Q1 Readme

1. Feature Engineering

We use Res-net to extract features with resnet50 pretrained features(dim=2048)

```
ResNet50(weights='imagenet', include_top=False, pooling='max')
```

It's a mature neuron network for image detection. On the ImageNet dataset, residual nets with a depth of up to 152 layers, which is useful for feature engineering. As the pre-trained model have a better performance, we use it directly.

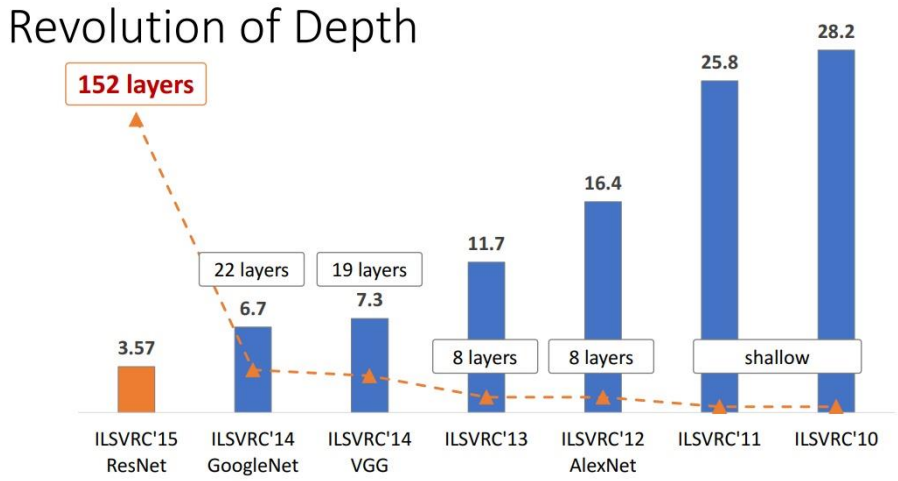

Figure 1 structure of res-net

2. Outlier detection

2.1 overview of data and local evaluation

Outlier percentage= 0.0792, I assume that the train set and test set is independently and identically distributed, there may exist 100 pictures with labeled "outlier".

In the exam paper, it mentions that evaluation based on recall and precision, so I decide to use ROC and precision@k as the loss function, (F1_score also makes

sense)

And as 2048 features seems large to train, I firstly try to PCA method and reduce the dimension to 500. However, the precision decreases as the component decreased. So, I keep all the features into the models.

2.2 Methods and score

I have trained 10 models with the package pyod, which focuses on outlier detection, including:

```
Model 1 Cluster-based Local Outlier Factor (CBLOF)
Model 2 Fast Angle-based Outlier Detector (FastABOD)
Model 3 Average KNN
Model 4 Local Outlier Factor (LOF)
Model 5 Median KNN
Model 6 Principal Component Analysis (PCA)
Model 7 Feature Bagging
Model 8 Isolation Forest
Model 9 K Nearest Neighbors (KNN)
Model 10 One-class SVM (OCSVM)
Model 11 Histogram-base Outlier Detection (HBOS)
```

The best models of those are One-class SVM(OCSVM), Principle Component Analysis(PCA), and K nearest Neighbors(KNN), but still get low rate of recall in the local validation.

```
In [12]:  1 from sklearn.metrics import confusion_matrix
          2 confusion_matrix(y_test,y_test_pred)

Out[12]: array([[1375,   56],
                [  63,   71]], dtype=int64)
```

2.3 Ensemble model

So I finally ensemble the top 3 best model and get the result.

References:

https://github.com/yzhao062/pyod/blob/master/notebooks/Model%20Combination.ipynb
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html