Document VQA

for Data Scientists

Visual NLP Team, John Snow Labs

Agenda

Main topic	Introduced Concepts
Introduction	What is Document Visual Question Answering. Examples. What is zero shot?
Common Architectures	Different architectures; with and without OCR.
Practical Considerations	Building real world document processing pipelines with VQA. Visual NLP. Table Detection.
Summary and next steps	JSL's roadmap on VQA.
Questions & Answers	Questions to discuss the content.

Introduction

- DocVQA is about answering questions in documents, when the visual clues are important.
- It is an *extractive(vs. abstractive) task*.
- May involve multi-step reasoning.
- Mix techniques of computer vision and natural language processing.
- table extraction or key-value pair extraction are blind to the end-purpose the extracted information will be used for.

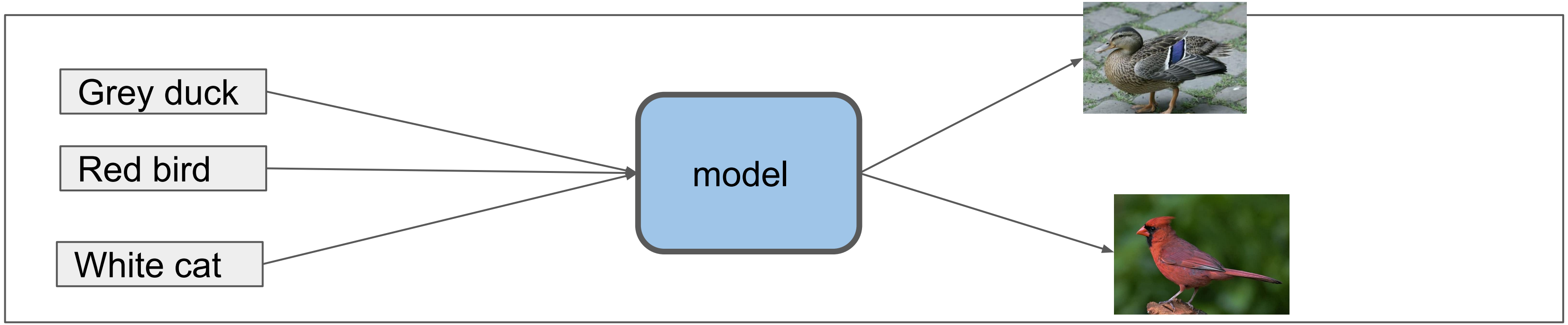
Zero shot what?

- Zero-Shot Learning is a Machine Learning paradigm where a pre-trained model is used to evaluate test data of classes that have not been used during training.

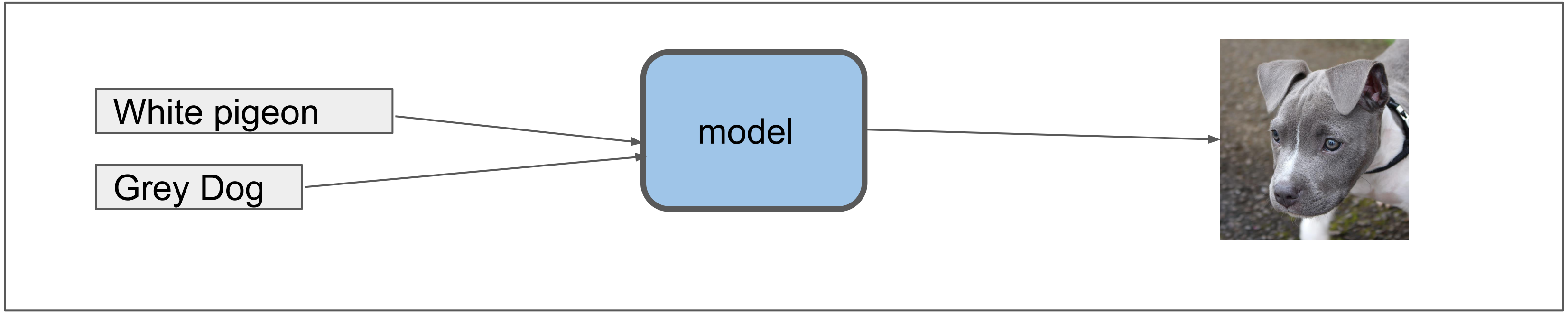
- Ability to predict the results without any training samples.

- **This is not magic!**

Oversimplified example

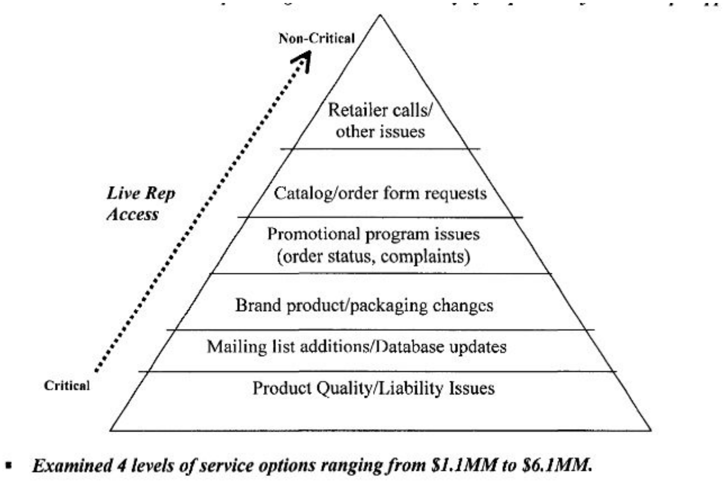


training



inference

Example #1


Non-Critical
Retailer calls/ other issues
Catalog/order form requests
Live Rep Access
Promotional program issues (order status, complaints)
Brand product/packaging changes
Mailing list additions/Database updates
Critical
Product Quality/Liability Issues

- *Examined 4 levels of service options ranging from $1.1MM to $6.1MM.*

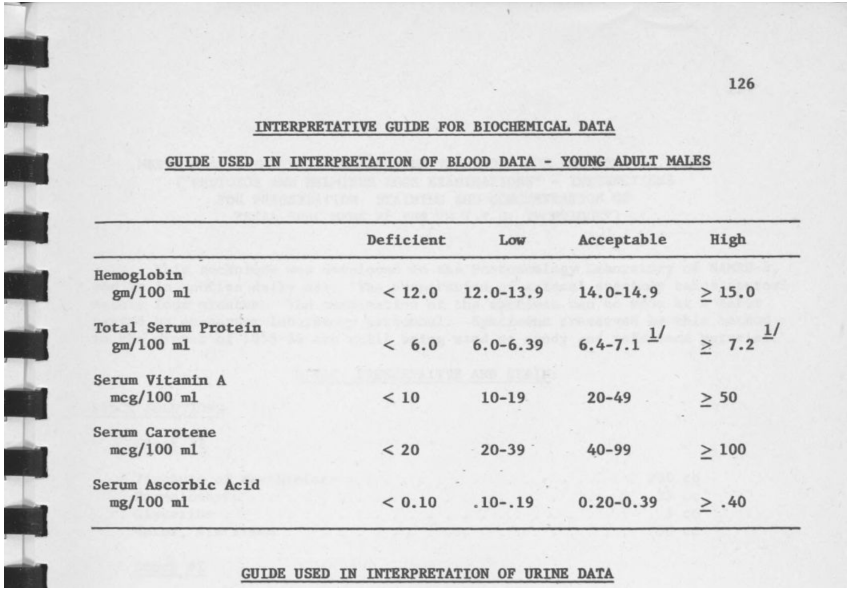
What is the issue at the top of the pyramid? Retailer calls/ other issues

Which is the least critical issue for live rep support? Retailer calls/other issues

Which is the most critical issue for live rep support? Product quality/liability issues

Not only extract and interpret the textual (handwritten, typewritten or printed) content of the document images, but also other visual cues including layout (page structure, forms, tables), non-textual elements (marks, tick boxes, separators, diagrams) and style (font, colours, highlighting).

Example #2



126

INTERPRETATIVE GUIDE FOR BIOCHEMICAL DATA

GUIDE USED IN INTERPRETATION OF BLOOD DATA - YOUNG ADULT MALES

	Deficient	Low	Acceptable	High
Hemoglobin gm/100 ml	< 12.0	12.0-13.9	14.0-14.9	≥ 15.0
Total Serum Protein gm/100 ml	< 6.0	6.0-6.39	6.4-7.1 [1]	≥ 7.2 [1]
Serum Vitamin A mcg/100 ml	< 10	10-19	20-49	≥ 50
Serum Carotene mcg/100 ml	< 20	20-39	40-99	≥ 100
Serum Ascorbic Acid mg/100 ml	< 0.10	.10-.19	0.20-0.39	≥ .40

GUIDE USED IN INTERPRETATION OF URINE DATA

What is the Acceptable Haemoglobin level(g/100ml)?

14.0-14.9

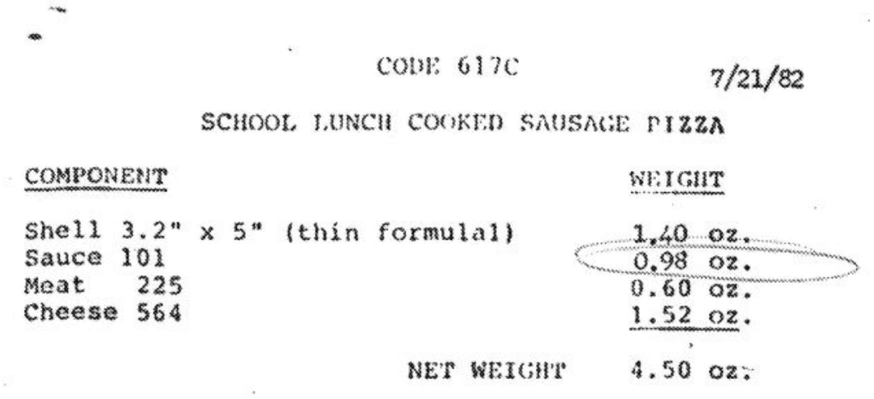
What is the deficiency level for Haemoglobin in blood?

<12.0

What is the acceptable level of Serum Carotene in blood?

40-99

Example #3



CODE 617C 7/21/82

SCHOOL LUNCH COOKED SAUSAGE PIZZA

COMPONENT	WEIGHT
Shell 3.2" x 5" (thin formulal)	1.40 oz.
Sauce 101	0.98 oz.
Meat 225	0.60 oz.
Cheese 564	1.52 oz.
NET WEIGHT	4.50 oz.

What is the number circled?

0.98 oz.

What is the net weight?

Retailer calls/other issues

What is the title of the table?

SCHOOL LAUNCH COOKED SAUSAGE PIZZA

Common Architectures

1. **Roughly 2 types of architectures:** OCR + decoder and encoder/decoder.
2. **Key Points:**
   a. Pre-training objectives.
   b. Attention type that they use.
   c. Order of processing the document.

3. **OCR + decoder:** They use 3 types of features: layout, text, and image.

4. **Encoder-decoder:** they use a visual transformer + language model decoding.

LayoutLM

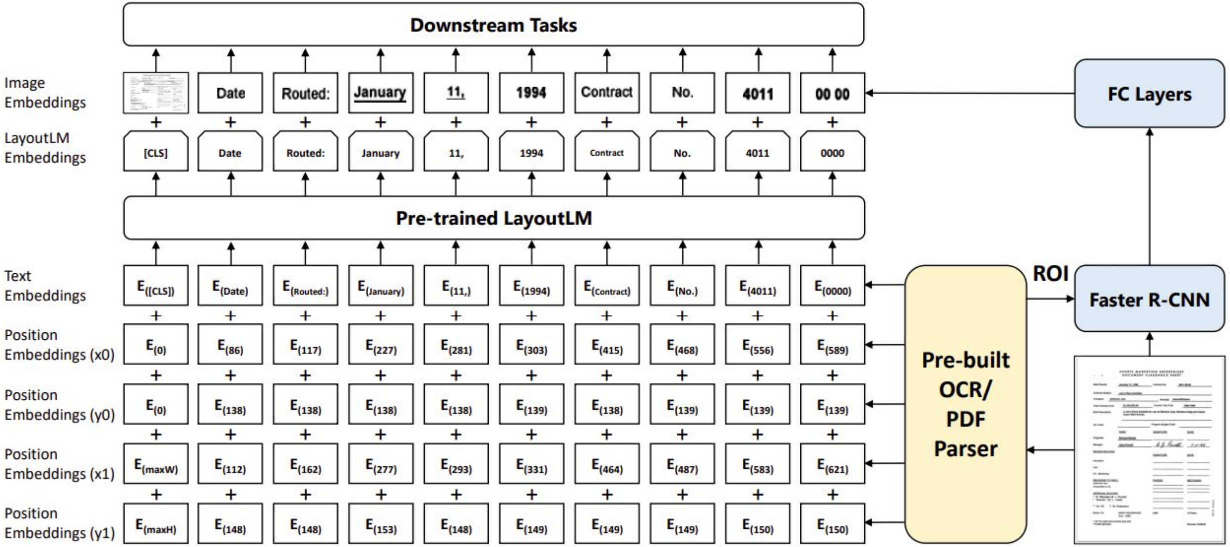
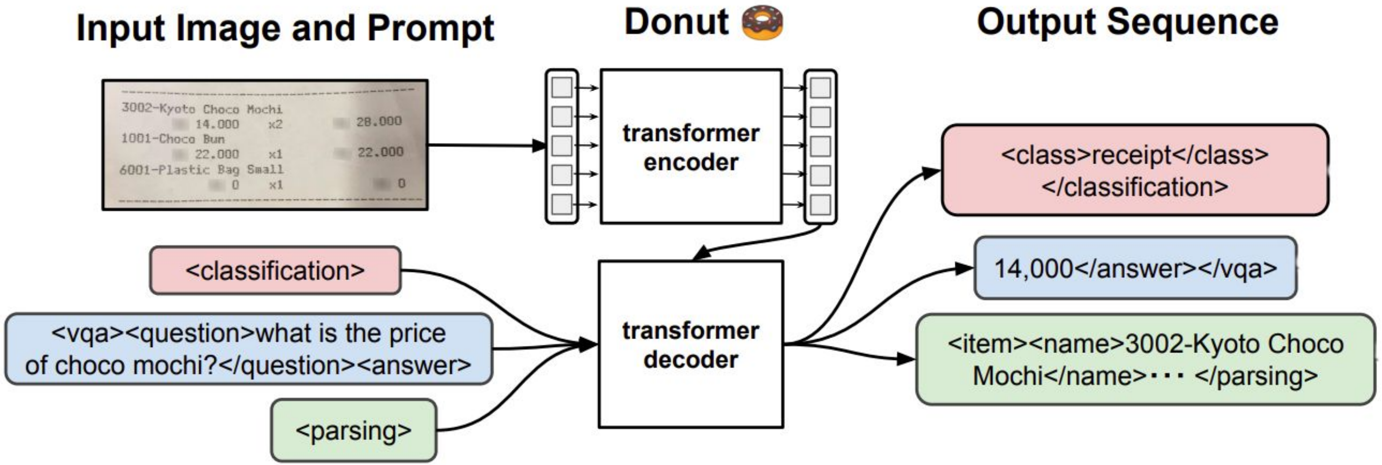


Image Embeddings		Date	Routed:	January	11,	1994	Contract	No.	4011	00 00
LayoutLM Embeddings	[CLS]	Date	Routed:	January	11,	1994	Contract	No.	4011	0000
Text Embeddings	$E_{([CLS])}$	$E_{(Date)}$	$E_{(Routed:)}$	$E_{(January)}$	$E_{(11,)}$	$E_{(1994)}$	$E_{(Contract)}$	$E_{(No.)}$	$E_{(4011)}$	$E_{(0000)}$
Position Embeddings (x0)	$E_{(0)}$	$E_{(86)}$	$E_{(117)}$	$E_{(227)}$	$E_{(281)}$	$E_{(303)}$	$E_{(415)}$	$E_{(468)}$	$E_{(556)}$	$E_{(589)}$
Position Embeddings (y0)	$E_{(0)}$	$E_{(138)}$	$E_{(138)}$	$E_{(138)}$	$E_{(138)}$	$E_{(139)}$	$E_{(138)}$	$E_{(139)}$	$E_{(139)}$	$E_{(139)}$
Position Embeddings (x1)	$E_{(maxW)}$	$E_{(112)}$	$E_{(162)}$	$E_{(277)}$	$E_{(293)}$	$E_{(331)}$	$E_{(464)}$	$E_{(487)}$	$E_{(583)}$	$E_{(621)}$
Position Embeddings (y1)	$E_{(maxH)}$	$E_{(148)}$	$E_{(148)}$	$E_{(153)}$	$E_{(148)}$	$E_{(149)}$	$E_{(149)}$	$E_{(149)}$	$E_{(150)}$	$E_{(150)}$

- *“first time that text and layout are jointly learned in a single framework for document level pre-training”*
- *“LayoutLM further adds two types of input embeddings: (1) a 2-D position embedding that denotes the relative position of a token within a document; (2) an image embedding for scanned token images within a document.”*

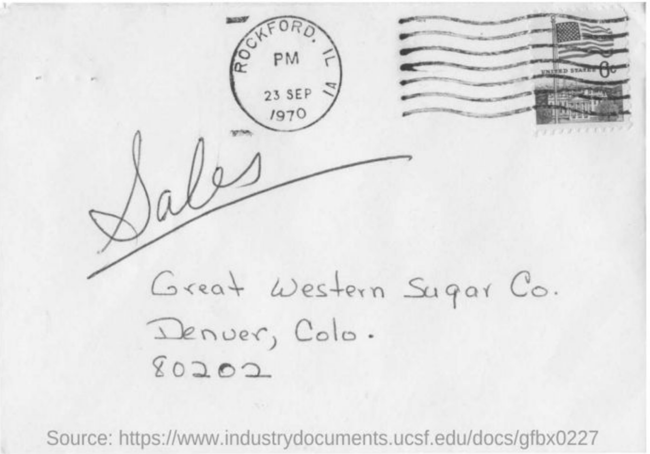
Donut 🍩



- OCR-free VDU model
- Swin Transformer is used for the encoder.
- BART is used as the language decoder.

Practical Considerations

- Creating the questions is difficult.
- You cannot create specific questions to each doc
- No applicable to millions of pages.



ROCKFORD, IL

PM

23 SEP

1970

UNITED STATES 6c

Sales

Great Western Sugar Co.
Denver, Colo.
80202

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

Q: Mention the ZIP code written?
A: 80202

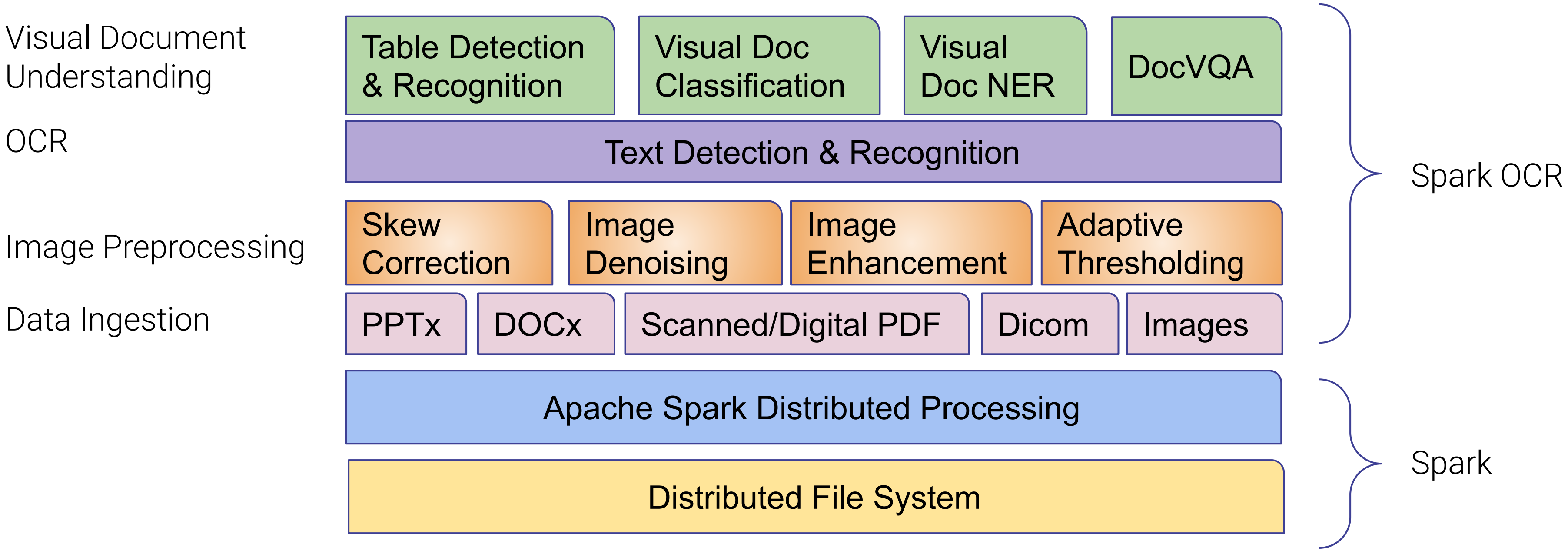
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970

Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

What can be done?

- Identify the documents/sections you care about.
- Apply a limited set of well defined questions covering the information you need.
- Rephrase the questions.
- Use confidence scores.

Visual NLP



Visual Document Understanding: Table Detection & Recognition | Visual Doc Classification | Visual Doc NER | DocVQA

OCR: Text Detection & Recognition

Image Preprocessing: Skew Correction | Image Denoising | Image Enhancement | Adaptive Thresholding

Data Ingestion: PPTx | DOCx | Scanned/Digital PDF | Dicom | Images

Apache Spark Distributed Processing

Distributed File System

Spark OCR

Spark

Examples

QuestionAnsweringOnTables

QuestionAnsweringOnInvoices

Trial license for 10 days, https://bit.ly/Zero_Shot_Visual_NLP

Summary and next steps

We’ve covered…

- The DocVQA task itself.
- The common architectures used to implement models.
- Practical problems & solutions.
- Two implementations of practical pipelines using Visual NLP models.

Summary and next steps

Next steps…

- Add new specialized models(e.g., bar chart understanding, genetic tests).
- Add new (OCR based) architectures.
- Continue to improve accuracy.
- Continue to improve performance and memory consumption.
- Integration with NLP Lab.
- Want to try it yourself? Ask your trial license!: enes@johnsnowlabs.com

Questions & Answers