Digital Humanities Assignment--Analyze and predict popular images on Pixabay

1. Introduction:

Every media or advertisement company should know what is popular and what can attract customers to watch or stay for a longer time to watch. And for international customers, the first problem for the enterprise is cultural barriers.

Entrepreneurs cannot just focus on what topic of texts are more attractive because people use different languages and a word that is popular in one culture sometimes may not be able to translate to another culture. For example, the English word "propaganda" has a relatively negative meaning but there is not an exact word with the total same meaning to translate propaganda into Mandarin. In Cambridge Dictionary, they use "宣传" to translate propaganda, but this word is neutral or positive in Mandarin. So, when a Chinese company translates "宣传" into English and put the word: propaganda in the advertisement, they may lose more money and customers in western society.

Instead of words, the image may be a better choice for corporations. Firstly, images do not need to be translated and they can be easily understood by any culture. Secondly, customers may need to take time to read the word in an advertisement, but the image can be understood in a moment, and this can improve the efficiency of advertising.

but when making the image in adverting, the company or self-media needs to know what kinds of pictures can attract customers to watch in normal life and whether a picture that they made in the same famous topics can also be very famous, which is the aim of this article.

This research uses Pixabay as a research platform because Pixabay is an image website that can provide 1,556,009 free images and all the pictures have Pixabay licenses, which means this research will be legal. This research will also use the CNN neural network, an unsupervised method of the image to analyze and predict the picture. This research used this unsupervised method to analyze because it is difficult to point out which features of a picture can lead it to become a popular picture. Another reason to choose Pixabay is that this website can provide API, and this will be easier for researchers to download data from the website for future research. One drawback is that the API only allows downloading 600 popular pictures for each topic in one day and there are 20 topics, which means the maximum number of pictures that can be downloaded is 12000 in one day. The research device is the laptop Legion R7000P which has next-gen AMD Ryzen 5000 H-series Mobile Processors and NVIDIA GeForce RTX 3060 GPU. If someone also wants to do similar research, they may need a better device because this device took at least 6 minutes for only one epoch, and this is too slow.

2. Analyze Data:

In the documentation of Pixabay API, there are two choices:" popular pictures", and "the latest picture", for people to choose. There are still many pictures that do not have a lot of views but are tagged "popular" in the "popular pictures" choice, and this is a good thing because they may be popular pictures before but it is not popular anymore so this can help the computer to learn their differences. When choosing "the latest picture", this research may meet some problems, for example, some pictures should be popular, but it is too latest to prove if they can be popular and they don't have too many views. This article chose the popular picture because if the Researcher downloaded the latest picture, these pictures may not have enough the number of views and the computer may identify them as unpopular pictures. At last, this research downloaded the data of the picture URL, the number of views, the number of downloads, the number of comments, the tag for each picture, and the topic for each picture. There are 20 topics: backgrounds, fashion, nature, science, education, feelings, health, people, religion, places, animals, industry, computer, food, sports, transportation, travel, buildings, business, music. Each topic can be downloaded for 600 pictures in one day. At first, this research tried to use Image and Tag to predict the number of views or downloads, or likes, but then found that this can give the computer an extraneous influence, which can cause the computer to analyze results that are not only focused on the image but also influenced by the tag. But the tag is describing the content of the image, which creates a multicollinearity problem and the main thing this research wants to find is to analyze the popular images. So, this research used the image as the independent variable and try to predict the number of views or downloads, or likes.

One topic has 3 pages, and each page has 200 pictures, so I created 20 big documents for each topic and 3 small documents for each big document. Each small document has 200 pictures and a JSON document that contains the information for each picture. Firstly, I made a dictionary to contain the necessary information in JSON.

The key name of the dictionary	The meaning of the key name
"Views"	The number of views for each picture
"likes"	The number of the likes for each picture
"downloads"	The number of the downloads for each picture
"comments"	The number of the comments for each picture
"Path"	The path of each picture on my laptop

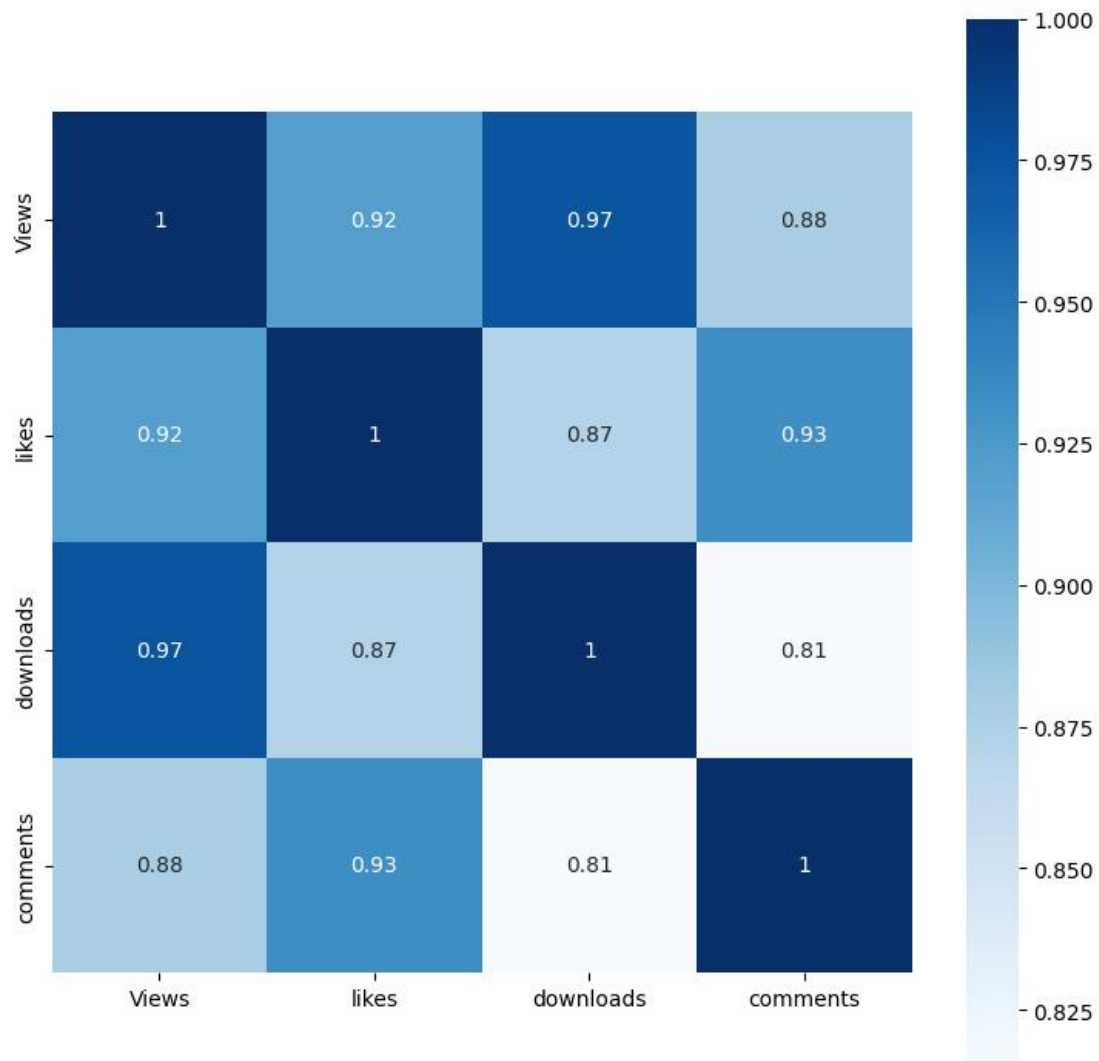Table 2.1 explains what each key name means in my dictionary

Figure 2.1 the heatmap is a covariance matrix to show the relationship among variable views, comments, downloads, likes

The heatmap shows that these four variables have a strong multicollinearity problem. All the scores are almost 1 and this means they are highly correlated. So, I can choose any of them for this research to be my dependent variable. I chose views as my dependent variable and tried to predict at last.
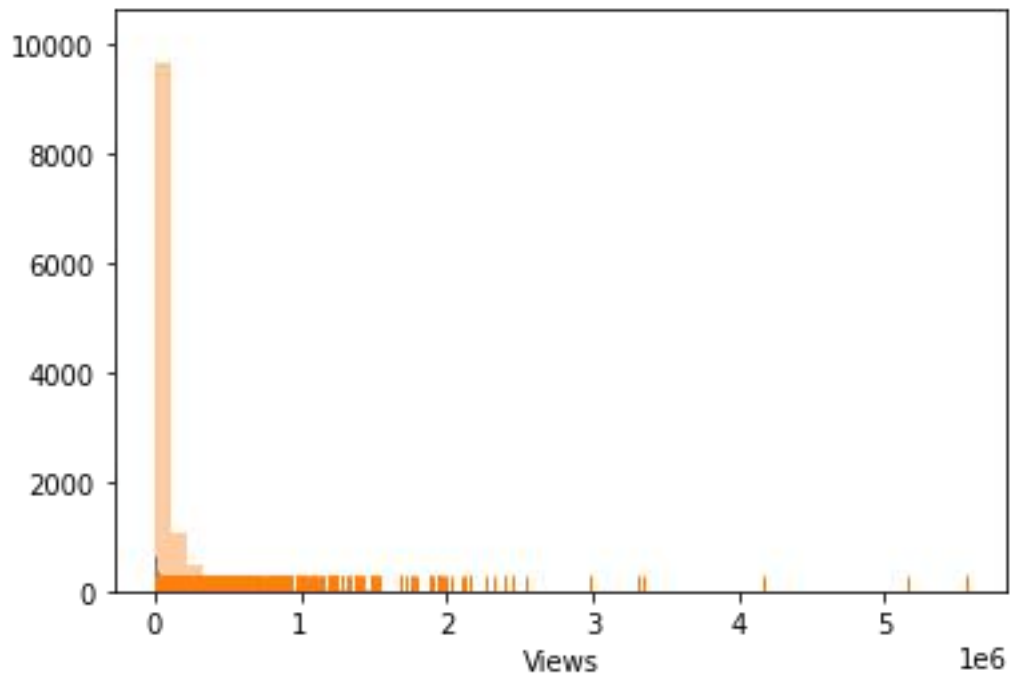
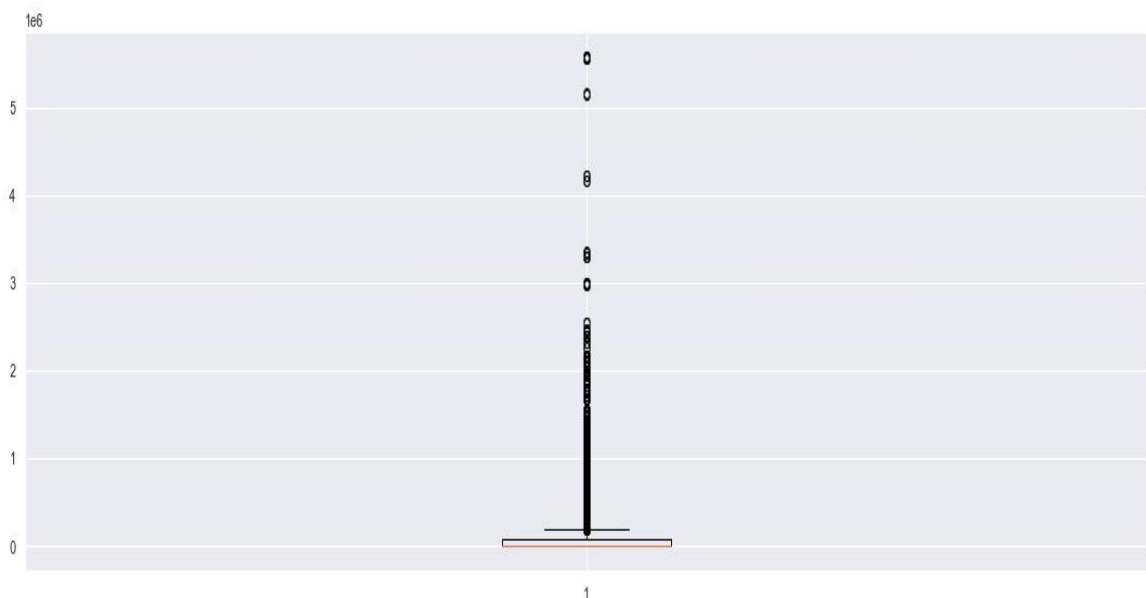Figure 2.2 shows the distribution of the views

Figure 2.3 a boxplot shows the distribution of views based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). The black circle points mean the outliers of views

In figure 2.2, 1e6 in the right bottom means six powers of 10. The views of most data are very low. In figure 2.3, the first and the third quartile is like a flat rectangle that is about to become a straight line. These two figures show that the distribution of views is very uneven. So, the number of three-quarters of the image's views is 77812.5 and the maximum number of views of one picture is 5585840.
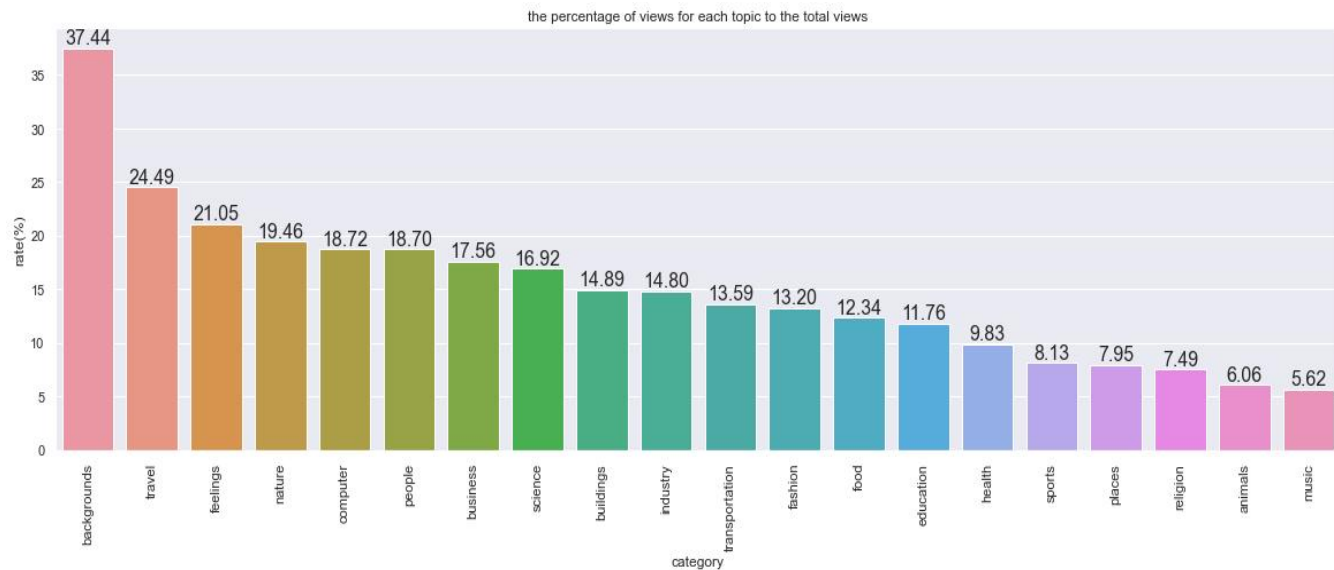
the percentage of views for each topic to the total views

Figure 2.4 shows the percentage of views for each topic to the total views

In figure 2.4, the rate of topic "background" is 37.44%, which ranks the number one, the topics "travel" and "feeling" are also very popular, with 24.49% and 21.05% respectively. The rates of topics "health", "sports", "religion", "animals", and "music" are below 10% and they are the least popular topics.
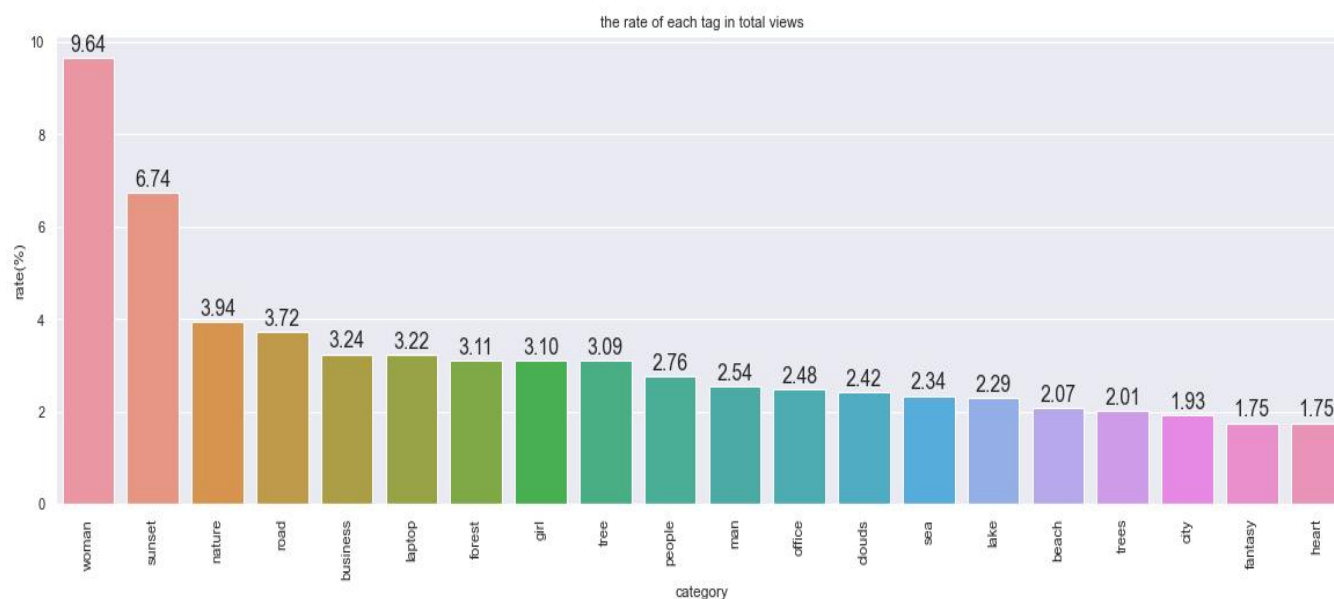

the rate of each tag in total views

Figure 2.5 shows the per tag as a percentage of total views (top 20 percentage)

The total views for these 24000 pictures are 1103587102 and there are 6135 tags in total. In figure 2.5, the "woman" tag ranks number one (9.64%), which is followed by the "sunset" tag and "nature", with 6.74% and 3.94% respectively. Among the top 20 tags, three tags which are "woman", "girl", "man", are related to humans, and nine tags "sunset", "nature", "forest", "tree", "clouds", "sea", "lake", "beach", and "trees", are related to nature. Three tags "business", "laptop", and "office" are related to the topic "business". It is difficult to intuitively determine which topic these tags ("road", "city", "fantasy", "heart") belong to. Because the topics "backgrounds" and "travel" are the most popular topics, the next step is to see what tags the two topics are made up of.
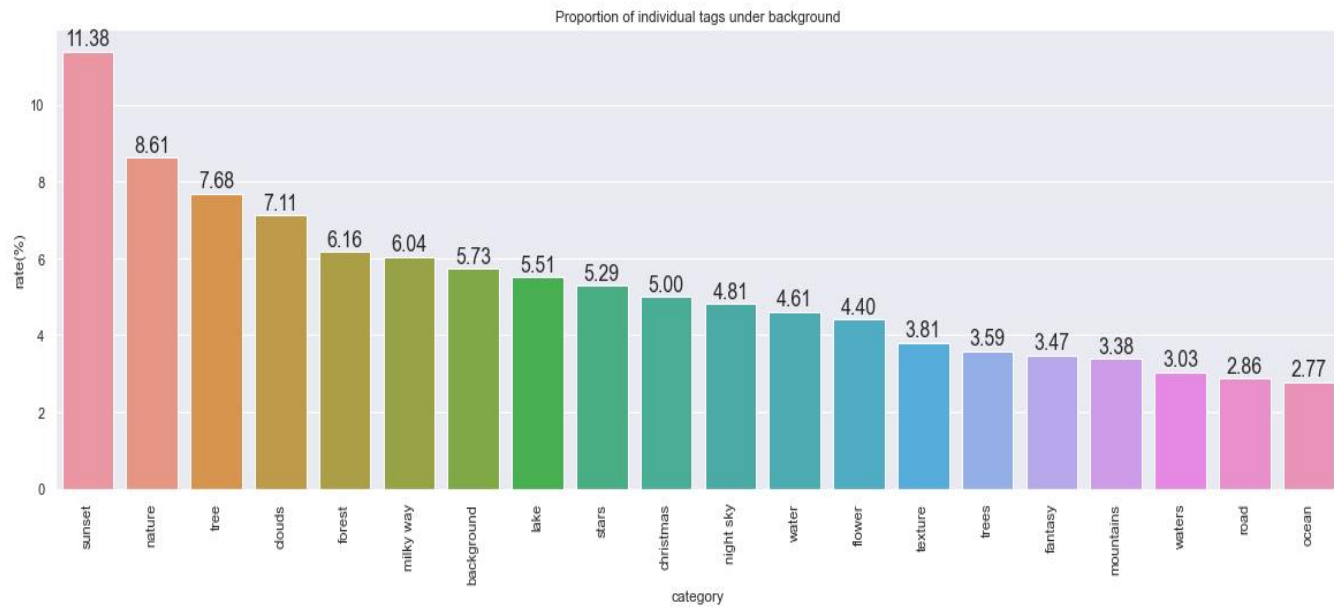
Figure 2.6 shows the top 20 proportion of individual tags under the topic "backgrounds"

Glance to figure 2.6, all these 20 tags in the topic "backgrounds" are related to nature stuff, and "sunset" is the most popular tag on this topic.
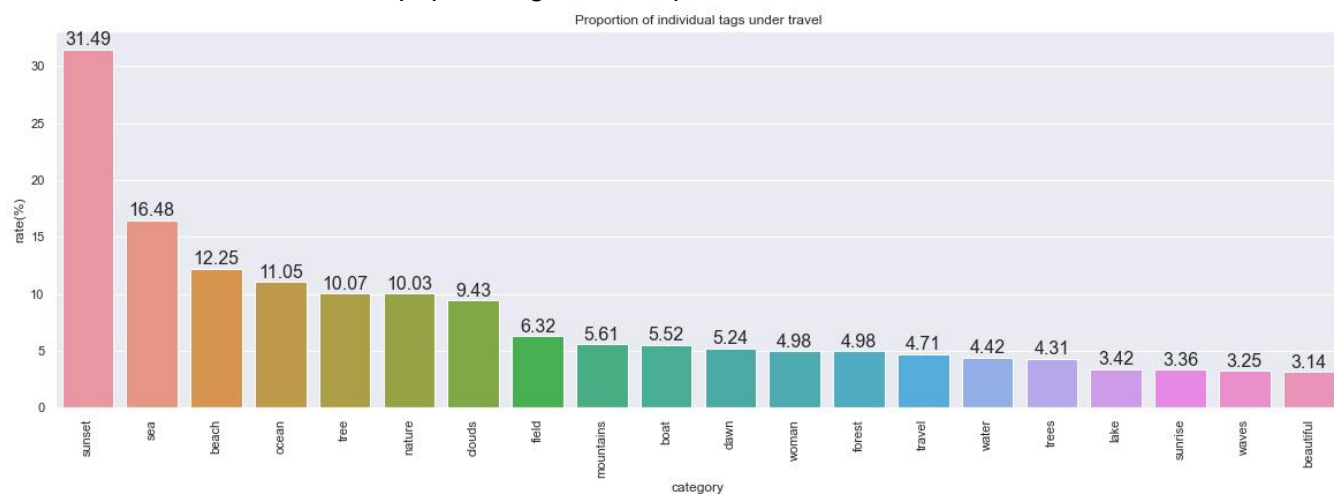

Figure 2.7 shows the top 20 proportion of individual tags under the topic "travel"

In figure 2.7, the second popular topic "travel" is like the topic "backgrounds" because many of the tags are related to nature and the tag "sunset" also ranks the number one. An interesting thing is that the tag "woman" which is the most popular tag also appears in the top 20 list After looking at the topics, this research started to focus on the popular tags: "woman" and "sunset".
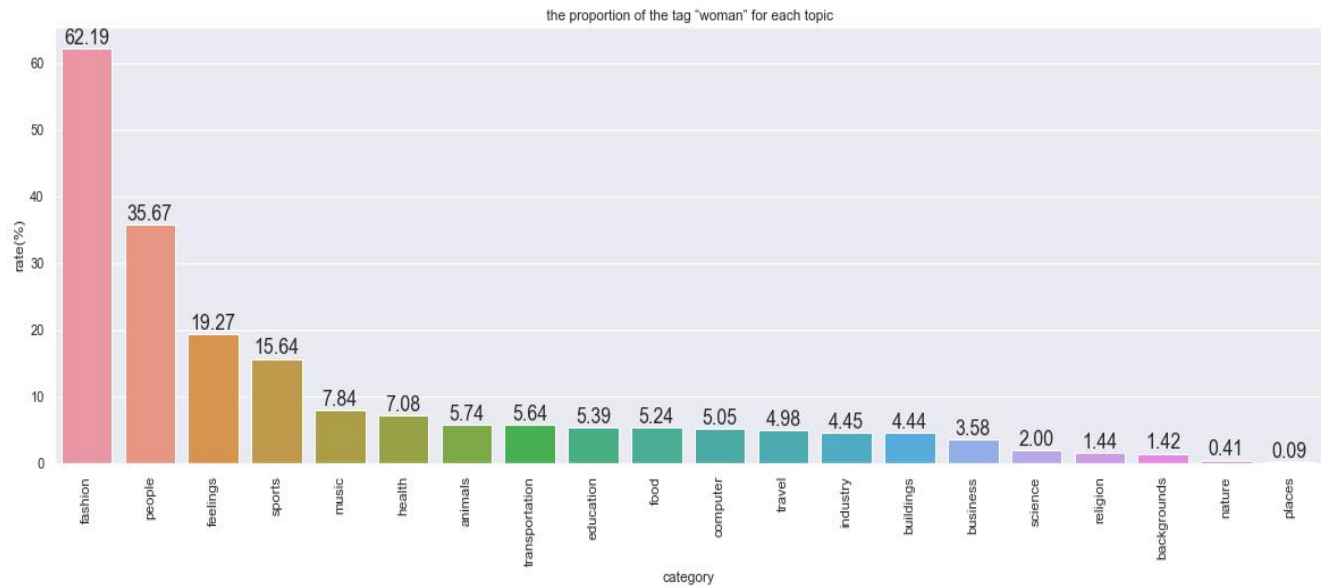
the proportion of the tag "woman" for each topic

Figure 2.8 shows the proportion of the tag "woman" for each topic

In figure 2.8, it is surprised to see that the tag "woman" occupied 62.19% of the topic "fashion" and 35.67% of the topic "people". This may prove that the female factor is bonded to fashion and people's life. On the topic "feeling", there are 19.27% of views related to women, which may be because women are more emotional. In stereotype, the woman does not always join in the sport, but woman contributes 15.64% of views on the "sport" topic. It can be understood that the topics "nature" and "places" are not related to the woman because they are more related to nature or some architecture. The topics "business", "science", and "religion" do not have a lot of views that are related to the woman, which means that females may not play an important role or not be popular in these zones in true life. To prove if this idea is true, I made figures to see what tag the 4 topics are made up of.
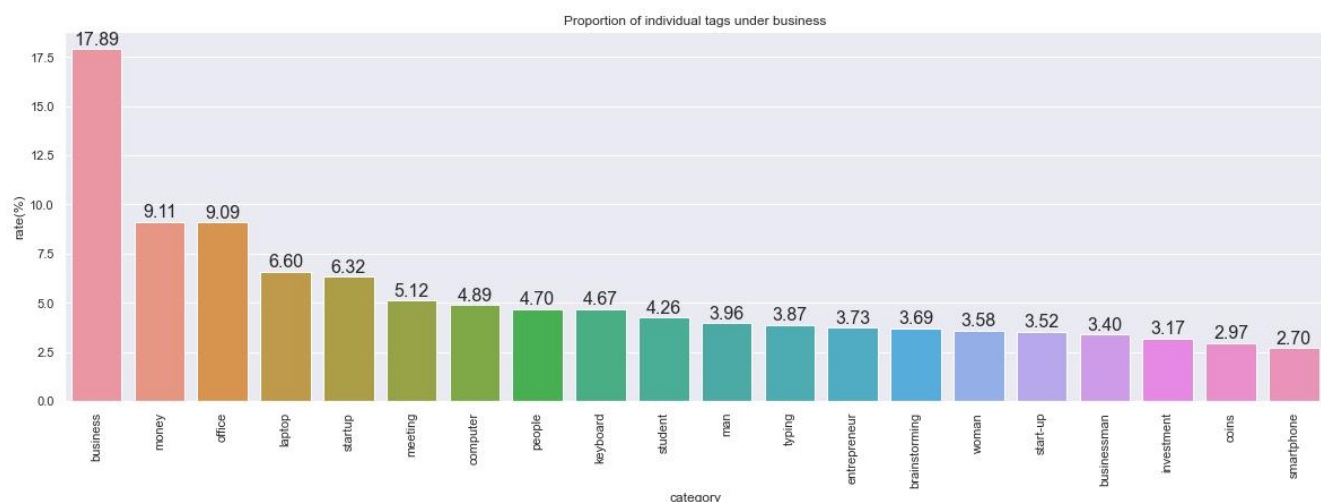

Proportion of individual tags under business

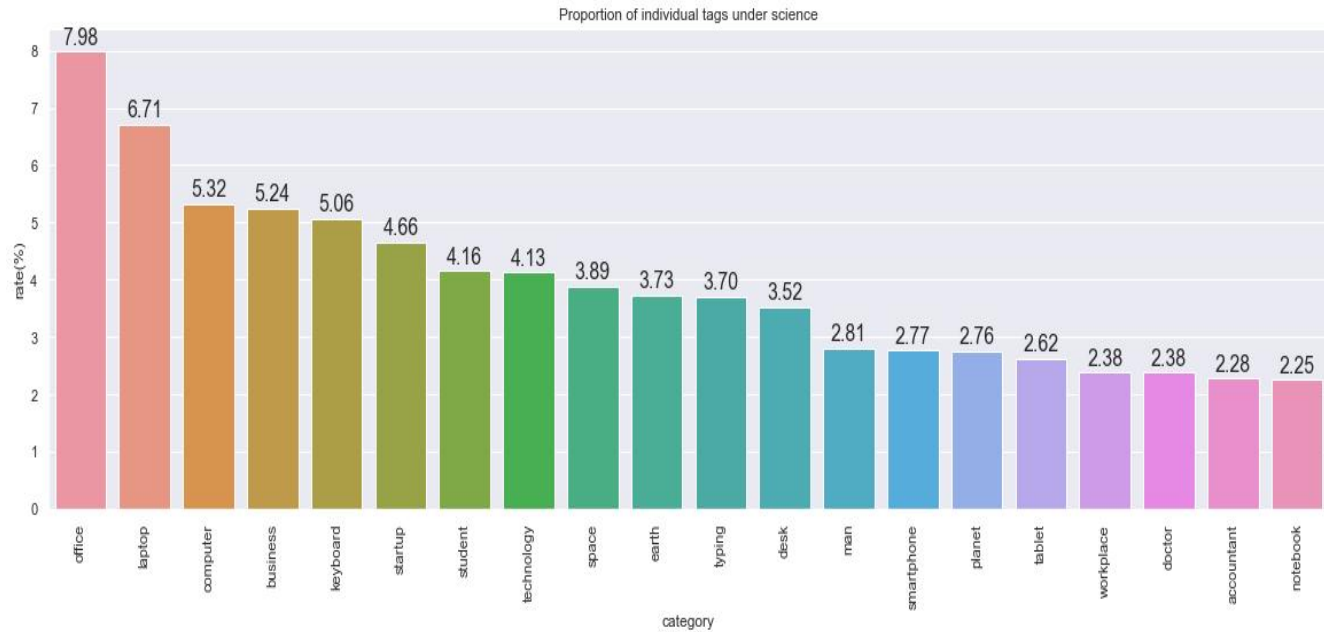Figure 2.9 shows the top 20 proportion of individual tags under the topic "business"

Figure 2.10 shows the top 20 proportion of individual tags under the topic "science"
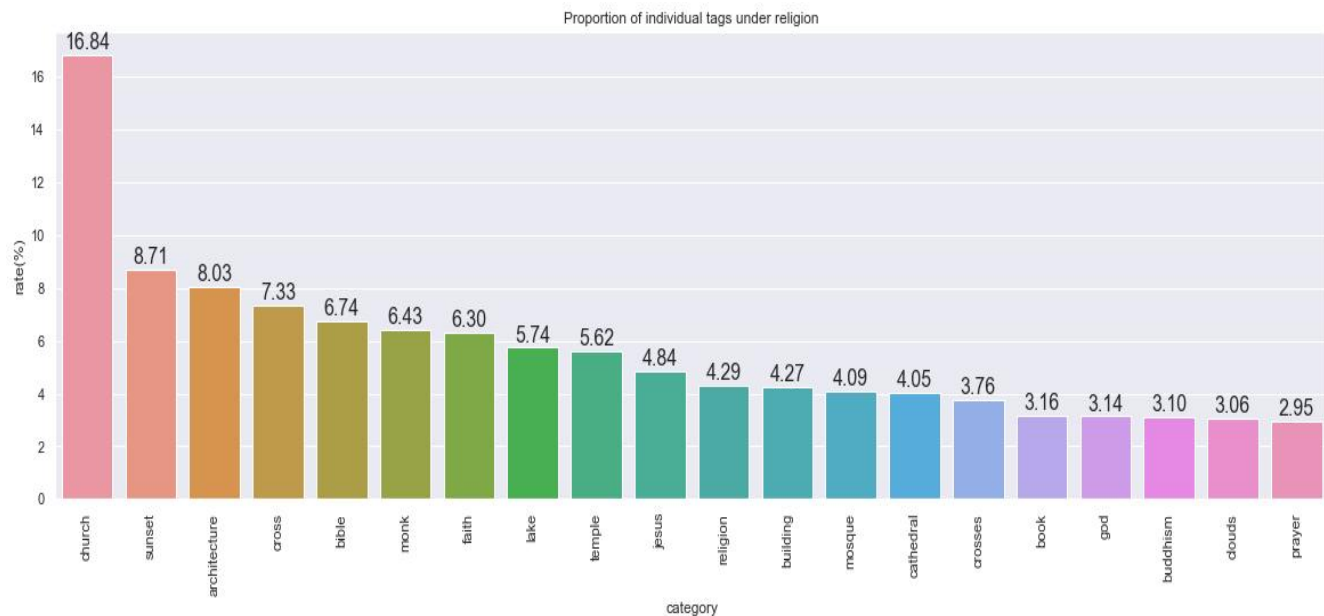

Figure 2.11 shows the top 20 proportion of individual tags under the topic "religion"

In figure 2.9, there are 3.96% of views that are related to "man" in the "business" and this rate is a little higher than "woman" (3.58%). In figure 2.10, In the topic "science", the rate of tag man is 2.81%, and the tag "woman" is not in the top 20 list. This means the "man" tag is more popular than the "woman" in the science zone. In figure 2.11, both these two genders are not in the top 20 list on the topic "religion". So, in some pictures that are related to science, and business, a picture with a "man" tag is more popular than a "woman".
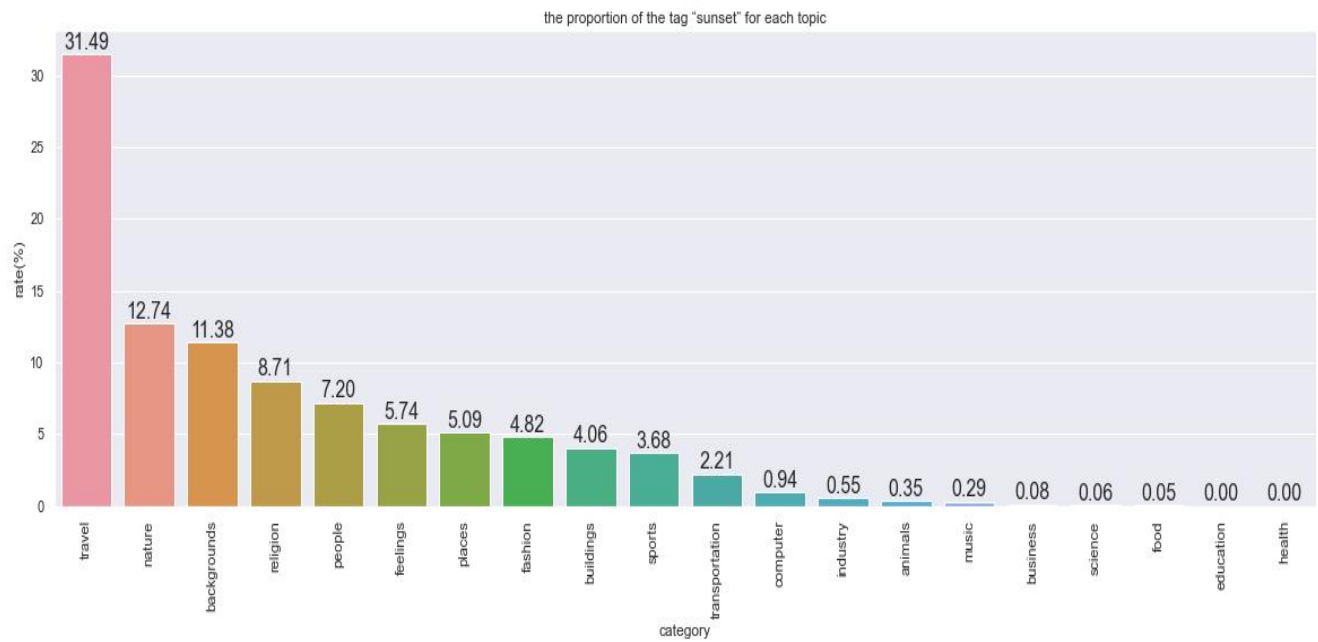
Figure 2.12 shows the proportion of the tag "sunset" for each topic

The second popular tag is "sunset". Figure 2.12 shows that sunset pictures are famous in the topic "background", "travel", and "nature", which means that people prefer to see the sunset in some natural topics and these three topics are the first, the second, and the fourth popular topics respectively. This may be the key reason for "sunset" ranks second.

However, a picture that is related to the female or sunset does not mean that this picture is a popular picture. People still need to care about if these tags' average number of views is also very high.
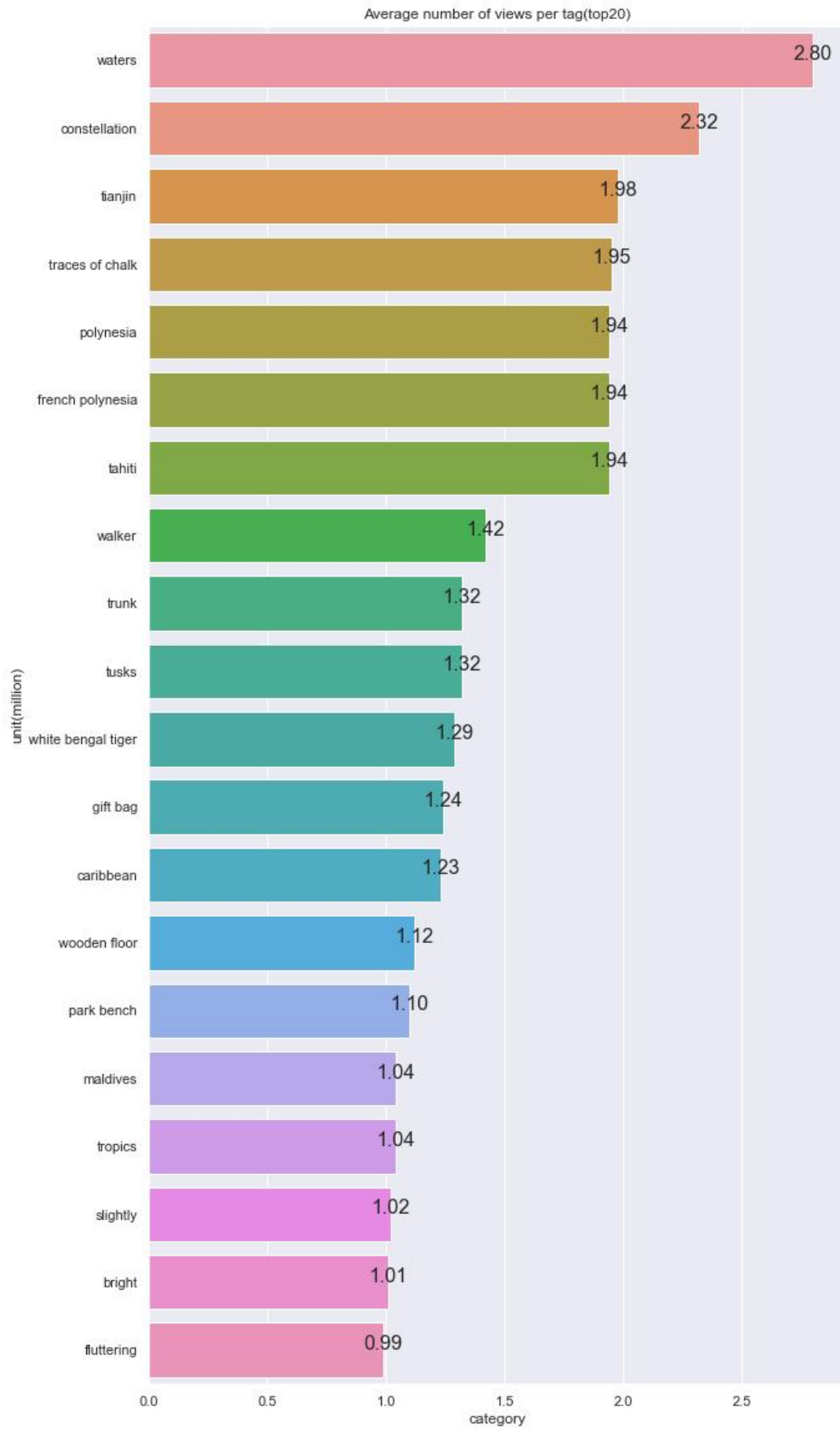
Figure 2.13 shows the average number of views per tag (top 20). unit is a million.

The previous famous tags cannot be seen in figure 2.13. but the tags in figure 2.13 cannot prove that a picture related to these tags is popular because their total number is small compared with the tags in figure 2.5 and it may be a coincidence or because of other factors, like the color of the picture.

Above all, the topic "background" gets the greatest number of views, and the tag "woman" also attracts a lot of views. In general, these two directions are more likely to have more views.

3. Predict Data

After analyzing the data, this step tried to predict the picture. Convolutional Neural Network (CNN) is a traditional way to analyze image data. However, this research used the Residual neural network (ResNet) rather than basic CNN. In general, a deeper network will be difficult to train, but a residual learning framework can help people train the data faster and more accurately (Kaiming He et al., 2015)
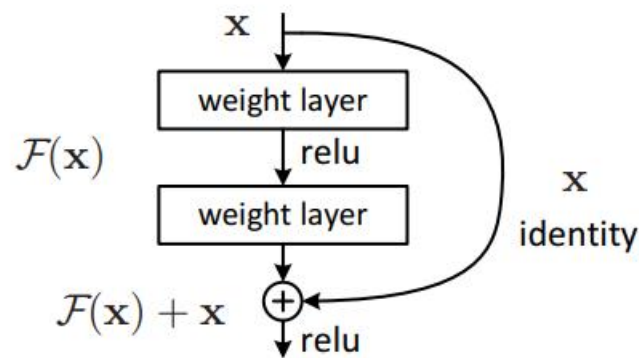

Figure 3.1 residual learning frame: a building block

In general, input an x into CNN, the CNN tries to analyze and gets a result H(x). But ResNet is trying to get the process F(x) in figure 3.1, and F(x)=H(x)-x. so, compared to calculating H(x) directly, Resnet tries to get F(x) at first, which can get results easier. All this theory has been proved by Kaiming He and his group members.

At first, I divided views into 7 ranges: "0~100", "100~500", "500~900", "900~1400", "1400~2000", "2000~3000", "3000~ ∞ " and identify them into 7 classes from 0 to 6. There is no reason why divide views into 7 classes because I wanted to see how the result was at first and I divide the data as casual. To get the result faster, I only chose 200 pictures at first. Then, this research used a CSV document that has two columns: the path of each picture and the number of views. To put the data into the neural network, I needed to transform data into the tensor format. and the laptop opened all these 200 pictures and saved them into a list then put them and the number of views into a custom class to resize pictures to 512X512 at first then transform them as a tensor. the reason I chose 512X 512 is that the size of many pictures is either 1280X823 or 823X 1280 and I do not want to lose too much content. By far, the length of the big tensor is 200 and for each of 200 parts, there are two small tensors: the picture matrix which is the RGB depth X Height X width, and the views. After this, 70% of data are identified as train data and others are validation data, then they are put in a data loader which can keep data manageable and helps to simplify the

machine learning pipeline. I chose resnet34 as my network frame because it is not too complicated and is easier for the device to run. Resnet34 means that this neural network has 34 layers. Because it is an experiment, so epoch was set as 2 at first. The device took 2 minutes, and the accuracy of validation was around 35%. The train and validation loss was more than 1.6 and could not be reduced. To reduce the loss, the learning rate was increased. The loss started to reduce but the speed of reduction is too slow, and the loss was still more than 1.6. One reason for so high loss was that the data size was too small. So, the next step was analyzing 12000 pictures.

It is worth noting that 12000 pictures are too large for many laptops and the python cannot open them at the same time. Instead of saving all pictures in a list, 200 pictures were chosen to open one time, and when the neural network finished analyzing 200 pictures, the laptop would open another 200 pictures automatically and analyze them, which was too inefficient. However, one problem is that it needed to wait for 3 or 4 hours every time to get the result. Instead of reading pictures directly, I finally found a new idea: let the custom class read the pictures and each picture would be saved as RGB content which can save a lot of storage. this time epoch was 30 and at last, one epoch was taken around 6 minutes and validation accuracy was about 0.46 which is better than the previous result. But both train loss and validation loss were around 1.52 and even if the accuracy and loss increased and decreased respectively in these 30 epochs, the step of increasing accuracy or decreasing loss was too small.

I tried to reduce the learning rate or change the kernel size, but the result was not better. Another reason is that the ResNet may not be able to find different features from 7 kinds of popular levels and I tried to reduce the number of the classes from 7 to 2. The picture will be identified "popular picture" if the views are higher than the average number of views and vice versa.
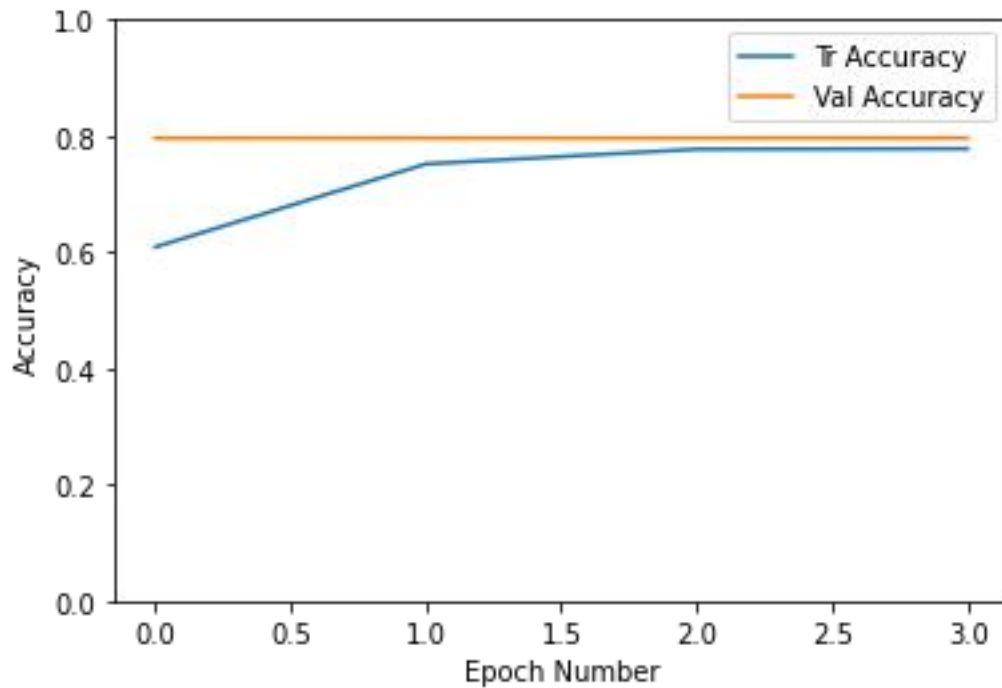
Figure 3.2 shows how validation accuracy and train accuracy changed in 3 epochs. The blue line means train dataset accuracy and the orange line means validation dataset accuracy

In figure 3.2, the blue line increased a bit, but the orange line did not change. However, the validation accuracy was much better than before, which was about 79%.
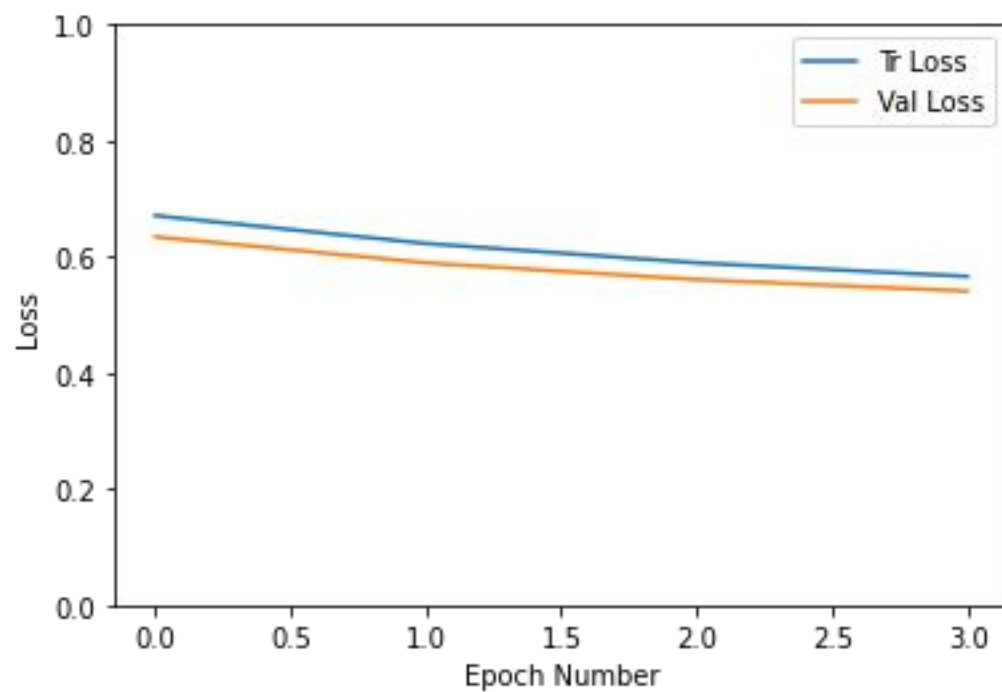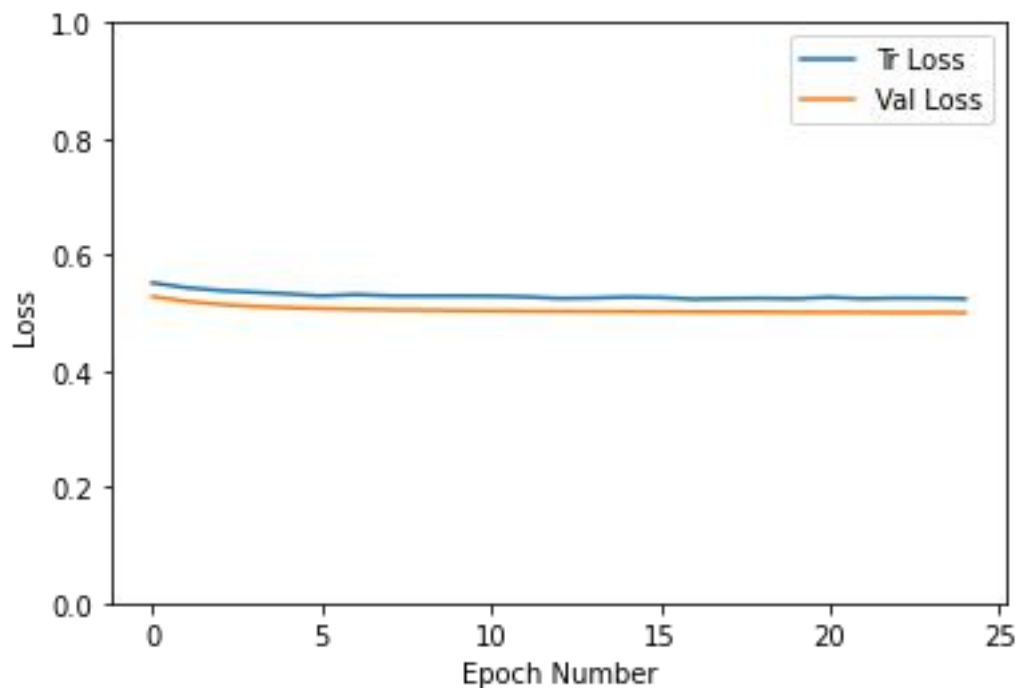

Figure 3.3 shows how validation loss and train loss changed in 3 epochs. The blue line means train dataset accuracy and the orange line means validation dataset accuracy

In figure 3.3, it was obvious that the two lines had a decreased trend. To see if the

result can be better, I set the learning rate smaller and increase the epoch to

25.

Figure 3.4 shows how validation loss and train loss changed in 25 epochs. The blue line means train dataset accuracy and the orange line means validation dataset accuracy.
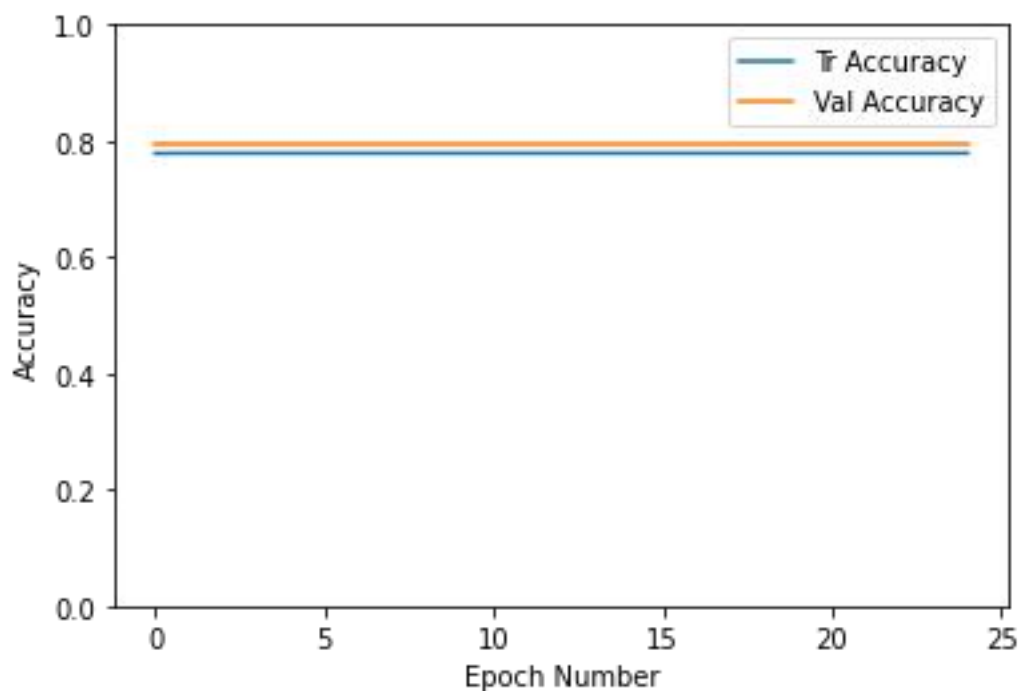

Figure 3.5 shows how validation accuracy and train accuracy changed in 25 epochs. The blue line means train dataset accuracy and the orange line means validation dataset accuracy

In figure 3.4, the loss became smaller after changing the learning rate, but it did not decrease a lot after 30 epochs. In figure 3.5, the result did not change a lot. To improve the result, I used the pre-train model from the Pytorch website, and I

downloaded another 12000 pictures to increase the number of pictures. However, there were many replicated pictures in these 24000 pictures because I could only choose what kinds of pictures to download rather than what pictures I could download and I did not know if they were different, for example, if I chose "background" popular pictures, the API may download same pictures that it had downloaded in the first time. To use the pre-train model, the parameter: the number of classes should be set as 1000 because the pretrain model output is 1000. But the output of this research is 2 and adding the linear layer can set the output of the model as 2. This model added two linear outputs:1000 to 256 and 256 to 2 because if just adding one output from 1000 to 2, the step is too big, and the result may be worse.

```
8] | ✓ 23m 58.7s

  Epoch: 1/1
  Epoch: 001, Training: Loss: 0.5572, Accuracy: 73.6190%,
              Validation: Loss: 0.5572, Accuracy: 79.3333%, Time: 1437.6708s
  Best Accuracy for validation : 0.7933 at epoch 001
```

Figure 3.6 shows the result of using the pre-train model

However, the result in figure 3.6 did not change a lot. I thought the 0.79 result is not bad. After downloading a new picture from pixabay, which has 700000 views, I inputted that picture in the model and the result is an unpopular picture. And I tried another picture which has more than 1000000 views and the result was still an "unpopular   picture".
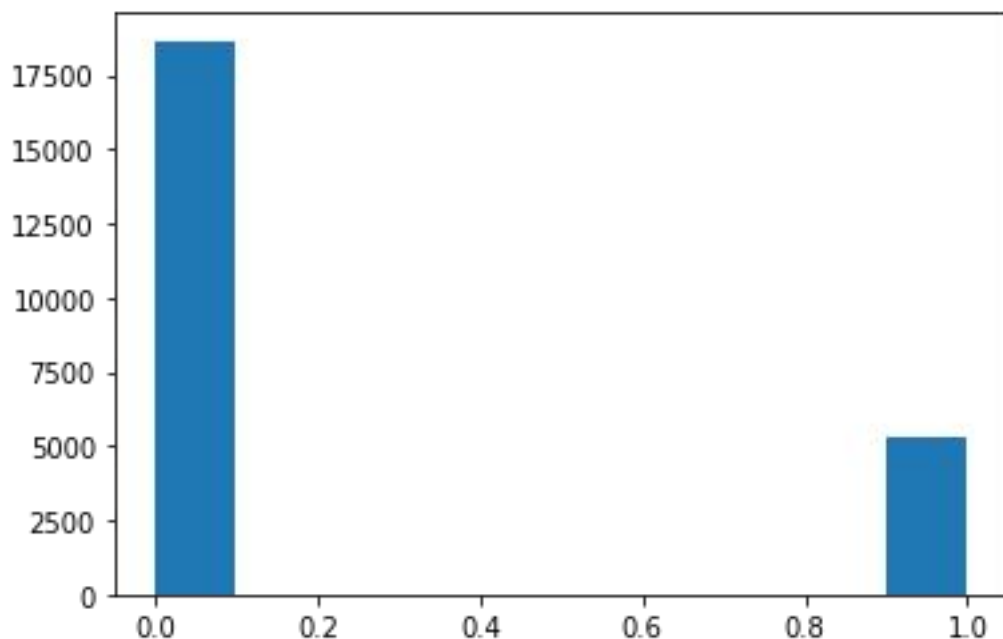

Figure 3.7 shows the distribution after the two classes were identified as 1 and 0

In figure 3.7, there are 77% of data were identified as 0 which means that they are below the mean. So, I guessed that if the model just predicted all the output is 0, the accuracy will still be 77%. This means the accuracy may be related to the distribution. So, I used the median to replace the mean.

Epoch: 1/1
Epoch: 001, Training: Loss: 0.6796, Accuracy: 51.5060%,
            Validation: Loss: 0.6796, Accuracy: 56.5833%, Time: 931.5171s
Best Accuracy for validation : 0.5658 at epoch 001

1.0
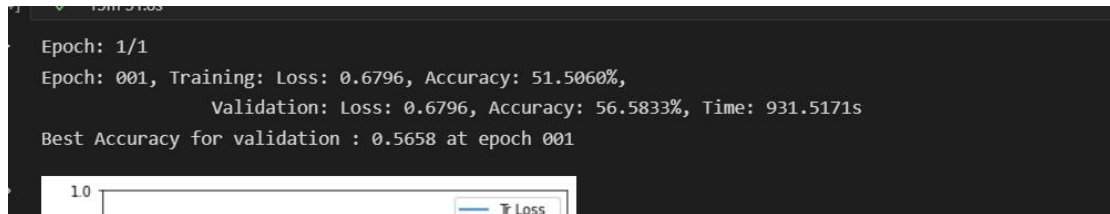   — Tr Loss

Figure 3.8 shows the result of ResNet when using the median to divide two classes

The validation result in figure 3.8 is 56% which means even if the distribution of data did impact the accuracy the ResNet still could find some features of the two classes and distinguished them.

Then I changed the number of classes to 4: "very unpopular picture"," unpopular picture", "popular picture", and "very popular picture". If the views were smaller than 1247 which is the first quartile, it would be identified "very unpopular picture". If the views were between1247 to 14122.5 which was the median and it would be identified "unpopular picture" and if the views were between14122.5 to 77812.5 which is the third quartile, it would be identified "popular picture". If the views were larger than 77812.5, it would be identified as a "very popular picture".
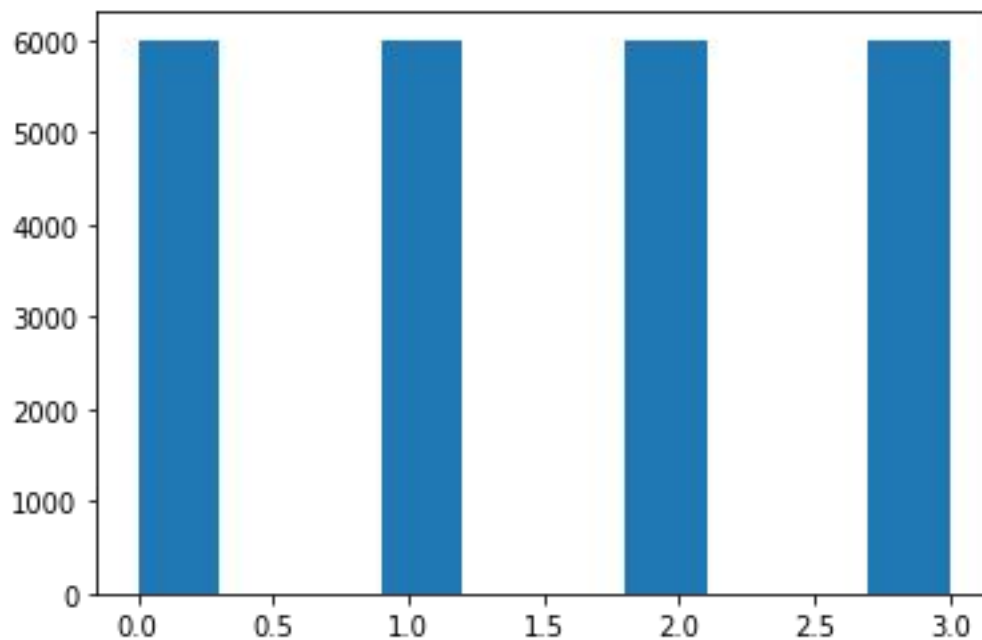

Figure 3.9 shows the distribution after the three classes were identified as 0,1,2 and 3

15m 20.4s
Epoch: 1/1
Epoch: 001, Training: Loss: 1.3702, Accuracy: 26.7143%,
            Validation: Loss: 1.3702, Accuracy: 30.1944%, Time: 919.6762s
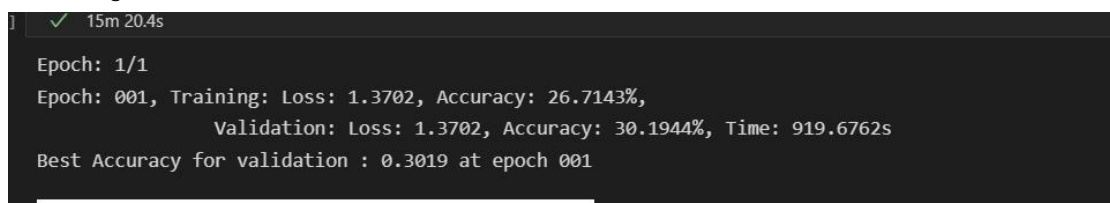Best Accuracy for validation : 0.3019 at epoch 001

Figure 3.10 shows the result of 4 classifications

In figure 3.9, there are 25% data in each class. And the accuracy in 3.10 is only 30% this time. Above all, the accuracy was impacted by the distribution of the data, but the ResNet can find some features of different levels of pictures.

4. Conclusion

In general, through the data in pixabay, the woman tag and sunset tag are more attractive now. When people want to take some pictures that are related to fashion or people, they should choose some pictures that are related to women, and when they are interested in some traveling or nature stuff, the sunset picture is the first choice. However, this research did not contribute a lot to predicting the picture part. The Resnet model did find some features in some popular pictures, but the result is still not good and can't help people to predict so it still needs to be developed. To continue to find whether the neural network can find some features of the popular pictures or not, researchers need to download much more different pictures to test.

Bibliography

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015. s.l.: Computer Vision Foundation.

Pixabay API documentation URL: https://pixabay.com/api/docs/
Pixabay API code URL: https://boook24.com/?p=1620
Resnet network code:
https://pytorch.org/vision/0.8/_modules/torchvision/models/resnet.html