VisionX

GENG TECHNOLOGY

EXTRACTING STRUCTURED DATA FROM SCANNED PDFS

Team: VisionX

Our Team

ANNY IRUMVA

Team Leader

MARINOS PAPAMICHAEL

Data Preparation

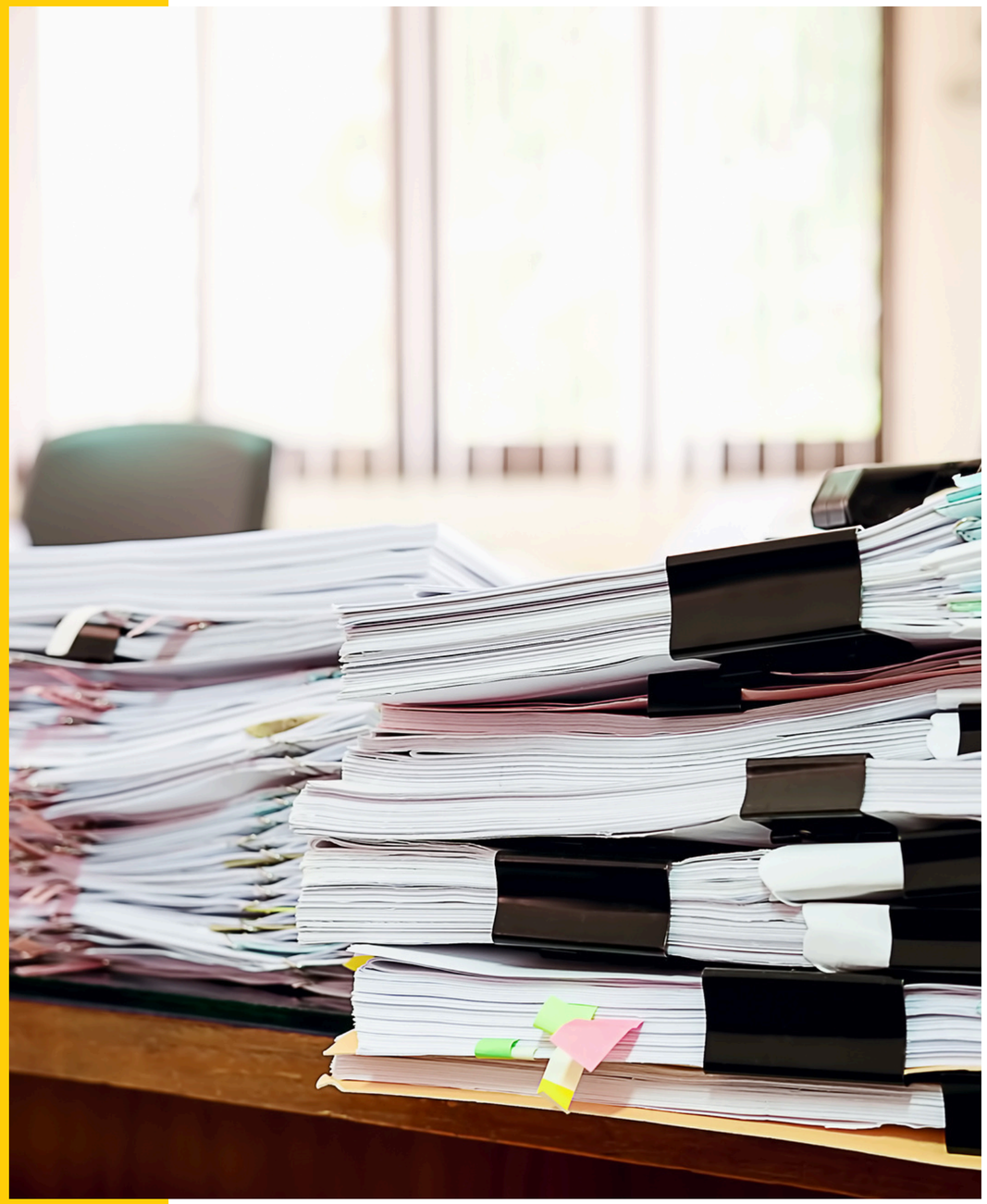
ELIE NIRINGIYIMANA

Modeling

YOUNES ABAROUDI

Deployment

Table of Content

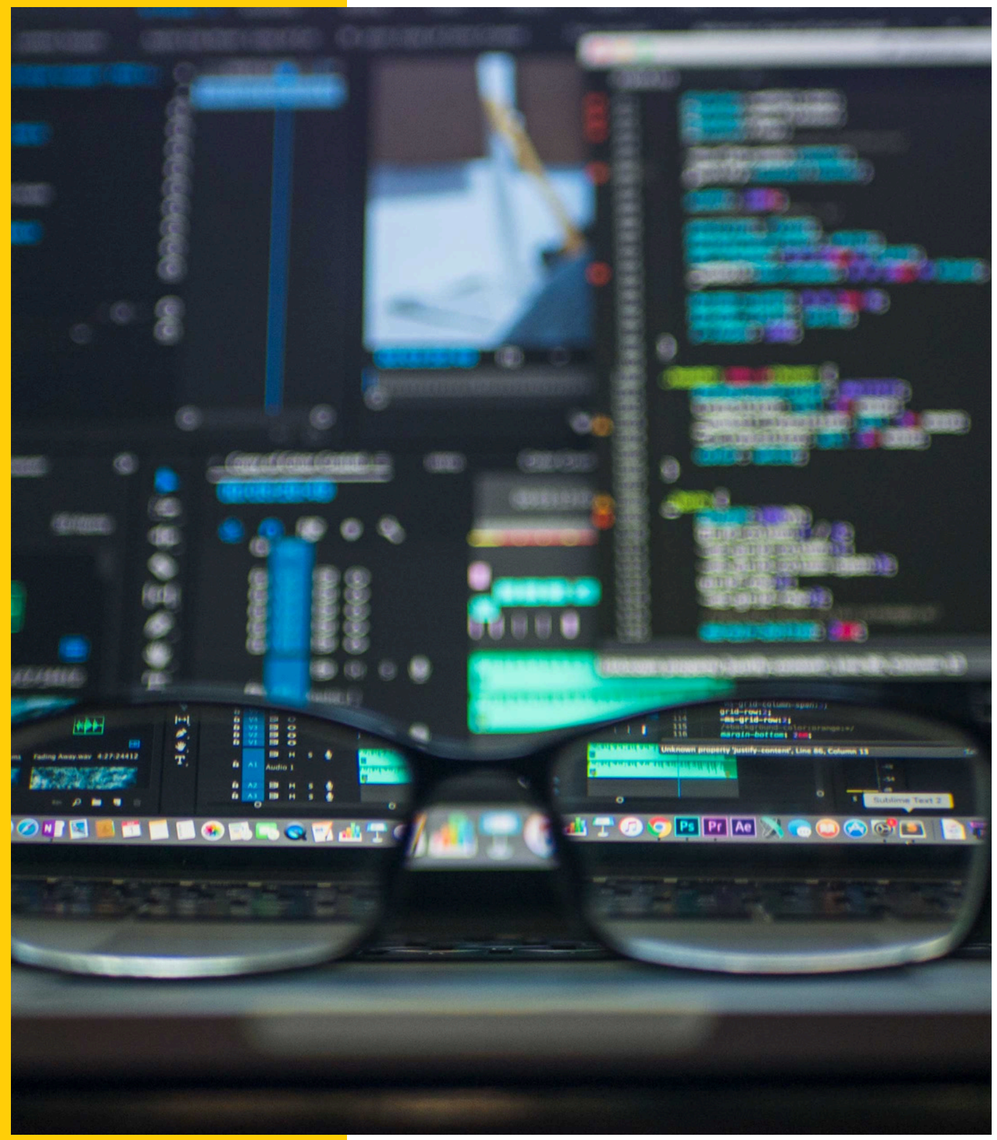


Business Understanding

The company holds a vast collection of scanned PDF documents, including forms, catalogs, and reports, that are difficult to use in their current form
The goal of this project is to unlock the value of this data by accurately extracting and organizing text, images, tables, and charts into structured formats for further analysis and monetization

Key requirements include:

- Extracting text, images, tables, and figures with captions
- Saving the output in organized folders, with JSON files linking related elements for easy reference



Data Preparation

- A range of documents was collected from public datasets available in PDF or JPG formats
- Approximately 2,000 pages were labeled
- Different sections of the documents were annotated using CVAT.ai, applying seven labels: Form, Figure, Image, Text, Table, Page Number, and Header/Footer
- The annotations were exported in COCO format
- Data augmentation was conducted to guarantee an adequate amount of data

Labeling Example

TABLE V
MANTISSAS AND EXPONENT PRE/POST PROCESSING COMPLEXITY OF COMPLEX BLOCK ALU

Block Addition	Mantissas Scaling	Exponents Arithmetic
Complex IEEE754	$4 * N$	$2 * N$
Common Exponent	$4 * N$	2
Exponent Box	$8 * N$	4
Block Multiplication	**Mantissas Scaling**	**Exponents Arithmetic**
Complex IEEE754	$8 * N$	$6 * N$
Common Exponent	$8 * N$	2
Exponent Box	$16 * N$	5
Convolution	**Mantissas Scaling**	**Exponents Arithmetic**
Complex IEEE754	$6 * N_1N_2 + 4 * (N_1 - 1)(N_2 - 1)$	$6 * N_1N_2 + 2 * (N_1 - 1)(N_2 - 1)$
Common Exponent	$6 * N_1N_2 + 4 * (N_1 - 1)(N_2 - 1)$	$3 * (N_1 + N_2 - 1) + 1$
Exponent Box	$10 * N_1N_2 + 8 * (N_1 - 1)(N_2 - 1)$	$3 * (N_1 + N_2 - 1) + 1$

of the complex block output. With Exponent Box Encoding in the worst case, we need eight more mantissas post-scaling. Also, the Shift Vectors allow for four possible intermediate exponent values instead of one intermediate exponent value in Common Exponent Encoding.

C. Complex Convolution

Let $\boldsymbol{X_1} \in \mathbb{C}^{1\times N_1}$, $\boldsymbol{X_2} \in \mathbb{C}^{1\times N_2}$, and $\boldsymbol{Y} \in \mathbb{C}^{1\times(N_1+N_2-1)}$ be complex-valued row vectors, where $*$ denotes convolution, such that,

$$\begin{aligned}\Re\{\boldsymbol{Y}\} &= \Re\{\boldsymbol{X_1} * \boldsymbol{X_2}\}\\ \Im\{\boldsymbol{Y}\} &= \Im\{\boldsymbol{X_1} * \boldsymbol{X_2}\}\end{aligned} \tag{3}$$

We assume $N_1 < N_2$ for practical reason where the model of channel impulse response has shorter sequence than the discrete-time samples. Each term in the complex block output is complex inner product of two complex block input of varying length between 1 and $min\{N_1, N_2\}$. Complex convolution is implemented as complex block multiplication and accumulation of intermediate results. We derive the processing complexity of mantissas and exponents in Appendix .

IV. System Model

We apply Exponent Box Encoding to represent IQ components in baseband QAM transmitter in Figure 5 and baseband QAM receiver in Figure 6. The simulated channel model is Additive White Gaussian Noise (AWGN). Table VI contains the parameter definitions and values used in MATLAB simulation and Table VII summarizes the memory input/output rates (bits/sec) and multiply-accumulate rates required by discrete-time complex QAM transmitter and receiver chains.

A. Discrete-time Complex Baseband QAM Transmitter

We encode complex block IQ samples in Exponent Box Encoding and retain the floating-point resolution in 32-bit IEEE-754 precision in our model. For simplicity, we select block size to be, $N_v = L^{TX} f_{sym}$. The symbol mapper generates a $L^{TX} f_{sym}$-size of complex block IQ samples that shares common exponent. Pulse shape filter is implemented as Finite Impulse Response (FIR) filter of N^{TX}-order and requires complex convolution on the upsampled complex block IQ samples.

TABLE VI
QAM TRANSMITTER, RECEIVER SPECIFICATIONS

QAM Parameters	Definition	Values / Types
Constellation Order	M	1024
Transceiver Parameters	**Definition**	**Values / Types**
Up-sample Factor	L^{TX}, L^{RX}	4
Symbol Rate (Hz)	f_{sym}	2400
Filter Order	N^{TX}, N^{RX}	32^{th}
Pulse Shape	g^{TX}, g^{RX}	Root-Raised Cosine
Excess Bandwidth Factor	α^{TX}, α^{RX}	0.2

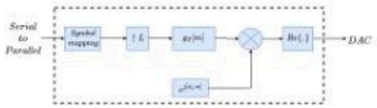


Fig. 5. Block diagram of discrete-time complex baseband QAM transmitter

B. Discrete-time Complex Baseband QAM Receiver

Due to the channel effect such as fading in practice, the received signals will have larger span in magnitude-phase response. The Common Exponent Encoding applied on sampled complex block IQ samples is limited to selecting window size of minimum phase difference. The Common Exponent Encoding must update its block size at the update rate of gain by the Automatic Gain Control (AGC). Instead, our Exponent Box Encoding could lift the constraint and selects fixed block size, $N_v = L^{RX} f_{sym}$ in this simulation. We simulate matched filter of N^{RX}-order.

V. Simulation Results

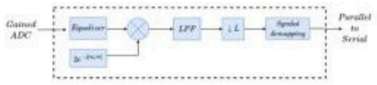
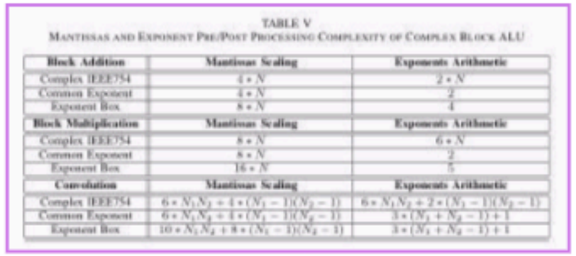
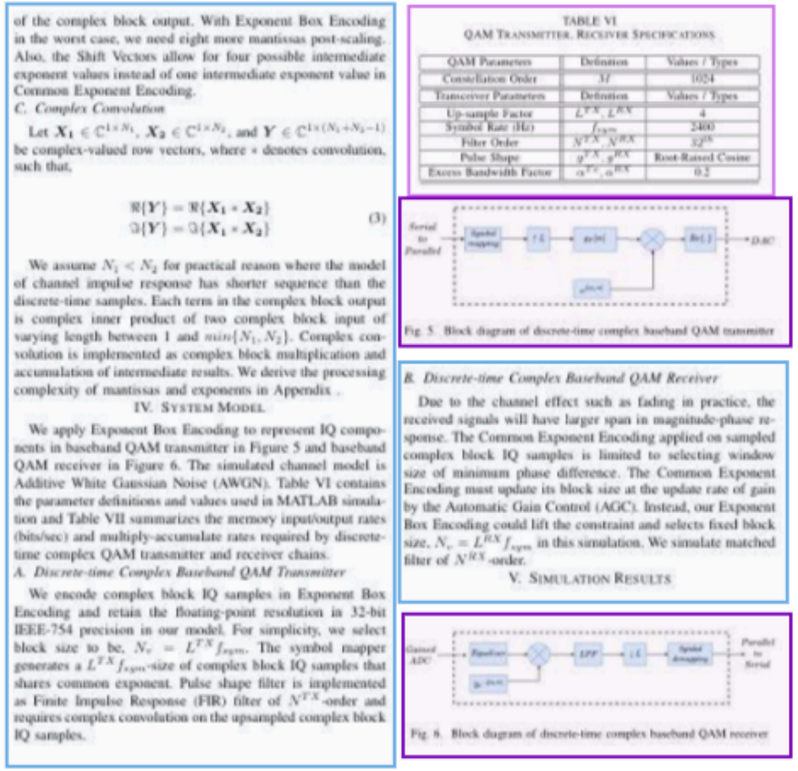
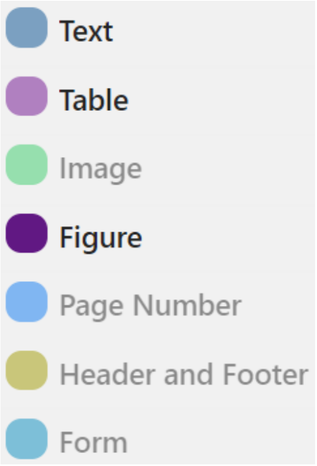


Fig. 6. Block diagram of discrete-time complex baseband QAM receiver







Modeling

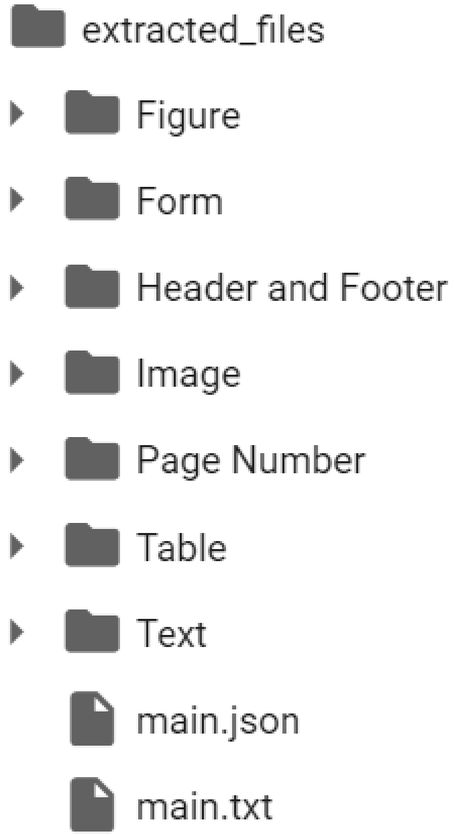
Model Overview:

The project utilizes a **Faster R-CNN** model with a **ResNet-50-FPN** backbone, fine-tuned for a custom dataset using **Detectron2**
This model is effective for detecting multiple objects in images, balancing speed and accuracy

Techniques & Modifications:

- **Early Stopping**: Stops training when validation loss stops improving
- **Data Augmentation**: Resizing, flipping, brightness/contrast adjustments, rotations
- **Gradient Clipping**: Stabilizes training, prevents gradient explosion
- **Warmup & LR Decay**: Gradual learning rate adjustments for smooth convergence
- **Increased Batch Size**: Optimizes training efficiency

Extracted Data Structure



Structure of the output

```
The company holds a large volume of scanned PDF documents that contain valuable
information (e.g., forms, catalogues, reports, financial documents, etc.). These documents are
challenging to use in their current form. The business wants to unlock the value of this
information by extracting it into structured formats that can be monetized, analysed, and
integrated into other tools. The challenge is to accurately extract and organize text, images,
tables. and captions, while preserving their context and relationships.

Sdfghjkoiuytr

corporation controlled directly or indirectly by sucn a creator or contributor nas or will
[Link to Figure 2: /content/extracted_document/Figure/Figure_2.jpg]

The initial extraction doesn't need to be highly granular (e.g., splitting all components). However,
the company may later want to further classify or divide extracted content into finer details.

Patent applications that landed on Einstein's desk for his evaluation included ideas for a gravel
sorter and an electric typewriter. His employers were pleased enough with his work to make his
position permanent in 1903, although they did not think that he should be promoted until he had
"fully mastered machine technology". It is conceivable that his labors at the patent office had a
bearing on his development of his special theory of relativity. He arrived at his revolutionary
ideas about space, time and light through thought experiments about the transmission of signal:
```

Example of main.txt

Value to Business

MONETIZATION OF DATA

The extracted and structured information can be monetized by integrating it into analytics platforms or selling it as valuable datasets, creating new revenue streams for the business

OPERATIONAL EFFICIENCY

Automating the extraction process reduces manual work, improving operational efficiency and allowing employees to focus on more strategic tasks

IMPROVED DATA ACCESSIBILITY

By transforming unstructured PDFs into structured formats, the company can easily access, analyze, and integrate the data into other tools, enhancing usability and decision-making

ENHANCED ANALYTICAL INSIGHTS

Extracted data from financial reports, forms, and catalogs can provide deeper insights and trends, helping the company make more informed business decisions

Challenges

Limited Computing Power: We were unable to test multiple models due to computing power constraints, limiting our ability to optimize the solution

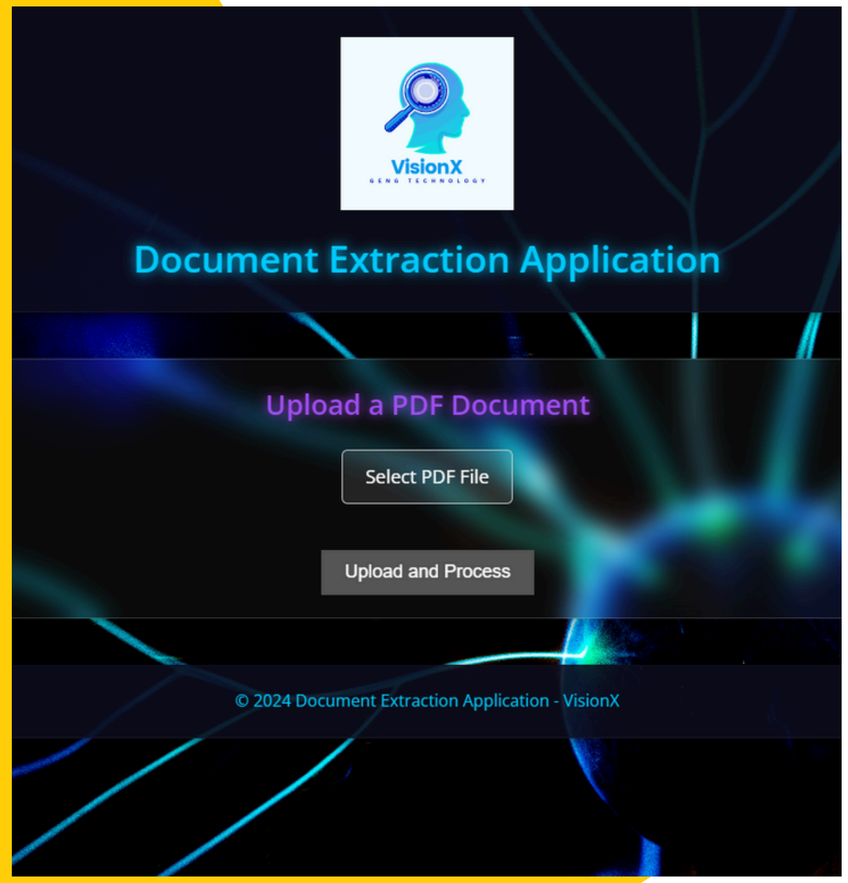
Time Constraints for Labeling: The project's short duration restricted the amount of labeled data we could produce, impacting the model's performance

Difficulty Obtaining Raw Data: Sourcing enough diverse and suitable raw data was a challenge, which limited the variability of our training set

Conclusion

- The goal was to develop a model capable of extracting and organizing text, images, tables, and figures along with their captions
- We utilized various documents from public datasets available in PDF or JPG formats
- The Faster R-CNN model was employed and fine-tuned using multiple methods
- We are delivering an effective model that successfully extracts text, images, figures, and tables along with their captions with minimal error
- This solution will provide significant value to the business by facilitating data monetization and enhancing data accessibility

Deployment and Demo



TECHNOLOGY CHOSED

- Backend: Flask
- Frontend: HTML, CSS, JavaScript

KEY FEATURES

- **File Upload**: Users can upload PDFs via a clean, intuitive interface
- **Output**: A downloadable ZIP file is generated with structured content (text, images, tables)
- **Automation**: Previous files are cleared automatically before each new upload to ensure clean results
- **UX Optimization**: Page automatically refreshes post-download to reset the interface for the next upload

Thank you
Any Question?