Document Understanding in the Age of AI

Kshitij Agrawal

28 June 2022

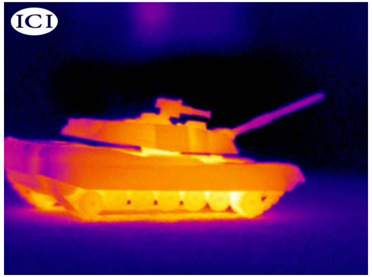
About Me

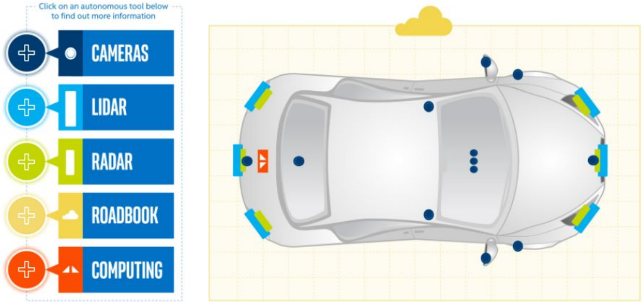
Seasoned computer vision scientist

Worked across applications in

- Document understanding
- Autonomous Driving
- In game advertising
- Defence Imaging

Multiple publications in conferences





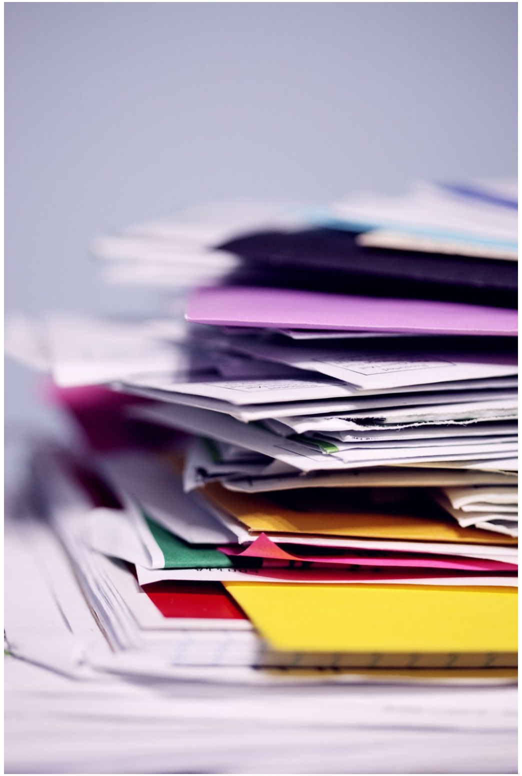
Overview

1. Need for Document Understanding
2. Stages of Document Processing
3. Text Extraction - In detail
4. Doc classification - In detail
5. Information Extraction - In detail
6. Conclusion

Why do we need Document Understanding

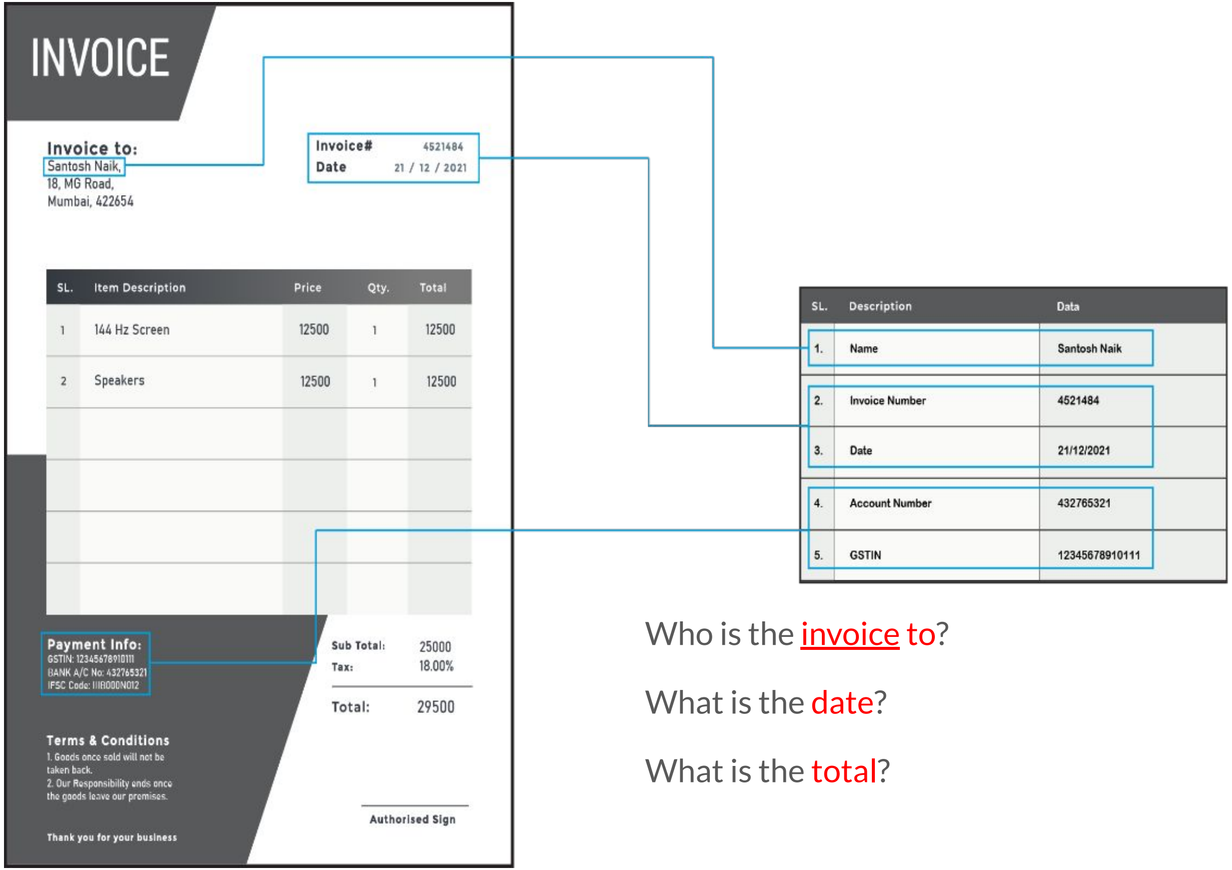
Documents are part of our everyday way of interacting and transacting with the world

- Many types of docs - Forms, texts, handwritten, certificates, legal documents
- Multiple formats - images, pdfs, raw text, books
- Varied content - tables, images, stylized text
- Challenges of lighting, shape, size and orientation



Information in Documents

What are the titles of the books?



INVOICE

Invoice to:
Santosh Naik,
18, MG Road,
Mumbai, 422654

Invoice# 4521484
Date 21 / 12 / 2021

SL.	Item Description	Price	Qty.	Total
1	144 Hz Screen	12500	1	12500
2	Speakers	12500	1	12500

Payment Info:
GSTIN: 12345678910111
BANK A/C No: 432765321
IFSC Code: IIIB000N012

Sub Total: 25000
Tax: 18.00%
Total: 29500

Terms & Conditions
1. Goods once sold will not be taken back.
2. Our Responsibility ends once the goods leave our premises.

Authorised Sign

Thank you for your business

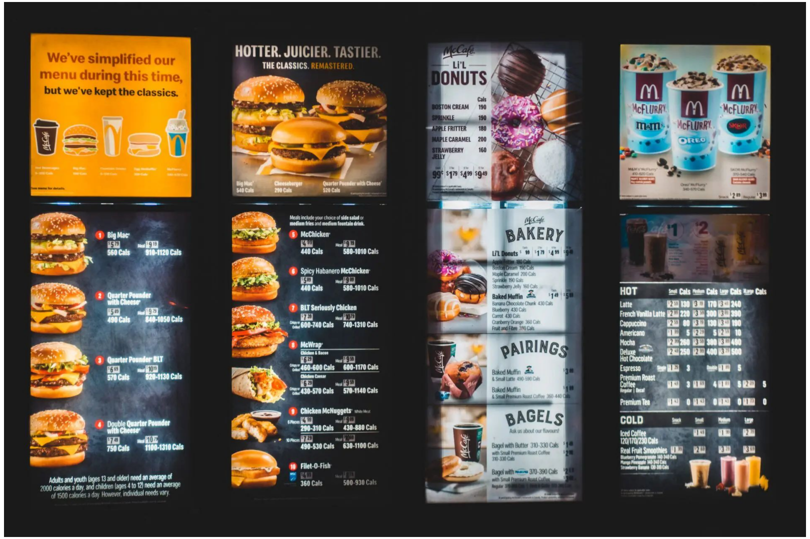
SL.	Description	Data
1.	Name	Santosh Naik
2.	Invoice Number	4521484
3.	Date	21/12/2021
4.	Account Number	432765321
5.	GSTIN	12345678910111

Who is the invoice to?

What is the date?

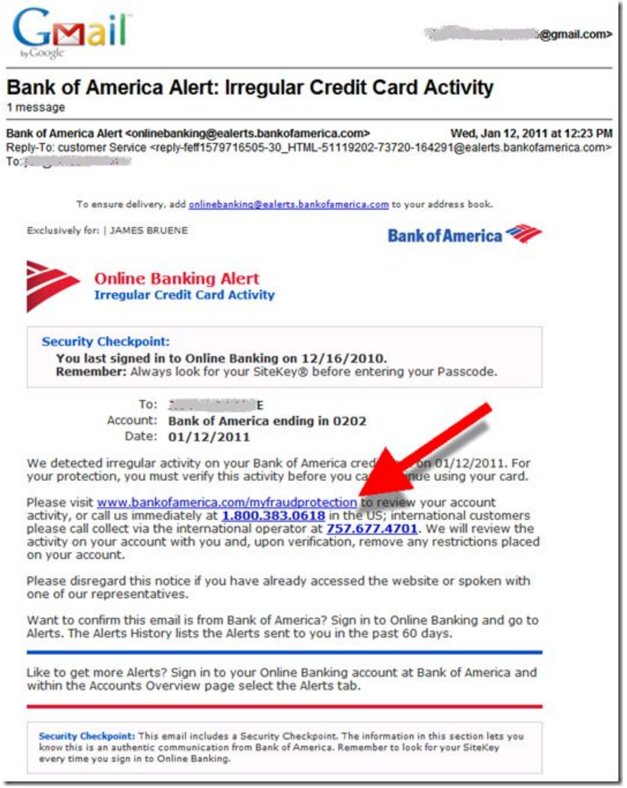
What is the total?

Information in Documents



Which restaurant's menu?

What is the price of a burger? ice cream?



Gmail by Google [redacted]@gmail.com>

Bank of America Alert: Irregular Credit Card Activity

1 message

Bank of America Alert <onlinebanking@ealerts.bankofamerica.com> Wed, Jan 12, 2011 at 12:23 PM
Reply-To: customer Service <reply-feff1579716505-30_HTML-51119202-73720-164291@ealerts.bankofamerica.com>
To: [redacted]

To ensure delivery, add onlinebanking@ealerts.bankofamerica.com to your address book.

Exclusively for: | JAMES BRUENE

Bank of America

Online Banking Alert
Irregular Credit Card Activity

Security Checkpoint:
You last signed in to Online Banking on 12/16/2010.
Remember: Always look for your SiteKey® before entering your Passcode.

To: [redacted]E
Account: **Bank of America ending in 0202**
Date: **01/12/2011**

We detected irregular activity on your Bank of America cred[...] on 01/12/2011. For your protection, you must verify this activity before you ca[...]nue using your card.

Please visit www.bankofamerica.com/myfraudprotection to review your account activity, or call us immediately at 1.800.383.0618 in the US; international customers please call collect via the international operator at 757.677.4701. We will review the activity on your account with you and, upon verification, remove any restrictions placed on your account.

Please disregard this notice if you have already accessed the website or spoken with one of our representatives.

Want to confirm this email is from Bank of America? Sign in to Online Banking and go to Alerts. The Alerts History lists the Alerts sent to you in the past 60 days.

Like to get more Alerts? Sign in to your Online Banking account at Bank of America and within the Accounts Overview page select the Alerts tab.

Security Checkpoint: This email includes a Security Checkpoint. The information in this section lets you know this is an authentic communication from Bank of America. Remember to look for your SiteKey every time you sign in to Online Banking.

What is the content of the email?

What is the transaction amount?

Is it fraud/alert? Should the account be blocked?

If we can extract information, we can search and automate processes

Journey of Document Digitization

Document is input to the system

Operator views and enters fields manually

Another agent verifies and approves

Digitized information stored in database for later use

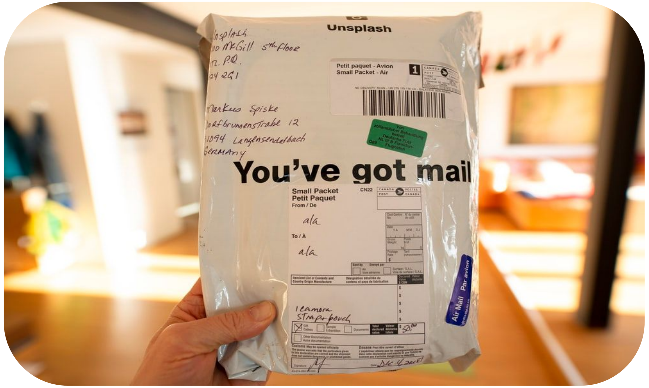
1. Time consuming
2. Error prone - if a business decision relies on it, can lead to mishaps!
3. High amount of variability -
   a. Varied formats - pdfs, jpegs, pngs etc
   b. Varied layouts - single column, multi column, tabular, forms
   c. Multiple pages

Applications

If we can ***automate*** information extraction then we can automate business processes.

1. Banking, Insurance,credit card
   Customer verification, Loan document processing, email classification

2. Ecommerce and logistics
   Delivery verification, shipment processing, payment processing
   invoice processing



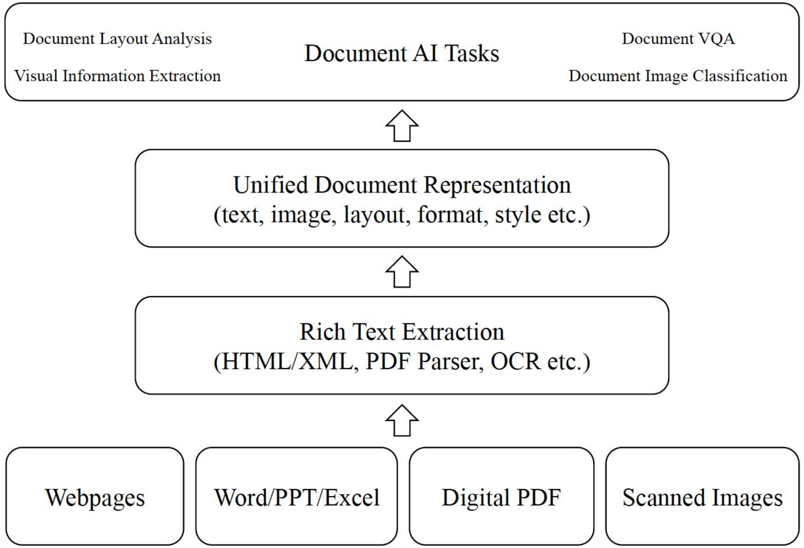
Document Processing Stages

Document understanding

Visual understanding / Text understanding / High level embeddings

Text Extraction

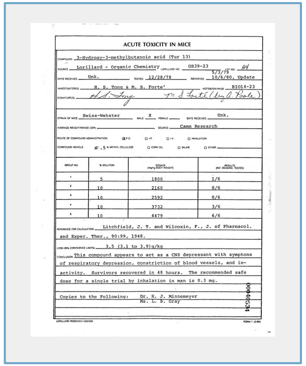
Input doc



Document Understanding Tasks

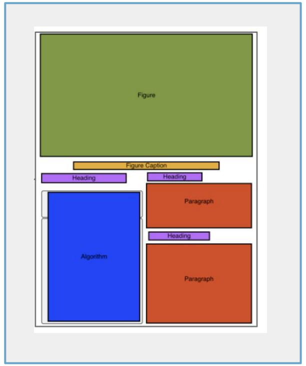
Information Extraction

Image Classification


ACUTE TOXICITY IN MICE

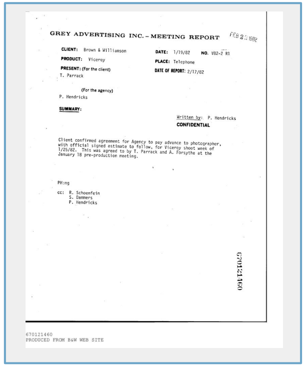
```
{
  "type": "form",
}
```

Layout Understanding



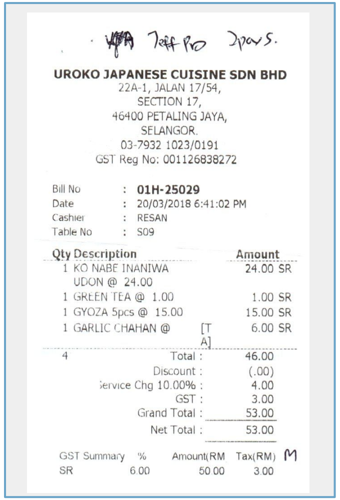
Layout segmentation
Table Detection
Image Detection

Form Extraction


GREY ADVERTISING INC. – MEETING REPORT

```
{
  "company": "Brown & Williamson",
  "date": "19/01/1982",
  "product": "viceroy"
}
```

Receipt Extraction


UROKO JAPANESE CUISINE SDN BHD
22A-1, JALAN 17/54,
SECTION 17,
46400 PETALING JAYA,
SELANGOR.
03-7932 1023/0191
GST Reg No: 001126838272

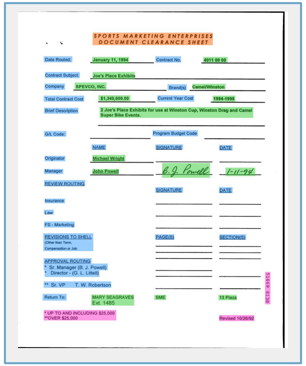
Bill No	:	01H-25029
Date	:	20/03/2018 6:41:02 PM
Cashier	:	RESAN
Table No	:	S09

Qty	Description	Amount
1	KO NABE INANIWA UDON @ 24.00	24.00 SR
1	GREEN TEA @ 1.00	1.00 SR
1	GYOZA 5pcs @ 15.00	15.00 SR
1	GARLIC CHAHAN @ [TA]	6.00 SR
4	Total :	46.00
	Discount :	(.00)
	Service Chg 10.00% :	4.00
	GST :	3.00
	Grand Total :	53.00
	Net Total :	53.00

GST Summary	%	Amount(RM)	Tax(RM)
SR	6.00	50.00	3.00

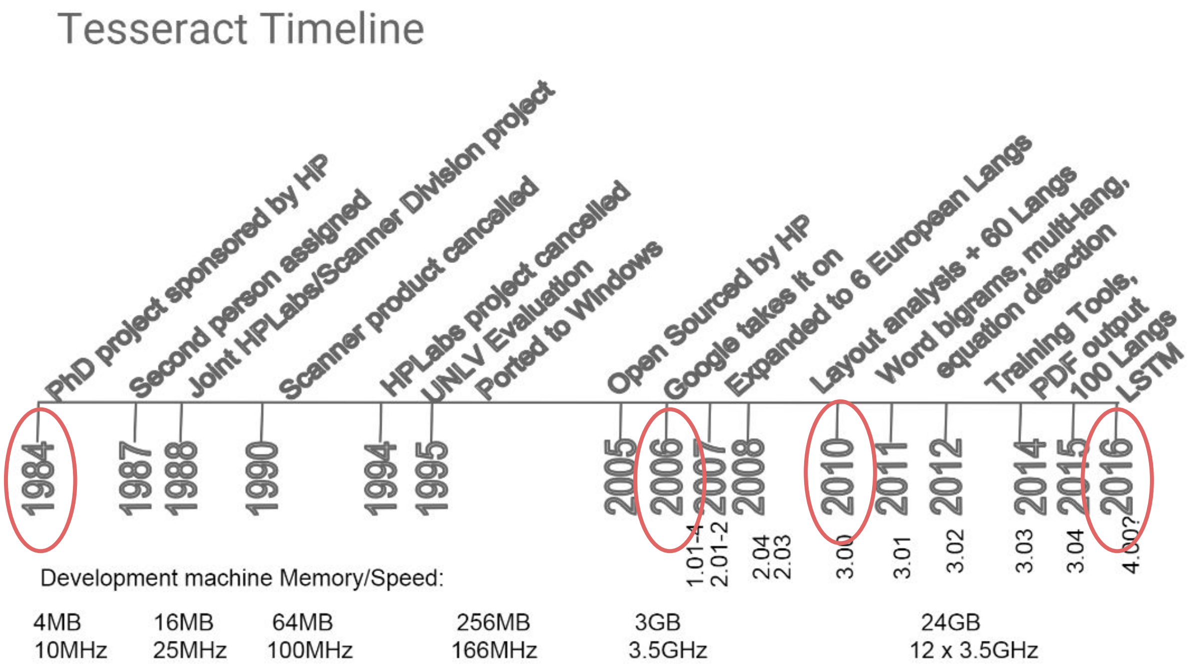
```
{
  "company": "Uroko Japanese Cusine SDN BHD",
  "date": "20/03/2018",
  "total": "53.00"
}
```

Form Understanding


SPORTS MARKETING ENTERPRISES DOCUMENT CLEARANCE SHEET

Entity detection
Entity Linking
Visual Question Answering

From Images to Text - Optical Character Recognition



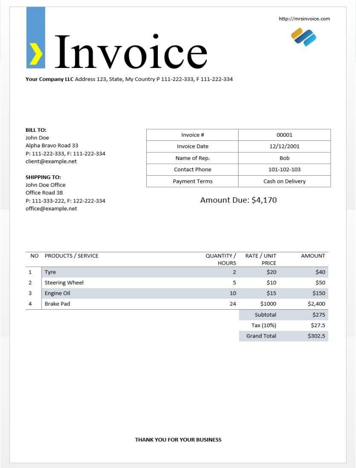
source: [Tesseract docs]

OCR - Process

Tesseract 4.0

- 100+ languages supported
- Full layout analysis
- Table detection
- Equation detection
- Better (Deeper) language models
- Improved segmentation search
- Word bigrams
- Training tools for custom datasets

```
apt install tesseract-ocr
```



http://mrsinvoice.com

Invoice

Your Company LLC Address 123, State, My Country P 111-222-333, F 111-222-334

BILL TO:
John Doe
Alpha Bravo Road 33
P: 111-222-333, F: 111-222-334
client@example.net

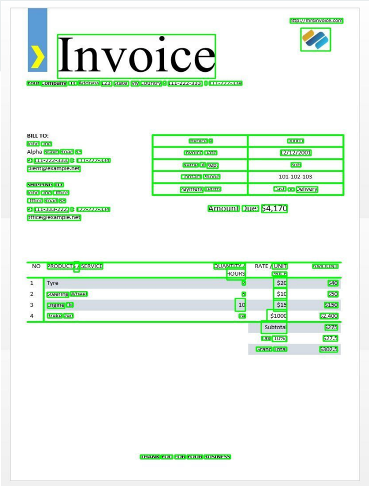
SHIPPING TO:
John Doe Office
Office Road 38
P: 111-333-222, F: 122-222-334
office@example.net

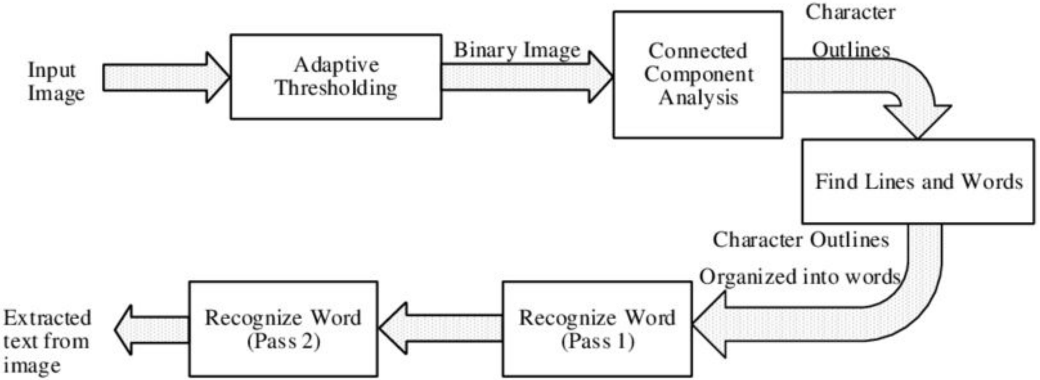
Invoice #	00001
Invoice Date	12/12/2001
Name of Rep.	Bob
Contact Phone	101-102-103
Payment Terms	Cash on Delivery

Amount Due: $4,170

NO	PRODUCTS / SERVICE	QUANTITY / HOURS	RATE / UNIT PRICE	AMOUNT
1	Tyre	2	$20	$40
2	Steering Wheel	5	$10	$50
3	Engine Oil	10	$15	$150
4	Brake Pad	24	$1000	$2,400
			Subtotal	$275
			Tax (10%)	$27.5
			Grand Total	$302.5

THANK YOU FOR YOUR BUSINESS

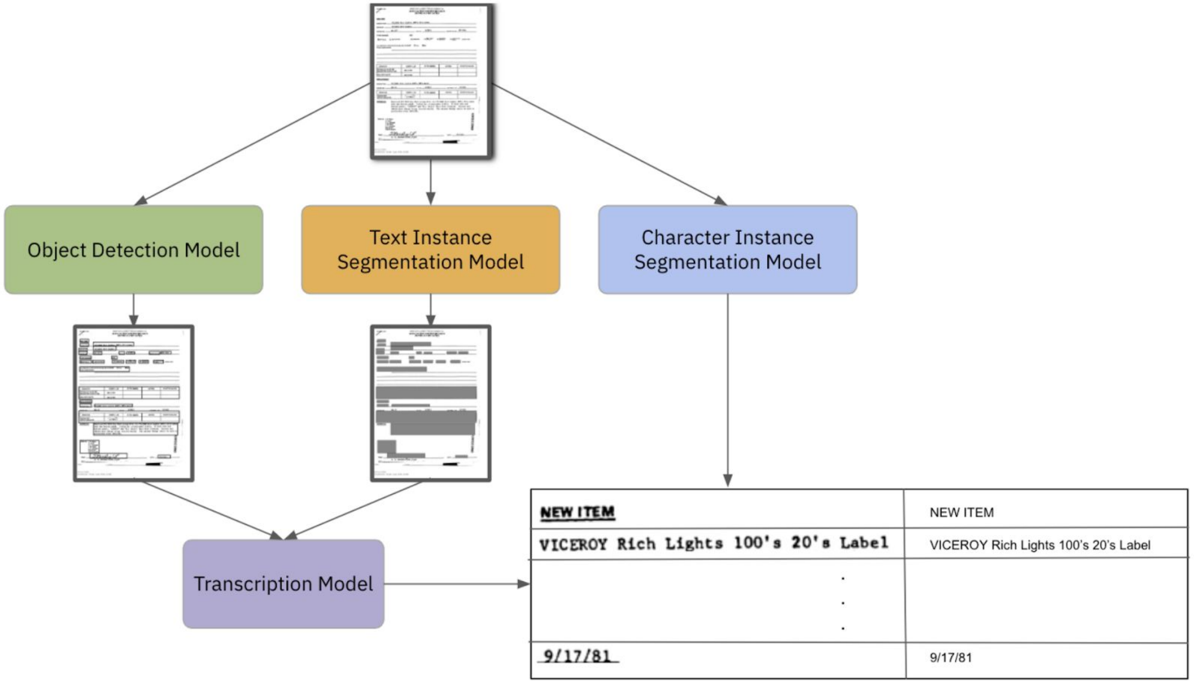




source: [Architecture of Tesseract]

Optical Character Recognition - Modern architectures

1. Detection Based
    a. EAST
    b. Faster RCNN/SSD
2. Text Instance Segmentation Model
    a. Borrowed from segmentation - FCN
3. Character Instance Segmentation Model

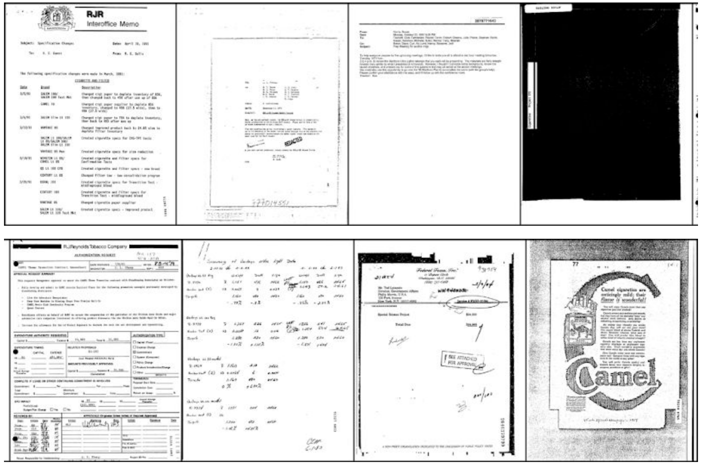


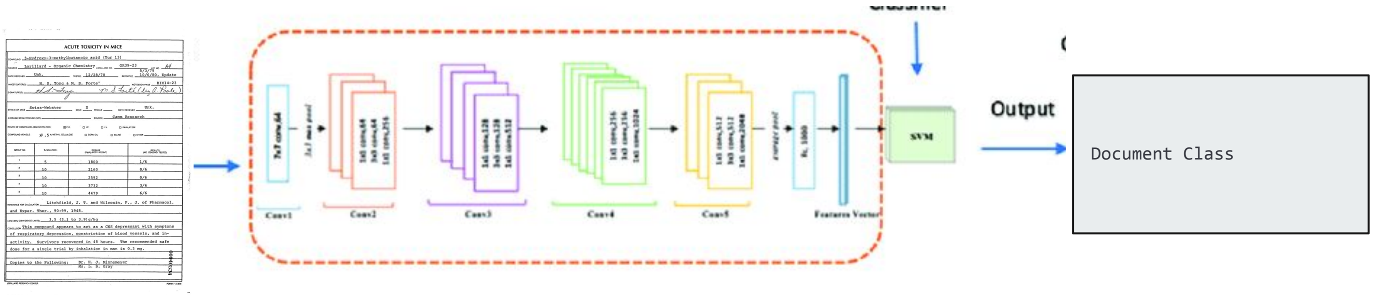
Document Image Classification

1. Dataset - RVLCPID
    a. 400k images
    b. 16 categories - letter, email, form, handwritten, invoice, advertisement etc.

2. Classification network - Resnet50

3. Accuracy - 90%





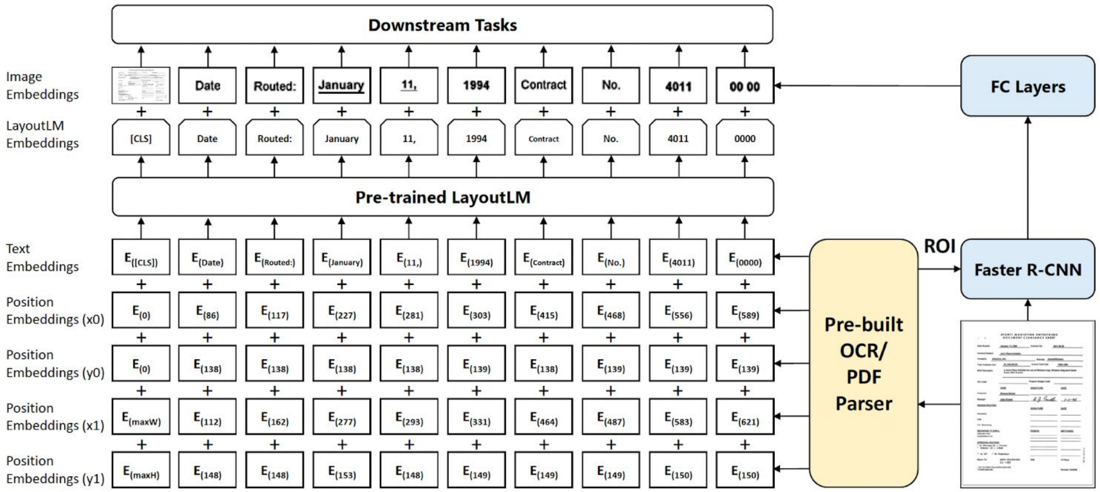
Vision + Language Models

Can we leverage vision and text to improve accuracy ?

1. Vision based image features with language models - LayoutLM

2. Leverages the design of BERT models

3. Document classification Accuracy - 94.5%



"LayoutLM: Pre-training of Text and Layout for Document Image Understanding", Y. Xu et al, KDD2020

LayoutLM

1. First time that text and layout are jointly learned in a single network for document level pre-training

2. Two types of input embeddings in BERT Model :
   a. 2D position embedding that denotes the relative position of a token within a document
   b. image embedding for scanned token text within a document

3. Architecture
   a. Multi-task learning objective -
      Masked Visual-Language Model (MVLM) loss and
      a Multi-label Document Classification (MDC) loss

Transformers ++ : Word + position embedding

Tasks

- form understanding (from 70.72 to 79.27),
- receipt understanding (from 94.02 to 95.24) and
- document image classification (from 93.07 to 94.42).

Where to get started?

Information extraction - track finances from purchase receipts

Tesseract + LayoutLM

Result - Key value pairs of the extracted information

SROIE dataset

Information extraction from receipts

OCR text with bounding boxes

Four fields for extraction - company, date, address, total

Extracted Information


```
{
  "company": "STARBUCKS STORE #10208",
  "address": "11302 EUCLID AVENUE, CLEVELAND, OH (216) 229-0749",
  "date": "14/03/2015",
  "total": "4.95"
}
```


Notebook: https://github.com/NielsRogge/Transformers-Tutorials/blob/master/LayoutLM/Fine_tuning_LayoutLMForSequenceClassification_on_RVL_CDIP.ipynb

Open Research Areas in Document Understanding

Self supervised learning from large scale data

Few shot learning from scarce data

Improving image capture and preprocessing

Dealing with noise in data

Fin.

Questions?