Introduction

In this assignment, we have 70000 images from twenties researchers. In this case, the quality of images is different as well as the size of images.

Besides, according to the introduction, we will have 28 kinds of species and more than 15,000 identities. What we need to do is find the feature of each identity and do prediction (on the identity level). We can have at most 5 identities for each test images.

Strategy

First, we find these things interesting:

1 data are extremely imbalanced. Some identities have more than 200 images whereas some of them only contains single image.

2 In the test data, there are some ‘new species’ which never show up in the training data. In this case, we need to determine these images as ‘new\_species’ in the inference stage.

3 The usage of species. We are going to focus on identity rather than species. So, what’s the use of species here?

Second, the things below are our strategies.

1 augmentation. We build a dictionary to memory identity and number of images. When the number of images is too few, we will produce some new images (By larger probability and range of rotating, changing brightness, saturation, and contrast….) and send enhanced image and original image into dataloader together. So, we will have more than 70000 images.

2 Choosing loss function. We have more than 15,000 classes here. So, normal cross-entropy cannot work well here. The Arc-face loss can somehow handle this situation.This algorithm set a margin between each class, currently, we consider it as advanced cosine distance loss.

3 We will extract features of each images and using KNN or other clustering algorithms in test stage to find the closest image in the training data and set corresponding identity to test images.

4 We will use model embedding here. Basically, there would be two kinds of predictions. One with the species features, another not. The one with species will first do the prediction of species of each test image, and then find closest identity among the same species group in the training data. Finally, we will combine the result based on CV scores .

5 Threshold of ‘new\_species’

As said before, there would be ‘new species’ in the test dataset, so how to pick the threshold is one thing important. We set a measure of confidence. When using clustering method, we will always have the distance. Based on distance, we can calculate confidence (In our strategy, 1/(1-distance)). We pick 5 top confidence value, e.g. confidence (1, 2, 3, 4, 5) identity(a,b,c,d,e).

When confidence value /sum(5 top confidence) smaller than threshold, than we put ‘new species’ here.(e.g. we have threshold 0.2, than it will be e, d, new species, c, b).

6 model embedding

We will have several different models based on included species or not (5 for each). We will combine them together based on CV-score. Here is the simple example about how we combine them:

The score of each ‘location’ will be [2,1.8,1.6,1.4,1.2]\*cv\_socre

file1: [a,b,c,d,e] cv:0.7

file2:[a,new\_species,c,d,f] cv:0.8

file3:[new\_species,a,d,e,c] cv:0.75

total score:

new\_species : 0\*0.7+1.8\*0.8+2\*0.75 = 2.94

a: 2\*0.7+2\*0.8+1.8\*0.75 = 4.35

b: 1.8\*0.7+0\*0.8+0\*0.75 = 1.26

….

In this way, for each picture, we pick the top 5 score and set them as the prediction.