Toward Content-Based Image Retrieval with Deep Convolutional Neural Networks

Judah E.S. Sklan[a*], Andrew J. Plassard[a], Daniel Fabbri[b], Bennett A. Landman[a,c]
[a] Computer Science, Vanderbilt University, Nashville, TN, USA 37235
[b] Biomedical Informatics, Vanderbilt University, Nashville, TN, USA 37235
[c] Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

ABSTRACT

Content-based image retrieval (CBIR) offers the potential to identify similar case histories, understand rare disorders, and eventually, improve patient care. Recent advances in database capacity, algorithm efficiency, and deep Convolutional Neural Networks (dCNN), a machine learning technique, have enabled great CBIR success for general photographic images. Here, we investigate applying the leading ImageNet CBIR technique to clinically acquired medical images captured by the Vanderbilt Medical Center. Briefly, we (1) constructed a dCNN with four hidden layers, reducing dimensionality of an input scaled to 128x128 to an output encoded layer of 4x384, (2) trained the network using back-propagation 1 million random magnetic resonance (MR) and computed tomography (CT) images, (3) labeled an independent set of 2100 images, and (4) evaluated classifiers on the projection of the labeled images into manifold space. Quantitative results were disappointing (averaging a true positive rate of only 20%); however, the data suggest that improvements would be possible with more evenly distributed sampling across labels and potential re-grouping of label structures. This prelimainry effort at automated classification of medical images with ImageNet is promising, but shows that more work is needed beyond direct adaptation of existing techniques.

Keywords: deep convolutional neural networks, content based image retrieval, medical images, unsupervised learning

1. INTRODUCTION

Over the last 18 months, the Vanderbilt Medical Center has compiled over 100 million anonymized medical images. If this database could be mined with a Content Based Image Retrieval System (CBIR), researchers and clinicians could retrieve images visually similar to those of their patients, allowing for a comparison of patient histories and treatment profiles, eventually leading to improved patient care.

Although CBIR is not a new idea for the medical field, recent efforts have not been particularly effective [1-3]. Broadly speaking, there are two perpetual challenges for computer vision: the sensory gap and the semantic gap.[3] The sensory gap refers to the fundamental limitations of imaging the physical world with a limited capture device. The semantic gap arises between the captured pixels and the objects that humans recognize and name within the image. The goal of CBIR is to bridge the semantic gap and enable the computer to extrapolate high order information of the image's subject.[4] CBIR garnered a lot of attention in the mid-1990s. At that time, however, the field suffered from slow database management. [1] Since then database management efficiency has vastly improved, allowing for implementation on image sets in the millions. In 2012, at the ImageNet competition, the group from the University of Toronto used a Deep Convolutional Neural Network (dCNN) trained on over 1 million images to achieve results far better than the previous state of the art. [5] The goal of the manuscript is to evaluate the ImageNet framework on unsupervised feature learning and subsequent classification on clinically acquired magnetic resonance (MR) and computed tomography (CT) images.

2. METHODOLOGY

Two datasets were randomly selected from the previous 9 months of the Vanderbilt Medical Center's anonymized image archive of all MR and CT images acquired for clinical care. First 1 million image slices were selected at random from a library of ~50 million images. Second, a non-overlapping random sample of 2,100 images was

* judah.e.sklan@vanderbilt.edu; http://masi.vuse.vanderbilt.edu; Medical-image Analysis and Statistical Interpretation Laboratory, Department of Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging,
edited by Barjor Gimi, Robert C. Molthen, Proc. of SPIE Vol. 9417, 94172C

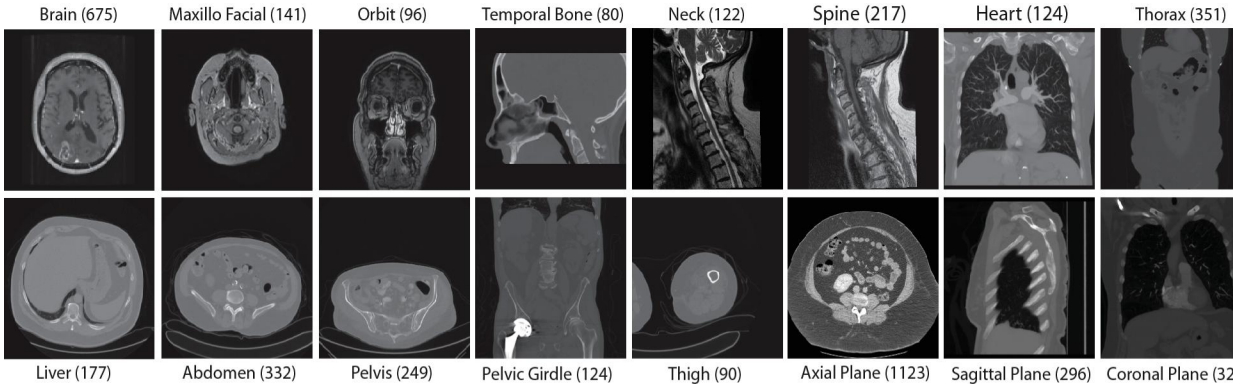

Figure 1 Illustration of the 16 non-mutually exclusive labels on normalized and resampled images. Name (#occurences)

retrieved. The images were normalized (Windsorized to the 5th and 95th percentile) and scaled using bicubic interpolation to 128x128 pixels.

2.1 Data labeling

In collaboration with a radiologist, the anatomy was divided for classification into 16 non-mutually exclusive sub-fields representing the human anatomy (Figure 1). A trained undergraduate rater then labeled the 2100 single slice images with a boolean for each of the 16 labels.

2.2 Manifold training

Using the 1 million unlabeled images, a manifold was trained with a series of autoencoders of increasing size (pylearn2 [10]) as illustrated in Figure 2. Each layer consisted of a square patch of the image and the corresponding channels of the image. The autoencoder was trained through back-propagation to minimize the mean square reconstruction (eq. 1) error between the input image patch and the learned image patch

$$MSE = \frac{1}{2}\sum_{i=1}^{P}(x_i - \hat{x}_i)^2 \quad (1)$$

where x_i is the input intensity value, $\hat{x}_i$ is the reconstructed intensity value, and P is the total number of points in the input. The first layer consisted of patches of size 13x13x1 with an autoencoder layer of size 36, the second was 5x5x36 with a hidden layer of size 96, the third was 5x5x96 with a hidden layer of size 256, the fourth and final was 3x3x256 with a hidden layer of size 384. To be clear, the first two dimensions describe the patch shape, and the third dimension describes the number of neurons in the subsequent hidden layer. Each layer of the manifold was trained with an independent set of 200,000 images with a total of 2,000,000 patches extracted from the images. Each layer was trained for 500 epochs.

2.3 Classifier training

First, multi-layer perceptrons (MLP) were trained to predict all 16 of label variables in binary classification problems (pylearn2 [10]). For each individual label 'a', a 10-fold cross validation was performed on a classifier that decided 'a' or '~a' for all samples in the 2100 image labeled set. As a

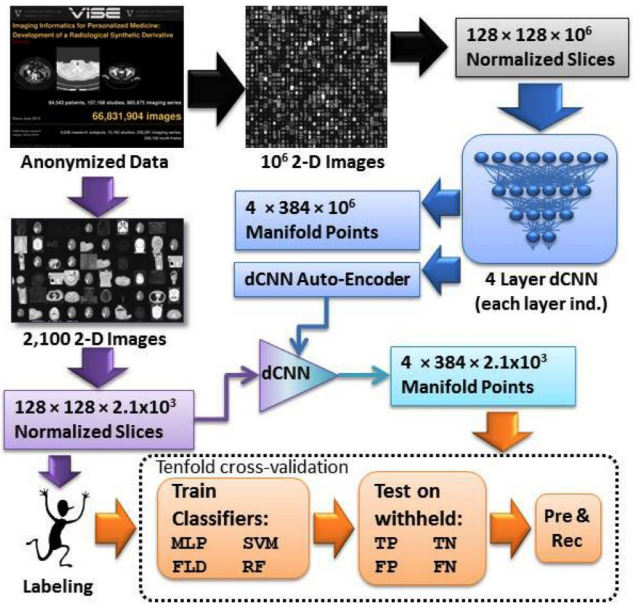

Figure 2 Flowchart of the ImageNet approach applied the the anonylized imaging data from clinical radiology.

secondary analysis, a reduced labeling scheme was created so that every image was a member of one and only one of the following five classes: {Head, Thorax, Abdomen, Pelvis, Legs}. This resulted in two reduced datasets: 'Axial Images Only' (350 total images, 70 per class), and 'All Images' (500 total images, 100 per class).

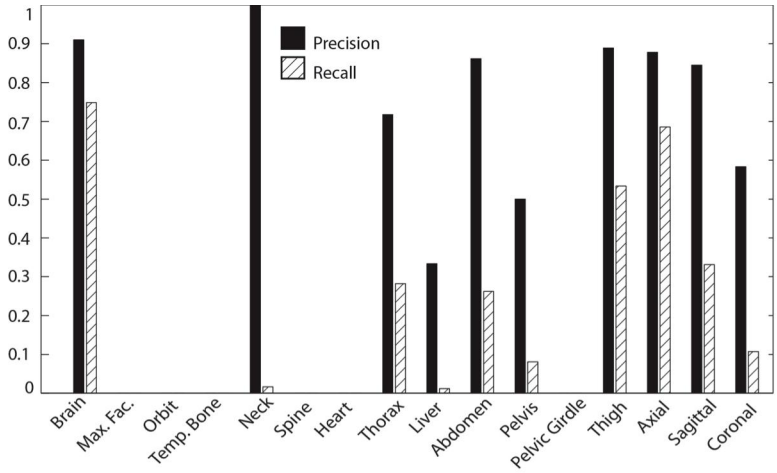

Figure 3 Precision was moderate (>0.7) for 7 of 16 labels, but recall was uniformly low.

Each of the five group classification tasks were evaluated with Fisher's linear discriminant analysis (FLDA), a support vector machine (SVM) with a radial basis kernel, and a random forest (RF) classier (Sci-kit Learn machine learning library [11]) The FLDA classifier was initialized without bound on the number of estimators. The SVM classifier had a kernel function of degree 3, a penalty of 1 for the error term, and a termination criterion tolerance of 0.001. The RF classifier had a bound of 100 estimators, determined quality of split on the Gini impurity, had no limit to depth, and used bootstrap samples to build the trees.

3. RESULTS AND DISCUSSION

Despite the promise of initial theory, overall results were disappointing. As can be seen in Figure 3, only 3 of the original 16 labels achieved a true positive rate greater than 50%. As can be seen in Figure 1, however, the distribution among labels is uneven, ranging from 80 to 1123 examples for a given label. This disparity might have had detrimental effects on the classifiers. Examining the two labels with the most samples (Axial Plane, Brain), we see that they are the top performing classes, with recall > 65% and precision > 85%. Figure 4 provides examples of each of three cases of main interest (false-positive, false-negative, true-positive) for each label that achieved observable results. Note that the false-positive for 'Neck' is N/A because there were no false positives for this label. It is also worth noting that the three most successfully classified structures (Brain, Thigh, and Axial Plane) have definite texture features. The brain has many curved lines, the thighs contain two grayish structure separated by black space whereas most contain only one, and axial scans are generally ovular and often contain an outline of the scanning table. It is possible that these obvious visual signifiers were also factors in the success of these three classifiers. Note especially that Thigh, despite its small number of samples (90) was able to achieve success comparable to Brain (675) and Axial Plane(1123), seeming to point toward unique texture as a factor of success. See Figure 1 for further comparison of class distribution.

By exploring the performance of alternative classifiers on the reduced datasets (Figure 5) we can infer likely performance of the complete system given more data. Results become volatile with a training set of size less than 10. However, as size of the training set increases, we see a dramatic increase in classifier performance. This shows that it is quite possible that given a greater number of labeled images, a classifier of might perform better. Note the substantively lower recall scores for the set of All Images versus Axial Only. The poor performance here could be the result of poor definition of classes, causing error in their reduction from a non-mutually exclusive label set to the reduced independent label set. Ensuring greater homogeneity of classes through synchronizing meta-data could help improve the label set. For example, one could possibly implement a hierarchical ontology.

4. CONCLUSION

Unlike the general photographs used in the ImageNet competition, there is very little variance of context for the subjects of medical images. All images are roughly gray ovals appearing on a black background. This is why the prevalent CBIR systems for medical images involved a coding system of metadata that must be input manually. [6] Unlike ImageNet, these systems have been limited in scale to a database size in the thousands or tens of thousands.

[7, 8] This tight bound on scope necessitated the combination of many techniques, often resulting in a complex framework. In dealing with clinically acquired images, some pre-regularization steps are often necessary to account for translations or rotations of regions on interest within the image. Towards simplifying this process, the use of a dCNN is promising because it is an unsupervised machine learning technique specifically designed to address these issues. [9] Future efforts at CBIR can make greater use of the existing automatically coded information in radiological databases.

Acknowledgments: This work was supported in part by the Robert J. Kleberg, Jr. and Helen c. Kleberg Foundation. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. The project described was supported by the National Center for Research Resources, Grant UL1 RR024975-01, and is now at the National Center for Advancing Translation Sciences, Grant 2 UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH,

REFERENCES

[1] Akgül, C.B., et al., *Content-based image retrieval in radiology: current status and future directions.* Journal of Digital Imaging, 2011. **24**(2): p. 208-222.

[2] Hsu, W., et al., *SPIRS: a Web-based image retrieval system for large biomedical databases.* International Journal of Medical Informatics, 2009. **78**: p. S13-S24.

[3] Müller, H., et al., *A review of content-based image retrieval systems in medical applications—clinical benefits and future directions.* International Journal of Medical Informatics, 2004. **73**(1): p. 1-23.

[4] Smeulders, A.W., et al., *Content-based image retrieval at the end of the early years.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000. **22**(12): p. 1349-1380.

[5] Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

[6] Lehmann, T.M., et al., *Automatic categorization of medical images for content-based retrieval and data mining.* Computerized Medical Imaging and Graphics, 2005. **29**(2): p. 143-155.

[7] Rahman, M.M., P. Bhattacharya, and B.C. Desai, *A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback.* Information Technology in Biomedicine, IEEE Transactions on, 2007. **11**(1): p. 58-69.

[8] Müller, H., et al., *Overview of the CLEF 2009 medical image retrieval track*, in *Multilingual Information Access Evaluation II. Multimedia Experiments* 2010, Springer. p. 72-84.

[9] Lee, H., et al. *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations*. in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009. ACM.

[10] Goodfellow, I.J., et al., *Pylearn2: a machine learning research library.* arXiv preprint arXiv:1308.4214, 2013.

[11] Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* The Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

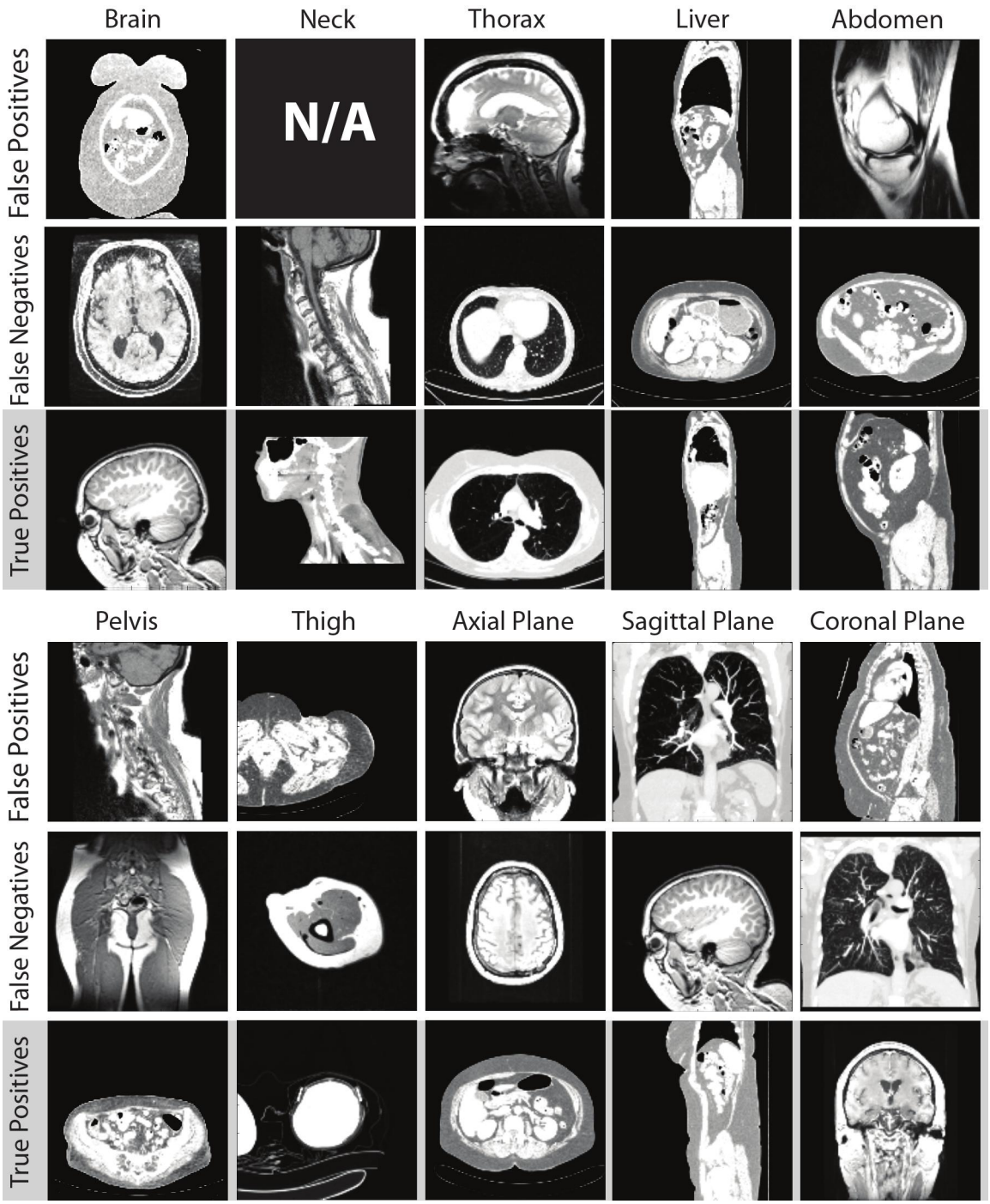

Figure 4 Representative examples of classifer failures for the 10 labels with non-null (true positive>0) classifiers.

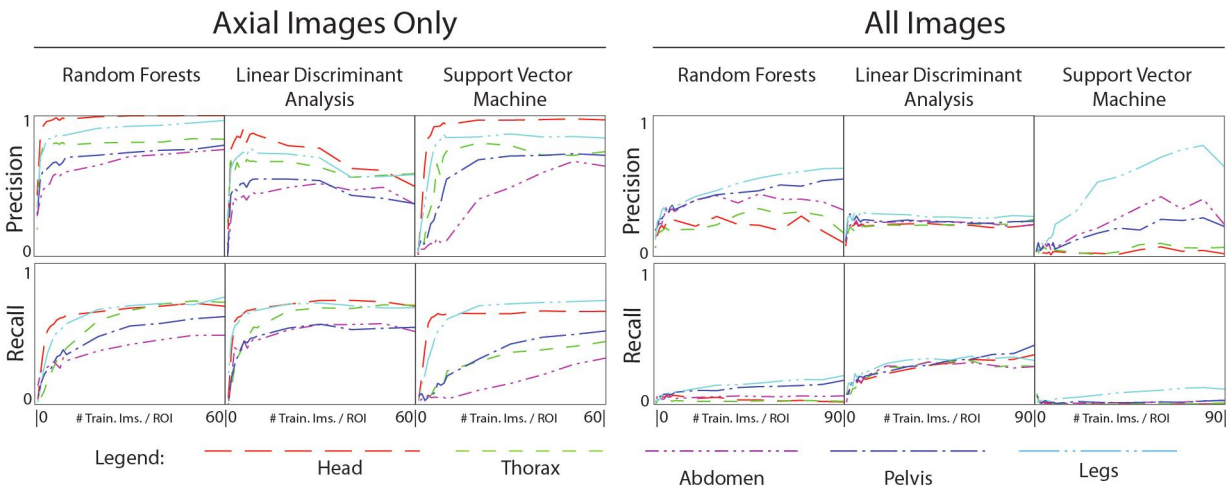

Figure 5 Results of image classification on the reduced label sets.