MIL-UT at ILSVRC2014

Senthil Purushwalkam, Yuichiro Tsuchiya,
Atsushi Kanehira, Asako Kanezaki and *Tatsuya Harada
The University of Tokyo
Pipeline of CLS-LOC task

1-1 Scoring each bounding boxes by RCNN

1-2 Scoring whole image by FV as contextual scores

Input image

Extract region proposals

Extract CNN features

Scoring regions by multiclass PA

Whole image

Extract FV with spacial information

Scoring whole image by multiclass PA

Late fusion

Score

Multiclass Object Detection with hard negative classes

Averaged multiclass Passive Aggressive with hard negative mining

Averaged multiclass Passive Aggressive
Region Proposals and Feature Extraction

• **R-CNN**

• **Region proposals**
  • Selective Search

• **CNN features**
  • Single CNN model (5 conv layers, 2 fully connected layers)
  • Pre-computed ILSVRC13 model
    • No fine-tuning
  • 4096 dim fc7 features
• **Hard negatives classes**
  • **Idea:** Create ‘negative’ classes and train on 2K classes


• Minimize detection errors as well as classification errors

• Passive Aggressive algorithm with hard negative mining
**Multiclass object detection**  
*(training with negative classes)*

We use **Passive Aggressive (PA)** [Crammer et al., 2006]  

to learn multi-class linear classifiers.

\[
W_{t+1} = \arg \min_W \frac{1}{2} \|W - W_t\|^2 + C \zeta \text{ s.t. } l(x_{i(t)}, y_{i(t)}; W) \leq \zeta, \zeta \geq 0
\]

\[
WX_t = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix} x_t = \begin{bmatrix} \text{Score of class 1} \\ \text{Score of class 2} \\ \vdots \\ \text{Score of class } K \end{bmatrix}
\]

\[
\begin{align*}
W_r^{(t+1)} &= W_r^{(t)} + \tau_t x_t \\
W_s^{(t+1)} &= W_s^{(t)} - \tau_t x_t
\end{align*}
\]

where \( \tau_t = \min \left\{ C, \frac{1 - (W_r^{(t)T} x_t - W_s^{(t)T} x_t)}{2\|x_t\|^2} \right\} \)
Multiclass object detection
(training with negative classes)

Core Idea: Hard negative classes

\[ l(x_i(t), y_i(t); W) \]

\( l = \text{Score of class 1}
\]
\( \vdots \)
\( l = \text{Score of negative class 1}
\]
\( l = \text{Score of negative class 2}
\]
\( l = \text{Score of negative class } K \)

\[
\begin{align*}
\mathbf{w}_{r}(t+1) &= \mathbf{w}_r(t) + \tau_t \mathbf{x}_t \\
\mathbf{w}'_{s}(t+1) &= \mathbf{w}'_s(t) - \tau_t \mathbf{x}_t \\
\end{align*}
\]

where \( \tau_t = \min \left\{ C, \frac{1 - (\mathbf{w}_r(t)^T \mathbf{x}_t - \mathbf{w}'_s(t)^T \mathbf{x}_t)}{2\|\mathbf{x}_t\|^2} \right\} \)

Core Idea: Hard negative classes

Cf.) single background class does not work.
Multiclass object detection
(training with negative classes)

Ex.) If a training sample $x_t$ is a positive sample of class 2,

\[
\begin{align*}
\begin{array}{c|c}
 w_1 & x_t \\
 w_2 & \\
 \vdots & \\
 w_K & \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c|c}
 w'_1 & \text{Score of negative class 1} \\
 w'_2 & \text{Score of negative class 2} \\
 \vdots & \\
 w'_K & \text{Score of negative class } K \\
\end{array}
\end{align*}
\]

\[
l(x_i(t), y_i(t); W) = \begin{cases} 
\frac{1}{2} \left( \frac{1 - (w_r(t)^T x_t - w_s(t)^T x_t)^2}{2||x_t||^2} \right) 
\end{cases}
\]

where $\tau_t = \min \left\{ C, \frac{1 - (w_r(t)^T x_t - w_s(t)^T x_t)^2}{2||x_t||^2} \right\}$

ERROR $\leftarrow$ Classification error

\[
\begin{align*}
\begin{array}{c|c}
 w_r(t+1) & = w_r(t) + \tau_t x_t \\
 w_s(t+1) & = w_s(t) - \tau_t x_t \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c|c}
 r & = \text{class 2} \\
 S : \text{Negative class with the highest score} \\
\text{Candidates of } S : \text{class1, 3, ..., or } K, \text{ or negative class 2} \\
\end{array}
\end{align*}
\]

Ex.) If a training sample $x_t$ is a positive sample of class 2,
**Multiclass object detection**

**(training with negative classes)**

Ex.)

If a training sample $x_t$ is a negative sample of class 2,

\[
\begin{align*}
\mathbf{w}_{r}^{(t+1)} &= \mathbf{w}_{r}^{(t)} + \tau_t x_t \\
\mathbf{w}_{s}^{(t+1)} &= \mathbf{w}_{s}^{(t)} - \tau_t x_t
\end{align*}
\]

where

\[
\tau_t = \min \left\{ C, \frac{1 - (\mathbf{w}_r^{(t)} x_t - \mathbf{w}_s^{(t)} x_t)^T}{2\|x_t\|^2} \right\}
\]

\[
l(x_i(t), y_i(t); W)\]

ERROR $\rightarrow$ Detection error

\[
S = \text{class 2}
\]

\[
r = \text{negative class 2}
\]
Features for Contextual Scores

- **Improved Fisher Vector**
  - INRIA's Fisher vector implementation
  - L2 normalization, Power normalization, Spatial pyramid

- **Parameters of IFV for all local features in our system**
  - Dimension reduction of local feature (D): 64 dim
  - # of components in GMM (K): 256
  - 5 scales of local patches
  - Spatial pyramid (P): 1x1 + 2x2 + 3x3 = 8
  - Dimension of IFK: 2PKD=262,144 dim

- **Local Descriptors**
  - SIFT
Classifiers for Contextual Scores

• Classifiers
  – Averaged multi-class Passive Aggressive Algorithm
  – Efficiency of averaging and multi-class setting for large-scale visual recognition
Online Learning for Large-Scale Visual Recognition

• Three guidelines

1. Perceptron can compete against the latest methods.
   • Provided that the second guideline is observed.

2. Averaging is necessary for any algorithm.
   • First-order algorithms w/o averaging cannot compete against second-order algorithms.
   • When averaging is used, the accuracies of all algorithms become very close to each other.

3. Investigate multiclass learning first.
   • Both one-versus-the-rest learning and multiclass learning achieve similar accuracy.
   • However, one-versus-the-rest takes much longer CPU time to converge than multiclass does.

\[ y'_i = \arg \max_{y \in Y \setminus y_i} \mu^y_i \cdot x_i \]

Averaging \[ \bar{\mu} = \frac{1}{T} (\mu_1 + \mu_2 + \cdots + \mu_T) \]

Figure 6. Comparison using ILSVRC 2010 1.2M dataset with SIFT+FV. The darker bar for each algorithm shows the accuracy with averaging. The brighter shows the accuracy without averaging for easy reference.
Late Fusion

1-1 Scoring each bounding boxes by RCNN

Input image

Extract region proposals

Compute CNN features

Multiclass PA for class 1

\( S_{i,1}^{CNN} \)

\( \vdots \)

Multiclass PA for class \( j \)

\( S_{i,j}^{CNN} \)

\( \vdots \)

Multiclass PA for class 1000

\( S_{i,1000}^{CNN} \)

Scoring regions by Multiclass PA for each class

1-2 Scoring whole image by FV as contextual scores

Whole image

Extract FV with spacial information

Multiclass PA for class 1

\( S_{1}^{FV} \)

\( \vdots \)

Multiclass PA for class \( j \)

\( S_{j}^{FV} \)

\( \vdots \)

Multiclass PA for class 1000

\( S_{1000}^{FV} \)

Scoring by linear classifier trained by PA for each class

2. Rescoring with combining RCNN feature and FV

For bounding box \( i \), class \( j \),

\[ S_{i,j}^{new} = S_{i,j}^{CNN} S_{j}^{FV} \]
# Results

## Validation dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Localization error</th>
<th>Classification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN feature + one-vs-all SVMs</td>
<td>0.631743</td>
<td>0.460080</td>
</tr>
<tr>
<td>R-CNN feature + multi-class PA</td>
<td>0.446121</td>
<td>0.285720</td>
</tr>
<tr>
<td>R-CNN feature + multi-class PA using hard negative classes</td>
<td>0.387516</td>
<td>0.227200</td>
</tr>
<tr>
<td>R-CNN feature + multi-class PA using hard negative classes, and FV</td>
<td>0.341743</td>
<td>0.18768</td>
</tr>
</tbody>
</table>

## Test dataset

<table>
<thead>
<tr>
<th>Team name</th>
<th>Localization error</th>
<th>Classification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>0.253231</td>
<td>0.07405</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>0.264414</td>
<td>0.14828</td>
</tr>
<tr>
<td>SYSU_Vision</td>
<td>0.31899</td>
<td>0.14446</td>
</tr>
<tr>
<td>MIL (our team)</td>
<td>0.337414 😊</td>
<td>0.20734 😞</td>
</tr>
</tbody>
</table>
Conclusion

- **Our pipeline**
  - R-CNN based region proposals and features with multi-class object detectors which create hard negative class for each positive class
  - Global features (FVs) with multi-class online-learning
  - Late fusion of region score and global score

- **Combining R-CNN with the contextual information improves the localization performance.**

- **Multi-class object detector trained with ‘hard negative classes’ outperforms one-vs.-the-rest SVMs.**