OCR and Document Understanding

CHA ZHANG

MICROSOFT CLOUD & AI

Microsoft

Azure Cognitive Services

Pre-Trained and Customizable models with your data

Vision

Identify and analyze content within images, videos, and digital ink

Speech

Integrate speech processing into apps and services

Language

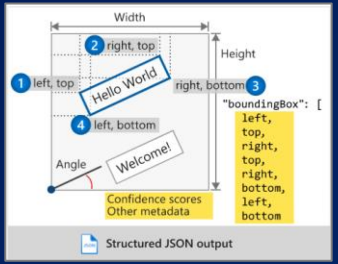
Extract meaning from unstructured text

Decision

Make smarter decisions faster

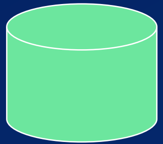
As Part of Cognitive Service:

OCR (Read API)

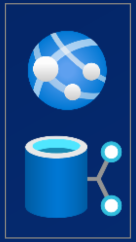


OCR includes:

- Pages
- Text lines
- Words
- Locations

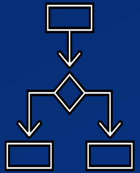


Unlocked text



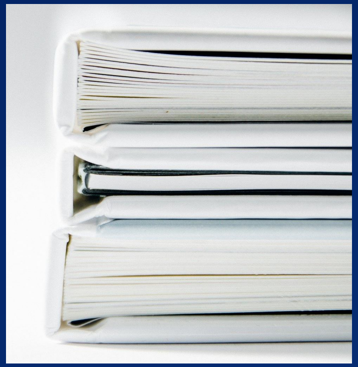
Customers and partners add processing to get intelligent insights

ENABLE



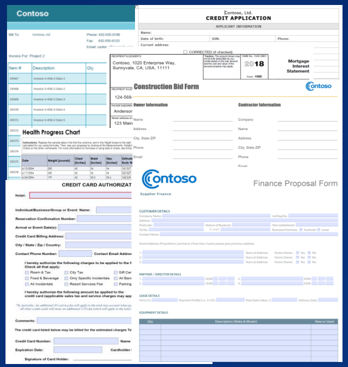
Process automation

Document Understanding (Form Recognizer APIs)



Layout

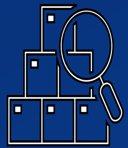
Pre-built



Custom

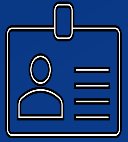
Form Recognizer includes:

- Text extraction
- Document structure
  - Tables
  - Selection marks
- Fields and values
- Other intelligence



Knowledge mining

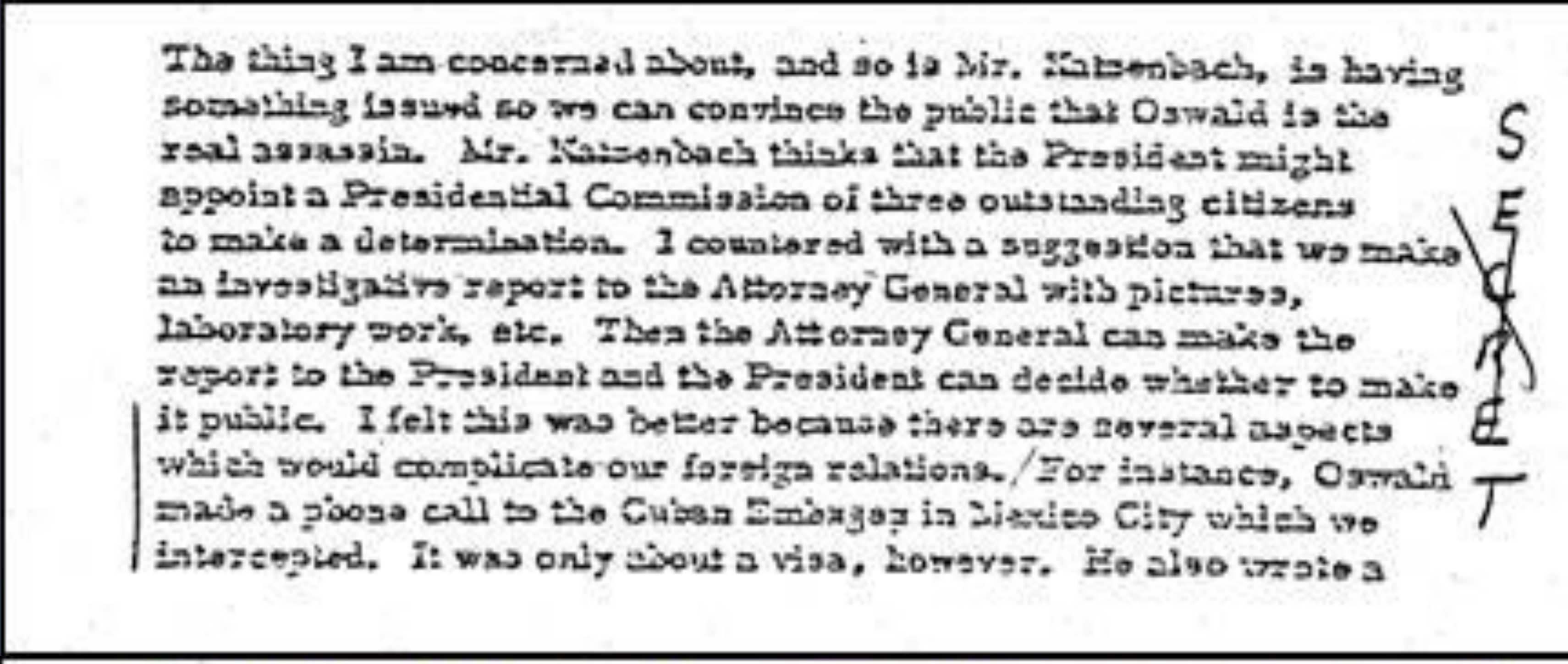
ENABLE



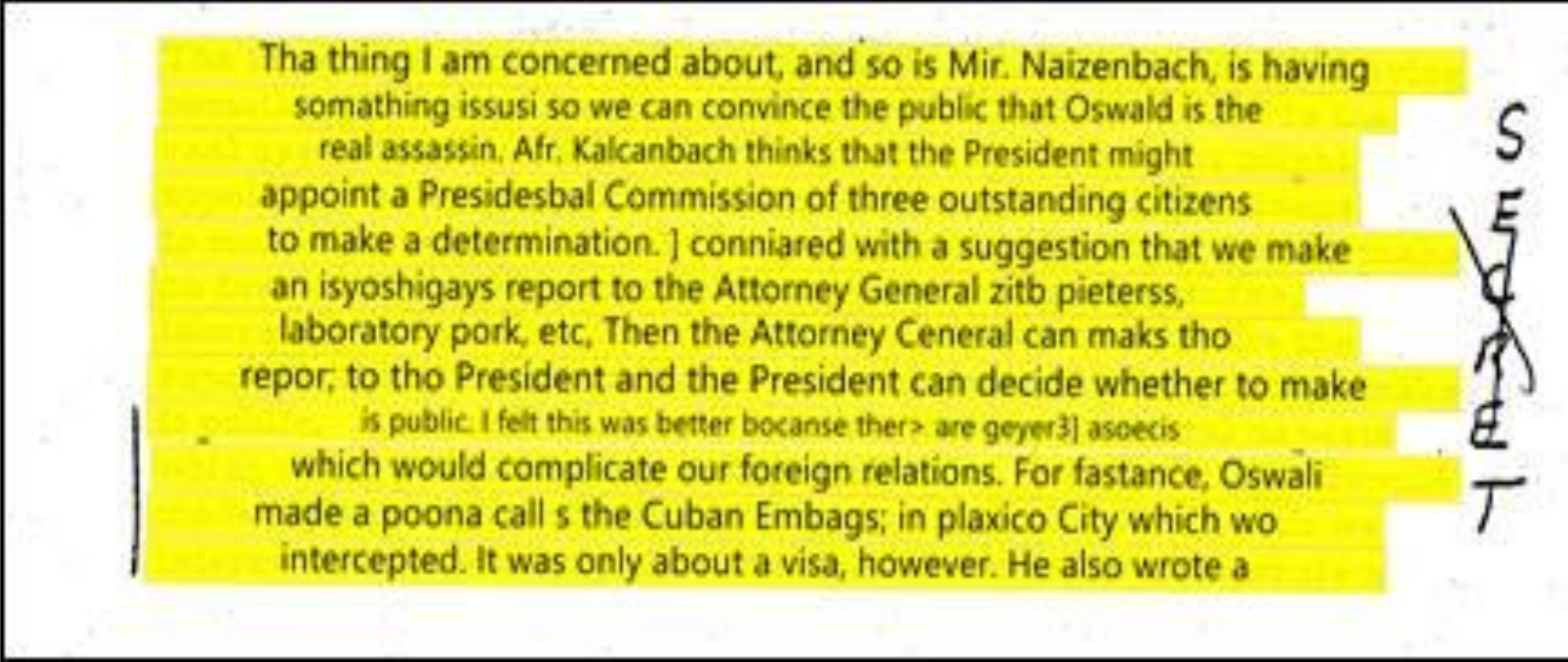
Industry specific applications

Microsoft

OCR
(Read API)


The thing I am concerned about, and so is Mr. Katzenbach, is having
something issued so we can convince the public that Oswald is the
real assassin. Mr. Katzenbach thinks that the President might
appoint a Presidential Commission of three outstanding citizens
to make a determination. I countered with a suggestion that we make
an investigative report to the Attorney General with pictures,
laboratory work, etc. Then the Attorney General can make the
report to the President and the President can decide whether to make
it public. I felt this was better because there are several aspects
which would complicate our foreign relations. For instance, Oswald
made a phone call to the Cuban Embassy in Mexico City which we
intercepted. It was only about a visa, however. He also wrote a

SECRET


Tha thing I am concerned about, and so is Mir. Naizenbach, is having
somathing issusi so we can convince the public that Oswald is the
real assassin. Afr. Kalcanbach thinks that the President might
appoint a Presidesbal Commission of three outstanding citizens
to make a determination.] conniared with a suggestion that we make
an isyoshigays report to the Attorney General zitb pieterss,
laboratory pork, etc, Then the Attorney Ceneral can maks tho
repor; to tho President and the President can decide whether to make
is public. I felt this was better bocanse ther> are geyer3] asoecis
which would complicate our foreign relations. For fastance, Oswali
made a poona call s the Cuban Embags; in plaxico City which wo
intercepted. It was only about a visa, however. He also wrote a

SECRET

Challenges for “Universal OCR”

Large scale variability

Large aspect ratio

Cannot be enclosed tightly by axis-aligned rectangles

- e.g., skewed/curved text-lines

Nearby small-size text-lines

- e.g., inter-line space could be less than 2 pixels

Complex/ambiguous layout

Text-like background

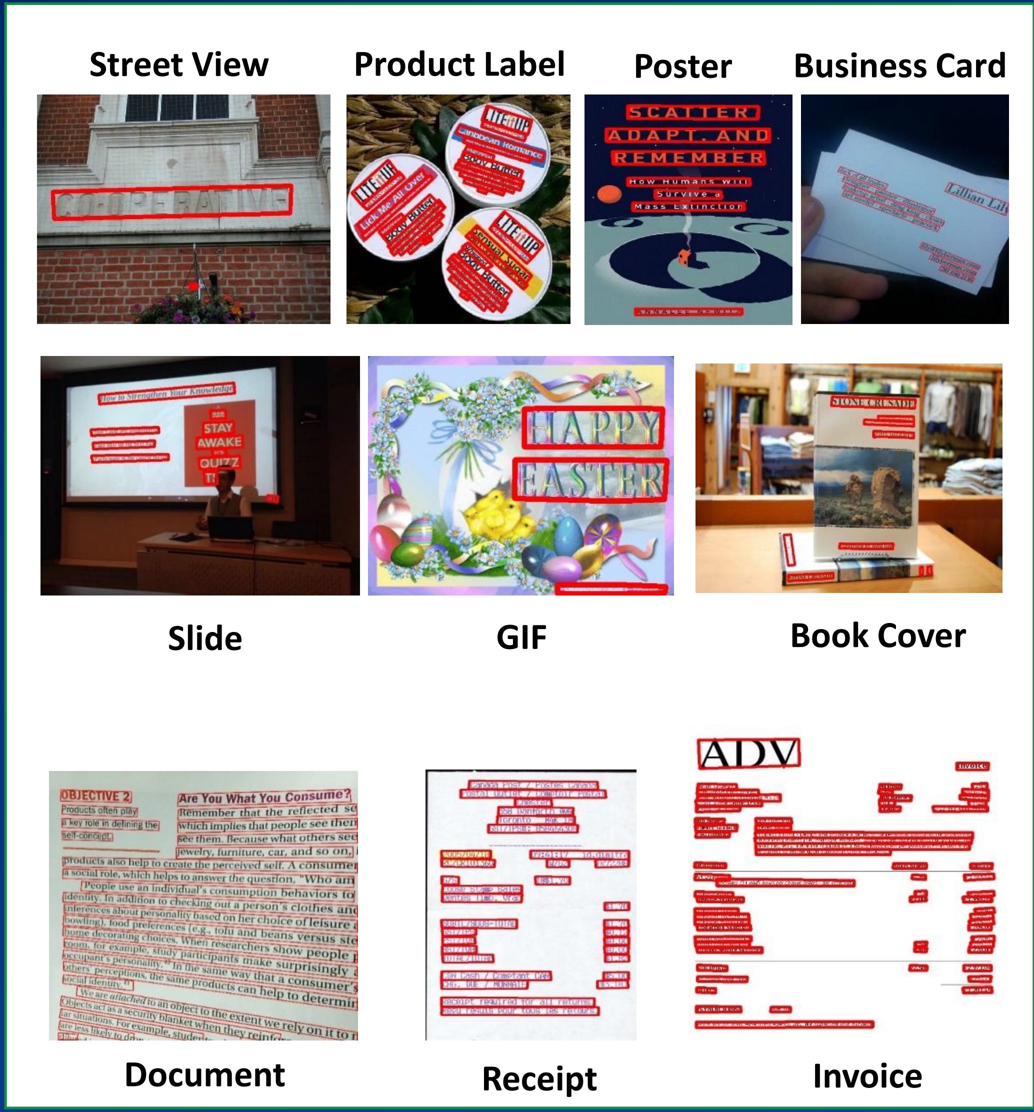
- e.g., fences, bricks, stripes

High localization accuracy is required for text recognition engine

- “IOU>0.5” criterion is far from enough

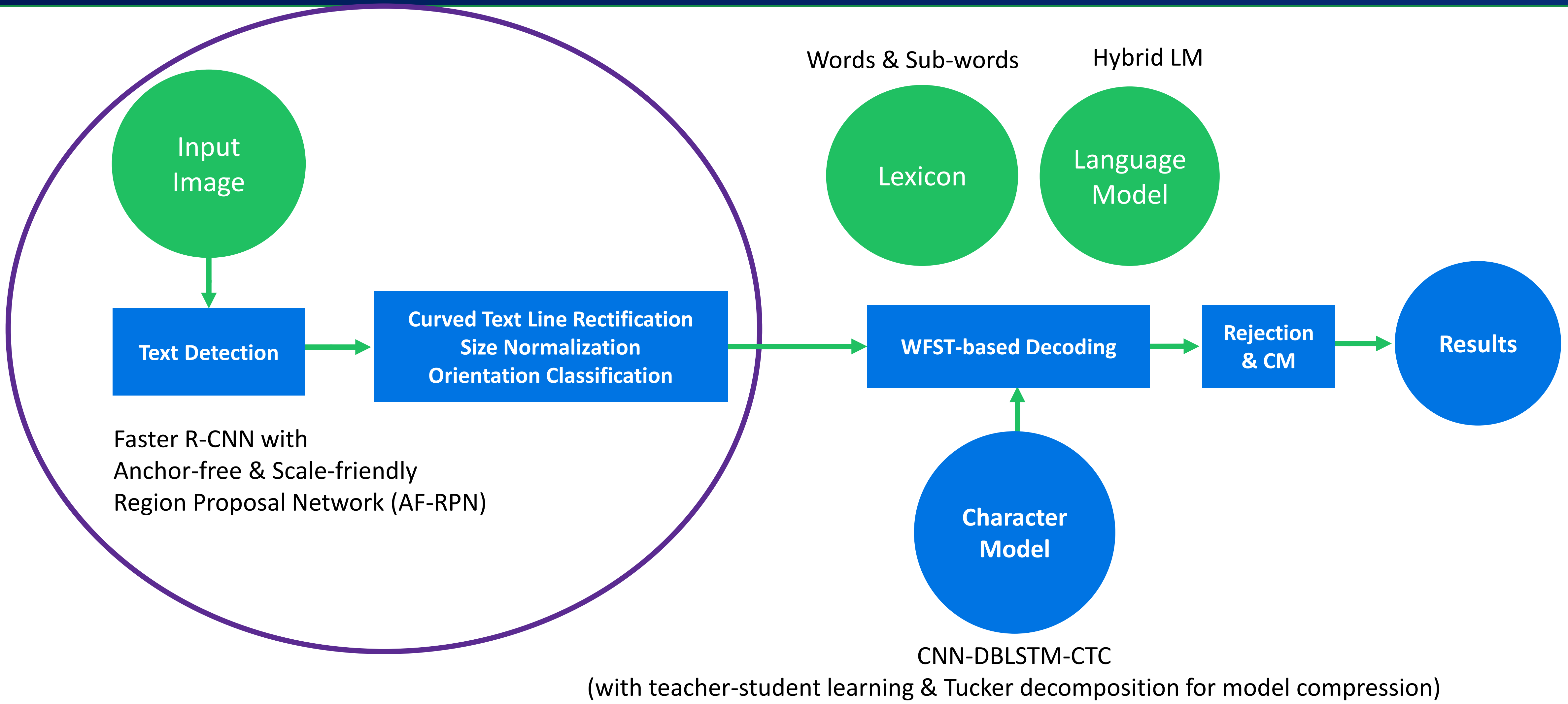
Resolutions of input images cannot be reduced aggressively

- to avoid excessive small text instances



Microsoft

Read API: Microsoft's New Generation OCR Engine



A **unified** engine to recognize mixed printed and handwritten text lines with arbitrary orientations (even flipped)

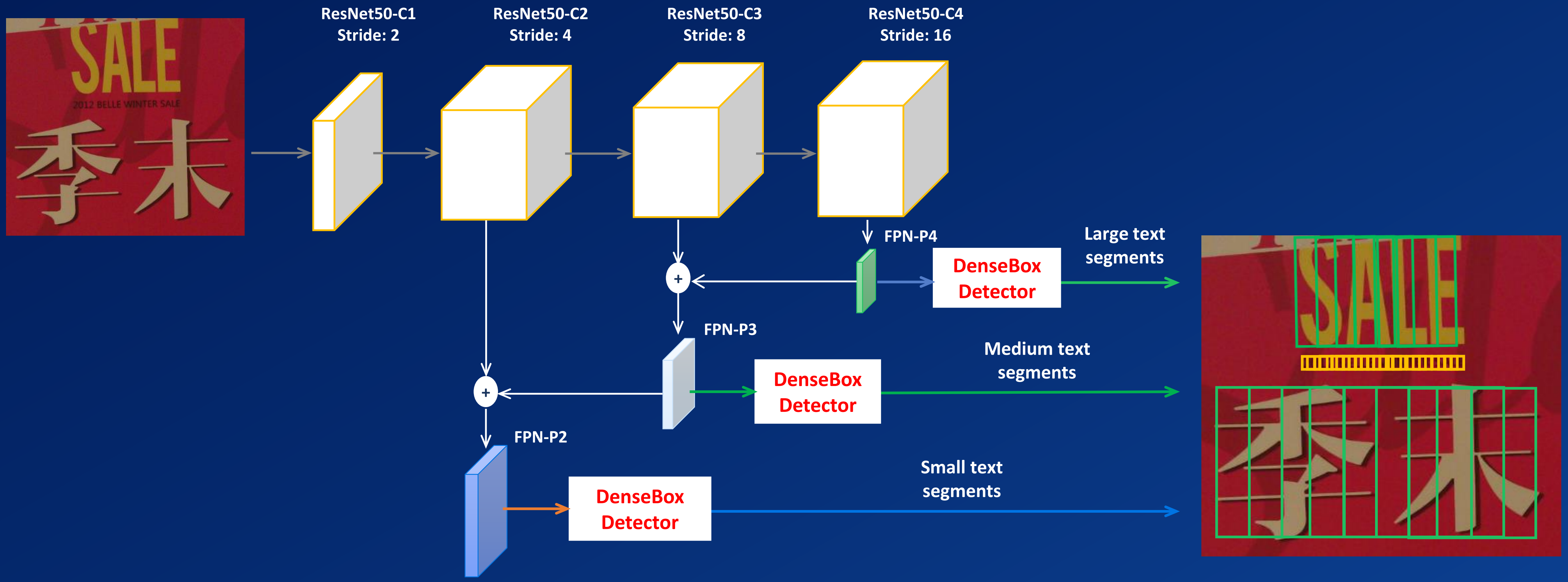
Our Text Detection Approach

Input Image → **Anchor-Free Region Proposal Network (AF-RPN)** [1] based Text Segment Detection → **Relation Network** based Line Grouping [2] → NMS → Text-lines

[1] JS1-1 "An Anchor-Free Region Proposal Network for Faster R-CNN based Text Detection Approaches," Zhuoyao Zhong, Lei Sun, Qiang Huo, ICDAR-2019 oral presentation of the IJDAR paper

[2] PS2-07 "A Relation Network Based Approach to Curved Text Detection," Chixiang Ma, Zhuoyao Zhong, Lei Sun, Qiang Huo, ICDAR-2019

Anchor-Free Region Proposal Network (AF-RPN)

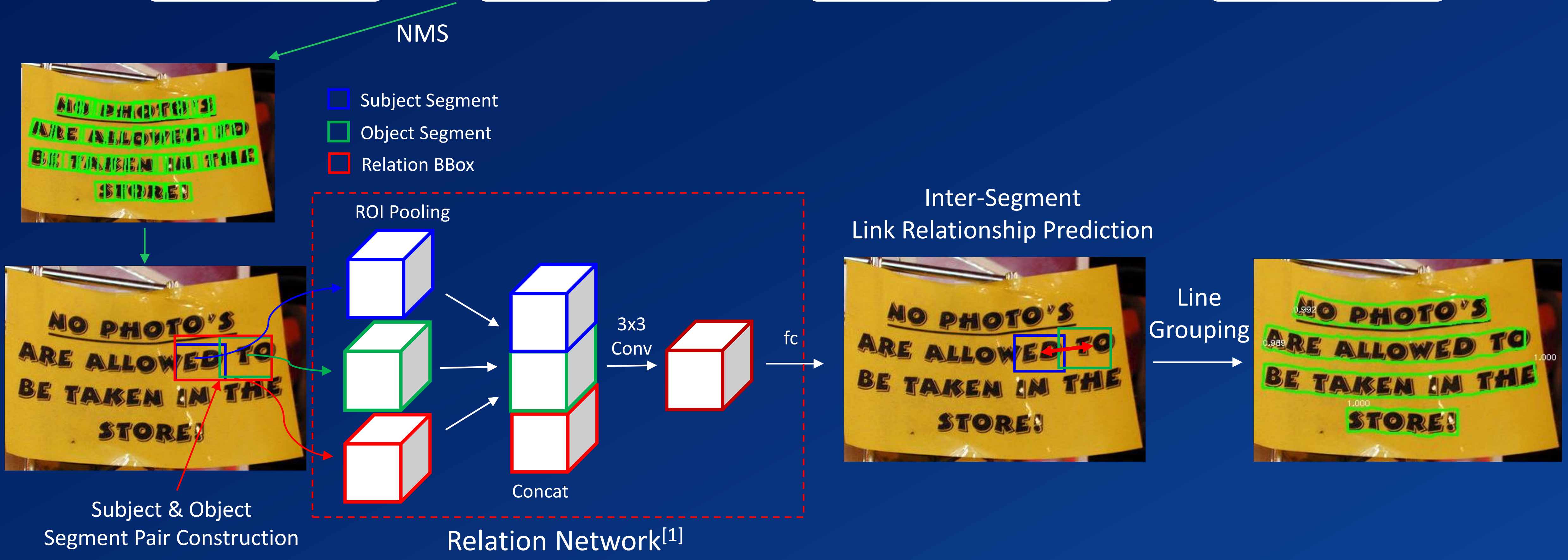


Scale-friendly learning: each DenseBox [1] only detects texts of scales within an appropriate range.

[1] L.-C. Huang, Y. Yang, Y.-F. Deng, and Y.-N. Yu, “Unifying landmark localization with end to end object detection,” arXiv, 2015.

Relation Network based Line Grouping

Input Image → Segment Detection with AF-RPN → **Relation Network** based Line Grouping → Text-Lines



[1] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, "Relationship proposal networks," in CVPR, 2017, pp. 5678-5686.

Advantages of Relation Network based Line Grouping

- Leverage the link between the pair of relatively distant segments
  - ➢ Able to detect text-lines with large inter-character spaces robustly
- Leverage wider context information to improve link prediction accuracy
  - ➢ More robust

Textness score map

Relation Network

SegLink *vs.* Relation Network

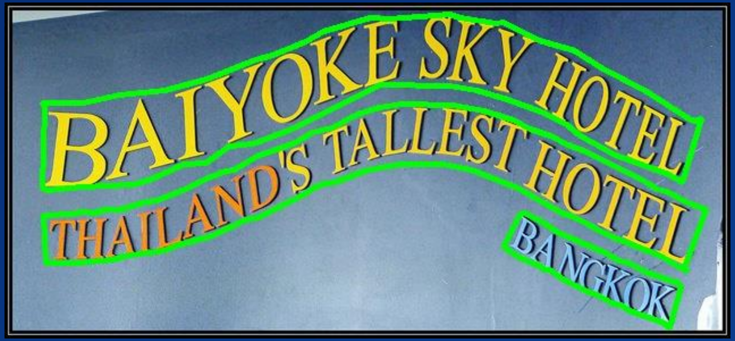
SegLink[1]

Relation Network

[1] B.-G. Shi, *et al.*, “Detecting oriented text in natural images by linking segments,” CVPR, 2017.

Microsoft

Examples on SCUT-CTW1500[1]



[1] Y.-L. Liu, L.-W. Jin, S.-T. Zhang, S. Zhang, “Detecting curve text in the wild: New dataset and new solution,” arXiv, 2017.

Challenge of Detecting Small Text in High Resolution Images

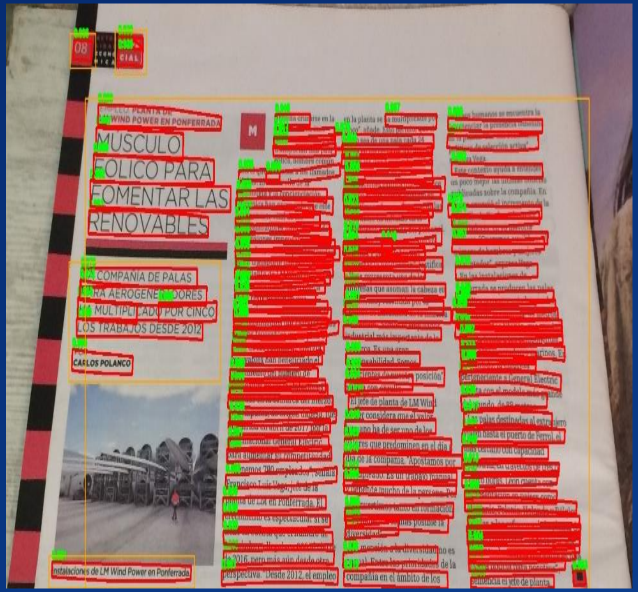
Raw high-resolution image

Resize Image

Text sizes $\ll$ 12px

Resized low-resolution image

Text Detection



Naive solution: use high-resolution image => Very high computation cost

How to detect small texts efficiently in high-resolution images?

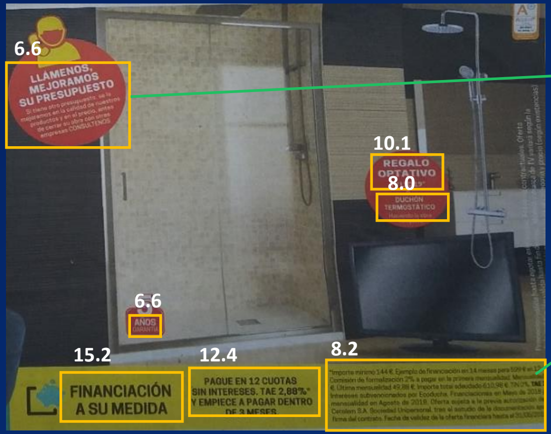
Microsoft

Our Solution: Region-wise Adaptive Scaling

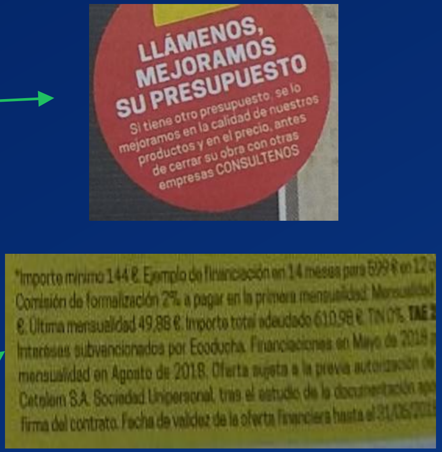
Input Image → 1^{st} Stage Text Block Detection and Scale Estimation → Region-wise Adaptive Scaling → 2^{nd} Stage Text Detection → Text-lines

- **Coarsely localize text-block regions from a low-resolution input image**
- **Resize each text-block to make shorter side lengths of contained texts in a working range**

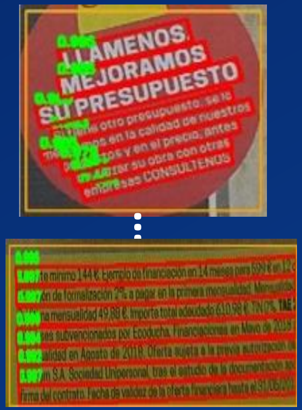
Resized low-resolution image



Detected text-blocks with estimated scales



Adaptively re-scaled text-blocks



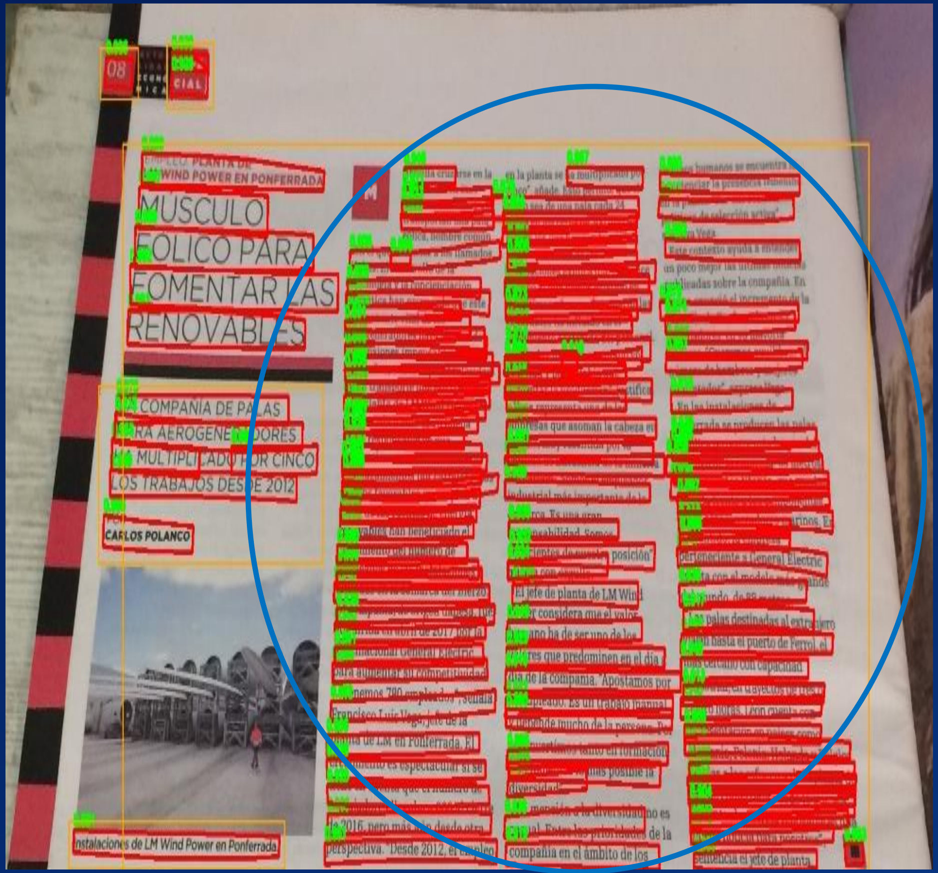
Detected texts on each Re-scaled Text-block

Final detected texts

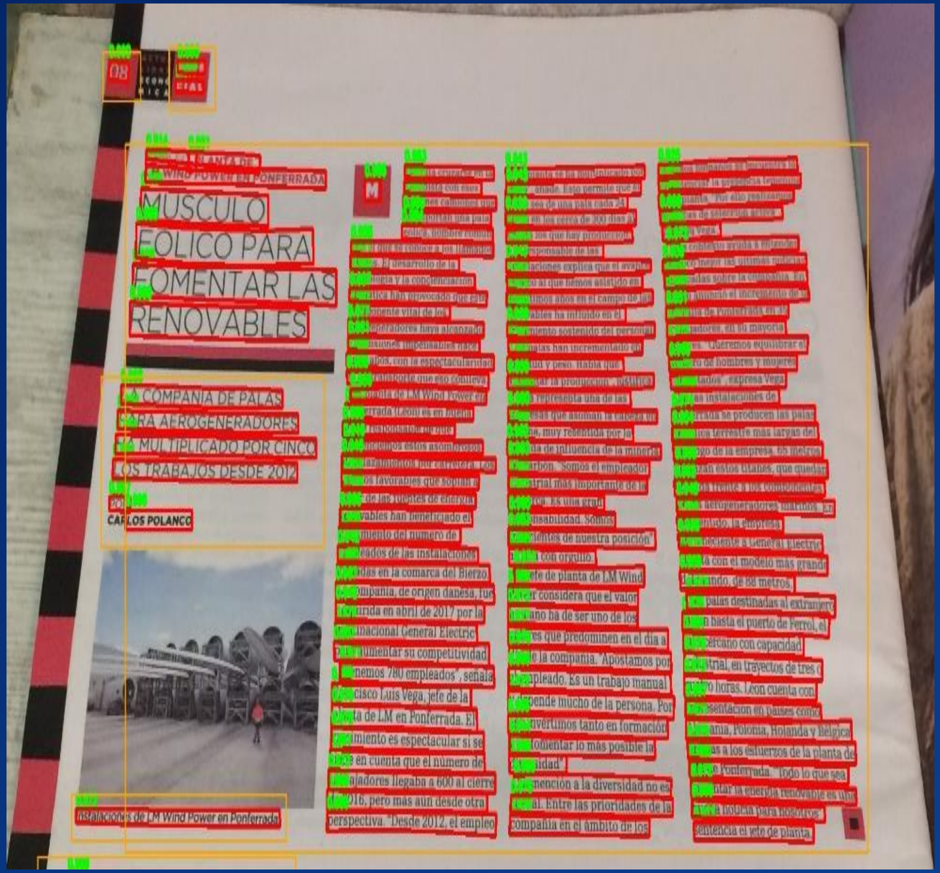
Coarse → **Fine**

A concurrent work: E. Richardson, et al., “It’s All About The Scale - Efficient Text Detection Using Adaptive Scaling,” arXiv:1907.12122, July 28, 2019

Example

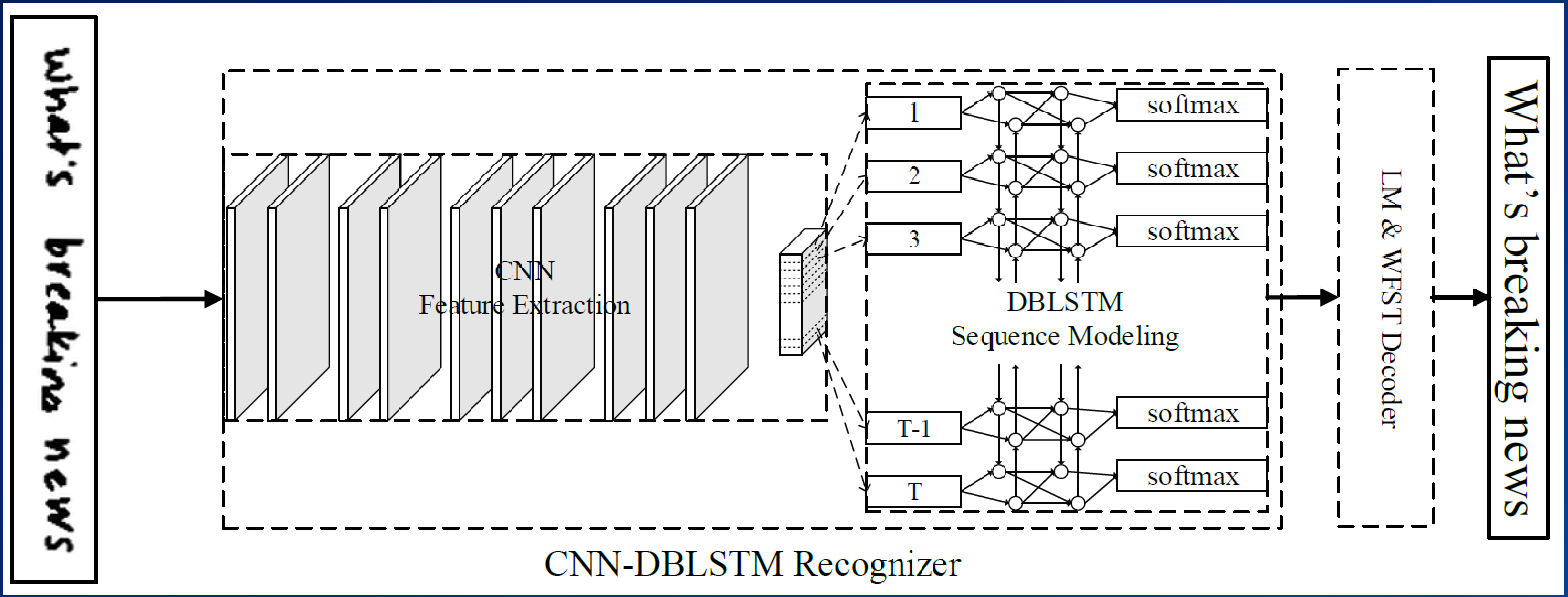


Without adaptive scaling



With region-wise adaptive scaling

CNN-DBLSTM based Text Decoder

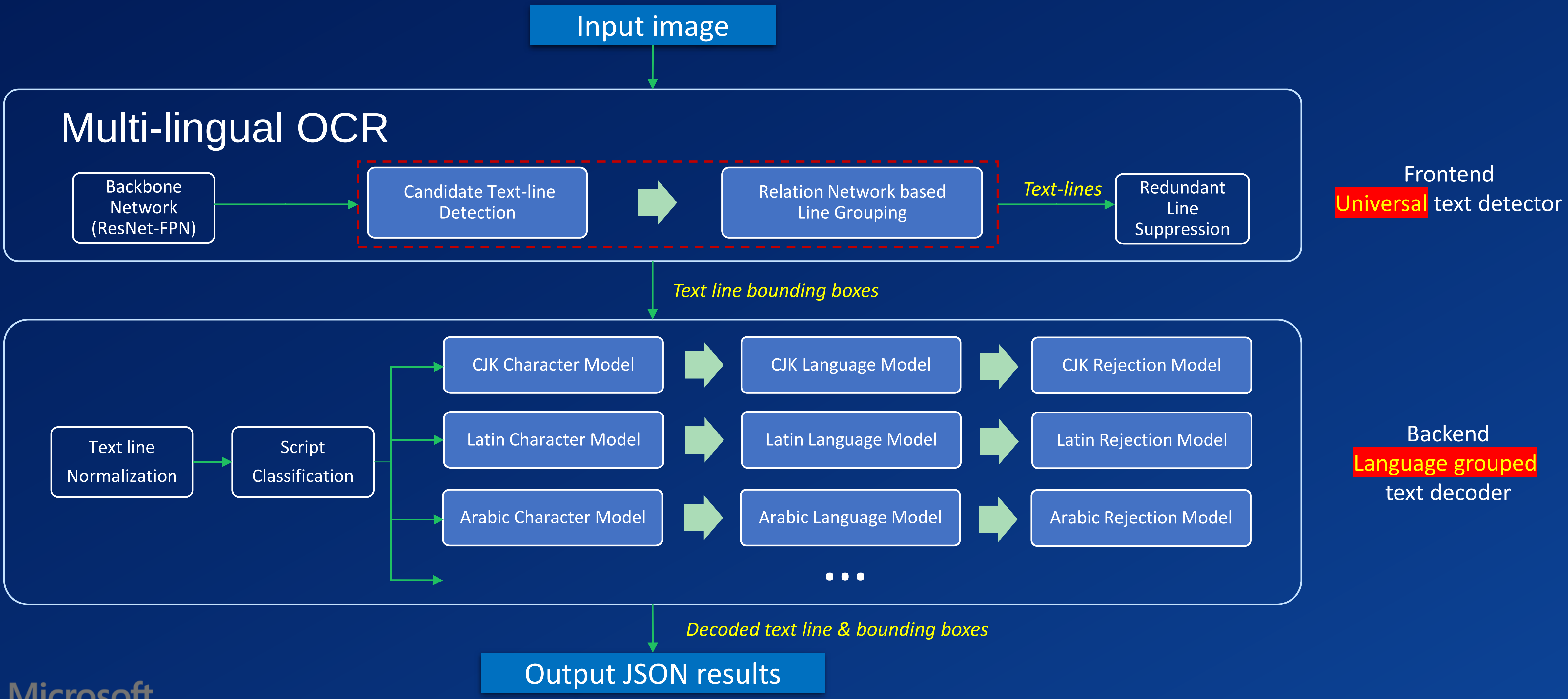


Input

WFST-based Decoder

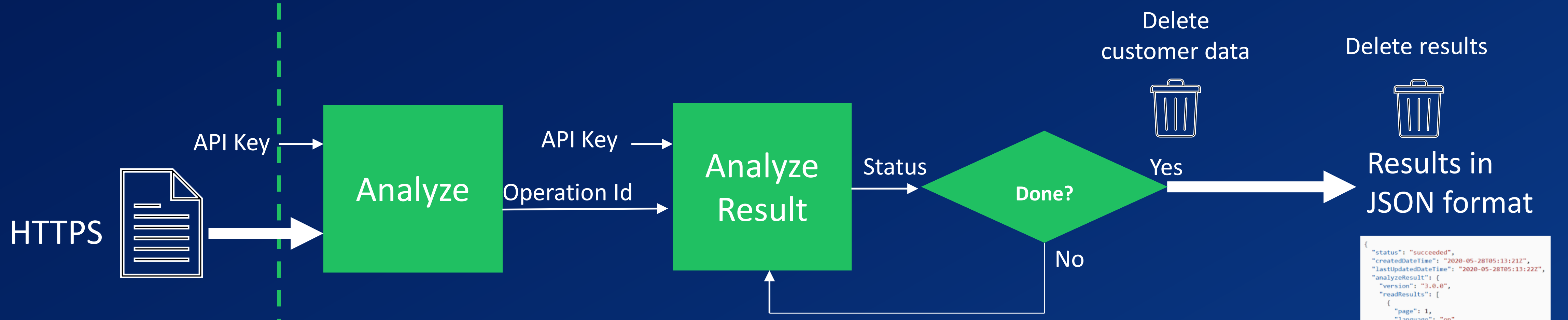
Microsoft

Language Expansion



Read API – Available in Cloud and on-Prem

Optional: Virtual networks, IP firewalls, and private endpoints




```json
{
  "status": "succeeded",
  "createdDateTime": "2020-05-28T05:13:21Z",
  "lastUpdatedDateTime": "2020-05-28T05:13:22Z",
  "analyzeResult": {
    "version": "3.0.0",
    "readResults": [
      {
        "page": 1,
        "language": "en",
        "angle": 0.8551,
        "width": 2661,
        "height": 1901,
        "unit": "pixel",
        "lines": [
          {
            "boundingBox": [
              67,
              646,
              2582,
              713,
              2580,
              876,
              67,
              821
            ],
            "text": "The quick brown fox jumps",
            "words": [
              {
                "boundingBox": [
                  143,
                  650,
                  435,
                  661,
                  436,
                  823,
                  144,
                  824
                ],
                "text": "The",
                "confidence": 0.958
              }
            ]
          }
        ]
      }
    ]
  }
}
```


- Supported file formats: JPEG, PNG, BMP, PDF, and TIFF
- For PDF AND TIFF, up to 2000 pages are processed. For free tier subscribers, only the first two pages are processed.
- The file size must be less than 50 MB and dimensions at least 50 x 50 pixels and at most 10000 x 10000 pixels.

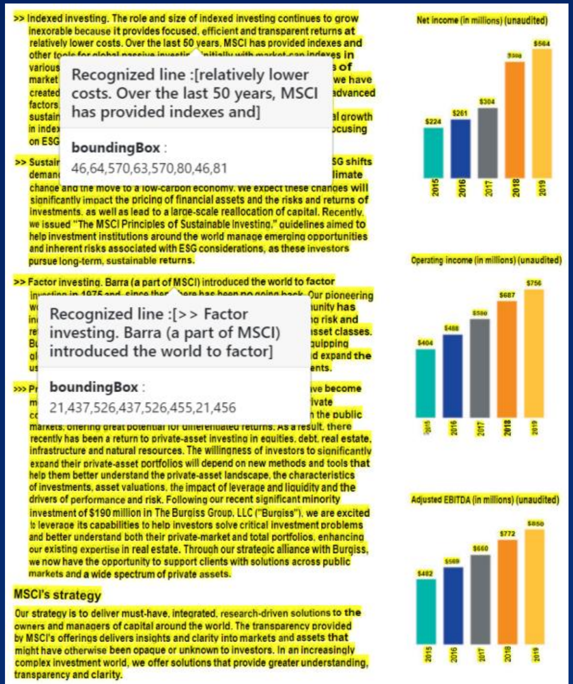
Mixed Languages Comparison (Example)

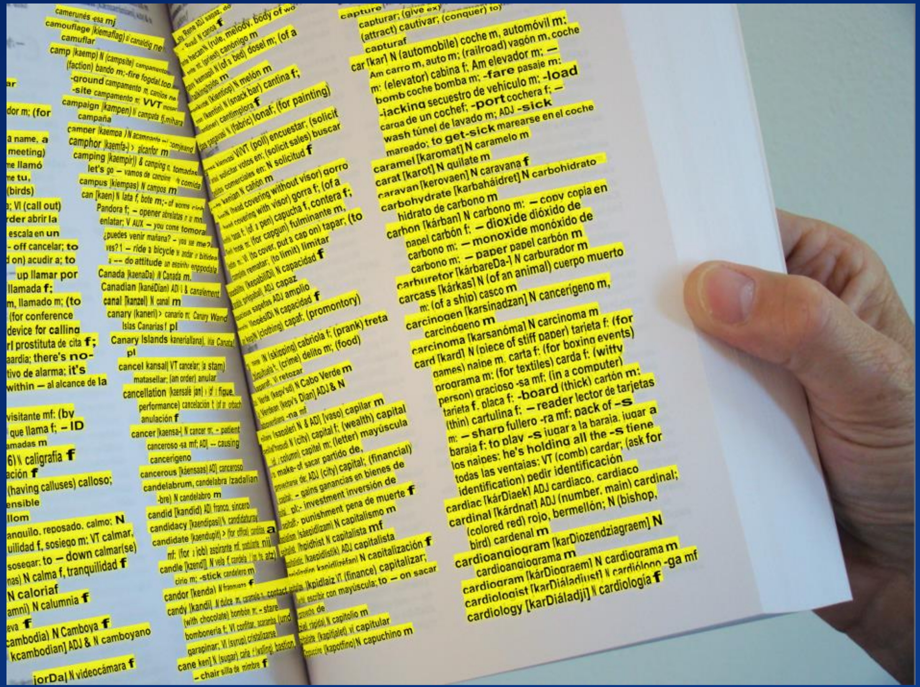
Other OCR auto mode

Microsoft OCR auto mode

Read 3.0+ Examples

1. Text in documents
2. Text in the wild
3. Languages



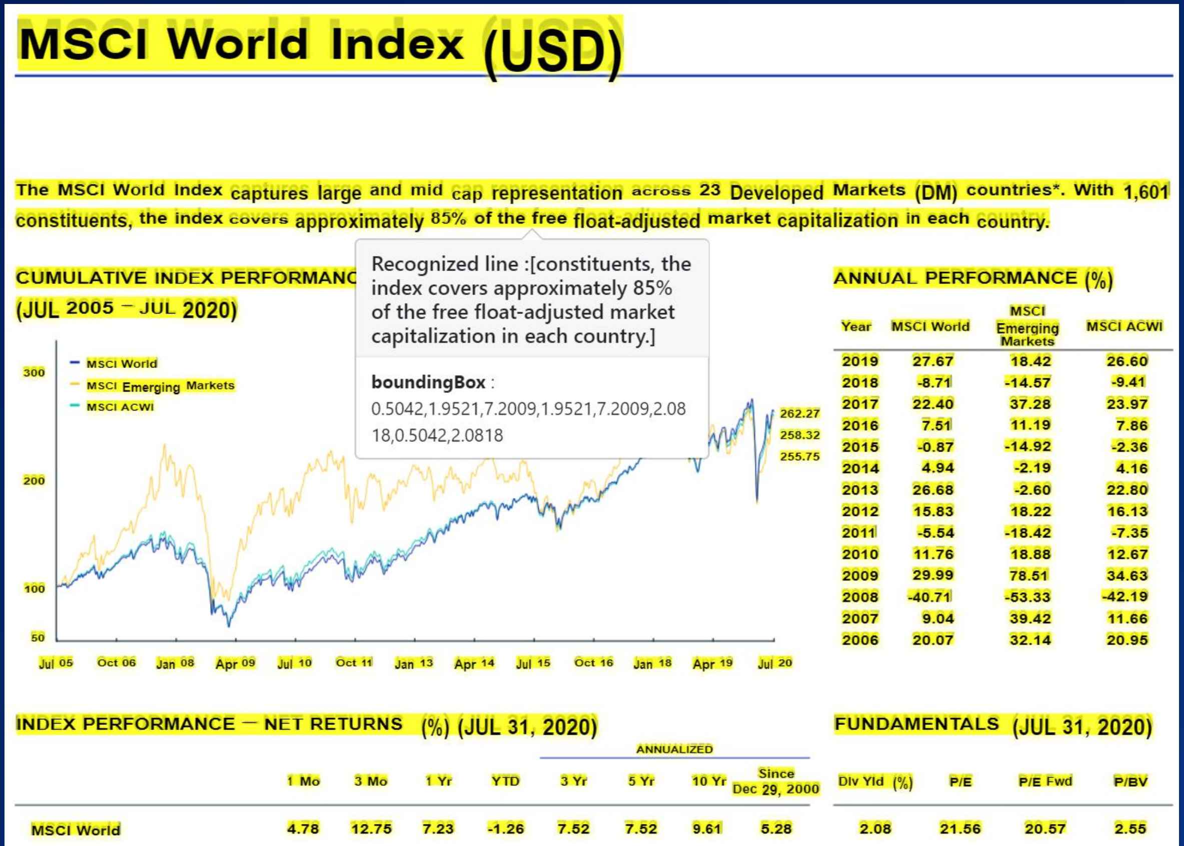


Images in the Wild

Documents

AUTOMATION ANYWHERE
Customer:
Automation Anywhere
Industry:
Partner Professional Services
Size:
1,000 – 9,999 employees
Country:
United States
Products and services:
Microsoft Azure
Azure Cognitive Services
Computer Vision
Form Recognizer
Optical Character Recognition
Read full story here
Microsoft
MENU
STARTERS
SEAFOOD
CHEF'S SPECIALS
DESSERTS
MAIN DISHES
una guida rapida a ....
VENEZIA
Non ci vorrà molto prima che ti innamori di Venezia, la meravigliosa città d'acqua. Sperimentale sua sorprendente architettura e arte che si estende in modo seducente attraverso canali scintillanti e ponti decorati.
a shortcut guide to ....
VENICE
It won't take long before you fall in love with Venice, the wonderful city of water. Experience its astonishing architecture and art that sprawl enticingly across sun-spangled canals and ornate bridges.
Auto
English
Computer Vision for Augmented Reality – Microsoft Tech Community - 318099
Length
Color
Port
Input
POPULAR MECHANICS
WIN
YOU CAN SEE THE FUTURE WITH THESE
-26BOLD IDEAS
That Are Changing Our Times
GPU
Hololens Graphics
GPU engine utilization
100.00
Frame rate
120.00
0.00
App frames per second:35.2
A Hololens running RS4 (1803) is not able to use the GPU, all predictions will fast. If your device is upgraded to RS5 (1805) you are able to utilize the GPU. output. There is one catch though: if you're using the CustomVision service t it as v1.2 otherwise MLGen will give you a completly random error message
My colleague Sebastian described the Windows ML implementation in great
pen and paper? I'd argue the biggest thing you loose is personality times New Roman looks the same on every paper, and frankly, it's boring to look at. Handwriting adds another dimension to your writing and make it truly yours. I can
TC# 5288 9647 9579 6887 9704
NEED A MONEY ORDER? WAL-MART HAS IT!
02/14/03 09:46:31
*** CUSTOMER COPY ***
February
March
April
May
June
July
August

Documents…



MSCI World Index (USD)

The MSCI World Index captures large and mid cap representation across 23 Developed Markets (DM) countries*. With 1,601 constituents, the index covers approximately 85% of the free float-adjusted market capitalization in each country.

Recognized line :[constituents, the index covers approximately 85% of the free float-adjusted market capitalization in each country.]

boundingBox : 0.5042,1.9521,7.2009,1.9521,7.2009,2.0818,0.5042,2.0818

CUMULATIVE INDEX PERFORMANC
(JUL 2005 – JUL 2020)

ANNUAL PERFORMANCE (%)

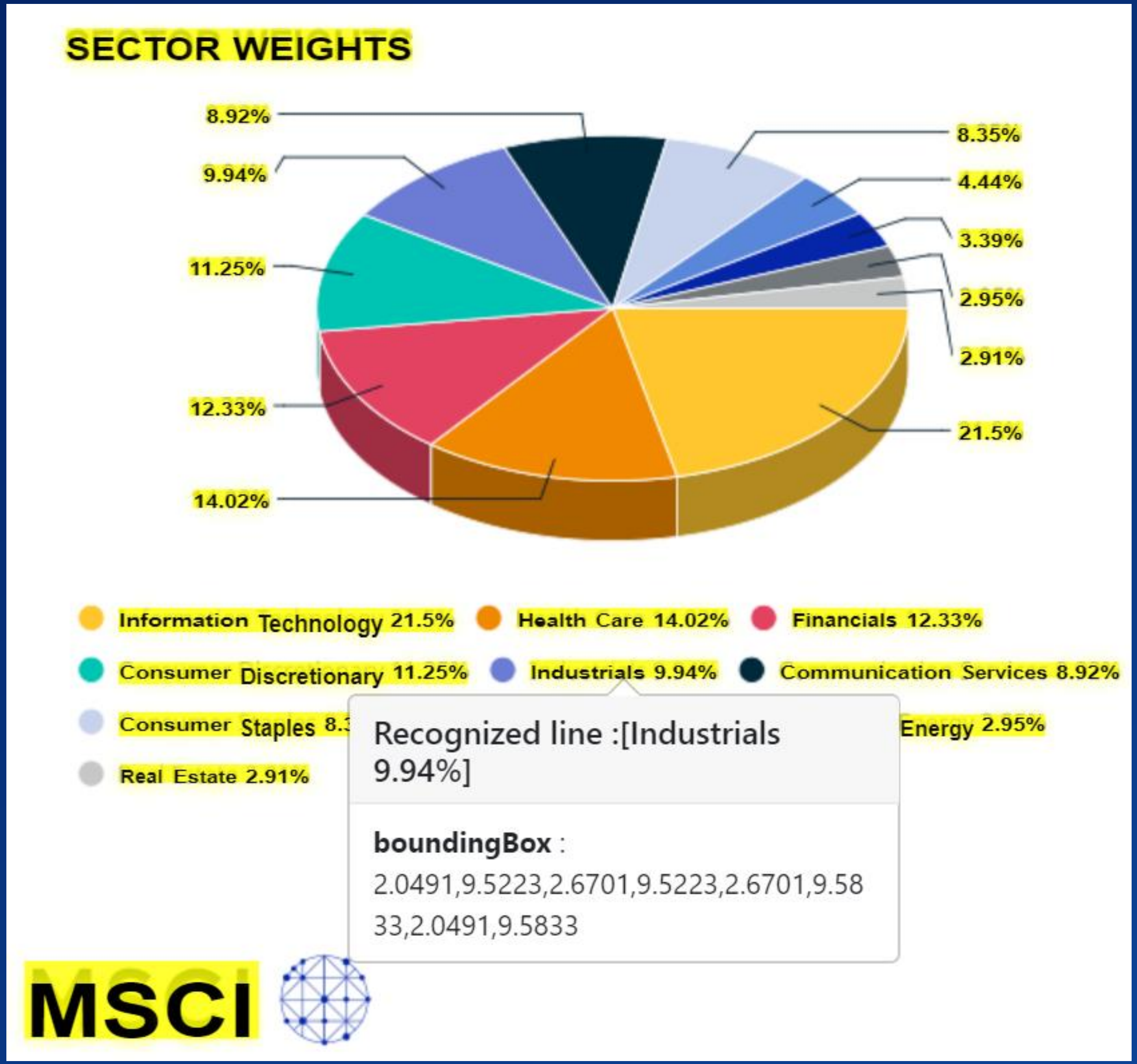
Year	MSCI World	MSCI Emerging Markets	MSCI ACWI
2019	27.67	18.42	26.60
2018	-8.71	-14.57	-9.41
2017	22.40	37.28	23.97
2016	7.51	11.19	7.86
2015	-0.87	-14.92	-2.36
2014	4.94	-2.19	4.16
2013	26.68	-2.60	22.80
2012	15.83	18.22	16.13
2011	-5.54	-18.42	-7.35
2010	11.76	18.88	12.67
2009	29.99	78.51	34.63
2008	-40.71	-53.33	-42.19
2007	9.04	39.42	11.66
2006	20.07	32.14	20.95

INDEX PERFORMANCE — NET RETURNS (%) (JUL 31, 2020)

					ANNUALIZED			
	1 Mo	3 Mo	1 Yr	YTD	3 Yr	5 Yr	10 Yr	Since Dec 29, 2000
MSCI World	4.78	12.75	7.23	-1.26	7.52	7.52	9.61	5.28

FUNDAMENTALS (JUL 31, 2020)

Div Yld (%)	P/E	P/E Fwd	P/BV
2.08	21.56	20.57	2.55



SECTOR WEIGHTS

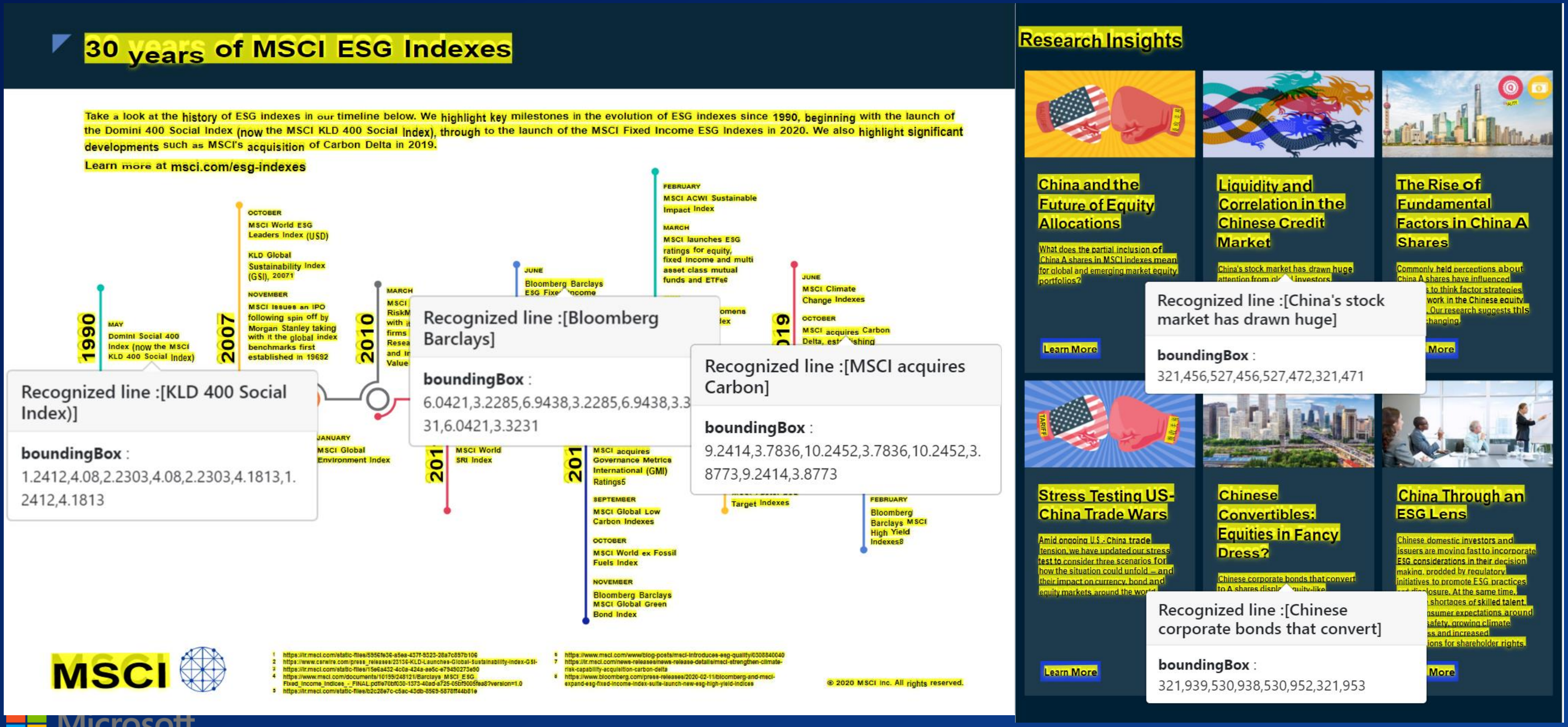
Recognized line :[Industrials 9.94%]

boundingBox : 2.0491,9.5223,2.6701,9.5223,2.6701,9.5833,2.0491,9.5833

MSCI

Microsoft

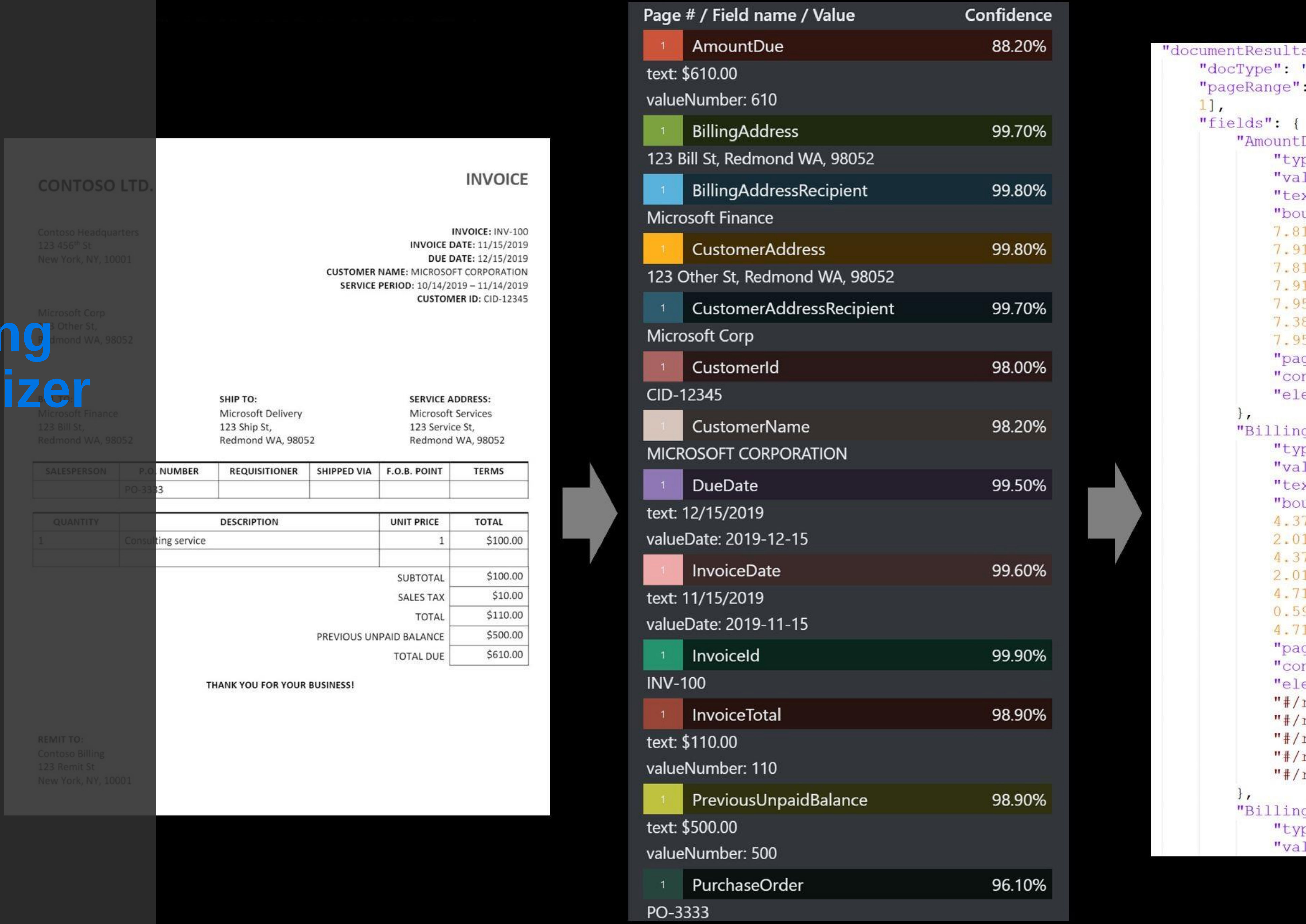
Visuals…



30 years of MSCI ESG Indexes
Take a look at the history of ESG indexes in our timeline below. We highlight key milestones in the evolution of ESG indexes since 1990, beginning with the launch of the Domini 400 Social Index (now the MSCI KLD 400 Social Index), through to the launch of the MSCI Fixed Income ESG Indexes in 2020. We also highlight significant developments such as MSCI's acquisition of Carbon Delta in 2019.
Learn more at msci.com/esg-indexes
1990
MAY
Domini Social 400 Index (now the MSCI KLD 400 Social Index)
2007
OCTOBER
MSCI World ESG Leaders Index (USD)
KLD Global Sustainability Index (GSI), 20071
NOVEMBER
MSCI issues an IPO following spin off by Morgan Stanley taking with it the global index benchmarks first established in 19692
2010
MARCH
JUNE
Bloomberg Barclays ESG Fixed Income
FEBRUARY
MSCI ACWI Sustainable Impact Index
MARCH
MSCI launches ESG ratings for equity, fixed income and multi asset class mutual funds and ETFs6
JUNE
MSCI Climate Change Indexes
OCTOBER
MSCI acquires Carbon Delta, establishing
Recognized line :[KLD 400 Social Index)]
boundingBox :
1.2412,4.08,2.2303,4.08,2.2303,4.1813,1.2412,4.1813
Recognized line :[Bloomberg Barclays]
boundingBox :
6.0421,3.2285,6.9438,3.2285,6.9438,3.331,6.0421,3.3231
Recognized line :[MSCI acquires Carbon]
boundingBox :
9.2414,3.7836,10.2452,3.7836,10.2452,3.8773,9.2414,3.8773
JANUARY
MSCI Global Environment Index
MSCI World SRI Index
MSCI acquires Governance Metrics International (GMI) Ratings5
SEPTEMBER
MSCI Global Low Carbon Indexes
OCTOBER
MSCI World ex Fossil Fuels Index
NOVEMBER
Bloomberg Barclays MSCI Global Green Bond Index
Target Indexes
FEBRUARY
Bloomberg Barclays MSCI High Yield Indexes8
MSCI
© 2020 MSCI Inc. All rights reserved.
Research Insights
China and the Future of Equity Allocations
What does the partial inclusion of China A shares in MSCI indexes mean for global and emerging market equity portfolios?
Learn More
Liquidity and Correlation in the Chinese Credit Market
China's stock market has drawn huge
Recognized line :[China's stock market has drawn huge]
boundingBox :
321,456,527,456,527,472,321,471
The Rise of Fundamental Factors in China A Shares
Commonly held perceptions about China A shares have influenced
More
Stress Testing US-China Trade Wars
Amid ongoing U.S.-China trade tension, we have updated our stress test to account for three scenarios for how the situation could unfold – and their impact on currency, bond and equity markets around the world
Learn More
Chinese Convertibles: Equities in Fancy Dress?
Chinese corporate bonds that convert
Recognized line :[Chinese corporate bonds that convert]
boundingBox :
321,939,530,938,530,952,321,953
China Through an ESG Lens
Chinese domestic investors and issuers are moving fast to incorporate ESG considerations in their decision making, prodded by regulatory initiatives to promote ESG practices
More

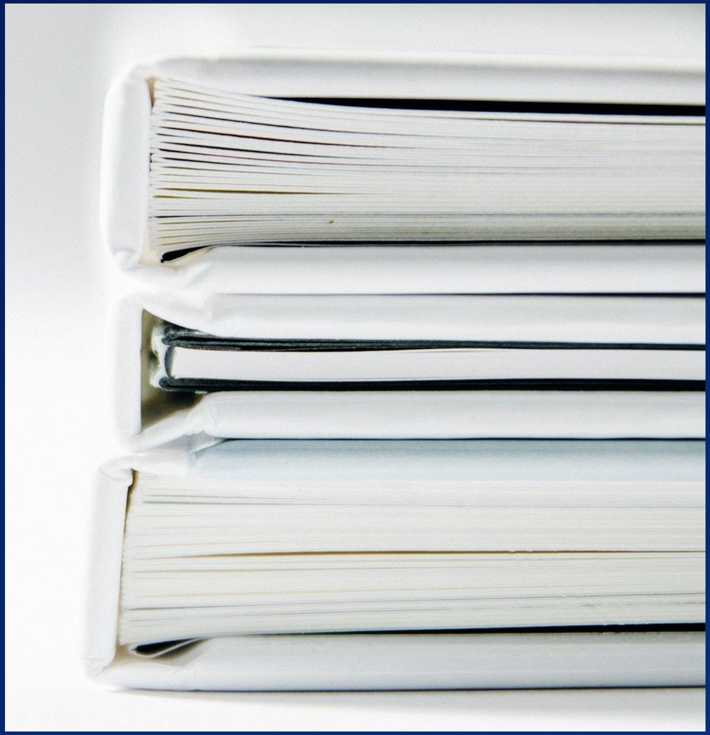
Mixed Languages…

Document Understanding (Form Recognizer API)



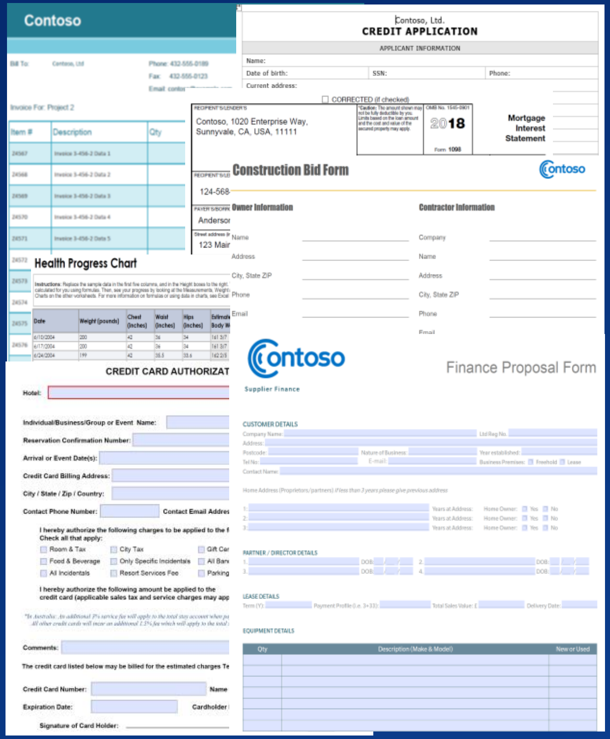
Form Recognizer

Data extraction in any business process that intakes forms and outputs structured data



Layout

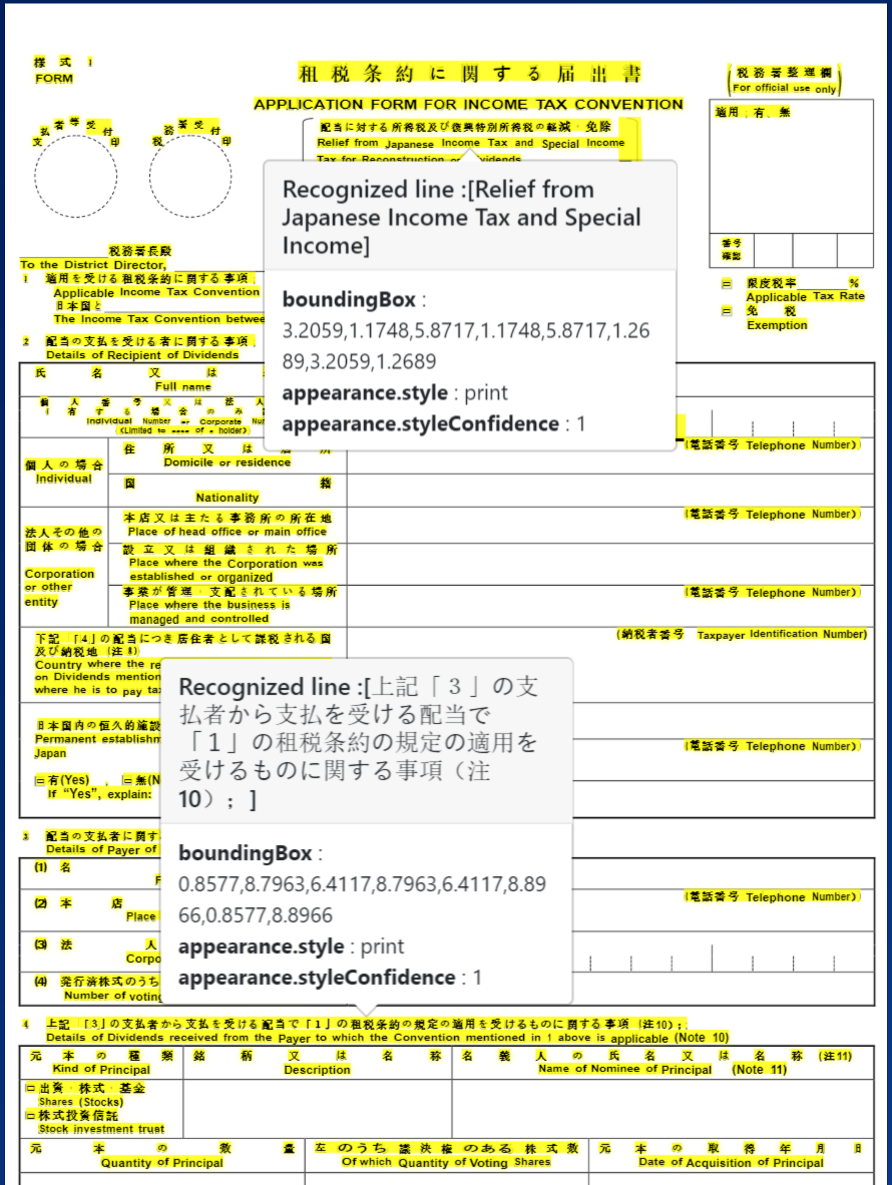
Prebuilt[s]



Custom

Microsoft

An Atypical Language Understanding Problem



Not all documents can have a clear read order… Can we still extract knowledge like key-value pairs?

Visual Linguistic Tasks

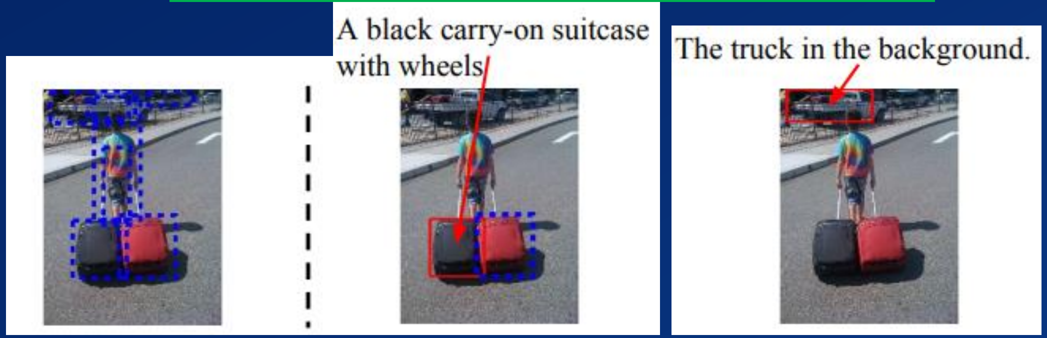
Visual Question Answering

Image Captioning

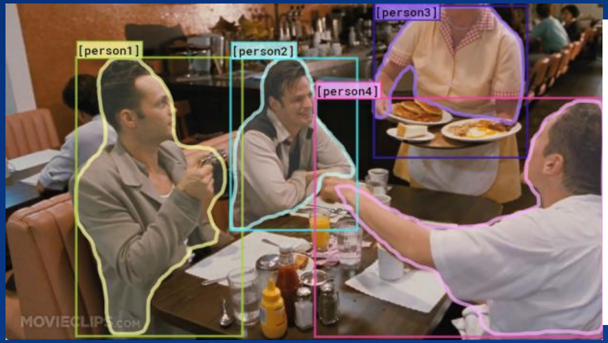
Image-Text Retrieval

1:A female runner dressed in blue athletic wear is running in a competition , while spectators line the street . ✓
2:A lady dressed in blue running a marathon . ✓
3:A young woman is running a marathon in a light blue tank top and spandex shorts . ✓
4:A lady standing at a crosswalk . ✗
5:A woman who is running , with blue shorts . ✓

Referred Expression Comprehension



Visual Commonsense Reasoning



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Microsoft

Some Representative Works on Visual-Linguistic Joint Modeling

- VideoBERT: A Joint Model for Video and Language Representation Learning, Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid, ICCV 2019.
- ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee, NIPS 2019.
- LXMERT: Learning Cross-Modality Encoder Representations from Transformers, Hao Tan, Mohit Bansal, EMNLP 2019.
- Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training, Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou, AAAI 2020.
- VL-BERT: Pre-training of Generic Visual-Linguistic Representations, Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai, ICLR 2020.

Focusing More on Documents

- Yang et al.[1] presented an end-to-end, multimodal, fully convolutional network for extracting semantic structures.
- Liu et al.[2] introduced a Graph Convolutional Networks (GCN) based model to combine textual and visual information.
- Davis et al. [3] proposed to use relationship classifier and neighbor prediction networks to identify key-value pairs.
- Sarkhel et al. [4] proposed a multi-scale classification method to classify the visually rich document.

[1] Yang, Xiaowei et al. "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks." CVPR (2017).
[2] Liu, Xiaojing et al. "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents." NAACL-HLT (2019).
[3] Davis, Brian et al. "Deep Visual Template-Free Form Parsing." ICDAR (2019).
[4] Sarkhel, Ritesh and Arnab Nandi. "Deterministic Routing between Layout Abstractions for Multi-Scale Classification of Visually Rich Documents." *IJCAI* (2019).

Report

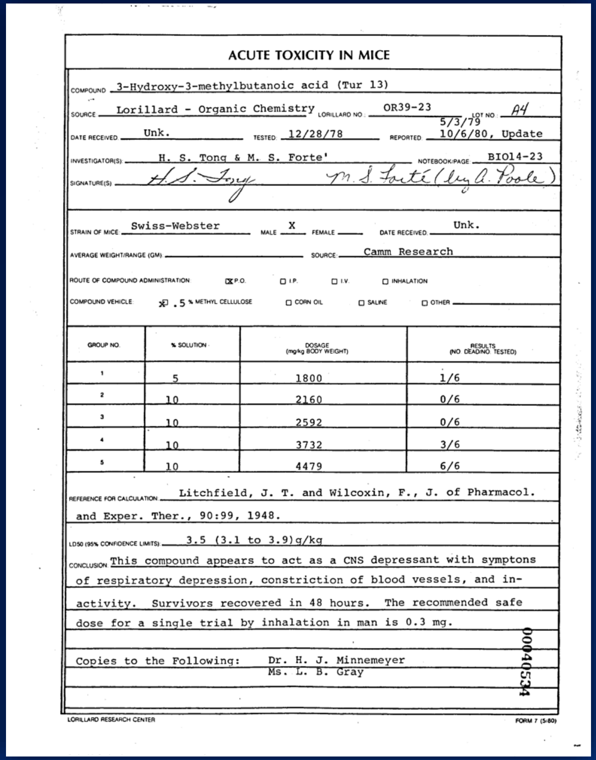

UBS | Global Research | 31 January 2018

First Read

Microsoft Corp.

Azure Acceleration Impresses in Solid Q2

Form


ACUTE TOXICITY IN MICE

COMPOUND 3-Hydroxy-3-methylbutanoic acid (Tur 13)

SOURCE Lorillard - Organic Chemistry LORILLARD NO. OR39-23 LOT NO. A4

DATE RECEIVED Unk. TESTED 12/28/78 REPORTED 5/3/79 10/6/80, Update

INVESTIGATOR(S) H. S. Tong & M. S. Forte' NOTEBOOK/PAGE BIO14-23

STRAIN OF MICE Swiss-Webster MALE X FEMALE DATE RECEIVED Unk.

AVERAGE WEIGHT/RANGE (GM) SOURCE Camm Research

GROUP NO.	% SOLUTION	DOSAGE (mg/kg BODY WEIGHT)	RESULTS (NO. DEAD/NO. TESTED)
1	5	1800	1/6
2	10	2160	0/6
3	10	2592	0/6
4	10	3732	3/6
5	10	4479	6/6

REFERENCE FOR CALCULATION Litchfield, J. T. and Wilcoxin, F., J. of Pharmacol. and Exper. Ther., 90:99, 1948.

LD50 (95% CONFIDENCE LIMITS) 3.5 (3.1 to 3.9)g/kg

CONCLUSION This compound appears to act as a CNS depressant with symptoms of respiratory depression, constriction of blood vessels, and in-activity. Survivors recovered in 48 hours. The recommended safe dose for a single trial by inhalation in man is 0.3 mg.

Copies to the Following: Dr. H. J. Minnemeyer Ms. L. B. Gray

LORILLARD RESEARCH CENTER FORM 7 (5-80)

Receipt


Morton's The Steakhouse
735 S Figueroa Street
Los Angeles, CA 90017
(213) 553-4566

Server: Sally DOB: 09/08/2016
06:23 PM 09/08/2016
404/1 1/10084

SALE

Visa 6291458
Card #XXXXXXXXXXXX8698
Magnetic card present: SICKAFOOSE/DANNY
Card Entry Method: S

Approval: 04702B

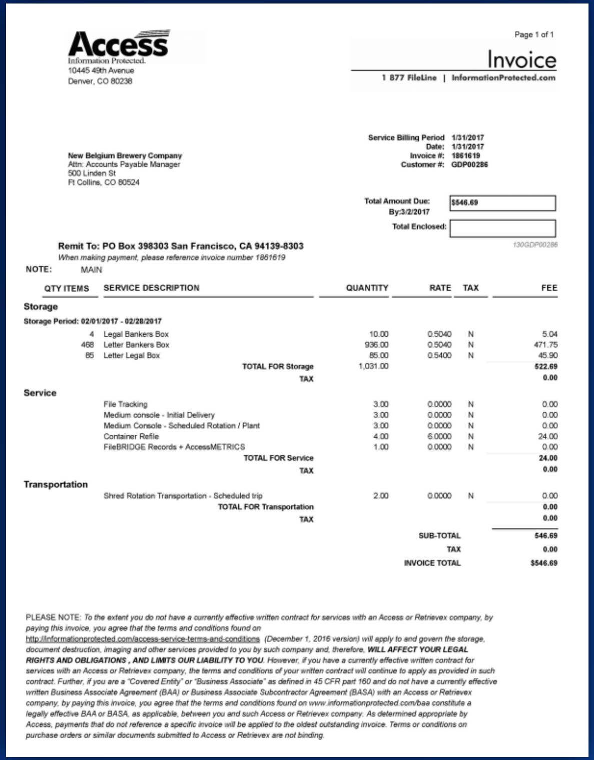
Amount: $33.79
+ Gratuity Not Inc: 6
= Balance Due: 39.79

I agree to pay the above total amount according to the card issuer agreement.

For banquet events, balance due includes suggested gratuity if accepted.

Guest Copy

Invoice


Access

Invoice

Page 1 of 1

Total Amount Due: $546.69

Remit To: PO Box 398303 San Francisco, CA 94139-8303

QTY ITEMS	SERVICE DESCRIPTION	QUANTITY	RATE	TAX	FEE
4	Legal Bankers Box	10.00	0.5040	N	5.04
468	Letter Bankers Box	936.00	0.5040	N	471.75
85	Letter Legal Box	85.00	0.5400	N	45.90

SUB-TOTAL 546.69

INVOICE TOTAL $546.69

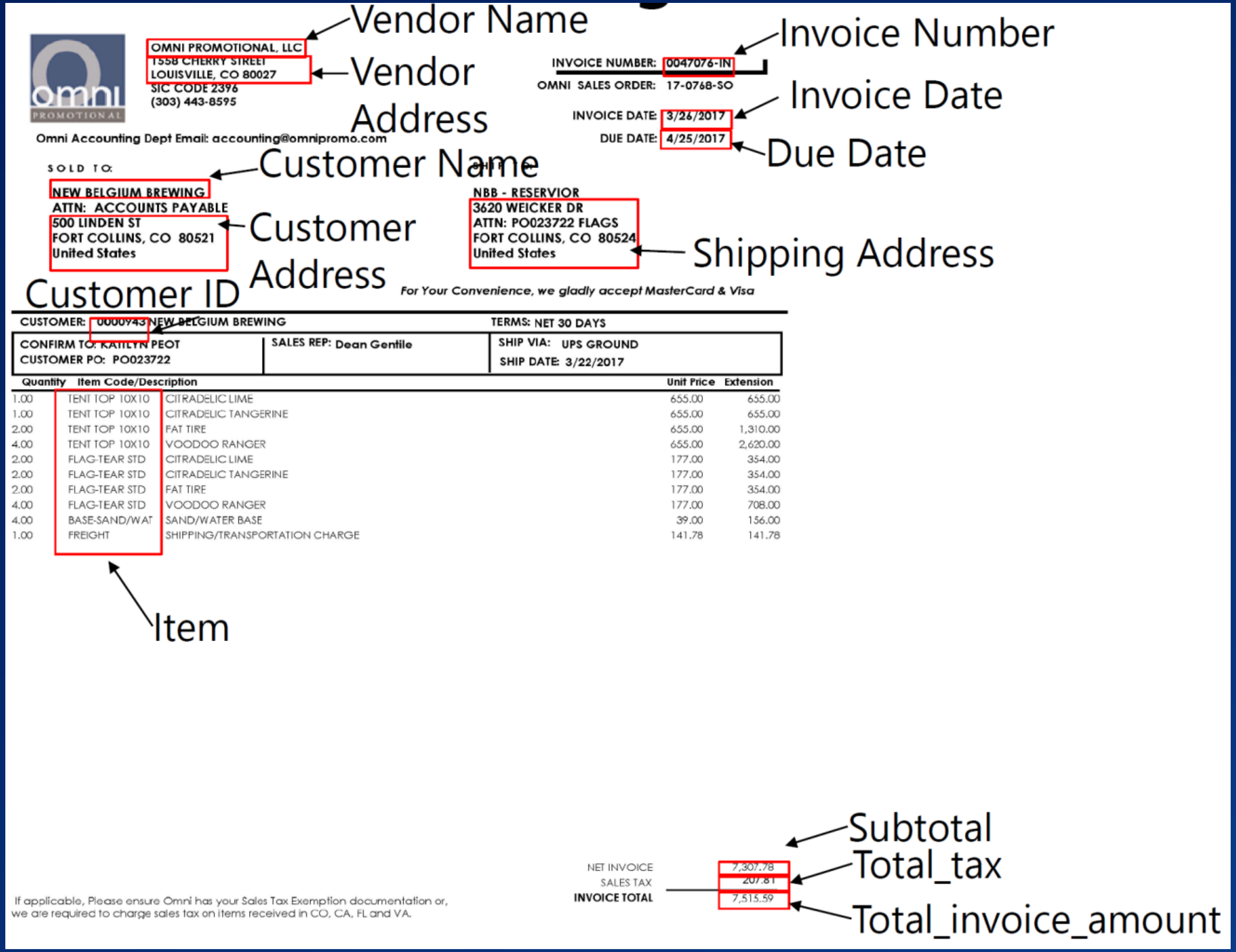
Document Understanding in Real World

LayoutLM: Pre-training for **Text** with rich **Layout** and **Style** information [1]

Microsoft

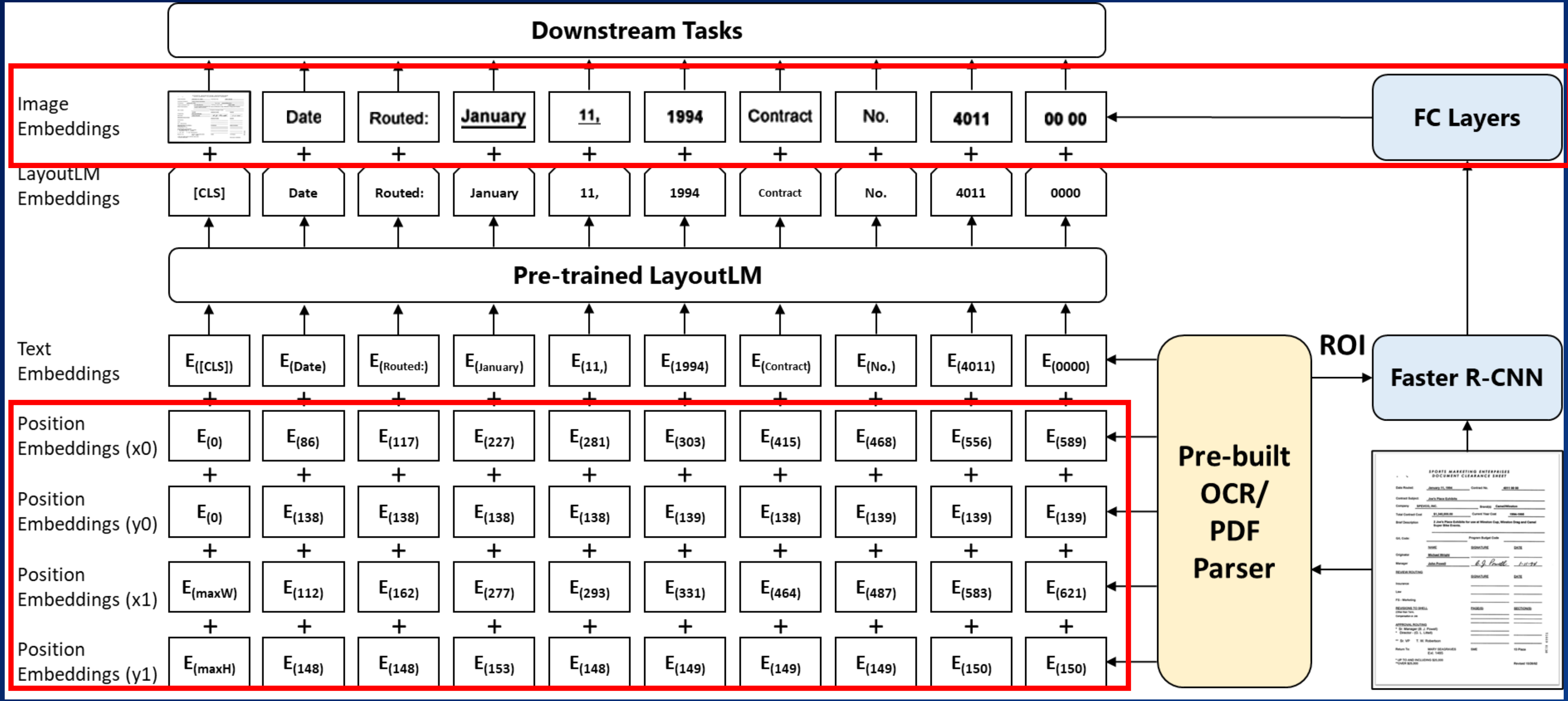
[1] Xu et al., LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD 2020.

Example: Invoice Understanding



- Date
- ID
- Number
- Address
- Name
- Item
- …

LayoutLM Architecture

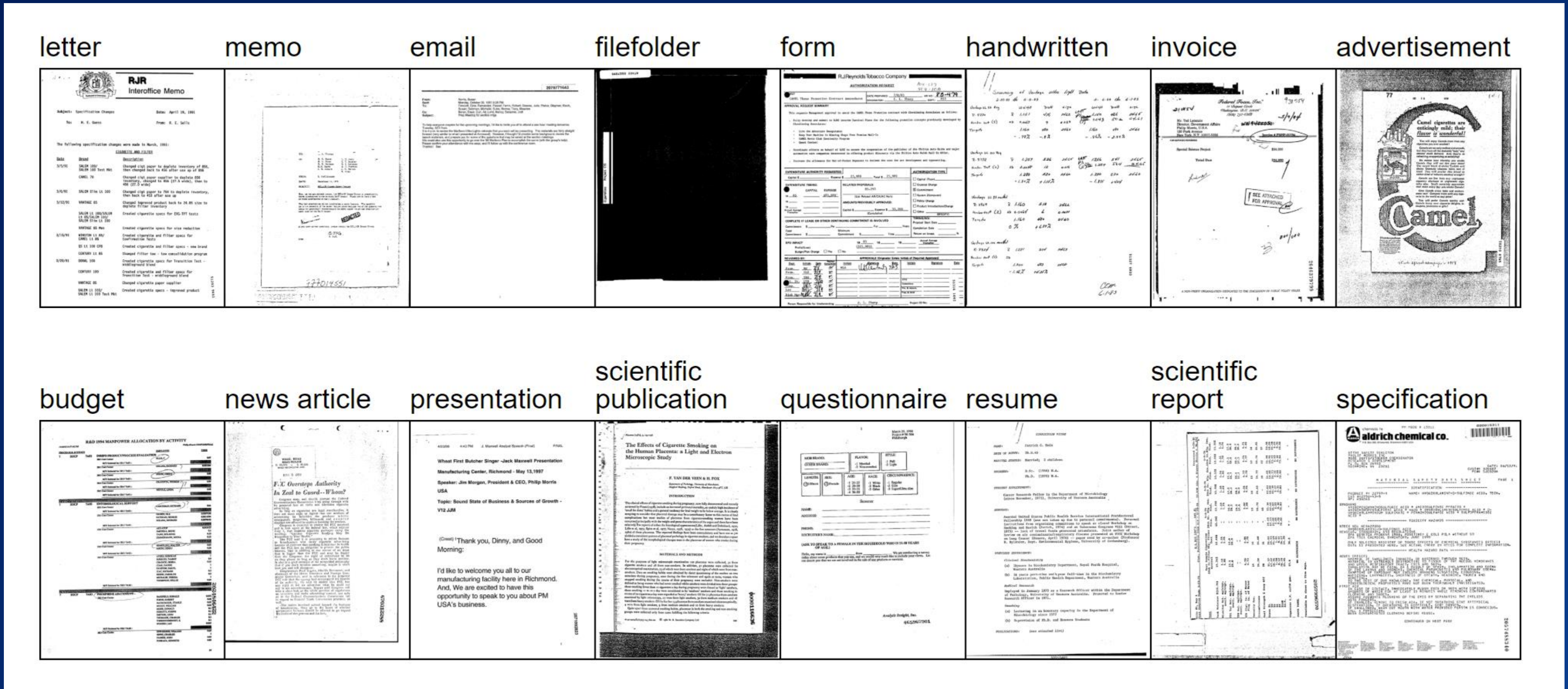


- Image Embeddings
- Position Embeddings

* Text embeddings initialized by BERT/UniLM

LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD'2020, https://arxiv.org/abs/1912.13318

Pre-training Data

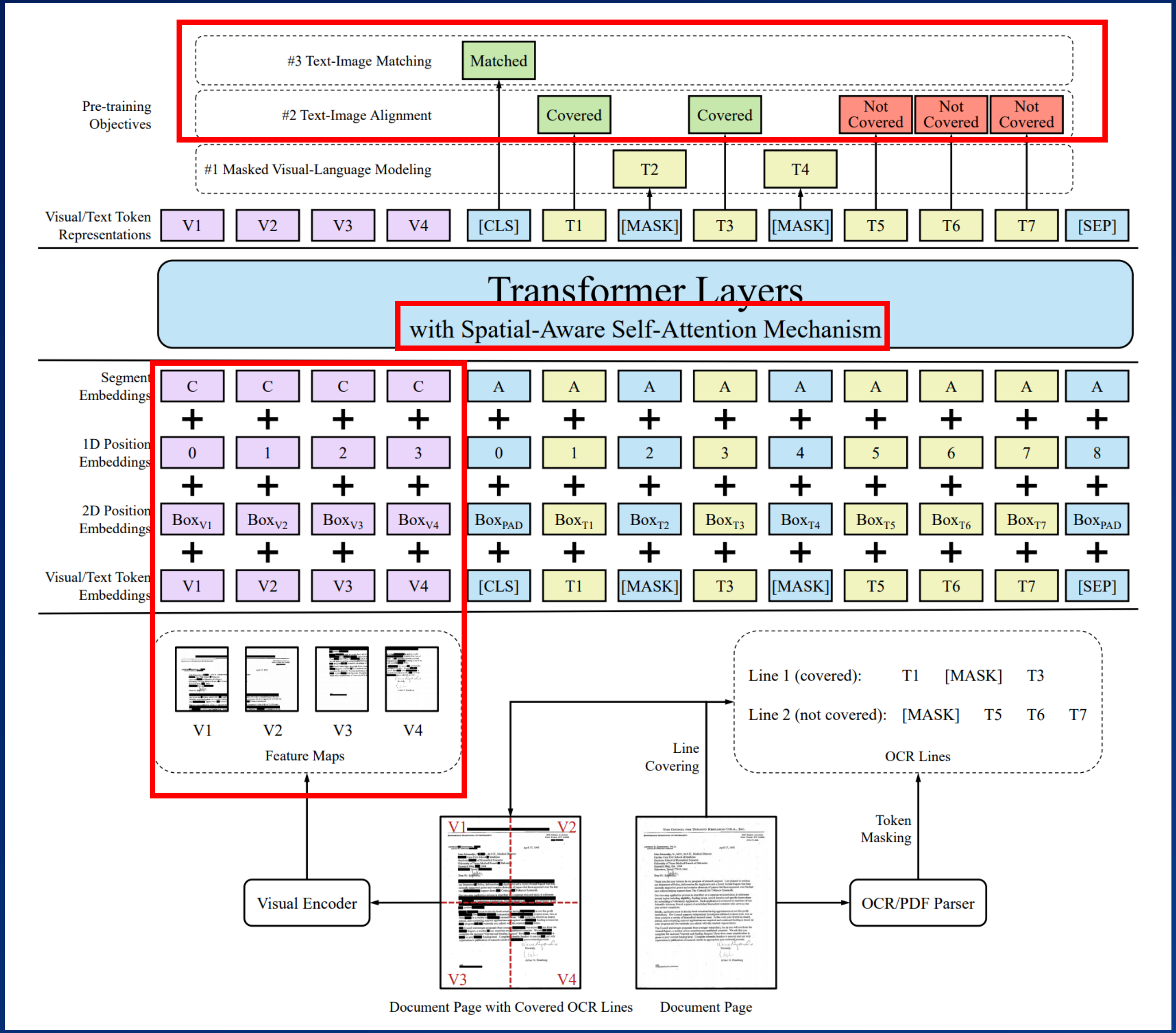


11 million scanned document images from IIT-CDIP Test Collection 1.0

https://ir.nist.gov/cdip/

LayoutLMv2

LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding, ACL 2021



- New pre-training tasks
- New self-attention mechanism
- Image features now go though transform layers

Semantic Entity Recognition

Model	FUNSD	CORD	SROIE	Kleister-NDA
$\text{BERT}_{\text{BASE}}$	0.6026	0.8968	0.9099	0.7790
$\text{UniLMv2}_{\text{BASE}}$	0.6648	0.9092	0.9459	0.7950
$\text{BERT}_{\text{LARGE}}$	0.6563	0.9025	0.9200	0.7910
$\text{UniLMv2}_{\text{LARGE}}$	0.7072	0.9205	0.9488	0.8180
$\text{LayoutLM}_{\text{BASE}}$	0.7866	0.9472	0.9438	0.8270
$\text{LayoutLM}_{\text{LARGE}}$	0.7895	0.9493	0.9524	0.8340
$\text{LayoutLMv2}_{\text{BASE}}$	0.8276	0.9495	0.9625	0.8330
$\text{LayoutLMv2}_{\text{LARGE}}$	**0.8420**	**0.9601**	**0.9781**	**0.8520**
BROS (Hong et al., 2021)	0.8121	0.9536	0.9548	–
SPADE (Hwang et al., 2020)	–	0.9150	–	–
PICK (Yu et al., 2020)	–	–	0.9612	–
TRIE (Zhang et al., 2020)	–	–	0.9618	–
Top-1 on SROIE Leaderboard (until 2020-12-24)	–	–	0.9767	–
$\text{RoBERTa}_{\text{BASE}}$ in (Graliński et al., 2020)	–	–	–	0.7930

Form Understanding (***FUNSD***)

https://guillaumejaume.github.io/FUNSD/

Receipt Understanding (***SROIE, CORD***)

https://rrc.cvc.uab.es/?ch=13

https://github.com/clovaai/cord

Document Information Extraction (***Kleister-NDA***)

https://github.com/applicaai/kleister-nda

Microsoft

Document Image Classification

Model	Accuracy	#Parameters
$\text{BERT}_{\text{BASE}}$	89.81%	110M
$\text{UniLMv2}_{\text{BASE}}$	90.06%	125M
$\text{BERT}_{\text{LARGE}}$	89.92%	340M
$\text{UniLMv2}_{\text{LARGE}}$	90.20%	355M
$\text{LayoutLM}_{\text{BASE}}$ (w/ image)	94.42%	160M
$\text{LayoutLM}_{\text{LARGE}}$ (w/ image)	94.43%	390M
$\text{LayoutLMv2}_{\text{BASE}}$	95.25%	200M
$\text{LayoutLMv2}_{\text{LARGE}}$	**95.64%**	426M
VGG-16 (Afzal et al., 2017)	90.97%	-
Single model (Das et al., 2018)	91.11%	-
Ensemble (Das et al., 2018)	92.21%	-
InceptionResNetV2[6] (Szegedy et al., 2016)	92.63%	-
LadderNet (Sarkhel & Nandi, 2019)	92.77%	-
Single model (Dauphinee et al., 2019)	93.03%	-
Ensemble (Dauphinee et al., 2019)	93.07%	-

Document Image Classification (***RVL-CDIP***)
https://www.cs.cmu.edu/~aharley/rvl-cdip/

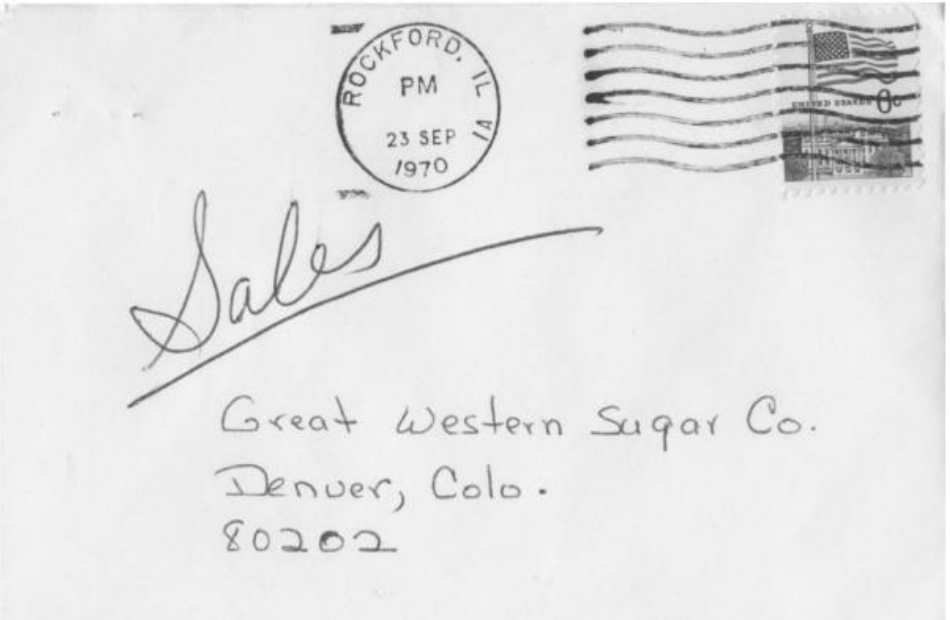
Document VQA

Model	Fine-tuning set	ANLS	#Parameters
$\text{BERT}_{\text{BASE}}$	train	0.6354	110M
$\text{UniLMv2}_{\text{BASE}}$	train	0.7134	125M
$\text{BERT}_{\text{LARGE}}$	train	0.6768	340M
$\text{UniLMv2}_{\text{LARGE}}$	train	0.7709	355M
$\text{LayoutLM}_{\text{BASE}}$	train	0.6979	113M
$\text{LayoutLM}_{\text{LARGE}}$	train	0.7259	343M
$\text{LayoutLMv2}_{\text{BASE}}$	train	0.7808	200M
$\text{LayoutLMv2}_{\text{LARGE}}$	train	0.8348	426M
$\text{LayoutLMv2}_{\text{LARGE}}$	train + dev	0.8529	426M
$\text{LayoutLMv2}_{\text{LARGE}}$ + QG	train + dev	**0.8672**	426M
Top-1 on DocVQA Leaderboard (30 models ensemble)[7]	-	0.8506	-

Document Visual Question Answering (***DocVQA***)
https://rrc.cvc.uab.es/?ch=17

Microsoft

DocVQA Leaderboard



Source: https://www.industrydocuments.ucsf.edu/docs/gfbx0227

Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

Ranking Table

Description Paper Source Code

Date	Method	Score	Figure/Diagram	Form	Table/List	Layout	Free_text	Image/Photo	Handwritten	Yes/No	Others
2020-06-13	Human Performance	0.9811	0.9756	0.9825	0.9780	0.9845	0.9839	0.9740	0.9717	0.9974	0.9828
2020-12-22	LayoutLM 2.0 (single model)	0.8672	0.6574	0.8953	0.8769	0.8791	0.8707	0.7287	0.6729	0.5517	0.8103
2020-08-16	Alibaba DAMO NLP	0.8506	0.6650	0.8809	0.8552	0.8733	0.8397	0.6758	0.7691	0.5492	0.7526
2020-05-16	PingAn-OneConnect-Gammalab-DQA	0.8484	0.6059	0.9021	0.8463	0.8730	0.8337	0.5812	0.7692	0.5172	0.7289
2020-05-14	Structural LM-v2	0.7674	0.4931	0.8381	0.7621	0.7924	0.7596	0.4756	0.6282	0.5517	0.6549
2020-05-15	QA_Base_MRC_2	0.7415	0.4854	0.8015	0.6738	0.7943	0.8136	0.5740	0.5831	0.5287	0.7161
2020-05-15	QA_Base_MRC_1	0.7407	0.4890	0.7984	0.6675	0.7936	0.8131	0.5854	0.6099	0.4943	0.7384
2020-05-15	QA_Base_MRC_4	0.7348	0.4735	0.8040	0.6647	0.7838	0.8043	0.5618	0.5810	0.4598	0.7332
2020-05-15	QA_Base_MRC_3	0.7322	0.4852	0.7958	0.6562	0.7842	0.8044	0.5679	0.5730	0.4511	0.7171
2020-05-15	QA_Base_MRC_5	0.7274	0.4858	0.7877	0.6550	0.7754	0.8047	0.5405	0.5619	0.4598	0.7084
2020-05-16	HyperDQA_V4	0.6893	0.3874	0.7792	0.6309	0.7478	0.7187	0.4867	0.5630	0.4138	0.5685
2020-05-16	HyperDQA_V3	0.6769	0.3876	0.7774	0.6167	0.7332	0.6961	0.4296	0.5373	0.4138	0.5650
2020-05-16	HyperDQA_V2	0.6734	0.3818	0.7666	0.6110	0.7332	0.6867	0.4834	0.5560	0.3793	0.5902
2020-05-09	HyperDQA_V1	0.6717	0.4013	0.7693	0.6197	0.7167	0.6922	0.3598	0.5596	0.4138	0.5504
2020-05-09	bert fulldata fintuned	0.5900	0.4169	0.6870	0.4269	0.6710	0.7315	0.5124	0.4900	0.4483	0.5907
2020-05-01	bert finetuned	0.5872	0.2986	0.7011	0.4849	0.6359	0.6933	0.4622	0.4751	0.4483	0.4895
2020-04-30	HyperDQA_V0	0.5715	0.3131	0.6780	0.4732	0.6630	0.5716	0.3623	0.4351	0.3793	0.4941
2020-04-27	bert	0.4557	0.2233	0.5259	0.2633	0.5113	0.7775	0.4859	0.3565	0.0345	0.5778
2020-05-16	UGLIFT v0.1 (Clova OCR)	0.4417	0.1766	0.5600	0.3178	0.5340	0.4520	0.2253	0.3573	0.4483	0.3356
2020-05-14	Plain BERT QA	0.3524	0.1687	0.4489	0.2029	0.4321	0.4812	0.3517	0.3096	0.0345	0.3747
2020-05-16	Clova OCR V0	0.3489	0.0977	0.4855	0.2670	0.3811	0.3958	0.2489	0.2875	0.0345	0.3062
2020-05-01	HDNet	0.3401	0.2040	0.4688	0.2181	0.4710	0.1916	0.2488	0.2736	0.1379	0.2458
2020-05-16	CLOVA OCR	0.3296	0.1246	0.4612	0.2455	0.3622	0.3746	0.1692	0.2736	0.0690	0.3205
2020-04-29	docVQAQV_V0.1	0.3016	0.2010	0.3898	0.3810	0.2933	0.0664	0.1842	0.2736	0.1586	0.1695
2020-04-26	docVQAQV_V0	0.2342	0.1646	0.3133	0.2623	0.2483	0.0549	0.2277	0.1856	0.1034	0.1635
2020-06-16	Test Submission	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=1&f=1&e=1 (Dec 22, 2020)

Microsoft

LayoutLM for Azure Form Recognizer

2019-03
TableBank
LREC'20

2019-12
LayoutLMv1
KDD'20

2020-06
DocBank
COLING'20

2020-08
Prebuilt Invoice
Private Preview

2020-11
Form Recognizer 2.0
Prebuilt Invoice
Public Preview

2020-12
LayoutLMv2
ACL'21

2021-03
Form Recognizer 2.1
Prebuilt Receipt
Public Preview

2021-04
Language expansion

2021-08
Prebuilt General Key-Value Extraction
Private Preview

LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD'20
LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding, ACL'21

TableBank: A Benchmark Dataset for Table Detection and Recognition, LREC'20
DocBank: A Benchmark Dataset for Document Layout Analysis, COLING'20

Microsoft

Invoice Demo

Concluding Remarks

"Universal OCR" is within our reach

- Representative training data
- Scalable computing platform and training tools
- Universal text detection and mixed language/style text recognition

Document understanding – a vision/language cross-field problem

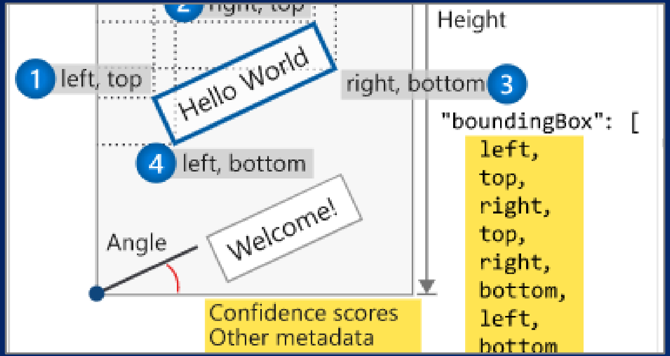
- Joint visual/language pre-training is a powerful idea
- Broadly applicable to many document understanding tasks

More researches on following topics for robotic process automation (RPA)

- Page object (especially table) detection
- Table structure recognition
- Customization

Thank you!

OCR



https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr

Form Recognizer

https://docs.microsoft.com/en-us/azure/cognitive-services/form-recognizer/overview

Contact us

formrecog_contact@microsoft.com