Ensemble of Exemplar-SVMs for Object Detection and Beyond

Tomasz Malisiewicz
September 20, 2011
Computer Vision Reading Group@MIT

Overview

• Motivation and Related Work
• Learning Exemplar-SVMs
• Results
  • PASCAL VOC Object Detection Results
  • Transfer and Prediction
Discriminative Object Detectors

Linear SVM on HOG
Hard-Negative Mining
Sliding Window Detection

DT
Dalal and Triggs 2005
Discriminative Object Detectors

- Linear SVM on HOG
- Hard-Negative Mining
- Sliding Window Detection
- Parts
- Mixtures

DT, LDPM

Dalal and Triggs 2005, Felzenszwalb et al. 2010
Discriminative Object Detectors

Parametric: A fixed number of models per category

Linear SVM on HOG
Hard-Negative Mining
Sliding Window Detection
Parts
Mixtures

DT
LDPM

Dalal and Triggs 2005, Felzenszwalb et al. 2010
Nearest Neighbor Approaches

• Non-parametric: keep all the data around
  • Enables **Label Transfer**

• However
  • No learning implies results depend on features and distance metric
  • Not shown to compete with discriminatively-trained LDPM on Pascal
Per-Exemplar Methods

- NN-method, where each exemplar has its own distance “similarity” function
- Better than using a single similarity measure across all exemplars

Frome et al. 2007, Malisiewicz et al. 2008
Exemplar-SVMs

- Combine
  - Effectiveness of discriminatively-trained object detectors
  - Explicit correspondence of Nearest Neighbor approaches
Exemplar-SVMs

- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)
Exemplar-SVMs

• Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)

• Each Exemplar-SVM is trained with a single positive instance
Exemplar-SVMs

• Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)

• Each Exemplar-SVM is trained with a single positive instance

• Each Exemplar-SVM is more defined by “what it is not” vs. “what it is similar to”
Because each Exemplar-SVM is defined by a single positive instance, we can use different features for each exemplar.
Because each Exemplar-SVM is defined by a single positive instance, we can use different features for each exemplar.

Adapt features to each exemplar’s aspect ratio.

- 7x4 HOG
- 4x8 HOG
Exemplar-SVMs

Exemplar E’s Objective Function:

\[ \Omega_E(w, b) = ||w||^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b) \]

\[ h(x) = \max(1-x,0) \text{ “hinge-loss”} \]
Exemplar-SVMs

Exemplar E’s Objective Function:

$$\Omega_E(w, b) = \|w\|^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b)$$

where

$$h(x) = \max(1-x, 0) \text{ “hinge-loss”}$$

Exemplar represented by $\sim 100$ HOG Cells ($\sim 3,100$ features)
Exemplar-SVMs

Exemplar E’s Objective Function:

\[ \Omega_E(w, b) = ||w||^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b) \]

\[ h(x) = \max(1-x, 0) \text{ “hinge-loss”} \]

- Exemplar represented by \(~100\) HOG Cells (\(~3,100\) features)
- Windows from images not containing any in-class instances (\(~2,000\) images \(\times\) \(~10,000\) windows/image = \(~2M\) negatives)
Large-scale training

- Each exemplar performs its own hard negative mining
- Solve many convex learning problems
- Parallel training on cluster
Exemplar-SVM Calibration

SVM after training
Exemplar-SVM Calibration

1) Apply ExemplarSVM to held-out negative images and all positive images
Exemplar-SVM Calibration

1) Apply ExemplarSVM to held-out negative images and all positive images

2) Fit sigmoid to responses [Platt 1999]

\[
f(x|w_E, \alpha_E, \beta_E) = \frac{1}{1 + e^{-\alpha_E(w_E^T x - \beta_E)}}
\]
Exemplar-SVM Calibration

1) Apply ExemplarSVM to held-out negative images and all positive images

2) Fit sigmoid to responses [Platt 1999]

\[ f(x|w_E, \alpha_E, \beta_E) = \frac{1}{1 + e^{-\alpha_E(w_E^T x - \beta_E)}} \]
Ensemble of Exemplar-SVMs
Ensemble of Exemplar-SVMs

Exemplars

Image + Detections

Learn an exemplar co-occurrence matrix
Qualitative Results

- Let’s take a look at some Exemplar-SVM results in PASCAL VOC dataset
Exemplar  w
Exemplar | $w$ | Averaged Detections

Average of first 10 detections

Average of first 20 detections
Evaluating Exemplar-SVMs

- **Nearest Neighbor**
- No Learning

- **Per-Exemplar Distance Functions**
  - Learning in distance-to-exemplar space [Malisiewicz et al. 2008]

- **Exemplar-SVMs**
Comparison of 3 methods

Exemplar  \( \mathbf{w} \)  Top 6 Detections from Testset

NN

Exemplar-SVM

*Learned Distance Function
Comparison of 3 methods

Exemplar $w$ Top 6 Detections from Testset

NN

* Exemplar-SVM

*Learned Distance Function
Comparison of 3 methods

Exemplar  \( w \)  Top 6 Detections from Testset

NN

* Learned Distance Function

Exemplar-SVM
Comparison of 3 methods

- Exemplar
- w
- Top 6 Detections from Testset

*Learned Distance Function
Quantitative: PASCAL VOC 2007 dataset

- A standard computer vision object detection benchmark
- 20 object categories
- Machine performance is far below human
## PASCAL VOC 2007 Object Category Detection Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>aeroplane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>diningtable</th>
<th>dog</th>
<th>horse</th>
<th>motorbike</th>
<th>person</th>
<th>pottedplant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tvmonitor</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>.006</td>
<td>.094</td>
<td>.000</td>
<td>.005</td>
<td>.000</td>
<td>.006</td>
<td>.010</td>
<td>.092</td>
<td>.001</td>
<td>.092</td>
<td>.001</td>
<td>.004</td>
<td>.096</td>
<td>.094</td>
<td>.005</td>
<td>.018</td>
<td>.009</td>
<td>.008</td>
<td>.006</td>
<td>.144</td>
<td>.039</td>
</tr>
<tr>
<td>NN+Cal</td>
<td>.056</td>
<td>.293</td>
<td>.012</td>
<td>.034</td>
<td>.009</td>
<td>.207</td>
<td>.261</td>
<td>.017</td>
<td>.094</td>
<td>.111</td>
<td>.004</td>
<td>.033</td>
<td>.243</td>
<td>.188</td>
<td>.114</td>
<td>.020</td>
<td>.129</td>
<td>.003</td>
<td>.183</td>
<td>.195</td>
<td>.110</td>
</tr>
<tr>
<td>DFUN+Cal</td>
<td>.162</td>
<td>.364</td>
<td>.008</td>
<td>.096</td>
<td>.097</td>
<td>.316</td>
<td>.366</td>
<td>.092</td>
<td>.098</td>
<td>.107</td>
<td>.002</td>
<td>.093</td>
<td>.234</td>
<td>.223</td>
<td>.109</td>
<td>.037</td>
<td>.117</td>
<td>.016</td>
<td>.271</td>
<td>.293</td>
<td>.155</td>
</tr>
<tr>
<td>E-SVM+Cal</td>
<td>.204</td>
<td>.407</td>
<td>.093</td>
<td>.100</td>
<td>.103</td>
<td>.310</td>
<td>.401</td>
<td>.096</td>
<td>.104</td>
<td>.147</td>
<td>.023</td>
<td>.097</td>
<td>.384</td>
<td>.320</td>
<td>.192</td>
<td>.096</td>
<td>.167</td>
<td>.110</td>
<td>.291</td>
<td>.315</td>
<td>.198</td>
</tr>
<tr>
<td>E-SVM+Co-occ</td>
<td>.208</td>
<td>.480</td>
<td>.077</td>
<td>.143</td>
<td>.131</td>
<td>.397</td>
<td>.411</td>
<td>.052</td>
<td>.116</td>
<td>.186</td>
<td>.111</td>
<td>.031</td>
<td>.447</td>
<td>.394</td>
<td>.169</td>
<td>.112</td>
<td>.226</td>
<td>.170</td>
<td>.369</td>
<td>.300</td>
<td>.227</td>
</tr>
</tbody>
</table>

| CZ [6]      | .262      | .409    | --   | --   | --     | .393| .432| --  | --   | --  | --          | --  | --    | .375      | --     | --          | --    | --   | --    | --       | --  |
| DT [7]      | .127      | .253    | .005 | .015 | .107   | .205| .230| .005| .021  | .128| .014        | .004| .122  | .103      | .101   | .022        | .056  | .050 | .120  | .248     | .097|

Table 1. **PASCAL VOC 2007 object detection results.** We compare our full system (E SVM+Co-occ) to four different exemplar based baselines including NN (Nearest Neighbor), NN+Cal (Nearest Neighbor with calibration), DFUN+Cal (learned distance function with calibration) and E SVM+Cal (Exemplar-SVM with calibration). We also compare our approach against global methods including our implementation of Dalal-Triggs (learning a single global template), LDPM [9] (Latent deformable part model), and Chum et al. [6]'s exemplar-based method. [The NN, NN+Cal and DFUN+Cal results for person category are obtained using 12500 exemplars]
# Object Category Detection

mAP on PASCAL VOC 2007 detection task

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN + Cal</td>
<td>0.110</td>
</tr>
<tr>
<td>DFUN + Cal</td>
<td>0.155</td>
</tr>
<tr>
<td>Exemplar-SVMs + Cal</td>
<td>0.198</td>
</tr>
<tr>
<td>Exemplar-SVMs + Co-occ</td>
<td>0.227</td>
</tr>
<tr>
<td>DT*</td>
<td>0.097</td>
</tr>
<tr>
<td>LDPM**</td>
<td>0.266</td>
</tr>
</tbody>
</table>

*Dalal et al. 2005  **Felzenszwalb et al. 2010
Beyond Object Category Detection

• Based on the idea of label transfer, ExemplarSVMs can be used for tasks which go beyond object category detection
Task 1: Geometry Transfer
Exemplar

Detector w

Appearance

Meta-data

Geometry
Task 1: Evaluation on Buses

- measure pixelwise accuracy on the 3-class geometric-labeling problem: “left,” “front,” “right”-facing

- 43.0% Hoiem et al. 2005
- 51.0% Category-SVM* + NN
- 62.3% Exemplar-SVMs

*Felzenszwalb et al. 2010
Task II: Person Prediction

Exemplar

Detector $w$

Appearance
Task II: Person Prediction

Exemplar

Detector $w$

Appearance

Meta-data

Person
Task II: Person Prediction

Exemplar

Appearance

Meta-data

Detector $w$

Person
Table 2. **Is there a person riding this horse?** We predict from our bicycle, motorbike, and horse detectors whether there is a person riding the object. Our approach is better than the majority vote baseline, suggesting that exemplars are useful at predicting nearby, related objects.
More Transfer Examples
3D Model Transfer

Manually align 3D model from Google 3D Warehouse with a subset of PASCAL VOC “chair” exemplars
Conclusion
Conclusion

• Exemplar SVMs can be used for recognition, label transfer, and complementary object prediction.
Conclusion

- ExemplarSVMs can be used for recognition, label transfer, and complementary object prediction.

- Large-scale negative mining is the key to learning a good ExemplarSVM.
Thank You

Questions?