Beyond Hard Negative Mining: Efficient Detector Learning via Block-Circulant Decomposition

Alberto Bietti

December 18, 2013

Motivation

For training an object detector, usually:

- Train with all positives and some randomly sampled negatives
- Repeat:
  - Add high-scoring false positives from training images (*hard negatives*) to the training set
  - Retrain classifier

Motivation

For training an object detector, usually:

- Train with all positives and some randomly sampled negatives
- Repeat:
  - Add high-scoring false positives from training images (*hard negatives*) to the training set **Slow!**
  - Retrain classifier

Motivation

Problems:

- Finding hard negatives is *slow*: sliding window, multiple scales, multiple rounds
- A lot of useful data remains unused (and exhaustive search is prohibitively expensive)

Idea

- Windows inside an image are translated versions of each other
- Constraints between these images reduces the degrees of freedom: exploit this to make learning easier

Idea

- Windows inside an image are translated versions of each other
- Constraints between these images reduces the degrees of freedom: exploit this to make learning easier
- **How?** The Gram matrix of the data $G = (x_i^\top x_j)_{ij}$ can be block-diagonalized

Why the Gram matrix

Many classifiers can be learned by solving

$$\min_{w} \|w\|^2 + C \sum_{i}^{n} L(w^\top x_i, y_i)$$

Dual formulation:

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

Why the Gram matrix

Many classifiers can be learned by solving

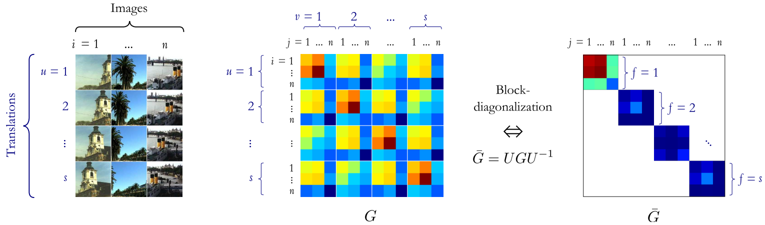
$$\min_w \|w\|^2 + C \sum_i^n L(w^\top x_i, y_i)$$

Dual formulation:

$$\min_\alpha \frac{1}{2}\alpha^\top G\alpha + \sum_i^n D(\alpha_i, y_i)$$

Only depends on the data through G.

Illustration



(a) (b) (c)

Structure of the Gram matrix

- Augmented dataset $\mathcal{X} = \{P^{u-1}x_i | i = 1, \ldots, n; u = 1, \ldots, s\}$
- We have
$$G_{(u,v),(i,j)} = x_i^\top P^{v-u} x_j = g_{v-u}(i,j)$$
- This gives correlations
$$\mathbf{g}(i,j) = x_i * x_j = \mathcal{F}^{-1}(\mathcal{F}^*(x_i) \cdot \mathcal{F}(x_j))$$

New formulation

If $G = U^{-1} diag(G_1, \ldots, G_s) U$,

$$\min_{\alpha} \frac{1}{2} \alpha^\top G \alpha + \sum_i^n D(\alpha_i, y_i)$$

is equivalent to the s independent problems:

$$\min_{\alpha_f} \frac{1}{2} \alpha_f^* G_f \alpha_f + \sum_i^n D(\alpha_{fi}, y_{fi}), \quad f = 1, \ldots, s$$

Algorithm

Inputs:

- X (m features on a $s_1 \times s_2$ grid for n samples, total size $s_1 \times s_2 \times m \times n$)
- Y (labels, size $s_1 \times s_2 \times n$)
- `regression` (a linear regression function)

Output:

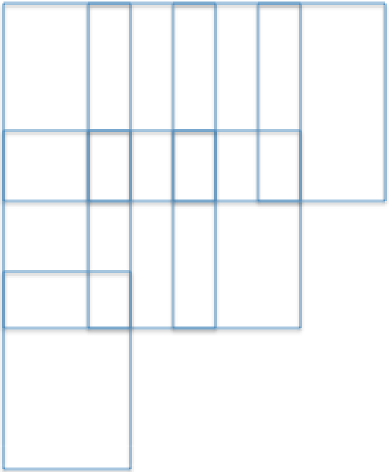
- W (weights, size $s_1 \times s_2 \times m$)

```
X = fft2(X) / sqrt(s1*s2);
Y = fft2(Y) / sqrt(s1*s2);
Y(1,1,:) = 0;
X = permute(X, [4, 3, 1, 2]);
Y = permute(Y, [3, 1, 2]);
for f1 = 1:s1
  for f2 = 1:s2
    W(f1,f2,:) = regression( ...
      X(:,:,f1,f2), Y(:,f1,f2) );
  end
end
W = real(ifft2(W)) * sqrt(s1*s2);
```

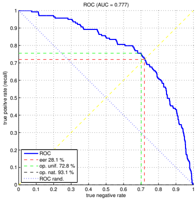
Experiments

Pedestrian detection, INRIA person dataset

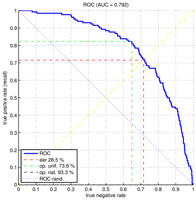
Experiments



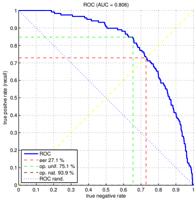
Resuls



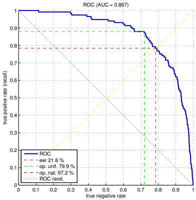
(a) Randomly sampled (0.777)



(b) 1 mining round (0.792)



(c) 2 mining rounds (0.806)



(d) Circulant decomposition (0.867)

Conclusions

Benefits of the method:

- Uses all the available image data
- Only one training phase, no slow hard-negative mining phase
- Easily parallelizable

Conclusions

Benefits of the method:

- Uses all the available image data
- Only one training phase, no slow hard-negative mining phase
- Easily parallelizable

Issues:

- Decomposition only works with some types of classifiers (SVR, Ridge regression), and filter-like features
- The training set can become quite large (small window at small scales)

References

J. F. Henriques, J. Carreira, R. Caseiro, J. Batista
Beyond Hard Negative Mining: Efficient Detector Learning via Block-Circulant Decomposition
Proceedings of ICCV, 2013.