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 $\rho$ 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 $\rho$, 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