**Object Detection**

- Generally reduced to Image Classification:
  - This reduction introduces new problems unique to detection
  - Huge imbalance b/w annotated Foreground (fg) objects & Background (bg) examples

**Bootstrapping to the rescue!**

- Simple, yet powerful algorithm:
  - Fix Training Set
  - Update Model
  - Freeze Model
  - Find Hard Detectables

**Mainstay in Object Detection for >2 years**

- Online Neural Networks:
  - [Girshick et al., 2014]
  - [SPPNet, et al., 2014]
  - [BG: Deformable ConvNet, et al., 2016]

**Why don’t state-of-the-art detectors use bootstrapping anymore?**

- Multi-stage pipelines are being replaced by end-to-end systems

**Hard Negative Mining for SVMs**

- Stochastic Gradient Descent (SGD)

**Why is standard bootstrapping not trivial in SGD?**

- Training Object Detector:
  - Fix Training Set
  - Update Model
  - Freeze Model
  - Find Hard Detectables
  - Iterate

**Why use OHEM?**

- Simple and easy to implement
  - Simplifies training: reduces costly to tune hyperparameters
  - Results in better training and higher performance

**Stochastic Gradient Descent (SGD) version:**

Object-agnostic method used in most Region-based Object Detectors, e.g., R-CNN, Girshick et al., 2014

**Online Hard Example Mining (OHEM) + SGD version:**

- Simple, effective, to implement, simplified (and improved) training, consistent and significant improvements

**Why just hard when you can see all?**

- When using all RoIs:
  - Too many easy RoIs (~0.99) dilute the impact of useful (hard) RoIs
  - Need to carefully adjust the LR to account for a larger batch-size
  - ~10% RoIs
  - OHEM operates this heuristic & is much faster to train w/o the need to tweak hyperparameters

**Robust Gradient Estimates**

- Fewer images in a mini-batch leads to correlated RoIs, unstable gradients and/or slower convergence
  - For FRNC: ~1MP for N=1 vs. N=2
  - No impact for OHEM, demonstrates robustness

**Training a ConvNet Detector: (Fast R-CNN)**

- Feature pyramid used in most Region-based Object Detectors, e.g., R-CNN, Girshick, 2015

**OHEM: Main Results (VOC07, VOC12, COCO)**

<table>
<thead>
<tr>
<th>Method</th>
<th>M</th>
<th>B</th>
<th>train set</th>
<th>07 mAP</th>
<th>Method</th>
<th>M</th>
<th>B</th>
<th>train set</th>
<th>12 mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRNC</td>
<td>07</td>
<td>69.9</td>
<td>07</td>
<td>69.8</td>
<td>0.75</td>
<td>FRNC</td>
<td>12</td>
<td>69.7</td>
<td>0.75</td>
</tr>
<tr>
<td>FRCN</td>
<td>07</td>
<td>72.4</td>
<td>07</td>
<td>72.7</td>
<td>0.75</td>
<td>FRCN</td>
<td>12</td>
<td>70.7</td>
<td>0.75</td>
</tr>
<tr>
<td>MR-CNN</td>
<td>07</td>
<td>74.9</td>
<td>07</td>
<td>72.9</td>
<td>0.75</td>
<td>MR-CNN</td>
<td>12</td>
<td>72.9</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours</td>
<td>07</td>
<td>75.1</td>
<td>07</td>
<td>72.4</td>
<td>0.75</td>
<td>Ours</td>
<td>12</td>
<td>71.9</td>
<td>0.75</td>
</tr>
<tr>
<td>MR-CNN</td>
<td>07</td>
<td>71.0</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>MR-CNN</td>
<td>12</td>
<td>71.9</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours</td>
<td>07</td>
<td>74.6</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>Ours</td>
<td>12</td>
<td>73.9</td>
<td>0.75</td>
</tr>
<tr>
<td>MR-CNN</td>
<td>07</td>
<td>74.6</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>MR-CNN</td>
<td>12</td>
<td>73.9</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours</td>
<td>07</td>
<td>74.6</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>Ours</td>
<td>12</td>
<td>73.9</td>
<td>0.75</td>
</tr>
<tr>
<td>RM-CNN</td>
<td>07</td>
<td>74.6</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>RM-CNN</td>
<td>12</td>
<td>73.9</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours</td>
<td>07</td>
<td>74.6</td>
<td>07</td>
<td>71.9</td>
<td>0.75</td>
<td>Ours</td>
<td>12</td>
<td>73.9</td>
<td>0.75</td>
</tr>
</tbody>
</table>

**Understanding OHEM**

- Ablation analysis and more

**Online Hard Mining vs. Heuristics**

| bg lo 0.1 used to approximate hard negative mining
| +1 MAP for VOC07, no impact VOC16
| Sub-optimal: ignores hard RoIs (e.g., paintings)
| OHEM naturally doesn't require this heuristic & automatically selects hard examples

**Bells and Whistles: Ablative Analysis**

- Impact of Multi-scale & Anchor Box Regression
- ROIs clean gradient-based loss mining

**Key Insights:**

- Even though a few images are sampled (N=2), each image has 1000s of RoIs
- Replace Heuristic Sampling with Hard Example Sampling
- How to freeze the model efficiently to find hard examples?
- Forward pass is already freezes the model, exactly for one SGD iteration
- How is it online?
  - Hard Examples sampling is performed inline with online SGD iteration
- Is this efficient?
  - ConvNet forward-backward pass and RoI Network backward-pass remain intact
  - Only addition is RoI Network forward-pass

**Online Training:**

- Study the empirical training loss (mean loss per RoI)
  - during mini-batch SGD

- Footer code available! https://github.com/