Image Retrieval of Traditional Chinese Painting Based on Convolutional Neural Network

Hongfan Mu[1], Yi Pang[2], and Yunsheng Jiang[1]

[1]School of Electrical Engineering and Computer Science, University of Ottawa
[2]Department of Systems and Computer Engineering, Carleton University

***Abstract*—Traditional Chinese painting, which dates back to almost five thousand years ago, can contribute to a great amount of amazing unique forms of arts. With the development of digital technologies, digital artwork is regarded as one of the most convenient and effective ways to help educate and improve the aesthetic tastes of the public. However, because of the long history of the development of traditional Chinese painting, each artist in a different time period and dynasty established multiple preferences, and skills on their artworks. This caused huge difficulties for the public to understand the multiple styles of the artworks and the connections among artists and periods in the history of this art. In our project, we developed an automatic analysis system to analyze the nuances of different artworks and output the genre and artist. We applied VGG-16 architecture to do the feature extraction of the input image and compare the extracted feature within the feature vector database. Finally, we present two results with the top similarity.**

***Keywords*—Traditional Chinese Painting, Digital Arts, Image Retrieval, Convolutional Neural Networks, Transfer Learning**

I. Introduction

Due to the continuous evolution of digital arts, the analysis of fine art paintings have been developed steadily in recent years, especially the western arts. Museums have digitalized their artworks to encourage the public interest and improve the public appreciation of fine arts in a more convenient way.

In Rodriguez's work, they established a new efficient method that improves the classification accuracy of fine-art paintings aimed at western artworks based on transfer learning and classification of sub-regions of the painting. They firstly resized the original image to the double of the input size required by the CNN model. For each image, it was divided into five sub-regions. The first four patches represented the image corners, and the last one represented the center of image. Secondly, a pre-trained CNN model was fine-tuned on the patch image data. Thirdly, the final stylistic class of the input image was determined by a combination of five patches. The final result indicates that this method achieved a computationally efficient way of improving the fine-art style classification accuracy without need to define and fully train new CNN model structures. [1].

As a unique and valuable form of art, the Traditional Chinese Painting should also gain more attention. With each brush stroke, the choice of pigment, rendering effect difference, Traditional Chinese paintings vary from period to period, and artist to artist. However, this is not just a random combination of these criteria. Each artist in a different time period has certain preferences. Our project here proposed an idea to tell the time period, artist and genre with a given Traditional Chinese Painting. This is the first step to have some understanding of the painting. We believe this would be a great help to improve the aesthetic consciousness of the public in the whole world. Hopefully, this would evoke their interest to learn more about Traditional Chinese Painting.

Our main method is content-based image retrieval, which directly analyzes the feature of the image, including colors, texture, shapes or other information derived from the image itself. After that, we output the artist and artworks with similar style of the input image . The structure of this paper is as follows: firstly, introducing the background knowledge we researched, secondly showing our method of achieving our goal, thirdly, presenting the result and discussion, finally, discussing about future work.

II. Background Knowledge

A. Image Retrieval

1) Text-Based Image Retrieval: One of the traditional methods is text-based image retrieval, started from 1970s, which utilizes some methods of adding metadata, such as keywords, captioning, or descriptions to the images, and searching is performed over those annotation words. Despite the fact that adding annotations would be time consuming, the result of the retrieval is too dependent on the annotation quality, completeness of search words. Thus, a method of retrieval relies on the image itself is evolved.

2) Content-Based Image Retrieval: Content-based image retrieval (CBIR) is one of the widely applied image retrieval methods, which has been investigated for classifying and search images according to similarities derived from extracted visual features, such as colors, shapes, and textures. CBIR aims to search for images through analyzing their visual contents, and thus image representation is the main point of CBIR. The image content can be understood as a simplified hierarchical model, as shown in the Figure 1.

The first layer is the original data layer, which is the original pixels of the image; the second layer is the physical feature

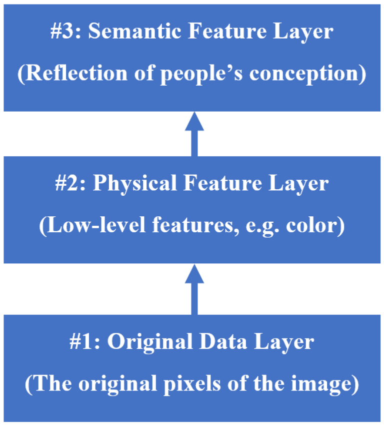

#3: Semantic Feature Layer
(Reflection of people's conception)
#2: Physical Feature Layer
(Low-level features, e.g. color)
#1: Original Data Layer
(The original pixels of the image)

Figure 1: Image Content Model

layer, which reflects the low-level physical characteristics of the image content, such as color, texture, shape, outline, etc. The third layer is the semantic feature layer, which is a reflection of people's conceptual level of image content, generally the description of the image content.

There are a variety of low-level feature descriptors have been proposed for image representation, ranging from global features representations (e.g. color, edge, texture, etc.) to local feature representations, such as the bag-of-words models using local feature descriptors (e.g. SIFT, SURF, etc.).

Firstly, CBIR skipped the extra step for adding metadata, especially when the database is relatively large. Secondly, metadata is always a partial description of the information in the image and depending on the purpose of the annotation. Furthermore, the CBIR has the potential to study relationships among paintings based on all the details from the image [2].

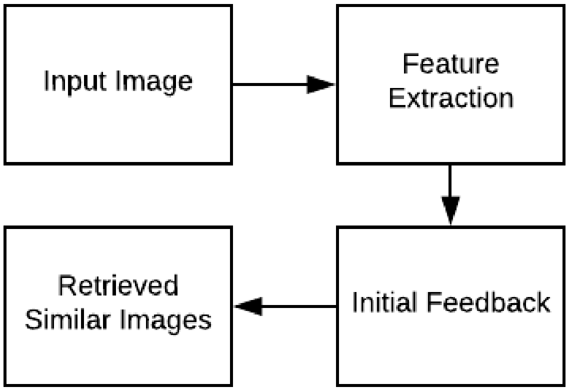


Figure 2: Content Based Image Retrieval

In most cases, Conventional CBIR approaches extract low-level features by using rigid distance functions to search the similarity of the contents, such as Euclidean distance or cosine similarity. The rigid distance function, however, may not be always optimal to the complex visual image retrieval tasks due to the huge challenge of the semantic gap between low-level visual features extracted by machines and the perception of human eyes.

3) Distance Metric Learning: Distance Metric learning (DML) for image retrieval has been studied in both conventional approach and machine learning communities. In terms of training data formats, There are two types of data that DML concerns: 1) pair-wise constraints where must-link constraints and cannot-link constraints are given. 2) triplet constraints that contain a similar pair and a dissimilar pair [3].

The core idea of distance metric learning is to learn an optimal metric which can minimize the distance between similar images, at the same time, maximize the distance between images that are dissimilar. Thus, DML is also closely related to another technique named similarity learning.

4) Deep Learning: Deep learning is the main branch of machine learning techniques, where many layers of information processing stages in hierarchical architectures play an important role in the pattern classification and feature learning. It lies in the crossing of several research areas, including neural networks, graphical modeling, optimization, pattern recognition, and digital signal processing, etc.

Deep learning has a long development history, and its basic concept starts with artificial neural networks (ANN) research. ANN is a simulation and approximation of biological neural networks. It is an adaptive nonlinear dynamic network composed of a huge number of neurons connected with each other. In 1943, psychologist McCulloch and mathematical logician Pitts proposed the first mathematical model of neurons, the M-P model [4]. The M-P model is a milestone and provides the basis for subsequent research work. By the end of the 1950s and the beginning of the 60s, Rosenblatt added learning functions based on the M-P model, and proposed a single-layer perceptron model, which put the research of neural networks into practice for the first time [5]. However, the single layer perceptron network model cannot handle linearity inseparable problems. Until 1986, Rumelhart and Hinton proposed a multi-layer feedforward network, namely, Back Propagation Network, which solved problems that single-layer perceptron could not solve [6].

In 2006, Hinton *et al.* [7] published a paper on Science, whose main points are: 1) multi-hidden artificial neural networks have excellent learning ability; 2) layer-wise pretraining can effectively overcome the difficulty of deep neural network training, which has set off another wave of studying deep learning and artificial neural networks. At present, the commonly used deep learning models include Deep Belief Network (DBN), Stacked Deoising Autoencoders (SDA), Convolutional Neural Network (CNN) and so forth.

From 1980s to 1990s, some researchers published relevant research works about CNN, and achieved good recognition results in several specific areas, especially handwritten digit recognition [8]. However, the CNN at that time was only suitable for the recognition of small-size pictures, and the result was not good for large-scale data. Until 2012, Krizhevsky *et al*. [9] used the extended depth CNN to achieve the best classification result in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), making CNN more and more valued by researchers.

B. Convolutional Neural Network

The basic structure of the convolutional neural network consists of an input layer, convolution layers (one or more layers), pooling layers (one or more layers), fully connected layers (one or more layers), and an output layer. In most cases, the convolutional layers and pooling layers are alternately arranged. Since each neuron of the output feature maps in convolutional layers is locally connected to its input, and the weighted sum is added to the local input by the corresponding connection weight and the offset value, the input value of the neuron is obtained, which is equivalent to the convolution process.

1) Convolutional Layers: A convolutional layer consists of feature maps, each of which is composed of a plurality of neurons. Every neuron is connected to a local region of the previous feature map by a convolution kernel. Essentially, the convolution kernel is a weight matrix. Convolutional neural network's convolutional layer extracts different features of the input through convolution processes. The first layer of convolutional layers extracts low-level features such as edges, contours, corners, and higher-level convolutional layers can extract more advanced features.

The convolutional layer neurons are organized into various feature maps, each of which is connected to a local area of the previous layer of the feature maps by a set of weights. It is then passed to a nonlinear activation function to obtain the output value of each neuron in the convolutional layer. In the conventional convolutional neural network, the activation function generally uses a saturating nonlinearity (i.e. sigmoid, tanh). Compared with the saturated nonlinear function, the non-saturating nonlinearity can solve the problem of gradient explosion and disappearance, and it can also accelerate the convergence speed [10].

In the current convolutional neural network structures, a non-saturating nonlinearity is commonly used as an activation function of convolutional layers such as a ReLU function.

In convolutional neural network structures, the deeper the depth and the greater the number of feature maps, the more complicate the network can represent and the stronger the learning ability of the network. However, the overfitting is also easier to occur. Therefore, in practical applications, the network depth, the number of feature maps, the size and step of the convolutional kernel should be appropriately chosen so that the training can obtain a good model.

2) Pooling Layers: The pooling layer (also known as the downsampling layer) immediately follows the convolutional layer and is also composed of a plurality of feature maps, each of which corresponds to a feature map of the previous layer. The number of feature maps unchanged during pooling process. The pooling layer is designed to achieve spatially invariant features by reducing the resolution of the feature maps. The commonly used pooling method includes 1) max pooling, which is taking the largest value in the pooling window; 2) mean pooling, which is averaging all the values in the pooling window; 3) stochastic pooling [11].

Boureau YL *et al*. [11] gave a detailed theoretical analysis of the maximum pooling and mean pooling. The following predictions are obtained through analysis: 1) Maximum pooling is particularly suitable for separating very sparse features. 2) It may not be optimal to use all sampling points in the local area to perform pooling operations. For example, the mean pooling takes advantage of all sampling points in the local acceptance region. It was found through experiments that: When the classification layer uses linear classifiers such as linear SVM, the maximum pooling method can get a better classification performance than the meaning pooling method.

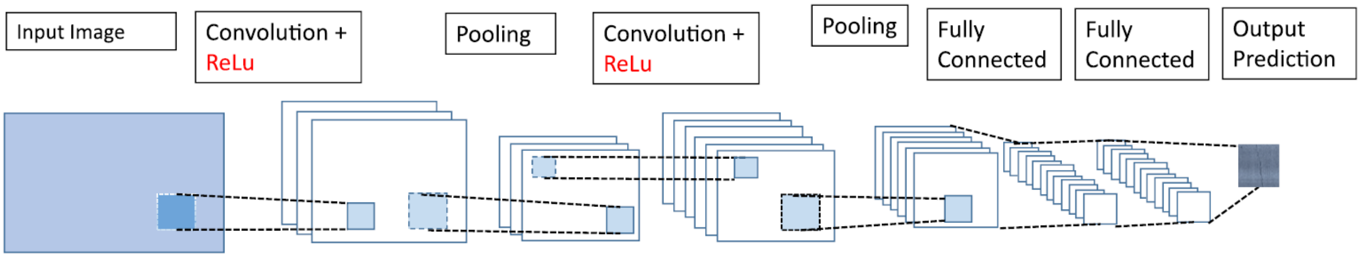


Figure 3: Flow Chart of Convolutional Neural Network

The random pooling method assigns probability values to the sample points of the local acceptance region according to their value, and then randomly selects according to the magnitude of the probability value. The random pooling has the greatest pooling advantage, and it avoids overfitting due to randomness. In the usual pooling method, the same feature map of the pooling layer does not overlap with the local acceptance region of the previous layer, but a method of overlapping can also be used, which means that there is an overlapping area between adjacent pooling windows.

3) Fully Connected Layers: In convolutional neural network structure, after a plurality of convolution layers and sampling layers, one or more fully connected layers are connected. Like the MLP, each neuron in the fully connected layer is fully connected to all neurons in its previous layer. The fully connected layer can integrate local information with discriminative classification in the convolutional layer or sampling layer. In order to improve convolutional neural network performance, the activation function of each neuron in the fully connected layer generally uses the ReLU function.

The output layer of the last fully connected layer is named sofmax layer, which can classify the output using softmax regression to make sure that each component is between 0 to 1.

C. Transfer Learning and Feature Extraction

Usually, a deep learning model has trained with millions of labeled images, they have achieved great performance on many image-related tasks. A key challenge in applying CNN Model to Traditional Chinese Painting retrieval is that the available labeled training samples are very limited. To overcome this difficulty and develop a universal representation for Traditional Chinese Painting retrieval, we proposed to employ transfer learning to transfer the trained knowledge from labeled image data.

The core idea of transfer learning is to improve the performance of our task by applying data acquired from different but correlated training samples, which has already yielded superior performance on related image recognition tasks.

In this work, we have explored whether the transfer learning of convolutional neural network model can be generalized to traditional Chinese paintings. The deep learning model we used was trained on the ImageNet data containing millions of labeled images with a thousand of classifications and used directly as a feature extractor to compute representations for traditional Chinese paintings. Specifically, we applied the pre-trained VGG model that has already trained on the ImageNet data to perform several computer vision tasks, such as classification, localization and detection [12]. There are two pre-trained models which are have 16 and 19 weight layers respectively. And we applied the 16 weight layers model in our project.

D. VGG-16

The Convolutional Neural Network architecture applied in this report is 16-layer VGGNet. It was a submission of ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2014 by Visual Geometry Group of University of Oxford, which has won the first place in localisation and the second place in classification[12].

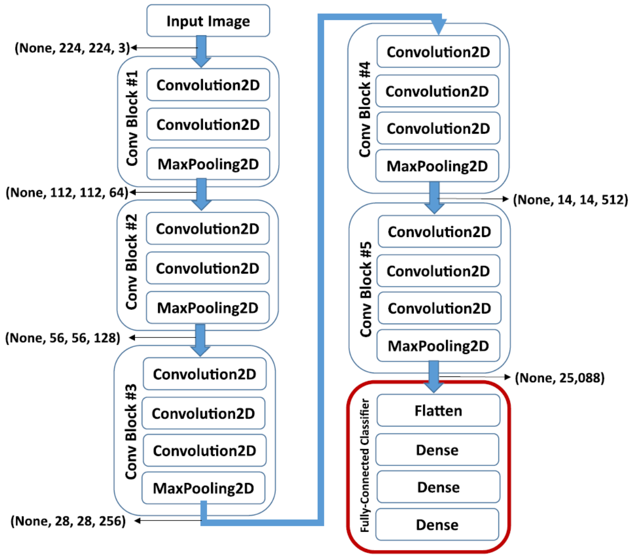


Figure 4: Architecture of VGG-16 Model

The input of VGG-16 is 224x224 RGB images. It uses consecutive two or three convolutional processes before a max-pooling layer. VGG-16 reduces the size of kernels and increase the amount of them. It uses 3x3 convolutional kernels, two of which have an effective receptive field of a 5x5 kernel. And three of them have the same effect of a 7x7 convolutional kernel. Compare to larger kernels, using small 3X3 kernels has several advantages. It makes the network has more activation functions, digs richer features and strengths discriminating ability. Because every convolution is accompanied by an activation function. The use of more convolution kernels makes the decision function more discriminating. In addition, in terms of the convolution itself, 3x3 is better than 7x7 to capture the change of features, which is deeper by two layers and two nonlinearity ReLU functions are added. Feature diversity and parameter quantity increase make the network capacity larger, and the distinguishing ability for different categories is strengthened.

The parameters of the convolutional layers are also reduced in VGG-16. For example, if the number of input channels and the number of output channels are both C, then 3 convolutional layer parameters required in terms of 3x3 kernels are: $3 \times (C \times 3 \times 3 \times C) = 27C^2$, and the number of parameters required for one convolutional layer of a 7x7 kernel is: $C \times 7 \times 7 \times C = 49C^2$, which is more than the former by 81%.

The max pooling kernel size used by VGGNet is 2x2 and stride is 2. The small kernel brings more detailed information capture. Compare to other pooling strategies, max pooling makes it easier to capture changes of features, gradients, and local information differences in images. And it has better description ability of the semantic details of edges, textures, etc., especially in network visualization.

VGG uses the 1x1 convolution kernel in the last three fully connected stages. The most direct reason for choosing the 1x1 convolution kernel is to inherit the dimension of the previous layer. The 1x1 convolution can increase the nonlinearity of the decision function (softmax). The nonlinearity is determined by the activation function ReLU. The 1x1 convolution process itself is a linear map, which reflects the feature maps of the previous layer to the next layer.

It was also found that Local Response Normalization (LRN), which is a method used by AlexNet[9] to normalize pixel values across channels, has no performance improvement, so it does not appear in the latter four groups of networks of VGGNet.

E. Oversampling

The imbalanced datasets seems to be an inevitable problem for the classification and image retrieval. A dataset is called imbalanced if one of the two classes having more samples than other classes. For example, the first class has 1000 data, while the second class just has 10. Class imbalance would lead difficulty in learning for retrieval.

In order to solve this kind of problem, undersampling and oversampling are considered as the most frequently used method. Undersampling requires a good sampling model to represent the whole population, while the oversampling requires reliable data generation tool [13]. However, in the previous example we mentioned, undersampling would waste lots of existing data, so that oversampling is a better choice under this circumstance. Oversampling is based on the pattern of the exist class with fewer sample labels to generate more data of that class, which makes the data tend to balance.

Synthetic Minority Oversampling Technique (SMOTE) is a very classic method in oversampling. It can control the number of examples and distribution to achieve the purpose of balancing the dataset through synthetic new examples. In K.Usha Rani's paper, they experimented SMOTE technique with diverse five classifiers on various breast cancer datasets to overcome high dimensionality and class imbalance problems. SMOTE method is verified to eliminate the biasedness towards the majority class and classified the data [14].

This method obtains a new class of samples by sampling on a line of a class of similar neighbors selected by a small class of samples. SMOTE overcomes the over-fitting problem of traditional oversampling methods to a certain extent.This would increase the reliability of our final model and increase the accuracy of image retrieval.

F. Data Augmentation

Data augmentation can be considered as a process of creating new similar samples to the training set. It's one of regularization technology mainly used to prevent overfitting, especially when the dataset is small. [15] In practice, we usually increase the depth and breadth of the neural network, so that the learning ability of the neural network is enhanced, and it is easy to fit the distribution of the training data. In convolutional neural networks, it has been experimentally found that depth is more important than breadth. However, as the neural network deepens, the parameters that need to be learned increase, which makes it easier to cause overfitting. When the data set is small, too many parameters will fit all the characteristics of the data set but not the commonality of data.

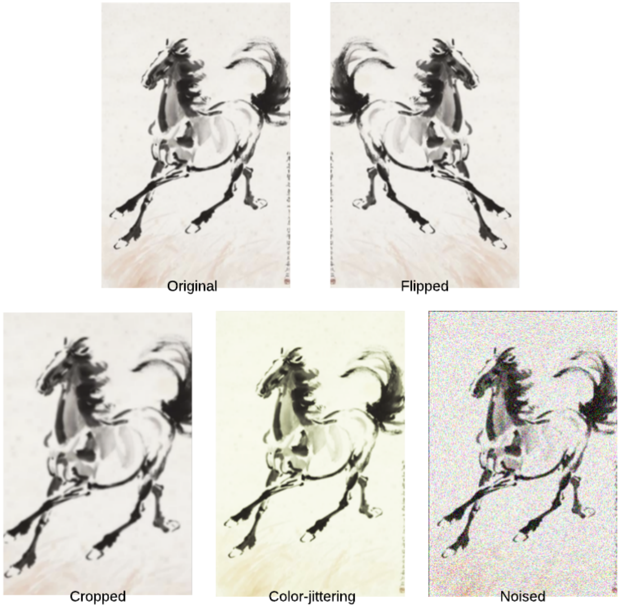


Figure 5: Original, Flipped, Cropped, Color-jittering, and Noised Image

Here are some main methods used in data augmentation. The first is unsupervised data augmentation, which means the augmentation methods are not related to data labels. The commonly applied image transformation methods are flipping, rotation, cropping, shifting, color-jittering, and noise. These are all within the category of unsupervised data augmentation. It's easy to use in the real world scenario. The second method is supervised data augmentation. GAN (Generative Adversarial Networks) and its improved methods can be categorized into supervised data augmentation. [15] These methods are all to increase the image diversity and keep the commonality. The Figure 5 shows some results of the unsupervised data augmentation.

G. *Django Framework*

Django is a Python-based web framework, which follows the model-view-template (MVT) architecture pattern. Model is responsible for managing the data of the application. It responds to the request from the view and it also responds to instructions from template to update. View is the presentation of data, triggered by templates. The template is a HTML file mixed with Django Template Language. As for the responding to the user input and perform interactions, Django itself takes care of that part. Django works well in agile web development. Its primary goal is to solve the problem caused by complex, database-driven websites.

The advantage of this framework is its complete functions and elements, including a large number of commonly used tools and frameworks, which is suitable for rapid development of websites. Django's advanced app design philosophy that it's pluggable, so that each function can be deleted directly when there's no need and has little effect on the whole system. Furthermore, it has powerful database access components. Django's model layer comes with a database ORM component, allowing developers to manipulate the database without having to learn the SQL language. This framework emphasizes reusability of components, less code, rapid development. We applied this framework to offer a comprehensive view of our retrieval result.

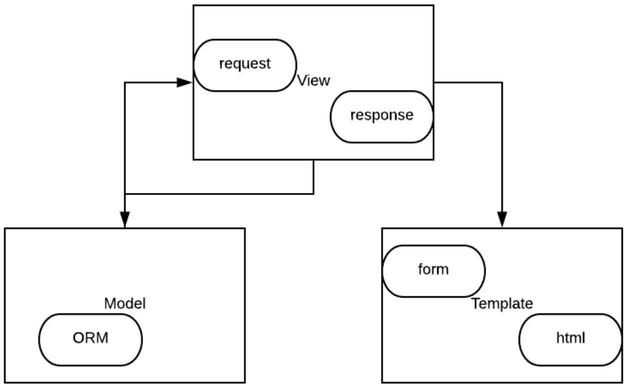
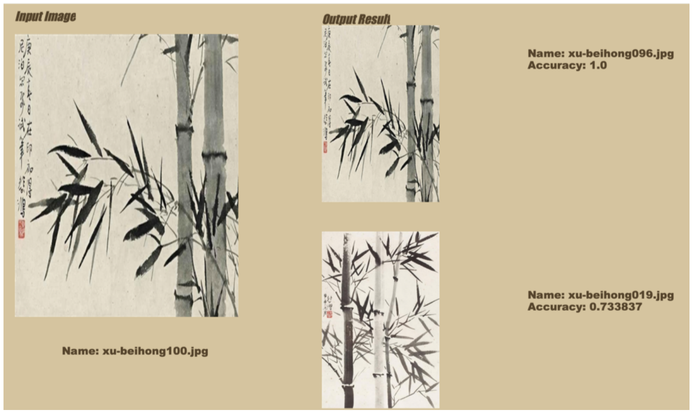


Figure 6: Model-View-Template Architecture

III. METHODS

The experiment aims to apply image retrieval of traditional Chinese artworks based on Convolutional Neural Network model. The application is in an effort to help people find more details about their favorite paintings, such as the artist of the artwork, the story behind it, the similar artworks which are from same artist and same genre. The experiment is based on Keras, which is the architecture of deep learning; Tensorflow, which is the backend of the deep learning environment. As initial experiments, combined with the papers we reached, we studied and compared the obtained Chinese artists' artwork - the art tastes and preferences of artist. The experiment has achieved the image retrieval of the traditional Chinese paintings based on pretrained VGG-16 model. We resized some of the artworks at spatial resolutions typically of 224*224 pixels. The outputs of the experiment are the paintings from the same artists, which contain similar contents, texture, brush strokes, or other characteristics of the uploaded image.



Name: xu-beihong096.jpg
Accuracy: 1.0

Name: xu-beihong019.jpg
Accuracy: 0.733837

Name: xu-beihong100.jpg

Figure 7: Output Sample with Django Framework

A. *Experimental Approach*

1) *Image Preprocessing:* The quality of each images effects the accuracy of the final output of image retrieval. All the images we obtained from online website were in multiple sizes and different qualities, which would present a huge difficulty in feature extraction modeling. The main purpose of image pre-processing is to eliminate irrelevant information among each images, restore useful real information, enhance the detectability of relevant information, and minimize data, thereby improving the reliability of feature extraction model. It is therefore essential to achieve image pre-processing before feature extraction modeling.

Since our dataset was from WikiArt.org (formerly known as WikiPaintings), which is an online, user-editable visual art encyclopedia can be used from modeling. A problem we had to focus on was that all the traditional Chinese paintings we obtained were labeled by volunteers. On one hand, it saved our time on labeling images, on the other hand, the images were labeled by the ones who did not receive any image labeling training, thus some of the images were not qualified enough to be used in model training. Here we just accepted part of the images from WikiArt.org and relabelled them after crawled from website.

The dataset we obtained exists an obvious problem, which might cause a huge difficulty in fellow steps, that is unbalanced images of each classes in the dataset. Seriously, it is a common problem that can happen if the dataset is from the real world.
We applied oversampling after relabeled the dataset we crawed from WikiArt.org. Since the dataset was quite small and the classes of it was not big as well, we accepted random oversampling to deal with imbalance classes problem. Random oversampling involves supplementing the training data with multiple copies of some of the minority classes. It randomly copies the images to ensure that each classes has the same number of data. Here, after applying oversampling to the dataset, we acquired a 500 images dataset from five different classes, in other words, each classes contained 100 images.

Even though applied oversampling, the number of the images in dataset was still not enough to reach a reliable output. Chances are that overfitting could occurs because of the limited number of the obainted dataset. Under this circumstance, in order to reducing overfitting on modeling, data augmentation should be considered during modeling process. Data augmentation increased the amount of training images by using the images that are from training dataset.

2) Feature Extraction: Three deep learning models had been tested in the experiment, an untrained three-layers model, an untrained six-layers model and a pre-trained VGG-16 model. The pre-trained weight of VGG-16 model was obtained from Imagenet, which is a large visual database designed for use in visual object recognition software research. Two dataset had been used in this experiment, dis-augmented dataset and augmented dataset.
Because of the best performance of pre-trained VGG-16 model with augmented dataset, we applied it as the feature extraction model for the next the design of image retrieval system.

3) Image Retrieval: How to define the metrics that match the similar feature vectors is the key common similarity measure in the image retrieval process: Euclidean distance method, Minkowski distance, Manhattan distance, weighted Euclidean distance and so on. In practical applications, the similarity measure of image retrieval mainly depends on two points: 1 the characteristics of data set; 2 the features types of the objected extracted image.
We applied cosine similarity in this experiment, which measures the cosine of the product space in two vectors to define the similarity between the artworks in our backend database and the image uploaded to the image retrieval system.

B. System Design

The system is designed in three modules:

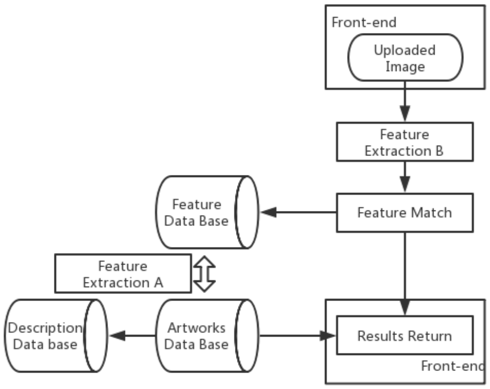


Figure 8: The Architecture of Image Retrieval System

- Input Module: Input module provides users with an inquiry interface. It also accepts the image uploaded by the user to the system. Through this module, the user submits the traditional Chinese painting image to the system (submitted from the front-end).

- Feature Extraction Module: Once feature extraction module is used, a series of image processing activities will be performed. Here we can see feature extraction module is used in two different process. The Feature Extraction A module will apply the trained convolutional neural network model to each images in the database, extracting the features of these images, and storing them in back-end Mysql database. The Feature Extraction B module will extract the features of the image which is uploaded to the front end by user. Both of the extracted features are used for subsequent image retrieval modules.

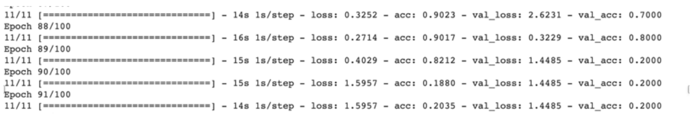
- Matching module: The task of this module is to compare the extracted features from the uploaded image with that of the images in the database according to the cosine similarity. The system retrieves top 3 similar images from the back-end database and returns the second and the third similar images to the front-end.

IV. RESULT AND DISCUSSION

A. Modeling

Due to the size of the dataset we obtained from WikiArt.org, the accuracy of the untrained three-layers model reached overfitting. Overfitting is a modeling error which occurs when the limited data points are too closely fit to each other. Overfitting refers to the problem in the process of model fitting. It happens just because a complex model has been applied to the limited data points. Since the training data contains sampling errors, the complex model takes the sampling error into consideration during training, and the sampling error also fits good in modeling.

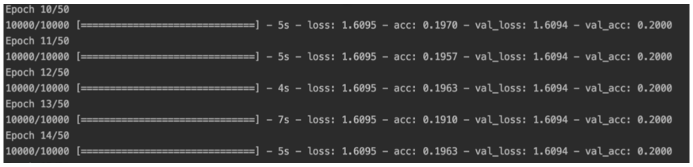
Overfitting occurred when the basic six-layers model applied to the non-augmented dataset. We can see from the figure 9, when model step reached epoch 90, compared with epoch 88, both training and validation loss of epoch 90 had been tripled, and the accuracy of them had been down from 0.27 and 0.32 to 0.18 and 0.20 respectively.



```
11/11 [==============================] - 14s 1s/step - loss: 0.3252 - acc: 0.9023 - val_loss: 2.6231 - val_acc: 0.7000
Epoch 88/100
11/11 [==============================] - 16s 1s/step - loss: 0.2714 - acc: 0.9017 - val_loss: 0.3229 - val_acc: 0.8000
Epoch 89/100
11/11 [==============================] - 15s 1s/step - loss: 0.4029 - acc: 0.8212 - val_loss: 1.4485 - val_acc: 0.2000
Epoch 90/100
11/11 [==============================] - 15s 1s/step - loss: 1.5957 - acc: 0.1880 - val_loss: 1.4485 - val_acc: 0.2000
Epoch 91/100
11/11 [==============================] - 14s 1s/step - loss: 1.5957 - acc: 0.2035 - val_loss: 1.4485 - val_acc: 0.2000
```

Figure 9: Loss and Accuracy of Untrained Six-layers Model with Non-augment Dataset

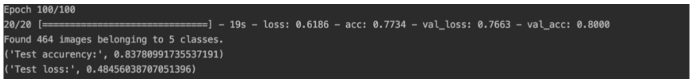
In order to reducing overfitting, we reduced the complexity of the model layers and applied the augmented dataset to the now model. We trained the untrained three-layers model with augmented dataset, however, the outcome of this model was not satisfied since the loss of it did not decrease any more. As what we can see from figure 10, the loss of both training and validation did not decrease any more, remaining at 1.61 for each. Also, the accuracy of training and validation were not satisfied as well, 0.2 for both.



```
Epoch 10/50
10000/10000 [==============================] - 5s - loss: 1.6095 - acc: 0.1970 - val_loss: 1.6094 - val_acc: 0.2000
Epoch 11/50
10000/10000 [==============================] - 5s - loss: 1.6095 - acc: 0.1957 - val_loss: 1.6094 - val_acc: 0.2000
Epoch 12/50
10000/10000 [==============================] - 4s - loss: 1.6095 - acc: 0.1963 - val_loss: 1.6094 - val_acc: 0.2000
Epoch 13/50
10000/10000 [==============================] - 7s - loss: 1.6095 - acc: 0.1910 - val_loss: 1.6094 - val_acc: 0.2000
Epoch 14/50
10000/10000 [==============================] - 5s - loss: 1.6095 - acc: 0.1963 - val_loss: 1.6094 - val_acc: 0.2000
```

Figure 10: Loss and Accuracy of Untrained Three-layers Model with Augment Dataset

The third model was a pre-trained VGG16 model with augmented dataset, which reached the best performance. It can be seen from figure 11, training accuracy and validation accuracy had increased to 0.77 and 0.8 respectively. The loss of training and validation had decreased to 0.62and 0.77, which are acceptable.
What's more, the test accuracy for the test dataset we prepared is 0.83, with the loss of 0.48.



```
Epoch 100/100
20/20 [==============================] - 19s - loss: 0.6186 - acc: 0.7734 - val_loss: 0.7663 - val_acc: 0.8000
Found 464 images belonging to 5 classes.
('Test accurency:', 0.83780991735537191)
('Test loss:', 0.48456038707051396)
```

Figure 11: Loss and Accuracy of Pre-trained VGG16 with Augmented Dataset

The picture of training accuracy and validation accuracy is listed:

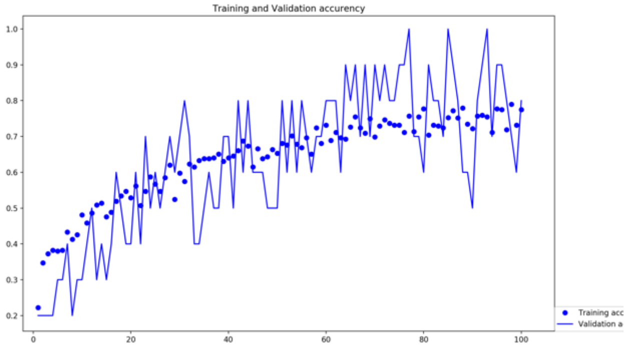


Figure 12: Accuracy of Pre-trained VGG16 with Augmented Dataset

From the figure 12, the trend of both training and validation accuracy increase, with training points and validation points fit quit well. Even though the trend of validation accuracy fluctuates, it is acceptable because we just focus on the whole trend.

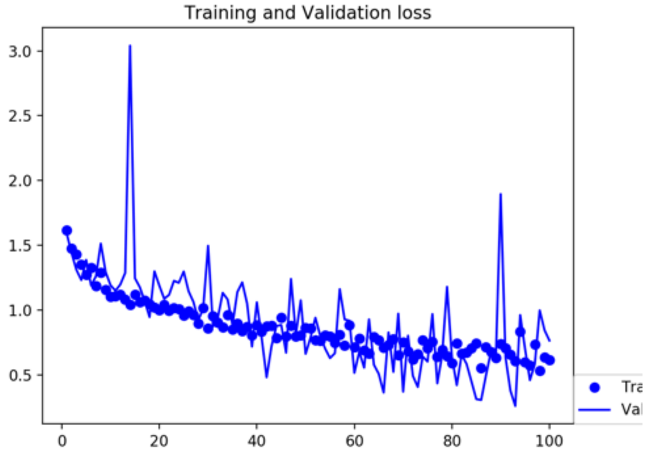
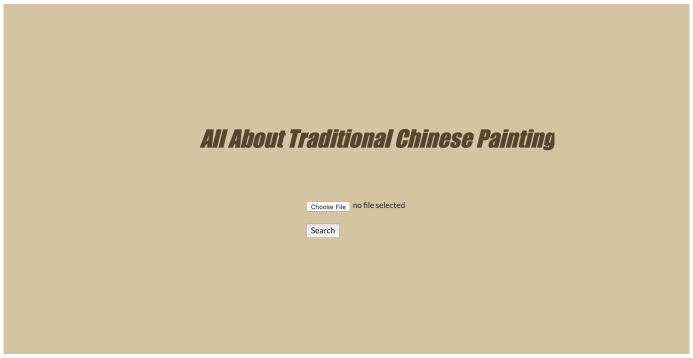


Figure 13: Loss of Pre-trained VGG16 with Augmented Dataset

From figure 13, we can see the trend of data points of training and validation loss decreases, with that of training points and validation points fit quit well as well. It is reasonable that some of points abruptly appears, since the whole tendency presents a softly decrease which has satisfied the expected result.

B. Image Retrieval System

1) Input Image: The front-end webpafe of image input:



All About Traditional Chinese Painting

Figure 14: The Webpage of Image Input

2) Image Retrieval: Since oversampling had been used in our dataset, it is common to see the same picture in a class. For example, as we can see from 15 xu-beihong96.jpg was the same as xu-beihong100.jpg. The expected result of our image retrieval system was to return the same picture of the uploaded picture except the uploaded picture itself, which was the second similar picture of the uploaded picture but the accuracy of which was in 1.0 as well. In order to simplify the display, the retrieval output of the uploaded picture itself had been manually dropped in front-end.

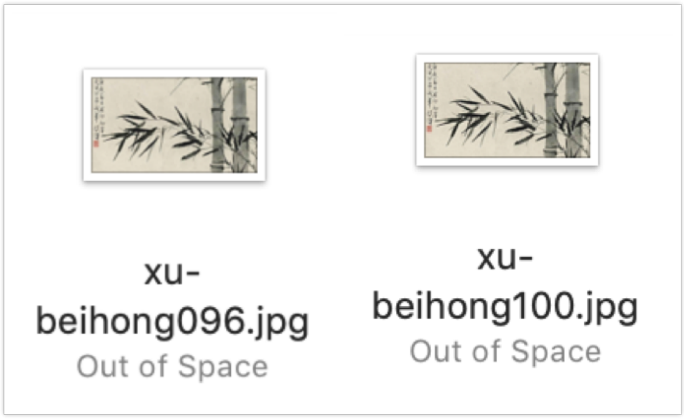
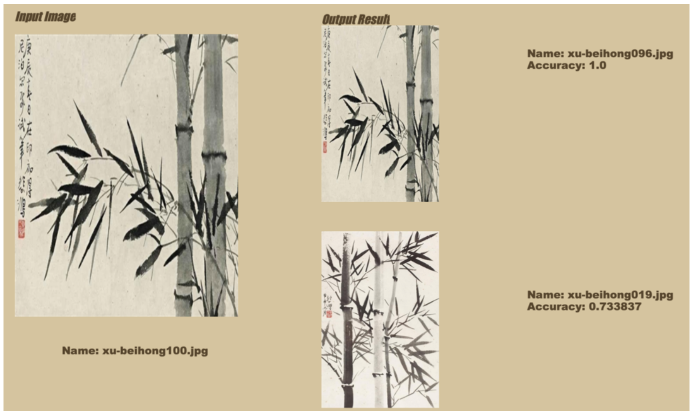


Figure 15: Oversampling Applied in Xu Beihong Class

We can see from the figure 16, xu-beihong100.jpg had been uploaded into image retrieval system. The output result of the system were xu-beihong96.jpg and xu-beihong019.jpg, which are the second and the third similar picture of the xu-beihong100.jpg with the accuracy of 1.0 and .73 respectively (except xu-beihong100.jpg itself). Obviously, both of the output have same painting object in the picture, which is Chinese bamboo. xu-beihong96.jpg and xu-beihong019.jpg have similar color (black and white), white physical space, and painting style as well.



Input Image

Output Result

Name: xu-beihong096.jpg
Accuracy: 1.0

Name: xu-beihong019.jpg
Accuracy: 0.733837

Name: xu-beihong100.jpg

Figure 16: Output Result of Image Retrieval in Xu Beihong Class

Oversampling was also used in Huang Yongyu class. huang-yongyu085.jpg is the same as huang-yongyu016.jpg. The output is:

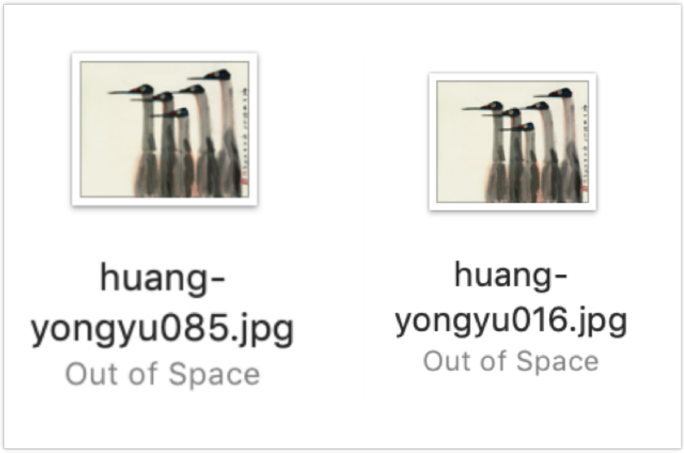


Figure 17: Oversampling Applied in Huang Yongyu Class

From figure 18, huang-yongyu016.jpg reaches the accuracy of 1.0 while that of huang-yongyu085,jpg is just 0.59. The same painting object - crane - can be seen in both of the output result. However, it can be seen that the two output results are not in the same painting style, one is in Chinese ink painting style and the other is a typical oil painting.

There was not any copy of xu-beihong089.jpg in the dataset of Xu Beihong class. It is clear to see from figure 19 that, after we uploaded xu-beihong089.jpg into the image retrieval system, the output results were xu-beihong097.jpg and xu-beihong027.jpg with accuracy in 0.85 and 0.79 respectively. Obviously, both of the pictures of output results are with the same painting object with the uploaded picture - a kind of magpie which is a popular bird in traditional Chinese painting. The similar background of the three paintings in front-end, plum tree, which is another popular object in traditonal Chinese painting. The main differences among xu-beihong089.jpg, xu-beihong097.jpg and xu-beihong029.jpg

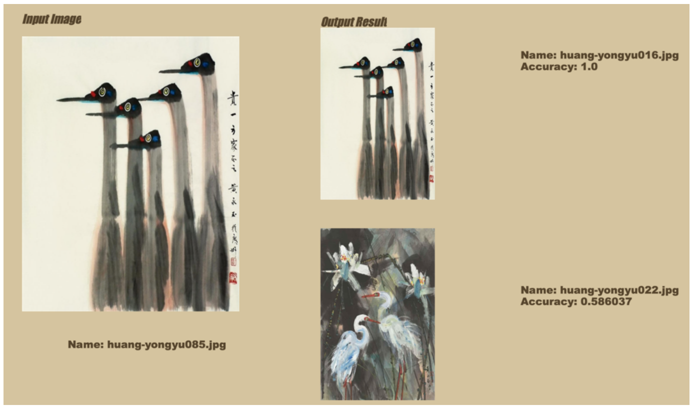

Input Image
Output Result
Name: huang-yongyu016.jpg
Accuracy: 1.0
Name: huang-yongyu022.jpg
Accuracy: 0.586037
Name: huang-yongyu085.jpg

Figure 18: Output Result of Image Retrieval in Huang Yongyu Class

are the number of magpies in the picture, the shape of plum trees, the motions of magpies.

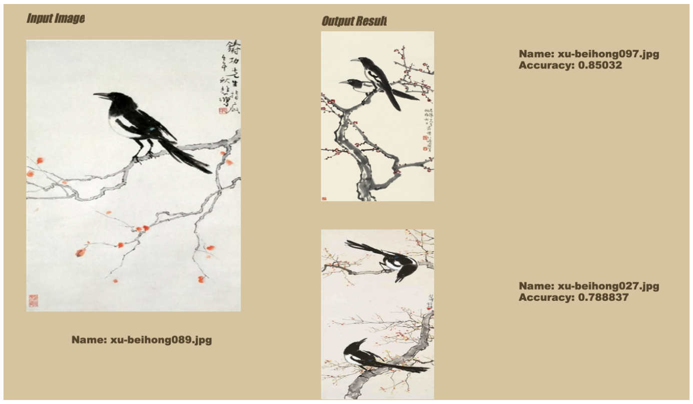


Figure 19: Output Result of Image Retrieval in Huang Yongyu Class without Copy of Original Picture

V. CONCLUSION AND FUTURE WORK

A. *Conclusion*

The present study investigated:

- A self-designed dataset from WikiArt.org which contains hundreds of images that filtered and relabeled by ourselves. More image reprocessing, such as oversampling and data augmentation have been applied before modeling as well.
- The application of a image retrieval technique based on traditional Chinese artworks. since it is a relatively rare field that not many scientists have contributed to it .
- Feature extraction model based on the pre-trained VGG-16 model based on transfer learning with our self-designed augment dataset. A relatively reliable feature extraction model with acceptable accuracy and loss.

B. *Future work*

Since all the images are directly obtained from WikiArt.org, multiple sizes and different qualities can raise the difficulty in model training. We apply image re-sizing and augment to avoid such difficulty. However, the image re-sizing process can introduce significant distortions, and loss of important details such as texture, brush strokes, or other characteristics which may be essential for automatic fine art analysis.

Inspired by [1], if each input image can be divided into five sub-images (patches) that have fixed locations within the image array, the accuracy of the model can be slightly improved.

Additionally, we hope our system can be applied to the real-world scenario more exactly, for example, in museums or art galleries, which requires that the dataset should be enlarged. We will seek to find more artwork related to Traditional Chinese Painting and put that into our database. We also wish to update our database as the new artworks appears, so that our system would not be out of date.

REFERENCES

[1] C. S. Rodriguez, M. Lech and E. Pirogova, “Classification of Style in Fine-Art Paintings Using Transfer Learning and Weighted Image Patches,” 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, Australia, 2018, pp. 1-7.

[2] Danging Zhang, B.Pham, Yuefeng Li, “Modelling traditional Chinese paintings for content-based image classification and retrieval”, 19 February 2004

[3] Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research. 2009;10(Feb):207-44.

[4] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 1943 Dec 1;5(4):115-33.

[5] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958 Nov;65(6):386.

[6] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Cognitive modeling. 1988 Oct;5(3):1.

[7] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006 Jul 28;313(5786):504-7.

[8] LeCun Y. Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature. 2015;521(7553):436-44.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. InAdvances in neural information processing systems 2012 (pp. 1097-1105).

[10] Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853. 2015 May 5.

[11] Boureau YL, Le Roux N, Bach F, Ponce J, LeCun Y. Ask the locals: multi-way local pooling for image recognition. InICCV’11-The 13th International Conference on Computer Vision 2011 Nov 6.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[13] T Antaresti, M I Fanany, A M Arymurthy, “Maintaining Imbalance Highly Dependent Medical Data Using Dirichlet Process Data Generation”, 1 December 2011

[14] K.Usha Rani, G. Naga Ramadevi, D. Lavanya, "Performance of Synthetic Minority Oversampling Technique on Imbalanced Breast Cancer Data", 31 Oct 2016.

[15] Jia Shijie, Wang Ping, Jia Peiyi, Hu Siping, “Research on Data Augmentation for Image Classification Based on Convolution Neural Networks”, 01 January 2018