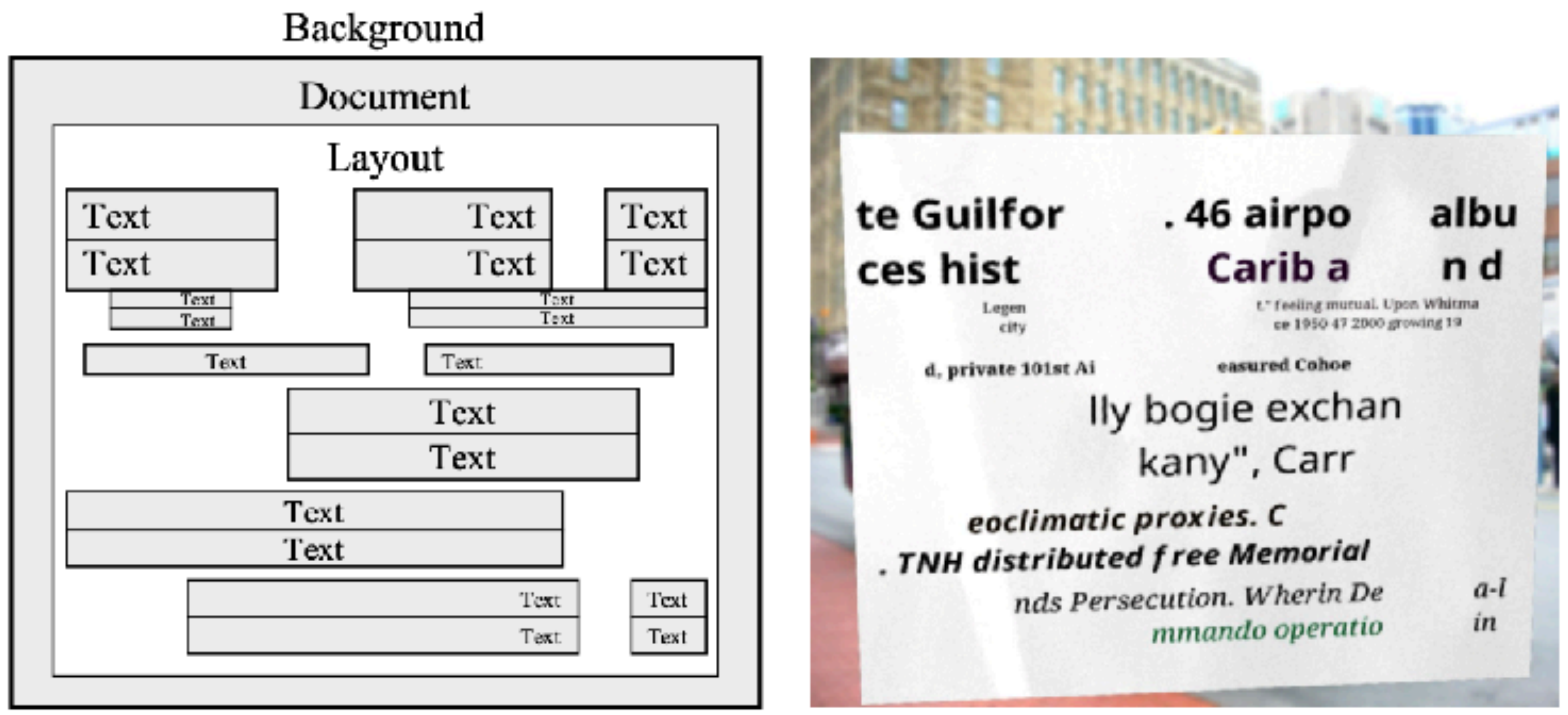
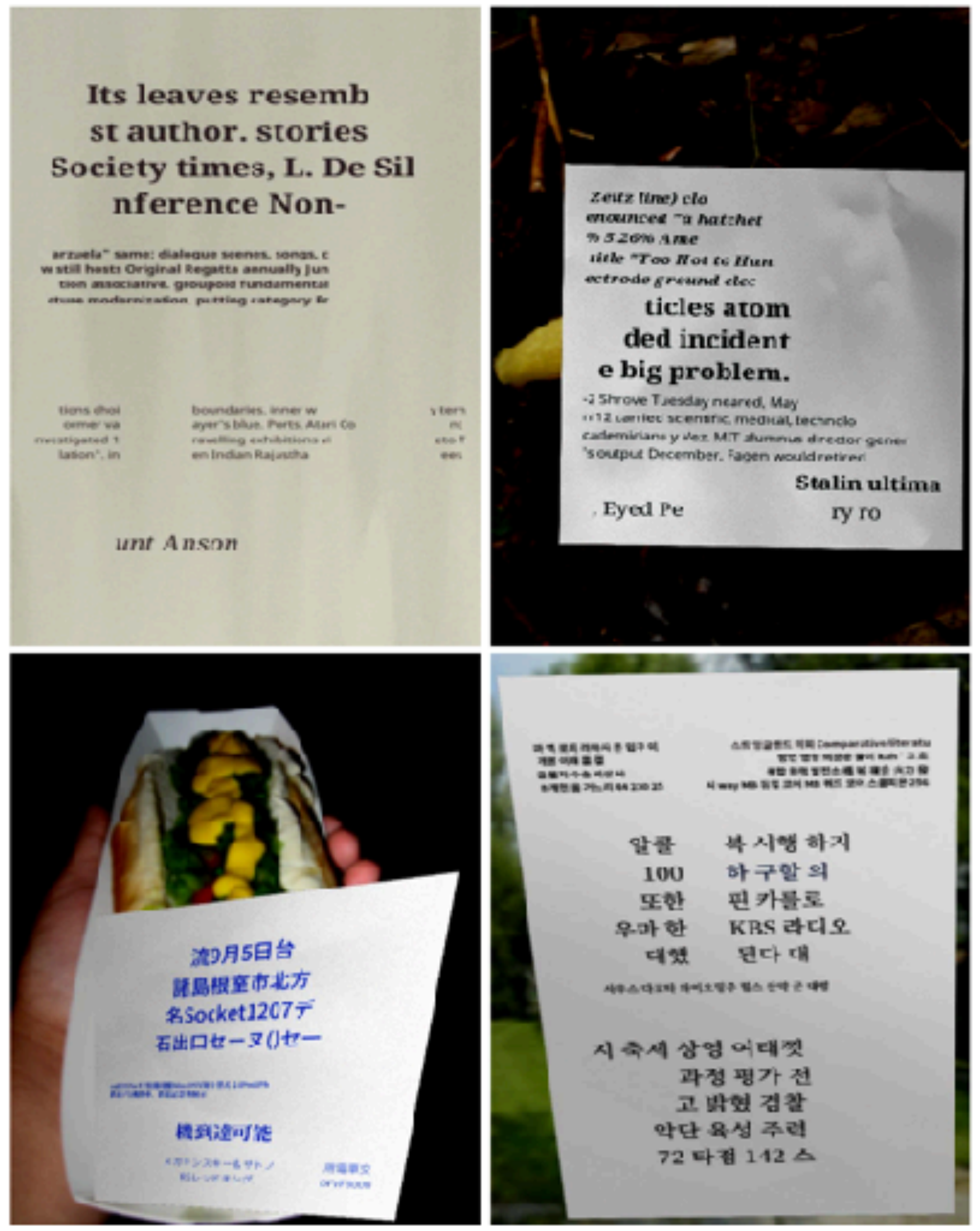
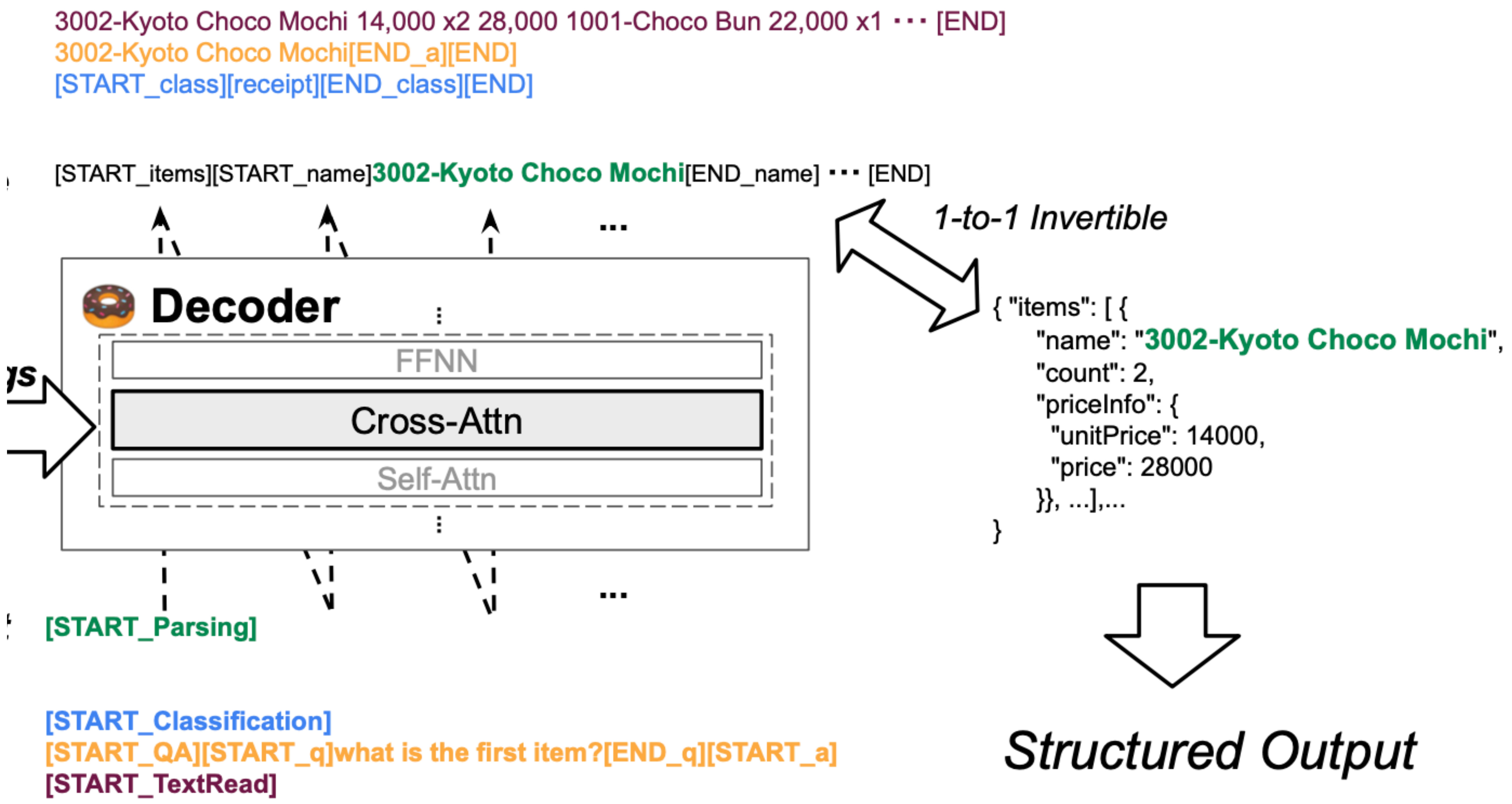


Figure 4: The components of **SynthDoG**.



Method

Application (i.e. Fine-Tuning) - How to Understand



Experiments

Downstream Tasks and Datasets

Document Classification	RVL-CDIP
Document Parsing	Indonesian Receipts Japanese Business Cards Korean Receipts
Document VQA	DocVQA

Experiments

Document Classification

	use OCR	#Params	Time(ms)	Accuracy (%)
$\text{BERT}_{\text{BASE}}$	✓	110M + n/a†	1392	89.81
$\text{RoBERTa}_{\text{BASE}}$	✓	125M + n/a†	1392	90.06
$\text{UniLMv2}_{\text{BASE}}$	✓	125M + n/a†	n/a	90.06
$\text{LayoutLM}_{\text{BASE}}$ (w/ image)	✓	160M + n/a†	n/a	94.42
$\text{LayoutLMv2}_{\text{BASE}}$	✓	200M + n/a†	1489	**95.25**
Donut (Proposed)		156M	**791**	94.50

† Parameters for OCR should be considered for the non-E2E models.

Experiments

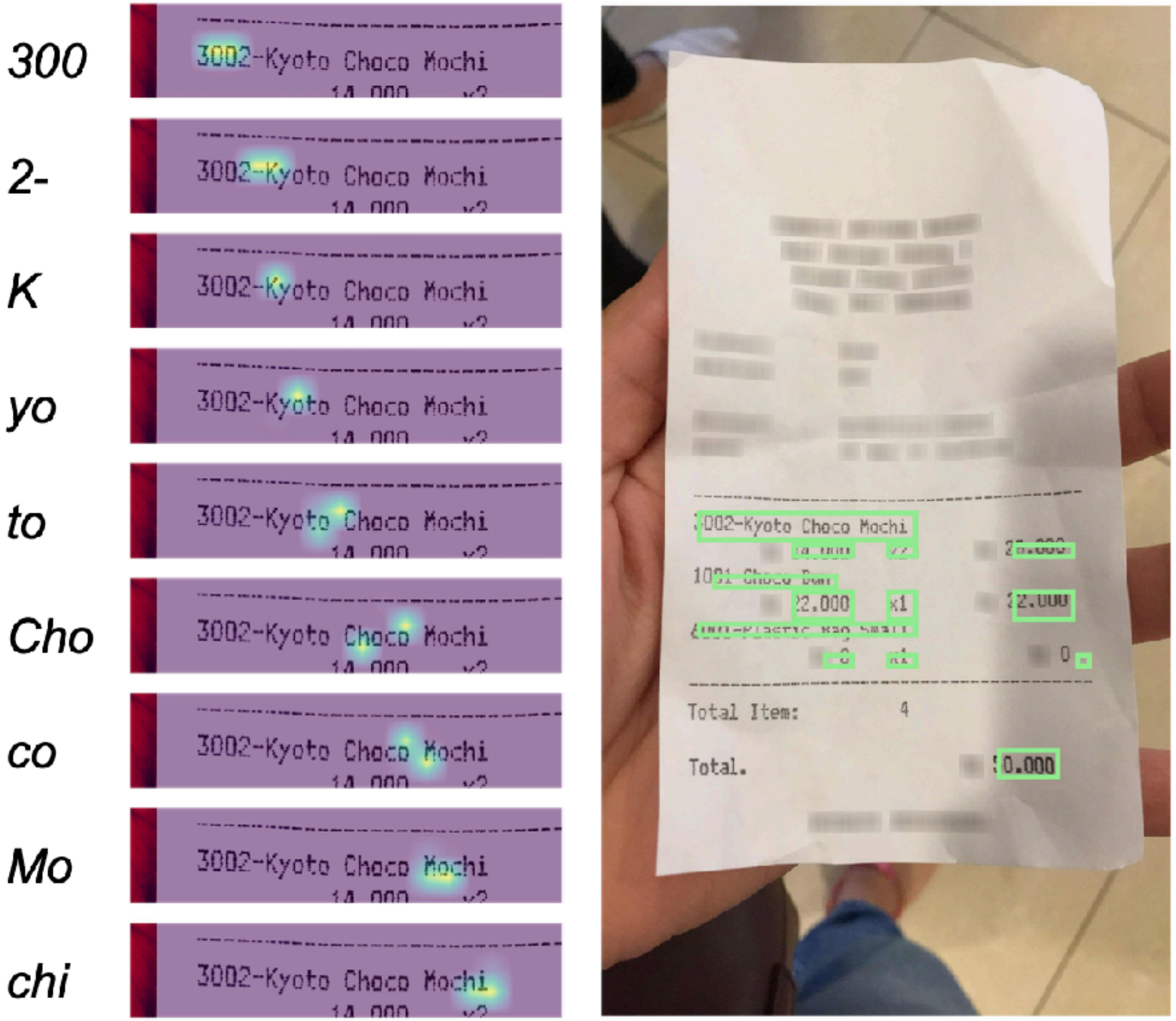
Document Parsing

	use OCR	Params	Indonesian Receipt		Korean Receipt		Japanese Business Card	
			Time (s)	nTED	Time (s)	nTED	Time (s)	nTED
BERT-based Extractor*	✓	86M† + n/a‡	0.89 + 0.54	11.3	1.14 + 1.74	21.67	0.83 + 0.50	9.56
SPADE (Hwang et al., 2021b)	✓	93M† + n/a‡	3.32 + 0.54	10.0	6.56 + 1.74	21.65	3.34 + 0.50	9.77
Donut (Proposed)		156M†	1.07	**8.45**	1.99	**5.87**	1.39	**3.70**

* Our currently-deployed model for parsing business cards and receipts in our real products. The pipeline is based on Hwang et al. (2019).
† Parameters for token (vocabulary) embeddings are omitted for a fair comparison.
‡ Parameters for OCR should be considered for non-E2E models.

Experiments

Document Parsing



Experiments

Document VQA

	OCR	Params‡	Time (ms)	ANLS
LoRRA	✓	∼223M	n/a	11.2
M4C	✓	∼91M	n/a	39.1
$\text{BERT}_{\text{BASE}}$	✓	110M	n/a	57.4
CLOVA OCR	✓	n/a	$\gtrsim$ 3226	32.96
UGLIFT v0.1	✓	n/a	$\gtrsim$ 3226	44.17
$\text{BERT}_{\text{BASE}}$	✓	110M + n/a†	1517	63.54
$\text{LayoutLM}_{\text{BASE}}$	✓	113M + n/a	1519	69.79
$\text{LayoutLMv2}_{\text{BASE}}$	✓	200M + n/a	1610	78.08
Donut		∼207M	809	47.14
+ 10K imgs of trainset				53.14

† Parameters for OCR should be considered for non-E2E models.
‡ Token embeddings for English is counted for a fair comparison.

Conclusion

Donut (End-to-End Method for VDU)

- maps an input document image into a desired structured output
- does not depend on OCR and large-scale real document images (unlike traditional)

SynthDoG (Synthetic Document Generator)

- important role in pre-training of the model
- gradually trained the model from how to read to how to understand