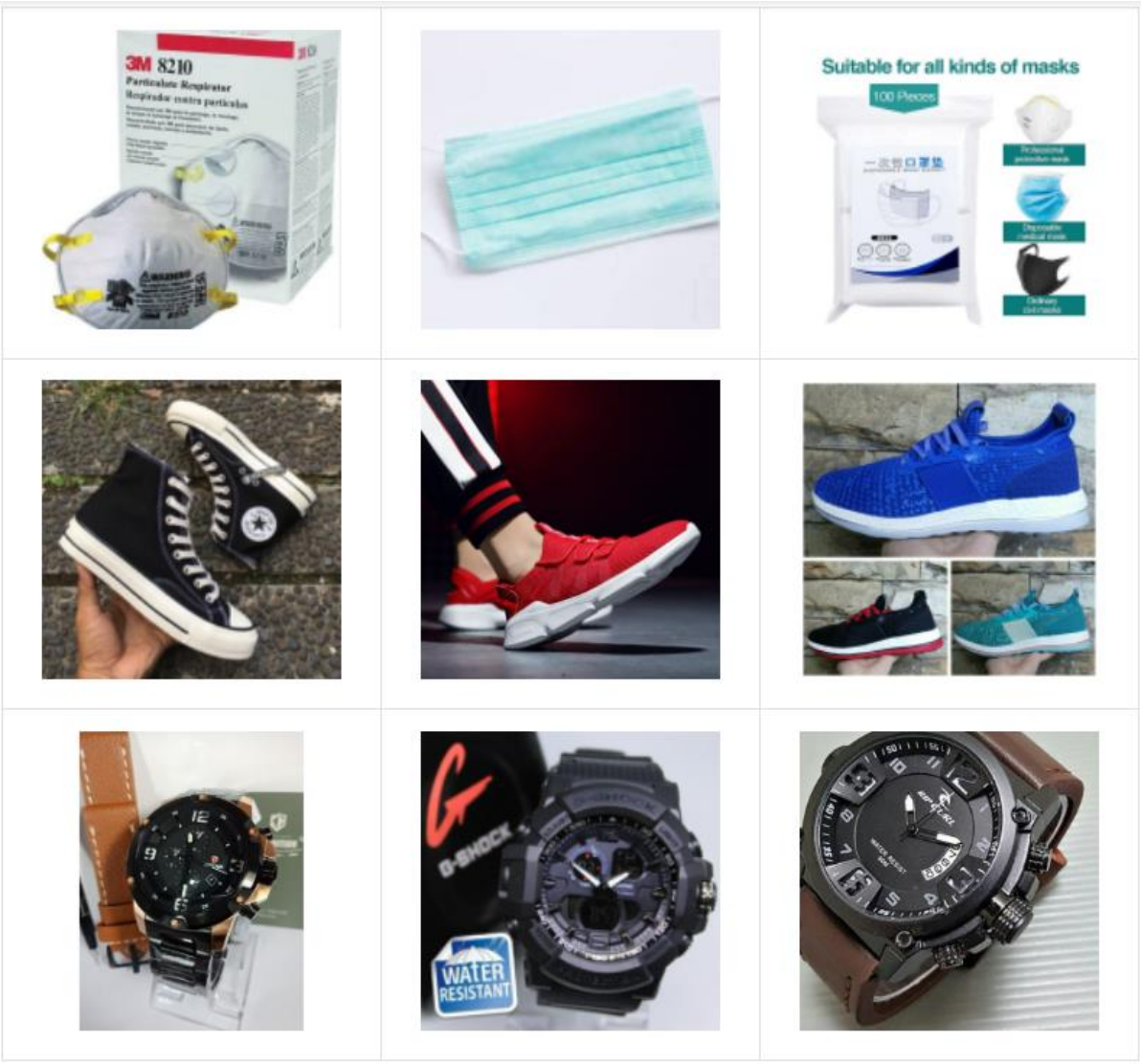
Product Recognition

Background

At Shopee, we always strive to ensure the correct listing and categorization of products. For example, due to the recent pandemic situation, face masks become extremely popular for both buyers and sellers, everyday we need to categorize and update a huge number of masks items. A robust product detection system will significantly improve the listing and categorization efficiency. But in the industrial field the data is always much more complicated and there exists mis-labelled images, complex background images and low-resolution images, etc. The noisy and imbalanced data and multiple categories make this problem still challenging in the modern computer vision field.



Task

In this competition, a multiple image classification model needs to be built. There are ~100k images within 42 different categories, including essential medical tools like masks, protective suits and thermometers, home & living products like air-conditioner and fashion products like T-shirts, rings, etc. For data security purpose, the category names will be desensitized. The evaluation metrics is top-1 accuracy.

Data Description

In Shopee Product Detection Dataset, there are more than 100k images directly from E-commercial industry field. You will be able to explore the real-world images which is noisy and long-tailed, and let your model predict the correct categories for the images. There contains 42 most popular categories product at Shopee.

File descriptions:
train.csv: training dataset.
test.csv: test dataset.

Columns of data fields:
filename: image file name(str).
category: image category(str).

Approach

Preprocessing:
Clean the dataset by some cluster techniques such as p-hash, image histogram and metric learning. Since we know that the most of images within one category are correct, can use such techniques to remove those outliers as much as possible. For some categories like class No.33 with only ~500 samples, may have less samples than others, you might need to use some augmentation to upsample those categories. Balanced dataset is extremely important for classification, it is always recommended to make dataset balance as possible as you can. The preprocessing and training strategy may count more than the model itself sometimes.

Data Augmentation:
Some useful techniques for image augmentation are color jittering, jpeg compression, cutout, mixup, label smoothing etc. Those techniques will help to make the model more robust which will let your score more stable for public leaderboard and private leaderboard.

Classification Model:
There are quite a few models that are suitable for this task. ResNet, ResNext, EfficientNet, ResNeSt, VGG, InceptionNet and etc. For such a 40-category classification task, it is recommended to use a relatively larger model to avoid overfitting and also for better performance.

Training strategy:
This part is another important part, but it totally depends on different conditions like different machines, different preprocessing and different optimizers. Basically, if you have a

strong machine which can support you to train with a large batch size, then you might use a relatively large learning rate and less epochs. But if your machine cannot support a large batch size training, to get the same level result, you might need to use a smaller learning rate and more epochs. Or you can also use the loss accumulation strategy to simulate larger batch size. And do note that the training strategy should be decided based on your analysis for validation/public leaderboard result. With some iterations of submissions and analysis, the strategy will be polished to a perfect state.

Submission Format

Submission file format should be `csv` file only. And for each `filename` in the test dataset, you must predict only one proper category name. The `csv` file should contain a header and have the following format:

```
filename, category
2f096e5e8e8955d43632be16e35993b5.jpg, 0
3d63a44c82c9d1299b5791bdd2c7a4e8.jpg, 1
```