Content based medical image retrieval using deep learning

Mohamed. A ELMarakbey[1], Eslam. Gamal[1], Mazen. Fouad[1], Ahmed. Gamal[1]
[1] *Systems & Biomedical Engineering Department, Cairo University, Cairo, Egypt*

***Abstract*— Learning effective hierarchal representation is crucial to the performance of any medical retrieval system. Deep Learning has been introduced as an unsupervised hierarchical feature extraction tool that outperformed most of the hand crafted based systems. Inspired by recent successes of deep learning techniques for unsupervised feature extraction. In this paper, we investigate the performance of state-of-the-art Convolutional Neural Networks (CNN) for learning hierarchical and discriminative feature representations for x-ray images. Several architectures are studied, to retrieve similar images using the IRMA images dataset, demonstrating the effect of different configurations, named as the number of layers, the filter size, the elimination of the max pooling layer and the addition of the drop put layer, on the performance of the retrieval system. Deep learning based systems were compared to the handcrafted based features and were found to outperform the later with accuracy of 92.69 % by the best performing configuration.**

I. INTRODUCTION

Content-based image retrieval has been an active area of research for more than two decades [1] [2]. As a branch of computer vision, CBIR aims to search in large databases based on the content of the image such as color, texture, shape and any other visual information extracted from the image itself [3]. Such systems can assist clinicians by inspecting a given case query image to a more precise reliable diagnosis process by comparing it with the past diagnosed cases and retrieving the most similar ones from the imaging archive.

Medical images have unique characteristics that make them different. For example, most medical images are single channel images (Gray scale). Another instance is that images captured from the close body regions (head, spine, pelvis etc.), to a great extent, present large global similarities, even from different individuals. Furthermore, the most valuable information in medical images usually appears to be located in a very small image region, known as the region of interest (ROI) [4]. Not only global properties but also specific local features should be considered when a medical image retrieval system to be designed, such systems require careful feature extraction process demanding extensive medical domain knowledge, which can be exhaustive design problem in a large image databases, where same features cannot be used for different types of tasks even on the same database.

State-of-the-art Deep Neural Networks [5] are very promising techniques that attempt to address this problem effectively. By modeling high-level abstractions in data employing deep architectures composed of multiple non-linear transformations to learn features at multiple level from data automatically [6], without relying on human-crafted features using domain knowledge or prior information about the image data, bridging the "semantic gap" issue that exists between low-level image pixels captured by machines and high-level semantic concepts perceived by human.

In this paper, we explore the performance of state-of-the-art Convolutional Neural Networks (CNN) as a feature extractor for learning hierarchical and discriminative image representations. Several architectures are investigated using the IRMA images [7] dataset demonstrating the effect of different CNN configurations on the performance of the retrieval system. In Section 2, a brief literature review on CBIR and CNN in medical imaging is presented. In Section 3, we describe the proposed architectures of the CNN's explored in this paper. Experiments of the proposed architectures are reported in section 4. Section 5 represents the results and discussion. In section 6, we conclude the paper with suggestions for future work.

II. LITERATURE REVIEW

In the beginning era of digital image search, various searching methods were investigated, although researchers were mainly focused on text-based search to retrieve images. Performing image search based on visual information, generally called Content Based Image Retrieval (CBIR), has increasingly become more difficult in recent years. Although CBIR systems differ in the methods applied to image in order to retrieve, store features and measure the similarity, the basic architectures of the systems are quite similar. Feature extraction and similarity measurement are considered as the main processes of most CBIR systems [8], [9].

Wei et al. [10] proposed a Gabor filtering method to extract the textural features for mammogram retrieval. Tommasi et al [11] [12] developed a multi-cue approach based on the Support Vector Machine (SVM) to annotate medical images automatically by combining the global and local features. Lehmann et al. [13] compared the performance of different approaches for automatic categorization of medical images. Recently, artificially Deep CNN has been gaining an increased use, since it was introduced by LeCun et al. [14] and Hinton [15]. Krizhevsky et al [6] was the first to use CNN in the Large Scale Visual Recognition Challenge. Sahiner et al. [16] was the first to use CNN in medical imaging applications for classification of mass and normal breast tissue. Cernazanu-Glavanet al. [17] used CNN for bone segmentation in X-ray images. Moreover, Prasoon et al. [18] employed CNNs to classify bone and non-bone areas in an X-ray image to help speed up extraction of features.

III. METHODS

A. convolutional neural network

A convolutional neural network (CNN) is a type of feed forward artificial neural network that is designed to use minimal image pre-processing and work on raw image data. After proper training, the CNN is usually able to extract hierarchical and discriminative feature representations for the images. On the contrary to hand crafted features, the parameters in the CNN that are used for feature extraction are automatically learned by the network in the process of training, without the need of extensive domain knowledge or prior information about the data. Although a very deep convolutional neural network has been shown to have better

performance on image classification [5], we use a relatively shallow architecture mainly for the sake of limited computational power and dataset size.

B. CNN Architecture

A typical CNN is formed of a repetitive series of Convolution, ReLu and pooling layers, inspired by the design principle of the VGGNet by Simonyan et al. [5].Multiple stacked Convolution filters extract features from the input image. The pooling layer between convolutional layers in a convnet architecture is used to reduce the spatial size of the image resulting in decreased amount of parameters and computational time. The Convnet is usually followed by a series of fully connected layer that uses high level features from the last convolutional layer to classify the input image into various classes. In this study, we used a convolutional layer, formed of filters sizes are of small receptive fields: 3 x 3 and 5 x 5 based on the layers configuration, the convolution stride is set to 1 pixel in all configurations. Spatial pooling in form of max pooling is used to reduce the dimensionality of feature maps, pooling is performed over a 2 x 2 pixel window with stride of two pixels. A stack of convolutional layers (with different order and depth based on the configuration) is followed by 3 fully connected layers of different channel sizes, having third layer classifying the image to one of the classes represented in the dataset. The final layer is a softmax layer. All hidden layers are followed by ReLu layer Krizhevsky et al [4].

C. Choice of filter sizes

The choice of filter sizes is inspired by the work of Simonyan et al. [19], where the authors use filters with smaller receptive field (3 x 3) and stride of one to convolve with every pixel in the image. They were able to represent larger filters by stacking multiples of (3 x 3), for an example stacking two (3 x 3) filters has the same receptive field as one (5 x 5) filter and stacking three (3 x 3) filters has the same effective receptive field as one (7 x 7). Stacking more convolution filters has the benefits of increasing the non-linearity (every conv layer is followed by a ReLu layer), which increase the capability of the convnet decision function as more discriminative through decreasing the number of parameters with a large factor.
In addition, increasing the depth of the network increases its capacity to learn more discriminative features and build a hierarchal representation from low level to high level features. [19].

IV. EXPERIMENTS

A. Dataset Description

The IRMA dataset [7], supplied by the imageCLEF organization, was employed in our study. It consists of a collection of 14410 x-ray image over different anatomical/biological/body orientation [7]. The dataset is annotated using IRMA code [20] which is a 13 character string containing four mono-hierarchal axes representing the information on technical, biological and anatomical details of the image. IRMA dataset is quite challenging since it is formed of low quality images due to imbalance of the data set size for each category, the poor contrast of the images, the low Signal to Noise Ratio, the huge interclass variability and small intraclass variability, the different image sizes, the different scale of the object in the image as well as the non-standardized directionality of the organs. Over the years, various works like [21], [22], [23], [24] have considered 1617 images in 6 classes, 9100 images in 40 classes, 6231 images, 5000 images in 20 classes respectively, from the IRMA dataset.
The Dataset is supplied by the organization in the form of train/test sets, training set is divided into train/validation sets with percentage 90%:10% respectively, and validation set is used to control number of iterations using early stop criteria.
The Choice of classes went from coarser to finer of three to five classes (Cranium, Spine, Upper extremity, chest, Lower extremity) from the anatomical point of view, discarding the classes with low sample density and high subclasses variability.

B. Preprocessing

All images are down-sampled to 128x 128 pixels. In addition, Normalization is used through subtracting the mean image, calculated across training set, from all training/validation/testing sets. In addition the Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to improve the performance of the system.

C. Network Training and Configuration

The training of the network is accomplished by optimizing a multinomial logistic regression function using minibatch (of size 256), gradient descent with momentum (set to 0.9) based on backpropagation introduced by LeCun et al. [14]. The training is then regularized with L_2 penalty set to $5 \, x \, 10^{-4}$. Drop out regularization layers are introduced in some networks configurations after the first two fully connected layers, drop ratio set to 0.5. The CNN was trained for 1000-2000 epoch using a NVIDIA tesla k80 GPU/NVIDIA GTX 960M GPU.

Several network configurations were employed in this study to studies the performance of the network, as presented in Table 1 and Table 2, one in each column, the nets will be referred to by their names (A-G). All the configurations follows the same design principle, with the same hyper-parameters such as the learning rate with momentum, the loss function, convolution stride etc. …

V. RESULTS AND DISCUSSION

A. Classification Platform

The performance of the different architectures under study is evaluated in terms of the classification accuracy defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

Where TP = true positive, FN = false negative, FP = false positive, and TN = true negative

The results obtained from the different networks configurations was compared to handcrafted based platforms relied on the texture, color and shape based features. As illustrated in table 6, the basic CNN configurations can outperform handcrafted features easily on the classification task in cbir system [9].

TABLE 1. Network configuration using (3 x 3) filters

A	B	C	D
Conv3-20	Conv3-20 **Conv3-20**	Conv3-20 Conv3-20	Conv3-20 Conv3-20
Maxpool			
Conv3-20	Conv3-20 **Conv3-20**	Conv3-20 Conv3-20	Conv3-20 Conv3-20
Maxpool			
Conv3-20	Conv3-20 **Conv3-20**	Conv3-20 Conv3-20 **Conv3-20**	Conv3-20 Conv3-20
Maxpool			
-	-	-	**Conv3-20** **Conv3-20**
-	-	-	**Maxpool**
FC-4500			FC-1280
FC-classes(5)			
Softmax			

TABLE 3. Network configuration using (5 x 5) conv

E	F	G
Conv5-20	Conv5-20	Conv5-20
-	**Maxpool**	Maxpool
Conv5-20	Conv5-20	Conv5-20
Maxpool		
Conv5-20	Conv5-20	Conv5-20
Maxpool		
-	-	**Conv5-20**
-	-	**Maxpool**
FC-5120		FC-1280
FC-800		
FC-classes(3,4)		
Softmax		

I. Effect of regularization

To prevent the neural network from overfitting, dropout (dpout) [25] is used in the first two fully-connected layers, where the output of each hidden neuron is set to zero with probability 0.5. In this way, the neural network will sample a different architecture when an input image is presented, benefiting the generalization of the neural network. Table 5 demonstrated the effect of dropout technique on the accuracy of the architectures.

II. Effect of Data Augmentation

Data augmentation is another technique that we use to prevent overfitting, by boosting the size of the training set so that the model cannot memorize all of it. Without the need to increase the dataset size. This can take several forms depending on the dataset. For instance, if the objects are supposed to be invariant to rotation such as x-ray images of upper and lower extremity, it is well suited to apply different kinds of rotations to the original images in order increase the model capacity to perceive variance in the images. The augmentation techniques used in this paper are randomly flipping the images with 90,180 and 270 degrees and random cropping. It is evident that the use of the regularization boosts the performance of the CNN by increasing the generalization ability for better classification results. Table 6 represents the effect adding the drop out layer and the usage of Data augmentation on the classification accuracy.

TABLE 4. Performance comparison of proposed architectures and handcrafted features

Feature	Accuracy [%]
Fourier Mellin	46.9
Gabor histogram	75.6
Tamura histogram	80.7
Color histogram	82.1
Local features	87
Network A	43.1
Network B	92.03
Network C	92.69
Network D	89.28
Network E	88.38
Network F	89.15
Network G	90.38

C. End to End CBIR

After training several networks, best performing models are used to test the effectiveness of the representations learned in an end to end cbir system. Precision and recall are chosen to evaluate system performance. Precision is the fraction of retrieved instances that are relevant to the query, while recall is fraction of the instances that are relevant to the query that are successfully retrieved, precision-recall curve of the end to end system is shown in figure 1.

VI. CONCLUSION

In this paper, we explored the performance of CNN to extract representative and discriminative features from x-ray images to be used in a cbir systems, different configurations for CNN with the study of effect of adding drop-out and data augmentation were tested to report the results, the best performing configuration demonstrated the ability to outperform hand crafted features. As a future work, we will focus on increasing the depth of the network, while using better data preparation and augmentation techniques and strong regularizes “for an example batch normalization” to avoid over fitting problems”

ACKNOWLEDGMENT

To Cherine, Ebtsam, Amal, Tabarak and Dr. Inas

REFERENCES

[1] C. Burak Akgül, D.L. Rubin, S. Napel, C.F. Beaulieu,H. Greenspan, and B. Acar, “Content-based image retrieval in radiology: current status and future directions,” *Journal of Digital Imaging*, vol. 24, no. 2, pp.208–222, 2011.

[2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *Pattern Analysis and Machine Intelligence,IEEE Transactions* on, vol. 22, no. 12, pp. 1349–1380,2000.

[3] Meenakshi Shruti Pal, and Dr Sushil Kumar Garg. "Image retrieval: A literature review." *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)* 2, no. 6 (2013): 2278-1323.

[4] Babaie, Morteza, Hamid R. Tizhoosh, Shujin Zhu, and M. E. Shiri. "Retrieving similar x-ray images from big image data

Figure 1. Precision-recall curve of the end to end cbir

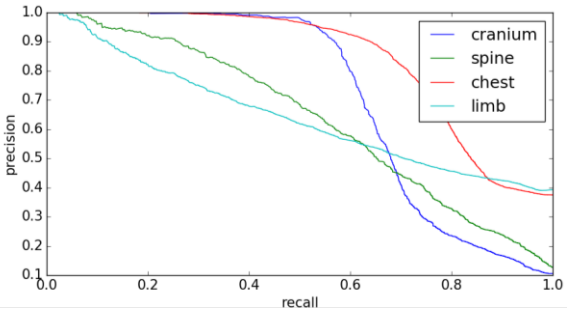

TABLE 5. Effect of dropout on network performance

Network	w/o dpout [%]	w/ dpout [%]
A	43.1	-
B	91.11	92.03
C	-	92.69
D	-	89.28
E	86.18	87.41
F	85.73	87.93

TABLE 6. Effect of augmentation on network performance

Network	w/o Aug [%]	w/ Aug [%]
A	43.1	-
B	91.11	-
C	-	-
D	-	-
E	86.18	87.54
F	85.73	87.09

Table 7. Effect of augmentation and dropout combined

Network	Accuracy [%]
E	88.38
F	89.15
G	90.38

using radon barcodes with single projections." *arXiv preprint arXiv:1701.00449* (2017).

[5] Andreas Veit, Michael Wilber, and Serge Belongie. "*Residual networks are exponential ensembles of relatively shallow networks*." arXiv preprint arXiv:1605.06431 1 (2016).

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097-1105. 2012.

[7] T.M. Lehmann, T. Deselaers, H. Schubert, M.O. Guld, C. Thies, B. Fischer, and K. Spitzer, "Irma – a content based approach to image retrieval in medical applications," *in IRMA International Conference*, 2006, vol. 5033, pp. 911–912.

[8] M.S. Meharban and Dr.S. Priya "A Review on Image Retrieval Techniques" *Bonfring International Journal of Advances in Image Processing*, Vol. 6, No. 2, April 2016.

[9] Thomas Deselaers, Daniel Keysers and Hermann Ney. "Features for image retrieval: A quantitative comparison." *Lecture Notes in Computer Science* (2004): 228-236.

[10]]C.-H. Wei, Y. Li, and C.-T. Li, "Effective extraction of gabor features for adaptive mammogram retrieval," *in Multimedia and Expo*, 2007 *IEEE International Conference on. IEEE*, 2007, pp. 1503–1506.

[11] T. Tommasi, F. Orabona, and B. Caputo, "Discriminative cue integration for medical image annotation*," Pattern Recognition Letters*, vol. 29, no. 15, pp. 1996–2002,2008.

[12] T. Tommasi, F. Orabona, and B. Caputo, "An svm confidence-based approach to medical image annotation," in *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 696–703. Springer,2009.

[13] T.M. Lehmann, M.O. G¨uld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B.B. Wein, "Automatic categorization of medical images for contentbased retrieval and data mining*," Computerized Medical Imaging and Graphics*, vol. 29, no. 2, pp. 143–155,2005.

[14] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] Sahiner Berkman, Heang-Ping Chan, Nicholas Petrick, Datong Wei, Mark A. Helvie, Dorit D. Adler, and Mitchell M. Goodsitt. "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images." *IEEE transactions on Medical Imaging* 15, no. 5 (1996): 598-610.

[17] ernazanu-Glavan, Cosmin, and Stefan Holban. "Segmentation of bone structure in X-ray images using convolutional neural network." *Adv. Electr. Comput. Eng* 13, no. 1 (2013): 87-94.

[18] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network." *In International conference on medical image computing and computer-assisted intervention*, pp. 246-253. Springer, Berlin, Heidelberg, 2013.

[19] Karen Simonyan, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition."*arXiv preprint* arXiv:1409.1556 (2014).

[20] T.M. Lehmann, T. Deselaers, H. Schubert, M.O. Guld, C. Thies, B. Fischer, and K. Spitzer, "The irma code for unique classification of medical images," *in SPIE Proceedings*. SPIE, 2003, vol. 5033, pp. 440–451.

[21] Daniel Keysers, Hermann Ney, Berthold B. Wein, and Thomas M. Lehmann. "Statistical framework for model-based image retrieval in medical applications." *Journal of Electronic Imaging* 12, no. 1 (2003): 59-68.

[22] Hossein Pourghassem, and Hassan Ghassemian. "Content-based medical image classification using a new hierarchical merging scheme." *Computerized Medical Imaging and Graphics* 32, no. 8 (2008): 651-661.

[23] Thomas M. Lehmann, Mark O. Güld, Thomas Deselaers, Daniel Keysers, Henning Schubert, Klaus Spitzer, Hermann Ney, and Berthold B. Wein. "Automatic categorization of medical images for content-based retrieval and data mining." *Computerized Medical Imaging and Graphics* 29, no. 2 (2005): 143-155.

[24] Md Mahmudur Rahman, Prabir Bhattacharya, and Bipin C. Desai. "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback." *IEEE transactions on Information Technology in Biomedicine* 11, no. 1 (2007): 58-69.

[25] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting*." Journal of Machine Learning Research* 15, no. 1 (2014): 1929-1958.