Pattern Recognition 62 (2017) 225–235

Contents lists available at ScienceDirect

Pattern Recognition

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

Combining local and global: Rich and robust feature pooling for visual recognition

Wei Xiong [a], Lefei Zhang [a], Bo Du [a,*], Dacheng Tao [b]

[a] Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, State Key Lab of Software Engineering, School of Computer Wuhan University, China
[b] Centre for Quantum Computation & Intelligent Systems University of Technology, Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 4 March 2016
Received in revised form
30 May 2016
Accepted 9 August 2016
Available online 17 August 2016

Keywords:
Unsupervised learning
Spatial pooling
Auto-encoder

ABSTRACT

The human visual system proves expert in discovering patterns in both global and local feature space. Can we design a similar way for unsupervised feature learning? In this paper, we propose a novel spatial pooling method within an unsupervised feature learning framework, named *Rich and Robust Feature Pooling* (R^2FP), to better extract rich and robust representation from sparse feature maps learned from the raw data. Both local and global pooling strategies are further considered to instantiate such a method. The former selects the most representative features in the sub-region and summarizes the joint distribution of the selected features, while the latter is utilized to extract multiple resolutions of features and fuse the features with a feature balance kernel for rich representation. Extensive experiments on several image recognition tasks demonstrate the superiority of the proposed method.

1. Introduction

The performance of modern machine learning and pattern recognition algorithms relies more and more on the quality of data representation, which embodies the inner structures and correlations of the input instances to make data more separable [1–4]. One of the most successful representation learning algorithm for visual recognition tasks is convolutional neural networks (CNNs) [5,6], which stack layers of neurons to learn deep features from the raw data. Though the CNNs have achieved satisfactory performance on some big database such as ImageNet [7], it's success requires tremendous amount of labeled data which is very expensive to obtain. To relieve the urgent need for labeled data, unsupervised feature learning systems [8–10] are built, aiming to learn features automatically without any supervisory information attached to the training data.

The generic pipeline of the unsupervised feature learning system is composed of two main modules: the encoder module and the pooling module. The encoder module is established to learn the feature detectors/weights and encode the input data into feature maps. Effective encoders such as sparse coding [11–13], k-means clustering [14], auto-encoders [15,16] and restricted Boltzmann machines [17,18], have been proved promising to learn useful feature detectors. In this paper, we build our framework based on a single-hidden-layer autoencoder. The hidden layer of the autoencoder is imposed on sparse constraint for better representation. In addition to the sparse constraint, ReLU (Restricted Linear Unit) [5,19] is utilized as the activation function to further increase the sparsity of the learned features. Following the encoder module, the pooling module [20–24], [22,25–29] is deployed to sub-sample the feature maps for compact and abstract representation. It removes the redundancy of the features and allows small transformations of the input. Generally, two types of pooling methods have been investigated in prior work, the global pooling and the local pooling. The global methods provides efficient strategies to divide the global feature space into different resolutions of sub-regions. Then the statistic value of the elaborately divided sub-region's features can be calculated by any type of local pooling methods embedded into the global pooling to form the pooled features. After that, the pooled features of the sub-regions at each resolutions are fused into the final representation. The fusion strategy is also defined by the global pooling.

One of the typical global pooling method is spatial pyramid matching (SPM) [30,31]. SPM partitions an image into spatial bins at different scales and computes the histograms of each spatial bin using the spatial pyramid matching kernel. Kaiming He et al. [32] utilized spatial pyramid pooling (SPP) based on SPM in convolutional neural networks to pool the convolutional feature maps at multiple resolutions and learn rich representations at the inference stage. They fuse the pooled features at each resolution by

* Corresponding author.
E-mail addresses: wxiong@whu.edu.cn (W. Xiong), zhanglefei@whu.edu.cn (L. Zhang), remoteking@whu.edu.cn (B. Du), dacheng.tao@uts.edu.au (D. Tao).

http://dx.doi.org/10.1016/j.patcog.2016.08.006
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

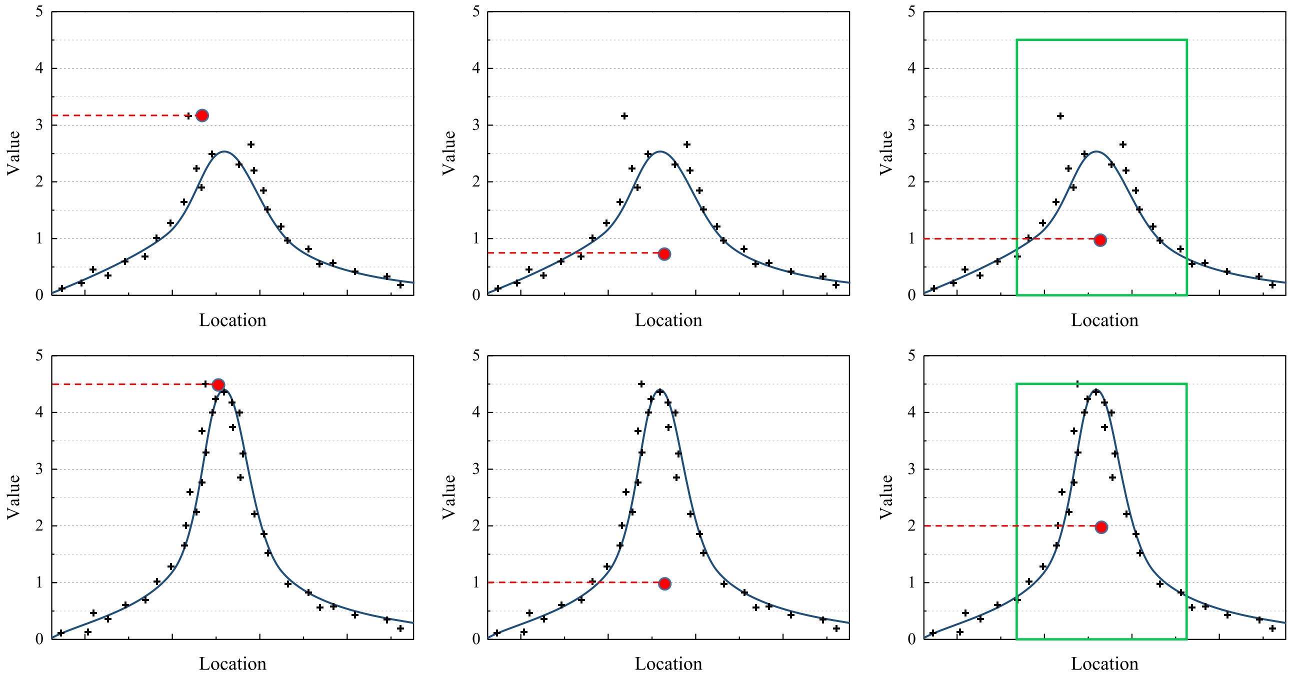
simply concatenating them into a single vector. Another alternative of global pooling is to merely use convolution instead of pooling. By setting different strides in the convolution procedure, we can also generate global features at different scales. This strategy is simple, but may not be effective as the stride will be very large, hence there will be a great information loss in the convolution procedure.

To generate suitable statistics of the local region, several impressive local pooling methods have been proposed. The most universal methods are max pooling and average pooling, which adopt the maximum and the mean of the local features to represent the sub-region, respectively. Boureau et al. [33] utilized the p-norm of the local elements to represent region of the interest. The p-norm pooling introduces a factor P that can be tuned to fit the input data with different distributions, thus better representations can be extracted. Similar to p-norm pooling, other norm-form pooling methods have been proposed. In these methods, the pooled features are further normalized by the L_1, L_2 norm and the power normalization technique [24]. Zeiler et al. [34] proposed stochastic pooling, which selects the pooled features by sampling from a multinomial distribution formed by the activations in the sub-region, imposing randomness to the generated representations. Boureau et al. [33] assumed the features were Bernoulli random variables and proposed smooth max pooling (SM) to combine the advantages of both average and max pooling.

Though the prior pooling methods have proved effective in current unsupervised feature learning systems, both the local and the global pooling methods have some disadvantages in learning more efficient features.

For the local methods, the typical ones are max pooling and average pooling as they are very simple and effective in many frameworks. However, we find in the unsupervised learning systems that these methods are not perfect enough to calculate the proper statistic of features in a sub-region. In terms of max pooling, it selects only one element of all the local features to represent the whole sub-region and discard the other features that may also have great impacts on the quality of the final representation. So the statistic of features generated by these methods may lose conductive information and perform poorly to describe the distribution of features in the sub-region. Furthermore, there may be some outliers and noise in the input data and they can be transformed to abnormal features that are improper to represent the local regions. If these noisy features are selected by the pooling module, the resulting features may be harder to be separated, as illustrated in Fig. 1 (a) (top and bottom). The same problem goes with stochastic pooling, which selects one single feature in each sub-region to conduct pooling operation. In terms of average pooling, it utilizes all the elements in the sub-region to calculate the statistic value of the local features. When dealing with features with large sparsity, say features generated by some of the unsupervised feature learning systems, these methods can lead to a situation that the means or norms of features in different sub-regions are close to each other and tend to be very small, as most elements in the local areas approach zero, as illustrated in Fig. 1 (b). The resulting pooled features are then much harder to be correctly separated. The same problem goes with p-norm and other norm-form pooling methods.

For the global methods, conventional methods such as spatial pyramid pooling simply concatenate features of each resolution into the final feature vector, without taking into consideration the importance of features at different resolutions. In terms of using larger stride in convolution to replace the global pooling, it also doesn't consider the weights of features at different scales. If equal importance is attached to different resolutions of features, the less vital features may take the leading place to greatly affect the fused features. Better fusion strategies can be discovered that will



(a) Max Pooling (b) Average Pooling (c) Local R^{2}FP

Fig. 1. Toy example showing the problems different methods may encounter when pooling the sparse features in the local region A and B. The top three sub-figures show the distribution of features in sub-region A, and the bottom three figures illustrate the feature distribution of sub-region B. In each sub-figure, the horizontal axis represents the location of a feature and the vertical axis stands for the value of each feature. Figure (a), (b) and (c) shows the possible pooling results of max Pooling, Average pooling and the local R^{2}FP, respectively. The pooled feature is marked as a red point. As shown in figure (c), the proposed local R^{2}FP selects part of the larger features (features inside the green box are selected), the resulting pooled features of region A and region B are much more separated than that pooled by average pooling. Compared with max pooling, the proposed local R^{2}FP automatically reduces the negative effects of the outliers and the feature generated by local R^{2}FP fits the distribution of the local elements much better.