



OCR-free Document Understanding Transformer

Geewook Kim¹ *, Teakgyu Hong⁴, Moonbin Yim², Jeongyeon Nam¹, Jinyoung Park⁵,
Jinyeong Yim⁶, Wonseok Hwang⁷, Sangdoo Yun³, Dongyoon Han³, Seunghyun Park¹

* gwkim.rsrch@gmail.com

¹NAVER CLOVA

²NAVER Search

³NAVER AI Lab

⁴Upstage

⁵Tmax

⁶Google

⁷LBox



OCR-free Document Understanding Transformer

Geewook Kim¹ *, Teakgyu Hong⁴, Moonbin Yim², Jeongyeon Nam¹, Jinyoung Park⁵,
Jinyeong Yim⁶, Wonseok Hwang⁷, Sangdoo Yun³, Dongyoon Han³, Seunghyun Park¹

* gwkim.rsrch@gmail.com

¹NAVER CLOVA

²NAVER Search

³NAVER AI Lab

⁴Upstage

⁵Tmax

⁶Google

⁷LBox

which will be presented at



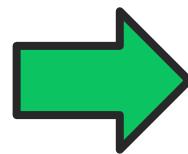
Agenda

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

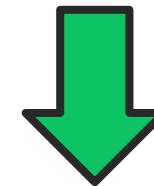
This is the table of contents. Let's start from the Introduction.



Visual Document Understanding (VDU)



VDU Model

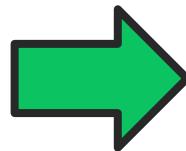


**Useful
Information**

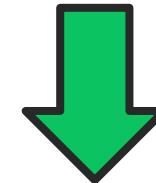
VDU aims to extract useful information from the document image. For example,



Example 1: Document Classification



VDU Model

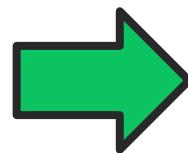


{ "class": "receipt" }

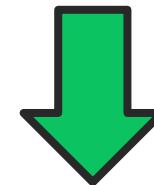
A document classifier aims to extract a category information from the image.



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.



Conventional VDU Model

Input



Output

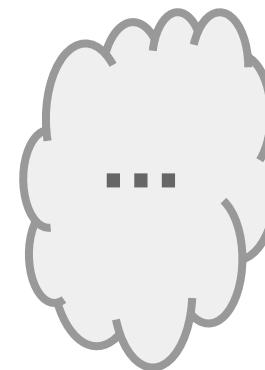
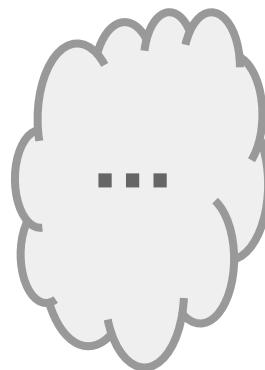
```
{ "items": [ { "name": "3002-Kyoto Choco Mochi", "count": 2, "priceInfo": { "unitPrice": 14000, "price": 28000 } }, { "name": "1001 - Choco Bun", "count": 1, "priceInfo": { "unitPrice": 22000, "price": 22000 } }, ... ], "total": [ { "menuqty_cnt": 4, "total_price": 50000 } ] }
```

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



Conventional VDU Model

Input



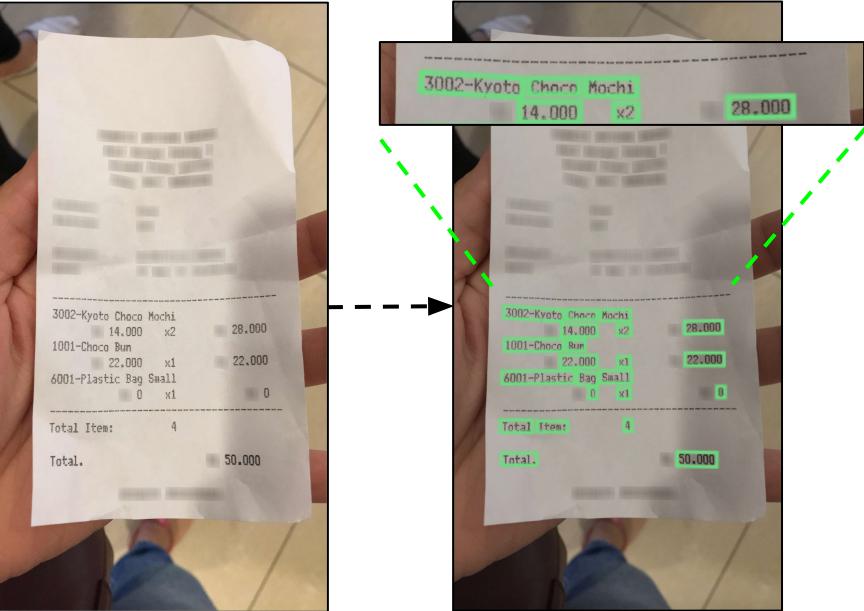
Output

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

Most conventional VDU methods share a similar pipeline.



Conventional VDU Model

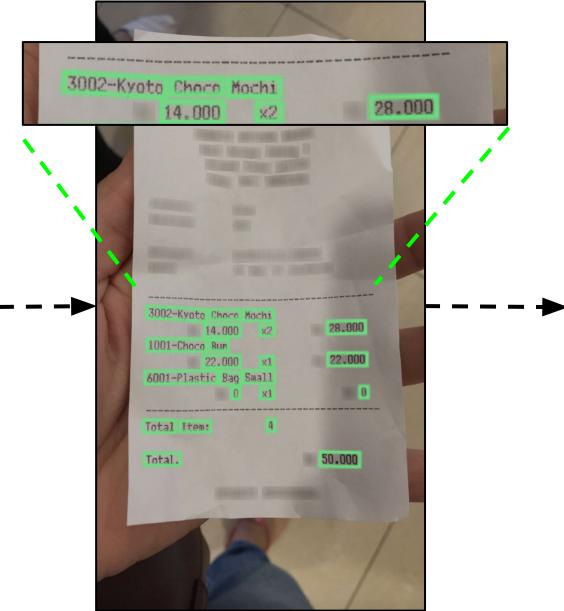


Detection!

First, a text detector finds all text boxes.



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"  
}  
]
```

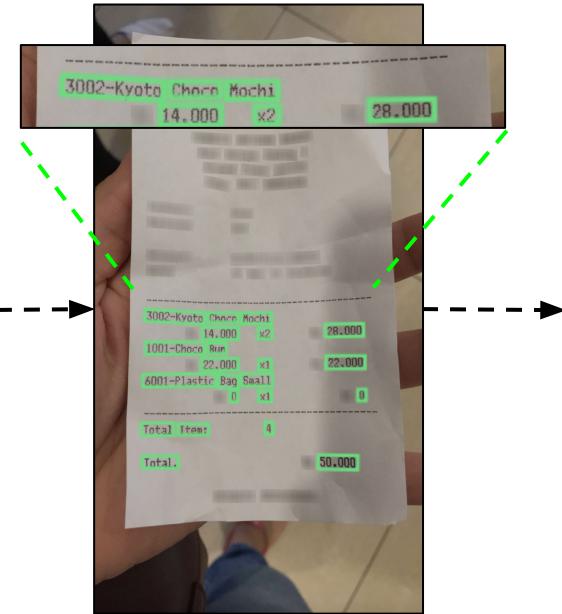
Detection!

Recognition!

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



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!

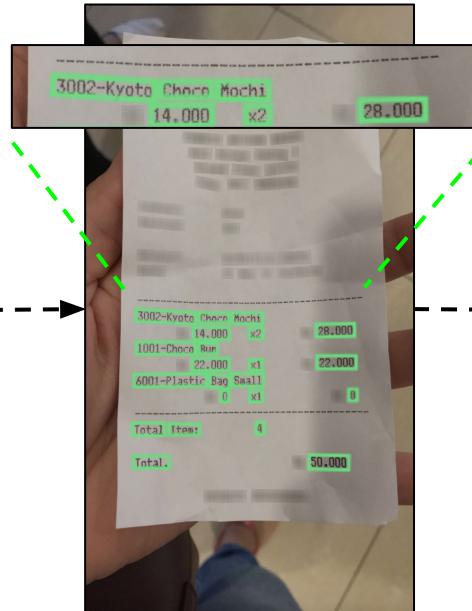
Recognition!

OCR

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



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!

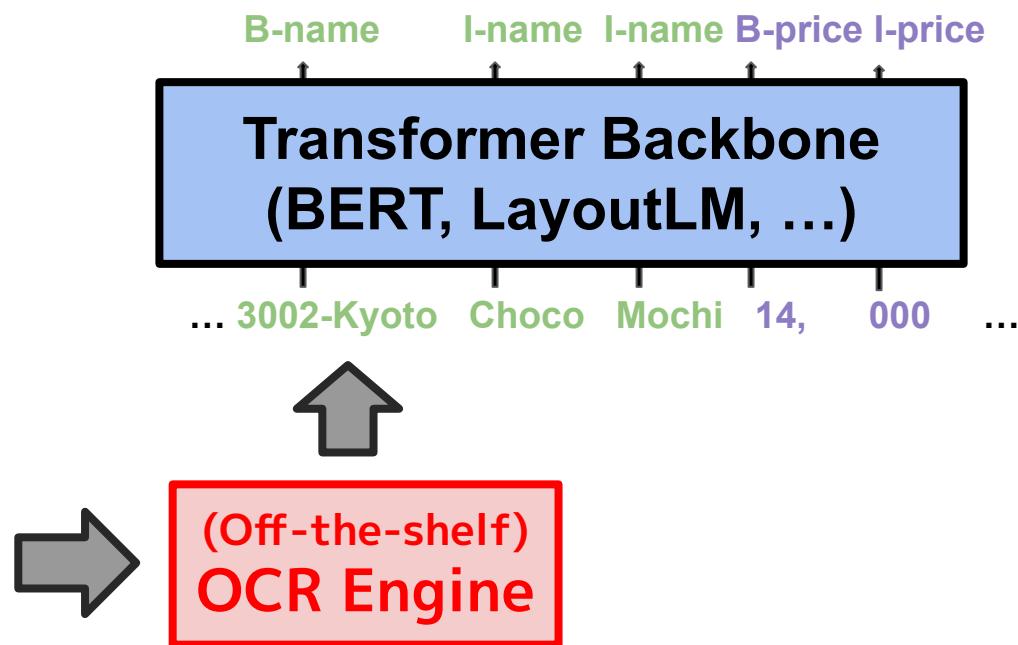
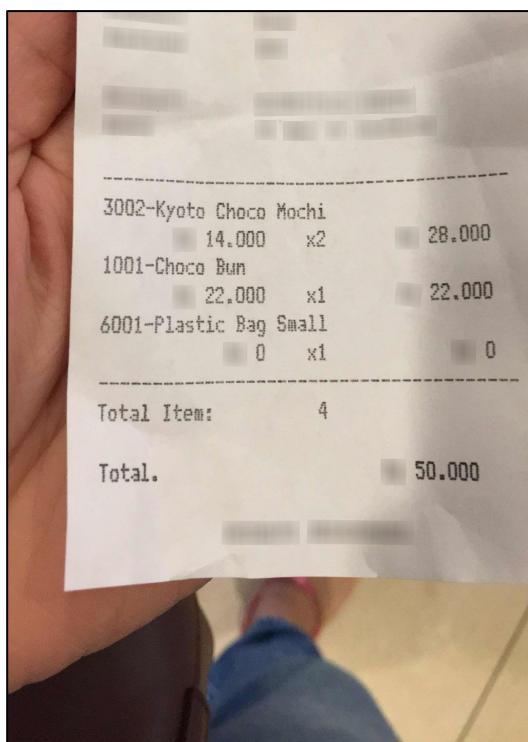
Parsing!

OCR

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



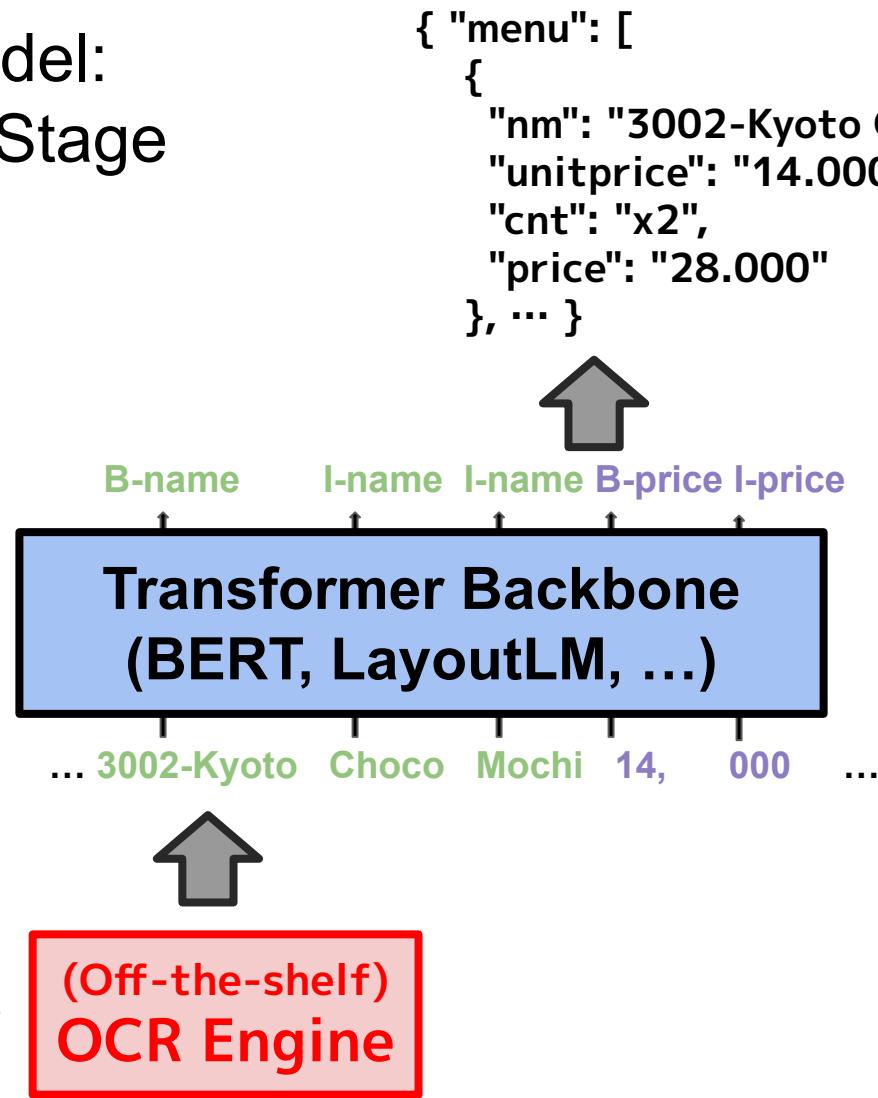
Conventional VDU Model: Details of the Parsing Stage



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



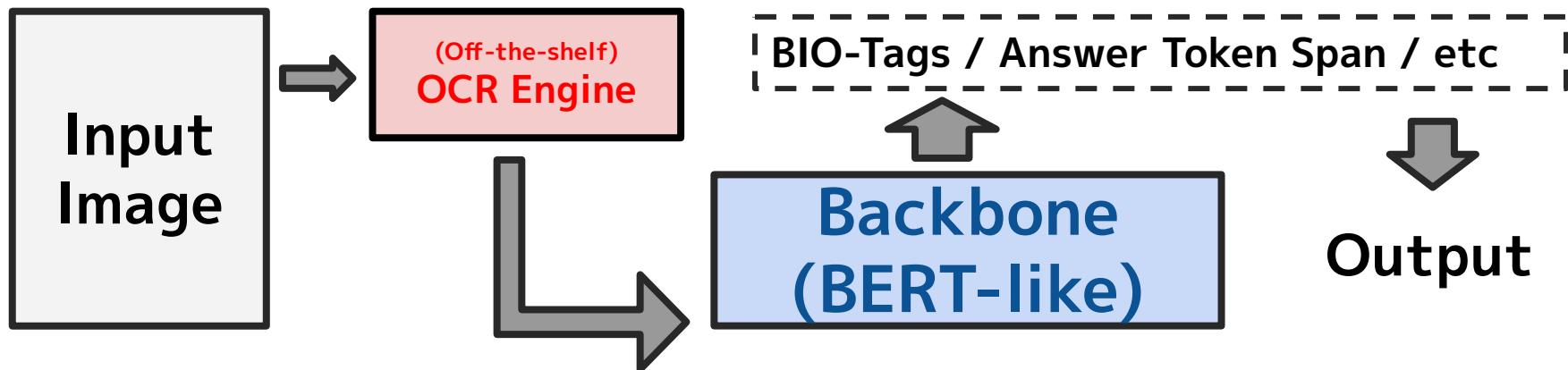
Conventional VDU Model: Details of the Parsing Stage



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



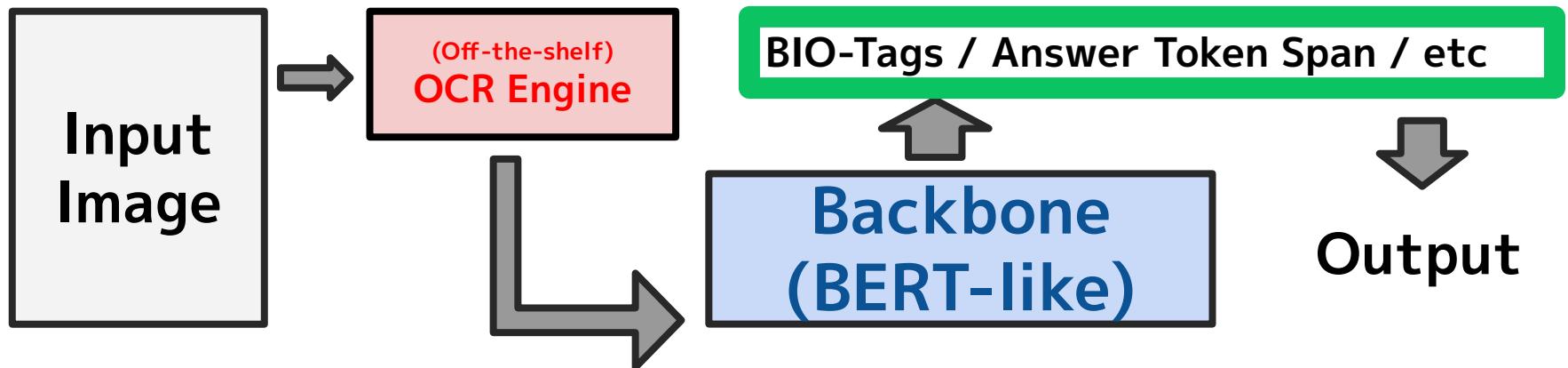
Conventional VDU Model: Overview



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



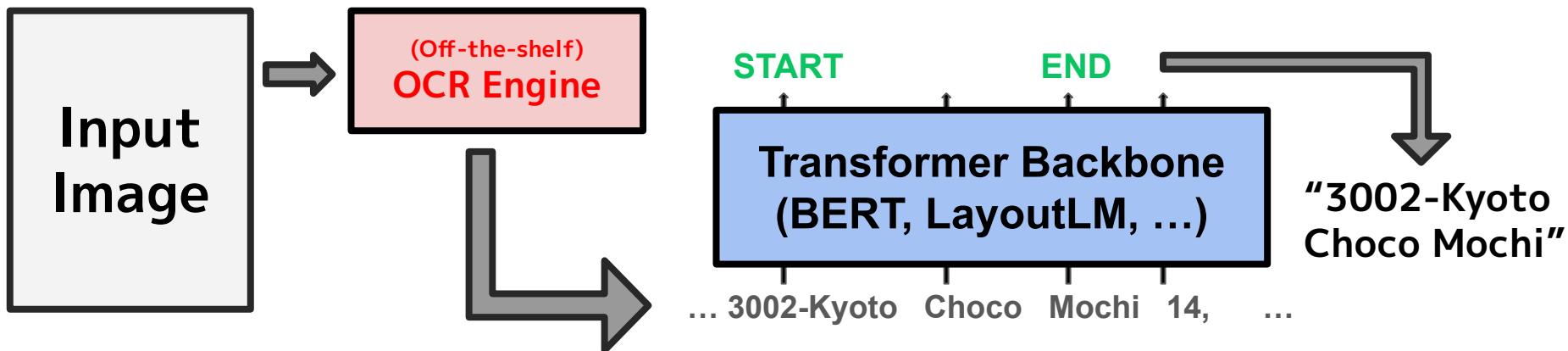
Conventional VDU Model: Overview



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



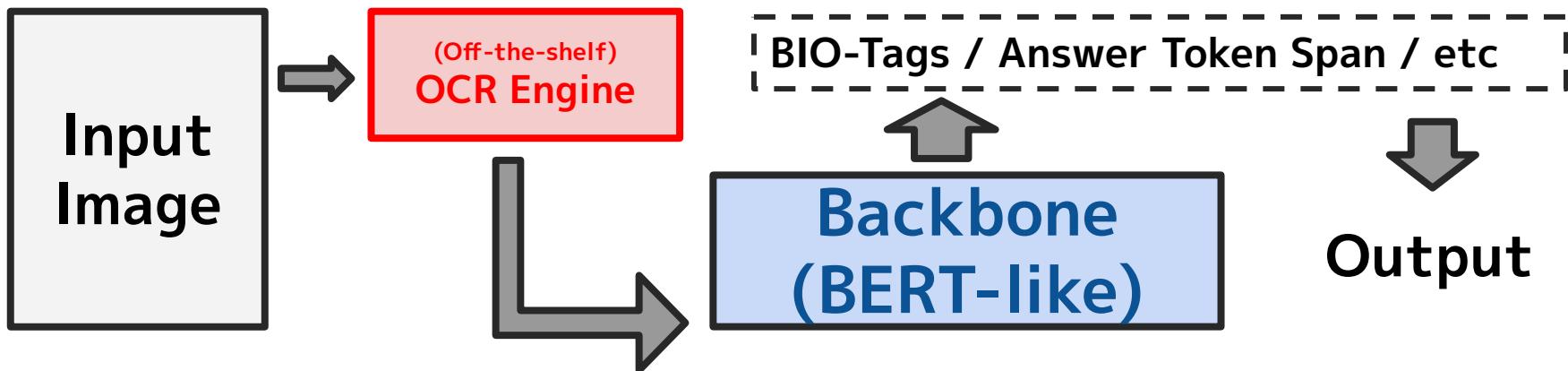
Conventional VDU Model: Overview



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



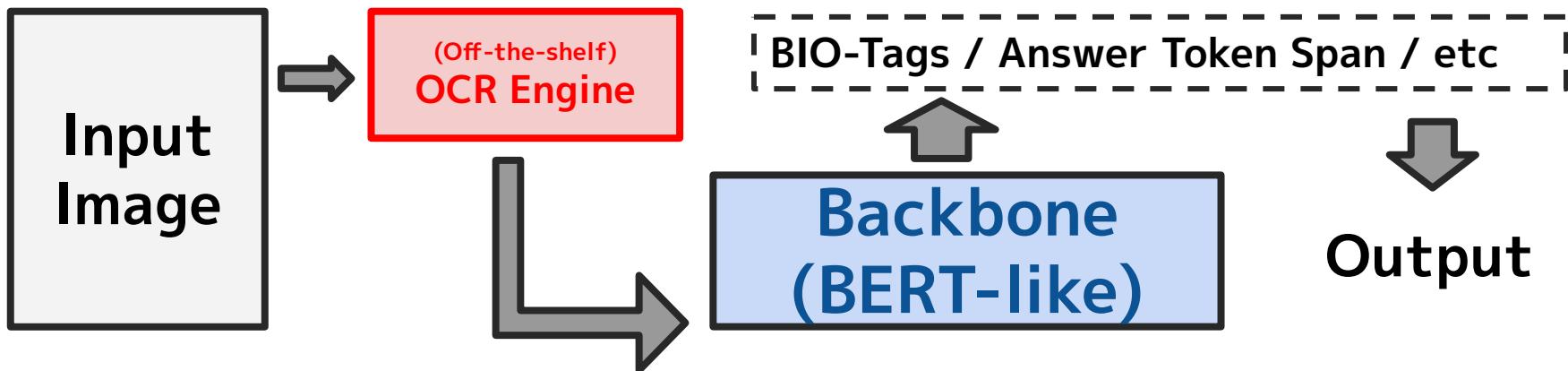
Conventional VDU Model: Overview



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



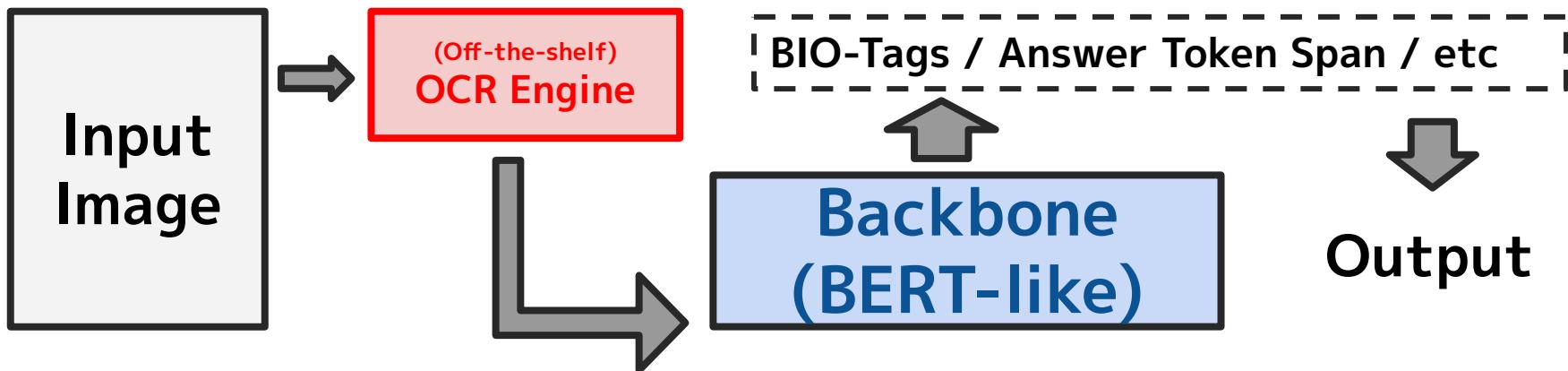
Conventional VDU Model: Overview



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



Conventional VDU Model: Overview

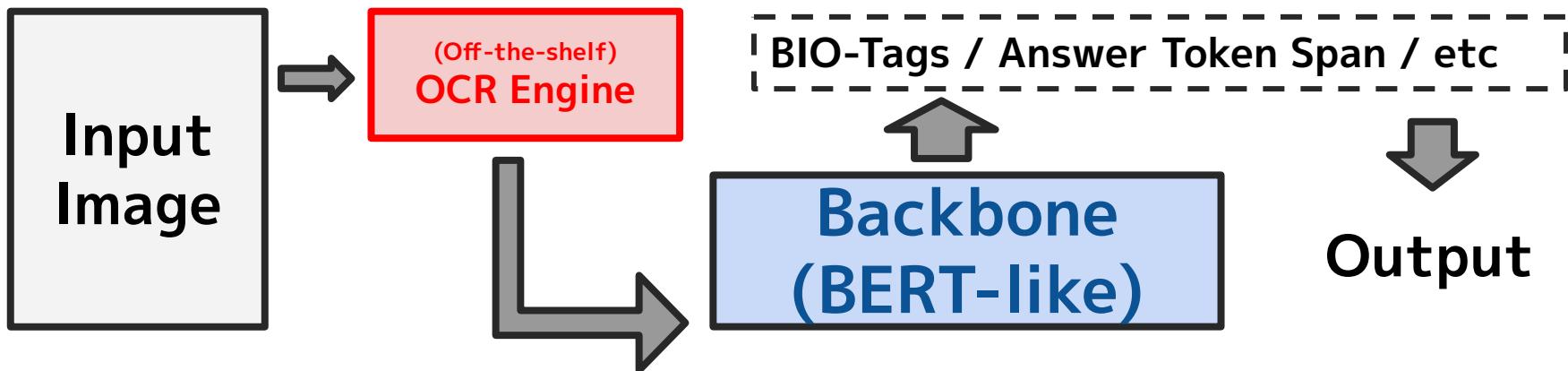


- high computational costs

First, OCR increase computational costs.



Conventional VDU Model: Overview

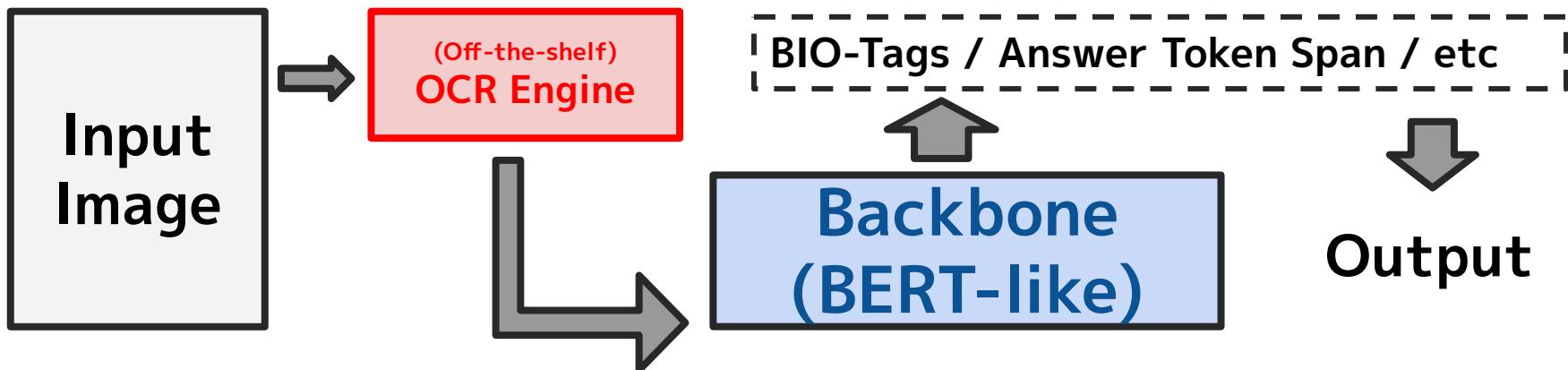


- high computational costs
- inflexibility of OCR on languages or document type

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



Conventional VDU Model: Overview

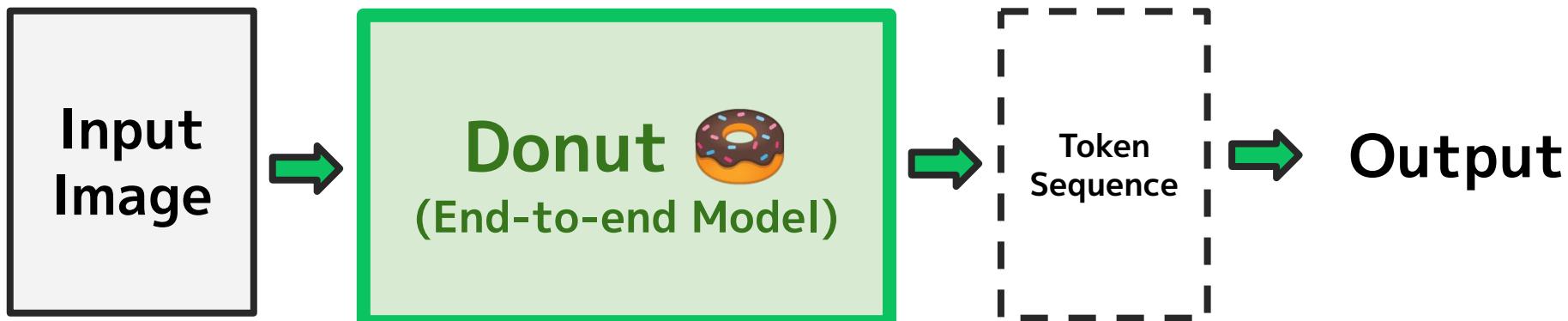


- high computational costs
- inflexibility of OCR on languages or document type
- OCR error propagation

Lastly, OCR errors are propagate to the subsequent process.



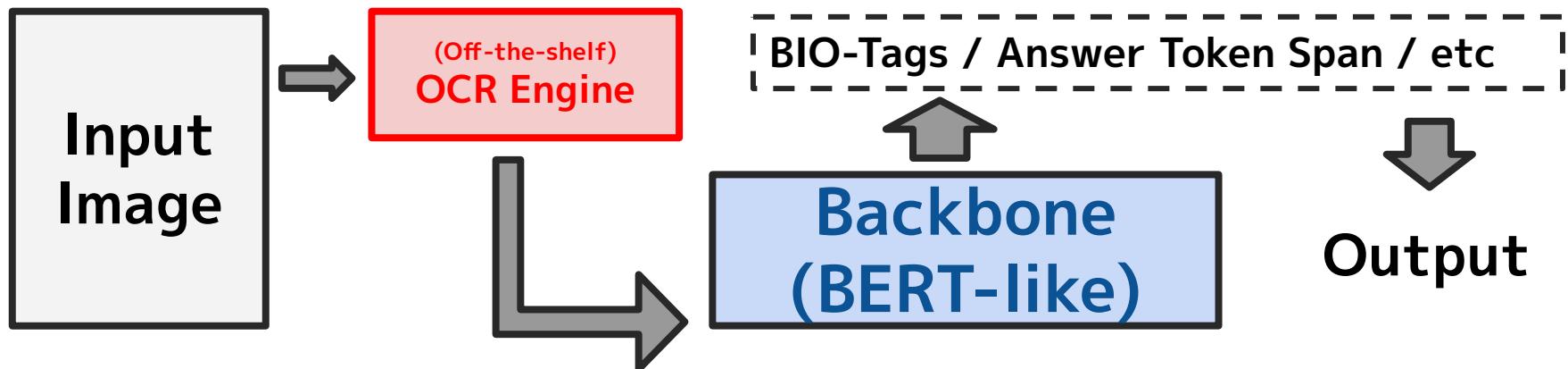
Proposal: OCR-free Approach



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



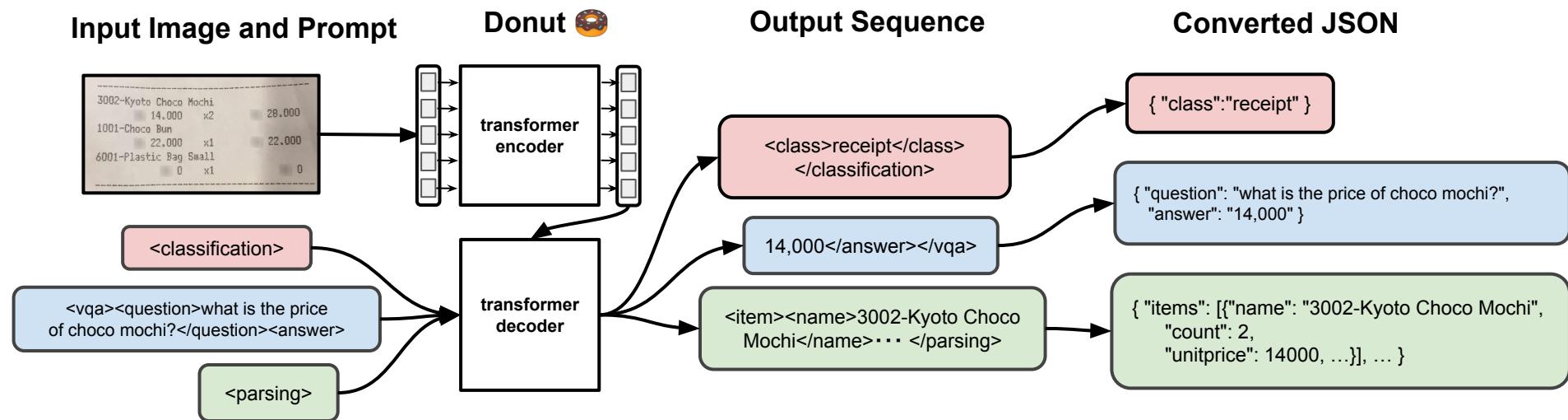
AS-IS v.s. TO-BE



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



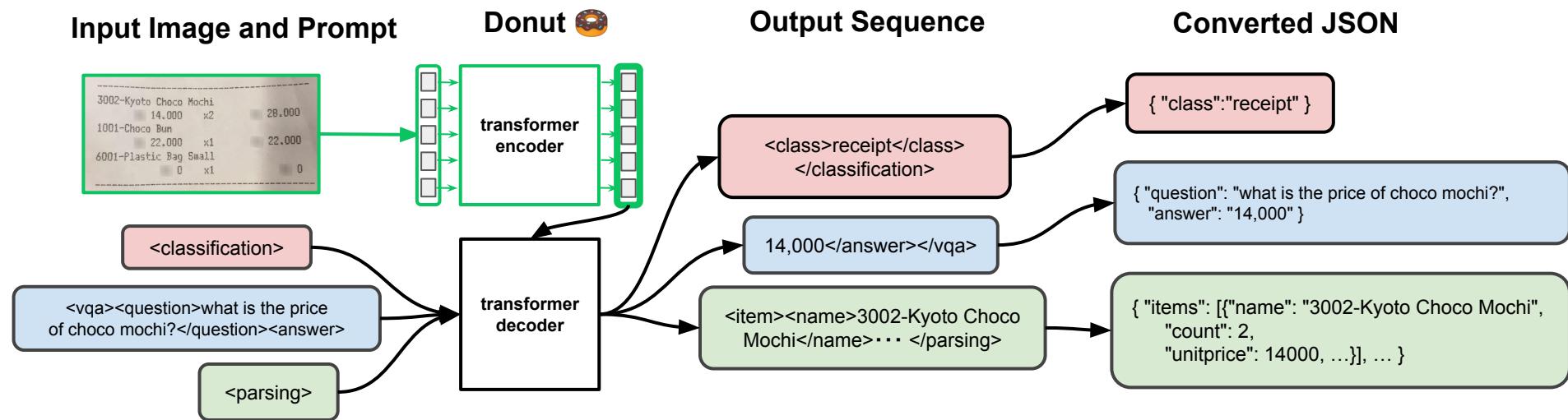
Overview



This is the overview of Donut.



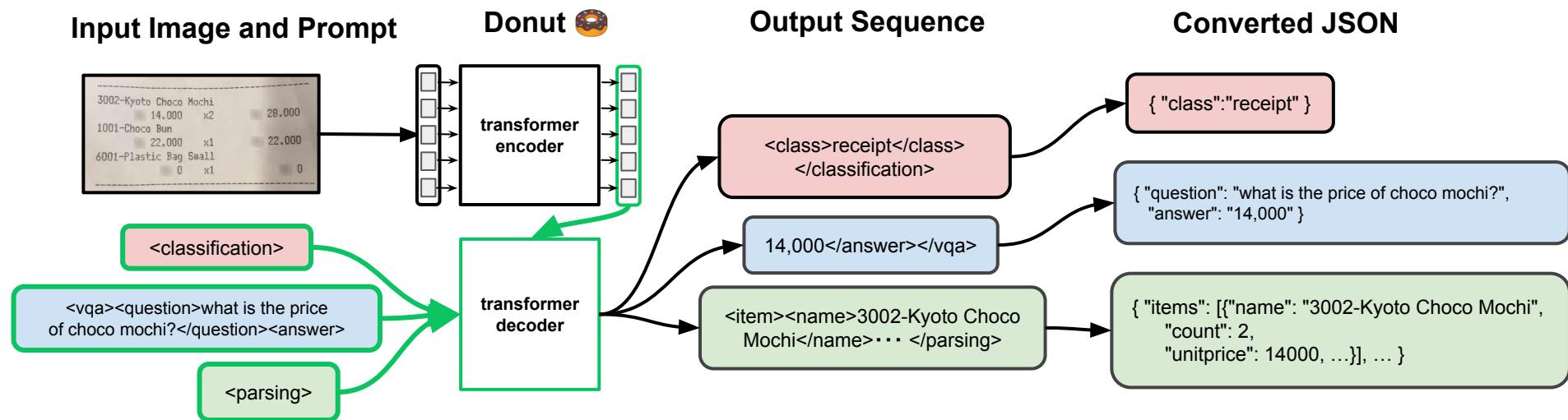
Overview



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



Overview

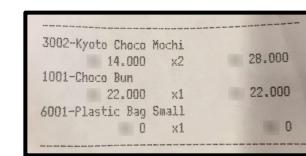


The textual decoder processes the image embeddings and prompt tokens.



Overview

Input Image and Prompt

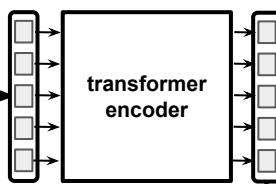


<classification>

<vqa><question>what is the price of choco mochi?</question><answer>

<parsing>

Donut 🍩



Output Sequence

<class>receipt</class>
</classification>

14,000</answer></vqa>

<item><name>3002-Kyoto Choco
Mochi</name>...</parsing>

Converted JSON

{ "class": "receipt" }

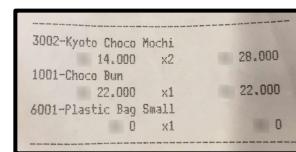
{ "question": "what is the price of choco mochi?",
"answer": "14,000" }{ "items": [{ "name": "3002-Kyoto Choco Mochi",
"count": 2,
"unitprice": 14000, ... }], ... }

Then, the decoder outputs token sequences,



Overview

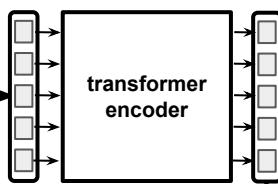
Input Image and Prompt



<classification>
<vqa><question>what is the price of choco mochi?</question><answer>

<parsing>

Donut 🍩



Output Sequence

<class>receipt</class>
</classification>

14,000</answer></vqa>

<item><name>3002-Kyoto Choco
Mochi</name>... </parsing>

Converted JSON

{ "class": "receipt" }

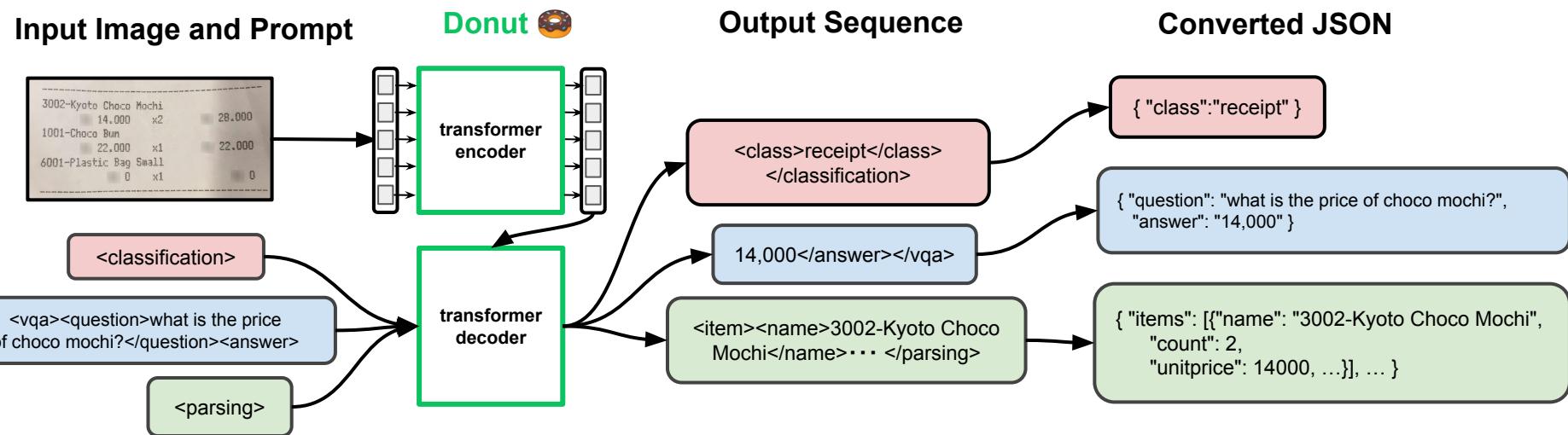
{ "question": "what is the price of choco mochi?",
"answer": "14,000" }

{ "items": [{ "name": "3002-Kyoto Choco Mochi",
"count": 2,
"unitprice": 14000, ... }], ... }

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



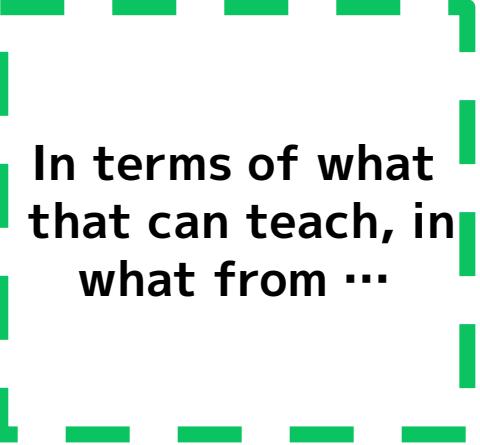
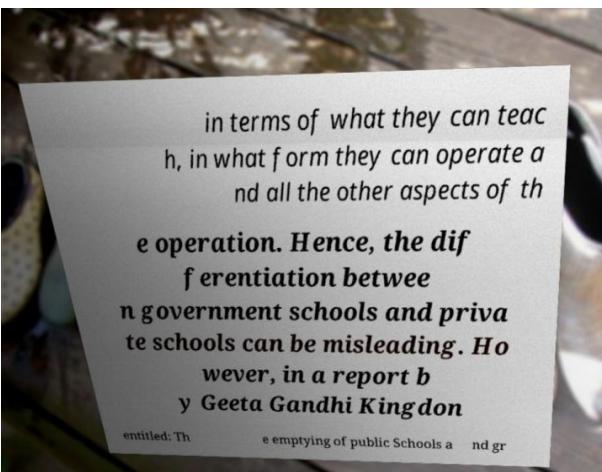
Overview: Model Architecture



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



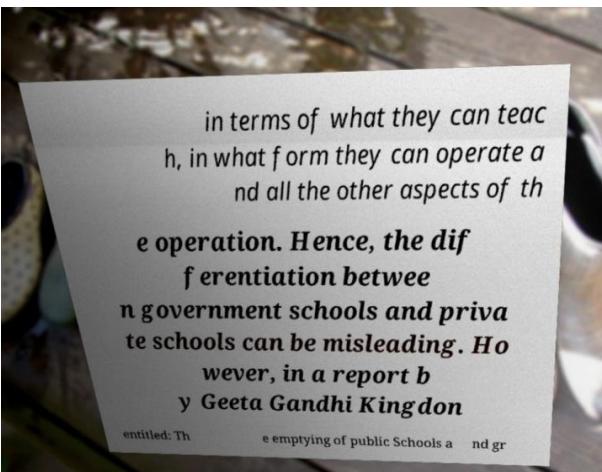
Pre-training Task



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



Pre-training Task

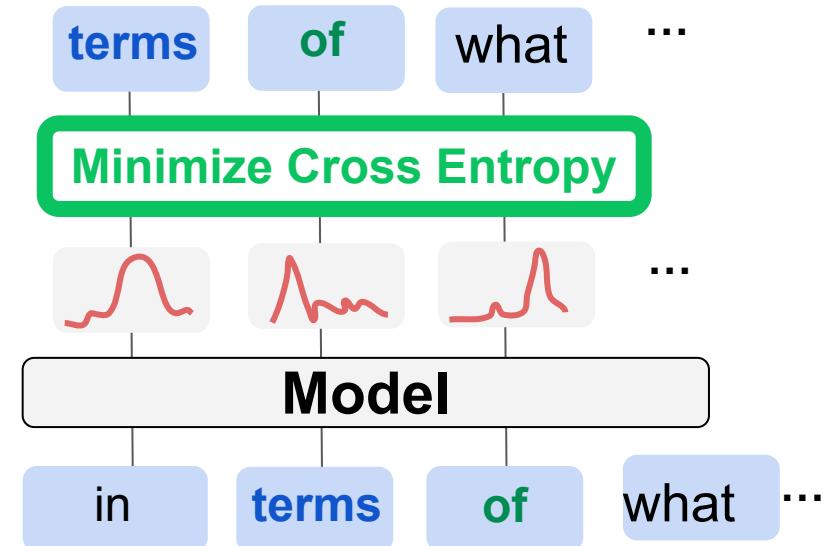
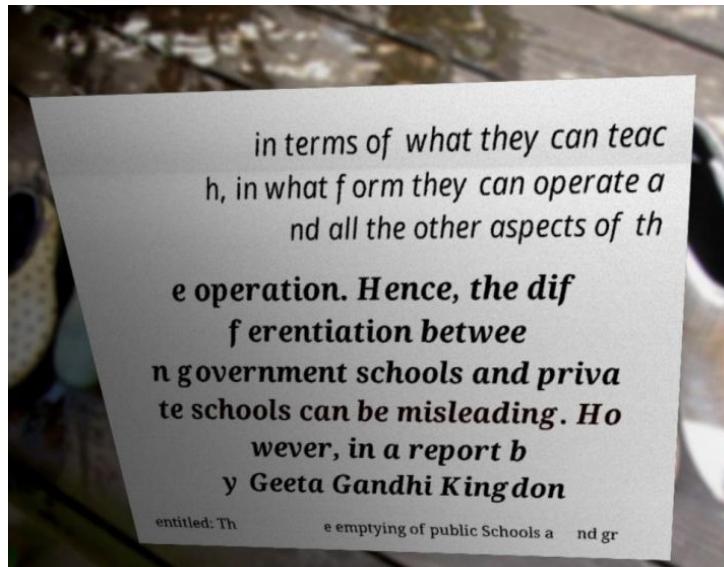


In terms of what that can teach, in what from ...

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



Training Strategy: Teacher-forcing Scheme

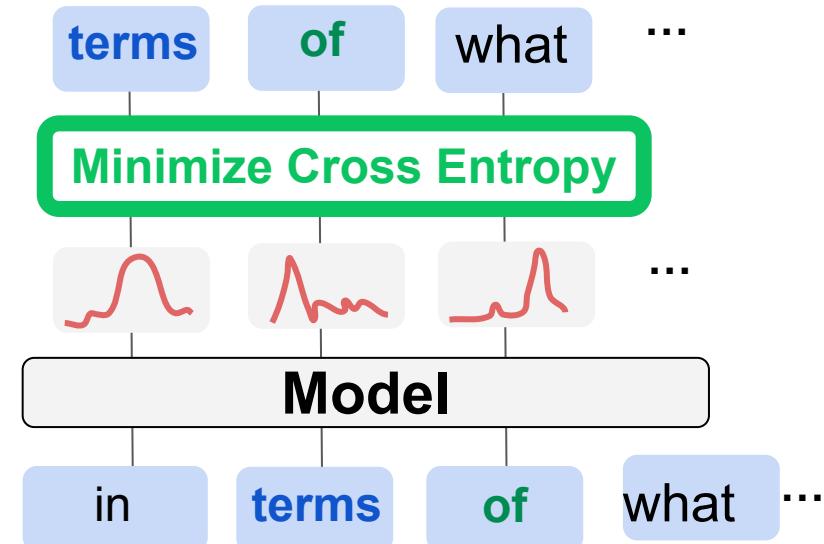
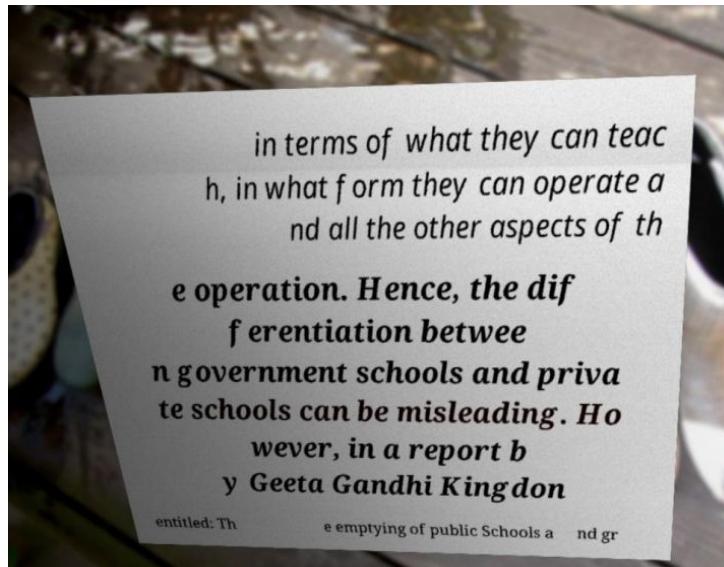


This can be interpreted as a token classification at each step.

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



Training Strategy: Teacher-forcing Scheme



This can be interpreted 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.



SynthDoG 🐶: Synthetic Document Generator



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



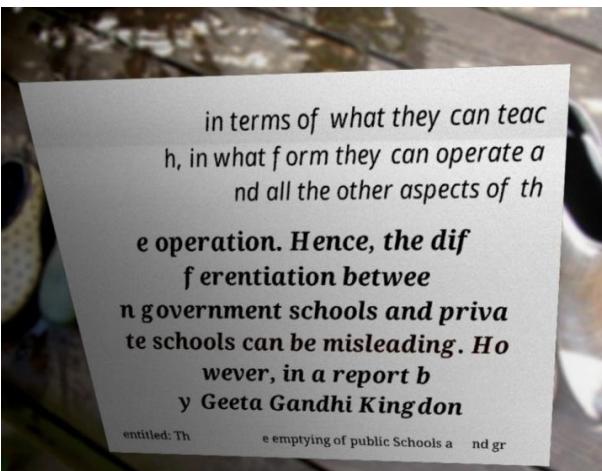
SynthDoG 🐶: Synthetic Document Generator



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



Pre-training Task

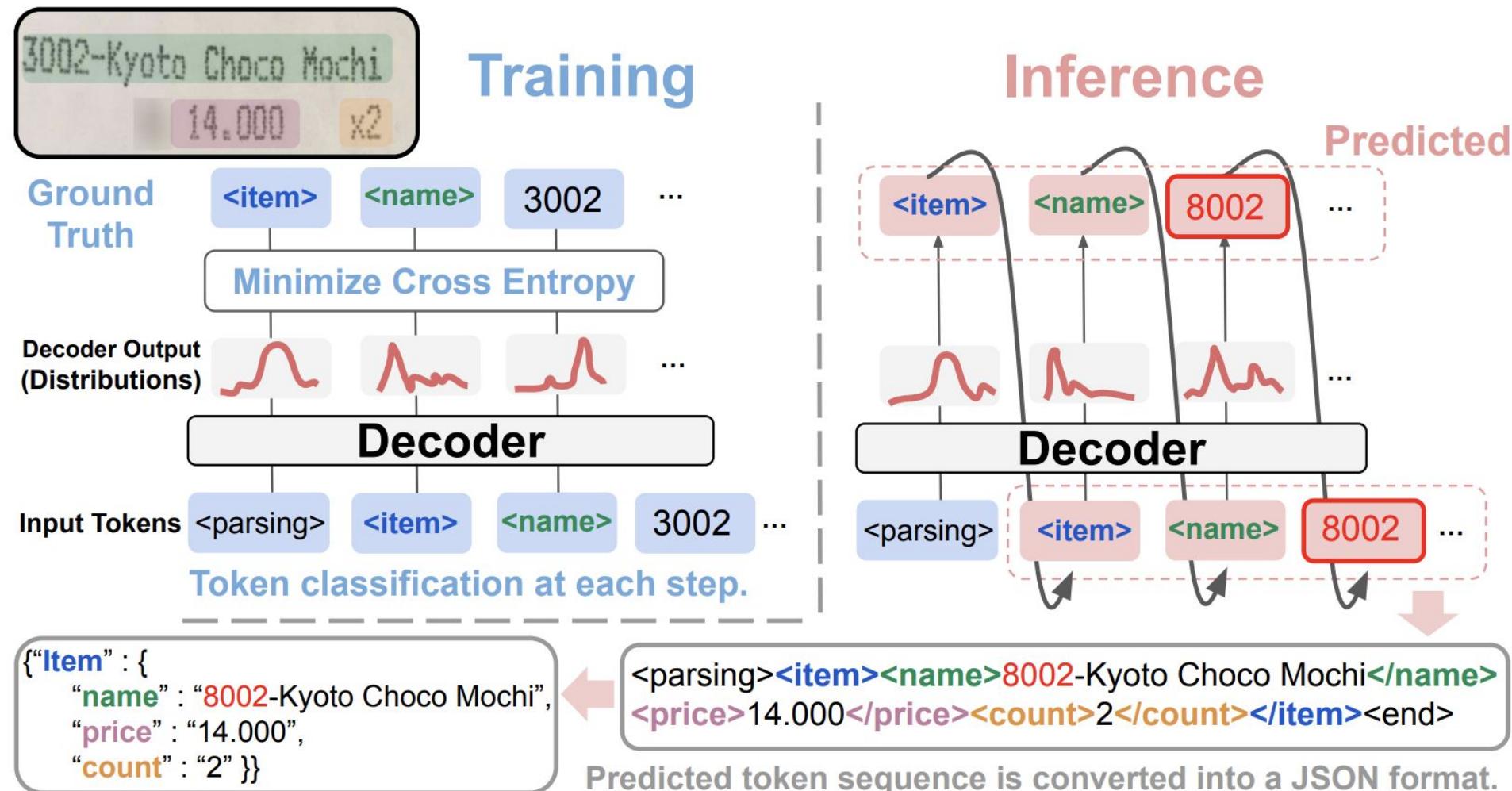


In terms of what
that can teach, in
what from ...

After the model learns “how to read”,



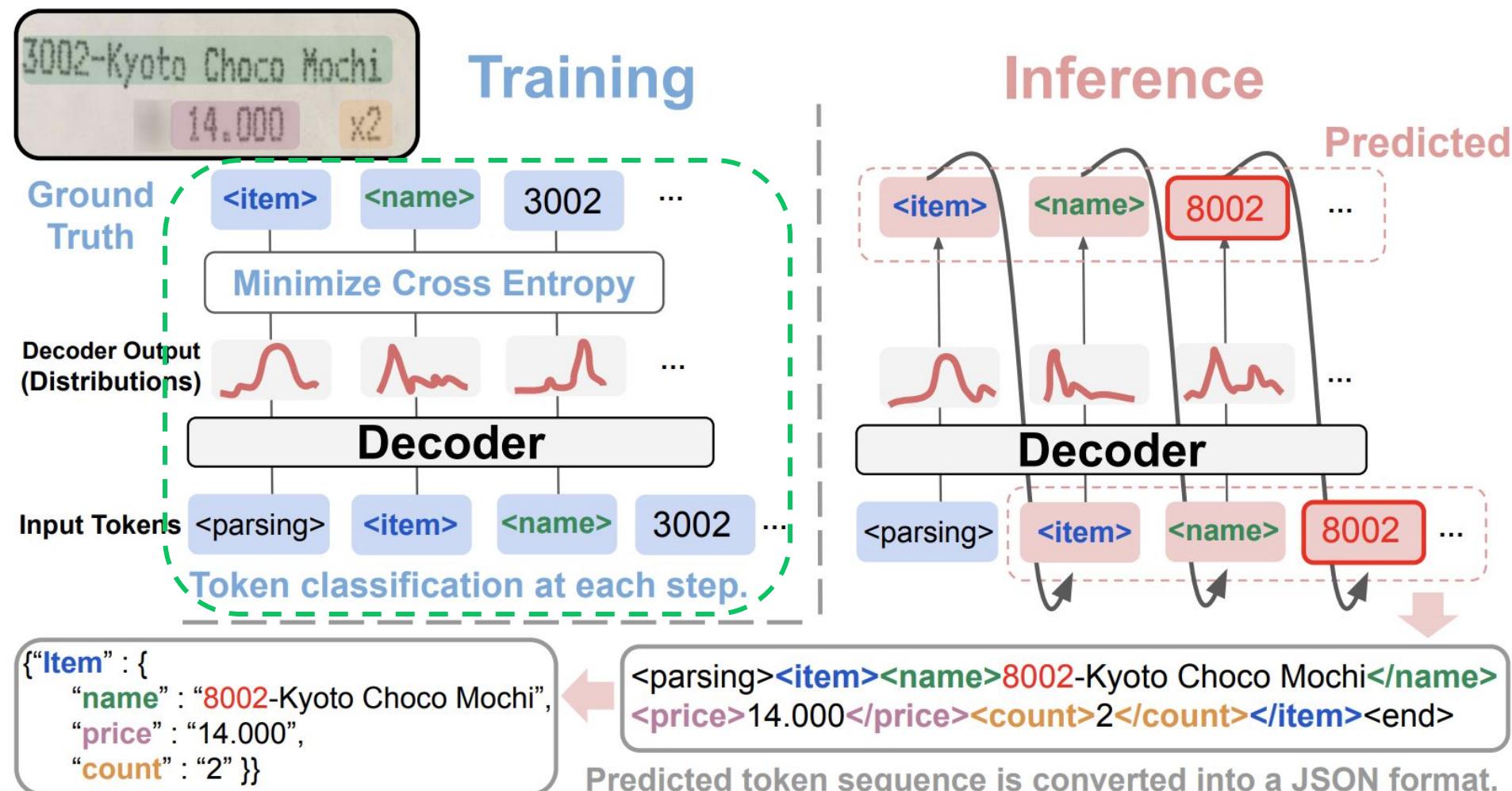
Model Fine-tuning and Inference Overview



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



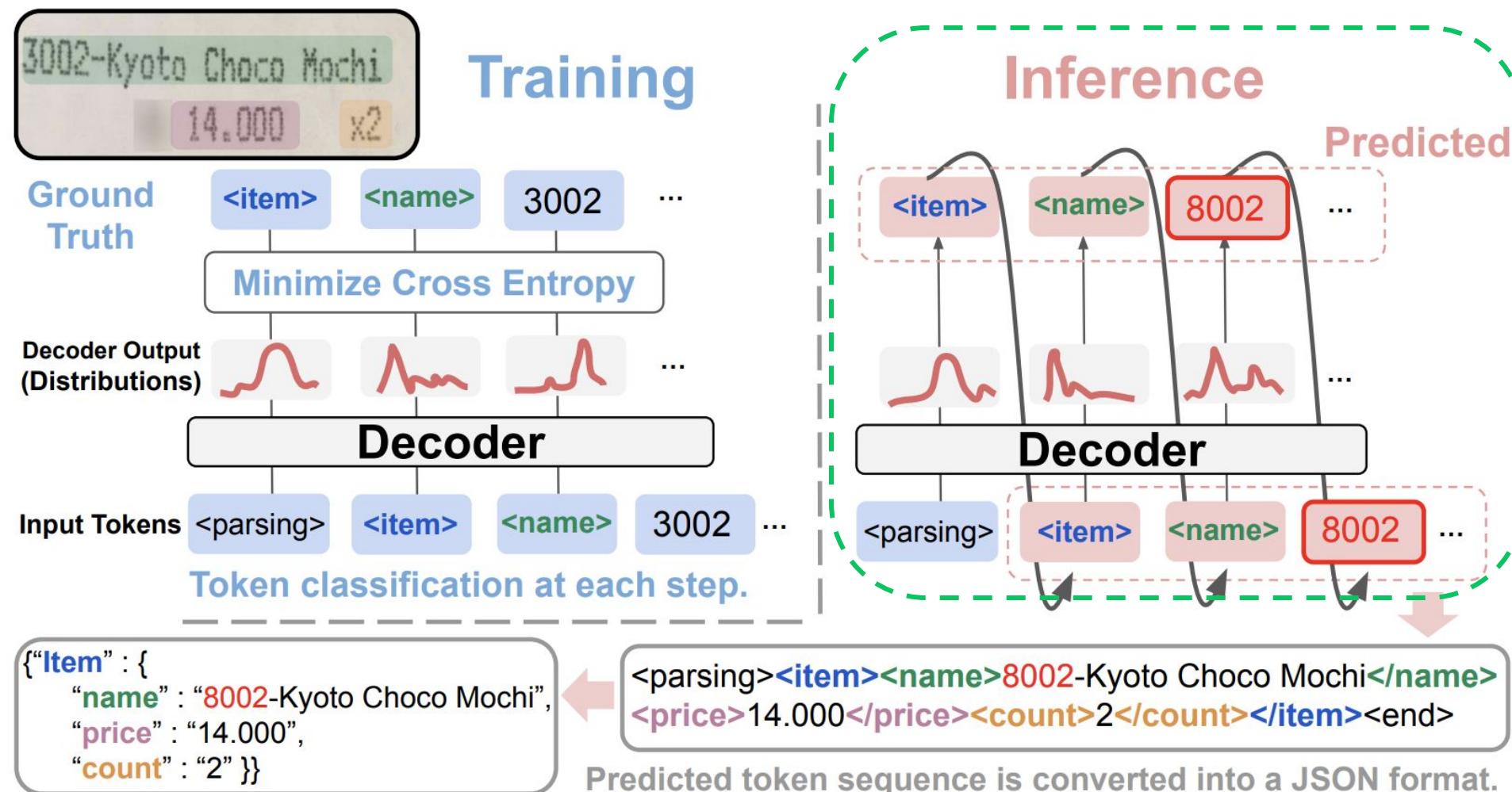
Model Fine-tuning and Inference Overview



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

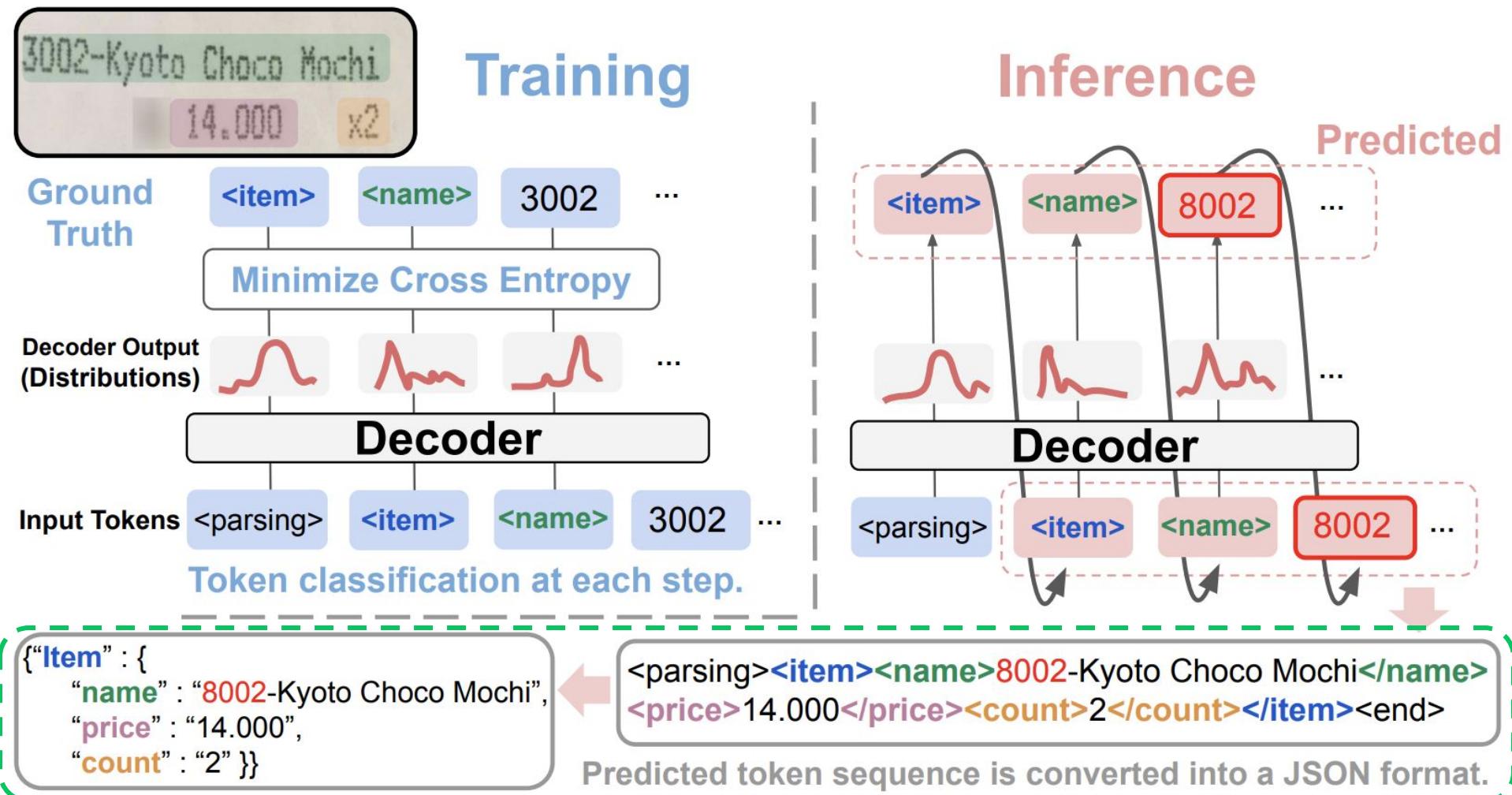


Model Fine-tuning and Inference Overview





Model Fine-tuning and Inference Overview



The predicted sequence is converted into a JSON format.



Experiments: Samples of Downstream Datasets

form

handwritten



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



Experiments: Document Classification

R.J.Reynolds Tobacco Company

AUTHORIZATION REQUEST

AR-N-137
SIC 4-107

Rec'd: CAPEL Theme Promotion Contract Amendment DATE PREPARED: 7/8/83 AR NO.: 75-479
ORIGINATOR: C. L. Sharp DEPT: 853

APPROVAL REQUEST SUMMARY

This request Management approval to amend the CAPEL Theme Promotion contract with Glandining Associates as follows:

- Fully review and submit to RURT Interne Tactical Plane for the following promotion concepts previously developed by Glandining Associates:
 - Live the Adventure Sweepstakes
 - Keep Your Machine in Winning Shape Free Premium Mail-In
 - CAPEL Movie Club Continuity Program
 - Quest Contest
- Coordinate efforts on behalf of RURT to secure the cooperation of the publisher of the Chilton Auto Guide and major automotive care companies interested in offering product discounts via the Chilton Auto Guide Mail-In Offer.
- Increase the allowance for Out-of-Pocket Expenses to include the cost for art development and typesetting.

EXPENDITURE AUTHORITY REQUESTED

Capital \$	Expense \$	25,989	Total \$	25,989
------------	------------	--------	----------	--------

EXPENDITURE TIMING:

19 - 85	Capital	Expense	85-293
19 -	(List Related AR/CA/CAC No's)		
19 -	AMOUNTS PREVIOUSLY APPROVED:		
Annual Average Thru Date	Capital \$	Expense \$	33,200
	(Cumulative)		

COMPLETE IF LEASE OR OTHER CONTINUING COMMITMENT IS INVOLVED:

Commitment \$	Per	For	Years
Total Commitment \$	Minimum Commitment \$	Time	%

EFO IMPACT

19 85	19	19	Annual Average Thru Date
Profit/Loss	(\$23,989)		
Budget/Pan Change	<input type="checkbox"/> Yes	<input type="checkbox"/> No	

REVIEWED BY:

Dept.	Initials	Date	Manager	Initials	Date	Manager	Initials	Date
Ex-Off.	SP	7/8/83		WHD	7/8/83			
Prod.	CLS	7/8						
Ex-Off.	TBO	7/8						
Ex-Off.	GTR	7/10						
Ex-Off.	DPS	7/10						
Law	PL	7/11						
Risk Mgt	MJ	7/14						

Person Responsible for Implementing: _____

APPROVALS (Originator Enters Initials of Required Approval)

Dept.	Initials	Date	Manager	Initials	Date	Manager	Initials	Date
EFO								
Operations								
Fin. & Admin.								
Pres. & CEO								
Other								

form

Person ID No.: _____ Project ID No.: _____



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



Experiments: Document Classification

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

This is the results on the RVL-CDIP dataset.

Donut achieves state-of-the-art scores with reasonable speed and efficiency.



Experiments: Document Classification

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

This is the results on the RVL-CDIP dataset.

Donut achieves state-of-the-art scores with reasonable speed and efficiency.



Experiments: Document Parsing

input_img

output

```
{...} 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.



Experiments: Document Parsing

	OCR	#Params	CORD [45]			Ticket [12]			Business Card			Receipt		
			Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.
BERT* [22]	✓	$86_M^\dagger + \alpha^\ddagger$	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]	✓	$86_M^\dagger + \alpha^\ddagger$	1.7	74.7	70.0									
LayoutLM [65]	✓	$89_M^\dagger + \alpha^\ddagger$	1.7	78.4	81.3									
LayoutLMv2* [64,66]	✓	$179_M^\dagger + \alpha^\ddagger$	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_M^\dagger	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]	✓	$93_M^\dagger + \alpha^\ddagger$	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]	✓	$106_M^\dagger + \alpha^\ddagger$	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 performance with significantly faster inference.



Experiments: Document Parsing

	OCR	#Params	CORD [45]			Ticket [12]			Business Card			Receipt		
			Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.	Time (s)	F1	Acc.
BERT* [22]	✓	$86_M^\dagger + \alpha^\ddagger$	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]	✓	$86_M^\dagger + \alpha^\ddagger$	1.7	74.7	70.0									
LayoutLM [65]	✓	$89_M^\dagger + \alpha^\ddagger$	1.7	78.4	81.3									
LayoutLMv2* [64,66]	✓	$179_M^\dagger + \alpha^\ddagger$	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_M^\dagger	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]	✓	$93_M^\dagger + \alpha^\ddagger$	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]	✓	$106_M^\dagger + \alpha^\ddagger$	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.



Experiments: Document VQA

BUSINESS EXPENSE VOUCHER		Date Prepared 10/03/97	Page 1 of 1	Disbursement Accounting Use ONLY 963.12				
Employee Name Charles A. Blixt	Account Number 7404	Return to:						
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 MM J.L. Strauch, MM R.C. Weber, MM P.G. Crist (Jones Day Reavis & Pogue), S.L. Temko (Covington & Burling), MM C.A. Blixt, MM T.F. McKim, MM D.W. Donahue, MM R.Johne (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)								
A I C H R L A R I G E N E S	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.



Experiments: Document VQA

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

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



Experiments: Document VQA

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

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



Analysis: VQA on Handwritten Documents

05/19/99 WED 10:10 FAX 513 489 9130

THE ANSWER GROUP

THE ANSWER GROUP
 4665 Cornell Road, Suite 160
 Corporate Headquarters
 Cincinnati, Ohio 45241

JOB NUMBER: 90514 DATE: 5/19/99
 # PAGES (INCL COVER SHEET): 6 TIME: 10:15
 TO: Lynn Buzzard

COMPANY: _____
 TELEPHONE #: 336-723-6100
 FAX NUMBER: 556-125-6105
 FROM: SHARON LALLY TELN: (513) 387-2232
 FAX#: (513) 489-9130

Source: <https://www.industrydocuments.ucsf.edu/docs/xynd0004>

COMPANY: _____

TELEPHONE #: 336-723-6100

Q: What is the phone number given?

Answer: 336-723-6100Donut: 336-723-6100LayoutLMv2-Large-QG: **336-723-4100**

001

NAME OF PASSENGER		NOT TRANSFERABLE
DR. William J. Darby		
1333-530004		
13 PLD 64		
C 181		
LAX 00 884.00		
CDVIA 319.00		
740.00		
American Airlines 001 8353530005 2		

See below for Airline Form, Serial Number

HEIGHTS TRAVEL SERVICE INC BROOKLYN N.Y. 33. 68924 2 B

1 2 3

1. DENVER Y CO 20147 KENOSHA 740.00

2. PORTLAND Y CO 20147 KENOSHA 740.00

3. CHICAGO - O'HARE Y CO 20147 KENOSHA 740.00

LAGUARDIA

704.76

35.24 740.00

Source: <https://www.industrydocuments.ucsf.edu/docs/ydcp0227>

NAME OF PASSENGER		NOT TRANSFERABLE
DR. William J. Darby		

Q: What is the name of the passenger?

Answer: DR. William J. DarbyDonut: DR. William J. DarbyLayoutLMv2-Large-QG: **DR. William J. Jarry**

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



Analysis: VQA on Handwritten Documents

05/19/98	WED 10:19	FAX 513 489 9130	THE ANSWER GROUP
			
THE ANSWER GROUP 4665 Cornell Road, Suite 160 Corporate Headquarters Cincinnati, Ohio 45241			
JOB NUMBER:		<u>90514</u>	
# PAGES (INCL COVER SHEET):		<u>6</u>	DATE: <u>5/19/98</u>
TO:		<u>Lynn Buzzard</u>	
TELEPHONE #:		<u>336-723-6100</u>	
FAX NUMBER:		<u>556-725-6105</u>	
FROM: SHARON LALLY		TEL#:	<u>(513) 387-2232</u>
		FAX#:	<u>(513) 489-9130</u>

Source: <https://www.industrydocuments.ucsf.edu/docs/xynd0004>

COMPANY: _____
TELEPHONE #: 336-723-6100

Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-OG: 336-723- 4100

NAME OF PASSENGER		LIA		LIA		See below for Airline Form, Serial Number	
Mr. William J. Derry		153-3-530004		153-3-530004			
CLASS OF PASSENGER		1ST CLASS		1ST CLASS		60.00	
CONTRACT AIRLINE		CARRIER		CARRIER		60.00	
CONTRACT AIRLINE		CODE		CODE		7.00	
CONTRACT AIRLINE		EXPIRY		EXPIRY		7.00	
CONTRACT AIRLINE		ADX/02		ADX/02		354.00	
CONTRACT AIRLINE		COUNTRY		COUNTRY		COUNTRY	
CONTRACT AIRLINE		CITY		CITY		CITY	
CONTRACT AIRLINE		FARE BASIS		FARE BASIS		FARE BASIS	
CONTRACT AIRLINE		CLASS		CLASS		CLASS	
CONTRACT AIRLINE		ROUTE CLASS		ROUTE CLASS		ROUTE CLASS	
CONTRACT AIRLINE		FARE		FARE		FARE	
CONTRACT AIRLINE		TAX		TAX		TAX	
CONTRACT AIRLINE		TOTAL		TOTAL		TOTAL	
CONTRACT AIRLINE		104.76		104.76		104.76	
CONTRACT AIRLINE		35.24		35.24		35.24	
CONTRACT AIRLINE		740.00		740.00		740.00	
PASSENGER TICKET & BAGGAGE CHECK ISSUED BY American Airlines 001 8353530005 2 □							
SUBJECT TO CONDITIONS OF CONTRACT ON PASSENGERS COUNTRY							
DO NOT WRITE OR DRAW IN THIS PRINTED AREA ABOVE							

Source: <https://www.industrydocuments.ucsf.edu/docs/tbkn032>

NAME OF PASSENGER	NOT TRANSFERABLE
<i>Dr. William T. Barry</i>	

Q: What is the name of the passenger?

Answer: DR. William J. Darby

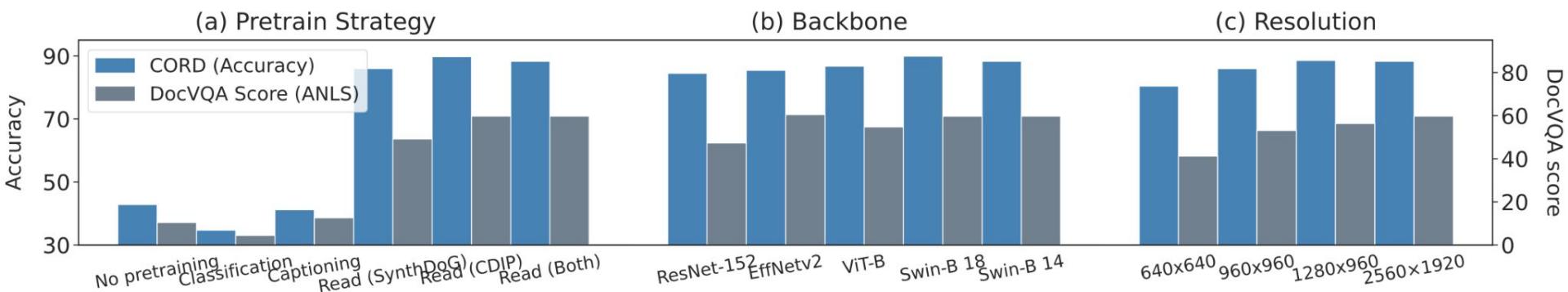
Donut: DR. William J. Darby

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

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



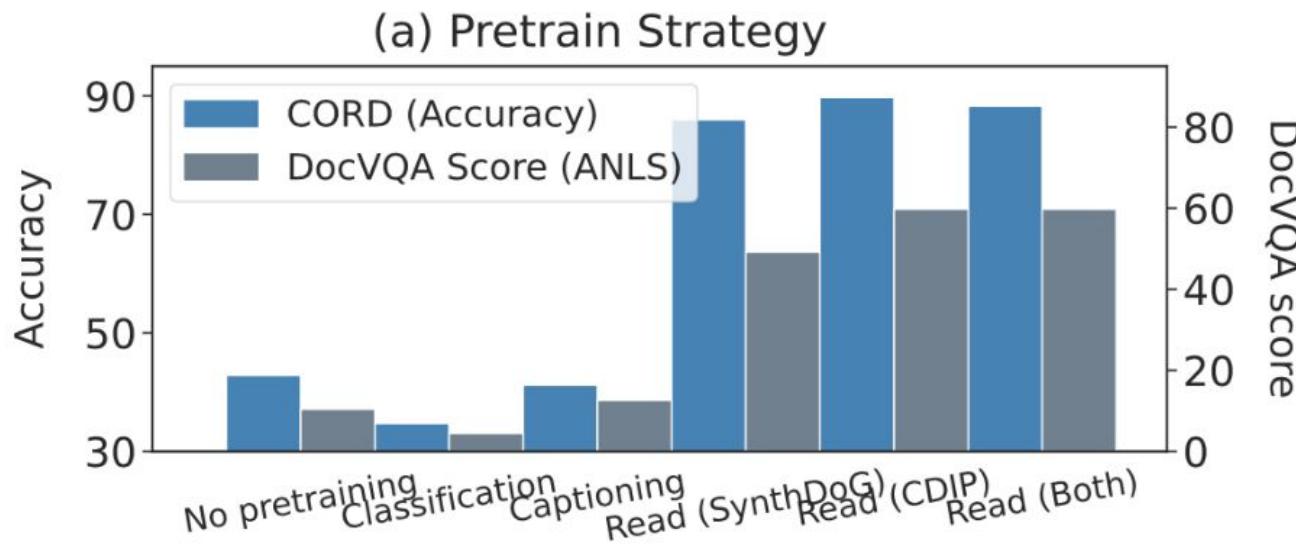
Analysis



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



Analysis: Pre-training Strategies

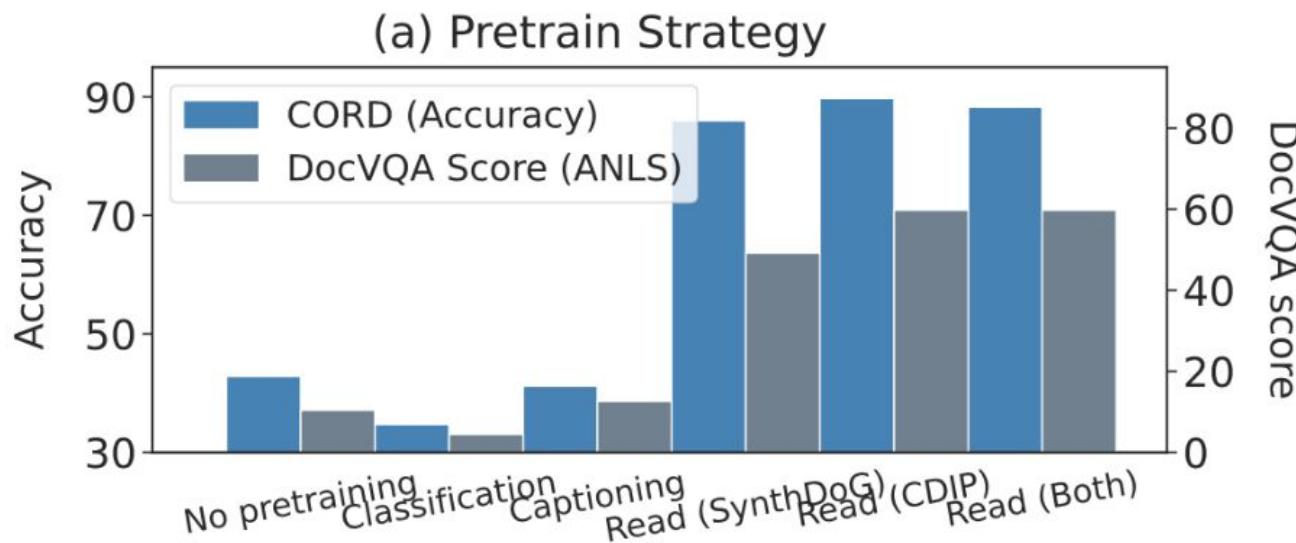


We tested several pre-training tasks.

We found that the proposed task is the most simple yet effective approach.



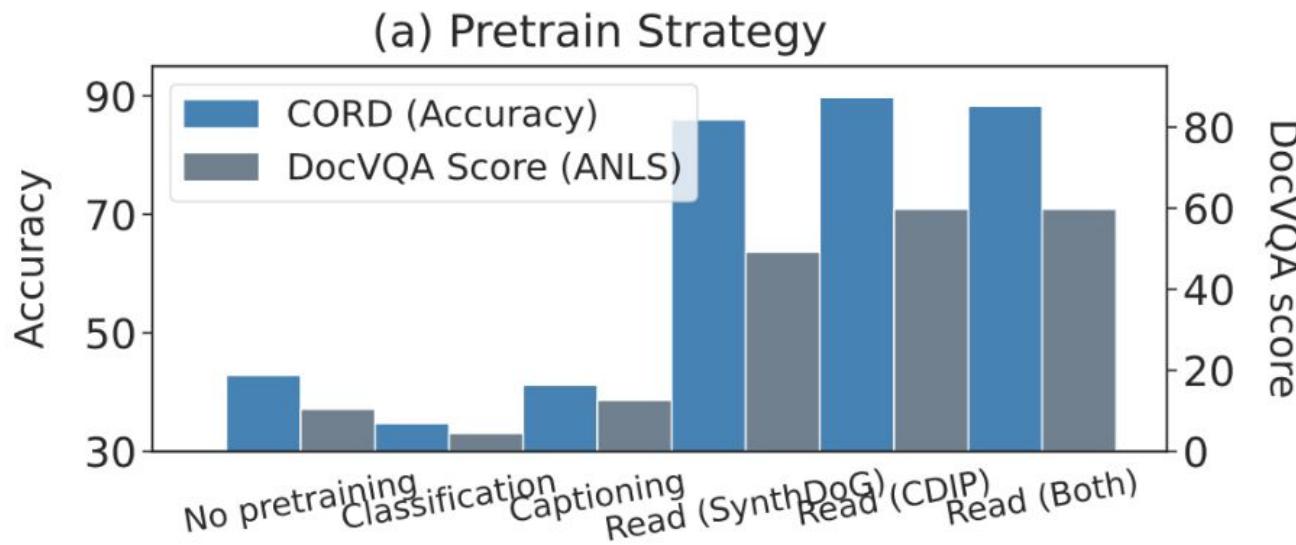
Analysis: Pre-training Strategies



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



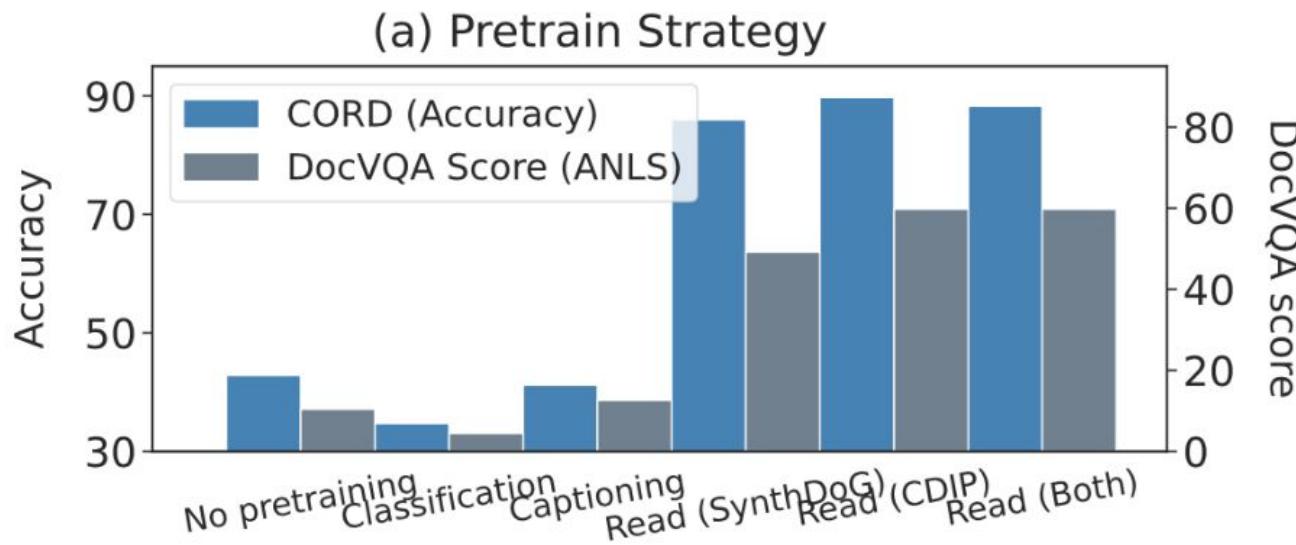
Analysis: Pre-training Strategies



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



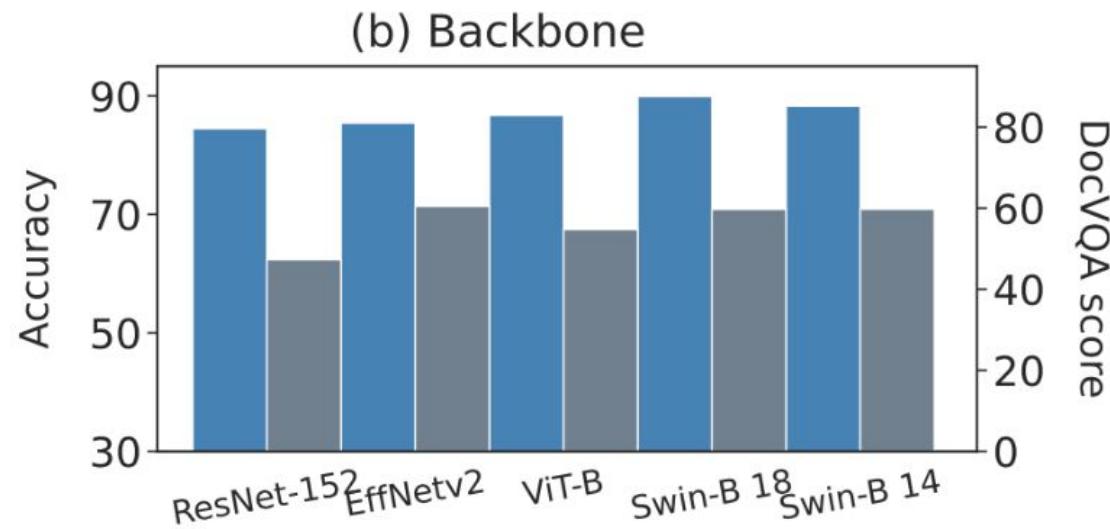
Analysis: Pre-training Strategies



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



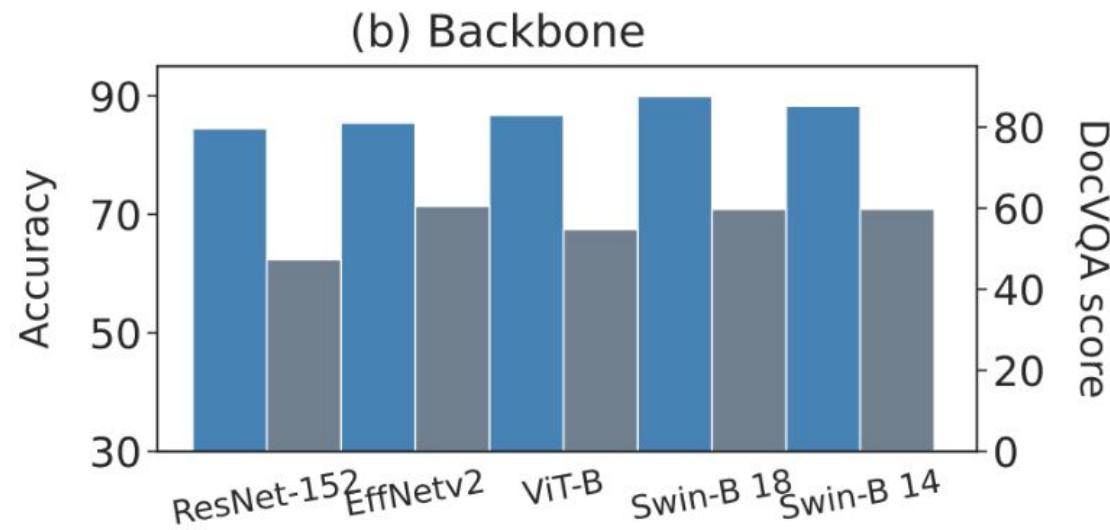
Analysis: Image Backbones



Next, we study popular image classification backbones.



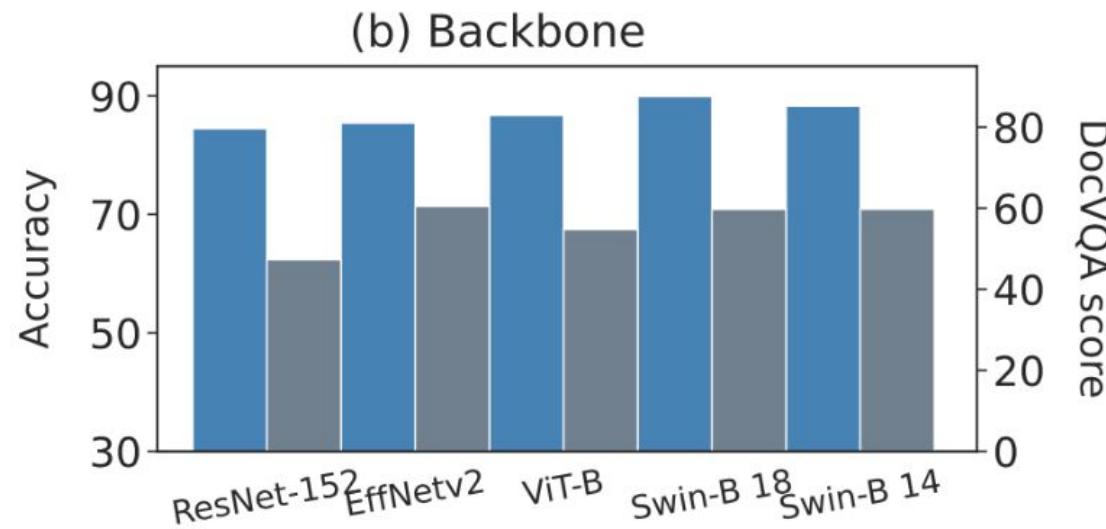
Analysis: Image Backbones



Overall, EfficientNetV2 and Swin Transformer outperform the others.



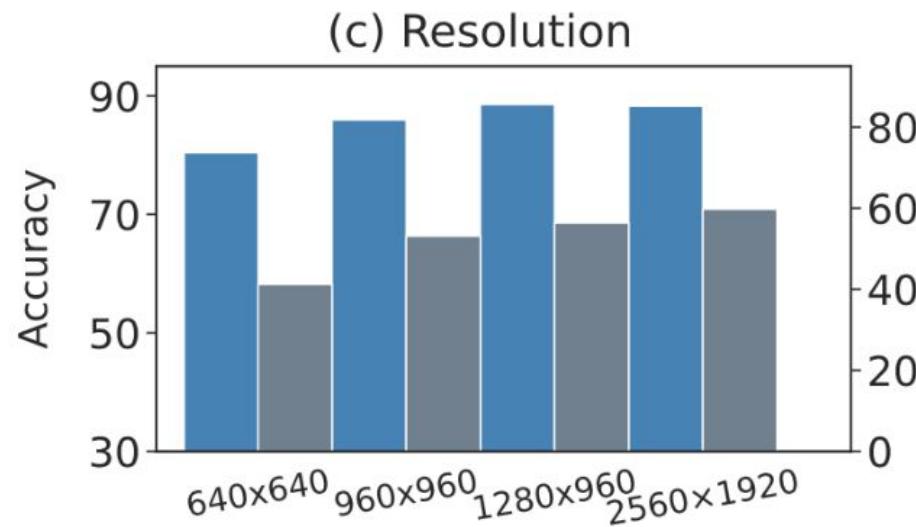
Analysis: Image Backbones



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



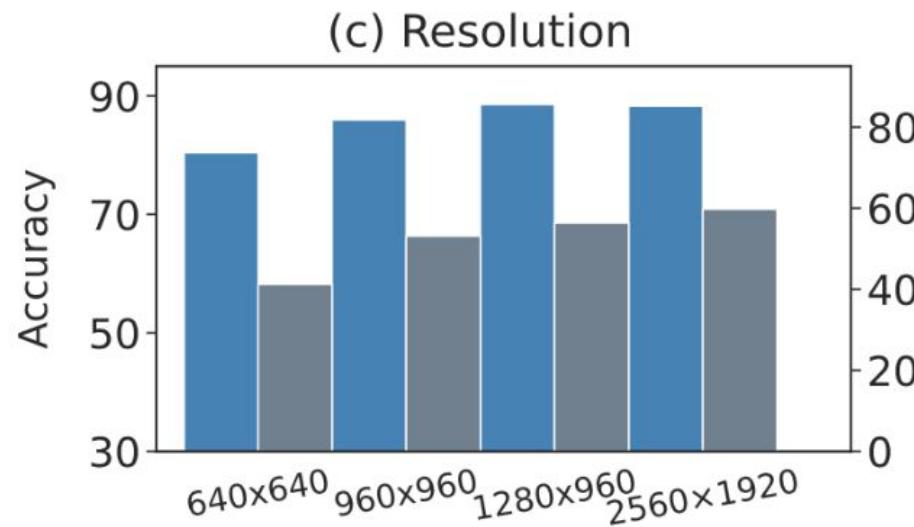
Analysis: Input Resolution



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



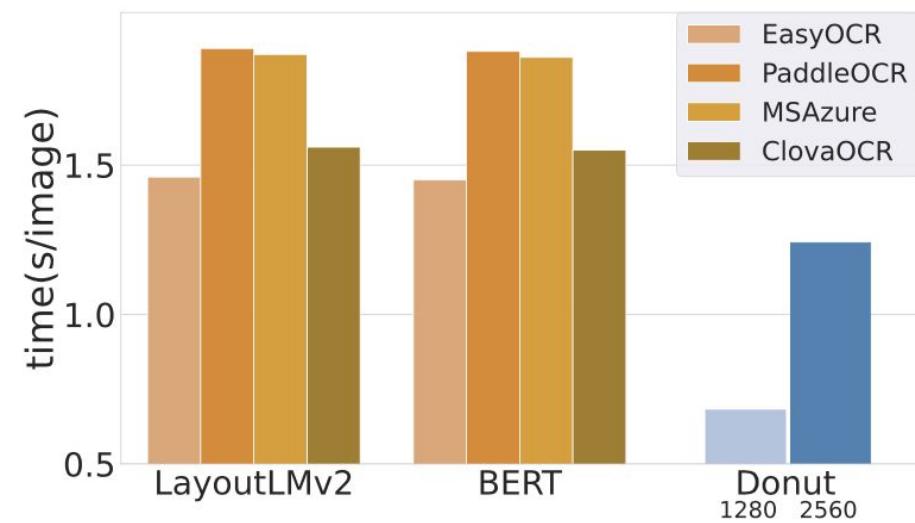
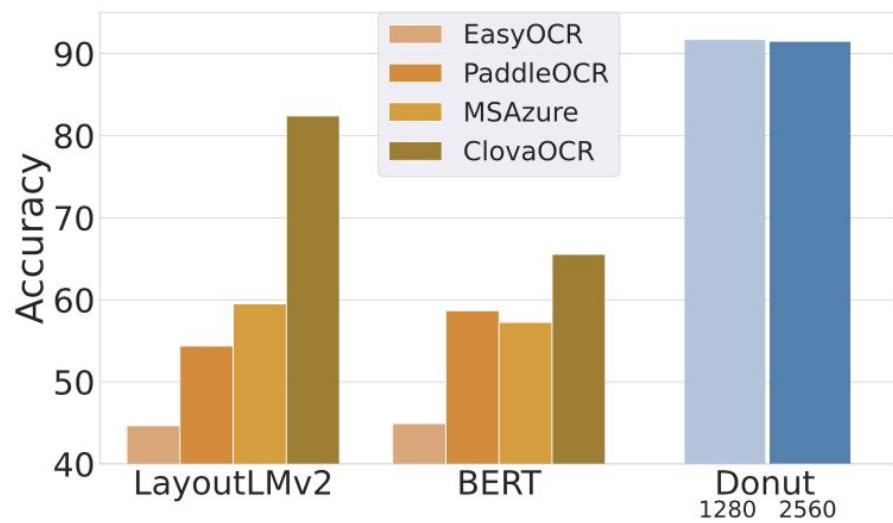
Analysis: Input Resolution



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



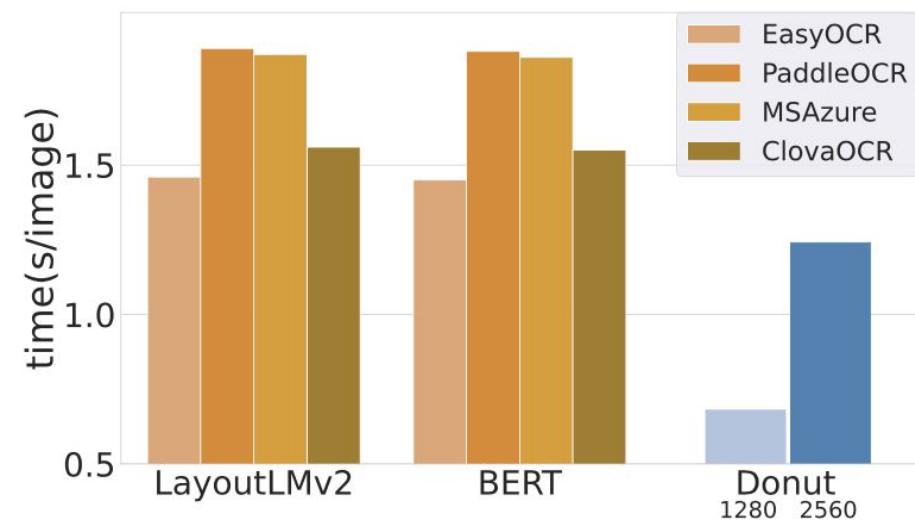
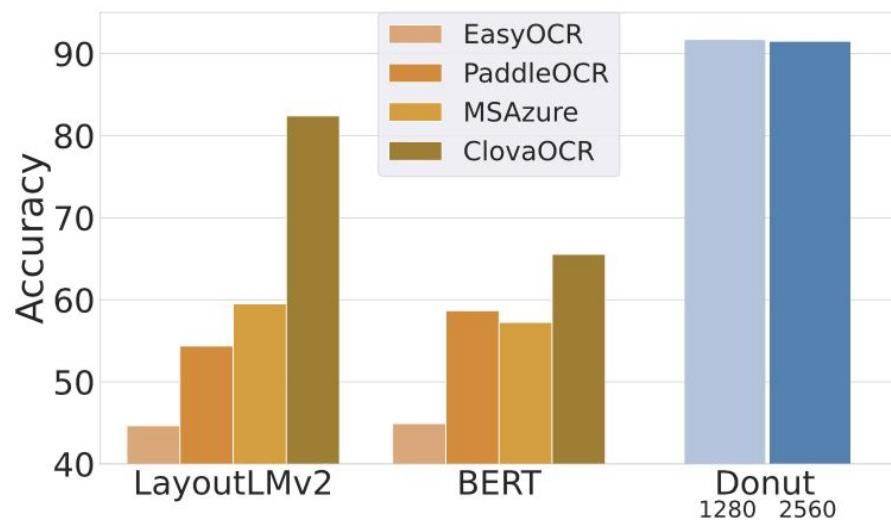
Analysis: OCR Engines



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



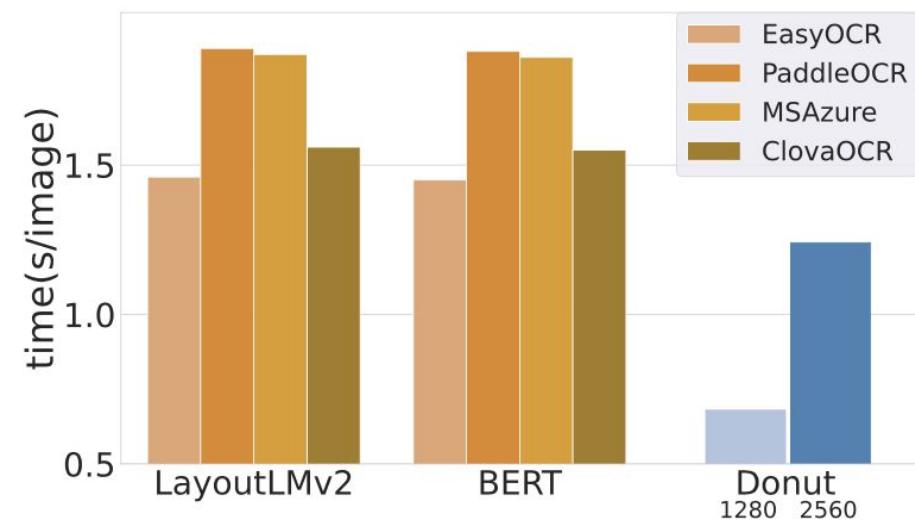
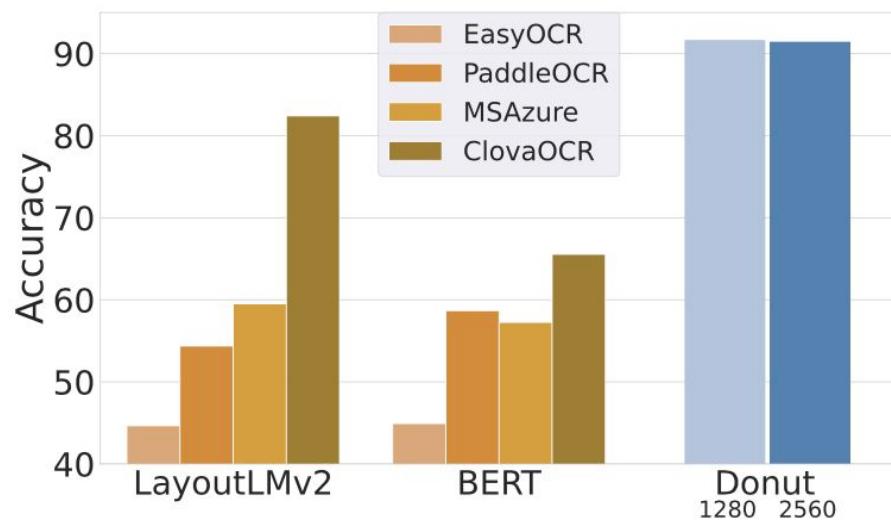
Analysis: OCR Engines



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



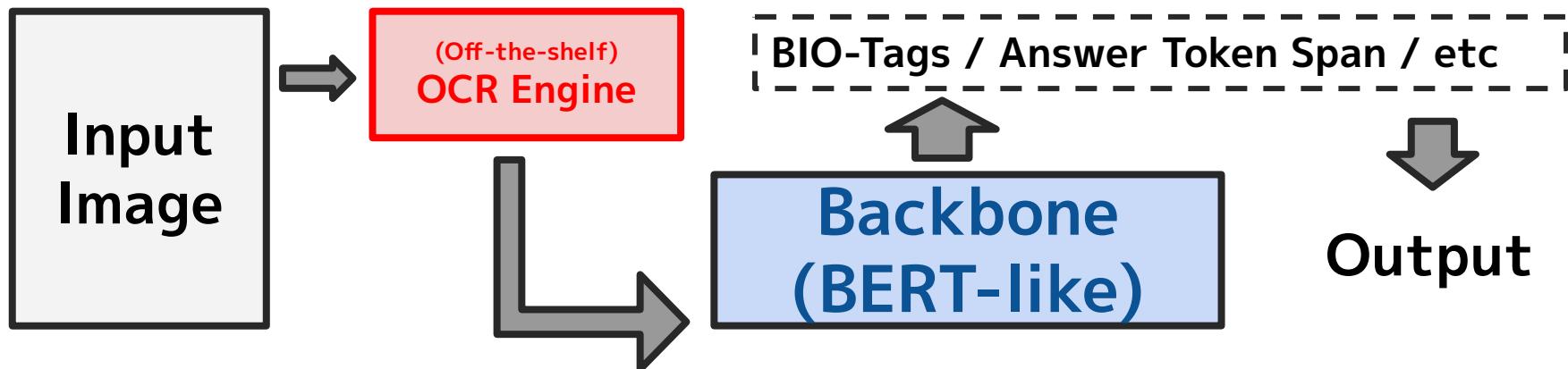
Analysis: OCR Engines



This shows why we need an OCR-free method.



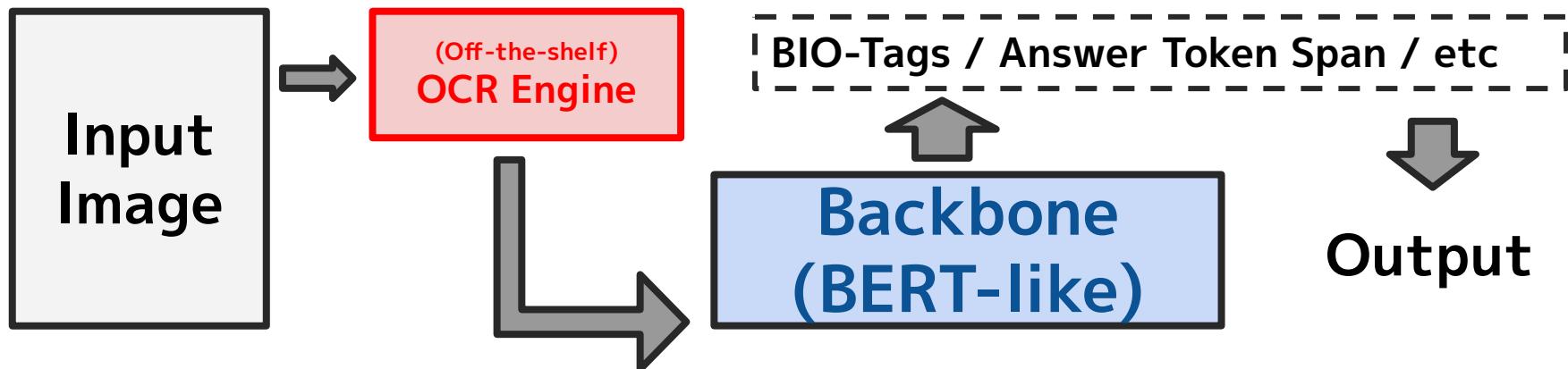
Conclusions



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



Conclusions



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



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.



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.



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.



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.



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.



Thank you!

Contact: gwkim.rsrch@gmail.com

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