Kuih Classification and Segmentation

Using Ensemble Learning: CNN Segmentation + Vision Transformer

NAIC (AI Technical) Team: CantByteUs

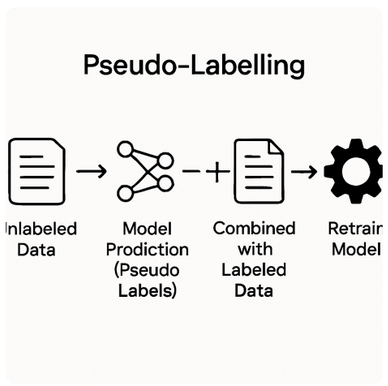
Chin Zhi Xian • Ng Tze Yang • Ong Chong Yao • Terrence Ong Jun Han

Data Collection & Preparation

- **Scraped ~2500 images/class** from Bing and Google search engines
- **Removed duplicates** by computing tensor differences between images
- **Manual filtering** of unrelated images and combined with original photos
- **Annotated original dataset** for segmentation using Label Studio
- **Varied image types:** High/low-res, varied lighting/angles, partially eaten kuih

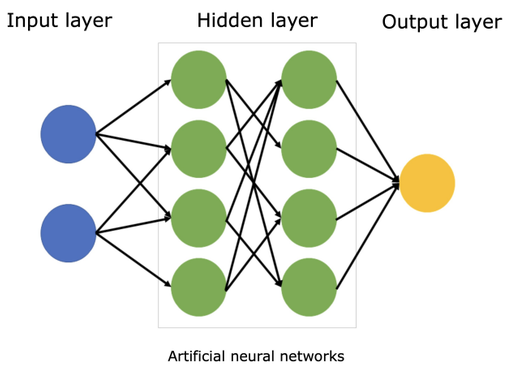
Pseudo Labelling Method

- **Semi-supervised annotation** technique for efficient dataset creation
- **Annotated 20-30 kuih/class** based on feature complexity
- **Trained small YOLOv11-seg** for faster annotation with manual verification
- **Iterative process:** Retrained larger model with combined data
- **Final dataset:** 98 images/class (perfectly annotated)



Model Development

- **Tools:** PyTorch, Ultralytics, CUDA for GPU acceleration
- **Initial approach:** Modified YOLOv11-seg configuration (depth, width, channels)
- **Challenge:** Large models overfit on small datasets
- **Solution:** Pre-trained YOLOv11x-seg on COCO 2017, then fine-tuned on kuih dataset
- **Preprocessing:** Normalized exposure for consistent lighting

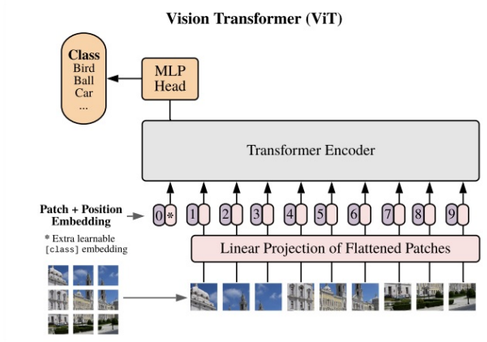


Why Segmentation?

- **"Robust segmentation inherently improves classification accuracy"**
- **Prioritizes core object** (the kuih) and reduces background distractions
- **Near-perfect confusion matrix** achieved with YOLOv11x-seg
- **Classification loss plummeted** after only a few epochs

Why Vision Transformer?

- **Splits images into patches** and transforms them into tokens (like LLMs)
- **Captures relationships** across entire image in every layer
- **Addresses visual similarities** (Kek Lapis vs. Kuih Lapis) through texture analysis
- **Self-attention mechanism** analyzes long-distance relationships between regions
- **Fast convergence** when pre-trained on large datasets (ImageNet 22k)



Final Model: Ensemble Approach

- **Combines CNN Segmentation + Vision Transformer**
- **Soft voting method:** If models agree → chosen class
- **If disagree:** Class with highest confidence from either model
- **Hybrid approach outperformed** solo models in classification tests
- **Leverages strengths** of both architectures for robust classification