Slide

NAVER CLOVA

OCR-free Document Understanding Transformer

Geewook Kim[1] *, Teakgyu Hong[4], Moonbin Yim[2], Jeongyeon Nam[1], Jinyoung Park[5], Jinyeong Yim[6], Wonseok Hwang[7], Sangdoo Yun[3], Dongyoon Han[3], Seunghyun Park[1]

* gwkim.rsrch@gmail.com

[1]NAVER CLOVA [2]NAVER Search [3]NAVER AI Lab [4]Upstage [5]Tmax [6]Google [7]LBox

This is a slide for a paper “OCR-free Document Understanding Transformer”

Slide

NAVER CLOVA

OCR-free Document Understanding Transformer

Geewook Kim[1] *, Teakgyu Hong[4], Moonbin Yim[2], Jeongyeon Nam[1], Jinyoung Park[5], Jinyeong Yim[6], Wonseok Hwang[7], Sangdoo Yun[3], Dongyoon Han[3], Seunghyun Park[1]

* gwkim.rsrch@gmail.com

[1]NAVER CLOVA [2]NAVER Search [3]NAVER AI Lab [4]Upstage [5]Tmax [6]Google [7]LBox

which will be presented at

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022 October 23-27, 2022, Tel Aviv

Slide

Agenda

1. Introduction: Background and Motivation
2. Proposal: Document Understanding Transformer (Donut 🍩)
3. Experiments and Analyses
4. Conclusions

This is the table of contents. Let’s start from the Introduction.

Slide

Visual Document Understanding (VDU)

VDU Model

Useful Information

VDU aims to extract useful information from the document image. For example,

Slide

Example 1: Document Classification

VDU Model

{ "class": "receipt" }

A document classifier aims to extract a category information from the image.

Slide

Example 2: Document Parsing

VDU Model

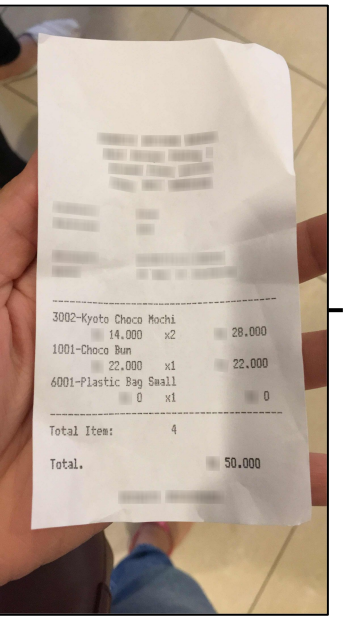

```
{ "menu": [
    {
      "nm": "3002-Kyoto Choco Mochi",
      "unitprice": "14.000",
      "cnt": "x2",
      "price": "28.000"
    }, … }
```


For another example, a document parser aims to get a data in a format, such as, JSON or XML, that contains full information.

Slide

Conventional VDU Model

Input



Output

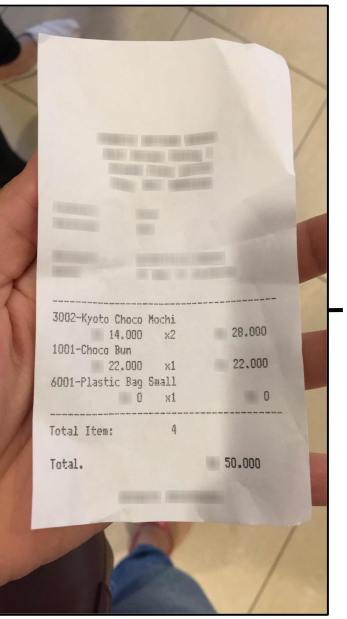

```
{ "items": [
    {
      "name": "3002-Kyoto Choco Mochi",
      "count": 2,
      "priceInfo": {
        "unitPrice": 14000,
        "price": 28000
      }
    }, {
      "name": "1001 - Choco Bun",
      "count": 1,
      "priceInfo": {
        "unitPrice": 22000
        "price": 22000
      }
    }, ...
  ],
  "total": [ {
      "menuqty_cnt": 4,
      "total_price": 50000
    }
  ]
}
```


Here, we show a representative pipeline of visual document parsing.

Slide

Conventional VDU Model

Input



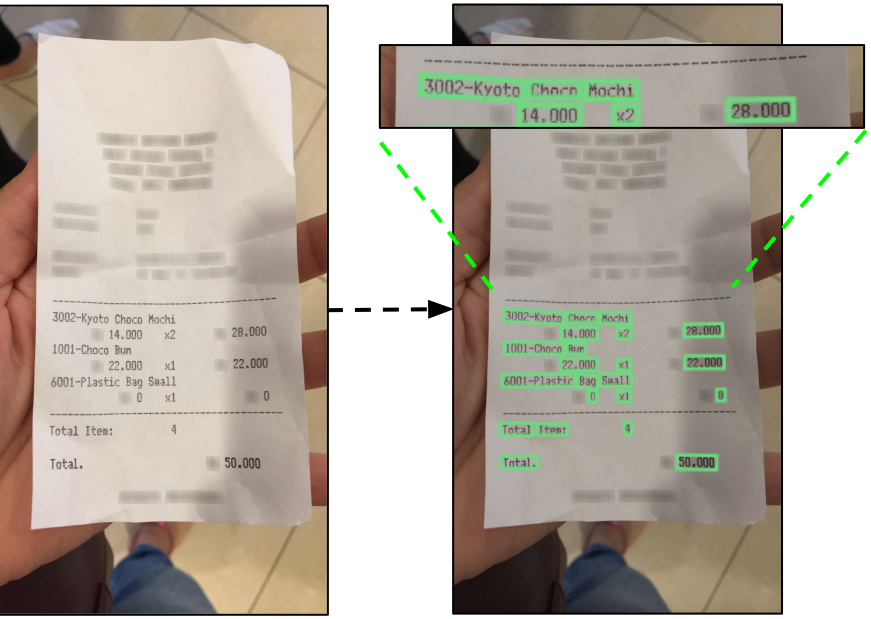
Output


```
{ "items": [
    {
      "name": "3002-Kyoto Choco Mochi",
      "count": 2,
      "priceInfo": {
        "unitPrice": 14000,
        "price": 28000
      }
    }, {
      "name": "1001 - Choco Bun",
      "count": 1,
      "priceInfo": {
        "unitPrice": 22000
        "price": 22000
      }
    }, ...
  ],
  "total": [ {
      "menuqty_cnt": 4,
      "total_price": 50000
    }
  ]
}
```


Most conventional VDU methods share a similar pipeline.

Slide

Conventional VDU Model

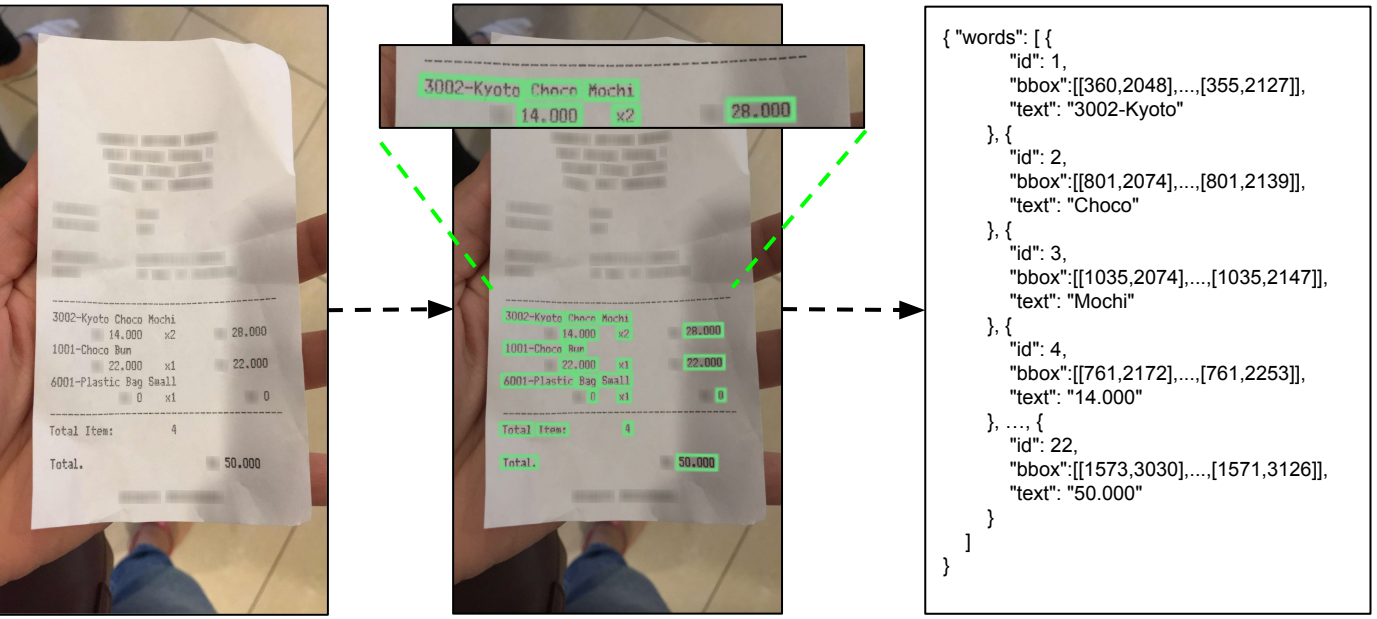


Detection!

First, a text detector finds all text boxes.

Slide

Conventional VDU Model

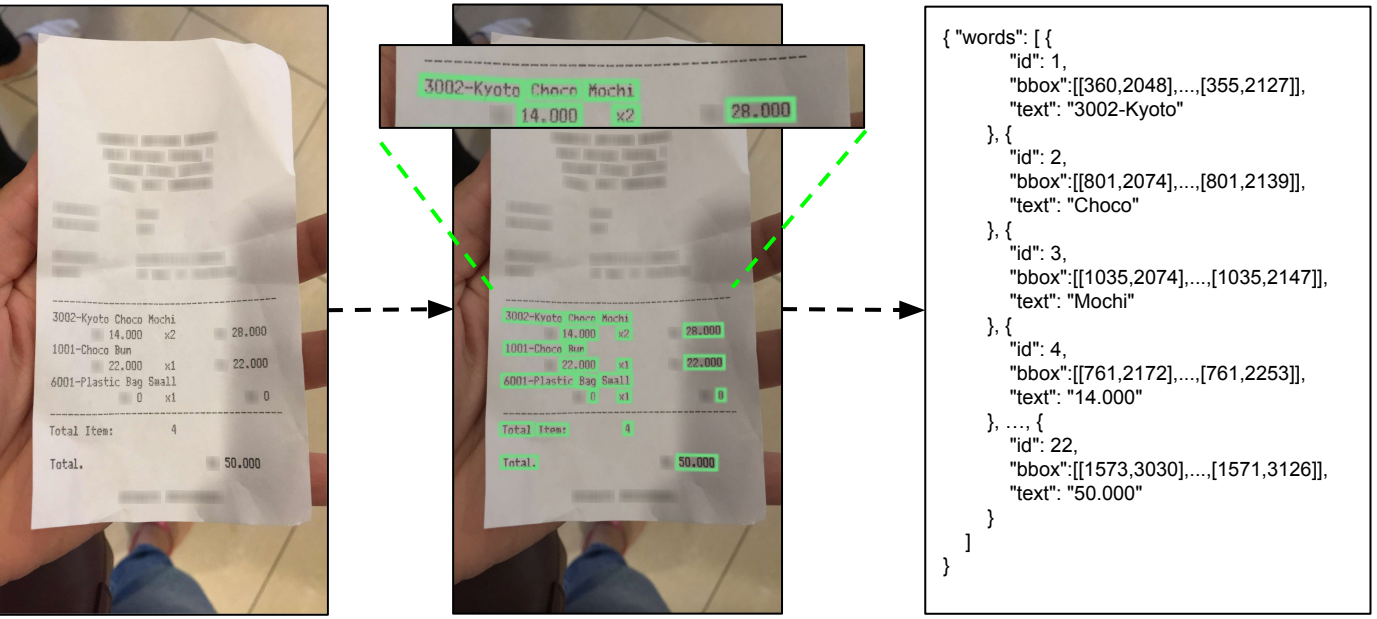


Detection! Recognition!

And then, a text recognizer reads all texts in the extracted boxes.

Slide

Conventional VDU Model




```
{ "words": [ {
        "id": 1,
        "bbox":[[360,2048],...,[355,2127]],
        "text": "3002-Kyoto"
    }, {
        "id": 2,
        "bbox":[[801,2074],...,[801,2139]],
        "text": "Choco"
    }, {
        "id": 3,
        "bbox":[[1035,2074],...,[1035,2147]],
        "text": "Mochi"
    }, {
        "id": 4,
        "bbox":[[761,2172],...,[761,2253]],
        "text": "14.000"
    }, …, {
        "id": 22,
        "bbox":[[1573,3030],...,[1571,3126]],
        "text": "50.000"
    }
  ]
}
```

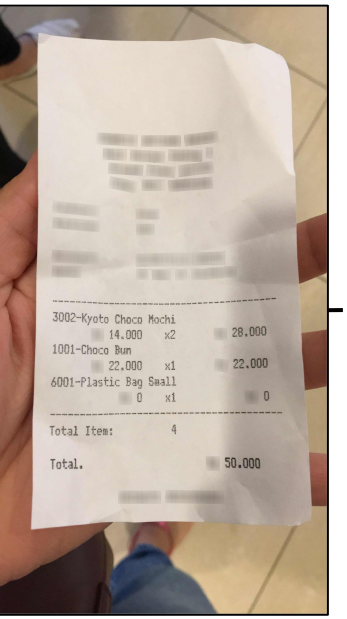

Detection! Recognition!

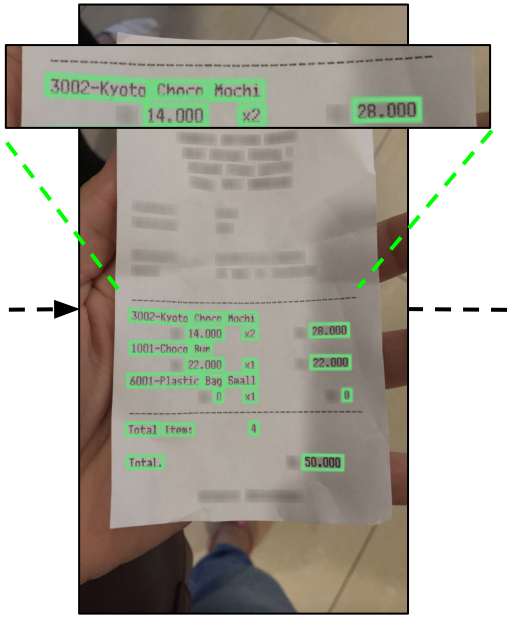
OCR

This two parts are also called as Optical Character Recognition (OCR).

Slide

Conventional VDU Model






```
{ "words": [ {
        "id": 1,
        "bbox":[[360,2048],...,[355,2127]],
        "text": "3002-Kyoto"
    }, {
        "id": 2,
        "bbox":[[801,2074],...,[801,2139]],
        "text": "Choco"
    }, {
        "id": 3,
        "bbox":[[1035,2074],...,[1035,2147]],
        "text": "Mochi"
    }, {
        "id": 4,
        "bbox":[[761,2172],...,[761,2253]],
        "text": "14.000"
    }, …, {
        "id": 22,
        "bbox":[[1573,3030],...,[1571,3126]],
        "text": "50.000"
    }
  ]
}
```



```
{ "items": [
   {
    "name": "3002-Kyoto Choco Mochi",
    "count": 2,
    "priceInfo": {
     "unitPrice": 14000,
     "price": 28000
    }
   }, {
    "name": "1001 - Choco Bun",
    "count": 1,
    "priceInfo": {
     "unitPrice": 22000
     "price": 22000
    }
   }, ...
  ],
  "total": [ {
    "menuqty_cnt": 4,
    "total_price": 50000
   }
  ]
}
```

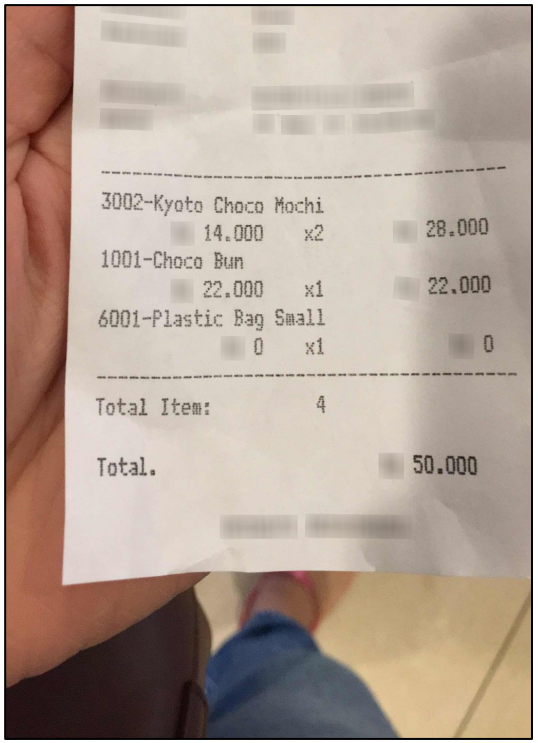

Detection! Recognition! Parsing!

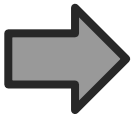
OCR

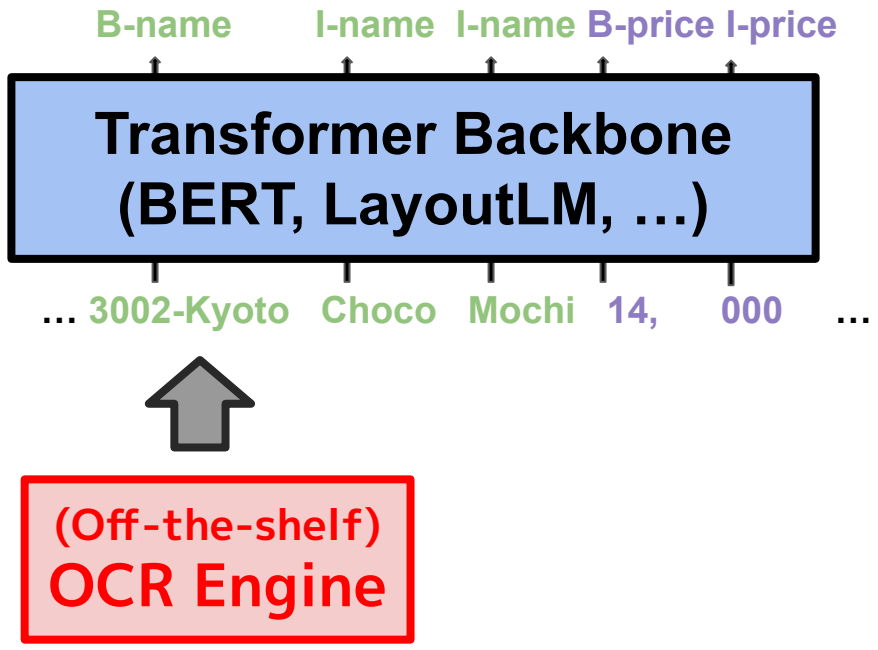
Finally, the OCR results are fed to a following module to get full information of the document.

Slide

Conventional VDU Model: Details of the Parsing Stage



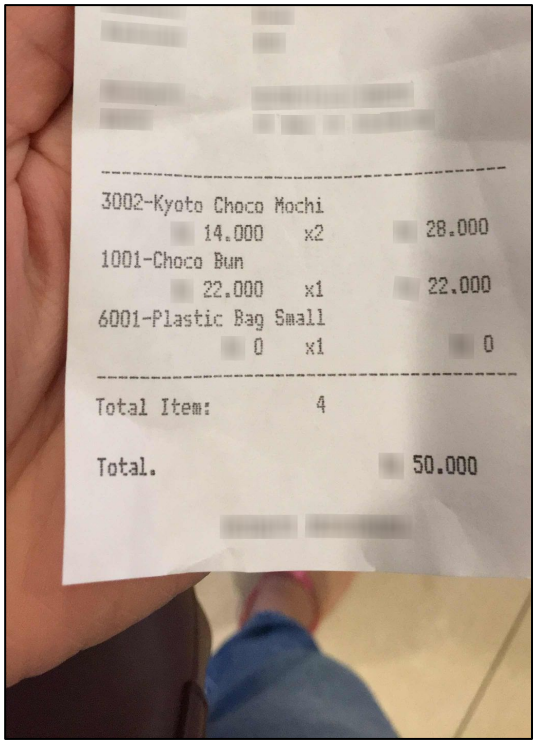


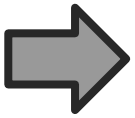


For example, in most methods, BIO-tags are predicted by a backbone.

Slide

Conventional VDU Model: Details of the Parsing Stage





(Off-the-shelf)
OCR Engine

… 3002-Kyoto Choco Mochi 14, 000 …

Transformer Backbone
(BERT, LayoutLM, …)

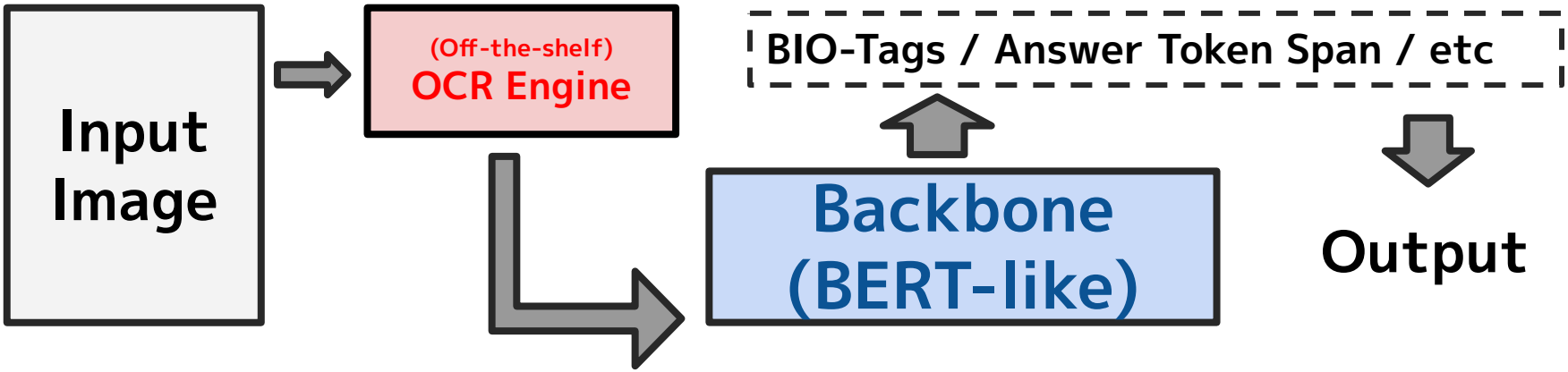
B-name I-name I-name B-price I-price

```
{ "menu": [
   {
    "nm": "3002-Kyoto Choco Mochi",
    "unitprice": "14.000",
    "cnt": "x2",
    "price": "28.000"
   }, … }
```

Then, the tag sequence is converted into a final data format (e.g., JSON).

Slide

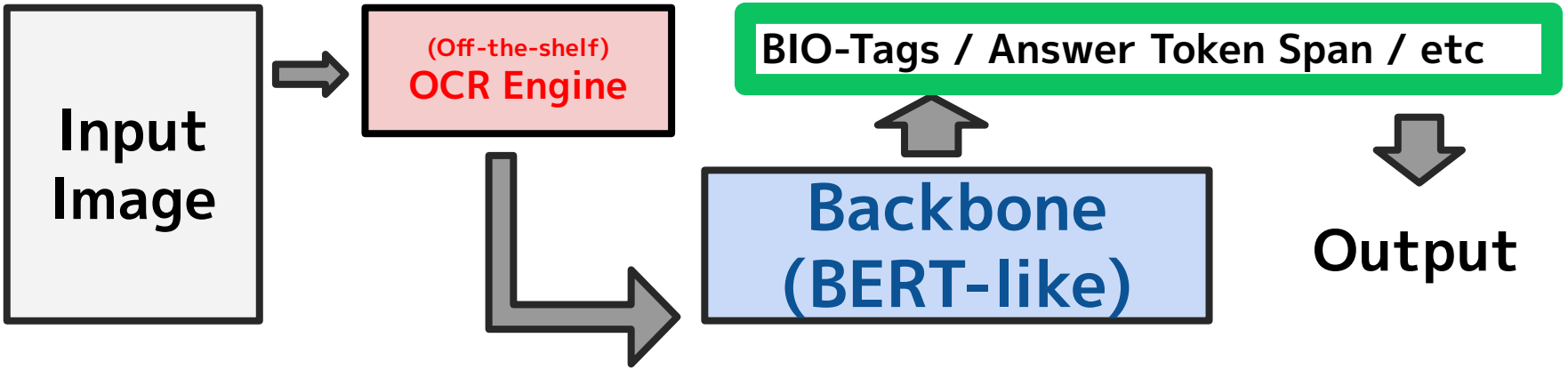
Conventional VDU Model: Overview



Overall, the conventional VDU methods can be summarized as shown in the figure.

Slide

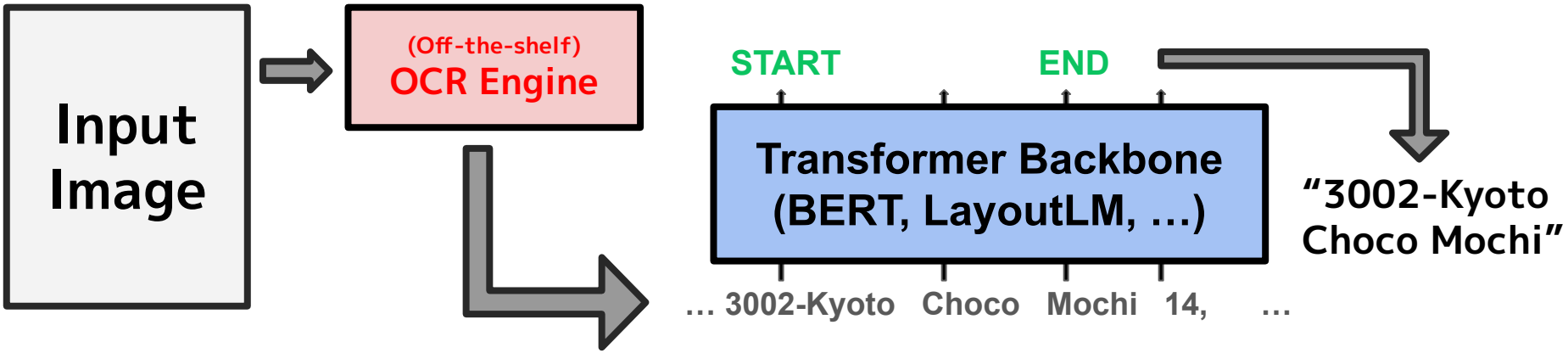
Conventional VDU Model: Overview



For each task, the backbone predicts a desired set of tokens/tags.

Slide

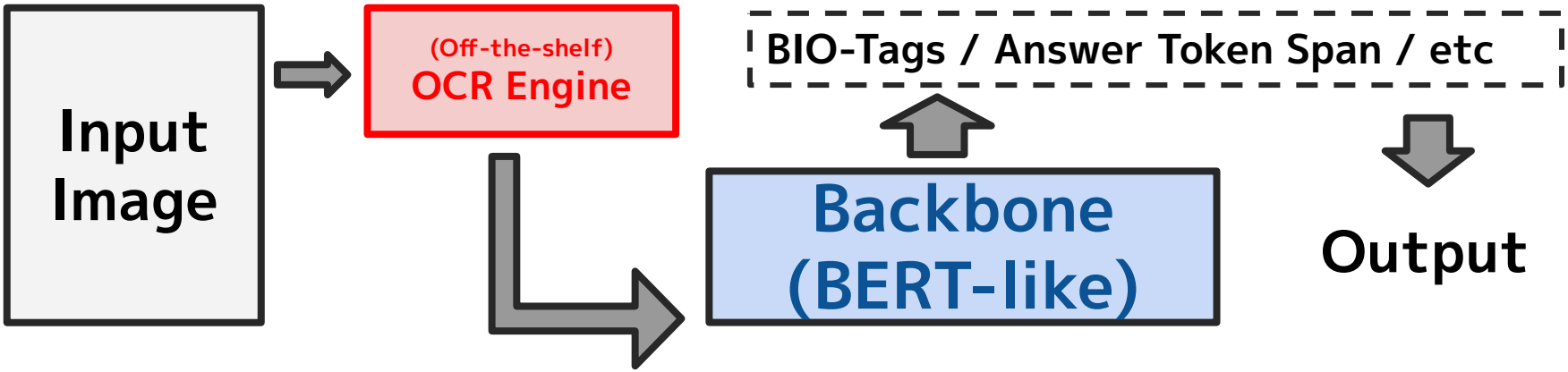
Conventional VDU Model: Overview



For example, in order to conduct VQA,
a span is predicted over the OCR-ed text tokens.

Slide

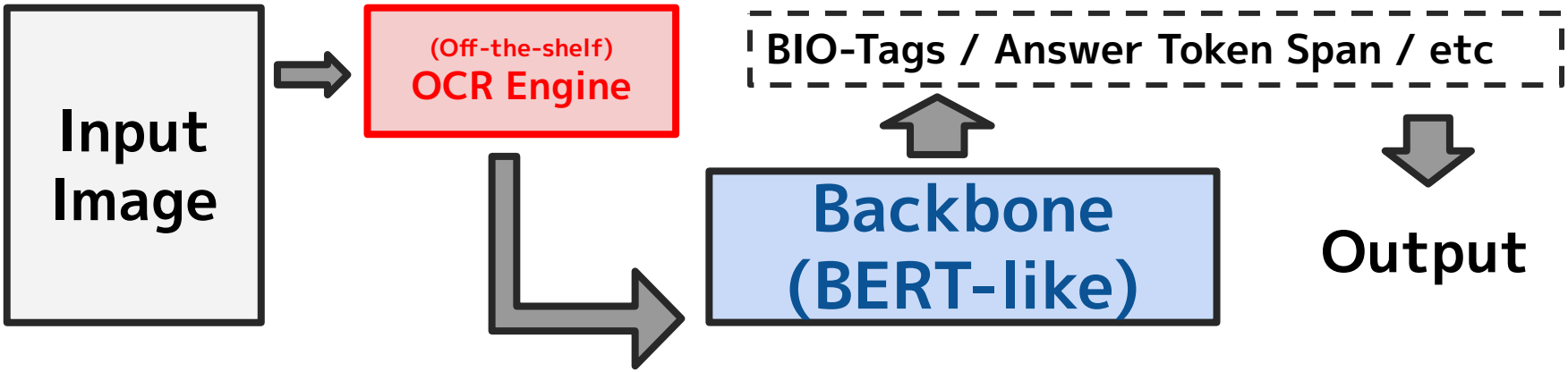
Conventional VDU Model: Overview



Although such OCR-based approaches have shown promising performance, they have several problems induced by OCR.

Slide

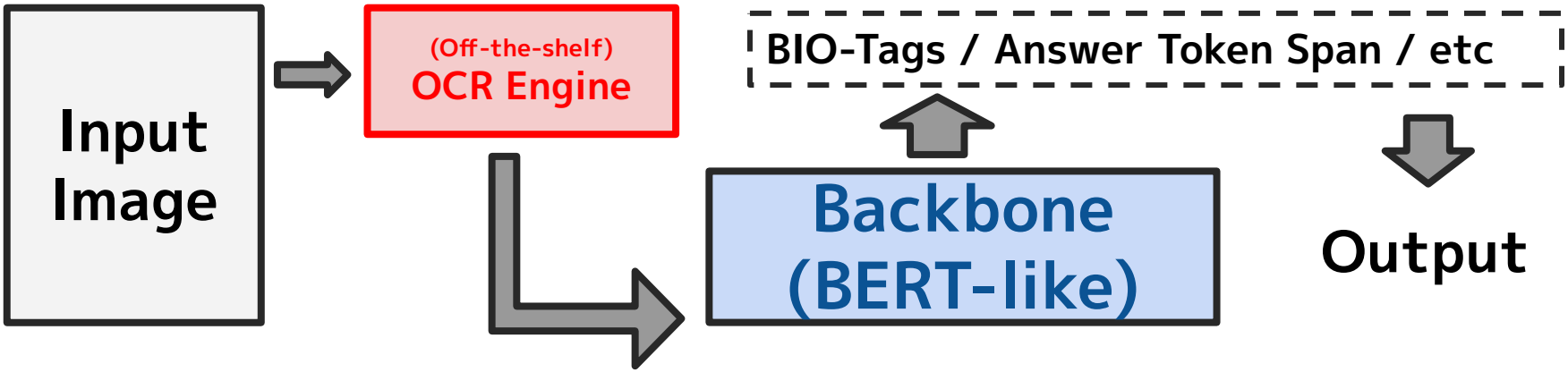
Conventional VDU Model: Overview



Although such OCR-based approaches have shown promising performance, they have several problems induced by OCR.

Slide

Conventional VDU Model: Overview

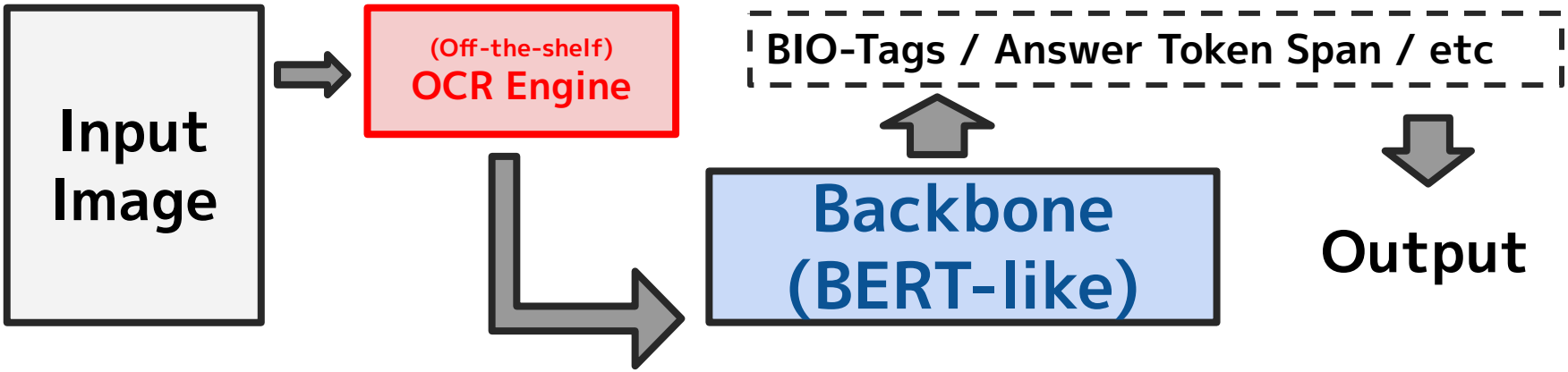


- high computational costs

First, OCR increase computational costs.

Slide

Conventional VDU Model: Overview

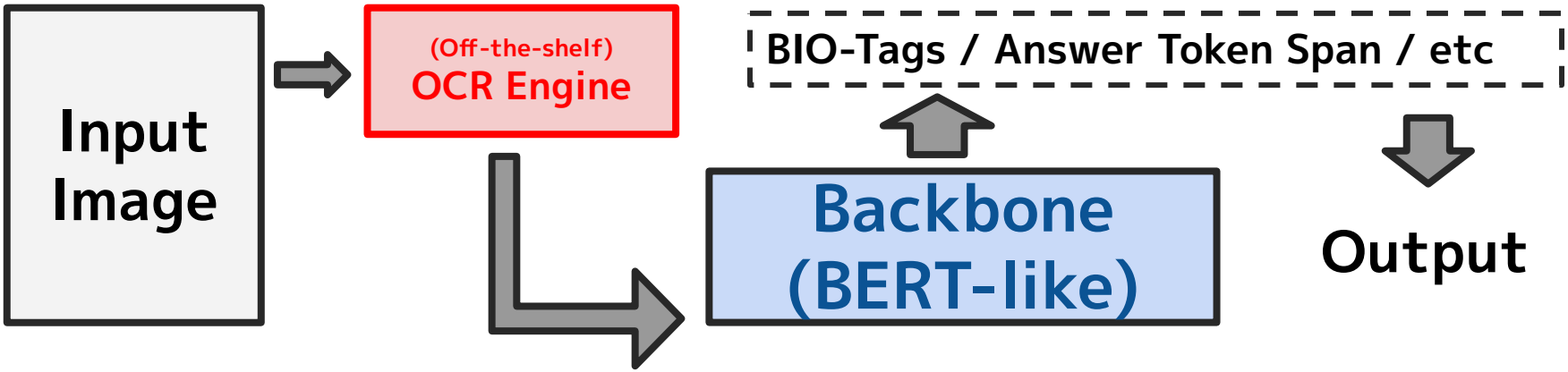


- high computational costs
- inflexibility of OCR on languages or document type

Second, OCR makes it hard to handle various languages/types of documents.

Slide

Conventional VDU Model: Overview

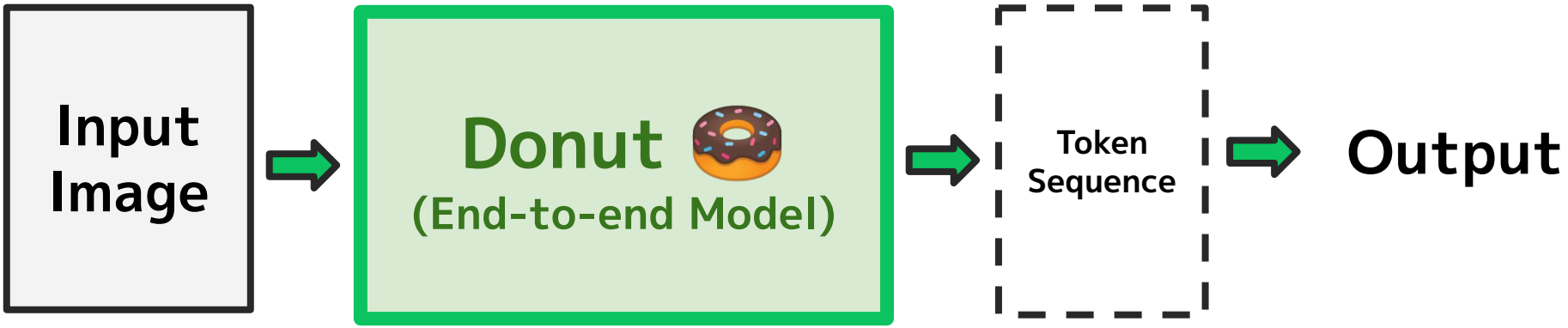


- high computational costs
- inflexibility of OCR on languages or document type
- OCR error propagation

Lastly, OCR errors are propagate to the subsequent process.

Slide

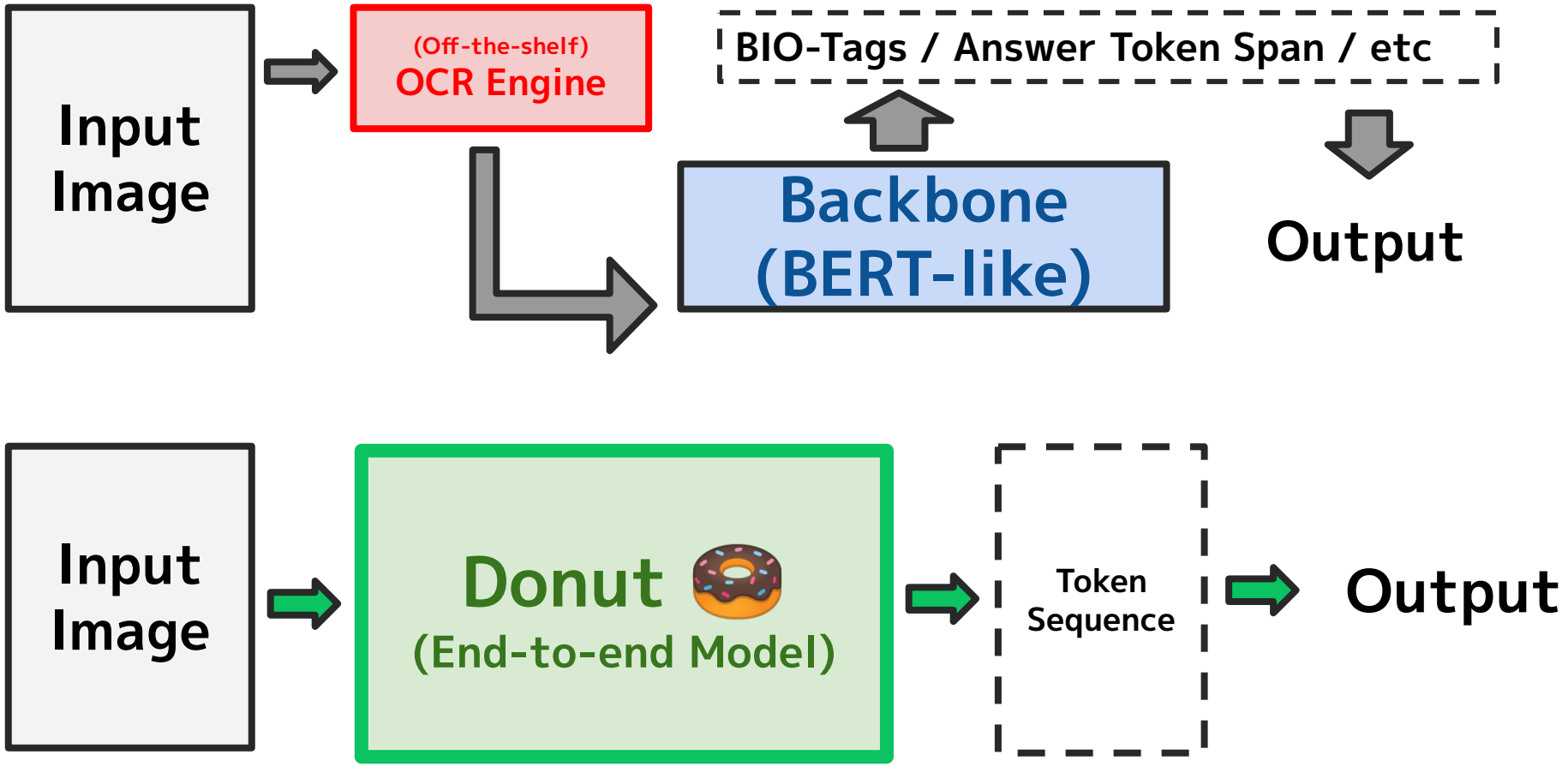
Proposal: OCR-free Approach



To address the issues, we introduce a novel OCR-free VDU model, Donut.

Slide

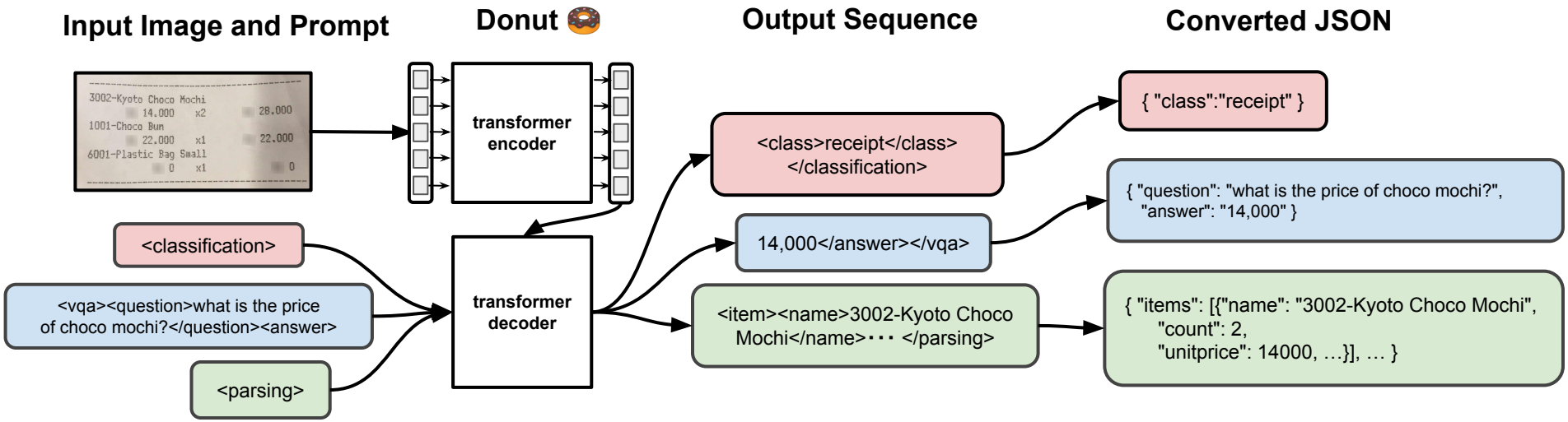
AS-IS v.s. TO-BE



Without OCR, Donut directly processes the input image
and gets an output that contains desired types of information.

Slide

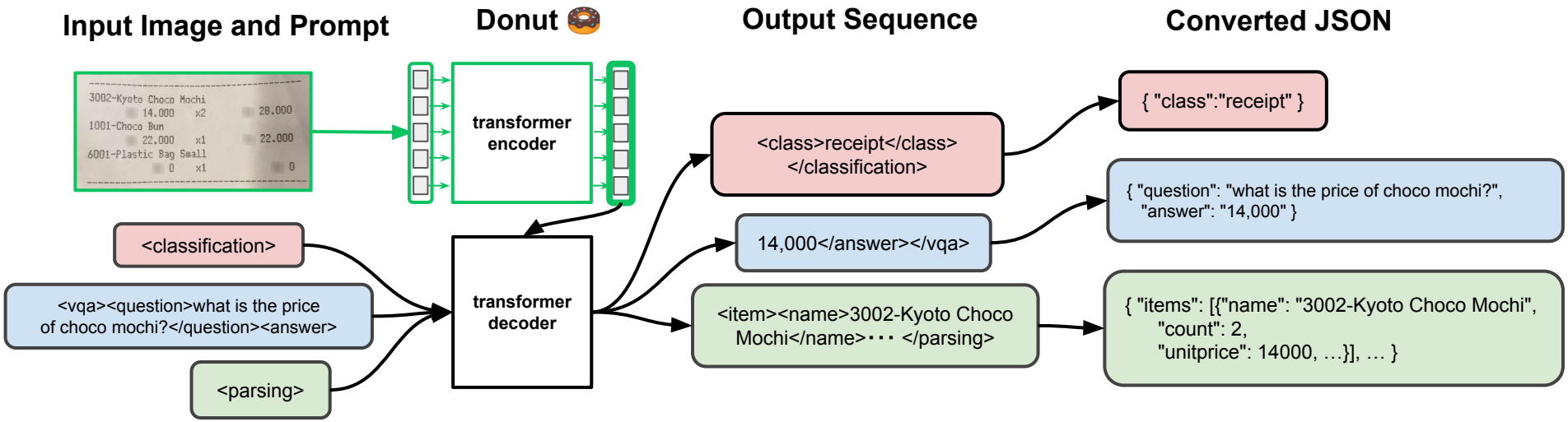
Overview



This is the overview of Donut.

Slide

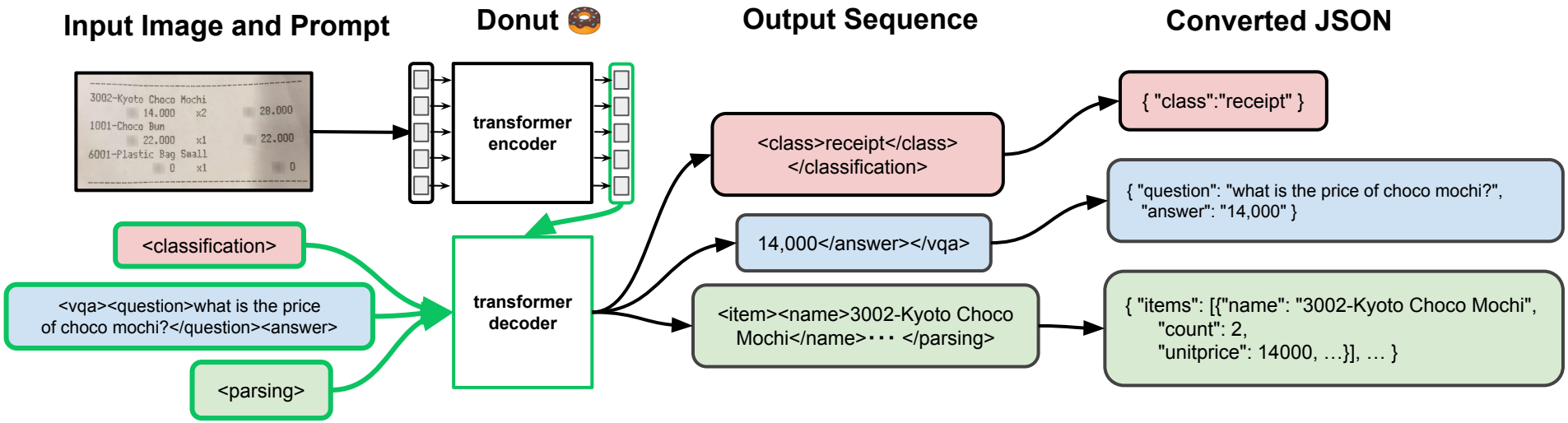
Overview



The visual encoder maps the input image into a set of embeddings.

Slide

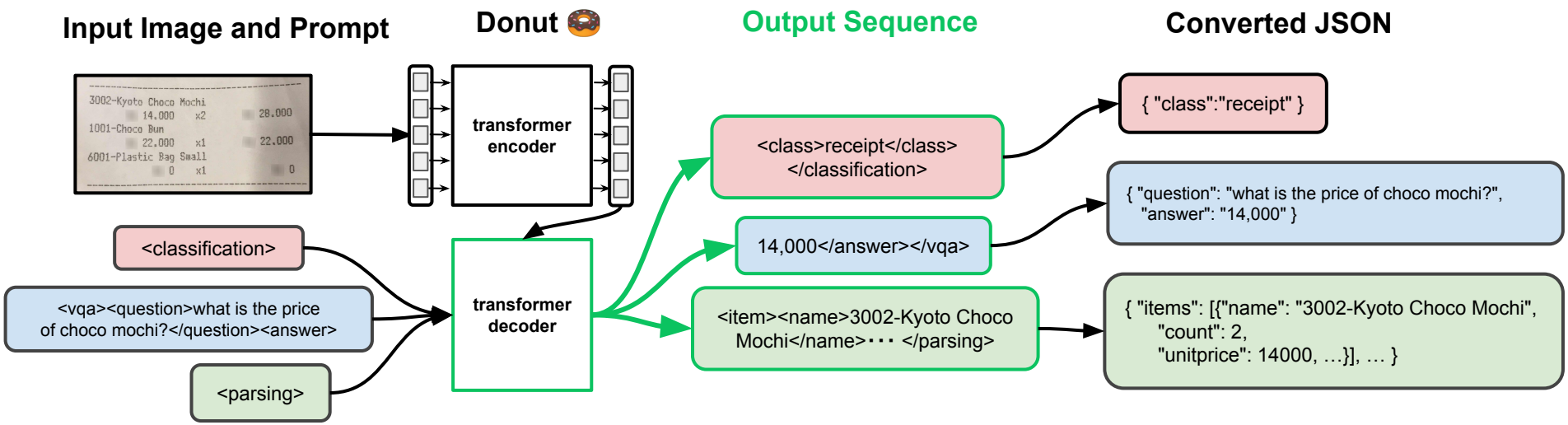
Overview



The textual decoder processes the image embeddings and prompt tokens.

Slide

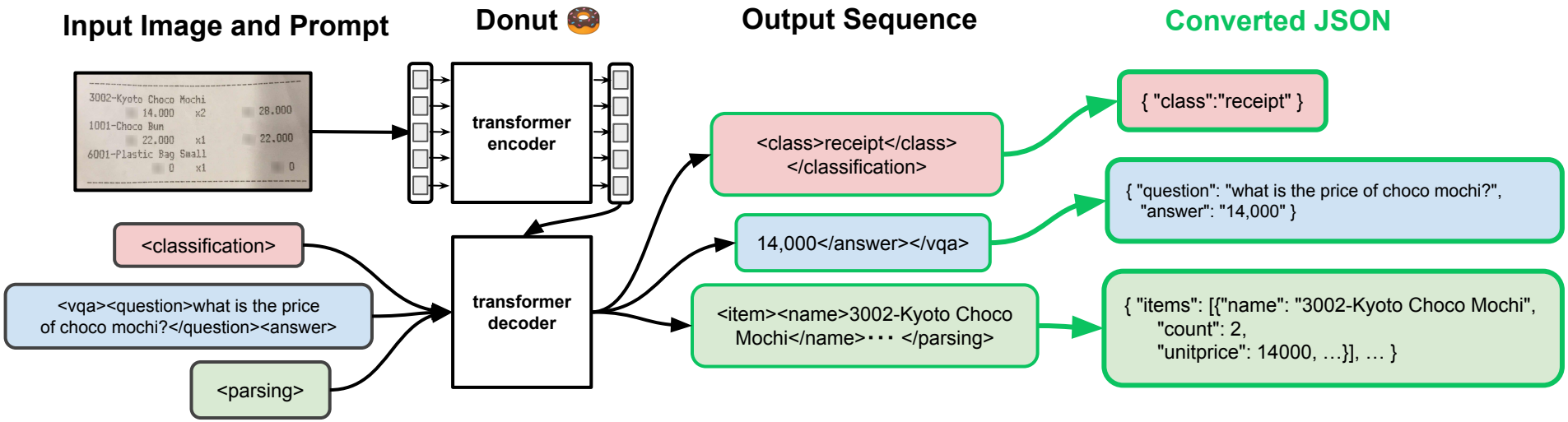
Overview



Then, the decoder outputs token sequences,

Slide

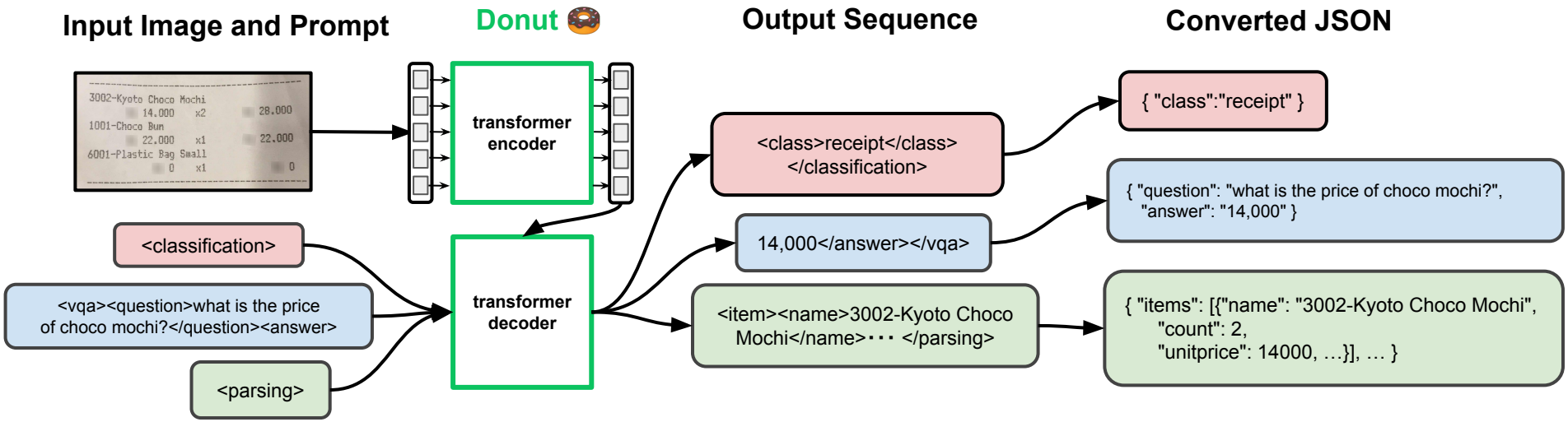
Overview



that can be converted into a desired data format, such as, a JSON format.

Slide

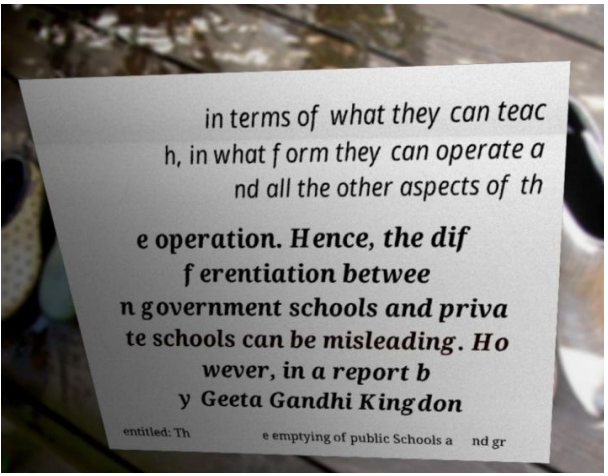
Overview: Model Architecture



Swin Transformer and BART are used as an encoder and decoder, respectively.
More details can also be found in the manuscript.

Slide

Pre-training Task



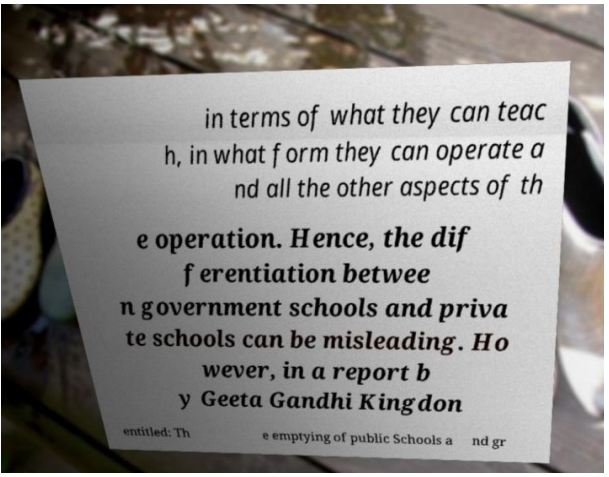
Donut 🍩
(End-to-end Model)

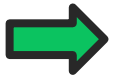
In terms of what that can teach, in what from …

To train Donut, we propose a simple pre-training task.

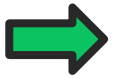
Slide

Pre-training Task





Donut 🍩
(End-to-end Model)

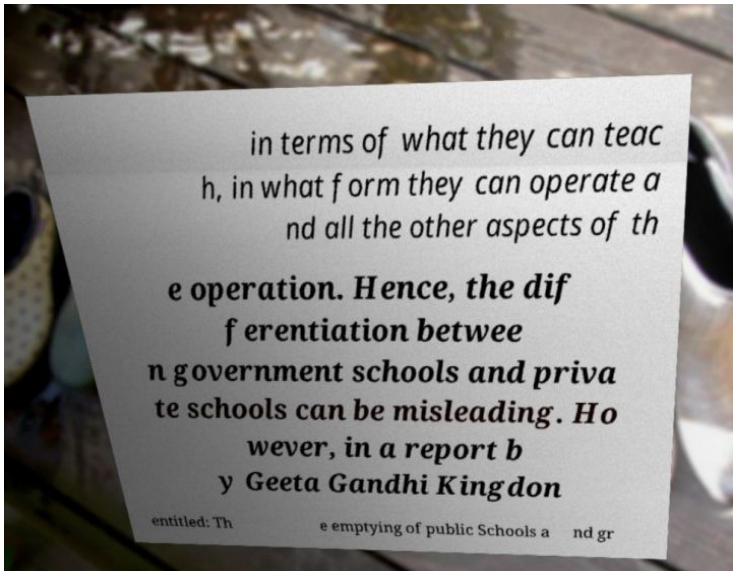


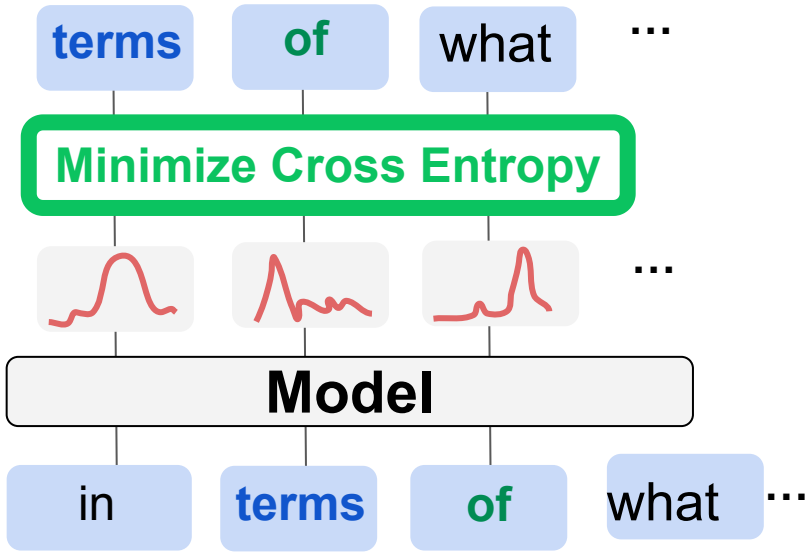
In terms of what that can teach, in what from …

The objective is to read all texts from the top-left to bottom-right.
This task can be interpret as a pseudo OCR task.

Slide

Training Strategy: Teacher-forcing Scheme



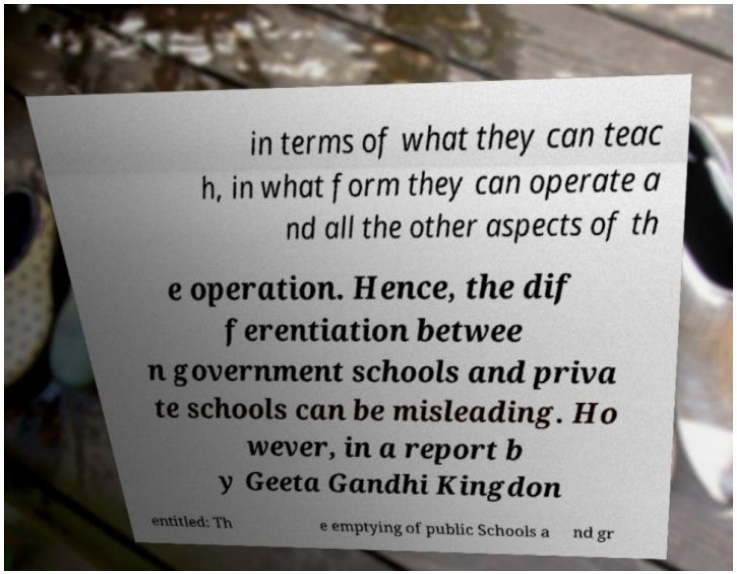


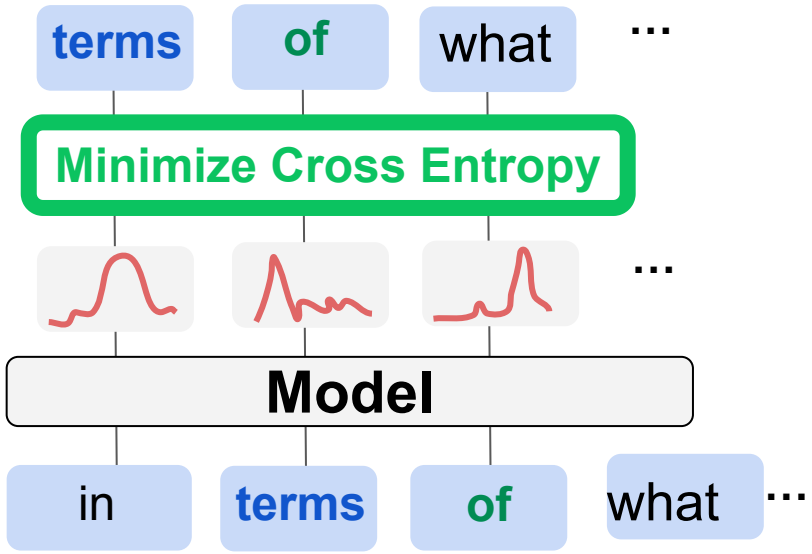
This can be interpret as a token classification at each step.

Following the original Transformer, the model training is done with Teacher-forcing scheme. More details can also be found in the manuscript.

Slide

Training Strategy: Teacher-forcing Scheme





This can be interpret as a token classification at each step.

Following the original Transformer, the model training is done with Teacher-forcing scheme. More details can also be found in the manuscript.

Slide

SynthDoG 🐶:Synthetic Document Generator

For the pre-training of Donut, we also present a data generator, SynthDoG 🐶.

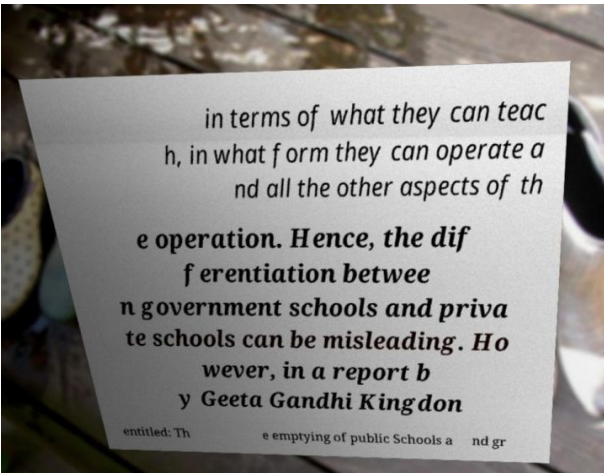
Slide

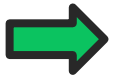
SynthDoG 🐶:Synthetic Document Generator

SynthDoG alleviates the dependency on large-scale real document images and enables the extension to a multilingual setting.

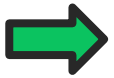
Slide

Pre-training Task





Donut 🍩
(End-to-end Model)

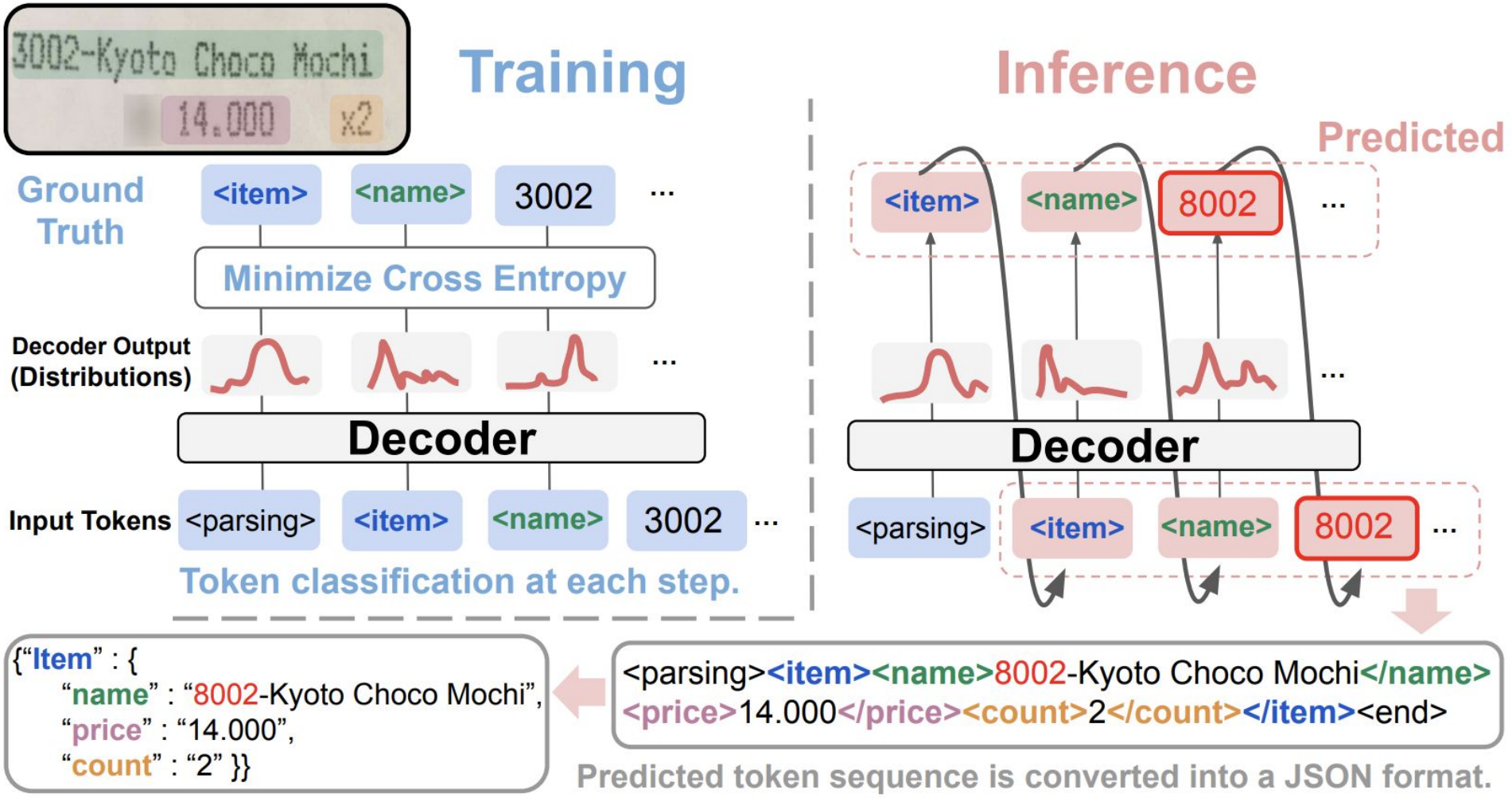


In terms of what that can teach, in what from …

After the model learns “how to read”,

Slide

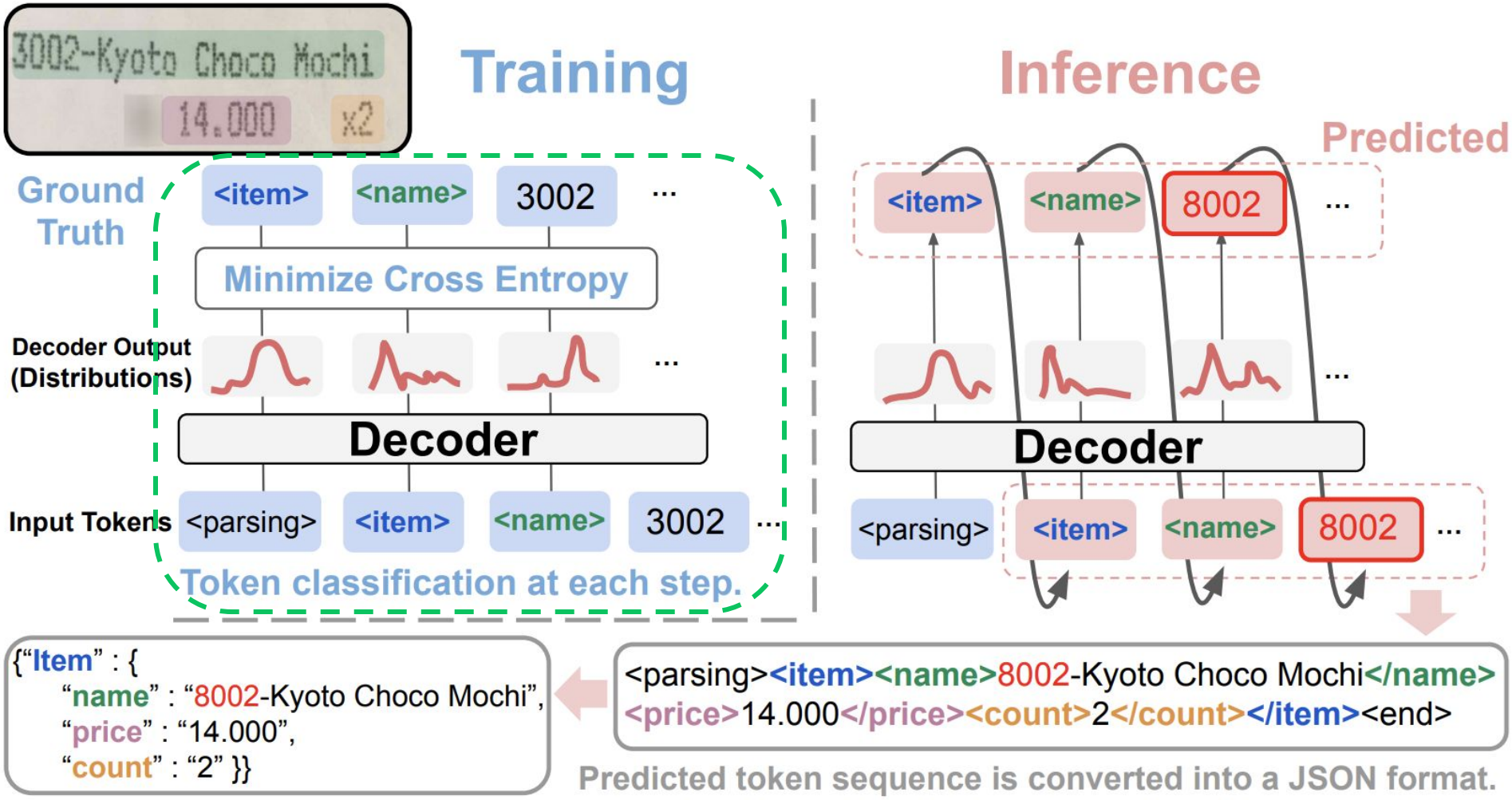
Model Fine-tuning and Inference Overview



in the fine-tuning, we teach the model "how to understand".

Slide

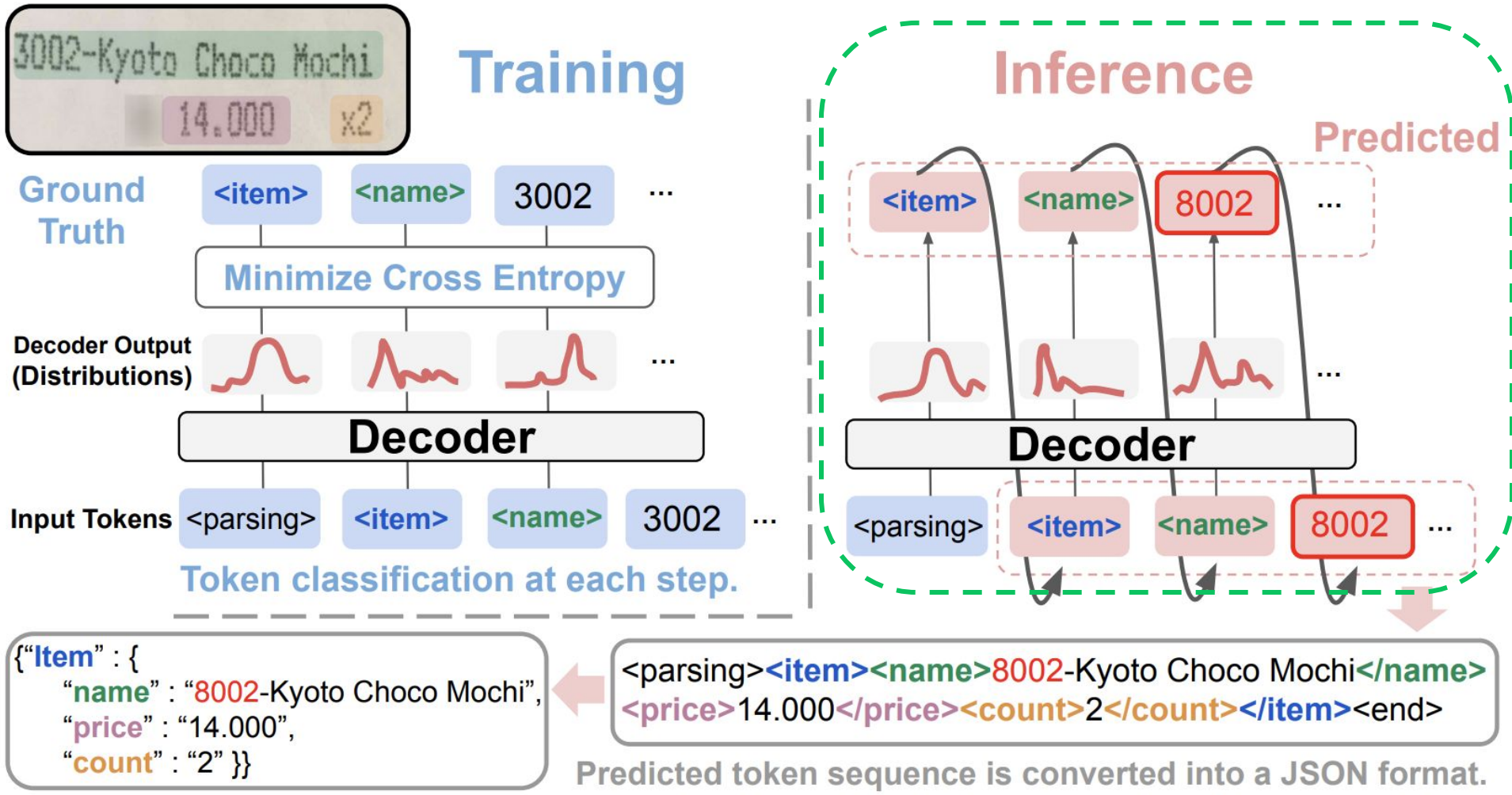
Model Fine-tuning and Inference Overview



The prediction target is set to a desired downstream token sequence, including some special tokens.

Slide

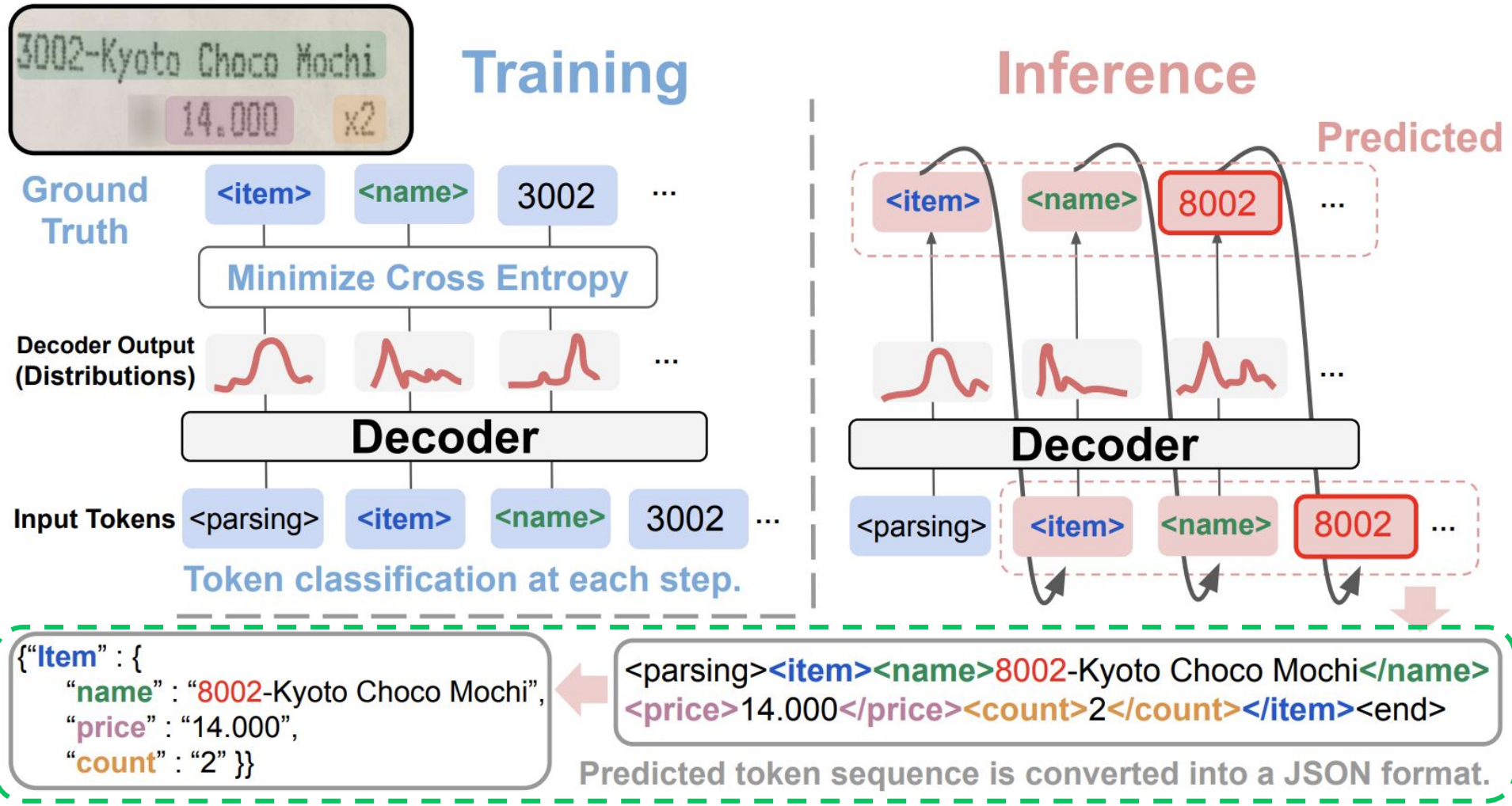
Model Fine-tuning and Inference Overview



At inference, the predicted token from the last step is fed to the next.

Slide

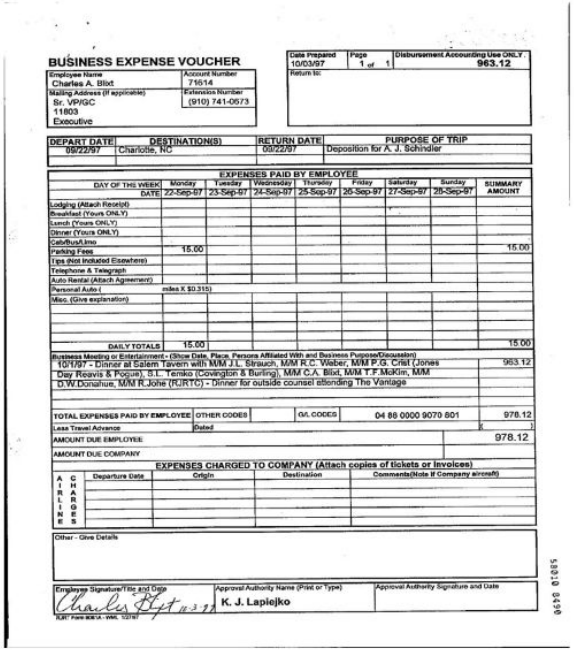
Model Fine-tuning and Inference Overview



The predicted sequence is converted into a JSON format.

Slide

Experiments: Samples of Downstream Datasets



Q: What is the Extension Number as per the voucher?
A: (910) 741-0673

We conducts 3 VDU applications on 6 different datasets. Some samples are shown.

Slide

Experiments: Document Classification

To see whether the model can distinguish across different types of documents, we test a classification task.

Slide

Experiments: Document Classification

	OCR	#Params	Time (ms)	Accuracy (%)
BERT	✓	110M + $\alpha^{\dagger}$	1392	89.81
RoBERTa	✓	125M + $\alpha^{\dagger}$	1392	90.06
LayoutLM	✓	113M + $\alpha^{\dagger}$	1396	91.78
LayoutLM (w/ image)	✓	160M + $\alpha^{\dagger}$	1426	94.42
LayoutLMv2	✓	200M + $\alpha^{\dagger}$	1489	95.25
Donut (Proposed)		143M	**752**	**95.30**

This is the results on the RVL-CDIP dataset.
Donut achieves state-of-the-are scores with reasonable speed and efficiency.

Slide

Experiments: Document Classification

	OCR	#Params	Time (ms)	Accuracy (%)
BERT	✓	110M + $\alpha^{\dagger}$	1392	89.81
RoBERTa	✓	125M + $\alpha^{\dagger}$	1392	90.06
LayoutLM	✓	113M + $\alpha^{\dagger}$	1396	91.78
LayoutLM (w/ image)	✓	160M + $\alpha^{\dagger}$	1426	94.42
LayoutLMv2	✓	200M + $\alpha^{\dagger}$	1489	95.25
Donut (Proposed)		143M	**752**	**95.30**

This is the results on the RVL-CDIP dataset.
Donut achieves state-of-the-are scores with reasonable speed and efficiency.

Slide

Experiments: Document Parsing

input_img

output

copy to clipboard

```
{
  menu:  [
    0:  {
      nm:  "0571-1854 BLUS WANITA",
      unitprice:  "@120,000",
      cnt:  "1",
      price:  "120,000"
    } ,
    1:  {
      nm:  "1002-0060 SHOPPING BAG",
      cnt:  "1",
      price:  "0"
    }
  ],
  total:  {
    total_price:  "120,000",
    changeprice:  "0",
    creditcardprice:  "120,000",
    menuqty_cnt:  "1"
  }
}
```

Next, to see the model fully understands the complex layouts and contexts, we test document parsing tasks.

Slide

Experiments: Document Parsing

			CORD [45]			Ticket [12]			Business Card			Receipt		
	OCR	#Params	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.
BERT* [22]	✓	$86^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.6	73.0	65.5	1.7	74.3	82.4	1.5	40.8	72.1	2.5	70.3	54.1
BROS [18]	✓	$86^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	74.7	70.0									
LayoutLM [65]	✓	$89^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	78.4	81.3									
LayoutLMv2* [64,66]	✓	$179^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	78.9	82.4	1.8	87.2	90.1	1.6	52.2	83.0	2.6	72.9	78.0
Donut		$143^{\dagger}_{\mathrm{M}}$	**1.2**	**84.1**	**90.9**	**0.6**	**94.1**	**98.7**	**1.4**	**57.8**	**84.4**	**1.9**	**78.6**	**88.6**
SPADE* [25]	✓	$93^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	4.0	74.0	75.8	4.5	14.9	29.4	4.3	32.3	51.3	7.3	64.1	53.2
WYVERN* [21]	✓	$106^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.2	43.3	46.9	1.5	41.8	54.8	1.7	29.9	51.5	3.4	71.5	82.9

For all domains, Donut showed the best with significantly faster inference

Slide

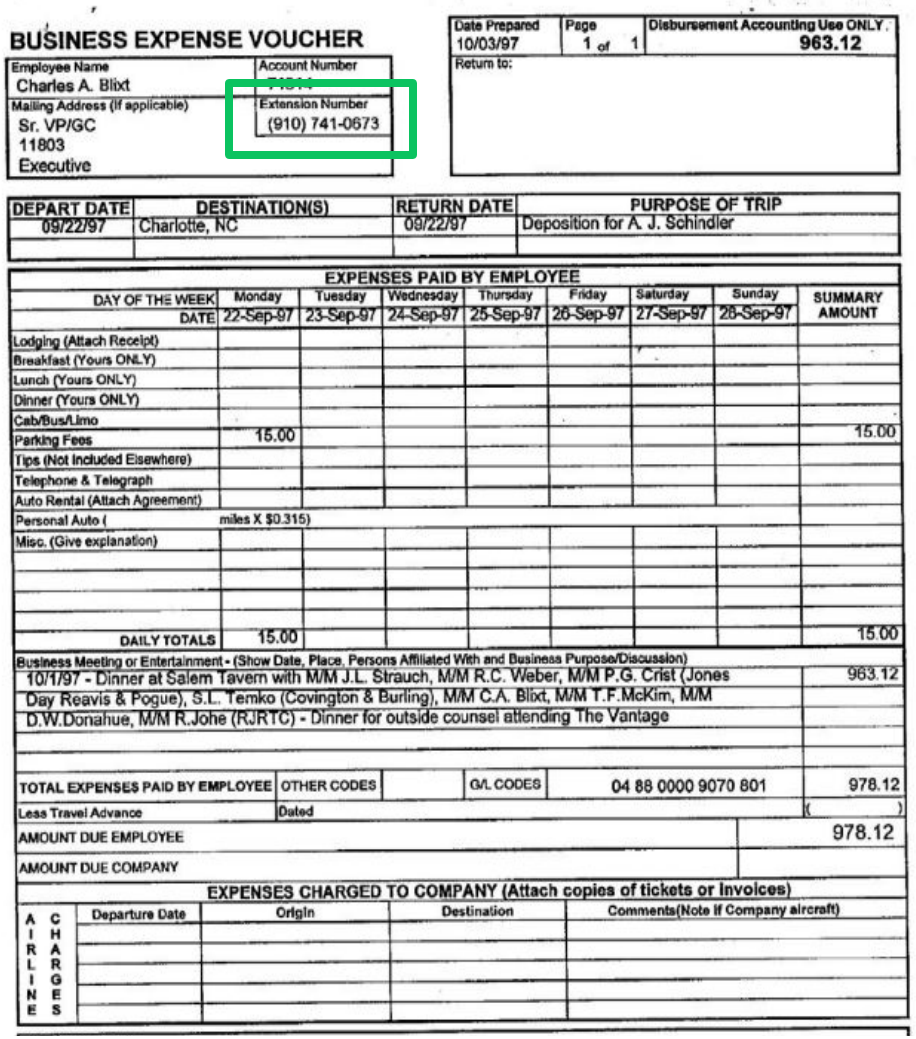
Experiments: Document Parsing

			CORD [45]			Ticket [12]			Business Card			Receipt		
	OCR	#Params	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.
BERT* [22]	✓	$86^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.6	73.0	65.5	1.7	74.3	82.4	1.5	40.8	72.1	2.5	70.3	54.1
BROS [18]	✓	$86^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	74.7	70.0									
LayoutLM [65]	✓	$89^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	78.4	81.3									
LayoutLMv2* [64,66]	✓	$179^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.7	78.9	82.4	1.8	87.2	90.1	1.6	52.2	83.0	2.6	72.9	78.0
Donut		$143^{\dagger}_{\mathrm{M}}$	**1.2**	**84.1**	**90.9**	**0.6**	**94.1**	**98.7**	**1.4**	**57.8**	**84.4**	**1.9**	**78.6**	**88.6**
SPADE* [25]	✓	$93^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	4.0	74.0	75.8	4.5	14.9	29.4	4.3	32.3	51.3	7.3	64.1	53.2
WYVERN* [21]	✓	$106^{\dagger}_{\mathrm{M}} + \alpha^{\ddagger}$	1.2	43.3	46.9	1.5	41.8	54.8	1.7	29.9	51.5	3.4	71.5	82.9

For all domains, Donut showed the best accuracies with significantly faster inference speed.

Slide

Experiments: Document VQA



BUSINESS EXPENSE VOUCHER

Date Prepared	Page	Disbursement Accounting Use ONLY
10/03/97	1 of 1	963.12

Return to:

Employee Name: Charles A. Blixt
Account Number: 71311
Mailing Address (If applicable): Sr. VP/GC 11803 Executive
Extension Number: (910) 741-0673

DEPART DATE	DESTINATION(S)	RETURN DATE	PURPOSE OF TRIP
09/22/97	Charlotte, NC	09/22/97	Deposition for A. J. Schindler

EXPENSES PAID BY EMPLOYEE

DAY OF THE WEEK	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	SUMMARY AMOUNT
DATE	22-Sep-97	23-Sep-97	24-Sep-97	25-Sep-97	26-Sep-97	27-Sep-97	28-Sep-97	
Lodging (Attach Receipt)								
Breakfast (Yours ONLY)								
Lunch (Yours ONLY)								
Dinner (Yours ONLY)								
Cab/Bus/Limo								
Parking Fees	15.00							15.00
Tips (Not Included Elsewhere)								
Telephone & Telegraph								
Auto Rental (Attach Agreement)								
Personal Auto (miles X $0.315)								
Misc. (Give explanation)								
DAILY TOTALS	15.00							15.00

Business Meeting or Entertainment - (Show Date, Place, Persons Affiliated With and Business Purpose/Discussion)

10/1/97 - Dinner at Salem Tavern with M/M J.L. Strauch, M/M R.C. Weber, M/M P.G. Crist (Jones Day Reavis & Pogue), S.L. Temko (Covington & Burling), M/M C.A. Blixt, M/M T.F.McKim, M/M D.W.Donahue, M/M R.Johe (RJRTC) - Dinner for outside counsel attending The Vantage — 963.12

TOTAL EXPENSES PAID BY EMPLOYEE	OTHER CODES		G/L CODES	04 88 0000 9070 801	978.12
Less Travel Advance	Dated				( )
AMOUNT DUE EMPLOYEE					978.12
AMOUNT DUE COMPANY					

EXPENSES CHARGED TO COMPANY (Attach copies of tickets or Invoices)

AIRLINE CHARGES	Departure Date	Origin	Destination	Comments(Note if Company aircraft)

Q: What is the Extension Number as per the voucher?

A: (910) 741–0673

To validate the further capacity of the model, we test a document VQA task.

Slide

Experiments: Document VQA

	Fine-tuning set	OCR	#Params†	Time (ms)	ANLS test set	ANLS* handwritten
BERT [64]	train set	✓	110M + $\alpha^{\ddagger}$	1517	63.5	n/a
LayoutLM[65]	train set	✓	113M + $\alpha^{\ddagger}$	1519	69.8	n/a
LayoutLMv2[64]	train set	✓	200M + $\alpha^{\ddagger}$	1610	78.1	n/a
Donut	train set		176M	**782**	67.5	**72.1**
LayoutLMv2-Large-QG[64]	train + dev + QG	✓	390M + $\alpha^{\ddagger}$	1698	**86.7**	67.3

For VQA, Donut showed a high score on the handwritten documents which are known to be challenging.

Slide

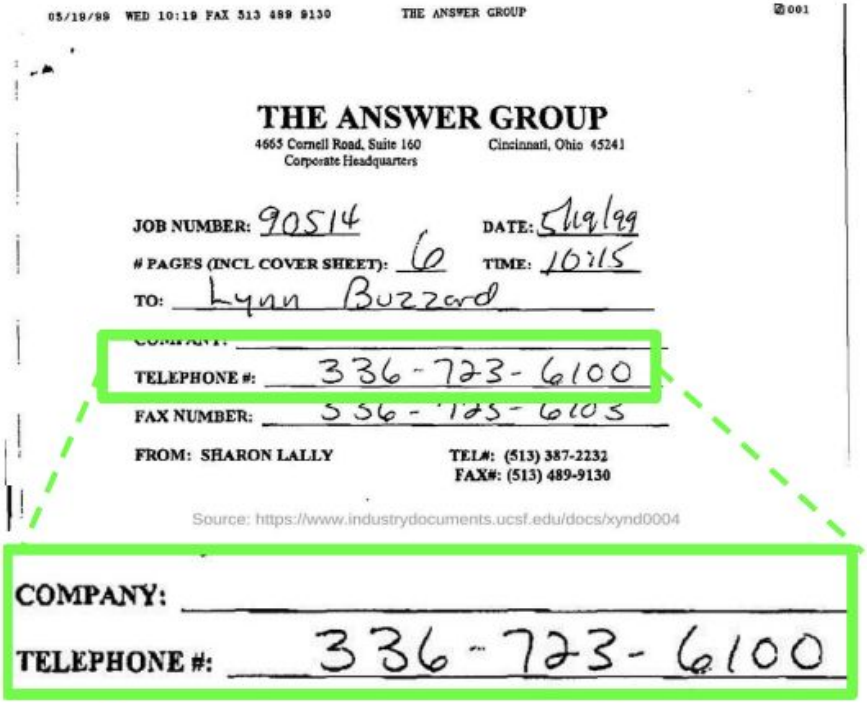
Experiments: Document VQA

	Fine-tuning set	OCR	#Params†	Time (ms)	ANLS test set	ANLS* handwritten
BERT [64]	train set	✓	110M + $\alpha^{\ddagger}$	1517	63.5	n/a
LayoutLM[65]	train set	✓	113M + $\alpha^{\ddagger}$	1519	69.8	n/a
LayoutLMv2[64]	train set	✓	200M + $\alpha^{\ddagger}$	1610	78.1	n/a
Donut	train set		176M	**782**	67.5	**72.1**
LayoutLMv2-Large-QG[64]	train + dev + QG	✓	390M + $\alpha^{\ddagger}$	1698	**86.7**	67.3

For VQA, Donut showed a high score on the handwritten documents which are known to be challenging.

Slide

Analysis: VQA on Handwritten Documents

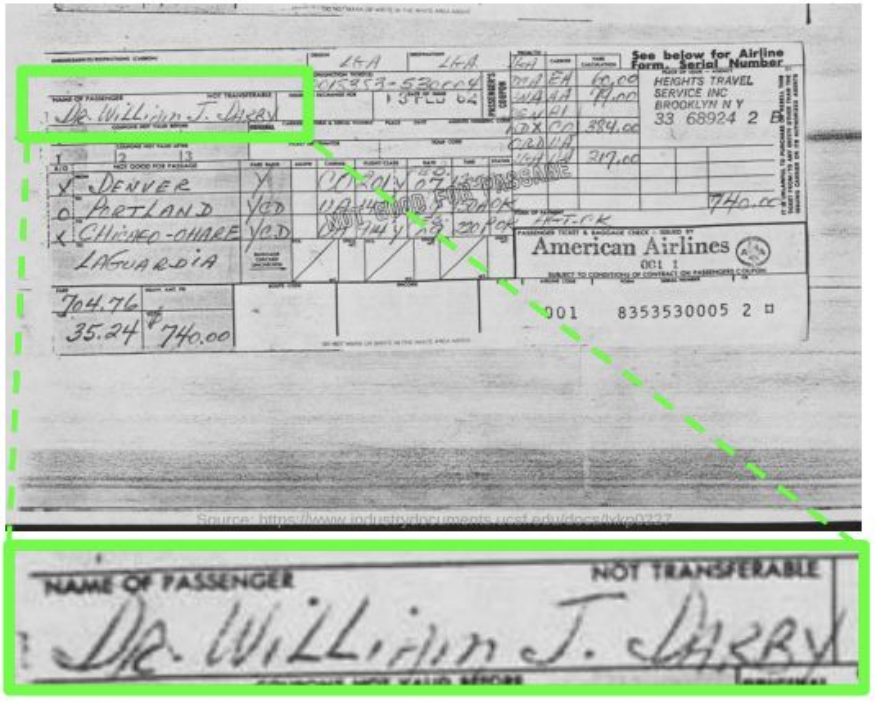


Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-QG: **336-723- 4100**



Q: What is the name of the passenger?

Answer: DR. William J. Darby

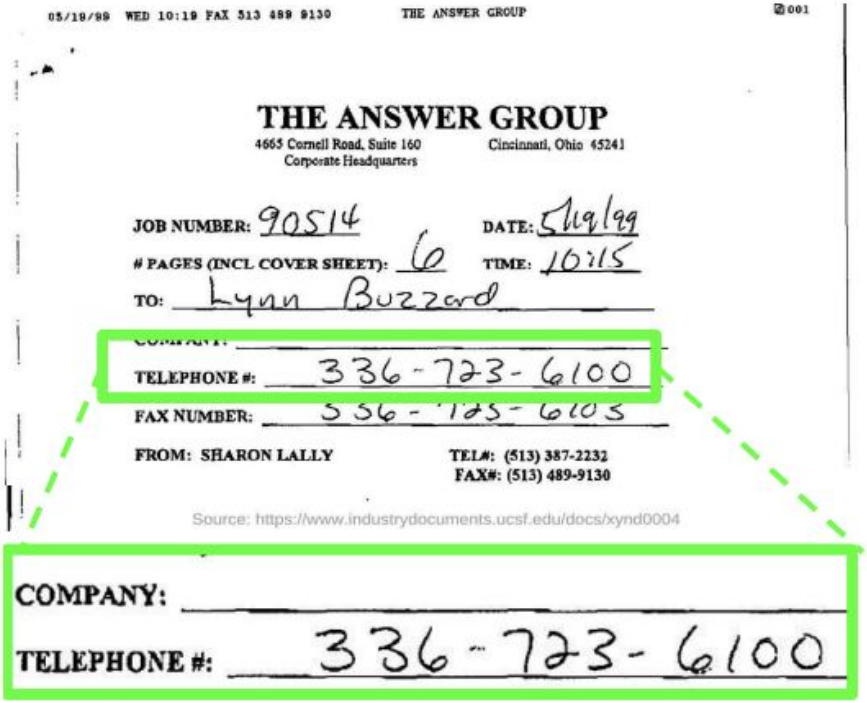
Donut: DR. William J. Darby

LayoutLMv2-Large-QG: **DR. William J. Jarry**

As can be seen, the OCR errors make the performance upper-bound for the conventional baselines.

Slide

Analysis: VQA on Handwritten Documents

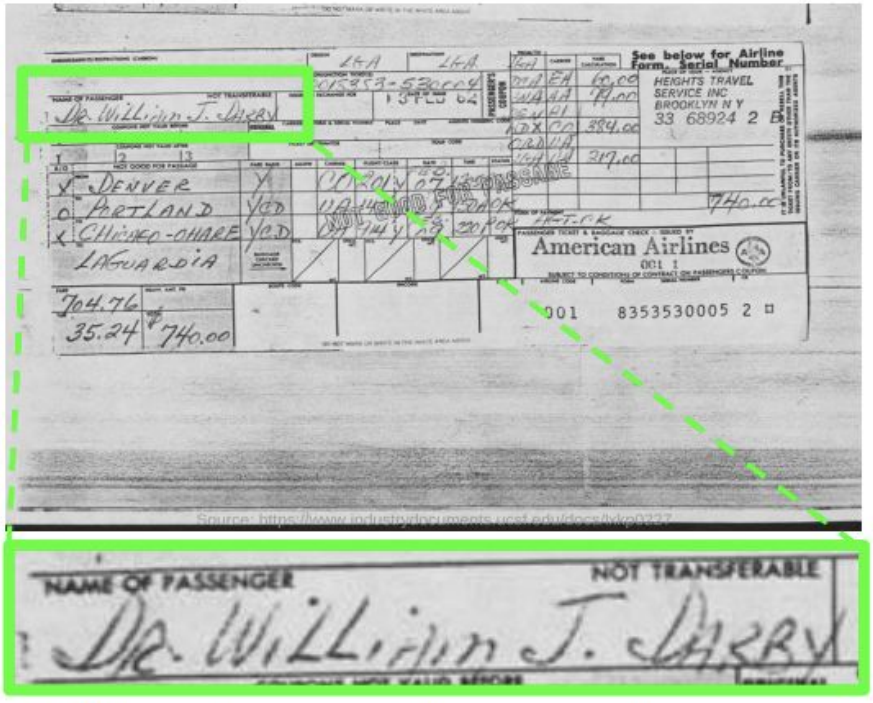


Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-QG: **336-723- 4100**



Q: What is the name of the passenger?

Answer: DR. William J. Darby

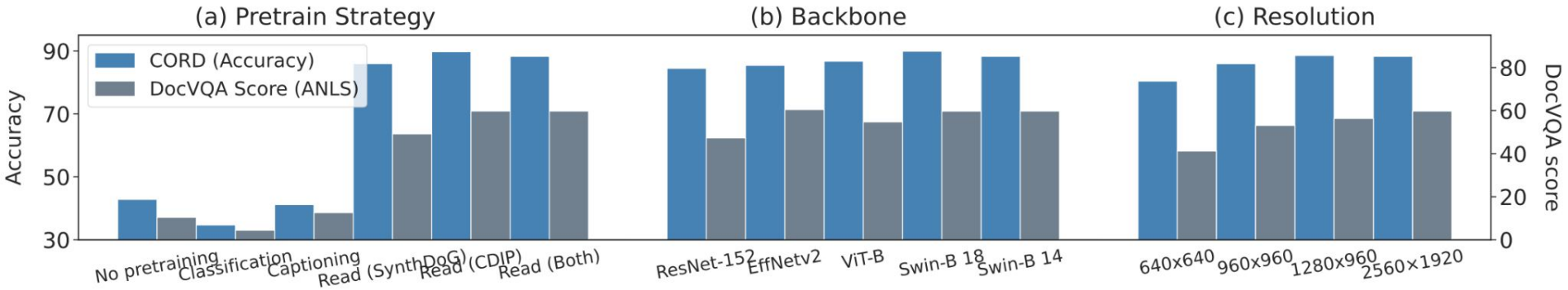
Donut: DR. William J. Darby

LayoutLMv2-Large-QG: **DR. William J. Jarry**

On the other hand, Donut seems robust to the handwritten documents.

Slide

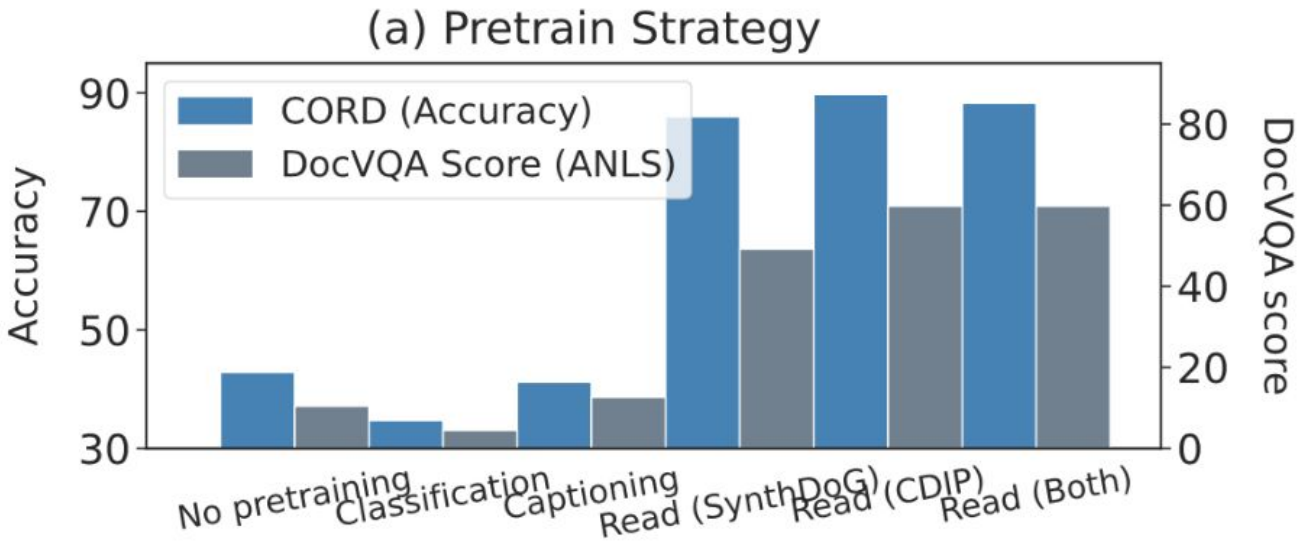
Analysis



Next, we show some main results of our analysis on Donut.

Slide

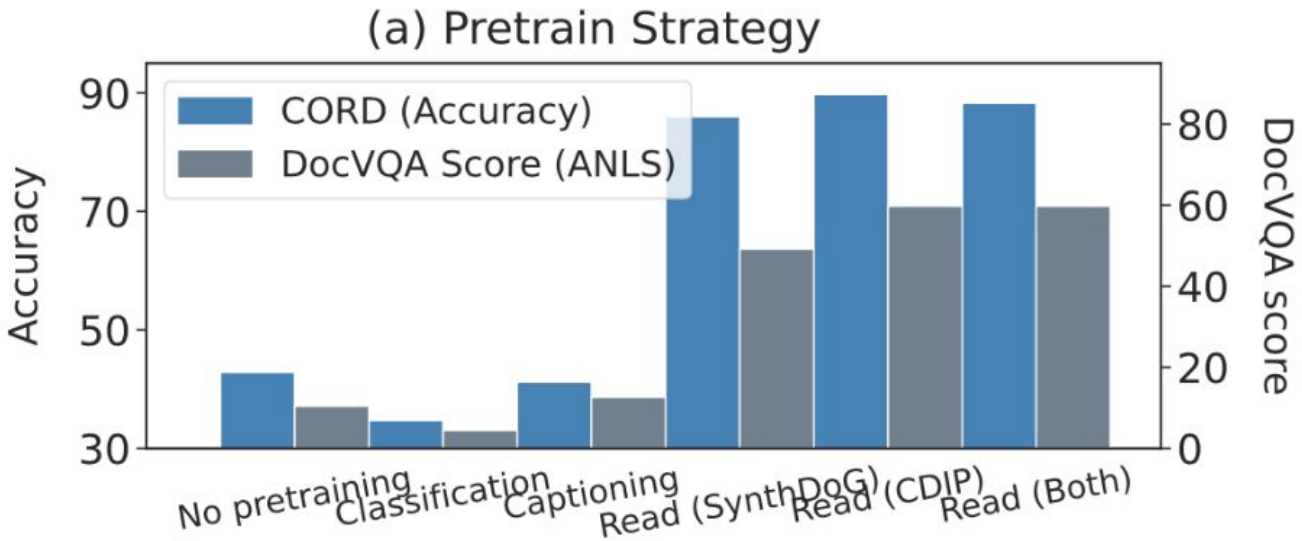
Analysis: Pre-training Strategies



We tested several pre-training tasks.
We found that the proposed task is the most simple yet effective approach.

Slide

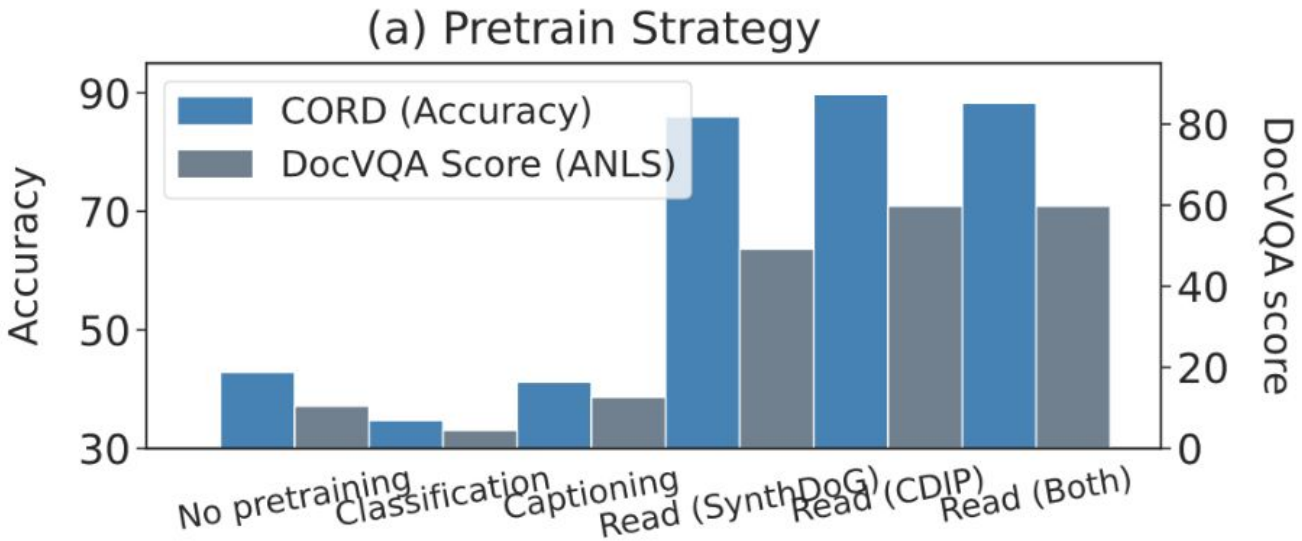
Analysis: Pre-training Strategies



Other tasks that impose a general knowledge of images and texts on models show little gains in the fine-tuning tasks.

Slide

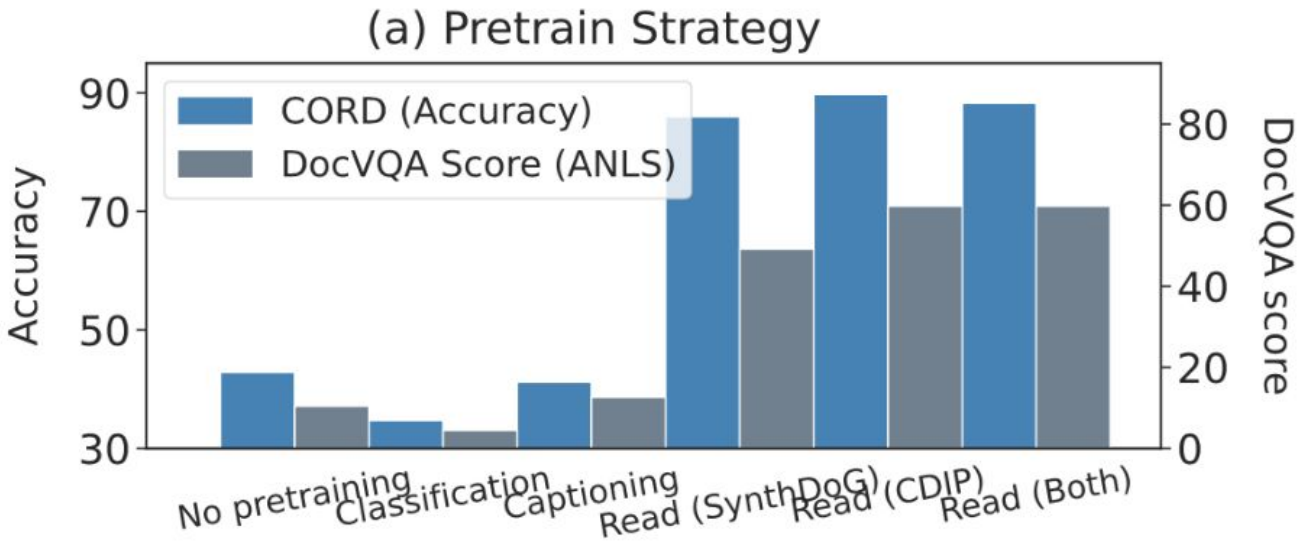
Analysis: Pre-training Strategies



For the text reading task, synthetic images were enough for the CORD task.

Slide

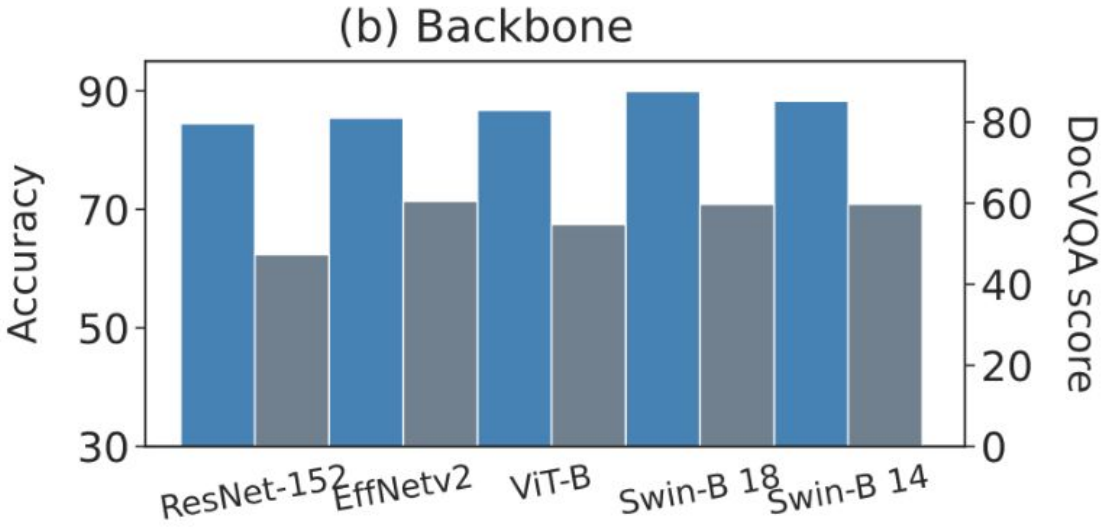
Analysis: Pre-training Strategies



But, in the DocVQA task, it was important to see the real images.

Slide

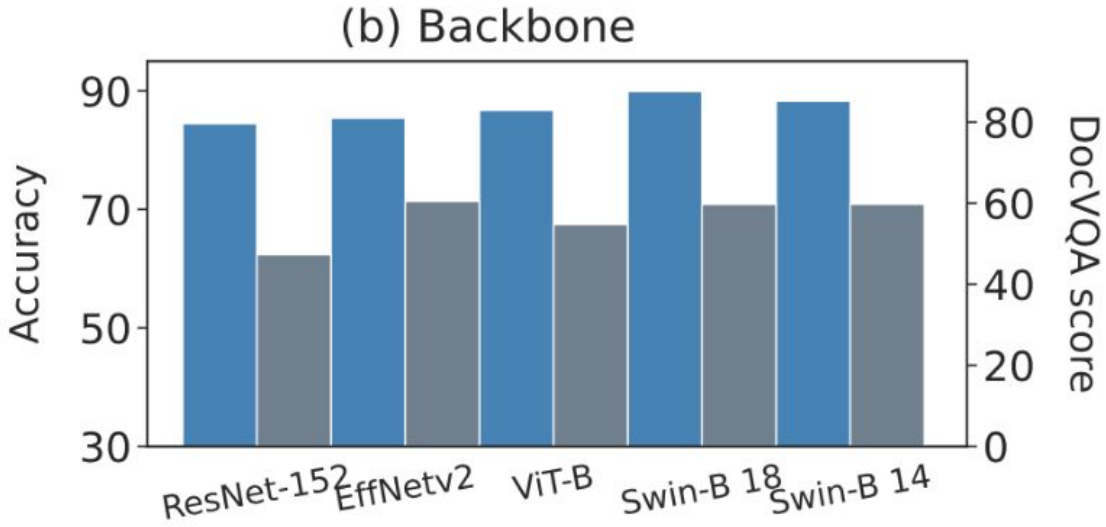
Analysis: Image Backbones



Next, we study popular image classification backbones.

Slide

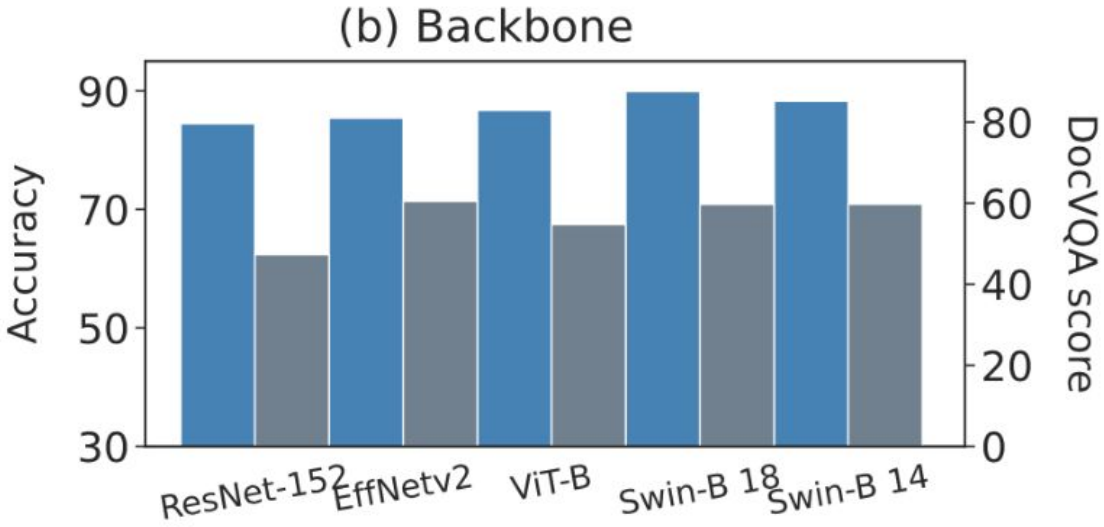
Analysis: Image Backbones



Overall, EfficientNetV2 and Swin Transformer outperform the others.

Slide

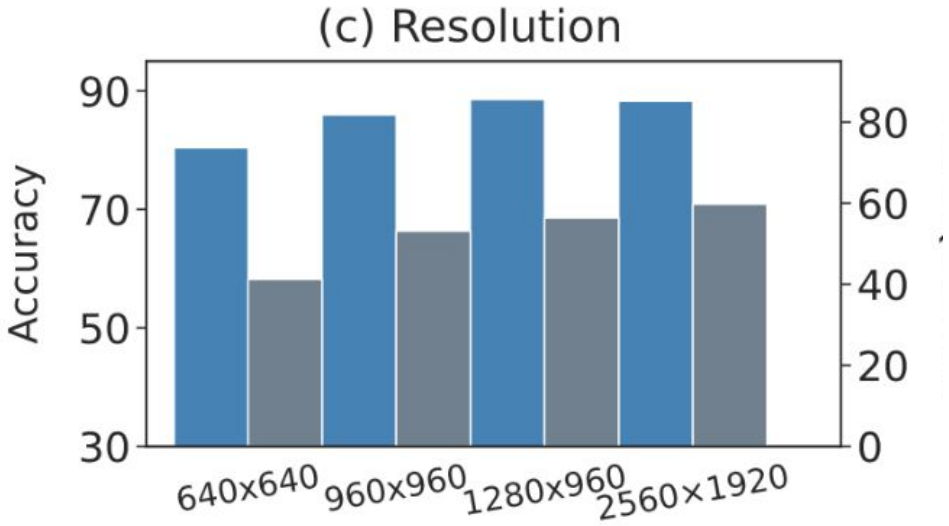
Analysis: Image Backbones



We choose Swin Transformer due to the high scalability and performance.

Slide

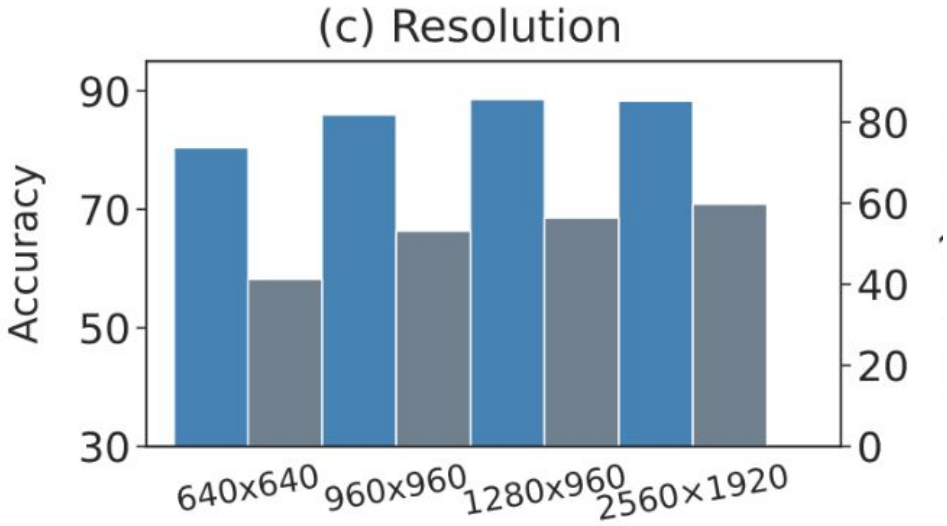
Analysis: Input Resolution



We observed that Donut grows rapidly as we set a larger input size.

Slide

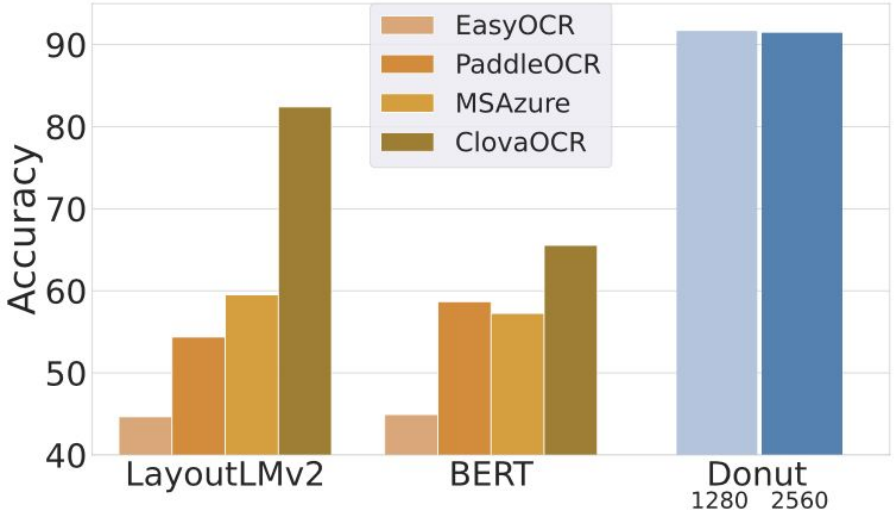
Analysis: Input Resolution

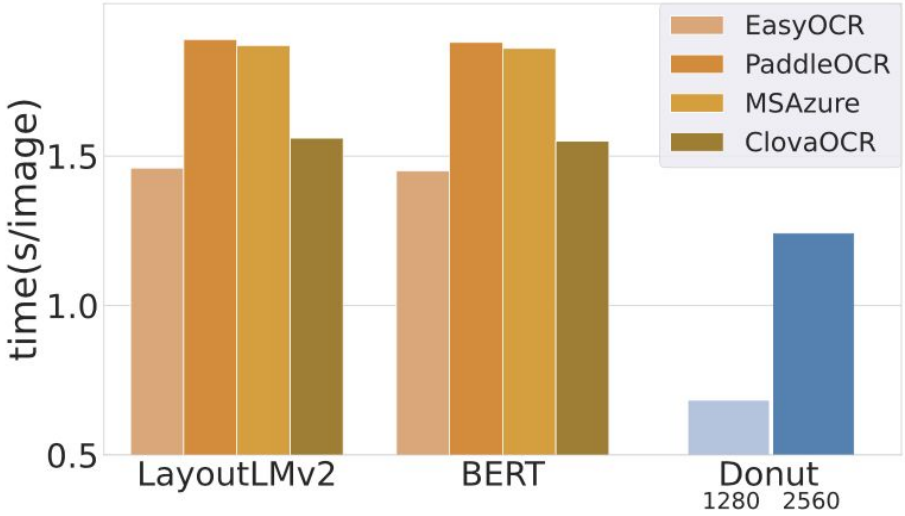


This gets clearer in the DocVQA where the images are larger with many tiny texts.

Slide

Analysis: OCR Engines

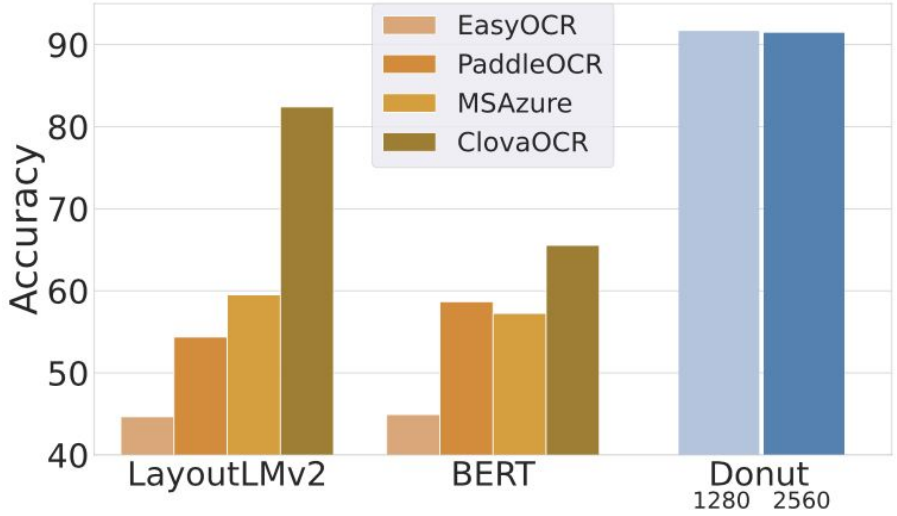


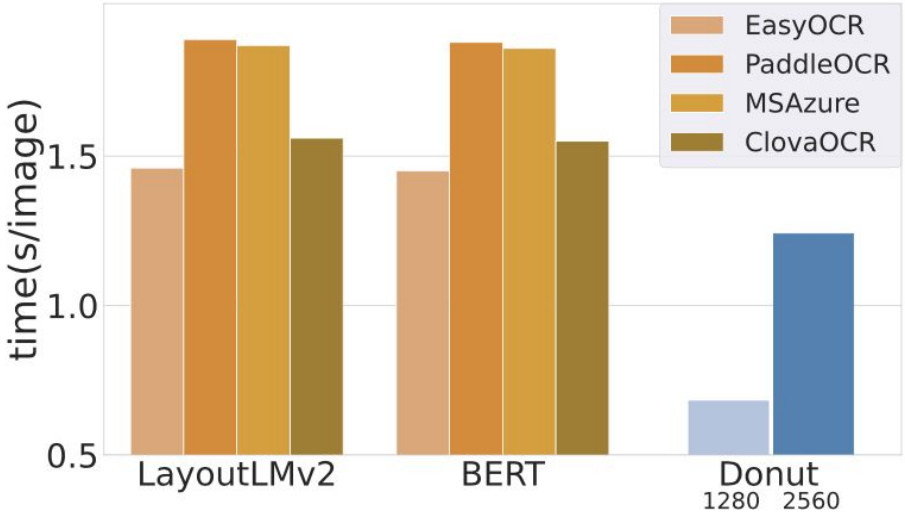


Next, we study the effects of OCR engines on the traditional baselines. We test four widely-used public engines.

Slide

Analysis: OCR Engines

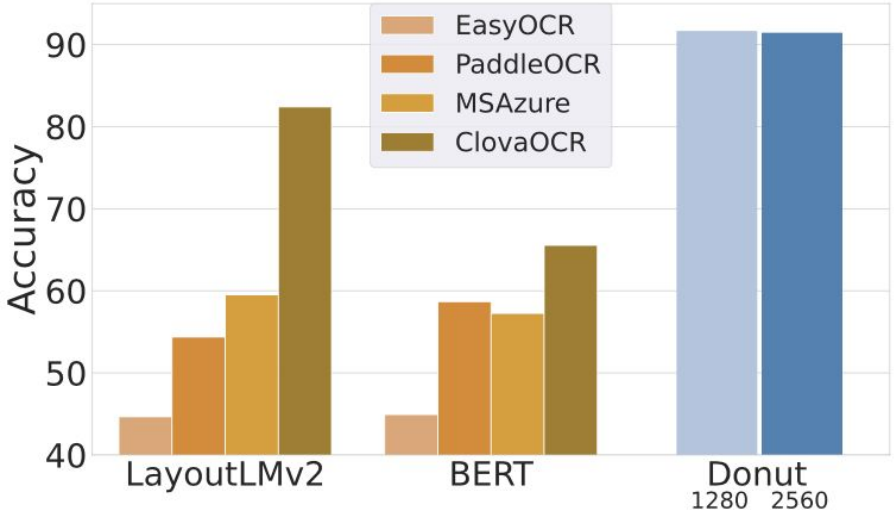


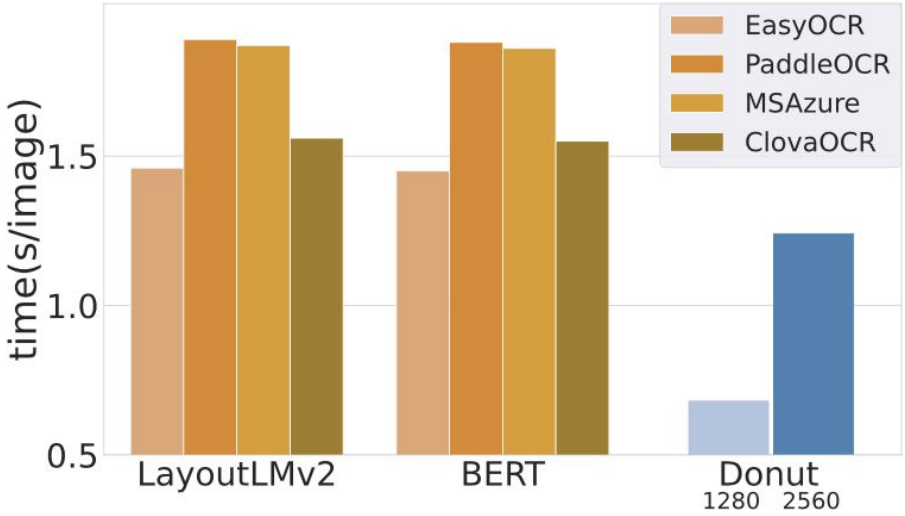


We observed that the scores heavily rely on the OCR engine.

Slide

Analysis: OCR Engines

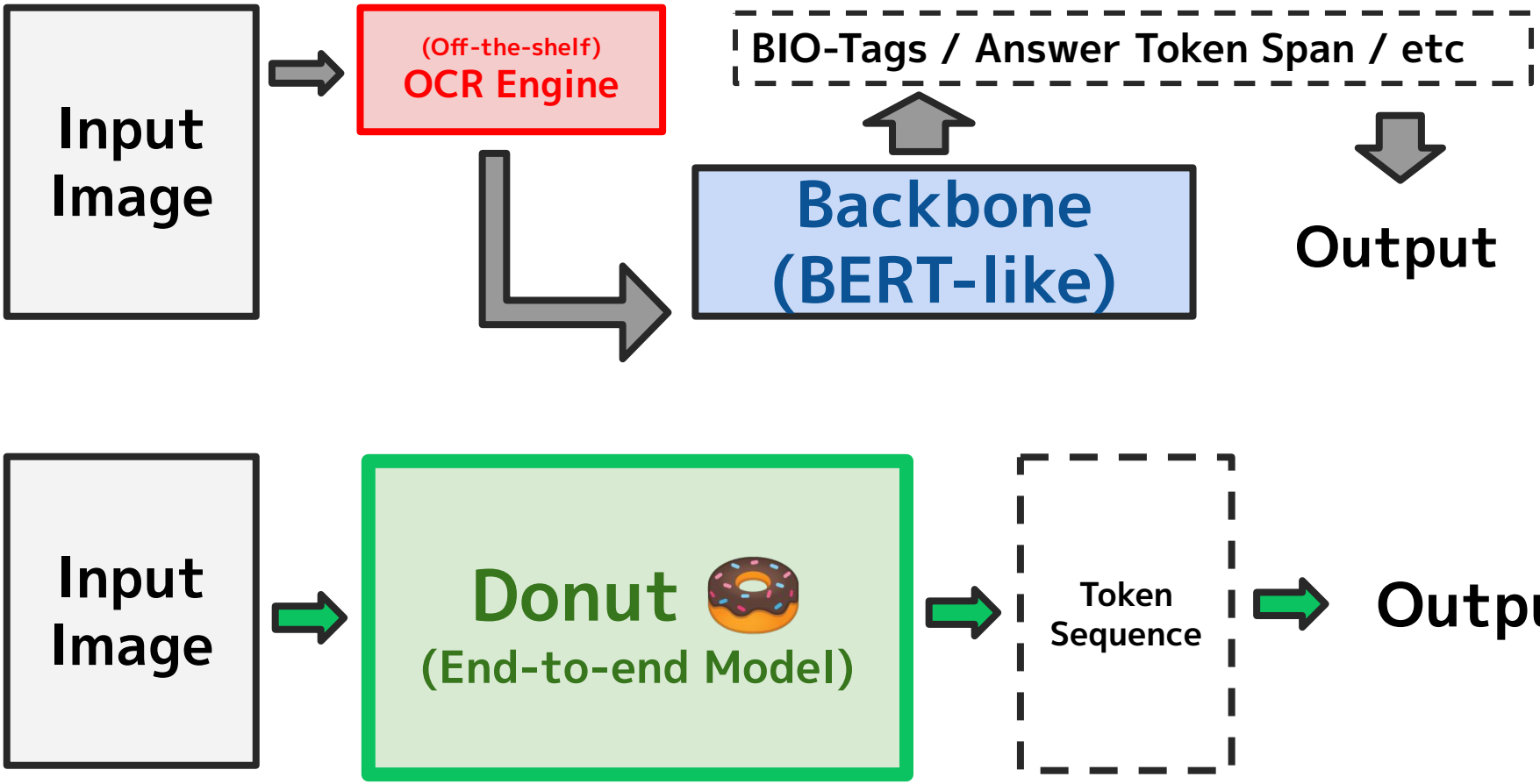




This shows why we need an OCR-free method.

Slide

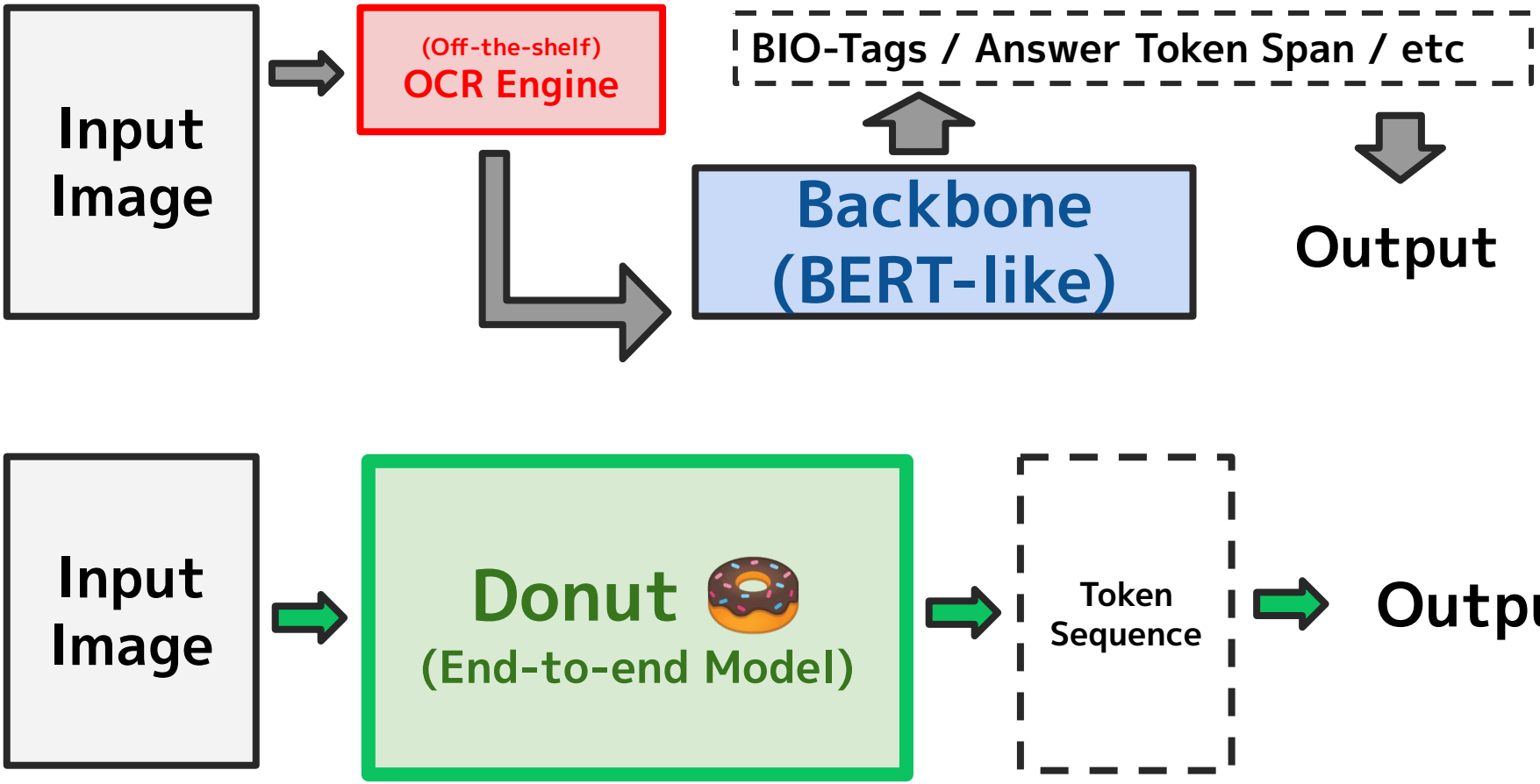
Conclusions



So far, we have introduced our new OCR-free method, Donut.
More experiments and analysis can be found in the manuscript.

Slide

Conclusions

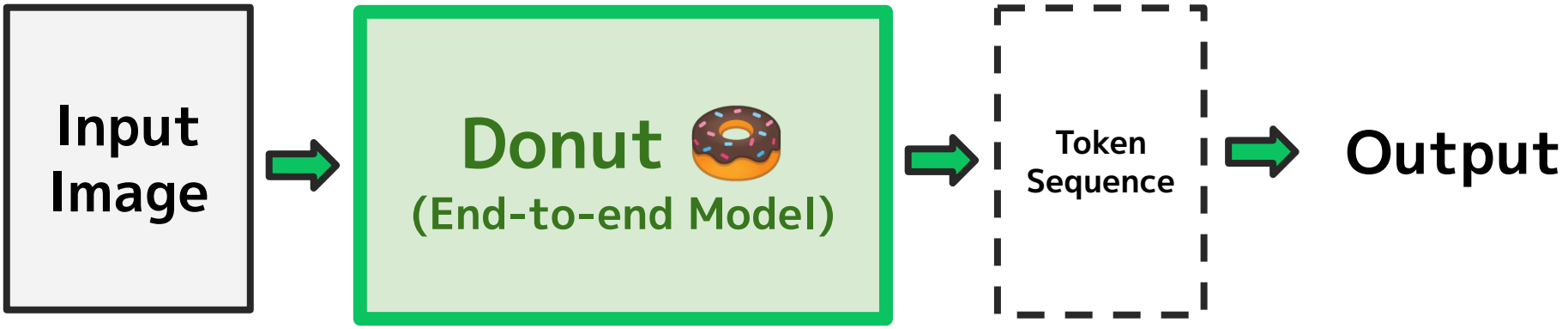


So far, we have introduced our new OCR-free method, Donut.
More experiments and analysis can be found in the manuscript.

Slide

Conclusions

- The proposed method, Donut, directly maps an input document image into a desired structured output.

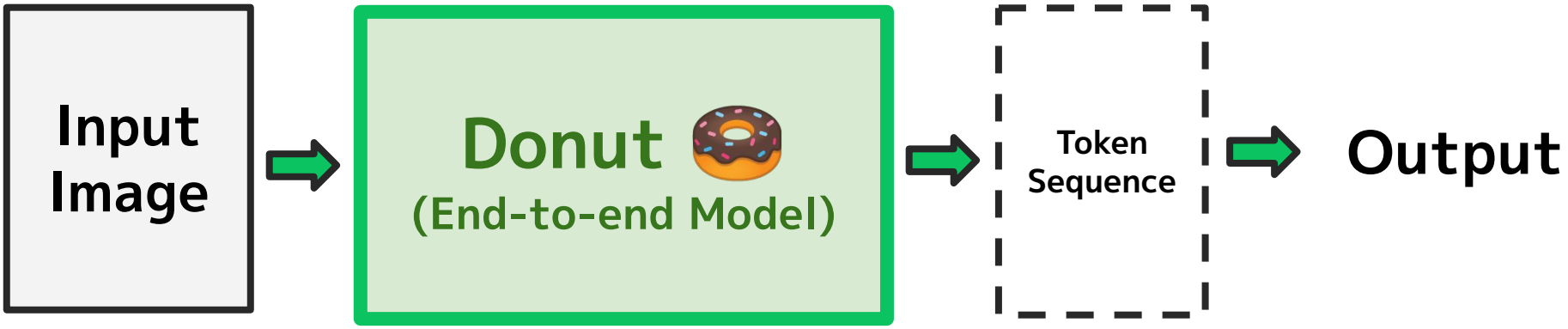


The proposed method, Donut, directly maps an input document image into a desired structured output.

Slide

Conclusions

- The proposed method, Donut, directly maps an input document image into a desired structured output.
- Unlike conventional methods, Donut does not depend on OCR and can easily be trained in an end-to-end fashion.



Unlike conventional methods, Donut does not depend on OCR and can easily be trained in an end-to-end fashion.

Slide

Conclusions

- The proposed method, Donut, directly maps an input document image into a desired structured output.
- Unlike conventional methods, Donut does not depend on OCR and can easily be trained in an end-to-end fashion.
- We also propose a synthetic document image generator, SynthDoG.

We also propose a synthetic document image generator, SynthDoG.

Slide

Conclusions

- The proposed method, Donut, directly maps an input document image into a desired structured output.
- Unlike conventional methods, Donut does not depend on OCR and can easily be trained in an end-to-end fashion.
- We also propose a synthetic document image generator, SynthDoG.
- Through extensive experiments and analyses, we show the high performance and cost-effectiveness of Donut.

Through extensive experiments and analyses,
we show the high performance and cost-effectiveness of Donut.

Slide

Conclusions

- The proposed method, Donut, directly maps an input document image into a desired structured output.
- Unlike conventional methods, Donut does not depend on OCR and can easily be trained in an end-to-end fashion.
- We also propose a synthetic document image generator, SynthDoG.
- Through extensive experiments and analyses, we show the high performance and cost-effectiveness of Donut.
- We believe our work can easily be extended to other domains/tasks regarding document understanding.

We believe our work can easily be extended to other domains/tasks regarding document understanding.

Slide

Thank you!

Contact: gwkim.rsrch@gmail.com

If you have any questions please feel free to contact me :)
Thank you!