Applied ML & Sense Tools

Charvi Shukla: B.S. Math-CS, 2025, M.S. CSE, 2027

Manager and Mentor: Tushar Jagtap, Harsh Karia

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Product Recognition (Espio) Scoring and Ranking



Problem: Current Product Recognition system is heavily reliant on closed-source frontier VLMs to aid in open-vocabulary product classification.
1. High token usage → High cost $
2. High reliance on third party VLM provider
3. Prone to hallucinations and unstructured outputs

For each crop:  → AI Powered Product Recognition System → Product GTIN: 70501013304

Solution 1: VLM Fine Tuning

Idea: Can use open-source VLMs as a transparent, and domain-adapted solution.
Solution: Fine tuned **Unsloth's** 4 Bit quantized models on **1x NVIDIA A100 40GB GPU** and used **vLLM** inference engine for fast multimodal inference.
Results:

Model	Top-1-acc	Inference time	Fine Tuned top-1-acc	Inference with vLLM
Open Source Model 1	43.43%	~17.00s/it	**74.63%** (↑31.2%)	**7.96s/it** (↓ 9.31s)
Open Source Model 2	58.33%	~19.60s/it	**89.00%** (↑30.7%) **(on a subset)**	**3.61s/it** (↓ 15.99s)
Open Source Model 3	37.33%	~35.00s/it	**76.52%** (↑39.2%)	**1.78s/it** (↓ 33.22s)

Personal and Professional Growth

Learnings:
- Working with proprietary datasets, SQL
- Interacting with the open source community and creating GitHub Issues
- Fine tuning and vLLM inference and meaningful experimentation

Useful CSE and Math classes:
- MATH 170 series: Numerical analysis
- MATH 102: Advanced Linear Algebra
- CSE 152A and 252D: Computer Vision
- CSE 110: Software Engineering

Ethics & Society

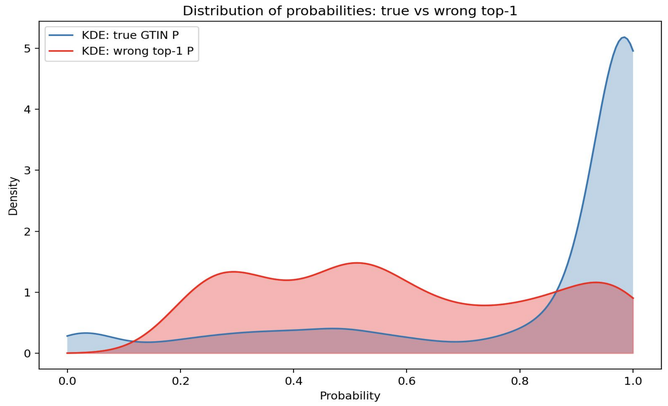
- **Fairness** in AI ensures product recognition works equally well across diverse brands and packaging. Thus, I always aim to make the model as generalizable as possible.
- Reducing reliance on large VLMs also lessens the **environmental impact** of massive data centers on our planet.

Solution 2: VLMs + Ranking Fusion Algorithm

Part 1: Proof of Concept

Idea: VLMs are better at image captioning tasks than reasoning tasks. Use them for structured information extraction and use a fusion algorithm to get final GTIN assignments and confidences
Why: (1) Faster inference due to smaller context and output length (2) Lower cost
How: Leverage different modalities as contributors a rank-fusion algorithm.

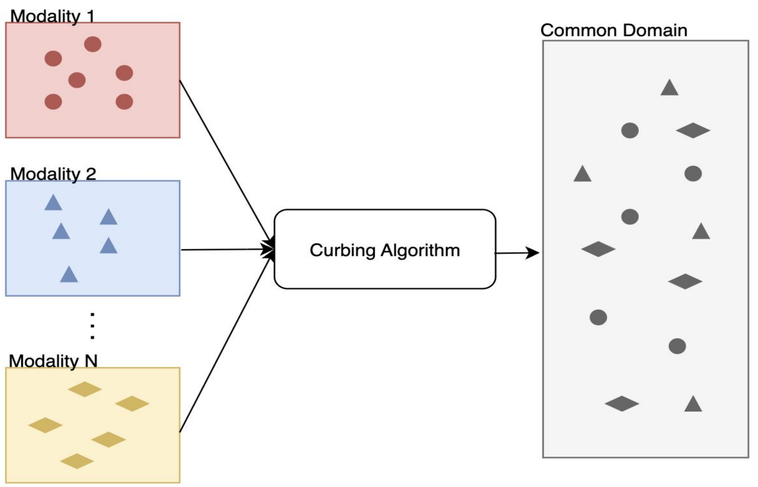
Impact:
- Found a meaningful distribution of confidences with high confidence when the correct GTIN was assigned and low confidence otherwise
- Significant cost reduction compared to the previous method



Model	Accuracy	Cost Impact
Model 1	87.12%	- 60.91%
Model 2	84.64%	- 41.51%
Model 3	87.79%	- 54.91%
Model 4	85.81%	- 41.39%
Model 5	88.44%	- 54.31%

Part 2: Introducing cross-modal fusion

Idea: Test combinations of different modalities as inputs to the ranking fusion algorithm to have additional signals
Problem: A **modality gap** arises when *equally* correct matches produce different evidence strengths across modalities. How can we curb this modality gap?



Solution: Used a custom curbing algorithm to map all modalities to the same domain

Impact: Ran the experiment with Model 5 and saw the accuracy increase to **90.62%**