Artificial Intelligence in Medicine 79 (2017) 71–78

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Medical image classification via multiscale representation learning

Qiling Tang[a,*], Yangyang Liu[a], Haihua Liu[b]

[a] College of Biomedical Engineering, South Central University for Nationalities, Wuhan 430074, PR China
[b] Huibei Key Laboratory for Medical Information Analysis and Tumor Treatment, Wuhan 430074, PR China

ARTICLE INFO

Article history:
Received 24 February 2017
Received in revised form 19 May 2017
Accepted 20 June 2017

Keywords:
Multiscale feature learning
Sparse autoencoder
Fisher vector
Image classification

ABSTRACT

Multiscale structure is an essential attribute of natural images. Similarly, there exist scaling phenomena in medical images, and therefore a wide range of observation scales would be useful for medical imaging measurements. The present work proposes a multiscale representation learning method via sparse autoencoder networks to capture the intrinsic scales in medical images for the classification task. We obtain the multiscale feature detectors by the sparse autoencoders with different receptive field sizes, and then generate the feature maps by the convolution operation. This strategy can better characterize various size structures in medical imaging than single-scale version. Subsequently, Fisher vector technique is used to encode the extracted features to implement a fixed-length image representation, which provides more abundant information of high-order statistics and enhances the descriptiveness and discriminative ability of feature representation. We carry out experiments on the IRMA-2009 medical collection and the mammographic patch dataset. The extensive experimental results demonstrate that the proposed method have superior performance.

1. Introduction

Image classification plays an important role in many medical imaging tasks, including diseases diagnosis, medical references and surgical planning. For example, searching similar cases from a huge medical image database to determine the diagnosis and treatment options for a patient, a successful classification can reduce the search scope by omitting the irrelevant categories and thus improves the speed and precision of the retrieval.

Medical image classification usually suffers from high inter-class similarity and intra-class variability. Additionally, medical images commonly possess a low level of contrast together with a large amount of noise, which makes more difficult to distinguish images of different categories. The challenging task is supported by the ImageCLEF project [1] that provides an evaluation forum and framework for visual information analysis, classification, and retrieval.

Image classification techniques comprise two main stages: image feature representation and classification methods. In this work, we focus on the first stage that is, designing representative visual features for medical images. Since the classification performance is heavily dependent on the extracted features, selecting sufficient and appropriate features to characterize the specific properties of each class is a significant process in medical image classification.

Over the past decade, considerable efforts have been devoted to design effective feature representation for automatic classification of medical imaging. Most studies focus on: I) utilizing various feature descriptors, such as scale-invariant feature transform (SIFT) [2], local binary patterns (LBP) [3], etc., to extract features from medical images; II) feature encoding techniques such as sparse coding [4] and locality-constrained linear coding [5], usually incorporating bag of words framework or histogram representation; III) multiple feature combination methods by aggregating different feature information [6,7].

Recently unsupervised feature learning techniques have been successfully applied to the computer vision domain [8]. Compared with the previous hand-crafted features, the advantage of feature learning is that large amounts of unlabeled data can be fully utilized to capture good feature representations, and that it does not require much prior knowledge about the data, so that the method can be generalized to different cases without making significant modifications. Feature learning algorithms have been employed to achieve state-of-the-art results on a number of image and speech benchmark datasets [9,10].

Sparse autoencoder (SAE) network is one of the most popular representation learning methods. The work by Coates et al. [11]

* Corresponding author.
E-mail addresses: qltang@mail.scuec.edu.cn (Q. Tang), yyliu@sohu.com (Y. Liu), lhh@mail.scuec.edu.cn (H. Liu).

http://dx.doi.org/10.1016/j.artmed.2017.06.009
0933-3657/

has demonstrated that the single-layer network can achieve high performance with less parameters to tune and low computational cost if choosing appropriate parameters. The single-layer SAE as the feature extractor is widely applied to image classification tasks. A hierarchical convolutional SAE model was proposed in [12] for high spatial resolution imagery scene classification. A single-layer SAE with a mean pooling operation was utilized to learn feature representation from scene data sets [13]. Arevalo et al. [14] used the unsupervised feature learning framework to detect the basal cell carcinoma in histopathology images. Luo et al. [15] introduced the locality-constrained linear coding into the SAE for image classification, which produces similar codes for similar feature descriptors.

Multi-layer SAE can be stacked to construct deep networks, called stacked sparse autoencoder (SSAE), which are more capable of capturing abstract information at high layers of feature representations. SSAE framework presents excellent performance on medical applications, such as prostate MR image segmentation [16] and Alzheimer's Disease staging analysis [17].

Although much progress has been made on SAE-based feature learning in the past few years, there is little research about multiscale spatial feature representation. Multiscale structure is an essential attribute of natural images [18]. Similarly, there exist also scaling phenomena in medical images, and therefore a wide range of observation scales would be useful for image measurements. The multiscale processing has been proven to be significantly superior to the single-scale version in the vision tasks of boundary detection [19], image segmentation [20], object recognition [21], etc. In this work, we explore the application of the multiscale single-layer SAE networks to medical image classification, and our main contributions are twofold:

I) Multiscale method is introduced to the SAE framework, achieving complementary feature representation in scale space. We use the image patches sampled at different scales as input to train the single-layer SAEs, and thus obtain the feature detectors with different receptive field sizes. These learned feature detectors are assembled into a set of filter bank, which are used to convolve the source images to generate the feature maps at different scales. This proposal is more beneficial for characterizing various size structures in medical imaging than single-scale methods. To our knowledge, our work is the first attempt to combine multiscale feature learning on the single-layer SAEs.
II) Fisher vector (FV) technique is used to encode the extracted features, which provides more abundant information of high-order statistics (up to the second order) and allows a variable length feature set to be transformed into a fixed-length image representation. The FV computes the feature descriptors by the deviation from the Gaussian mixture model (GMM) distribution. Compared to the pooling methods (max-pooling and mean-pooling) adopted commonly in the above mentioned SAE models, the FV encoding greatly increases the feature dimension and enhances the descriptiveness and discriminative ability of feature representation. The combination of SAE networks and Fisher vector encoding can further improves the classification performance.

The rest of the paper is organized as follows. In Section 2, we describe the details of this proposed approach, including feature learning based on the multiscale SAEs, feature encoding with the FV and image classification by the multi-class SVM. Section 3 introduce the Image Retrieval in Medical Applications (IRMA) database and the evaluation metrics of classification performance, and conducts classification experiments on the IRMA database [22] to validate the effectiveness of our method. Finally, we conclude with a discussion in Section 4.

2. Method

In this section, we first utilize the multiscale SAEs to extract the features of images, and the features are then encoded into a FV representation with 128 Gaussian components for training Support Vector Machine (SVM) classifiers.

2.1. Multiscale feature learning

Feature representation plays a pivotal role in image classification task. Recently, unsupervised feature learning techniques have been successfully applied to a variety of domains. The feature learning can automatically extract features from raw unlabeled data, which is especially beneficial in medical applications because it is expensive to achieve a labeled medical dataset due to its large volume. In this work, a multiscale single-layer SAE network is employed to build sufficient and appropriate features from unlabeled medical images with low computational cost.

A single-layer SAE is a kind of self-learning neural network, consisting of only one hidden layer, in which sparse representation is achieved by constraining the average activation of each hidden node to a small value close to 0. The SAE is composed of two main modules: the encoder module and the decoder module. In the encoding stage, an input data $\boldsymbol{x} \in \mathbb{R}^n$ is mapped to a new feature representation $h(\boldsymbol{x}) \in \mathbb{R}^m$ (i.e. activation of the hidden nodes) by the following function,

$$h(\boldsymbol{x}) = f(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \tag{1}$$

where $f(\boldsymbol{z}) = 1/(1+\exp(-\boldsymbol{z}))$ is the nonlinear sigmoid activation function. $\boldsymbol{W}_1 \in \mathbb{R}^{m\times n}$ is a mapping matrix to be learned, and $\boldsymbol{b}_1 \in \mathbb{R}^m$ is a bias vector.

In the decoding stage, the hidden representation $h(\boldsymbol{x})$ is used to reconstruct the original input $\boldsymbol{x}$, and the reconstruction value $\tilde{\boldsymbol{x}}$ is computed by a linear activation function with mapping matrix $\boldsymbol{W}_2 \in \mathbb{R}^{n\times m}$ and bias $\boldsymbol{b}_2 \in \mathbb{R}^n$,

$$\tilde{\boldsymbol{x}} = \boldsymbol{W}_2 h(\boldsymbol{x}) + \boldsymbol{b}_2 \tag{2}$$

The weight matrices $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and bias vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ are computed by means of the back-propagation algorithm to minimize the reconstruction error. The cost function is given as follows,

$$J_{sparse} = \frac{1}{2N}\sum_{i=1}^{N}\|\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2 + \frac{\lambda}{2}\sum_{l=1}^{2}\|\boldsymbol{W}_l\|^2 + \beta\sum_{j=1}^{m}KL(\rho||\hat{\rho}_j) \tag{3}$$

where the first term is an average sum-of-squares error term which describes the reconstruction error between input x and its reconstruction $\tilde{\boldsymbol{x}}$ over the entire data, and N is the number of samples. The second term is a weight decay term which tends to decrease the magnitude of the weight, and helps prevent overfitting. The third term is sparsity penalty term which provides the sparsity constraint. $KL\left(\rho||\hat{\rho}_j\right)$ is the Kullback-Leibler divergence between the average activation $\hat{\rho}_j$ of hidden node j over the training set and the desired activation ρ defined as follows:

$$KL(\rho||\hat{\rho}_j) = \rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{4}$$

By setting a considerably small value to ρ, the model ensures that for a given input vector $\boldsymbol{x}$, only a small fraction of the hidden nodes are highly activated, while the majority of the activations of the hidden nodes are limited to values close to zero.

Patches of dimension w-by-w sampled from raw images are used as the SAE input, and w is referred to as the receptive field size. The SAE can effectively learn the feature representation through unlabeled image patches and has been successfully applied to many

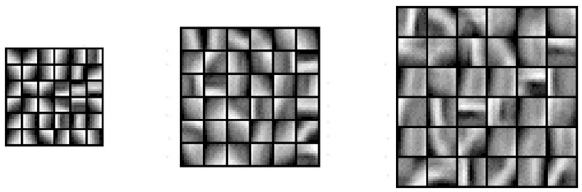

Fig. 1. For each receptive field size, randomly selected 36 features learned on medical images. Panels from left to right correspond to 6 pixel, 9 pixel, and 12 pixel receptive field.

vision tasks [12–17]. However, these works only use a single-scale receptive field, without taking into account the multiscale effect on the performance of the feature learning system. Image structures at different scales exhibit different characteristics, and the multiscale structure is an intrinsic attribute of images. Multiscale representation has extensively explored in boundary detection [19] and mid-level grouping [20].

We utilize the multiscale SAEs to learn features that then served as input to a linear SVM, improving classification performance. Three scales were employed in this work, with receptive field sizes of 6, 9, and 12 pixels, corresponding to the number of hidden nodes in the single-layer SAE to be 36, 81, and 144, respectively. We randomly select 200,000 sample positions from the IRMA medical image collection, and then obtain local patches of different sizes taken at every sample position for different scale receptive fields to train the multiscale SAEs. The learned features can be visualized according to the weights assigned to the connections in the network, as shown in Fig. 1. A joint representation is formed by combining multiscale features, which thus allows interactions between scales.

The convolution operation can capture effective features from source images, which has been well demonstrated in many works on convolutional neural networks [9,23]. It can be observed from Fig. 1 that the learned features over small patches sampled randomly from images are representatives of image structures at different orientations, scales, contrasts etc. These feature detectors that can be regarded as a set of filter bank are used to convolve the whole source image to produce new feature maps.

2.2. Fisher vector encoding

Images of different sizes will form feature maps of different lengths through the convolution operation. However, a fixed size input is usually required for training a classifier model. Fisher vector (FV) as an encoding method has been widely applied to image classification [24] and object recognition [25]. The FV with attractive properties can not only effectively characterize the local features of images, but also can produce a fixed length output regardless of the input size, which is very important to deal with variable length data for classification task.

The implementation of the FV encoding includes two main steps: First utilize a parametric generative model, e.g. the Gaussian Mixture Model (GMM), to fit the probability distribution of local features, and then construct the feature representation using the gradient vector with respect to the model parameters. Given a GMM with K modes, the set of the parameters is denoted by $\{\pi_k, \mu_k, \sigma_k\}_{k=1,\cdots K}$, where π_k, μ_k, and σ_k are respectively the mixture weight, mean vector and standard deviation vector of the k-th Gaussian. Assuming the GMM with diagonal covariance matrices, the representation vector is achieved by computing the average first and second order differences between the features and each of the GMM centers [24], which is described as follows:

$$\varphi_\mu^k = \frac{1}{T\sqrt{\pi_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right) \tag{5}$$

$$\varphi_\sigma^k = \frac{1}{T\sqrt{2\pi_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right) \tag{6}$$

where $\gamma_t(k)$ denotes the soft assignment of feature x_t to Gaussian k, defined as $\gamma_t(k) = \pi_k P_k(x_t) / \sum_l \pi_l P_l(x_t)$ in which $P_k(x_t)$ is the likelihood of Gaussian k. The FV encoding φ is obtained by concatenating the φ_μ^k and φ_σ^k, i.e.,$\boldsymbol{\varphi} = [\varphi_\mu^1, \varphi_\sigma^1, \cdots \varphi_\mu^K, \varphi_\sigma^K]$.

We use 128 Gaussian components in the GMM, and the dimensionality of the features obtained by the multiscale SAE is 261. The resulting FV is 66.8K-dimensional, which would lead to very high computation cost and memory requirement. To address this issue, the principal component analysis procedure (PCA) is performed to reduce the data dimensionality. The FV performance is further improved by adding signed square-rooting operation and L_2 normalization [26].

The image feature dimension is greatly improved after the FV encoding, which is useful and helpful for mining the more abundant information from image characteristics, including 1st-order and 2nd-order statistics (i.e., expectation and variance information), and form a fixed-length image representation. The FV is efficient to compute, and presents excellent results even with efficient linear classifiers [27].

2.3. Image classification

The unsupervised feature learning framework directly learns feature representation from raw data through an optimization process, which can explain better the content of the data. We employ the SAEs with various receptive field sizes to capture a set of multiscale feature detectors. An input image is convolved with the feature detectors, and each pixel produces a 261-element response. The convolved feature maps are then encoded by Fisher vector method. Finally, the FV encoding is fed to a discriminative classifier. As an illustration of our scheme, the flow chart is shown in Fig. 2.

In the experiments, the multi-class SVM proposed by Crammer and Singer [28] is used as the supervised learner. Given a training set with labels $S = \{(\varphi_1, y_1), \cdots, (\varphi_i, y_i)\}$ where φ_i is feature representation obtained by the FV encoding, and $y_i \in \{1, ..., N\}$ is the corresponding label and N is the number of image classes. For optimization in training, the objective function of the multi-class SVM is defined by,

$$\min_{\{\boldsymbol{\omega}_n\}} \sum_{n=1}^{N} ||\boldsymbol{\omega}_n||^2 + C \sum_i \max(0, 1 + \boldsymbol{\omega}_{r_i}^T \boldsymbol{\varphi}_i - \boldsymbol{\omega}_{y_i}^T \boldsymbol{\varphi}_i) \tag{7}$$

where $r_i = \arg\max_{n \in \{1,\cdots,N\}, n \neq y_i} \boldsymbol{\omega}_n^T \boldsymbol{\varphi}_i$. The first term in the formulation is a regularization term, and the second term is a sum of ranking hinge losses. The parameter C controls the tradeoff between the hinge loss and regularization terms. For fast computation, we utilize Liblinear with a linear kernel [29] to solve the SVM optimization problem. In the testing stage, the predicted label $\hat{y}$ can be obtained by,

$$\hat{y} = \arg\max_n \boldsymbol{\omega}_n^T \boldsymbol{\varphi} \tag{8}$$

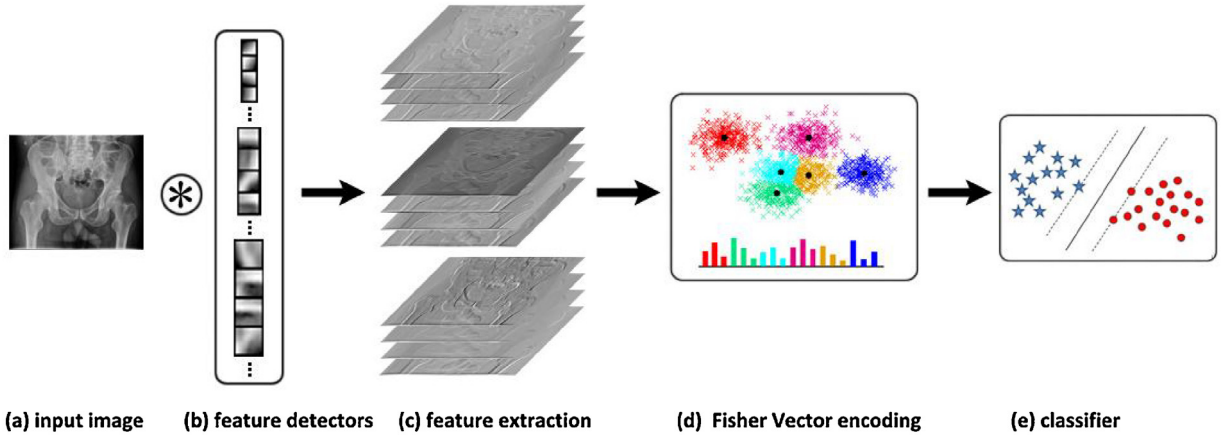

Fig. 2. The flow chart describes the process of our approach on image classification task.

3. Experiments

In this section, we verify the proposed method on publicly available medical image collections. First, we first present the benchmark data and the evaluation metrics used to assess the classification performance. Then, we describe the details of experiments and report the performance comparison with other methods.

3.1. IRMA database and evaluation metrics

We evaluate the proposed approach on the IRMA-2009 medical collection which contains 14410 radiographs taken randomly from daily routine work at the Department of Diagnostic Radiology of the RWTH Aachen University [22]. This collection offers 12677 images for training and 1733 images for testing, and these images were scaled to fit into a 512×512 maximum pixel window keeping the original aspect ratio. All images in the archive are annotated according to the IRMA coding system, which consists of four mono-hierarchical axes describing the imaging modality, the body orientation, the body region examined and the biological system examined. The IRMA-2009 collection images are subdivided into 193 distinct categories, and each image can be identified its category by the associated code with regard to the mono-hierarchical coding scheme. Some classes have large intra-variability and inter-similarity, making the IRMA classification task much difficult. Sample images from the IRMA-2009 are shown in Fig. 3. More information on the IRMA database and code can be found in [22].

Both the error score proposed by the ImageCLEF campaign [30] and classification accuracy are taken as the evaluation criteria in order to measure the performance of category recognition results in this work. Since in IRMA dataset the four IRMA code axes (the technical, directional, anatomical and biological axes) is separate, the error score E is computed by adopting a hierarchical approach according to the IRMA axes, which is expressed as follows:

$$E = \sum_{i=1}^{I} \frac{1}{b_i} \frac{1}{i} \delta(l_i, \hat{l}_i) \tag{9}$$

$$\text{with} \quad \delta(l_i, \hat{l}_i) = \begin{cases} 0 & if l_j = \hat{l}_j \quad \forall j \le i \\ 0.5 & if l_j = * \quad \exists j \le i \\ 1 & if l_j \neq \hat{l}_j \quad \exists j \le i \end{cases}$$

where b_i is the number of possible labels at position i, and I are the depth of the hierarchy. For every axis, the maximal possible error is calculated and the errors are normalized to a range of 0–0.25. So a completely wrong decision for that axis gets an error count of 0.25 and a completely correctly predicted axis has an error of 0.00. If all positions in all axes are wrong, error value is 1. Furthermore, this measure suggests that the decisions at an early stage in the code are more important than those at a later stage.

The classification accuracy is a very common and widely used evaluation measure, which is the percentage of the number of correctly classified images over total number of testing images.

3.2. Classification results on the IRMA-2009

To verify the benefits of multiscale feature representations on classification tasks, we test the proposed approach on the IRMA-2009 medical database. In order to find an appropriate number of Gaussians for the FV encoding, we first empirically analyze the impact of the number of Gaussian components in GMM on the classification performance in the different scale processing, as shown in Fig. 4. As can be seen, the classification performance generally improves as the number of Gaussian components increases and the performance reaches a plateau when Gaussians are more than 128. This count is a good choice as larger values only contribute a slight improvement, but at the expense of higher computational complexity.

Meanwhile, we also assess the effect of the different scale features captured by the SAEs of different receptive field sizes. The experimental results in Fig. 4 show that the 9 pixel receptive field works best and the 6 pixel is worst on a single scale. The experiment suggests that medium scale receptive field work better than too small or too large receptive fields, which is different from the conclusion of small receptive field being better in [11]. There may be two reasons: 1) the data source processed is not the same (natural images from the CIFAR and NORB datasets in [11]); 2) feature calculation method is different (Coates et al. [11] adopted the sliding window approach followed by pooling operation, whereas we used the convolution operator followed by the FV encoding). Finally, we combine signals from multiple scales in a joint representation, which produces a richer description for feature representation and further improves the classification performance over the single-scale SAEs.

Fig. 5 shows the mean and standard deviation of the classification accuracy for each category obtained by various single-scale and multiscale SAEs. It can be observed that the multiscale method outperforms the single-scale SAEs on the statistical average accuracy for each category and, although the average accuracy of the SAE

Fig. 3. Examples from the IRMA-2009 database.

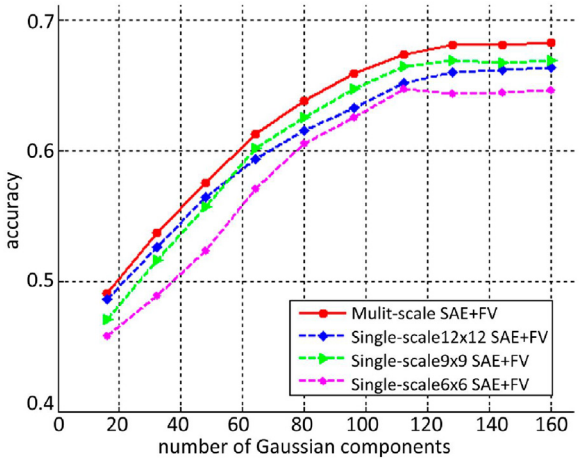

Fig. 4. Effect of the number of Gaussian components on classification performance.

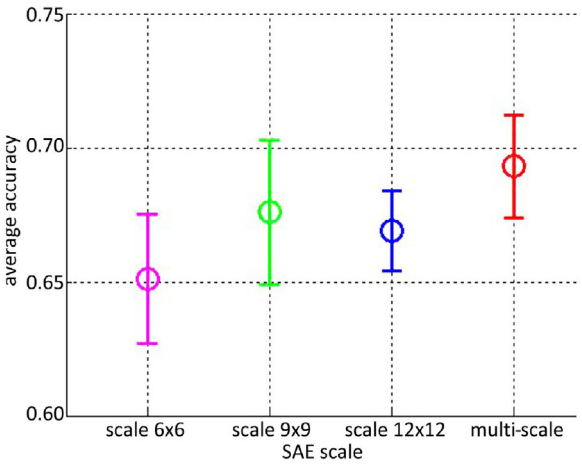

Fig. 5. The average of accuracy for each category obtained with different scale SAEs.

with 12 pixel receptive field is not the highest, its standard deviation is the smallest compared with the others, which seems to suggest that large receptive field has better robustness to different classes.

Various classification models based on SAE feature representations have been investigated in recent years. We compared our approach to some related works, including the hierarchical convolutional SAE with max-pooling strategy [12], the single-layer SAE with mean-pooling operation [13], the SAE framework integrated with topographic independent component analysis (TICA) and bag of features (BOF) [14], and the SAE network with the locality-constrained linear coding (LLC) [15], as shown in Fig. 6. These approaches all applied the SAE feature learning to image classification tasks. Different methods are tested on different databases. For comparisons, we repeat these methods on IRMA database and compare their performance in terms of the classification accuracy and the error score. The experimental results show that the proposed scheme can capture complementary information in the scale-space and can provide better performance compared to the single-scale versions and other SAE-based classification methods.

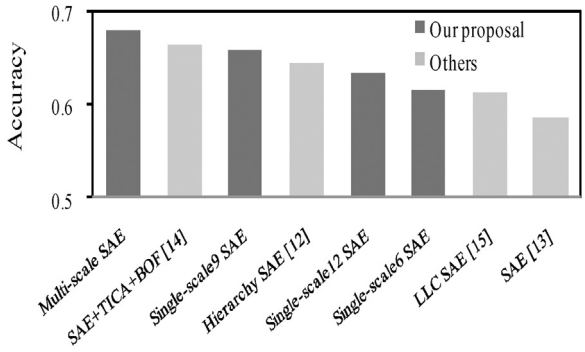

Fig. 6. Accuracy of different classification models based on SAE feature representations on the IRMA-2009 medical database.

We further evaluate the classification performance of our proposal with the error score defined in Eq. (9), and present more compared results, as shown in Table 1. According to the error score, the multiscale approach outperforms most reported results on the IRMA-2009 Benchmark, and is highly competitive with the one with the lowest error score reported as 146.55 in [3]. These experiments demonstrate that multiscale SAE is useful for describing the multiscale nature of image structures. Note that the images in [3] are required a preprocessing procedure of detecting salient regions, and then the data are folded to reduce the effect of irrelevant regions. Thus their classification performance is highly dependent on the detection results of salient regions, while our method does not require the saliency-based folded data.

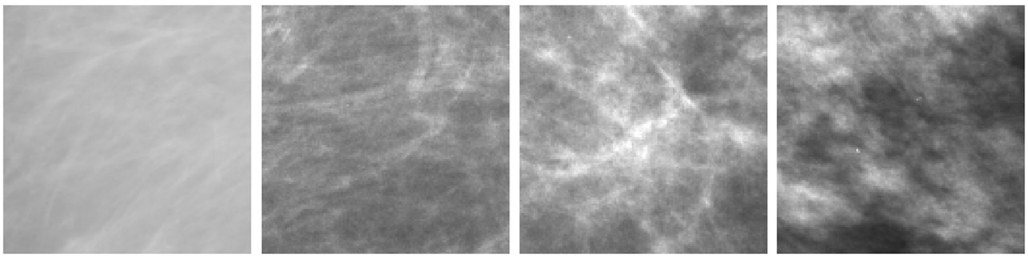

Fig. 7. Four classes of BI-RADS breast tissue density. The breast density increases from left to right, corresponding to BI-RADS I, II, III, and IV.

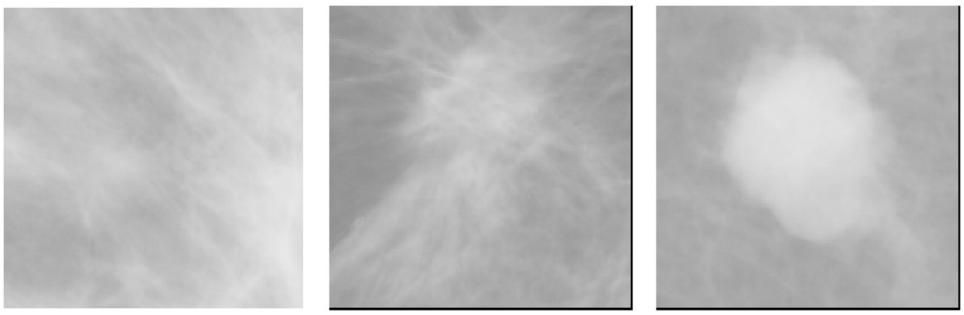

Fig. 8. Panels from left to right correspond to BI-RADS lesion assessment categories: 1. no finding, 2. benign, 5. malignant, respectively.

Table 1
Quantitative comparison of various methods on the IRMA-2009 benchmark with the error score. A smaller error score indicates a better classification performance.

	Method	Error score
SAE based methods:	**Multi-scale SAE + FV/SVM**	**148.76**
	Single-scale$^{9\times9}$ SAE + FV/SVM	**154.53**
	Single-scale$^{12\times12}$ SAE + FV/SVM	**158.11**
	Single-scale$^{6\times6}$ SAE + FV/SVM	**164.28**
	SAE + TICA + BOF/Softmax [14]	153.49
	Hierarchy SAE + Max-pooling/Softmax [12]	156.26
	LLC SAE + Max-pooling/SVM [15]	162.57
	SAE + Mean-pooling/SVM [13]	179.73
Other methods:	MS LBP/SVM [3]	146.55
	LBP + EHD + SIFT/RF [31]	153.21
	TAUbiomed [32]	169.50
	VPASabanci [33]	261.16
	Radon Barcode/SVM [34]	294.83

The lower part of this table lists some results reported in the literature.

3.3. Classification on the mammographic dataset

The proposed approach is also applied to classify the mammographic images. Mammographic patch dataset (*12er-patches*) from the IRMA project is used for experimental test. According to metadata annotations, the patches centering the lesion are extracted from screening mammographies of different BI-RADS (Breast Imaging Reporting Data System) classes. These patches are divided into four classes in terms of the BI-RADS breast tissue density, corresponding to I. almost entirely fatty, II. scattered fibro glandular tissue, III. heterogeneously dense tissue, and IV. extremely dense tissue [36], as shown in Fig. 7. For each of the four tissue density classes, according to BI-RADS lesion assessment, suspicious regions are labeled as three categories of pathology: 1. no finding (negative), 2. benign, and 5. highly suggested malignant. Fig. 8 exemplifies the resulting patches. The dataset of *12er-patches* contains 2796 patches of 12 classes, each with 233 images.

The multiscale SAE approach is utilized to describe breast density and lesion characteristics in mammographic patch dataset and the classification accuracy is assessed using ten-fold cross validation of all the patches. Three other criteria (i.e., precision, recall, and F-measure) are utilized to evaluate the performance on classifying mammographic images. Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. These two metrics are calculated as $Precision = TP/(TP + FP)$ and $Recall = TP/(TP + FN)$, where TP, FP, and FN are the number of true positives, false positives, and false negatives respectively [35]. The F-measure is defined as the harmonic mean of precision and recall, i.e., $F\text{-}measure = 2 \times Precision \times Recall/(Precision + Recall)$.

Table 2 shows the confusion matrix of the 12 class experiments and the precision, recall, and F-measure of each class obtained by the multiscale SAE method. In the confusion matrix the rows correspond to the true (prelabeled) classes and the columns correspond to the predicted classes. Investigation of the misclassified patches reveals that classes within the same category of pathology are more difficult to distinguish from each other than those from different categories, and furthermore, categories of more translucent tissue seem to be easier to identify, which are consistent with the experimental findings in [38].

We compare the overall classification accuracy of the proposed multiscale SAE with other SAE-based methods and with the existing techniques on classifying the mammographic patch dataset, including PCA [37], 2DPCA [38], Wavelet and Haralick (WA-HA) descriptors [39]. Table 3 presents the classification accuracy of different methods on the *12er-patches* dataset. Visually, breast tissues of various densities with or without lesions are distinguished through their texture patterns in mammographies. Experimental results reveal that the proposal of multiscale SAE with FV encoding can achieve competitive performance on the classification of mammographic texture patterns.

4. Conclusion

The present study proposed a multiscale representation learning method to classify medical X-ray images. In our scheme the sparse autoencoders with different receptive field sizes are utilized to learn multiscale feature detectors directly from raw medical data, and generate feature maps by the convolution operation. Fisher vector technique is then used to encode the extracted fea-

Table 2
Confusion matrix of 12 classes, and the precision, recall, F-measure per class.

Predicted TRUE	I-1	I-2	I-5	II-1	II-2	II-5	III-1	III-2	III-5	IV-1	IV-2	IV-5	Precision	Recall	F-measure
I-1	184	12		5	7		13	2		10			78.30%	79.00%	0.786
I-2	4	181			6	9	5	10		7	11		74.50%	77.70%	0.761
I-5	2	5	171		5	12	8		10		7	13	80.30%	73.40%	0.767
II-1	12	1		178	8	4	7	3		11		9	83.20%	76.40%	0.797
II-2		6		3	175	2	14	12		6	15		72.30%	75.10%	0.737
II-5		7	14	4	7	164			20		8	9	71.60%	70.40%	0.71
III-1	13	5		10		4	148	17	4	22	10		63.50%	63.50%	0.635
III-2	8	11		3	9		14	152	7	10	17	2	68.80%	65.20%	0.67
III-5		4	9		1	21	2	12	162		11	11	74.70%	69.50%	0.72
IV-1	12	5		2	11		16			164	21	2	67.50%	70.40%	0.689
IV-2		6	5		13	3	5	8		10	155	28	56.40%	66.50%	0.61
IV-5			14	9		10	1	5	14	3	20	157	68.00%	67.40%	0.677

Rows: true (pre-labed) classes; Columns: predicted classes by our method.

Table 3
Classification accuracy of various methods on the *12er-patches* dataset.

Computational methods	Accuracy
Proposed method	**71.2%**
SAE + TICA + BOF [14]	69.4%
Hierarchy SAE + Max-pooling [12]	68.7%
SAE + Mean-pooling [13]	66.5%
WA-HA [39]	66.0%
2DPCA [38]	61.6%
PCA [37]	53.7%

tures and achieve a fixed-length image representation that is fed to the multi-class SVM classifier.

Extensive experiments are performed on the IRMA medical database to verify the effectiveness of our approach. We explored the effects of different single-scale and multi-scale combination detectors on the classification tasks, and compared the results to those obtained by other SAE-based models and the existing medical image classification techniques. Experimental results show that multiscale combination is helpful for capturing the intrinsic scales of image structures, leading to superior performance over single-scale versions, and that our method of combining SAE networks and Fisher vector encoding outperformed other SAE-based models.

In future work, we intend to extend the multiscale shallow representation to the deeper architecture (i.e., the stacked sparse autoencoder) to learn high-level feature representation. We are also interested in applying the multiscale strategy on the multi-modality medical imaging.

Acknowledgment

The main image dataset used in this study is courtesy of the IRMA Group, Aachen, Germany, http://irma-project.org.

References

[1] Müller H, Clough P, Deselaers T, Caputo B. ImageCLEF: experimental evaluation in visual information retrieval. Berlin, Heidelberg: Springer; 2010.
[2] Caicedo JC, Cruz A, Gonzalez FA. Histopathology image classification using bag of features and kernel functions. In: Proc. 12th Conf. Artificial Intelligence in Medicine. 2009. p. 126–35.
[3] Camlica Z, Tizhoosh HR, Khalvatid F. Medical image classification via SVM using LBP features from saliency-based folded data. Proc. IEEE Int. Conf. Mach. Learn. Appl 2015:128–32.
[4] Weiss N, Rueckert D, Rao A. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. Proc. MICCAI 2013:735–42.
[5] Zhang P, Wee CY, Niethammer M, Shen D, Yap PT. Large deformation image classification using generalized locality-constrained linear coding. Med Image Comput Assist Interv 2013;16(1):292–9.
[6] Alberdi A, Aztiria A, Basarab A. On the early diagnosis of Alzheimer's Disease from multimodal signals: a survey. Artif Intell Med 2016;71:1–29.
[7] Tabesh A, Teverovskiy M, Pang H-Y, Kumar VP, Verbel D, Kotsianti A, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE Trans Med Imaging 2007;26(10):1366–78.
[8] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798–828.
[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proc. Adv. Neural Inf. Process Syst 2012:1097–105.
[10] Zeiler MD, Ranzato MA, Monga R, et al. On rectified linear units for speech processing. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing 2013:3517–21.
[11] Coates A, Ng AY, Lee H. An analysis of single-layer networks in unsupervised feature learning. J Mach Learn Res 2011;15:215–23.
[12] Han X, Zhong Y, Zhao B, Zhang L. Unsupervised hierarchical convolutional sparse auto-encoder for high spatial resolution imagery scene classification. Proc. Int. Conf. Natural Computation 2015:42–6.
[13] Yin H, Jiao X, Chai Y, Fang B. Scene classification based on single-layer SAE and SVM. Expert Syst Appl 2015;42:3368–80.
[14] Arevalo J, Cruz-Roa A, Arias V, Romero E, González FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med 2015;64:131–45.
[15] Luo W, Yang J, Xu W, Fu T. Locality-constrained sparse auto-encoder for image classification. IEEE Signal Proc Lett 2015;22(8):1070–3.
[16] Guo Y, Gao Y, Shen D. Deformable MR prostate segmentation via deep feature learning and sparse patch matching. IEEE Trans Med Imaging 2016;35(4):1077–89.
[17] Shi B, Chen Y, Zhang P, Smith CD, Liu J. Nonlinear feature transformation and deep fusion for Alzheimer's Disease staging analysis. Pattern Recogn 2017;63:487–98.
[18] Ruderman DL, Bialek W. Statistics of natural images: scaling in the woods. Phys Rev Lett 1994;73(6):814–7.
[19] Ren X. Multi-scale improves boundary detection in natural images. Proc. Eur. Conf. Comput. Vis 2008:533–45.
[20] Arbeláez P, Pont-Tuset J, Barron J, Marques F, Malik J. Multiscale combinatorial grouping. Proc. IEEE Conf. Comput. Vis. Pattern Recogn 2014:328–35.
[21] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Proc. IEEE Conf. Comput. Vis. Pattern Recogn 2008.
[22] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. Proc. SPIE – Med. Imaging 2003;5033:440–51.
[23] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 2015;37(9):1904–16.
[24] Perronnin F, Dance CR. Fisher kernels on visual vocabularies for image categorization. Proc. IEEE Conf. Comput. Vis. Pattern Recogn 2007.
[25] Savelonas MA, Pratikakis I, Sfikas K. Fisher encoding of differential fast point feature histograms for partial 3D object retrieval. Pattern Recogn 2016;55:114–24.
[26] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. Proc. Eur. Conf. Comput. Vis 2010:143–56.
[27] Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the Fisher vector: theory and practice. Int J Comput Vis 2013;105(3):222–45.
[28] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2001;2:265–92.
[29] Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Liblinear Lin C-J. A library for large linear classification. J Mach Learn Res 2008;9:1871–4.
[30] Tommasi T, Caputo B, Welter P, Guld MO, Deserno TM. Overview of the CLEF 2009 medical image annotation track. CLEF 2009 Workshop, Lecture Notes in Computer Science 2010;6242:85–93.
[31] Dimitrovski I, Kocev D, Loskovska S, Džeroski S. Hierarchical annotation of medical images. Pattern Recogn 2011;44:2436–49.

[32] Avni U, Greenspan H, Konen E, Sharon M, Goldberger J. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. IEEE Trans Med Imaging 2011;30(3):733–46.
[33] Unay D, Soldea O, Ozogür-Akyuz S, Cetin M, Ercil A. Medical image retrieval and automatic annotation: VPA-SABANCI at image CLEF 2009. Working Notes for CLEF 2009 Workshop 2009.
[34] Zhu S, Tizhoosh HR. Radon features and barcodes for medical image retrieval via SVM. Proc. IEEE Int. Joint Conf. Neural Networks 2016:5065–71.
[35] Zhu C, Wang Z. Entropy-based matrix learning machine for imbalanced data sets. Pattern Recogn Lett 2017;88:72–80.
[36] De Oliveira JEE, Lopes A, Camara-Chavez G, De Araújo AA, Deserno T. MammoSVD: a content-based image retrieval system using a reference database of mammographies. Proc. IEEE Int. Symp. Biomed. Comput.-based Med. Syst 2009:1–4.
[37] Oliver A, Lladó X, Freixenet J, Martí R, Pérez E, Pont J, et al. Influence of using manual or automatic breast density information in a mass detection cad system. Acad Radiol 2010;17(7):877–83.
[38] Deserno TM, Soiron M, De Oliveira JEE, Araújo AA. Computer-aided diagnostics of screening mammography using content-based image retrieval. Proc. SPIE – Opt. Eng 2012;8315:211–9.
[39] Azevedo WW, Lima SML, Fernandes IMM, Rocha ADD, Cordeirol FR, Da Silva-Filhol AG, et al. Morphological extreme learning machines applied to detect and classify masses in mammograms. Proc. Int. Joint Conf. Neural Networks 2015.