UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Student: Khanh Huyen Bui
Student ID: 19021307

AN ENSEMBLE OF NEURAL NETWORK MODELS FOR PRESCRIBED PILL RECOGNITION

SCIENCE PROJECT

Major : Computer Science

Hanoi, 2022

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Student: Khanh Huyen Bui
Student ID: 19021307

AN ENSEMBLE OF NEURAL NETWORK MODELS FOR PRESCRIBED PILL RECOGNITION

SCIENCE PROJECT

Major : Computer Science

Supervisor: Dr. Quoc Long Tran

Hanoi, 2022

Acknowledgments

I would like to express my deep appreciation to Pill Gates team, my supervisor Dr. Quoc Long Tran for accompanying me during the competition and throughout my internship at the Institute for Artificial Intelligence. I would also like to give a big thank to the Mr. Thuc Pham for guiding our team on the basics of the competition.

Student: Khanh Huyen Bui
Supervisor : Dr. Quoc Long Tran Student ID: 19021307

An Ensemble of Neural Network Models for Prescribed Pill Recognition

Abstract

Medicine usage is a common treatment that supports patients in overcoming illness. Taking the wrong medication, on the other hand, can have serious consequences, such as decreasing treatment effectiveness, causing severe side effects, or even fatal. According to the WHO, drug misuse, rather than disease, accounts for one-third of all death cases. Because of the widespread development and increasing demand for drugs, there is a greater demand for applications that aid in drug identification.

Stemming from the actual situation, it is necessary to develop effective solutions which utilize AI technology to tackle the prescribed pill identification problem, which might be integrated into a mobile device application for practical usage. In addition, the objective of the system is to identify the pills that do not exist in the prescriptions captured by mobile devices; and to raise a warning of wrong usage. Therefore, the topic is about localizing and recognizing prescribed pill instances in real-world images, and out-of-prescription pills recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem

Object Detection is the task that involves detecting instances of objects from a particular class in an image:The goal of object detection is to detect all instances of objects from a known class, such as people, cars or faces in an image. Typically only a small number of instances of the object are present in the image, but there is a very large number of possible locations and scales at which they can occur and that need to somehow be explored.

The Object Detection problem is divided into 2 sub-tasks, with images as input and the location of the objects as output:

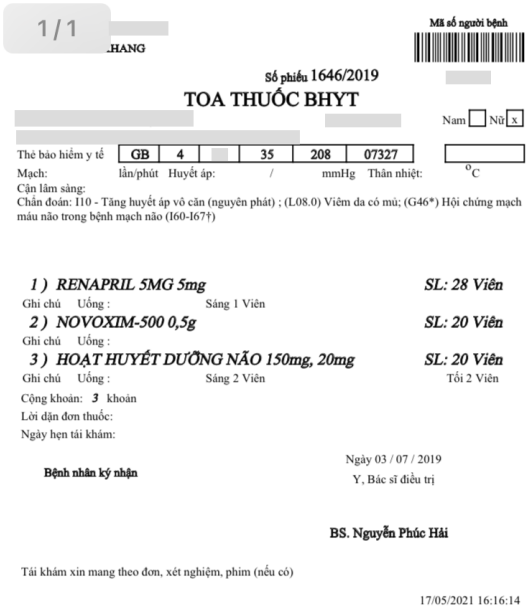
- Detect bounding boxes for each object – the rectangle drawn around the object in order to locate the object.
- When the bounding boxes of objects are detected, they are then classified with confidence score.

Chapter 2

Problem specification

2.1 Problem specification

Within the scope of this problem, the main task is pills Detection and Out-of-prescription pills Recognition. The proposed solution must be capable of detecting and identifying drugs from pill images, as well as utilizing the prescription information associated with those images (Prescription information). Data is gathered from medical centers (prescriptions images) and actual photos of prescription pills (pills images).



1 / 1

Mã số người bệnh

Số phiếu 1646/2019

TOA THUỐC BHYT

Nam ☐ Nữ ☒

Thẻ bảo hiểm y tế	GB	4		35	208	07327	

Mạch: lần/phút Huyết áp: / mmHg Thân nhiệt: °C

Cận lâm sàng:

Chẩn đoán: I10 - Tăng huyết áp vô căn (nguyên phát) ; (L08.0) Viêm da có mủ; (G46*) Hội chứng mạch máu não trong bệnh mạch não (I60-I67†)

1) RENAPRIL 5MG 5mg — *SL: 28 Viên*
Ghi chú Uống : Sáng 1 Viên

2) NOVOXIM-500 0,5g — *SL: 20 Viên*
Ghi chú Uống :

3) HOẠT HUYẾT DƯỠNG NÃO 150mg, 20mg — *SL: 20 Viên*
Ghi chú Uống : Sáng 2 Viên Tối 2 Viên

Cộng khoản: *3* khoản

Lời dặn đơn thuốc:

Ngày hẹn tái khám:

Ngày 03 / 07 / 2019

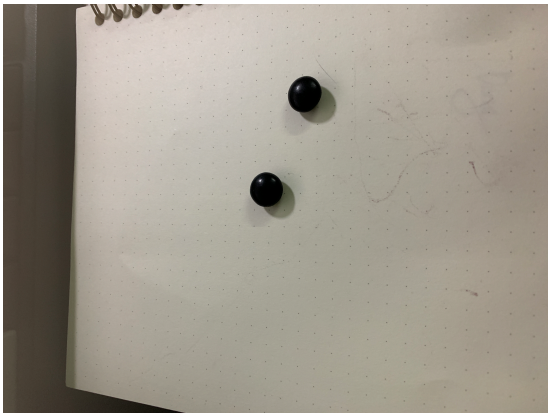
Bệnh nhân ký nhận

Y, Bác sĩ điều trị

BS. Nguyễn Phúc Hải

Tái khám xin mang theo đơn, xét nghiệm, phim (nếu có)

17/05/2021 16:16:14



(a) Prescription **(b)** Pill

Figure 2.1.1: Data sample

Labels of pills range from 0 to 107, in which 0-106 are the ID of the pills (regarding drug-name in the prescriptions) and 107 is the ID for any pills that are not in the prescription. In other words, the 107 label assigns to many different pills, both the drugs with the actual labels 0-106 and the drugs without labels.

2.2 Data

2.2.1 Overview

The dataset consists of 3 volumes:

- Training set includes 9.500 pills images và 1.171 prescription images
- Validation set includes 1.500 pills images và 172 prescription images
- Test set includes 4178 pills images và 829 prescription images

Inputs to the model are raw images without bounding boxes. The output of the model consists of the bounding boxes (coordinates of the two left-top and right-most points), the pills' corresponding labels, and the confidence scores of the predictions.

2.2.2 Data analysis

Regarding the data, after surveying the dataset with 9500 drug images on the training set, the following information was obtained. The number of drugs on one image ranges from 3 to 4 pills/prescription, in which for 1 prescription, the corresponding drug image appears at least 1 pill, at most 11 pills.

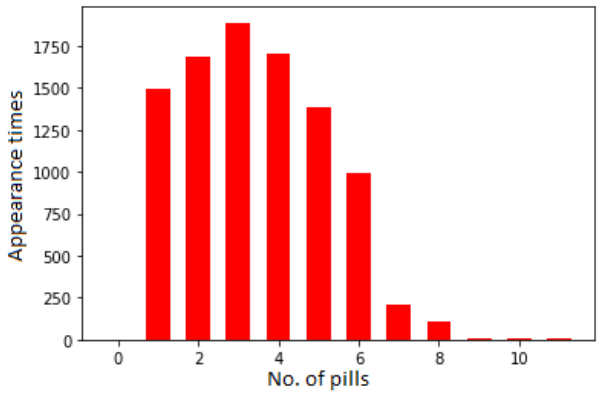


Figure 2.2.1: No. pills/images distribution

Fig 2.2.1 shows the distribution of the number of labels on the train set. A data imbalance can be seen here. Some labels appear a lot while other labels appear very little, for example, labels 25, 49. In terms of image size, mostly images are sized [3024, 4032] and vice versa. 412 / 5,000.
Furthermore, there are anomalies in the data (as shown in **Fig 2.2.2**) that can be identified as:

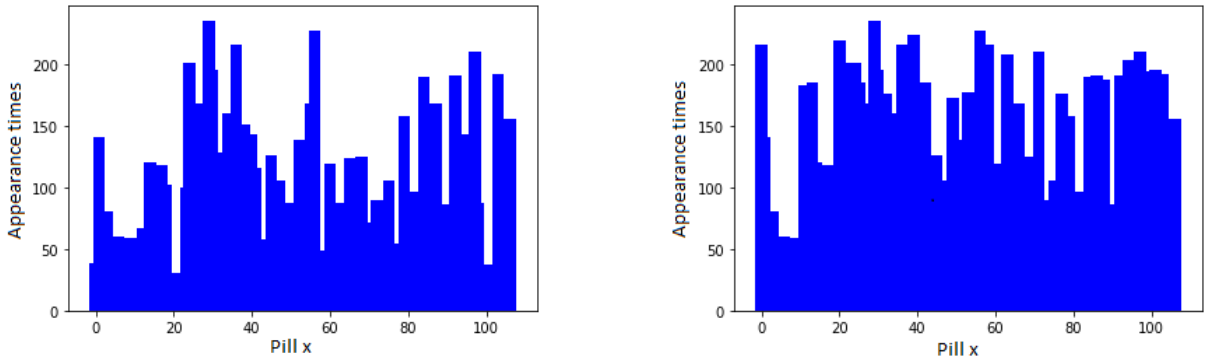
- The pills are nearly identical.
- The pill's color is altered by light.
- Medicines are packaged in both transparent and opaque bags.
- Some tablets are only half-filled.

- Medicines piled on top of one another.
- Sometimes drugs are close together, and sometimes they are far apart.
- The drug image is blurry.
- A single drug name corresponds to several IDs, and vice versa.

Figure 2.2.2: Data Abnormalities

This is what makes the problem challenging.

2.2.3 Data generation



(a) Before **(b)** After

Figure 2.2.3: Data before and after generating ≈ 6000 images

From the information collected after the survey and data analysis, new data is generated and augmented based on the provided data set. Label distribution before and after generating ≈ 6000 images is shown in **Fig 2.2.3**

2.3 Evaluation Metrics

Results will be evaluated based on the key metric **wmAP0.5-0.95** (weighted COCO mean Average Precision). **wmAP** is the weighted average of the APs over the labels, where the AP for the non-prescription drug label is given higher weight in this problem. The higher the **wmAP**, the better the recognition model and the greater the number of non-prescription drugs it can identify.

wmAP is calculated by the formula:

$$\mathbf{wmAP} = \frac{1}{10 * (N + \alpha)} \sum_{\substack{IoU \rightarrow 0.5+0.05*i \\ i=0}}^{i=9} \sum as \tag{2.1}$$

In addition, the following metrics were also calculated to show the effectiveness of the model when changing the IoU threshold as well as with the drug image sizes:

- **wmAP50**: wmAP calculated for threshold IoU=0.5
- **wmAP75**: wmAP calculated for threshold IoU=0.75

- **wmAPs**: wmAP calculated for small bounding boxes ($S < 32^2$)
- **wmAPm**: wmAP calculated for bounding box mediums ($32^2 < S < 96^2$)
- **wmAPl**: wmAP for large bounding boxes ($96^2 < S$)

Chapter 3

Solution

3.1 Method

3.1.1 Detection model

YOLOv7 [1] is the new state-of-the-art object detector in the YOLO family. According to the YOLOv7 paper, it is the fastest and most accurate real-time object detector to date. YOLOv7 established a significant benchmark by taking its performance up a notch. Starting from YOLOv4, new entries can be seen in the YOLO family one after another in a very short period of time. Each version introduces something new to improve the performance.

1. YOLO Architecture in General
   YOLO architecture is FCNN(Fully Connected Neural Network) based. However, Transformer based versions have recently been added to the YOLO family as well. We will discuss Transformer based detectors in a separate post. For now, let's focus on FCNN (Fully Convolutional Neural Network) based YOLO object detectors. The YOLO framework has three main components.

   - Backbone
   - Head
   - Neck

   The Backbone mainly extracts essential features of an image and feeds them

to the Head through Neck. The Neck collects feature maps extracted by the Backbone and creates feature pyramids. Finally, the head consists of output layers that have final detections.

2. YOLOv7

   YOLOv7 improves speed and accuracy by introducing several architectural reforms. Similar to Scaled YOLOv4, YOLOv7 backbones do not use ImageNet pre-trained backbones. Rather, the models are trained using the COCO dataset entirely. The similarity can be expected because YOLOv7 is written by the same authors of Scaled YOLOv4. The following major changes have been introduced in the YOLOv7 paper.

   Architectural Reforms:

   - E-ELAN (Extended Efficient Layer Aggregation Network)
   - Model Scaling for Concatenation-based Models

   Trainable BoF (Bag of Freebies):

   - Planned re-parameterized convolution
   - Coarse for auxiliary and Fine for lead loss

3. Experiments

   The experiments were carried out using various architectures, including standard architecture, tiny architecture, fine-tuned width architecture, and running many different experiments to select the parameters appropriate for the data set.

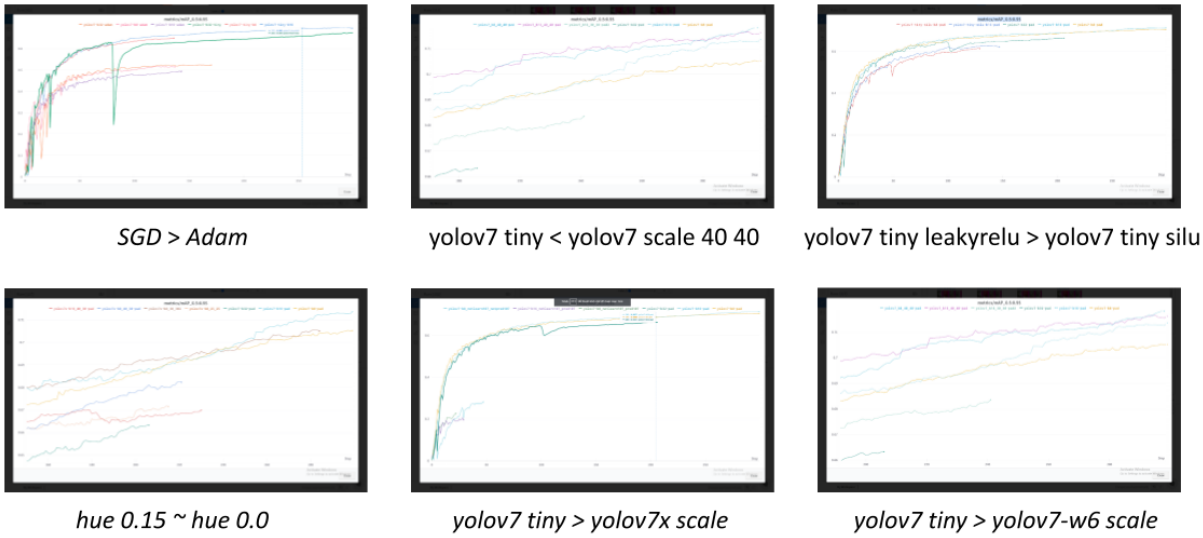


Figure 3.1.1: YOLOv7 Experiments

After the experiment, the best architectures were filtered out, with the mAP as follows:

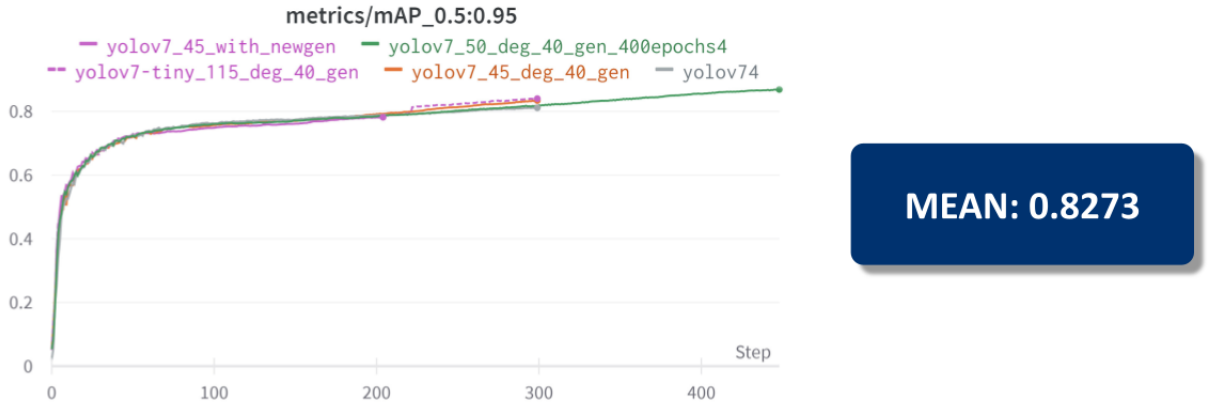


Figure 3.1.2: Best architectures

3.1.2 OCR model

OCR is formerly known as Optical Character Recognition which is revolutionary for the digital world nowadays. OCR is actually a complete process under which the images/documents which are present in a digital world are processed and from the text are being processed out as normal editable text.

OCR is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera into editable and searchable data.

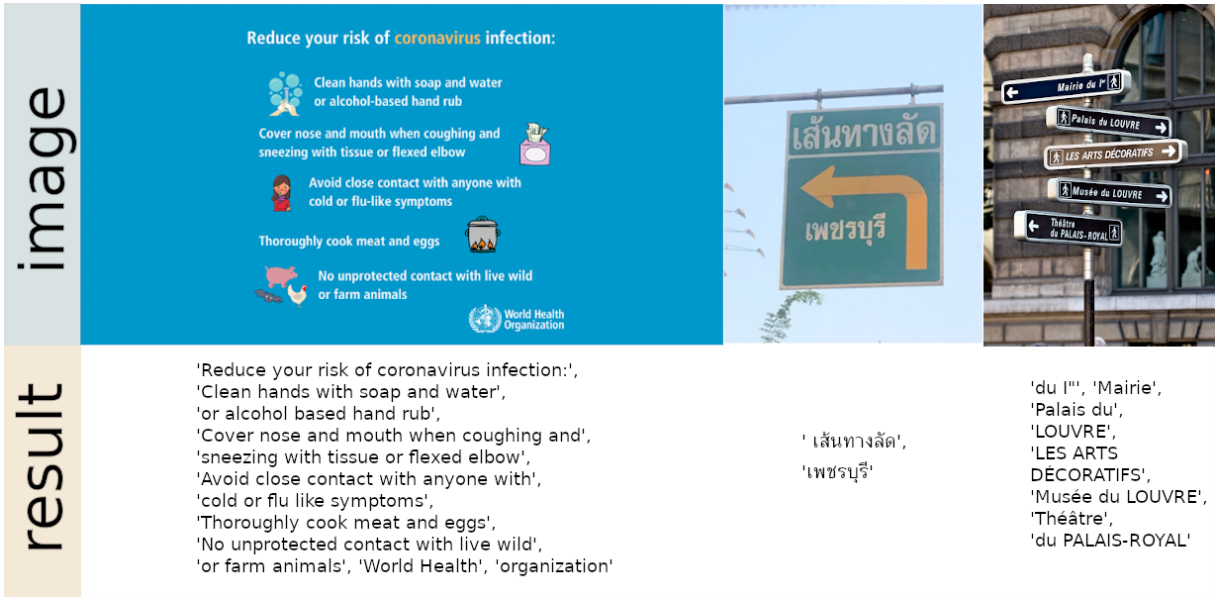


Figure 3.1.3: OCR task example

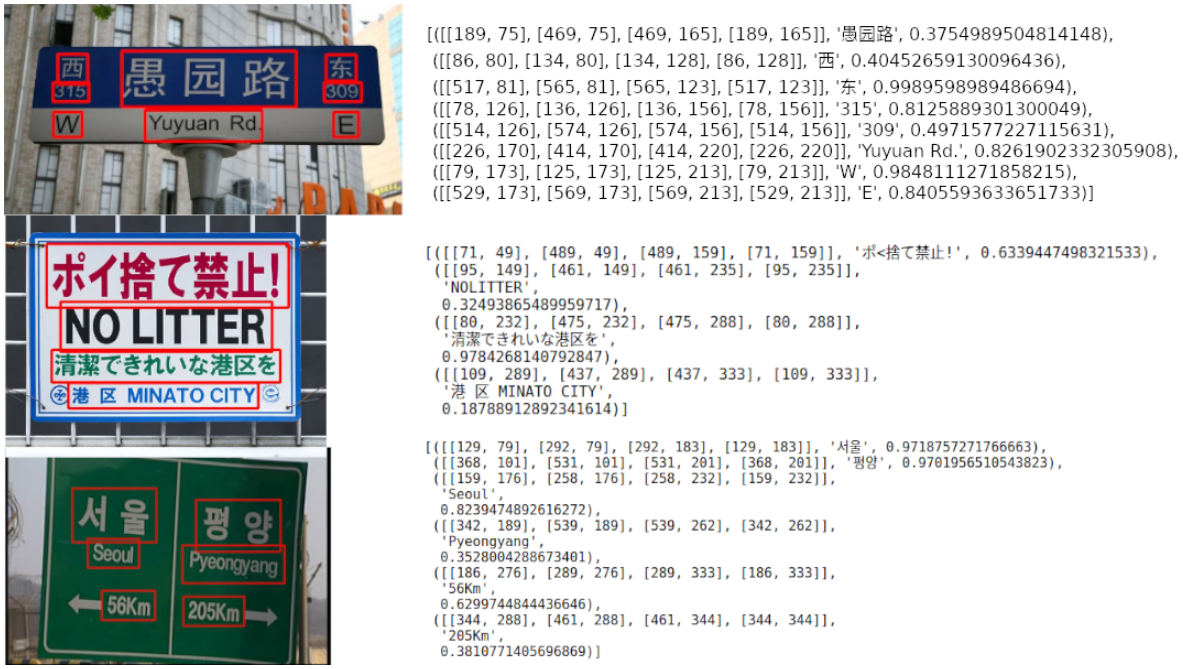


Figure 3.1.4: OCR task example (2)

EasyOCR [2] is actually a python package that holds PyTorch as a backend handler. EasyOCR like any other OCR(tesseract of Google or any other) detects the text from images but in my reference, while using it I found that it is the most straightforward way to detect text from images also when high end deep learning library(PyTorch) is supporting it in the backend which makes it accuracy more credible. EasyOCR supports 42+ languages for detection purposes. EasyOCR is

created by the company named Jaided AI company.

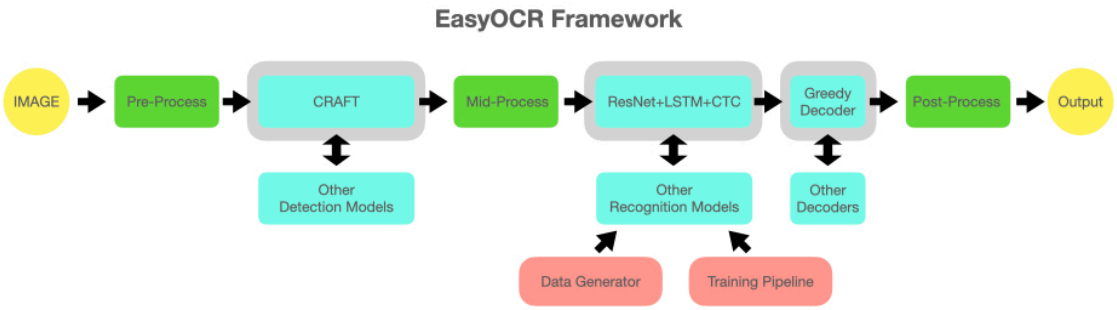


Figure 3.1.5: EasyOCR Framework

Results after postprocess archieved upto 99% accuracy.

3.1.3 Inference flow

The workflow of the inference phase is describe as follows:

1. OCR module is used to extract useful information from prescription
2. Multi-class object detection module detects and recognizes pills with label 0-106
3. Single-class object detection module detects pills (2 labels are 1 for object and 0 for background)
4. The predicted bounding boxes then are post-processed with OCR results and final output is obtained.

In **Fig 3.1.6** the dict_drugname_id is the mapping between ID and drug name in the prescription obtain from the train set; the extracted drugname is the OCR results from prescription image. These information are used to suggest potential ids in drug images.
base and **adv** are ensemble results from multi-class models (base) and single-class models (adv), respectively.

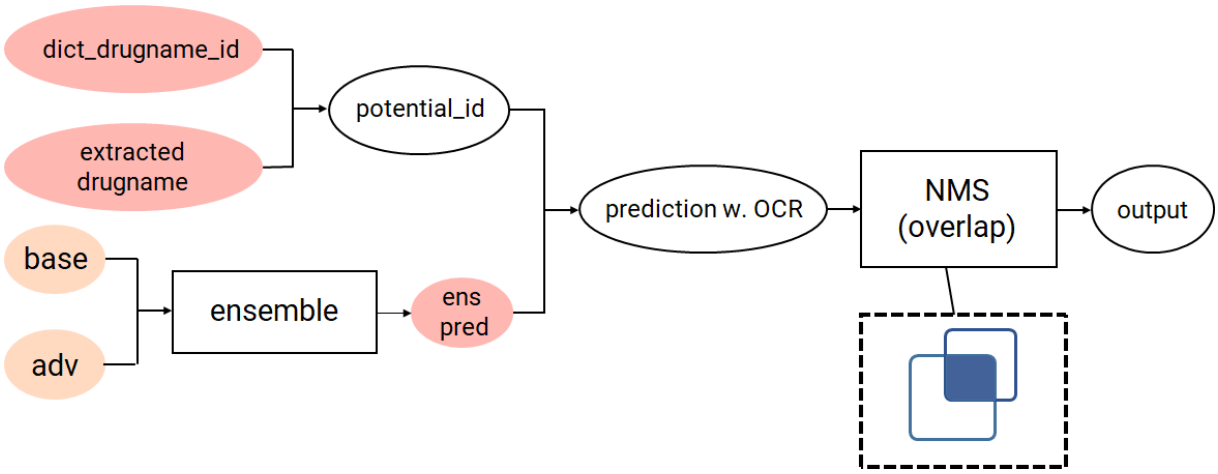


Figure 3.1.6: Inference baseline

In terms of drug image preprocessing (**Fig 3.1.7**), information from the drug image's metadata, specifically the orientation in the exif meta data, is used to transpose the drug image to match the provided label. This task is only performed once during the train and infer phases.

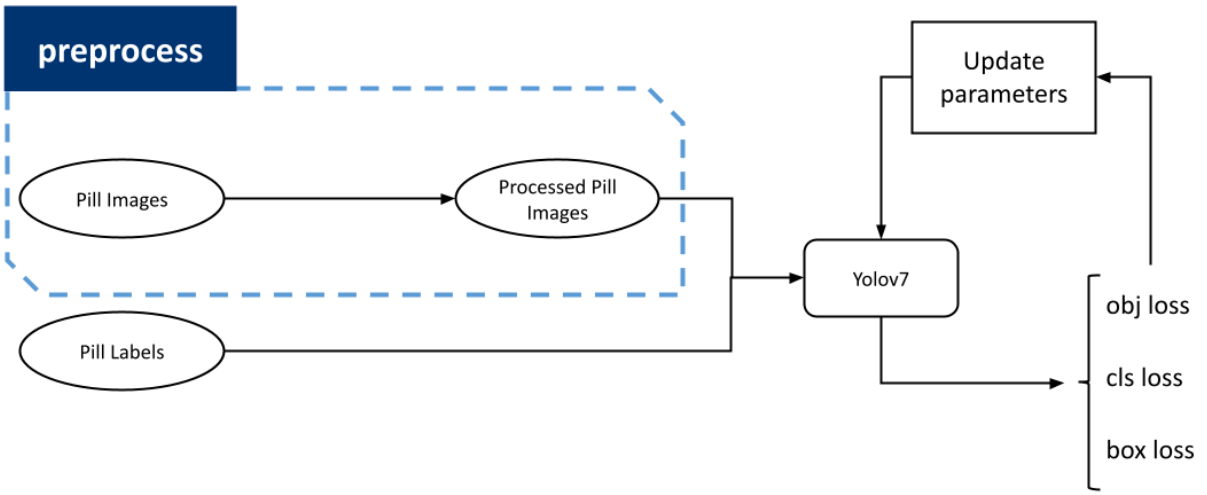


Figure 3.1.7: Preprocessing images

Regarding the pill recognition module **Fig**, the model family Yolo (yolov7) with different architectures is used for recognition. Architectures are classified into two categories: multi-class and single class recognition models. Here multi-class will identify drugs with id from 0 -106, single class will identify object (drug) and background

In the training phase, the drug after being preprocessed will be included in the model, predicted output (y_pred) and ground truth (gt) are used to calculate the

loss values (box loss, class loss, ...) and update the parameters. The architectures used were mentioned in the Dectection Model (**3.1.1**) section.

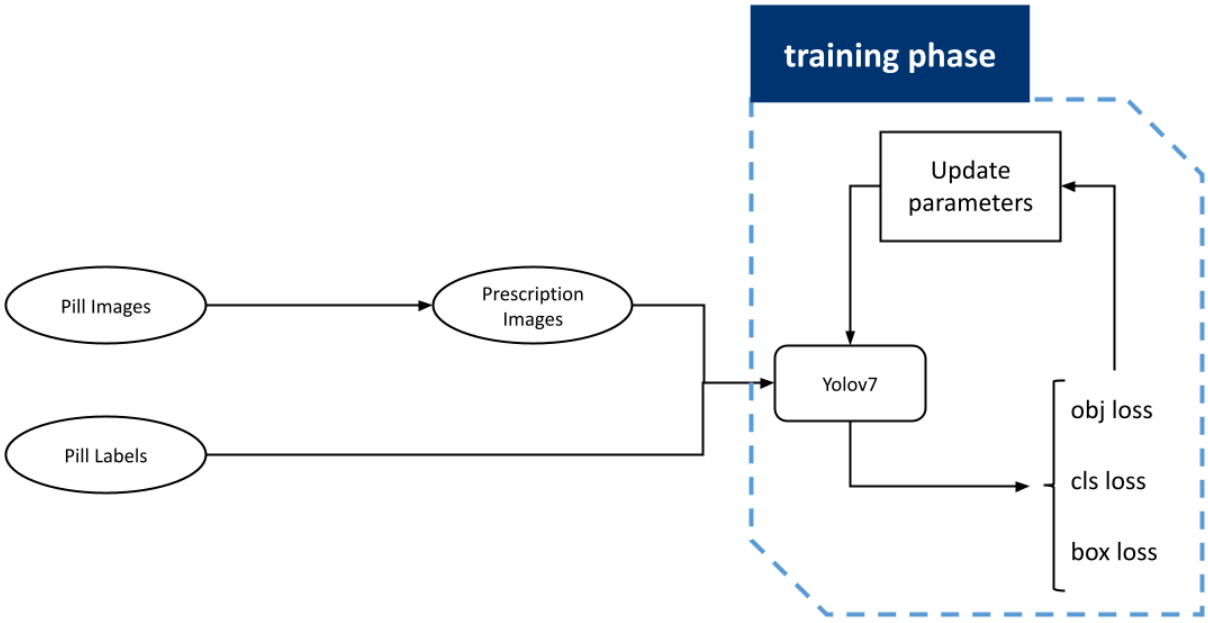


Figure 3.1.8: Object recognition module

When the trained weights of multi-label (hereinafter referred to as base) and single-label (hereinafter referred to as adv) models are available (hereinafter 5 base models and 2 adv models), in the infer pipeline phase will proceed as follows **Fig 3.1.9**:

1. The preprocessed images are passed through the models for predictions.
2. Base predictions are concatenated, and an NMS (Non-Maximum Suppression) technique is used to select a bounding box from the overlapping bounding boxes.
   Specifically, the bounding boxes will be sorted by confidence score. The bounding box with the highest confidence score among bounding boxes of the same class that overlap with a specific threshold will be chosen.
   Typically, the adv model can detect more regions than reality. In other words, the model's prediction will cover all pills. The next NMS is similar to that of base prediction, however this time NMS will select the model's bounding box with the best coverage. This reduces the possibility of the model guessing too many bounding boxes.
3. The output of these 2 processes is then ensemble.

Specifically, those with each bounding box in adv that do not overlap with any pill in the base will be assigned 107. In other words, the pill that adv can detect but the base cannot detect will be labeled as 107.

4. Combined with OCR results (as shown in **Fig 3.1.6**).
   The predictions are combined with the potential id information generated by the OCR extraction results and the drugname id dictionary before going through the NMS overlap function again to handle the 107 overlapping bounding boxes. The final output will be the result. The NMS function is used to process the 107 overlapping tablets in this case, so the overlap area is used to narrow the bounding boxes.

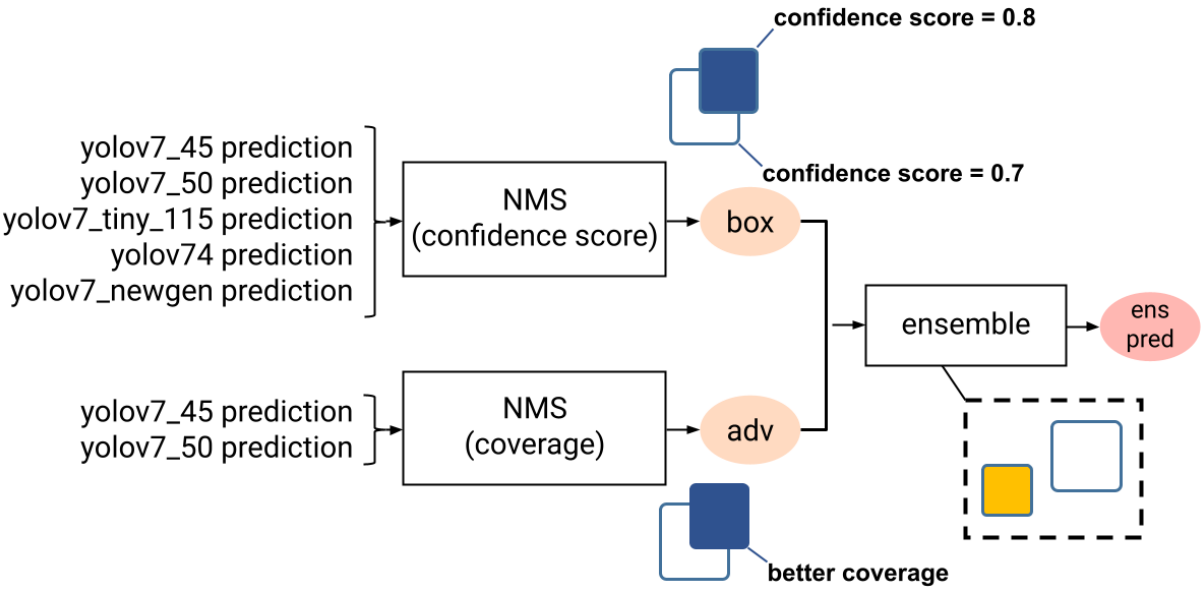


Figure 3.1.9: Ensemble multiple recognition models

Regarding the module to extract information from prescriptions, which hereinafter referred to as OCR for short, use open source EasyOCR to perform OCR on the entire prescription.
The OCR model extracts the drug name from the prescription image. This output is then matched with the dictionary (of drug names and the corresponding id generated from the prescription label of the train set), the module outputs the ids 0-106 that are likely to appear in the corresponding drug image as shown in **Fig 3.1.10**.

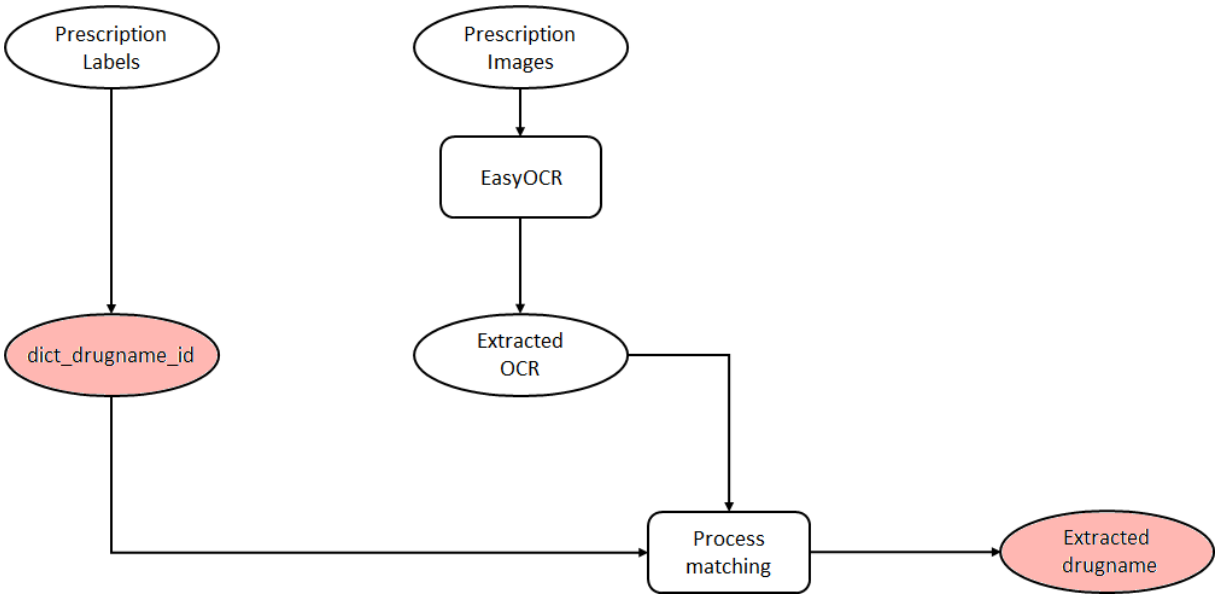


Figure 3.1.10: Optical Character Recognition module

Chapter 4

Conclusion

The conclusions obtained after several experiments were conducted:

1. Training with generated data is more effective than without generating data.
2. Ensemble of various multi-class models is better than only one multi-class model.
3. Ensemble with single-class models is better than without single-class model.
4. NMS shows significant improvement.
5. Noise training data.
6. Metric is trivial.

Table 4.0.1: Validation of YOLO models and Results on test set

(a) Validation of each YOLO models

Model	Epochs	Width scale	mAP
yolov7_45	450	45%	0.84
yolov7_50	450	50%	0.87
yolov7_tiny_115	300	115%	0.84
yolov74	300	100%	0.81
yolov7newgen	204	45%	0.78

(b) Results on test set

Metric	**Value**
wmAP	0.4919
wmAP50	0.7245
wmAP75	0.5656
wmAPs	-1.0000
wmAPm	0.0408
wmAPl	0.4924

Bibliography

[1] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[2] JaidedAI. Jaidedai/easyocr: Ready-to-use ocr with 80+ supported languages and all popular writing scripts including latin, chinese, arabic, devanagari, cyrillic and etc..