Neurocomputing 144 (2014) 198–207

Contents lists available at ScienceDirect

Neurocomputing

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

Sparse-based neural response for image classification

Hong Li [a], Hongfeng Li [a,*], Yantao Wei [b], Yuanyan Tang [c], Qiong Wang [a]

[a] School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department of Information Technology, Central China Normal University, Wuhan 430079, China
[c] Faculty of Science and Technology, University of Macau, Macau, China

ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form
15 March 2014
Accepted 22 April 2014
Communicated by D. Wang
Available online 2 June 2014

Keywords:
Neural response
Sparse coding
Maximum pooling operation
Robust
Image classification

ABSTRACT

Image classification is a popular and challenging topic in the computer vision field. On the basis of advances in neuroscience, this paper proposes a sparse-based neural response feature extraction method for image classification. The approach extracts discriminative and invariant representations of images by alternating between non-negative sparse coding and maximum pooling operation with effectiveness. Additionally, effective template selection methods are proposed to further enhance the performance of the algorithm. In comparison with traditional hierarchical methods, our proposed model accounts for the neural processing of visual cortex in human brain, which appears to gain more beneficial discriminative and robust properties for image classification tasks. A variety of benchmarks are used to evaluate the algorithm. The experiment results demonstrate that our proposed algorithm achieves quite excellent or state-of-the-art performance compared with other popular methods.

1. Introduction

Nowadays, with the rapid development of video surveillance, Internet technology and unmanned technology, etc., there are overwhelming amounts of images available to us, which makes it imperative to develop novel image classification algorithms to sift images efficiently. However, images are high dimensionality and variable signals that change dramatically under varying scales, lighting conditions, viewpoints, context variation, etc., which makes them difficult for classifying. Robust and discriminative algorithms should be designed to complete the classification tasks [1–3] for a specific problem. Therefore, how to extract effective features from images has become an essential problem for image classification.

Over the past few decades, there are a large number of feature extraction methods have been proposed. These methods can be divided into supervised methods, semi-supervised methods and unsupervised methods. Supervised feature extraction methods involve learning features from labeled image data. Examples of these supervised methods include scale invariant feature transform (SIFT) [4], histogram of oriented gradients (HOG) [5], and speeded up robust features (SURF) [6], multiple kernel learning [7], etc. These feature representations are expensive to obtain due to the significant human labor for a large amount of labeled data. Therefore, it is desirable to learn feature representations using a large amount of unlabeled data with only a small amount of labeled data. There has been research on utilizing unlabeled data to enhance supervised methods, such as latent Dirichlet allocation (LDA) [8] and alternating structural optimization [9]. In contrast, unsupervised feature extraction methods can learn features from unlabeled data. These methods include clustering methods (e.g. K-means), matrix factorization methods (e.g. principal component analysis (PCA), independent component analysis (ICA), non-negative matrix factorization (NMF) [10], and sparse coding [11]). The above-mentioned matrix factorization methods can be viewed as some sort of linear embeddings. There are also more flexible representations obtained by nonlinear embeddings, which include Isomap [12], locally linear embedding (LLE) [13] and Laplacian eigenmaps (LE) [14], etc. Particularly, sparse coding can be viewed as a matrix factorization where the coefficients are assumed to be sparse. One advantage of the sparse coding is that it can explore the succinct representations of the input data and it is robust to noise in the data or “change” in the input distributions.

Recently, deep learning approaches have gained wide interest as a way of constructing hierarchical structures to extract features from unlabeled data. Hierarchical structures are promising in learning simple concepts first and then constructing more complex ones by combining simple ones together. For example, Bengio et al. [15] proposed a greedy algorithm based on autoencoders. Spatial pyramid matching (SPM) algorithms [16,17] and its variants were proposed by introducing global information via combining local features with a finite multilayered structure. With the development of neuroscience, Riesenhuber and Poggio [18] described a hierarchical model that is consistent with the real

* Corresponding author.
E-mail addresses: hongli@hust.edu.cn (H. Li), lihongfeng@hust.edu.cn (H. Li), yantaowei@mail.ccnu.edu.cn (Y. Wei), yytang@umac.mo (Y. Tang), wangqiong701@hust.edu.cn (Q. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.04.053
0925-2312/

visual processing in human cortex, called HMAX. After that, based on the HMAX model, Ref. [19] described a hierarchical system inspired by the organization of visual cortex and constructed an increasingly complex and invariant feature representation for realistic object recognition. Based on [20], many new works were proposed by incorporating additional properties, such as sparsity and localized features [21]. The neural response (NR) model [22] proposed by Smale et al. is a hierarchically constructed natural image representation model. It has the potential to be further developed to enhance the performances for various pattern recognition tasks [23,24]. More recently, there have been a few novel works [25,26] based on the NR model for feature extraction.

In this paper, based on the investigation of the NR model and properties of sparse coding, we propose a semi-supervised sparse-based neural response algorithm for image classification. The model is built by alternating between local non-negative sparse coding and global maximum pooling operation in a multi-layer way, which is more faithful to the hierarchical visual processing in the human visual cortex. The non-negative sparse coding is utilized to extract salient features of local regions on different layers, which leads to the proposed algorithm gaining more discriminative characters than the NR model, while the max-pooling operation gains the model more invariant properties.

Compared with other previous works, our contributions in this paper can be summarized as follows. Firstly, we introduce a recursively constructed sparse-based neural response algorithm for feature extraction. Then, the proposed model is constructed by alternating between non-negative sparse coding and maximum pooling operation based on the dense SIFT features of images. The spare coding operation can explore the salient features of the images, while the maximum pooling operation makes the algorithm invariant to transformations, rotations, etc. In addition, simple and effective template selection methods are proposed for the algorithm construction, which further enhances the performance of the proposed algorithm. The proposed model has advantage that we have no previous assumption to positions where objects may appear, while the hierarchical algorithms based on the SPM model assume that similar parts of an object should appear at similar positions or grids, which is not always true in practical tasks. All the above contributions lead to the discriminative and invariant properties of our model, which are beneficial to image classification tasks that objects in images vary in illumination, translations, scalings, rotations, and so on.

The remaining part of the paper is organized as follows: Section 2 describes the construction of the sparse-based neural response in detail. In Section 3, we propose two effective methods for selecting the two layer templates. Experiments are given to evaluate the effectiveness of our model in Section 4. Finally, in Section 5, we draw conclusions and suggest more work should be done in the future.

2. Sparse-based neural response

In this section, we will describe the proposed sparse-based neural response algorithm in detail in a bottom-up way. The algorithm is constructed by alternating between sparse coding and max-pooling operation on the dense SIFT features of the original images. Our proposed algorithm is supposed to mimic the visual process of human visual cortex by constructing a hierarchical architecture based on the non-negative sparse coding, consequently, it is efficient for image classification tasks which have large variation in images and a small number of labeled images.

2.1. Preliminaries

Before giving a formal description of the proposed hierarchical algorithm, we firstly introduce some preparations which are necessary for constructing the model.

In this paper, we introduce a hierarchical architecture which includes three nested layers: u, v, and S_q in R^2, with $u \subset v \subset S_q$, as Fig. 1 shows.

We assume that a function space is defined on S_q, denoted by $I_{S_q}(S_q) = \{f : S_q \to [0,1]\}$ as well as function spaces $I_u(u)$, $I_v(v)$ defined on u, v respectively. We interpret the functions as grayscale images when working with a vision problem. Given a patch S_q, we can define an image I on it and, similarly, image patches I_u and I_v can be defined on u and v respectively, which Fig. 2 illuminates.

Then we further assume a set $H_u = \{h_u^1, h_u^2, \ldots, h_u^{|H_u|}\}$ of transformations which are maps from the smallest patch to the next larger patch $h_u^i : u \to v, i = 1, 2, \ldots, |H_u|$, where $|H_u|$ denotes the cardinality of H_u, and similarly, $H_v = \{h_v^1, h_v^2, \ldots, h_v^{|H_v|}\}$ with $h_v^i : v \to S_q, i = 1, 2, \ldots, |H_v|$, and $|H_v|$ is the cardinality of H_v. The sets of transformations are supposed to be finite. Examples of transformations are translations, scalings, rotations, and so on. Combining the first two,

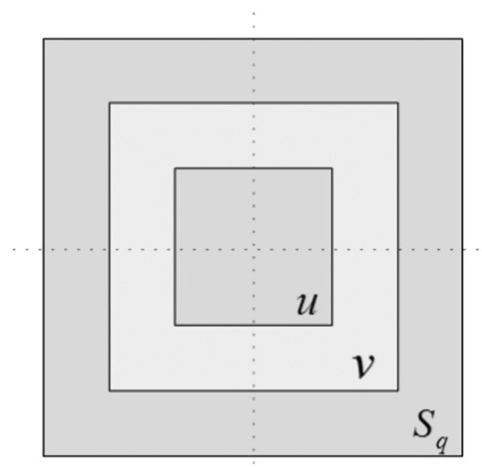

Fig. 1. The three layers architecture.

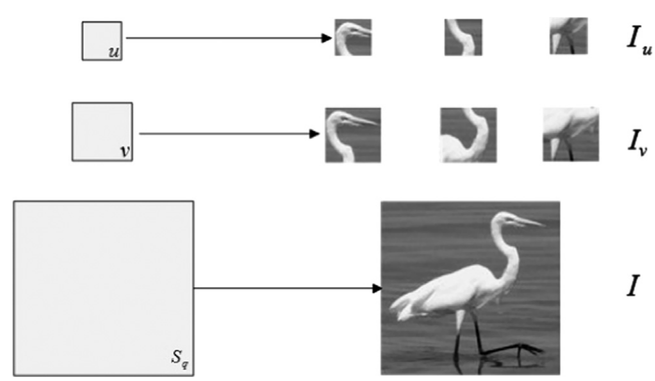

Fig. 2. Examples of image patches defined on the three layers.

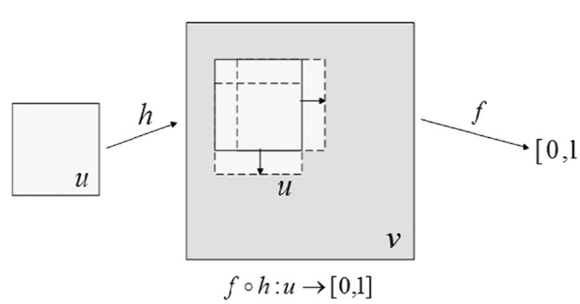

Fig. 3. The sample of transformation.