Comparing CNN Detection and GPT-4o for Dietary Intake Reporting: Toward a Semantically Grounded Ensemble Framework

Ying-Chieh Liu
Industrial Design Department, Chang Gung University, Tao-Yuan, Taiwan

Chien-Hung Chen
Cybersecurity Technology Institute, Institute for Information Industry, Taipei, Taiwan

Mekhla Sarkar
Computer Science and Information Engineering Department, Chang Gung University, Tao-Yuan, Taiwan

Prasan Kumar Sahoo
Computer Science and Information Engineering Department, Chang Gung University, Tao-Yuan, Taiwan

Abstract

Automated dietary intake reporting depends on reliable multi-dish food-image recognition. While CNN-based object detectors have improved classification accuracy, they still struggle to generalize across variations in ingredients, containers, and presentation styles—especially for culturally specific dishes. In our previous work, EfficientDet-D1 achieved 0.92 mAP on 87 Taiwanese dishes, yet recognition accuracy dropped substantially in the presence of atypical serving conditions. For example, beef noodle soup served in a tall mug rather than a traditional wide bowl poses difficulties for CNNs that rely heavily on spatial and visual cues.

To better understand these limitations, we compare EfficientDet-D1 with GPT-4o, a multimodal vision–language model. EfficientDet-D1 delivers high spatial precision in standard settings but falters with visually ambiguous inputs. GPT-4o, by contrast, demonstrates semantic flexibility—successfully associating non-standard presentations and alternate dish names (e.g., "ramen" vs. "Japanese noodle soup")—yet it suffers from occasional hallucinations and imprecise localization.

This comparative evaluation highlights the trade-off between perceptual robustness and semantic reasoning in food image recognition. The findings help clarify whether the two models offer complementary strengths, potentially guiding future research on hybrid frameworks that integrate spatial precision with contextual understanding.

Keywords

Automatic Image Recognition, Multimodal Large Language, CNN-based detector, GPT-4o.