Content-Based Image Retrieval using Deep Learning: A Comprehensive Survey

Latika Pinjarkar [1*], Manisha Sharma [2], Smita Selot [3]

Shri Shankaracharya Technical campus, Bhilai(C.G.),490020,India,

latikabhorkar@gmail.com[1]

Bhilai Institute of Technology, Bhilai(C.G.), 490020,India [2]

Shri Shankaracharya Technical campus, Bhilai(C.G.), 490020,India [3]

Abstract: Learning effectual feature representations and similarity measures are essential to the retrieval performance of a content-based image retrieval (CBIR) system. In spite of wide research hard work for decades, it remnants one of the most challenging open problems that considerably hold back the achievements of real-world CBIR systems. The key challenge has been credited to the well-known "semantic gap" issue that exists between low-level image pixels captured by machines and high-level semantic concepts perceived by human. Amongst various techniques, machine learning has been actively examined as a probable way to bridge the semantic gap in the long term. There are recent successes of deep learning techniques for computer vision and other applications. This paper provides a detail survey on: if deep learning is a hope for bridging the semantic gap in CBIR and how much improvements in CBIR tasks can be achieved by exploring the state-of-the-art deep learning (Convolutional Neural Networks) techniques for learning feature representations and similarity measures.

Keywords: Content-Based Image Retrieval (CBIR), Deep Convolutional Neural Networks (CNN), Feature representation

1. Introduction

The studies by multimedia researchers over decades proved that the retrieval performance of a content-based image retrieval system significantly depends on the feature representation and similarity measurement. Even though a variety of techniques have been anticipated, it leftovers one of the most challenging problems in current content-based image retrieval (CBIR) research, which is mostly due to the eminent "semantic gap" issue that exists between low-level image pixels captured by machines and high-level semantic concepts perceived by human. From a high-level perspective, such challenge can be rooted to the fundamental challenge of Artificial Intelligence (AI) that is, how to build and train intelligent machines like human to tackle real-world tasks. Machine learning is one capable practice that attempts to address this imposing challenge in the extensive term.

Recent years have observed some significant advances of new techniques in machine learning. One important infiltrate technique is recognized as "deep learning", which includes a family of machine learning algorithms that attempt to mold high-level abstractions in data by utilizing deep architectures compiled of multiple non-linear transformations [1, 2]. Contrasting usual machine learning methods, deep learning imitates the human brain that is arranged in a deep architecture and processes information through multiple stages of transformation and representation. By exploring deep architectures to learn features at multiple level of abstracts from data automatically, deep learning methods permit a system to learn complex functions that directly map raw sensory input data to the output, without relying on human-crafted features using domain knowledge. Many recent studies have reported hopeful results for employing deep learning techniques to a variety of applications, including speech recognition [3, 4], object recognition [5, 6], and natural language processing [7, 8], along with others.

The researches issues open for addressing in this field are:

a. How deep learning methods can be efficient for learning superior feature representations from images to deal with CBIR jobs?
b. How much enhancements can be attained by deep learning techniques when evaluated with traditional features representation by experts in multimedia and computer vision?
c. How to pertain and adapt an accessible deep learning model trained in one domain to a new CBIR job in a different domain efficiently?

In this paper pragmatic studies for comprehensive assessments of deep convolutional neural networks with application to learn feature representations for a variety of CBIR tasks under varied settings are reported. The rest of this paper is organized as follows. Section 2 provides the reviews of related work. Section 3 introduces the structure of deep learning for CBIR. Section 4 gives the application area of deep learning to information retrieval domain. Section 5 concludes this paper.

2. Reviews of the related work in the area

The following section describes the contributions of the various researchers in the area based upon the two described learning techniques.

2.1 Distance metric Learning based systems

Distance metric learning for image retrieval has been extensively studied in both machine learning and multimedia retrieval communities [9, 10, 11, 12, 13, 14, 15, and 16]. In the following, we briefly discuss different groups of existing work for distance metric learning organized by different learning settings and principles. In terms of training data formats, most existing DML studies often work with two types of data (a.k.a. side information): pairwise constraints where must-link constraints and cannot-link constraints are given and triplet constraints that contains a similar pair and a dissimilar pair. There are also studies that directly use the class labels for DML by following a typical machine learning scheme, such as the LargeMargin Nearest Neighbor (LMNN) algorithm [11], which however is not essentially different. In terms of different learning approaches, distance metric learning techniques are typically categorized into two groups: the global supervised approaches [10, 17] that learn a metric on a global setting by satisfying all the constraints simultaneously, the local su- pervised approaches [9, 11] that learn a metric on the local sense by only satisfying the given local constraints from neighboring information. In terms of learning methodology, most existing DML studies generally employ batch learning methods which often assume the whole collection of training data must be given before the learning task and train a model from scratch. Unlike the batch learning methods, in order to handle large-scale data, online DML algorithms have been actively studied recently [18, 19]. The key idea of distance metric learning is to learn an optimal metric which minimizes the distance between similar images and simultaneously maximizes the distance between dissimilar images. In this condition, another technique named similarity learning is closely related to distance metric learning.

2.2 Deep Learning based systems

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. [5] 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 stat-of-the art performance for various tasks. The deep convolutional neural networks (CNNs) proposed in [5] 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 [22] 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 [23]. 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. Over the past several years, a rich family of deep learning techniques has been proposed and extensively studied, e.g., Deep Belief Network (DBN) [24], Boltzmann Machines (BM) [25], Restricted BoltzmannMachines (RBM) [26], Deep BoltzmannMachine (DBM) [27], Deep Neural Networks (DNN) [3], etc. More detailed survey of latest deep learning studies can be found in [11]. Among various techniques, the deep convolutional neural networks, which is a discriminative deep architecture and belongs to the DNN category, has found state-of-the-art performance on various tasks and competitions in computer vision and image recognition [5,20, 28, 29]. Specifically, the CNN model consists of several convolutional layers and pooling layers, which are stacked up with one on top of another. The convolutional layer shares many weights, and the pooling layer sub-samples the output of the convolutional layer and reduces the data rate from the layer below. The weight sharing in the convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some "invariance" properties (e.g., translation invariance). Zeiler and Fergus [6] introduced a novel visualization technique that gives insight into the function of intermediate feature layers and the operation of the classifier. For feature representation, they keep the top $1 - 7$ layers of the ImageNet-trained model fixed and retrain a new softmax classifier on top using the training images in the new database. Our work is fundamentally different from these studies in that we focus on evaluating feature representation performance on CBIR tasks, where we aim to learn an effective distance measure for retrieval tasks instead of classifiers in recognition tasks. Finally, we note that our work is also very different from another recent study in [30] which aims to address multimodal image retrieval using deep learning and their raw input still rely on human-crafted features. By contrast, we aim to learn features directly from images without domain knowledge.

3. Deep learning for CBIR

This section, introduces the deep learning framework for CBIR, which consists of two stages: (1) training a deep learning model from a large collection of training data; and (2) applying the trained deep model for learning feature representations of CBIR tasks in a new domain.

3.1 Deep Convolutional Neural Networks

In general, the deep convolutional network, (a), consists of two parts: 1) the convolution layers and maxpooling layers, and 2) the fully connection layers and the output layers. Specifically, the first layer is the input layer which adopts the mean-centered raw RGB pixels in intensity value.

To reduce overfitting, two data augmentation tricks are performed: first, the input images are generated with translation and horizontal reflections by extracting random 224×224 patches from the original 256×256 images and our network is trained on these extracted patches; second, to capture the invariance in illumination and color, they add random multiples of the principle components of the RGB pixel values throughout the dataset.

Following the input layers, there are five convolutional layers. The first and the second convolution layers are following with a response normalization layers and a max pooling layers, while the third, fourth, and fifth convolution layers are connected to one another without any intervening pooling or normalization. There are several novel or unusual features in Krizhevsky's convolutional neural network, which makes it work better than previous convolutional neural networks. First, the neuron output function f is the nonlinear function: Rectified Linear units (ReLUs), which can reduce the training time of the deep convolutional neural networks several times than the equivalents with "tanh"' units. Second, they adopt the "local response normalization", which is helpful for generalization. Last but not least, they adopt the

"overlapping pooling" scheme. Max pooling layers are very common in general convolutional neural networks, which summarize the outputs of neighboring groups of neurons in the same kernel map. The max pooling step can enhance the transformation invariance of the feature mapping. Traditionally, the neighborhoods summarized by adjacent pooling units do not overlap. By adopting overlapped neighborhoods, they can reduce the top-1 and top-5 error rates by 0.4% and 0.3%, respectively.
Following the convolutional layers, there are two more fully connected layers with 4, 096 neurons, denoted as "FC1" and "FC2". The last output layer, which is fed by the "FC2" layer, is a 1000-way softmax layer which produces a distribution over the 1, 000 class labels. In the whole deep convolutional neural network, there are about 60 million parameters in total.

3.2 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.
Specifically, to apply a trained CNNs model for direct feature representation, the activations of the last three fully connected layers (FC1, FC2, and FC3) are taken as the feature representations for CBIR tasks.
In the following, three kinds of feature generalization schemes are discussed in detail.

1. **Direct Representation**

This is the direct feature representation as discussed above. We assume the retrieval domain is similar to the original dataset for training the CNN model. In this scenario, we will simply adopt one of the activation features DF.FC1, DF.FC2, and DF.FC3, directly. To obtain the feature representation, we directly feed the images in new datasets into the input layer of the pre-trained CNN model, and then take the activation values from the last three layers. Since we only need to compute the feed forward network based on the matrix multiplication for one time, the whole scheme will be very efficient. In our experiment, we also normalize the feature representation with l2-norm.

2. **Refining by Similarity Learning**

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 in scheme 1.Various distance metric learning or similarity learning algorithms can be used, according to the training data available in the new CBIR tasks.

3. **Refining by Model Retraining**

Scheme 3 will re train 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. Depend on the available label information; there are two ways to retrain the CNN model. 1) Refining with class labels. For datasets with class labels, we can retrain the model by optimizing the classification objective function. In this case, all layers of the new model will be initialized based on our ImageNet-trained model except the last output layer, which is adapted to the number of class labels of the new dataset and initialized randomly. Then, we update the whole convolutional neural networks by training on images from the new dataset. We can retrain the CNN model with similarity learning objective function, like what we do in scheme 2, and back-propagate the errors to previous layers in order to refine the entire model on the new dataset.

4. **Applications of deep learning**

1. Speech and audio Recognition
2. Image, video, and multimodality

3. Language modeling
4. Natural language processing
5. Information retrieval

5. Conclusion

This paper studies the works proposed by various researchers in the area of CBIR based upon deep learning. Deep learning is a promising technique for CBIR. This learning can be employed to improve the retrieval results of the CBIR systems. The integration of deep learning with CBIR application can be experimented and tested as a future research direction.

References:

1. Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.
2. L. Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing, 3:e2, 2014.
3. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6):82–97, 2012.
4. D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide. Feature learning in deep neural networks - a study on speech recognition tasks. CoRR, abs/1301.3605, 2013.
5. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012.
6. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.
7. E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *ACL (1)*, pages 873–882, 2012.
8. T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.
9. C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptivemetric nearest-neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1281–1285, 2002.
10. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *ICML*, pages 11–18, 2003.
11. K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.
12. J.-E. Lee, R. Jin, and A. K. Jain. Rank-based distance metric learning: An application to image retrieval. In *CVPR*, 2008.
13. M. Guillaumin, J. J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009.
14. Z. Wang, Y. Hu, and L.-T. Chia. Learning image-to-class distance metric for image classification. *ACM TIST*, 4(2):34, 2013.
15. S. Mian, Y. Hu, R. Hartley, and R. A. Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *IEEE Transactions on Image Processing*, 22(12):5252–5262, 2013.
16. D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. He, and C. Miao.Learning to name faces: a multimodal learning scheme for search-based face annotation. In *SIGIR*, pages 443–452, 2013.
17. S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *CVPR (2)*, pages 2072–2078, 2006.

18. P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *NIPS*, pages 761–768, 2008.
19. R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS*, pages 862–870, 2009.
20. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
21. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.
22. A. K. Jain and A. Vailaya. Image retrieval using color and shape. Pattern Recognition, 29(8):1233–1244, 1996
23. . B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.
24. G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.
25. D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines*. *Cognitive science*, 9(1):147–169, 1985.
26. R. Salakhutdinov, A. Mnih, and G. E. Hinton. Restricted boltzmann machines for collaborative filtering. In *ICML*, pages 791–798, 2007.
27. R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *AISTATS*, pages 448–455, 2009.
28. D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *NIPS*, pages 2852–2860, 2012.
29. J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In *NIPS*, pages 1232–1240, 2012.
30. P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao. Online multimodal deep similarity learning with application to image retrieval. In *ACM Multimedia*, pages 153–162, 2013.