Amazon ML Challenge Finale

NeuralNinjas

Kushal Agrawal

Indian Institute of
Technology Jodhpur

Alli Khadga Jyoth

Indian Institute of
Technology Jodhpur

Nachiketa Purohit

Indian Institute of
Technology Jodhpur

Ritu Singh

Indian Institute of
Technology Jodhpur

Problem Statement

The challenge is to build a machine learning model that extracts entity values, such as weight, volume, and dimensions, etc., directly from product images

Why not OCR+NER?

- Unable to interpret unlabeled values
- Fails to capture spatial context & positional meaning

Data Preprocessing:

- Replaced entity values having invalid units with “NA“ (e.g., horsepower)
- Replaced range with max value (e.g., [24, 30] volt with 30 volt)

Image				
Entity Name	**Depth**	**Width**	**Voltage**	**Wattage**
Train Label	15 cm	9 cm	50 volt	3.0 horsepower
Accurate Label	NA	32 cm	NA	NA

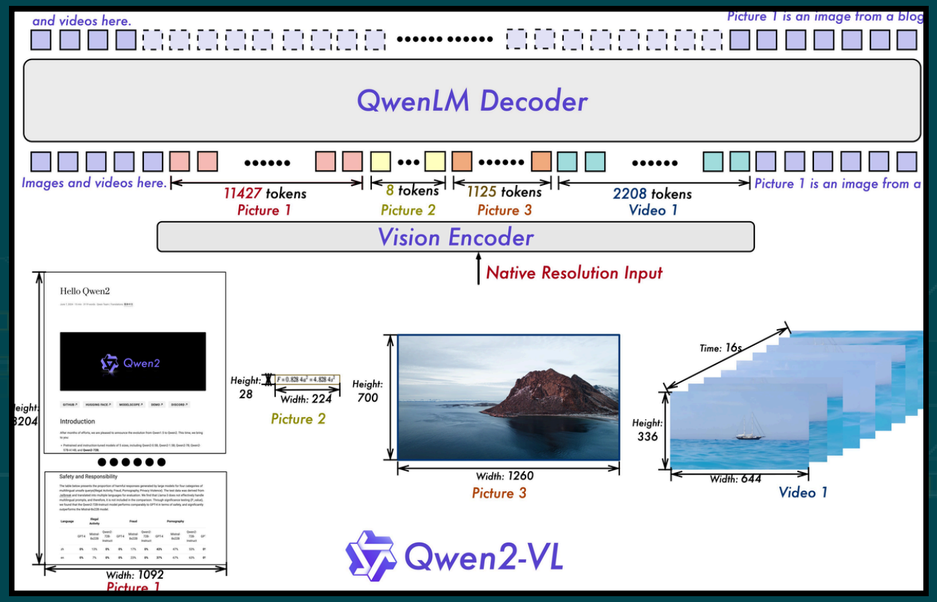
Our Approach

Model

Why Qwen2VL model?

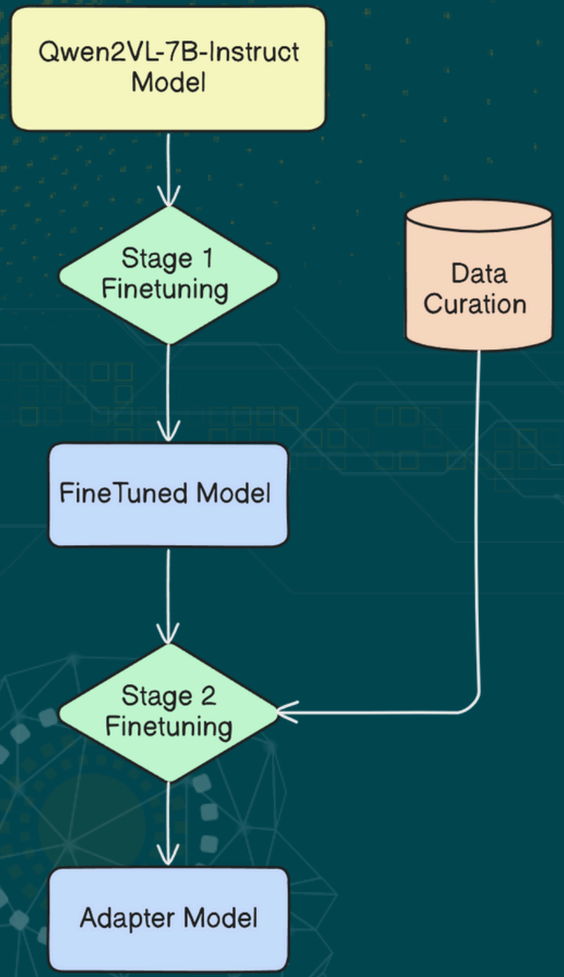
1) Pre-trained on VQA and OCR datasets
2) Vision-Language Alignment
3) Dynamic Resolution
4) Open-Source

Base Model: **Qwen2VL-7B-Instruct**



Benchmark	InternVL2-8B	MiniCPM-V2.6	GPT-4o-mini	Qwen2-VL-7B
$DocVQA_{test}$	91.6	90.8	-	94.5
OCRBench	794	852	785	845

Two Stage Learning



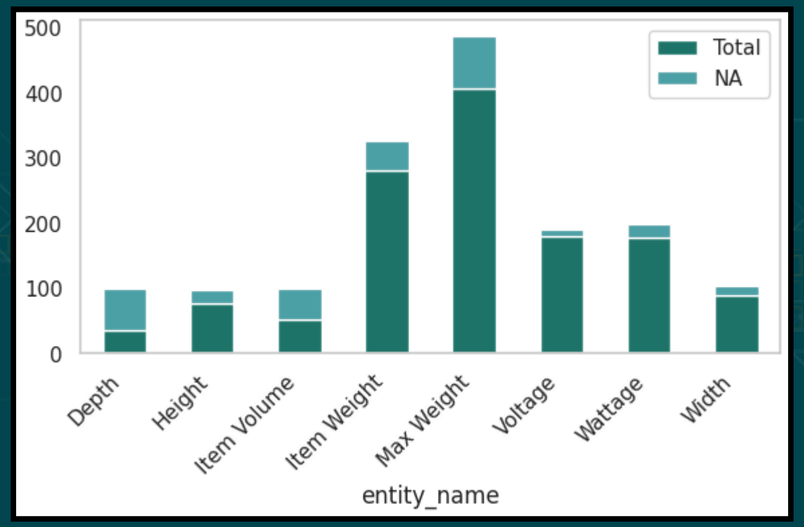
First Stage: Fine-tuned on 20k samples

- Adapted to the domain of product images and entity extraction
- Learned broad patterns and relationships

"Garbage in, Garbage out"

Data Curation:

- Curated 1600 samples
- Assigned "NA" when entity value was absent in an image
- Corrected inaccurate entity values



Second Stage: Fine-tuned on curated Dataset

- Refined model's understanding
- Corrected biases introduced by noisy labels

FineTuning Setup

Base Model: Qwen2VL-7B-Instruct
Fine-Tuning Method: QLoRA- 8 bit
Prompt: What is the {entity_name}?
Learning Rate: 5e-5
Compute: 2 A100 40GB
Post Processing:

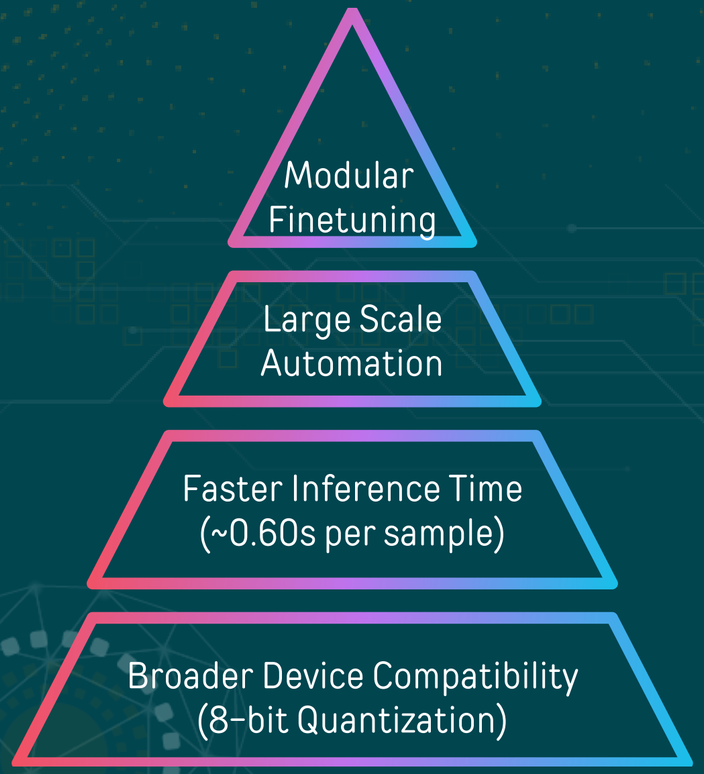
- Corrected invalid entity units
- Replaced “NA” values with empty string

Hyperparameter	Stage 1	Stage 2
Epochs	3	30
Batch Size	8	4
Gradient Accumulation	8	4
LR Scheduler	Cosine	ReduceLROnPlateau

Results

Model / FineTuning Method Qwen2VL-7B-Instruct (M0)	F1 Score
Baseline Qwen2VL-7B-Instruct AWQ (M1)	0.617
Fine-Tuned M0 model on 10k samples (M2)	0.678
Fine-Tuned M0 model on 20k samples (M3)	0.679
Fine-Tuned M3 model on curated 1600 samples (M4)	**0.865**

Scalability



Future Scope