Information Extraction from Invoices

By-
Adarsh Ghimire
Nhujaw Tuladhar

Problem Statement

- Manual invoice processing is time-consuming and error-prone.
- Variations in invoice templates make rule-based methods brittle.
- Need for an automated solution to extract structured data from invoices.

Objectives

- Develop an AI-based information extraction system that can parse invoice images and extract key-value pairs and tabular data.
- Leverage AI models to handle invoices of varying layouts and formats.
- Output structured JSON matching the dataset's defined schema.
- Build a UI for interactive demonstration of invoice parsing.

Dataset

- So for training <u>Data</u> for Information Extraction we used katanaml-org (invoices-donut-data-v1). Source: Hugging Face.
- Each sample includes an invoice image and its structured JSON ground truth.
- Enables supervised fine-tuning for accurate field extraction.
- Total Samples: 501 (Train: 425, Val: 50, Test: 26)

Invoice no: 40378170

Date of issue: 10/15/2012

Seller:

Patel, Thompson and Montgomery
356 Kyle Vista
New James, MA 46228

Tax Id: 958-74-3511
IBAN: GB77WRBQ31965128414006

Client:

Jackson, Odonnell and Jackson
267 John Track Suite 841
Jenniferville, PA 98601

Tax Id: 998-87-7723

ITEMS

No.	Description	Qty	UM	Net price	Net worth	VAT [%]	Gross worth
1.	Leed's Wine Companion Bottle Corkscrew Opener Gift Box Set with Foil Cutter	1,00	each	7,50	7,50	10%	8,25

SUMMARY

	VAT [%]	Net worth	VAT	Gross worth
	10%	7,50	0,75	8,25
Total		$ 7,50	$ 0,75	$ 8,25

{"gt_parse": {"header": {"invoice_no": "40378170", "invoice_date": "10/15/2012", "seller": "Patel, Thompson and Montgomery 356 Kyle Vista New James, MA 46228", "client": "Jackson, Odonnell and Jackson 267 John Track Suite 841 Jenniferville, PA 98601", "seller_tax_id": "958-74-3511", "client_tax_id": "998-87-7723", "iban": "GB77WRBQ31965128414006"}, "items": [{"item_desc": "Leed's Wine Companion Bottle Corkscrew Opener Gift Box Set with Foil Cutter", "item_qty": "1,00", "item_net_price": "7,50", "item_net_worth": "7,50", "item_vat": "10%", "item_gross_worth": "8,25"}], "summary": {"total_net_worth": "$7,50", "total_vat": "$0,75", "total_gross_worth": "$8,25"}}}

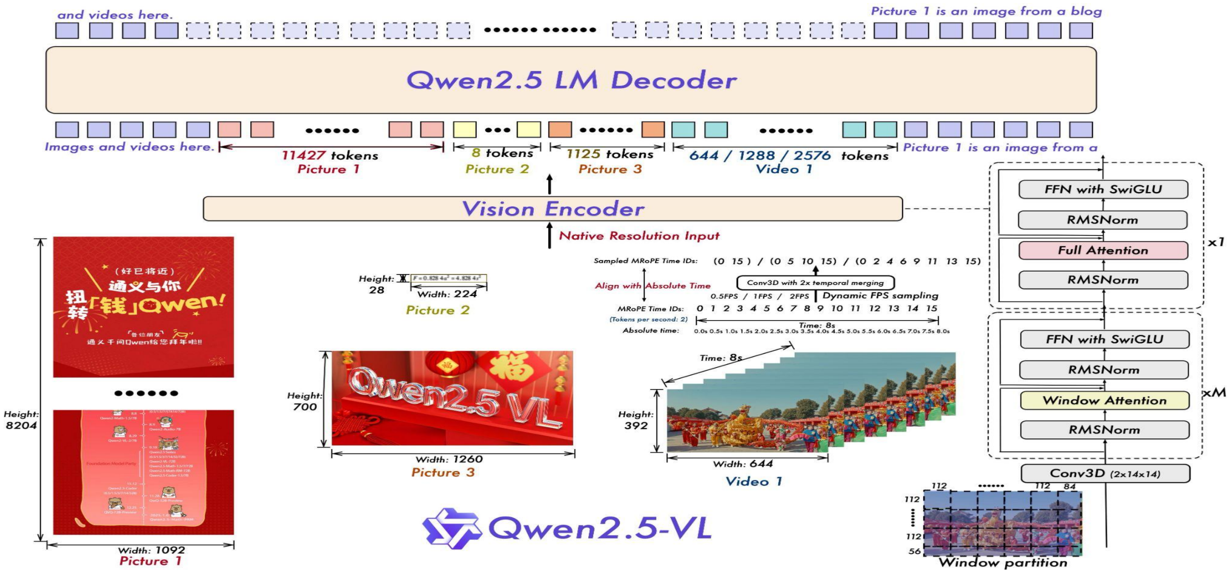
Model

We chose Qwen 2.5 VL because it is a cutting-edge Multimodal Large Language Model (MLLM) that processes both images and text.

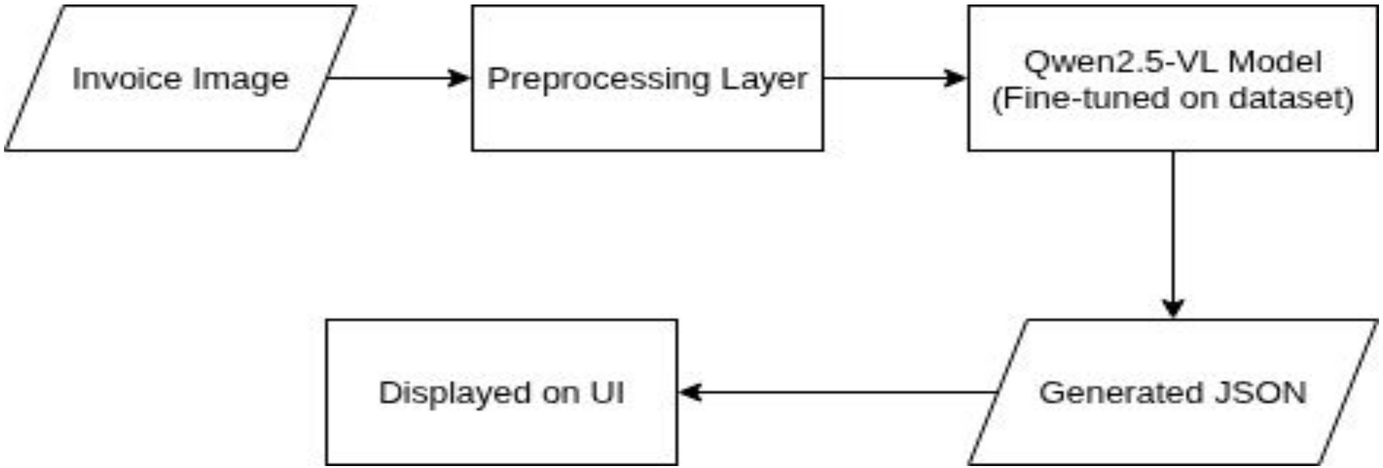
Ideal for Invoices due to:

- **Visual Understanding:** Interprets invoice layouts, spatial relationships, and text appearance.
- **Language Understanding:** Extracts specific entities and interprets textual context.
- **Contextual Reasoning:** Fuses visual and text cues for robust, accurate extraction across variations.
- **Efficient Fine-tuning:** Leverages strong pre-trained knowledge for rapid task adaptation.

Model Architecture



System Architecture



Fine-Tuning Process: Overview & Base Model

- **Base Family:** Qwen2.5-VL (multimodal vision-language model).
- **Checkpoint Used:** Pre-trained *Qwen/Qwen2.5-VL-3B-Instruct* as foundation.
- ~3B parameters; strong visual understanding while GPU-friendly.
- Accepts image + instruction text.
- Loaded in **4-bit precision** for efficiency (***`load_in_4bit=True`***).
- Enables training on modest GPUs without severe performance loss.

Fine-Tuning Process: PEFT with LoRA

Challenge: Full fine-tuning large VLMs is compute-heavy.

So, we used **Unsloth FastVisionModel + LoRA (PEFT)** to update only lightweight adapter ranks instead of all weights.

Comprehensive Adaptation: We enabled LoRA for both *vision* and *language* paths:

- *`finetune_vision_layers`*
- *`finetune_language_layers`*
- *`finetune_attention_modules`*
- *`finetune_mlp_modules`*

LoRA Config: *`r=32`*, *`lora_alpha=32`*.

Fine-Tuning Process: Hyperparameters

We used:

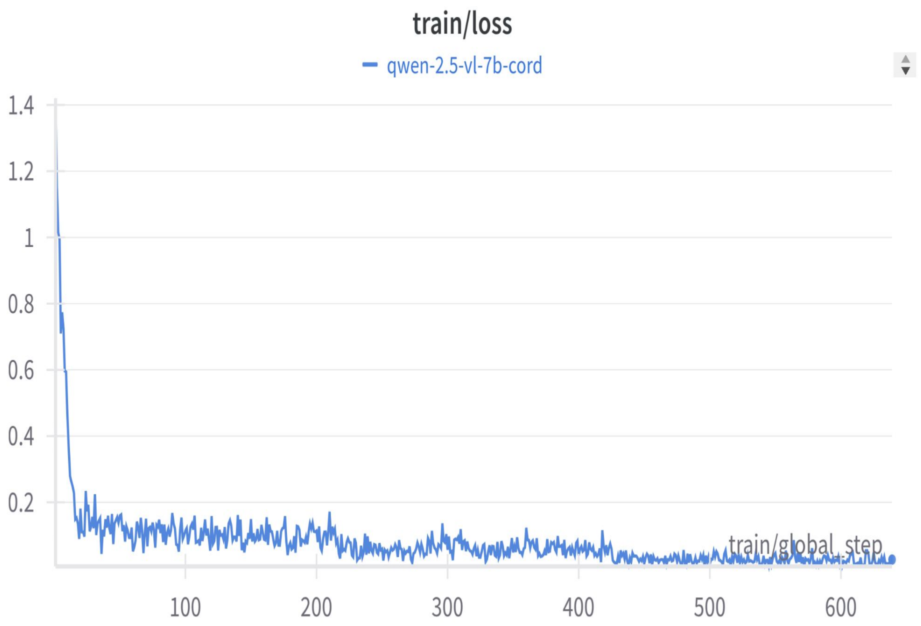
- Hugging Face **TRL** `SFTTrainer` for supervised fine-tuning.
- **UnslothVisionDataCollator** for multimodal batching.

Key Hyperparameters:

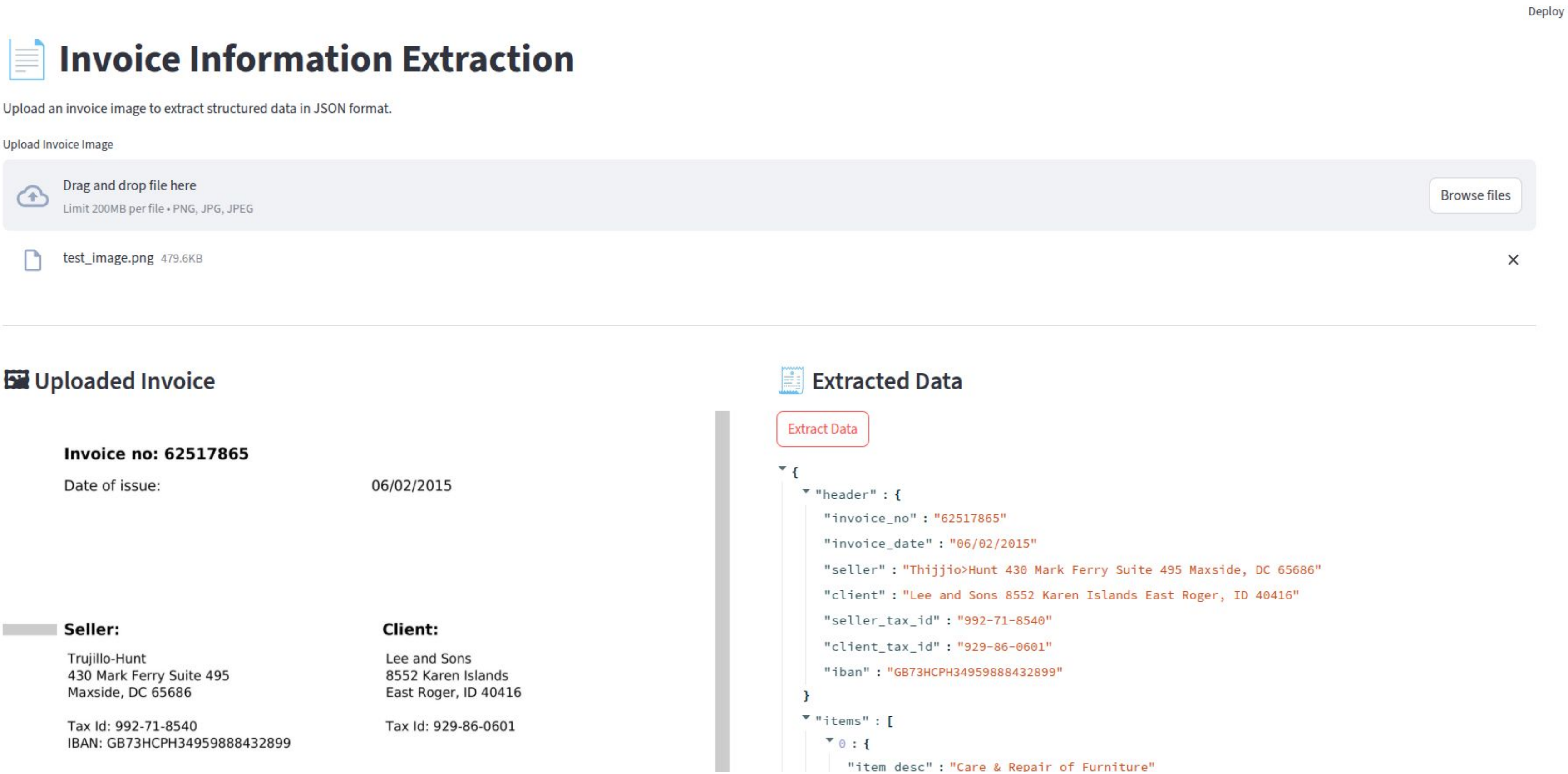
- Epochs: **3**
- Learning Rate: **2e-4**
- Optimizer: **adamw_8bit**
- Batch Size: **1 / device** (gradient_accumulation=1)
- Max Seq Len: **1024 tokens**
- Mixed Precision: **bf16** or **fp16** fallback

Training Loss

- **Observed a rapid initial decrease followed by a steady decline in training loss**, indicating the model quickly adapted and continuously refined its understanding of invoice data.
- **The curve's eventual flattening demonstrates successful convergence and effective fine-tuning**, confirming the model's robust learning of information extraction patterns.



User Interface



Evaluation

We employed a multi-faceted evaluation strategy combining text similarity metrics with structured field accuracy.

ROUGE Scores:

- Used to evaluate overlap and quality of generated text vs. ground truth.
- Achieved:
  - **ROUGE-1:** 0.9718
  - **ROUGE-2:** 0.9518
  - **ROUGE-L:** 0.9706
  - **ROUGE-Lsum:** 0.9705
- These high ROUGE values indicate strong textual fidelity.

Evaluation [contd..]

Custom Structured Accuracy Metrics:

- Developed a field-level evaluation script to measure exact matches, data type handling, and normalization.
- Provides granular insight into which fields are extracted correctly.

Overall Structured Extraction Accuracy (Test Set – 26 invoices):

- **Micro Accuracy:** 0.8939 (≈89.39%)
  - Percentage of all individual fields correctly extracted across all test invoices.
- **Macro Accuracy:** 0.8606 (≈86.06%)
  - Average accuracy per field type, preventing dominant fields from skewing results.
- **Prediction Parse Failures:** 0
  - The model always produced valid JSON, ensuring reliability in downstream pipelines.

Future Work & Enhancements

- Enhance accuracy for challenging invoices and refine error handling.
- Broaden the dataset to include more diverse invoice types and support additional languages/regions.
- Implement strategies for ongoing model improvement with new data and feedback.

Thank You