DOI: 10.54254/2755-2721/53/20241282

Multimodal bird information retrieval system

Hong Xu

Digital Media Technology, School of Information Science and Technology, Beijing Forestry University, 100091, Beijing

xh15860746281@163.com

Abstract. Multimodal bird information retrieval system can help people popularize bird knowledge and help bird conservation. In this paper, we use the self-built bird dataset, the ViT-B/32 model in CLIP model as the training model, python as the development language, and PyQT5 to complete the interface development. The system mainly realizes the uploading and displaying of bird pictures, the multimodal retrieval function of bird information, and the introduction of related bird information. The results of the trial run show that the system can accomplish the multimodal retrieval of bird information, retrieve the species of birds and other related information through the pictures uploaded by the user, or retrieve the most similar bird information through the text content described by the user.

Keywords: multimodal recognition, bird retrieval, CLIP model, graphic retrieval.

1. Introduction

In recent years, China has strengthened the ecological and environmental protection, but most people still lack the awareness of environmental protection. Especially for some endangered animals lack of specific understanding, unable to distinguish which animals for national protected animals, leading to some illegal hunting. At the present, our country existing 1445 kinds of birds, including 394 species of wild birds, for the endangered birds countries also use issued law for the protection of the birds, but because in the daily life people for birds knowledge contact with little. In the era when the country attaches great importance to the protection of ecological environment and the harmonious coexistence between man and nature, the protection of natural environment has been paid more and more attention by people, but most people have a little understanding of the natural environment and natural animals, leading to the frequent occurrence of many cases of rare animals, especially birds.In recent years, there have been many cases of illegal hunting and cherishing birds, and the fundamental reason is that the public does not know much about birds. Based on the CLIP model developed by OpenAI company, this paper designs the [1] multimodal bird retrieval system for bird knowledge popularization, which realizes the specific functions of map search and text search. Users only need to upload pictures of birds or describe the appearance characteristics of birds, and can match the closest birds and provide relevant information to determine whether the bird belongs to the national protected animal, so as to popularize the bird knowledge to the public while protecting and cherishing birds.

2. Image Retrieval

Text-based image retrieval mainly includes catalog retrieval and keyword retrieval. In this paper, we adopt the keyword retrieval method, and label more than 2,000 pictures of birds with corresponding names and put them into folders with corresponding labels, so that we can get the picture paths and bird category names through the labeling of the corresponding pictures in the retrieval process, which is convenient for the retrieval process. For the acquisition of text information, this paper will use PyQT5 to design the front-end interface, so that the user can input text information from the front-end text box, and then let the system read the text information in the text box to carry out the subsequent text-based image retrieval work.

Content-based image retrieval enables users to search images more directly, reflecting their search intent. In this paper, the algorithmic process of content-based bird image retrieval is divided into three steps: in the first step, the bird images uploaded by users are feature extracted. In the second step, the obtained image feature information is encoded, and a lookup table is produced based on the image encoding. For the target image, reduced sampling can be used for images with larger resolution so as to reduce the number of operations, and then the image is processed for feature extraction and coding. In the third step, the similarity matching calculation is performed, using the encoded value of the target image, and then the image database in the image search engine is used for local or global similarity calculation, and then the threshold is set according to the desired robustness, and the images with higher similarity are saved, and finally the closest images are filtered.

3. CLIP model

CLIP model (Contrastive Language-image Pre-training) is OpenAI company in early 2021 open source a large-scale graphic pre-training model based on comparative learning for matching images and text, through the image and the corresponding text of the comparative training, to achieve the purpose of the relationship between the two matched.[2] CLIP model will be the text as a label of the image for training, when the prediction task, the model only needs to obtain the text information corresponding to the image description, you can carry out zero-sample inference migration, zero-sample analysis is the use of the training set of data to train the model, so that the model can be able to classify the object of the test set, but there is no intersection between the categories of the training set and the categories of the test set; during the period of time you need to rely on the description of the categories to build the training set, to build the training set of objects, but the training set and test set of categories, the training set and test set of categories. The CLIP model architecture is shown in Figure 1.

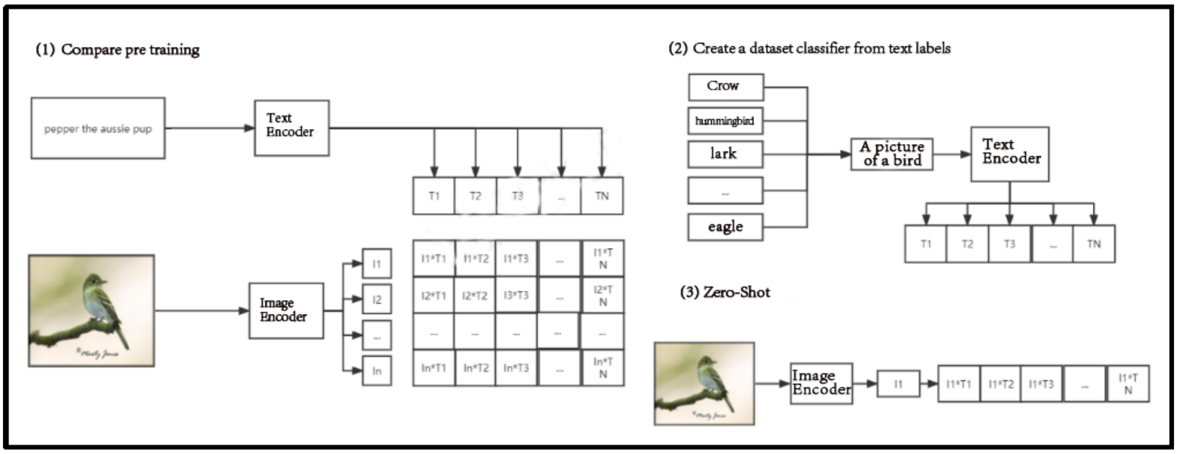


Figure 1. CLIP Model Basic Architecture

Although CLIP has some limitations today, it is also the most popular model for multimodal research, and there are many models developed based on CLIP, such as Chinese-CLIP developed by Ali Dharmo Academy, which can achieve optimal performance in Chinese cross-modal retrieval, in which Chinese

CLIP achieves optimal performance on the Chinese native e-commerce image retrieval dataset MUGE at multiple scales, such as the Chinese-CLIP, the Chinese CLIP, and the Chinese-CLIP. On the Chinese native e-commerce image retrieval dataset MUGE, Chinese-CLIP achieves the best performance in this scale. On the English-native Flickr30K-CN dataset, Chinese CLIP significantly outperforms domestic baseline models such as Wukong, Taiyi, and R2D2, no matter under zero-sample or fine-tuning settings. This is largely attributed to Chinese-CLIP's larger Chinese pre-trained graphic corpus, as well as the fact that Chinese-CLIP differs from some of the existing graphic representation models in China by freezing the image side throughout the whole process in order to minimize the training cost, and instead adopts a two-phase training strategy in order to better adapt to the Chinese domain.

4. Design of Bird Retrieval System Based on CLIP Modeling

4.1. Data set creation and image preprocessing

The dataset used in this paper is the bird pictures collected on the network, the picture format contains PNG and JPG format, there are 2138 bird pictures in total, including 36 kinds of birds, each kind of picture is about 40 pictures, these bird pictures are labeled and classified accordingly, the pictures are named by the kind of bird and classified into different folders, each folder is named according to the label of the bird, so as to facilitate the later retrieval process can index the location of the pictures according to the bird labels. Each folder is named according to the bird's label, to facilitate the later retrieval process can be indexed according to the bird's label the location of the pictures, of which the test set of 1075 pictures, the training set of 1063 pictures.

Pre-processing needs to be all the images are normalized to 128 * 128 uniform size, first read the size of the original image, and then the target image size of 128 on the original size of the width and height of the maximum value of the division to obtain a proportionality coefficient, and then traverse all the images, all the images of the length and width of the coefficient can be multiplied by the 128 * 128 size of the image, and finally all the images for the RGB conversion, so that all the images into three-channel image, three-channel data stored in the array, and then set its type to uint8 type (0-255 between the number), then divided by 255 will be between 0 ~ 1, and finally by calling np.mean and np.std functions to obtain the mean and standard deviation of the picture, and then finally the normalization process.

4.2. CLIP Model Graph Search Function Implementation

There are 9 models in the CLIP model library, including RN50, RN101, ViT-B/32, ViT-B/16, Vit-L/14 and so on.[3] In this paper, we use the ViT-B/32 model to realize the implementation of the map search function. The first stage is the extraction of image representations, traversing all the images in the dataset and representing each image with a vector to generate a dictionary, which includes the representations of all the photos.[4] In the second stage, the dictionary is imported and the representation vectors of the test photos are obtained, and then the cosine similarity of the dictionary representations is computed, the cosine similarity formula is shown in equation (1).

$$\cos\theta = \frac{A \cdot B}{||A| \cdot |B||} \tag{1}$$

The specific operation of the image representation extraction stage is as follows, first define a dictionary to store the representation information of all the photos in the dataset, the key of the dictionary is the file path of each photo, and the value of the dictionary is the representation vector of the image obtained through the ViT-B/32 model, and then iterate through all the bird images in the file, feed all the images into the preprocess operation of the pre-process function of CLIP model and then expand its dimension, and then use the encode_image function of CLIP model to obtain the representation vector of the image, and finally reduce the dimension of the representation vector and return it, after obtaining all the images, we will reduce the dimension and return it. function of the CLIP model and then expand its dimension, then use the encode_image function of the CLIP model to obtain the representation vector of the image, and finally reduce the dimension of the representation vector and return it, and after obtaining the representation information of all the images, the dictionary with the representation information of the image is stored to disk.

The specific operation of the image retrieval stage is as follows, firstly, the dictionary obtained in the image representation extraction stage is imported, and the bird image to be retrieved is read in, and then the image representation information is obtained and put into the dictionary, and then the next step is to calculate the cosine similarity of the image representation information in the dictionary. In this paper, we use the matrix for the batch operation to simplify the calculation. The first step is to store all the picture representation vectors as an array, then multiply the matrix with the transpose of the matrix, the number of the Nth row of the result represents the inner product of the Nth picture and all the vectors of all the other pictures, and this result serves as the numerator of the cosine similarity computation, and similarly the denominator of the cosine similarity is also used as the matrix for the batching operation, the number of the Nth row of the result represents the product of the number of the Nth picture's paradigm with the number of the paradigms of all the other pictures, and this two steps lead to the cosine similarity. two steps to derive the cosine similarity. Then the cosine similarity size is sorted, the larger the cosine similarity means the more similar the images are, and then the image with the largest cosine similarity is output through the image path in the dictionary to complete the retrieval.

4.3. CLIP Model Text Search Function Implementation

The CLIP model uses comparative learning to predict whether the image and text are paired or not. The implementation can be divided into three steps in general, firstly, the text and image are encoded with features respectively, secondly, the features of the text and the image are projected from their respective unimodal feature spaces to a multimodal feature space, and thirdly, the distance between the feature vectors of the original pair of images and texts is closer and closer while the distance between the feature vectors of the original unpaired images and texts is farther and farther in the multimodal feature space. feature vectors of originally paired image texts are closer to each other, while the feature vectors of originally unpaired image texts are farther away from each other. The unimodal features of images and texts can be projected into a multimodal feature space through the projection layer, so that the semantics of images and texts can be mapped in the same high-dimensional space, in which the similarity between the semantics of the text and the image can be judged to determine whether there is a correspondence between the graphic and the text.[5].

In this paper, we first convert the image dataset into a csv file, which contains some basic information of 2138 images, including the number of the image, the path name of the image, and the label of the image. First create a three-dimensional data used to save the csv text image number, the path name of the image, and the label of the image, and then read the folder of each bird, get the file path information of each bird, and convert the bird name in the file name to the label of the image, and then converted into a numpy array, and finally generate the csv file.

After completing the preparation of the dataset, the features of the images are encoded into vectors, in this paper, the Towhee library is utilized for the encoding of the image features, and the vectors of the images are generated by inference of the CLIP model called by Towhee. First read the data in the csv file, read the image file through the file path in the data, store the data into an array, then encode the semantic features of each image in the array into vectors, then normalize all the vector data, then construct an index for the vectors of the image and associate each image path with the vector. The next step is the vectorization of the query text, the process is similar to the process of image semantic vectorization, first encode the text into vectors, and then normalize the process, and finally by using the vector corresponding to the text to the vector index of the image to query, and by comparing the computed similarity value and finding the three photos that are closest to the semantics of the altered text for output. For example, by inputting the text yellow body and black head of the bird we can get the result as in Figure 2.

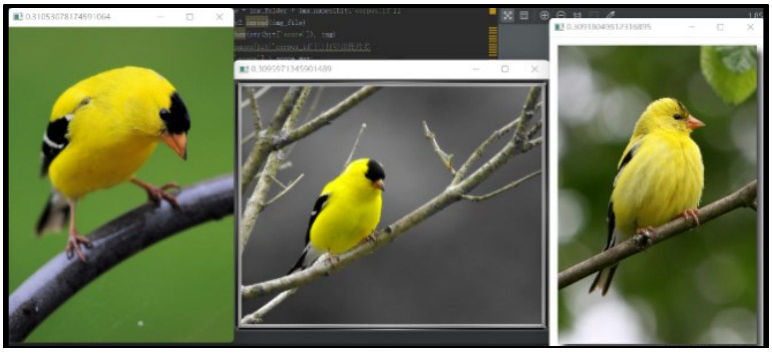
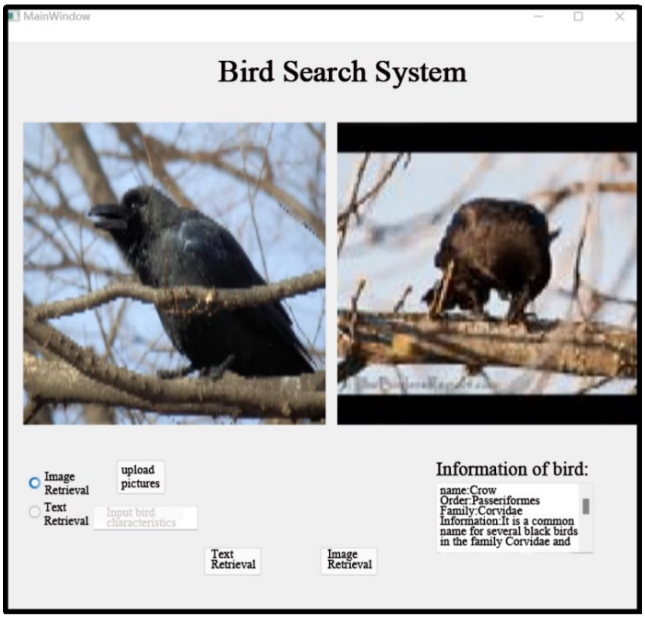


Figure 2. Displayed through text search and image results

4.4. Implementation of Content-based Image Retrieval Interface Functions

The content-based bird image retrieval interface function is divided into three steps. The first step is that the user uploads the bird images that need to be retrieved locally, and the second step is the image retrieval. After the user clicks on the image retrieval button, the system will call the back-end image search function through the button click event, which will carry out a series of operations on the images uploaded by the user such as feature extraction, image retrieval, and image similarity computation and output the image with the highest similarity in the front-end interface for display. After the user clicks the image search button, the system will call the back-end image search function through the button click event to perform a series of operations such as feature extraction, image retrieval and image similarity calculation, and then output the highest similarity image in the front-end interface for display. The back end will match the tags of the resultant images with the tags of the bird information prepared by the system, and then output the corresponding name of the bird, the bird order and the family of birds, the bird's habits and whether it is included in the list of protected animals in the Bird Information text box, as shown in Figure 3.

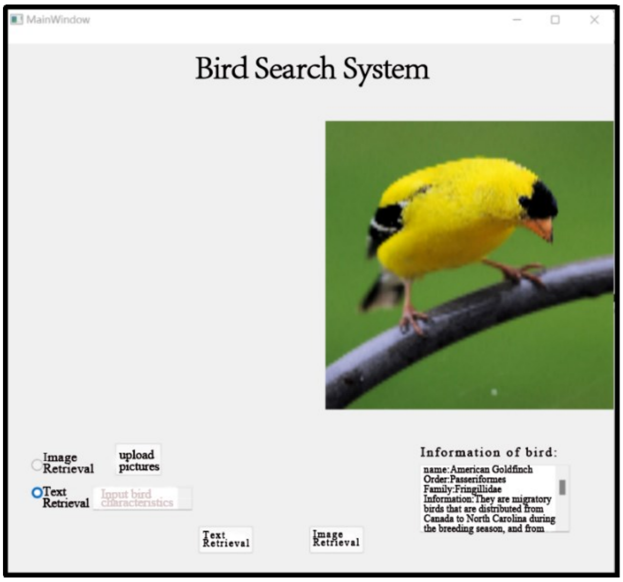


name:Crow
Order:Passeriformes
Family:Corvidae
Information:It is a common name for several black birds in the family Corvidae and

Figure 3. Display of search results

4.5. Implementation of Text-based Image Retrieval Interface Functions

The text-based bird image retrieval interface function is mainly divided into three steps. In the first step, the user selects text retrieval, and then inputs the characteristics of the bird that needs to be retrieved into the text box. In the second step, the user clicks on the text retrieval button, and the front-end will call the text-based image retrieval function according to the button click event, reads the user's input from the front-end text box to the back-end, and then carries out a series of operations such as text vectorization [6], similarity calculation [7], and so on, similarity calculation and a series of operations [7], the third step, the system through the calculation of the highest similarity of the image output in the front-end interface for display, in addition, the back-end will be through the results of the image label and the system written in the bird information label to match, in the bird information text box to output the name of the corresponding bird, the bird order and family, some of the bird's habits of the information, as well as whether or not to list the protection of animals, as shown in Figure 4.



name:American Goldfinch
Order:Passeriformes
Family:Fringillidae
Information:They are migratory birds that are distributed from Canada to North Carolina during the breeding season, and from

Figure 4. Display of search results

5. Conclusion

In this paper, a multimodal bird retrieval system is designed based on the CLIP model to retrieve bird information through content-based image retrieval and text-based image retrieval, so that the user only needs to upload the retrieved images in the retrieval process or input the characteristics of the bird to be retrieved into the text box of the system to retrieve the bird and give the specific information about the bird, which realizes the goal of helping the user to popularize bird-related knowledge. In this paper, we collected and created our own bird dataset on the Internet, with a total of more than 2,000 bird images, added labels to the images, classified the images, unified the image size and RGB conversion, and called the CLIP model to carry out feature extraction and cosine similarity computation pairing of the images to realize the image retrieval function.

The retrieval system in this paper realizes the two functions of searching map by map and searching map by text, but the whole system still has room for improvement. Firstly, each time the retrieval system carries out the retrieval function, the system has to read the dataset once, which makes the retrieval take more time, and we hope that there will be a new way of thinking to solve the problem later on; secondly, the bird dataset created in this paper covers a small number of birds, which is only 36 birds, and it cannot contain most of the birds. Secondly, the bird dataset created in this paper covers fewer bird species, only 36 bird species, which is not able to include most of the bird species, and there is still room for improvement in the expansion of the dataset.

References

[1] CLIP: Connecting Text and Image [EB/OL]. [2022-06-30]. https://openai.com/blog/clip/.

[2] ZHAO Jinwei, LIU Xiaopeng, LUO Wei et al. Research on a multimodal search tool for image resources in military domain based on CLIP model[J]. Chinese Journal of Medical Library Intelligence, 2022,31(08):14-20.

[3] Gao Yunlong. Research on deep local feature extraction model based on image retrieval [D]. Wuhan Light Industry University, 2021.10.27776/d.cnki.gwhgy.2021.000043.

[4] Wang, Xincheng. Research on content-based image retrieval system [D]. Donghua University, 2021.10.27012/d.cnki.gdhuu.2021.001043.

[5] Zhang Zhiqi, Yuan Xinpan, Zeng Zhigao. Research on the performance strategy of multimodal model for image-scene text fusion - an example of cross-modal retrieval[J]. Modern Information Technology, 2023,7(09):166-168+172.

[6] Sangers, J. , et al. "Semantic Web service discovery using natural language processing techniques." Expert Systems with Applications An International Journal 40.11 (2013): 4660-4671.

[7] Sun Qian, Wei Mingqiang, Yan Xuefeng et al. Searching for graphs with graphs: a saliency-attention based retrieval network for beauty products[J]. Journal of Computer-Aided Design and Graphics, 2023,35(03):383-391.