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

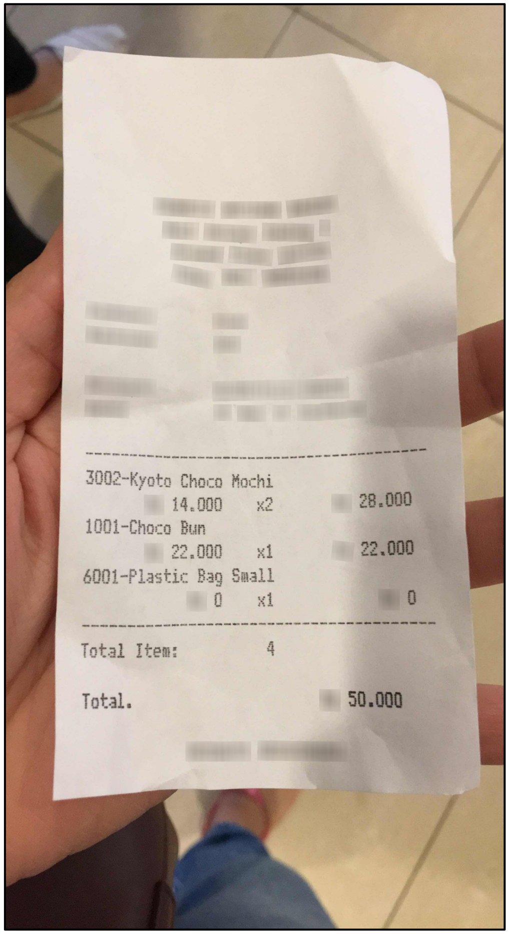
Agenda

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)



[redacted]

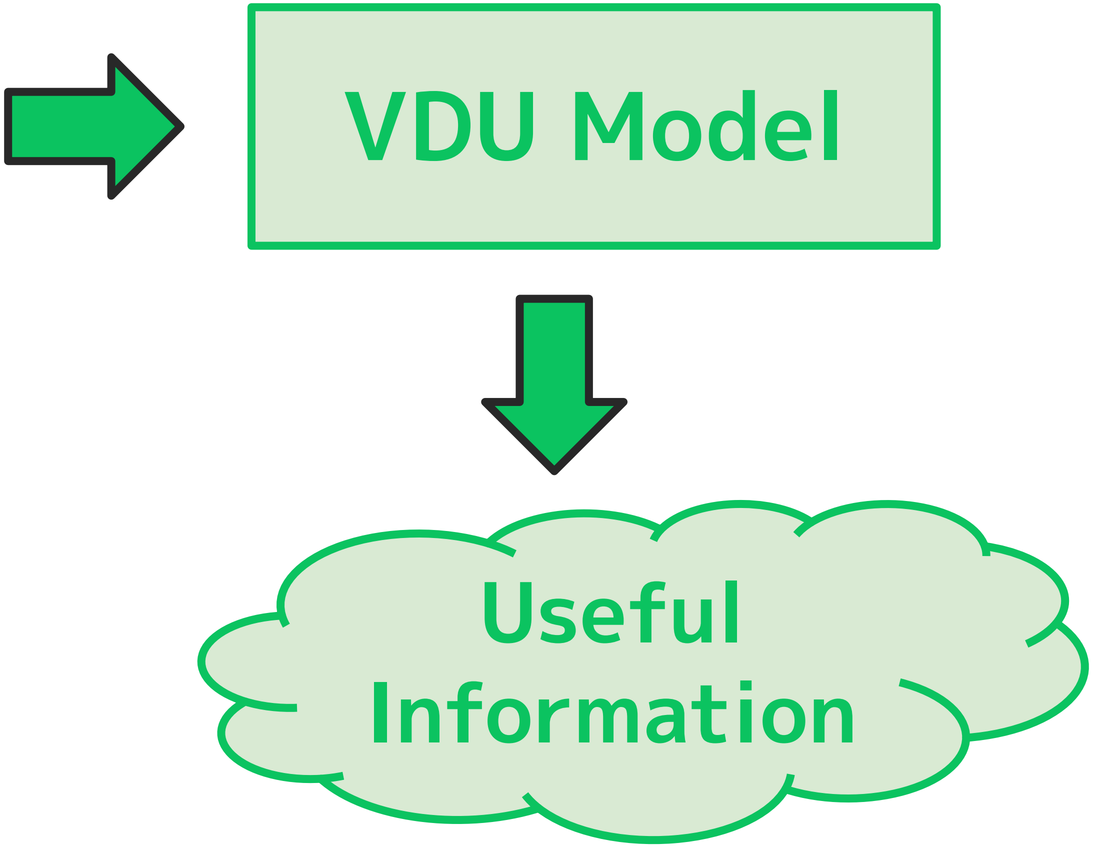
[redacted]

3002-Kyoto Choco Mochi
[redacted] 14.000 x2 [redacted] 28.000
1001-Choco Bun
[redacted] 22.000 x1 [redacted] 22.000
6001-Plastic Bag Small
[redacted] 0 x1 [redacted] 0

Total Item: 4

Total. [redacted] 50.000

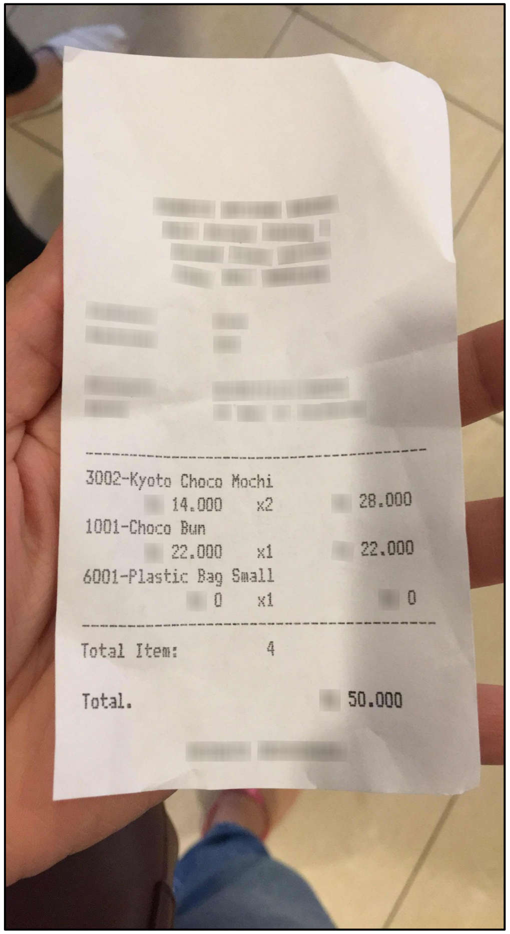
[redacted]



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

Slide

Example 1: Document Classification



[redacted]

3002-Kyoto Choco Mochi
[redacted] 14.000 x2 [redacted] 28.000
1001-Choco Bun
[redacted] 22.000 x1 [redacted] 22.000
6001-Plastic Bag Small
[redacted] 0 x1 [redacted] 0

Total Item: 4

Total. [redacted] 50.000

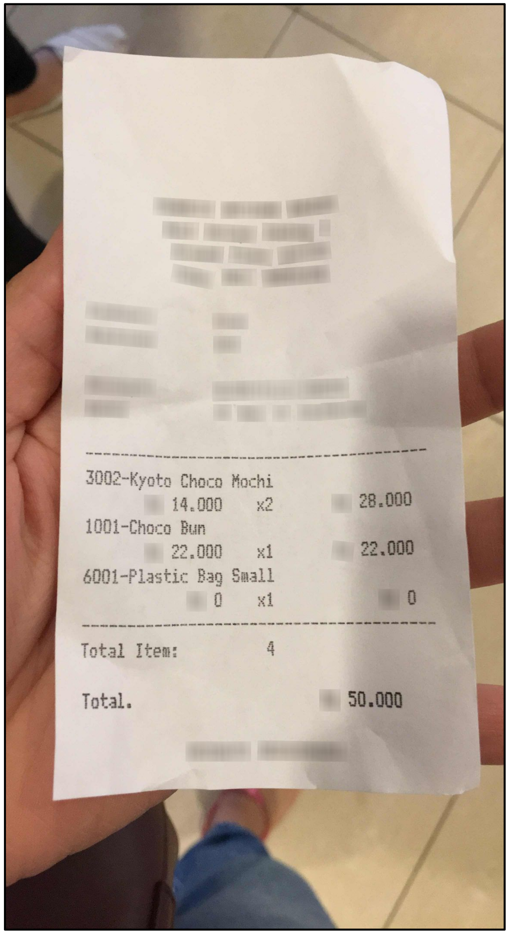
[redacted]

VDU Model

{ "class": "receipt" }

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

Example 2: Document Parsing

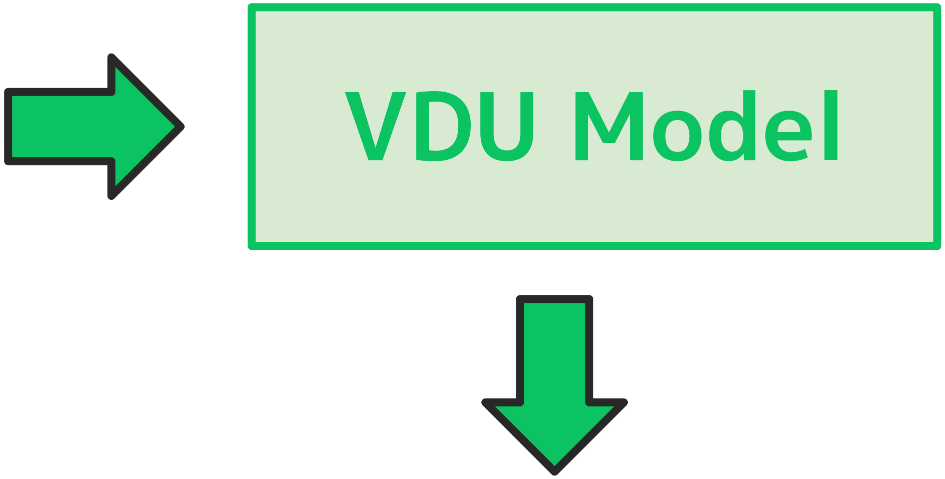


[redacted]

```
3002-Kyoto Choco Mochi
      [redacted] 14.000   x2        [redacted] 28.000
1001-Choco Bun
      [redacted] 22.000   x1        [redacted] 22.000
6001-Plastic Bag Small
             [redacted] 0   x1           [redacted] 0
--------------------------------------------------
Total Item:          4

Total.                        [redacted] 50.000
```

[redacted]



VDU Model

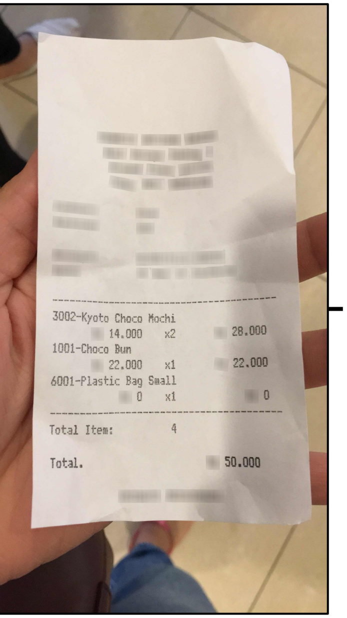
```
{ "menu": [
    {
      "nm": "3002-Kyoto Choco Mochi",
      "unitprice": "14.000",
      "cnt": "x2",
      "price": "28.000"
    }, … }
```

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



[redacted]

```
3002-Kyoto Choco Mochi
     14.000   x2        28.000
1001-Choco Bun
     22.000   x1        22.000
6001-Plastic Bag Small
          0   x1             0

Total Item:        4

Total.                  50.000
```

[redacted]

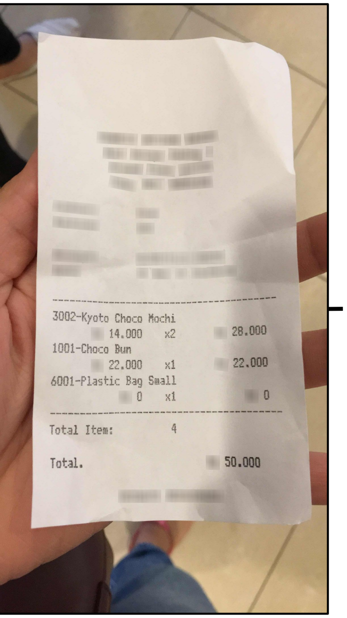
Output

```
{ "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



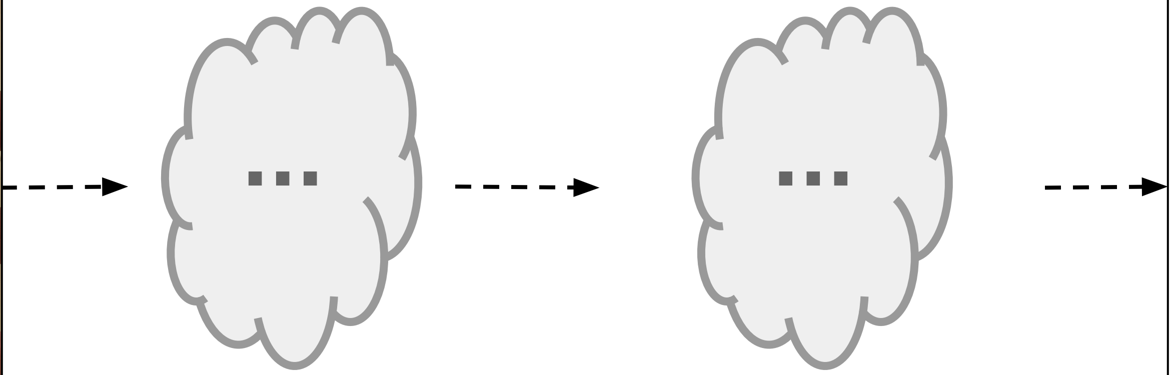
[redacted]

3002-Kyoto Choco Mochi
[redacted] 14.000 x2 [redacted] 28.000
1001-Choco Bun
[redacted] 22.000 x1 [redacted] 22.000
6001-Plastic Bag Small
[redacted] 0 x1 [redacted] 0

Total Item: 4

Total. [redacted] 50.000

[redacted]

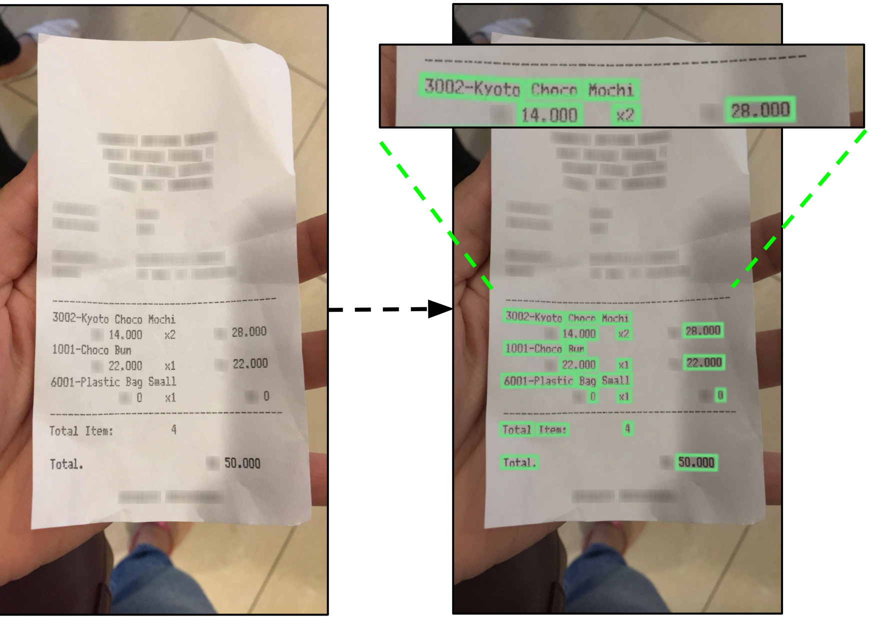


Output

```
{ "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.

Conventional VDU Model



[redacted]

3002-Kyoto Choco Mochi
14.000 x2 28.000
1001-Choco Bun
22.000 x1 22.000
6001-Plastic Bag Small
0 x1 0

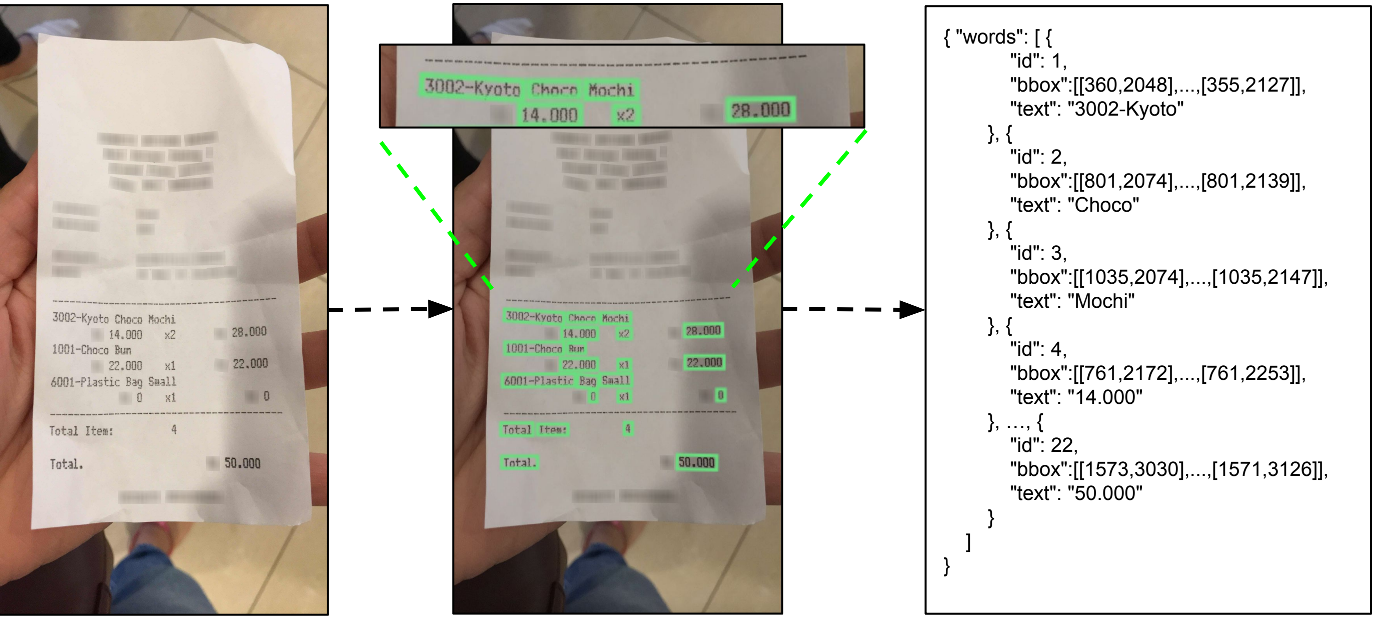
Total Item: 4

Total. 50.000

Detection!

First, a text detector finds all text boxes.

Conventional VDU Model



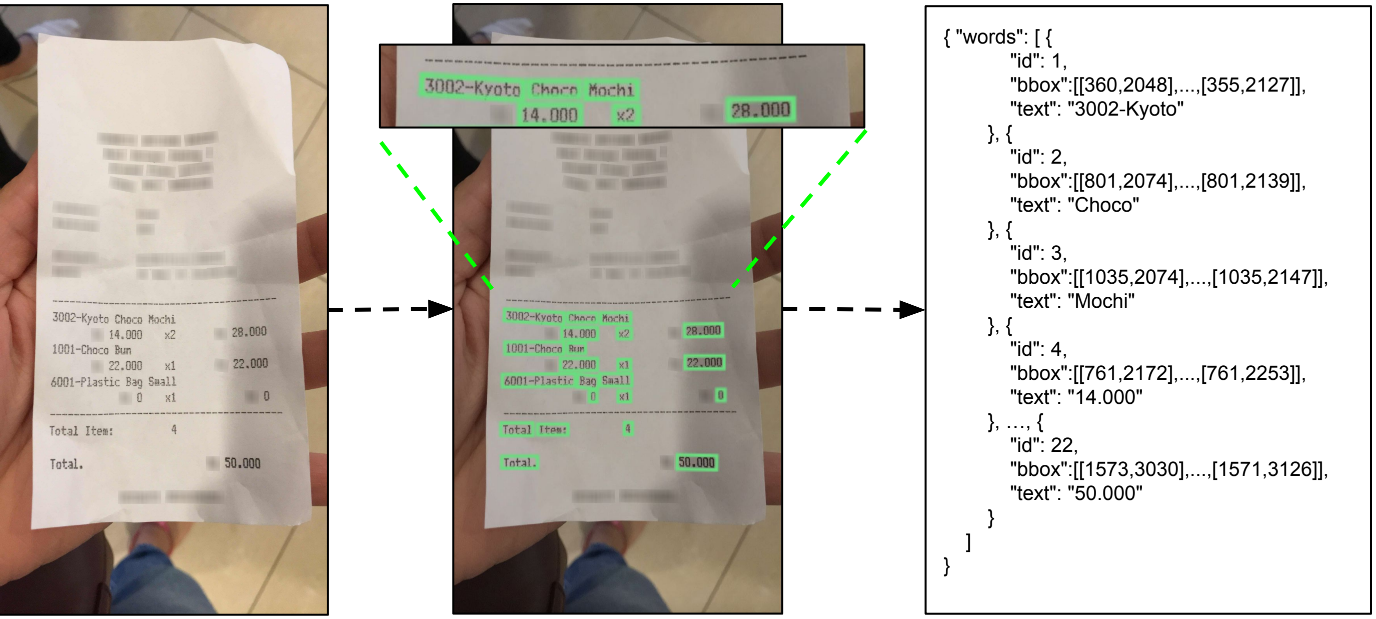
```
{ "words": [ {
          "id": 1,
          "bbox":[[360,2048],...,[355,2127]],
          "text": "3002-Kyoto"
     }, {
          "id": 2,
          "bbox":[[801,2074],...,[801,2139]],
          "text": "Choco"
     }, {
          "id": 3,
          "bbox":[[1035,2074],...,[1035,2147]],
          "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.

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"
      }
   ]
}
```

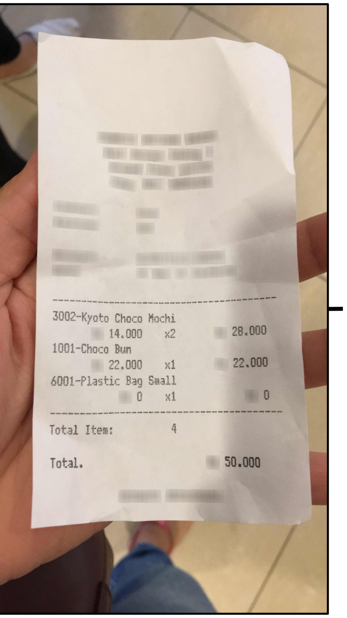
Detection! Recognition!

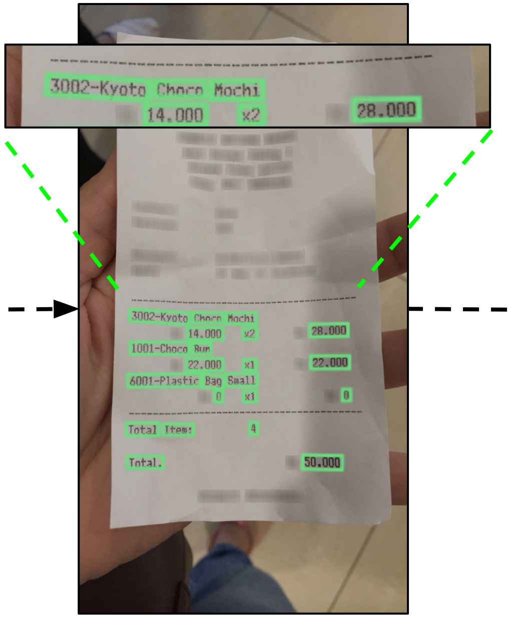
OCR

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
    }
  ]
}
```

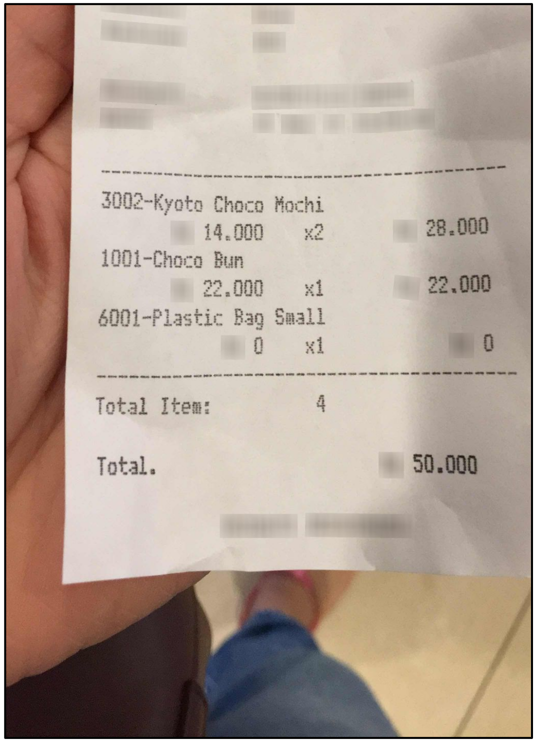
Detection! Recognition! Parsing!

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



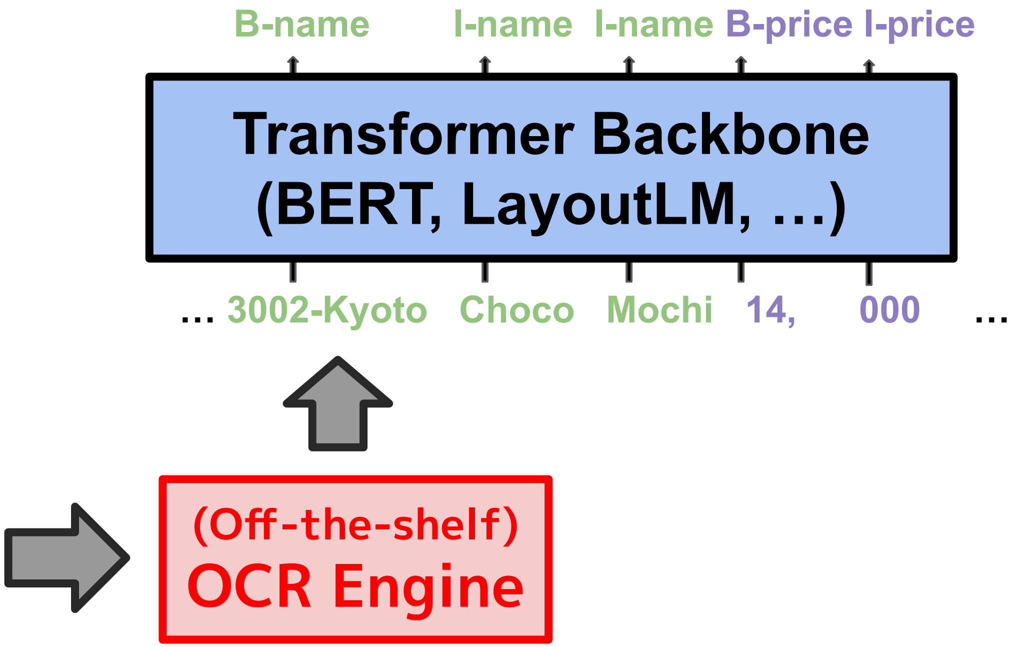
[redacted]

3002-Kyoto Choco Mochi
[redacted] 14.000 x2 [redacted] 28.000
1001-Choco Bun
[redacted] 22.000 x1 [redacted] 22.000
6001-Plastic Bag Small
[redacted] 0 x1 [redacted] 0

Total Item: 4

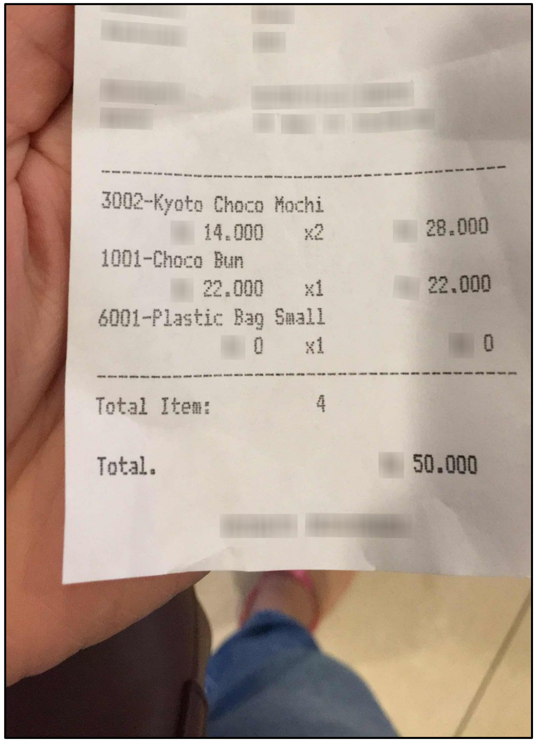
Total. [redacted] 50.000

[redacted]



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

Conventional VDU Model: Details of the Parsing Stage



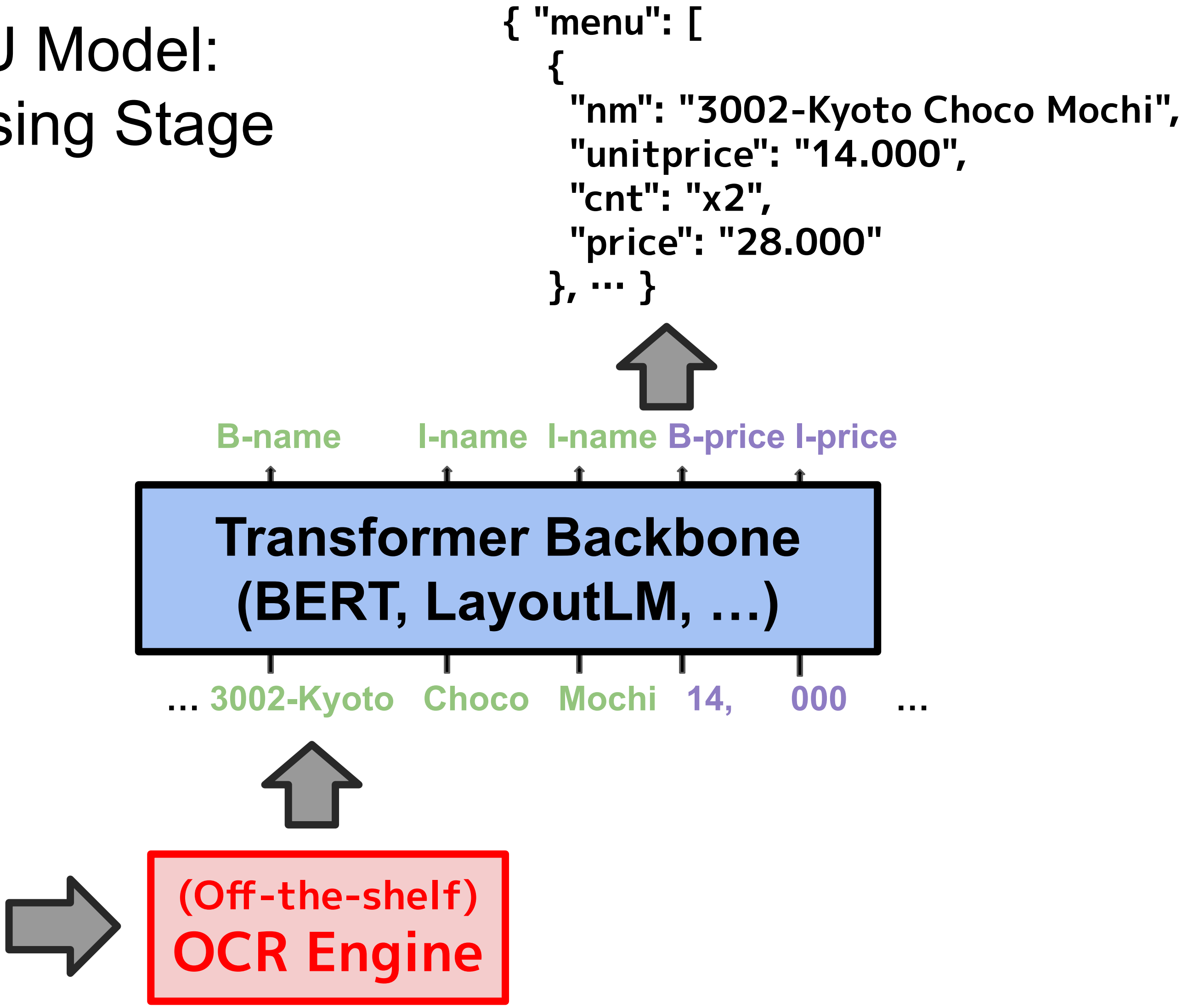
[redacted]

```
3002-Kyoto Choco Mochi
      14.000    x2          28.000
1001-Choco Bun
      22.000    x1          22.000
6001-Plastic Bag Small
           0    x1               0

Total Item:         4

Total.                      50.000
```

[redacted]



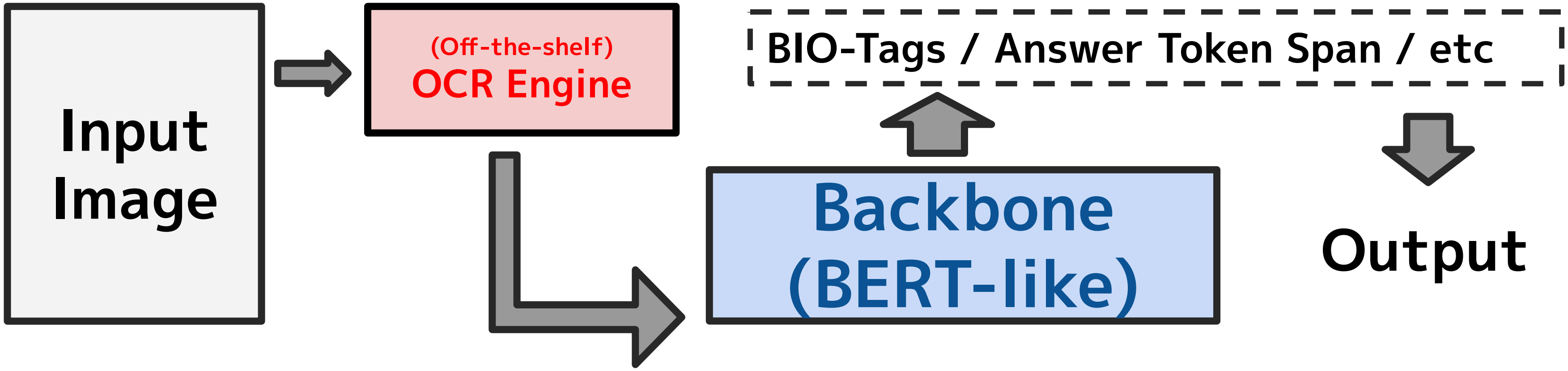
```
{ "menu": [
   {
     "nm": "3002-Kyoto Choco Mochi",
     "unitprice": "14.000",
     "cnt": "x2",
     "price": "28.000"
   }, … }
```

B-name
I-name
I-name
B-price
I-price
Transformer Backbone (BERT, LayoutLM, …)
… 3002-Kyoto
Choco
Mochi
14,
000
…
(Off-the-shelf) OCR Engine

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

Slide

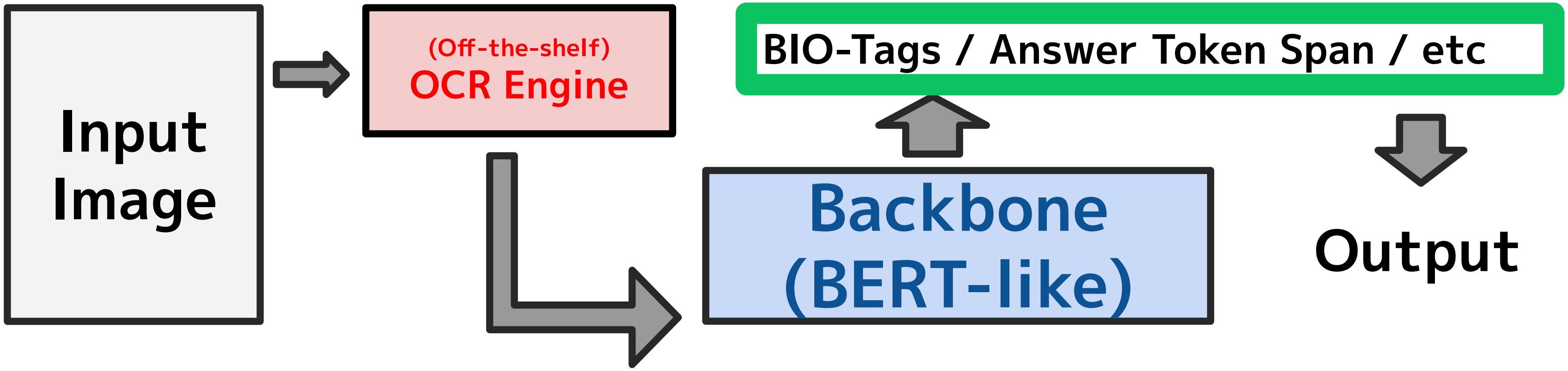
Conventional VDU Model: Overview



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

Slide

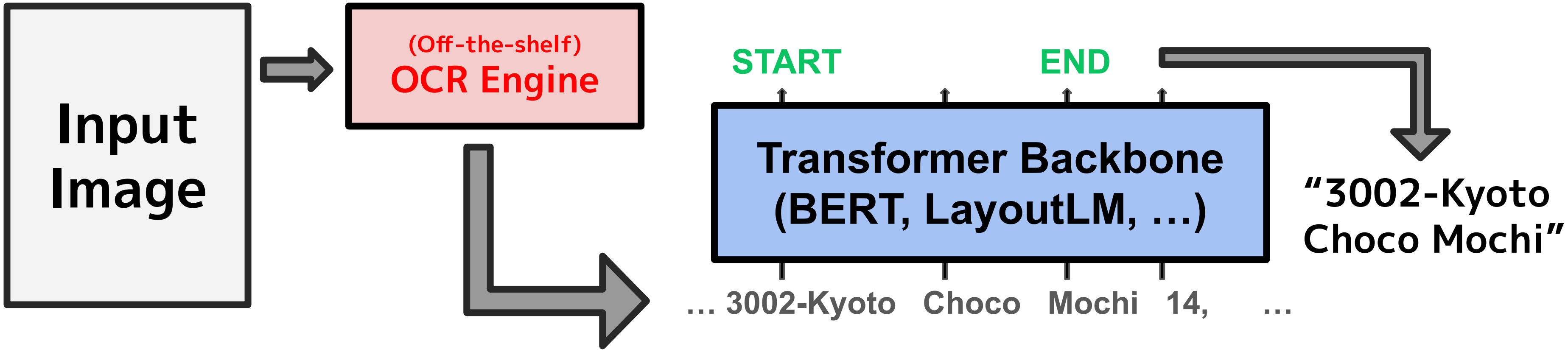
Conventional VDU Model: Overview



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

Slide

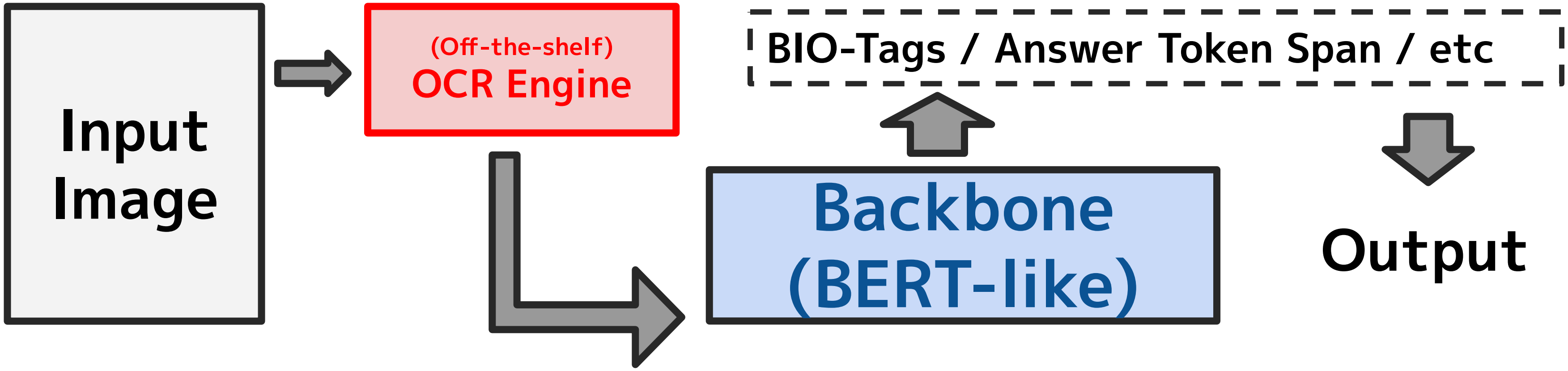
Conventional VDU Model: Overview



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

Slide

Conventional VDU Model: Overview



Input Image

(Off-the-shelf) OCR Engine

Backbone (BERT-like)

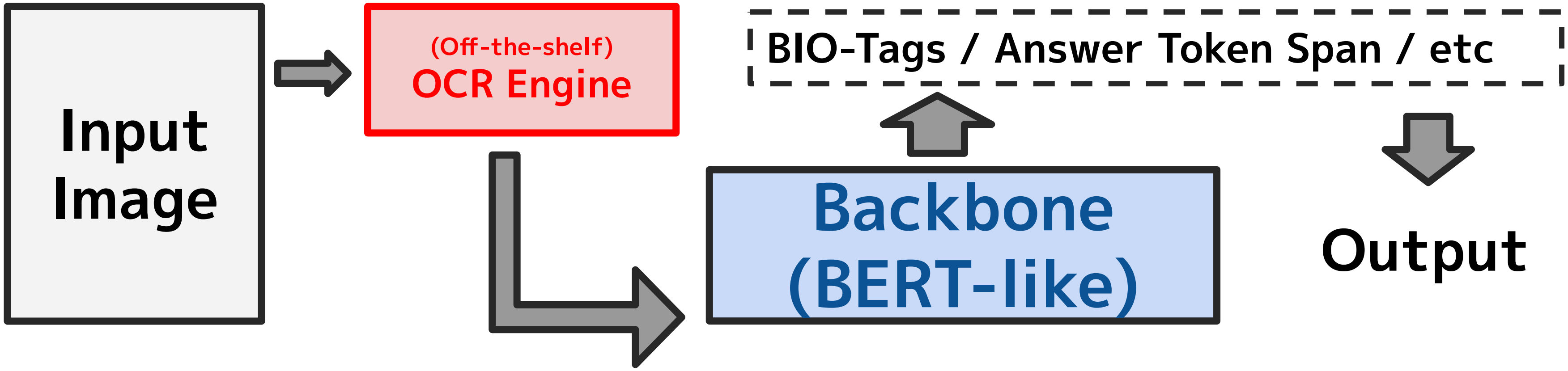
BIO-Tags / Answer Token Span / etc

Output

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

Slide

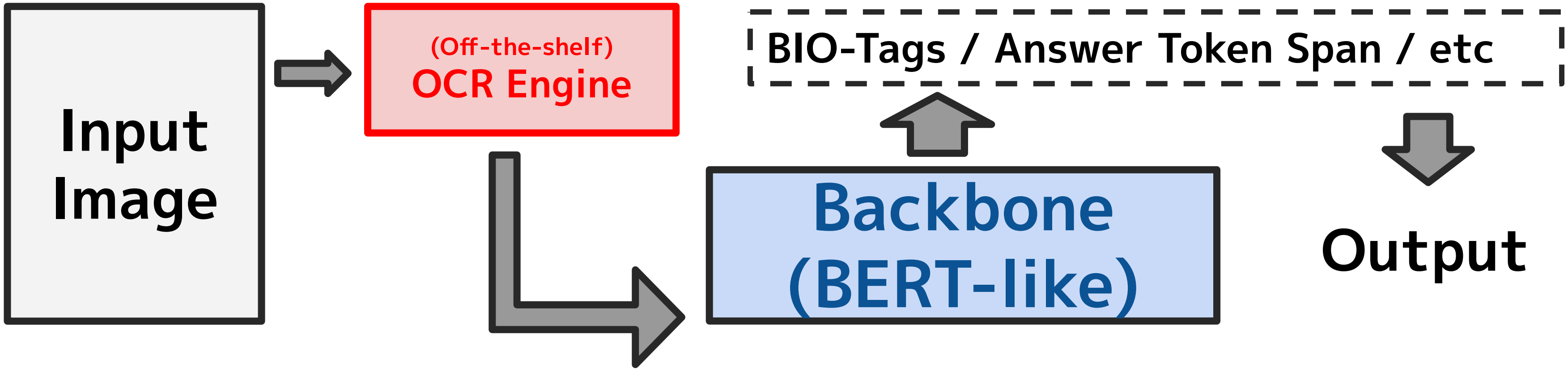
Conventional VDU Model: Overview



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

Slide

Conventional VDU Model: Overview

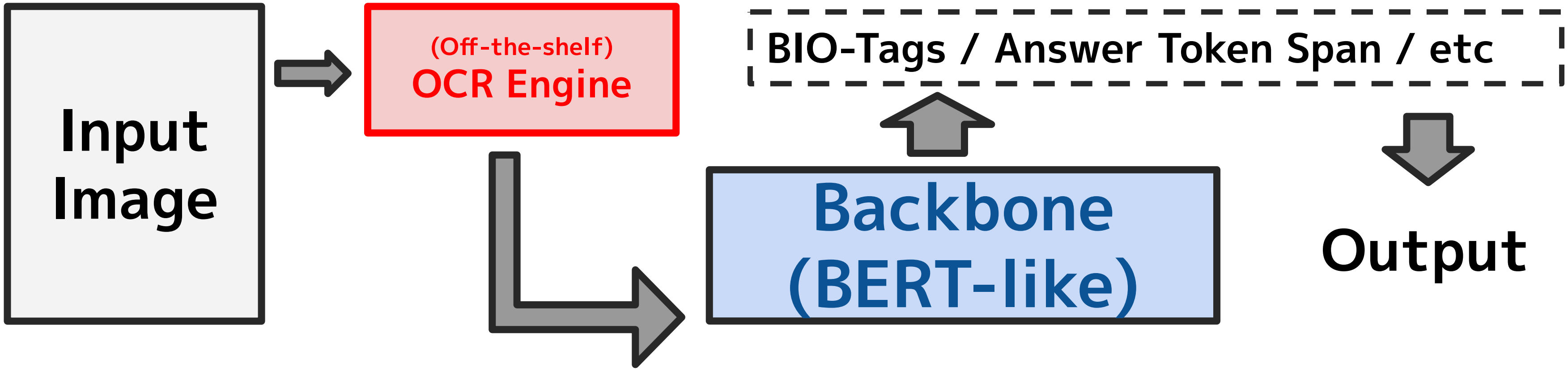


- high computational costs

First, OCR increase computational costs.

Slide

Conventional VDU Model: Overview

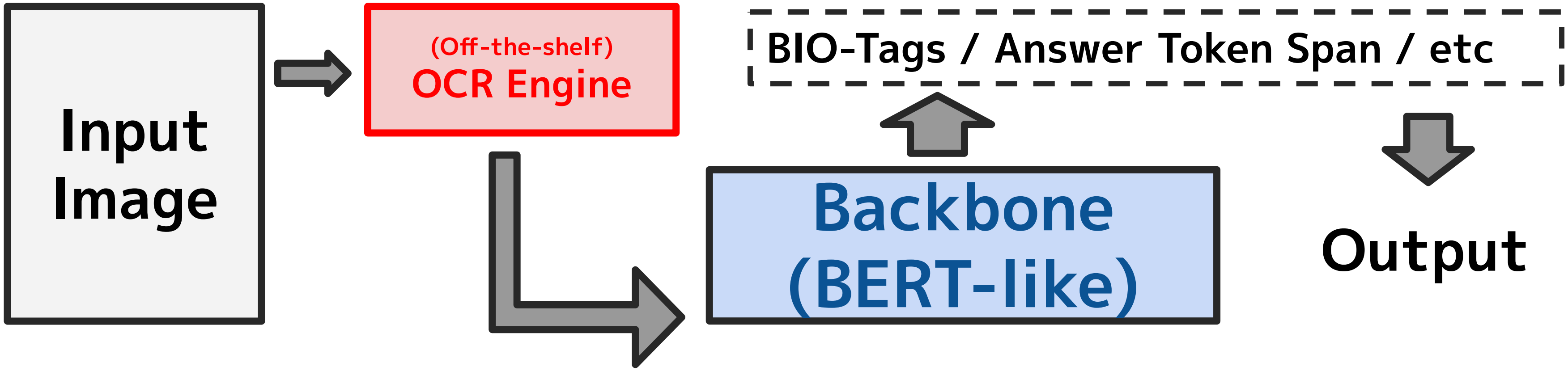


- high computational costs
- inflexibility of OCR on languages or document type

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

Slide

Conventional VDU Model: Overview

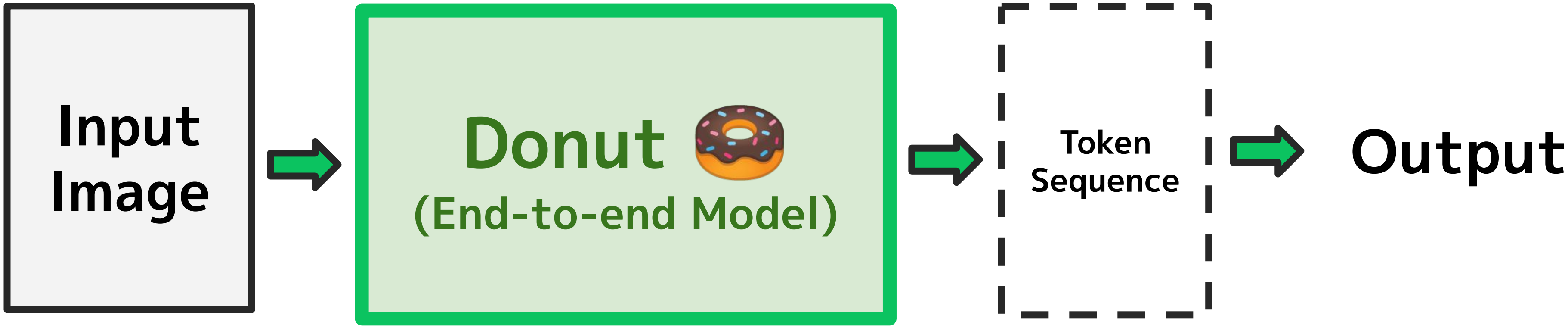


- high computational costs
- inflexibility of OCR on languages or document type
- OCR error propagation

Lastly, OCR errors are propagate to the subsequent process.

Slide

Proposal: OCR-free Approach



Input Image

Donut 🍩

(End-to-end Model)

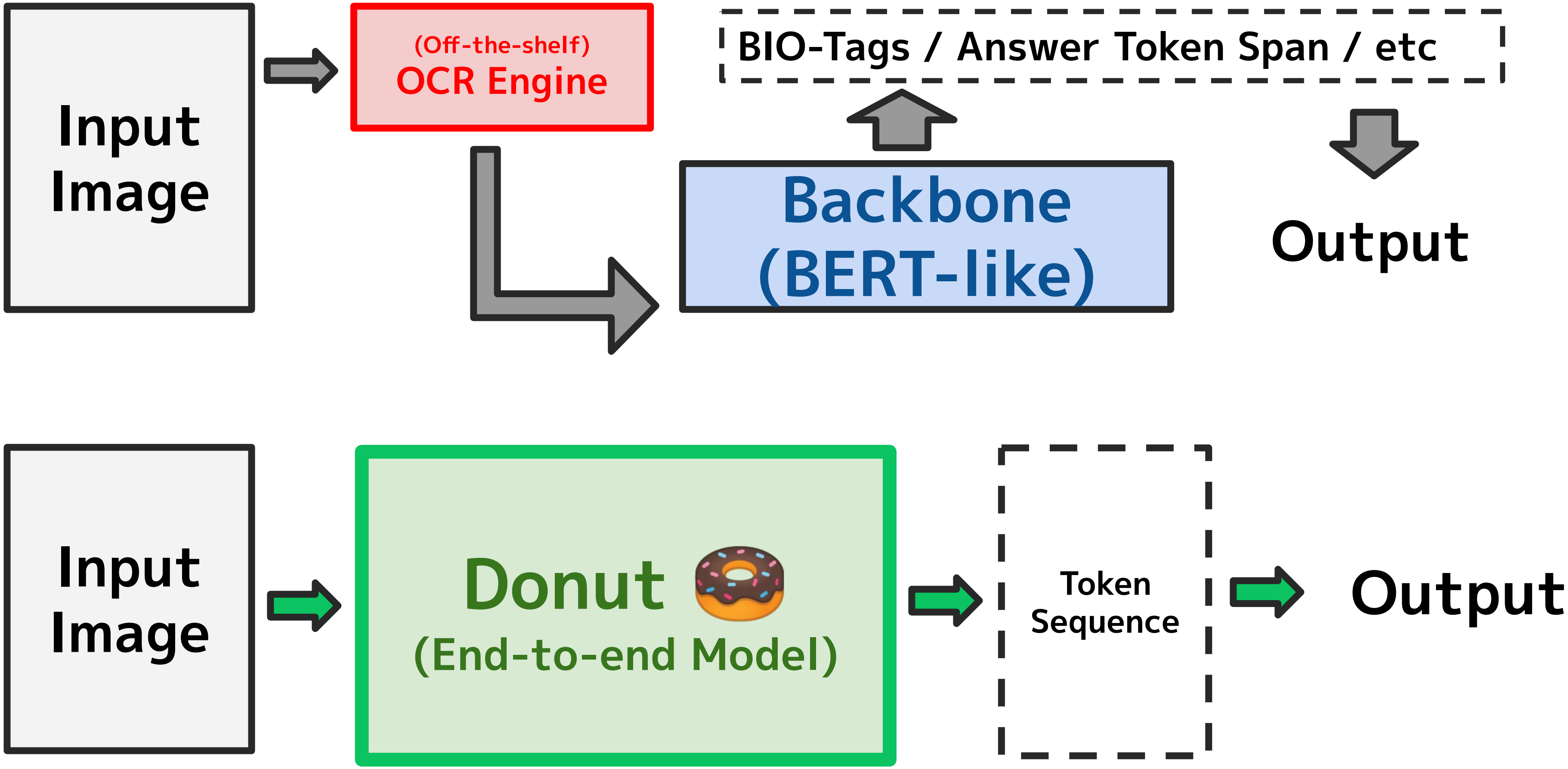
Token Sequence

Output

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

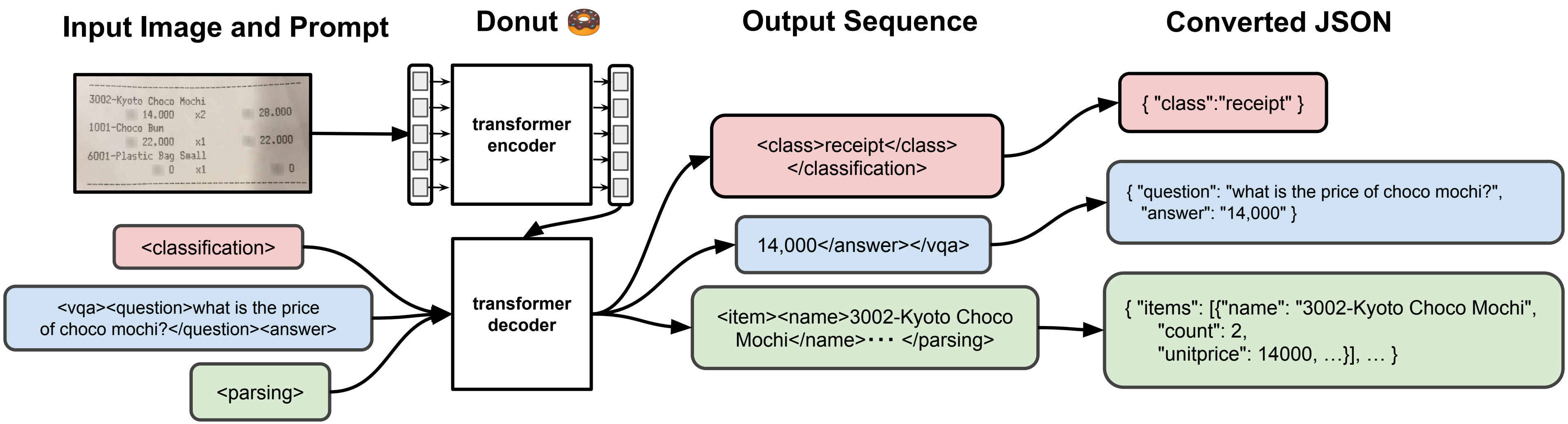
Slide

AS-IS v.s. TO-BE



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

Overview



Input Image and Prompt

Donut 🍩

Output Sequence

Converted JSON

transformer encoder

transformer decoder

<classification>

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

<parsing>

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

14,000</answer></vqa>

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

{ "class":"receipt" }

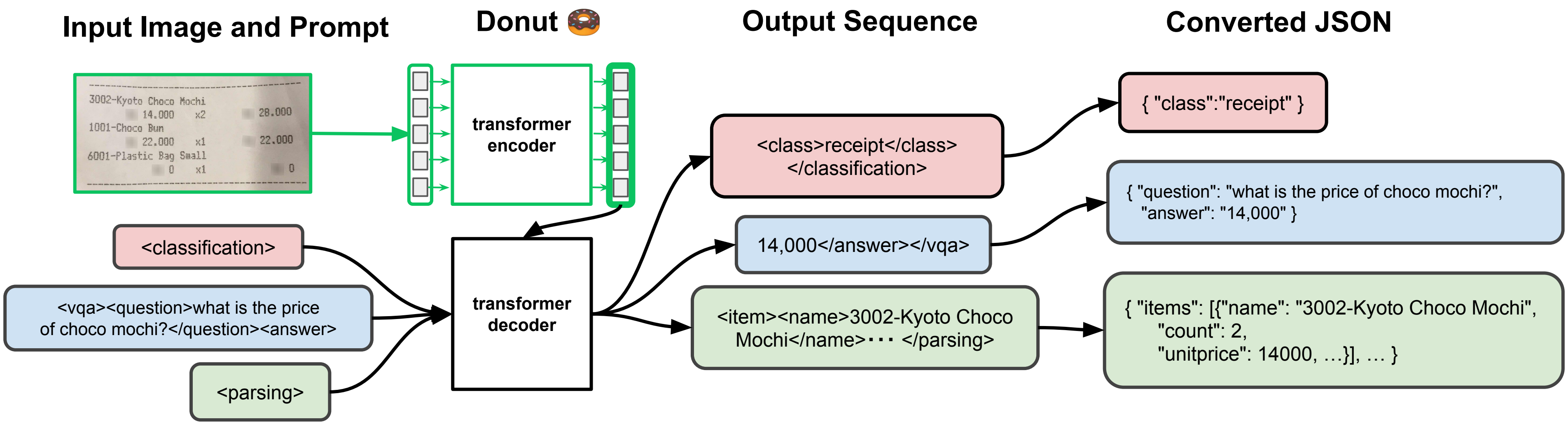
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

This is the overview of Donut.

Slide

Overview



Input Image and Prompt

3002-Kyoto Choco Mochi
14.000 x2 28.000
1001-Choco Bun
22.000 x1 22.000
6001-Plastic Bag Small
0 x1 0

<classification>

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

<parsing>

Donut 🍩

transformer encoder

transformer decoder

Output Sequence

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

14,000</answer></vqa>

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

Converted JSON

{ "class":"receipt" }

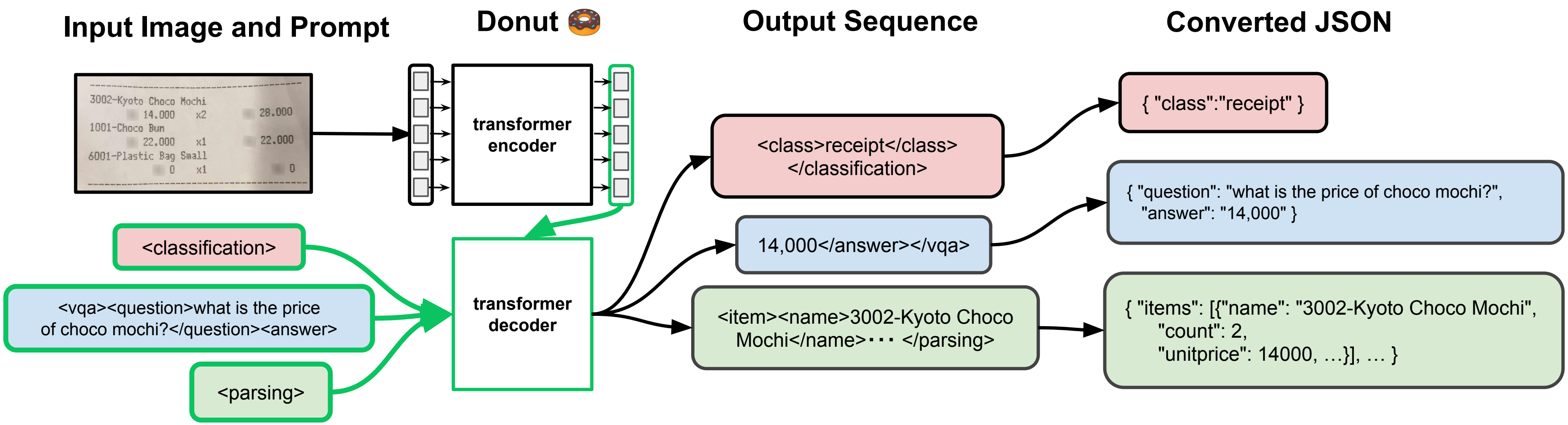
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

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

Slide

Overview



Input Image and Prompt

<classification>

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

<parsing>

Donut 🍩

transformer encoder

transformer decoder

Output Sequence

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

14,000</answer></vqa>

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

Converted JSON

{ "class":"receipt" }

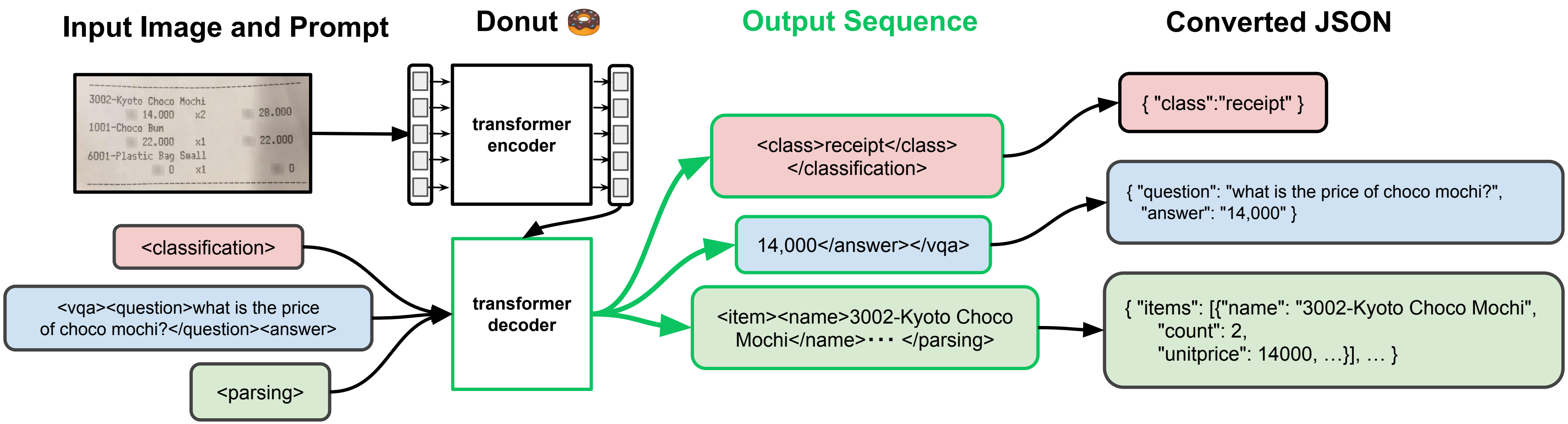
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

The textual decoder processes the image embeddings and prompt tokens.

Slide

Overview



Input Image and Prompt

```
3002-Kyoto Choco Mochi
     14.000    x2        28.000
1001-Choco Bun
     22.000    x1        22.000
6001-Plastic Bag Small
          0    x1             0
```

<classification>

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

<parsing>

Donut 🍩

transformer encoder

transformer decoder

Output Sequence

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

14,000</answer></vqa>

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

Converted JSON

{ "class":"receipt" }

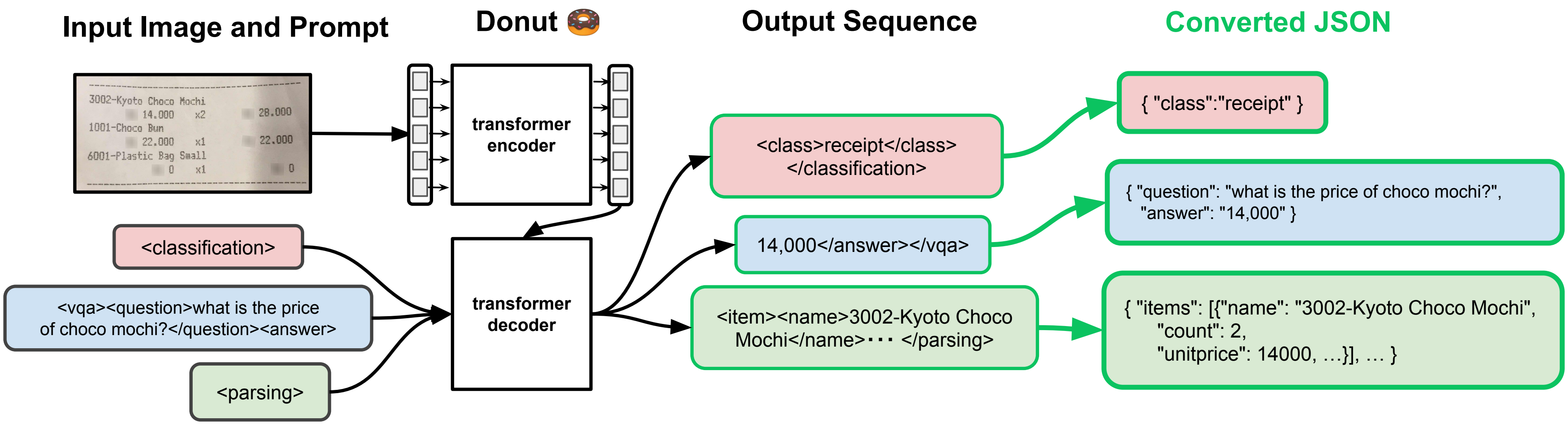
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

Then, the decoder outputs token sequences,

Slide

Overview



Input Image and Prompt

<classification>

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

<parsing>

Donut 🍩

transformer encoder

transformer decoder

Output Sequence

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

14,000</answer></vqa>

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

Converted JSON

{ "class":"receipt" }

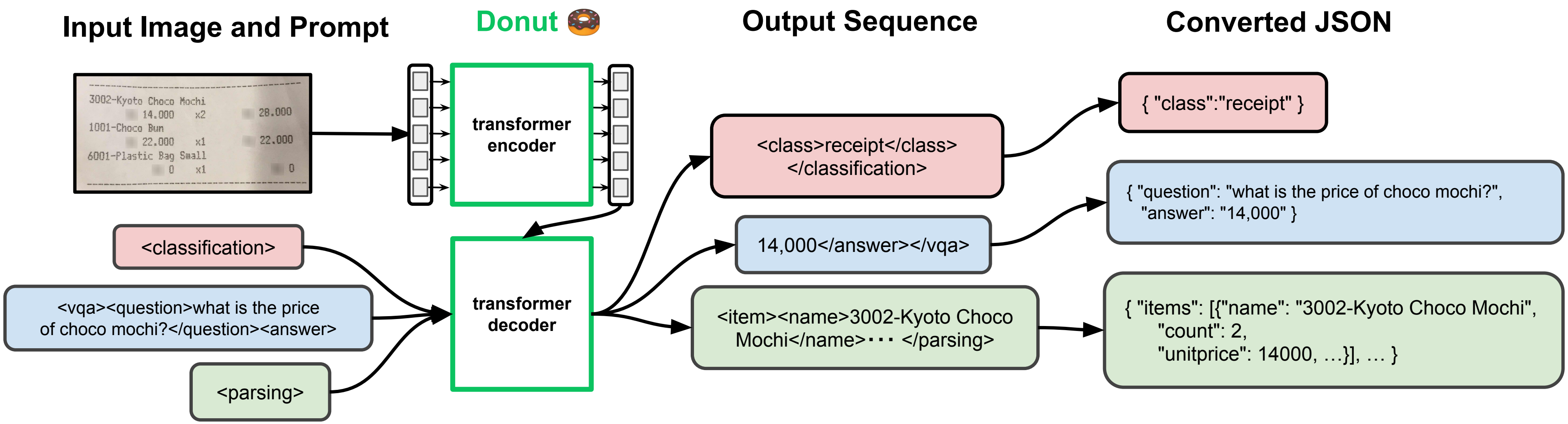
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

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

Slide

Overview: Model Architecture



Input Image and Prompt

<classification>

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

<parsing>

Donut 🍩

transformer encoder

transformer decoder

Output Sequence

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

14,000</answer></vqa>

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

Converted JSON

{ "class":"receipt" }

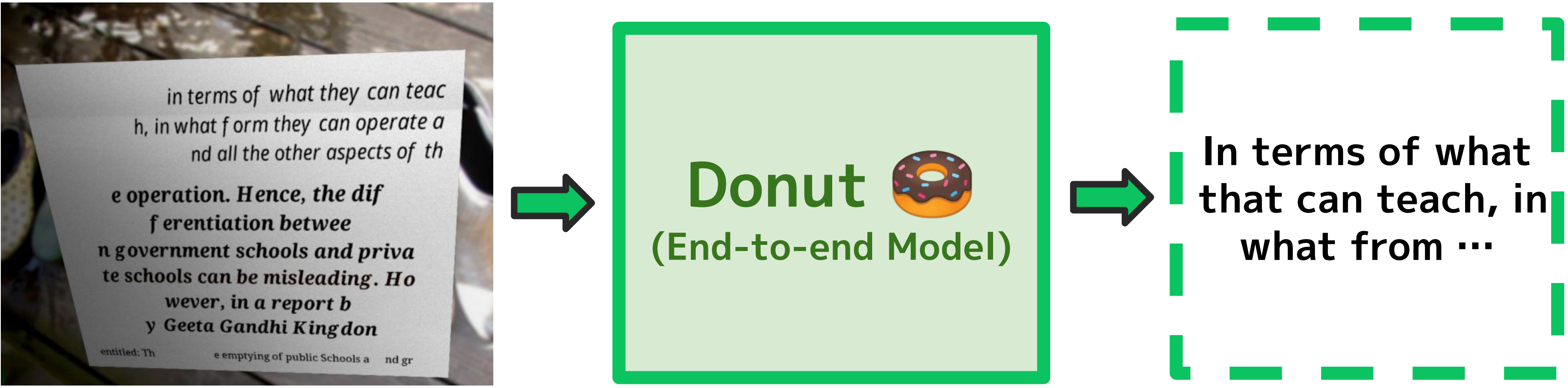
{ "question": "what is the price of choco mochi?", "answer": "14,000" }

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

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



in terms of what they can teac
h, in what f orm they can operate a
nd all the other aspects of th
e operation. Hence, the dif
f erentiation betwee
n government schools and priva
te schools can be misleading. Ho
wever, in a report b
y Geeta Gandhi Kingdon
entitled: Th e emptying of public Schools a nd gr

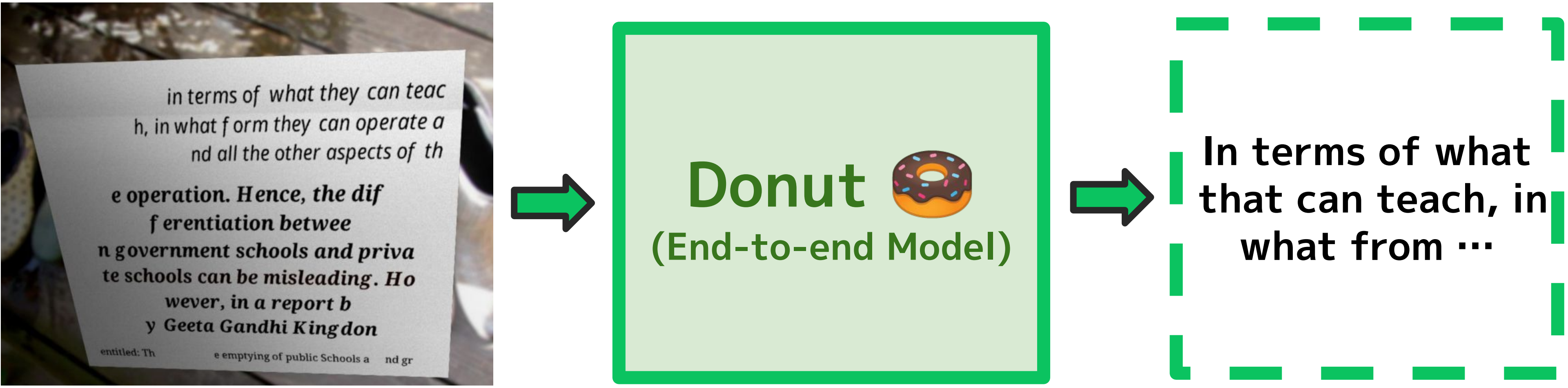
Donut 🍩
(End-to-end Model)

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

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

Slide

Pre-training Task



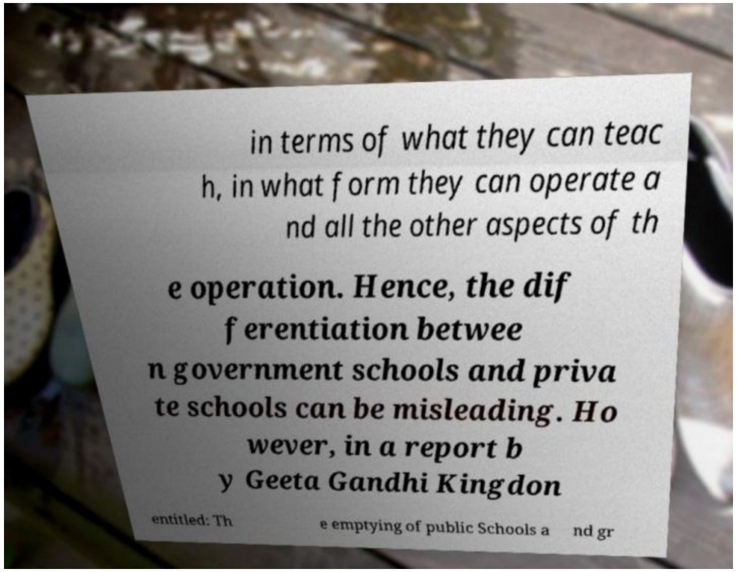
in terms of what they can teac
h, in what form they can operate a
nd all the other aspects of th
e operation. Hence, the dif
ferentiation betwee
n government schools and priva
te schools can be misleading. Ho
wever, in a report b
y Geeta Gandhi Kingdon
entitled: Th e emptying of public Schools a nd gr

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

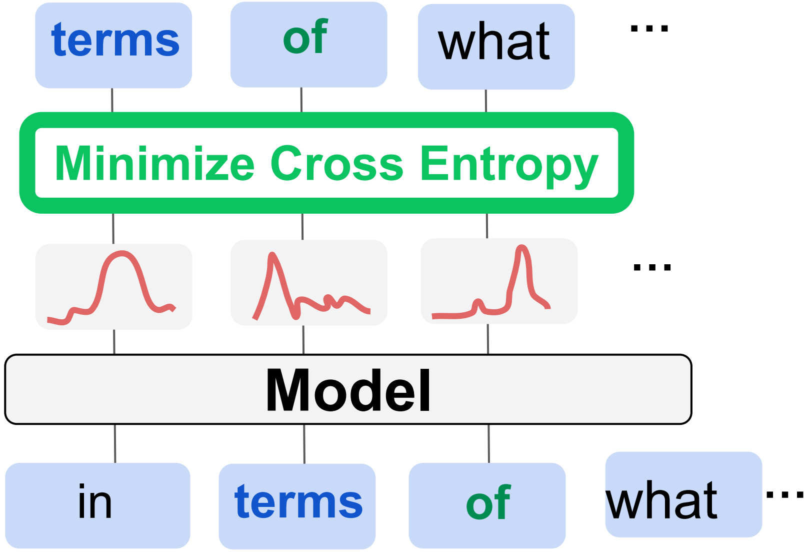
The objective is to read all texts from the top-left to bottom-right.
This task can be interpret as a pseudo OCR task.

Slide

Training Strategy: Teacher-forcing Scheme



in terms of what they can teac
h, in what f orm they can operate a
nd all the other aspects of th
e operation. Hence, the dif
f erentiation betwee
n government schools and priva
te schools can be misleading. Ho
wever, in a report b
y Geeta Gandhi Kingdon
entitled: Th e emptying of public Schools a nd gr

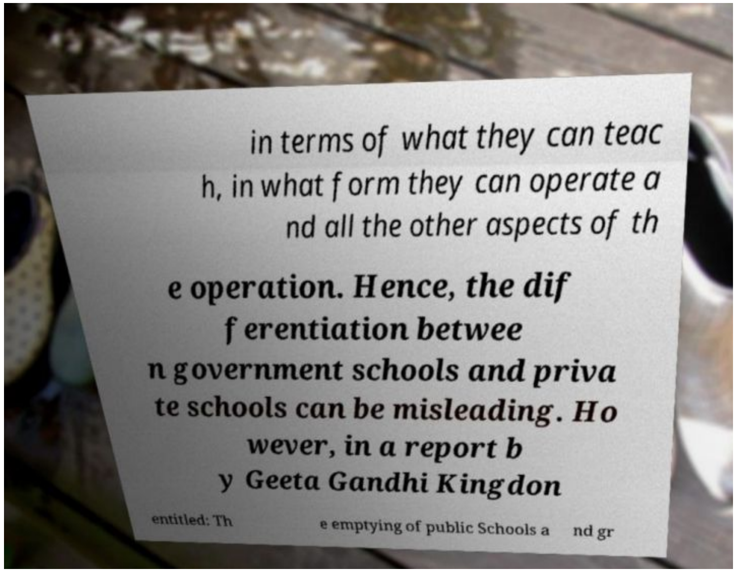


This can be interpret as
a token classification at each step.

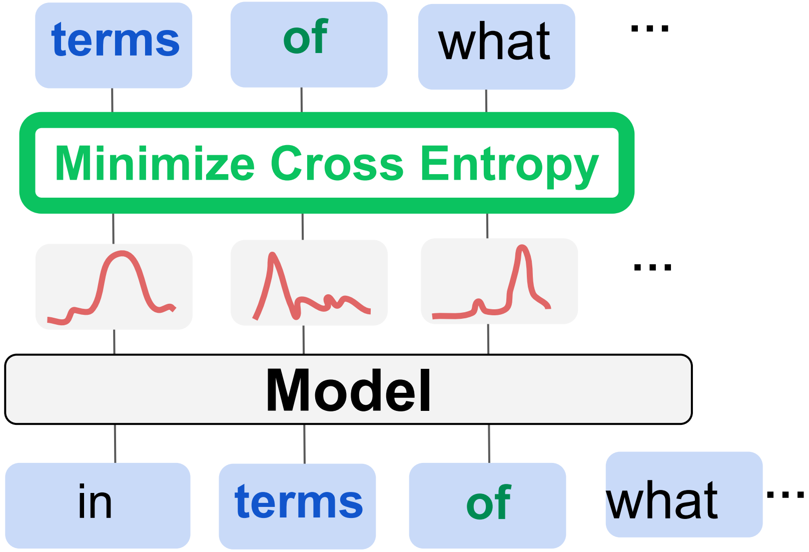
Following the original Transformer, the model training is done with Teacher-forcing scheme. More details can also be found in the manuscript.

Slide

Training Strategy: Teacher-forcing Scheme



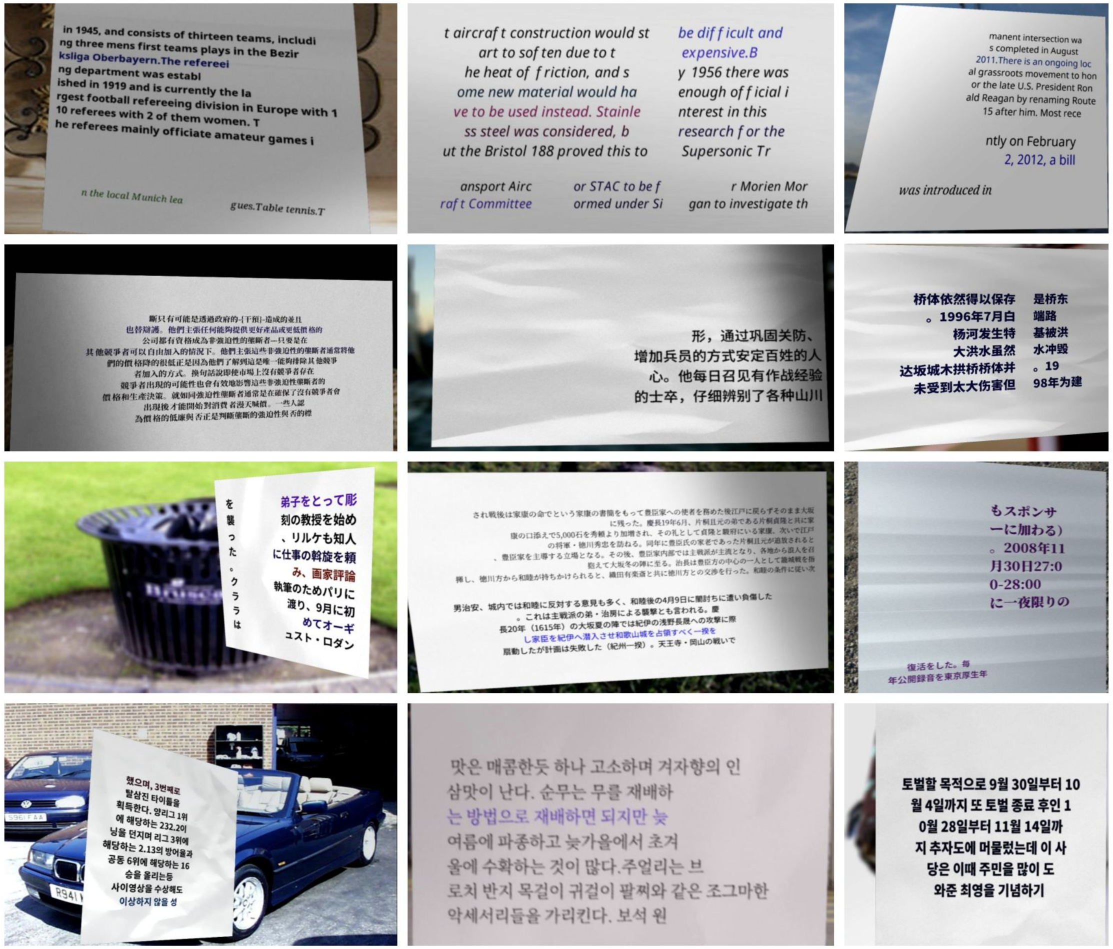
in terms of what they can teac
h, in what f orm they can operate a
nd all the other aspects of th
e operation. Hence, the dif
f erentiation betwee
n government schools and priva
te schools can be misleading. Ho
wever, in a report b
y Geeta Gandhi Kingdon
entitled: Th e emptying of public Schools a nd gr



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.

SynthDoG 🐶:Synthetic Document Generator



in 1945, and consists of thirteen teams, includi
ng three mens first teams plays in the Bezir
ksliga Oberbayern.The refereei
ng department was establ
ished in 1919 and is currently the la
rgest football refereeing division in Europe with 1
10 referees with 2 of them women. T
he referees mainly officiate amateur games i

n the local Munich lea
gues.Table tennis.T

t aircraft construction would st
art to soften due to t
he heat of friction, and s
ome new material would ha
ve to be used instead. Stainle
ss steel was considered, b
ut the Bristol 188 proved this to

be difficult and
expensive.B
y 1956 there was
enough official i
nterest in this
research for the
Supersonic Tr

ansport Airc
raft Committee

or STAC to be f
ormed under Si

r Morien Mor
gan to investigate th

manent intersection wa
s completed in August
2011.There is an ongoing loc
al grassroots movement to hon
or the late U.S. President Ron
ald Reagan by renaming Route
15 after him. Most rece

ntly on February
2, 2012, a bill

was introduced in

斷只有可能是透過政府的-{干預}-造成的並且
也替辯護。他們主張任何能夠提供更好產品或更低價格的
公司都有資格成為非強迫性的壟斷者—只要是在
其他競爭者可以自由加入的情況下。他們主張這些非強迫性的壟斷者通常將他
們的價格降的很低正是因為他們了解到這是唯一能夠排除其他競爭
者加入的方式。換句話說即使市場上沒有競爭者存在
競爭者出現的可能性也會有效地影響這些非強迫性壟斷者的
價格和生產決策。就如同強迫性壟斷者通常是在確保了沒有競爭者會
出現後才能開始對消費者漫天喊價。一些人認
為價格的低廉與否正是判斷壟斷的強迫性與否的標

形，通过巩固关防、
增加兵员的方式安定百姓的人
心。他每日召见有作战经验
的士卒，仔细辨别了各种山川

桥体依然得以保存
。1996年7月白
杨河发生特
大洪水虽然
达坂城木拱桥桥体并
未受到太大伤害但

是桥东
端路
基被洪
水冲毁
。19
98年为建

弟子をとって彫
刻の教授を始め
、リルケも知人
に仕事の斡旋を頼
み、画家評論
執筆のためパリに
渡り、9月に初
めてオーギ
ユスト・ロダン

を襲った。クララは

され戦後は家康の命でという家康の書簡をもって豊臣家への使者を務めた後江戸に戻らずそのまま大坂
に残った。慶長19年6月、片桐且元の弟である片桐貞隆と共に家
康の口添えで5,000石を秀頼より加増され、その礼として貞隆と駿府にいる家康、次いで江戸
の将軍・徳川秀忠を訪ねる。同年に豊臣氏の家老であった片桐且元が追放されると
、豊臣家を主導する立場となる。その後、豊臣家内部では主戦派が主流となり、各地から浪人を召
抱えて大坂冬の陣に至る。治長は豊臣方の中心の一人として籠城戦を指
揮し、徳川方から和睦が持ちかけられると、織田有楽斎と共に徳川方との交渉を行った。和睦の条件に従い次

男治安、城内では和睦に反対する意見も多く、和睦後の4月9日に闇討ちに遭い負傷した
。これは主戦派の弟・治房による襲撃とも言われる。慶
長20年（1615年）の大坂夏の陣では紀伊の浅野長晟への攻撃に際
し家臣を紀伊へ潜入させ和歌山城を占領すべく一揆を
扇動したが計画は失敗した（紀州一揆）。天王寺・岡山の戦いで

もスポンサ
ーに加わる）
。2008年11
月30日27:0
0-28:00
に一夜限りの

復活をした。毎
年公開録音を東京厚生年

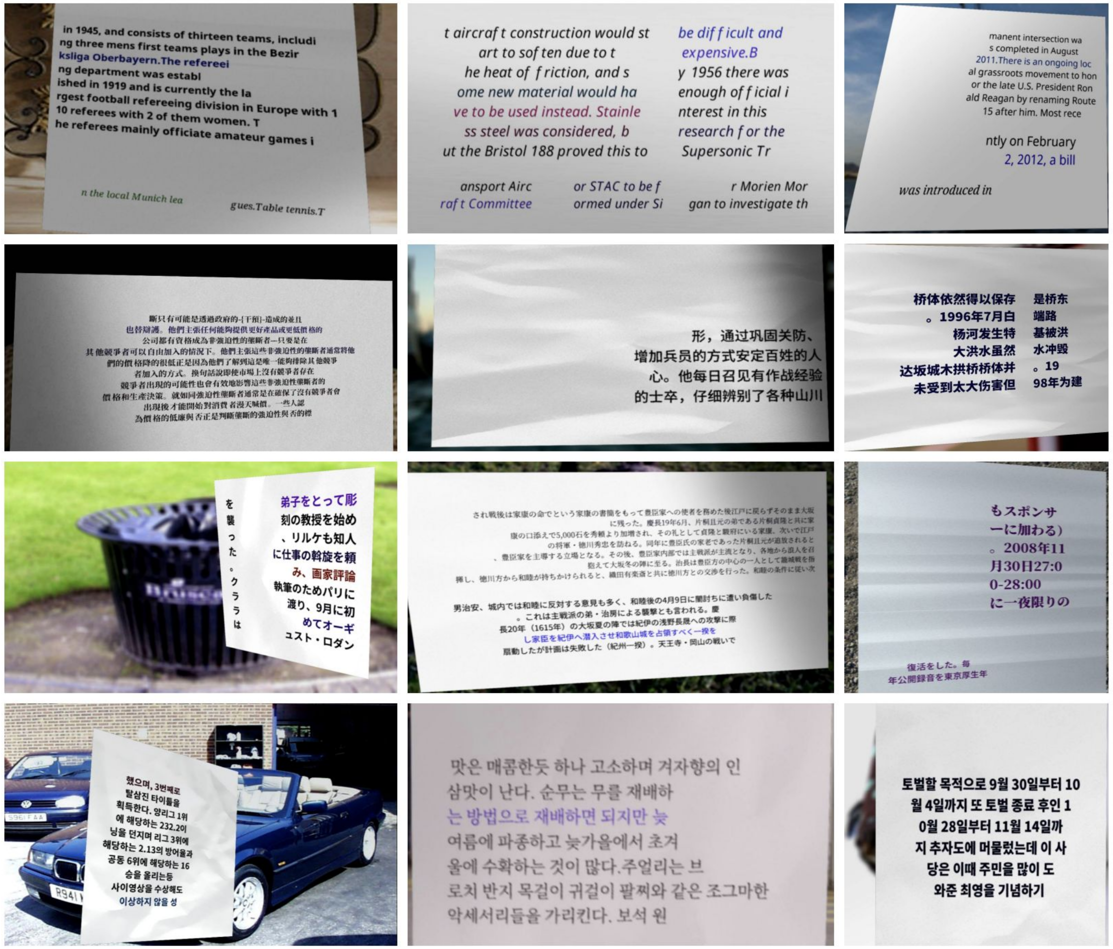
했으며, 3번째로
탈삼진 타이틀을
획득한다. 양리그 1위
에 해당하는 232.2이
닝을 던지며 리그 3위에
해당하는 2.13의 방어율과
공동 6위에 해당하는 16
승을 올리는등
사이영상을 수상해도
이상하지 않을 성

맛은 매콤한듯 하나 고소하며 겨자향의 인
삼맛이 난다. 순무는 무를 재배하
는 방법으로 재배하면 되지만 늦
여름에 파종하고 늦가을에서 초겨
울에 수확하는 것이 많다.주얼리는 브
로치 반지 목걸이 귀걸이 팔찌와 같은 조그마한
악세서리들을 가리킨다. 보석 원

토벌할 목적으로 9월 30일부터 10
월 4일까지 또 토벌 종료 후인 1
0월 28일부터 11월 14일까
지 추자도에 머물렀는데 이 사
당은 이때 주민을 많이 도
와준 최영을 기념하기

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

SynthDoG 🐶:Synthetic Document Generator



in 1945, and consists of thirteen teams, includi
ng three mens first teams plays in the Bezir
ksliga Oberbayern.The refereei
ng department was establ
ished in 1919 and is currently the la
rgest football refereeing division in Europe with 1
10 referees with 2 of them women. T
he referees mainly officiate amateur games i

n the local Munich lea

gues.Table tennis.T

t aircraft construction would st
art to soften due to t
he heat of friction, and s
ome new material would ha
ve to be used instead. Stainle
ss steel was considered, b
ut the Bristol 188 proved this to

be difficult and
expensive.B
y 1956 there was
enough official i
nterest in this
research for the
Supersonic Tr

ansport Airc
raft Committee

or STAC to be f
ormed under Si

r Morien Mor
gan to investigate th

manent intersection wa
s completed in August
2011.There is an ongoing loc
al grassroots movement to hon
or the late U.S. President Ron
ald Reagan by renaming Route
15 after him. Most rece

ntly on February
2, 2012, a bill

was introduced in

斷只有可能是透過政府的-{干預}-造成的並且
也替辯護。他們主張任何能夠提供更好產品或更低價格的
公司都有資格成為非強迫性的壟斷者—只要是在
其他競爭者可以自由加入的情況下。他們主張這些非強迫性的壟斷者通常將他
們的價格降的很低正是因為他們了解到這是唯一能夠排除其他競爭
者加入的方式。換句話說即使市場上沒有競爭者存在
競爭者出現的可能性也會有效地影響這些非強迫性壟斷者的
價格和生產決策。就如同強迫性壟斷者通常是在確保了沒有競爭者會
出現後才能開始對消費者漫天喊價。一些人認
為價格的低廉與否正是判斷壟斷的強迫性與否的標

形，通过巩固关防、
增加兵员的方式安定百姓的人
心。他每日召见有作战经验
的士卒，仔细辨别了各种山川

桥体依然得以保存
。1996年7月白
杨河发生特
大洪水虽然
达坂城木拱桥桥体并
未受到太大伤害但

是桥东
端路
基被洪
水冲毁
。19
98年为建

弟子をとって彫
刻の教授を始め
、リルケも知人
に仕事の斡旋を頼
み、画家評論
執筆のためパリに
渡り、9月に初
めてオーギ
ュスト・ロダン

を襲った。クララは

され戦後は家康の命でという家康の書簡をもって豊臣家への使者を務めた後江戸に戻らずそのまま大坂
に残った。慶長19年6月、片桐且元の弟である片桐貞隆と共に家
康の口添えで5,000石を秀頼より加増され、その礼として貞隆と駿府にいる家康、次いで江戸
の将軍・徳川秀忠を訪ねる。同年に豊臣氏の家老であった片桐且元が追放されると
、豊臣家を主導する立場となる。その後、豊臣家内部では主戦派が主流となり、各地から浪人を召
抱えて大坂冬の陣に至る。治長は豊臣方の中心の一人として籠城戦を指
揮し、徳川方から和睦が持ちかけられると、織田有楽斎と共に徳川方との交渉を行った。和睦の条件に従い次

男治安、城内では和睦に反対する意見も多く、和睦後の4月9日に闇討ちに遭い負傷した
。これは主戦派の弟・治房による襲撃とも言われる。慶
長20年（1615年）の大坂夏の陣では紀伊の浅野長晟への攻撃に際
し家臣を紀伊へ潜入させ和歌山城を占領すべく一揆を
扇動したが計画は失敗した（紀州一揆）。天王寺・岡山の戦いで

もスポンサ
ーに加わる）
。2008年11
月30日27:0
0-28:00
に一夜限りの

復活をした。毎
年公開録音を東京厚生年

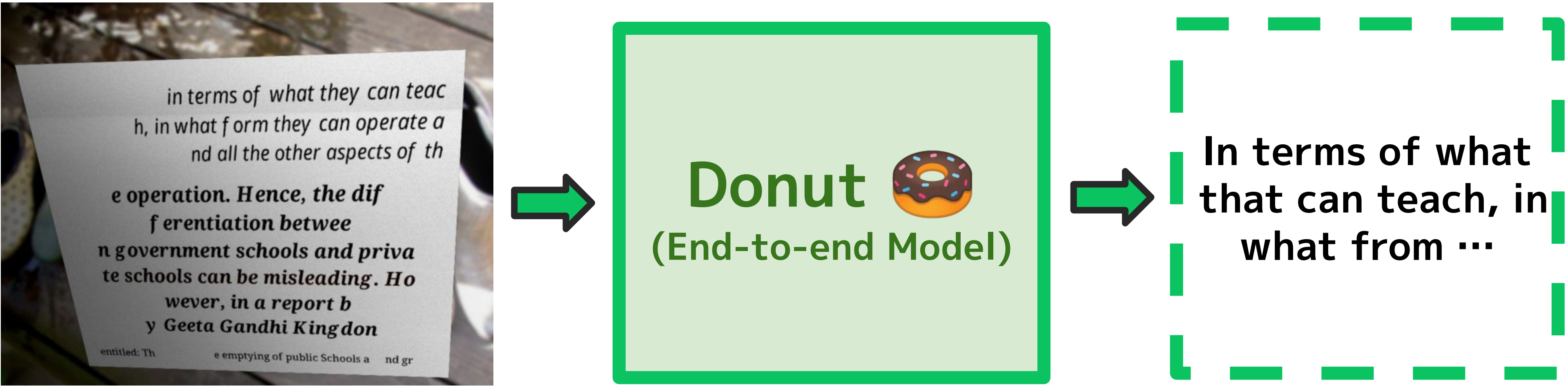
했으며, 3번째로
탈삼진 타이틀을
획득한다. 양리그 1위
에 해당하는 232.2이
닝을 던지며 리그 3위에
해당하는 2.13의 방어율과
공동 6위에 해당하는 16
승을 올리는등
사이영상을 수상해도
이상하지 않을 성

맛은 매콤한듯 하나 고소하며 겨자향의 인
삼맛이 난다. 순무는 무를 재배하
는 방법으로 재배하면 되지만 늦
여름에 파종하고 늦가을에서 초겨
울에 수확하는 것이 많다.주얼리는 브
로치 반지 목걸이 귀걸이 팔찌와 같은 조그마한
악세서리들을 가리킨다. 보석 원

토벌할 목적으로 9월 30일부터 10
월 4일까지 또 토벌 종료 후인 1
0월 28일부터 11월 14일까
지 추자도에 머물렀는데 이 사
당은 이때 주민을 많이 도
와준 최영을 기념하기

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 they can teac
h, in what form they can operate a
nd all the other aspects of th
e operation. Hence, the dif
ferentiation betwee
n government schools and priva
te schools can be misleading. Ho
wever, in a report b
y Geeta Gandhi Kingdon
entitled: Th e emptying of public Schools a nd gr

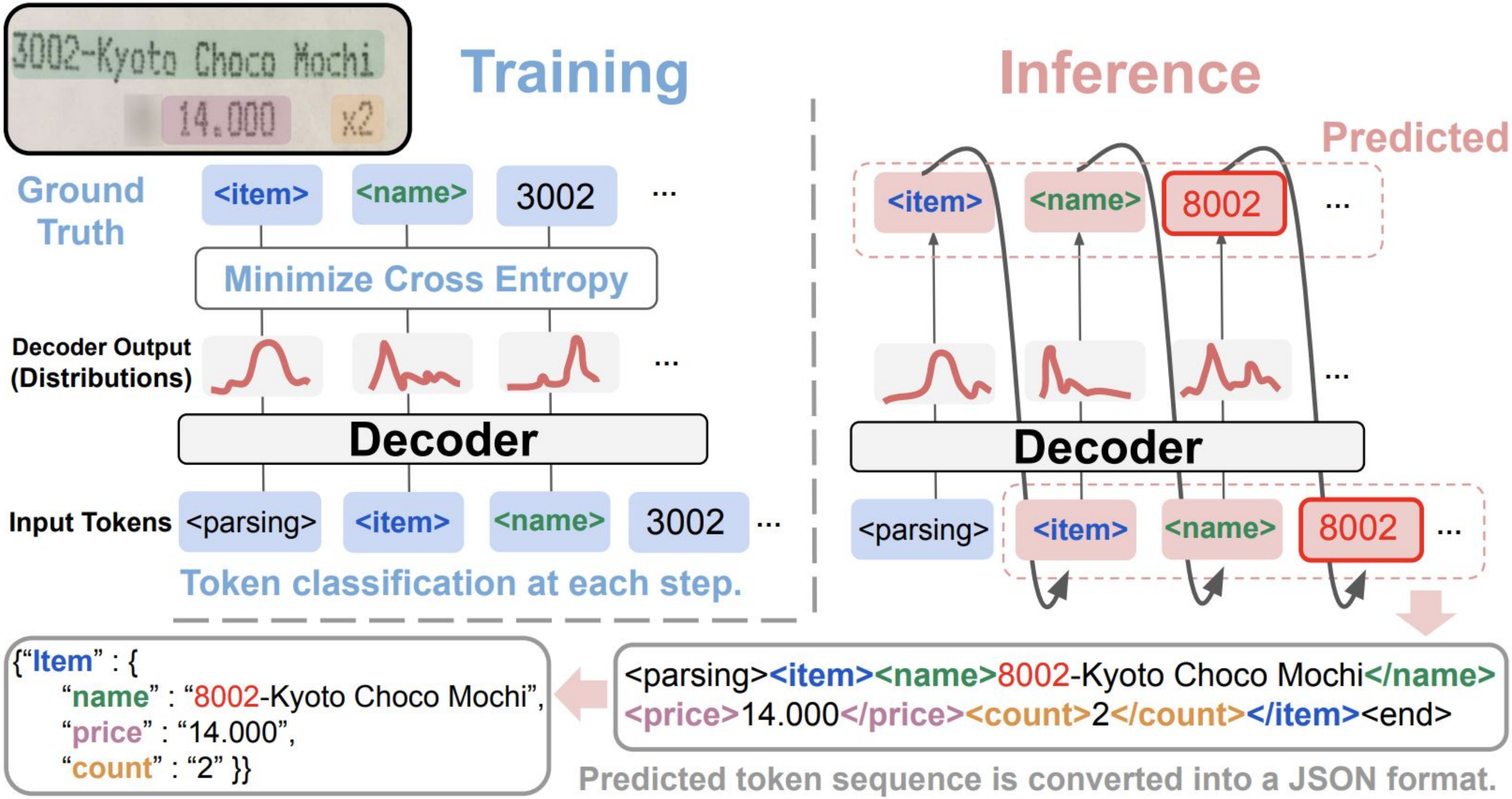
Donut 🍩
(End-to-end Model)

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

After the model learns “how to read”,

Slide

Model Fine-tuning and Inference Overview



Training

Ground Truth <item> <name> 3002 ...

Minimize Cross Entropy

Decoder Output (Distributions) ...

Decoder

Input Tokens <parsing> <item> <name> 3002 ...

Token classification at each step.

Inference

Predicted <item> <name> 8002 ...

Decoder

<parsing> <item> <name> 8002 ...

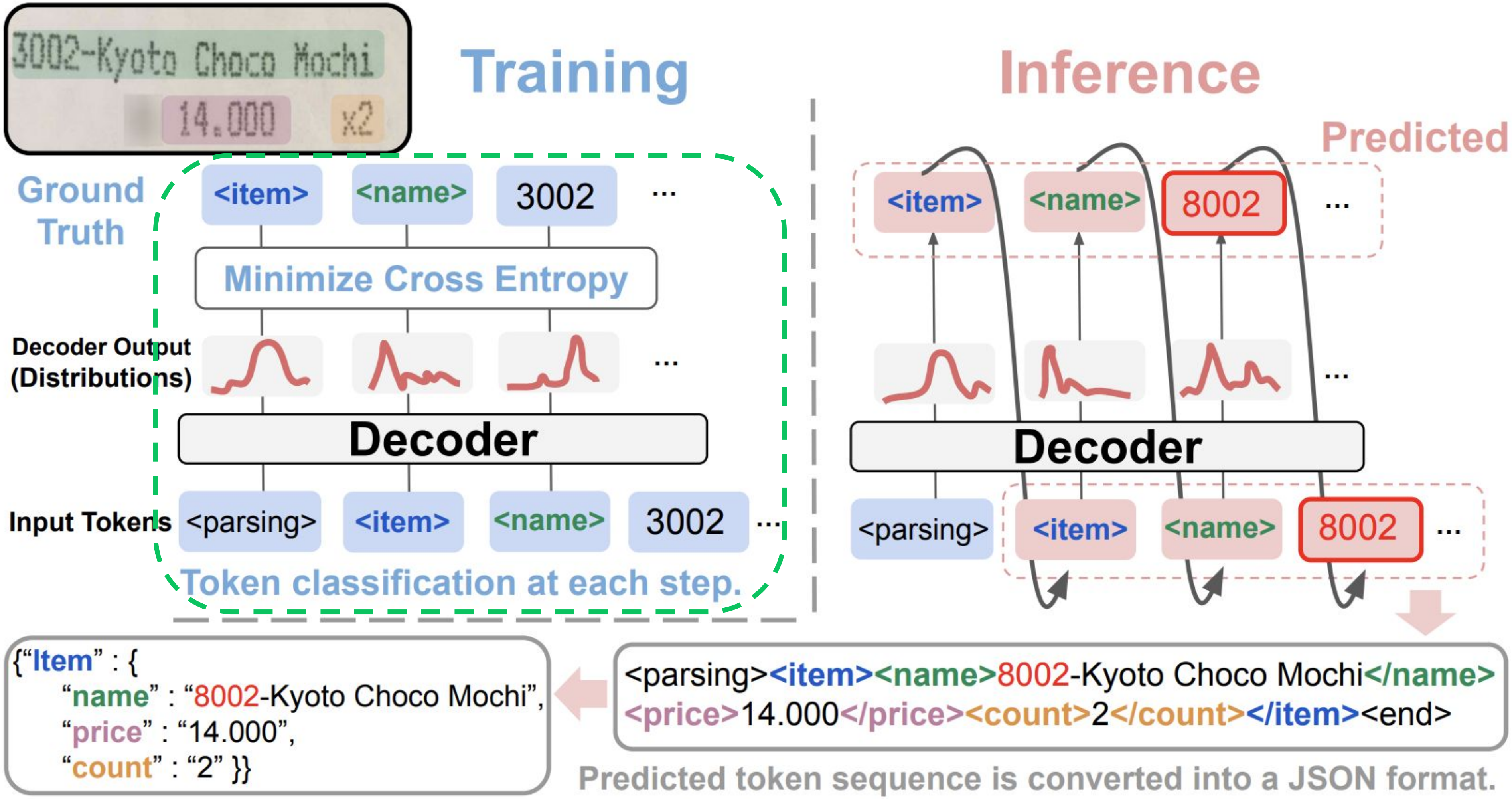
{"Item" : {
"name" : "8002-Kyoto Choco Mochi",
"price" : "14.000",
"count" : "2" }}

<parsing><item><name>8002-Kyoto Choco Mochi</name><price>14.000</price><count>2</count></item><end>

Predicted token sequence is converted into a JSON format.

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

Model Fine-tuning and Inference Overview



Token classification at each step.

```
{"Item" : {
    "name" : "8002-Kyoto Choco Mochi",
    "price" : "14.000",
    "count" : "2" }}
```

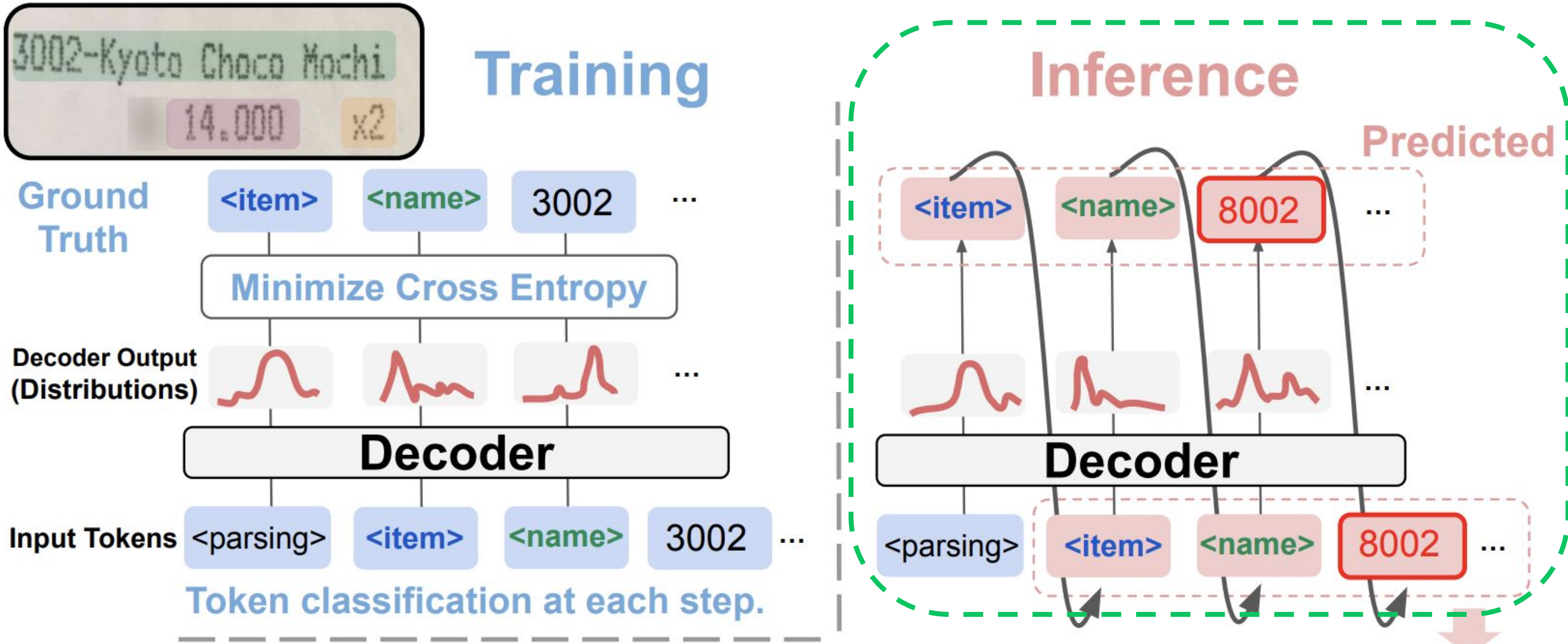
<parsing><item><name>8002-Kyoto Choco Mochi</name><price>14.000</price><count>2</count></item><end>

Predicted token sequence is converted into a JSON format.

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

Slide

Model Fine-tuning and Inference Overview



Token classification at each step.

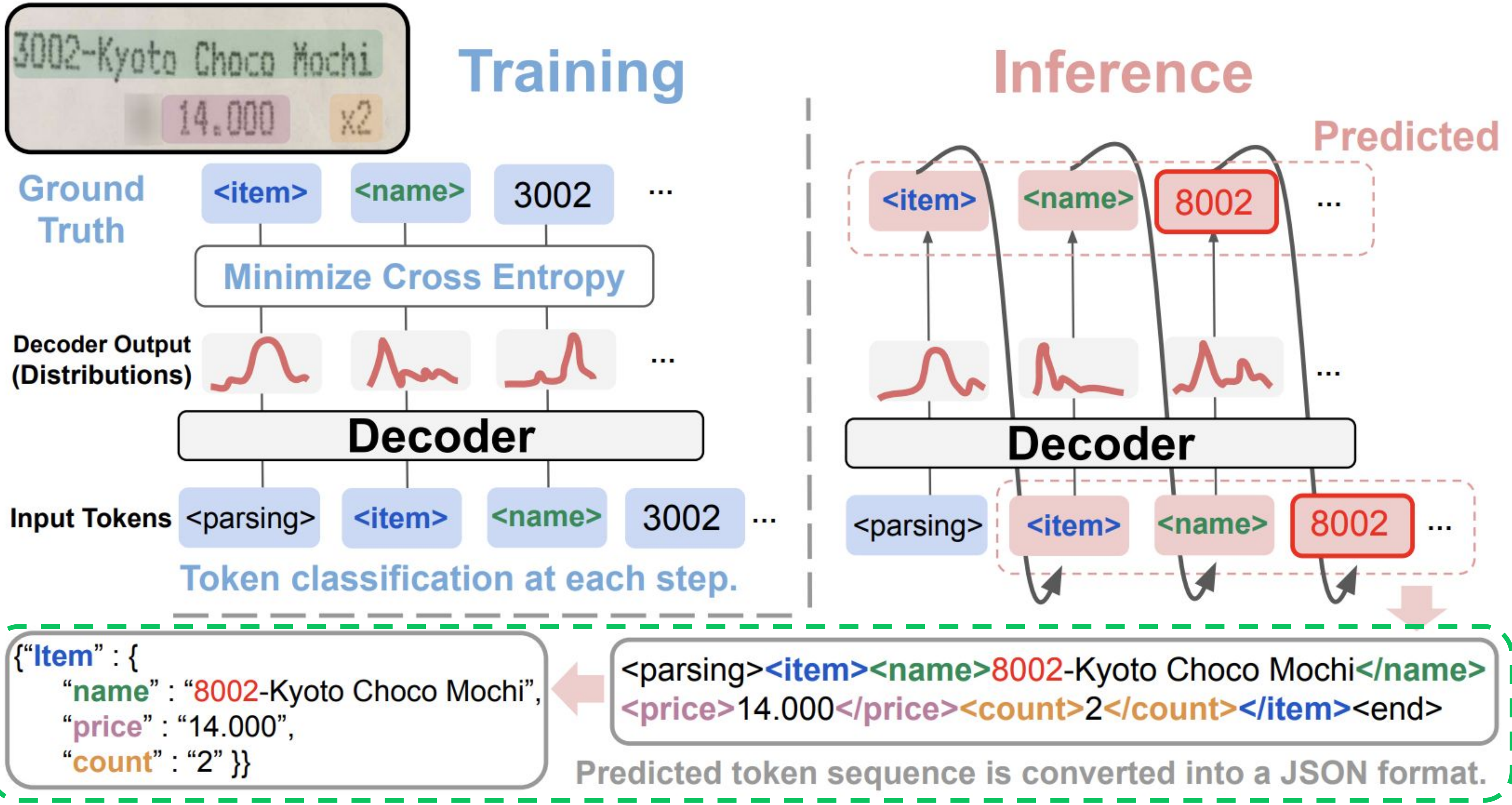
```
{"Item" : {
    "name" : "8002-Kyoto Choco Mochi",
    "price" : "14.000",
    "count" : "2" }}
```

```
<parsing><item><name>8002-Kyoto Choco Mochi</name>
<price>14.000</price><count>2</count></item><end>
```

Predicted token sequence is converted into a JSON format.

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

Model Fine-tuning and Inference Overview



Token classification at each step.

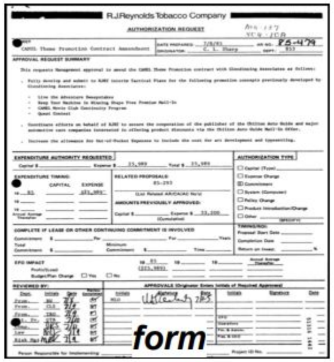
```
{"Item" : {
    "name" : "8002-Kyoto Choco Mochi",
    "price" : "14.000",
    "count" : "2" }}
```

```
<parsing><item><name>8002-Kyoto Choco Mochi</name>
<price>14.000</price><count>2</count></item><end>
```

Predicted token sequence is converted into a JSON format.

The predicted sequence is converted into a JSON format.

Experiments: Samples of Downstream Datasets


[redacted] R.J.Reynolds Tobacco Company [redacted]

AUTHORIZATION REQUEST

CAMEL Theme Promotion Contract Amendment

AR NO. 85-479

APPROVAL REQUEST SUMMARY

- Keep Your Machine in Winning Shape Free Premium Mail-In
- CAMEL Movie Club Continuity Program
- Quest Contest

EXPENDITURE AUTHORITY REQUESTED

AUTHORIZATION TYPE

□ Capital (Type)
□ Expense Charge
☒ Commitment
□ System (Computer)
□ Policy Change
□ Product Introduction/Change
□ Other

EXPENDITURE TIMING

CAPITAL EXPENSE

RELATED PROPOSALS: 85-293

AMOUNTS PREVIOUSLY APPROVED:

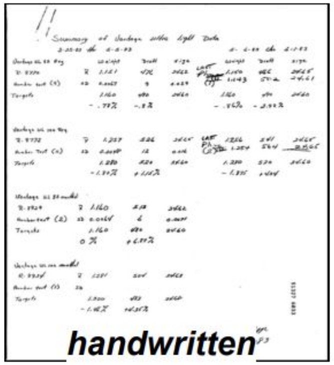
COMPLETE IF LEASE OR OTHER CONTINUING COMMITMENT IS INVOLVED

EPO IMPACT

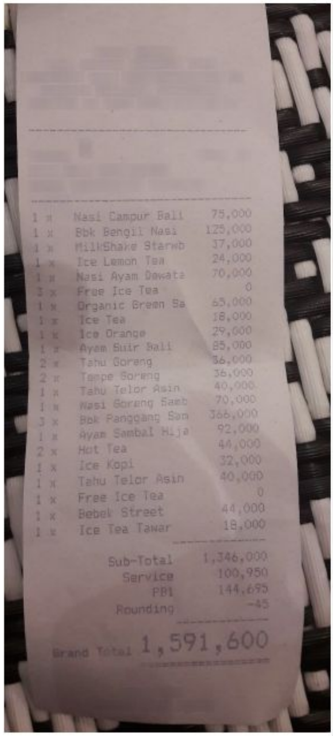
REVIEWED BY:

APPROVALS (Originator Enters Initials of Required Approvers)

form

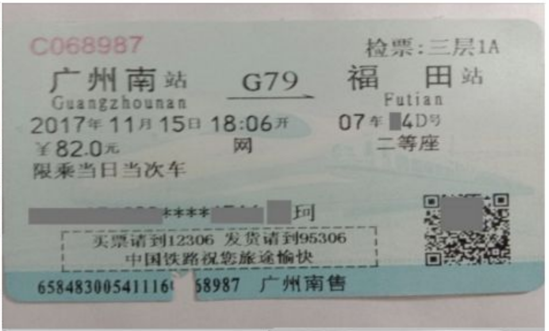

Summary of [illegible] [illegible] light Data

handwritten



1	x	Nasi Campur Bali	75,000
1	x	Bbk Bengil Nasi	125,000
1	x	MilkShake Starwb	37,000
1	x	Ice Lemon Tea	24,000
1	x	Nasi Ayam Dewata	70,000
1	x	Free Ice Tea	0
1	x	Organic Green Sa	65,000
1	x	Ice Tea	18,000
1	x	Ice Orange	29,000
1	x	Ayam Suir Bali	85,000
2	x	Tahu Goreng	36,000
2	x	Tempe Goreng	36,000
1	x	Tahu Telor Asin	40,000
1	x	Nasi Goreng Samb	70,000
3	x	Bbk Panggang Sam	366,000
1	x	Ayam Sambal Hija	92,000
2	x	Hot Tea	44,000
1	x	Ice Kopi	32,000
1	x	Tahu Telor Asin	40,000
1	x	Free Ice Tea	0
1	x	Bebek Street	44,000
1	x	Ice Tea Tawar	18,000

Sub-Total	1,346,000
Service	100,950
PB1	144,695
Rounding	-45
Grand Total	1,591,600


C068987 检票：三层1A

广州南站 G79 福　田站

Guangzhounan Futian

2017年11月15日 18:06开 07车 [redacted]4D号

¥82.0元 网 二等座

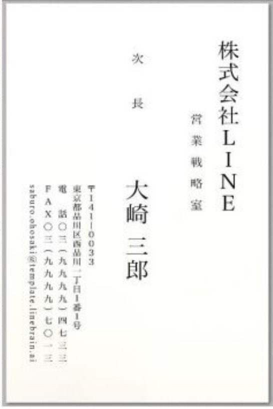
限乘当日当次车

[redacted]****[redacted]珂

买票请到12306 发货请到95306

中国铁路祝您旅途愉快

65848300541116 58987 广州南售


株式会社LINE

営業戦略室

次長

大崎　三郎

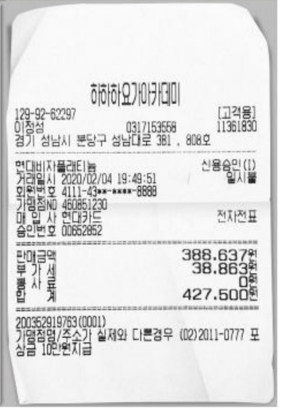
〒141-0033

東京都品川区西品川一丁目1番1号

電　話〇三（九九九九）四七三三

FAX〇三（九九九九）七〇一三

saburo.ohosaki@template.linebrain.ai


하하하요가아카데미

129-92-62297 [고객용]

이정성 0317153558 11361830

경기 성남시 분당구 성남대로 381 , 808호

현대비자플래티늄 신용승인(I)

거래일시 2020/02/04 19:49:51 일시불

회원번호 4111-43**-****-8888

가맹점NO 460851230

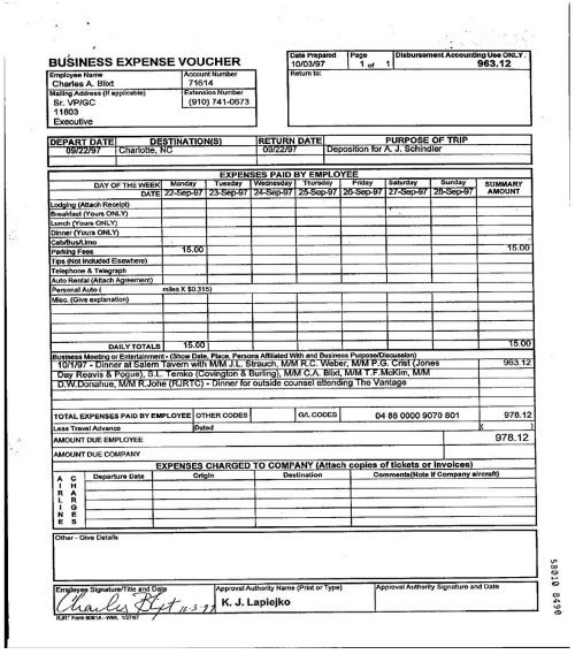
매 입 사 현대카드 전자전표

승인번호 00652852

판매금액	388.637원
부 가 세	38.863원
봉 사 료	0원
합　　계	427.500원

200352919763(0001)

가맹점명/주소가 실제와 다른경우 (02)2011-0777 포상금 10만원지급


BUSINESS EXPENSE VOUCHER

Employee Name Charles A. Blixt	Account Number 71614
Mailing Address (If applicable) Sr. VP/GC 11803 Executive	Extension Number (910) 741-0673

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

Return to:

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)

Other - Give Details

Employee Signature/Title and Date	Approval Authority Name (Print or Type)	Approval Authority Signature and Date
Charles Blixt 10-3-97	K. J. Lapiejko	

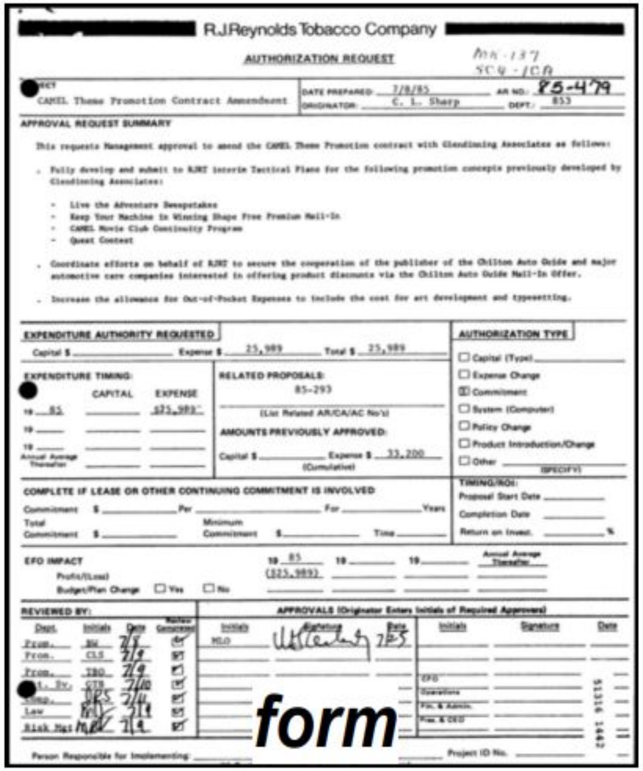
RJRT Form 8061A - WML 1/27/97

58010 8490

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.

Experiments: Document Classification



[redacted] R.J.Reynolds Tobacco Company [redacted]

AUTHORIZATION REQUEST

MK-137
5C4-ICA

SUBJECT: CAMEL Theme Promotion Contract Amendment

DATE PREPARED: 7/8/85 AR NO: 85-479
ORIGINATOR: C. L. Sharp DEPT: 853

APPROVAL REQUEST SUMMARY

This requests Management approval to amend the CAMEL Theme Promotion contract with Glendinning Associates as follows:

- Fully develop and submit to RJRT interim Tactical Plans for the following promotion concepts previously developed by Glendinning Associates:
  - Live the Adventure Sweepstakes
  - Keep Your Machine in Winning Shape Free Premium Mail-In
  - CAMEL Movie Club Continuity Program
  - Quest Contest
- Coordinate efforts on behalf of RJRT 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:

	CAPITAL	EXPENSE
19 85		$25,989
19		
19		
Annual Average Thereafter		

RELATED PROPOSALS:
85-293
(List Related AR/CA/AC No's)

AMOUNTS PREVIOUSLY APPROVED:
Capital $ ______ Expense $ 33,200
(Cumulative)

AUTHORIZATION TYPE
□ Capital (Type) ______
□ Expense Change
☒ Commitment
□ System (Computer)
□ Policy Change
□ Product Introduction/Change
□ Other ______ (SPECIFY)

COMPLETE IF LEASE OR OTHER CONTINUING COMMITMENT IS INVOLVED
Commitment $ ______ Per ______ For ______ Years
Total Commitment $ ______ Minimum Commitment $ ______ Time ______

TIMING/ROI:
Proposal Start Date ______
Completion Date ______
Return on Invest. ______ %

EFO IMPACT

	19 85	19	19	Annual Average Thereafter
Profit/(Loss)	($25,989)			
Budget/Plan Change □ Yes □ No				

REVIEWED BY:

Dept.	Initials	Date	Review Completed
Prom.	BW	7/5	☑
Prom.	CLS	7/9	☑
Prom.	TBO	7/9	☑
Mkt. Sv.	CTB	7/10	☑
Comp.	DPS	7/11	☑
Law	RPD	7/11	☑
Risk Mgt	MBV	7/11	☑

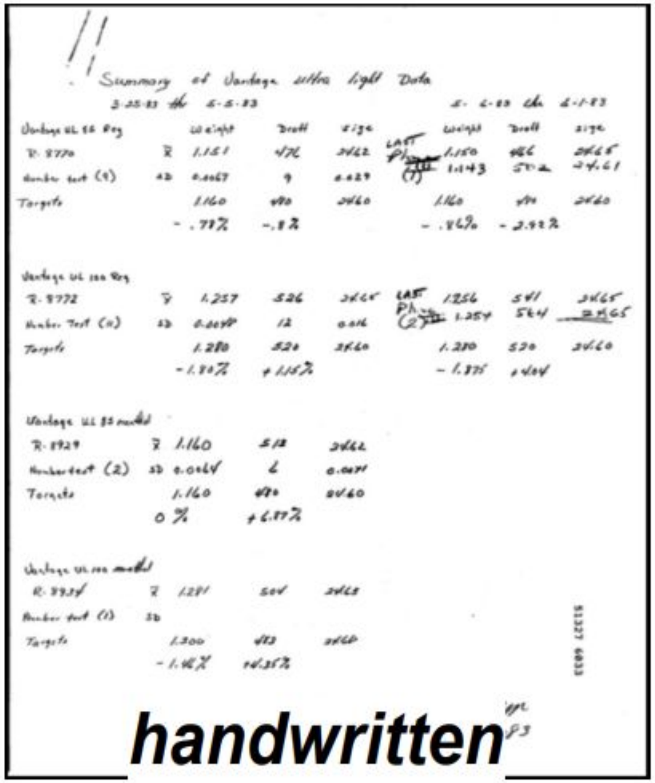
APPROVALS (Originator Enters Initials of Required Approvers)

Initials	Signature	Date	Initials	Signature	Date
MLO		7/25			
			CFO		
			Operations		
			Fin. & Admin.		
			Pres. & CEO		

Person Responsible for Implementing: ______ Project ID No. ______

51316 1442

form



Summary of Vantage ultra light Data

Vantage UL 85 Reg		3-25-83 thru 5-5-83				5-6-83 thru 6-1-83		
		Weight	Draft	Size		Weight	Draft	Size
R-8770	x̄	1.151	476	24.62	LAST Phase (1)	1.150 1.143	486 502	24.65 24.61
Number test (9)	SD	0.0067	9	0.029				
Targets		1.160	480	24.60		1.160	480	24.60
		-.78%	-.8%			-.86%	-2.92%	

Vantage UL 100 Reg								
R-8772	x̄	1.257	526	24.65	LAST Phase (2)	1.256 1.254	541 544	24.65 24.65
Number Test (8)	SD	0.0048	12	0.016				
Targets		1.280	520	24.60		1.280	520	24.60
		-1.80%	+1.15%			-1.875	+4.04	

Vantage UL 85 menthol				
R-8929	x̄	1.160	512	24.62
Number test (2)	SD	0.0064	6	0.0071
Targets		1.160	480	24.60
		0%	+6.67%	

Vantage UL 100 menthol				
R-8934	x̄	1.281	504	24.65
Number test (1)	SD			
Targets		1.300	483	24.60
		-1.46%	+4.35%	

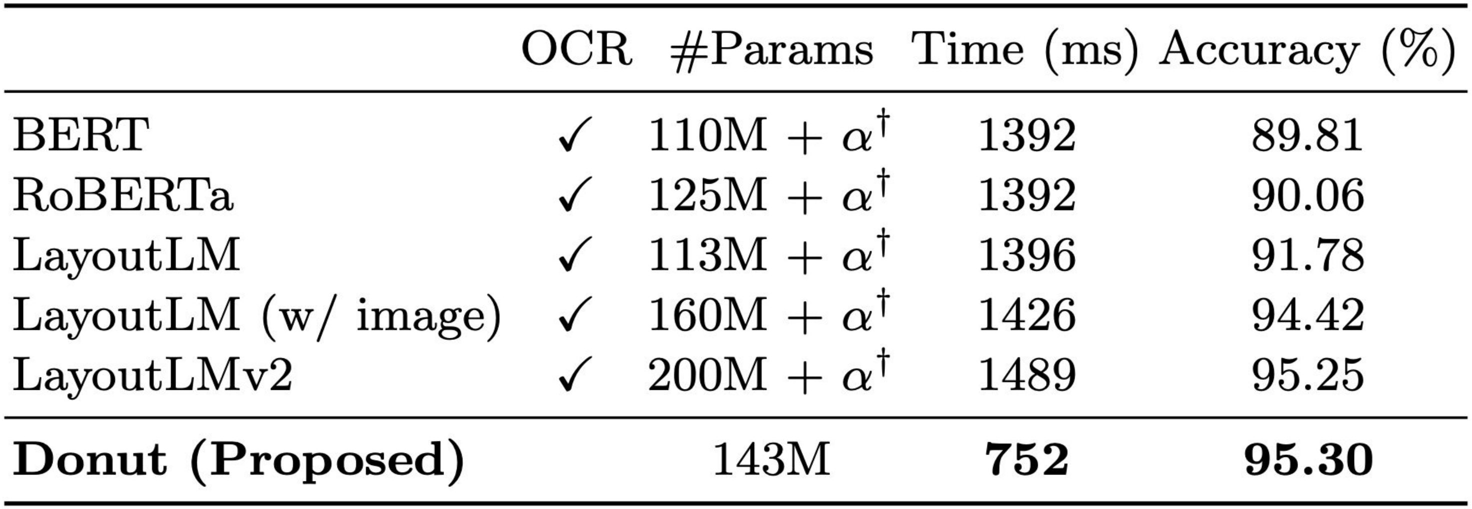
51327 6033

handwritten

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

Slide

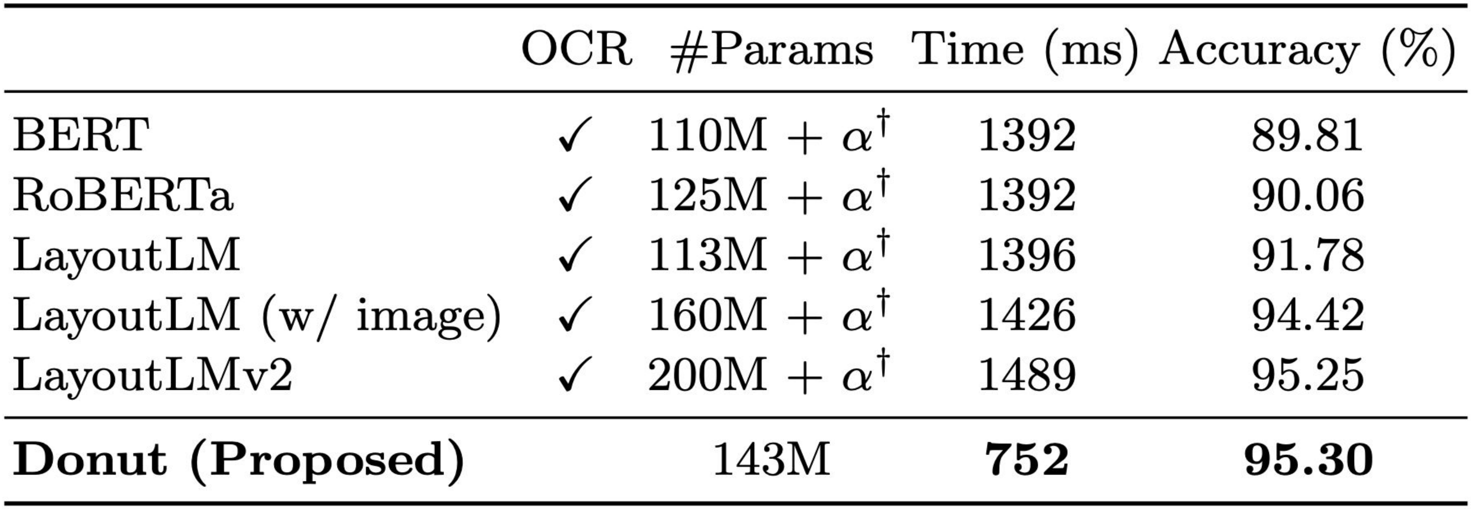
Experiments: Document Classification



	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.

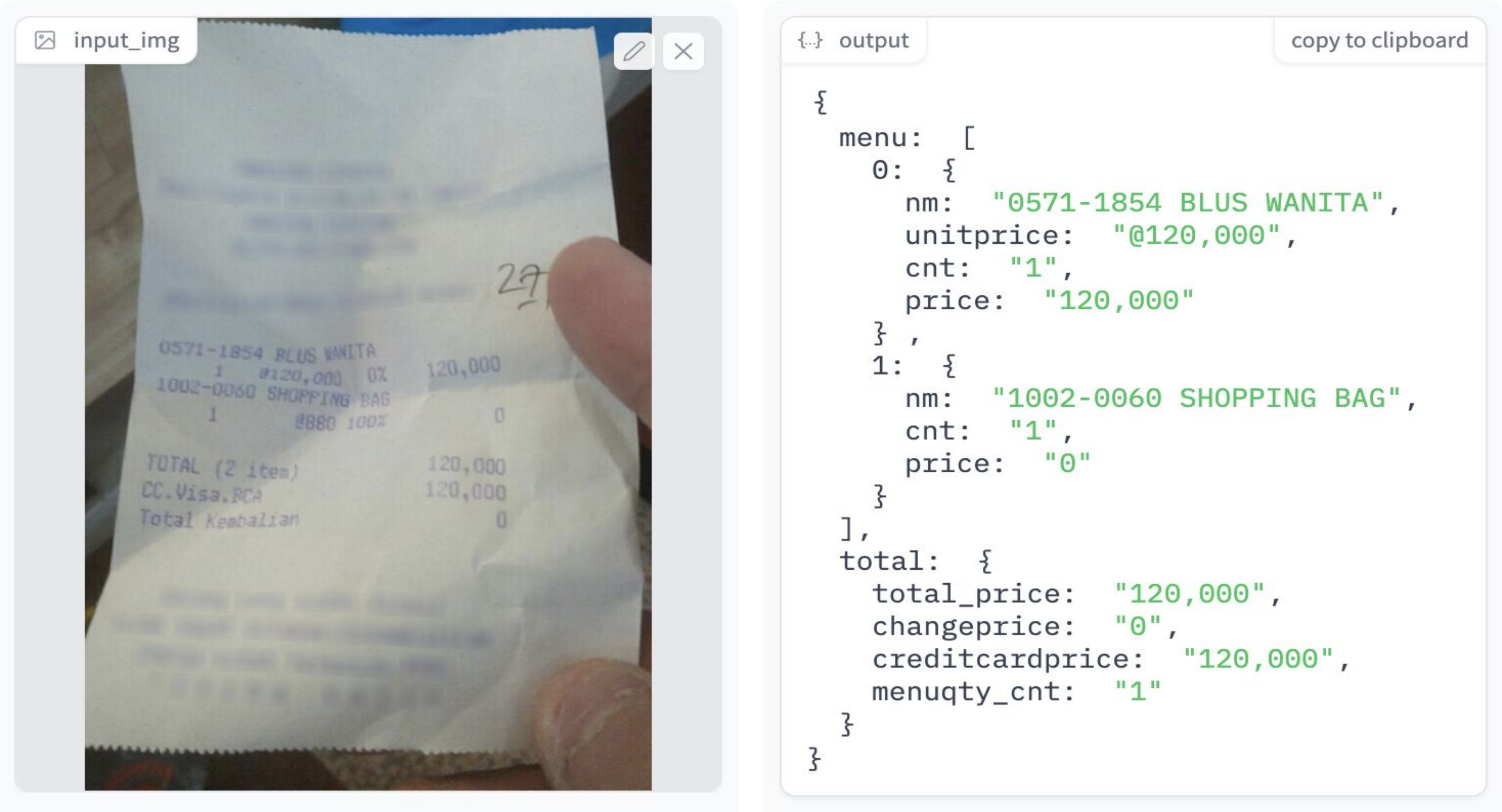
Experiments: Document Classification



	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.

Experiments: Document Parsing



[redacted]

29

0571-1854 BLUS WANITA
1 @120,000 0% 120,000
1002-0060 SHOPPING BAG
1 @880 100% 0

TOTAL (2 item) 120,000
CC.Visa.BCA 120,000
Total Kembalian 0

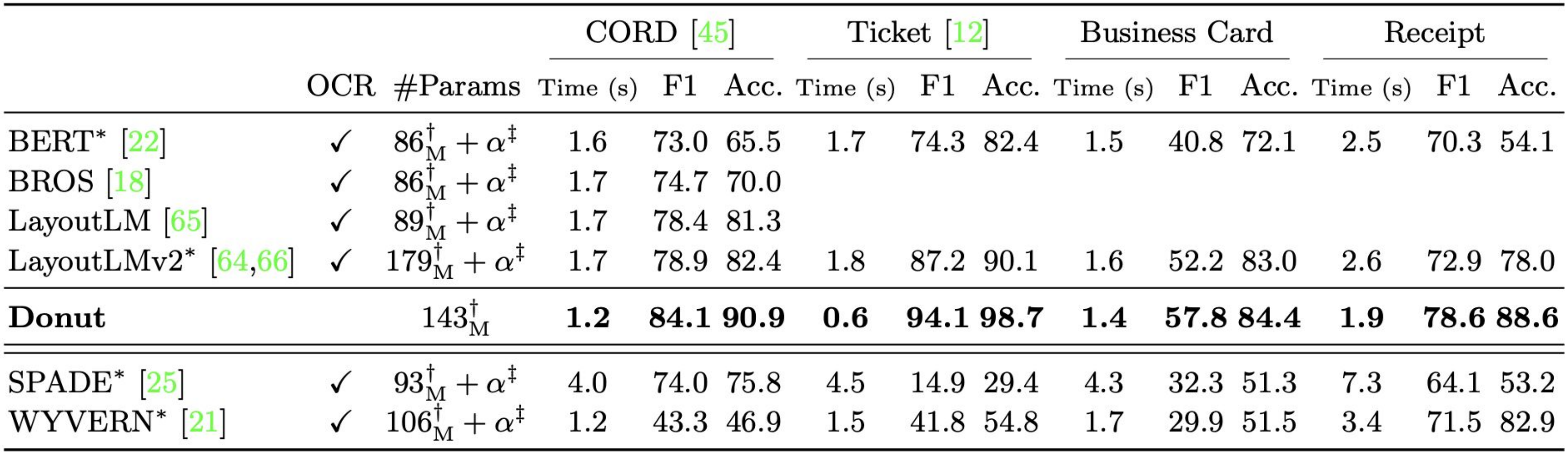
[redacted]

```
{
  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

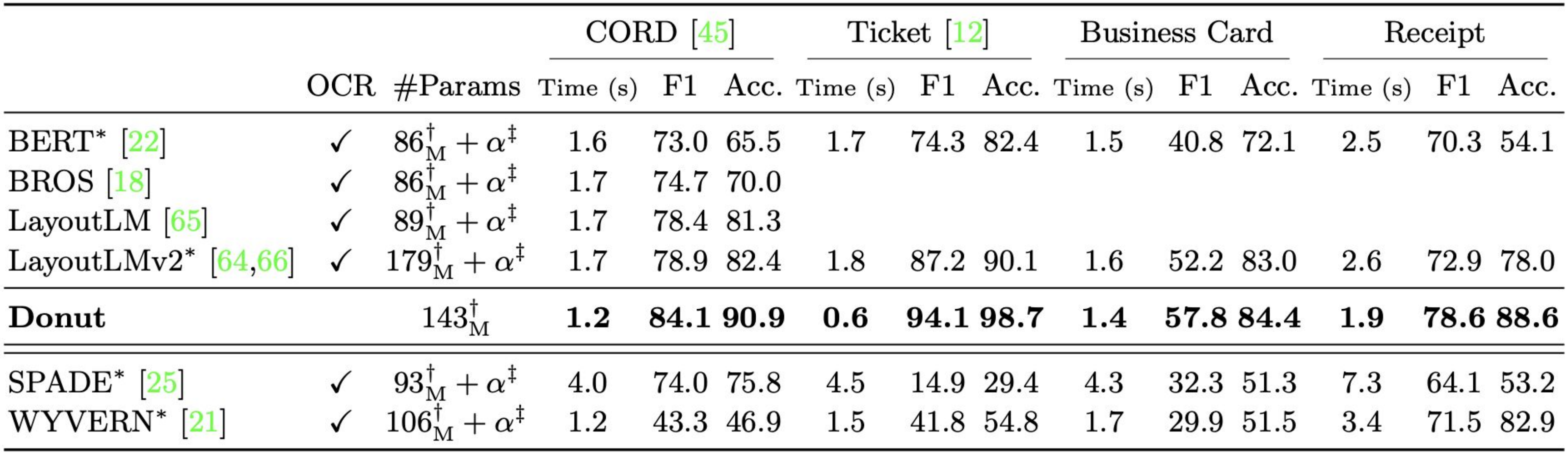
Experiments: Document Parsing



			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

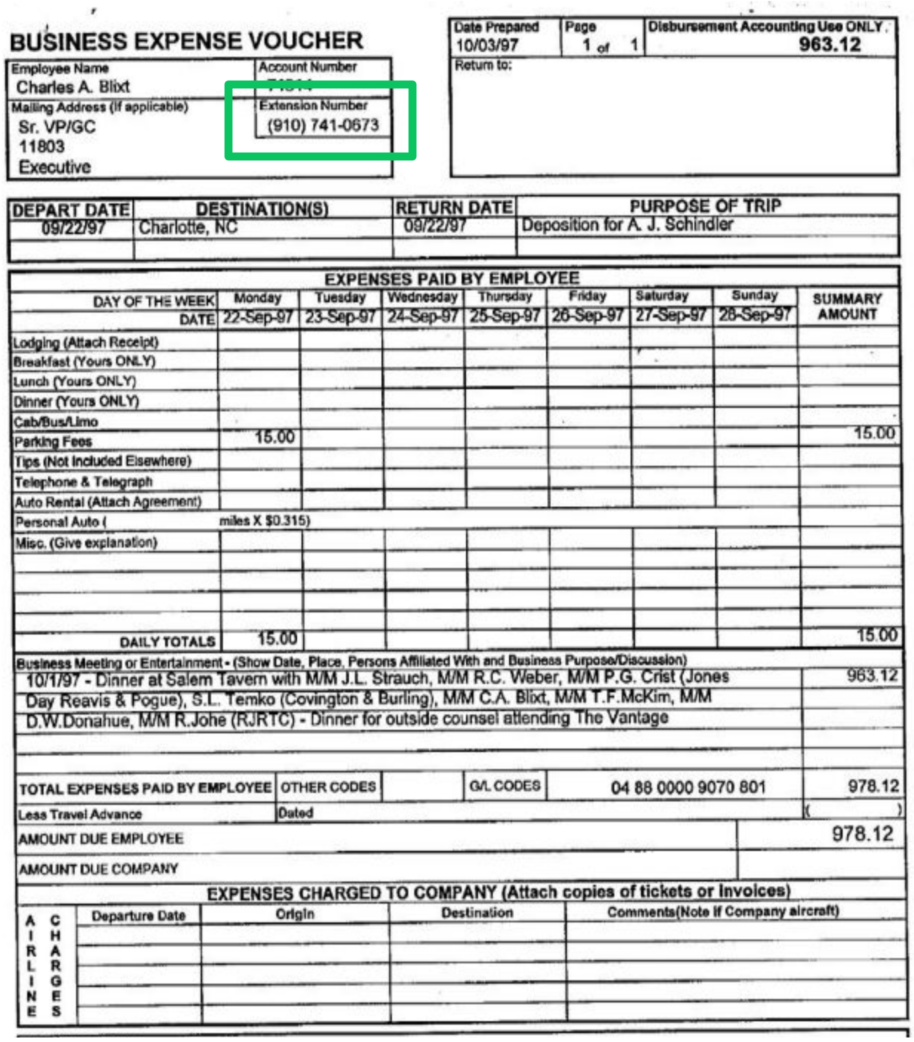
Experiments: Document Parsing



			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]	✓	$86^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	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]	✓	$86^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	1.7	74.7	70.0									
LayoutLM [65]	✓	$89^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	1.7	78.4	81.3									
LayoutLMv2* [64,66]	✓	$179^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	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		$143^{\dagger}_{\text{M}}$	**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]	✓	$93^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	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]	✓	$106^{\dagger}_{\text{M}} + \alpha^{\ddagger}$	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.

Experiments: Document VQA



BUSINESS EXPENSE VOUCHER

Employee Name Charles A. Blixt	Account Number 71814	
Mailing Address (If applicable) Sr. VP/GC 11803 Executive	Extension Number (910) 741-0673	

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

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	963.12
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	

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)

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

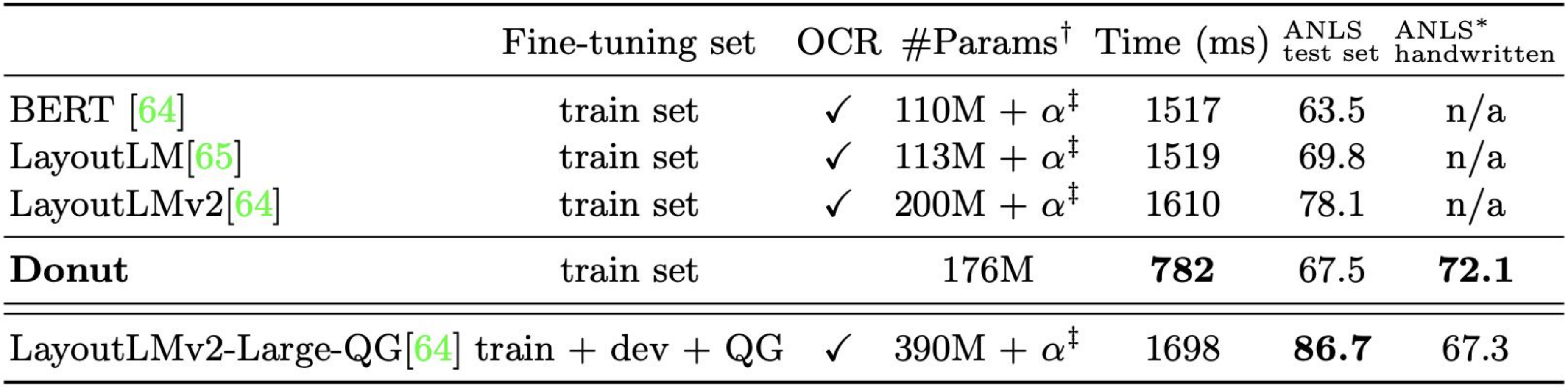
Q: What is the Extension Number as per the voucher?

A: (910) 741–0673

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

Slide

Experiments: Document VQA

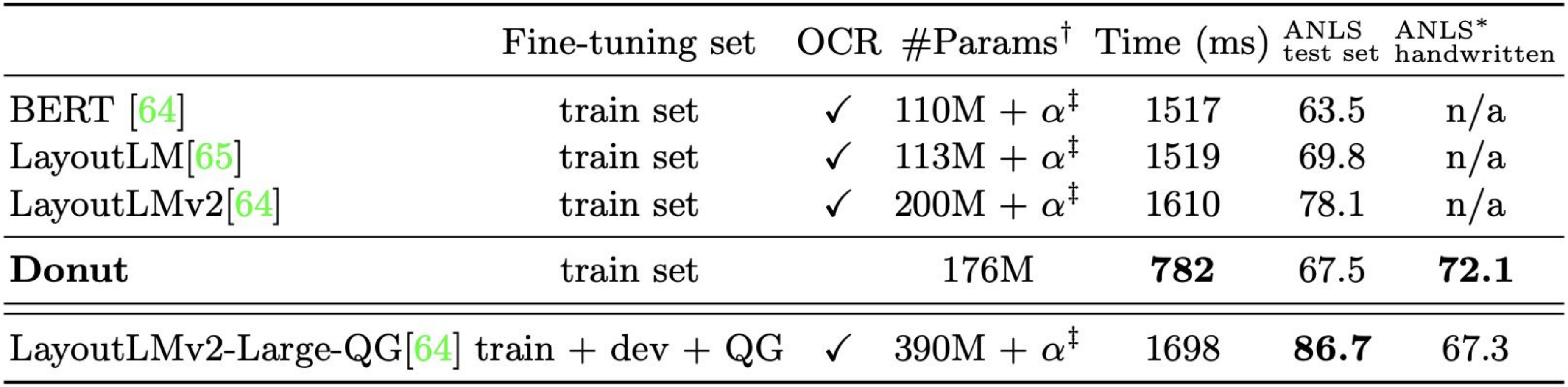


	Fine-tuning set	OCR	#Params†	Time (ms)	ANLS test set	ANLS* handwritten
BERT [64]	train set	✓	110M + $\alpha^\ddagger$	1517	63.5	n/a
LayoutLM[65]	train set	✓	113M + $\alpha^\ddagger$	1519	69.8	n/a
LayoutLMv2[64]	train set	✓	200M + $\alpha^\ddagger$	1610	78.1	n/a
Donut	train set		176M	**782**	67.5	**72.1**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

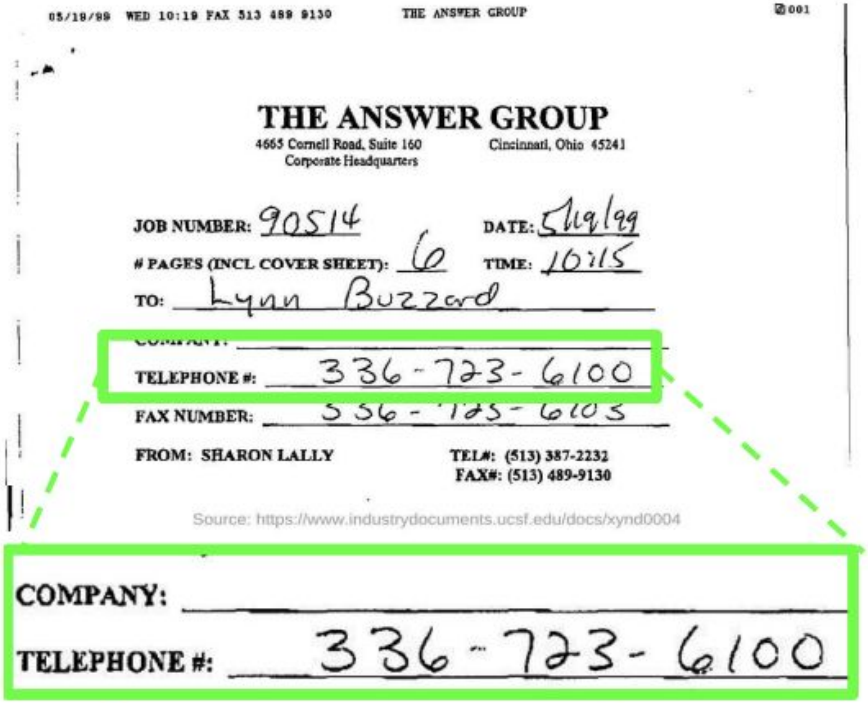


	Fine-tuning set	OCR	#Params†	Time (ms)	ANLS test set	ANLS* handwritten
BERT [64]	train set	✓	110M + $\alpha^\ddagger$	1517	63.5	n/a
LayoutLM[65]	train set	✓	113M + $\alpha^\ddagger$	1519	69.8	n/a
LayoutLMv2[64]	train set	✓	200M + $\alpha^\ddagger$	1610	78.1	n/a
Donut	train set		176M	**782**	67.5	**72.1**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



05/19/99 WED 10:19 FAX 513 489 9130 THE ANSWER GROUP 001

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: 336-723-6103
FROM: SHARON LALLY
TEL#: (513) 387-2232
FAX#: (513) 489-9130

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

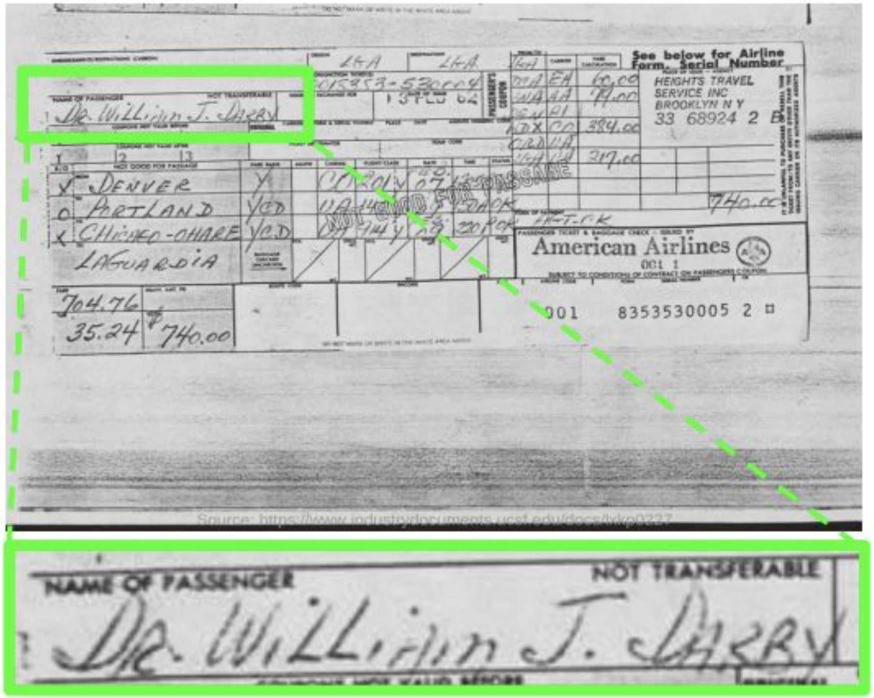
COMPANY:
TELEPHONE #: 336-723-6100

Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-QG: 336-723- 4100



NAME OF PASSENGER NOT TRANSFERABLE
Dr. William J. Darby

LGA LGA
13 FEB 62
See below for Airline Form, Serial Number
HEIGHTS TRAVEL
SERVICE INC
BROOKLYN N Y
33 68924 2 B

X DENVER
O PORTLAND
X CHICAGO-OHARE
LAGUARDIA

60.00
74.00
384.00
317.00
740.00

NOT GOOD FOR PASSAGE

American Airlines
001 1

704.76
35.24
740.00

001 8353530005 2

Source: https://www.industrydocuments.ucsf.edu/docs/[illegible]

NAME OF PASSENGER NOT TRANSFERABLE
Dr. William J. Darby

Q: What is the name of the passenger?

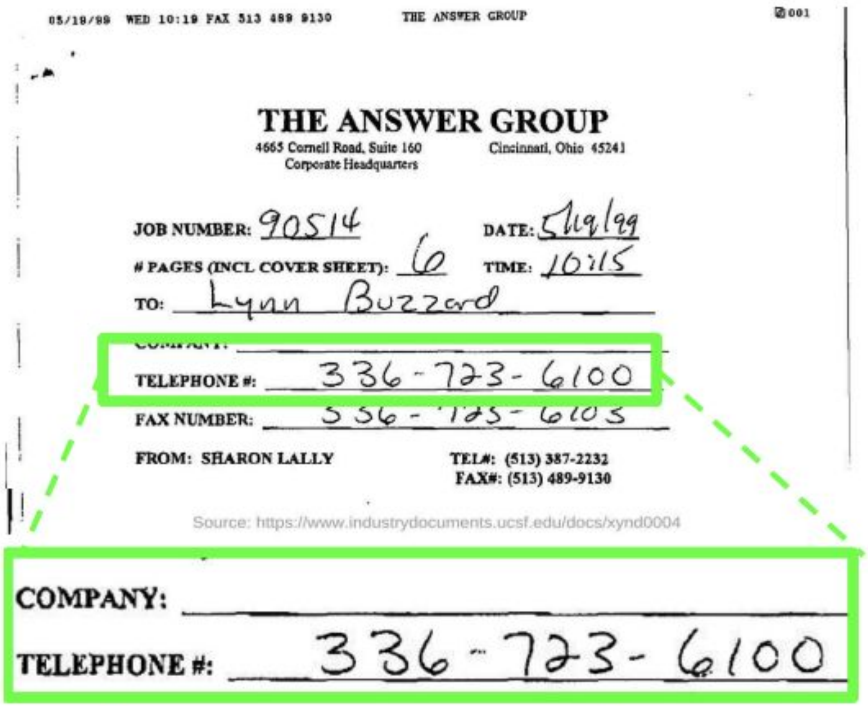
Answer: DR. William J. Darby

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.

Analysis: VQA on Handwritten Documents



05/19/99 WED 10:19 FAX 513 489 9130 THE ANSWER GROUP 001

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: 336-723-6103

FROM: SHARON LALLY TEL#: (513) 387-2232
FAX#: (513) 489-9130

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

COMPANY:

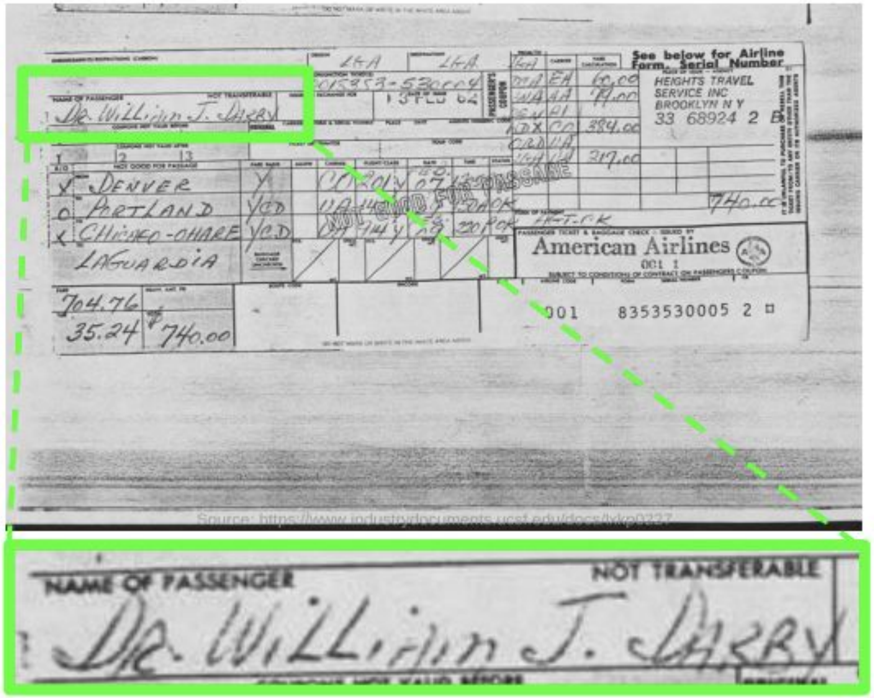
TELEPHONE #: 336-723-6100

Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-QG: 336-723- 4100



NAME OF PASSENGER NOT TRANSFERABLE

Dr. William J. Darby

LGA LGA

15353-530004

13 FEB 62

See below for Airline Form, Serial Number

HEIGHTS TRAVEL
SERVICE INC
BROOKLYN N Y
33 68924 2 B

X DENVER Y CO 201 Y 07
O PORTLAND YCD UA 444 OK
X CHICAGO-OHARE YCD AA 194 Y 8 230 OK
LAGUARDIA

60.00
79.00
384.00
217.00
740.00

704.76
35.24 $ 740.00

NOT GOOD FOR PASSAGE

American Airlines
001 1

001 8353530005 2

Source: https://www.industrydocuments.ucsf.edu/docs/lykp0227

NAME OF PASSENGER NOT TRANSFERABLE

Dr. William J. Darby

Q: What is the name of the passenger?

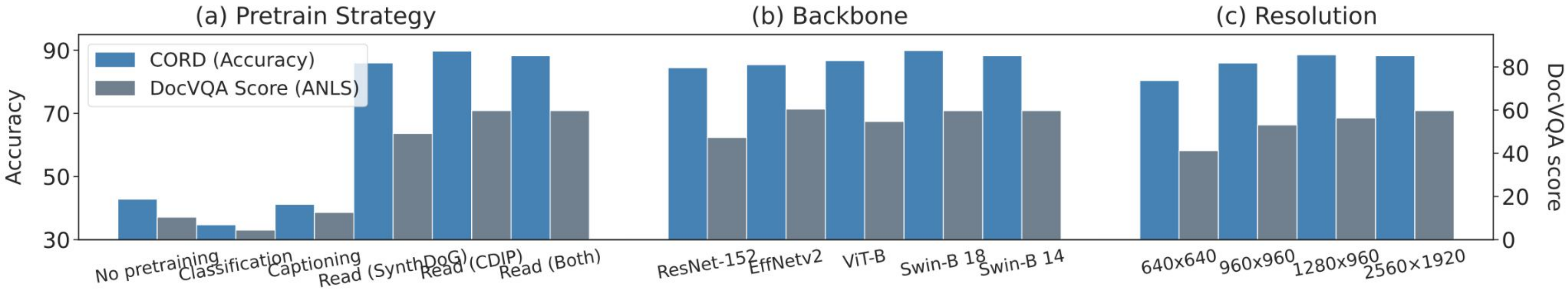
Answer: DR. William J. Darby

Donut: DR. William J. Darby

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

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

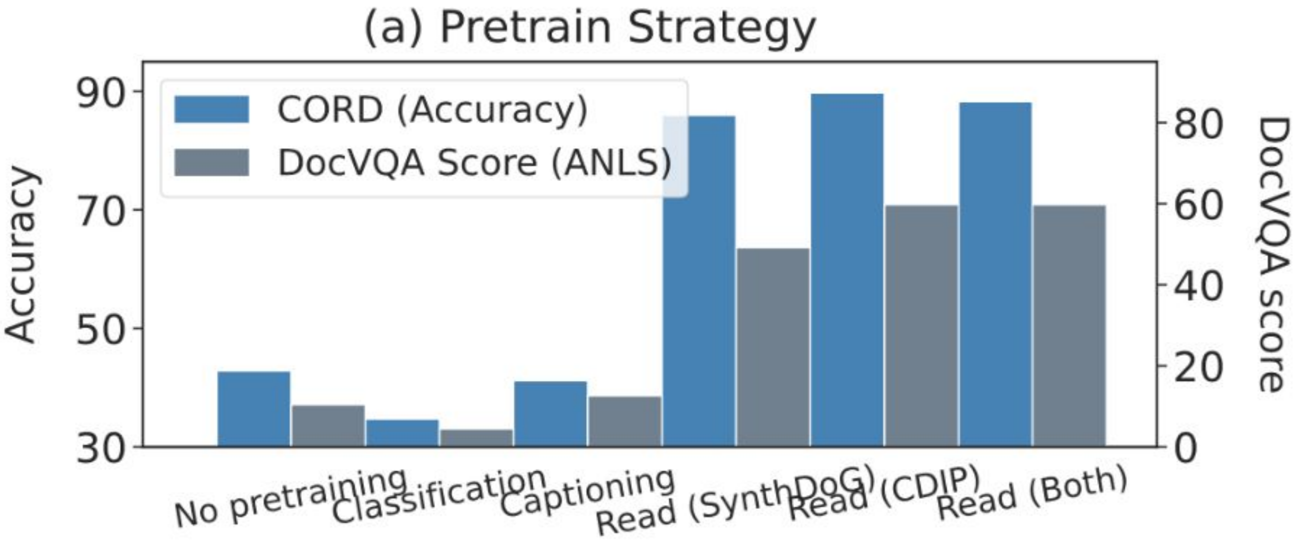
Analysis



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

Slide

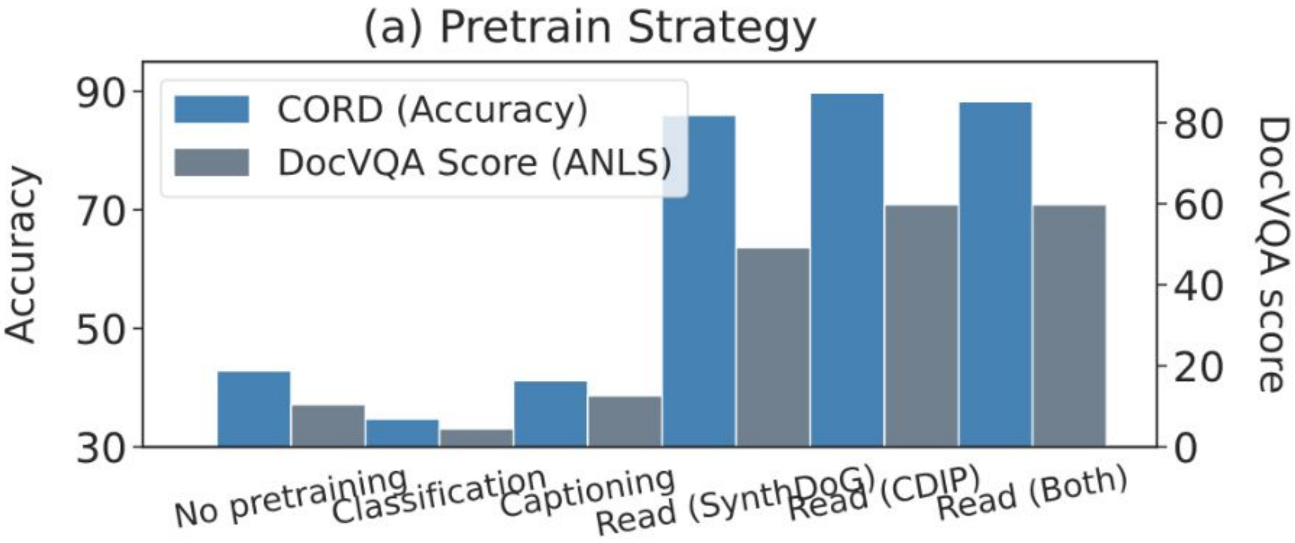
Analysis: Pre-training Strategies



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

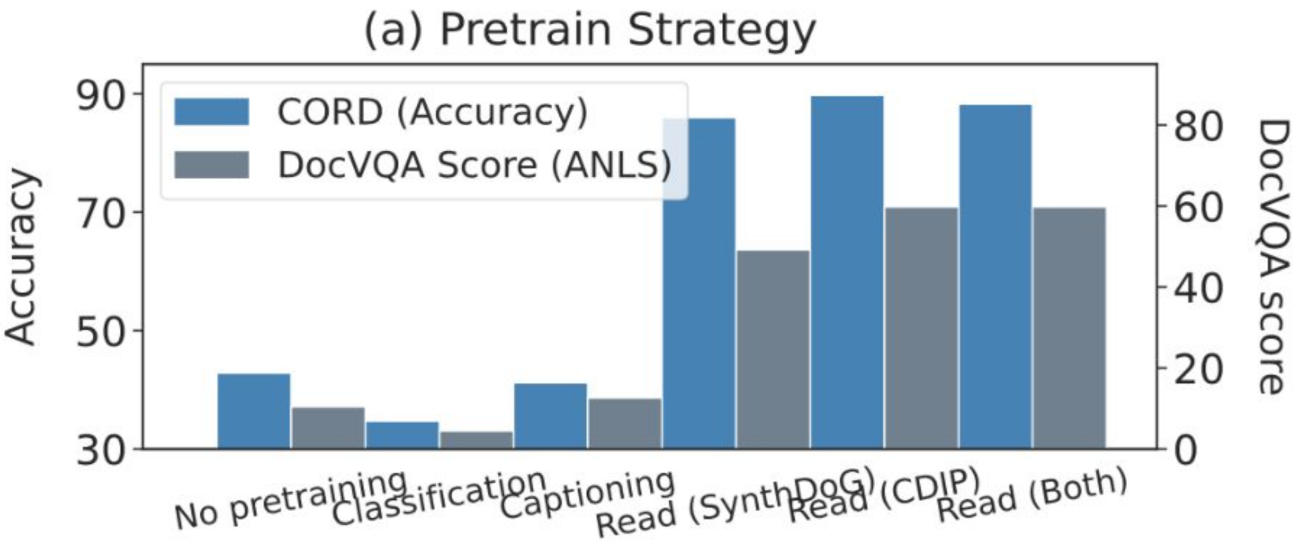
Slide

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.

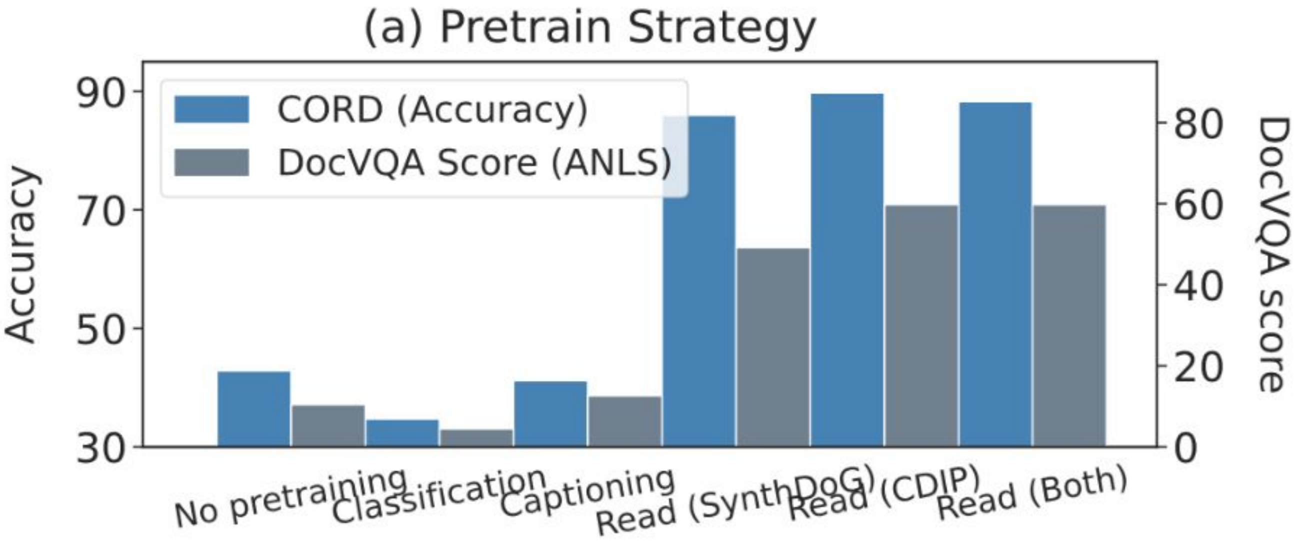
Analysis: Pre-training Strategies


(a) Pretrain Strategy

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

Slide

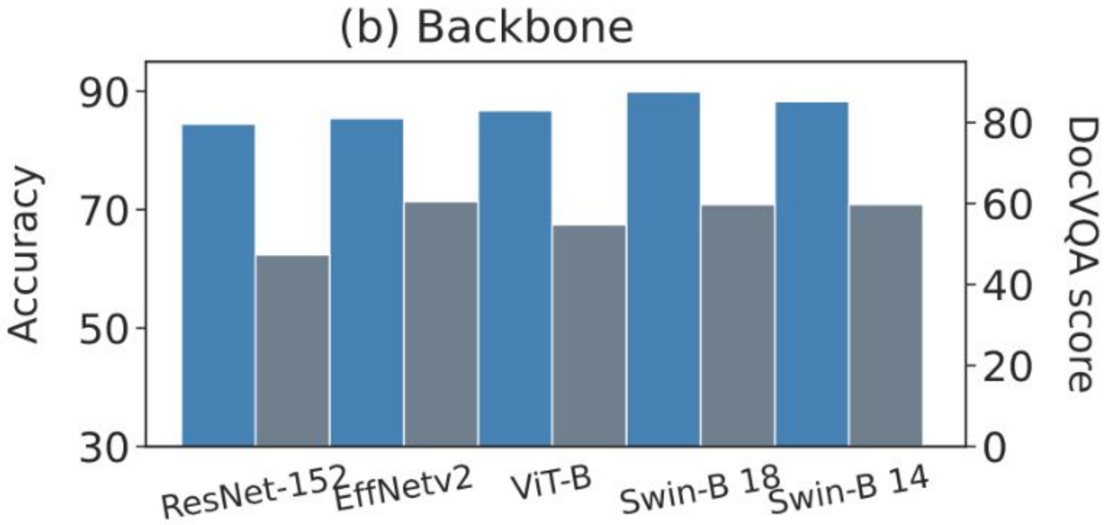
Analysis: Pre-training Strategies



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

Slide

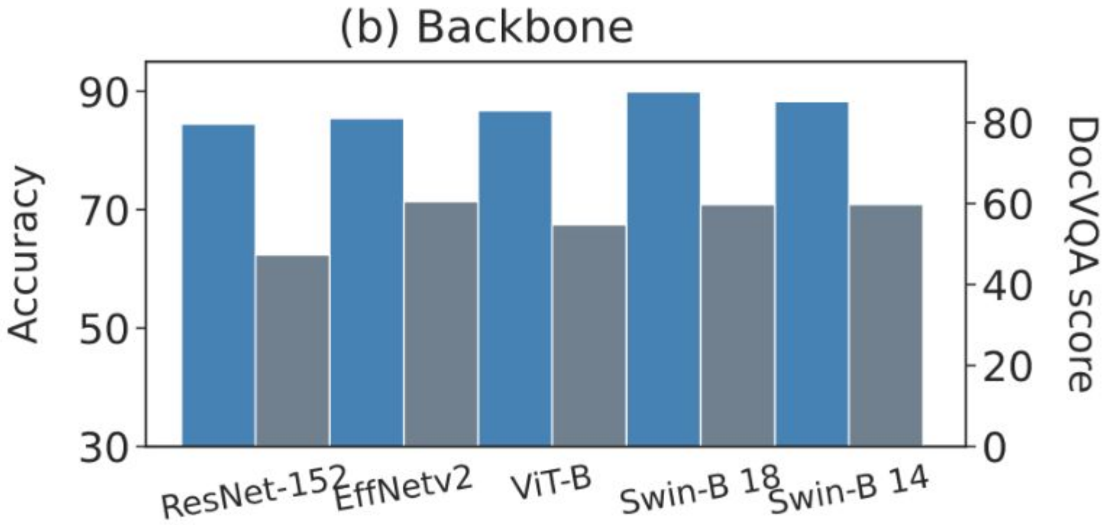
Analysis: Image Backbones



Next, we study popular image classification backbones.

Slide

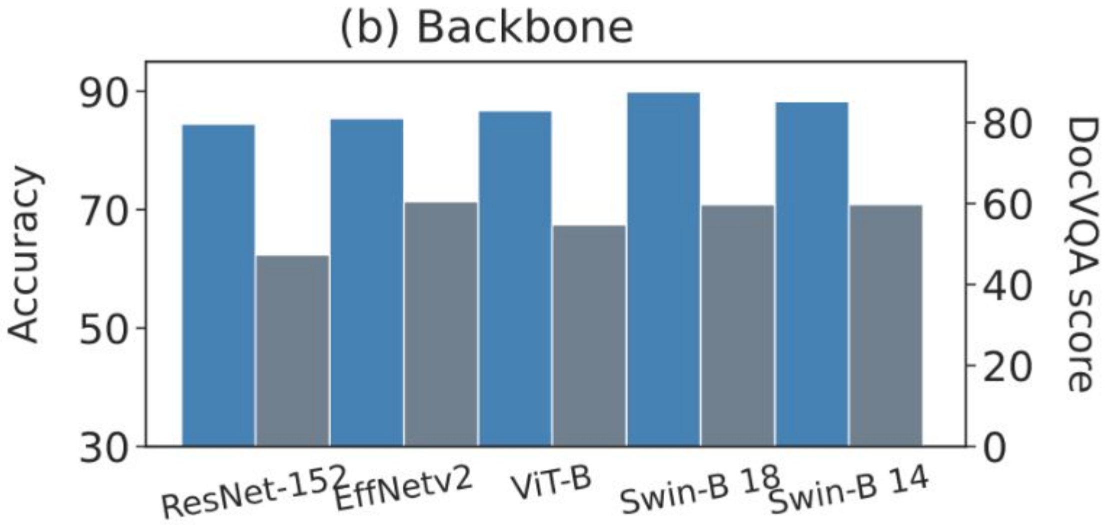
Analysis: Image Backbones



Overall, EfficientNetV2 and Swin Transformer outperform the others.

Slide

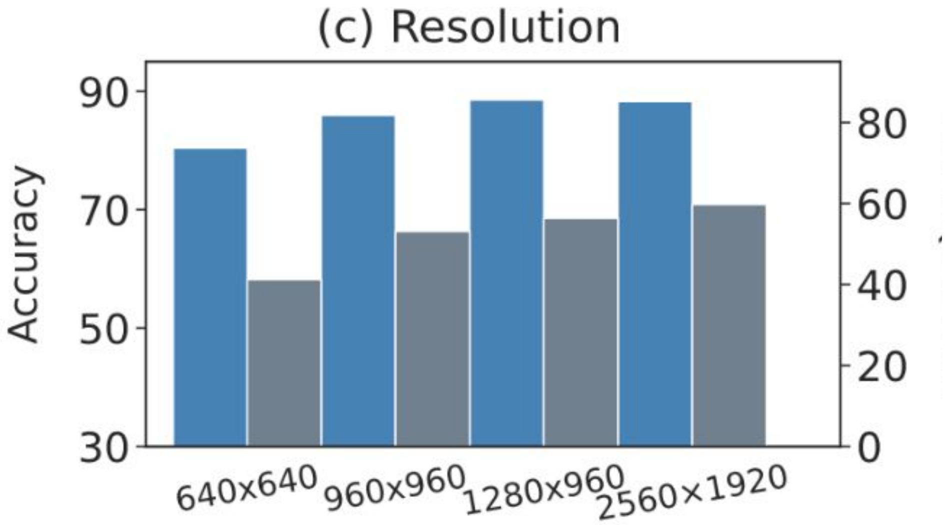
Analysis: Image Backbones



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

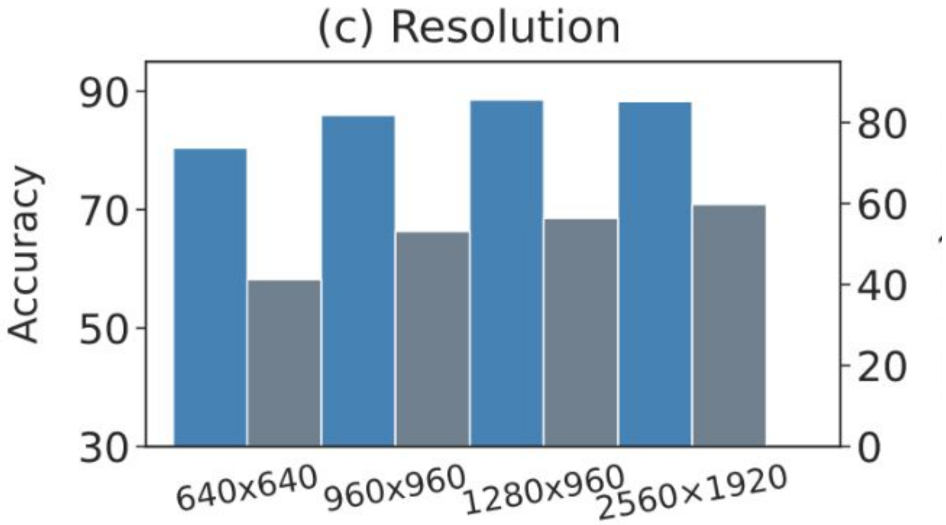
Slide

Analysis: Input Resolution



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

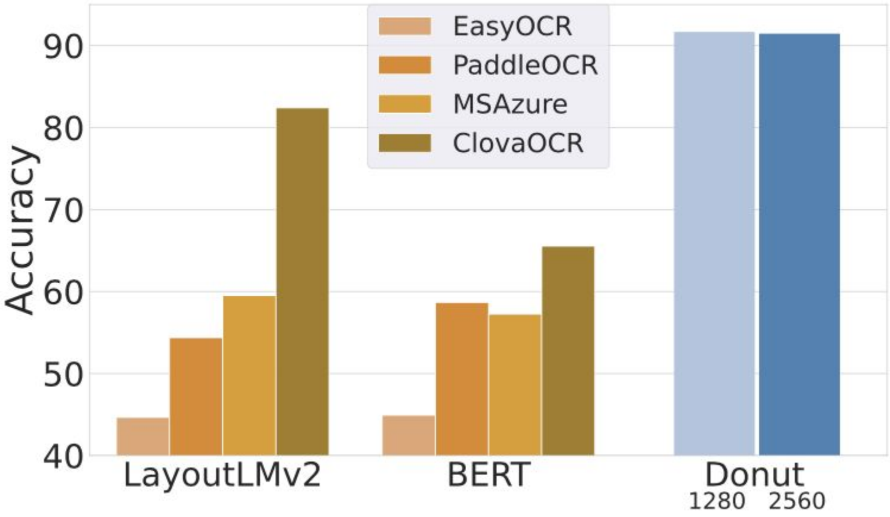
Analysis: Input Resolution

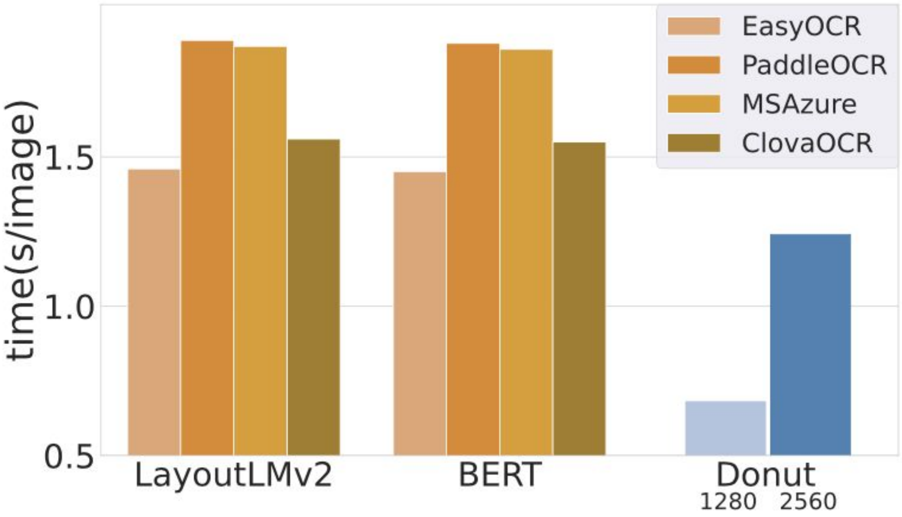


This gets clearer in the DocVQA where the images are larger with many tiny texts.

Slide

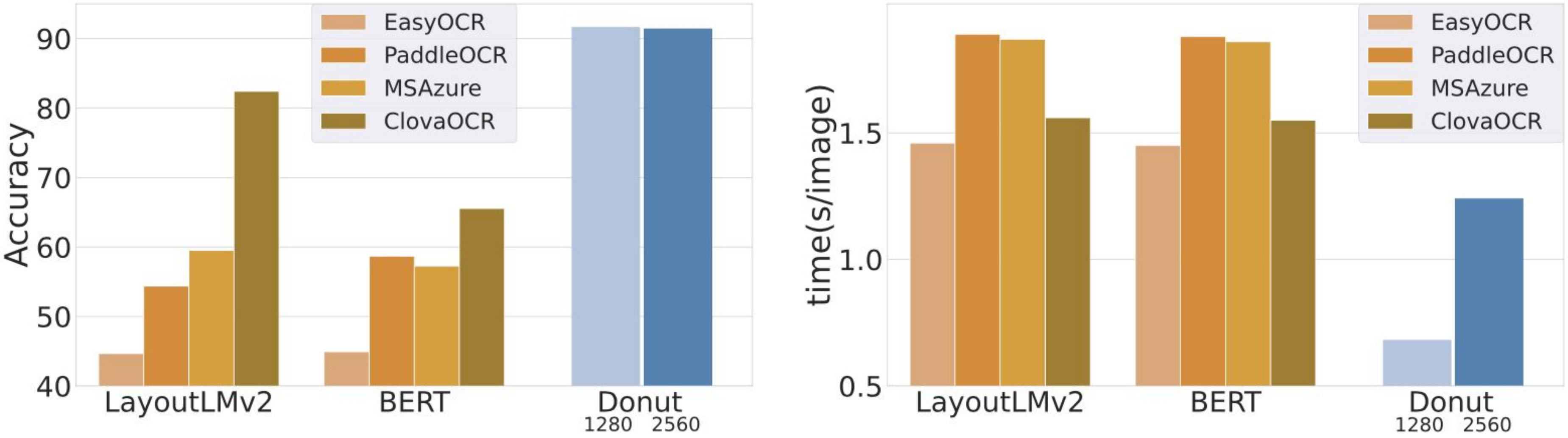
Analysis: OCR Engines





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

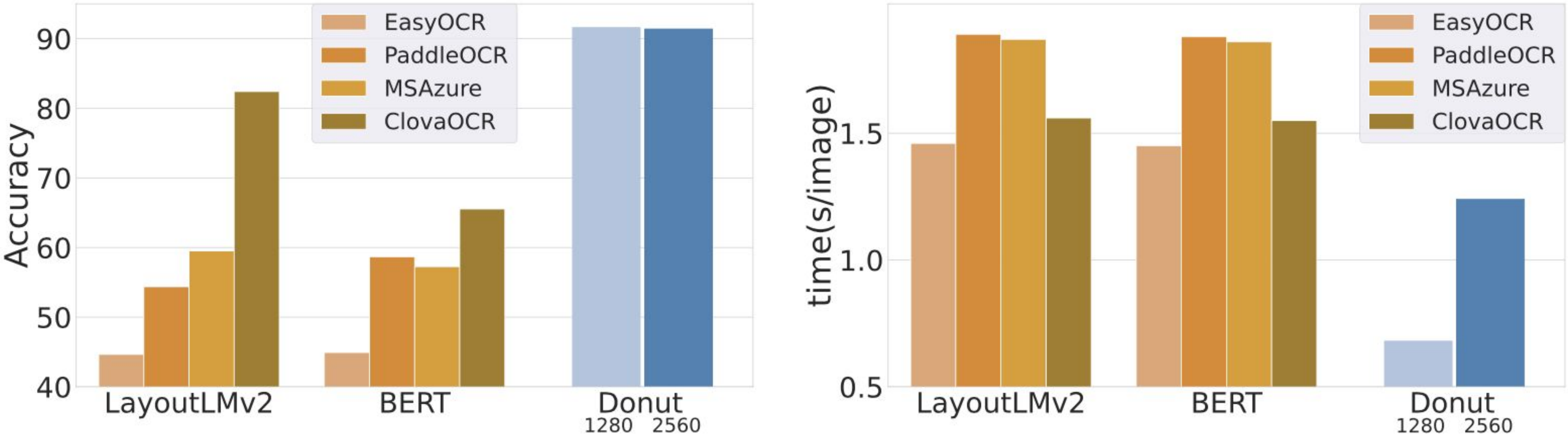
Analysis: OCR Engines



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

Slide

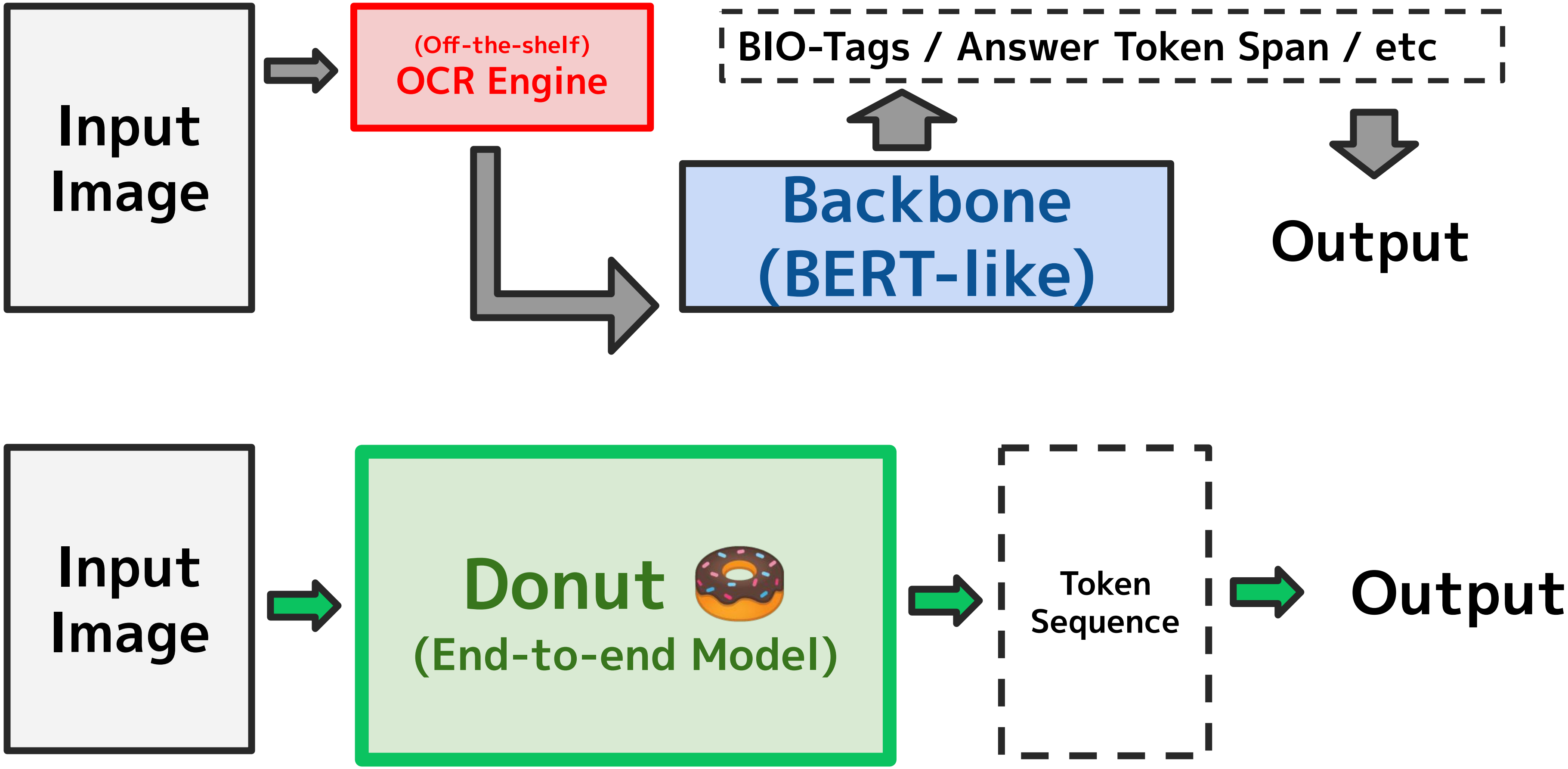
Analysis: OCR Engines



This shows why we need an OCR-free method.

Slide

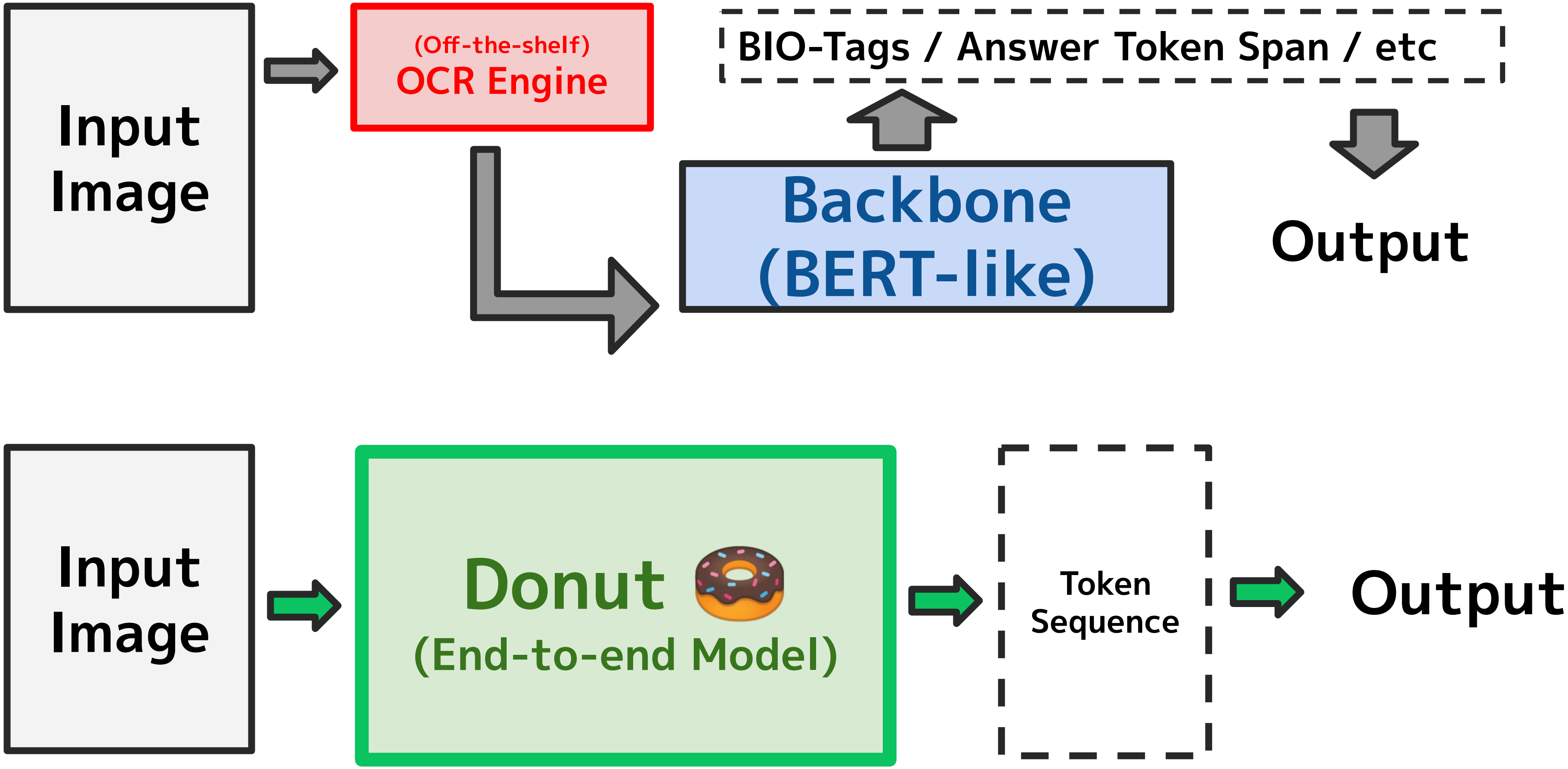
Conclusions



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

Slide

Conclusions

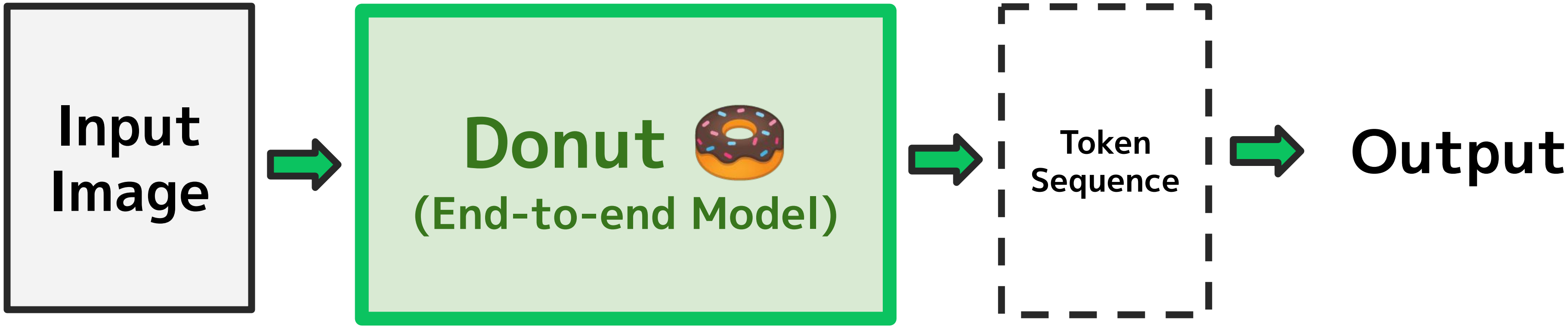


So far, we have introduced our new OCR-free method, Donut.
More experiments and analysis can be found in the manuscript.

Slide

Conclusions

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

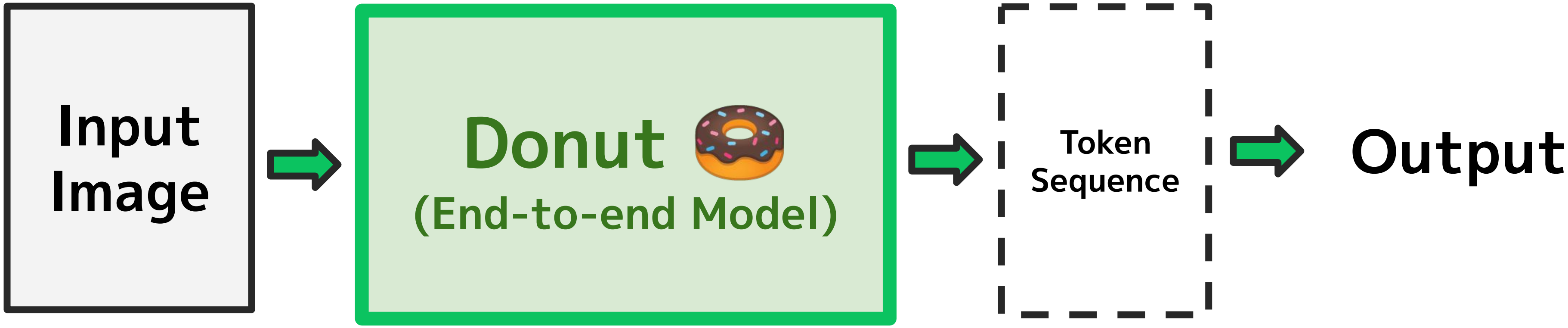


The proposed method, Donut, directly maps an input document image into a desired structured output.

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.

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.

Slide

Thank you!

Contact: gwkim.rsrch@gmail.com

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