AI/ML Internship Report
Medvarsity, Apollo Hospitals

Aryaman Bahl
Guided by Dr. Sujoy Kar

1 July - 31 July 2025

1. Initial Data Exploration and Preparation

The project began with extensive data wrangling on raw medical data provided by Apollo Hospitals. This included cleaning, editing, and merging multiple CSV files into a consolidated dataset suitable for model training. Processed outputs were validated by Dr. Sujoy Kar before proceeding further.

2. OCR Model Benchmarking

To establish a performance baseline, four OCR models—TrOCR, Donut, Pix2Struct, and EasyOCR—were benchmarked generating synthetic medical documents using faker library from python.

Metrics: Word Error Rate (WER), Character Error Rate (CER), Fuzzy Accuracy, and Exact Match Accuracy.

Findings:

- **EasyOCR** performed best out-of-the-box (WER: 0.40, CER: 0.10, Fuzzy Accuracy: 70%).
- Transformer-based models performed poorly on long, structured text without domain-specific fine-tuning.
- Exact match accuracy was 0 across all models, as expected given document-level complexity.

Conclusion: EasyOCR provides a strong baseline. However, advanced models require fine-tuning with task-specific formatting (e.g., line-level crops) for real-world deployment.

3. Florence-2 Fine-Tuning Methodology

3.1 Environment Setup

Fine-tuning was performed on a T4 GPU using the following libraries:

- `torch==2.3.0`, `transformers==4.41.2`
- `peft==0.11.1`, `bitsandbytes==0.43.1`

3.2 Data Preprocessing

Each image (.png/.jpg) was paired with a .json annotation. A custom function `create_dataset_df` created a DataFrame, later converted into a HuggingFace `DatasetDict`.

A key function, `process_example`, converted annotation coordinates and text into Florence-2 compatible format: `<OD><x1><y1><x2><y2>text...`.

Invalid or empty annotations were filtered out to ensure data integrity.

3.3 PEFT-Based Fine-Tuning

An iterative fine-tuning strategy was used to accommodate hardware constraints.

Initial Attempt (4-bit QLoRA):

- Used 4-bit quantization with `fp16=True`.
- Training failed due to a gradient unscaling error caused by incompatibility between FP16 gradients and the 4-bit optimizer.

Revised Strategy (8-bit LoRA):

- Switched to 8-bit precision using `load_in_8bit=True`.
- Applied Low-Rank Adaptation (LoRA) using `peft`, reducing trainable parameters to 7.3M.
- Config: `r=16`, `lora_alpha=32`, target: all linear layers.
- Enabled gradient checkpointing for VRAM optimization.

3.4 Error Diagnosis

The training failure stemmed from a type conflict within PyTorch's AMP system. The GradScaler expected `float32` gradients but received `float16`, due to the interaction between 8-bit quantization and FP16 training.

4. Resolution and Future Work

4.1 Immediate Solutions

- **Disable AMP:** Setting `fp16=False` enables stable FP32 training on the T4 GPU.
- **Use better hardware:** Transitioning to an A100 or H100 GPU would enable efficient mixed-precision training with quantization.

4.2 Research Direction: V-LoRA

For more robust fine-tuning, implementing V-LoRA (Han Wang et al., 2025) is recommended. Unlike standard LoRA, V-LoRA allows adaptive decomposition across vision and language components, potentially improving downstream performance in OCR tasks.

4.3 Evaluation Metrics (Post-Training)

Once training is successful:

- **Object Detection:** Evaluate using Intersection over Union (IoU).
- **OCR Accuracy:** Assess using CER and WER.

5. Conclusion

This internship involved full-cycle OCR benchmarking and VLM fine-tuning tailored to medical document understanding. While EasyOCR serves as a strong baseline, Florence-2 requires tailored tuning. The project's infrastructure and codebase are complete, with clear pathways identified for final model training, evaluation and clear research direction.

Appendix

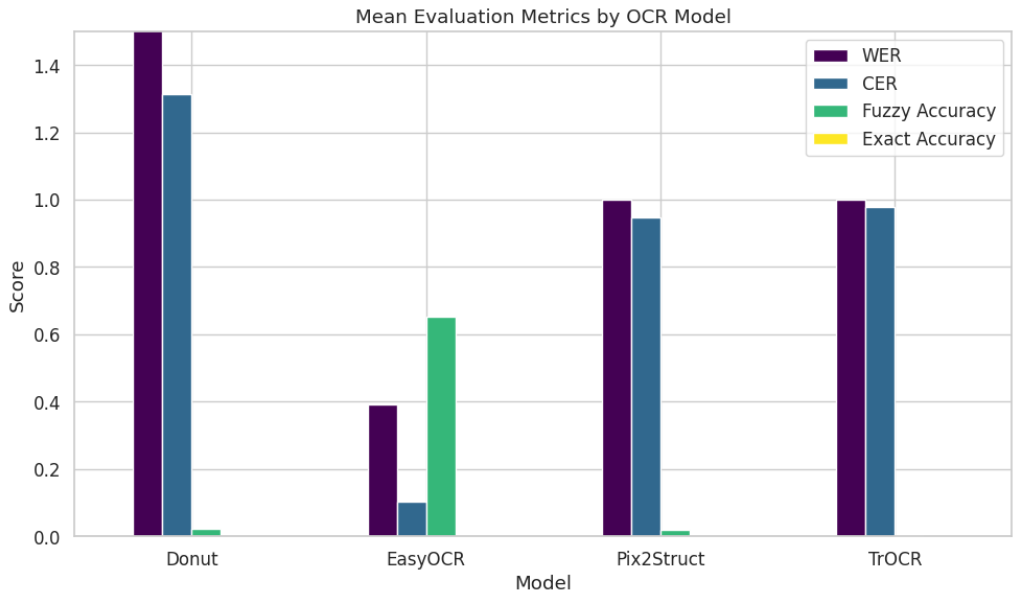

Figure A1: Mean evaluation metrics (WER, CER, Fuzzy Accuracy, Exact Accuracy) across OCR models.

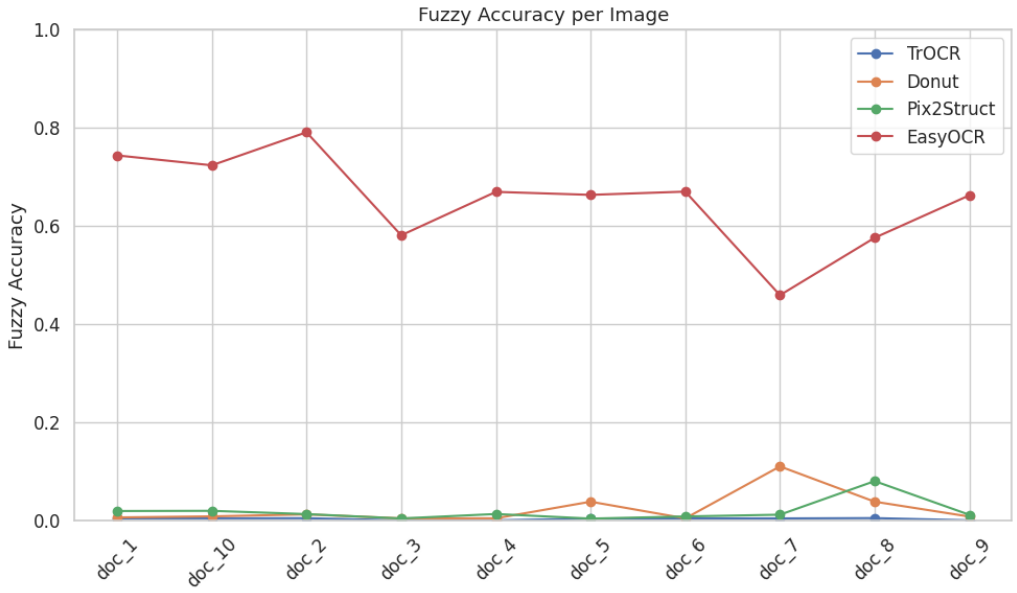

Figure A2: Fuzzy Accuracy across 10 synthetic documents. EasyOCR maintains high consistency; transformer models underperform.

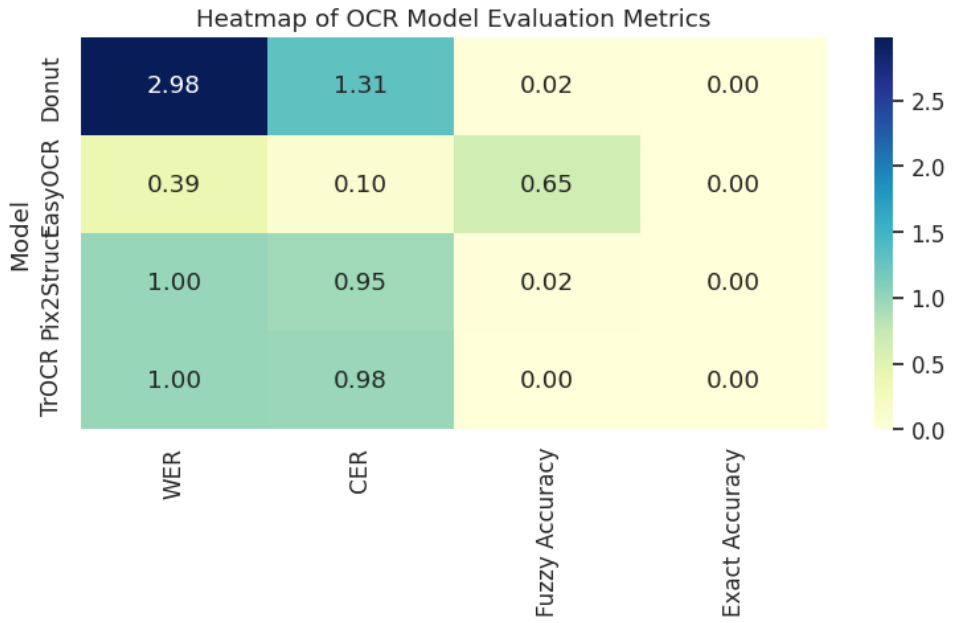

Figure A3: Attention Heatmap

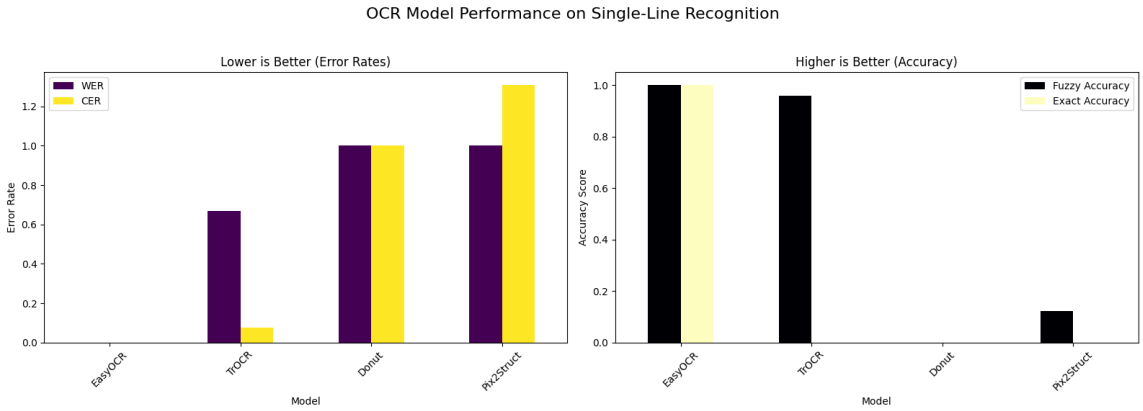

Figure A4: OCR performance on single-line text: error rates (left) and accuracy (right). EasyOCR clearly dominates on this simpler task.