Pattern Recognition 83 (2018) 161–173

Contents lists available at ScienceDirect

Pattern Recognition

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

Discriminatively boosted image clustering with fully convolutional auto-encoders

Fengfu Li[a,b], Hong Qiao[c,d,e], Bo Zhang[a,b,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049,China
[c] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[d] University of Chinese Academy of Sciences, Beijing 100049, China
[e] CAS Centre for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 23 March 2017
Revised 10 January 2018
Accepted 20 May 2018
Available online 21 May 2018

Keywords:
Image clustering
Fully convolutional auto-encoder
Representation learning
Discriminatively boosted clustering

ABSTRACT

Traditional image clustering methods take a two-step approach, feature learning and clustering, sequentially. However, recent research results demonstrated that combining the separated phases in a unified framework and training them jointly can achieve a better performance. In this paper, we first introduce fully convolutional auto-encoders for image feature learning and then propose a unified clustering framework to learn image representations and cluster centers jointly based on a fully convolutional auto-encoder and soft k-means scores. At initial stages of the learning procedure, the representations extracted from the auto-encoder may not be very discriminative for latter clustering. We address this issue by adopting a boosted discriminative distribution, where high score assignments are highlighted and low score ones are de-emphasized. With the gradually boosted discrimination, clustering assignment scores are discriminated and cluster purities are enlarged. Experiments on several vision benchmark datasets show that our methods can achieve a state-of-the-art performance.

1. Introduction

Clustering methods are very important techniques for exploratory data analysis with wide applications ranging from data mining [1,2], image segmentation [3–5] and so on. Their aim is to partition data points into clusters so that data in the same cluster are similar to each other while data in different clusters are dissimilar. Approaches to achieve this aim include partitional methods such as fuzzy c-means [6], hierarchical methods like agglomerative clustering and divisive clustering, methods based on density estimation such as DBSCAN [7], and recent methods based on finding density peaks such as CFSFDP [8].

Image clustering [9,10] is a special case of clustering analysis that seeks to find compact, object-level models from many unlabeled images. Its applications include automatic visual concept discovery, content-based image retrieval, image annotation, and so on. However, image clustering is a hard task mainly owning to the following two reasons: 1) images often are of high dimensionality, which will significantly affect the performance of clustering methods such as k-means [11], and 2) objects in images usually have two-dimensional or three-dimensional local structures which should not be ignored when exploring the local structure information of the images.

To address these issues, many representation learning methods have been proposed for image feature extractions as a preprocessing step. Traditionally, various hand-crafted features such as SIFT [12], LBP [13], NMF [14], and CW-SSIM similarity [15] have been used to encode the visual information. Recently, many approaches have been proposed to combine clustering methods with deep neural networks (DNN), which have shown a remarkable performance improvement over hand-crafted features [16,23]. Roughly speaking, these methods can be categorized into two groups: 1) sequential methods that apply clustering on the learned DNN representations, and 2) unified approaches that jointly optimize the deep representation learning and clustering objectives.

In the first group, a kind of deep (convolutional) neural networks is first trained in an unsupervised manner to approximate the non-linear feature embedding from the raw image space to the embedded feature space (usually being low-dimensional) [17]. And then, either k-means or spectral clustering or agglomerative clustering can be applied to partition the feature space. However, since the feature learning and clustering are separated from each other, the learned DNN features may not be reliable for clustering.

[*] Corresponding author at: Institute of Applied Mathematics, 55 Zhongguancun Donglu, Beijing 100190, China.
E-mail address: b.zhang@amt.ac.cn (B. Zhang).

https://doi.org/10.1016/j.patcog.2018.05.019

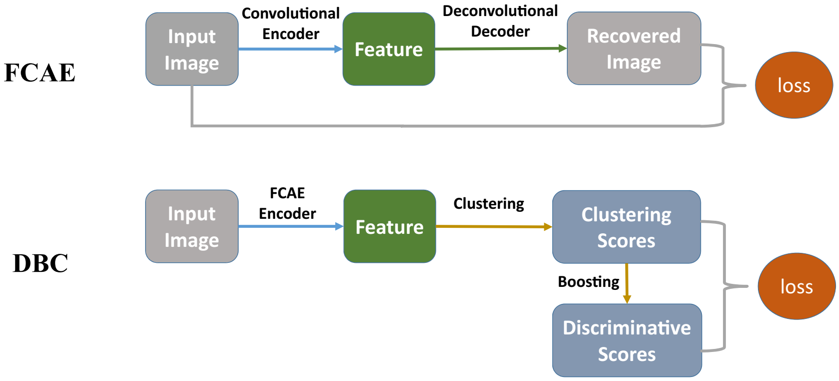

Fig. 1. Diagram of the proposed fully convolutional auto-encoders and discriminatively boosted clustering.

There are a few recent methods in the second group which take the separation issues into consideration. In [18], the authors proposed deep embedded clustering that simultaneously learns feature representations with stacked auto-encoders and cluster assignments with soft *k*-means by minimizing a joint loss function. In [19], joint unsupervised learning was proposed to learn deep convolutional representations and agglomerative clustering jointly using a recurrent framework. In [20], the authors proposed an infinite ensemble clustering framework that integrates deep representation learning and ensemble clustering. The key insight behind these approaches is that good representations are beneficial for clustering and conversely clustering results can provide supervisory signals for representation learning. Thus, two factors, designing a proper representation learning model and designing a suitable unified learning objective will greatly affect the performance of these kind of methods.

In this paper, we follow recent advances to propose a unified clustering method named discriminatively boosted clustering (DBC) for image analysis based on fully convolutional auto-encoders (FCAE). See Fig. 1 for the diagram of the them. We first introduce a fully convolutional encoder-decoder network for fast and coarse image feature extraction. We then discard the decoder part and add a soft *k*-means model on top of the encoder to make a unified clustering model. The model is jointly trained with gradually boosted discrimination where high score assignments are highlighted and low score ones are de-emphasized. The our main contributions are summarized as follows:

- We propose a fully convolutional auto-encoder (FCAE) for image feature learning. The FCAE is composed of convolution-type layers (convolution and de-convolution layers) and pool-type layers (pooling and un-pooling layers). By adding batch normalization (BN) layers to each of the convolution-type layers, we can train the FCAE in an end-to-end way. This avoids the tedious and time-consuming layer-wise pre-training stage adopted in the traditional stacked (convolutional) auto-encoders. To the best of our knowledge, this is the first attempt to learn a deep auto-encoder in an end-to-end manner.
- We propose a discriminatively boosted clustering (DBC) framework based on the learned FCAE and an additional soft *k*-means model. We train the DBC model in a self-paced learning procedure, where deep representations of raw images and cluster assignments are jointly learned. This overcomes the separation issue of the traditional clustering methods that use features directly learned from auto-encoders.
- We show that the FCAE can learn better features for clustering than raw images on several image datasets include MNIST, USPS, COIL-20 and COIL-100. Besides, with discriminatively boosted learning, the FCAE based DBC can outperform several state-of-the-art analogous methods in terms of *k*-means and deep auto-encoder based clustering.

The remaining part of this paper is organized as follows. Some related work including stacked (convolutional) auto-encoders, deconvolutional neural networks, and joint feature learning and clustering are briefly reviewed in Section 2. Detailed descriptions of the proposed FCAE and DBC are presented in Section 3. Experimental results on several real datasets are given in Section 4 to validate the proposed methods. Conclusions and future works are discussed in Section 5.

2. Related work

Stacked auto-encoders have been studied in the past years for unsupervised deep feature extraction and nonlinear dimensionality reduction. In [21], Restricted Boltzmann Machines (RBMs) were stacked layer-by-layer to form a deep auto-encoder for dimensionality reduction. In [28], Stacked Autoassociators Network was proposed to learn hierarchical features. Vincent et al. [22] introduced stacked denoising auto-encoders (SDAEs), which corrupts the input with random noise to make the algorithm more robust to variation than ordinary SAEs and gets superior classification performance. Though the above SAEs can learn hierarchical features from raw data, they are not suited to deal with structural image data. Masci et al. [25] and Lee et al. [26] proposed convolutional extensions of the SAEs (namely stacked convolutional auto-encoders, or SCAEs) to address the issue. They take 2-D or 3-D image as the input and adopt convolution operation along with pooling operation to train the SCAEs. Compared with the SAEs, the SCAEs can reserve more structural information. Though the SAEs and SCAEs are powerful unsupervised feature extraction methods, their training usually contain a tedious two-stage procedure [21,24]: one is layer-wise pre-training and the other is overall fine-tuning. One of the significant drawbacks of this learning procedure is that the layer-wise pre-training is time-consuming and tedious, especially when the base layer is a RBM rather than an ordinary auto-encoder, or when the overall network is very deep.

Recently, there is an attempt to discard the layer-wise pre-training procedure and train a deep auto-encoder type network in an end-to-end way. In [29], a deep deconvolution network was learned for image segmentation. The input of the architecture is an image and the output is a segmentation mask. The network achieves the state-of-the-art performance compared with analogous methods thanks to three factors: 1) introducing a deconvolution layer and a unpooling layer to recover the original image size of the segmentation mask, 2) applying the batch normalization strategy [30] to each convolution layer and each deconvolution layer to reduce the internal covariate shifts, which not only