Beyond Hard Negative Mining

Alberto Bietti
École Normale Supérieure, Cachan
alberto.bietti@gmail.com
January, 2014

Abstract

Training object detectors usually involves hard negative mining, which consists in repeatedly finding high scoring false positive detections and using them to train a new classifier. This process can be very expensive, and most negative windows won't be considered for training. Recent work [5, 6] has shown that it is possible to obtain good object detectors with no hard negative mining, by using all the available training data and exploiting its particular structure. We mainly focus on the work of Henriques et al. *[6], which uses the block-circulant structure of the Gram matrix describing an augmented dataset to efficiently train an object detector. We evaluate this method on the INRIA person dataset and the Caltech pedestrians dataset, showing improved performance of the detector compared to standard hard negative mining approaches, as well as speed improvements by about an order of magnitude.*

1. Introduction

Most recent approaches for training an object detector rely on training a linear classifier (typically a linear SVM) on filter-like features such as Histogram of Oriented Gradients (HOG) features from Dalal and Triggs [1]. The popular technique of deformable part models [3] is also based on training these linear classifiers on HOG features for different parts of an object template. Despite their success, there is one significant problem with these methods, which is that training these models is a slow procedure, mostly because of expensive rounds of what is commonly referred to as *hard negative mining*. Training with all possible windows (at multiple scales) from all training images is considered impractical since the number of training samples would be too large. Instead, one typically keeps a pool of training windows, which initially includes all positive windows and a set of randomly sampled negative windows, and repeats the process of training a classifier, running the detector at multiple scales on training images to find high-scoring false positives (referred to as *hard negatives*) and adding them to the pool of training windows. This repeated process of *hard negative mining* can be very slow since it involves evaluating a classifier in a sliding-window manner at multiple scales over many training images.

In addition to being computationally expensive, hard negative mining ends up using only a few windows per training image, thus not exploiting the abundance of negative windows in the training images, which could help to improve the performance of the classifier. A recent idea has been that of using the structure of the full training images and their redundancies to build a classifier from all the available data. This can lead to better performance since more training data usually implies better generalization, as well as much faster training time, since hard negative mining isn't needed anymore.

One approach introduced by Hariharan *et al*. [5] is to model the data generatively with Gaussian means and covariances, using Linear Discriminant Analysis (LDA). The method can represent all training image data with its statistics, and hence avoids the problem of hard negative mining. This work has recently been applied deformable part models through a generalization of LDA to latent variable models called *latent LDA* [4].

Another approach introduced by Henriques *et al*. [6], which will be our main focus, is based on the observation that neighboring negative windows are almost translated versions of each other, and thus share a lot of structure. This observation is brought further by noticing that the Gram matrix representing an augmented dataset containing translated versions of the samples has a block-circulant structure, and its block-diagonalization leads to a set of subproblems which can be optimized independently for some commonly used classifiers like Support Vector Regression (SVR) or Ridge Regression.

We will describe this method in detail and present some results on how it improves performance and speed over classical hard negative mining on standard pedestrian detection dataset: the INRIA person dataset and the challenging Caltech pedestrian dataset. We will then discuss the benefits of this method, as well as some of the issues it presents and

possible improvements.

2. Related work

We start by briefly describing the method introduced by Hariharan *et al*. [5] to avoid hard negative mining. The method is based on LDA and models the data with Gaussians by computing means for both positive (μ_1) and negative (μ_0) HOG windows, and a single covariance matrix Σ for both classes. Because we consider all negative samples, the number of positive windows is very small compared to that of all negatives, therefore it won't hurt to compute μ_0 and Σ from all available training images even though it includes some positives, with the additional benefit that μ_0 and Σ will describe generic, object-independent backgrounds, and can be used against different object classes. Using invariance properties of image statistics and properties of Gaussian distributions, one can reduce the number of parameters needed and simply compute the mean on a single HOG cell, and the covariance between HOG cells at every offset smaller than the dimensions of the detection window. The resulting LDA linear classifier is given by the weight vector $w = \Sigma^{-1}(\mu_1 - \mu_0)$, which intuitively centers μ_1 with respect to μ_0 and then decorrelates (or "whitens") this difference by multiplying by Σ^{-1}.

Although this decorrelation method succeeds at training a classifier efficiently using all available data, without needing hard negative mining rounds, it models the data generatively relying on the assumption that the HOG features are distributed as a Gaussian, which might not be a faithful representation. In contrast, the circulant decomposition approach manages to train discriminative models such as SVR and ridge regression on the full set of windows, without making assumptions on the distribution of the data.

3. Block-circulant decomposition

3.1. Learning problem and separability

Most of the commonly used classifiers can be represented as the solution of the following regularized empirical risk minimization problem:

$$\min_w \|w\|^2 + C \sum_{i=1}^{N} \ell(y_i, w^\top x_i), \tag{1}$$

where C is a regularization parameter, and ℓ is a loss function, specific to each algorithm. In many cases, this optimization problem has a dual formulation of the form

$$\min_\alpha \frac{1}{2}\alpha^\top G\alpha + \sum_{i=1}^{N} D(\alpha_i, y_i), \tag{2}$$

where α is such that $w = \sum_i \alpha_i x_i$, $G = (x_i^\top x_j)_{ij}$ is the Gram matrix of the examples, and D is a function which depends on the algorithm. The main result of [6] is that this optimization problem can be decomposed in easier subproblems by considering a decomposition of a particular Gram matrix G.

If G is block-circulant[1] with $s \times s$ blocks of size $n \times n$ ($N = ns$), then it can be shown that there exists a unitary matrix U such that $G = U^* \hat{G} U$, where $\hat{G}$ is block-diagonal (in fact, $U = F_s \otimes I_n$, where F_s is the Discrete Fourier Transform (DFT)). Problem 2 becomes:

$$\min_\alpha \frac{1}{2}(U\alpha)^* \hat{G}(U\alpha) + \sum_{i=1}^{N} D(\alpha_i, y_i), \tag{3}$$

where $\hat{\alpha} = U\alpha$. Let's assume that D can be written only in terms of inner products of its arguments. Then $\hat{D}(\alpha, y) = \sum_i D(\alpha_i, y_i)$ can be written in terms of inner products between α and y. Since U is unitary, these inner products are conserved when multiplying by U, the problem becomes:

$$\min_{\hat{\alpha}} \frac{1}{2}\hat{\alpha}^* \hat{G}\hat{\alpha} + \hat{D}(\hat{\alpha}, \hat{y}), \tag{4}$$

with $\hat{y} = Uy$. Since $\hat{G}$ is block-diagonal and the $\hat{D}$ term only consists of inner products, the problem can be easily decomposed in s independent sub-problems, one for each block (corresponding to each Fourier frequency):

$$\min_{\hat{\alpha}_f} \frac{1}{2}\hat{\alpha}_f^* \hat{G}\hat{\alpha}_f + \sum_{i=1}^{n} D(\hat{\alpha}_{fi}, \hat{y}_{fi}), \tag{5}$$

where we have partitioned $\hat{\alpha}$ and $\hat{y}$ in to s blocks $\hat{\alpha}_f$ and $\hat{y}_f$. On can directly solve these problems in the dual formulation by using matrix $\hat{G}$, or try to identify $\hat{G}$ as the Gram matrix of some vectors $\hat{x}_i$, and use these in the primal formulation. In our case, it is easy to identify these vectors, and therefore we use the primal solver.

In the case of **Ridge Regression**, the loss is the square loss and the dual has $D(\alpha_i, y_i) = \frac{1}{2C}\alpha_i^2 - \frac{1}{c}\alpha_i y_i$. D only depends on inner products, so the decomposition is exact. In the case of **Support Vector Regression** (L^2-SVR), $\ell(y, f(x)) = \max(0, |w^\top x - y| - \epsilon)^2$ and $D(\alpha_i, y_i) = \frac{1}{2C}\alpha_i^2 - \alpha_i y_i + \epsilon|\alpha_i|$. The last term in D isn't a dot product, therefore the decomposition isn't exact, and there will be an approximation error, bounded by $\epsilon|\|\hat{\alpha}\|_1 - \|\alpha\|_1|$. This can be applied to other models with the appropriate dual formulation with different approximation errors, including Logistic Regression and L^1-SVR. Regression models are preferable because most solvers for classification algorithms such as SVM and Logistic Regression require labels to be in $\{-1, 1\}$, but the $\hat{y}_i$s won't since they are Fourier transforms. The issue of solving the optimization problems with complex variables can be easily managed (see [6]).

[1] A matrix is called *block-circulant* if every row of blocks is a periodic shift of the row above it

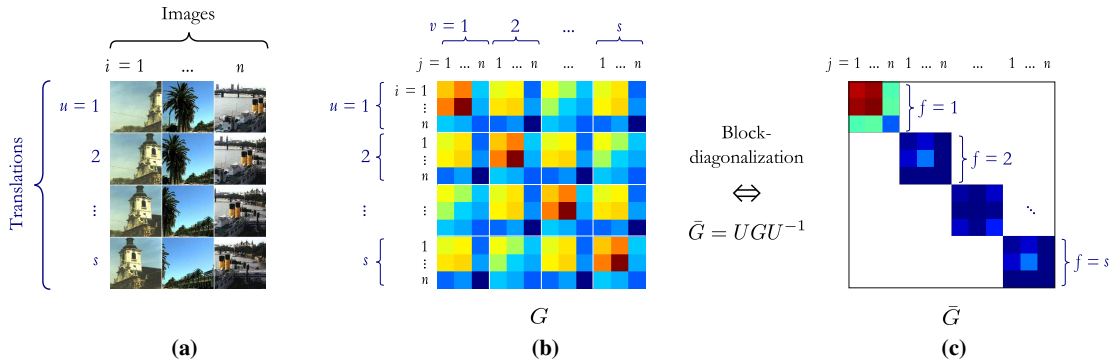


Figure 1: The augmented training set and the block-diagonalization of its Gram matrix. Source: [6]

3.2. Gram matrix on the augmented training set

We would like our system to use all the available image data to train the object detector. To do that, we consider an augmented training set, which includes the translations of each training window. We assume that features are filter-like (e.g. HOG) and we will consider 1D images for clarity; the generalization to 2D is straightforward. For a given image window x of size s, we consider its translations $P^u x$, where P is the cyclic permutation matrix $P = \begin{pmatrix} 0_{s-1}^\top & 1 \\ I_{s-1} & 0_{s-1} \end{pmatrix}$. The goal then is to train with all samples from the augmented training set $\mathcal{X} = \{P^{u-1}x_i | i = 1, \ldots, n; u = 1, \ldots, s\}$ (see Figure 1a). The Gram matrix on this set is of size $ns \times ns$ and its elements are given by

$$\begin{aligned} G_{(u,v),(i,j)} &= (P^{u-1}x_i)^\top P^{v-1}x_j \\ &= x_i^\top P^{v-u} x_j =: g_{v-u}(i,j) \end{aligned}$$

The blocks (u, v) of G only depend on their relative offset, therefore G is a *block-circulant matrix* and can be block-diagonalized as explained in Section 3.1. This is illustrated in Figure 1bc.

The diagonal blocks of $\hat{G}$ are given by $\hat{G}_f = (\hat{g}_f(i,j))_{ij}$, where $\hat{g}(i,j)$ is the DFT of $g(i,j)$. Because $g(i,j)$ represents correlations between x_i and x_j, we have $\hat{g}(i,j) = \hat{x}_i^* \odot \hat{x}_j$ (where $\hat{x}$ is the Fourier transform of x and $\odot$ is the element-wise product). This gives an explicit representation of the vectors giving rise to the Gram matrices of each sub-problem $\hat{G}_f$ ($\hat{x}_{if}$ for $i = 1, \ldots, n$) and allows us to use these vectors directly in a primal solver without having to store the $\hat{G}_f$ matrices.

4. Experiments

We evaluated the block-circulant decomposition method on two standard pedestrian detection datasets: the INRIA Person dataset [1] and a subset of the Caltech Pedestrians dataset [2]. Figure 2 shows precision-recall curves comparing circulant decomposition to standard hard negative mining. On the INRIA dataset, we also compare SVR and Ridge Regression for the circulant case and observe quite similar results, with just a slight improvement using SVR, despite the fact that the decomposition isn't exact. We used the author's code (recently made available on their website), since it gave better results than our own code. The total training time for the INRIA dataset was of just about 2 minutes for the circulant approach (about 1 minute for the FFTs and 1 minute for training), versus 33 minutes for hard negative mining (about 10 minutes for each mining round), on a dual core Macbook Pro with appropriate parallelization.

5. Conclusion

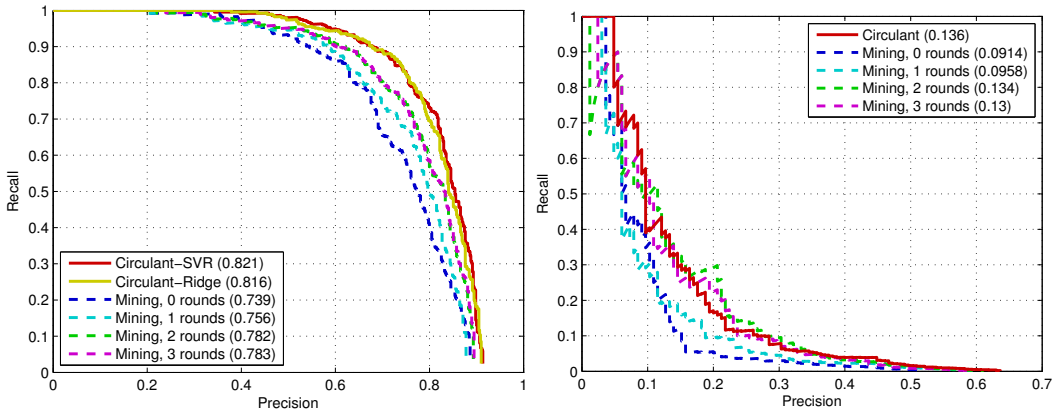
We have seen how a block-circulant decomposition of a Gram matrix can allow us to efficiently train an object detector using all the available training image windows and without the need of expensive hard negative mining rounds. All this can be done with a simple algorithm involving a Fourier transform which can be trivially parallelized, and it gives better test performance while reducing the training time by an order of magnitude. Nevertheless, the limitations on the types of models and features that can be used might be impractical for novel applications, and the technique needs to be generalized beyond single-HOG detectors in order to compete with state-of-the art detectors [3].

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 2012.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.



(a) INRIA person dataset

(b) Subset of Caltech pedestrians dataset

Figure 2: Test performance on standard pedestrian datasets.

[4] R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *ICCV*, 2013.

[5] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.

[6] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *ICCV*, 2013.