# Large-scale Dictionary Learning for Local Coordinate Coding

Bo Xie[1]
boxie.ntu@gmail.com
Mingli Song[2]
brooksong@ieee.org
Dacheng Tao[1]
dctao@ntu.edu.sg

[1] School of Computer Science
Nanyang Technological University
Singapore 639798

[2] College of Computer Science
Zhejiang University
Hangzhou 310027, China

Dictionary learning is a method to learn dictionary items adapted to data of a given distribution. It is shown that dictionary learned from data is more suited for vision task than universal dictionaries [4]. Traditionally, Vector Quantization (VQ), or using k-means to learn data cluster centroids, is a simple and popular method in the bag-of-features framework [5]. Recently, sparse coding is used in visual dictionary learning and achieves lower reconstruction error [6]. To capture manifold geometry of the data distribution, local coordinate coding (LCC) [7] is proposed and it achieves state-of-the-arts performance on PASCAL VOC 2009 challenge.

One problem with the original LCC for learning a visual dictionary is that the time complexity grows linearly with the number of samples. For large-scale datasets of millions of samples, the computational cost becomes unacceptable. In this paper, we propose an online LCC dictionary learning algorithm that only processes one or a small mini-batch of random samples at every iteration round based on [3]. This stochastic approach converges almost surely and can scales up gracefully to large-scale datasets.

To learn a visual dictionary $D$ from data samples $x_i$, LCC optimizes the following objective function [7]

$$\min_{D,\alpha_i} \sum_i \left( \frac{1}{2} \|x_i - D\alpha_i\|^2 + \mu \sum_j |\alpha_i^j| \|d_j - x_i\|^2 \right) \tag{1}$$

where $d_j$ is the $j$-th dictionary item and $\alpha_i$ is the coding coefficient for $x_i$. $\alpha_i^j$ is the $j$-th component of $\alpha_i$ and $\mu$ is the regularization coefficient. The objective function consists of two parts: The first term of the objective function measures reconstruction error and the second term preserves locality in coding.

The optimization is solved by alternating between $D$ and $\alpha_i$. Since at every iteration all $\alpha_i$ need to be updated given the current $D$, the complexity is linear with the number of samples. Our approach minimizes an upper bound of the original objective function which only requires a random sample per iteration.

Suppose at iteration $t$, we randomly draw a sample $x_t$ from the distribution $p(x)$ and now have a sequence of random samples $x_1, \ldots, x_t$. The objective function at iteration $t$ is defined as

$$\tilde{f}_t(D) = \frac{1}{t} \sum_{i=1}^{t} \left( \frac{1}{2} \|x_i - D\alpha_i\|^2 + \mu \sum_j |\alpha_i^j| \|d_j - x_i\|^2 \right). \tag{2}$$

And

$$\alpha_i = \operatorname*{argmin}_{\alpha} \frac{1}{2} \|x_i - D_{i-1}\alpha\|^2 + \mu \sum_j |\alpha^j| \|(D_{i-1})_j - x_i\|^2 \tag{3}$$

where $(D_{i-1})_j$ denotes the $j$-th dictionary item from $D_{i-1}$. Note that $\alpha_t$ is computed from $D_{t-1}$ and $x_t$ and thus decouples from future dictionary values $D_k$, $(k \geq t)$.

Optimizing over $D$ given fixed $\alpha_i$ is a quadratic programming problem, which can be solved efficiently by block-coordinate descent [3]:

$$\begin{aligned} D =& \operatorname*{argmin}_{D} \sum_i \left( \frac{1}{2} \|x_i - D\alpha_i\|^2 + \mu \sum_j |\alpha_i^j| \|d_j - x_i\|^2 \right) \\ =& \operatorname*{argmin}_{D} \frac{1}{2} \operatorname{tr}\left(D^T DA\right) - \operatorname{tr}\left(D^T B\right). \end{aligned} \tag{4}$$

And

$$A = \sum_i \alpha_i \alpha_i^T + 2\mu\Sigma_i, \tag{5}$$

$$B = \sum_i x_i \alpha_i^T + 2\mu x_i \bar{\alpha}_i^T, \tag{6}$$

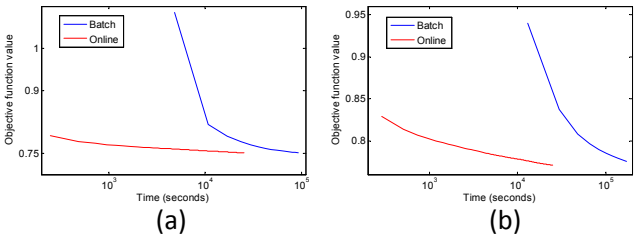


Figure 1: Objective value v.s. time (a) VOC dataset; (b) Caltech 256 dataset.

where $\bar{\alpha}_i$ is component-wise absolute value of $\alpha_i$, i.e. $\bar{\alpha}_i^j = \left|\alpha_i^j\right|$ and $\Sigma_i$ is a diagonal matrix constructed from $\bar{\alpha}_i$. The history informaiton is accumulated in matrices $A$ and $B$.

The upper bound converges to the desired objective function almost surely at infinity. Thus, by minimizing the upper bound, we approximately learn a visual dictionary by LCC, with much less computational cost.

We evaluated our method by object recognition task in VOC 2007 [1] and Caltech 256 [2] datasets. Fig. 1 illustrates how objective values descend over time in batch and online versions. Our conclusion is that the proposed online learning method can scale up to large-scale datasets with comparable performance with batch LCC dictionary learning algorithm.

[1] Mark Everingham, Luc Van Gool, Chris Williams, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. URL `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`.

[2] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

[3] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

[4] Joseph F. Murray and Kenneth Kreutz-Delgado. Learning sparse overcomplete codes for images. *J. VLSI Signal Process. Syst.*, 46 (1):1–13, 2007. ISSN 0922-5773.

[5] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*. Springer, 2006. URL `http://lear.inrialpes.fr/pubs/2006/NJT06`.

[6] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801. IEEE, 2009.

[7] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2223–2231. 2009.