Kuih Classification and Segmentation

Using Ensemble Learning: CNN Segmentation + Vision Transformer

NAIC (AI Technical) Team: CantByteUs

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

DATA COLLECTION

1 High and low-resolution images
Varied lighting and angles
Partially eaten kuih

1 SELF — SONY, Canon, Xiaomi, DJI

1 ONLINE — roboflow

1 ‘better-bing-image-downloader’ library in PyPi

‘Download All Images’ extension in Chrome

3 Annotated original dataset for segmentation using Label Studio due to a lack of existing segmentation-formatted kuih datasets.

2 Filter

(1) Convert images into tensors
(2) Comput tensor cosine difference between images
more effective than using pixel similarity to remove duplicates

Duplicate Image Finder by elisemercury

3 Annotation in: Label Studio

4 Augmentation

roboflow

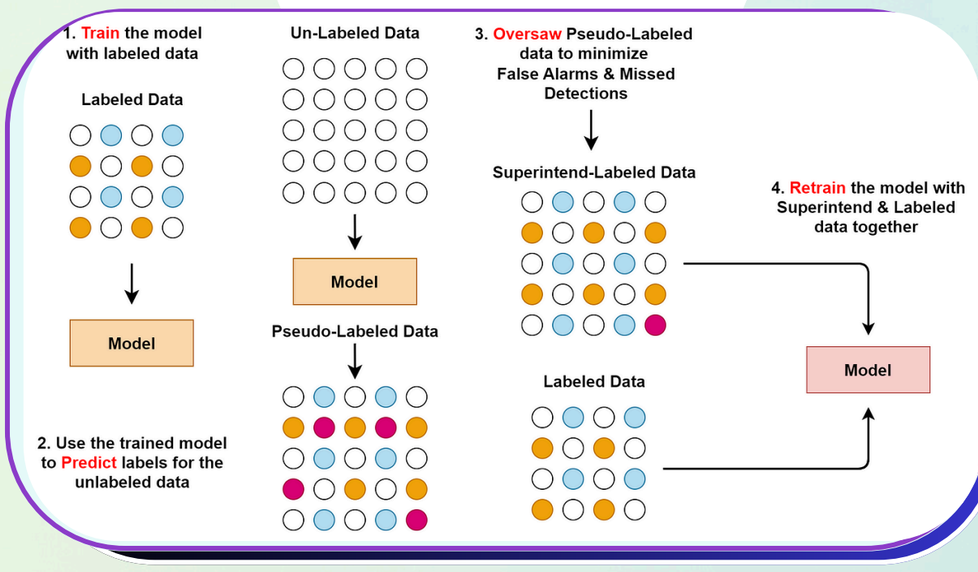
Tripled the dataset size

Cropping and added some noise

Excluded hue/color adjustments to preserve color-sensitive features.

6 Dataset

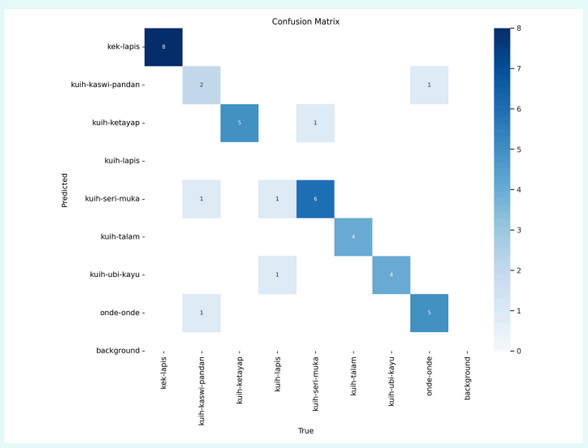
5 Pseudo labelling

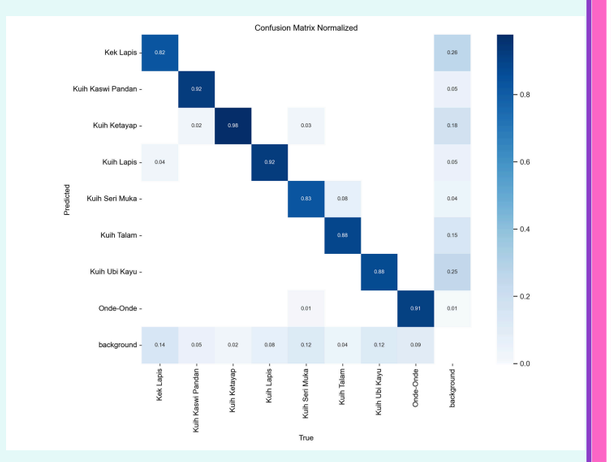


7 TRAINING

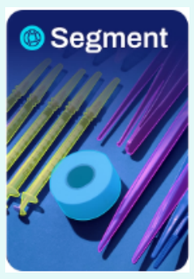
CNN
YOLOv11x-seg
ultralytics YOLO

“A robust segmentation model inherently improves classification accuracy" - It prioritises the core image (the kuih) and reduces distractions.





NVIDIA CUDA



PyTorch

ViT
VisionTransformer
Hugging Face

used the ‘eva02_base_patch14_224.mim_in22k’ model pretrained on ImageNet 22k

```
...  c:\Users\ochon\.conda\envs\naic\Lib\site-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found
     from .autonotebook import tqdm as notebook_tqdm
Classes: ['Kek Lapis', 'Kuih Kaswi Pandan', 'Kuih Ketayap', 'Kuih Lapis', 'Kuih Seri Muka', 'Kuih T
Epoch 1/10: 100%|██████████| 270/270 [01:23<00:00,  3.23it/s]
Epoch 1: Train Loss = 0.8603
Epoch 1: Val Accuracy = 91.89%
Epoch 2/10: 100%|██████████| 270/270 [01:25<00:00,  3.14it/s]
Epoch 2: Train Loss = 0.1891
Epoch 2: Val Accuracy = 96.68%
Epoch 3/10:   4%|▌          | 10/270 [00:17<01:54,  2.27it/s]
```

ViT’s fast convergence tends to bring the risk of overfitting

ViTs use a 'self-attention' mechanism
to analyse the entire image, capturing textures and long-distance relationships between image regions.

Thus, even if a kuih looks slightly different across images, the ViT can still recognise it based on learnt patterns.

8 ENSEMBLE

▶ **A soft voting method:** If the models agree, that class is chosen.
▶ **Decision Logic**: If models disagree, the class with the highest confidence score from either model is chosen.
▶ **Superior Performance**: This hybrid approach significantly outperformed solo models in classification tests

Our writeup and Git repository