YOLOv5 vs. YOLOv8: Comparative Evaluation for Face Mask Detection

Cho Hye Won
Computer Vision
AIFFEL Research
Gyeonggi-do, South Korea
c09789@naver.com

Abstract

In this paper, we present a comparative analysis of YOLOv5 and YOLOv8 for face mask detection using a dedicated dataset. With the emergence of the COVID-19 pandemic, the importance of automated face mask detection systems has become paramount. YOLOv5 and YOLOv8, both being state-of-the-art object detection models, offer promising capabilities in this domain. However, their performance in the context of face mask detection remains to be explored comprehensively. To address this gap, we conducted experiments using a standardized dataset, evaluating the detection accuracy, speed, and robustness of both models. Our results indicate that while YOLOv8 exhibits superior performance in bounding box creation, YOLOv5 demonstrates stronger capabilities in determining mask usage. The findings of this study provide valuable insights for the development of efficient face mask detection systems and contribute to the ongoing efforts to combat the spread of infectious diseases.

1 Introduction

The COVID-19, first identified in December 2019 in Wuhan, Hubei Province, China, has spread worldwide, becoming a pandemic. COVID-19 is a respiratory infection disease primarily transmitted among humans through respiratory droplets, leading to global dissemination. Major symptoms include fever, cough, and difficulty breathing, while severe cases can result in pneumonia and respiratory distress. Some individuals may experience mild symptoms, but it can pose serious health issues for the elderly or those with underlying health conditions. The virus causing COVID-19 continues to mutate over time, occasionally giving rise to new variant viruses. For this reason, government authorities issued guidelines for wearing masks at that time. While the infection control guidelines have been relaxed to a large extent, many facilities such as hospitals, nursing homes, and childcare centers still adhere to these policies. Consequently, this paper aims to develop a model using the object detection algorithm YOLO to detect the position of faces and determine mask wearing status. Additionally, we seek to compare YOLOv5 and YOLOv8, the latest object detection models, to determine which is more suitable for the proposed model and suggest a better model for the system.

2 Related Work

One of the major milestones in the field of Object Detection is the emergence and distinction between One-Stage Detector and Two-Stage Detector. This reflects the difference in the object detection process, whether it is performed in one stage or two stages, involving Region Proposal and Detection. First, the 2-stage object detector is an approach consisting of two stages. In the first stage, it selects candidate objects (proposals) from the input image, and in the second stage, it accurately predicts the position and class of the selected object proposals. These models exhibit high accuracy and stability but require relatively slow speed and high computational resources. The 1-stage object detector is an approach consisting of a single stage. It directly predicts the position and class of objects from the input image, requiring fast speed and low computational resources but potentially lower accuracy

and stability. Representative examples of One-Stage Detector methods include YOLO, Retina-Net, SSD, and EfficientDet, while Two-Stage Detector methods include the RCNN series and SPPNet.

In this paper, we decided to utilize the YOLO algorithm, a One-stage Detector, for swiftly distinguishing mask-wearing status. YOLO, introduced in 2016 with its initial version, v1, has undergone continuous performance improvements up to v8 by 2023. This evolution showcases a transition in the approach from traditional object detection methods like Hog and DPM to CNN-based Two-Stage Detectors initiated with AlexNet, and eventually to the One-Stage Detector approach of the YOLO series. The remarkable performance enhancements in recent YOLO versions have sparked new interest in object detection technology. This progress reaffirms the significance of object detection technology and has spurred various research and applications. It drives innovation and advancement in the field of object detection, significantly amplifying expectations for practical applications.

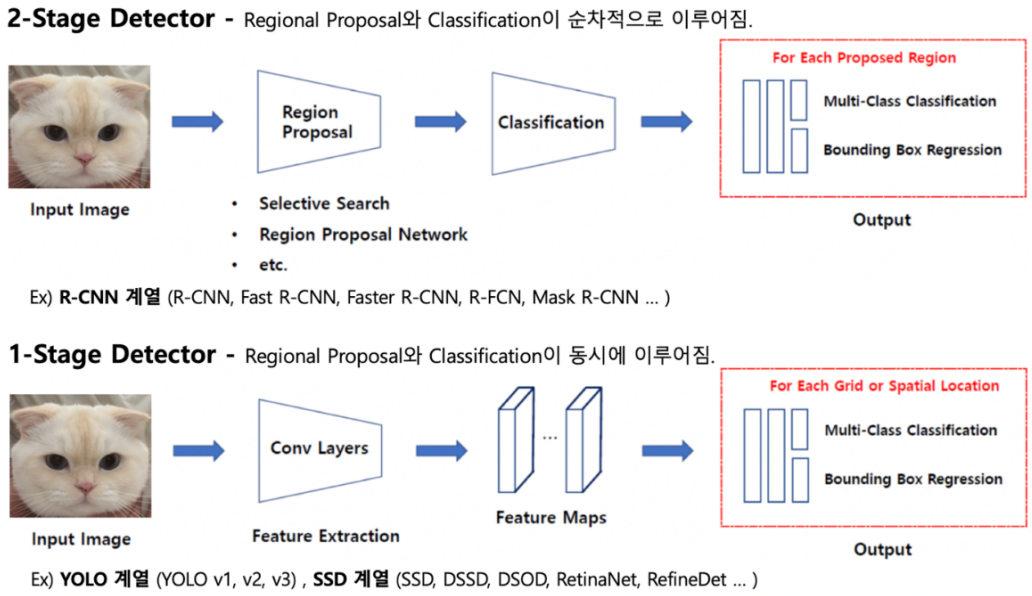

Figure 1: a comparison between the 1-stage detector and the 2-stage detector

3 METHODS

3.1 DATASET IN EXPERIMENT

Since the outbreak of COVID-19, numerous face mask datasets(1) have been proposed. Among them, the ”Face Mask Detection” dataset provided by Kaggle comprises a total of 853 images annotated in PASCAL VOC format, including bounding boxes and class information, suitable for training on YOLO platforms. The dataset is classified into three categories based on class information: ’With mask’, ’Without mask’, and ’Mask worn incorrectly’.

When examining the Face Mask Detection dataset, you will find two folders: ”images” and ”annotations”. The ”images” folder contains image files numbered from 0 to 852, while the ”annotations” folder contains XML files numbered from 0 to 852. The XML files within the ”annotations” folder contain information about each respective image file.

```
#include <stdio.h>
<annotation>
    <folder>images</folder>
    <filename>maskssskskssss0.png</filename>
    <size>
        <width>512</width>
```

```
        <height>366</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>without_mask</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <occluded>0</occluded>
        <difficult>0</difficult>
        <bndbox>
            <xmin>79</xmin>
            <ymin>105</ymin>
            <xmax>109</xmax>
            <ymax>142</ymax>
        </bndbox>
    </object>
```

Upon examining the contents, the dataset includes folder names and file names, along with information about image sizes. Inside the "Object" section of the code, it is evident that objects are categorized into three classes: "mask−weared−incorrect," "with−mask," and "without−mask." "mask−weared−incorrect" contains information about objects not wearing masks properly, "with−mask" contains information about the positions of objects wearing masks, and "without−mask" contains information about objects not wearing masks. The bounding box information is provided for each object in the order of $xmin, ymin, xmax, ymax$, specifying the bounding box's area.

3.2 Model in Experiment

3.2.1 YOLOv5

YOLOv5 was released in June 2020, offering lower capacity and faster speed compared to v4. As depicted in the figure, it can be divided into Backbone, Neck, and Head. The Backbone serves as the architecture or network for feature extraction, typically composed of ordinary convolution layers and modules like C3 and Spatial Pyramid Pooling-Fast (SPPF). The Neck is a feature fusion network that performs multi-scale fusion of features extracted from the backbone network. The Head includes three detection heads of different scales, outputting predictions of the category with the highest confidence score and the location of the target. YOLOv5 is introduced in four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The models are divided based on performance and time. "S" is the lightest model with the highest frame rate, while "x" is the heaviest model with the lowest frame rate. In this paper, due to the real-time detection requirement, we selected YOLOv5s, which is the fastest and lightest model, for analysis and comparison.(2)

3.2.2 YOLOv8

In YOLOv5, the CSP layer used was replaced with the new C2f module in the YOLOv8 model. This model utilizes a modified CSPDarknet53 backbone. The C3 module contains 3 ConvModules and n BottleNecks, and ELAN introduced in YOLOv7 is combined in the C2f module (2 ConvModules and n BottleNecks). Batch normalization and SiLU (Sigmoid Linear Units) activation functions are used in each convolution layer. The backbone is connected to the neck at three different levels, and the neck combines information from the network's various layers and transmits it to the head. The decoupling of the head allows for the separation of classification, regression, and detection heads. Instead of estimating center offsets from anchors, the head predicts the object's center and bounding box offsets directly, making the YOLOv8 method anchor-free. YOLOv8 uses soft NMS instead of NMS, allowing overlapping bounding boxes to remain partially, which enhances model performance.(3)

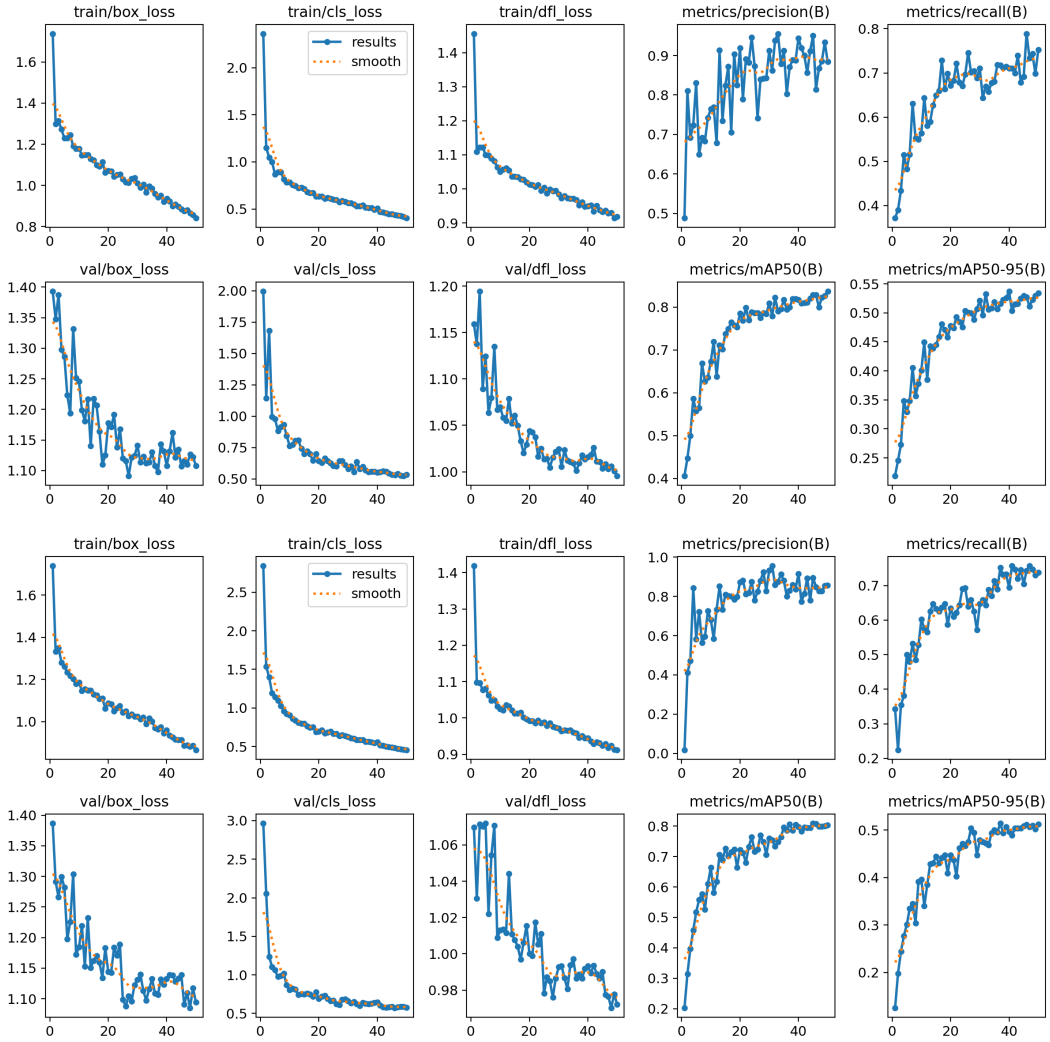

Figure 2: Performance evaluation of YOLOv5 and YOLOv8

4 RESULT

Figure 2 illustrates the training results of YOLOv5s and YOLOv8 models under the same training environment. From the training results, it can be observed that YOLOv8 shows a faster and more stable increase in metrics/precision and metrics/recall compared to YOLOv5. In terms of Validation dfl (Distribution Focal Loss), YOLOv5 exhibits a rapid decrease instead. Apart from this, there are no significant differences observed.

Figure 3: Using YOLOv5 and YOLOv8 to experiment with actual mask-wearing status

The above Figure 2 shows, from left to right, the original image, YOLOv5 results, and YOLOv8 results. Both models accurately measure whether a mask is worn. However, in the second image, YOLOv5 creates bounding boxes in non-facial areas for measurement, whereas YOLOv8 detects faces even at a distance, indicating mask usage. Conversely, in the fourth image, YOLOv5 accurately judges mask wearing, while YOLOv8 fails to do so. This experiment highlights that YOLOv8 demonstrates a significant advantage in creating bounding boxes, while YOLOv5 exhibits strength in determining mask usage.

5 RESULT

In this study, we conducted a comparative analysis of the performance of object detection models YOLOv5 and YOLOv8. To ensure comparability, we utilized the same dataset (with the same number of classes and training images) and evaluated them under identical conditions. Upon examining the results, it was evident that YOLOv8 demonstrated significantly better performance in terms of

speed and accuracy of bounding boxes. However, when it came to object recognition, YOLOv5 exhibited superior performance. Nevertheless, considering that YOLOv8 excels in detecting objects with high precision in bounding boxes, one might speculate that with a larger dataset, the results could potentially differ.

Due to memory constraints, it was challenging to utilize larger datasets for training and testing computer vision models, limiting our ability to explore and optimize hyperparameter values effectively. In the future, we hope to overcome this limitation by leveraging more extensive datasets and applying optimization algorithms to select optimal hyperparameters for the YOLO architecture, thereby enhancing its performance.

REFERENCES

[1] K Suresh, MB Palangappa, and S Bhuvan. Face mask detection by using optimistic convolutional neural network. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1084–1089. IEEE, 2021.

[2] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[3] Luka Jovanovic, Nebojsa Bacanin, Miodrag Zivkovic, Joseph Mani, Ivana Strumberger, and Milos Antonijevic. Comparison of yolo architectures for face mask detection in images. In *2023 16th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, pages 179–182. IEEE, 2023.