LSDA: Large Scale Detection through Adaptation

Judy Hoffman*, Sergio Guadarrama*, Eric Tzeng*, Ronghang Hu*, Jeff Donahue*, Ross Girshick*, Trevor Darrell*, Kate Saenko°

*UC Berkeley, °UMass Lowell
ImageNet

Dataset: Millions of images, >10K classes
State of the art: 200 class detection

bear

bear
What we want

Grizzly bear

Teddy bear
State of the art: 200 class detection
Our Model: 7.5K classes

- taillight: 0.9
- wheel and axle: 1.0
- car: 6.0
Object classification

\{ airplane, bird, motorbike, person, sofa \}
Object detection

\{ airplane, bird, motorbike, person, sofa \}
ImageNet-LSVRC Classification

- 1000 classes
- 1.2 million training images
- Image classification

bus anywhere?

[Deng et al. CVPR’09]
Multi-layer feature learning

“SuperVision” Convolutional Neural Network (CNN)

ImageNet Classification with Deep Convolutional Neural Networks.

ImageNet-LSVRC Detection

Dataset: 400k images, 350k objects, 200 classes

Detect: people, horses, sofas, bicycles, pizza, ...
R-CNN: “Regions with CNN features”

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

[ van de Sande et al. 2011 ]

Alternative approach: “overfeat”
[ Sermanet et al. ICLR’14 ]

“selective search” [ van de Sande et al. 2011 ]

[ Girshick et al. CVPR’14 ]
Can we produce detectors for all of ImageNet?
And use *all* available labeled image data?

Bounding boxes for only 200 classes
Nearly all models only use 1K class images
Adaptation Paradigm

• Weak-label learning commonly considered as a MIL problem
  • requires inference per category.

• Consider a domain adaptation paradigm
  • Is there something common to “detection”?
Transform Classifiers into Detectors

\{I, y^c\} 

\(y^c \in \{A \cup B\}\) 

\(\phi^c(I)\) 

\(\phi^d(I(b))\) 

\(\{I, y^d, b\}\) 

\(y^d \in B\) 

\(W^c\) 

\(W^d\)
LSDA Overview

Input image → Region Proposals → Warped region

LSDA Net (det layers 1-5) → Predictions

- cat: 0.90
- dog: 0.45
- background: 0.25

Produce Predictions

cat? yes → LONSA Net

dog? no
LSDA Overview

Input Image
LSDA Overview

Region Proposals
Select and Warp a Region
LSDA Overview

Input image → Region Proposals → Warped region → LSDA Net

LSDA Net:
- det layers 1-5
- det fc6
- det fc7
- fcA
- fcB
- bkgrnd

LSDA Net:
- \[ \text{cat: 0.90} \]
- \[ \text{dog: 0.45} \]
- \[ \text{background: 0.25} \]

Produce Predictions:
- cat? yes
- dog? no

Compute Region Scores
Training LSDA

Classification Net

Krizhevsky et al. NIPS'12.
Training LSDA

Fine-tune Representation with Detection Data
Training LSDA

Fine-tune Representation with Detection Data
Training LSDA

LSDA Net

det layers 1-5

det fc6

det fc7

Learn a background class detector

dog: 0.45

bkgrnd: 0.25

cat: 0.90

Detection Data

Warped Region
Training LSDA

Detection Data

Warped Region

LSDA Net

det fc6
det fc7

δB

fcA

cat: 0.90

dog: 0.45

background: 0.25

Fine-tune Category Specific Parameters
Training LSDA

Adapt output layer for held-out classes
For category $i$ in set $A$ find $j^{th}$ nearest neighbor in set $B$:

$$\delta A_i = \frac{1}{K} \sum_{j=1}^{K} \delta B_{NB}(A_i, j)$$
ImageNet-LSVRC Detection

Dataset: 400k images, 350k objects, 200 classes

Detect: people, horses, sofas, bicycles, pizza, ...

We hold out bounding boxes from 100 classes to evaluate performance.
Evaluating a detector

Test image (previously unseen)

[Slide credit Ross Girshick]
First detection ...

[Slide credit Ross Girshick]
Second detection ...

[Slide credit Ross Girshick]
Third detection ...

'person' detector predictions

[Slide credit Ross Girshick]
Compare to ground truth

[Slide credit Ross Girshick]
Sort by confidence

0.9

true positive (high overlap)

0.8

false positive (low overlap or duplicate)

0.6

0.5

0.2

0.1

[Slide credit Ross Girshick]
Evaluation metric

Average Precision (AP)
0% is worst
100% is best

mean AP over classes (mAP)

[Slide credit Ross Girshick]
mAP (%) Held-out 100 Categories

Oracle performance on this set is 26.25
Improved Localization

Classification Network

LSDA Network
False Positives

- microphone
- miniskirt
- motorcycle
- mushroom
- nail
- laptop
- lemon
7K class detector!

- Public release of a 7604 category detector trained using this method
- 200 ILSVRC2013 classes trained with bounding box data
- 7404 ImageNet leaf nodes trained with adaptation

http://lsda.berkeleyvision.org/
Detection with 7K Classes

200 trained with bbox

7K trained without bbox
Funny wagon
An ambulance used to transport patients to a mental hospital
Conclusions

• Domain adaptation approach to weak-label detector learning
• “Detection” can transfer across categories
• 7.4K SPP R-CNN Detector runs ~< 1 fps
• Model and code available at lsda.berkeleyvision.org
• Future work: context, hierarchical backoff
Conclusion

- Domain adaptation approach to weak-label detector learning
- 7.4K SPP R-CNN Detector runs \(\leq 1\) fps
- Model and code available at [lsda.berkeleyvision.org](http://lsda.berkeleyvision.org)
- Future work: context, hierarchical backoff