Transfer Representation Learning for Medical Image Analysis

Chuen-Kai Shie, Chung-Hisang Chuang, Chun-Nan Chou, Meng-Hsi Wu, and Edward Y. Chang

Abstract — **There are two major challenges to overcome when developing a classifier to perform automatic disease diagnosis. First, the amount of labeled medical data is typically very limited, and a classifier cannot be effectively trained to attain high disease-detection accuracy. Second, medical domain knowledge is required to identify representative features in data for detecting a target disease. Most computer scientists and statisticians do not have such domain knowledge. In this work, we show that employing *transfer learning* can remedy both problems. We use Otitis Media (OM) to conduct our case study. Instead of using domain knowledge to extract features from labeled OM images, we construct features based on a dataset entirely OM-irrelevant. More specifically, we first learn a codebook in an unsupervised way from 15 million images collected from *ImageNet*. The codebook gives us what the encoders consider being the fundamental elements of those 15 million images. We then encode OM images using the codebook and obtain a weighting vector for each OM image. Using the resulting weighting vectors as the feature vectors of the OM images, we employ a traditional supervised learning algorithm to train an OM classifier. The achieved detection accuracy is 88.5% (89.63% in sensitivity and 86.9% in specificity), markedly higher than all previous attempts, which relied on domain experts to help extract features.**

I. INTRODUCTION

Otitis media (OM) is any inflammation or infection of the middle ear and consumes significant medical resources each year [1] [2]. Several symptoms such as redness, bulging, and a perforation may suggest an OM condition. Color, geometric, and texture descriptors may help in recognizing these symptoms. However, specifying this kind of features involves a hand-crafted process, and thereby requires domain expertise. Often times, human heuristics obtained from domain experts may not be able to capture the most discriminative characteristics, and hence the extracted features cannot achieve high detection accuracy. Besides the problem of feature representation, developing a good disease-diagnosis classifier also faces the challenge of limited amount of labeled training data. Under such constraint, even an effective model like deep neural network cannot learn discriminative features. Inevitably, the lack of labeled data is a common issue for almost all medical analysis.

In this work, we tackle the twin problems of feature representation and training-data scarcity using transfer learning. We use OM to conduct our case study. Instead of directly identifying the most discriminative features from the 1,195 OM images collected by seven otolaryngologists at Cathay General Hospital, Taiwan [7], we performed representation learning on an entirely irrelevant image set: ImageNet [6], which is the largest image dataset (15 million images over 22,000 categories) of daily objects (e.g., animals, cars, and people). Our approach consists of four steps:

1. Unsupervised codebook construction. We learn a *codebook* from ImageNet images. We call our codebook construction unsupervised one in the sense that our construction method is unsupervised with respect to OM.
2. Encoding OM images using the codebook. Each OM image is encoded into a weighted combination of the pivots in the codebook. The weighting vector is the feature vector of the OM image.
3. Supervised learning. Using the transfer-learned feature vectors, we then employ supervised learning to learn a classifier from the 1,195 labeled OM instances.
4. Feature fusion. We also combine some heuristic features (published in [7]) with features learned via transfer learning, and show that further improvement can be achieved.

In summary, this work demonstrates that while the amount of labeled medical images for conducting statistical analysis is typically limited, the underline structures of the images can be learned and transferred from a large, though semantically unrelated, dataset. Not only is the lack of labeled data problem mitigated, but also the lack of domain knowledge to extract features can be remedied. Our scheme outperforms all prior attempts using human heuristics to extract discriminative features.

The rest of the paper is organized into four sections. Section II discusses related work. Section III details our transfer learning scheme. Section IV reports experimental results. Finally, in Section V, we offer our concluding remarks.

II. RELATED WORK

Biomedical image processing is an emerging research topic. Most approaches utilize hand-crafted feature extraction, which encodes human domain knowledge into computer algorithms to extract features. Several OM image-processing papers have proposed such features for OM detection. Mironică et al. [3] compared several color descriptors, and used color coherence histograms to recognize OM with an accuracy of 73.11%. Kuruvilla et al. [4] developed a feature named "vocabulary" to describe the symptoms, which doctors usually use to diagnose whether a patient is OM infected or not. The proposed tree structure "grammar" simulates the decision process used by the otolaryngologist, and achieved 89.9% in recognition accuracy. However, the algorithm only diagnoses acute OM and OM with effusion, but not the more complicated chronic OM cases. Our developed classifier covers all OM cases.

Deep learning in general and deep convolutional neural networks (CNN) [3] [4] in particular, is a composite model of neural networks. Deep learning enjoys good success since 2006, and is shown to achieve improvement in classifying images, audio, and speech data. All deep learning models require a substantial amount of training instances to avoid the problem of over-fitting. Some research works in the medical field have started employing a deep architecture [11] [12]. In this work, we do not use deep learning to directly training an OM classifier. Instead, we use the unsupervised layers (unsupervised with respect to OM) of CNN to learn the structures of non-OM-related images, and then use that learned representation to model OM data. In other words, we use CNN to perform transfer representation learning.

Transfer learning, defined as the ability of a system to recognize and apply knowledge and skills learned in previous tasks to a novel task, is on the rise in recent years and has been applied in several data mining tasks, e.g., [2] [8] [9] [10]. Our work transfers representations learned from one domain to detect OM, and we refer this approach as transfer representation learning.

III. METHODS

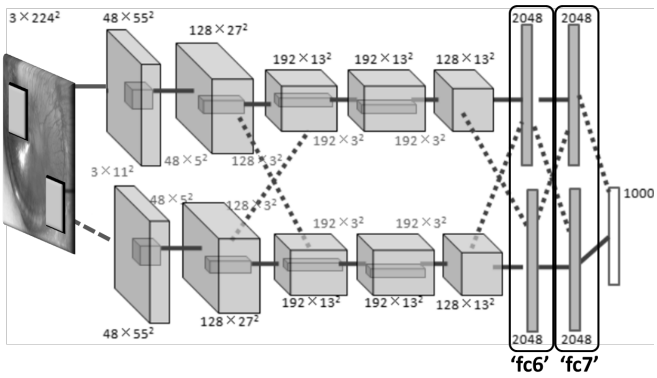
As depicted in the introductory section, our scheme is composed of four steps. This section details these steps.

We start with unsupervised codebook construction. On a large ImageNet dataset, we learn the representation of these images using a variant of deep CNN, Alexnet [4]. Alexnet, the winner of the world-wide image recognition competition (ILSVRC) in 2012, contains eight neural network layers. The first five are convolutional and the remaining three are fully-connected. Different layers represent different levels of abstraction concepts. We utilize Alexnet in Caffe [5] as our foundation to build our encoder to capture generic visual features.

For each image input, we obtain a feature vector using the codebook. The information of the image moves from the input layer to the output layer through the inner layers. Each layer is a weighted combination of the previous layer and stands for a feature representation of the input image. Since the computation is hierarchical, higher layers intuitively represent higher concepts. For images, the neurons from lower levels describe rudimental perceptual elements like edges and corners, while higher layers represent object parts such as contours and categories. To capture higher-level abstractions, we extract transfer-learned features of OM images from the fifth, sixth and seventh layer, denoted as pool5, fc6 and fc7 in Fig.1, respectively. The reason why we don't extract OM features from the eighth layer is that it only produces the probability of the class prediction, and it is not a representation of the input image.

Once we have transfer-learned feature vectors of the 1,195 collected OM images, we perform supervised learning by training a Support Vector Machine (SVM) classifier [14]. We choose SVM to be our model since it is an effective classifier widely used by prior work. As usual, we scale features to the same range and find parameters through cross validation. For comparing to the previous work fairly, we select radial basis function (RBF) kernel.

Figure 1. The flowchart of our transfer representation learning algorithm (otitis media photo is from [17])



To further improve classification accuracy, we experiment with two feature fusion schemes, which combine features hand-crafted by human heuristics in [7] with features learned from our codebook. In the first scheme, we combine transfer-learned and hand-crafted OM features to form fusion feature vectors. We then deploy the supervised learning upon the fused feature vectors to train a SVM classifier. In the second scheme, we use a two-layer classifier fusion structure mentioned in [7]. Concisely, in the first layer, we train different classifiers upon different feature sets separately. We then combine all the outputs from the first layer to train the classifier in the second layer.

Fig. 2 summarizes our transfer representation learning approaches. On the top of the figure depict two feature-learning schemes: the transfer-learned on the left-hand side and the hand-crafted on the right. The solid lines depict how OM features are extracted via the transfer-learned codebook, whereas the dash-dot lines represent the flow of hand-crafted OM feature extraction. The bottom half of the figure describes two fusion schemes. Whereas the dashed lines illustrate the feature fusion by concatenating two feature sets, the dotted lines show the second fusion scheme at the classifier level. At the bottom of the figure, the four classification flows yield their respective OM-prediction decision. From the left-hand side to the right-hand side, they are transfer-learned features only, feature-level fusion, classifier-level fusion, and hand-crafted features only.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We conducted two sets of experiments. Subsection Ⅳ-A reports OM classification performance by using our proposed transfer representation learning approach. The effect of transfer representation learning on improving OM classification is evaluated in Subsection Ⅳ-B.

A. Results of Transfer Representation Learning

Table 1 compares OM classification results for different feature representations. All experiments are conducted by using 10-fold SVM classification. The measures of results reflect the discrimination capability of the features.

The first two rows in Table 1 show the results of human-heuristic methods (hand-crafted), followed by our proposed transfer-learned approach. The eardrum

segmentation, denoted as 'seg', identifies the eardrum by removing OM-irrelevant information such as ear canal and earwax from the OM images [7]. The best accuracy achieved by using human-heuristic methods is around 80%. With segmentation (the first row), the accuracy improved 3% over that without segmentation (the second row).

Rows three to six show results of applying transfer representation learning. All results outperform the results in rows one and two, suggesting that the features learned from transfer learning are superior to the human-crafted ones. Interestingly, segmentation does not help improve accuracy for learning representation via transfer learning. This indicates that the transfer-learned feature set is not only more discriminative but also more robust. Among three transfer-learning layer choices (layer five (pool5), layer six (fc6) and layer seven (fc7)), fc6 yields slightly better prediction accuracy for OM. We believe that fc6 provides features that are more general or fundamental to transfer to a novel domain than pool5 and fc7 do.

The seventh row in Table 1 shows the result of applying deep learning directly to the 1,195 OM images to train a classifier. The resulting classification accuracy is only 71.8%, much lower than applying transfer representation learning. This result confirms our speculation that even CNN is a good model, with merely 1,195 OM images (without the ImageNet images to facilitate feature learning), it cannot learn discriminative features due to under-fitting.

Two fusion methods, combining both hand-crafted and transfer learning features, achieve slightly higher OM-prediction F1-score (0.9 over 0.895) than using transfer-learned features only. This statistically insignificant improvement suggests that hand-crafted features do not provide much help.

Figure 2. Four classification flows
(otitis media photos are from [19])

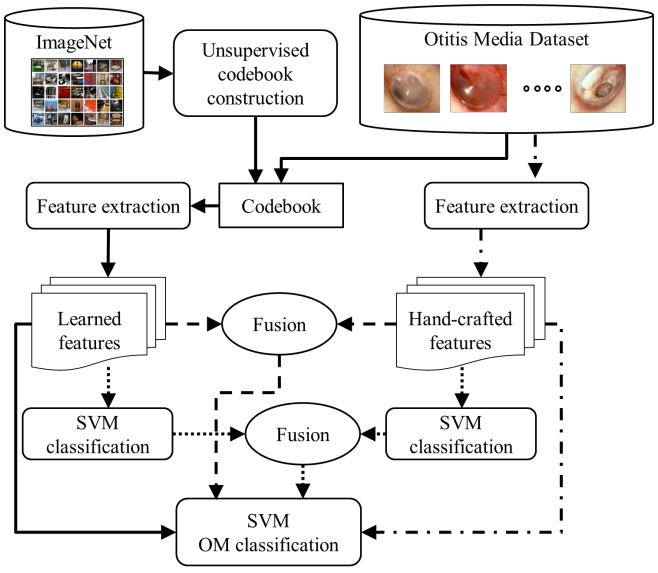


TABLE I. OM CLASSIFICATION EXPERIMENTAL RESULTS

Method	Measures			
	Accuracy	***Sensitivity***	***Specificity***	***F_1-score***
Heuristic w/ seg	80.11%	83.33%	75.66%	0.822
Heuristic w/o seg	76.19%	79.38%	71.74%	0.79
Transfer w/ seg (pool5)	87.86	89.72%	86.26%	0.89
Transfer w/o seg (pool5)	88.37%	89.16%	87.08%	0.894
Transfer w/ seg (fc6)	87.58%	89.33%	85.04%	0.887
Transfer w/o seg (fc6)	**88.5%**	89.63%	86.9%	0.895
Transfer w/ seg (fc7)	85.6%	87.5%	82.7%	0.869
Transfer w/o seg (fc7)	86.9%	88.5%	84.9%	0.879
OM-trained codebook	71.8%	95.06%	41.66%	0.818
Feature fusion	89.22%	90.08%	87.81%	0.90
Classifier fusion	**89.87%**	89.54%	90.2%	0.898

B. Codebook Performance Evaluation with Cifar-10

We are curious as to why transferring learning from one domain to a novel domain helps? In this subsection, we evaluate codebook performance by varying the training data quantity and diversity. Since running Alexnet on ImageNet takes weeks to complete one experiment, we used a simpler CNN architecture with Caffe on a smaller dataset, Cifar-10 dataset (10 classes and 5,000 images per class), to achieve our purpose.

In experiment (1), we retain the class number to 10 and assess the OM classification accuracy by increasing the training data quantity in each class. In Fig. 3, '1p-class' refers to one image per class, '2p-class' refers two images per class, etc. As we increase the image number of each class, the accuracy of the OM classification also increases. We can conclude that the more training data that transfer learning can use, the more discriminative features we can obtain to help the target classification task.

In experiment (2), we maintain the same total training data size and assess the OM classification performance by changing the diversity of the training data. Our hypothesis is that the number of classes suggests the diversity of the data, i.e., the more image classes in the training data the higher diversity we obtain from their perceptive features. Our experiment increases the number of classes from one to ten. To maintain the total number of training instances to be the same, we use 5,000 images when dealing with one class, and 2,500 images from each class when dealing with two classes, and so on and so forth.

Fig. 4 suggests that having more diversity in the training dataset can help improve feature quality. When the codebook is constructed by using only one class, the features derived from codebook are much less discriminative. The codebook can offer richer representation when there are at least two image classes in the training data (as shown in Fig. 4). To further verify our diversity hypothesis, we prepared two classes with similar perceptual features – sky and ocean. We used 2,500 images per class in this experiment. The accuracy drops down to 61.5%, which is significantly lower than the

previous accuracy (75%) reported in Fig. 4 of using images from two random classes. Based on this experiment, we conjecture that increasing the diversity of the training data can help the effectiveness of transfer representation learning for our purpose of classifying OM.

Figure 3. Relationship between the OM classification accuracy and the training data quantity

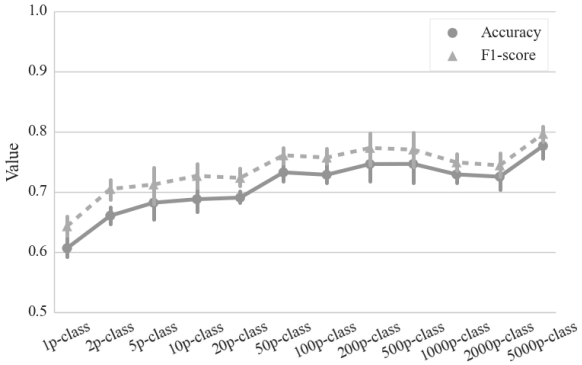
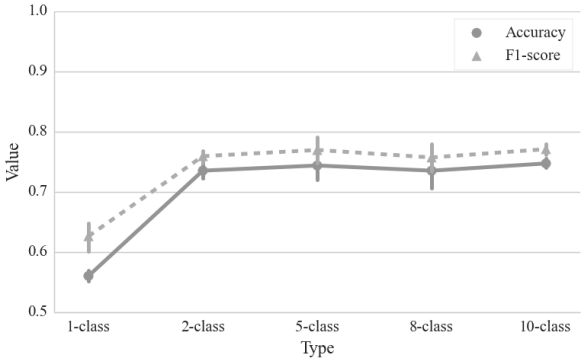


Figure 4. Result of diversity of pre-train data experiment



V. Conclusions and Future Work

In summary, this work demonstrates the transfer representation learning can remedy two major challenges of medical image analysis – labeled data scarcity and medical domain knowledge shortage. We constructed codebook based on a large OM-irrelevant dataset and obtained feature representation of OM images via transfer learning. We employed the traditional SVM as our classifier and achieved 88.5% OM detection accuracy (89.63% in sensitivity and 86.9% in specificity). Using transfer representation learning to analyze medical data is very promising, and we consider the potential to be immense. Our ongoing work is using time series data collected from speech and music to help analyzing ECG signals.

Another critical area of further research is improving speed of deep learning training. As we mentioned in the experiment of training data quantity and diversity, it took us weeks to train one cycle using the full set of ImageNet data. We have begun to use distributed CPU and multiple GPUs configured by fast interconnects to speed up our training time. Only under such speed improvement (as previous works speeding up other algorithms, e.g., [15][16][17][18]), we will be able to verify hypotheses quickly for many potential medicine applications.

References

[1] Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S.,. "CNN Features off-the-shelf: an Astounding Baseline for Recognition," In Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on (pp. 512-519). IEEE, (2014).

[2] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." Knowledge and Data Engineering, IEEE Transactions on 22.10 (2010): 1345-1359.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In ICML, 2014.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[5] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[7] Chuen-Kai Shie, Hao-Ting Chang, Fu-Cheng Fan, Chung-Jung Chen, Te-Yung Fang and Pa-Chun Wang, "A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media." In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.

[8] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. Technical Report HAL-00911179, INRIA, 2013.

[10] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arxiv:1311.2524 [cs.CV], 2013.

[11] Dan Claudiu Ciresan, Alessandro Giusti, Luca Maria Gambardella, Jürgen Schmidhuber, "Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks," MICCAI (2013).

[12] Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., and Nielsen, M., "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013 (pp. 246-253). Springer Berlin Heidelberg, (2013).

[13] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." Computer Science Department, University of Toronto, Tech. Rep 1.4 (2009): 7.

[14] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[15] PSVM: Parallelizing support vector machines on distributed computers, Edward Y. Chang, Kaihua Zhu, Hao Wang, Hongjie Bai, Jian Li, Zhihuan Qiu, Hang Cui, NIPS 2007.

[16] Foundations of Large-Scale Multimedia Information Management and Retrieval, Edward Y. Chang, Springer 2011.

[17] Parallel Spectral Clustering in Distributed Systems, Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, E Y. Chang, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2010.

[18] PLDA+, Zhiyuan Liu, Yuzhou Zhang, Edward Y Chang, Maosong Sun, ACM Transactions on Intelligent Systems and Technology (TIST), 2011.

[19] Gallery. http://codex.galleryproject.org/Main_Page