Shopee - Price Match Guarantee

Bing Liu

Jilin University
College of Computer Science and Technology

2021-04-17

Overview

Overview of this competition

- Problem background
- Problem description

Data

- Data description
- Files

Algorithm

Conclusion

Overview of this competition

Problem background

- This problem is an active competition, which prize money is $30000.
- The competition purpose is to determine if two products are the same by their images.

Problem description

Retail companies use a variety of methods to assure customers that their products are the cheapest. Among them is product matching, which allows a company to offer products at rates that are competitive to the same product sold by another retailer. To perform these matches automatically requires a thorough machine learning approach, this is the mainly problem we need to solution! Two different images of similar wares may represent the same product or two completely different items. Retailers want to avoid misrepresentations and other issues that could come from conflating two dissimilar products. Currently, a combination of deep learning and traditional machine learning analyzes image and text information to compare similarity. But major differences in images, titles, and product descriptions prevent these methods from being entirely effective. In this competition, we'll apply our machine learning skills to build a model that predicts which items are the same products.

Data

Data description

Task is to identify which products have been posted repeatedly. The differences between related products may be subtle while photos of identical products may be wildly different! only the first few rows or images of the test set are published; the remainder are

only available to your notebook when it is submitted. Expect to find roughly 70,000 images in the hidden test set. The few test rows and images that are provided are intended to illustrate the hidden test set format and folder structure.

Files

- train/test.csv - the training set metadata. Each row contains the data for a single posting. Multiple postings might have the exact same image ID, but with different titles or vice versa.
  - posting_id - the ID code for the posting.
  - image - the image id/md5sum.
  - image_phash - a perceptual hash of the image.
  - title - the product description for the posting.
  - label_group - ID code for all postings that map to the same product. Not provided for the test set.
- train/test images - the images associated with the postings.
- sample_submission.csv - a sample submission file in the correct format.
  - posting_id - the ID code for the posting.
  - matches - Space delimited list of all posting IDs that match this posting. Posts always selfmatch. Group sizes were capped at 50, so there is no need to predict more than 50 matches.

Overview of this competition

Data

Algorithm

Conclusion

Algorithm

Algorithm

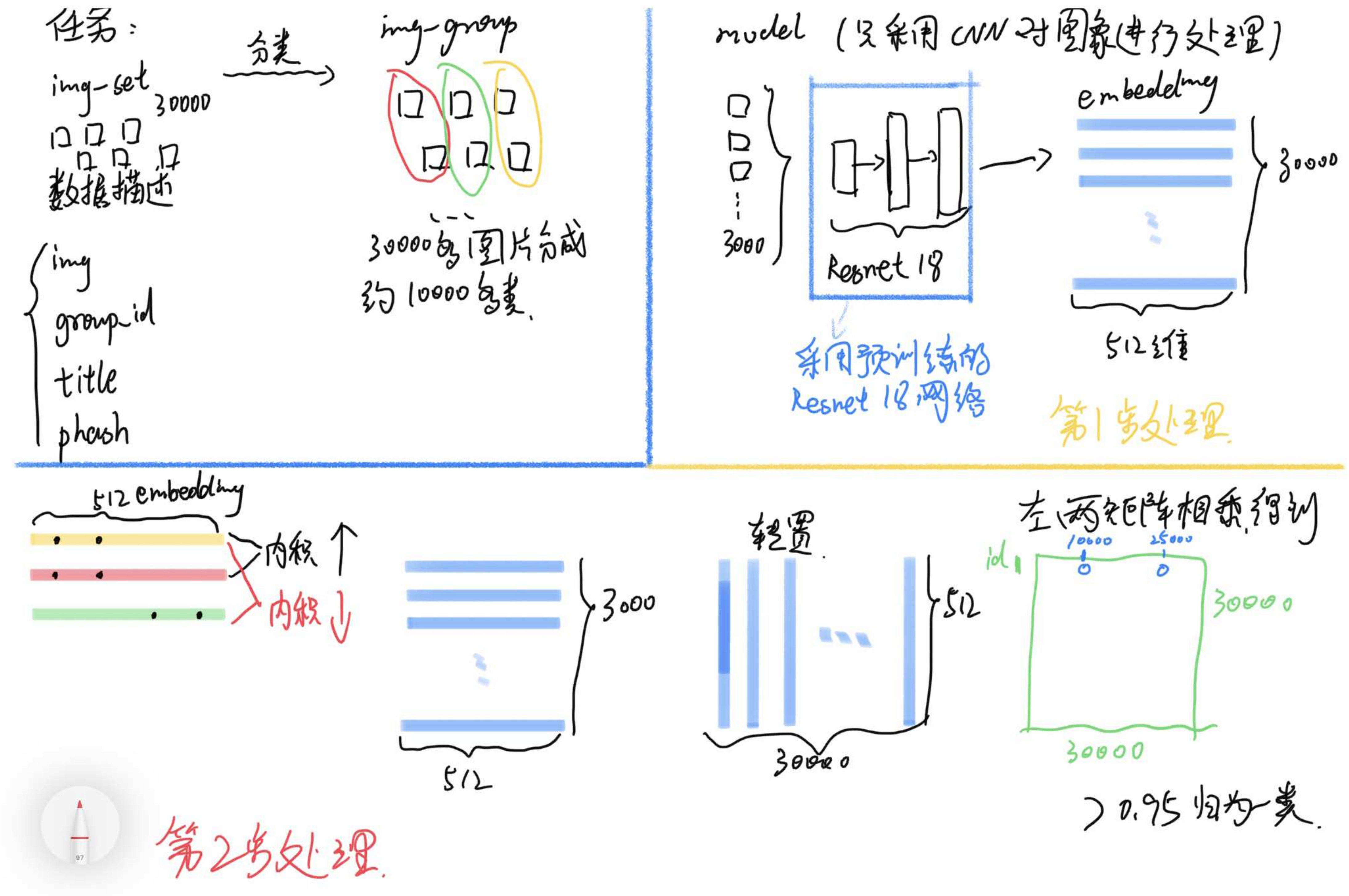
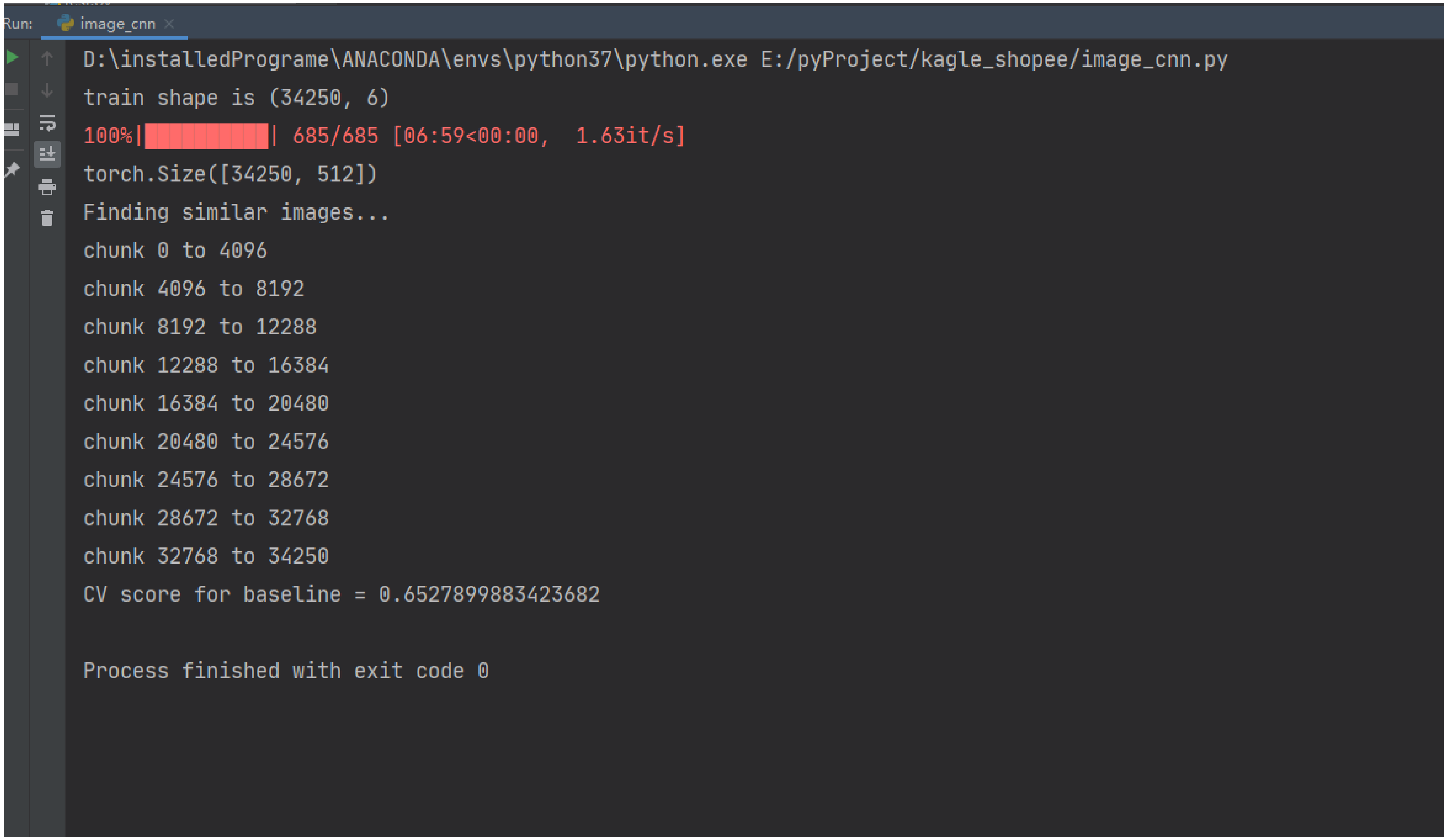


Figure 1: model

Conclusion

Conclusion



```
D:\installedPrograme\ANACONDA\envs\python37\python.exe E:/pyProject/kagle_shopee/image_cnn.py
train shape is (34250, 6)
100%|██████████| 685/685 [06:59<00:00,  1.63it/s]
torch.Size([34250, 512])
Finding similar images...
chunk 0 to 4096
chunk 4096 to 8192
chunk 8192 to 12288
chunk 12288 to 16384
chunk 16384 to 20480
chunk 20480 to 24576
chunk 24576 to 28672
chunk 28672 to 32768
chunk 32768 to 34250
CV score for baseline = 0.6527899883423682

Process finished with exit code 0
```

Figure 2: result

Conclusion

In the first algorithm, although the CV score for baseline is 0.6528, we only use the image information to match the products whether they are belong to the same product. In the next week, we will try to add the image_phash and title information to our model

to imporve score.

Contact Information

Bing Liu
College of Computer Science and Technology
Jilin University, China

✉ BLIU@TULIP.ACADEMY

🏠 TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING