Information Sciences 385–386 (2017) 338–352

Contents lists available at ScienceDirect

Information Sciences

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

Canonical correlation analysis networks for two-view image recognition

Xinghao Yang[a], Weifeng Liu[a,*], Dapeng Tao[b,*], Jun Cheng[c,d]

[a] College of Information and Control Engineering, China University of Petroleum (East China), Qingdao 266580, Shandong, China
[b] bSchool of Information Science and Engineering, Yunnan University, Kunming 650091, Yunnan, China
[c] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[d] The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 2 August 2016
Revised 14 December 2016
Accepted 2 January 2017
Available online 4 January 2017

Keywords:
Canonical correlation analysis
Deep learning
Convolutional neural networks
Image classification

ABSTRACT

In recent years, deep learning has attracted an increasing amount of attention in machine learning and artificial intelligence areas. Currently, many deep learning network-related architectures such as deep neural networks (DNNs), convolutional neural network (CNN), wavelet scattering network (ScatNet) and principal component analysis network (PCANet) have been proposed. The most effective network is PCANet, which has achieved promising performance in image classification, such as for face, object and handwritten digit recognition. PCANet can only handle data that are represented by single-view features. In this paper, we present a canonical correlation analysis network (CCANet) to address image classification, in which images are represented by two-view features. The CCANet learns two-view multistage filter banks by a canonical correlation analysis (CCA) method and constructs a cascaded convolutional deep network. Then, we incorporate filters with binaryzation and block-wise histogram processes to form the final depth structure. In addition, we introduce a variation of CCANet—dubbed RandNet-2—in which the filter banks are randomly generated. Extensive experiments are conducted using the ETH-80, Yale-B, and USPS databases for object classification, face classification and handwritten digits classification, respectively. The experimental results demonstrate that the CCANet algorithm is more effective than PCANet, RandNet-1 and RandNet-2.

1. Introduction

In real world image classification tasks, a crucial problem is intra-class modifiability, which is derived from the variation in lighting, rotation and deformation. Numerous efforts have been made to eliminate the variability within image classes, such as low-level features and deep learning network structures. Although many handcrafted low-level image features, such as local binary patterns (LBPs) [20], salient features [40] and scale-invariant feature transforms(SIFT) [19], can extract the shape and texture features of a digital image in a valid manner, direct application to the new data sets is difficult [1,28]. Thus, new domain knowledge, such as multiview techniques [21,29], hashing algorithm [41], dictionary learning [23,25,44], manifold learning [24,36] and subspace selection [35], are usually needed when generalizing manually designed features to

* Corresponding authors.
E-mail addresses: liuwf@upc.edu.cn, liuwfxy@gmail.com (W. Liu), dapeng.tao@gmail.com (D. Tao).

http://dx.doi.org/10.1016/j.ins.2017.01.011
0020-0255/

new missions, including image classification [22,28,39], action retrieval [29], image super resolution [25] and efficient image search [21].

Deep learning (DL), which has rapidly developed in recent years, and many types of deep learning network-related algorithms have been successfully applied to image recognition tasks [2,3,15,30,33,37–39,42,43]. The main idea of a deep network structure entails the use of features at different levels to represent different degrees of abstract semantics of images, such as pixels, margins, motifs, parts, objects, and scenes [15]. These layered features that are learned from training data as a remedy to low-level features can effectively guarantee invariants to intra-class variability. Representative deep learning methods include DNNs [6,33], CNN [5,10,11,14,17,18,32,34], ScatNet [2,27,31] and PCANet [3].

Deep neural networks (DNNs) [6,33] employ a hierarchical structure to extract a multistage representation of data. Hinton et al. [6] utilized complementary knowledge to derive a fast, greedy algorithm that can rapidly learn parameters. Sun et al. [33] proposed two very deep neural networks that are based on stacked convolution architecture [32] and inception layers [34] for face recognition.

A convolutional neural network (CNN) [5,10,11,14,17,18,32,34] incorporates a convolution structure in each trainable stage that is usually composed of three layers: a convolutional filter layer, a nonlinearity process layer, and a feature merging layer. In a convolutional layer, the filter kernel is generally learned by a stochastic gradient descent (SGD) method [14], and each filter can detect a particular feature of the input image. Therefore, the output of each convolutional layer will have a corresponding change to the translation of the input image [15]. In the CNN method, parameters tuning is a time-consuming task that requires some specific techniques. Krizhevsky et al. [11] designed an expertise network for a large image dataset that contains 650,000 neurons and 60 million parameters to train. Additionally, high recognition accuracy is certified by an adequately deep structure [32,34]. For example, Simonyan et al. [32] researched the influence of the depths of convolutional networks in large-scale image recognition tasks, and ideal results are obtained when the model structure contains 16–19 layers. Convolutional based deep networks did not have an explicit mathematical explanation due to the nonlinearity process.

The wavelet scattering network (ScatNet) [2,27,31] is the first algorithm with a distinct mathematical basis. Bruna et al. [2] accomplished a scattering transform with a deep convolutional network that is composed of a cascaded wavelet transform and a modulus pooling operator. Compared with a CNN, ScatNet uses prefixed filters that are wavelet operators. Therefore, the filters are obtained without learning in ScatNet [2,27,31]. Although the filter bank is predetermined, the experimental results of ScatNet are remarkable and are superior to DNNs and CNN in some visual based recognition tasks, including handwritten digit recognition, texture discrimination [2,31] and object classification [27]. However, when a pre-fixed structure is extended to face recognition, in which the intra-class variation is significant, the results are not satisfactory [3].

Chan et al. [3] built a principal component analysis network (PCANet) that employs a cascaded PCA to learn two layers of filter banks and follows by binaryzation and block-wise histogram to pool the final feature. The architecture of PCANet is very simple without numerous parameters to tune in the training stage. This seemingly naive structure performs equal to or more commonly better than well-designed low-level features, DNNs, CNN and ScatNet in several well-known databases that entail LFW, MultiPIE, Extend Yale-B, AR, FERET and MNIST [3].

These deep learning network-related methods can only handle circumstances, in which the input images are represented by a single view. To surmount two view cases and achieve a more robust performance, we propose a canonical correlation analysis network (CCANet) in this paper. Two-view multilayer filter banks are learned by a CCA method, which finds the principle filters by maximizing the correlation of the projected two-view variables. Thus, the filters can reflect more comprehensive information of the same object compared with PCANet. Fig. 1 illustrates the framework of a two-convolutional stage CCANet. In the output stage of CCANet, binaryzation is adopted as a nonlinear process instead of a rectified sigmoid function [15] or ReLU function [11], and a block-wise histogram method is employed to form the final feature representation. Our proposed CCANet model has three significant advantages. (1) CCANet can simultaneously consider two-view features of one image, which is considered to be more robust than the use of a single view in classification tasks regarding intra-class variance. (2) The number of convolutional stages of CCANet is less than the number of convolutional neural networks [11,14,34]. An unsupervised learning method is adopted in CCANet instead of the backpropagation algorithm in a typical CNN [12,13]. The number of parameters in CCANet is small. (3) We also introduce a variation of the CCANet—named RandNet-2—which employs randomly generated filter banks (consider that the filters obey a Gaussian i.i.d.) to replace the filter banks in the CCANet structure. To verify the effectiveness of the proposed CCANet and RandNet-2, we conduct extensive experiments using the ETH-80 database for object recognition, using the Yale-B database for face verification and using the USPS database for handwritten digits classification. The experimental results demonstrate that CCANet achieves a higher recognition accuracy than the accuracy of the representative deep learning network-related methods, including PCANet, for object, face and handwritten digit recognition.

The remainder of this paper is arranged as follows: several types of related networks are described in Section 2. Section 3 presents details of the proposed CCANet. The experimental results are provided in Section 4. The conclusions are presented in Section 5.

2. Related works

In this section, we summarize several related networks of the CCANet, including the principle component analysis network (PCANet), two-dimensional principle component analysis network (2DPCANet), discrete cosine transform network