Department of Electrical and Computer Engineering

EL-GY 6123 INTRODUCTION TO MACHINE LEARNING

CROSS-MODAL RETRIEVAL

By

Guyu Liu
Han Jiang
Jiacheng Wang
Zhong Han

Spring 2019

Advisor: Prof. Sundeep Rangan

May 16th, 2019

ABSTRACT

This project aims to address the issue of cross-modal search of similar content on social networking sites and online shopping platforms. Through this technology, users can retrieve interrelated multimedia data between multiple modalities, such as text, images, and so on. The core of our project is to create a common subspace in which the items of different modalities can be directly compared or transferred to each other. We used an adversarial cross-modal retrieval method. First, we used the vgg16 model, which we learned in the course, to extract features from the input image or text. Then we performed triplet constraints on the collected feature set to obtain more precise feature information. These features were then compared to the images and text in the database. Experimental results on cross-modal retrieval tasks demonstrate that the proposed method is effective. In addition, we will further utilize this model to predict similar content of videos, audios and 3D models corresponding to images.

1. INTRODUCTION

1.1 Problem statement

With the rapid development of mobile devices, social networks, and self-media platforms, multimedia data such as text, images, and videos increasing exponentially in recent years. On social media websites such as Facebook and Twitter, users share a huge amount of text and pictures every day. These different types of data usually describe the same objects or events. Therefore, it is crucial important for researchers to have a novel method to collect these different types of data that have similar content.

1.2 General information

Text-based retrieval and content-based retrieval are two of the most common image retrieval methods. If using semantic information based retrieval, it is necessary to mark the semantic attributes of a large number of images before retrieval. This kind of labeling has subjective deviation, and its time cost is high. In addition, the semantic attributes cannot fully express the rich information contained in the image, so that the retrieval effect is limited. As for content-based retrieval—searching by image—has its unique advantages. For example, Google has already supported this kind of searching these years. It searches according to the description of the image features, and then compares the similarities between the features. After that it sorts according to the degree of similarity and gives the final search results.

The general process of the image retrieval system is to first pre-process the image after the user inputs a picture. The pre-processing includes image enhancement, binarization, and cropping to a uniform size, etc. The system then imports the normalized processed image into the feature extraction model—in practice, the researchers typically use the trained Convolutional Neural Network (CNN) model—again, the same processing is done for retrieving images from the database. In addition to the CNN model, models such as vgg, inception, alxnet, and resnet are often used to extract features. As a result, both the image to be retrieved by the user and the image in the database are represented by features of a certain dimension. Then we use Euclidean distance or cosine distance based on metric learning to compare the features of the images that need to be retrieved with the features of all the images in the database. The higher the feature similarity, the more similar we can think of the two images.

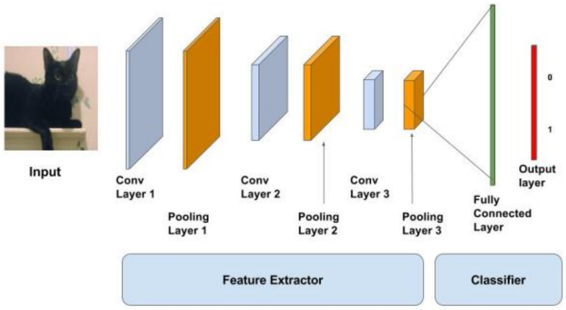


Figure 1.1 Model visualization of typical CNN

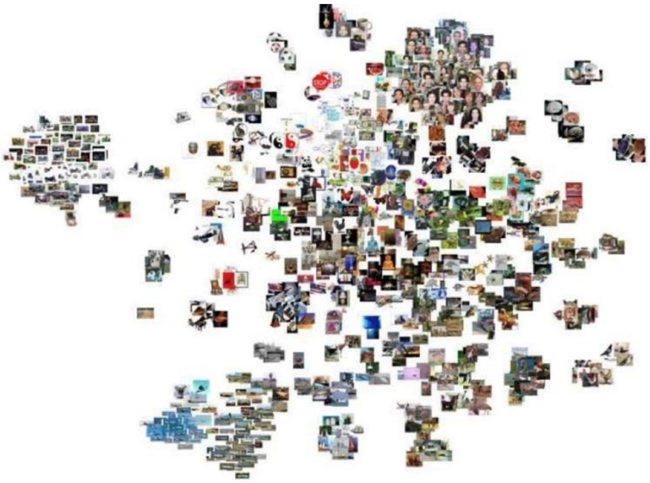


Figure 1.2 An example of feature clustering t-SNE visualization

We used to learn to classify with CNN in class, so it was originally used for image recognition. Taking Figure 1.1 as an example, CNN can be regarded as two parts, feature extractor and classifier. Through the multi-layered neural network, the abstract features are gradually extracted from the image, so there are indicators that can be distinguished and judged, and the classifier can also recognize what information the image is expressing. If the classifier is removed, the remaining feature extractors are used to extract features from the massive images, and with the help of cosine similarity and other indicators, we can achieve the image retrieval effect shown in the figure.

Figure 2 is an example of visualizing the clustering result by using the t-SNE clustering tool after all the images have been extracted from the CNN. t-SNE is a dimensionality reduction algorithm, which can map data in high-dimensional space into two-dimensional space. The dimension reduction process does not destroy the relationship between data itself. In this way, the relationship between high-dimensional data can be displayed in a two-dimensional space, which is easier to understand intuitively. Later we will also use this tool in the study of cross-modal retrieval. In the Figure 1.2, we can see that similar pictures are clustered together. For instance, the images of many faces in the upper right corner are gathered together.

We all know the network is fulfilled with multimodal multimedia data from different data sources, so it is necessary to have an information retrieval system that can adapt to different modalities. For example, when someone visits a museum and is interested in painting, he can take a photo and enter it into a cross-modal retrieval system to output a textual description associated with the painting.

1.3 Goal

The goal of this project is to Use one of the modalities (such as text) to simultaneously retrieve other modal results (such as images, video, audio, etc.) associated with it, that is, cross-modal retrieval.

2. DESIGN PROCEDURE

2.1 Simple Feature Extraction Test

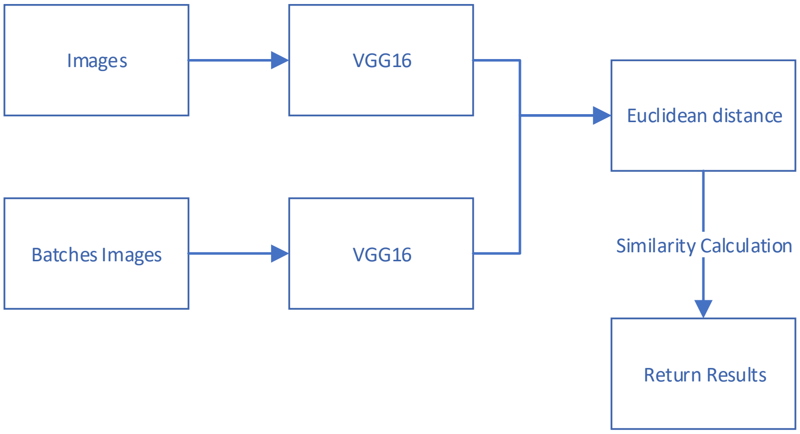


Figure 2.1 Schematic diagram of the basic process

As a preliminary transition of the project, our team first used the pre-trained vgg16 model on ImageNet to extract features from the image, and then used Euclidean distance to calculate the eigenvalues, thus implementing a content-based image retrieval program. Because the ImageNet dataset has more than 14 million images covering more than 20,000 categories, the pre-trained vgg16 model on ImageNet can effectively extract the features of the images, making the distance between similar images more similar. Figure 2.2 and 2.3 are the results of our program running, looking for a picture of basketball on the Internet, retrieving and displaying the four most similar pictures in the data set, and printing out their European distance.

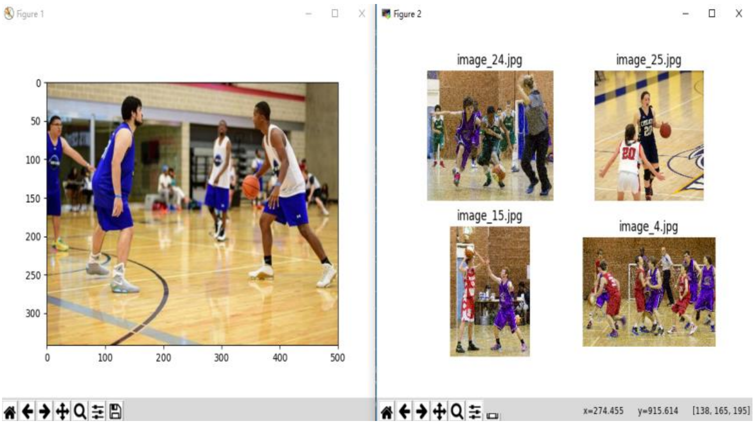
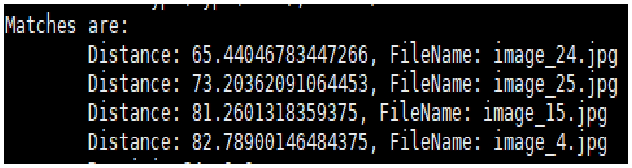


Figure 2.2 One retrieved image and the four most similar search results



```
Matches are:
        Distance: 65.44046783447266, FileName: image_24.jpg
        Distance: 73.20362091064453, FileName: image_25.jpg
        Distance: 81.2601318359375, FileName: image_15.jpg
        Distance: 82.78900146484375, FileName: image_4.jpg
```

Figure 2.3 Four Euclidean distances corresponding to the search results

2.2 Design Decision

In cross-modal retrieval, different modal data presents low-level feature heterogeneity and high-level semantic correlation. When processing different modal data of the same category or meaning, the features of the text are usually represented by Feature Hashing, Word2vec, Global Vectors (GloVe), etc., and the image is represented by visual features such as Scale-Invariant Feature Transform (SIFT) and CNN features. We visually display the text features and image features in the wiki dataset using t-SNE clustering, and obtain the results shown in Figure 2.4. The blue and red dots represent the image and text features, respectively, in completely different feature spaces but represent the same semantic topic, so different forms of data and models are heterogeneous, and it is difficult to directly measure the similarity between them. So the main problem of cross-modal retrieval is how to measure the similarity between different modes. The key lies in how to link different modalities across the semantic gap.

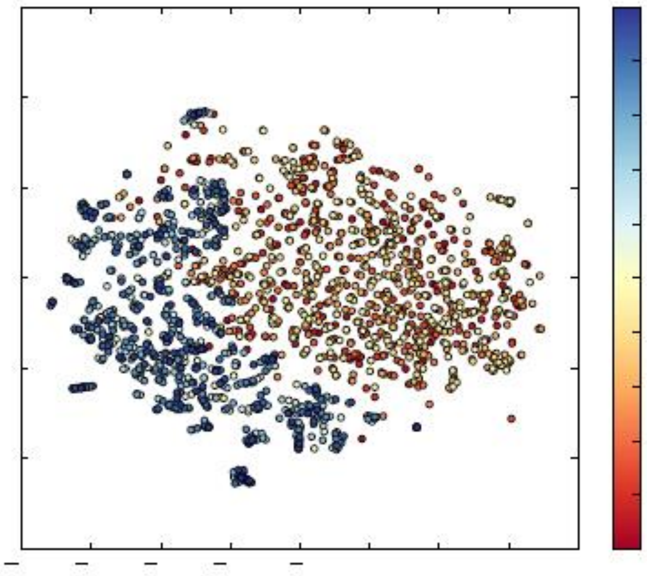
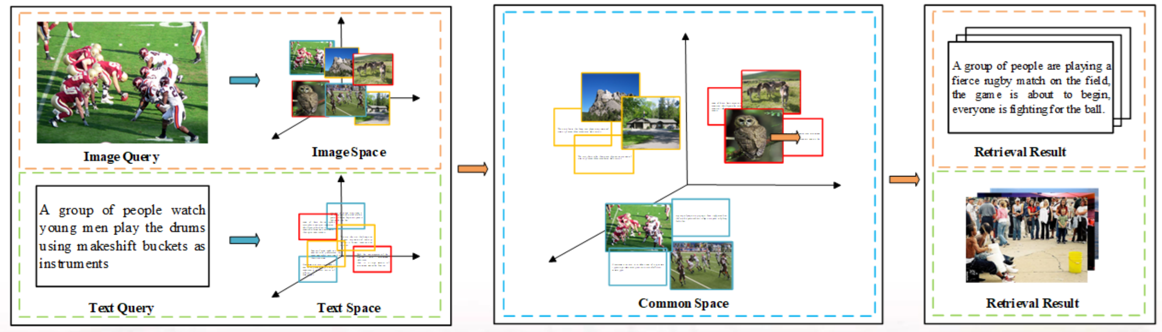


Figure 2.4 Visualization of image and text feature distribution

One novel approach is Joint Feature Selection and Subspace Learning. It learns the representative features of different modalities and at the same time models the correlation of similar data between different modalities, and then performs similarity metrics in the common potential space to achieve cross-modal retrieval. As can be seen from Figure 2.5, originally, both the image and the text have their own feature space. Through the learning of the common subspace, the associated text and images are closer in the common subspace, while the unrelated text and images are farther apart.



A group of people watch young men play the drums using makeshift buckets as instruments

A group of people are playing a fierce rugby match on the field, the game is about to begin, everyone is fighting for the ball.

Figure 2.5 Method for distinguishing multi-module information by content

3. DESIGN DETAILS

3.1 Problem Formulation

Focus on cross-modal retrieval problem, we regard the images and text as the bimodal data. Mathematically, we assume that there is a collection of n instances of image-text pairs, denoted as $O=\{o_i\}_{i=1}^{n}, o_i=(v_i,t_i)$.where $v_i\in\mathbb{R}^{d_v}$ is an image feature and $t_i\in\mathbb{R}^{d_t}$ is a text feature vector. d_V and d_t are the feature dimensions with, usually, $d_V\neq d_t$. In additional, each instance o_i is also assigned a semantic label vector $y_i=[y_{i1},y_{i2},\dots,y_{ic}]\in\mathbb{R}^{C}$, where c is the total number of semantic categories.

Since the image features V and text feature T have different statistical properties and distributions, we can't compare against them each other in cross-modal retrieval problem. In order to make the data comparable, we give a common subspace S the image features and text features can be projected to as $S_V = f_V(V;\theta_V)$ and $S_T = f_T(V;\theta_T)$, where the mapping functions $S_V\in\mathbb{R}^{m\times n}$ and $\mathrm{S}_T\in\mathbb{R}^{m\times n}$ are the transformed features of images and texts in subspace.

3.2 Label Prediction

In order to ensure that the intra-modal discrimination in data is preserved after feature projection, a classifier is deployed to predict the semantic labels of the items projected in the common subplace. For this purpose, a feed-for-forward network activated by softmax was added on top of each subplace embedding neural network. This classifier takes the projected features of the instances o_i of coupled images and texts as training data and generates as output a probability distributions $\hat{p}$ to formulate the intra-modal discrimination loss as follows:

$$L_{imd}(\theta_{imd})=-\frac{1}{n}\sum_{i=1}^{n}\left(y_i\cdot\left(\log\hat{p}_i(v_i)+\log\hat{p}_i(t_i)\right)\right)$$

where θ_{imd} denotes the parameters of the classifier, n is the number of instances within each mini-batch, y_i is the groundtruth of each instance, while $\hat{p}$ is the generated probability distribution per item(text or image).

3.3 Structure Preservation

In order to ensure the preservation of inter-modal invariance, we aim at minimizing the gap among the representations of all semantically similar items from different modalities, while maximizing the distance between semantically different items of the same modality. Inspired by the ranking-based cross-media retrieval approaches[1][2], we enforce *triplet Constraints* onto the embedding process via a triplet loss term we formulated for this purpose.

All distances between the mapped representations $S_V = f_V(V;\theta_V)$ and $S_T = f_T(V;\theta_T)$ per coupled item pair were computed and sorted using the l_2 norm:

$$L_2(v,t)=\left\|f_V(v;\theta_v)-f_T(t;\theta_T)\right\|_2$$

Then, we also select negative samples from unmatched image-text pairs having different semantic labels to build the sets of triplet samples per semantic label l_i: $\left\{\left(v_i,t_i^{+},t_j^{-}\right)\right\}_i$ and $\left\{\left(t_i,v_i^{+},\mathrm{v}_j^{-}\right)\right\}_i$. In this way of sampling, we can ensure that non-empty triplet sample sets will be constructed independently of how samples in the original dataset were organized into the mini-batches.

Finally, we compute the inter-modal invariance loss across image and text modalities using the following expressions that take as input the sample sets, respectively:

$$L_{imi,V}\left(\theta_{\mathrm{V}}\right)=\sum_{i,j,k}\left(l_2\left(v_i,t_j^+\right)+\lambda\max\left(0,\mu-l_2\left(v_i,t_k^0\right)\right)\right)$$

$$L_{imi,\mathrm{T}}\left(\theta_{\mathrm{T}}\right)=\sum_{i,j,k}\left(l_2\left(t_i,v_j^+\right)+\lambda\max\left(0,\mu-l_2\left(t_i,v_k^0\right)\right)\right)$$

Then the overall inter-modal invariance loss can now be modeled as a combination of $L_{imi,V}\left(\theta_V,\theta_{\mathrm{T}}\right)$ and $L_{imi,T}\left(\theta_V,\theta_{\mathrm{T}}\right)$:

$$L_{imi}\left(\theta_V,\theta_{\mathrm{T}}\right)=L_{imi,\mathrm{V}}\left(\theta_V\right)+L_{imi,\mathrm{T}}\left(\theta_{\mathrm{T}}\right)$$

In addition, the regularization term below is introduced to prevent the learned parameters from overfitting, where F denotes the Frobenius norm and W_v^l, W_t^l represent the later-wise parameters of DNNs.

$$L_{reg}=\sum_{l=1}^{L}\left(\left\|W_{\mathrm{V}}^l\right\|_F+\left\|W_t^l\right\|_F\right)$$

3.4 Feature Projector

Based on the above, the loss function of the feature projector, referred to as embedding loss, is formulated as the combination of the intra-modal discrimination loss and the inter-modal invariance loss with regularization:

$$L_{emb}\left(\theta_V,\theta_{\mathrm{T}},\theta_{imd}\right)=\alpha L_{imi}+\beta L_{imd}+L_{reg}$$

Where the hyper-parameters α and β control the contributions of the two terms.

Adversarial Learning : Optimization: The process of learning the optimal feature representation is conducted by jointly minimizing the adversarial and embedding losses. Since the optimization goals od these two objective functions are opposite, the process runs as a minimax game of the two concurrent sub-processes:

$$\left(\theta_V,\theta_{\mathrm{T}},\theta_{imd}\right)=\arg\min_{\theta_V,\theta_{\mathrm{T}},\theta_{imd}}\left(L_{emb}\left(\theta_V,\theta_{\mathrm{T}},\theta_{imd}\right)\text{-}L_{adv}\left(\hat{\theta}_D\right)\right)$$

$$\hat{\theta}_D=\arg\max_{\theta_D}\left(L_{emb}\left(\hat{\theta}_V,\hat{\theta}_{\mathrm{T}},\hat{\theta}_{imd}\right)\text{-}L_{adv}\left(\theta_D\right)\right)$$

4. DESIGN VERIFICATION

4.1 Testing

Wikipedia is the most widely used dataset for the cross-modal retrieval problem, we collected 2866 images from *Wikipedia*, 10 class total. We choose 2173 images to be our traning set and 693 to be our test set. In addition, we use the *Pascal Sentence* to be our second dataset, because it has more class number and every image is labeled by 5 sentence.

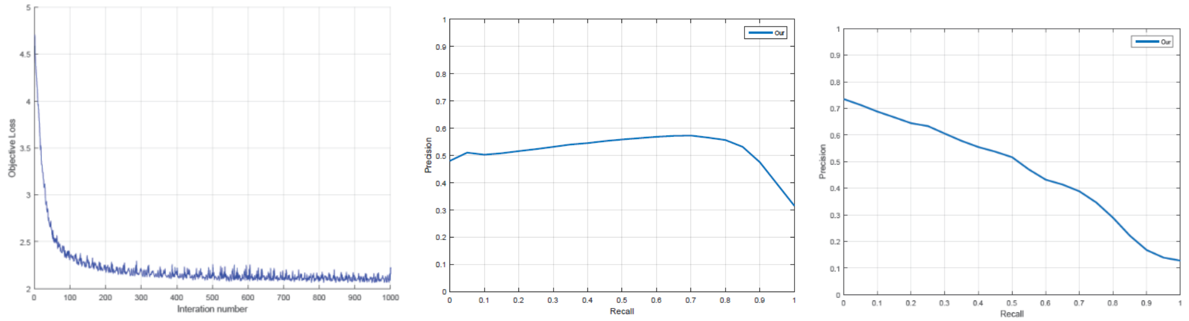
Dataset name	Test/train	Number of class
Wikipedia	2173/693	10
Pascal Sentence	800/200	20

In this project, we use Mean Average Precision to estimate the perfomence of results. We compare the results from different dataset and plot the loss curve from the Wikipedia dataset. As we can see from the figure 2.6, the performance on Pascal Sentence is better than on Wikipedia.

Dataset name	Retrieval the texts based on images(MAP)	Retrieval the images based on texts(MAP)	Average MAP
Wikipedia	0.53	0.46	0.495
Pascal Sentence	0.52	0.54	0.530

Table 1: Performance cross retrieval using different datasets

After the discussion, we think that the reason is that Pascal Sentence has better training data, because every image has their sentence label, so the neural network can have a better learning on the Pascal Sentence.



Wiki loss

Wiki i2t pr

Wiki t2i pr

Figure 2.6 Loss curve from Wikipedia dataset

In addition, we also compared the results from different text features and different transfer learning loss function, we found that the glove and adversarial loss has an better performance in this problem, and the performance will be improved when we introduce the adversarial loss.

Text features	Retrieval the texts based on images(MAP)	Retrieval the images based on texts(MAP)	Average MAP
Bow	0.50	0.42	0.460
Word2vec	0.52	0.44	0.480
Glove	0.53	0.46	0.495

Table 2: MAP performace using different texts features

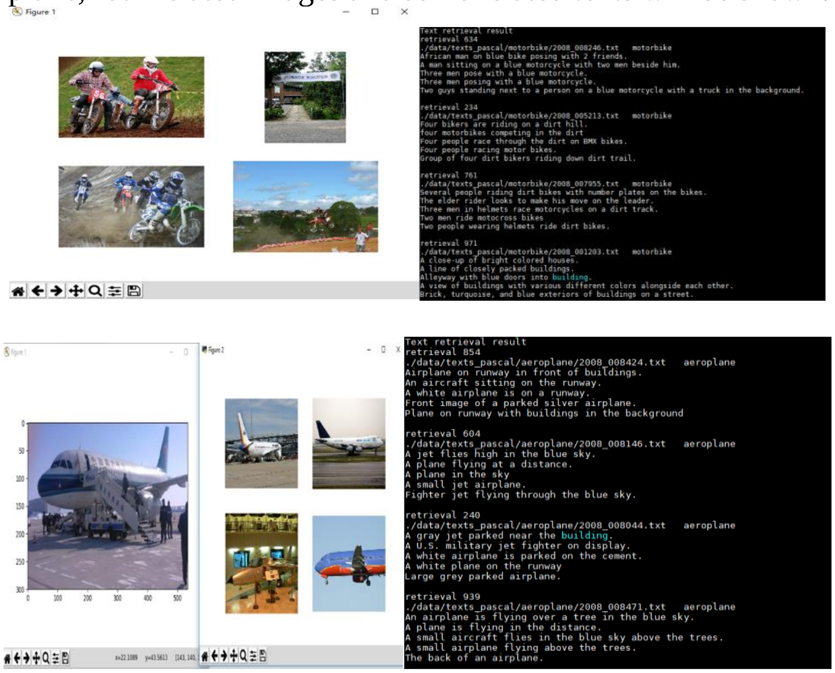
Loss name	Retrieval the texts based on images(MAP)	Retrieval the images based on texts(MAP)	Average MAP
Adversarial l0ss	0.53	0.46	0.495
MMD_loss	0.51	0.47	0.490
Deep_CORAL_loss	0.51	0.44	0.475
Correlation_loss	0.47	0.40	0.435

Table 3: MAP performace using different loss functions

Loss	Retrieval the texts based on images(MAP)	Retrieval the images based on texts(MAP)	Average MAP
Only loss1	0.35	0.40	0.375
Loss1+loss2	0.44	0.42	0430
Add adversarial loss	0.53	0.45	0.495

Table 4: MAP performace after introducing adversarial loss functions

Finally, we made some simple tests, from the figure 2.7 , we can see that when we type "Four bikes are riding on a dirt hill", four related images will be shown to us, on the other hand, when we input an airplane, four related images and some related texts will be shown.



```
Text retrieval result
retrieval 634
./data/texts_pascal/motorbike/2008_008246.txt   motorbike
African man on blue bike posing with 2 friends.
A man sitting on a blue motorcycle with two men beside him.
Three men pose with a blue motorcycle.
Three men posing with a blue motorcycle.
Two guys standing next to a person on a blue motorcycle with a truck in the background.

retrieval 234
./data/texts_pascal/motorbike/2008_005213.txt   motorbike
Four bikers are riding on a dirt hill.
four motorbikes competing in the dirt
Four people race through the dirt on BMX bikes.
Four people racing motor bikes.
Group of four dirt bikers riding down dirt trail.

retrieval 761
./data/texts_pascal/motorbike/2008_007955.txt   motorbike
Several people riding dirt bikes with number plates on the bikes.
The elder rider looks to make his move on the leader.
Three men in helmets race motorcycles on a dirt track.
Two men ride motocross bikes
Two people wearing helmets ride dirt bikes.

retrieval 971
./data/texts_pascal/motorbike/2008_001203.txt   motorbike
A close-up of bright colored houses.
A line of closely packed buildings.
Alleyway with blue doors into building.
A view of buildings with various different colors alongside each other.
Brick, turquoise, and blue exteriors of buildings on a street.
```

```
Text retrieval result
retrieval 854
./data/texts_pascal/aeroplane/2008_008424.txt   aeroplane
Airplane on runway in front of buildings.
An aircraft sitting on the runway.
A white airplane is on a runway.
Front image of a parked silver airplane.
Plane on runway with buildings in the background

retrieval 604
./data/texts_pascal/aeroplane/2008_008146.txt   aeroplane
A jet flies high in the blue sky.
A plane flying at a distance.
A plane in the sky
A small jet airplane.
Fighter jet flying through the blue sky.

retrieval 240
./data/texts_pascal/aeroplane/2008_008044.txt   aeroplane
A gray jet parked near the building.
A U.S. military jet fighter on display.
A white airplane is parked on the cement.
A white plane on the runway
Large grey parked airplane.

retrieval 939
./data/texts_pascal/aeroplane/2008_008471.txt   aeroplane
An airplane is flying over a tree in the blue sky.
A plane is flying in the distance.
A small aircraft flies in the blue sky above the trees.
A small airplane flying above the trees.
The back of an airplane.
```

Figure 2.7 Test results

4.2 Conclusions

In this project, we implemented an adversarial learning method on the cross-retrieval modal, we used two different dataset to verify our results. We compared the influence by different loss functions and different text features, finally we made some simple tests and had a good performance. An interesting issue is to further adjust this method on more different modals, such as video, 3D models or audio, we can use this method to retrieve these modal each other, this will be an interesting problem for people to research.

REFERENCES

[1] L. Wang, Y. Li, and S. Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*. 5005-5013.

[2] T. Yao, T. Mei, and C.-W. Ngo. 2015. Learning query and image similarities with ranking canonical correlation analysis. In *ICCV*. 28-36.