Department Of Computer Science & Information Systems

Multimedia Computing Assignment

Content-Based Image Retrieval using Convolutional Neural Networks

Under the supervision of

Prof. Sundaresan Raman

BITS F447 - Multimedia Computing

Submitted By

Ameesha Mittal (2014A7PS107P)

M Sharat Chandra (2014A7PS108P)

Birla Institute of Technology & Science

Pilani

Acknowledgement

We would like to express our special thanks of gratitude to our Professor Dr. Sundaresan Raman who gave us the opportunity to do a project assignment focussed on solving challenging problems in the field of multimedia computing. His mentorship motivated us throughout the project and guided us in proper direction. We gained a lot of knowledge and deep understanding regarding the subject especially image processing and retrieval through this project. We also gained insights about the research topics and activities in the field of multimedia computing. This research has also greatly helped us in understanding the application of multimedia computing techniques and the kind of challenges that exist in this field.

Table of Contents

1. Introduction

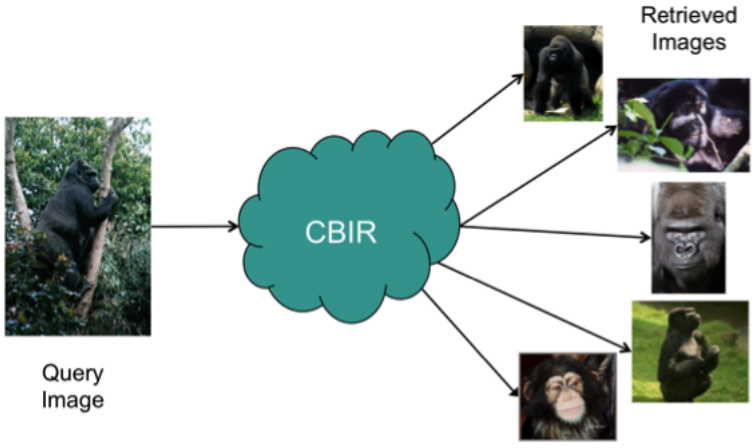
The last decade has witnessed great interest in research on content-based image retrieval(CBIR). This has paved the way for a large number of new techniques and systems, and a growing interest in associated fields to support such systems. Until 2010, older methods like Region based image retrieval, Vector quantization (VQ) on image blocks, Windowed search, Anchoring based image retrieval, Probabilistic frameworks for image retrieval etc. were used to identify content in an image. In spite of broad research endeavors for decades, it stays a standout amongst the most difficult open issues that considerably hinders the achievements of real-world CBIR systems.

The key challenge has been ascribed to the outstanding "semantic gap" issue that exists between low-level image pixels caught by machines and high-level semantics captured by humans. Among different systems, machine learning has been effectively researched as a conceivable heading to reduce the semantic gap in the long haul. Lately, Neural Networks (a set of models in Machine Learning) are being used to perform various tasks. Specifically, Convolutional Neural Networks appear to give astounding results in the context of content-based Image Classification.

2. What is CBIR?

CBIR or content-based image retrieval is the application of multimedia and computer vision techniques to retrieve images similar to an image from a large database of images. This approach is also referred to as query by image content(QBIC) and content-based visual information retrieval (CBVIR). As opposed to concept-based approaches(traditional approaches), this approach aims at searching images by analyzing the contents of the reference image instead of the metadata like tags, keywords, or descriptions of the image. The contents of an image refer to properties like color, textures, shapes etc. i.e. any information that can be extracted from the image itself. Thus the retrieval primarily relies on deriving the appropriate characteristic quantities describing the desired contents of images along with proper and efficient indexing, querying, searching and matching techniques.

Figure : Illustration of a typical CBIR system (Image source: [21])



3. Need and Applications CBIR

CBIR is desirable because searches that rely purely on metadata are dependent on annotation quality and completeness. Having humans manually annotate images by entering keywords or metadata in a large database can be time consuming and may not capture the keywords desired to describe the image. Thus the annotation process is usually inefficient because users, generally, do not make the annotation in a systematic way. In fact, different users tend to use different words to describe a same image characteristic. The lack of systematization in the annotation process decreases the performance of the keyword-based image search [1]. These shortcomings have been addressed by Content Based Image Retrieval.

Content-Based Image Retrieval is being widely used in several diverse applications like medicine, fingerprint identification, biodiversity information systems, museums, crime prevention, historical research etc. A few have these are described below:

- Medical Applications: The number of medical images generated each day at different hospitals and medical centres is far too much to be efficiently used at maintained using metadata. For instance, a medium-sized hospital usually performs procedures that generate medical images that require hundreds or even thousands of gigabytes within a small space of time [2]. Thus, medical domain one of the main areas where CBIR finds its application.
- Art collections and Museums: Fine Arts Museum of San Francisco is an excellent example of application of CBIR. Another example is the digital museum of butterflies, aimed at building a digital collection of Taiwanese butterflies. This digital library includes a module responsible for content-based image retrieval based on color, texture, and patterns [14].
- Commercial: Many modern commercial systems are based on CBIR. IBM's QBIC, Excalibur's Image RetrievalWare, VisualSEEk and WebSEEk and iSearch - PICT to name a few.
- World-wide web: Recent explosive progress of WWW (World-Wide Web),has resulted in a huge image database online. However, most of those images on WWW are not classified with appropriate keywords. The conventional keyword-based image search methods require appropriate keywords, attached to all images in a database, which have to be made by hand in advance, whereas CBIR does not require such keywords.[6]

4. Major challenges in CBIR

The biggest issue for CBIR system is to incorporate versatile techniques so as to process images of diversified characteristics and categories. [7] The performance of CBIR is challenged by various factors like image size, image resolution, illumination variations, multiple objects, nonhomogeneity of intra-region and inter-region textures etc. The other major difficulty is the "semantic-gap". Semantic gap is the gap between inferred understanding by image processing using low level features and human perceptions of the features of given image. In other words, there exists a gap between mapping of extracted features and human perceived semantics[7].

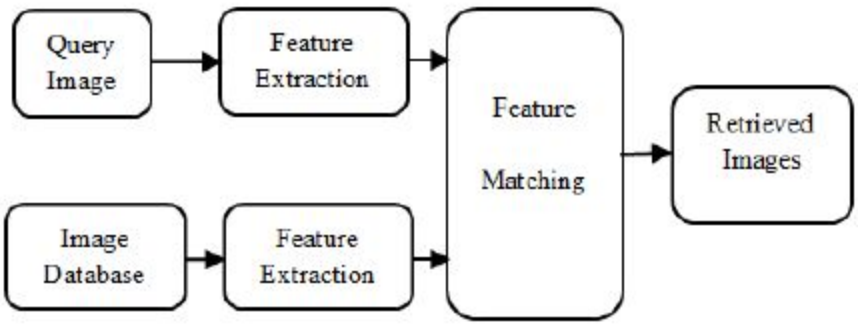
Moreover, Choice of techniques, parameters and threshold-values are mostly requirement specific e.g. the techniques producing good results on an image database of normal WWW images may not produce equally good results for medical images.
Thus the major challenges that need to be addressed for efficient Content based image-retrieval include:
(a) Image Representation (size and resolution)
(b) Feature extraction
(b) Image Similarity Evaluation
(c) Image Annotation/classification
(d) Image Indexing and Database Organization
(e) Query Formulation
(f) Query Result Display and Assessment
(g) Users' Feedbacks & Updating

5. Approaches for Content-Based Image Retrieval

5.1 Traditional approach:

Traditional architecture of CBIR systems is quite similar to that of classical retrieval systems. It includes two basic modules: indexing module and retrieval module. The former is responsible for data processing and constructing indices, which considerably speed up the search. The latter takes care of the retrieval itself by the user request.[8]
The traditional approach to the content-based retrieval tries to search independently by different image features. Each of image feature is represented by a point in the corresponding feature space. Some systems may use several feature spaces to represent the same feature. This may improve retrieval accuracy a bit. Search in each feature space is also performed independently. Then the retrieved sets (intermediate outputs) are merged using data fusion methods into one common output. An output here is a ranked set of retrieved objects closest to the given query. To merge the results of retrieval in different feature spaces in the last step, it is common to use linear combinations of ranks of an element in each intermediate output as its rank in the common output. In the literature, one can find data fusion methods designed specially for merging intermediate outputs in text retrieval [8].

Figure : Illustration of a typical CBIR Architecture



5.2 Developments in the Past Decade:

The last decade has witnessed great interest in research on content-based image retrieval. This has paved the way for a large number of new techniques and systems, and a growing interest in associated fields to support such systems. All approaches to CBIR rely on the following:

5.2.1 Approaches to Feature extraction

Most systems perform feature extraction as a preprocessing step. Feature extraction is used to obtain global image features like color histogram or local descriptors like shape and texture. Some of the key developments and methods suggested are:

- Color and texture: A region based dominant color descriptor indexed in 3-D space along with their percentage coverage within the regions is proposed in [15] which is argued to be more efficient than than high dimensional histograms in terms of search and retrieval. In [16], a multi-resolution histogram capturing spatial image information has been shown to be effective in retrieving textured images.
- Shape: Shape Context for shape matching as proposed in [17] is fairly compact yet robust to a number of geometric transformations.
- Segmentation: Normalized cuts for image segmentation [18].
- Other: 2-d multiresolution hidden markov models for characterizing spatial arrangements of color and texture [19].

5.2.2 Approaches to Retrieval

After the decision of the feature set has been made, we need to steer them towards accurate image retrieval. There has been a large number of fundamentally different frameworks proposed in the last few years like:

- Region based image retrieval
- Vector quantization (VQ) on image blocks
- Windowed search
- Anchoring based image retrieval
- Probabilistic frameworks for image retrieval

5.2.3 Annotation and Concept Detection

While image retrieval has been active over the years, an emerging new and possibly more challenging field is automatic concept recognition from visual features of images. This involves using the following techniques:

- Supervised classification
- Translation Approach

5.2.4 Approach using Convolutional Neural Networks

Until 2010, older methods like the ones mentioned above were used to identify content in an image. Lately, Neural Networks (a set of models in Machine Learning) are being used to

perform various tasks. Specifically, Convolutional Neural Networks appear to give astounding results in the context of content-based Image Classification.

6. DEEP LEARNING FOR CBIR

6.1 Overview

Deep learning refers to a class of machine learning techniques, where many layers of information processing stages in hierarchical architectures are exploited for pattern classification and for feature or representation learning. It lies in the intersections of several research areas, including neural networks, graphical modeling, optimization, pattern recognition, and signal processing, etc. Deep learning has a long history, and its basic concept is originated from artificial neural network research. The feed-forward neural networks with many hidden layers are indeed a good example of the models with a deep architecture. Back-propagation, popularized in 1980's, has been a well-known algorithm for learning the weights of these networks. For example, LeCun et al. [28] successfully adopt the deep supervised back-propagation convolutional network for digit recognition. Recently, it has become a hot research topic in both computer vision and machine learning, where deep learning techniques achieve state-of-the art performance for various tasks.

The deep convolutional neural networks (CNNs) proposed in [26] came out first in the image classification task of ILSVRC-2012. The model was trained on more than one million images, and has achieved a winning top-5 test error rate of 15.3% over 1, 000 classes. After that, some recent works got better results by improving CNN models. The top-5 test error rate decreased to 13.24% in [43] by training the model to simultaneously classify, locate and detect objects. Besides image classification, the object detection task can also benefit from the CNN model, as reported in [14]. Generally speaking, three important reasons for the popularity of deep learning today are drastically increased chip processing abilities (e.g., GPU units), the significantly lower cost of computing hardware, and recent advances in machine learning and signal/information processing research.

6.2 Convolutional Neural Networks (CNN)

6.2.1 Overview

Convolutional neural networks or CNNs, are a specialized kind of neural network for processing data that has a known, grid-like topology[3]. CNNs are quite similar to normal Neural Networks. They are made up of neurons having learnable weights and biases. However, CNN architecture makes the explicit assumption that the inputs are images. This enables encoding of certain special properties into the architecture. These, in turn, make the forward function more efficient to implement.

Thus, these networks use a special architecture which is particularly well-adapted to classify images. Using this architecture makes convolutional networks fast to train. This, in turns, helps us train deep, many-layer networks, which are very good at classifying images. Today, deep convolutional networks or some close variant are used in most neural networks for image recognition.[7]

There are four main steps in CNN: convolution, subsampling, activation and full connectedness. They are briefly described below.

- Convolution can be understood as application of filters over the images.
- Subsampling is done in order to converge the network into smaller portion by extracting the important information from the lower layers.
- The activation layer controls how the signal flows from one layer to the next, emulating how neurons are fired in our brain. Most common activation function used with CNN is ReLU (Rectified Linear Unit).
- The fully connected layer aggregates all the information and is then passed through a softmax function to provide the class probabilities.

Small (often minimal) receptive fields of convolutional winner-take-all neurons yield large network depth, resulting in roughly as many sparsely connected neural layers as found in mammals between retina and visual cortex. Only winner neurons are trained. Several deep neural columns become experts on inputs preprocessed in different ways; their predictions are averaged.

Figure : Brief description of Convolution

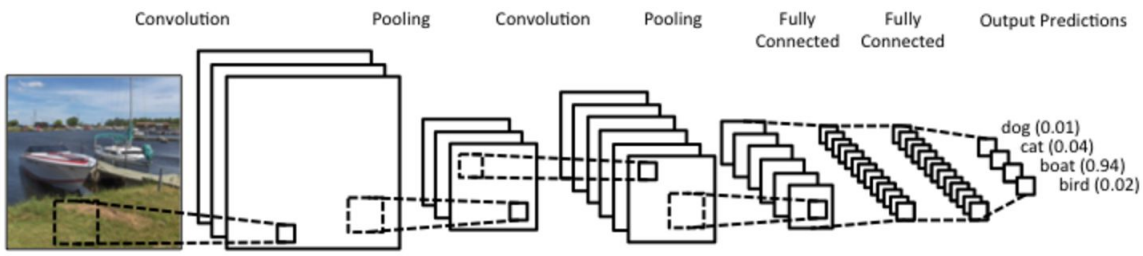


Image source: [9]

6.2.2 Why Convolutional Neural Networks?

As we know, Neural Networks in general are a models based on the structure of human brain. Convolutional Neural Network try to model the Visual Cortex part of the brain and hence perform extremely well in the area of Computer Vision. These networks use a special architecture which is particularly well-suited to classify images. Using this architecture makes convolutional networks fast to train. This, in turn, helps us train deep, many-layer networks, which are very good at classifying images. Today, deep convolutional networks or some close variant are used in most neural networks for image recognition.

There are two main aspects of this computation: **Location Invariance** and **Compositionality**. For example suppose we want to classify whether there's a car in an image or not . Now we don't really really care where the car occurs and thus we are sliding our filters all over the image. In practice, pooling also gives us invariance to rotation, translation and scaling. The second important aspect is compositionality i.e. each filter composes a local patch of lower-level features into higher-level representation. This is what makes CNNs so powerful for Computer Vision. It makes intuitive sense that we build our image step by step i.e. first edges from pixels, the shapes from edges, and finally more complex objects from shapes.

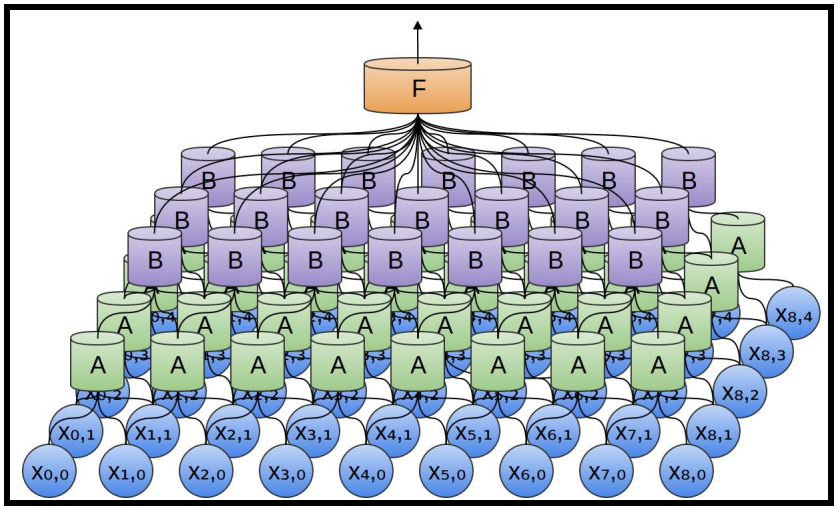
6.2.3 Architecture

The architecture of a typical CNN is composed of multiple convolutional and subsampling layers. Each layer perform a specific function to transform its input into more useful representation.

Input to Convolutional Layer A ***m x m x r Image*** where ***m*** is the height and width of the image and ***r*** is the number of channels.

Within the Layers A number of filters (or Kernels) of size ***n x n x q*** where ***n*** is smaller than the dimension of the image. The size of the filters gives rise to the locally connected structure which are each convolved with the image to produce ***k*** feature maps of size ***m-n+1***.
Each map is subsampled typically with **mean** or **max pooling** over ***p x p*** contiguous regions, where ***p*** ranges between 2 (for small images) to 5 (for larger inputs).
An additive bias and sigmoidal nonlinearity is applied to each feature map. The figure below describes a 2D CNN with Convolutional and Subsampling Layers. The units with same color have their weights tied. After convolutional layers, there are many fully connected layers, which appear identical to standard multilayer neural network.[11]

Figure : A 2D Convolutional Neural Network with Convolutional and Subsampling Layers



Source UFLDL [11]

6.3 Feature Representation for CBIR

Although CNNs have been shown with promising results for classification tasks, it remains unknown how it can perform for CBIR tasks.
Various techniques proposed in [13] discuss various schemes of feature representation as follows:

- Direct Representation
  In this the images are fed directly in new datasets into the input layer of the pre-trained CNN model, and then take the activation values from the last n layers.
- Refining by Similarity Learning

In this instead of directly using the features extracted by the pre-trained deep model, we attempt to explore similarity learning (SL) algorithms to refine the features.

- Refining by Model Retraining
  In this scheme, we will retrain the deep convolutional neural networks on the new image dataset for different CBIR tasks by initializing the CNN model with the parameters of the ImageNet-trained models by the method of Transfer Learing

The following sections of the report describe the approach taken to accomplish the task and the results obtained.

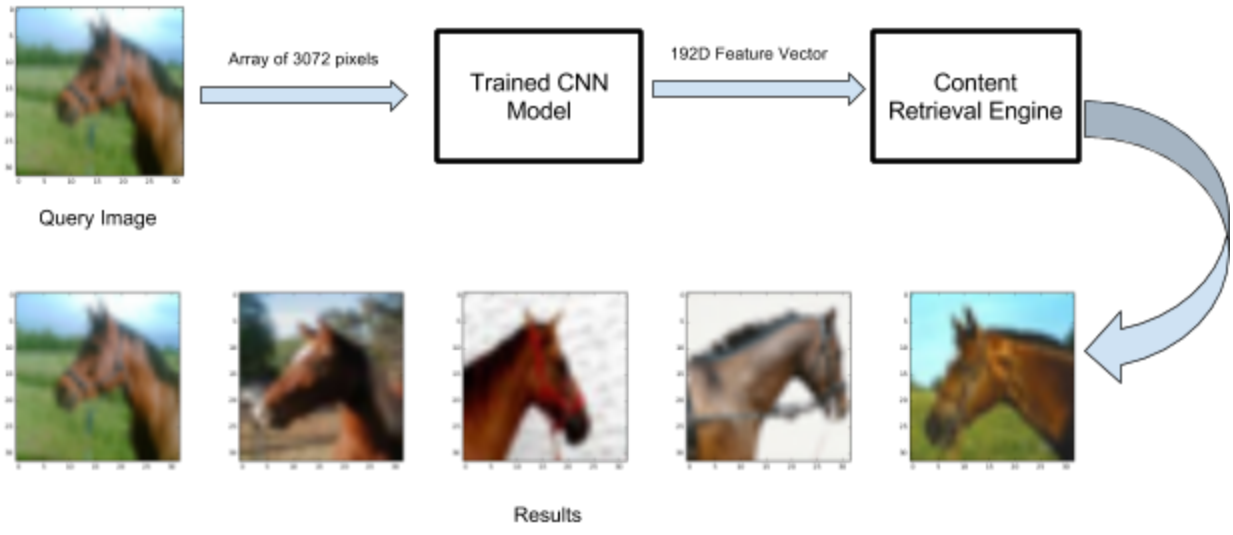
7. Our approach

We aim to build a Neural network architecture which will serve as a model to perform the task of Content based Image Retrieval. This model will be a modification of the currently existing architectures built for the task of image classification. We are using a slightly modified version of the Direct Representation Scheme. The activation values from the penultimate fully connected layer will be used to generate an image descriptor [A feature vector of 192 dimensions that represents each image]. This image descriptor will then be used to find out similarity between images.The similarity coefficient will determine the retrieval of the closest matching images for the given query image from our dataset.

7.1 Proposed Architecture of CBIR Model

The figure below elucidates the overview of our CBIR model. We first transform the query image to a 32x32 pixel JPEG image.We then feed in the transformed image to the trained CNN model which produces the activation values of the penultimate layer as the feature vector which is taken in by the Content Retrieval Engine which looks for the images in the database and retrieves the top matches.

Figure : Overview of our architecture



7.2 Trained CNN Model

The CNN model is a multi-layer architecture consisting of alternating convolutions and nonlinearities. These layers are followed by fully connected layers leading into a final layer which produces the image feature vector.With a few changes in the top few layer, this model follows the architecture described by Alex Krizhevsky [5].

Layer Name	Description
conv1	convolution and rectified linear activation.
pool1	max pooling.
norm1	local response normalization.
conv2	convolution and rectified linear activation.
norm2	local response normalization.
pool2	max pooling.
local3	fully connected layer with ReLU
local4	fully connected layer with ReLU

7.3 Content Retrieval Engine

The Content retrieval engine mainly performs two actions. It initially builds the database of images and indexes them appropriately. Then it provides function calls which take in the feature vector of query image and retrieves similar images based on the features.

(1) Database Building

The database is built by passing all the training images through the trained model and saving their feature vectors. To support efficient querying, we have used KDTree data structure to store the image vectors and their ids which are the indices.

(2) Query

The query interface loads the KDTree which is saved into the memory. When we supply the feature vector of the query image, the KDTree returns ids of similar images. These ids are used to index into the training images. The images are then retrieved and displayed.

7.4 Image Dataset

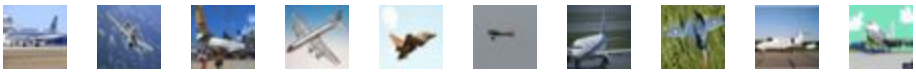
The CIFAR-10[20] are labeled subsets of the 80 million tiny images dataset. Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton collected them.The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is split into multiple batches, each with 10000 images. There are five training batches and one test batch. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another.

This data is available as a pickle file which when decoded, produces images in the form of 2d arrays. Each image is represented by an array of 3072 pixels. These pixels are arranged as 1024 Red, 1024 Green and 1024 Blue pixel values respectively.

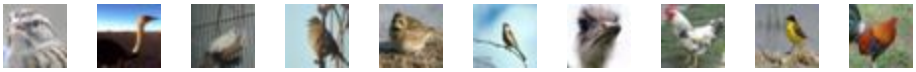
Here are the classes in the dataset, as well as 10 random images from each:

airplane



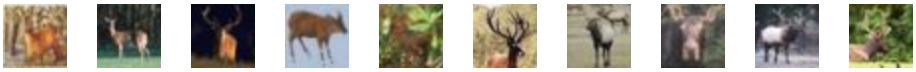
automobile

bird

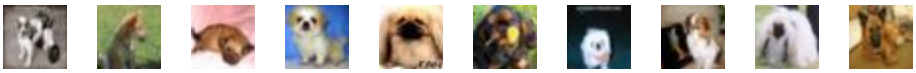


cat

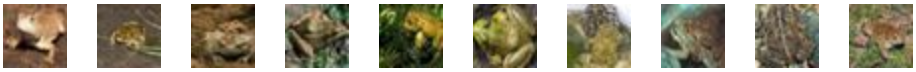
deer



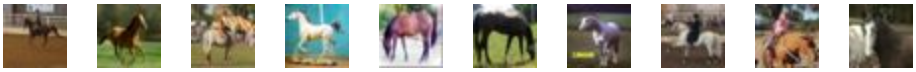
dog



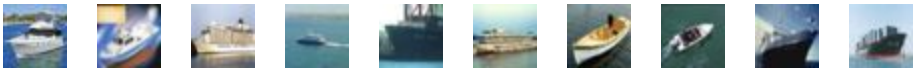
frog



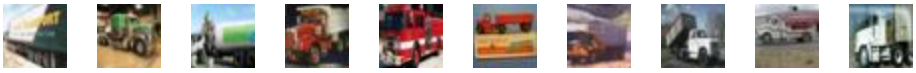
horse



ship



truck



7.5 Results

With 50000 train images, 10000 test images the model achieved an accuracy of 85% for the task of classification. After the training phase, this model was used to perform the task of Content Based Image Retrieval. Input query images from multiple sources were given to the model and the results were observed to be quite accurate. Given below are the results of three samples, one taken from the training data itself, one from the test data and one taken randomly from internet.

Figure : Query image from train image set

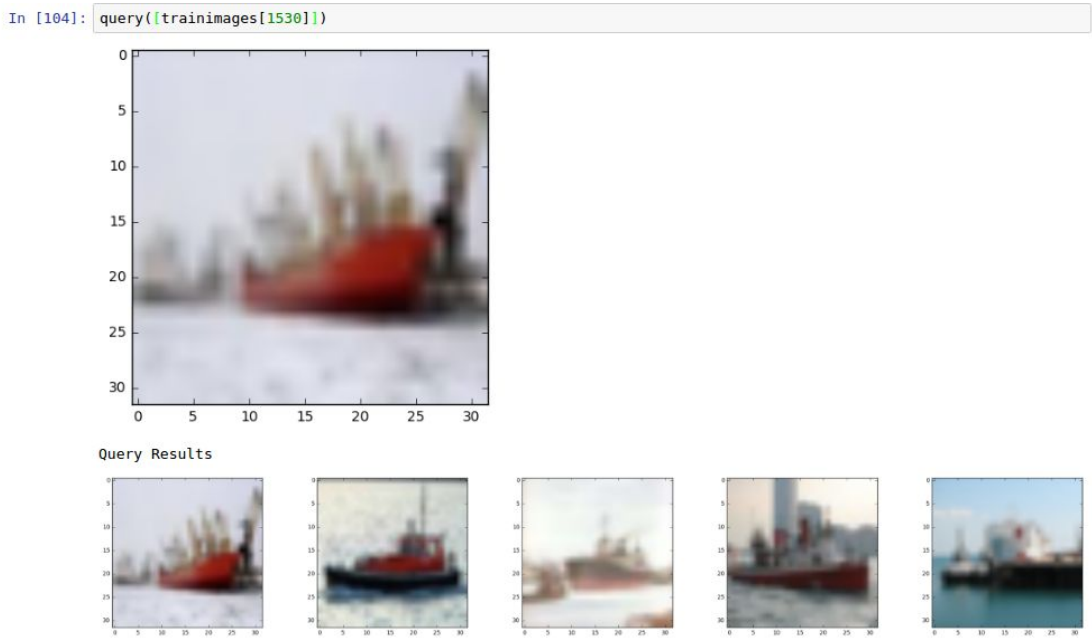
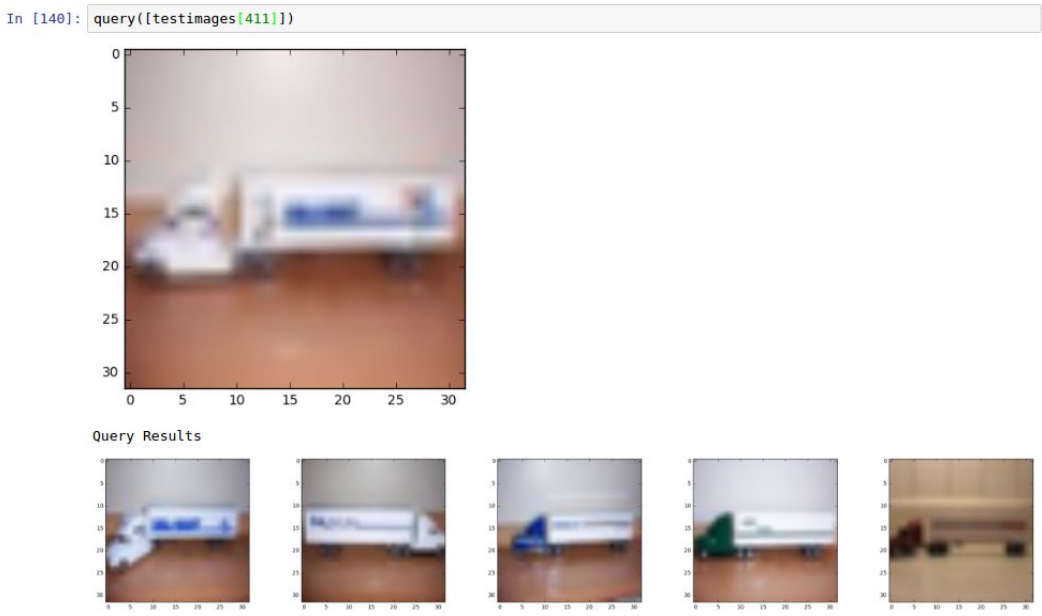


Figure : Query images from test image set

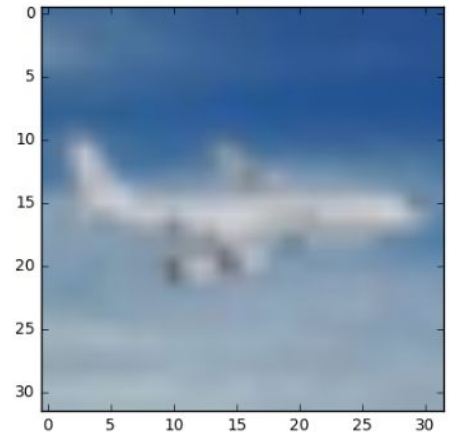


```
In [124]: query([testimages[1290]])
```

Query Results

Figure : Query image from the internet

```
In [116]: query(getarray("airplane.jpg"))
```



Query Results

8. Conclusions

Inspired by recent successes of deep learning techniques, we attempted to address the long-standing fundamental feature representation problem in Content-based Image Retrieval (CBIR). We tried to understand if deep learning is a hope for bridging the semantic gap in CBIR for the long term, and how much empirical improvements in CBIR tasks can be achieved by exploring the state-of-the-art deep learning techniques for learning feature representations and similarity measures.

In particular, we developed a model of deep learning with application to CBIR tasks by examining a state-of-the-art deep learning method (convolutional neural networks) for CBIR task using the CIFAR 10 dataset.

9. References

1. Ricardo da Silva Torres, Alexandre X. Falcão: Content-Based Image Retrieval: Theory and Applications. RITA 13(2): 161-185 (2006)
2. João Augusto da Silva Júnior, Rodiney Elias Marçal, Marcos Aurélio Batista: Image Retrieval: Importance and Applications. X Workshop de Vis˜ao Computacional - WVC 2014
3. Ian Goodfellow and Yoshua Bengio and Aaron Courville Deep Learning Book in preparation for MIT Press, Last accessed online at http://www.deeplearningbook.org, on 14.11.2016.
4. Ciresan, Dan, Ueli Meier, and Jürgen Schmidhuber. “Multi-column deep neural networks for image classification.” Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
5. Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, (NIPS 2012), 2012.
6. Keiji Yanai, Masaya Shindo, Kohei Noshita: A Fast Image-Gathering System on the World-Wide Web Using a PC Cluster. Web Intelligence 2001: 324-334
7. Darshak G. Thakore , A. I. Trivedi: Content based image retrieval techniques – Issues, analysis and the state of the art
8. N. S. Vassilieva: Content-based image retrieval methods. Programming and Computer Software 35(3): 158-180(2009)
9. Denny Britz Understanding Convolutional Neural Networks for NLP; Image online:http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/
10. Ji Wan, Dayong Wang, Steven Chu-Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, Jintao Li: Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. ACM Multimedia 2014: 157-166
11. Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, Caroline Suen, Adam Coates, Andrew Maas, Awni Hannun, Brody Huval, Tao Wang, Sameep Tandon, Deep Learning Tutorial, Online: http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/ , Accessed on 14.11.2016.

12. Understanding Convolutional Neural Network: http://colah.github.io/posts/2014-07-Conv-Nets-Modular/
13. Ji Wan, Dayong Wang, Steven Chu-Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, Jintao Li: Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. ACM Multimedia 2014: 157-166
14. R.S. Torres and A.X. Falcão. Content-Based Image Retrieval: Theory and Applications. Revista de Informática Teórica e Aplicada, nro 2, volume 13, 2006, pags 161–185.
15. Y. Deng, B. S. Manjunath, C. Kenney, M. S. Moore, and H. Shin, “An Efficient Color Representation for Image Retrieval,” IEEE Trans. Image Processing, 10(1):140–147, 2001.
16. E. Hadjidemetriou, M. D. Grossberg, and S. K. Nayar, “Multiresolution Histograms and Their Use for Recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, 26(7):831–847, 2004.
17. S. Belongie, J. Malik, and J. Puzicha, “Shape Matching and Object Recognition Using Shape Contexts,” IEEE Trans. Pattern Analysis and Machine Intelligence, 24(4):509–522, 2002
18. J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Trans. Pattern Analysis and Machine Intelligence, 22(8):888–905, 2000.
19. J. Li, R. M. Gray, and R. A. Olshen, “Multiresolution Image Classification by Hierarchical Modeling with Two Dimensional Hidden Markov Models,” IEEE Trans. Information Theory, 46(5):1826–1841, 2000
20. Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, CIFAR 10 Dataset. Url: https://www.cs.toronto.edu/~kriz/cifar.html
21. https://blog.pivotal.io/data-science-pivotal/features/content-based-image-retrieval-using-pivotal-hd-or-pivotal-greenplum-database